On the Analysis and Design of the Locust Olfactory System

Thesis by

Sina Tootoonian

In Partial Fulfillment of the Requirements for the Degree

of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2013

(Defended November 9, 2012)

Acknowledgements

First I would like thank my parents. I don't know where I'd be without your faith in me, and your years of hard work and sacrifice. I know it's a cliché, but I can never thank you enough. And I would like to thank my sister for always making me laugh, and for keeping me grounded.

I would like to thank my advisor Gilles for his constant support and encouragement over these last five years; for always letting me pursue my ideas; for always having an open door; for his clarity of thought; for teaching me to separate good data from bad; for his patience as I tried out yet another analysis; and for teaching me to see the beauty in small systems without losing sight of the big picture. I can't imagine a better environment in which to have been a graduate student. Oh, and thanks for bringing me to Germany, it's been a blast!

Thanks to my committee for their questions, their ideas, and for keeping me on course when I needed it. Your comments, insights, and feedback directly and significantly improved my work.

Thanks to Kai for his kindness and patience, for the conversations about biology and life, and for his generosity in letting me join him in analyzing such beautiful data.

Thanks to Mala for introducing me to the world of fly audition (who knew they sang!), for her patience with my ideas, and for also teaching me to keep my eyes on the big picture.

Thanks to the lab at Caltech, for making a theorist feel so comfortable among experimenters. And thanks to the Frankfurt lab for making me feel so at home in Germany! Your constant support and your insights have grounded my work in biological reality. It is a privilege to call you my friends.

And finally I'd like to thank Tanja for your support, for putting up with all those working holidays, and for the richness you've brought to my life.

Abstract

The ~ 830 projection neurons (PNs) of the locust antennal lobe respond to odors with dense, odor-specific spatio-temporal activity patterns that are mapped via intrinsic and circuit properties into a sparse representation by the Kenyon cells of the mushroom body, which are in turn read out by the beta-lobe neurons (bLNs). In this thesis we present several analyses of this system. First, we describe metrics for quantifying the geometric properties of PN population responses in the full response space that verify the structures revealed by locally linear embedding. Second, we analyze the mixture responses of single PNs and find that in many cases the mixture response can be explained using one of the component responses. Grouping PNs by their single component preferences reveals a potentially simple substrate for olfactory computations. Third, we look for evidence of cycle-by-cycle decoding of PNs by KCs. We show that much of the variance in single KC responses can be explained using small numbers of PNs, and conversely, that PN odor response trajectories can be reconstructed using KC responses. Finally, in a theoretical/ computational analysis, we assemble some of the basic biological facts about the locust olfactory system into an architecture for the online learning of arbitrary mappings from odors to valences.

Table of Contents

List of Figures

Introduction

Olfactory computations are interesting to study for several reasons. As primarily visual animals we maybe tempted to study the visual system, but olfaction is in some ways an easier computation than vision because the anatomical structure of olfactory circuits discards most if not all of the spatial information, doing away at a stroke with the problems of rotation and translation invariance that visual systems must solve: object recognition is reduced to histogram recognition. Second, olfactory systems are grossly similar across phyla (Eisthen, 2002), suggesting the existence of an optimal method of processing olfactory signals that has repeatedly been found by evolution, and can perhaps be derived from first principles. Third, the similarity across phyla means that the insights we gain in studying olfaction in the 'simpler' circuits of insects are likely to generalize to the circuits in more complex animals.

The work presented in this thesis studies the locust olfactory system. The structure of the locust olfactory circuit follows the basic plan found in other insects and in vertebrates. Odor information is received by 90,000 ORNs (per antenna) whose axons converge onto ~ 1000 spherical neuropilar structure called glomeruli in the antennal lobe (Laurent and Naraghi, 1994). In other species as disparate as mice and flies, ORNs typically express only one type of olfactory receptor (Vosshall and Stocker, 2007), and the axonal convergence onto glomeruli follows the '1 olfactory-receptor to 1 glomerulus' rule. The computational advantage of such a convergence would presumably be an increase in the signal-to-noise

ratio if the olfactory receptors are far enough apart to experience uncorrelated noise. It is currently not known whether this rule also holds in locust.

Glomeruli are sampled by the ~ 830 projection neurons (PNs) and ~ 300 inhibitory local interneurons (LNs) of the antennal lobe. The PNs are spiking neurons and form the sole output of the antennal lobe. Locust LNs are non-spiking, unlike those in bees and flies. The PNs and LNs both sample the glomeruli directly and are also densely interconnected in a recurrent circuit that produces oscillations when excited by olfactory input (Laurent and Davidowitz, 1994). Inhibition is thought to contribute to gain control and decorrelation of PN odor responses (Bhandawat et al., 2007; Mazor and Laurent, 2005), and to increase the precision of PN responses (Bazhenov et al., 2001) to presumably facilitate learning and memorization downstream.

The PNs send their axons to the lateral horn via the mushroom body (Laurent and Naraghi, 1994) where they synapse densely (~ 50%, (Jortner et al., 2007)) onto the 50,000 Kenyon cells, before continuing on to the lateral horn. The mushroom body is a structure that is heavily implicated in learning and memory (Heisenberg, 2003). One of the striking facts about the anatomy of this olfactory circuit is this very large fanout from ~ 1000 PNs onto about ~ 50,000 Kenyon cells, and this is a fact that we will try to explain the model of the olfactory circuit that we propose in Chapter 2 of this thesis.

Kenyon cells interact indirectly via the action of the giant GABA-ergic neuron, which samples the entire KC population and provide global feedback inhibition to the mushroom body (Papadopoulou et al., 2011), presumably to help maintain the sparseness that characterizes KC odor responses (Perez-Orive et al., 2002). This global feedback by the GGN will play a key role in the model circuit we propose.

KC axons leave the MB via the mushroom body peduncle, which branches into the alpha and beta lobes, and, at least in flies, the alpha', beta', gamma and heel lobes (Crittenden et al., 1998). In flies there is evidence mainly from genetic mutants that the different lobes play various roles in the process of memory acquisition and consolidation (Davis, 2005). In the locust only the beta lobe has been studied extensively. The ~ 100 beta lobe neurons (bLNs) are mutually inhibitory and respond densely to odors (Cassenaer and Laurent, 2012), in line with the massive fan-in of inputs they receive from the mushroom body. KC-to-bLN synapses have been shown to be sites of STDP (Cassenaer and Laurent, 2007). STDP appears to help maintain the relative timing of KC and bLN spikes, perhaps to maintain the temporal fidelity of information flow through the system. Additionally, the action of STDP seems to tag synapses for adjustment in response to a reinforcement signal (Cassenaer and Laurent, 2012). The branching of the MB peduncle, the large fan-in from KCs to bLNs, the mutual inhibition between bLNs, and the sensitivity of the KC to bLN synapse to neuromodulators inspired the design of the readout stage of the olfactory system model we propose in Chapter 2.

In addition to the basic anatomy of the locust olfactory system, much has been learned about its physiology, in particular about the responses of PNs and KCs. The first key observation made was that of odor-evoked oscillations, recorded in the mushroom body calyx but sourced ultimately to the recurrent excitatory/inhibitory circuit of the antennal lobe (Laurent and Naraghi, 1994). The ~ 20 Hz frequency of these oscillations was independent of odor and synchronized across the mushroom body, ruling out wave propagation. Thus these oscillations provided a natural global clock signal to which PN responses could be aligned. Doing so revealed that PNs responded reliably during cell- and odor-specific oscillation cycles, giving rise to the concept of odor coding by transient synchrony between odor- and oscillation-cycle specific subsets of PNs (Laurent and Davidowitz, 1994). PN odor responses were found to be dense, with ~ 50% of the population responding during any given time bin, with active subset changing entirely every ~ 300 ms (Mazor and Laurent, 2005). The density of the responses and the relative accessibility of PNs for recording meant that the responses of dozens of cells could be recorded over the course of several experiments. Examining the responses of these cells required the application of dimensionality reduction techniques, the most successful of which was locally linear embedding (LLE) (Roweis and Saul, 2000). When applied to the PN population responses to odors at a range of concentrations, LLE revealed the presence of odor specific manifolds within which responses at increasing concentrations appeared as concentric loops or wings (Stopfer et al., 2003). The appearance of these manifolds suggested that with an appropriate readout both odor identity and concentration could be

readily extracted. However the nonlinear nature of LLE made it difficult to translate the results directly into the desired readout.

Numerous recordings were also made from the KC population one synapse downstream, revealing a dramatic reformatting of responses. These recordings showed the KCs to be an order of magnitude sparser in their responses compared to the PNs (Perez-Orive et al., 2002), but nevertheless capable of reliable odor and concentration specific responses. Examining the precise timing of PN and KC responses revealed a phase preference for each that provided KCs with a brief integration window in which to sum PN responses before the delayed arrival of inhibition from the lateral horn neurons would close the window (Perez-Orive et al., 2002). Coupled with intrinsic voltage gated channels that amplified the effect of the coincident arrival of PN spikes (Perez-Orive et al., 2004), dense PN to KC connectivity and a high KC threshold (Jortner et al., 2007), a picture emerged of KCs as coincidence detectors performing a memory less readout of transient synchrony in the PN population, and mapping dense PN odor representations into sparse ones suitable for learning and memorization.

Summary of Contributions

In this thesis I present two investigations of the locust olfactory system. The first is part of a collaboration with Dr. Kai Shen, a former PhD student in the Laurent Lab. Kai recorded the responses of several hundred PNs and KCs to a range of mixtures. I have helped Kai analyze the data he recorded, and Chapter 1 is a revised manuscript describing the results that we will shortly resubmit for publication to *Neuron*. My contribution to the manuscript has been threefold. First, I have developed a number of simple metrics that characterize the responses of PN populations in the full response space, rather than the dimension reduced LLE space. These metrics allowed us to make quantitative statements about the geometry of odor representations in the antennal lobe, such as the extent to which odor representations clustered by odor rather than by concentration, or the whether the evolution of the representation as one mixture was morphed into another was indeed smooth. The second part of this work is an analysis of the responses of single PNs to mixtures. Specifically, I determined the extent to which mixture response could be explained by responses to the components. I found that often a significant fraction of the response variance could be explained by the response to just one single component response, and that most PNs had a preferred component that explained the majority of their responses. The PNs could be grouped by their single component preference, revealing PNs sensitive to each of the components present in mixture and with a range of concentration sensitivities, potentially providing a simple substrate for olfactory computations. Finally, I looked for evidence of the decoding of PN responses by KCs. I found that ~ 50% of the variance in the responses of single KCs could be explained by using at most $1/6^{th}$ of the PN population

for any given KC response. Conversely, I found a fixed basis in which PN responses could be reconstructed as linear combinations of KC responses, and used it to reconstruct PN trajectories from the KC data. My contributions to the paper are shown in Figures 1-2G to 1-2M, Figure 1-3G to 1-3I, Figure 1-4A to 1-4H, and 1-4L to 1-4N, and Figure 1-6F–K. Because the work presented involved both PN and KC aspects of the data, the entire paper in its current state of revision has been provided to supply the necessary context.

In Chapter 2 I present the results of a purely theoretical/computational study in which I assemble some of the known biological properties of the system into an architecture for learning arbitrary mappings between valences and odors. I begin by approximating the function of the antennal lobe as assigning to each odor a random, dense, binary vector, approximating the PN population representations after decorrelation. I then show that a simple KC-inspired nonlinear projection of these representations into a high dimensional KC-space, coupled with weight normalization (modeling a fixed input receptor pool for each KC), and a *k*-Winners-Take-All (modeling the action of the GGN) can produce an ultra sparse (within the biologically observed range) representation of the input odors that is full-rank, so that arbitrary mappings from odors to valences can be learned. Then I show through simulations that Hebbian learning at PN-to-KC synapse can make the representation robust to noise. I show both through theoretical derivations and numerical simulations how the odor representations can be read out in a near Bayes-optimal way using a pair of mutually inhibitory bLNs whose synapses are gated by the presence of the

dopamine or octopamine reinforcer signal presented with each odor. Finally, I demonstrate the operation of the entire system as it learns online and show that its error rate approaches an asymptote near zero as learning progresses. Thus I show how the known facts about the architecture of the locust olfactory system can be used to build a machine for learning mappings between odors and valences, while also providing a framework for understanding the biology.

Chapter 1: Encoding of Odor Mixtures: Identification, Categorization, and Generalization in an Olfactory System

1.1 Highlights

- We examined the responses of ~ 550 neurons (projection neurons or PNs, and Kenyon cells or KCs) in the olfactory system of locusts to 8 single odors and 32 of their possible mixtures (of 2 to 8 components).

- Responses of single PNs to mixtures could often be at least partially explained by responses to one of the components in mixture. The majority of PNs expressed a preference for one single component over the others. Component preference was distributed among the PNs and was different between the binary mixtures and complex mixtures experiments, suggesting PN adaptation.

- PN encoding space for binary mixtures did not appear to be discretized, and allowed the spread of odor representation to accommodate even fractional changes in input stimuli. For mixtures of n > 2 components, representations by PNs were

ordered by odor similarity, and population response vectors reflected similarity for several seconds following stimulus offset.

- The responses of many KCs signaled the presence of single odor components in mixtures, even when those single odors were part of an 8-component mixture. Individual KCs were significantly better classifiers for single odor components than individual PNs.

- Linear classifiers trained on instantaneous PN and KC population activity patterns performed odor identification, categorization, and generalization with high accuracy.

1.2 Summary

Natural odors are usually mixtures, sometimes composed of hundreds of analytes, and intrinsically variable (Wright and Thomson, 2005). Yet humans and animals can experience them as unitary percepts (Jinks and Laing, 1999): olfaction is a synthetic sense. Olfaction also enables stimulus categorization and generalization and, in some cases, component segmentation (Reinhard et al., 2010). How these complementary and sometimes contradictory computations are carried out remains unknown. We addressed these questions with the responses of 168 locust antennal lobe projection neurons (PNs) to mixtures of two monomolecular odors at varying ratios. We found that the mixture responses of single PNs could in most cases be explained by the response to one of the

components, and that the majority of the PN population could be split based on their preference for one of the two odors. Population responses clustered by concentration and evolved smoothly with concentration ratio as one pure odor was progressively morphed into the other. We next analyzed the responses of another 175 PNs and 209 mushroom body Kenyon cells (KCs) to 8 monomolecular odors and 32 of their possible mixtures (of 2 to 8). We again found that the mixture responses of single PNs could in many cases be explained by their responses to one of the components, and that the majority of the population could be decomposed based on single component preference. We found strong correlations between stimulus composition and PN population response, persisting over seconds after stimulus offset. The responses of individual KCs, much sparser on average than those of PNs, often signaled the presence of single components in odor mixtures. ROC analysis confirmed that KCs are significantly better classifiers for single odor components than PNs. Although the responses of individual KCs were brief, the KC population contained responding cells at every instant of the presynaptic PN population response, consistent with piecewise decoding of the PN output by KCs. As further evidence supporting piecewise decoding of PNs by KCs, we could explain ~ 50% of the KC response variance using small numbers of PNs, and conversely, we could reconstruct PN population trajectories from KC responses. We assessed the information available for such piecewise decoding by training linear classifiers on time-binned PN and KC population responses. We found that the responses of both populations could be read out in single time bins to perform odor identification, categorization, and generalization at high accuracy. Our results suggest that odor representations in the mushroom body result from competing

constraints to keep representations sparse while optimizing for odor memorization, identification and generalization. These rules may be relevant for pattern classifying circuits in general.

1.3 Introduction

The main computational problems of olfaction include discrimination (Abraham, 2004; Linster et al., 2002; Lu and Slotnick, 1998; Rubin and Katz, 1999; Uchida and Mainen, 2003), concentration invariance (Bhagavan and Smith, 1997; Stopfer et al., 2003; Uchida and Mainen, 2007) categorization (grouping of stimuli by shared features), generalization (assignment of novel stimuli to a group, based on shared features), and segmentation (of components from a mixture, or of signal from background) (Mainen, 2006; Wang et al., 1990; Wilson and Mainen, 2006). These object recognition problems (DiCarlo and Cox, 2007) are not specific to olfaction but they are interesting to study there, because olfactory systems solve them in very few neural steps. Using locusts as models, we gained some understanding of the representation formats for simple odors in the first three relays of its olfactory system—the antennal lobe (AL), mushroom body (MB) and beta lobe (bL)—and of the computations carried out by these circuits (Cassenaer and Laurent, 2007; Mazor and Laurent, 2005; Perez-Orive et al., 2002; Stopfer et al., 2003). We also discovered that odors at different concentrations generate families (low-dimensional manifolds) of spatio-temporal representations (Stopfer et al., 2003), providing a neural substrate for concentration invariance. In this study, we turn to odor mixtures. Most natural odors

comprise many components, usually mixed in particular ratios. Mixtures can be perceived as wholes (“coffee”, “grapefruit”) (Jinks and Laing, 1999), but they can also be classified into categories, with various degrees of refinement (“fruity” → “citrusy” → “grapefruit”). Humans can typically identify no more than ~ 3 components, but sometimes as many as 8–12 familiar components in a blend (Jinks and Laing, 1999) and insects and rodents can likely do better (Hurst and Beynon, 2004; Reinhard et al., 2010). These observations are interesting, because the computational constraints on generating a unitary percept and on segmenting a stimulus into its components are contradictory. Also, natural odors such as floral scents can vary from one flower to the next, or from one time of the day to another (Wright and Thomson, 2005). For foraging insects, this necessitates that animals be able to identify individual flowers (to prevent costly repeated visits), and that they generalize (so as to sample flowers of the same variety, species or type) (Reinhard et al., 2010; Wright et al., 2008; Wright and Thomson, 2005). How does the brain solve both discrimination and generalization problems? Our goal was to find out, using the locust system, whether and how the formats of representations for odors might be consistent with these competing requirements. We begin with binary mixtures (Figure 1-1A–D, Methods), and then expand to multi-component mixtures with a set of eight monomolecular odors, paraffin oil (their dilution substrate) and 32 of the 211 possible mixtures of two, three, four, five and eight of those odors (44 stimuli in all, see Figure 1-1E, Methods). We recorded from 343 projection neurons (PNs, the analog of vertebrate mitral cells) and 209 Kenyon cells (KCs, the mushroom body neurons) in 61 animals.

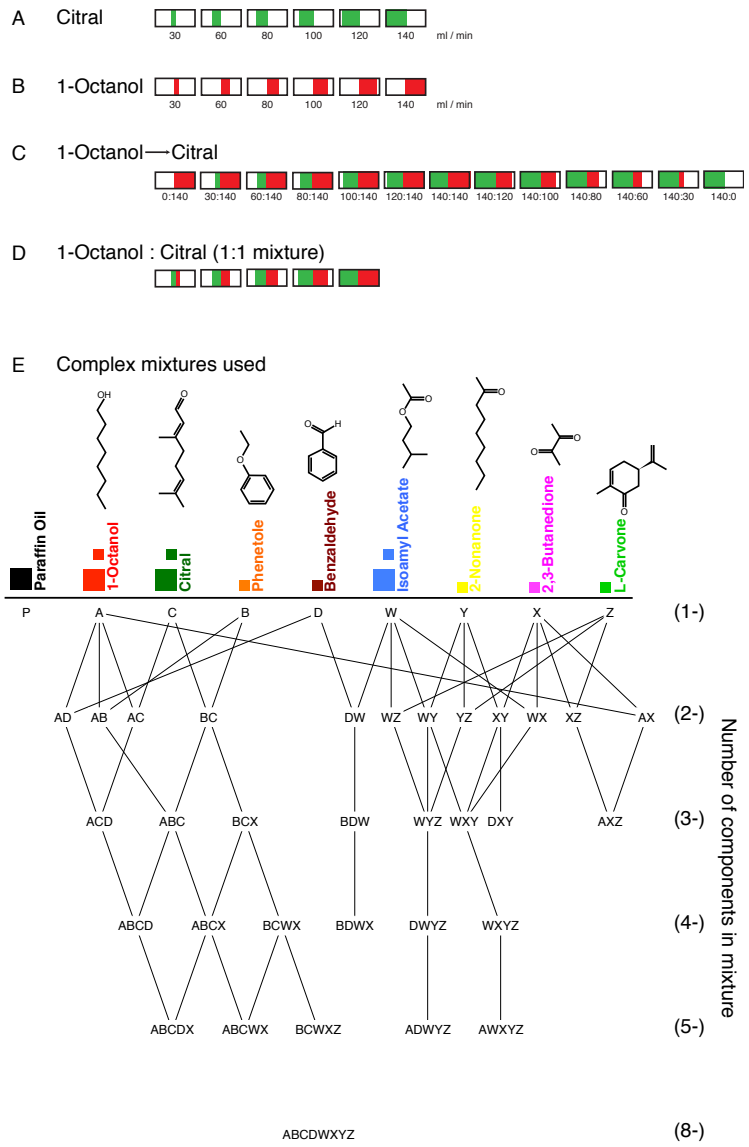


Figure 1-1 Stimulus descriptions.

For the binary mixture experiments, six different concentrations of pure citral and 1-octanol were used individually and in various mixtures. Odor pulse was 300 ms, stimulus repeated for 10 trials, each trial was 14 s. For the complex mixture experiments, 8 different molecules and a subset of their mixtures were used. Odor pulse was 500 ms, stimulus repeated for 7 trials, each trial was 14 s. **(A, B)** Concentrations and schematic representations of the pure odors used for the binary mixture experiments. **(C)** Mixture ratios used when morphing octanol to citral. **(D)** Mixture ratios for the concentration series experiments. **(E)** Complex mixtures: Eight odor components were presented individually and in combination as 2-, 3-, 4-, 5-, 8- mixtures. Paraffin oil was also presented. In addition, three individual components (1-octanol, citral, isoamyl acetate) were also presented at 4x the component concentration, comparable in concentration to 4-mixtures. Right most column (1-, 2- ,3-, 4-, 5-, 8-) indicates number of components.

1.4 Results

1.4.1 Definitions

Our primary data are neural responses to odor mixtures of up to 8 pure components. We define the *segmentation* of an odor mixture by an observer as its decomposition into its pure components. (We will describe a neuron as an *odor segmenter* for pure component *X* if its output in response to a mixture indicates the presence of *X* in the mixture). We define *categorization* as the grouping of a set of mixture stimuli based on the presence or absence of a particular odor component. *Generalization* is the correct categorization of a previously unknown stimulus, based on similarity to known stimuli.

Unless otherwise noted, the *response window* is defined as the one-second period from 0.1 s to 1.1 s following odor onset (the 0.1 s offset is to account mostly for stimulus delays external to the animal), and the *baseline window* is the one-second period from -1.1 s to -0.1 s before odor onset. Neural population responses were binned in time and metrics were computed either *locally* (i.e., separately for each bin) or *globally*, by temporally concatenating the binned response vectors in the time window of interest before computing the metric.

1.4.2 Representations of Binary Mixtures by Single PNs

We first examined the responses of single PNs to binary mixtures of octanol and citral. Figure 2A shows the response of a sample PN to the mixtures tested. The responses are mixture specific, reliable, and temporally patterned, as previously observed for PN odor responses (Perez-Orive et al., 2002; Stopfer et al., 2003).

We examined the extent to which mixture responses (of single PNs) could be explained by component responses. The insets in Figure 1-2A, C illustrate that mixture responses can deviate from the arithmetic sums of the corresponding component responses. We therefore tested whether weighted sums of the form

$$\text{fit}(t) = a\ \text{octanol}(t) + b\ \text{citral}(t) + c,$$

might generally describe mixture responses well. The citral and octanol response functions [octanol(t) and citral(t)] were expressed as mean firing rates across trials in consecutive 50 ms bins, in response to the concentrations present in the mixture. We tested the following models, which differ by the constraints on the fit coefficients (*a* and *b*):

Model	**Constraint**
Constant	$a = b = 0$
Unit Citral	a=1, b=0.
Unit Octanol	a=0, b=1.
Unit Mixture	a=1, b=1.
Scaled Citral	b=0.
Scaled Octanol	a=0.
Scaled Mixture	a=b.
Free Mixture	No constraints.

We allowed the component responses used a temporal jitter of up to 3 time bins. We fit all mixture responses of every PN to each one of these models and used Bayesian model selection to select the best model while simultaneously penalizing complexity (see Methods for details).

The results are shown in Figures 1-2G–M. The top panel in Figure 1-2G shows the result of the fitting procedure when applied to the response of the PN in Figure 1-2A–F to the mixture cit140:oct140. The mixture response is in black and component responses are in red (oct140) and green (cit140). The best fit selected for this mixture response was a scaled and lagged version of the response to citral. This produces an adequate fit ($R^2 = 0.42$). The panel below shows another example of fit for a different PN and mixture. The best model in this case was a scaling of lagged versions of the component responses, producing a good fit ($R^2 = 0.72$).

We summarize the results of this fitting procedure over the population in Figure 1-2H–J. The two columns in Figure 1-2H are PN x mixtures grids. Grid cells are colored according to which coefficients in the fit were non-zero: red for octanol, green for citral, yellow for both, and black for neither (i.e., constant model). For example, a red cell indicates that the response of the corresponding PN to the corresponding mixture was best described using the PN's response to pure octanol at the concentration corresponding to that contained in that mixture. We sorted the cells by their preference for responses of one type or another, and then by their center of gravity within each group. The grid on the left shows all cells, the majority of whose responses were best fit using their responses to pure octanol (octanol-type responses). The second grid shows the cells for which the majority of mixture responses were best fit using the cells' responses to citral (citral-type responses). The smaller grid below shows those cells the majority of whose responses were best fit using both component responses (mixture-type responses), or for which neither component response dominated. Not shown are the 24 cells for which all mixture responses were best fit by a constant model. Figure 1-2I indicates the quality of these fits. The two columns are arranged as in Figure 1-2H, but colored according to the R^2 value of their fits. The two columns in Figure 1-2J are arranged as in Figure 1-2H–I, but color colored to indicate the signal-to-noise ratio (SNR) of the response, defined as the ratio of the mean squared error when using the baseline mean to predict the response, to the baseline variance (see Methods).

Comparing Figures 1-2H and 1-2J suggests that non-constant fits were found whenever the response SNR was sufficiently high. Non-constant models fit approximately 62% of responses in the mixture morph experiments overall, but 89% of those for which response SNR > 3 dB (mean energy of deviations from baseline twice that at baseline). The mean ± SEM of R^2 was 0.35 ± 0.01 overall, but was 0.53 ± 0.01 when response SNR > 3 dB. Thus in most cases where a PN responded reliably to a binary mixture, the response to that mixture could be explained in terms of the component responses in a manner that accounted for more than half of the response variance, on average.

Figure 1-2H shows that the majority of fits were of the citral or octanol type— the mixture response was best fit by scaling the response to one of the components. Most cells exhibited one response-type over mixtures (component dominance), though there are several examples of PNs switching from citral-type to octanol-type responses as the mixture morph progressed. Varying degrees of concentration sensitivity were observed, with some PNs responding at all dilutions of their preferred component, while others doing so only above a PN-specific concentration. Such responses would clearly facilitate concentration-invariant (for the former type) and concentration-sensitive (for the latter type) computations.

Figure 1-2K shows the distribution of best models for the mixture responses of citral-, octanol-, and mixture-type PNs. For all three response-types, the majority of responses were best fit by scaling the inputs. The mean and standard deviations of the scaling factors were (0.54, 0.26), (0.51, 0.17), and (0.54, 0.18), respectively, with less than 5% of values less than zero or greater than 1. Hence most responses were best fit by scaling one of the components or the sum of the two responses by about half. We then asked whether there is an effect of mixture composition on the scaling coefficients. At each dilution, we looked for all citral-type PNs that produced a citral-type response at the dilution in question as well as at cit140:oct30 (the mixture closest to citral), and subtracted the scaling factor of the latter from the former, including unit-type responses in this analysis with a scaling factor of 1. We repeated this procedure for the octanol type responses. In Figure 1-2L, the mean and S.E.M.s of these differences are plotted as a function of dilution. The data have been rearranged to plot the same relative dilution at the same x-value for both citral- and octanol-type responses, so that the citral-type values plotted at 140:60 are for the cit140:60oct, while the octanol type values plotted are for the cit60:oct140 mixture. Stars indicate significant differences from zero ($p < 0.05$, paired t-test), indicating increased suppression as the concentration of the complementary odor is increased. An overall trend with mixture dilution was present in both traces, and could be fit with sinusoids (citral: $R^2 = 0.92$, $p < 10^{-5}$; octanol: $R^2 = 0.76$, $p < 10^{-3}$). This analysis shows an increase in suppression of the component response in binary mixture whenever the complementary (weaker) odor was at an intermediate, lower concentration and a reduction in suppression when the proportions

were reversed. These observations may be single-neuron hallmarks of gain control in the antennal lobe.

As shown in Figure 1-2H, the majority of binary-mixture responses were best fit by scaling the response to one or the other component; true mixture-type responses were relatively rare. This could have resulted from the suppression of the response to one of the components when presented in the mixture, indicating a strong nonlinearity. But this result could have other explanations: for example, if a PN responded to both components but to each with very similar response profiles, only one would have been selected to contribute to the mixture fit, causing the other to be artificially eliminated during model selection. We thus computed the 'SNR-angle' of each fit response. The SNRs of the component responses were computed as for the response SNR, substituting the component responses for the mixture response. The resulting two SNR values (in dB) defined a 2D vector whose angle to the x-axis we define as the SNR angle: an angle of zero indicates a response to octanol and no response to citral; an angle of $\pi/2$ indicated the converse. In Figure 1-2M we plot the histogram of these angles for the different response types. It shows that the majority of octanol-type responses have an SNR angle near zero, while the majority of citral-type responses have an angle of near $\pi/2$. This indicates that many citral- and octanol-type responses were for cell-mixture pairs in which there was a strong response to only one component in the mixture. Hence despite the apparent suppressive nonlinearity implied by a single-component fit, the majority of such responses were in conditions where

the complementary response was weak, making suppression unnecessary to explain the fit. Hence a simplified description of the results above is that ~ 80% of the PNs could be split into two groups based on their affinity for octanol or citral, and that their responses to the binary mixtures of these components were most simply explained by scaling their responses to one of the components.

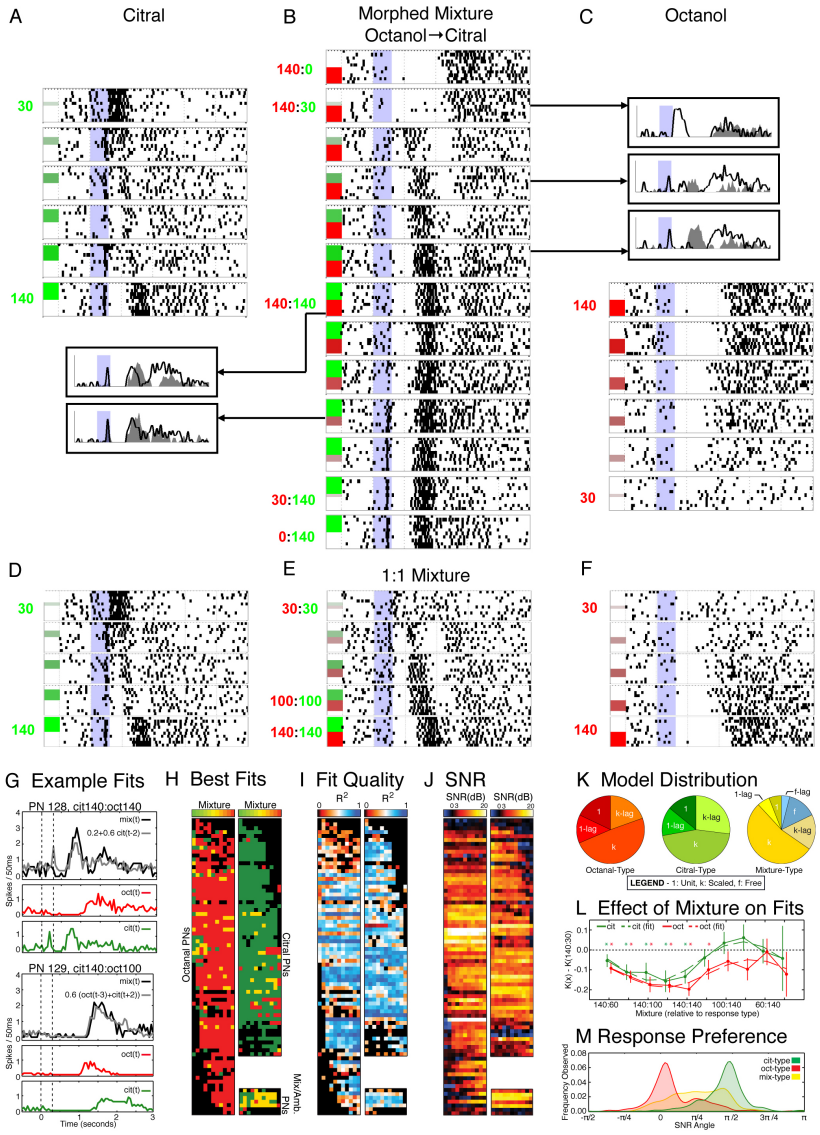


Figure 1-2 PN responses to binary odor mixtures can exhibit nonlinearities.

Responses of one PN to mixture series that morphs odor citral into odor octanol. Odor pulses are 300 ms; 10 trials per condition. **(A)** Responses of one PN to six concentrations of pure citral (30, 60, 80, 100, 120, 140 ml/min). **(B)** Responses of PN to mixture series, from pure octanol (top) to pure citral (bottom). Inserts: overlay of observed (filled PSTHs) and expected (by arithmetic sum of responses to components; black lines) PN responses to morphed mixtures. Mixture conditions are: 140:0, 140:30, 140:60, 140:80, 140:100, 140:120, 140:140, 120:140, 100:140, 80:140, 60:140, 30:140, 0:140 ml/min. **(C)** Response of PN to six concentrations of octanol. Concentrations as in (A). **(D–F)** Responses to 1:1 mixtures of the same odors, at different concentrations. **(D, F)** Same as (A, C). **(E)** Responses to mixture ratios 30:30, 60:60, 80:80, 100:100, 140:140 ml/min. **(G)** Two examples of fitting a PN's mixture response (black) to its component responses at the concentrations present in the mixture (octanal: red, citral: blue). Traces indicate mean firing rate in adjacent 50 ms bins. Dotted lines indicate the odor window. **(H)** PN x Mixture grids indicating the response type of each PN to each mixture. Mixture changes from mostly citral (cit140:oct30) to mostly octanal (cit30:oct140) from left to right, as per the legend at the top. The colors of grid-cells indicate the presence of the component responses in the fit: red if only the octanal response was present, green if only the citral response was present, yellow if both, and black if neither. PNs have been sorted based on which type of response dominated their mixture responses. Octanal-type PNs are in the left

column, citral-type PNs are in the right, and mixture-type PNs are in the bottom grid, along with two PNs in which neither citral or octanal type responses dominated. **(I)** Fit quality. Grids are arranged in exactly the same way as in (H), but colored according to the R^2 value of the fit from red to blue according to the legend at the top. **(J)** Mixture response SNR. Grids are arranged exactly the same way as in (I) and (J), but colored according to SNR(dB) according to the legend at the top. Black indicates 0 dB: mean response energy is equal to mean baseline energy. Dark-red is 3 dB: mean energy of response deviation from baseline mean is twice that at baseline. Comparing this panel to the previous two shows that high SNR typically produced good quality fits, and conversely. **(K)** Distribution of the forms of the models fit for each type of response. 1-: unit model, the fit is equal to the component response (or the sum of the component responses for mixture-type responses) plus a constant offset; k-: scaled model, the fit is equal to a scaled version of the component response (or the sum of the component responses for mixture-type responses); f: free model – the fit was an unconstrained linear combination of the component responses; -lag: At least one of the components used was lagged in time. Note the predominance of the scaled model in all three types of responses. **(L)** Effect of the mixture on the scaling factor of the fits. The change in the scaling coefficient relative to the value at the strongest dilution as a function of the mixture is plotted using data from all citral-type and octanal-type PNs. The data have been rearranged so that the x-axis labels correspond to the same dilution for each type of response, e.g., the data plotted at 140:60 correspond to cit140:oct60 for the citral data, and cit60:oct140 for the octanal data. Error bars are ± 1 S.E.M. Stars indicate significant difference from zero ($p < 0.05$, paired t-test). Dashed lines are sinusoidal fits (citral: $R^2 = 0.92$, $p < 10^{-5}$; octanol: $R^2 = 0.76$, $p < 10^{-3}$). **(M)** Smoothed histogram of SNR angles for mixture responses of each type. Values near 0 indicate that the corresponding response to pure citral was very weak relative to the response to pure octanal, and vise-versa for values near $\pi/2$. Note that that most octanal-type responses cluster around 0, and most citral type responses cluster around $\pi/2$, indicating a lack of response to the complementary component.

1.4.3 Representations of Binary Mixtures by PN Populations

Because odor representations by PNs are highly distributed and varied in time (Figure 2 and (Laurent and Davidowitz, 1994; Mazor and Laurent, 2005; Stopfer et al., 2003)), because their activity patterns are decoded by individual KCs on which converge many PNs (Jortner et al., 2007) and because KCs have very short effective temporal integration windows (Perez-Orive et al., 2004; Perez-Orive et al., 2002), it is useful to examine PN responses as time-series of instantaneous population vectors, or *trajectories*, and visualize them in an appropriately reduced state space (Broome et al., 2006; Brown et al., 2005; Mazor and Laurent, 2005; Stopfer et al., 2003). Figure 1-3A illustrates these PN population trajectories in response to 1-octanol (red), citral (green), and their 1:1 mixture (yellow), after dimension reduction by locally linear embedding (LLE) (Roweis and Saul, 2000). The non-linear nature of LLE makes quantitative comparisons between trajectories difficult. Therefore, we also show correlation measures between the trajectories in the full space. The nine matrices in Figure 1-3B plot the correlation distance (D_c, see Methods) within (matrices along diagonal) and between the responses (168-PN-vector time series over 3 s) to the three stimuli in Figure 1-3A. A correlation distance of zero indicates perfect correlation. Autocorrelations (diagonals) are not zero and cross-correlations are not symmetric around the diagonal because correlations are calculated across different trials with each stimulus. To summarize these matrices, we extracted the minimum value of each one of the nine sub-matrices (e.g., min(cit,cit)) and plotted these values as a matrix of minimum correlation distances (Figure 1-3C). We then combined these minimum correlation distance matrices with each corresponding set of trajectories (Figure 1-3D–F, 1-

4A–C). The combination of correlation measures and LLE trajectories provide both quantitative and qualitative descriptions of the data.

Figure 1-3D shows the same data as Figure 1-3A with the associated min. correlation matrix. As intuition might have predicted, the dimension-reduced mixture trajectory (yellow) lies in between those for the two components, an impression supported by intermediate correlation values (inset). We then computed, for consecutive 100 ms time bins, the projection of the (full-space) population vector for the mixture onto the plane spanned by the vectors for the two components. Then, we computed the angle between this mixture projection and the time-matched vector for citral, as a fraction of the angle between the simultaneous vectors for the two single odors. This yielded the "projection angle fraction" (PAF) with respect to citral, which is by definition 0 for citral, 1 for octanol, and intermediate for mixtures (values less than 0 are not possible; values greater than 1 are possible, though rare). Figure 1-3G shows the time course of the mean (traces) ± S.E.M. (shading) computed over the 10 available trials, for 'mostly citral' (140 ml/min citral:30 ml/min octanol, green), 1:1 mixture (140cit:140oct, gold), and 'mostly octanol' (30cit:140oct), for consecutive 100 ms time bins. During baseline, all three odors have a PAF of ~ 0.5 (as expected from cells firing randomly and independently at the average baseline rate). At odor onset, however, the PAF for 'mostly citral' quickly drops to ~ 0.1 while that of 'mostly octanol' rises to ~ 0.9, indicating high similarity to the representations of pure citral and pure octanol, respectively. Confirming the impression from Figure 1-3D,

the PAF for the 1:1 mixture mostly stays near 0.5, and the low variability indicates that this value is stimulus-driven and not simply due to noise. The PAFs computed with respect to octanol were equal to 1 minus the PAFs to citral, and nearly 75% of the magnitude of the population vector lies in the projection during the early phase of the response (Figure S1-3). Hence we conclude that the trajectory for the 1:1 mixture indeed lies almost exactly in between the trajectories for the two components.

Figure 1-3E represents concentration series for the three stimuli (cit, cit:oct, and oct). Extending previous results (Stopfer et al., 2003), we find that concentration series for 1:1 mixtures, as for single odors, generate families of closely related trajectories (lower-dimensional manifolds), clustered by odor rather than concentration (see also matrix inset). We quantified this impression by computing Rand indices (Rand, 1971) globally on the full-space data, measuring the agreement between clustering by correlation distance and clustering by odor, or between clustering by correlation distance and clustering by concentration (Figure 1-3H, see Methods for details) (range 0–1, higher meaning better agreement). At baseline, both comparisons yielded values close to chance (dashed line). During the response window, clustering was clearly by odor (Figure 1-3H).

In a final experiment, we “morphed” one odor into the other in 11 intermediate concentration steps (Figure 1-3F). Qualitatively, the population trajectory corresponding to

one odor appeared to shift gradually towards that for the other odor, passing through their 1:1 mixture trajectory. We quantified this impression by fitting, in consecutive 100 ms time bins, the correlation distance between each PN vector (full space) for the mixture and that for citral, as a function of the concentration ratio ($\log_{10}$ of the ratio of the concentrations in the mixture), to constant, linear, one-, and two-step functions (Figure 1-3I). We then used Bayesian model selection ((MacKay, 2003), see Methods) to rank the models at each point in time by their fit to the data while simultaneously penalizing them for complexity. In Figure 1-3I the time course of the logarithms of the resulting posterior probabilities for each model relative to that for the linear model are shown (traces are means over trials, shadings are S.E.M.s). At baseline, the constant model is best, indicating no relation between distance and mixture level. Upon odor onset, the linear model quickly dominates and remains superior for most of the response window (see Figure S1-3 for results using fraction-octanol as the independent variable, and further details). The superiority of the linear model over the step models suggests that the encoding space defined by PNs is not discretized (at least within the range and resolution of concentrations tested), and allows the spread of odor representations to accommodate fractional changes in the stimulus.

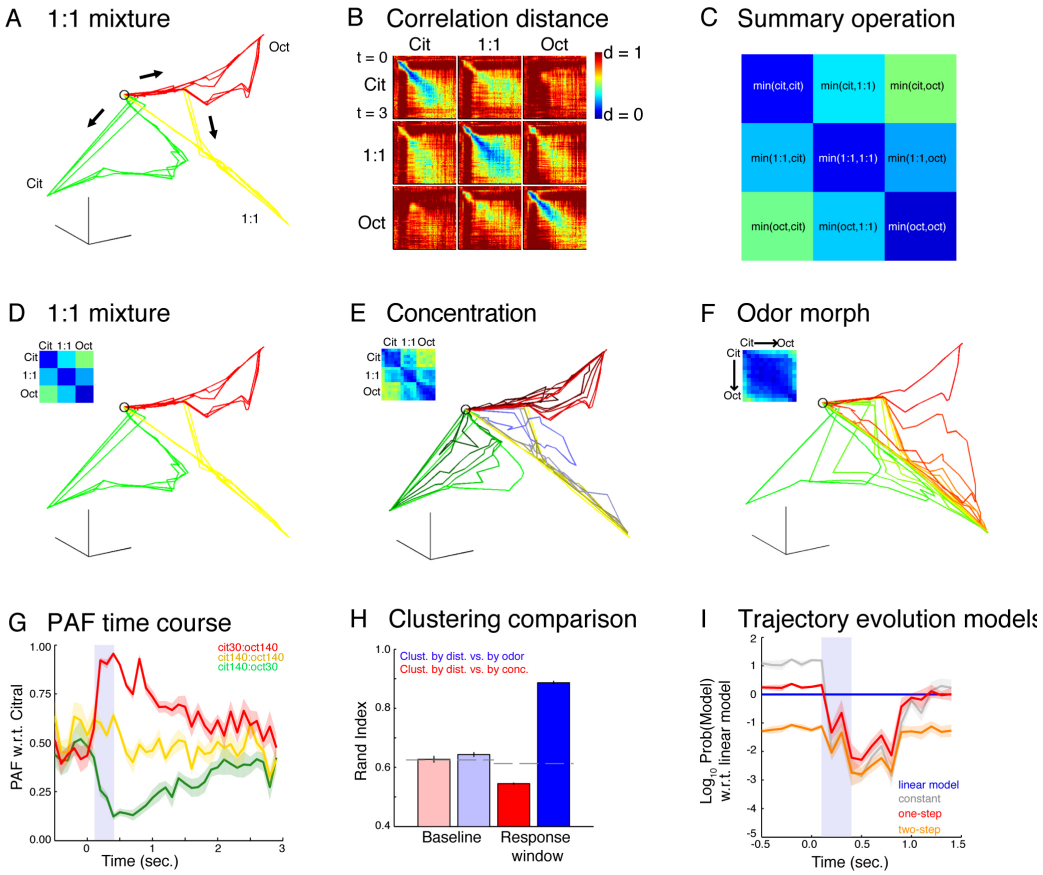


Figure 1-3 PN representations of binary mixtures are structured.

(A) Dimension reduced (LLE) population trajectories in response to pure citral (140 ml/min, green), pure octanol (140 ml/min, red), and the 1:1 mixture (yellow). **(B)** Correlation distances between average PN activity patterns in trials 3–6 and trials 7–10 evoked by pure citral, pure octanol, and the 1:1 mixture in adjacent 50 ms time bins in the 3 seconds following odor onset. The color range has been clipped from [0–2] to [0–1] for clarity. **(C)** The data in (B) is summarized by taking the minimum over each of the 9 odors comparisons. Colors are the same as in (B). **(D)** Same as (A), but now with correlation matrix inset as in (C). **(E)** Concentration series trajectories and correlation summary matrix. Concentrations were 30, 60, 80, 100, and 140 ml/min of citral (dark to light green), octanol (dark to light red), and the 1:1 mixture (purple to yellow). **(F)** Trajectories and correlation summary as pure citral (green) is morphed to pure octanol (red). The concentrations used were 140:0, 140:30, 140:60, 140:80, 140:100, 140:120, 140:140, 120:140, 100:140, 80:140, 60:140, 30:140, 0:140 in units of ml/min citral : ml/min octanol. **(G)** Mean (traces) ± S.E.M. (bands) over trials of the projection angle fraction (PAF) relative to citral in 100 ms time for 'mostly octanol' (red), 1:1 mixture (yellow), 'mostly citral' (green), computed in the full (non-dimension-reduced) PN space. Concentrations were 30:140, 140:140, and 140:30 in units of ml/min citral : ml/min octanol, respectively. The PAF for the 1:1 mixture remains close to 0.5 after onset, indicating that its trajectory lies almost exactly in between that of citral and octanol, confirming the impression from (A, D). **(H)** Mean ± S.E.M. over trials of Rand indices comparing clustering of global (binned in 100 ms bins and temporally concatenated) trajectories by correlation distance with clustering by concentration (red), or with clustering by odor (blue), in a 1 second baseline period (-1.1 s to -0.1 s relative to odor onset) (left), and in the response window (0.1 s to 1.1 s relative to odor onset) (right). Dashed lines are the chance levels for each time window (see Methods). At baseline, the agreement of clustering by distance with both clustering by odor and by concentration is near the chance level, but in the response window, the agreement with

clustering by odor is very high (0.887 ± 0.006), and much higher (one-sided t-test, $p < 0.001$) than that with clustering by concentration (0.545 ± 0.003). **(I)** Mean (traces) ± S.E.M. (bands) over trials of the base-10 logarithm of the posterior probability of the constant (gray), linear (blue), one-step (red), and two-step (orange) models relative to that of the linear model, for the evolution of the correlation distance between the trajectories for non-pure mixtures and pure citral with the base-10 logarithm of the concentration ratio of the mixture, in 100 ms time bins aligned to odor onset. The superiority of the linear model over the constant and the step models after odor onset indicates that PN mixture representations spread smoothly rather than discretely.

1.4.4 Representations of Complex Mixtures by Single PNs

Eight molecules were chosen to be chemically distinct (Figure 1-1E) and their concentrations adjusted to evoke electro-antennograms that were reliable, small and comparable in amplitude (Figure S1-1C, D), to compensate for differences in vapor pressure or receptor activation and to ensure operation far from saturation. As done with the responses of single PN to binary mixtures, we determined the extent to which the mixture responses of single PNs could be explained in terms of their responses to the single components in the mixture. The response of a PN to an n-component mixture was regressed on the constant model (no inputs), and all 2^n - 1 possible combinations of single component responses. For each such input combination, we computed the regression for the unit-, scaled- and free-coefficient models, as well as for lagged versions of these. We then determined the best model using Bayesian model selection (see Methods).

In Figure 1-4A we show an example fit for a response of a sample PN to the mixture ABCWX. The top panel shows the PN's response (black) and the fit produced by the best model (gray). The other panels show that PN's responses to the components in the mixture. The best model in this case was a mixture of octanol (A), citral (C) and isoamyl acetate (W), and fit the data well ($R^2 = 0.68$).

In Figure 1-4B we summarize the results of the model fitting procedure over all PNs and all mixtures. The layout is similar to that used for binary mixtures (Figure 1-2H–J), though extended to 8 components. The first column contains nine PN x mixture grids. The grid-cells are colored according to the component responses used in each fit: black when no combination of component responses was sufficient—the constant model was then chosen; white if n > 1 components were used in the best fit; and the color corresponding to one component if that single component response was chosen. The PNs have been sorted first according to the single component response (if any) that explained the majority of their responses, and then by the center of mass of their responses within each group. The first 8 grids show the PNs for which one of the 8 component responses dominated, yielding 6 A-type, 15 B-type, 10 C-type, 5 D-type, 17 W-type, 44 X-type, 21 Y-type, and 7 Z-type PNs. The columns within each grid have been arranged so that odors containing the component come first (in increasing order of mixture complexity), followed by those that don't contain the component, and in order of increasing mixture complexity. The colored row above each grid indicates the boundary between the two groups (mixtures containing the odor, mixtures not containing it). Note that by definition, no fits containing the component are possible past this boundary. The grid at the bottom shows the 6 PNs for which more than one component responses was required to explain the majority of the mixture responses, and the 10 PNs in which no single component response dominated. The columns have been left unsorted, with the legend-rows above the grid indicating the components present in the mixture corresponding to each. Not shown are the 34 PNs for which the single components did not provide an adequate fit for any mixture. Figure 1-4C is arranged as in Figure 1-4B,

but each grid cell is colored according to the quality of the corresponding fit measured using the R^2 value of the fit according to the color scale shown at the top. Figure 1-4D, again arranged as Figures 1-4A–B, encodes response SNR, computed as for binary mixtures, and colored according to the scale on top.

Overall, best fits involving one or several components were found for 31% of the PN-mixture conditions. Comparing Figures 1-4B and 1-4D suggests that such fits were possible whenever the SNR was sufficiently high (61% of PN-mixture conditions in which SNR > 3dB). The mean ± SEM of R^2 overall was 0.15 ± 0.0034, but 0.30 ± 0.0050 when SNR > 3dB. These values are significantly lower than for the binary mixture experiments (0.35 ± 0.0075, and 0.53 ± 0.0080, respectively, see above). We also compared the results of two sets of separate experiments directly (i.e., based on different PN recordings) at the mixture condition they had in common (cit100:oct100), tested in the binary mixture concentration series experiments and called odor AC in the complex-mixture experiments. In the latter, 21% of PN responses to AC could be fit to a non-constant model (45% of those in which response SNR > 3 dB). In the binary-mixture experiments, 60% of cases of odor AC could be fit overall (85% of those above 3 dB).

To understand the potential source of the discrepancy between the two sets of experiments, we again compared the odor-AC results to those with cit100:oct100 (binary-mixture set).

For each PN, we computed the maximum SNR of its component responses, and the SNR of its mixture response. We then categorized the PN responses into four groups based on their component and mixture responses being above or below 3dB: (1) Silent: Component and mixture SNR < 3 dB; (2) Suppression: Component SNR > 3 dB, mixture SNR < 3dB; (3) Emergent: Component SNR < 3dB, mixture SNR > 3 dB; (4) Full response: Component and mixture responses > 3dB. The percentages of PNs with suppression- or emergent-type responses were similar in the binary mixture and complex mixture experiments: 4.2%, and 15% for suppression- and emergent-type responses in the former case, vs. 8.0% and 19% in the latter. A large difference was found in the percentage of silent cells: 24% for cit100:oct100, vs. 47% for odor AC. This suggests that the main reason for an overall lower fraction of responses that could be fit in the complex mixture conditions is that these mixture conditions engaged fewer cells, possibly due to adaptation. Note that this explanation does not account for the lower fraction of responses above 3 dB that could be fit in the complex mixture conditions, which must be due to an increase in gross nonlinearities in the complex mixture conditions. These results will be examined in the Discussion.

Figure 1-4B showed that, for each of the 8 components, there exist PNs the majority of whose responses can best be fit using the single response to that component. Dilution effects are present, with some PNs yielding fits only at the binary mixture level, others up to the 3-mixture level, etc., with some PNs allowing fits at all dilutions of a component.

Such a distribution of response types would allow both concentration-invariant and concentration-sensitive types of olfactory computation. Note also that some PNs have a secondary preference. For example, there is a C-type PN that has a secondary preference for component W (a "Cw" type response), a W-type PN with a secondary preference for component C (a "Wc" type response). This is particularly interesting because these components are chemically similar (C = citral, W = isoamyl acetate). Other pairings can also be found, such as Xw, Wb, etc.

Figure 1-4E shows the distribution of model fits for each response type. The majority of the responses were scaled (57%) and un-lagged (53%). The mean ± S.E.M. of the scaling weights computed over all such PNs for all fits using the preferred component of each was 0.72 ± 0.0093, similar to what was found with binary mixtures when pooling over the preferred-component fits for citral- and octanol-type PNs (0.74 ± 0.0034).

We next looked for systematic trends in the scaling coefficients of the fits with mixture level in those PNs for which one type of response dominated. We computed, for each of these PNs, the average value of the scaling coefficients used at each mixture level in which the best fit was a unit- or scaled-version of the preferred component response. This yielded up to 5 values for each PN: the mean scaling factors at mixture levels 2–5, and 8 (not exactly 5 whenever a PN didn't have the required type of fit for any of the mixtures at

some mixture level). Of the 125 available PNs, we kept the 99 for which at least 3 of the 5 values were available. The data for these PNs are plotted in Figure 1-4F according to their preferred odor component; no clear trend could be detected. We then computed the Spearman rank correlation of scaling coefficient with the mixture level for these PNs. For 8 of the 99 PNs there was no change in scaling coefficient with mixture level, and to these we assigned a correlation value of 0. The mean and median of the correlation coefficients were -0.15, and -0.20, respectively, suggesting a slight negative trend, though it was not significant (median not significantly different from zero; p-value = 0.093, sign-test). Hence, PNs "preferring" a single component will respond to mixtures containing their preferred component by scaling their component response to ~ 3/4 of its unmixed magnitude, on average.

Finally we examined the extent to which a PN's mixture response best explained by a particular component implied a suppression of the other component responses. For each of the 1545 responses that were fit by single components (of 5600 total), we computed the SNR of the response to the favored single component, and the maximum SNR of the responses to the other components present in the corresponding mixture. We thus positioned each response in 2D space, and computed the SNR angle such that an angle near zero meant that the preferred response was much stronger than all of the others, while an angle near $\pi/2$ meant that at least one of the component responses was much stronger than the preferred response. Figure 1-4G plots the distribution of these angles (blue). Note a

clear peak near 0 degrees, and only a small one near π/2. However, there is a large secondary peak near π/4, suggesting that in many of the responses fit by a single component, at least one secondary component was suppressed, or had a response time course sufficiently similar to the preferred component that its contribution to the fit was minimal. Limiting the analysis to the 917 responses from A- to Z-type PNs that were fit by their preferred-components slightly increased the height of the peak at zero (red curve), reduced the one at π/2, but did not alter that at π/4. The preferred component was dominant (|SNR angle|< π/8) in only about half of the responses (42% overall, 49% for preferred-component responses). These values are lower than those (66%) for the binary-mixture experiments, indicating that component suppression (or redundancy) is greatly increased in the presence of complex mixtures.

The SNR angle cannot distinguish between a secondary component response that is not included in the fit because it appears redundant (e.g., present but similar to and overlapping with the response to the primary component), and one that is actively suppressed. To make this distinction we examined all 184 cases in which $\pi/8 < \text{SNR} < 3\,\pi/8$, i.e., those for which the secondary response was of similar magnitude to the primary. For each of these cases, we recomputed the fit but using only the secondary component response. We reasoned that if the secondary response was indeed redundant, the coefficient of this new fit would be similar to the first one, while if it was being actively suppressed, the coefficient would lower, and even possibly negative. In Figure 1-4H we plot the distribution of the

ratio of the weight for the fit using the secondary component to that using the primary component. Values range from -0.5, indicating strong suppression, to ~ 1, indicating redundancy. We took the threshold between suppression and redundancy to be the dashed line at a ratio of ~ 0.2 because manual inspection of the fits showed that when the ratio was small but positive, the best fit was essentially a constant function. 67% of the fits were below this threshold, indicating an active suppression of the secondary component. The mean ± S.E.M. of the correlation coefficient between the primary and secondary component responses in these cases was -0.10 ± 0.017, and -0.12 ± 0.012 between the secondary component and the mixture response. For the 33% of cases in which the responses were above threshold (i.e., redundant), the mean ± S.E.M. was 0.37 ± 0.031, and 0.36 ± 0.021 between the second component and the mixture response. Thus, when the fit considering a secondary component required only a moderate scaling, that component response was positively correlated with that to the primary component; conversely, when the fit required a suppression or subtraction of the secondary component response, the correlation was weak and/or negative, as should be expected.

In summary, our results suggest that a significant fraction of the PN population can be split into 8 groups based on which single component response explained best the majority of the mixture responses. The average fit was a scaling of the component response by a factor of ~ 3/4, which did not vary much with mixture complexity. PNs varied in their dilution sensitivity, from those responding when only one other component was present in mixture,

to those responding to all mixtures containing their “preferred” component. Responses requiring more than one component existed but were rare; several PNs did however exhibit a secondary odor “preference”. In about half of the single-component fits, the response to the “preferred“ component dominated the others. In about 20% of cases a second component response was also present but was absent in the fit because it was redundant with the single component response (~ 1/3 of the time), or because it was weakly correlated with the mixture response (~ 2/3 of the time). Fewer responses could be fit than in the binary mixtures experiments, partly due to there being fewer strong responses overall, but also due to increased nonlinearity in the mixture responses.

1.4.5 Representations of Complex Mixtures by PN Populations

The LLE trajectories corresponding to the eight component odors are shown in Figure 1-4I. Consistent with the odors' distinct chemical composition, these trajectories did not cluster (see also minimum correlation distances in inset). The observed lack of clustering suggests large differences between the evoked PN response patterns, as desired. Adding components to a single odor, W (Figure 1-4J), caused the mixture trajectories to deviate from that for W. Incremental changes in the population trajectory, however, decreased as the number of components in the mixture increased (see also minimum correlation distance matrix). This is consistent with the fact that the fractional change to the stimulus decreased with each single component addition. This observation was repeated with the other odors and quantified by analysis in full PN space (not shown).

While mixture representations deviated from those of their components, they still formed clusters of trajectories, well segregated from those corresponding to non-overlapping mixtures. In Figure 1-4K, sets of all single- and mixed-odor trajectories for odors containing only components W, X, Y, and Z and those containing only components A, B, C, and D are plotted, revealing two non-overlapping manifolds, seen as two low-distance clusters in inset (A and W).

These results suggest that the representations of complex mixtures by PNs are ordered: the more similar odors are, the more similar their corresponding representations by PNs. To test this hypothesis, we computed the dissimilarity between odors (represented as 8D binary vectors whose coordinates indicate the presence or absence of each of the 8 components) using the Jaccard distance (Deza and Deza, 2009). We computed all pairwise distances D_j between odors, and all correlation distances D_c between the PN population vectors corresponding to those odors (calculated globally over the entire response). In Figure 1-4L, we plot the Spearman rank correlation between D_j and D_c calculated over the response (blue), over baseline (red), and a control (gray). During the response window there is a strong tendency for trajectory distances to increase with odor distances, while during baseline this trend is very weak (see also Figure S1-4). We conclude that the evolving odor representations by the PN population continuously contain information about stimulus composition in their global inter-trajectory distances. When observed bin-by-bin, distances between trajectories can vary: Figure 1-4M, for example, shows the evolution of D_c calculated for odors WYZ and DWYZ over time. But information about odor composition is preserved also in the inter-trajectory distances measured instantaneously. Figure 1-4N shows the time course of the Spearman rank correlation (calculated as in Figure 1-4L) measured piecewise in time. Quickly following odor onset, the correlation reaches a high value of ~ 0.7 (blue), indicating a strong tendency for trajectory distances to increase with increasing odor distance. The correlation remains high and above baseline value for several seconds after odor offset (see also Figure S1-4). We conclude that the responses to mixtures of the PN population continues to contain information about mixture

composition in their pairwise distances for several seconds following odor offset. PN trajectories do not spread randomly in representation space, a result consistent with those obtained with binary mixtures.

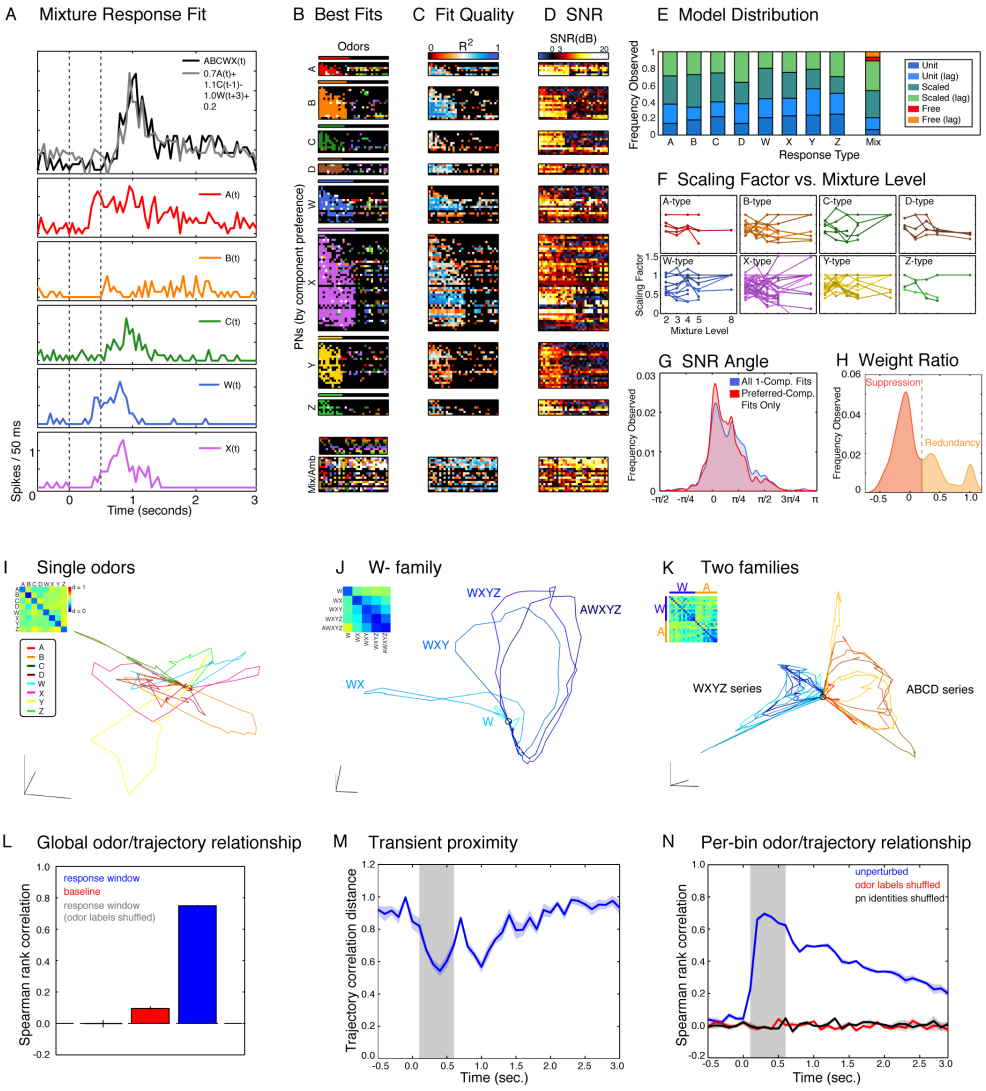


Figure 1-4 PN representations of complex mixtures reflect odor similarity.

(A) Example mixture response fit. Response (mean firing rate in adjacent 50 ms bins) of a cell to the 5-component mixture ABCWX (black), and the best fit (gray). Dashed lines indicate the odor window. **(B)** Component-type of best fits describing PN mixture responses. Each row is one PN, and each column is a single mixture. Cells are colored according to the single component that was used in the best fit, black if no fit was found, or white if more than one single component was used. PNs are grouped based on which component, if any, dominates the fits of their responses. PNs within each group are sorted according to the center- of-mass of their in-group responses. Odors (columns) for each group are arranged so that mixtures containing the group's component come first, followed by mixtures not containing the group's component, both in increasing order of mixture complexity. The legend above each grid indicates the boundary between the two groups. The grid and legend at the bottom are for those PNs in which the majority of fits required more than one component, or in which no single component dominated. The rows and columns have been left unsorted, and the legend above indicates the components present in the mixture represented by each column. **(C)** Fit quality. The grids are arranged identically to B, but colored according to the R^2 value of each fit, according to the legend at the top. **(D)** Response SNR. The grids are arranged identically to B, but colored according to the SNR of the mixture response of a PN measured in dB, according to the legend at the top. SNR is defined as the ratio of the average energy of deviations from the baseline mean during the response, to the baseline variance. **(E)** Distribution of model types according to components used. Each

stacked bar plot shows the relative frequency of unit- and scaled-coefficient models used in the corresponding single component fits, and also the distribution of free-coefficient models used in fits where more than one component was used. **(F)** Scaling factor in fits as a function of mixture complexity. Each panel plots the variation of the mean scaling factor used in the corresponding single component fits as the mixture level is increased. Only PNs are used which had at least one preferred-component response in at least three mixture levels. No obvious trend with mixture level is present in any of the panels. **(G)** Smoothed histograms of SNR angle computed over all single-component fits (blue), or limited to PNs with a preferred component, for fits using that component (red). Angles near 0 indicate that the single component response for the component used in the fit was dominant; those near $\pi/2$ indicate that a secondary component response was dominant; those near $\pi/4$ indicate that the primary component response was approximately matched in SNR by at least one of the secondary responses. **(H)** Smoothed histogram of the ratio of the scaling weight for the best single component fit using the secondary component, to that when using the primary component, for the fraction of data points whose SNR angle (red curve in panel G) was near $\pi/4$. Values below 0.2 were likely to be actively suppressed in the mixture response; those above were likely removed in the fit due to redundancy with the primary component response. **(I–N)** Multi-component-mixture LLE trajectories (dataset different from that in Figure 3: 175 other PNs, stimulated with 44 different odor conditions, see Methods). Four and a half seconds are represented, beginning at odor onset; odor pulses: 500 ms long; 50 ms bins, each averaged over three 2-trial averages. (**I)** Trajectories in LLE space for the 8 single-odor components. No obvious clustering is present either in LLE space or in the correlation distance matrix inset. **(J)** Starting from single odor W, trajectories increasingly deviate as components are added (W→WX→WXY→WXYZ→AWXYZ). **(K)** Mixtures form ordered trajectory clusters: family of {W,X,Y,Z} (W, X, Y, Z, WX, WY, WZ, XY, XZ, YZ, WXY, WYZ, WXYZ) well separated from family of {A,B,C,D} (A, B, C, D, AB, AC, AD, BC, ABC, ACD, ABCD). **(L)** Mean ± S.E.M. over trials of Spearman rank correlation (ρ_{sp}) between Jaccard distance between odors represented as binary vectors, and the correlation distance between global (temporally concatenated in the response window) activity patterns (blue), baseline (red), and the response window but with the odor labels on trajectories shuffled (gray, near zero). During the response window trajectory distances increase with odor distance ($\rho_{sp} = 0.749 \pm 0.004$), while at baseline they do so to a much lesser extent ($\rho_{sp} = 0.096 \pm 0.017$), and not at all if odor labels during the stimulus presentation are shuffled ($\rho_{sp} = 0.003 \pm 0.028$). **(M)** Mean (trace) ± S.E.M. (band) over trials of the correlation distance between activity patterns evoked by WYZ and DWYZ in adjacent 100 ms time windows, starting at t = -0.5 sec. The distance between the two odors is multiphasic in time. **(N)** Mean (trace) ± S.E.M. (band) over trials of the Spearman rank correlation between the Jaccard distance between odors represented as binary vectors, and the corresponding activity patterns in adjacent 100 ms time windows starting at t = -0.5 sec (blue), the activity patterns with odor labels shuffled randomly for each time bin (red), and activity patterns with PN identities shuffled for each bin and each odor (but fixed across trials for a given bin and odor, black). The relationship between odor distances and trajectory distances is strong shortly after odor onset and remains above baseline for several seconds past odor offset, but is absent in baseline and for both shufflings of the data.

1.4.6 Kenyon Cell Responses to Mixtures

Because Kenyon cells are the direct targets of PNs in the mushroom bodies, because mushroom bodies are a site for associative memory (Heisenberg, 2003; Masse et al., 2009) and because KC output synapses are plastic (Cassenaer and Laurent, 2007, 2012) KCs are a likely repository of olfactory memories. It is therefore important to determine the stimulus features that they extract from PNs. For comparison, we show first the responses of one representative PN to our 44 stimuli (Figure 1-5A). As is typical of PNs (Perez-Orive et al., 2002), this neuron responded to about half of the stimuli with a variety of discharge patterns. KCs, by contrast, responded very rarely to odors (Figure S1-5) but when they did, they did so with very high specificity (KCs1–3, Figures 1-5B, C). Surprisingly, KCs that responded to a component also often responded to many—sometimes all—of the mixtures containing it. KC 1 (Figure 1-5B), for example, fired in response to odor D, and responded to all mixtures containing D (though not necessarily at the same time following onset or for the same durations). The same can be seen with KCs 2 and 6 for odor W (Figure 1-5C, D). KCs 5 and 6 were recorded simultaneously, and each responded to a different molecule (C, W). When both odors were included in the mixture (e.g., BCWX, ABCWX, BCWXZ), both KCs responded, with a few exceptions (ABCDWXYZ). We found KCs specific to all 8 single odors. By chance (see Methods) we also found a few KCs specific for binary mixtures but not their components (not shown).

To quantify the response specificity of the cells in the PN and KC populations, we next analyzed the difference between PN and KC responses using receiver-operator-characteristic (ROC) analysis (Fawcett, 2006), measuring a neuron's ability to separate stimuli into sets "containing *i*" and "not containing *i*", as response threshold is varied (see Methods). On a true-positive (TP) vs. false-positive (FP) plot, selective neurons are identified by ROC curves that tend towards the corners, while unselective ones run along the diagonal (Figure 1-5E, upper panels). The area under the curve (AUC) thus measures selectivity (near 1 or 0 for high, near 0.5 for low) (Figure 1-5E, lower panels). This analysis indicated that individual KCs are significantly better ($p < 10^{-7}$, Wilcoxon rank sum test) than individual PNs at component segmentation. Hence, in addition to being highly selective and thus, rare responders, single KCs can categorize odors (e.g., as containing *X*) by extracting component information from PN population vectors.

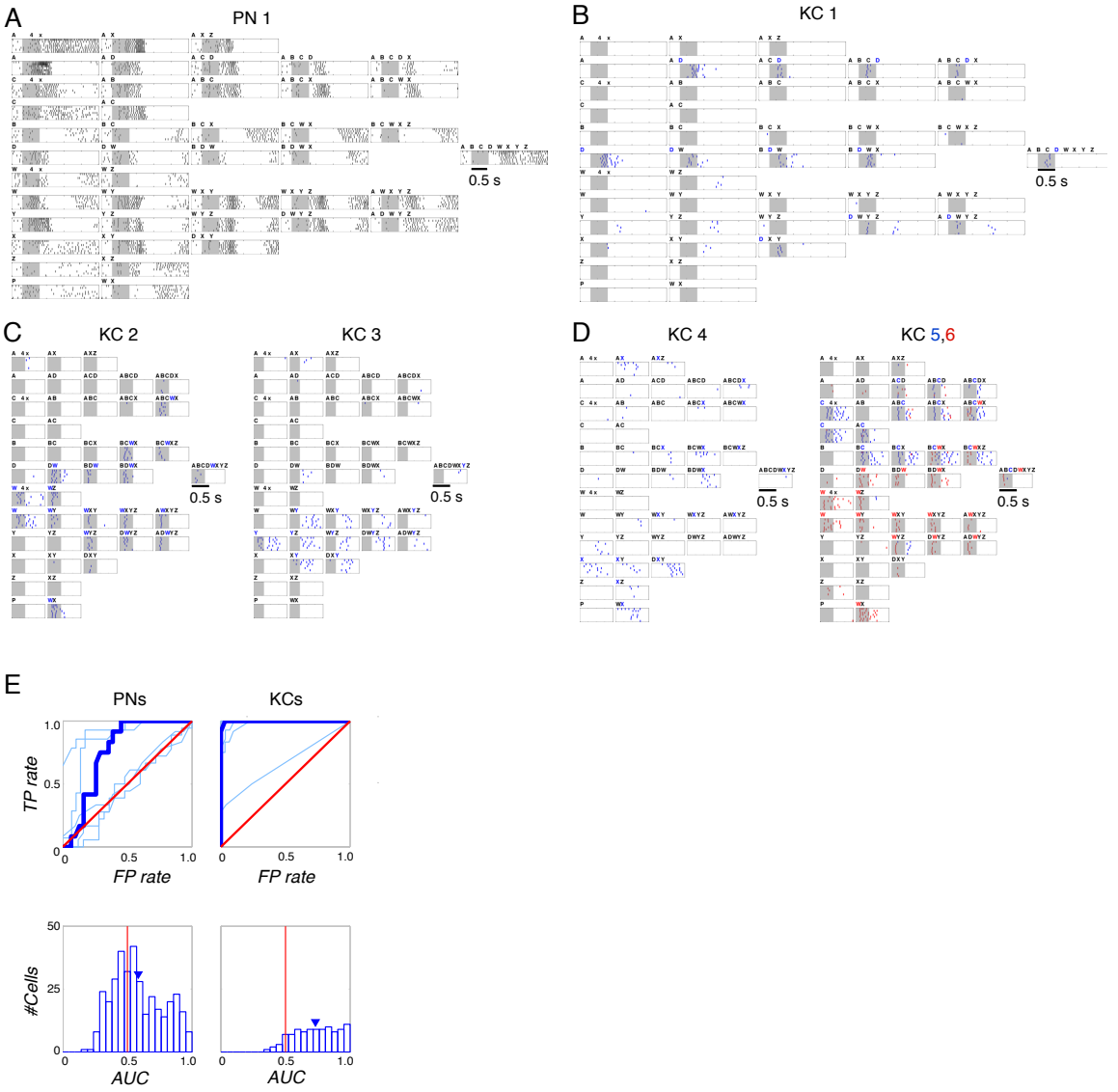


Figure 1-5 Single KCs segment components out of odor mixtures better than single PNs.

(A) Spike rasters of a representative PN to single and mixed odors (see Methods for 44 stimuli, 7 trials, 500 ms stimulus at shaded area, 2.5 s shown). Numbers of components organized by column, conditions arranged to make overlapping mixtures adjacent wherever possible. **(B–D)** Spike rasters of six representative KCs. **(B)** D-segmenting KC, with weak late response to unrelated mixtures Y, WZ, YZ. Same scale as in (A). **(C, D)** Other representative KCs, showing segmentation of different odors (in order of KCs: W, Y, X, C and W). KCs 5 and 6 recorded simultaneously; both responded to mixtures containing both C and W (e.g., BCWX), but at different times. Only 1 s shown, centered on KC response times; Time scale as in (B). **(E)** Top panel: ROC evaluation of component selectivity by PNs and KCs. True and false positive rates (TP, FP) determined by sliding response threshold; response based on spike counts within 1 s window summed over 7 trials. Red diagonal: chance performance. Blue lines: results for several example PNs (the curve for the PN 1 is highlighted), and for the KCs in (B–D) (the curve for KC 1 is highlighted; see Methods for class partitions). Bottom panel: Distribution of area-under-curve (AUC) values for KC-odor class pairs significantly shifted to the right of PN-odor class pairs ($p < 10^{-7}$, Wilcoxon rank sum test). Arrows indicate means: 0.74 (KCs; SD = 0.17); 0.59 (PNs; SD = 0.20).

1.4.7 Decoding PN Trajectories Over Time

Individual KCs are odor classifiers, that each make binary decisions based on the response state of a subset of all possible PN inputs. What is the collective decoding capability of KCs as a population? We examined the sequential organization of the KC population response. Figure 1-6A shows three W-responding KCs. As described before, each responded at a particular time during the stimulus with just a few action potentials on a baseline of 0. These three KCs also responded to mixtures. Their mixture responses, however, were not identical to their component responses (Figure 1-6A). They could vary in intensity, duration and in timing. Figure 1-6B sorts and plots the PSTHs of KCs that responded to W (grayscale, averaged over 7 trials and normalized) and the peaks of those PSTHs (red). As components were added, the number of mixture-responding KCs increased, in part due to new KCs that detected the other components of the mixture (black dots). The number of responding KCs, while always low (0.5–1% of all recorded KCs in any 50 ms bin), thus increased with mixture size (Figure S1-6A). This was correlated not with increased total PN activity—which varies little with concentration (Stopfer et al., 2003) or mixture complexity (Figure S1-6A), but rather with increased PN synchronization (Figure S1-6B, C). KC responses were distributed during and after the stimulus, with a peak within 200–300 ms of stimulus onset on average (Figure 1-6C). The response times of these KCs have been superimposed on the corresponding PN trajectories for these three stimuli (Figure 1-6D–E), demonstrating that KCs are decoding PN trajectories throughout their entire duration.

If single KCs decode PN trajectories, then, given our large dataset of PNs, we should be able to use PN responses to reconstruct those of the KCs. To test this possibility, we formed odor response vectors for each PN and KC by computing its average spike count binned in 100 ms bins in the 1 second window following odor onset, and concatenating across the 44 odors tested. This yielded, for each PN and each KC, a 440-element (10 bins x 44 odors) response vector. We then used the multi-stage adaptive lasso (Buhlmann, 2011) to linearly regress the response vector of each KC on those of the PNs, while constraining the regression weights to be positive, to reflect the excitatory nature of connections from PNs to KCs (Jortner et al., 2007). This is shown schematically in Figure 1-6F.

Figure 1-6G shows three of the best reconstructions: in all three, more than half of KC-response variance could be explained using 30 or fewer of the 175 available PNs. Figure 1-6H summarizes the distribution of reconstruction results over the KC population. The median value of the fraction of variance un-explained (SSE/SST) is 0.52. Twenty six percent of the KC responses could not be regressed at all on the PN responses. The mean ± S.E.M. of the number of connections used in the remaining reconstructions was 10±0.6, with a maximum value of 30.

We then tested the extent to which the PN and KC datasets were consistent with each other by computing the quality of reconstructions if one or the other set had been shuffled (see

Methods for details). If shuffled PN responses were used (with the same number of regressors as in control), the median of the SSE/SST distribution was 0.82, indicating a significant loss in performance. Conversely, we used the PN responses to reconstruct the shuffled responses of KCs, yielding a median value of SSE/SST of 0.86, again, indicating a loss in performance. In short, our results show that the responses of single KCs can be reconstructed from pooled subsets of PNs, consistent with the cycle-by-cycle decoding of the PN population by KCs.

Finally we examined whether the PN trajectories could be reconstructed from the fragments decoded by single KCs. In other words, we looked for a fixed basis set such that the PN population activity in each time bin could be expressed as a linear combination of basis elements with the KC responses as coefficients (Figure 1-6I). We found such a basis by minimizing the sum of the squared error between the recorded PN responses and the reconstructions. Due to the assumption of linearity, this procedure was equivalent to minimizing the error in reconstructing the responses of single PNs from those of the KCs (see Methods); we simply applied the procedure applied for reconstructing KC responses from PNs, in reverse. Figure 1-6J represents the PN odor trajectories corresponding to the odor ABC reconstructed directly from the PN data (blue) or indirectly using the KC data (red), illustrating a good qualitative fit.in a principle-components-space computed using the data for all odors. Figure 1-6K summarizes the fits over all stimuli and reconstructions. For each time bin along each trajectory, we computed the fraction of the response variance

unexplained (SSE/SST), and then averaged this value over the trajectory. The distribution of mean values over the 44 odors is plotted in in red. The median value was 0.44, the value for the example shown in Figure 1-6J. Similarly, we computed the mean over time bins of the correlation between the trajectories and their reconstructions (red, right panel of Figure 1-6K). The median was 0.76, the same as that for the example in Figure 1-6J.

We next measured the extent to which the recorded KC population was informative about the PNs. We generated sets of shuffled KC responses and for each of 100 shuffles, computed reconstructions of PN trajectories as done with the un-shuffled responses. Three of the reconstructions for odor ABC are plotted in Figure 1-6J (green), illustrating inferior fits, as confirmed in Figure 1-6K. The median value of SSE/SST for the fits using shuffled KC responses is 0.67, significantly higher than that obtained using the unshuffled data ($p = 0$, rank-sum test), although there is some overlap between the distributions. Finally, we tested the extent to which the KCs were well suited for reconstructing the trajectories of recorded PNs. We produced 50 fake PN populations by shuffling the odor responses of single PNs, and reconstructed the trajectories as above using the unshuffled KCs. The distribution of the resulting fit metrics are plotted in gray in Figure 1-6I. The median of SSE/SST over all shuffles and all odors is 0.83, again significantly higher ($p = 0$, rank-sum test) than for the unshuffled data.

In summary, we conclude that all fragments of any PN-population trajectory are decoded by at least some KCs, and conversely, that PN trajectories can be reconstructed from the recorded KC population responses. Thus PN trajectories (formed of dense PN vectors) are mapped onto new trajectories (formed of sparse KC vectors) in KC space, so that each odor (whether it is a mixture or not) is represented by a time series of sparse KC activity vectors.

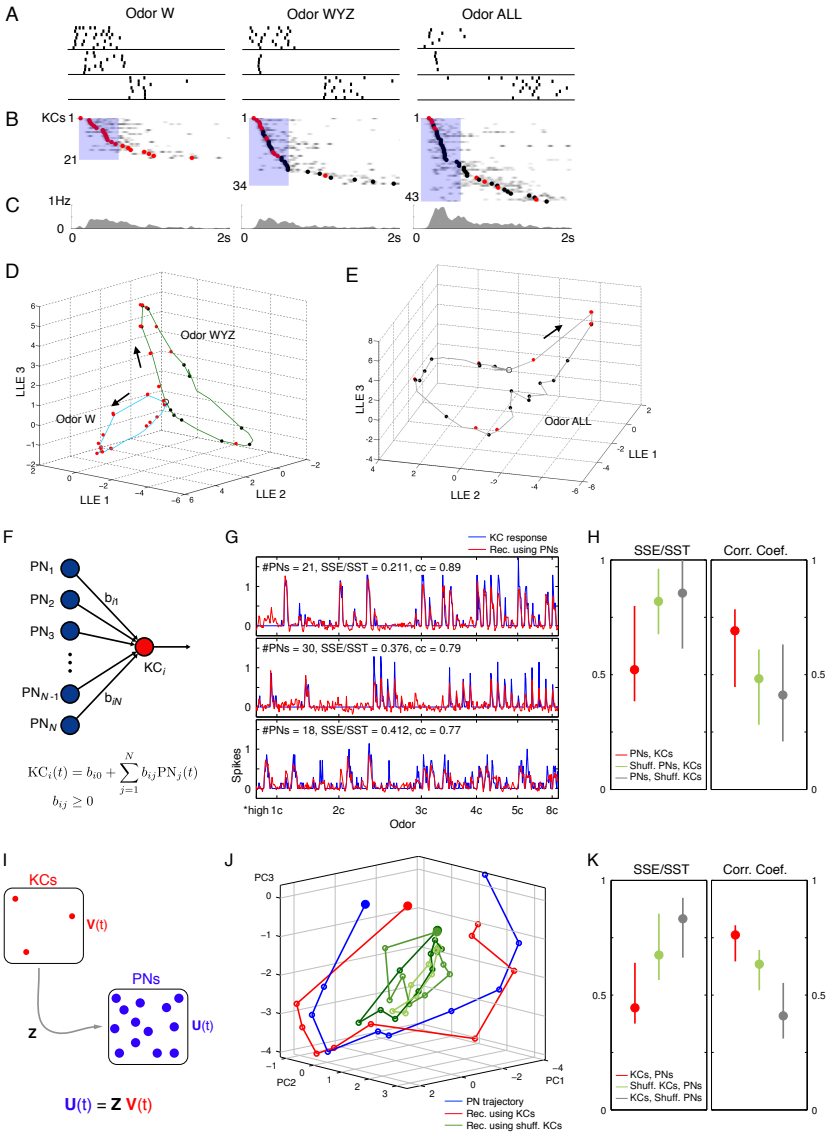


Figure 1-6 Distribution of KC response times.

(A) KC rasters in response to W, WYZ and ALL (ABCDWXYZ). KCs respond reliably across trials and differ in response duration and timing. **(B)** Activity of all recorded KCs that responded to W, WYZ, and ALL. Gray: PSTH averaged across 7 trials, normalized between 0 and 1; dots denote peak of corresponding PSTH. KCs ordered by time of peak, illustrating sequential spread of activity, especially tight within first 500 ms. Red dots: W-responding KCs; black dots: KCs that did not respond to W presented alone. Shaded bar: odor, 500 ms. **(C)** Instantaneous firing rates, averaged across all trials and all 209 KCs to W, WYZ, and ALL. Values are greater with larger mixtures because, on average, more KCs respond to a mixture than to individual components. **(D**, **E)** Two-trial averages of LLE trajectories evoked by W, WYZ and ALL, analyzed over 175 PNs (50 ms consecutive bins, 4.5 s from odor onset), plotted in LLE 1–3 space. LLE computed separately for (D) and (E) (i.e., different coordinates) and plotted separately for clarity. KC PSTH peaks from (B) are superimposed on PN trajectories at corresponding times. **(F)** Schematic representation of PN trajectory reconstruction from KC responses. We searched for a fixed PN x KC basis matrix Z which would transform instantaneous KC population response vectors V(t) into the simultaneous PN population response vectors U(t). **(G)** Example trajectory reconstructions. The PN population trajectory in response to odor ABC is shown in PCA-space in blue. The fit using the KCs is shown in red. Three

reconstructions using three different shuffles of the KCs are shown in green. **(H)** Median, 5th and 95th percentiles of the mean fraction of response variance unexplained (SSE/SST), and mean correlation coefficient between responses and reconstructions, over each trajectory. The distribution over the 44 odors, using unshuffled PNs and KCs is in red. The distribution using shuffled KCs to reconstruct trajectories of unshuffled PNs are shown in green, over all 44 odors and 100 shuffles. The distribution using unshuffled KCs to reconstruct trajectories of shuffled PNs is shown in gray, over all 44 odors and 50 shuffles. Best performance is when unshuffled KCs are used to reconstruct unshuffled PN trajectories. **(I)** Schematic of KC response reconstruction. We search for a set of positive weightings of PN responses that when summed, explain the response of a given KC. **(J)** Best fits. The three best fits are shown. Responses consist of the mean firing rate in each of 10 adjacent 100 ms bins starting at odor onset, and concatenated over the 44 odors. The KC responses are in blue, and the fits are in red. The number of PNs used in each fit is indicated, along with the fraction of unexplained variance, and the correlation coefficient. **(K)** Median, 5th and 95th percentiles of the fraction of response variance unexplained (SSE/SST), and mean correlation coefficient between responses and reconstructions for each KC. The best results are when using the unshuffled PN to reconstruct the unshuffled KCs, in red. The results when using shuffled PNs to reconstruct the KCs are in green for 1000 shuffle of the PNs, and those using unshuffled PNs to reconstruct shuffled KC responses are in gray, for 250 shuffles of the KCs.

1.4.8 Odor Identification, Categorization, and Generalization from Population Activity

Because KCs are in turn read out by downstream decoders (Cassenaer and Laurent, 2007, 2012; Heisenberg, 2003; MacLeod et al., 1998; Masse et al., 2009), we examined the information present in the KC population vectors in appropriately short time bins. Using a linear classifier (regularized-least-squares, see Methods), we compared the decoding of odor identity, category and generalization using instantaneous PN and KC population output. Decoding accuracy in the identification and categorization tasks was based on trials excluded from the classifier training; for the odor generalization task, all trials with the tested odor were excluded from training. Thus, the measured accuracy was what real downstream neurons might achieve in single trials by computing a weighted sum of spikes in each measurement bin (see Methods). *Identification* required attributing a particular KC or PN response vector to the correct odor (all-vs-all, 44 classes, chance = 2.27%). *Categorization* consisted in discriminating mixtures containing a given component from those that did not contain it (balanced sets, repeated over all 8 components and averaged; chance = 50%; see Methods). *Generalization* required the categorization of a previously "unknown" odor (repeated over all positive and negative examples for each of 8 components and averaged; chance = 50%; see Methods). The results are plotted in Figure 1-7 as a function of time around the stimulus. As expected from our large sample of PNs (~ 20% of the entire population), identity (Figure 1-7A, red) and category (Figure 1-7B, red) assignment were nearly perfect (peaks of ~ 100% and ~ 90%, respectively) with this PN set, while peak generalization performance (Figure 1-7C, red) was slightly lower at ~ 85%,

consistent with it being the more difficult task. KCs could also be read out to perform identification, categorization and generalization. Owing to the small size of our KC sample relative to KC population size (~ 0.4% vs. 20%), accuracy was expectedly lower than with PNs. Nevertheless, this performance was only about half that obtained with PN vectors, despite a 40-fold difference in sample size. Peak performance was obtained at ~ 300 ms on average after stimulus onset and remained high for ~ 500 ms beyond odor offset. Peak categorization and generalization performances (averaged over odors) were reached at similar times with KCs and PNs. The correlation between the time courses of PN and KC read-out accuracy was even more striking when performance profiles were considered individually for each odor (Figure S1-7). These observations are consistent with the instantaneous, piecewise decoding of PN output by KCs (Perez-Orive et al., 2004). They also indicate that peak accuracy is not reached with uniform dynamics for all stimuli.

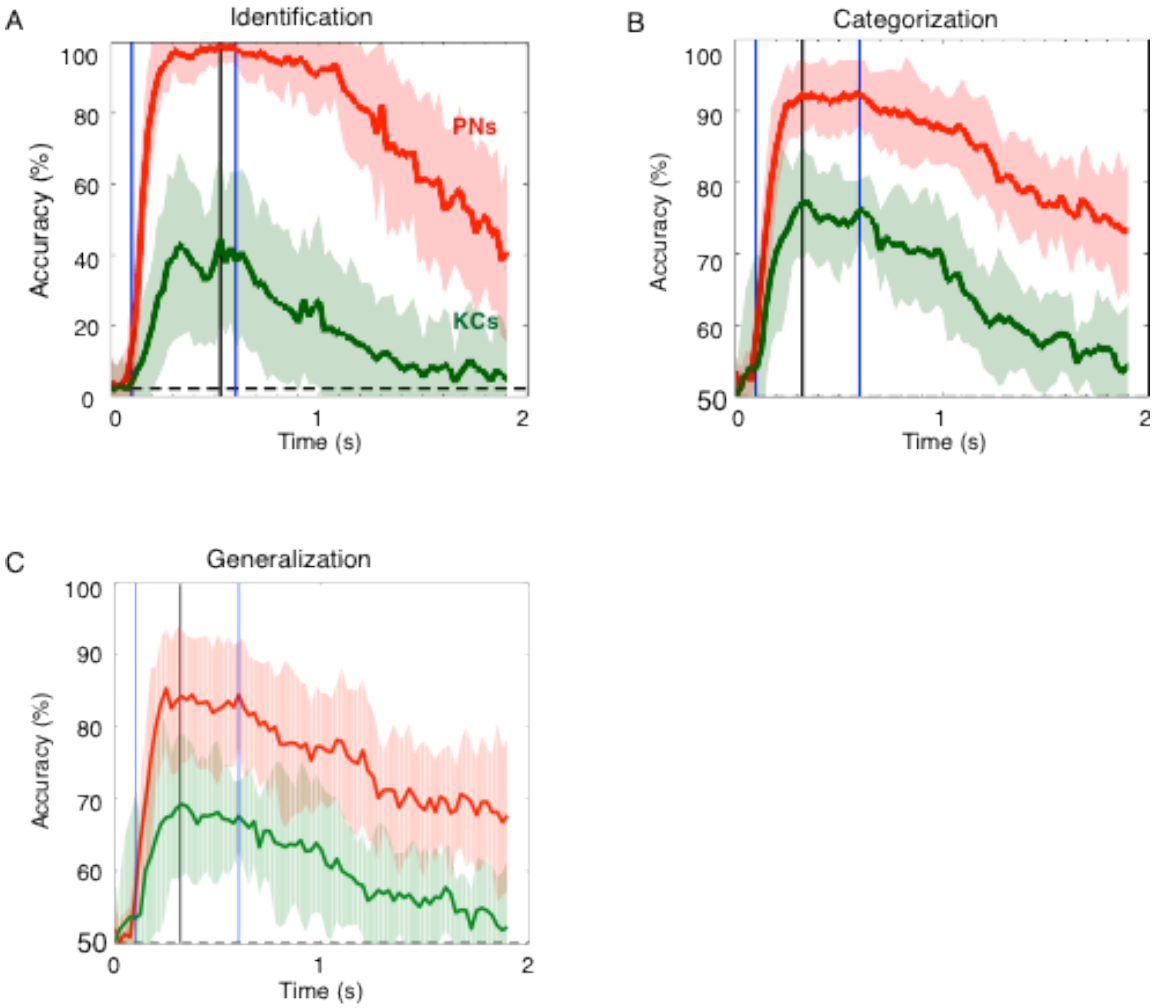


Figure 1-7 PNs and KCs can be linearly decoded to perform odor identification, categorization, and generalization.

Performance of a linear classifier at decoding odor identity (dashed, chance = 2.27%), and category (dashed, chance = 50%), and generalizing category (dashed, chance = 50%), as functions of time. Odor pulse between blue lines (500 ms). Peak of KC-decoding accuracy at black line. Means (traces) and SDs (bands) across all odor conditions. For categorization and generalization, the mean performance in response to in-, and not-in-category odors, for the 8 components were first computed (yielding 16 values for each time bin), and mean and SD of these were then plotted. See Figure S1-7 for the breakdown of the mean responses (averaged over in- and not-in-category) for each of the 8 components.

1.5 Discussion

1.5.1 Mixture Representations by Single PNs are Partially Explained by the Responses of those PNs to the Components

By linearly regressing the mixture responses of PNs on to their component responses, we found that the component responses could explain a significant fraction of the mixture response variance, particularly when a clear mixture response was present: ~ 50% in the binary morph experiments, and ~ 30% in the complex mixture experiments. Most fits used only a single component response, even for complex mixtures. This enabled us to decompose the PN population for each experiment by component preference, showing both that each component is represented in the population, and that a full spectrum of dilution sensitivity exists for each component. Such a distribution of response profiles would clearly facilitate olfactory computation.

The fact that "mixture responses" were rarely identified could be a result of the inherent bias of our model selection procedure against complex models. Indeed, it will always select a single-component model if the addition of other components does not improve the fit by a value greater than the cost of the model's increased complexity, even if the physiological truth indicates a mixture response. For binary-mixture experiments, this potential bias was unlikely, because the majority of cells responded strongly to one component or the other,

with relatively few responding to both. In the complex-mixture condition, we examined the ~ 20% of cases in which at least two equally strong component responses were present but only one was ultimately used in the fit. We found that in 2/3 of cases this was because the second component was poorly correlated with the response and was likely to be physiologically suppressed. In only about 1/3 of cases did we observe a response strongly correlated with the primary component response, and that may have been rejected by the fitting procedure due to its redundancy with the primary response, thus potentially hiding a true mixture response. We also checked many of the fits "manually", especially those with high response SNR that were poorly fit by single components, inspecting the models considered and the models finally selected. These inspections almost always confirmed the decision made. The procedure itself also had very few parameters and required very little tuning. Thus, we believe that the model selection procedure did not introduce an unjustified bias against mixture responses.

Our results suggest that component sensitivity in the antennal lobe is distributed among the PNs. How could this be accomplished? One possibility is that sensitivity to components is built-in, but our data argues against this. One of the odors conditions, cit100:oct100 was present in both the binary mixtures and the complex mixtures experiments. If component sensitivity were built in, and because PN sampling was performed in the same, unbiased manner in both sets of experiments, we would have expected to observe approximately the same number of fits to this mixture response in both sets of experiments. In fact, fits were

found for three times as many mixture responses in the binary morph experiments as in the complex-mixtures experiments. A possible explanation is that the antennal lobe tuned its sensitivity to the odors encountered most often during the course of each experiment. The fact that we found more PNs sensitive to the component most commonly used in mixtures (2,3 butanedione) is consistent with this explanation.

Conversely, the fraction of PNs none of whose mixture responses could be fit was about the same in both sets of experiments (~ 17%). These PNs may have been sensitive to a different set of components than those tested. Incorporating this non-responsive population, and assuming that the sensitivity of the remaining PNs is split evenly between the 8 components tested allows us to predict the number of mixture responses that could be fit. Splitting the PNs among the 8 components used in the mixture experiments, our model would predict that a fraction (1–0.17) *m*/8 of responses could be fit. For the two-component mixture cit100:oct100, this predicts that 0.87*2/8 x 100% = ~ 21% of responses could be fitted using the responses to the components; this is indeed the observed fraction. Averaged over all mixtures, this predicts that ~ 34% of responses could be fitted, a value close to the observed one (31%). Thus our results suggest a first-order approximation of the AL in which component sensitivity is distributed adaptively (i.e., based on immediate or recent experience) among the PNs and in which PNs will respond within a cell-specific range of dilutions to their "preferred" component.

1.5.2 Odor Representations by PN Populations are Ordered

With binary mixtures, measures of similarity between high-dimensional activity vectors—constructed from a large sample of antennal lobe PNs—confirmed visual inspection of corresponding trajectories projected into a reduced space after dimensionality reduction: (1) the representation of a 1:1 mixture lies approximately in between those of the components of that mixture; (2) trajectories cluster by odor rather than by concentration; (3) trajectories change smoothly as one odor is morphed into another. Similar experiments on binary mixtures were recently carried out by Niessing and Friedrich in zebrafish olfactory bulb, based on simultaneous calcium imaging over large numbers of mitral cells (Niessing and Friedrich, 2010). In that work, two dissimilar amino-acid odors (arginine and histidine) were morphed into one another (similar to what we did here), and the corresponding representations compared across all mixture levels. The authors showed that representations shifted abruptly from that for either component to a mixture cluster—a result in apparent disagreement with ours. Leaving aside the difficulties in comparing two model systems that operate in different media, we believe that the two sets of results are not inconsistent. This is because the abrupt transition to a mixture cluster described by Niessing and Friedrich occurs at mixture ratios between 99:1 and 90:10 (or 1:99 and 10:90), ratios just outside those that we tested (90:10 to 10:90 in steps of 10%). The smooth changes we describe concern trajectories that are entirely within Niessing and Friedrich's central/mixture cluster, in which average inter-trajectory correlation was about 0.6.

With more complex mixtures, we found a positive correlation between chemical similarity (i.e., the number of shared mixture components) and representation similarity. This was true whether representations were measured over the entire response window or piecewise, over individual time bins. While such a relationship might have been expected very early in the response, when PN activity is dominated by receptor input, we observed that it obtains throughout the odor presentation and for several seconds following odor offset. Together these results indicate that odor representations are not randomly distributed in PN space, but are ordered so that chemical similarities are reflected in similarities in the evoked neuronal activity patterns. While random representations can in principle be as useful as ordered ones for the decoding of odor identity by downstream cells, ordered representations can make the computation of categorization and generalization easier by representing similar odors in similar ways, and may be the substrate for the categorization and generalization performance we observe in KCs, the PNs' targets in mushroom bodies.

1.5.3 Subspace Readout of PNs by KCs

We used the multi-stage adaptive lasso (Buhlmann, 2011) to regress the binned odor responses of KCs on those of the PNs, and found that by using up to 30 of the available 175 PNs, we could explain ~ 50% of the variance in KC responses on average. This effective connectivity of ~ 17% would seem to contradict previous electrophysiological results that indicate 50 ± 15% connectivity from PNs to KCs (Jortner et al., 2007). These two results can be reconciled due to our finding above: the sampled PN population was split

approximately eight ways according to component sensitivity. Assuming some redundancy between the responses of PNs within a group, a small number of 'basis-PNs' would be required to capture the variability of all the responses within the group, and only those PNs would show up in the regression (due to the sparsity prior of the lasso). Thus this low apparent connectivity could be explained by the redundancy of PN responses.

1.5.4 Individual KCs are Better Odor Segmenters than Individual PNs

Surprisingly, Kenyon cells—which are directly postsynaptic to PNs—are individually much better than PNs at detecting a component in a mixture of up to eight odors (ROC analysis). This observation might be explained using a simple abstraction: odor representations are spread orderly in a high-dimensional PN space (Figure 1-3,1-4); because of their 50±15% connectivity to PNs (Jortner et al., 2007), however, the KCs each sample a different, lower-dimensional, subspace of PN space. After appropriate choice of these subspaces, individual KCs could recognize relationships between the trajectories of odors that contain a common component, even though those relationships are not detectable in the full PN space (Figure 1-4B). Alternatively, given that experiments involving KC recordings only commenced if a response had first been elicited in one KC by at least one of the eight odor components (see below), it is also possible that a yet-to-be-established learning rule fine-tuned PN-KC connectivity during this testing phase, such that a KC was

more likely to respond to a mixture if it had first been exposed to a component of that mixture.

1.5.5 Population Decoding from PNs and KCs

Our results (Figure 1-7, S1-7) show that both the PN and KC populations can be read out by linear classifiers in single time bins to perform odor identification, segmentation, and generalization, and that the time course of the performance is similar in both populations. Although we sampled ~ 20% of the full PN population, but only ~ 0.4% of the KC population, readout performance from KCs was usually only slightly worse than from PNs. One explanation for this observation is that by pooling information across the PN population, individual KCs are more informative than individual PNs. However, another explanation is that our bias in KC selection due to experimental design may have skewed our KC dataset towards particularly informative cells. The collection of larger and less biased KC datasets will be needed to elucidate this point.

1.5.6 Experimental Sampling Bias

Because odor representations by PNs are dense, finding responding PNs in an experiment was always guaranteed: recordings could start immediately after tetrodes had been inserted into the antennal lobes and yielded high S/N data. By contrast, the sparse responses of KCs forced us to do a targeted search on KC signal prior to initializing an experiment. This

search was done using only the eight single odors in our stimulus set. As soon as high S/N signal was detected on at least one tetrode in response to at least one of the eight odors, the experiment could commence. Our KC dataset is thus somewhat biased towards KCs that respond to the odors present in our single-component stimulus set. In a typical recording session, however, it was common to record KCs that did not respond to any of the single components. Among those were some that responded to "low-*n*" mixtures—two or three components. Those KCs then often responded to high-*n* mixtures containing the low-*n* ones, practically segmenting these lower-order mixtures, just as other KCs detect single components. Hence, while our KC data are biased towards KCs that responded to the 8 single components in our stimulus set, it contains no experimental bias towards mixture-responsive or component-segmenting KCs, for those were all discovered *post-hoc*, during data analysis. It is possible, however, that our initial screening procedure with the eight single odors introduced an acquired (though unconditioned) selectivity for these components. Whether the segmenting properties of KCs we describe here are intrinsic or learned via a fast non-associative process (see, for example, (Stopfer and Laurent, 1999)) is unknown thus far.

1.5.7 Functional Consequences of Odor Segmentation

That KCs responding to a mono-molecular odor respond also to mixtures containing that odor is important for our understanding of computation in this system. Our results are illustrated in Figure 1-8A. Each row represents a KC (taken from our dataset) that

expressed good segmenting properties (one KC for each one of the eight single odors). Each column represents one of the 43 stimulus conditions (paraffin oil not shown). The color of each square identifies the odor component that the corresponding KC detected. The circles indicate a response. For example, KC3 was a nearly perfect segmenter for citral, with only one false positive (last of the *3*-mixtures) and one false negative (*8*-mixture). Thus, any odor—whether simple or composite—can be represented by a unique 8-D activity vector. The first *4*- and *5*-mixtures, for example, are represented by KC activity vectors that differ simply by the activation of KC6. To the limit, if every KC was a perfect detector for only one feature, then *n* KCs could encode 2^n-1 different odor feature combinations, plus baseline (0,0,…0). By contrast, a "grandmother" scheme whereby each odor is represented by a unique neuron would require 2^n-1 KCs to represent this many odors and mixtures. Hence, KCs implement a clever strategy. Odor representation is sparse (effective for memory formation and recall, yet not maximally sparse), but distributed such that the coding capacity for related stimuli (mixtures) is greatly increased.

One could argue that the coding strategy of PNs is superior, because it engages far fewer neurons (800 *vs.* 50,000) to accomplish the same goal (information captured by KCs is obviously present across the PN population). However, because PN codes are dense, they overlap extensively. This scheme is economical for encoding, but bad for storage (Field, 1994; Foldiak, 2003): among other problems, a synapse modified to encode one memory will interfere with the encoding of other memories. Mushroom body representations are

therefore the expression of a trade-off: by using odor-segmenting neurons (explicit feature-representation scheme), they maximize capacity while minimizing total KC number. By using sparse representation vectors, they minimize overlap and interference. Both attributes are desirable for a memory system. This mixture-encoding format has yet another advantage: it leads naturally to stimulus generalization. The logic is illustrated in Figure 1-8B to 1-8F. The odor encoding process maps each odor to a pattern of KC population activity. Decoding can be achieved by linear separation of KC population activity, formalized as hyperplanes. In principle, the mapping from odors to KC population activity could be arbitrary, but some mappings make odor generalization easy to compute. Figures 1-8B and 1-8C show one such mapping, in which each KC signals the presence of one odor component. In this mapping a decoder that clusters the KC activity vectors for A, AC and AB (red plane, Figure 1-8B) will naturally group the vector for ABC with them, thus generalizing. This results from the order in representation space. In Figure 1-8D to 1-8F, the same cells and odors are plotted but each KC no longer represents a single odor component. In Figure 1-8D, generalization is just as easy as in Figure 1-8B,C, but the absence of odor is now represented by spikes in KC3, and A by (0,0,0), both of which contradict experimental data. In Figure 1-8E, AB, AC and BC can no longer be separated from, respectively, (not-AB), (not-AC) and (not-BC) using single linear classifiers. A hyperplane exists that can separate B from (not-B) (example shown), but this requires reading out 3 KCs rather than one. In Figure 1-8F, C and (not-C) are not separable with a single linear classifier. We conclude, therefore, that stimuli are not represented by sparse and random sets of KCs (like in the ASCII code). Each KC represents a meaningful feature,

and each stimulus is encoded by the combination of relevant feature-selective KCs (Barlow, 1972; Foldiak, 2003). In principle, this ordered scheme allows decoders of KC activity to determine not only the degree of similarity between stimuli, but also the assignment of category for novel stimuli (generalization). Hence, the scheme we observed for mixture coding by KCs is consistent with the fulfillment of several concurrent requirements: economy of size, maximization of capacity for that size, minimization of overlap between memories, and generalization. The rules observed here for a simple olfactory system could, in principle, form the basis for the encoding of multi-dimensional signals in any sensory system with comparable requirements.

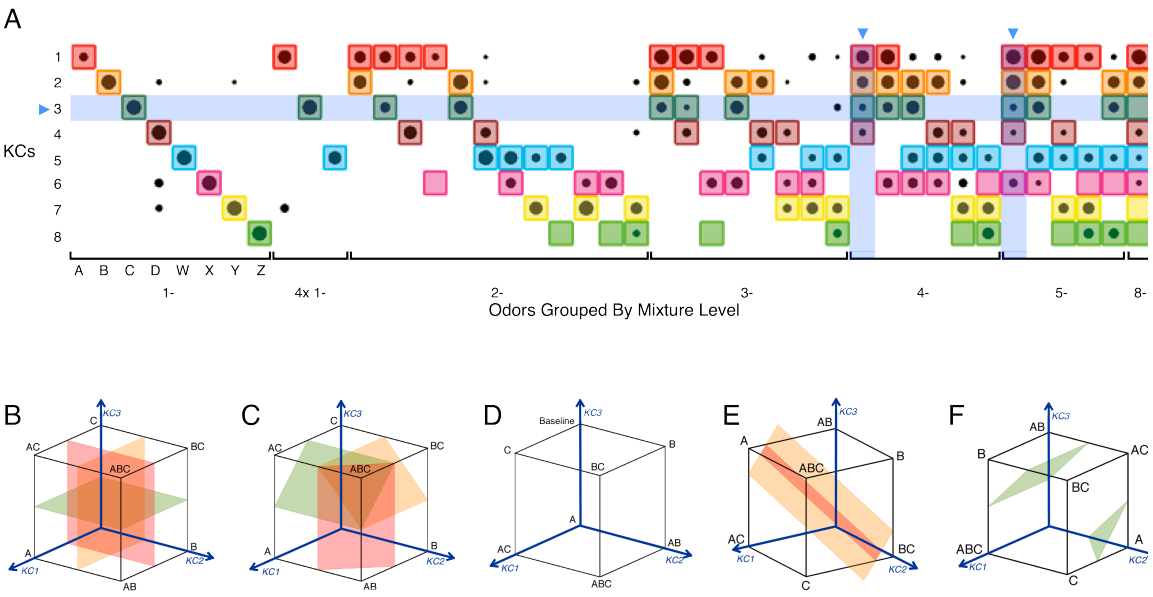


Figure 1-8 Coding principles for odor identification and generalization by KC assemblies.
(A) Diagram indicating the responses of eight of our recorded component-detecting KCs (one for each of the 8 tested) to all odor conditions. Response measured as spike counts in 1 s window from odor onset summed over all 7 trials, normalized between 0 and 1 for each KC. Filled circle area represents response. Each color represents one odor. A filled circle on a colored square indicates a "true positive" (a KC recognizes the component to which it is tuned from within a mixture). An empty colored square indicates a "false negative". Black circle alone indicates "false positive". KC 5 detects W without mistake. In the absolute, downstream decoders need only read one KC to recognize a category: for instance, a response in KC 3 indicates the presence of odor C. Identification is possible by observing across the 8-D KC vector (e.g., discrimination between mixtures ABCD and ABCDX, vertical arrows). With perfect component-detectors, n KCs can discriminate between 2n-1 odor mixtures (all combinations) and baseline. The code is sparse for single components and more distributed (though still sparse) for mixtures. **(B, C)** Linear classification of odors and mixtures using ordered KC encoding scheme. Schematic of KC coding space where each KC represents an odor component. With this scheme, generalization is simple: **(B)** shows hyperplanes that separate mixtures into A vs. not-A, B vs. not-B, and C vs. not-C; **(C)** shows hyperplanes that generalize mixtures into AB vs. not-AB, BC vs. not-BC, and AC vs. not-AC. Information is represented such that different kinds of generalization are easy to compute with linear classifiers**. (D–F)** Coding strategies with a scrambled scheme. In **(D)**, the presence of odors A, B, C, AB, AC and BC can be computed with a single linear classifier in each case, but A is represented by [0 0 0] and no-odor is represented by activation of KC3. This strategy is incompatible with experiments; KC baseline activity is 0, precluding signaling by inhibition. In **(E)**, although a hyperplane exists that separates B from not-B, odors AB/not-AB, AC/not-AC, and BC/not-BC each require multiple hyperplanes for separation. In **(F)**, odors C and not-C are not separable by a single hyperplane (as shown). See text for details.

1.6 Methods

1.6.1 Preparation and Stimuli

Results were obtained from 61 locusts (*Schistocerca americana*) in a crowded colony. We recorded from 168 PNs for the binary mixture experiments, 175 PNs for the multi-component mixture experiments and 209 KCs from 13 (42 groups), 11 (36 groups) and 37 locusts (53 groups), respectively. Experiments were typically conducted using left and right antennal lobes (ALs) and mushroom bodies (MBs) in each animal. Young adults of either sex were immobilized, with one or two antennae intact for olfactory stimulation. The brain was exposed, desheathed and superfused with locust saline, as previously described (Laurent and Davidowitz, 1994). Odors were delivered by injection of a controlled volume of odorized air within a constant stream of desiccated air. Teflon tubing was used at and downstream from the mixing point (see below) to prevent odor lingering and cross-contamination. Due to different odor stimulus requirements between the binary morphing and the multi-component mixture experiments, we built two independent odor delivery systems (Figure S1A,B).

1.6.2 Binary Mixture Experiments

Two odorants (1-octanol and citral, Sigma, Figure 1-1A, 1-1D) were stored as pure solutions in independent 500 ml custom-made bubblers. The odor nozzle (1 cm diameter, Teflon) was placed 1 cm from the antenna and supplied a constant 1 l/min carrier stream of

desiccated, filtered air. The flow of each odor was controlled by an independent electronic flow meter with feedback control (Aalborg, 2–200 ml/min) and a solenoid placed upstream of custom-made bubbler (Figure S1-1A). Relative odor concentrations were varied by controlling flow through each bubbler (30, 60, 80, 100, 120 and 140 ml/min). These concentrations spanned the dynamic range of electro-antennal responses recorded in recordings from isolated antennae (electro-antennograms, or EAGs). A large vacuum hose placed behind the antenna guaranteed the quick removal of odorants from the space surrounding the antenna. Odor puffs were triggered automatically using a custom computer interface (LabView, National Instruments Inc.). Trials were 14 s long, with 300-ms-long odor puffs presented 2 s after trial onset; each odor condition was repeated 10 times.

1.6.3 Complex Mixture Experiments

Individual odors (Figure 1-1E) were chosen to be chemically different and their respective concentrations adjusted to ensure that no one odor dominated over the others due to intrinsic differences in vapor pressure. In practice, odor concentrations were calibrated by dilution in paraffin oil to equalize EAG responses, recorded from isolated antennae (Figure S1-1D). Paraffin oil alone elicited negligible EAG response. The odors were: 1-octanol (A), diluted 0.7 ml/10 ml; phenetole (B), diluted 0.15 ml/15 ml; citral (C), pure 10 ml; benzaldehyde (D), diluted 0.02 ml/15 ml; iso-amyl-acetate (W), diluted 0.1 ml/10 ml; 2,3-butanedione (X), 0.04 ml/15 ml; 2-nonanone (Y) diluted 2 ml/15 ml; L-carvone (Z), pure 10 ml; paraffin oil (P), pure 15 ml. The individual odors were each placed into a glass vial

(60 ml). The headspace content was carried by puffs of desiccated and filtered air, with a flow rate of 100 ml/min for individual odors and 400 ml/min for paraffin oil. Three odors: 1-octanol (A), citral (C), and iso-amyl-acetate (W) were also presented at a second, higher concentration, by increasing flow rate to 400 ml/min. Odor mixtures were presented by combining the single odorants. For example, AB was the combination of 100 ml/min of A with 100 ml/min of B, with a total odor flow of 200 ml/min; thus, total odor concentration was higher during mixture conditions. A compensating stream of desiccated air was used to ensure that total air-flow remained constant throughout the experiment. The odors were mixed in a custom-built corrugated glass tube (~ 8 cm long, 1.5 cm diameter), with a total flow of 2 l/min to ensure turbulent mixing. The individual odor lines were arranged along the circumference of the mixer. Trials were 14 s long, with odor puffs presented for 500 ms, 2 s after trial onset, and repeated 7 times with each stimulus. To minimize the potential effects of priming (Bäcker, 2002), single odorants were presented first, followed by 2-, 3-, 4-, 5- and 8- mixtures. The order of presentation within each mixture group was pseudo-random.

1.6.4 Electrophysiology

Two types of tetrodes were used for extracellular recordings. Silicon probes were obtained from NeuroNexus. Wire tetrodes were constructed with insulated 0.0005" and 0.0004" wire (REDIOHM wire with PAC insulation). Four strands of wire were twisted together and heated to partially melt the insulation. The tip was cut with fine scissors and each channel

tip was electroplated with a gold solution to reduce the impedance to between 200 and 250 kΩ at 1 kHz. The same custom-built 16-channel preamplifier and amplifier were used for both types of tetrodes. Two to four tetrodes were used simultaneously. The preamplifier had a gain of 1, and the amplifier gain was set to 10,000x. Because of low baseline activity and low response probability in KCs, fewer KCs than PNs were usually isolated in a typical recording session. Tetrodes were placed within the AL or MB soma clusters, peripheral to the neuropils at depths between 50 and 200 μm. For some MB recordings (KCs, LFP), probes were pressed on the surface of the MB. Cell identification was unambiguous because PNs are the only spiking neurons in the locust AL—LNs do not produce sodium action potentials (Laurent and Davidowitz, 1994)—and because all the somata located dorsal to the MB calyx belong to KCs. Recording locations were tested randomly across the MB and selected if activity could be elicited by any of the 44 odor conditions. Identical stimuli were presented at the beginning, middle and end of the experiment to check that clusters had not drifted significantly over the course of the experiment. Drift was estimated qualitatively by determining if a given neuron's responses to each odor were similar across these three sampling periods. Hints of drift then led to examination of the waveform clusters.

1.6.5 Recording Constraints and Sampling Biases

Because PNs respond very promiscuously to odors (Perez-Orive et al., 2002), no effort was made to find PNs that responded to our stimuli. As soon as good signals suggestive of

separable PN clusters could be seen, recordings started and responsive PNs were always found. Our estimates of PN response-probabilities are therefore likely close to true values. By contrast, KCs respond very rarely to odors and responsive KCs had to be found through an active search. Experiments would commence only if a response was elicited by at least one of the 8 monomolecular odors. See the Discussion and supplementary text for more details on the search procedure and the resulting biases introduced.

1.6.6 Extracellular Data Analysis

Tetrode recordings were analyzed as described in (Pouzat et al., 2002). Briefly, data from each tetrode were acquired continuously from the four channels (15 kHz/channel, 12 bits/sample), filtered (custom-build amplifiers, band-pass 0.3-6 kHz) and stored. Events were detected on all channels as voltage peaks above a pre-set threshold (usually 2.5-3.5 times each channel's signal SD for PNs, and 4-5 SDs for KCs). For any detected event on any channel, the same 3 ms window (each containing 45 samples) centered on that peak was extracted from each one of the four channels in a tetrode. Each event was then represented as a 180-dimensional vector (4 × 45 samples). Noise properties for the recording were estimated from all the recording segments between detected events, by computing the auto- and cross-correlations of all four channels. A noise covariance matrix was computed and used for noise whitening. Events were then clustered using a modification of the Expectation-Maximization algorithm. Because of noise whitening, clusters consisting of, and only of, all the spikes from a single source should form a

Gaussian (SD = 1) distribution in 180-dimensional space. This property enabled us to perform several statistical tests to select only units that met rigorous quantitative criteria of isolation (Pouzat et al., 2002).

1.6.7 Computational Analysis

MATLAB (The MathWorks, Inc.) was used for all data analyses.

1.6.7.1 Dimensionality Reduction

For nonlinear dimensionality reduction with locally linear embedding (Roweis and Saul, 2000), we used code from Sam Roweis (http://www.cs.toronto.edu/~roweis/lle/) with Gerard Sleijpen's code for the JDQR eigensolver (http://www.math.uu.nl/people/vorst/JDQR.html). In the figures shown for nonlinear dimensionality reduction with LLE, we used as inputs 168-D time slices, 50 ms long each, averaged over 3 trials (binary mixtures), and 175-D time slices, 50 ms long each, averaged over 2 trials (multi-component mixtures). Other details are as described in (Stopfer et al., 2003).

1.6.7.2 Correlation Distance Insets

To complement the LLE trajectories in low–D space for PN odor responses, we computed correlation distances (1 minus linear correlation) between mean odor responses (N-dimensional rate vectors for each time bin) in trials 3–6 vs. trials 7–10 for the binary

mixtures, and trials 2–4 vs. 5–7 for the complex mixtures, in consecutive 50 ms time bins starting at odor onset. Figure 3B shows the resulting distances for the responses to pure octanol, pure citral, and their 1:1 mixture. To summarize this information, we then computed the minimum value of the distance for each odor comparison (Figure 1-3C), and included this summary inset with each one of our LLE figures. The displayed range of the distances was clipped for clarity to 0–1 from the full 0–2.

1.6.7.3 Bayesian Model Selection

We employed Bayesian model selection (MacKay, 2003) to select between regression models when fitting the mixture responses of single PNs to their component responses, and to determine whether PN population trajectories in response to binary mixtures evolved gradually or stepwise as the mixture was varied from pure octanol to pure citral. Given two models M_1 and M_2 of data D, Bayesian model selection computes posterior probabilities $P(M_1|D)$ and $P(M_2|D)$ for the models given the data and selects M_1 as a better description of the data than M_2 if the ratio $P(M_1|D)/P(M_2|D)$ is greater than 1. Assuming equal priors on the models, i.e., $P(M_1) = P(M_2)$ (as we do when describing trajectories), this ratio is

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)P(M_1)/P(D)}{P(D \mid M_2)P(M_2)/P(D)} = \frac{P(D \mid M_1)}{P(D \mid M_2)} = \frac{\int_{\theta_1} P(D, \theta_1 \mid M_1)\, d\theta_1}{\int_{\theta_2} P(D, \theta_2 \mid M_2)\, d\theta_2} = \frac{\int_{\theta_1} P(D \mid \theta_1, M_1) P(\theta_1 \mid M_1)\, d\theta_1}{\int_{\theta_2} P(D \mid \theta_2, M_2) P(\theta_2 \mid M_2)\, d\theta_2},$$

where θ_1 and θ_2 are place-holders for the parameters of the two models. The parameter likelihoods $P(D|\theta,M)$, and the parameter priors $P(\theta|M)$ depend on the form of the model. Unless otherwise noted, the integrals were computed using Laplace's method (MacKay, 2003), in which a Gaussian is fit to the integrand at its peak and the integral is estimated as

that of the fit Gaussian. The peak of the integrand was found numerically using the *fminsearch* function in MATLAB's Optimization Toolbox, and the Hessian of the logarithm of the integrand (required for the Gaussian fit) was estimated using the *hessian* function provided in John D'Errico's DERIVEST toolbox, available on the MATLAB file exchange.

1.6.7.4 Explaining Single PN Mixture Responses using Component Responses

The procedure for fitting single PN binary mixture morph responses was the same as when fitting complex mixture responses. We first binned the responses into consecutive 50 ms time bins starting at 0.5 sec before odor onset and ending at 3 seconds after odor onset, for a total of 70 bins. We chose a window of this size so that the models fit would have to explain the transitions to and from baseline as well as the response itself. The spike counts in these bins were averaged across trials, yielding, for an *n*-component mixture, *n* 70-element regressors (the responses to the single components), and 1 70-element response to be fit. We then computed posterior probabilities for the constant model (no component responses used) and each of the 2^n-1 regressor configurations in which at least one of the component responses was used. For of these configurations, we tested the following three types of models:

Name	**Form**	**Parameters**
Unit	$Fit(t) = b_0 + x_1(t) + x_2(t) + \ldots$	b_0.
Scaled	$Fit(t) = b_0 + b_1 (x_1(t) + \ldots + x_k(t))$	b_0, b_1.
Free	$Fit(t) = b_0 + b_1 x_1(t) + \ldots + b_k x_k(t)$	b_0, b_1, b_2, …, b_k.

For computing the model posteriors, we used the standard linear regression assumptions that the data points in each time bin are independent, and normally distributed around their predicted values with variance *v* (provided as an additional parameter). We could then computed the likelihood of the parameters (**b**, v) given the data y and the model M as

$$P(y|\mathbf{b},v,M)=\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi v}}\exp\left(-\frac{(y(t)-fit(t))^2}{2v}\right)=(2\pi v)^{-T/2}\exp\left(-\frac{SSE}{2v}\right)$$

where SSE is the sum of the squared difference between the fit and the data. We used the same, Gaussian, prior for the regression coefficients, and a Gamma prior for the variance:

$$\begin{aligned}P(\mathbf{b},v|M)&=P(\mathbf{b}|M)P(v|M)\\P(\mathbf{b}|M;\sigma_b^2)&=\prod_{i=1}^{k}\frac{1}{\sqrt{2\pi\sigma_b^2}}\exp\left(-\frac{b_i^2}{2\sigma_b^2}\right)=(2\pi\sigma_b^2)^{-k/2}\exp\left(-\frac{\sum_{i=1}^{k}b_i^2}{2\sigma_b^2}\right)\\P(v|M;\alpha,\beta)&=\beta^{\alpha}\frac{v^{\alpha-1}e^{-\beta v}}{\Gamma(\alpha)}\end{aligned}$$

We set the variance of the regression prior to 1 because earlier experimentation with unconstrained fits had shown that the coefficients typically fell in the [-1.5, 1.5] range, and a variance of 1 would allow the regressions freedom to achieve such fits. For the Gamma prior on the noise variance we set alpha and beta to 1, in which case the prior reduces to a decaying exponential with unit constant. This is the simplest case of the prior and has a broad enough support to accommodate most fits while avoiding the very large variances that are often associated with poor fits. We could then compute the likelihood of each model by using Laplace's method to approximate the integration over model parameters

$$P(y|M) = \int_\theta P(y, \theta|M) d\theta = \int_\theta P(y|\theta, M) P(\theta|M) d\theta$$

We were then in a position to compute the posterior on each model given the data, to within a constant of proportionality (the same for all models):

$$P(M|y) = \frac{P(y|M)P(M)}{P(y)} \propto P(y|M)P(M)$$

In our first attempt we used the same prior for all models, so that the model with the highest likelihood was selected. However, manual inspection of the fits showed that non-constant models were being chosen where the constant model would suffice, and spurious secondary components were frequently included in the fits. After an initial period of experimentation with manual tuning of the model priors, we took inspiration from the minimum description length principle (Grünwald, 2007), and assigned a model fitting an *n*-component mixture with k components a prior of n^{-k}, since in a naïve encoding it would require log(n) bits to specify each of the components used and k log(n) bits in total. The log probability of the prior on a model would then reflect its description cost, so that log(P(M)) = -k log n, yielding the expression above for the prior probability on the model. Using this prior on models immediately produced very good results and obviated manual tuning of the model priors.

Finally, to fit lagged versions of the models, we first estimated the best lag values by trying all possible lag combinations for the single components being considered in a given fit (lagging each component by up to 3 time bins in either direction), and taking the

combination that produced the best ordinary least-squares fit to the data. We then substituted these lagged versions of the regressors in the fitting procedure above, while also including terms for the priors on lags. We used a Gaussian prior with mean zero and variance 1 for each lag term, to represent our belief that small lag values may be possible due to jitter in responses and odor delivery, but that large lag values are likely to be spurious.

Using the procedure described above we were able to compute posteriors for all of the models given the observed data (to within a constant of proportionality that is the same for all models), and we chose the model with the highest posterior.

1.6.7.5 Single PN Response SNR

To compute the SNR of a single PN's mixture response we first compute the average noise power, by computing the variance of each component response and the mixture response during the baseline period. The baseline period was defined as the two seconds before odor onset, and the 5 second interval starting 8 seconds after odor onset (this latter interval was selected because the average activity computed over the population had settled to near its baseline value by this point). The baseline variance is then averaged over the components and the mixture response to yield an estimate of the noise power. The signal power for a given response is defined as the average of its squared deviation from its mean value during the baseline period, over the two seconds following odor onset, i.e., the mean squared-error when using the baseline mean to predict the response. The signal power is then divided by the noise power to yield the SNR. To get the SNR in dB, its base 10 logarithm is multiplied

by 10. SNRs for component responses are computed in the same way, but substituting the component response for the mixture response when computing the signal power.

1.6.7.6 PN Population Odor Representation Metrics

1.6.7.6.1 Conventions

Unless otherwise noted, we used the following conventions for the PN odor representation metrics. The *response window* refers to the 1 second period [0.1 s to 1.1 s] following odor onset. The 100 ms offset is to account for the time needed by odors to reach the antenna. The *baseline window* refers to the 1 second period [-1.1 s to -0.1 s] before odor onset. Responses were binned in 100 ms consecutive bins aligned with odor onset. Metrics were computed for single trials, and means and S.E.M.s were computed across trials. Metrics were computed either *locally*, meaning independently for each time bin, or *globally,* by first temporally concatenating the binned responses in the window into a single vector (the *global trajectory*), and then computing the metrics.

1.6.7.6.2 Projection Angle Fraction

To compute the projection angle fraction (PAF) of a binary mixture response with respect to citral, the mixture response in a given bin was projected onto the plane spanned by the simultaneous responses to pure octanol and to pure citral. The angle of the projection to citral was computed, and divided by the angle between octanol and citral. The PAF with

respect to octanol was computed in the same way, except that the angle of the projection to octanol was used instead of that to citral.

1.6.7.6.3 Concentration Series Clustering

We used the Rand index (Rand, 1971) to determine whether the population responses to the binary mixture concentration series clustered by odor or by concentration. Given two partitions (non-overlapping and exhaustive clusterings) L_1 and L_2 of a finite dataset, the Rand index is defined as the ratio of the sum of the number of pairs of elements that are in the same partition in both datasets and the number of pairs that are in different partitions in both datasets, to the total number of pairs of elements. It ranges in value from 0 to 1, with higher values indicating greater agreement between the two partitions. Intuitively, it can be interpreted as the probability that a randomly selected pair of elements will be grouped the same way in both clusterings.

In line with the dimensionality reduction results (Figure 1-3E), we assumed that the global trajectories in a given trial for the three odors and the five concentrations would form three clusters when clustered by correlation distance. Our aim was to use the Rand index to measure the agreement of such a clustering with clustering by odor (by labeling each trajectory with its corresponding odor), and with clustering by concentration (by labeling each trajectory with its corresponding concentration). Because trajectories for 5

concentrations were available, such a procedure would introduce a bias against clustering by concentration due to the mismatch in the number of clusters (3 for clustering by distance, 5 for clustering by concentration). Hence we first split the data into all 10 possible subsets of 3 of the 5 concentrations (no such procedure was necessary for the clustering by odor, since only three odors were used). For each concentration subset, we used 10 runs of k-means clustering to cluster the nine global trajectories (9 = 3 odors x 3 concentrations) in a time window of interest for a single trial into three clusters by correlation distance. For each such run, we computed the Rand index measuring the agreement of this clustering with clustering by odor, and another Rand index measuring the agreement with clustering by concentration. We then computed the averages of these two indices over concentration subsets and k-means runs to yield two summary indices for each trial. Finally, we computed the means and S.E.M.s of these two summary indices over trials to yield the bar plots in Figure 1-3H.

Because the Rand index does not correct for chance (e.g., by assigning the chance level of agreement a value of 0), we estimated the chance level directly. (An "adjusted Rand index" exists that performs such a correction (Hubert and Arabie, 1985), at the cost of making the index itself more complex. For clarity of interpretation, we chose to use the simple Rand index and compute the chance level directly.) For each of the concentration subset and k-means runs above, we computed a chance Rand index by randomly shuffling the odor (or equivalently, concentration) labels of the trajectories, and computing the index between the

resulting clustering and that by correlation distance. We repeated this procedure 1000 times for each run. We then averaged the indices over these repetitions, the 10 concentration subsets and the 10 k-means runs to yield a summary index for the chance level in each trial. We then computed the mean and S.E.M. across trials, and displayed the means for the two different time windows of interest in Figure 1-3H (the S.E.M.s were negligible and were left out for clarity). The slight difference in the chance levels between the baseline and response windows is due to the differences in clustering by correlation distance in the two time windows.

1.6.7.6.4 Evolution of PN Population Trajectories for Binary Mixtures

We used Bayesian model selection to determine whether binary mixture trajectories evolved gradually or stepwise as the mixture was varied from pure octanol to pure citral. We computed the posteriors for four different models: constant, linear, one-step, and two–step. All models assume that the data points are independent and are normally distributed with unknown variance v around a model-dependent function of the independent variable x: a constant value yl for the constant model, a line for the linear model, a constant level yl_1 for x < a threshold θ, and another constant level yl_2 beyond for the one-step model, and a constant level yl_1 for x < a threshold θ_1, a constant level yl_2 for $\theta_1 \leq x <$ a threshold θ_2, and a second constant level yl_3 beyond for the two-step model. Priors on the constant levels and the intercept of the linear model were uniform over the [0,2] range of correlation distance. The prior on the slope of the linear model was taken to yield a uniform prior over [-π/2,

$\pi/2]$ on the angle of the line (relative to the x-axis). For the one- and two-step models, the prior on the first threshold value was uniform over the x-range of the data, and that on the second was uniform over the range from the first threshold to the upper limit of the x-range. An improper prior (Sivia and Skilling, 2006) of 1/v was used for the variance. Additional details including a description of integration methods used to compute the posteriors are provided in the supplementary text. The base-10 logarithm of the posterior probability for the linear model was subtracted from that of the other models, and the mean and S.E.M. over trials of the result are plotted in Figure 3I.

1.6.7.6.5 Complex Mixture Trajectory Relationships

We assessed the relationship between complex odor mixtures and the responses they evoked by computing the Spearman rank correlation between the set of all pairwise distances between the odors with the correlation distances between the evoked activity patterns. We represented odors as 8-bit binary vectors and used the Jaccard distance (Deza and Deza, 2009) to measure distances between them, defined in our context as 1 minus the ratio of the number of components two odors share to the total number of components present pooled over the two odors (see supplementary text for a justification of this metric). For a global measure (Figure 1-4D), we correlated the odor distances and the correlation distances between the trajectories in the response window, in the baseline window, and for the response window but after shuffling the odor labels (separately for each trial) of the responses. For a local measure (Figure 1-4F), we computed the Spearman rank correlation

between the odor distances and the correlation distances between the evoked population activity patterns, for the evoked activity patterns but with odor labels shuffled once per bin and per trial (red), and for the evoked activity but with PN identities shuffled for each odor and each bin, but using the same shuffling across trials for a given odor and bin (black).

1.6.7.7 PN and KC Responsivity

A PN or KC was classified as responding if its firing behavior during the response window met two independent criteria of response *amplitude* and *reliability*.

Amplitude: the neuron's firing rate (measured in successive 200 ms bins, averaged across all trials) had to exceed the mean baseline and *n* standard deviations of the baseline rate (measured across 200 ms bins and all 7 trials) in at least one bin within the response window. Baseline rate was measured for each cell-odor pair over a period of 600 ms preceding stimulus onset and over all 7 trials. Values of *n* of 1.5 or 2 gave low rates of both false positives (during baseline) and false negatives (during stimulation) in PNs. Values of *n* between 0 and 4 made no significant difference with KCs. We show results with $n = 1.5$.

Reliability: to ensure that responses detected were reliable even at low firing rates in KCs, we required that at least one trial more than 50% of all trials (i.e., at least 4/7) with each odor contained at least one spike during the response window. Our metric for

responsiveness is extremely conservative, because it measures PN activity for only 1 s shortly after odor onset. In reality the dynamics of PNs last for as long as 3–4 s after odor offset (e.g., rebound excitation that occurs later in PNs is not captured by our metric).

1.6.7.8 ROC Analysis

We used ROC analysis (Fawcett, 2006) to evaluate individual PNs and KCs as classifiers for the presence of single odor components in mixtures. In the ROC framework, a binary classifier computes a score for each input that is thresholded to assign a class (positive or negative) to the input. The fraction of positive inputs that are correctly classified yields the true positive rate (TPR) of the classifier at the given threshold, while the fraction of negative inputs that are incorrectly classified yields the false positive rate (FPR). Plotting the TPR against the FPR as the threshold is varied yields the *ROC curve*, and the area under this curve is the *AUC* score which is a measure of the performance of the classifier (0 for perfect reverse classification, 0.5 for chance performance, 1 for perfect classification). Applied to the PNs and KCs, we first found cells that responded (using our response criteria above) to at least one of the single odor components. For each single odor component, we partitioned all odor conditions into two classes. For example, for the case of component A, the "positive" class consisted of all odor conditions including A (A-high, A, AB, AC, AD, AX, ABC, ACD, AXZ, ABCD, ABCX, ABCDX, ABCWX, ADWYZ, AWXYZ, ALL) and the "negative" class of all conditions without A (C-high, W-high, B, C, D, W, X, Y, Z, BC, DW, WX, WY, WZ, XY, XZ, YZ, BCX, BDW, DXY, WXY, WYZ, BCWX, BDWX, DWYZ, WXYZ, BCWXZ), not including paraffin oil. We then

used ROC analysis to evaluate each responding cell as a classifier for the presence of each of the single odors it responded to (i.e., if a cell responded to components A and B, it was evaluated as classifier for A, and again for B), using the total number of spikes produced in the response window across all trials of a given odor as the classifier's score for the odor. The top panel of Figure 1-5E shows some sample ROC curves, and the bottom panel shows the distribution of AUC scores for all cell-odor pairs in the two populations. We repeated the analysis for longer response windows of 1.4 s and 2 s, and for single trials, and observed no significant differences.

1.6.7.9 Multi-Stage Adaptive Lasso

Lasso regression (Tibshirani, 1996) regularizes ordinary least squares regression by constraining the sum of the absolute value of the weights. Its advantage over ridge regression is that it can shrink weights identically to zero, while ridge-regression only scales the weights. The adaptive lasso improves on the lasso by penalizing large coefficients less than small ones, and the multi-stage adaptive lasso tries to reduce the probability of including spurious regressors by running the adaptive lasso repeatedly, each time using the previous run's best estimate of the fit coefficients (typically estimated using cross validation), until the form of the model has stabilized. Our implementation of the adaptive lasso was based heavily on the *lasso* function included in MATLAB's Statistics toolbox, and allowed us to provide weights for the coefficients and to limit the weights to be strictly positive by removing them from the active set when they became negative.

1.6.7.10 Constructing Single KC Responses from PNs

To regress a KC's responses on the PNs, we first computed its mean firing rate in 10 consecutive 100 ms time bins start at 0.1 seconds following odor onset. We concatenated these responses to form a 440 = 44 odors x 10 bins / odor response vector for the KC. We performed the same procedure for the PNs, and then ran the multi-stage adaptive lasso algorithm for 10 stages or until the number of regressors in the model didn't change, whichever came first. We then computed the fraction of the variance left unexplained by the fit (SSE/SST) and the correlation coefficient between the response and the fit, since due to regularization the residuals are not necessarily uncorrelated with the fit and SSE does not determine the correlation coefficient, as it does in ordinary least-squares regression.

Shuffled populations of cells were created from un-shuffled ones by reassigning the odor responses 'whole' among the cells, i.e., without perturbing their temporal structure neither within a trial nor across trials. Thus a shuffled cell's responses to a given odor are some other cell's responses to a different odor.

We measured how well the PNs were suited to reconstructing the *recorded KC* responses by producing 250 shuffled KC populations using the procedure above, reconstructing them using the un-shuffled PN population, and reported the SSE/SST and correlation coefficient distributions for the resulting over all cells and all shuffles in Figure 1-2H. We also

measured how well suited the *recorded PNs* in particular were to reconstructing the KC responses by producing 1000 shuffled PN populations and using them to reconstruct the un-shuffled KCs, while constrained to using the same number of shuffled PNs in each reconstruction as were required when using the un-shuffled PNs. The reconstruction metrics were computed for all KCs and all shuffles and the distributions summarized in Figure 1-2H.

1.6.7.11 Constructing PN Population Trajectories from KC Responses

We searched for a fixed basis in which to describe the response of the PN population in any given time bin as a linear combination of basis elements, with the coefficients of combination determined by the KCs (plus a constant term). We can write this as looking for a solution to U = ZV, where U is a # PNs x # time bins matrix whose columns contain the PN population responses for each of the 440 time bins formed from 10 response bins for each of 44 odors, the # KCs x time bins matrix V is the corresponding matrix for the KC responses, and Z is the # PNs x # KCs matrix whose columns are the basis vectors that each KC corresponds to. While in general we won't find an exact solution to the equation above, we can 'settle' for the least squares solution found using the Moore-Penrose pseudo-inverse: $Z = UV'(VV')^{-1}$. However, the resulting solution will likely contain many spurious small but non-zero weights, making it harder to interpret the resulting basis matrix. To remedy this situation, let Y(Z) = ZV be the estimated PN reconstructions for a given basis matrix Z. We want to find the matrix Z such that the sum of the squared error (SSE)

between U and Y(Z) is reduced. Clearly the SSE will be left unchanged if we instead consider U' and Y(Z)' = V'Z'. But now the columns of U' are just the responses of single PNs over the 440 bins, the columns of V' are the corresponding KC responses, and the columns of Z' are the reconstruction weights for each of the PNs. Since there is no requisite dependency between the column of Z', they can be treated independently, and we can minimize the SSE by setting these columns to minimize the reconstruction error of each PN using the KCs. Hence we can apply the multi-stage adaptive lasso algorithm we used to construct KCs from PNs in reverse (and without the constraint on strictly positive weights), to find the columns of Z' one at a time, and transpose the result to get the desired weight basis matrix. Having done so, we computed the average value of the unexplained variance in explaining the PN population response at each of the 10 bins in a given odor response, as well as the corresponding correlation coefficients between the PN population vector and the fits. We plotted the distribution over the 44 odors in Figure 1-2K.

Just as when reconstructing single KCs, we then measured how well matched the KCs were to the PNs. First, we produced 100 shuffled KC populations and for each, computed fits to the un-shuffled PN responses, computed the fit metrics as before, and plotted them in Figure 1-2K. As described above, our fits were computed by reconstructing the single PN responses, and we constrained the reconstructions to require the same number of weights as required by the unshuffled KCs. We also used the unshuffled KCs to reconstruct each of

the odor trajectories in 50 shuffled PN populations, computed the fit metrics as before and plotted them in Figure 1-2K.

1.6.7.12 Mean KC Spike Latency

To quantify the preservation of temporal structure across the KC population to different odor conditions, we computed for each KC a measure of mean spike latency. For each KC, PSTHs were computed using a 20 ms Gaussian smoothing kernel, averaged across 7 trials and baseline subtracted. Mean latency was defined as the peak of the PSTH.

1.6.7.13 Population Decoding

To estimate the information carried by PN and KC ensembles about odor component and identity in single trials, we used a decoding-based approach (Hung et al., 2005; Meyers et al., 2008). A linear classifier was provided with spike counts in 4 consecutive bins (25 ms each bin) across all PNs (175) and KCs (209) and computed over 2 s shortly after odor onset. The classifier consisted of a weighted sum of PN or KC inputs. The weights were estimated using regularized least squares regression (Rifkin et al., 2003). This approach can be thought of as multiple linear regression with a constant term. Multiple linear regression cannot determine the weights unambiguously if the sample matrix is ill conditioned, which is often the case with few trials or few spikes (as with KCs). The formulation for RLSC is below:

$$w = X^T(XX^T + \lambda I)^{-1}y$$

The T x n matrix X contains spike counts across all cells; each row is one trial (T contains $T/2$ positive trials and $T/2$ negative trials); each column represents spike counts in one cell (n columns represents n cells). w is the n x 1 weight vector, a unique weight for each cell. y is the T x 1 vector of class labels (+1 and -1). I is the T x T identity matrix, and λ is the scalar regularizer. The larger λ is, the more constraints are placed on the solution; the smaller λ is, the closer the solution is to multiple linear regression. Even a small value of the regularizer punishes unrealistically large weights and ensures stability. Regularization becomes particularly important when the number of input variables (neurons) outnumbers the number of training examples, as was the case here. There usually is an optimal value for λ. We tried values of 0.01, 0.1 and 1 but observed no significant difference. Therefore λ was kept constant at 1 throughout. The number of trials in each class during training was always kept the same to avoid decoding bias. Where the numbers of positive and negative trials were different, we repeated k (20 or 50 bootstraps) random sampling to equalize the number of positive and negative trials. The decoding accuracy of the classifier was estimated using leave-one-out cross validation for all training samples available.

1.6.7.13.1 Decoding Odor Identity

To decode odor identity (i.e., which of the 44 odors had been presented), we used all-*vs*-all multiclass decoding, one time bin at a time. The number of spikes was counted in each 100 ms time bin sampled at 25 ms intervals with data from each time bin being classified

independently, leading to a slight temporal smoothing. We built 44x43 binary classifiers (e.g., A vs. Ahigh, A vs. B, A vs. C, A vs. D, A vs. AB, … Z vs. A, Z vs. AB, etc.) for every time bin and every trial using all trials minus one; the tested trial was classified using all binary classifiers and assigned the class with the maximum votes across classifiers. In cases where there was a tie, the tested trial was assigned randomly to one of the leading vote getters. Chance performance was 1/44 or 2.27%.

1.6.7.13.2 Decoding Odor Category

To decode odor component or category information, we built 8 different classifiers (A *vs* not-A, B *vs* not-B, C *vs* not-C, D *vs* not-D, W *vs* not-W, X *vs* not-X, Y *vs* not-Y, Z *vs* not-Z), one for each odor component. We selected the odors for each classification task by attempting to satisfy the following two constraints: (1) that the two classes differ only by the odor component (e.g., A, AB, ABC, AC vs. B, C, BC) since the difference between e.g. ABC and WXY is more than just the presence of A, and (2) that the two classes have approximately the same number of *n*-level mixtures for each value of *n*, since there is a positive correlation between the number of KCs activated and the number of odor components in the mixture *n* (see Figure S1-6A). The resulting class partitions for each classifier are listed in the supplementary text. We performed *k* bootstraps (20 or 50) of all available trials, where in each bootstrap, the number of positive and negative trials were kept equal. Because this classification is binary, chance performance was at 50%.

1.6.7.13.3 Decoding Odor Generalization

The procedure for decoding odor generalization was exactly the same as for decoding odor category, except that during training, the data for one entire odor, instead of one single trial, was withheld. The classifier was then tested on the withheld odor and its performance, defined as the fraction of 7 withheld trials correctly classified, was recorded. This procedure was repeated for all odors in the positive and negative class for the category, and the average over all these odors (and all bootstrap runs when positive and negative groups differed in size) was reported as the generalization performance for the category.

1.7 Supplementary Text

1.7.1 KC Search

Unlike PNs, KCs respond very sparsely to odors and their individual baseline activity is ~ 1 spike every 30 s on average (Perez-Orive et al., 2002). Hence, significant effort was made to find KCs that responded to some at least of the odors in our panel prior to initializing an experiment. Due to the large number of conditions in our experiments, we did not (nor did we wish to) pre-test all 44 stimuli. Rather, we searched for responsive KCs by presenting the eight monomolecular odors; we selected a recording position from which some spikes could be recorded in response to any one of these stimuli. Due to the rarity of KC spikes, KC-spike cluster models were defined using all trials (usually ~ 50 conditions, 7 trials each, 14 s per trial). The condition in the middle of the set was used to calculate the noise covariance matrix (Pouzat et al., 2002). The threshold was set typically at 4–5 times each channel's signal SD. The model generated by this method was refined using criteria identical to those used with the PN data. Stability over the course of the experiment was assessed after sorting and was based on a stable baseline firing-rate over the course of the experiment.

Finding responding KCs always necessitated an active search. Indeed, most experiments yielded not a single KC responding to 8-single odor set. Because KC recordings started with an active search for cells responding to our stimuli, our KC dataset is biased towards

cells selective for one of those primary odors. Hence, the true response-probabilities of KCs are even lower than we presently estimate. Based on the fraction of recordings in which KCs responding to our 8 single odors were found, the fractions of responding cells we measured in each 50 ms bin (0.5–1%; 5% maximum of our recorded set for one large mixture when integrated over the entire response, Figure S1-6) was likely overestimated 10–20-fold. For example, the 5% maximum response measured for an 8-mixture would in reality correspond to a total of 125–250 KCs (of 50,000) per MB.

Yet, not all the KCs analyzed here were recorded because they had been selected during the initial search process: when some signal deemed hopeful had been detected on at least a couple of channels of one tetrode, recording started. Upon *post hoc* analysis, several KCs were identified that happened to respond to those stimuli, or in some cases, that appeared to recognize a particular mixture (e.g., 'WX') delivered alone or in a mixture (as in 'AWXYZ'). Such KCs often responded to neither of the components alone. For them, the simple mixture appeared to be a basis vector. Our database with such KCs is too small, however, to propose statistical estimates of their frequency.

1.7.2 Bayesian Model Selection

To characterize the spread of PN odor representations as pure citral was morphed to pure octanol, we fit the correlation distances from the non-pure odors to pure-citral as a function of either the base-10 logarithm of the concentration of citral to octanol, or the fraction of

octanol in the mixture, using four different models: a constant model, a linear model, a one-step model, and a two-step model, and used Bayesian model selection (MacKay, 2003) to select between them.

Given observed data D, Bayes' rule can be applied to assign a posterior probability P(M|D) to each model M given the data by combining the likelihood of the model P(D|M) with its prior P(M): P(M|D) = P(D|M)P(M)/P(D). When comparing several models $M_1,M_2,\ldots,M_n$ of data D, Bayesian model selection selects the model with the highest posterior probability. If we assume that all models have equal prior probability (as we do), then since P(D) is the same for all models, model selection is reduced to selecting the model with the highest likelihood P(D|M). The likelihood can be computed by integrating over the parameters of the model:

$$P(D \mid M) = \int_{\Theta} P(D,\theta \mid M)\, d\theta = \int_{\Theta} P(D \mid \theta, M) P(\theta \mid M)\, d\theta,$$

where θ is a place-holder for the model parameters. P(D|θ,M) is the likelihood of the parameters θ, and P(θ|M) is their prior.

All of the models we used assume that the data points are normally distributed independently with variance v around some model-dependent mean function of the

independent variable x. The form of each model, along with the prior on its parameters and the integration method used to compute the posterior, are tabulated below:

Model	**Parameters**	**Form**	**Prior^{-1}**	**SS$_T$**	**Integration Method**
Linear	$v \in (0,\infty)$ $m \in (-\infty,\infty)$ $b \in [y_l, y_u]$	$y[i] \sim N(mx[i]+b, v)$	$v(\Delta y)\pi(1+m^2)$	$\sum_{i=1}^{K}(y[i]-mx[i]-b)^2$	Laplace's Method
Constant	$v \in (0,\infty)$ $y_1 \in [y_l, y_u]$	$y[i] \sim N(y_1, v)$	$v\Delta y$	$\sum_{i=1}^{K}(y[i]-y_1)^2$	Laplace's Method
One-step	$v \in (0,\infty)$ $y_1 \in [y_l, y_u]$ $\theta \in [x_l, x_u]$	$y[i] \sim \begin{cases} N(y_1, v) & x[i] < \theta \\ N(y_2, v) & x[i] \geq \theta \end{cases}$	$v(\Delta y)^2\Delta x$	$\sum_{i=1}^{K_1}(y[i]-y_1)^2$ $+\sum_{i=K_1}^{K}(y[i]-y_2)^2$	Laplace's Method over y_1, y_2 and v (jointly) to yield $Q(\theta)$, followed by direct integration of $Q(\theta)$ over θ.
Two-step	$v \in (0,\infty)$ $y_1 \in [y_l, y_u]$ $y_2 \in [y_l, y_u]$ $y_3 \in [y_l, y_u]$ $\theta_1 \in [x_l, x_u]$ $\theta_2 \in (\theta_1, x_u]$	$y[i] \sim \begin{cases} N(y_1, v) & x[i] < \theta_1 \\ N(y_2, v) & \theta_1 \leq x[i] < \theta_2 \\ N(y_3, v) & x[i] > \theta_2 \end{cases}$	$v(\Delta y)^3\Delta x(x_u-\theta_1)$	$\sum_{i=1}^{K_1}(y[i]-y_1)^2$ $+\sum_{i=K_1}^{K_2}(y[i]-y_2)^2$ $+\sum_{i=K_2}^{K}(y[i]-y_3)^2$	Laplace's Method over y_1, y_2, y_3 and v (jointly) to yield $Q(\theta_1,\theta_2)$, followed by Monte Carlo integration of $Q(\theta_1,\theta_2)$ over θ_1 and θ_2.

where the data points are (x[i],y[i]) for I = 1,2,…, K, xl and xu are the lower and upper limits of the prior range on x and similarly for yl and yu, Δx = xu – xl is the width of the prior range of x values and similarly for Δy, v is the variance, m and b are the slope and offset of the Linear model, y_1, y_2, and y_3 are the various constant values of the constant and step models, θ_1 and θ_2 are threshold values defining step boundaries, and K_1 and K_2 are the number of elements constituting each step. Expressions for priors are derived by assuming that the parameters are selected independently and uniformly over their respective ranges, except for v, for which as a scale parameter the "improper" prior 1/v better reflects our lack of prior knowledge about it (Sivia and Skilling, 2006), and θ_2, which is selected uniformly over the range (θ_1, xu] once θ_1 has been selected, to enforce the requirement that the boundary of the second step in the two step model must be ahead of the first. The form of

the models and the assumed independence of the data points yield likelihoods for each model of $(2\pi v)^{-K/2} \exp(-SS_T/2v)$. For the constant and linear models, the integral in the likelihood equation above is computed using Laplace' method, in which an integral is approximated by the volume of a multivariate normal distribution fit to the integrand at its peak (MacKay, 2003). The thresholds in the one- and two-step models make their integrands only piecewise continuous, so the integration is performed by first applying Laplace's method to the pieces and then integrating the result over the pieces either directly in the case of the one-step model, or using Monte Carlo integration (with 10,000 randomly selected points in the integration range) in the case of the two-step model.

1.7.3 Odor Metrics

There is no obvious metric for measuring similarity between odors. Hence for our complex mixtures, which we represented as 8-bit binary vectors whose elements indicate the presence of each of the 8 single odor components, we tried three different metrics that satisfy two intuitive criteria: (1) odors that don't share any components should be maximally dissimilar, and (2) for a fixed amount of overlap between two odors, the distance should increase as the number of components in the odors increases. The metric we used in the main text is the Jaccard distance (Deza and Deza, 2009), defined for two binary vectors x and y as:

$$d_J(x,y) = 1 - \frac{|x \wedge y|}{|x \vee y|},$$

where ∧ is bit-wise logical AND, ∨ is bit-wise logical OR, and |x| is the L_1 norm (the number of 1's in the binary vector x). When comparing two odor vectors, this distance is 1 minus the ratio of the number of odor components the two odors share to the total number of components present pooled over the two odors. This metric satisfies both of our criteria above, since (a) if two odors don't share any components, then the numerator of the ratio will be zero, and the distance assigned will be 1 (maximal), and (b) if the numerator is held constant while the number of components in either odor is increased, the denominator in the ratio will increase, reducing the ratio and increasing the distance.

The cosine distance also satisfies our criteria:

$$d_C(x,y) = 1 - \frac{x \bullet y}{|x||y|}.$$

Here the norms |x| are Euclidean norms. Again, if the odors don't share components then their dot product will be zero and the distance will be maximal, while if the dot product is kept constant but the number of components in either odor is increased, then its norm will increase, the ratio will decrease, and the distance will increase.

Finally, we tried the Braun-Blanquet (Deza and Deza, 2009) distance,

$$d_{BB}(x,y) = 1 - \frac{x \bullet y}{\max(|x|,|y|)},$$

where the norms here are the same as for the Jaccard distance. In terms of odors, this distance is 1 minus the ratio of the total number of components two odors share to the maximum number of components present in either odor. It satisfies our first criterion as before, and partially satisfies our second criterion because if the size of the more complex mixture is increased, the distance between the two odors will increase.

1.7.4 Balanced Odor Classes for Decoding Categorization and Generalization

For categorization and generalization, the following odor groupings were used to ensure as much as possible that for each odor X vs. not-X classification task, (a) the number of n-level mixtures was the same in both the X and not-X categories, and (b) the odors in the two groups differed only by the presence or absence of the component X.

Odor A vs. not-A

A: A (4x), A, AB, AC, AD, AX, ABC, ACD, AXZ, ABCD, ABCX, ABCDX, ABCWX, ADWYZ, AWXYZ

A': C (4x), B, C, D, X, BC, DW, XZ, BCX, BDW, DXY, WXY, WYZ, BCWX, BDWX, DWYZ, WXYZ, BCWXZ

Odor B vs. not-B

B: B, AB, BC, ABC, BCX, BDW, ABCD, ABCX, BCWX, BDWX, ABCDX, ABCWX, BCWXZ

B': A (4x), A, C (4x), C, X, AC, AD, AX, DW, WX, ACD, AXZ, DXY, WXY, DWYZ, WXYZ, ADWYZ, AWXYZ

Odor C vs. not-C

C: C (4x), C, AC, BC, ABC, ACD, BCX, ABCD, ABCX, BCWX, ABCDX, ABCWX, BCWXZ

C': A (4x), W (4x), A, B, X, AB, AD, WX, AXZ, BDW, DXY, BDWX, DWYZ, WXYZ, ADWYZ, AWXYZ

Odor D vs. not-D

D: D, AD, DW, ACD, BDW, DXY, ABCD, BDWX, DWYZ, ABCDX, ADWYZ

D': W (4x), A, B, W, AC, XY, ABC, WXY, WYZ, ABCX, BCWX, WXYZ, ABCWX, AWXYZ, BCWXZ

Odor W vs. not-W

W: W (4x), W, DW, WX, WY, WZ, BDW, WXY, WYZ, BCWX, BDWX, DWYZ, WXYZ, ABCWX, BCWXZ

W': A (4x), C (4x), B, D, X, Y, Z, BC, XY, XZ, YZ, ABC, ACD, AXZ, BCX, DXY, ABCD, ABCX, ABCDX

Odor X vs. not-X

X: X, AX, WX, XY, XZ, AXZ, BCX, DXY, WXY, ABCX, BCWX, BDWX, WXYZ, ABCDX, AWXYZ

X': A (4x), C (4x), W (4x), A, D, W, Y, Z, BC, WY, WZ, YZ, ABC, ACD, BDW, WYZ, ABCD, DWYZ, ADWYZ

Odor Y vs. not-Y

Y: Y, WY, XY, YZ, DXY, WXY, WYZ, DWYZ, WXYZ, ADWYZ, AWXYZ

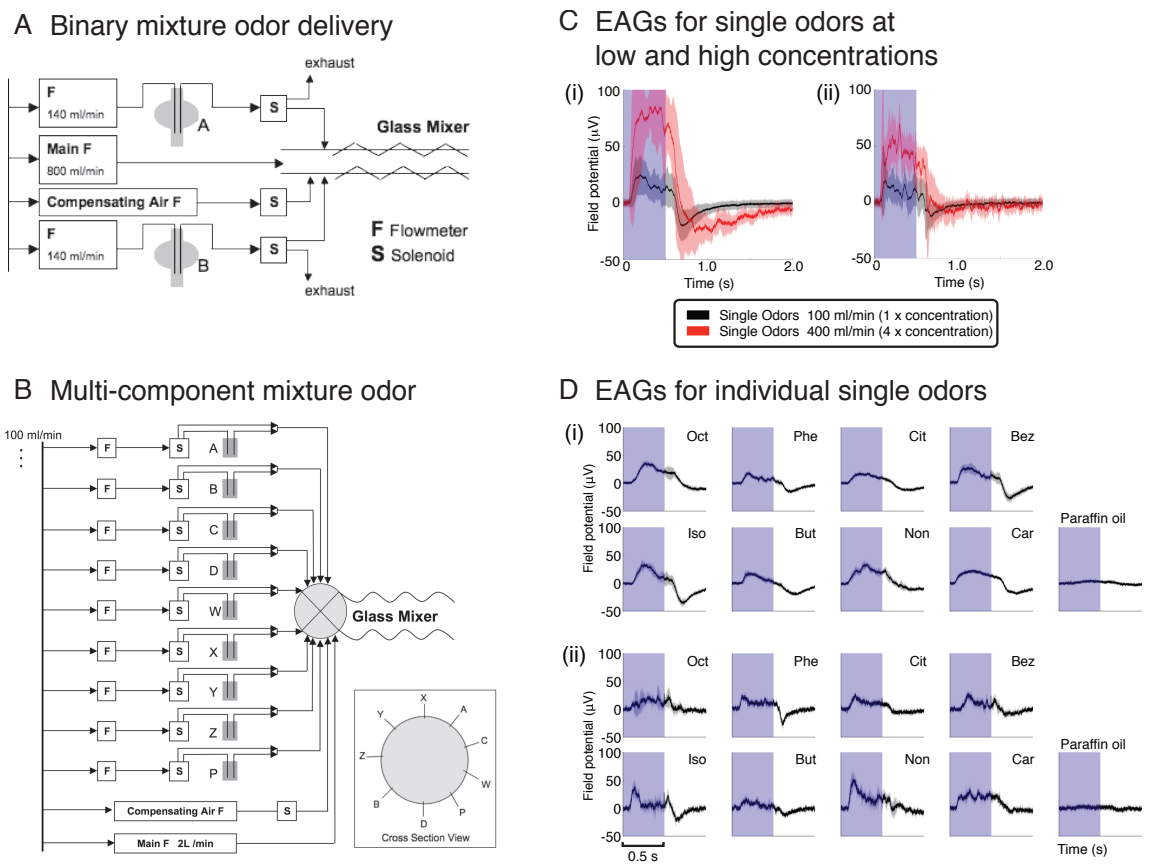
Y': D, W, X, Z, DW, WX, WZ, XZ, AXZ, BDW, BCWX, BDWX, ABCWX, BCWXZ

Odor Z vs. not-Z

Z: Z, WZ, XZ, YZ, AXZ, WYZ, DWYZ, WXYZ, ADWYZ, AWXYZ, BCWXZ

Z': W, X, Y, AX, DW, XY, WY, BDW, WXY, DXY, BCWX, BDWX, ABCDX, ABCWX

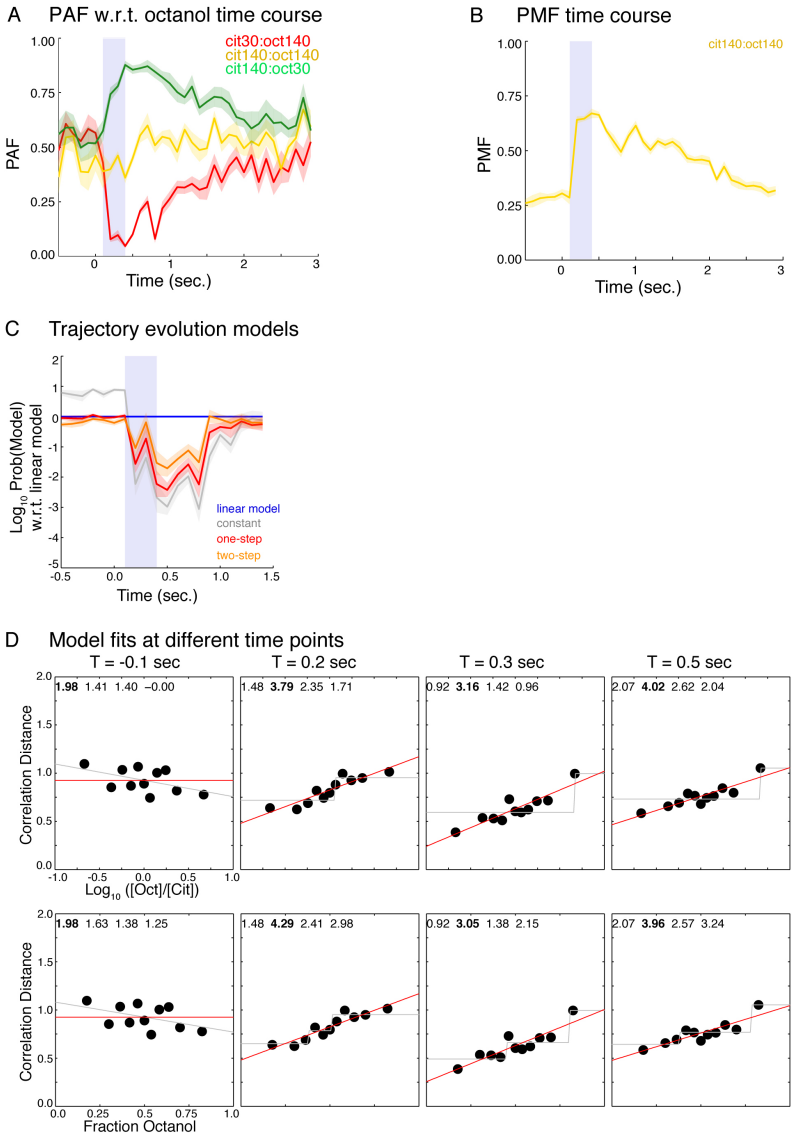
1.8 Supplementary Figures



Supplementary Figure 1-1: Olfactometer setup and EAGs.

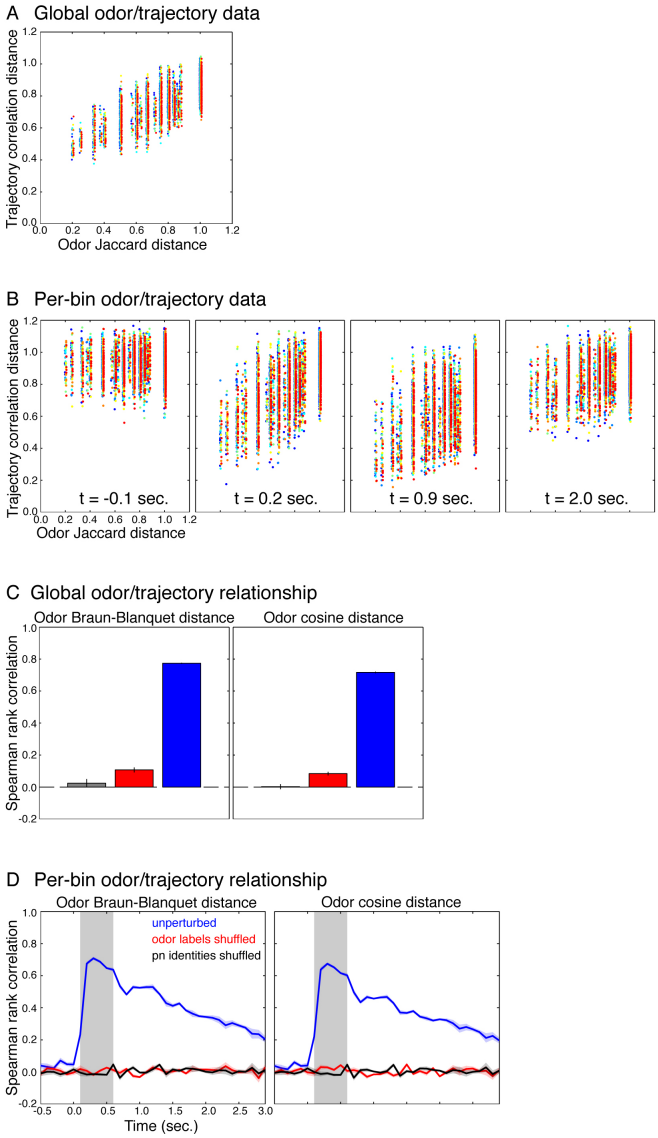
(A, **B)** An olfactometer was constructed using small diameter (1/32" inner diameter) Teflon tubing and compression fittings to minimize dead space and delay times. Flow rates of independent air streams were independently controlled by mass flow controllers (range 20–200 ml/min). By mixing odorized air streams with defined flow rates, different mixture ratios were achieved. **(A)** Binary mixture experiments; **(B)** Multi-component mixture experiments. **(C)** Comparison of EAG of single odor concentration (8 single odors: 1-Octanol,Phenetole, Citral, Benzaldehyde, Isoamyl Acetate, 2,3-Butanedione, 2-Nonanone, L-Carvone; 5 trials each; 100 ml/min) in black, with EAG of odors presented at 4 times the single odor concentration (3 single odors: 1-Octanol, Citral and Isoamyl Acetate; 5 trials each; 400 ml/min) in red. Shaded region indicates 1 standard deviation above and below the mean. As shown here, single odors were calibrated to elicit minimal EAG response, at the lower end of its dynamic range. EAG shown for two different antennae **(i** and **ii)**. **(D)** Comparison of EAG of 8 different single odors, as in (B), and paraffin oil. Concentrations were calibrated to evoke as similar an EAG as possible across single odors, so that no one single odor is dominant during odor mixture conditions. Interestingly, paraffin oil evokes no EAG response in isolated antenna, but can evoke PN and KC responses (see Figure S1-5C for KC population response). Shaded region indicates 1 SD above and below the mean (5 trials each). EAG shown for two different antennae **(i** and **ii)**.

Supplementary Figure 1-2: No supplementary figure to accompany Figure 1-2.



Supplementary Figure 1-3: Additional details for PN binary mixtures metrics.
(A) Time course of the mean (traces) ± S.E.M. (bands) over trials of projection angle fraction with respect to octanol for adjacent 100 ms bins starting at -0.5 seconds. The odor onset is at 0.0 seconds, arrives at the antenna 0.1 seconds later, and is presented for 0.3 seconds (blue patch). The values are 1 minus the values for the PAF with respect to citral (Figure 3G), indicating that the population vectors for the mixtures shown are between those of the pure components. **(B)** Time course of the mean (traces) ± S.E.M. (bands) over trials of the projection magnitude fraction (PMF, defined as the length of the projection of the mixture vector on the plane spanned by the simultaneous vectors for the two components, divided by length of the mixture vector) for the 140:140 binary mixture for adjacent 100 ms bins starting at -0.5 seconds. Odor presentation parameters are as in (A). **(C)** Mean (traces) ± S.E.M. (bands) over trials of the relative $\log_{10}$ posterior probability of each of the four models relative to the linear model as descriptions of the evolution of the correlation distance of responses to non-pure mixtures from the response to citral,

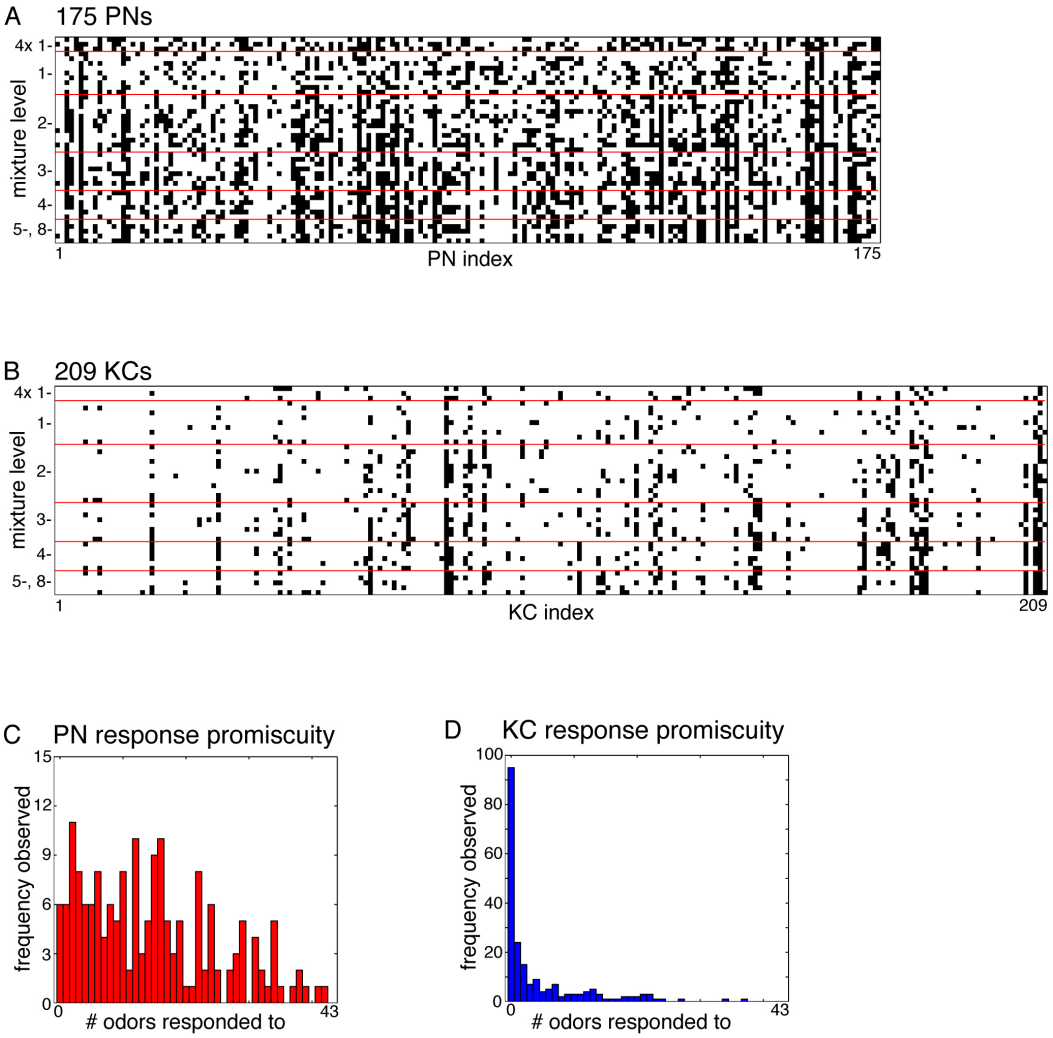
as a function of the fraction of octanol in the mixture. The probabilities are computed for adjacent 100 ms bins, starting at -0.5 seconds. Odor presentation parameters are as in (A). The linear model is still the best description over the response window, although now the two-step model has a larger posterior than when using the $\log_{10}$ ratio of octanol to citral as the independent variable due in part to the smaller prior range on the independent variable (~ 1 vs. ~ 2). **(D)** Representative samples of the raw data and the fits used to produce Figure 3I and (C). The data are for the responses in the fourth trial for the time bins starting at the times indicated above each plot. The top row uses the $\log_{10}$ ratio of octanol to citral as the independent variable (as in Figure 3I), the bottom row uses the fraction of octanol (as in (C)). The two sets of plots are very similar because the independent variables are approximately linearly related over the range of concentrations tested. The best and second best fits are shown in red and gray, respectively. The $\log_{10}$ posterior probabilities of each model (minus a data-dependent constant term common to all the models) are indicated at the top left of each plot, in order of constant, linear, one-step, two-step model, with the value for the best model in bold.



Supplementary Figure 1-4: Additional details for complex mixtures metrics.

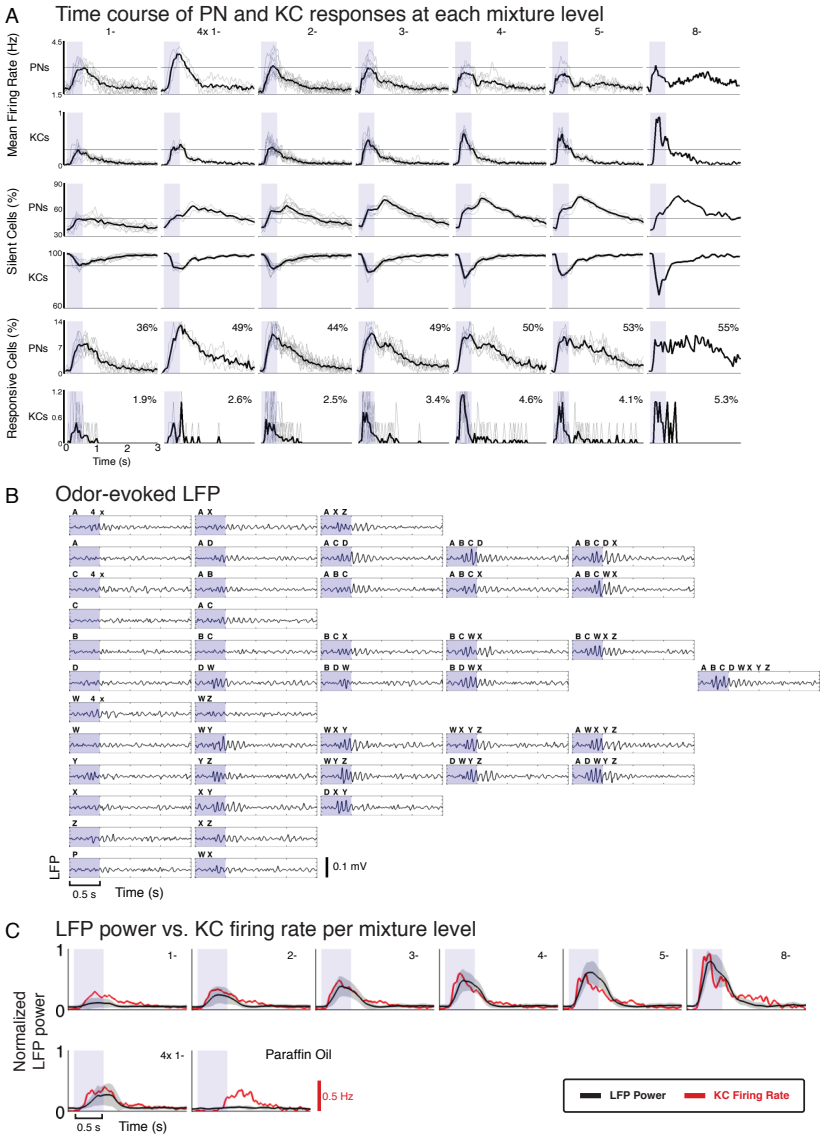
(A) Raw data showing the relationship between odor Jaccard distance and the global trajectory correlation distance for each of trials 1–7 (rainbow colored from blue to red) in the response window. **(B)** Raw data for the per-bin relationship between odor Jaccard distance and trajectory correlation distance, at the time points indicated (odor onset is at 0.0 seconds, reaches the antenna at 0.1 seconds, and remains on for 0.5 seconds). The trials are colored as in (A). **(C)** Spearman rank correlation between odor and global trajectory distances, using the Braun-Blanquet distance, and the cosine distance to measure distances between odors (see above). The mean over trials for the 1-second response window (t = 0.1–1.1 sec) is in blue, for 1-second baseline window (t = -1.1 to -0.1 sec) is in red, and 1 second response window but with the odor labels shuffled (differently for each trial and for the two panels) for the trajectory data is in gray. The vertical lines at the top of each bar indicate ± S.E.M. The results are qualitatively similar to those using the Jaccard distance between odors (Figure 4D): during the response window, trajectory distances increase with increasing odor distance, while at baseline, this is true to a much lesser extent. **(D)** Per-bin mean (blue traces) ± S.E.M. (blue bands) over trials of the Spearman rank correlation between odor distances measured using the Braun-Blanquet distance (left panel) or the cosine distance

(right panel) and the trajectory correlation distance between responses in 100 ms bins starting at t = -0.5 seconds. The data in red is computed the same way as the data in blue, but with the odor labels on the trajectories shuffled (differently for each time bin and for the two panels). The data in black is computed the same way as the data in blue, but with PN identities shuffled per odor and per bin (but fixed across trials for a given odor and bin). As in Figure 4F, immediately after odor onset there is a strong relationship between odor distances and trajectory distances, which persists for several seconds. The unperturbed traces are very similar in the two panels because Spearman rank correlation is sensitive to the rank-order of the data being correlated, not the actual numerical values.



Supplementary Figure 1-5: PN and KC responsivity.

The responsivity (see Methods) of PNs **(A)**, and KCs **(B)**, to each of the 43 odors (Paraffin oil is not included). Each column represents a single cell, and the rows are odors organized according to mixture level. **(C)** Distribution of the promiscuity values for PNs, where the promiscuity of a cell is the total number of odors it responded to (equivalent to summing the cell's column in (B)). **(D)** Same as (C), but for the KCs. KCs are much less responsive than PNs. For example, nearly half respond to none of the odors, while for the PNs only ~ 3% respond to no odors.



Supplementary Figure 1-6: Time courses of aggregate PN, KC, and LFP responses.
(A) All statistics were computed from 175 PNs and 209 KCs; each presented seven times with 44 different stimuli. Thin gray lines show averages for each odor, thick lines are averages across odors, for each mixture level. Blue bar is 500 ms odor pulse, 3 s shown. Horizontal lines in the top four rows indicate maximum or minimum value reached during the single component conditions.

Mean firing rate: Mean PN and KC firing rate as a function of the number *n* of odor components in the mixture. Firing rate was computed by convolving 10 ms binned spikes with a 20 ms width Gaussian filter. Mean firing rate for PNs remains approximately constant with increasing mixture level *n* (with a slight decrease). In comparison, KC mean firing rate clearly increases as a function of *n*. Interestingly, PN firing for single components at 4x concentration is higher than 1x, but no significant differences were detected in KC firing rate.

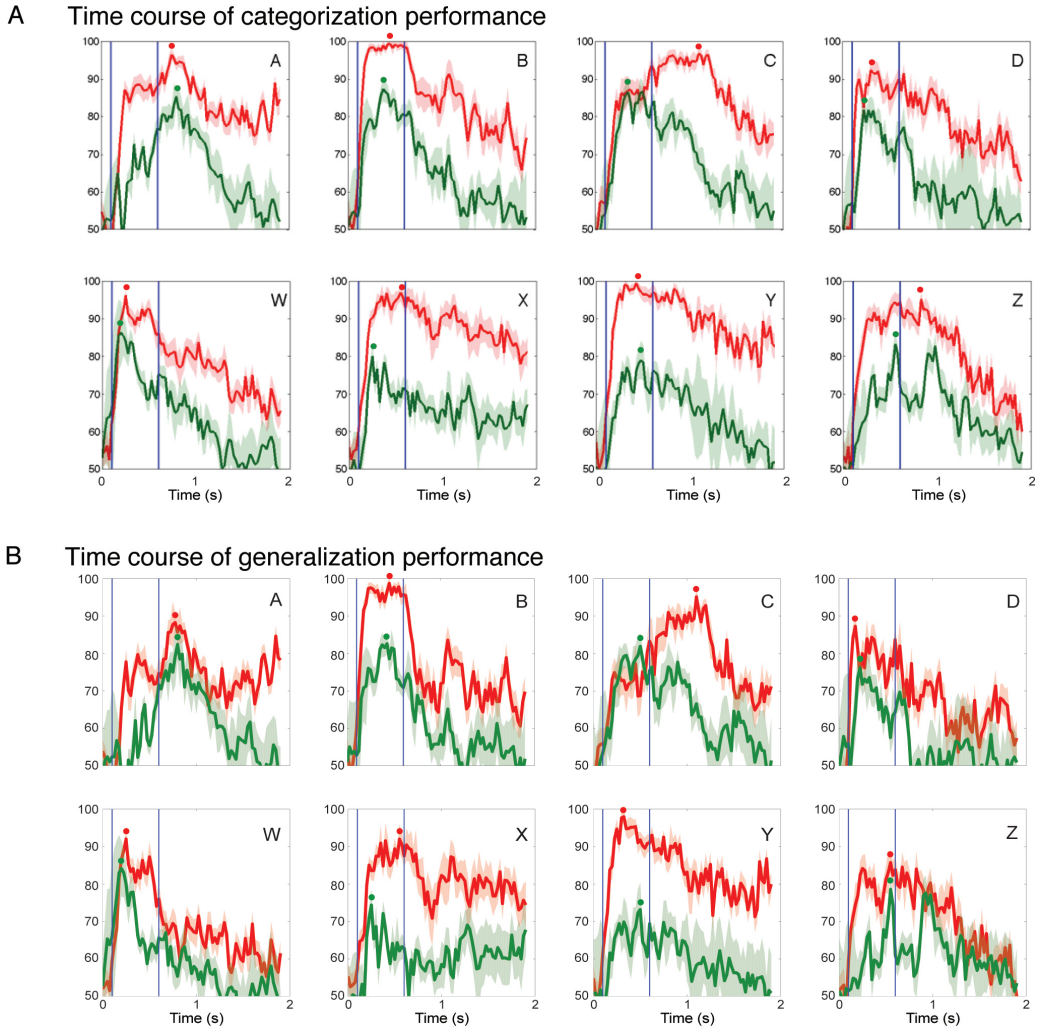
Percentage of silent cells for each time bin: A cell is defined as silent during a 100 ms time bin if it fired no spikes in that time bin in all 7 stimulus trials. Notice that percentage of silent PNs increases as a function of *n*, indicating increased inhibition from LNs, as a form of gain control on

the output of the PNs. The peak of the percentage of silent PNs is reached ~ 200–300 ms later than the peak of PN firing. Four part odor mixtures elicited greater inhibition than single components at comparable concentration, suggesting that the form of gain control observed here is specialized for mixtures. In comparison, at baseline all KCs are silent. The silent fraction dips to ~ 90% for single components and returns to near complete silence within 500 ms of odor onset, which we attribute to feedback gain mechanism of the GGN which keeps KCs sparse.

Percentage of responsive PN- and KC- odor pairs as a function of odor components in consecutive 50 ms time bins: Very few PNs and KCs were responsive at baseline by our measure (see response metric in Methods). However, shortly after odor onset, 7–14% of PNs are responsive, in comparison to 0.5–1% of KCs in any 50 ms time bin. The number to the upper right of each panel shows the cumulative proportion of responding cells over 3 s.

(B) Odor-evoked LFP. Larger odor mixtures (as components in mixture increases) elicit greater oscillatory power in the local field potential (LFP). Single odor components (at 1x concentration) elicit minimal (if any) power in LFP (left most column). Shown single trial (from 1 experiment) to all odor conditions. LFP was band-pass filtered, 5–35 Hz. Odor bar indicates time since odor onset; odors arrive at the antenna about 100 ms later.

(C) LFP power vs. mean KC firing rate at each mixture level. Mean normalized LFP power in 5–35 Hz band (mean over all odors and 6 locusts, 200 ms sliding window in 50 ms steps) in black line (SD in gray). Mean instantaneous KC firing rate (10 ms binned spikes convolved with 20 ms Gaussian) is superimposed in red. Increases in both KC firing rate and LFP power are well matched. Notice that increases in LFP power with *n* components in the mixture cannot be explained by concentration alone, as 4-part odor mixtures (4-) elicit still greater power than single odors at 4x the concentration (4x 1-) (the two are at equivalent concentration and elicit equivalent EAG responses, not shown). Interestingly, Paraffin oil does not elicit power in the LFP or EAG responses (see Figure S1-1), but does lead to increases in KC firing rate. Odor bar shifted by 100 ms to better align with odor arrival at the antenna.



Supplementary Figure 1-7: Time course of decoding accuracy for PNs and KCs.
In each panel, the mean (traces) and SD (bands) over 50 bootstraps of the mean categorization or generalization performance for each component is shown for PNs (red) and KCs (green). **(A) Time course of categorization performance.** Decoding accuracy of KCs follows faithfully that of PNs, the only exception is for odor component 'C', where the peak accuracy for KCs occurs ~ 500 ms before that of PNs. The timing of the peak accuracy varies across odor component categories, occurring within 100 ms of odor onset for 'W', but as late as ~ 200 ms after odor offset for 'A'. Thus KCs as a population could extract different stimulus features at different favored times. Dots indicate time of peak performance. Vertical blue lines indicate odor onset and offset (500 ms pulse). Chance performance at 50%. **(B) Time course of generalization performance.** Time course of decoding accuracy is very similar to that for categorization, but with lower performance due to the increased difficulty of the generalization task relative to categorization. Dots indicate time of peak performance. Vertical blue lines indicate odor onset and offset (500 ms pulse). Chance performance at 50%.

Supplementary Figure 1-8: No supplementary figure to accompany Figure 1-8.

Chapter 2: Full-Rank, Ultra-Sparse Odor Representations

2.1 Introduction

Sensory systems in both vertebrates and invertebrates often contain "fan-out" stage in which stimulus representations are distributed from a small population of neurons to a much bigger one. Although several possible explanations for this fan-out have been proposed, such as (Maass et al., 2002), its precise function remains unclear. Here we study an instance of this fan-out transformation between the antennal lobe and the mushroom body in the locust. We propose that the function of the locust olfactory system is to learn mappings between odors and valences, and that the observed fan-out between the AL and the MB greatly increases the number of odors whose valences can be learned. We show that

- A neurally plausible non-linear embedding is sufficient for capacity expansion,
- Increased capacity of the MB can be maintained while using known circuit properties to sparsify the representation to biologically observed levels,
- MB representations can be made robust to noise using Hebbian learning,

- Near Bayes-optimal readout of MB representations can be performed using neurally plausible linear readout.

Our work makes a number of predictions both about the architecture and learning rules at work in the system, and about the behavioral effects of anatomical and physiological perturbations of the system.

2.2 The Biological Circuit

Figure 2-1A shows parts of the locust olfactory circuit. It consists of 90,000 olfactory receptor neurons (ORNs) converging onto ~ 1000 spherical neuropilar structures called glomeruli (Laurent and Naraghi, 1994). The glomeruli are sampled by the ~ 1000 projection neurons (PNs) and about ~ 300 local interneurons (LNs, not shown) of the antennal lobe (AL) (Laurent and Naraghi, 1994; Leitch and Laurent, 1996). The antennal lobe is a highly recurrent circuit (Leitch and Laurent, 1996), part of whose function appears to be decorrelation/gain control of ORN inputs (Bhandawat et al., 2007; Mazor and Laurent, 2005). The PNs form the sole output of the AL and densely project to the ~ 50,000 Kenyon cells (KCs) of the mushroom body (MB) (Jortner et al., 2007; Laurent and Naraghi, 1994). In contrast to the promiscuous responses of PNs, KC odor responses are very sparse, due to intrinsic cellular properties (Perez-Orive et al., 2004; Perez-Orive et al., 2002) and to global feedback inhibition by the giant GABA-ergic neuron (GGN) (Papadopoulou et al., 2011). The KCs are read out by the ~ 100 inhibitory beta lobe neurons (bLNs) (MacLeod et al., 1998). bLN odor responses are dense and PN-like, and bLNs inhibit each other (Cassenaer and Laurent, 2012). bLN output is then sent on to the lateral horn and the central complex, where it ultimately influences the behavioral decisions of the animal.

2.3 Problem Formulation

Our aim was to explain the gross anatomical features of this circuit: (1) the ~ 50x increase in the number of cells from the AL to the MB, the (2) the ~ 500x decrease in number of cells from the MB to the bLN, (3) global feedback inhibition of the MB by the GGN, and (4) mutual inhibition of the bLNs.

We began by assuming that the fundamental problem of the olfactory system is to map odor inputs to valences. We made the simplifying assumption that each odor is mapped to a single input vector, and formulated the problem as finding a solution to the system of linear equations $\mathbf{Zw} = \mathbf{v}$, where $\mathbf{Z}$ is the N x M *encoding matrix*, where N is the number of odors the animal will encounter in the environment, and M is the number of sensory input channels. Each of its rows represents the inputs to the system (via, e.g., ORNs) due to a single odor. The vector $\mathbf{w}$ is the M x 1 vector of *readout weights*, and $\mathbf{v}$ is the N x 1 *valence vector*, delivered by the environment via neuromodulators such as dopamine and octopamine.

Under what conditions does the equation above have a solution, in general? Using a basic result of linear algebra, only if the *rank* of the encoding matrix equals the number of odors, N. This is because $\mathbf{Z}$ can be seen as a linear mapping from R^M to R^N, and the rank of the mapping is the dimension of its range. Since we want the animal to be able to learn arbitrary valence mappings, a necessary condition is

that the range of the mapping must equal the dimension of the target space. Another basic result of linear algebra is that the rank of an N x M dimensional matrix is less than or equal to the minimum of M and N. Thus the rank of **Z** is less than or equal to the lesser of the number of odors and the number of input channels. Hence, if the number of channels is ≥ than the number of odors, a solution to the equation can in principle be found (unless there is degeneracy in the matrix **Z**), for any valence vector **v**. On the other hand, if the rank is less than the number of odors, then no solution exists in general because there exist valence assignments outside of the range of the mapping. We will call a matrix '*full-rank*' if its rank is the maximum possible for its size. Hence the system above always has a solution if the number of input channels ≥ the number of odors, and the encoding matrix is full- rank.

This condition on the rank means that the animal can learn at most as many odors as it has input channels, which may be restrictive. One way the animal can expand its capacity is to use more complex readout functions. An equivalent method, which may be easier to implement biologically, is to first *nonlinearly* project the inputs to a higher dimensional space before performing the linear readout. Figure 2-1 demonstrates this in a toy example.

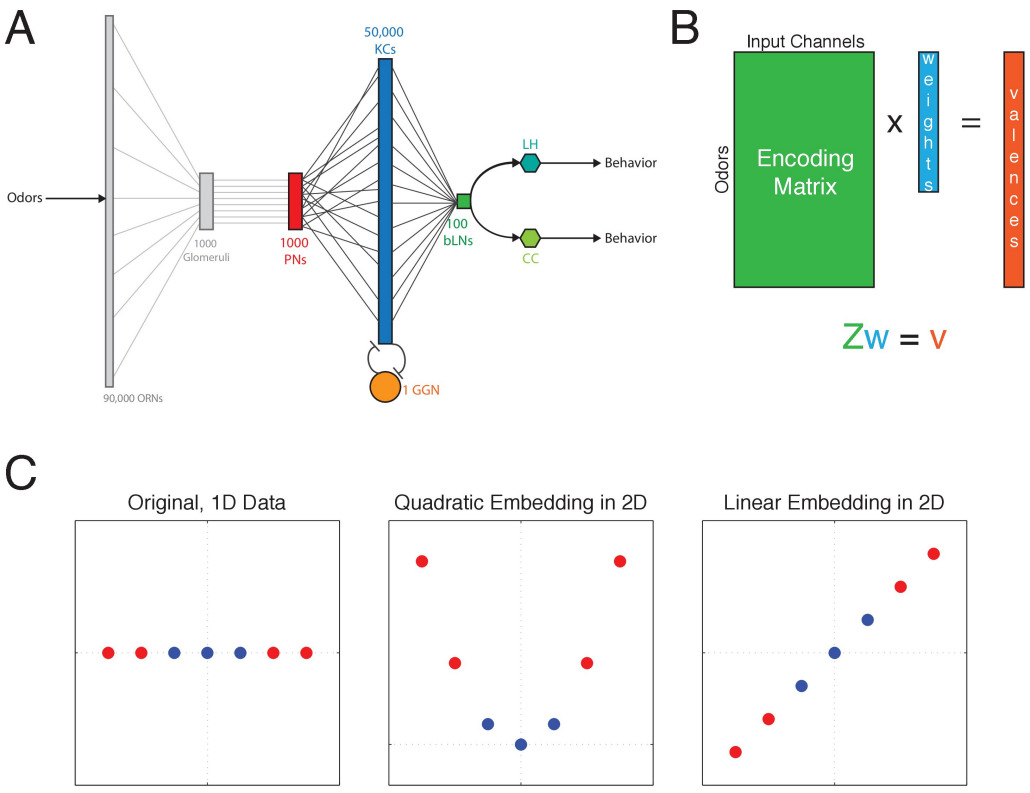


Figure 2-1 Biological circuit and problem formulation.

(A) Partial schematic of the locust olfactory system. Odors excite ~ 90,000 olfactory receptor neurons (ORNs). ORN axons converge on ~ 1000 spherical glomeruli, and are sampled by the 830 projection neurons (PNs) and ~ 300 local interneurons (LNs, not shown) of the antennal lobe (AL). The PNs form the only output of the AL and send their axons to the 50,000 Kenyon cells (KCs) of the mushroom body (MB), and to the lateral horn (connection not shown). The KCs in turn are readout by the ~ 100 beta-lobe neurons (bLNs), alpha-lobe neurons (not shown), and a single giant GABA-ergic neuron (GGN), which provides global feedback inhibition onto the MB. The bLNs send axons to the lateral horn and to the central complex, where olfactory information is presumably combined with information from other modalities to generate behavior. **(B)** Problem formulation: olfactory learning is simplified to a system of linear equations. We're looking for a set of weights to map odors into valences. **(C)** The blue and red points cannot be separated by a point or line in the original 1-D space, but can be separated if nonlinearly projected into 2D space. A linear embedding is not sufficient.

2.4 Full-Rank Ultra-Sparse Representations

Many nonlinear projections to higher-dimensional spaces can expand the capacity of the system. Because the present system is modeled on the projection from PNs to KCs, we chose to use a perceptron nonlinearity (binary response based on a thresholded linear summation), as a simple model of single-KC responses to PN input. Below we use simulations to demonstrate the performance of our system. In all simulations, the inputs and outputs are binary vectors. The input consists of N d-dimensional binary vectors, representing the single-oscillation-cycle responses of the d-PNs to N different odors, *after* decorrelation of receptor inputs. The output consists of N D-dimensional binary vectors, for $D \gg d$, representing the corresponding responses of the KCs.

We began by selecting each PN to KC weight randomly from the uniform distribution on the unit interval, and set the response threshold of each KC to $d/4$, its expected input current in response to an odor. Figure 2-2A shows the results for $N = 200$ odors, $d = 50$ PNs, $D = 400$ KCs. The rank of the PN encoding matrix is full, but is bounded by the number of PNs, which at 50 is much less than the number of odors. Hence arbitrary valence mappings cannot be learned with a linear readout. However, after nonlinear projection to the 400 dimensional KC space, the encoding matrix remains full rank, but this rank is now determined by the number

of odors. Hence, readout weights can be found for any arbitrary assignment of valences to odors.

Unfortunately, a KC encoding matrix such as in Figure 2-2A is unrealistically dense, at about ~ 50%. That is, a KC will respond to about ~ 50% of the odors, on average. The biological value is less than 1% (Perez-Orive et al., 2002). To increase the sparsity of the encoding matrix, we next increased the threshold of the perceptron nonlinearity. As is shown in the blue curve in Figure 2-2D, increasing the threshold does increase the sparseness of the KCs, as desired. But it also causes the rank to drop significantly, even with small increases in the threshold. Hence, although we can increase the encoding sparseness, we can no longer learn arbitrary mappings between odors and valences.

To understand the cause of the drop in rank, we examined the encoding matrix for a sparsity level at which the rank was greatly reduced. We observed prominent vertical streaks in the matrix, indicating that some odors were being represented by many KCs. Less prominent, but much more prevalent, were empty columns, i.e., odors that were not encoded by any KCs. To remedy this problem, we took inspiration from the presences of the GGN in the biological circuit, and augmented the encoding with a *k*-winners-take-all operation (*k*-WTA), so that the response to

any odor consisted of the *k* most active KCs. This procedure would by design solve the sparsity problem, fixing sparsity at a level of *k*/*D*. What was surprising was that rank of the resulting matrix remained full, even when the encoding was made very sparse (Figure 2-2D, red curve). The *k*-WTA operation allowed the encoding matrix to maintain full rank but with nearly two orders of magnitude higher sparseness.

To see if we could improve the results even further, we again looked at the KC encoding matrix at a level of sparseness at which the rank had dropped significantly. This time, we observed prominent horizontal streaks, and even more prevalent empty rows, indicating that some KCs were responding to many odors, while most KCs were responding to none. To remedy this situation we again appealed to biology and assumed that each KC has a fixed excitatory receptor pool, so that the sum of the weights to each KC must be the same. Implementing this 'weight normalization' procedure (a portion of an example encoding matrix is shown in Figure 2-2C) increased performance by nearly another order of magnitude, and into the biological range (Figure 2-2D, green curve).

We demonstrate successful readout explicitly in Figure 2-2E, in which we've used the perceptron learning rule (PLR, (Hertz et al., 1991)) to train the readout of the system on randomly assigned +1/-1 valence vectors (each trace is the performance

for one such assignment, the red trace is the mean). As the figure shows, the errors drop to zero after ~ 50 repetitions of each odor. Hence, a simple perceptron non-linearity coupled to weight normalization and a *k*-WTA transformation can produce encoding encodings that are not only very sparse, but also full-rank, so that arbitrary mappings of odors to valences can be learned by the system.

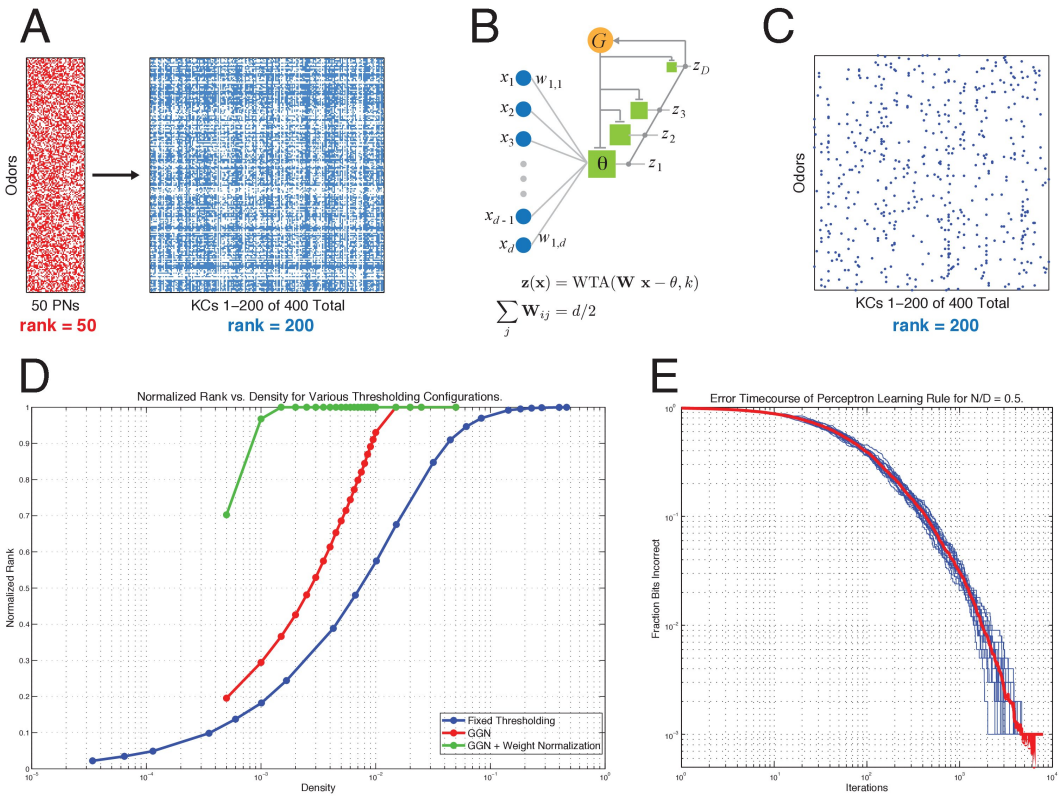


Figure 2-2 Full-rank, ultra-sparse odor representations.

(A) A representation of 200 odors by 50 PNs has a maximum rank of 50 (achieved in this case), but is not enough to learn arbitrary valence assignments to the 200 odors. A mapping to 400 dimensional KC space increases the rank to the 200 required, but the representation is much denser than biologically observed. **(B)** Nonlinear projection via *k*-WTA and weight normalization. Each KC computes a weighted sum of its inputs and the GGN allows only the *k* most excited KCs to respond. The sum of the input weights to every KC is kept constant. **(C)** A full rank, but also ultra-sparse, KC odor representation achieved using the projection in (B). **(D)** Performance of different nonlinear projections, measured as normalized rank vs representation density. If the nonlinear projection is the same fixed threshold for all KCs, then full rank can only be achieved at near full density (blue curve). Addition of the GGN allows full rank to be maintained at ~ an order of magnitude higher sparseness. The addition of weight normalization provides another order of magnitude increase in performance, allowing full-rank representations but with sparseness in the biological range. **(E)** The perceptron learning rule (PLR) can be used to learn readout weights from full-rank representations. The PLR was trained to learn a mapping to a random, binary valence vector, and the fraction of bits incorrect was recorded as learning progressed. The results for several trials (blue), the overall mean (red) indicate that the mapping is eventually learned perfectly.

2.5 Robustness to Input Noise via Hebbian Learning

In the examples above, we used fixed input vectors to represent the response of the antennal lobe to each odor after decorrelation. However, neural responses can be noisy. Hence we next measured the robustness of our encoding system to input noise, by flipping some bits at random and measuring the deviation of the output vector from the uncorrupted response. Figure 2-3A shows a typical response: A single bit flip on the input, representing a < 1% change in the in the input vector, often caused single bit flips on the output, which, due to the sparseness of the representation, corresponding to ~ 20% changes in the active bits of the output vector. Figure 2-3B plots the average normalized Hamming distance from the response to the noisy vs. clean inputs (the *smear radius*) as a function of the level of the level of input noise, demonstrating this *encoding fragility*, and showing that it exists across a wide range of output sparseness levels.

The encoding is fragile because there is no correspondence between input vectors and the weights from PNs to KCs. To create such a correspondence we implemented simple Hebbian learning at the PN-to-KC synapse. All *k*-KCs responding to an odor adjusted their input weights towards the input vector by a fixed, small amount, followed by weight normalization. Figure 2-3D demonstrates that this procedure led to a rapid drop in the smear radius. Figure 2-3E summarizes

the performance before and after learning, and shows that the smear radius has dropped to nearly zero, while the rank has remained full. Hence, full-rank, ultra-sparse representations can be learned that are also robust to input noise.

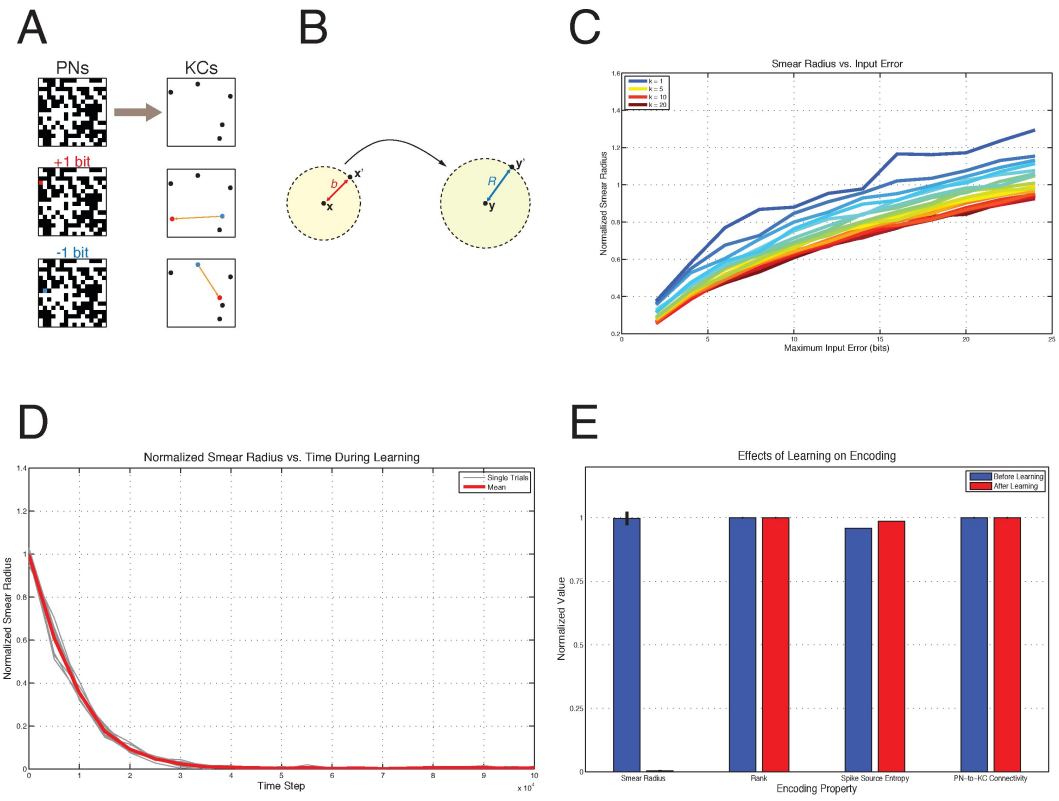


Figure 2-3 Noise robustness via Hebbian learning.

(A) Fragility of encoding. Changing 1 bit in the dense input representation often changes the output representation by at least one bit, a much larger relative change given the sparsity of the latter. **(B)** The *smear radius* measures the fragility of encoding by corrupting the input with noise and measuring the average distance of the resulting representation from that of the uncorrupted encoding. **(C)** Average smear radius as a function of input noise. A 200 dimensional input was corrupted by input noise and projected into a 2000 dimensional space and the distance from the uncorrupted encoding computed as a function of the maximum amount of noise added (in bits flipped). The smear radius normalized by the density of the encoding. Flipping a maximum of 10% of the inputs results in a change in the output representation equal to the density of the encoding. **(D)** Hebbian learning at PN-to-KC synapses makes the encoding noise robust. Plotted are single trials (gray) and the mean (red) of the normalized smear radius sampled at different points during learning, showing a rapid decrease to near zero. **(E)** Hebbian learning reduces smear radius to nearly zero, while still maintaining full rank representations, near maximum spike source entropy (KCs are active nearly equally often). But note that PN-to-KC connectivity is nearly 100%, about twice what is biologically observed.

2.6 Robustness to Pruning

The PN-to-KC connectivity in our system is nearly 100%, even after learning (red trace, (Figure 2-4A, red trace). This contradicts the known 50 ± 15% connectivity in the locust (Jortner et al., 2007). Hence we next measured the effect of pruning the weakest 50% of connections (Figure 2-4, purple trace), followed by weight normalization (Figure 2-4, orange trace). Computing the same response metrics as before (Figure 2-4B), we found that the encoding remains full rank, while its noise-robustness, as measured by smear radius, decreases only slightly. On the other hand, if 50% of the weights are pruned at random, noise-robustness decreases substantially. Experimentation with training rules that would automatically prune the lowest 50% of weights during learning remains a task for future work.

Figure 2-4 Robustness to pruning.

(A) Probability densities of PN-to-KC weight distribution before and after learning, after pruning and normalization. Before learning weights are essentially uniformly distributed over the unit interval. After learning prominent peaks are introduced into the distribution, and almost all weights are non-zero. The pruning procedure eliminates the weakest 50% of weights. The subsequent weight normalization at each KC smooths out the weight distribution. **(B)** Pruning only increases the smear radius slightly, while keeping rank full and spike-source-entropy high. Connectivity is now in the biologically observed range. The latter three properties can be maintained even if 50% of weights are pruned randomly (teal bars), but the smear radius increases drastically to near its value before learning.

2.7 Online Learning with Near Bayes-Optimal Readout

We have demonstrated that our system produces full-rank encoding matrices, so that arbitrary mapping between odors and valences can, *in principle*, be learned. We've also shown that a batch algorithm such as the perceptron learning rule can, *in fact*, learn the weights. However, batch learning is unrealistic, as it requires the animal to first gather information about the environment (e.g., record the valences of all observed odors), and to then learn the appropriate weights all at once. This approach is clearly unrealistic – what is required is a means for the animal to learn the weights 'on-the-fly'. Below we demonstrate that, at least for binary valences, a very simple and biologically plausible mechanism can perform nearly Bayes-optimal online learning.

A first approach could to be to implement the PLR at the synapses between the KCs and a proposed output neuron, and to simply have the updates occur online, instead of in batch mode. However, this approach is problematic because the PLR generally requires both positive and negative weights, while the KCs are thought to have only an excitatory effect on the beta-lobe neurons (bLNs) (Cassenaer and Laurent, 2007), which are immediately downstream.

We avoided this problem by taking inspiration from several facts about the biological system:

- Insect brains use different neuromodulators to signal aversive and appetitive stimuli.
- KC axons branch into the alpha, beta and gamma lobes,
- Beta lobe neurons inhibit each other.

Our proposed readout architecture is shown in Figure 2-5. In this system, each KC axon branches to form synapses with two beta lobe neurons: one signaling positive odors (the +bLN, green), and one signaling negative ones (the –bLN, red). These two bLNs are mutually inhibitory, so that the animal's valence decision when presented with an odor is determined by the stronger of these two neurons. Whenever the animal receives an odor with a positive valences, the synapses from active KCs to the +bLN are strengthened slightly, while those onto the –bLN remain unchanged. The opposite occurs when the animal receives an odor with a negative valence. The sum of the synaptic weights onto each bLN is kept constant, modeling a receptor pool of fixed (though possibly different) size for each neuron.

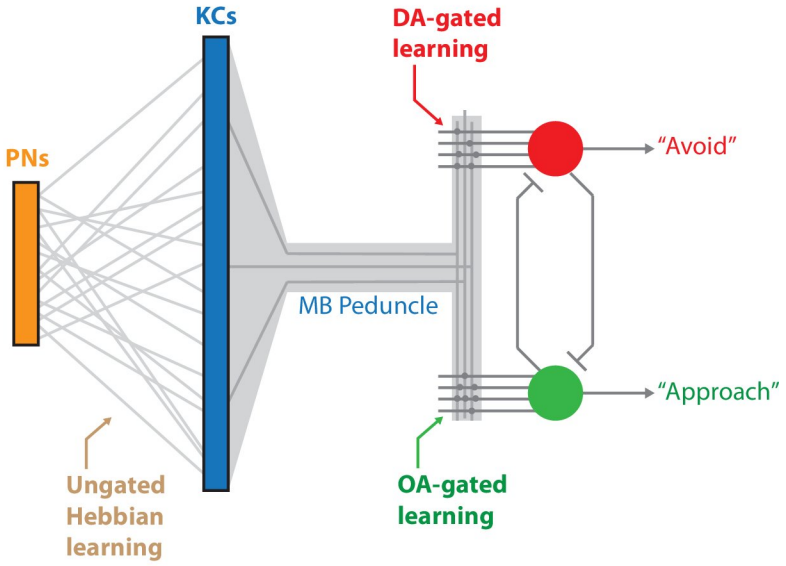


Figure 2-5 Proposed architecture for online learning.
Hebbian learning occurs at the PN-to-KC synapses as described above, to provide noise robustness. KC axons leave mushroom body along the peduncle and fork into excitatory and inhibitory readout branches. Learning at these synapses is a simple imprinting of the input pattern gated by the presence of the associated reward or punishment signal. Learning in the two stages occurs simultaneously and online. The readout neurons mutually inhibit each other, and the animal's decision is made according to which readout neuron is more strongly activated.

Now we show that the procedure above produces near Bayes-optimal learning. We proceed by

- Defining Bayes-optimal,
- Showing that the learning rule above leads to each KC-to-bLN synapse reflecting the likelihood of the corresponding KC being active during the presentation of excitatory or inhibitory odors (depending on the bLN).
- The net excitation of each bLN by an odor is proportional to an approximation of the likelihood of the corresponding valences,
- The proportionality constant is the size of the receptor pool. If these are made proportional to the prior probability of each valence, then the net excitation of each bLN by an odor reflects the posterior on the corresponding valence, given that odor.

Assuming (without loss of generality) that the animal receives rewards and punishments of equal magnitude, it will maximize its expected reward (Kay, 1993) if it makes a decision in favor of the odor whenever the posterior probability of positive valence given the odor, $P(+|\text{odor})$, is greater then the posterior on negative valence $P(-|\text{odor})$, and vise-versa. Hence if one of the bLNs computes $P(+|\text{odor})$, and the other $P(-|\text{odor})$, and the animal acts on the decision of the bLN with stronger excitation, it will maximize its expected reward. We will now show that

our system does indeed result in the bLN excitation reflecting these posteriors. We will focus on the excitation of the positive valence bLN.

2.8 Asymptotic Mean and Variance of Readout Weights

First, we note that our learning rule, coupled with synaptic normalization, leads to a KC-to-+bLN synapse storing the likelihood of positive reward given the activation of that KC. Let $w_i(n)$ be the weight of the synapse from KC *i* onto the +bLN, and $y_i(n)$ the (binary) activity of that KC, and $[OA](n)$ the (binary) positive-reward signal, all at time step *n*. Then our learning rule has the following effect:

$$u_i(n+1) = w_i(n) + \beta y_i(n)[OA](n)$$
$$w_i(n+1) = \frac{S}{\sum_{j=1}^{D} u_j(n+1)} u_i(n+1).$$

The first statement describes the update of the weight by the learning rule: the weight is increased by a small amount β if both the corresponding KC is active, and the reward signal is present. The second statement describes synaptic normalization, and ensures that the sum of the weights after each update sum to *S*. Indeed

$$\sum_{i=1}^{D} w_i(n+1) = \sum_{i=1}^{D} \frac{S}{\sum\limits_{j=1}^{D} u_j(n+1)} u_i(n+1) = \frac{S}{\sum\limits_{j=1}^{D} u_j(n+1)} \sum_{i=1}^{D} u_i(n+1) = S.$$

Next we rewrite the normalization update by decomposing the denominator:

$$
\begin{aligned}
w_i(n+1) &= \frac{S}{\sum_{j=1}^{D} u_j(n+1)} u_i(n+1) \\
&= \frac{S}{\sum_{j=1}^{D} w_j(n) + \beta y_j(n)[OA](n)} u_i(n+1) \\
&= \frac{S}{\left(\sum_{j=1}^{D} w_j(n)\right) + \beta[OA](n)\left(\sum_{j=1}^{D} y_j(n)\right)} u_i(n+1) \\
&= \frac{S}{S + \beta[OA](n)k} u_i(n+1).
\end{aligned}
$$

Here we've used the fact that sum of the weights at each time step is *S*, and the sum of the inputs at each time step is *k*, the sparseness level of the mushroom body.

Since the weights only change when the binary reward signal [*OA*](*n*) is active, we can re-index time to the indices when [*OA*](n) is 1. Then we have

$$
w_i(n+1) = \frac{S}{S+\beta k} u_i(n+1) = \frac{S}{S+\beta k} w_i(n) + \frac{S\beta}{S+\beta k} y_i(n).
$$

We can now look at the time course of the development of the weight $w_i(n+1)$

$$
\begin{aligned}
&w_i(0) = w_0 \\
&w_i(1) = aw_0 + by_i(0) \\
&w_i(2) = a(aw_0 + by_i(0)) + by_i(1) = a^2 w_0 + aby_i(0) + by_i(1) \\
&w_i(3) = a(a^2 w_0 + aby_i(0) + by_i(1)) + by_i(2) = a^3 w_0 + a^2 by_i(0) + aby_i(1) + by_i(2) \\
&\quad \vdots \\
&w_i(n) = a^n w_0 + b\sum_{j=0}^{n-1} a^{n-1-j} y_i(j) = a^n w_0 + ba^{n-1} \sum_{j=0}^{n-1} a^{-j} y_i(j).
\end{aligned}
$$

Here we've set $a = S / (S + \beta k)$, and $b = S\beta / (S + \beta k)$ for clarity. Now we can take the expectation over odor presentation schedules:

$$\begin{aligned}
E(w_i(n)) &= E\left(a^n w_{i,0} + ba^{n-1}\sum_{j=0}^{n-1} a^{-j} y_i(j)\right) \\
&= E\left(a^n w_{i,0}\right) + E\left(ba^{n-1}\sum_{j=0}^{n-1} a^{-j} y_i(j)\right) \\
&= a^n E(w_{i,0}) + ba^{n-1}\sum_{j=0}^{n-1} a^{-j} E\left(y_i(j)\right) \\
&= a^n \overline{w}_{i,0} + \overline{y}_i b a^{n-1}\sum_{j=0}^{n-1} a^{-j} \\
&= a^n \overline{w}_{i,0} + \overline{y}_i b a^{n-1}\frac{1}{a^{n-1}}\frac{1-a^n}{1-a} \\
&= a^n \overline{w}_{i,0} + \overline{y}_i b \frac{1-a^n}{1-a}.
\end{aligned}$$

The main assumption we make is that odors arrive randomly in time, so that the expectation of y is independent of time. Taking the limit as n goes to infinity, and noting that $a < 0$, yields the steady-state mean value of the weight as

$$\overline{w}_i = \frac{b}{1-a}\overline{y}_i = \frac{S\beta/(S+\beta k)}{1-(S/(S+\beta k))}\overline{y}_i = \frac{S\beta}{\beta k}\overline{y}_i = \frac{S}{k}\overline{y}_i,$$

But since KC activity is binary,

$$\overline{y}_i = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} y_i(n) = P(y_i = 1|+),$$

i.e., the likelihood of a positive odor given that the *i*'th KC is active. Hence, the steady-state value of the synaptic weight from the *i*'th KC onto the +bLN is proportional to the likelihood of a positive odor, given that the KC is active.

We can also compute the asymptotic variance of each weight:

$$\begin{aligned}
\mathrm{var}\left(w_i(n)\right) &= \mathrm{var}\left(a^n w_{i,0} + ba^{n-1}\sum_{j=0}^{n-1} a^{-j} y_i(j)\right) \\
&= \mathrm{var}\left(a^n w_{i,0}\right) + \mathrm{var}\left(ba^{n-1}\sum_{j=0}^{n-1} a^{-j} y_i(j)\right) \\
&= a^{2n}\mathrm{var}\left(w_{i,0}\right) + b^2 a^{2(n-1)}\sum_{j=0}^{n-1} a^{-2j}\mathrm{var}\left(y_i(j)\right) \\
&= a^{2n}\mathrm{var}\left(w_{i,0}\right) + \mathrm{var}\left(y_i\right) b^2 a^{2(n-1)}\sum_{j=0}^{n-1} a^{-2j} \\
&= a^{2n}\mathrm{var}\left(w_{i,0}\right) + \mathrm{var}\left(y_i\right) b^2 a^{2(n-1)}\frac{1}{a^{2(n-1)}}\frac{1-a^{2n}}{1-a^2} \\
&= a^{2n}\mathrm{var}\left(w_{i,0}\right) + \mathrm{var}\left(y_i\right) b^2 \frac{1-a^{2n}}{1-a^2}.
\end{aligned}$$

For the second equality, we've assumed that the initial value of the weight and the KC activity on that synapse are independent, allowing the variances to sum. More importantly, in the third equality we've used the assumption of temporal independence of inputs, which again allows the variances to sum. In the fourth equality, we've used the temporal symmetry assumption, that the input statistics at any point in time are the same, so that the variance is independent of time.

Now recalling that $a = S / (S + \beta k) < 1$, we can take the limit as n goes to infinity,

$$\begin{aligned}
\lim_{n\to\infty} \operatorname{var}(w_i(n)) &= \lim_{n\to\infty} a^{2n}\operatorname{var}(w_{i,0}) + \operatorname{var}(y_i)\, b^2 \frac{1-a^{2n}}{1-a^2} \\
&= \frac{b^2}{1-a^2}\operatorname{var}(y_i) \\
&= \frac{(S\beta)^2/(S+\beta k)^2}{1-S^2/(S+\beta k)^2}\operatorname{var}(y_i) \\
&= \frac{(S\beta)^2}{(S+\beta k)^2 - S^2}\operatorname{var}(y_i) \\
&= \frac{(S\beta)^2}{(S+\beta k - S)(S+\beta k + S)}\operatorname{var}(y_i) \\
&= \frac{S^2\beta}{k(2S+\beta k)}\operatorname{var}(y_i)\,.
\end{aligned}$$

Hence a non-zero variance is eventually attained, which is inversely proportional to the sparseness.

2.9 Expected Drive of Excitatory bLN

We can now compute the expected net current into the +bLN when the animal is presented with a positive odor:

$$
\begin{aligned}
E(z^+(u|u \in Y^+)) &= \sum_{i=1}^{D} \overline{w}_i u_i \\
&= \sum_{i=1}^{k} \overline{w}_{I_u(i)} u_{I_u(i)} = \sum_{i=1}^{k} \overline{w}_{I_u(i)} \\
&= \frac{S}{k} \sum_{i=1}^{k} P(y_{I_u(i)}|+) \\
&= \frac{S}{k} \sum_{i=1}^{k} \sum_{v \in Y^+} P(y_{I_u(i)}, v|+) \\
&= \frac{S}{k} \sum_{i=1}^{k} \sum_{v \in Y^+} P(y_{I_u(i)}|v,+)P(v|+) \\
&= \frac{S}{k} \sum_{i=1}^{k} \left(P(y_{I_u(i)}|u,+)P(u|+) + \sum_{\substack{v \in Y^+, \\ v \neq u}} P(y_{I_u(i)}|v,+)P(v|+) \right) \\
&\approx \frac{S}{k} \sum_{i=1}^{k} \left(1 \cdot P(u|+) + \sum_{\substack{v \in Y^+, \\ v \neq u}} \frac{k}{D} P(v|+) \right) \\
&\approx \frac{S}{k} \sum_{i=1}^{k} \left(P(u|+) + \sum_{\substack{v \in Y^+, \\ v \neq u}} \frac{k}{D} \cdot \frac{1}{|Y^+|} \right) \\
&= \frac{S}{k} k \left(P(u|+) + \sum_{\substack{v \in Y^+, \\ v \neq u}} \frac{k}{D} \cdot \frac{1}{|Y^+|} \right) \\
&= \frac{S}{k} k \left(P(u|+) + \frac{k\left(|Y^+| - 1\right)}{D|Y^+|} \right) \\
&\approx S \left(P(u|+) + \frac{k}{D} \right).
\end{aligned}
$$

Hence the net input into the +bLN on presentation of an excitatory odor u is approximately proportional to the likelihood of positive valence given the odor, plus an interference term due to overlap between KC odor representations. If this overlap is made small, either by increasing the sparseness of the code (reducing k), or increasing the number of KCs (D), then the input to the bLN is approximately proportional to the likelihood of positive valence given the odor.

We've now shown that the input to the +bLN in response to an excitatory odor u can be made approximately proportional to the likelihood of positive valence given that odor. To have this bLN compute the posterior in positive valence given the odor, we need a multiplicative term to reflect the prior on positive valences. But S, the normalization constant of the weights into the bLN, is just such a term! It models the size of the receptor pool available to this bLN. The size of this pool can be adjusted over the lifetime of the animal, or even over evolutionary time, to reflect the prior on positive odors. The result will be that the excitation of the +bLN will be proportional to the posterior on positive valence given the odor. Similarly, the excitation of the -bLN will be proportional to the posterior on negative valence given the odor. Hence, if the animal makes its decision based on whichever bLN is more excited, via, e.g., mutual inhibition between these bLNs, then its decisions will be near Bayes-optimal.

2.10 Numerical Verification

We demonstrate these results numerically in the following figures. Our derivations above were based on approximating the encoding process as a mapping of each *d*-dimensional binary odor input vector into a *D*-dimensional binary vector with *k* elements set randomly and independently to 1. We first numerically verify our derivations under this assumption. We fixed the set of *N* encoded input vectors and ran our algorithm 96 times, each time using a different set of initial weights onto the excitatory bLN, and a different presentation schedule for the *N* odors. We then estimated the expectation of each weight at each time point by averaging over the 96 runs.

In Figure 2-6 the evolution of the weight distribution for the excitatory bLN is plotted. The weights are sorted by their theoretically predicted expected values. The figure shows that the distribution converges to the theoretically predicted one within ~ 50 excitatory updates.

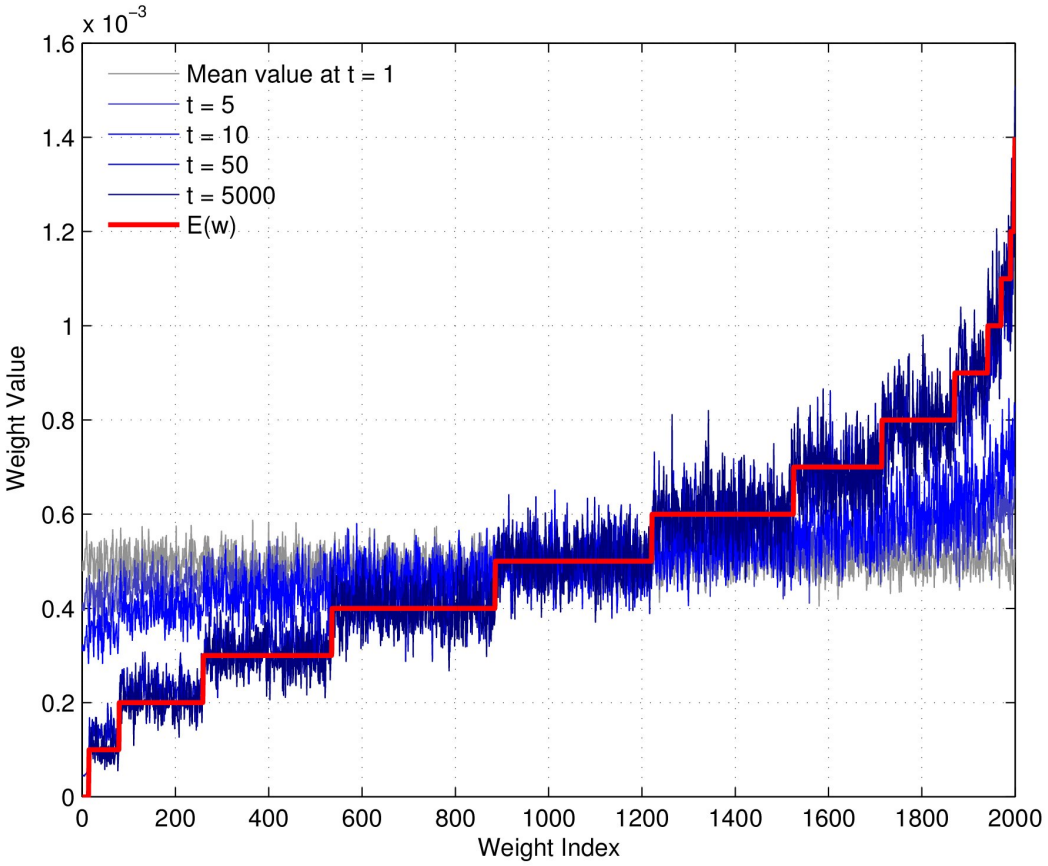


Figure 2-6 Evolution of readout weights.
The PN-to-KC encoding process was approximated as a mapping from odors to binary vectors in a 2000 dimensional KC space, with the active 50 KCs in each encoding chosen at random. Readout weights were trained and their mean value over 96 trials plotted as a function time, showing the distribution converges very quickly to the theoretically predicted values.

In Figure 2-7 we track the updates to a single weight. The mean value of this weight over all 96 runs, at each point in time, is plotted in the top panel in blue, and is clearly centered around the theoretically predicted value, shown in red. Similarly, the bottom panel shows the variance at each time point in blue, and the theoretically expected value in red, again showing a good match.

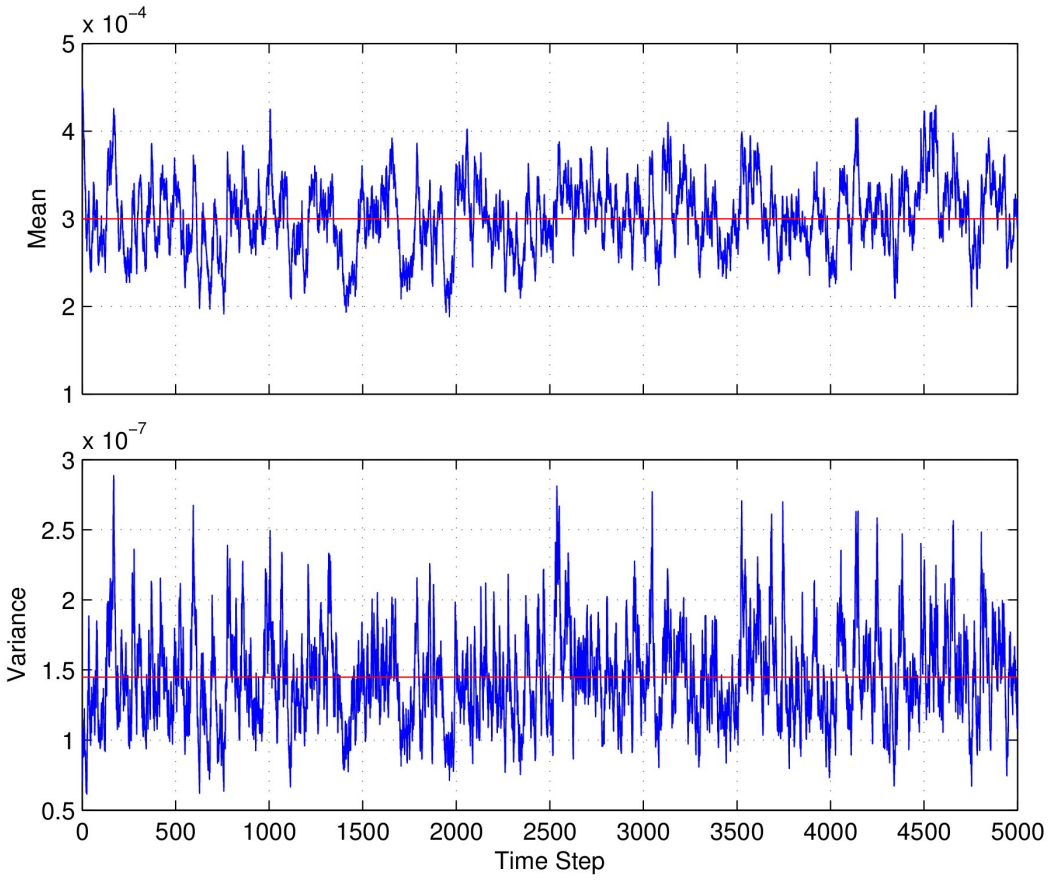


Figure 2-7 Mean and variance of an example readout weight.
A single readout weight was tracked during learning and its mean and variance over 96 trials computed and plotted as a function of time (blue). The observed values are clearly centered around their theoretically predicted expected values (red).

In Figure 2-8 we show the average drive received by the excitatory bLN near the end of learning, for each of 200 equiprobable excitatory inputs. Again, the values of the weights are approximately centered around their theoretically predicted value, shown in red.

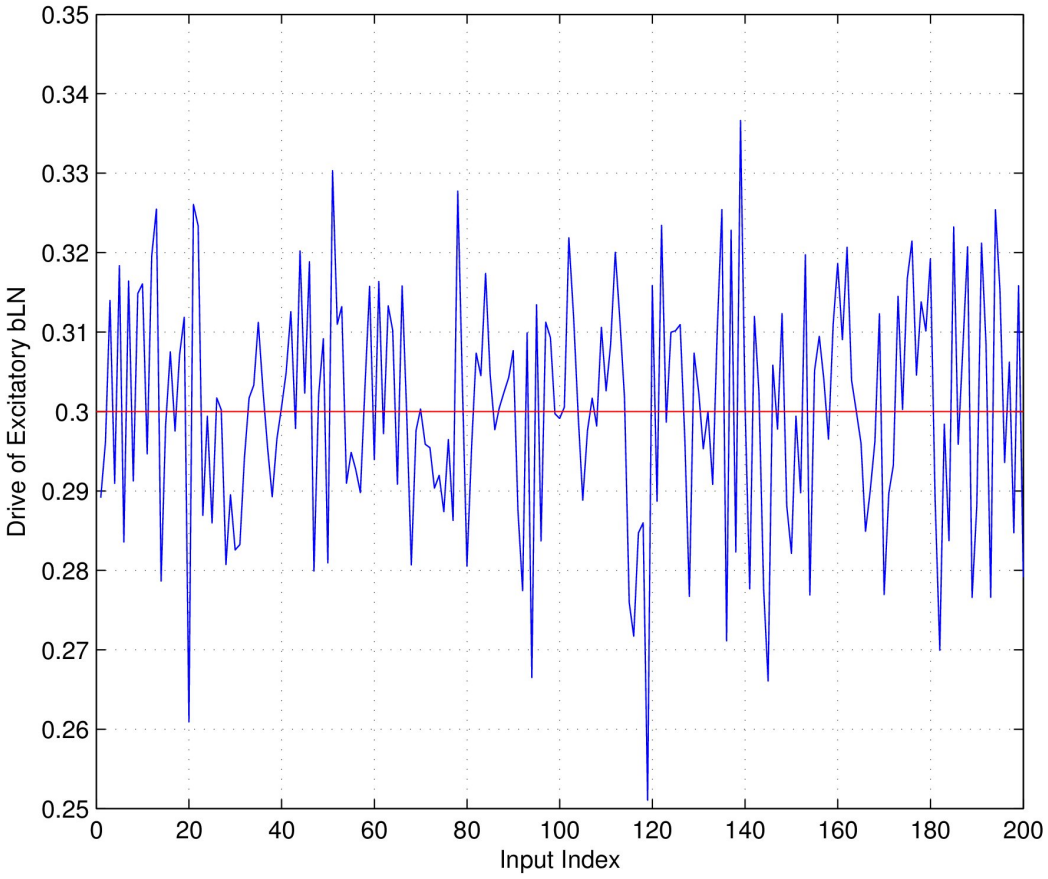


Figure 2-8 Excitation of +bLN for 200 equiprobable inputs.
The mean value over 96 trials of the drive to the excitatory bLN in response to the presentation of each of 200 equiprobable excitatory odors, sampled near the end of learning. Because the odors are equiprobable, they all have the same theoretically predicted expected value (red), and are clearly centered around it.

Next we perform the same procedure as above, but instead use an actual learned encoding of 1000 200-dimensional odor inputs into a 4000-dimensional output space with 5 elements active in each encoding. The results are shown in Figure 2-9, Figure 2-10, and Figure 2-11, and are a good match between to the theoretical predictions, indicating that the random encoding approximation is sufficiently accurate.

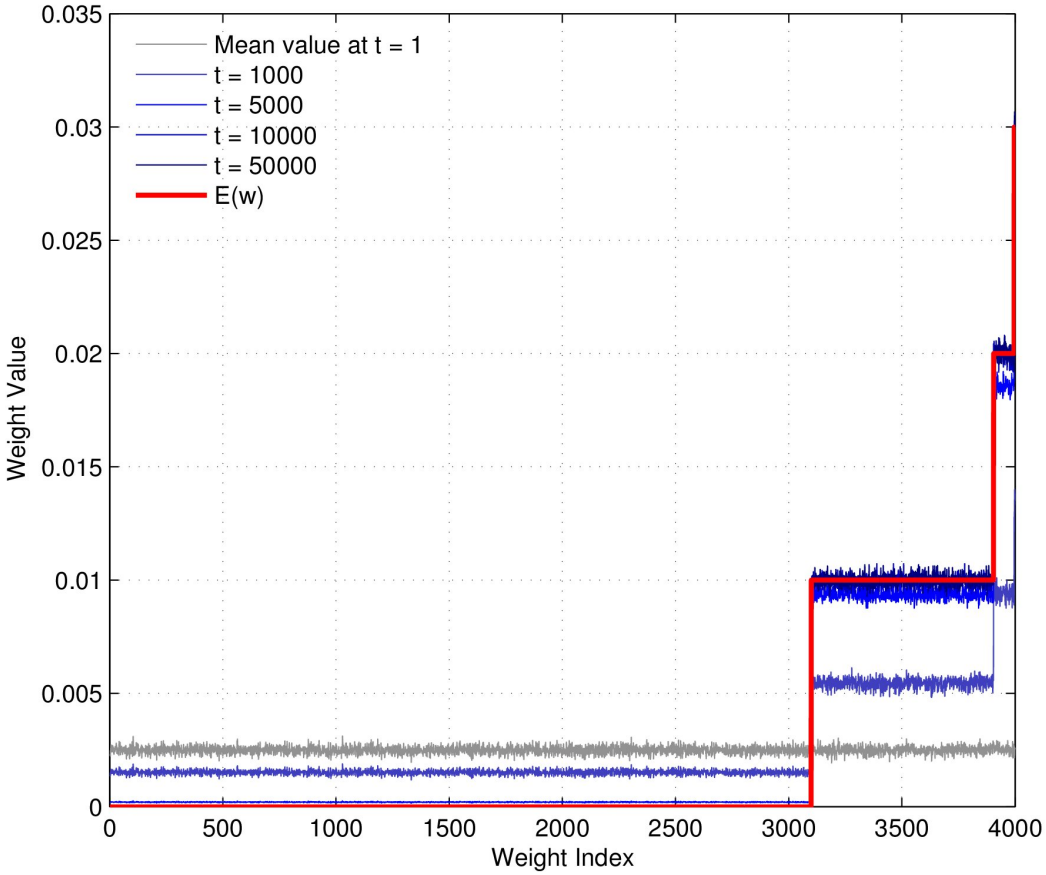


Figure 2-9 Evolution of readout weights using an actual odor encoding as input.
Same procedure as in Figure 2-6, but using an actual encoding matrix learned by the system. As before, the weight distribution converges to its expected form. The expected value of the weight distribution is different from that in Figure 2-6 and more trials are required for convergence because we used a much higher and more realistic value of sparseness (5 active KCs in each representation instead of the 50 used previously).

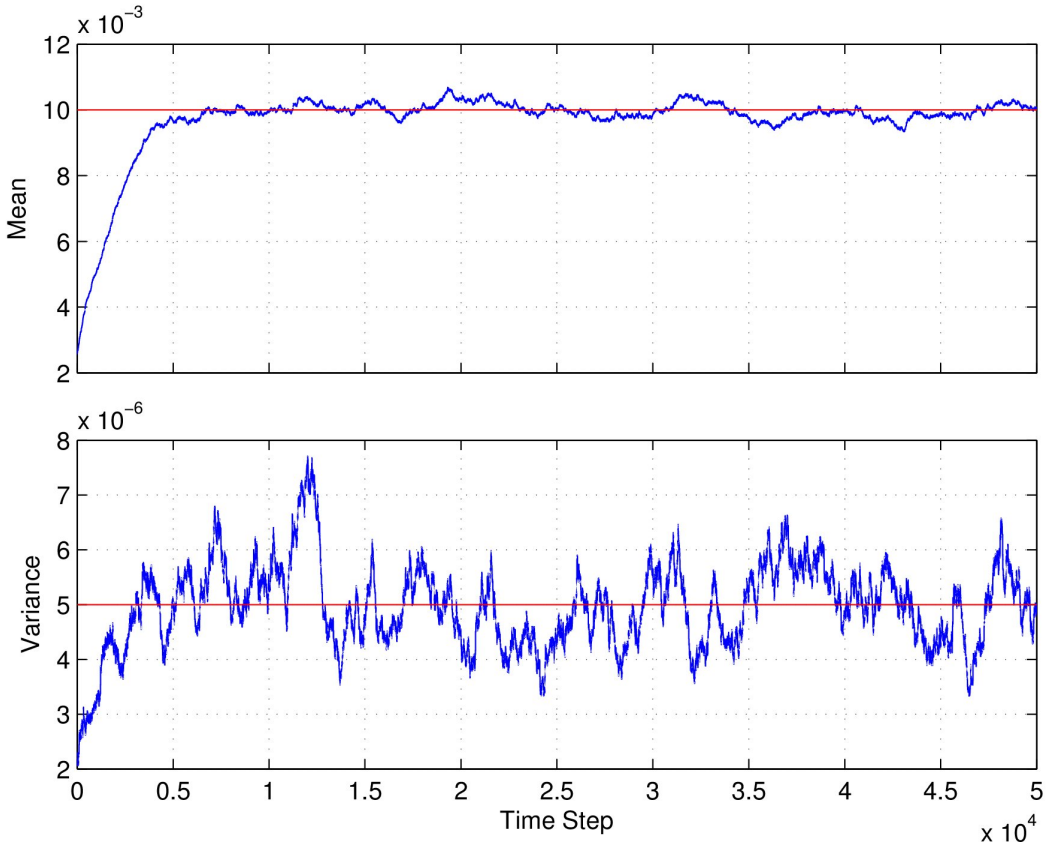


Figure 2-10 Mean and variance of an example readout weight when using actual odor encodings as input.

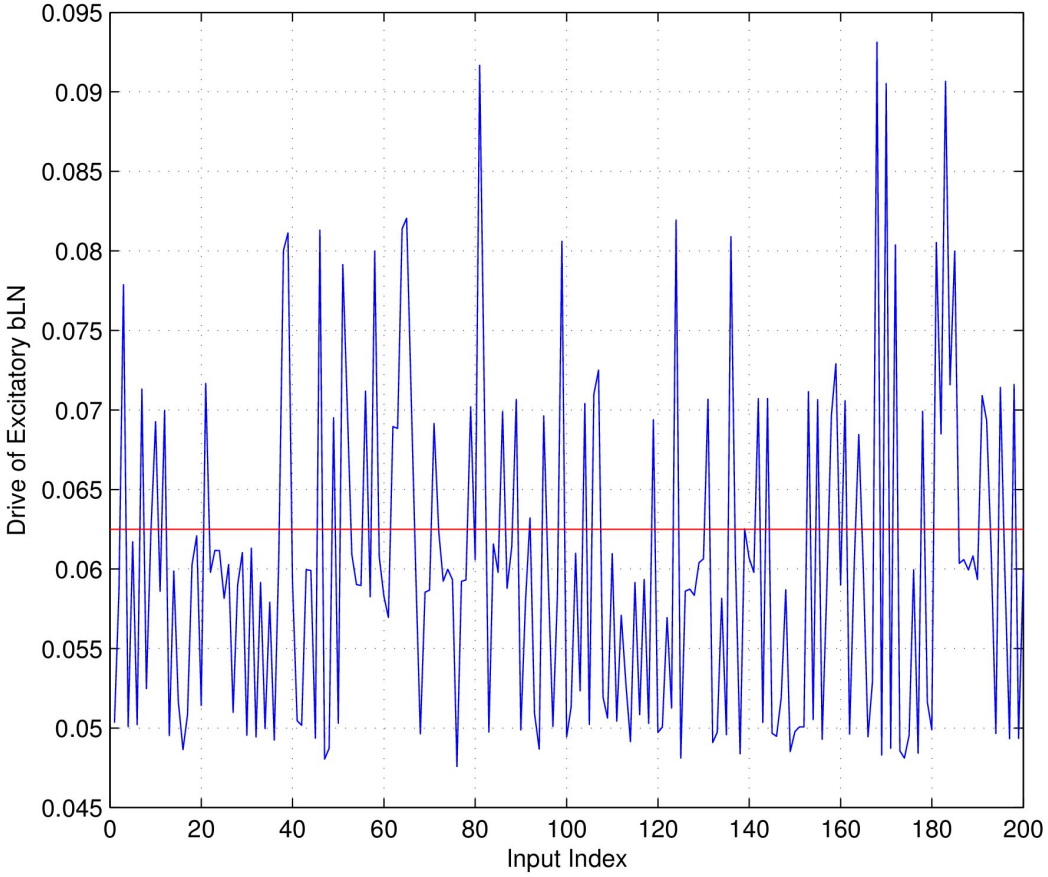


Figure 2-11 Excitation of +bLN using an actual encoding of 200 equiprobable excitatory odors.

Finally, we demonstrate the performance of the entire system. A set of 500 200-dimensional binary vectors were randomly generated with mean density 0.5, and assigned to be excitatory or inhibitory according to a pre-specified bias. The system weights were initialized randomly. For each of 120,000 time steps, an input was selected at random and presented to the system along with the valence signal. The system projected the inputs into a 2000-dimensional KC space, with the GGN allowing only 5 KCs to respond. Weights from PNs to KCs and from KCs to the readout neurons were updated as above, after which the next odor was presented. The system was tested at various points in time on the entire odor repertoire and its performance recorded. The procedure was repeated 5 times, using a different set of odors each time. In Figure 2-12 we show the mean (thick curves) ± 3 S.E.M. (thin curves) of the error rate of the system as a function of time and for three different values of the prior on positive odors. In all cases tested, the errors drop to near zero after ~ 80,000 time steps, implying that each odor needs to be encountered on the order of 100 times before learning is complete. The asymptotic error rate is not exactly zero, and this is presumably due to errors caused by variance in the excitations of the readout neurons. Future work will derive this variance and provide an estimate of the steady-state error.

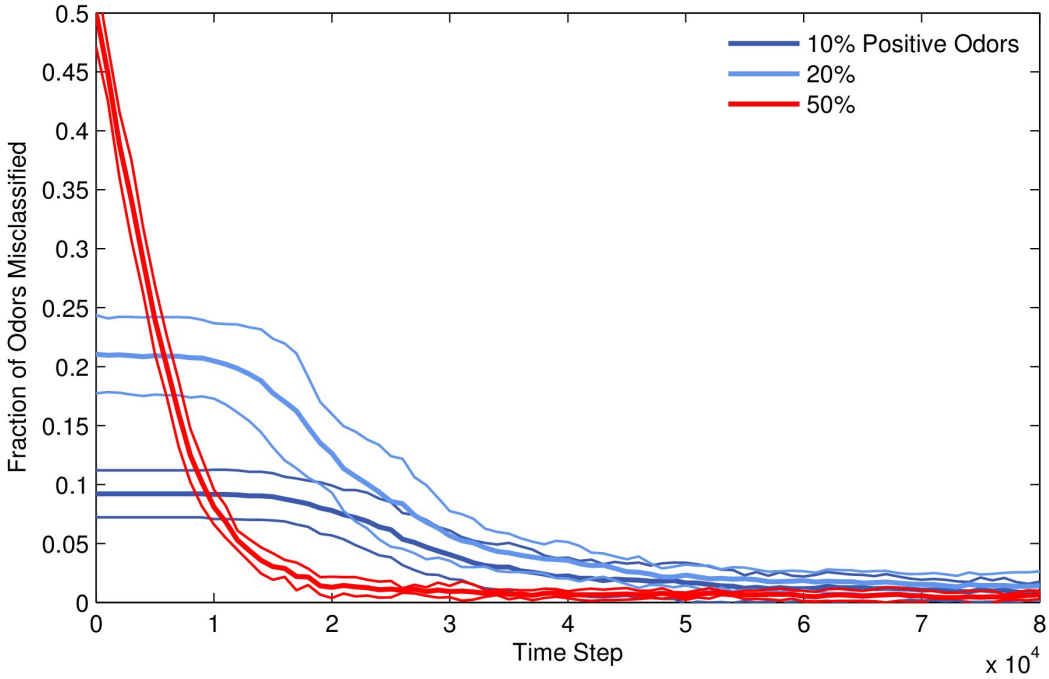


Figure 2-12 Performance of online learning.

The system was trained on 500 200-dimensional input vectors, with a fixed fraction being positive and the rest being negative. Inputs were projected into 2000-dimensional KC space, with 5 KCs active per response. The input at each time step was corrupted by up to 20 bits of noise. Hebbian learning at PN-to-KC synapses and gated learning at KC-to-bLN synapses took place simultaneously online. The classification performance of the system was measured on the entire uncorrupted odor set periodically during learning, and the mean ± 3 S.E.M.s computed over 5 trials are plotted above, for three different fractions of excitatory odors. In all cases learning reduces errors to near zero after about 80,000 time steps.

2.11 Discussion

In summary, we propose that the olfactory system is designed to allow mapping of odors to valences arbitrarily. We propose that the large fan-out from PNs to KCs is to expand the number of odors that can be mapped in this way. We showed through simulations that such a full-rank odor representation can be maintained while using known circuit mechanisms to achieve biological levels of sparseness, and that Hebbian learning at the PN-to-KC synapse to stabilize odor representations. Finally we proposed a biologically plausible architecture for *online* readout of the KCs representation by bLNs which performs at a near Bayes-optimal level.

Our model makes several anatomical and physiological predictions, including:

- Weight normalization by KCs and bLNs;
- Hebbian learning at the PN-to-KC synapse;
- Branched readout of KCs, with different populations learning excitatory vs. inhibitory weights, and their mutual inhibition deciding the animal's decision on the odor;
- Storage of priors on excitation and inhibition in the sizes of the Ach receptor pools for the two readout branches;
- Short-term disabling of the GGN will lead to decisions being made based on the priors on excitation vs. inhibition, as the both branches will receive roughly equal excitation;

- Long-term disabling of the GGN will lead to forgetting/garbling of odor representations as Hebbian plasticity at the input overwrites established PN-to-KC weights;
- Removing the environmental feedback signal (dopamine or octopamine) once learning is complete should have no effect on behavioral performance; the odor itself should be enough.

There are a number of areas in which this work can be expanded. One of the simplifying assumptions we made was to approximate antennal lobe function as a mapping of odors into single, dense, binary vectors each of whose elements are independently generated. This approximation disregards the temporal dynamics of the antennal lobe, which produce decorrelated (though not necessarily independent) odor representations over the course of several oscillation cycles(Mazor and Laurent, 2005). Thus it is important to investigate how our proposed architecture deals with more complex mappings between odors and AL representations, in which the PNs are often not even uncorrelated, much less independent.

A second question regards our simplification of PN spiking dynamics into binary responses representing excitation or inhibition within an oscillation cycle. It would be interesting to implement our model with spiking neurons, particularly because such an implementation may shed light on the potential role played by STDP. Our

present framework cannot incorporate STDP directly because our implicit temporal resolution of a single oscillation cycle (~ 50 ms) is much longer than the resolution in which STDP operates (< 10 ms, REF). Yet STDP has experimentally been found to play a key role in KC-bLN interactions, so it is important to account for it within our architecture.

A third question to be investigated is whether the valence readout scheme can be extended to a wider range of valences than just positive and negative. This should be possible by adding additional branches to the KC readout, each with its own readout neuron and with learning gated by a dedicated neurotransmitter, but the precise details need to be worked out and the scaling limits determined.

References

Abraham, N.M. (2004). Maintaining accuracy at the expense of speed: stimulus similarity defines odor discrimination time in mice. Neuron *44*, 865-876.

Bäcker, A. (2002). Pattern recognition in locust early olfactory circuits: Priming, gain control, and coding issues. PhD Thesis, Computation and Neural Systems, California Institute of Technology.

Barlow, H.B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology. Perception *1*, 371-394.

Bazhenov, M., Stopfer, M., Rabinovich, M., Huerta, R., Abarbanel, H.D., Sejnowski, T.J., and Laurent, G. (2001). Model of transient oscillatory synchronization in the locust antennal lobe. Neuron *30*, 553-567.

Bhagavan, S., and Smith, B.H. (1997). Olfactory conditioning in the honey bee, Apris mellifera: effects of odor intensity. Physiol Behav *61*, 107-117.

Bhandawat, V., Olsen, S.R., Gouwens, N.W., Schlief, M.L., and Wilson, R.I. (2007). Sensory processing in the Drosophila antennal lobe increases reliability and separability of ensemble odor representations. Nature neuroscience *10*, 1474-1482.

Broome, B.M., Jayaraman, V., and Laurent, G. (2006). Encoding and decoding of overlapping odor sequences. Neuron *51*, 467-482.

Brown, S.L., Joseph, J., and Stopfer, M. (2005). Encoding a temporally structured stimulus with a temporally structured neural representation. Nat Neurosci *8*, 1568-1576.

Buhlmann, P. (2011). Statistics for high-dimensional data (New York: Springer).

Cassenaer, S., and Laurent, G. (2007). Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. Nature *448*, 709-713.

Cassenaer, S., and Laurent, G. (2012). Conditional modulation of spike-timing-dependent plasticity for olfactory learning. Nature *482*, 47-52.

Crittenden, J.R., Skoulakis, E.M.C., Han, K.A., Kalderon, D., and Davis, R.L. (1998). Tripartite mushroom body architecture revealed by antigenic markers. Learning & Memory *5*, 38-51.

Davis, R.L. (2005). Olfactory memory formation in Drosophila: from molecular to systems neuroscience. Annual review of neuroscience *28*, 275-302.

Deza, M., and Deza, E. (2009). Encyclopedia of distances (Dordrecht New York: Springer Verlag).

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn Sci *11*, 333-341.

Eisthen, H.L. (2002). Why are olfactory systems of different animals so similar? Brain, behavior and evolution *59*, 273-293.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recogn Lett *27*, 861.

Field, D.J. (1994). What Is the Goal of Sensory Coding. Neural Comput *6*, 559-601.

Foldiak, D.J. (2003). Sparse coding in the primate cortex. In The Handbook of Brain Theory and Neural Networks, M.A. Arbib, ed. (MIT Press), pp. 895-898.

Grünwald, P.D. (2007). The minimum description length principle (MIT press).

Heisenberg, M. (2003). Mushroom body memoir: from maps to models. Nature reviews Neuroscience *4*, 266-275.

Hertz, J., Krogh, A., and Palmer, R.G. (1991). Introduction to the theory of neural computation (Redwood City, Calif.: Addison-Wesley Pub. Co.).

Hubert, L., and Arabie, P. (1985). Comparing Partitions. J Classif *2*, 193-218.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identiy from macaque inferior temporal cortex. Science *310*, 863-866.

Hurst, J.L., and Beynon, R.J. (2004). Scent wars: the chemobiology of competitive signalling in mice. Bioessays *26*, 1288-1298.

Jinks, A., and Laing, D.G. (1999). A limit in the processing of components in odour mixtures. Perception *28*, 395-404.

Jortner, R.A., Farivar, S.S., and Laurent, G. (2007). A simple connectivity scheme for sparse coding in an olfactory system. J Neurosci *27*, 1659-1669.

Kay, S.M. (1993). Fundamentals of statistical signal processing (Englewood Cliffs, N.J.: Prentice-Hall PTR).

Laurent, G., and Davidowitz, H. (1994). Encoding of olfactory information with oscillating neural assemblies. Science *265*, 1872-1875.

Laurent, G., and Naraghi, M. (1994). Odorant-induced oscillations in the mushroom bodies of the locust. The Journal of neuroscience : the official journal of the Society for Neuroscience *14*, 2993-3004.

Leitch, B., and Laurent, G. (1996). GABAergic synapses in the antennal lobe and mushroom body of the locust olfactory system. J Comp Neurol *372*, 487-514.

Linster, C., Johnson, B.A., Morse, A., Yue, E., and Leon, M. (2002). Spontaneous versus reinforced olfactory discriminations. J Neurosci *22*, 6842-6845.

Lu, X.C., and Slotnick, B.M. (1998). Olfaction in rats with extensive lesions of the olfactory bulbs: implications for odor coding. Neuroscience *84*, 849-866.

Maass, W., Natschlager, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput *14*, 2531-2560.

MacKay, D.J.C. (2003). Information theory, inference, and learning algorithms (Cambridge, UK ; New York: Cambridge University Press).

MacLeod, K., Backer, A., and Laurent, G. (1998). Who reads temporal information contained across synchronized and oscillatory spike trains? Nature *395*, 693-698.

Mainen, Z.F. (2006). Behavioral analysis of olfactory coding and computation in rodents. Curr Opin Neurobiol *16*, 429-434.

Masse, N.Y., Turner, G.C., and Jefferis, G.S. (2009). Olfactory information processing in Drosophila. Curr Biol *19*.

Mazor, O., and Laurent, G. (2005). Transient dynamics versus fixed points in odor represenetations by locust antennal lobe projection neurons. Neuron *48*, 661-673.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neuropyhsiol *100*, 1407-1419.

Niessing, J., and Friedrich, R.W. (2010). Olfactory pattern classification by discrete neuronal network states. Nature *465*, 47-52.

Papadopoulou, M., Cassenaer, S., Nowotny, T., and Laurent, G. (2011). Normalization for sparse encoding of odors by a wide-field interneuron. Science *332*, 721-725.

Perez-Orive, J., Bazhenov, M., and G., L. (2004). Intrinsic and circuit properties favor coincidence detection for decoding oscillatory input. J Neurosci *24*, 6037-6047.

Perez-Orive, J., Mazor, O., Turner, G.C., Cassenaer, S., Wilson, R.I., and Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. Science *297*, 359-365.

Pouzat, C., Mazor, O., and Laurent, G. (2002). Using noise signature to optimize spike-sorting and to assess neuronal classification quality. J Neurosci Methods *122*, 43-57.

Rand, W.M. (1971). Objective Criteria for Evaluation of Clustering Methods. J Am Stat Assoc *66*, 846-&.

Reinhard, J., Sinclair, M., Srinivasan, M.V., and Claudionos, C. (2010). Honeybees learn odour mixtures via a selection of key odorants. PLoS One *5*.

Rifkin, R., Yeo, G., and Poggio, T. (2003). Regularized least squares classification, Vol 190 (Amsterdam: VIOS Press).

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323-2326.

Rubin, B.D., and Katz, L.C. (1999). Optical imaging of odorant representations in the mammalian olfactory bulb. Neuron *23*, 499-511.

Sivia, D.S., and Skilling, J. (2006). Data analysis : a Bayesian tutorial, 2nd edn (Oxford ; New York: Oxford University Press).

Stopfer, M., Jayaraman, V., and Laurent, G. (2003). Intensity versus identity coding in an olfactory system. Neuron *39*, 991-1004.

Stopfer, M., and Laurent, G. (1999). Short-term memory in olfactory network dynamics. Nature *402*, 664-668.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological), 267-288.

Uchida, N., and Mainen, Z.F. (2003). Speed and accuracy of olfactory discrimination in the rat. Nat Neurosci *6*, 1224-1229.

Uchida, N., and Mainen, Z.F. (2007). Odor concentration invariance by chemical ratio coding. Front Syst Neurosci *1*, 3.

Vosshall, L.B., and Stocker, R.F. (2007). Molecular architecture of smell and taste in Drosophila. Annual review of neuroscience *30*, 505-533.

Wang, D.L., Buhmann, J., and von der Malsburg, C. (1990). Pattern segmentation in associative memory. Neural Comput *2*, 94-106.

Wilson, R.I., and Mainen, Z.F. (2006). Early events in olfactory processing. Annu Rev Neurosci *29*, 163-201.

Wright, G.A., Kottcamp, S.M., and Thomson, M.G. (2008). Generalization mediates sensitivity to complex odor features in the honeybee. PLoS One *3*.

Wright, G.A., and Thomson, M.G. (2005). Odor perception and variabiltiy in natural odor scenes (Amsterdam: Elsevier).