Representation and learning in cerebellum-like structures

Ann Kennedy

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Representation and learning in cerebellum-like structures

Ann Kennedy

Animals use their nervous system to translate signals from their sensory environment into appropriate behavioral responses. In some cases, these responses are hard-wired through genetic sculpting of neural circuits, such that certain stimuli drive innate behavioral responses in the absence of prior experience [Ewert, Burghagen, and Schurg-Pfeiffer 1983; Yilmaz and Meister 2013; Wu et al 2014]. But most often, responses to stimuli are modified over the course of an organism's lifetime via associative learning, in which past experience is used to adaptively modify the neural circuits controlling behavior.

The remarkable regularity of cerebellar circuitry made it an early target of experiments seeking a link between neural circuit structure and computational function [Eccles, Ito, and Szentágothai, 1967]. These efforts led to a first generation of models describing cerebellar cortex as a device for associative learning, remarkable for their focus on linking each cell type of cerebellar cortex to a computational aspect of associative memory formation and adaptive control [Marr 1969; Albus 1971; Ito 1972]. In subsequent decades, specialized neural architecture resembling that of the cerebellum has been identified in several other brain regions, including the dorsal cochlear nucleus of most mammals [Oertel and Young, 2004], the mushroom body of the insect olfactory system [Farris, 2011], and a region evolutionarily and developmentally related to the cerebellum in the brains of weakly electric fish, the electrosensory lobe [Bell, Han, and Sawtell, 2008]. This has raised the hope that a similar computational mechanism is at work in these structures.

It is not easy to find behavioral paradigms that isolate learning in the cerebellum, and a complete mechanistic account of learning during commonly studied behaviors has remained elusive. In this thesis, I analyze two cerebellum-like structures– the electrosensory lobe of the mormyrid fish and

the mushroom body of the fly olfactory system– in which mapping out associative learning is more tractable, due to the availability of well-controlled learning paradigms and the development of powerful biochemical and genetic techniques.

With the help of my experimental collaborators, I constructed computational models of the electrosensory lobe and mushroom body from electrophysiological and anatomical data, and studied the process of associative learning in these models. In both systems, an initial sensory representation is first projected up into a high dimensional space, and then read out via convergent input onto individual neurons. Learning adjusts the input to readout neurons, causing changes in their responses to future stimuli that alters their drive to downstream nuclei. Two details shape how each circuit handles associative learning: the way in which sensory inputs are represented, and the mechanism of learning. Together, these two pieces determine what transformations each circuit is able to learn and how it generalizes after learning.

In the four chapters of this thesis I present four related projects dealing with sensory representation and learning in cerebellum-like structures. The first chapter has previously been published as a paper and describes a model for cancellation of self-generated sensory input in the passive electrosensory system of the mormyrid fish. In the second chapter, I adapt this model to a more high-dimensional cancellation problem in the fish's active electrosensory system, which deals with the effects of the fish's body on the electric fields it generates. In the next two chapters, I construct a network model of odor representation in fly olfactory system, terminating at the mushroom body. Finally, I use this model in conjunction with recent experimental findings on the output of the mushroom body, to build a model of associative odor learning in the fly.

Table of Contents

List of Figures

Acknowledgments

This thesis is a direct product of the support I've received from both my experimental collaborators and my fellow theorists at Columbia. The graduate students, postdocs, faculty, and alumni of the theory center are a unique community, and the daily discussion over the past few years is something I will miss very much. I thank my advisor, Larry Abbott, for his patience, insight, and guidance throughout my PhD, and for placing his trust in me as I gained my footing in the field. I am happy to have spent a PhD in a lab so full of exciting projects and good discussion (thanks to Brian DePasquale, Greg Wayne, Saul Kato, Patrick Kaifosh, Jeff Seely, Merav Stern, and other past members of the Abbott lab!) I am especially grateful to Nate Sawtell, an incredible experimentalist and collaborator, for many thoughtful discussions and attention to detail throughout my work on the ELL and electric fish models. Greg Wayne and Patrick Kaifosh were my fellow theorists working on the first of the fish projects; both are extremely talented and the work presented in that chapter is also theirs in no small part. Tim Requarth conducted experiments critical to the second fish project, and contributed his thoughts during my work. Daisuke Hattori enthusiastically shared his extensive knowledge of olfactory learning in fly, as well as exciting data and observations from his own research. Richard Axel was a font of critical and engaging discussion of olfactory learning, and his group meetings were invaluable in getting my model off the ground. Peter Wang contributed data from his imaging work and worked with me to bring the mushroom body model closer to biological reality, and Evan Schaffer helped me to get a theoretical handle on computation in the mushroom body. Finally, Wujie Zhang was a very patient and understanding deskmate, and a good friend.

I dedicate this thesis to my family, and the inspiration and support they gave me. Thank you Mom and Dad, for teaching me to code and to think critically about the world. Thank you Dad and Daniel, for these past five years that I've been able to spend away exploring this wonderful field. And Ralf, thank you for your love, for keeping me going and for giving me a life. This thesis never would have happened without you.

Chapter 1

Introduction and background

The cerebellum has been implicated in motor learning [Llinas and Welsh, 1993], adaptive control [Dean, Porrill, Ekerot, and Jörntell, 2010], motor timing [Ivry, 1996], and even routing of sensory information [Bower 1997]. Uniting all of these functions is a characteristic computation performed by its remarkably regular circuit architecture. Because of its interest to broader neuroscience, I will frame this computation as a form of associative learning. In this chapter, I will review associative learning, using a well-studied example from the cerebellar literature for context. I also introduce the two model organisms I studied during my thesis, and the computational role cerebellum-like structures play for each of them.

1.1 Associative learning in the cerebellum

In associative learning, a connection between two stimuli is learned, leading to an adaptive modification of an organism's neural (and often behavioral) response. For example, in classical conditioning, a neutral conditioned stimulus (CS) such as a tone is made to be predictive of a subsequent appetitive or aversive unconditioned stimulus (US), such as food or a shock. Animals that have undergone associative learning will react to the CS in anticipation of the US.

Associative learning is composed of three parts. First, a neural representation of the sensory world (and the CS) must acquire valence via association with a behaviorally meaningful US. Next,

the learned relationship must be stored as a stimulus-independent long-term memory. Finally, subsequent presentations of the CS alone must recall the stored memory and initiate the appropriate behavioral response.

1.1.1 A cerebellar mechanism for associative learning

1.1.1.1 Cerebellar anatomy

The cerebellar cortex is a laminar structure composed of a highly stereotyped repeating arrangement of cells [Eccles, Ito and Szentágothai, 1967]. Sensory information from the brain stem and spinal cord enter the cerebellum via mossy fibers, which form synapses with cells of the deep cerebellar nucleus as well as the intrinsic neurons of the cerebellum, called granule cells. Granule cells are tiny and extremely numerous (there are around 50 billion in the human brain, receiving input from 200 million mossy fibers). Each granule cell has 4-5 dendrites, each of which terminates in a dendritic claw that synapses with a mossy fiber. Granule cells project their axons to the upper, molecular layer of the cerebellum, where they extend as long parallel fibers, forming synapses with the dendritic arbors of Purkinje cells.

Purkinje cell dendrites form intricate, two-dimensional sheets aligned perpendicular to the parallel fibers, allowing them to receive massively convergent input from on the order of 200,000 granule cells. In addition, Purkinje cells receive very strong input from a single climbing fiber. Climbing fiber spikes trigger complex action potentials in Purkinje cells that propagate through the dendrites. Climbing fiber spikes play an important role in cerebellar learning: high climbing fiber firing rates induce LTD at granule cell-Purkinje cell synapses [Sakurai, 1987; Hirano, 1990], while low climbing fiber firing rates induce LTP [Salin, Malenka, and Nicoll 1996]. Purkinje cells axons inhibit the deep cerebellar nuclei (DCN), which form the output of the cerebellum and are involved in gating motor behavior. The DCN also inhibits climbing fiber spikes to gate their control of learning.

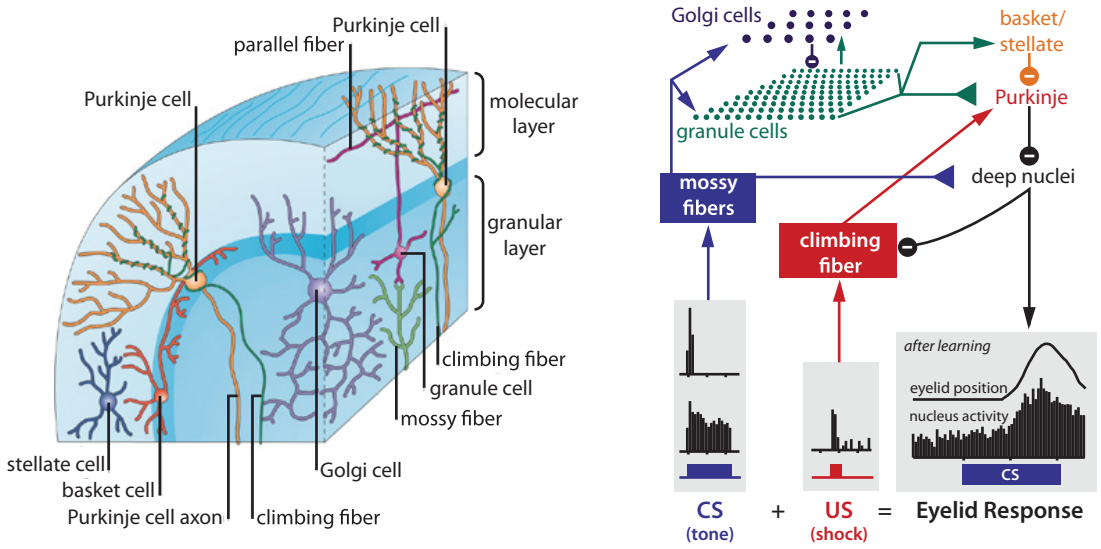


Figure 1.1: **Left:** Basic anatomy of the cerebellum showing mossy fibers, granule cells, Purkinje cells and climbing fibers, reproduced from Dean et al [Dean, Porrill, Ekerot and Jörntell, 2010]. Also shown are basket cells, stellate cells, and Golgi cells, inhibitory interneurons which (with the exception of Golgi cells) I will not discuss here. **Right:** Cerebellar circuitry and its role in classical conditioning of the eyeblink response, reproduced from Medina and Mauk [Medina and Mauk 2000]. Mossy fibers (blue) convey information about the conditioned stimulus, in this case a tone, and the climbing fiber (red) conveys information about the unconditioned stimulus, an aversive air puff to the eye. Changes in Purkinje cell activity during learning alter activity in the deep cerebellar nucleus, which after learning resembles a ramping up of activity that drives an eyeblink at the appropriate time to block the unconditioned stimulus.

1.1.1.2 Cerebellar learning

Eyeblink conditioning is intimately linked with cerebellar inputs and outputs [Gormezano, Schneiderman, Deaux, and Fuentes, 1962]. In this paradigm, playing of a tone is followed by an air puff directed at the eye. At the beginning of training, the air puff drives a reflexive blink response; however after 100-200 trials of pairing, the tone is sufficient to drive an anticipatory eyeblink timed to match the presentation of the air puff. Substantial experimental investigation has determined that mossy fiber inputs to the cerebellum encode information about the tone [Aitkin and Boyd, 1978;

Steinmetz, Lavond, and Thompson, 1985], and climbing fibers encode timing of the air puff [Sears and Stenmetz, 1991; Mauk, Steinmetz and Thompson, 1986; McCormick, Stenmetz and Thompson, 1985], while the output of the DCN is required to drive the conditioned eyeblink response [McCormick, Clark, Lavond, and Thompson, 1982; McCormick and Thompson, 1984].

In a large-scale computational model, Medina and Mauk showed how the components of cerebellar circuitry could give rise to the conditioned response [Medina and Mauk, 2000]. Among mossy fiber inputs to the cerebellum are fibers that respond either to the tone onset, or are active for the duration of the tone. This representation of the tone drives activity of a subpopulation of granule cells receiving input from one or more tone-driven mossy fibers. Because granule cells receive combinations of inputs from mossy fibers encoding a variety of sensory stimuli, their responses are theorized to encode mixtures of stimuli as sparse, high-dimensional patterns of activation, a phenomenon referred to as expansion recoding [Marr, 1969; Albus, 1971].

When the air puff is presented at the end of the tone, climbing fibers fire a burst of spikes. Granule cells that are active just before the climbing fiber burst undergo LTD at their synapses onto Purkinje cells, causing a drop in tone-evoked excitatory drive to Purkinje cells at the time of the airpuff in future trials. Purkinje cells are spontaneously active at high rates, driven by their granule cell inputs, thus a drop in granule cell input leads to a transient pause in Purkinje cell activity. This pause disinhibits the DCN, leading to initiation of a motor response (the conditioned eyeblink). The increase in DCN activity also inhibits the climbing fibers at the time of the air puff, providing a feedback mechanism for termination of learning: when DCN inhibition is strong enough to balance the excitatory drive to climbing fibers elicited by the air puff, climbing fiber modulation of granule cell-Purkinje cell synapses will cease. At the end of learning, the conditioned eyeblink will be a purely feedforward response: rather than being initiated by the air puff, presentation of the tone alone will result in an eyeblink, due to the effect of plasticity on granule cell drive of Purkinje cells.

The learned response is specific to the conditioned tone: other stimuli will activate different sets of mossy fibers, and thus drive different patterns of granule cell activity. Provided there is little overlap between the trained set of granule cells and those activated by another stimulus, the Purkinje cell response to the second stimulus should remain unaltered. Due to the large number of granule cells

synapsing onto Purkinje cells, and the sparseness of each stimulus-evoked pattern of granule cell activity, it is assumed that the correlation between granule cell representations of different stimuli is vanishingly small [Tyrell and Willshaw, 1992; Itskov and Abbott, 2008].

1.1.2 Missing pieces in the learning model

This model relies on the granule cells to provide a sufficient basis for generating a temporally specific Purkinje cell response: for the eyeblink to occur at/before the offset of the tone, excitatory input to the Purkinje cell must drop in a window close to tone offset, meaning there must be granule cells that are active specifically at the end of the tone presentation. The mossy fiber input to the cerebellum has no representation of time, other than marking tone onset, and granule cells do not form synapses with each other, thus ruling out feedforward input and granule cell recurrence as sources of temporal diversity. One hypothesis is that granule cell responses acquire additional temporal diversity via recurrent inhibition from Golgi cells [Medina and Mauk, 2000]. Unfortunately, the technical challenge of recording granule cell responses has left the nature of the granule cell basis uncertain in the cerebellum. In the next chapter, I will present a set of granule cell responses recorded in the electrosensory lobe of mormyrid fish by my collaborator Nate Sawtell, which suggest that another interneuron, the excitatory unipolar brush cell, is responsible for creating the necessary temporal diversity in the mormyrid granule cell population.

A second problem deals with specificity of the conditioned response. If the granule cell representation of stimulus A is highly correlated with that of stimulus B, altering synaptic weights of the stimulus A cells to change the Purkinje cell response will also affect the cell's response to stimulus B. This is not necessarily a bad thing: part of the value of learning is the ability to generalize from previous experiences. However excessive overlap between stimulus representations will make it difficult to form a large number of associative memories. In a related problem, different stimulus conditions have the potential to drive different number of mossy fibers. If some stimuli activate many more granule cells than others, it becomes difficult for Purkinje cells to maintain a constant baseline firing rate across stimuli, making changes in Purkinje cell firing rate due to learning difficult to detect. Marr and Albus hypothesized that recurrent Golgi cell inhibition could normalize granule cell representations of different stimulus conditions, so that the same number of granule

cells are active for any given stimulus and patterns of activation have little overlap [Tyrell and Willshaw, 1992]. This raises the question of how recurrent normalization of stimulus representations affects generalization during learning. Again, the answer to this question lies in the details of how mossy fibers and granule cells encode stimuli, information which has been difficult to obtain through direct experimental study.

Finally, it must be noted that the model of climbing fiber-mediated plasticity at granule cell to Purkinje cell synapses is an oversimplification of cerebellar plasticity. Climbing fiber activity is not always correlated with learning, and in some paradigms learning occurs even when climbing fiber activity is unmodulated [Ke, Guo and Raymond, 2009]. Experimental investigation has turned up multiple sites of learning in the cerebellum, including plasticity at mossy fiber-DCN synapses that is gated by Purkinje cell activity [Boyden, Katoh and Raymond, 2004; Medina, Garcia and Mauk, 2001]. Interestingly, Purkinje cells are known to be recurrently connected into inhibitory networks whose function is unknown; there is evidence that they play a role in cerebellar development [Watt et al, 2009], however their role in the mature cerebellum is unclear. As I will later discuss, recurrent networks and multiple sites of plasticity are also found in the mushroom body and the mormyrid electrosensory lobe, raising the question of whether this conserved anatomy reflects an important second stage of cerebellar learning.

1.2 Negative image formation in weakly electric fish

The weakly electric mormyrid fish uses its active and passive electrosensory systems to navigate its environment. While passive electroreception is not uncommon in fish, the mormyrid is one of a small number of species that also uses electroreception for active sensing, by generating electric fields using a specialized muscle-like organ in its tail. The active system creates two computational challenges for the fish: first, the fish's electric organ creates sensory artifacts that impair the function of the passive electrosensory system. And second, the sensory image of the world produced by the active system is distorted by effects of the fish's posture on the shape of its electric field. The fish solves both of these problems in the electrosensory lobe, an extension of the cerebellum in which granule cell representations of the fish's own movements and electric organ discharge commands

are used by Purkinje-like cells to construct "negative images" that cancel self-generated effects in the electrosensory system.

1.2.1 Anatomy and origin of the passive and active electrosensory systems

The mormyrid fish I study here (*Gnathonemus petersii*) has two distinct sensory systems that respond to electric fields. The first of these, called the passive system, is shared with lampreys and early fish, including sharks, rays, lungfish, and sturgeons. (Dolphins and monotremes also have a passive electroreceptive system, but this evolved separately.) In early vertebrates, skin-bound electrosensory structures called ampullary organs are thought to have evolved from the lateral line system (which senses water currents and vibration) to detect the weak, low-frequency electric fields generated by the bioelectric processes of other aquatic organisms. Ampullary organs are flask-shaped ducts filled with conducting jelly, that open to the water via a pore and are lined at the base with hair-cell-like ampullary cells. Ampullary cells respond to external field frequencies from DC up to 50 Hz, encoding the amplitude of fields in their firing rates. [Baker, Modrell, and Gillis 2013; Wilkens and Hofmann 2005]

Ampullary organs were lost in teleosts, and are not expressed in most modern fish, but reemerged independently in two groups: Gymnotiformes in South America, and Mormyriformes in Africa. In these two groups, reemergence of ampullary organs was followed (or possibly accompanied) by development of the electric organ. In mormyrids, the electric organ is a muscle-like structure composed of electrocytes, stocky cylinder-shaped electrically excitable cells arranged in series. Both faces of electrocytes are innervated by a pool of spinal motorneurons; a burst of spikes in the motorneuron pool evokes a single simultaneous discharge on one face of each electrocyte; this discharge subsequently evokes a second discharge on the opposite face, reversing the flow of current through the electric organ to give the electric organ discharge its characteristic biphasic shape. (The biphasic waveform reduces the DC component of the EOD waveform, an adaptation that reduces interference with the passive system, and may also reduce visibility to predators.) [Hopkins 1980]

Development of the electric organ was accompanied by formation of a second electrosensory structure, the tuberous organs. Similar in structure to ampullary organs, but with some differences in

shape and a loose plug of epithelial cells covering their pore, tuberous organs are specialized for detecting signals in the higher frequency range of the electric organ discharge. Together, the electric organ and tuberous organs constitute the second, active electrosensory system, so called because the fish uses its electric organ-generated field as read out by the tuberous organs to actively probe its environment. Thus, unlike the passive system, the active system allows detection of objects that don't generate fields on their own. In mormyridae, tuberous organs are further divided into two types: mormyromasts, used for active electrolocation, and knollenorgans, used for communication; I will focus only on the former here. Unlike ampullary cells, mormyromasts encode the amplitude of the local electric field using a latency code, although this signal is converted to a combination of latency and amplitude encoding in the central nervous system. [Kawasaki 2005]

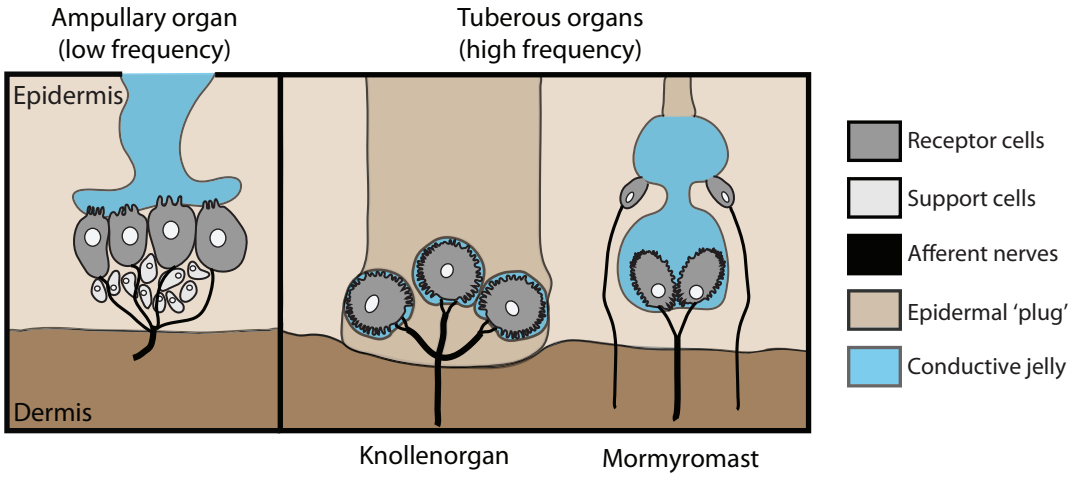


Figure 1.2: Receptors of the passive and active electrosensory systems. Ampullary cells (left) and mormyromasts (right) will be addressed here.

1.2.2 Function of the passive and active electrosensory systems

In some fish, passive electroreception is also thought to aid in long-distance navigation via detection of the earth's magnetic field [Kalmijn 1974; Collin and Whitehead 2004]. More commonly, the passive electrosensory system is used for prey localization. Aquatic organisms generate weak electric fields due to muscle and nerve activity, ventilation, and osmoregulation [Bodznick, Montgomery

and Bradley 1992]. The low-frequency detection range of ampullary cells is tuned to detect these signals, called bioelectric fields, which are typically DC or low frequency fields.

From a distance, the electric field generated by a fish can be approximated by a dipole, and field strength drops with distance cubed [Bodznick and Montgomery 2005]. Unlike sound waves, there is little propagation delay in generated electric fields, and the wavelength of low frequency fields generated by organisms is extremely large, making direct source localization difficult; instead, the emphasis of the electrosensory system is on detection of very small fluctuations in field intensity [Hopkins 2005]. Behavioral analysis in weakly electric fish suggests that they are unable to directly localize the dipole source of a generated field, but rather that they find field sources by moving parallel to the local electric field vector, and follow this path to the field source– a behavior referred to in early literature as galvanotaxis [Fraenkel and Gunn, 1940; Hopkins, Shieh, McBride and Winslow 1997].

Active electroreception is so called because it is an active sensing system, one in which self-generated energy is used to probe the surrounding environment. When the fish activates the muscle of its electric organ, it generates a transient, high-frequency electric field around its body. Objects near the fish with different conductivity than the surrounding water distort the amplitude of the fish's generated field. Conducting objects have lower impedance and thus more current flow than the water around them, leading to a higher current density at the fish's skin, making the EOD amplitude appear larger. Nonconducting objects have the reverse effect, with lower current flow than water leading to a lower current density at the skin, and smaller observed EOD amplitude.

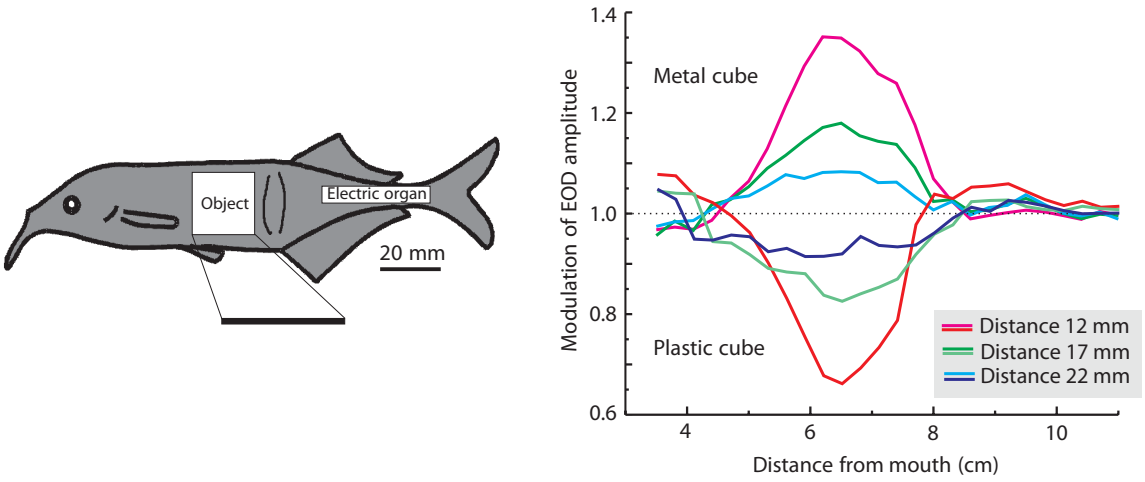


Figure 1.3: Electrosensory images of insulating (plastic) and conducting (metal) objects measured at the fish's skin; reproduced from von der Emde [von der Emde 1999].

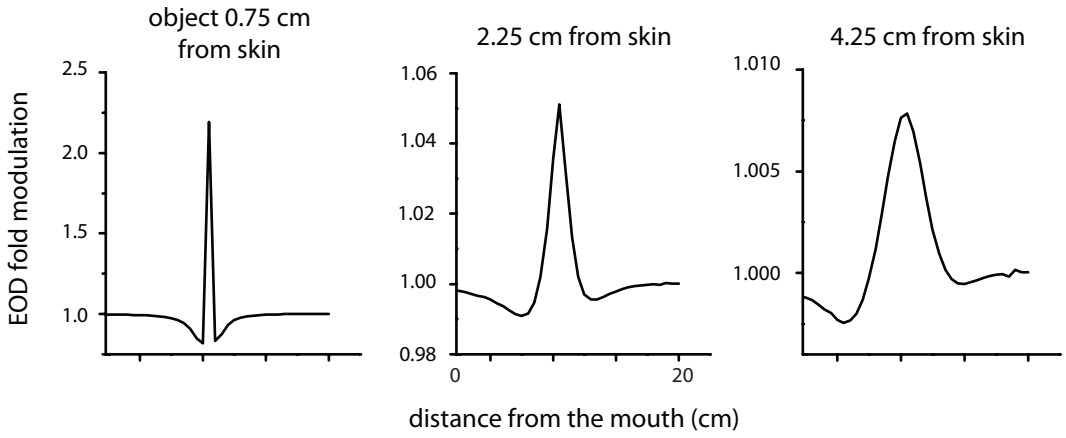


Figure 1.4: Effect of a resistive object on EOD field modulation at the skin, from a 2d electric circuit model of the fish's environment; reproduced from Caputi et al [Caputi, Burdelli, Grant and Bell 1998]. Both the amplitude and shape of the object image change with distance from the skin.

In behavioral studies, fish have been found to discriminate objects based on their conductivity [Lissmann and Machin 1958], capacitance [von der Emde 1990], distance, and shape [von der Emde et al, 1998]. These properties can be extracted from the amplitude, slope, and size of the electrosensory image.

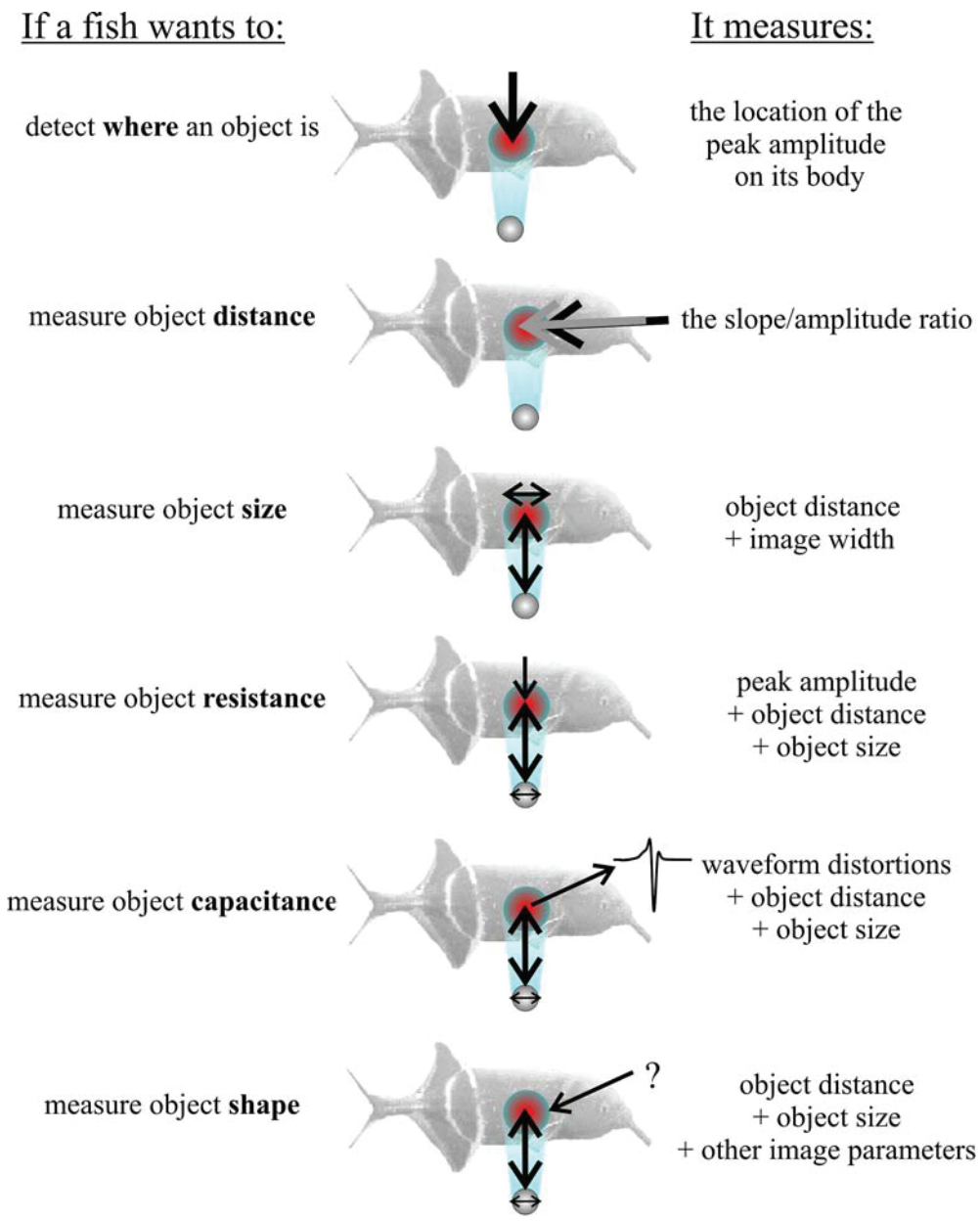


Figure 1.5: Proposed mechanisms for feature detection in active electrosensation, reproduced from von der Emde [von der Emde, 2006].

The fish's field evokes a burst of spikes in mormyromasts at short latency to the EOD. Modulation in field amplitude at the skin change the latency of spikes in mormyromasts, which are tuned to detect signals in the frequency range of the EOD [Bell 1990].

1.2.3 Representation of electrosensory information in the central nervous system

The first stage of electrosensory processing in the central nervous system is the electrosensory lobe (ELL), a six layer structure that receives input from all three types of electroreceptors [Meek, Grant, and Bell 1999]. Inputs from the passive and active system are segregated into different anatomical regions of the ELL: ampullary cells project to the ventrolateral zone, and mormyromasts to the dorsolateral and medial zones.

Sensory input to the ELL from mormyromast afferents is processed by an elaborate and poorly-understood layer of interneurons, depicted in the bottom half of Figure 1.6. The computational role of these interneurons is unknown, as many have only been identified anatomically, however they seem to transform the sensory input from mormyromasts from a spike latency code into a combination of a firing rate and latency code. The transformed sensory input is ultimately received by the basal dendrites of four key intrinsic cells of the ELL, the Large Ganglionic (LG), first and second types of Medium Ganglionic (MG_1 and MG_2), and Large Fusiform (LF) cells; the first two of these cells are "off" cells that are inhibited by sensory input, while the second two are "on" cells that are excited by sensory input. The LG and LF cells form the two main outputs of the ELL, while the two MG cell types are inhibitory interneurons that stabilize the responses of LG and LF cells. In addition to sensory input on their basal dendrites, all four cell types have extensive apical dendrites that form 10,000 (for LG and LF cells) to 20,000 (for MG cells) synapses with parallel fibers from the eminentia granularis posterior (egp), a layer of granule cells that the ELL shares with the cerebellum. This input plays an important role in sensory processing, as will be discussed in the next section.

In the active system, the four key intrinsic cells have center-surround receptive fields on the skin, with center widths averaging around 5.5-6 mm, and max widths averaging 8-11.2 mm; on cells have slightly larger centers on average and do not always have an inhibitory surround [Metzen et al 2008]. Receptive field size does not seem to vary along the length of the body, or possibly grows slightly narrower towards the tail. Mormyromast density varies along the length of the fish, from 65 mormyromasts/mm^2 at the tip of the chin organ (which operates as an electric fovea) to 10/mm^2 at the head, down to less than 2/mm^2 at the back of the fish [von der Emde et al 2008]. Thus the

number of mormyromasts pooled by an intrinsic cell varies with the location of the cell's receptive field.

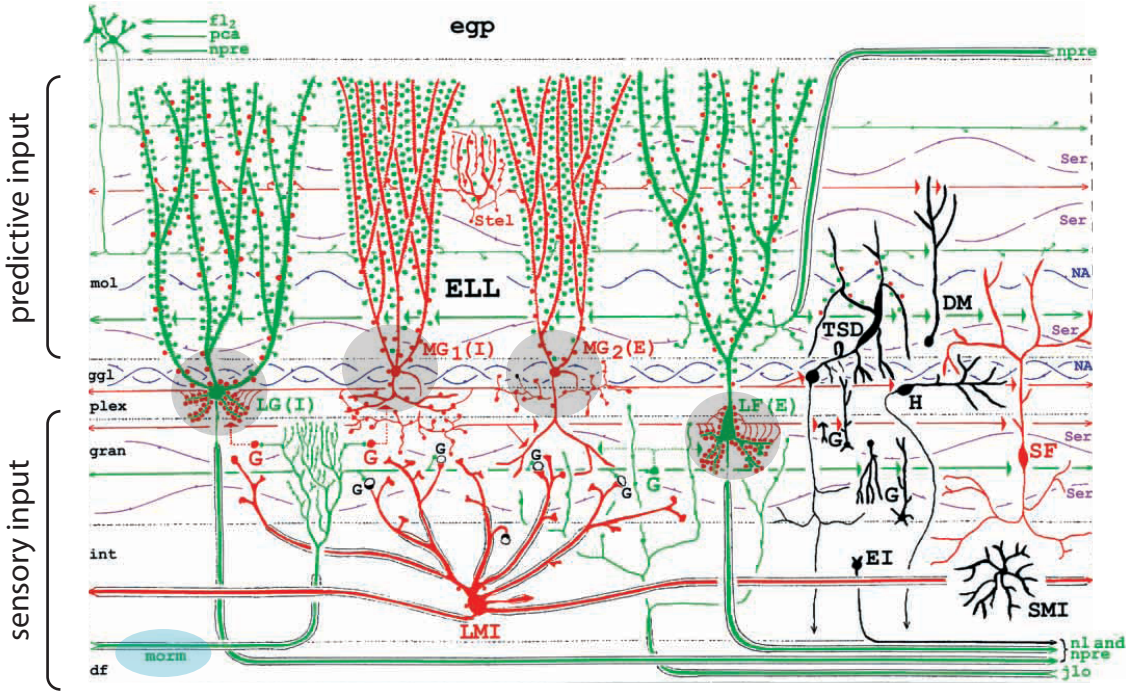


Figure 1.6: Intrinsic circuitry of the ELL showing 14 identified cell types, most of which are interneurons, reproduced from Meek et al [Meek, Grant, and Bell 1999]. Inhibitory cells are shown in red, and excitatory cells in green; cells with unknown transmitters are indicated in black. Many of the interneurons have not been characterized extensively, and will not be discussed here. The four cells highlighted in gray operate akin to Purkinje cells in the cerebellum: from left to right, they are the Large Ganglionic (LG), first and second types of Medium Ganglionic (MG_1 and MG_2), and Large Fusiform (LF) cells. LG and MG_1 cells are inhibited by sensory input (as indicated by the (I) next to their names), while MG_2 and LF cells are excited by sensory input. Highlighted in blue is the mormyromast afferent that relays sensory information from the periphery.

1.2.4 Subtraction of self-generated signals in the electrosensory lobe

The fish's own actions can hinder the performance of its active and passive electrosensory systems. In the passive system, the electric organ discharge generates large, ringing fluctuations of ampullary

cell firing rates that last on the order of 200 ms [Bell and Russell, 1978]. In the active electrosensory system, changes in the fish's posture alter the amplitude of the EOD field at the fish's skin [Sawtell and Williams, 2008]. A major function of the ELL is to cancel these self-generated artifacts, via the widely-employed mechanism of corollary discharge [Bell, 1981; Requarth and Sawtell, 2014]. Corollary discharge is a system of motor-to-sensory feedback, proposed as a mechanism for distinguishing self- versus externally-generated sensory input [von Holst and Mittelstaedt, 1950; Sperry, 1950]. By keeping track of motor commands, the nervous system can inform the sensory processing stream about the effects of movements, allowing self-generated inputs to be filtered or ignored by the sensory system [Crapse and Sommer, 2009].

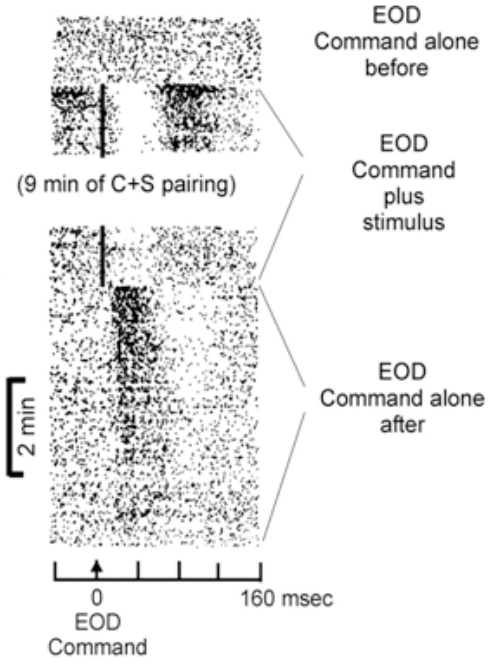


Figure 1.7: Negative images of externally driven MG cell spiking following a period of stimulation, reproduced from Bell [Bell, 1981]. At the start of the experiment, the fish's EOD command is artificially paired with an externally applied electrical stimulus. While MG cell spiking is initially strongly affected by the stimulus, the response is reduced following nine minutes of pairing with the EOD. At the end of pairing, the stimulus is turned off, and a temporally specific negative image of the paired signal can be seen in the MG cell firing rate.

When the fish discharges its electric organ, it relays a corollary discharge signal from the electric or-

gan command nucleus to the ELL, via a population of mossy fibers originating in nuclei downstream of the command nucleus [Meek, Grant, and Bell, 1999]. Like the cerebellum, the four Purkinje-like intrinsic cells of the ELL (which I will refer to as efferent cells) receive massive convergent input from a population of granule cells, which each receives input from a small number of mossy fibers (granule cell-efferent cell synapses are indicated by green dots in Figure 1.6.) Like Purkinje cells, efferent cells fire both normal action potentials and a second response called a broad spike; rather than being driven by a climbing fiber as in the cerebellum, efferent cell broad spikes are evoked by input from the electrosensory system onto the cell's basal dendrite. If broad spikes evoked by fluctuations in sensory input are time-locked to the EOD, they will, over time, drive synaptic depression in granule cell-efferent cell synapses [Bell, Han, Sugawara and Grant, 1997], reducing granule cell input within a temporally-specific window. Similarly, a drop in broad spike rates that is time-locked to the EOD will drive potentiation of granule cell synapses. These two effects drive the formation of stable, temporally-specific negative images of self-generated sensory input, such that when granule cell and electrosensory input are summed in the efferent cells, only sensory signals not predicted by the corollary discharge signal remain [Roberts and Bell, 2000; Williams, Roberts, and Leen, 2003].

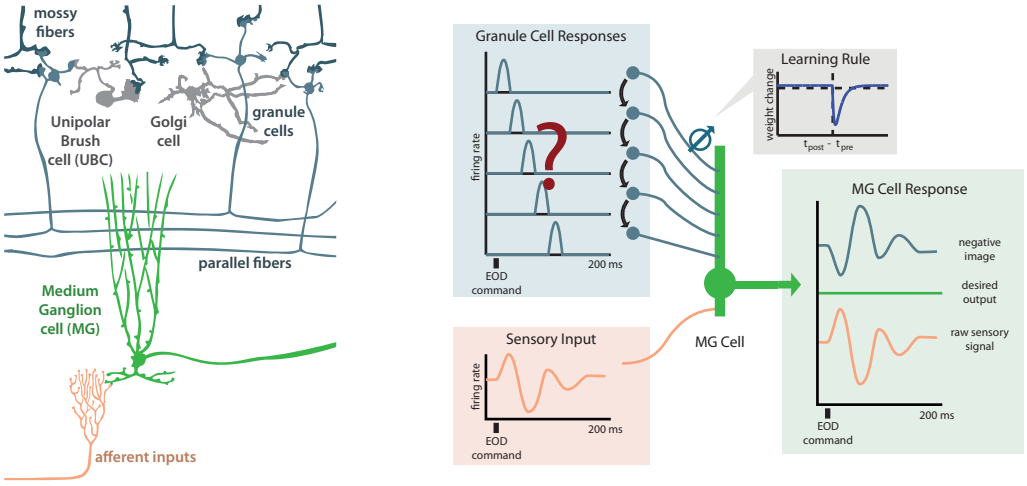


Figure 1.8: **Left:** Anatomy of the ELL, this time highlighting the inputs to efferent cells that underlie their role in negative image formation. This circuit will be discussed in more depth in the next two chapters. **Right:** Proposed mechanism for cancellation of self-generated sensory artifacts in the passive electrosensory system, adapted from Roberts and Bell [Roberts and Bell, 2000]. The granule cells form a set of temporal basis functions (here shown as a hypothesized delay line), that are sculpted via anti-Hebbian plasticity at their synapses onto efferent cells to form a negative image of any sensory input that is time-locked to the EOD. This model will be discussed further in the next chapter.

While the learning mechanism for negative image formation is well understood, the representation of corollary discharge information by the granule cell population had not been previously explored: early models relied on idealized assumptions of granule cell responses, such as the delay line portrayed in the right panel of Figure 1.8. In the following chapters, I will discuss two forms of information represented by the granule cells—temporally-expanded encoding of the EOD command, and proprioceptive signals—and how the representation of these signals impact negative image formation by the ELL.

1.3 Associative odor learning in insects

In chapters 4 and 5 of this thesis, I will discuss another structure involved in associative learning, the mushroom body of the insect olfactory system. Remarkably, the anatomy of the mushroom body and its readouts again resembles the architecture of the cerebellum [Farris, 2011]. In these chapters, I discuss how the canonical associative learning circuit of the cerebellum can be mapped onto the elements of the mushroom body, and investigate this model as a mechanism for formation of odor-specific associative memories. Because I focus extensively on the anatomy of the insect olfactory system in later chapters, I restricted this section to a review of associative conditioning and olfactory memory in insects.

1.3.1 Classical conditioning with olfactory stimuli in insects

Aside from flies, olfactory learning and memory has been studied in a diverse group of insects, including honeybees [Giurfa and Sandoz, 2011; Menzel, Erber, and Masuhr, 1974; Müller 2002], locusts [Laurent and Naraghi, 1994] and cockroaches [Mizunami, Iwasaki, Nishikawa, and Okada, 1997]. Classical conditioning is often studied using the proboscis extension reflex, in which an olfactory stimulus is followed by a sugar reward; after pairing, the insect learns to anticipate the reward and will extend it proboscis upon presentation of the olfactory stimulus [Takeda 1961]. A second approach that has seen considerable success in flies is conditioned odor avoidance in a T-maze, illustrated below [Jellies, 1981; Tully and Quinn, 1985].

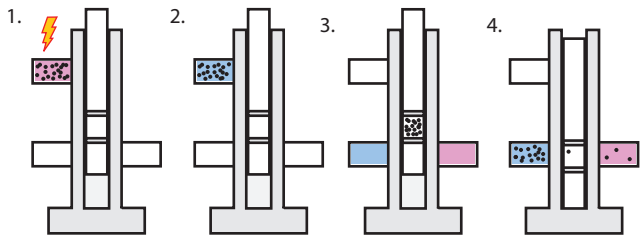


Figure 1.9: Flies placed in one arm of a T-maze are presented with one of two behaviorally-neutral odors, one of which is paired with a shock (panel 1), and the other not (panel 2). Flies are then moved to the choice point of the T-maze and exposed to both odors *(continued on following page)*

Figure 1.9: *(continued)* (panel 3); flies with working associative memory systems will avoid the arm with the odor that has been paired with shock. Behavior is measured as the number of flies either arm of the maze (panel 4).

In addition to conditioned odor attraction and avoidance, several variations of classical conditioning have been studied in insects, particularly the honeybee. In one such example, the subject is first presented with odor A paired with shock or reward, until a learned association is formed. The subject is then presented for a comparable number of trials with a mixture of odors A and B, followed by the same shock or reward. After pairing, subjects presented with odor B alone will not exhibit the conditioned response, because they have not learned to associate odor B with shock/reward. This effect is called blocking. In honeybees, there is some evidence that blocking between odors in binary mixtures exists [Smith and Cobey, 1994], although it is not found in all circumstances [Gerber and Ullrich, 1999]; the most recent evidence indicates that the degree of olfactory blocking that occurs is proportional to the similarity between two odors in a mixture [Guerrieri, Lachnit, Gerber, and Giurfa, 2005b].

Odor mixtures have also been used to study compound processing, a behavioral take on the XOR problem [Rescorla and Wagner, 1972; Pearce 1987]. Odor mixtures could be represented in the brain in two ways: as a sum of their constituent elements, or as a unique sensory object. In olfactory conditioning experiments, honeybees were able to learn both "negative patterning", in which a single odor is rewarded, but not a mixture, and "positive patterning", in which mixtures were rewarded and individual odors were not [Deisig, Lachnit, Giurfa, and Hellstern, 2001]. An investigation using pairs and triplets of odors found that the bee's performance could be matched by a model in which mixture representations were given by the sum of the representations of the mixture elements, plus an extra term unique to the mixture [Deisig et al 2003].

1.3.1.1 Generalization of conditioned responses

In both flies and fish, I will discuss generalization and how it arises from stimulus representations. Generalization is rarely studied in classical cerebellar learning experiments, although it does occur

in some instances. For example, in eyeblink conditioning, altering the conditioned stimulus, for instance by changing the pitch of the tone, will still elicit the conditioned response [Michael Mauk, private communication]. Other changes of the conditioned stimulus, such as replacing the tone with a light or changing the duration of the tone, fail to elicit the conditioned response.

Von Frisch observed that freely behaving bees trained to associate an odor with food reward tended to generalize this association to other odors that, at least to humans, smelled similar [von Frisch 1919]. Systematic study of generalization using the proboscis extension reflex has confirmed that while the reflex is evoked most strongly by a conditioned odor, bees do generalize the reflex behavior to chemically similar odors- see Figure 1.10.

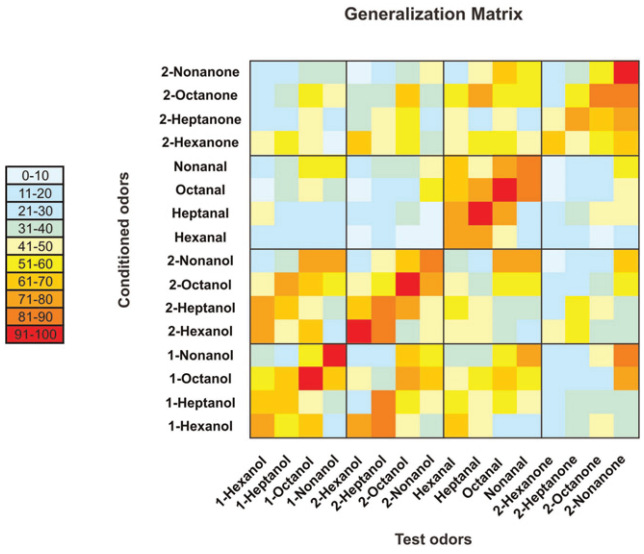


Figure 1.10: Probability of generalization of the conditioned proboscis extension reflex by honeybees, reproduced from Guerrieri et al [Guerrieri, Schubert, Sandoz, and Giurfa, 2005a]. Bees were conditioned to produce the proboscis extension reflex to one of a panel of odors (vertical axis), and then tested with the remaining odors (horizontal axis); the probability that a tested odor elicited a conditioned response is indicated by color. Odors are arranged into chemically similar groups (divided by black lines)– the roughly block-diagonal structure of the matrix suggests that bees were more likely to respond with the conditioned response to chemically similar odors.

A conditioned response should presumably also be preserved when the conditioned odor is presented at different concentrations, or in different mixtures of odors. Conversely, insects should be able

to refine the specificity of their behavior: if odor A is consistently associated with reward/shock, but odor B is not, the insect should be able to learn to respond only to odor A. In chapters 4 and 5, I will discuss how generalization of odor response could relate to their representation in the mushroom body, and discuss learning rules that lead to stronger or weaker generalization across odors.

1.3.2 Computational role of the mushroom bodies

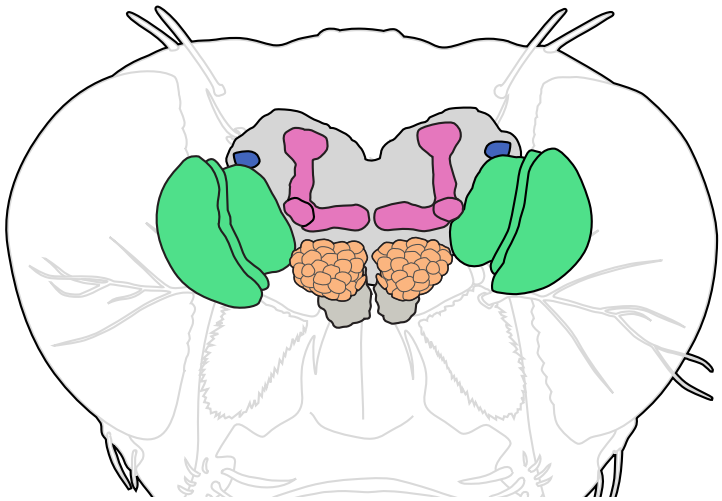


Figure 1.11: Anatomy of the drosophila central nervous system showing prominent structures, adapted from Heisenberg [Heisenberg, 2003]. The mushroom bodies are shown in pink, optic lobes in green, and antennal lobes in orange. A second key structure of the olfactory processing stream, the lateral horn, is indicated in blue.

The mushroom bodies are large bilateral structures found in the brains of most insects. They are the first central stage of olfactory sensory processing, receiving input from the antenna lobe (the insect equivalent of the olfactory bulb). While predominantly studied in the context of olfaction, the mushroom body in some insects also receives input from visual, gustatory, mechanosensory and proprioceptive systems [Farris, 2011], and has been implicated in tasks as diverse as spatial learning [Mizunami, Weibrecht, and Strausfeld, 1998], temperature preference behavior [Hong et al 2008], and conditioning of courtship behavior [McBride et al, 1999].

1.3.2.1 Genetic methods for investigating circuit function in drosophila

Drosophila neuroscience has benefited immensely from the sophisticated genetic methods available for manipulation of neural activity. In chapter 5 I will discuss one such tool, the GAL4-UAS transgene system, that allows extremely precise genetic labeling of small populations of neurons. Unlike mammals, computation in the central nervous system of flies is typically mediated by small sets of genetically distinct neurons; many neurons in the fly can be individually identified (the intrinsic cells of the mushroom body, called Kenyon cells, are a rare exception.)

Another genetic tool that has seen great success when used in conjunction with GAL4 lines is the temperature-sensitive dynamin allele *shibire* [Kitamoto, 2001]. Dynamin is a protein involved in synaptic vesicle recycling; the *shibire* allele ceases to function when flies are exposed to elevated temperatures (around 29° C), causing neurotransmission to cease within 1-2 minutes in all cells expressing the *shibire* allele. Restoration to normal temperatures quickly reverses this effect. Selective expression of *shibire* in a subset of neurons thus allows their function to be studied during specific stages of learning and memory.

1.3.2.2 Evidence for involvement of the mushroom bodies in associative learning

Chemical ablation of the mushroom bodies in drosophila larva yields adult flies that behave normally, but are unable to learn conditioned odor avoidance in the T-maze task [de Belle and Heisenberg, 1994]. This result was confirmed by a gentler study in which disruption of synaptic transmission in the mushroom body using the *shibire* transgene specifically impaired retrieval of memories, but not memory formation [Dubnau, Grady, Kitamoto and Tully, 2001; McGuire, Le, and Davis, 2001]. Several drosophila mutants showing impaired memory performance, such as *dunce*, *rutabaga*, *DC0*, *PKA-RI*, *leonardo*, *Volado*, *fasciclinII*, and *pumilio* have gene products that are preferentially expressed within the mushroom bodies [Davis, 2005].

Refinement of genetic targeting to subsets of neurons in the mushroom body suggests that different populations of Kenyon cells, the granule-cell-like intrinsic neurons that represent odors, are involved in memory acquisition, consolidation, and retrieval [Krashes et al 2007]. Shibire-mediate disruption of neurons peripheral to the mushroom body, including a giant serotonergic neuron (DPM) [Keene

et al 2004; Tanaka, Tanimoto, and Ito 2008] and a giant gabaergic neuron (APL) [Wu et al 2011], has implicated them in learning and memory as well. I will discuss the hypothesized role of mushroom body neurons in associative memory formation in depth in chapter 5.

Chapter 2

Predicting the sensory consequences of motor commands in the mormyrid passive electrosensory system

Weakly electric mormyrid fish emit brief EOD pulses for communication and active electrolocation. However, the fishs own EOD also affects passive electroreceptors tuned to detect external fields. Previous studies have shown that such interference, a ringing pattern of activation that may persist for as long as the interval between EODs [Bell and Russell, 1978], is cancelled out in MG cells through the generation of motor corollary discharge responses that are temporally-specific negative images of the sensory consequences of the EOD [Bell, 1981]. Elegant theoretical studies [Roberts and Bell, 2000; Williams, Roberts, and Leen, 2003] have suggested that anti-Hebbian spike timing-dependent plasticity known to exist at synapses from GCs onto MG cells [Bell, Han, Sugawara and Grant, 1997] could provide a basis for negative image formation, but this work depends on the untested assumption that GC corollary discharge responses exhibit a rich temporal structure spanning the approximately 200 ms period over which negative images can be generated [Bell 1981; Bell, 1982; Bell, Caputi, Grant, and Serrier, 1993] (Fig. 1a). GCs, located in the eminentia granularis posterior (EGp) overlying the electrosensory lobe (ELL) molecular layer, receive excitatory input from extrinsic mossy fibers (MFs) originating from neurons in a number of brain regions and

from UBCs located within EGp itself (Fig. 1b). Though there are a small number of published recordings of delayed corollary discharge responses from unidentified elements in the EGp itself [Bell, Grant, and Serrier, 1992], corollary discharge responses of MFs appear to be extremely brief and minimally delayed, resembling literal copies of the EOD motor command [Bell, Grant, and Serrier, 1992; Bell and von der Emde, 1995; Sawtell, Mohr, and Bell 2005; von der Emde and Bell, 1996]. Moreover, delayed or temporally diverse corollary discharge responses have not been reported for GCs. Therefore, we set out to determine: 1) whether delayed and temporally diverse GC responses exist and, if they do, 2) how they are generated and 3) if they are sufficient to support negative-image formation.

As in previous studies, we take advantage of an awake preparation in which fish continue to emit the motor command to discharge the electric organ, but the EOD itself is blocked by neuromuscular paralysis, allowing corollary discharge responses, i.e. neural activity in sensory areas that is time-locked to the EOD motor command, to be studied in isolation from sensory effects.

2.1 Methods

2.1.1 Classifying Mossy Fibers

The diversity of mossy fiber responses readily suggests division into four categories based on the timing and reliability of their EOD-triggered responses. We named these categories early, medium, late and pause, after the timing of their activity relative to the EOD. Two of these categories (early and medium) have responses which resemble those of cells recorded in two extrinsic sources of mossy fibers, PCA and PE. The other two categories (late and pause) do not resemble any known extrinsic regions, but do show strong similarity to responses recorded from unipolar brush cells, an interneuron located in EGp which also produces mossy fibers that synapse onto granule cells.

To make the mossy fibers classes more concrete, we fit a multinomial logistic regression model to the EOD-triggered responses of cells recorded in PCA and PE, and non-pausing UBCs, and used this model to assign labels of early, medium, or late to the recorded mossy fibers (the pause mossy fibers were hand-classified, as their responses were clearly distinct from the other three.) To train

the model, we took a set of 12 cells recorded in PCA, 28 from PE, and 10 UBCs, labeling these as early, medium, and late respectively. We tested several methods of parameterizing cell responses, and arrived at a set of three parameters which minimized the classifier error on holdout data (2.67% error in a set of 10 cells.) Specifically, these parameters were:

1. Time of first rise: the fist time relative to the EOD at which the smoothed, trial-averaged firing rate of the cell achieves 75% of its maximum rate.
2. Half-width of response: the width of the first peak for which the cell's smoothed, trial-averaged firing rate is above 50% of its maximum rate.
3. Spiking variability: the total variance of the cell's spike times across trials and time.

2.1.2 Data Collection and Model Setup

We collected data from 135 mossy fiber and 170 granule cells. For each cell we have two simultaneously-recorded channels: the first channel contains either spike times of an extracellularly-recorded mossy fiber or the membrane potential of a patched granule cell, and the second channel contains EOD command times recorded from the fish's command nucleus. Mossy fiber spiking was recorded at 40kHz, and granule cell membrane potentials at 20kHz. For the fitting of mossy fiber inputs to granule cells, we interpolated both sets of recordings up to 200kHz, so as to more accurately simulate our synaptic filters. During simulation of synthetic granule cells, we precomputed the convolution of synaptic and membrane filters, and were able to relax our sampling rate back to 20kHz.

For the purposes of this study we will only consider granule cell responses to single EOD commands. For each recorded mossy fiber and granule cell, we identified EOD command events for which there were more than 200 milliseconds following and preceding the command. These are termed *well-isolated* commands. We broke the mossy fiber/granule cell recordings into segments occurring from 25 milliseconds before to 200 milliseconds after each well-isolated command, and averaged across segments from each cell to obtain its command-triggered average response. Both mossy fiber and granule cell responses to the EOD command are highly stereotyped, and granule cell responses resemble weighted sums of a small number of mossy fiber inputs. This observation is consistent with the anatomy of granule cells, which have a small number of dendritic claws, each of which

forms a microglomerulus with a single mossy fiber.

To study negative image formation on the scale of an MG cell, we sought to generate 20,000 model granule cells with response properties similar to experimentally-recorded granule cells. We first constructed a current-based integrate-and-fire model granule cell and selected inputs and synaptic weights to fit its command-triggered average membrane potential to the responses of recorded granule cells. Model cell inputs were selected from among the population of recorded mossy fibers, allowing up to three inputs to each model granule cell and constraining input weights to be non-negative, reflecting the strictly excitatory nature of mossy fiber - granule cell synapses. The distribution of inputs from the four mossy fiber classes to the model fits of recorded granule cells was consistent with a random mixing model in which each input to the model granule cell was selected independently from the pool of recorded mossy fibers, with a fixed probability of input selection from each of the four mossy fiber classes. We then used the mossy fiber class-dependent input probabilities and synaptic weights to generate 20,000 synthetic granule cells. Distributions of spikes per EOD command and time of peak EOD-triggered membrane potential were not statistically different in recorded and generated cells, supporting the conclusion that the response properties of generated cells were consistent with those of the recorded cells. We therefore assert that the model granule cells we developed provide a representative example of the pool of granule cells available to an MG cell as temporal basis functions for negative image formation.

2.1.3 Fitting the Model to Recorded Granule Cells

Inputs to our model granule cell were selected from the population of recorded mossy fibers so as to fit the model's response to the trial-averaged membrane potentials of each of 170 recorded granule cells. We notionally separate the problem of fitting into the two problems of 1) finding a likely set of mossy fiber inputs from our population and then 2) determining the weights of the slow and fast components of each model synapse. Fitting is constrained by the fact that granule cells receive a small number of inputs (here restricted to 3 or fewer), all of which are excitatory.

A subpopulation of mossy fibers, all of them late-spiking, did not spike on every trial. Similarly, a small number of recorded granule cells had late EPSPs on a fraction of the total recorded trials; these late EPSPs often had an impact on the cell's spiking, but because they were infrequent they were

difficult to detect in the cell's trial-averaged membrane potential. To prevent underestimation of the synaptic weight on these late inputs, we discarded non-spiking trials when computing the trial-averaged responses of mossy fibers, as well as trials without late input in the affected granule cells. During simulation of synthetic granule cells with late input, non-spiking trials were included.

2.1.4 Finding Synaptic and Membrane Time Constants

Granule cells were modeled as a single-compartment current-based leaky integrate-and-fire neuron receiving 1-3 excitatory inputs from recorded mossy fibers. The mossy fiber-granule cell synapse was modeled as a weighted sum of two exponential filters, a fast filter ($\tau_{\text{fast}} = 0.2$ms) and a slow filter ($\tau_{\text{slow}} = 37.8$ms), where the weights on the fast and slow filters were fit independently. This combination of timescales allowed us to fit both the fast rise of granule cell EPSPs and the observed integration of tonically-spiking proprioceptive and pause inputs. The values of the two time constants fit were consistent with other models of granule cell response properties [Schwartz et al 2012].

We hand-fit the fast synaptic time constant ($\tau_{\text{fast}} = 0.2$ms) to match the EPSP width in granule cells receiving input from early mossy fibers: these inputs spiked in high-frequency bursts ($\sim$600Hz) but individual EPSP peaks could still be readily resolved even in the command-triggered average granule cell responses. This suggests a fast synaptic time constant of below 1ms; the chosen value of 0.2ms best fit the shape of the fast early EPSPs.

To fit the slow synaptic time constant, we took single-trial traces recorded from granule cells receiving input from a single mossy fiber, and inferred the input spike train semi-automatically by low-pass filtering the second derivative of the membrane potential to identify inflection points arising from EPSPs; because EPSPs were large and temporally sparse, these points could be identified with high confidence. Using the inferred spike train as input to our integrate-and-fire model, we constructed the slow and fast filtered components of the synaptic input for fixed τ_{slow}, and weighted the two components by least-squares to fit the recorded granule cell trace. We repeated this for a range of τ_{slow} in a total of 360 traces from 10 granule cells, and chose for each trace the τ_{slow} which minimized the mean squared error of the fit. The error of the least-squares fit for each trace had a definite minimum with respect to the time constant τ_{slow}; averaging across the 10 fit cells gave our

τ_{slow} of 37.8ms.

The membrane time constant of the recorded granule cells was determined experimentally to be $\tau_{\text{membrane}} = 8.7\text{ms}$.

2.1.4.1 Input selection

Restricting the number of mossy fiber inputs to our model granule cell would ideally be accomplished via L^0 optimization, which constrains the number of active (nonzero) inputs, but such an approach is in general computationally intractable. As a work-around, we use a more tractable L^1 optimization problem to cut our collection of mossy fibers down to a small number of candidate inputs; we may then solve the L^0 problem on this restricted pool through brute force. Defining the membrane potential of the granule cell we are trying to fit as $v(t)$, we seek up to three mossy fiber inputs $x^i(t)$ and weights w^i_{fast} and w^i_{slow} that minimize:

$$\begin{aligned} C(w^i_{\text{slow}}, w^i_{\text{fast}}) &= \frac{1}{2T}\sum_{t=0}^{T}\left((v(t)-\bar{v}) - \Big(\sum_i w^i_{\text{slow}}(h^i_{\text{slow}}(t)-\bar{h}^i_{\text{slow}}) + \sum_i w^i_{\text{fast}}(h^i_{\text{fast}}(t)-\bar{h}^i_{\text{fast}})\Big)\right)^2 \\ &\quad +\alpha_1 \cdot \left(\sum_i |w^i_{\text{slow}}| + \sum_i |w^i_{\text{fast}}|\right). \end{aligned}$$

Where $h^i_{\text{slow/fast}}(t) = g_{\tau_{\text{membrane}}} * g_{\tau_{\text{slow/fast}}} * x^i(t)$, the inputs and outputs have been mean-subtracted, and T is the length of the trial-averaged membrane potential (here 225 ms sampled at 20kHz.) The value of α_1 determines the magnitude of the L^1 penalty on fits; we choose α_1 such that we find up to 10 candidate mossy fiber inputs. We implement the constraint that the weights must be non-negative using an adjusted Least-Angle Regression (LARS) solution of the LASSO problem that selects only inputs which are positively correlated with the target granule cell trace.

2.1.4.2 Weight fitting

After using LARS to select up to 10 candidate inputs for each granule cell, we refine the input weights to produce our final fit. We did not use LARS to choose the 1-3 mossy fiber inputs directly, because sparse solutions to the cost function above are dominated by the L^1 penalty term, which reduces the number of fit inputs at the cost of overly penalizing the input weight magnitudes and

therefore yields poor quality fits. By instead using LARS to reduce the pool of candidate inputs to 10, we can solve our original L^0 optimization problem on this restricted pool of possible inputs, and find the best fit by exhaustively enumerating possible input combinations.

To solve the L^0 optimization problem, we enumerate all combinations of one, two, and three inputs selected from among our 10 candidate mossy fibers. This gives $\binom{10}{3} + \binom{10}{2} + \binom{10}{1} = 175$ possibilities, and we need a fast method to search through them to find adequate fits to the granule cell. Consequently, we set up the following objective:

$$\begin{aligned} C(w^i_{\text{slow}}, w^i_{\text{fast}}) &= \frac{1}{2T}\sum_{t=0}^{T}\left((v(t)-\bar{v}) - \Big(\sum_i w^i_{\text{slow}}(h^i_{\text{slow}}(t)-\bar{h}^i_{\text{slow}}) + \sum_i w^i_{\text{fast}}(h^i_{\text{fast}}(t)-\bar{h}^i_{\text{fast}})\Big)\right)^2 \\ &\quad +\alpha_0 \cdot \sum_i \mathbb{I}\left(|w^i_{\text{slow}}| + |w^i_{\text{fast}}| > 0\right). \end{aligned}$$

Where $\mathbb{I}$ is the indicator function. Adjusting the value of α_0 gives the lowest-MSE model fits that use one, two, or three mossy fiber inputs. For each granule cell fit, we manually selected the value of α_0 that best captured all features of the recorded cell's command-triggered response, confirming the quality of the fit by eye.

2.1.5 Random mixing test

Granule cell categories (E, M, L, P, EM, EL, EP, ML, MP, LP, EML, EMP, ELP, MLP, and N, where E =is early, M = is medium, L = is late, P = is pause, and N = is none) were assigned to the 169 recorded granule cells depending on the mossy fibers selected as inputs to the fit model granule cells. We constructed a random mixing model with the following assumptions: (1) Each granule cell has three sites for mossy fiber synaptic inputs; (2) The probabilities of a given input being of E, M, L, and P type are given by P_E, P_M, P_L, and P_P, with $P_E + P_M + P_L + P_P \leq 1$. (3) The type of input received at one mossy fiber-granule cell synapse is independent of that received at any other synapse. We fit the input type probabilities to the model granule cell fits by minimizing the Chichi-squared statistic. The category frequencies included all possible input combinations that produced a granule cell of a given category; for example, EEM was calculated as $3P_E^2P_M + 3P_EP_M^2 + 6P_EP_M(1 - P_E - P_M - P_L - P_P)$.

2.1.6 Generating model cells

We introduced two sources of variability to our model based on observed sources of variability in recorded cells. We found trial-to-trial variability in peak height of recorded single EPSPs to be normally distributed with $\sigma = 0.224$ mV; during simulation of model granule cells, we sampled this distribution for each mossy fiber spike. As shown previously [Sawtell, 2010], in addition to receiving corollary discharge inputs, some granule cells (84 of 212 recorded in the present study) also receive input from mossy fibers that fire at high rates, independent of the EOD command. Many such 'tonic' mossy fibers convey proprioceptive information [Bell and von der Emde, 1995; Sawtell 2010]. We added tonic input to our model based on 72 tonic mossy fibers recorded in a previous study15. The probability of a granule cell receiving tonic input was computed under the assumption of random mossy fiber mixing, and we set the synaptic weight of tonic inputs to model granule cells by sampling a Gaussian distribution fit to observed tonic EPSP sizes (2.5 ± 0.9 mV). Using the mossy fiber input probabilities fit from our random mixing model, we randomly determined whether each "dendrite of a given model granule cell received early ($P_E = 0.425$), medium ($P_M = 0.075$), late ($P_L = 0.050$), pause ($P_P = 0.050$), tonic ($P_T = 0.157$), or no input ($PN = 0.243$). We then chose a particular mossy fiber response of the previously-determined class as the source of input to that "dendrite"; we assumed that a dendrite is equally likely to choose any of the mossy fibers within a given class. For each synapse, we set the fast and slow components of the synaptic weight by randomly sampling from the pool of all fast+slow weight pairs obtained from fitting the granule cell model to recorded granule cell responses. Finally, if a model granule cell received input from one or more late mossy fibers, we set for each such fiber a probability of that mossy fiber being active after a given command; this probability was drawn from a uniform distribution. This choice was motivated by the observation that the probability of spike firing varied widely across recorded late mossy fibers (unlike the other response classes, which fired after every command).

We then added a spiking threshold V_{thresh} to model cells, measured relative to the average granule cell membrane potential measure before the EOD command, V_{rest}. In model granule cells receiving only early, medium and/or late mossy fiber input, $V_{\text{rest}} = E_L$. In model granule cells receiving pause or tonic input, $V_{\text{rest}} = E_L + \sum_i \bar{r}^i (w^i_{\text{slow}} \int \mathcal{E}_{\text{slow}}(t)dt + w^i_{\text{Fast}} \int \mathcal{E}_{\text{fast}}(t)dt$, where $\bar{r}^i$ is the average

firing rate of each pause/tonic input. We measured V_{rest} and V_{thresh} in 196 granule cells, and fit the distribution of Vthresh - Vrest with a Gaussian with $\mu = 20.2$ mV, $\sigma = 5.97$. To set the threshold of model granule cells, we calculated V_{rest} and then sampled granule cell $V_{\text{thresh}} = V_{\text{rest}} + \mathcal{N}(\mu, \sigma)$, resampling if $V_{\text{thresh}} < V_{\text{rest}}$. Upon spiking, the cell was clamped to E_L for 4 ms. To simulate the activity of our model granule cell on a single trial, we randomly drew one recorded trial (25 ms before to 200 ms after the EOD command) from each of its presynaptic mossy fibers to be used as input.

2.1.7 Simulating negative image formation

We modeled the medium ganglion cell as a passive, current-based leaky unit receiving excitatory input from 20,000 model granule cells ($r^i(t)$) and sensory input ($s(t)$), with anti-Hebbian spike timing-dependent plasticity at granule cell-medium ganglion cell synapses (w^i), and EPSP ($\mathcal{E}$) fit to granule cell-evoked EPSPs recorded intracellularly in medium ganglion cells [Grant et al 1998]. Because the timescale of learning is slow, we assumed the w^i's to be constant over a single command cycle. $s(t)$ was taken from Fig. 1b of Bell and Russell [Bell and Russel, 1978].

The granule cell- medium ganglion cell learning rule has the form $\mathcal{L}(t) = \Delta^+ - \Delta^- \mathcal{L}_0(t)$ where $t = t_{\text{MG spike}} - t_{\text{GC spike}}$ and $\mathcal{L}_0(t)$ determines the time dependence of associative depression. Theoretical analysis (see section 2.1.8) has shown that negative images are guaranteed to be stable when $\mathcal{L}_0 = \mathcal{E}$. The timescale of $\mathcal{E}$ agrees learning rules fit to experimental data [Roberts and Bell, 2000], thus we use here. Scaling of the weights by $\mathcal{L}$ was chosen to be multiplicative; because the change in synaptic weights during negative image formation was small, we simply scale by the weight before learning ($w^i 0$) for each synapse. We set $w_0^i = \left(\sum_i (\mathcal{E} * r^i)(t^i_{\max})\right)^{-1}$ where $t^i_{\max} = \text{argmax}_t(r^i(t))$, which brings the weighted granule cell input to the medium ganglion cell close to constant over time. Thus, with the approximation of linearizing the medium ganglion cell spiking response about the equilibrium voltage V_0 (see section 2.1.8), w^i evolves as $dw^i/d\tau = -w_0^i(\Delta^+ \int r^i(t)dt - \Delta^- \int V(t)(\mathcal{L}_0 * r^i)(t)dt)$, where τ is the period of each EOD cycle. We fit Δ^+ and Δ^- to negative images recorded experimentally: given experimentally- recorded membrane potential changes $\Delta V_{t_D}(t)$ induced by a broad spike at time $t_D \in \{0, 25, 50, 75, 100, 125, 150\}$ ms, and predicted membrane potential change $\Delta\tilde{V}_{t_D}(t) = \sum_i w_0^i (\mathcal{E} * r^i)(t)\big(\Delta^+ \int r^i(s)ds + \Delta^- \int \delta(s - t_D)(\mathcal{L}^0 * r^i)(s)ds\big)$, we chose Δ^+ and Δ^- to

minimize $C(\Delta^+, \Delta^-) = \sum_{t_D} \min_{C_{t_D}} \int \left(\Delta V_{t_D}(t) - \Delta\tilde{V}_{t_D}(t) + c_{t_D}\right)^2 dt$ using standard linear least-squares, where c_{t_D} is a constant offset term used to remove the effect of any net drift in membrane potential.

To monitor the degree of negative image formation during simulation, given a total change to each weight, Δw^i, we defined the residual signal error as $\left[\int \left(s(t) + \Delta w^i(\mathcal{E} * r^i)(t)\right)^2 dt\right] / \int s(t)^2 dt$.

2.1.8 Stability Analysis

We follow the approach of Williams, Roberts, and Leen [William, Roberts, and Leen, 2003]. The MG cell spiking rate is a function of the MG cell voltage, $f(V)$, which we linearize around the steady state value V_0 (defined below). $V(t)$ and $r^i(t)$ are periodic on a fast time scale (the EOD response), denoted by t, whereas the w^i's change on the slower timescale of many EODs, denoted by τ.

We thus obtain the following dynamical system for $V(t)$ and w^i over many EODs:

$$\frac{dV(t)}{d\tau} = \sum_i (\mathcal{E} * r_i)(t) \frac{dw_i}{d\tau} \tag{2.1}$$

$$\begin{aligned} \frac{dw^i}{d\tau} &= -w_0^i \Delta^- \int [f(V(t)) - f(V_0)](\mathcal{L}_0 * r_i)(t) dt \\ &= -\left.\frac{df}{dV}\right|_{V_0} w_0^i \Delta^- \int [V(t) - V_0](\mathcal{L}_0 * r_i)(t) dt + \mathcal{O}((V(t) - V_0)^2), \end{aligned} \tag{2.2}$$

where $V_0 = f^{-1}\left(-\Delta^+/(\Delta^- \int \mathcal{L}_0(t)\, dt)\right)$ is the voltage at which the nonassociative and associative plasticities balance[4].

To confirm that our system will converge to form a stable negative image, we next generalized a result proved in [4] for the case of a delay line basis, that negative images are stable when $\mathcal{L}_0$ has the same shape as $\mathcal{E}$, to the case of an arbitrary GC basis.

Given a voltage perturbation $V(t)$ arising from sensory input to the MG cell, we defined $\tilde{V}(t) \equiv V(t) - V_0$, linearized around $V(t) = V_0$, and substituted equation (2) into (1) to obtain the dynamics of the voltage perturbation from V_0.

$$\frac{d\tilde{V}(t)}{d\tau} = -\left.\frac{df}{dV}\right|_{V_0} \sum_i w_0^i \Delta^- (\mathcal{E} * r_i)(t) \int \tilde{V}(s)(\mathcal{L}_0 * r_i)(s) ds \tag{2.3}$$

In the case that the learning rule matches the shape of the EPSP, i.e. $\mathcal{L}_0 = \mathcal{E}$, we have

$$\frac{d\tilde{V}(t)}{d\tau} = -\frac{df}{dV}\bigg|_{V_0} \sum_i w_0^i \Delta^-(\mathcal{E} * r_i)(t) \int \tilde{V}(s)(\mathcal{E} * r_i)(s)ds$$

We evaluated the stability by taking the inner product of a displacement from the equilibrium with the derivative.

$$\begin{aligned}\int \tilde{V}(t)\frac{d\tilde{V}(t)}{d\tau}dt &= -\frac{df}{dV}\bigg|_{V_0} \sum_i w_0^i \Delta^- \int \tilde{V}(t)(\mathcal{E} * r_i)(t)dt \int \tilde{V}(s)(\mathcal{E} * r_i)(s)ds \\ &= -\frac{df}{dV}\bigg|_{V_0} \sum_i w_0^i \Delta^- \left(\int \tilde{V}(t)(\mathcal{E} * r_i)(t)dt\right)^2 \leq 0 \end{aligned} \tag{2.4}$$

We thus see that any voltage perturbation within the subspace spanned by the EPSP-convolved firing rates will decay back to the equilibrium (note that Δ^- and all w^i are positive). Voltage perturbations outside of this subspace will be left unaltered by the learning rule.

2.1.8.1 Learning Dynamics Analysis

Discretizing t and s, Equation 2.3 may be written

$$\frac{d\tilde{V}(t)}{d\tau} = \sum_s \mathbf{A}(t,s)\tilde{V}(s) \tag{2.5}$$

where

$$\mathbf{A}(t,s) = -\frac{df}{dV}\bigg|_{V_0} \sum_i w_0^i \Delta^-(\mathcal{E} * r_i)(t)(\mathcal{L}_0 * r_i)(s)$$

The matrix $\mathbf{A}$ determines the dynamics by which the voltage perturbation $\tilde{V}$ decays due to learning. Eigenvectors of $\mathbf{A}$ reflect temporal patterns which can be cancelled by the GC basis $\{r_i\}$; we subsequently refer to these as eigenmodes. The eigenvalue corresponding to each eigenmode reflects the rate at which that pattern is cancelled by the basis, also referred to as its rate of decay.

2.1.8.2 Learning Rate Normalization

On a given trial, the sensory input to the MG cell is a noisy observation $\tilde{h}(t,\tau)$ of the true EOD-driven input $h(t)$:

$$\tilde{h}(t,\tau) = (\mathcal{E} * h)(t) + (\mathcal{E} * \eta)(t,\tau) - V_0,$$

where η is the sensory observation noise that reflects sensory signals and spiking noise uncorrelated with the EOD. Plugging into the linearized weight dynamics,

$$\frac{dw^i}{d\tau} = -\frac{df}{dV}\bigg|_{V_0} w_0^i \Delta^- \left(\int ((\mathcal{E} * h)(s) - V_0)(\mathcal{L}_0 * r_i)(s) ds + \int (\mathcal{E} * \eta)(s,\tau)(\mathcal{L}_0 * r_i)(s) ds \right)$$

The first of these terms decays to zero provided $(\mathcal{E} * h)(t)$ is in the span of $\{(\mathcal{E} * r_i)(t)\}$. The second term does not decay, but rather introduces a small trial-to-trial fluctuation in the weights w^i due to interaction of $\eta(t,\tau)$ with the learning rule.

We next introduce a positive learning rate λ, substituting $\Delta^+ \to \lambda\Delta^+$ and $\Delta^- \to \lambda\Delta^-$ in the learning dynamics (note that introducing λ does not change the stability of learning or the value of V_0 provided λ is small). The weight dynamics become

$$\frac{dw^i}{d\tau} = -\lambda\frac{df}{dV}\bigg|_{V_0} w_0^i \Delta^- \left(\int ((\mathcal{E} * h)(s) - V_0)(\mathcal{L}_0 * r_i)(s) ds + \int (\mathcal{E} * \eta)(s,\tau)(\mathcal{L}_0 * r_i)(s) ds \right)$$

Smaller values of λ decrease the amplitude of the noise-induced weight fluctuations, but also increase the number of trials required for negative image formation. Thus given some assumption on the structure of η (such as its power spectrum), we may choose λ for a given basis such that the magnitude of noise-induced weight fluctuations is fixed.

2.2 Results

2.2.1 Corollary discharge responses in MFs, UBCs, and Golgi cells

Consistent with previous studies [Bell, Grant, and Serrier 1992; Bell and von der Emde, 1995; Sawtell, Mohr, and Bell, 2005; von der Emde and Bell, 1996], extracellular recordings from two midbrain nuclei that are the main sources of corollary discharge input to GCs revealed responses restricted to short delays after the EOD motor command (Figure 2.1c, PCA, n=12; PE, n=31). To further characterize corollary discharge inputs to GCs we used high-impedance glass microelectrodes to record from putative MF axons within EGp itself (see Methods for details of MF recordings). Most MFs recorded in EGp exhibited responses restricted to short delays, termed early and medium, that closely resembled the responses recorded in midbrain neurons that send MFs to EGp (Figure 2.11d,e; early, n=54; medium, n=28). Thus corollary discharge inputs to EGp

appear insufficient for canceling the effects of the EOD over their entire duration. However, we also found other putative MFs within EGp, termed late and pause, that exhibited far more delayed and diverse corollary discharge responses (Fig. 1d,e; late, n=26; pause, n=27). Late MFs fire bursts or single action potentials at long delays after the EOD command (>50 ms), while pause MFs show highly regular tonic firing that ceases abruptly around the time of the command. Resumption of firing is often marked by precise time-locking of spikes at long delays relative to the EOD command (Figure 2.1d, bottom).

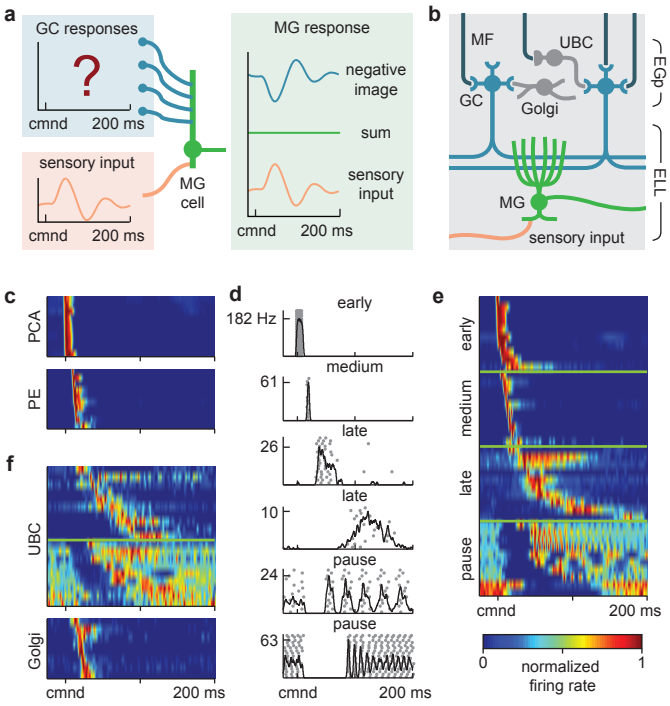


Figure 2.1: Corollary discharge responses in MFs, UBCs, and Golgi cells. **a)** Schematic of negative image formation and sensory cancellation in an MG cell. The question mark indicates that temporal patterns of corollary discharge response in GCs are the critical unknown in current models of sensory cancellation. **b.** Schematic of the circuitry of the EGp and ELL. Corollary discharge signals related to the EOD motor command are relayed via several midbrain nuclei (not shown) and terminate in EGp as MFs. UBCs give rise to an intrinsic system of MFs that provide additional excitatory input to GCs. Golgi cells inhibit GCs and UBCs. MG cells in ELL receive both sensory input and GC input via parallel fibers. **c.** Corollary discharge responses of units recorded in the paratrigeminal command associated nucleus (PCA) and the preeminential nucleus (PE). Each row shows the smoothed (5 ms Gaussian kernel) and normalized average firing rate of a single unit. In this and subsequent figures time is defined relative to the EOD motor command (cmnd), which is emitted spontaneously by the fish at 2-5 Hz. Color bar in e applies also to c and f. **d.** Example spike rasters (grey dots) and smoothed firing rates (black curves) for putative MFs recorded extracellularly in EGp illustrating four temporal response classes (early, medium, late, and pause). **e.** Corollary discharge responses of putative MFs recorded extracellularly in EGp. Each row represents the smoothed and normalized average firing rate of a single MF, with *(continued on following page)*

Figure 2.1: *(continued)* 10 examples of each class shown. **f.** Corollary discharge responses of UBCs (n = 19) and Golgi cells (n = 8) recorded intracellularly. Each row represents the smoothed and normalized average firing rate of a single cell. Note the similarity with late and pause MFs, shown in e.

A candidate for the source of late and pause responses recorded in EGp are the UBCs that, in mormyrid fish as in the mammalian cerebellum and dorsal cochlear nucleus [Mugnaini, Sekerkova and Martina, 2011], give rise to an intrinsic system of MF axons that provides additional excitatory input to GCs [Campbell, Meek, Zhang and Bell, 2007]. Whole-cell recordings from UBCs provided direct support for this idea. UBCs (n=54), GCs (n=184), and Golgi cells (n=11) could be clearly distinguished on the basis of their electrophysiological properties and morphology. Strikingly, corollary discharge responses in UBCs are delayed and diverse, and they closely resemble late and pause responses recorded extracellularly (compare Figure 2.1e and Figure 2.1f). An objective classification algorithm supports our conclusion that early and medium responses are extrinsic MF axons originating from midbrain nuclei while late and pause responses are intrinsic MF axons originating from UBCs.

Possible mechanisms for generating diverse and delayed responses in UBCs were revealed by our intracellular recordings. Prominent post-inhibitory rebound firing was observed in a subset of UBCs (Figure 2.2a), so rebound evoked by an inhibitory input arriving at a short delay after the EOD command (Figure 2.2b) could account for their delayed firing. Suggestive of such a mechanism, the morphologically identified UBC shown in Figure 2.2a,b fired bursts at a long delay after the command that were stronger when the preceding membrane potential was more hyperpolarized. Other UBCs exhibit regular tonic firing that, when terminated by hyperpolarization, is followed by precisely time-locked spikes (Figure 2.2c). This firing pattern is similar to pause responses recorded extracellularly and could also be explained by inhibition arriving at a short delay after the command (Figure 2.2d). Golgi cells respond at short delays after the EOD command (Figure 2.1f) and could be the source of such inhibition.

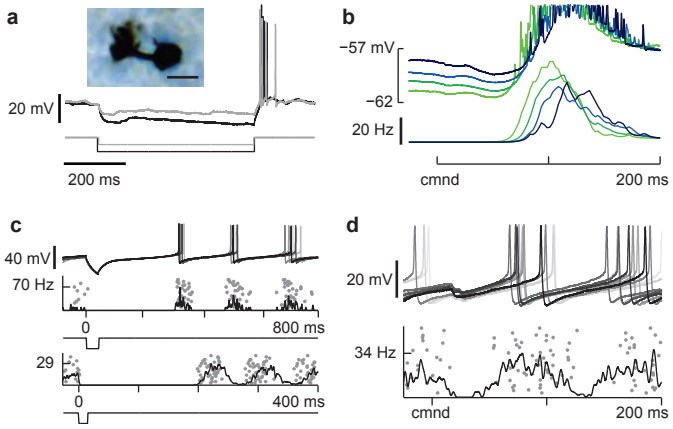


Figure 2.2: Mechanisms for delayed and diverse corollary discharge responses in UBCs. **a.** Two overlaid traces illustrating prominent rebound firing in response to hyperpolarizing current injections (-10 and -20 pA) in a UBC. This cell was filled with biocytin allowing for post-hoc morphological identification (inset, scale bar 10 μM). **b.** Late corollary discharge response in the same UBC recording shown in a. The strength of late action potentials bursts (bottom traces) is related to the degree of preceding membrane potential hyperpolarization (top traces), suggesting rebound from command-locked hyperpolarization as a possible mechanism underlying late responses observed in UBCs. **c.** Two UBCs in which a brief hyperpolarizing current injection (-50 pA, top; -200 pA, bottom) results in an entrainment of tonic firing, similar to temporal patterns of action potential firing observed in pause MFs. Similar effects were seen in 7 additional UBCs. **d.** Pause-type corollary discharge response in a UBC, note the small hyperpolarization time-locked to the command and the entrainment of tonic action potential firing after the pause.

2.2.2 Experimental characterization and modeling of corollary discharge responses in GCs

We next catalogued corollary discharge responses in a large number of GCs using whole-cell recording. Corollary discharge responses were observed in 170 of 184 GCs and consisted of prominent depolarizations with temporal patterns that are highly consistent across commands (Figure 2.3a,b-left). GC depolarizations closely resemble early, medium, late, pause, or in some cases, apparent mixtures of these responses (Figure 2.3b-right). Action potential firing consistent mainly of single spikes (1.25 spikes/command for the roughly 20% of GCs that fired on greater than 10% of

commands) and always occurred at the peak of the subthreshold membrane potential (Figure 2.4). These observations led us to hypothesize that the temporal structure of subthreshold GC corollary discharge responses is shaped primarily by summation of excitatory inputs, rather than by phasic Golgi cell inhibition or the intrinsic properties of the GCs themselves. To test this, we modeled GC depolarizations as sums of excitatory postsynaptic potentials (EPSPs) computed from the spike trains of up to three of the recorded EGp MFs (Figure 2.1e), including UBCs (Figure 2.1f). The small number of excitatory inputs is consistent with anatomical observations that mormyrid GCs have, on average, three claw-like dendritic endings [Nate Satell, unpublished observations] and previous physiological observations indicating that GCs receive other sources of MF input in addition to corollary discharge, e.g. proprioceptive input from spinocerebellar MFs [Sawtell, 2010]. By choosing an appropriate set of inputs from the recorded data and adjusting their excitatory synaptic strengths within a reasonable range (Figure 2.3b-right; see Methods), we were able to fit the membrane-potential responses of the recorded GCs with high accuracy (average MSE = 4.6%, n = 169; Figure 2.3b-left). This provides strong support for the view of GC recoding as excitatory input summation stated above.

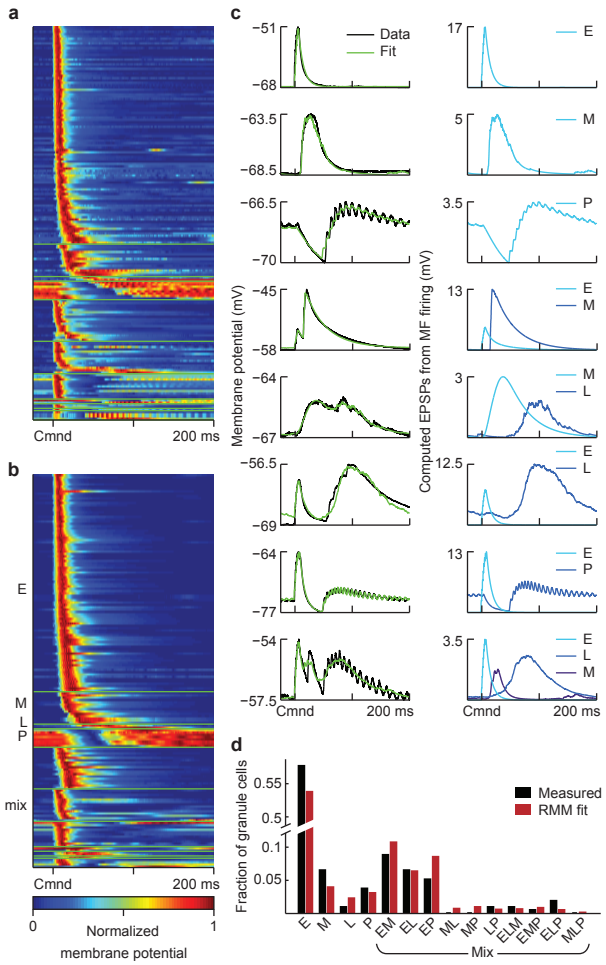


Figure 2.3: Experimental characterization and modeling of corollary discharge responses in GCs. **a.** Average subthreshold corollary discharge responses of 170 GCs. Responses are grouped by category (see d) and then sorted by the latency of their peak membrane potential. **b.** left, examples of recorded GC subthreshold responses (black trace) and model fits (green). Right, EPSPs computed from the recorded MF inputs used to fit each GC, labeled according to the class to which they belong. **c.** The distribution of response categories assigned to recorded GCs based on model fits (black bars). Bars labeled E, M, L and P indicate the fraction of early, medium, late and pause inputs used to fit the recorded GC responses. Mixed bars show these fractions for combinations of inputs used in the same way. These fractions are consistent with a four-parameter random mixing model (RMM; parameters are the probability of early, medium, late, and pause inputs) in which each input to a GC is assigned independently of the others (red bars). This suggests that the combinations of inputs GCs receive are random. **d.** Average *(continued on following page)*

Figure 2.3: *(continued)* subthreshold corollary discharge responses of 170 randomly constructed model GCs selected from a total of 20,000. In this sample, the number of model cells from each GC category was matched to the experimental data, but the selection process was otherwise random. Note that the temporal response properties of the model GCs closely resemble those of the recorded GC shown in a.

The similarity of the constructed and recorded GC responses also provides a powerful tool for addressing the central question of whether GC responses can support negative image formation and sensory cancellation. Given the sparseness of GC firing that we observed (both a small percentage of GCs that fire and a small number of spikes per EOD in those that do), cancellation likely depends on large numbers of GC inputs; indeed, anatomical estimates are on the order of 20,000 GC inputs per MG cell [Meek et al, 1996]. To expand the data to this number, we constructed model GCs. This was aided by the fact that the distribution of inputs found in our fits of recorded GC responses is consistent with a random mixing process in which each GC dendrite samples the different functional input classes (early, medium, late, and pause) independently (Figure 2.3c). We extracted the probability of a GC receiving an input from each functional class from these fits (see Methods). Drawing randomly from these input probabilities and from the distribution of synaptic weights obtained during fitting allowed us to construct model GCs with corollary discharge responses that closely match those of the recorded GCs (Figure 2.3d; note that these are not GCs fit to the data, but randomly constructed and sampled model cells). The remarkable similarity between these model responses and the data provides additional support for the hypothesis that GC recoding of corollary discharge inputs can be explained by random mixtures of small numbers of excitatory inputs conveyed by extrinsic MFs and UBCs.

Spiking in model GCs was implemented by randomly assigning action potential thresholds sampled from a normal distribution fit to the thresholds of the recorded GCs (average distance to threshold 20.2 5.97 mV). The resulting temporal firing patterns and distribution of average spike counts per EOD in the model GCs were statistically consistent with those of the recorded GCs (Figure 2.4).

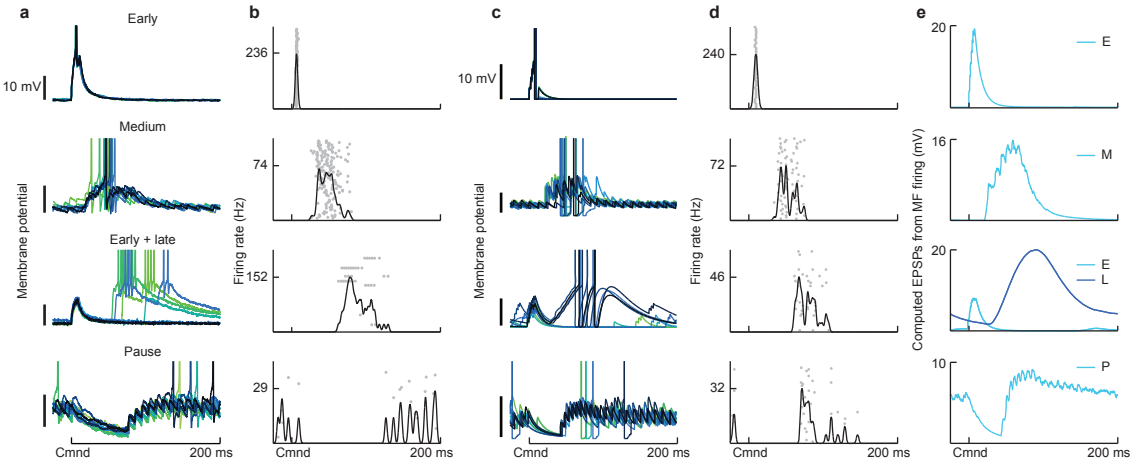


Figure 2.4: Patterns of corollary discharge-evoked action potential firing in recorded and model GCs. **a.** Corollary discharge responses of four recorded GCs that spiked in response to the EOD command. GC membrane potentials from several commands are shown overlaid. Spikes are truncated to show details of subthreshold membrane potentials. **b.** Spiking responses of the recorded GCs shown in a. Spike trains on 50 individual trials are shown in gray, and the smoothed (5 ms Gaussian kernel) trial-averaged firing rate of the cell is overlaid in black. **c, d.** Corollary discharge responses of four model GCs selected from among the pool of 20,000 generated cells. Displays for model GCs are the same as for recorded cells. **e.** Sources of MF input to each model GC, as computed EPSPs from the trial-averaged MF firing rates. Both subthreshold corollary discharge responses and spiking in model GCs closely resembles that seen in recorded GCs.

2.2.3 GC corollary discharge responses provide an effective basis for canceling natural patterns of self-generated input

Previous experimental work has revealed a combination of anti-Hebbian spike-timing dependent long-term synaptic depression and non-associative long-term potentiation at GC-MG cell synapses [Bell, Han, Sugawara and Grant, 1997; Han, Grant and Bell, 2000]. To determine whether this form of plasticity can cancel self-generated sensory input using realistic GC responses, we drove a passive model MG cell with 20,000 model GC inputs through plastic synapses. The GC to MG cell synapses were strictly positive and their strengths were initially set so that GC responses in the

absence of EOD-driven sensory input generated a roughly flat MG membrane potential, consistent with recordings from MG cells in regions of ELL involved in passive electrosensory processing [Bell 1981; Bell 1982; Bell, Caputi, Grant and Serrier, 1993]. Next, we added a temporally varying sensory input to the model MG cell to mimic the responses of passive electroreceptors to the fish's own EOD recorded in a previous study [Bell and Russell, 1978] (Figure 2.5a). As in previous modeling work [Roberts and Bell, 2000], the strength of the GC-MG synapses evolved according to the experimentally described plasticity rule [Bell, Han, Sugawara and Grant, 1997; Han, Grant and Bell, 2000]: synaptic strength is increased for each presynaptic action potential, corresponding to experimentally described non-associative potentiation, and decreased when a postsynaptic action potential occurs shortly after a presynaptic action potential, corresponding to experimentally described associative depression (Figure 2.6c). Over the course of about 1000 EOD commands (approximately 5 minutes at EOD command rates typical of paralyzed fish), the membrane potential fluctuation caused by the sensory input is canceled by the corollary discharge inputs conveyed by GCs (Figure 2.5a), consistent with the time-course over which negative images are formed in vivo [Bell 1981; Bell 1982]. The resulting negative image closely matches the inverse of the sensory input (Figure 2.5b) and has small command-to-command variations (Figure 2.5b, blue shading; standard deviation of $\sim$1 mV) despite the sparseness of the GC firing. We also confirmed the stability of negative images formed using the GCs as a temporal basis and that the changes in synaptic strength underlying negative images were within a physiologically plausible range. Finally, because our estimates of both the number of GCs active at long delays and the number of command-locked action potentials fired by GCs were based on limited data, we tested the effects of systematically varying these properties of the model GCs on negative images and sensory cancellation. Rapid cancellation and negative images with small command-to-command variations were observed even when numbers of late and pause inputs used to generate model GCs were reduced and when the number of action potentials fired by GCs was reduced.

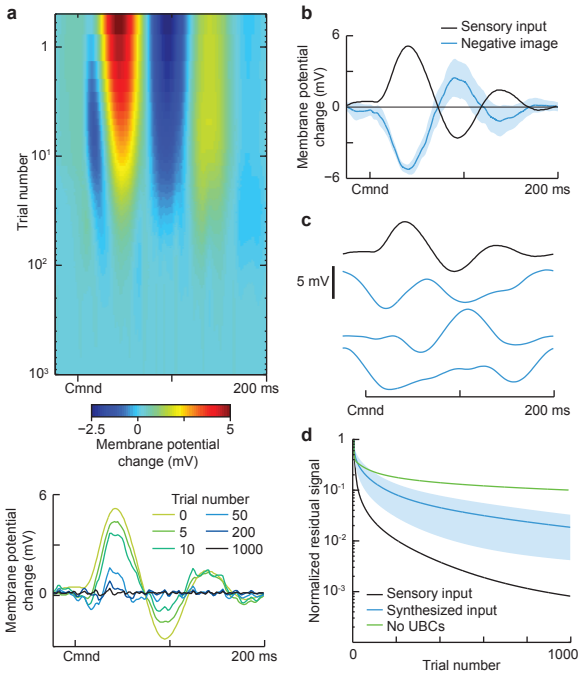


Figure 2.5: GC corollary discharge responses provide an effective basis for canceling natural patterns of self-generated sensory input. **a.** top, Cancellation of the change in membrane potential caused by sensory input locked to the EOD motor command in a model MG cell. The MG cell receives 20,000 model GC inputs with synaptic strengths that are adjusted by anti-Hebbian spike-timing dependent plasticity. Bottom, select trials showing the time course of cancellation. The temporal profile of the sensory input (trial 0) was chosen to resemble the effects of the EOD on passive electroreceptors recorded in a previous study1. **b.** The negative image (blue line) effectively cancels the sensory input (black line), with small command-to-command variability (shaded region shows 1 std across trials.) **c.** Different input signals used for the tests of sensory cancellation rates shown in d. The top trace is the same input used in a resembling natural self-generated inputs due to the EOD. The blue traces are selected from a set of 1,000 synthesized inputs with the same power spectrum as the natural input but with randomized phases. **d.** Comparison of the time course of cancellation for the natural sensory input (black) versus the synthesized inputs (blue; shaded region is 1 std). Note that cancellation is faster for the natural input, suggesting that the structure of GC responses is matched to the temporal pattern of the self-generated signal. Cancellation is also much slower and less effective if the model GCs are generated without UBC inputs (green line).

The effectiveness of the cancellation in the model is notable given the highly non-uniform temporal structure of the GC population response, in particular the fact that most GCs are active at short delays. Rather than a general-purpose temporal basis, such as the delay-line model considered in previous theoretical work [Roberts and Bell, 2000], the structure of GC corollary discharge responses appears to be matched to the temporal patterns of self-generated sensory input that the fish encounters in nature, i.e. the particular pattern of ringing that the large EOD evokes in electroreceptors tuned to detect much smaller signals1. To test this idea more directly we generated synthetic inputs with different temporal profiles but the same power spectrum as the electroreceptor response (Figure 2.5c). Synthetic inputs are cancelled more slowly than inputs resembling the electroreceptor response (Figuref 2.5d), suggesting that the structure of GC responses is particularly suited to natural patterns of self-generated input. Furthermore, the rate and accuracy of sensory cancelation in the GC basis is comparable to that of an idealized uniform delay line basis with tuning widths approximately equal to that of GCs receiving medium and late inputs. Finally, we note that model GC populations lacking late and pause inputs provide a far less effective basis for cancellation (Figure 2.5d, green line), indicating an important role for the temporally diverse and delayed corollary discharge responses generated by UBCs.

2.2.4 Non-uniform temporal structure of GC responses predicts paradoxical features of negative images

Our knowledge of the temporal structure of GC responses allowed us to make specific predictions about the shapes of negative images induced in experiments in which a single dendritic spike in an MG cell is paired with the EOD command at fixed delays. Previous studies have shown that such dendritic spikes are the key triggers for associative depression at GC synapses [Bell, Han, Sugawara, and Grant, 1997; Englemann et al, 2008]. Our predictions based on the measured GC responses were twofold (Figure 2.6a, green curves). MG cell spikes evoked at short delays should induce a brief hyperpolarization peaked around the spike time due to associative depression of early GCs inputs. More complex, bi-phasic changes were predicted for MG cell spikes evoked either at longer delays or at zero delay. At such delays, associative depression should induce a hyperpolarization around the MG cell spike time, while non-associative potentiation of the numerous early GC inputs

should cause a "paradoxical" MG cell depolarization at short delays after the EOD command. In contrast, a model with a temporally uniform delay line basis predicts that negative images induced by MG cell spikes at different delays would always have the same shape, determined by the temporal window of the synaptic plasticity, and differ only by time translation (Figure 2.6a, bottom right, dashed curve).

To test these model predictions, we recorded intracellularly from MG cells, and compared corollary discharge responses before and after 3 minutes of pairing (approximately 600 commands) with a brief current injection that evoked a dendritic spike at a fixed delay after the EOD command (Figure 2.6b). The shapes of the resulting negative images exhibit a strong dependence on the delay during pairing, in agreement with our qualitative predictions (Figure 2.6a, black traces). Furthermore, close quantitative agreement between our model and the experimental MG cell response changes (Figure 2.6a, compare green and black traces) could be achieved by fitting just two parameters of the synaptic plasticity rule (Figure 2.6c; see Methods). The similarity of the modeled and measured changes in MG cell responses indicates that the measured GC responses and previously measured anti-Hebbian plasticity at GC-MG cell synapses accurately describe negative image formation.

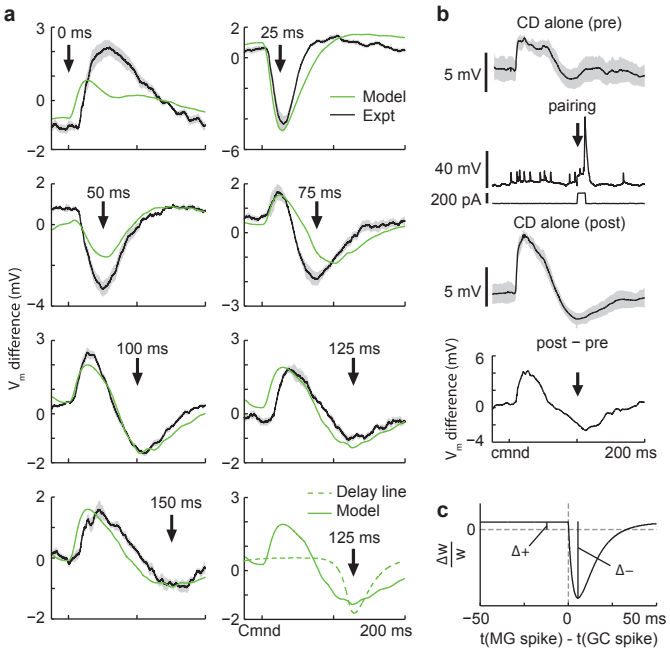


Figure 2.6: Non-uniform temporal structure of GC responses predicts specific features of negative images in MG cells. **a.** Changes in corollary discharge responses induced by pairing with MG dendritic spikes at 7 different delays after the EOD command. Green traces are membrane potential differences derived from the model with fitted values for the magnitudes of associative depression and non-associative potentiation (panel c). Black traces are experimentally observed membrane potential differences averaged across MG cells (outlines represent SEM; 0 ms, n = 6; 25 ms, n = 8; 50 ms, n = 6; 75 ms, n = 6; 100 ms, n = 10; 125 ms, n = 4; 150 ms, n = 3). The bottom right panel compares these predictions with those for a delay line basis (dashed green line). **b.** Design of the pairing experiment. Intracellular traces from an MG cell showing the average (black) and standard deviation (gray outline) of the corollary discharge response before (pre) during (pairing), and after (post) three minutes of pairing during which a brief (12 ms) intracellular current injection evoked a dendritic spike at a fixed delay after the EOD command (arrow). The small spikes are axonal spikes and do not contribute to plasticity5. The bottom trace (post-pre) shows the difference in the membrane potential induced by the pairing, corresponding to the traces shown in a. Note the complex pattern of changea relative hyperpolarization around the time of the paired spike as well as a large relative depolarization just after the command, as predicted by the model. **c.** Synaptic plasticity rule and parameters used for the fits shown in a.

2.3 Discussion

Using intracellular recordings and modeling of GCs in mormyrid fish we provide a relatively complete description of GC recoding, far more complete than that available in other systems. The remarkably close agreement between recorded and model GCs (shown in Figure 2.3) strongly suggests that the simple rules we used to transform MF inputs into GC responses, i.e. summation of randomly selected excitatory inputs, are essentially correct and complete. Such a complete understanding of how inputs are transformed into output in vivo is remarkable in its own right and places us in a unique position to explore the relationships between input coding, an experimentally defined synaptic plasticity rule [Bell, Han, Sugawara and Grant, 1997; Han, Grant, and Bell, 2000], and a well characterized adaptive network output in the form of negative images [Bell 1981; Bell 1982; Bell, Caputi, Grant and Serrier, 1993]. Though input coding and plasticity are the critical elements for the functioning of many neural circuits, including other cerebellum-like circuits [Bell, Han, and Sawtell 2008; Farris 2011; Oertel and Young 2004] and the cerebellum itself [Albus 1971; Marr 1969; Medina and Mauk, 2000; Dean, Porrill, Ekerot, and Jörntell 2010] there are few cases in which these elements are understood so thoroughly.

The function of EGp circuitry demonstrated here closely parallels longstanding, but still untested, expansion recoding schemes posited for the granular layer of the mammalian cerebellum [Albus, 1971; Marr, 1969]. Whereas most models of cerebellar granular layer function, posit pivotal roles for Golgi cell inhibition of GCs in expansion recoding [Medina and Mauk, 2000], our study suggests a key role for UBCs. Though we had no way to specifically target UBCs and hence cannot provide a complete account of their properties, our in vivo intracellular recordings suggest that they generate temporally diverse and delayed responses that are faithfully recoded in GCs. Though in vivo responses to discrete inputs have yet to be described for UBCs in the mammalian cerebellum or dorsal cochlear nucleus, in vitro studies have documented a variety of synaptic and intrinsic mechanisms capable of generating prolonged and/or delayed responses [Diana et al, 2007; Locatelli et al, 2013; Rossi, Alford, Mugnaini and Slater, 1995; Russo, Mugnaini, and Martina, 2007; Rousseau et al, 2012]. These include rebound firing [Russo, Mugnaini, and Martina 2007], regular tonic firing [Russo, Mugnaini, and Martina 2007], and inhibitory synaptic input from Golgi cells [Rousseau et al, 2012]the key ingredients for delayed responses suggested by our in vivo intracellular recordings.

Hence the functions for UBCs established here may extend to other circuits in which they are found. Finally, though the capacity to generate temporally diverse responses in GCs may be useful for a variety of cerebellar computations, the density of UBCs varies widely across different regions of the cerebellum and across different species [Mugnaini, Sekerková, and Martina, 2011]. Whether other circuit mechanisms, e.g. phasic Golgi cell inhibition of GCs, function to generate temporally diverse GC responses for regions of the cerebellum in which UBCs are scarce is an important question for future studies.

An unexpected finding of this study is that rather than a general temporal basis, such as delay-line models considered in previous theoretical work [Roberts and Bell, 2000], the temporal structure of GC responses is highly non-uniform. Despite the preponderance of GCs active at short delays, our modeling suggests that they provide a highly effective basis for sensory cancellation. The explanation to this apparent paradox is that the temporal structure of GCs is matched to natural patterns of self-generated sensory input. How such matching might occur and whether it could be modified by experience are interesting questions for future investigations. The non-uniform temporal structure of GC responses also provides a simple explanation for unusual features of MG cell negative images formed in response to artificial inputs. The ability to accurately predict detailed features of negative images based on modeled GC responses, and previously described anti-Hebbian plasticity, also provides an additional experimental validation for the links we establish between input coding, plasticity, and adaptive network output. Finally, the apparent matching between GC responses and natural patterns of self-generated sensory input does not imply that the system cannot provide effective cancellation when conditions change. Indeed, EOD amplitude along with passive electroreceptor responses to the EOD are expected to change on multiple timescales due to growth of the fish, seasonal changes in water conductivity, and the presence of large nonconducting objects near the fish. However, as has been shown in a previous study on the effects of water conductivity on passive electroreceptor responses [Bell and Russell, 1978], such changes will primarily affect the size rather than the temporal structure of sensory responses to the EOD. Hence the matching between the temporal structure of GC responses and self-generated sensory input described here is expected to hold over a wide range of behaviorally relevant conditions.

Though the notion that motor corollary discharge signals could be used to predict and cancel the

sensory consequences of an animals own behavior has a long history [Sperry, 1950; von Holst and Mittelstaedt, 1950], there are few cases in which such functions have been characterized at the level of neural circuits [Crapse and Sommer, 2008]. In particular, it has proven challenging to understand how copies of motor commands are translated into an appropriate format to cancel sensory inputs. This problem takes a particularly clear and tractable form in the case of mormyrid ELL, where copies of a brief, highly stereotyped motor command must be delayed and diversified in order to provide a basis for cancelling sensory effects that are extended in time. A major contribution of the present study is to directly demonstrate that such a temporal expansion indeed occurs in GCs and that, along with previously described anti-Hebbian plasticity [Bell, Han, Sugawara and Grant, 1997; Han, Grant, and Bell, 2000], is sufficient to account for negative images. Our results hence provide the critical missing piece in a relatively complete mechanistic account of how motor commands are used to predict sensory consequences at the levels of synaptic plasticity, cells, and circuits.

Chapter 3

Mechanisms for internal model learning in an electric fish

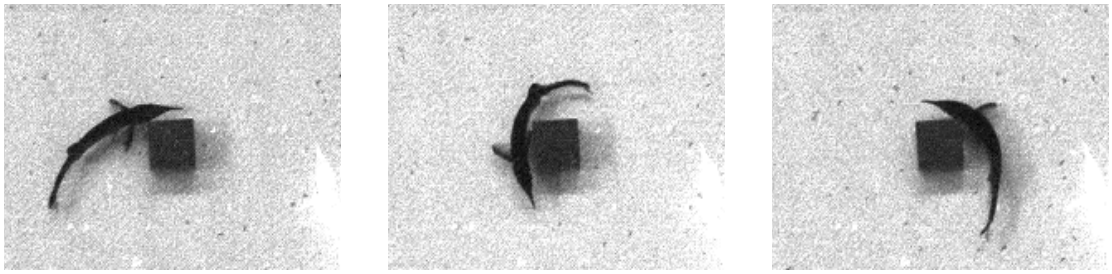


Figure 3.1: Mormyrid fish maneuver their electric organ and chin appendage to investigate novel objects in their environment. The chin appendage is densely packed with electroreceptors, and acts like a fovea of the electrosensory system.

In active electrosensing, the mormyrid fish uses its electric organ in conjunction with mormyromast electroreceptors to detect objects in its environment that do not necessarily generate their own electric field. The electric organ discharge (EOD) forms a transient field around the fish's body, the amplitude of which is encoded by mormyromasts found throughout the skin of the fish. Adding a conducting or insulating object to the water surrounding the fish distorts its generated field, changing field amplitude at the skin. Behavioral studies show that fish can use their active

electrosensory system to distinguish objects of different sizes, shapes, distances, and conductivities [von der Emde, Schwarz, Gomez, Budelli, and Grant 1998; von der Emde 1999].

Reliable detection of objects requires mormyromasts to be able to perceive changes of a few percent in the strength of the field at the skin. In addition to being modulated by objects, the amplitude of the fish's field is modulated by the fish's own movements, which change the location of the electric organ in the tail relative to the skin. Fish presented with novel objects in their environment are quite active, and approach the object with different body postures during their investigation, suggesting that effects of posture on the fish's field do not interfere with electrosensory perception. A mechanism for distinguishing self-generated vs external EOD modulations appears to exist in the fish's electrosensory lobe: extracellular recordings show that effects of objects and body bends are comparable in size in mormyromasts- but in the efferent cells of the electrosensory lobe, posture effects are removed while effects of objects persist.

In this chapter, I set out to determine how complex postures affect the amplitude of the fish's field at mormyromasts throughout the body, and whether the model of negative image formation established in the previous chapter can be extended to account for the transformation from a posture-dependent sensory response in mormyromasts to a posture-invariant representation of objects in the efferent cells of the electrosensory lobe. Following a review of the active electrosensory system, I will incorporate data from posture-encoding mossy fibers into the granule cell model, and study the capacity of the granule cell basis to form posture-specific negative images in model efferent cells. I then present an electrostatic model of the fish's field which I used to study the effect of posture on sensory representations, and study the capacity of model efferent cells to form large families of posture-specific negative images.

3.0.1 A note on terminology

As discussed in the previous chapter, EOD corollary discharge and electrosensory pathways converge in the electrosensory lobe, where the corollary discharge signal is sculpted via synaptic plasticity at granule cell to MG cell synapses to cancel effects of the EOD from the electrosensory response. This chapter refers to efferent cells, rather than MG cells, due to a difference in experimental methods. The electrosensory lobe has four types of Purkinje-like efferent cells: LG, LF, and two types of MG

cell (called MG_1 and MG_2, which are inhibited and excited by sensory input, respectively) [Meek, Grant and Bell 1999]. In the previous chapter, MG cells were recorded intracellularly, allowing cell type to be verified anatomically; because the experiments in this chapter relied on extracellular recording, cell type could not always be determined, and I will use the more general term efferent cell. While negative images formed in the four types of efferent cells are similar, there are some interesting functional differences between the four cell types. These will not be relevant here, but I will review them briefly in the general discussion section.

3.1 Introduction: Cancellation of posture effects in active electrosensing

This section reviews experimental study of the active electrosensory system, the computational challenges body movements introduce for active electrosensing, and the neural mechanism by which the electrosensory system appears to solve these challenges.

3.1.1 Sensory encoding in the active system

Unlike ampullary cells in the passive system, mormyromasts are specialized to operate in conjunction with the fish's EOD. The bioelectric fields generated by organisms in the water typically have frequencies below 50 Hz, the detection range of ampullary cells. By contrast, mormyromasts respond to frequencies between 50 Hz and 10 kHz, with a preferred frequency of around 2500 Hz, where the power spectrum of the fish's EOD waveform peaks [von der Emde 1999; Bell 1990]. They are therefore much more sensitive to the fish's own EOD than to external fields generated by other organisms.

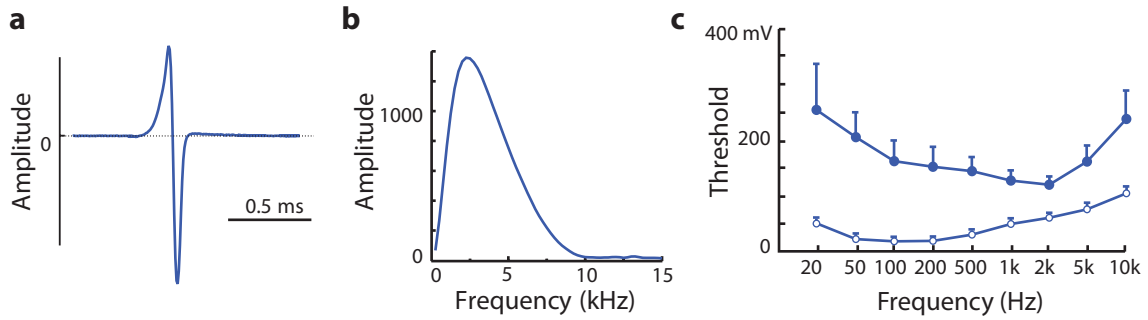


Figure 3.2: **a.** Waveform of the electric organ discharge of *Gnathonemus petersii*, the mormyrid species used in this study. **b.** Power spectrum of the EOD waveform. This panel and panel a are reproduced from [von der Emde 1999]. **c.** Frequency sensitivity of type A (open dots) and B (filled dots) sensory cells of mormyromasts, reproduced from [Bell 1990]. Both cell types have a higher preferred frequency than ampullary cells. (I will disregard differences between type A and B cells here.)

Objects in the water around the fish change the observed EOD amplitude in nearby portions of the fish's skin by distorting the EOD field. Mormyromasts encode local amplitude of the EOD-driven field via a latency code in which they fire a burst of 2-3 spikes between 3 and 10 ms after the EOD [Sawtell and Williams 2008]. This activity is translated to a firing rate code in the electrosensory lobe via a layer of sensory processing neurons prior to the efferent cells [Meek, Grant and Bell 1999; Metzen et al 2008].

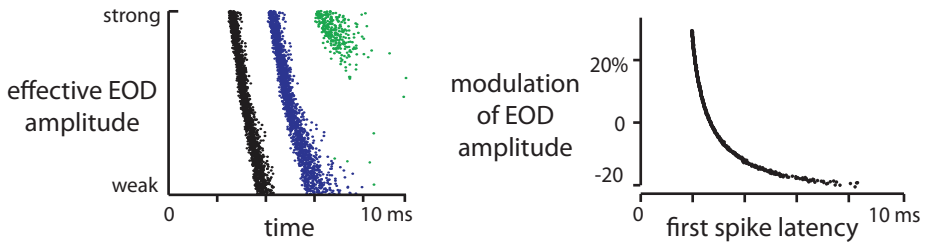


Figure 3.3: Relationship between EOD amplitude and mormyromast spiking, reproduced from [Sawtell and Williams 2008]. **Left:** spiking response of mormyromasts to the fish's own EOD, as EOD amplitude is varied. First, second, and third spikes per EOD are colored black, blue and green respectively. **Right:** Latency of the first EOD-evoked spike as a function of fold modulation of the EOD amplitude.

3.1.2 Effects of posture on the EOD field, and their cancellation

In previous work, Sawtell and Williams studied the joint effects of posture and object position on the amplitude of the EOD-evoked field at the fish's skin [Sawtell and Williams 2008]. They found that modulation due to tail bends of $\pm 30^\circ$ were of comparable of greater magnitude than modulations due to a conducting metal rod placed 5 mm from the surface of the fish's skin.

Both tail bends and object placement have a strong effect on mormyromast spike latency, as seen in Figure 3.4d. But in efferent cells of the fish's electrosensory lobe, which receive both sensory and corollary discharge input, the effect of tail bends is absent, while the cell remains responsive to the metal rod. (Note however that the response is not fully independent of tail angle.) This suggests that the efferent cell is able to form a posture-specific negative image, which cancels EOD modulation caused by the position of the fish's tail while leaving intact any modulation arising from external objects. If the mechanism of posture-specific negative image formation is the same as the temporally-structured negative image framework introduced in the previous chapter, we would guess that the efferent cell must receive input from granule cells that encode both the timing of the EOD command and the position of the fish's tail.

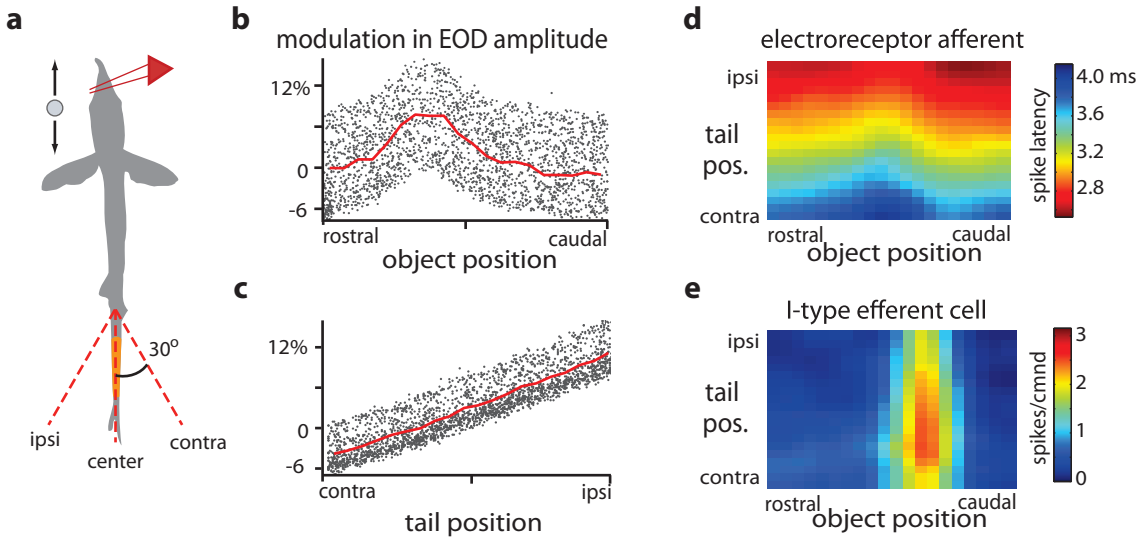


Figure 3.4: Combined effects of object location and tail bends on neural representation of objects, reproduced from [Sawtell and Williams 2008]. **a.** Experimental setup. A 2mm-diameter metal cylinder held 5mm from the fish's skin was moved alongside the head between the tip of the chin appendage and the gill cover, while the tail was moved between $\pm 30°$. EOD amplitude was measured using a recording dipole (in red) placed rostral to the fish's eye. **b.** The EOD amplitude at the dipole location as modulated by the metal rod. Each point corresponds to a single tail angle + object position pair; the response to the object averaged over tail angles is given by the red line. **c.** EOD amplitude at the dipole as modulated by tail movements. The red line is the response to the tail averaged over object locations. **d.** Spike latency of a mormyromast near the fish's head. Color indicates time from EOD to first spike for each combination of tail angle + object position tested. There is a clear effect of both the object and the fish's tail on the mormyromast response. **e.** In the firing rate of efferent cells of the cerebellum-like electrosensory lobe, the effects of tail angle are largely removed, while object location is still reflected.

3.1.3 Evidence for posture-specific negative images in efferent cells

In his recent thesis work, Tim Requarth investigated formation of negative images that varied in amplitude as a function of tail position. As in the previous chapter, Tim used an awake preparation in which the fish continued to issue the EOD command, but the EOD itself was blocked by neuromuscular paralysis. At the same time, the EOD-triggered spiking of efferent cells were recorded extracellularly. With the EOD blocked, the only input to efferent cells in the electrosensory lobe is from granule cells, which receive corollary discharge input relaying information about the timing of the EOD command, as well as other sources which will be discussed in the following section. Each efferent cell receives input from on the order of 20,000 granule cells.

Next, the EOD command was used to trigger an externally-applied electric field mimicking the fish's own EOD, restoring sensory input to the efferent cell. While moving the fish's tail slowly back and forth, the amplitude of the applied field was scaled as a function of tail angle, recreating the effect of posture on EOD amplitude. After ten minutes of pairing, the external field was turned off, and the EOD-triggered spiking of the efferent cell was again measured.

As seen below, pairing evoked a clear tail angle-specific change in efferent cell spiking. Subtracting the tail angle dependent firing rate of the efferent cell prior to pairing, and summing over time to get the average number of spikes per command, we obtain the change in corollary discharge-driven input to the efferent cell due to learning, which forms a negative image of the tail angle/field strength relationship imposed during pairing.

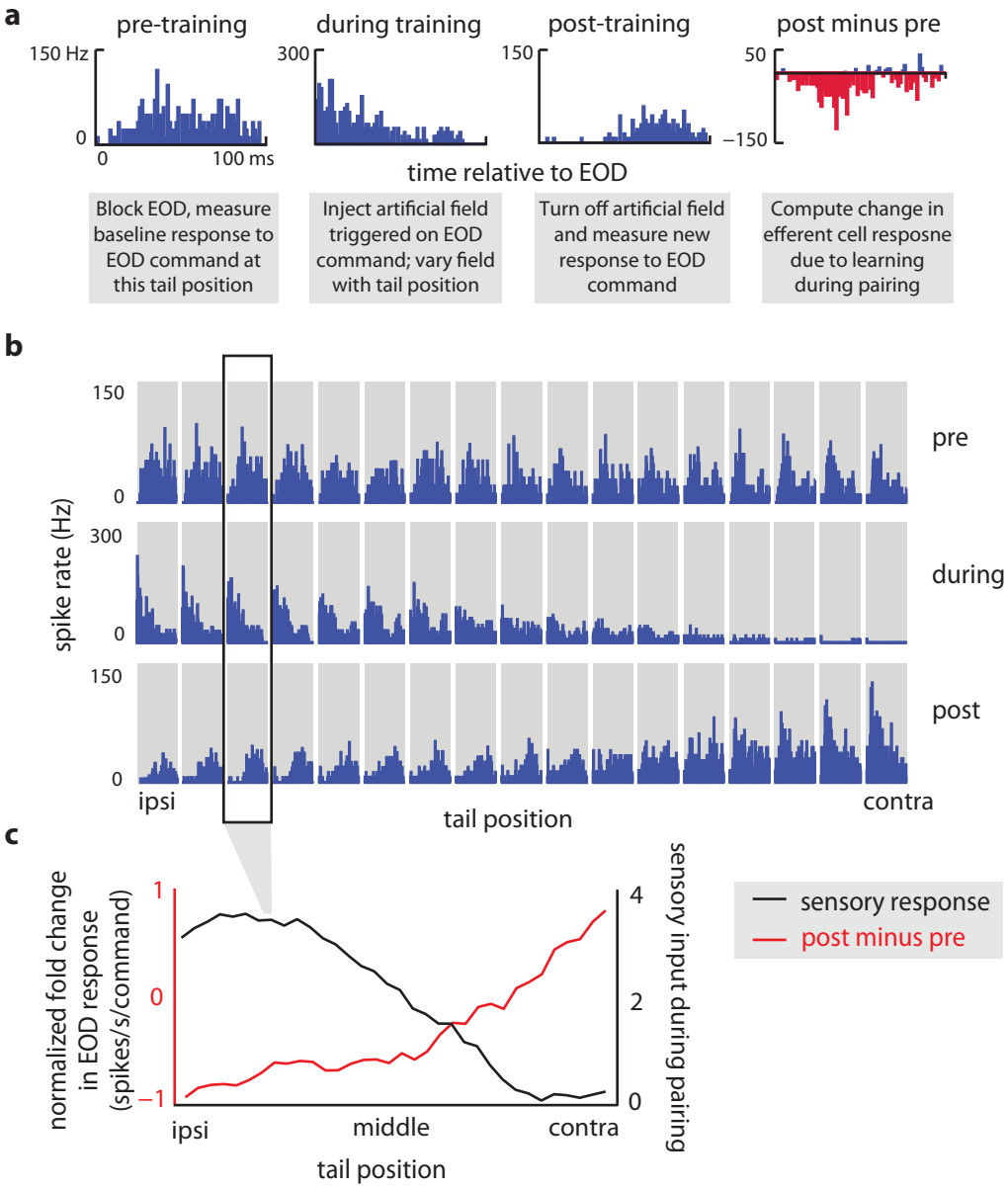


Figure 3.5: **a.** Example firing rate from an efferent cell of the electrosensory lobe before, during, and after pairing of the EOD command with the externally applied field, as well as the difference between the firing rate before and after pairing (far right). Firing rates are averaged over trials in which the tail was at the position highlighted in panel b. **b.** Top row: the firing rate of the efferent cell as a function of tail angle, triggered on the EOD command (x axis on each subplot is time relative to EOD command, as in panel a.) Middle row: firing of the efferent cell during pairing of an external field with the EOD command. The amplitude of the field *(continued on next page)*

Figure 3.5: *(continued)* was varied as a function of tail angle, driving tail-position-dependent spiking in the efferent cell. Bottom row: EOD-triggered spiking in the efferent cell after ten minutes of pairing shows adaptation to the paired signal. **c.** In black, the amplitude of the externally-applied field used during pairing, as a function of tail angle. The marked portion of the curve corresponds to the highlighted plots in panel b. In red, the time-averaged change in EOD-triggered spiking of the efferent cell is a learned posture-specific negative image of the black trace.

3.1.4 Encoding of posture in mossy fibers and granule cells

Negative image formation in efferent cells is mediated via anti-Hebbian synaptic plasticity on the set of granule cell - efferent cell synapses [Bell, Han, Sugawara and Grant, 1997; Roberts and Bell 1999]. Thus for negative images to be posture dependent, there must be some encoding of posture in the granule cell population.

Granule cells receive input from mossy fibers, which enter the electrosensory lobe from a variety of other brain regions. In addition to the mossy fibers conveying corollary discharge information from the EOD, previous studies have found tonically active mossy fibers that are unaffected by the EOD command, but carry information about posture. Some of these mossy fibers relay proprioceptive information from the spinal cord, and have tonic firing rates that can vary with either tail position or tail velocity. Other mossy fibers originate from motor control regions and carry corollary discharge signals of non-EOD-related motor commands [Requarth and Sawtell 2014].

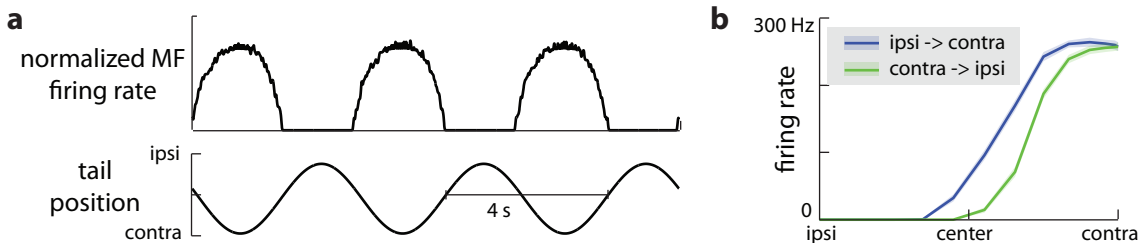


Figure 3.6: **a.** Firing rate of a tonically mossy fiber during sinusoidal movement of the tail by a manipulator. This mossy fiber responded to contralateral bends; other fibers preferred ipsilateral bends or had more complicated responses. **b.** Tuning curves computed from the firing rate in panel a. Tuning was significantly different if computed from ipsi-to-contra vs contra-to-ispi movements, though this could be an artifact of how the fish was restrained during tail manipulation.

Proprioceptive and EOD-related signals are combined in the granule cells, which receive input from 1-3 randomly selected mossy fibers. Tonic input from proprioceptive mossy fibers is not strong enough to drive granule cell spiking, but instead the tonic firing rate modulates the membrane potential of the granule cell. If a granule cell synapsing with both types of mossy fibers receives EOD command-driven input at a time when the firing rate of the tonic mossy fiber is high, the combined effect of the two inputs can be sufficient to evoke one or more spikes. Tonic mossy fibers can therefore modulate granule cell activity to a population of granule cells whose response to the EOD command is posture-specific.

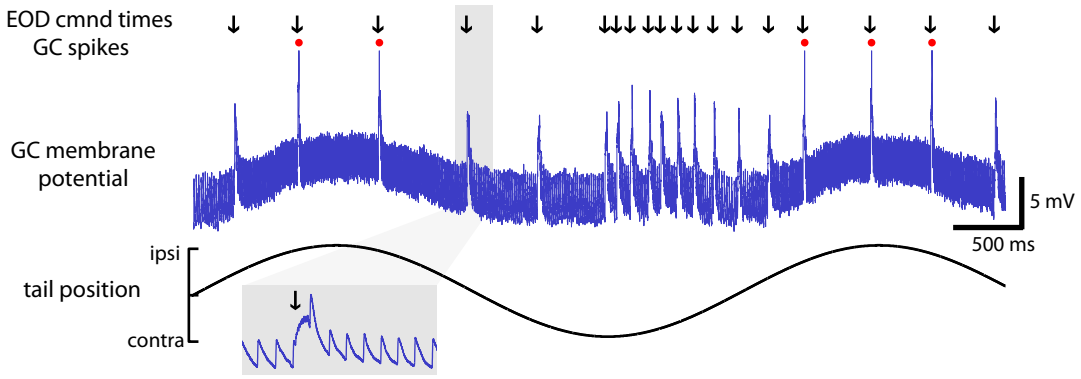


Figure 3.7: Intracellular recording from a granule cell receiving input from two mossy fibers, one conveying proprioceptive information and the other conveying the timing of the EOD command.Tail position was controlled by a manipulator and is plotted below the membrane potential, while EOD motor commands were recorded from the EOD command nucleus, and are indicated with arrows above. The tonic firing rate of the proprioceptive mossy fiber increases when the tail is ipsilateral, depolarizing the cell enough that EOD command-driven inputs evokes a spike (red dots). Highlighted in gray is a magnified portion of the membrane potential trace, showing EPSPs evoked by spikes in the tonic proprioceptive mossy fiber. The arrow marks the time of an EOD command, following which the granule cell receives a burst of EPSPs from the command-driven mossy fiber.

3.1.5 Objectives

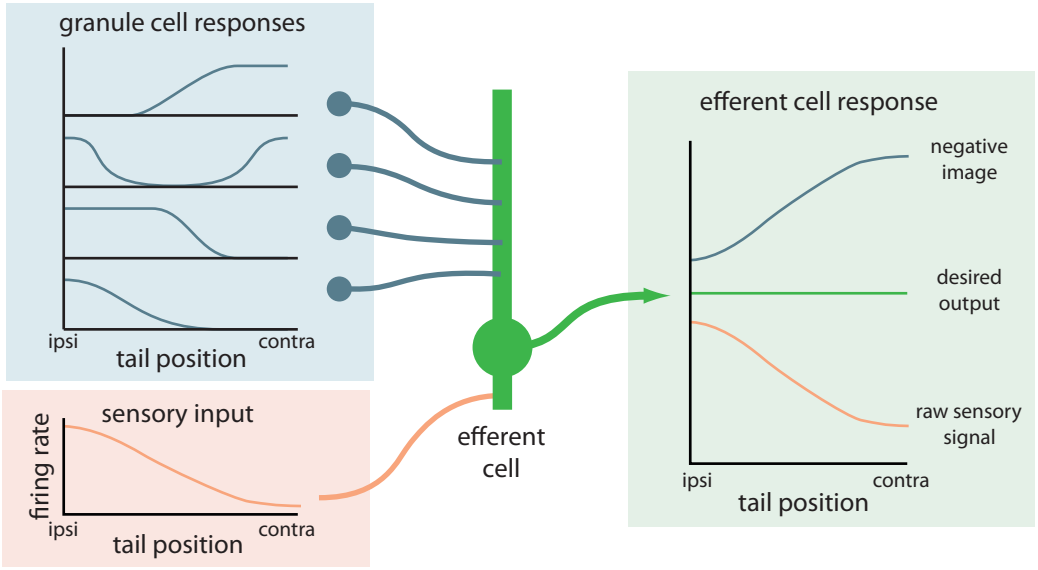


Figure 3.8: Circuit model for proprioceptive negative image formation, adapting the framework from the previous chapter. A basis of proprioceptively-modulated EOD command-driven granule cells (cartooned here in blue) allows the efferent cell to form a negative image that cancels the effect of tail position on sensory input from mormyromasts.

These results from previous studies suggest that sensory consequences of posture are cancelled in efferent cells via learning of posture-specific negative images from a set of proprioception and EOD command-driven granule cells. In this chapter, I will test this hypothesis directly by examining whether the granule cell model developed in the previous chapter provides a sufficient basis for forming posture-specific negative images, when input to model cells is expanded to include recordings from proprioceptive mossy fibers. This analysis is divided into two parts: first, I compare negative images formed by the model granule cell population to negative images recorded experimentally in a preparation similar to that of Figure 3.5, where the tail angle/EOD amplitude relationship was manipulated artificially. Negative images in the fish were found to be remarkably flexible, and efferent cells were able to fully or partially learn negative images to sensory input they would never experience in nature.

Next, I will address the broader role of proprioceptive negative images. It is difficult to experimen-

tally record EOD amplitude at the skin in freely behaving fish, therefore studies of posture-specific negative images have predominantly been restricted to single joints. By constructing an electrostatic model of the fish's field in a manipulable 3D mesh, I explore the effect of more complicated families of bends on the amplitude of the fish's field. I then study the capacity of the granule cell basis to form families of negative images across postures, focusing on the different computational problems that must be solved by efferent cells with receptive fields at different locations on the fish's body.

3.2 Methods: Modeling negative image formation in the active system

The granule cell model and the model of learning in efferent cells are predominantly the same as in the previous chapter. In this section I discuss proprioceptive mossy fiber input to granule cells, as well as a fitting technique I used to find the input-output function of model granule cells. The large families of postures I wished to study in this chapter would require thousands of simulations of the model granule cell population; because full simulation was intractable, I fit the input-output functions of model granule cells using a restricted set of simulations, and used these fits to generate granule cell activity for study of negative image formation.

3.2.0.1 Granule cell model

The full construction of the granule cell model is described in the previous chapter. Briefly: I model granule cells as leaky integrate-and-fire cells with three dendritic claws that form synapses at random with five classes of mossy fiber, early, medium, late, pause, and tonic, named for the time they spike relative to the EOD (random here means that each synapse is assigned independently). I fit the model to a population of 170 recorded granule cells by finding the mossy fiber inputs and synaptic weights that best accounted for the recorded response, obtaining a distribution of connection probabilities and fast and slow synaptic weights from mossy fibers to granule cells. I can then sample these distributions to generate an arbitrary number of synthetic granule cells.

Unlike the passive system, sensory responses to the EOD in the active system are restricted to short delays. I will therefore ignore medium, late, and pause mossy fibers in this chapter, and only model granule cells receiving early and tonic input.

3.2.1 Adding proprioceptive mossy fibers as model inputs

The tonic mossy fibers of the first chapter are in reality tonically-active proprioceptive fibers. Because I only considered negative image formation in a paralyzed, stationary fish, tonic mossy fiber firing rates did not vary in the model, and they only served to add some variability to granule cell spiking. Tonic inputs could not be fit in the same way as EOD-triggered inputs to granule cells, because their effect is washed out in the EOD-triggered average response of recorded cells. Instead, the probability and strength of tonic input in synthetic granule cells were obtained in two steps. First, probability of a granule cell getting tonic input was computed from the recorded granule cell population, in which cells receiving tonic input could be easily identified. And second, mean and variance of EPSP amplitudes were measured experimentally under postures in which tonic firing rates were low and single EPSPs were well isolated. To generate fast and slow synaptic weights for a model tonic input, I first randomly generated an EPSP amplitude from the measured distribution, then drew a ratio of fast to slow synaptic weights from the distribution of fit weights, and scaled these to match the drawn EPSP amplitude.

In the previous chapter, granule cell spiking thresholds were measured experimentally from the baseline membrane potential of the cell- which, if the cell is receiving tonic input, is above its resting potential. I used this same convention here, and defined granule cell thresholds relative to the depolarization due to tonic input, calculated using the tonic firing rate when the fish's tail is straight. Because proprioceptive input is not observed to drive granule cell spiking on its own, I discarded model cells in which tonic input drove granule cell spiking (which occasionally happened if a model cell received very sharply posture-modulated, strong tonic input and had a low spiking threshold.)

3.2.2 Fitting granule cell input-output functions using Bézier splines

Simulating spiking responses of granule cells over large families of postures is prohibitively slow: running enough simulations to get average responses of 20,000 granule cells at a given posture takes several hours on a desktop computer, and studying interaction of bends at multiple locations and angles quickly accumulates to thousands of simulations. But because I am not interested in the temporal aspects of negative images, I can simplify the problem to calculating the granule cells' spikes per command, a value that depends only on the firing rate of its proprioceptive mossy fiber input or inputs, a 1-2 dimensional parameter. Rather than re-simulate granule cells at each posture, I simulated granule cell spiking for a range of proprioceptive mossy fiber firing rates, and use the results to fit an input-output function for each granule cell. To generate granule cell responses to arbitrary postures, I then had only to determine the appropriate mossy fiber firing rates at that posture, and calculate the corresponding granule cell responses from the fit functions.

I reduced the number of simulations to run/parameters to fit further by assuming that in granule cells receiving input from two tonic mossy fibers, the two inputs could be approximated by a single effective tonic input, with firing rate given by the sum of the two mossy fibers' firing rates weighted by the relative amplitude of their EPSPs. Thus granule cell firing rate at a given posture is simply a function of the effective mossy fiber firing rate. This assumption seemed to match granule cell responses well, though the methods described here could be expanded to the case of two-dimensional inputs if desired.

The model granule cell firing rates were low (typically 0-3 spikes per EOD command), thus their input-output relationships resemble a monotonically increasing series of plateaus connected by smooth steps; see Figure 3.10 for examples. I fit these shapes with a set of plateaus of variable width placed at integer firing rates, connected by cubic Bézier curves, with a few restrictions imposed to reduce the number of free parameters and ensure the fit granule cell firing rate is a well-defined function of the input mossy fiber firing rate. Figure 3.9 shows an example with the fit parameters labeled in red:

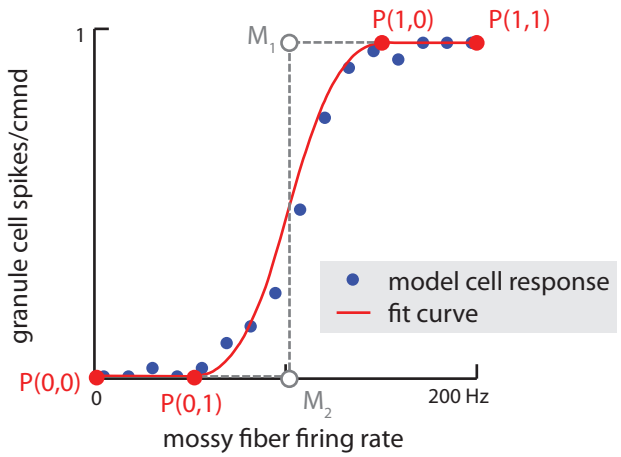


Figure 3.9: In blue, firing rate of a model granule cell plotted against the firing rate of its mossy fiber input. Superimposed in red is the Bézier spline fit: a cubic Bézier curve connecting plateaus at $r_{\mathrm{GC}} = 0$ and $r_{\mathrm{GC}} = 1$. Points labeled in red are fit to the model cell responses, while points in gray control the shape of the spline, and are fixed based on the values of the red points.

Given a set of mossy fiber firing rates $r_{\mathrm{MF}} \in [0, r_{\max}]$, model granule cell responses fall within a range $r_{\mathrm{GC}} \in [a, b]$ with $a, b \in \mathbb{N}_0$. For each natural number $n \in [a, b]$, I define two endpoints $x_{(n,0)}$ and $x_{(n,1)}$ between which $r_{\mathrm{GC}} = n$. Interpolating between plateaus is a cubic Bézier curve is defined by the parametric function:

$$B(t) = (1-t)^3 \mathbf{P_{(n,1)}} + 3(1-t)^2 t \mathbf{M_1} + 2(1-t)t^2 \mathbf{M_2} + t^3 \mathbf{P_{(n+1,0)}}, \quad t \in [0, 1]$$

where

$$\mathbf{P_{(n,1)}} = \begin{pmatrix} x_{(n,1)} \\ n \end{pmatrix}, \mathbf{P_{(n+1,0)}} = \begin{pmatrix} x_{(n+1,0)} \\ n+1 \end{pmatrix}$$

Points $\mathbf{P_{(n,1)}}$ and $\mathbf{P_{(n+1,0)}}$ mark the start and end of the curve from the plateau where $r_{\mathrm{GC}} = n$ to the plateau where $r_{\mathrm{GC}} = n+1$, while $\mathbf{M_1}$ and $\mathbf{M_2}$ are points in $(r_{\mathrm{MF}}, r_{\mathrm{GC}})$ that control the curve's trajectory. To ensure the fit granule cell firing rate is a well-defined function of mossy fiber firing rate, I define $\mathbf{M_1}$ and $\mathbf{M_2}$ to be:

$$\mathbf{M_1} = \begin{pmatrix} (x_{(n,0)} + x_{(n+1,1)})/2 \\ n \end{pmatrix}, \mathbf{M_2} = \begin{pmatrix} (x_{(n,0)} + x_{(n+1,1)})/2 \\ n+1 \end{pmatrix}$$

this leaves only the set of points $\{x_{(n,0)}, x_{n,1}\}$, $n \in [a, b]$ to be fit. Restricting the placement of $\mathbf{M_1}$ and $\mathbf{M_2}$ allows me to convert the parametric function $B(t)$ to a function $B(r_{\mathrm{MF}})$, by solving the cubic equation

$$(1-t)^3 x_{(n,1)} + 3(1-t)^2 t \frac{x_{(n,1)} + x_{(n+1,0)}}{2} + 2(1-t)t^2 \frac{x_{(n,1)} + x_{(n+1,0)}}{2} + t^3 x_{(n+1,0)} = r_{\mathrm{MF}} \quad (3.1)$$

for t, and plugging this value into

$$r_{\mathrm{GC}} = (1-t)^3 n + 3(1-t)^2 tn + 2(1-t)t^2(n+1) + t^3(n+1)$$

(while not pretty, Equation 3.1 has an analytical solution with a single real root for the restricted values of $\mathbf{M_1}$ and $\mathbf{M_2}$ used here.) I initialized values of $\{x_{(n,0)}, x_{(n,1)}\}$, $n \in [a, b]$ near the plateau endpoints in the input-output function obtained from simulation of granule cell responses, then fit them by solving for the values which minimized the mean squared error between r_{GC} and the data. For curves whose endpoints were outside of the simulated range (eg the transition from 1 to 2 spikes/command in the leftmost cell in Figure 3.10), I fit a tanh function to the simulated points, and used the endpoint of the fit tanh function to initialize the Bézier curve fit.

Example fits in Figure 3.10 show the diversity of nonlinearities in the model granule cell population.

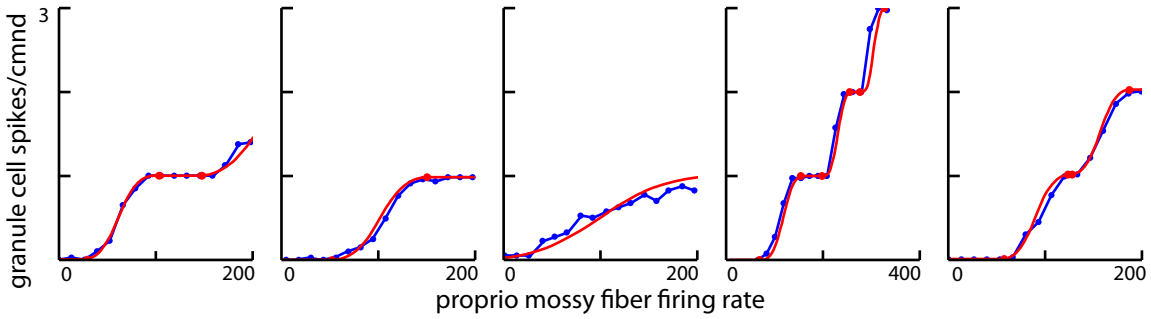


Figure 3.10: In blue, input-output functions from five example model cells, computed from the spiking granule cell model using 16 values for tonic mossy fiber firing rates, spaced evenly from 0 to 200 Hz. The fourth cell from the left received two proprioceptive inputs, so its response is plotted against the effective mossy fiber firing rate, computed as described above. In red, the Bézier splines fit to each cell, with fit points marked by dots.

3.2.3 Learning rule analysis

This is a simple adaptation of the stability analysis of the previous chapter, to examine how canceling a sensory perturbation at posture p affects the shape of the negative image at posture p'. Because I ignore time in this chapter and look only at spikes per EOD command in model granule cells and efferent cells, the shapes of the efferent cell EPSP and the anti-Hebbian learning rule are unimportant, and do not factor into the calculation.

Given a vector of synaptic weights w from model granule cells to a model efferent cell, a change Δw in synaptic weights yields a change in the membrane potential $V(p)$ of the model efferent cell:

$$\Delta V(p) = \sum_{i=1}^{N} \Delta w_i \cdot r_i(p)$$

where p is tail angle (posture), and $r_i(p)$ is the firing rate of the i^{th} granule cell at posture p. Anti-Hebbian learning rules drive the efferent cell membrane potential to a stable fixed point [Roberts and Bell 1999]; when sensory input perturbs the membrane potential from that point, the change in weights due to cancellation of a perturbation $\tilde{V}(p)$ is

$$\Delta w_i \sim -\tilde{V}(p) \cdot r_i(p)$$

Plugging in this equation for the effect of a perturbation $\tilde{V}$ on w, we get the effect of a perturbation at posture p on the negative image at each other posture p':

$$\Delta V(p') = \sum_{i=1}^{N} \tilde{V}(p) \cdot r_i(p) \cdot r_i(p')$$

Thus the matrix $M(p, p') = \sum_{i=1}^{N} r_i(p) \cdot r_i(p')$ describes the effect of a perturbation at posture p on the negative image at posture p'.

3.3 Methods: Modeling the fish's field

Numerous models of the fields of weakly electric fish exist in the literature, though previous studies have not investigated the effect of body bends systematically [Babineau, Longtin and Lewis, 2006; Chen, House, Krahe and Nelson 2005; Assad 1997; Englemann et al 2008]. I implemented two

such models: one which uses point charges to recreate the field at the fish's skin (Point Charge model)[Chen, House, Krahe and Nelson 2005], and one which models the fish's body explicitly as a high-conductance object containing a charge dipole, separated from its environment by the fish's skin, which forms a thin, resistive barrier (Body Mesh model)[Assad 1997].

The two models make very different predictions about the effects of bends on the field at the fish's skin. While bend effects have not been exhaustively studied experimentally, the Body Mesh model seems like a closer fit to available data, and I use this model for all study of negative image formation unless otherwise noted.

3.3.1 Justification of an electrostatic model

Both the Point Charge and Body Mesh model make the simplifying assumption that the fish's field can be modeled as an electrostatics problem. Generally, a time-varying electric field like the field generated by the EOD induces formation of a spatially-varying magnetic field, which in turn affects the form of the electric field. The electrostatic approximation assumes that the electric potential in the fish's body and surrounding water responds instantaneously to the changes in charge distribution which constitute the EOD waveform, and that as a result the induced magnetic field is negligible. While the EOD of mormyrid fish is a narrow pulse with duration of a few milliseconds, the dielectric relaxation time (the time it takes to respond to an induced field) of water is still very small compared to the frequency range of the EOD (<10 kHz); therefore the electrostatic approximation should be appropriate to the field formed by the fish.

3.3.2 The Point Charge model

I first modeled the fish's field as the sum effect of a set of point charges located inside the fish's body, based on the work of Chen et al [Chen, House, Krahe and Nelson 2005]. In this model, the potential at location p is simply the sum of the potential induced by each point charge:

$$\phi(p) = \sum_i \frac{q_i}{|p - x_i|}$$

where q_i is the charge of the i^{th} point charge, and x_i is its location (for simplicity, I modeled the field in a 2d slice and restricted all body bends to the plane of the slice.) The values of the charges

were constrained according to $\sum_i q_i = 0$, so that the net flux through the fish's skin fish was zero. I placed a set of 100 point charges along the midline of the fish, and adjusted their values to fit measurements of the fish's field along the length of its body, taken from [Sawtell 2006]. To reduce the number of free parameters, I adopted the approach used by [Babineau, Longtin and Lewis 2006] in a similar model, and defined the charge q_i as a function of position along the fish's central axis using a sum of Gaussians:

$$q_i = \frac{1}{\sqrt{2\pi}}\Big(\sigma_A^{-1}\, e^{-\frac{(x_i-\mu_A)^2}{2\sigma_A^2}} + \sigma_B^{-1}\, e^{-\frac{(x_i-\mu_B)^2}{2\sigma_B^2}} - 2\sigma_C^{-1}\, e^{-\frac{(x_i-\mu_C)^2}{2\sigma_C^2}}\Big)$$

where the parameters μ_{A-C} and σ_{A-C} were set to match the fish's field: (μ_C, σ_C) defined a narrow negative Gaussian centered over the electric organ in the fish's tail, while (μ_A, σ_A) and (μ_B, σ_B) characterized the positive charge over the remainder of the fish's body. I use two Gaussians for positive charge because I found that in models of charge distribution with a single Gaussian, the field strength fell off too rapidly in the direction of the fish's head (the measured field of the fish is elongated in the direction of the head, due to the fact that the fish's body has a higher conductance than the water around it.)

Most experimental measurements of the fish's field were collected in immobilized fish in a small recording tank, and the walls of the tank and surface of the water create insulating boundaries that distort the fish's field. As in Chen et al, I used the method of image charges [Jackson 1975] to model the effect of the fish's tank. This method recreates the effect of planar nonconducting boundaries by mirroring the charges in the fish's body across each wall of the recording tank. The addition of the tank distorted the field close the tank walls and magnified the effect of tail bends on field strength at the fish's skin, but otherwise did not alter the bend effects observed in the tank-free model.

Having placed the set of point charges according to the fish's posture, the field at the fish's skin is determined as follows. First, I computed the electric field generated by the point charges, defined at a point p to be the negative gradient of the sum of the point charge potentials:

$$E(p) = -\nabla\phi(p) = \sum_i \frac{q_i}{|p - x_i|^3}(p - x_i)$$

Then given the field, the voltage across the fish's skin is the inner product of the electric field at the fish's skin and a vector normal to the skin's surface, scaled by the ratio of skin and water

resistivities, ρ_{skin} and ρ_{water}, and the skin thickness t:

$$V(p_{\text{skin}}) = E(p_{\text{skin}}) \cdot \hat{n}(p_{\text{skin}}) \frac{\rho_{\text{skin}}}{\rho_{\text{water}}} t$$

The mesh developed in the Body Mesh model below could be used here to determine the normal vector $\hat{n}$, but for simplicity I assumed that $\hat{n}$ was orthogonal to the central axis of the fish, and constrained to the XY plane; this should hold for locations along the side of the fish's body and away from the face.

While this model is simple to set up and quick to simulate, it generated complex predictions of the effect of bends on the fish's field, which don't match well with experimental observations (see Results). Bends in the fish's body under this model affect the field in two competing ways. First, a bend simply changes the position of the point charges, which changes the magnitude of ϕ at the fish's skin. And second, bends change the angle the field vector E forms with the fish's skin, which depending on bend location can either enhance, cancel, or reverse the first effect. This second effect seems to be an artifact of the model not representing the fish's body conductance explicitly: because the fish's body conductance is higher than that of the water around it, the electric field at the skin should in fact be close to perpendicular with the skin, regardless of the fish's posture. I therefore switched to the Body Mesh approach described below, which models the effect of the fish's body more explicitly.

3.3.3 Electrostatic formulation of Body Mesh model

This model is adapted from Chris Assad's thesis work[Assad 1997], incorporating adjustments to make the system solvable using the Finite Element Method, as implemented in the solver environment GetDP [Dular and Geuzaine, 1997].

As in the previous model, I will work under the assumption that the fish's field can be approximated with an electrostatic model. Under these conditions, a distribution ρ of charge in space gives rise to an electric field, with associated potential ϕ related to f by Poisson's equation:

$$-\nabla^2 \phi = \frac{\rho}{\varepsilon}$$

where ε is permittivity of the medium. Multiplying both sides by the conductance σ gives

$$-\sigma\nabla^2\phi = \frac{\rho\sigma}{\varepsilon} = f \tag{3.2}$$

where f is the current source density. Note here that σ, ϕ, and f are all spatially-varying functions (since σ depends on whether a point is inside the fish's body, in the water, or in the fish's skin). I found that the fish's field could be reasonably approximated by defining f to be a dipole in the fish's tail, though a more advanced model of the current source density could also be used with this approach.

The problem domain over which Eq. 3.2 must be solved consists of two volumes within which the conductance σ is constant (inside and outside the fish), with constraints on two surfaces (the skin and the edges of the fish's tank).

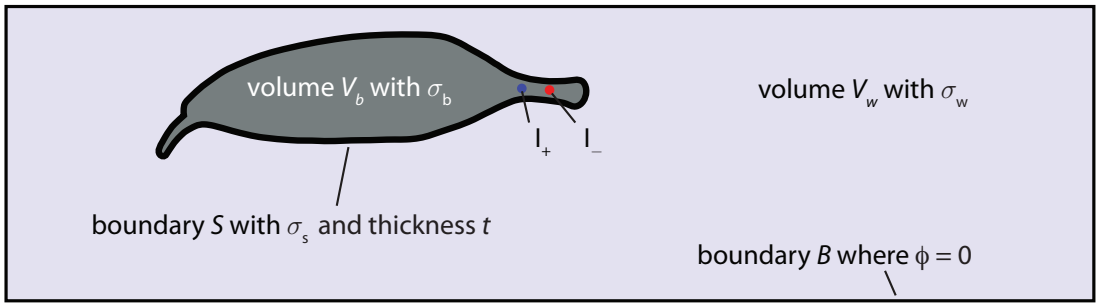


Figure 3.11: Schematic of the mesh model indicating regions of constant conductance, adapted from [Assad 1997]. I_+ and I_- are point charges in the fish's tail which give rise to the EOD field.

The skin was approximated as an infinitely thin, low-conductance boundary between the body and water, with potential ϕ_b at the inner surface of the skin, and ϕ_w at the outer surface. The voltage drop across the fish's skin is given by Ohm's law as $t\frac{\sigma_w}{\sigma_s}\nabla\phi_w$, where t is skin thickness and $t\frac{\sigma_w}{\sigma_s}$ is thus the effective resistance of the fish's skin. This gives the boundary condition

$$\phi_w - t\frac{\sigma_w}{\sigma_s}\nabla\phi_w = \phi_b$$

Noting that $\sigma_w\nabla\phi_w = -\sigma_b\nabla\phi_b$ on S, an equivalent boundary condition is $\phi_b - t\frac{\sigma_b}{\sigma_s}\nabla\phi_b = \phi_w$.

3.3.4 The variational method

To solve Poisson's equation over the unusually-shaped problem domain of the fish in its environment, I used Finite Element Methods (FEM), a numerical technique for finding solutions to differential equations with imposed boundary conditions, using the variational method. The two steps of FEM are 1) recasting the original differential equation into an equivalent weak formulation in which continuity requirements on the solution are weakened (see below), followed by 2) discretizing the problem domain into small regions over which the solution is roughly constant, allowing the reformulated problem to be solved numerically.

3.3.4.1 Weak formulation of Poisson's equation

The solution ϕ to Poisson's equation is the value of ϕ that minimizes the electrostatic potential energy of the system, $U_E = \frac{1}{2}\int_V \rho\phi\, dV$, where ρ is the charge density. The calculus of variations deals with problems in which a quantity to be minimized (here U_E) appears as an integral [Arfken, 1970]. Assuming a solution to the system exists, we define a test function ψ on the same domain as ϕ, and with the same boundary conditions. Then it can be shown that solving the Poisson equation is equivalent to finding ϕ such that

$$\langle \psi, -\sigma\nabla^2\phi - f\rangle = 0$$

for all values of ψ, where $\langle\cdot,\cdot\rangle$ denotes the inner product. Here $\langle u, v\rangle = \int_V uv\, dV$, giving

$$\int_V -\sigma\psi\nabla^2\phi - \psi f\, dV = 0$$

applying the product rule: $-\psi\nabla^2\phi = -\nabla(\psi\nabla\phi)+\nabla\psi\nabla\phi$, and the divergence theorem: $\int_V \nabla(\psi\nabla\phi)\, dV = \oint_S \psi\nabla\phi\, dS$, yields the weak formulation of Poisson's equation:

$$\int_V \sigma\nabla\psi\nabla\phi\, dV - \oint_S \sigma\psi\nabla\phi\, dS - \int_V \psi f\, dV = 0 \tag{3.3}$$

I broke this into domains over which σ is constant: the body of the fish and the water surrounding the fish:

$$\int_{V_b} \sigma_b\nabla\psi_b\nabla\phi_b\, dV - \oint_S \sigma_b\psi_b\nabla\phi_b\, dS - \int_{V_b} \psi f\, dV = 0$$

$$\int_{V_w} \sigma_w \nabla\psi_w \nabla\phi_w \, dV - \oint_S \sigma_w \psi_w \nabla\phi_w \, dS = 0$$

(since there are no external charges, f is only nonzero inside the body of the fish.) On S we have boundary condition $\phi_w - t\frac{\sigma_w}{\sigma_s}\nabla\phi_w = \phi_b$; or, noting that $\sigma_w \nabla\phi_w = -\sigma_b \nabla\phi_b$ on S, the equivalent $\phi_b - t\frac{\sigma_b}{\sigma_s}\nabla\phi_b = \phi_w$. Plugging this in gives

$$\int_{V_b} \sigma_b \nabla\psi_b \nabla\phi_b \, dV - \oint_S \sigma_b \psi_b (\frac{\sigma_s}{t\sigma_b})(\phi_b - \phi_w) \, dS - \int_{V_b} \psi f \, dV = 0$$

$$\int_{V_w} \sigma_w \nabla\psi_w \nabla\phi_w \, dV - \oint_S \sigma_w \psi_w (\frac{\sigma_s}{t\sigma_w})(\phi_w - \phi_b) \, dS = 0$$

Combining these two parts and noting that $\psi_w\phi_w$ and $-\psi_b\phi_b$ terms in the two surface integrals cancel, I arrive at the equation

$$\Big[\int_{V_b} \sigma_b \nabla\psi_b \nabla\phi_b \, dV - \int_{V_b} \psi_b f \, dV\Big] + \Big[\int_{V_w} \sigma_w \nabla\psi_w \nabla\phi_w \, dV\Big] - \oint_S \frac{\sigma_s}{t}(\psi_b\phi_w + \psi_w\phi_b) \, dS = 0 \quad (3.4)$$

These integrals can then be passed on to GetDP to solve numerically for ϕ_b and ϕ_w.

3.3.4.2 Discretizing the problem domain

To describe the geometry of the fish in its recording box, I constructed a 3d mesh of the fish's body in the open-source rendering tool Blender [Blender Foundation, 2014] from top- and side-view photos; I assumed transverse slices of the body were oval-shaped. The resulting mesh contained 1693 vertices and 3382 triangular faces: side, top, and perspective views of the mesh are shown below. I placed a rectangular box around the fish's body (not pictured) to represent the tank in which field measurements were taken, both to capture the effects of the recording box on the fish's field, and because this box is used to impose boundary conditions by the FEM solver. I then imported this mesh into Gmsh, a finite-element grid generator designed to operate in conjunction with GetDP [Geuzaine and Remacle, 2009]. I used Gmsh to cleanly tile the problem domain with a much higher-resolution 3d mesh. I then used GetDP to numerically evaluate Equation 3.4 by the finite element method, in which the test function ψ becomes the interpolating function between nodes in the 3d mesh (which approaches a delta function as the mesh becomes infinitely dense.)

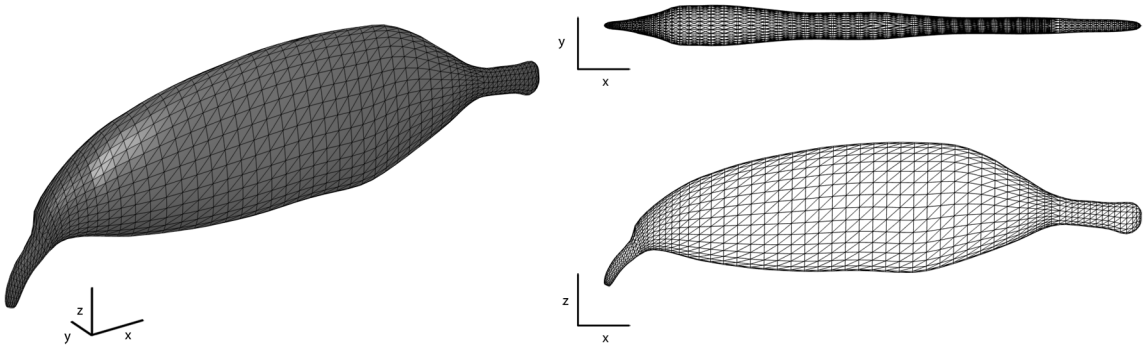


Figure 3.12: Mesh of fish body generated in Blender; the head with chin organ is facing left in all views. The fins and tail do not affect the electric field of the fish, and were not rendered.

3.3.5 Simulating body bends

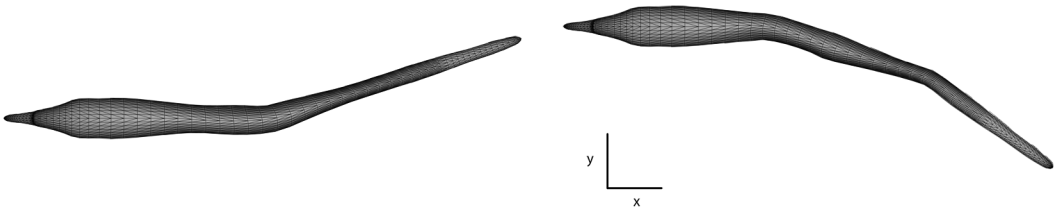


Figure 3.13: Two example bends of the fish mesh. **Left**, a single 20° bend, **right**, two 20° bends. Bends appear stiff because single joints are being affected, whereas naturalistic postures likely involve the correlated bending of multiple joints. Blender also has the capacity to distribute bends over multiple joints, but because effects on the fish's field are likely small, I did not investigate these here.

To simulate body bends in the fish, I rigged the Blender mesh to a 16-joint armature, an animation tool composed of a chain of rigid "bones" connected by ball-and-socket joints. The rigging process allows bends in the armature to be translated to deformations of the body mesh, producing realistic bending effects in the body. Pairs of joints were systematically bent through combinations of angles in the XY plane, and the resulting posed meshes were exported from Blender to Gmsh for modeling of the fish's field.

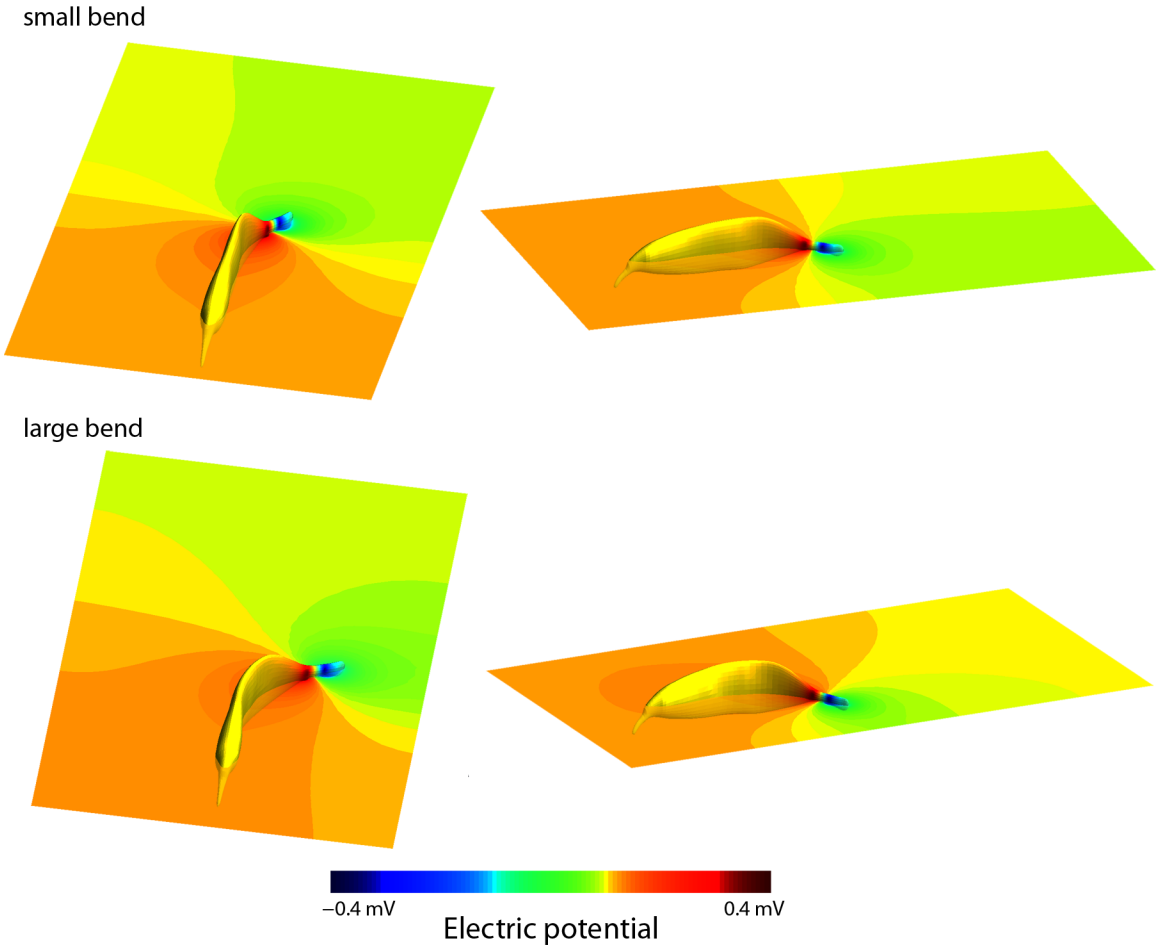


Figure 3.14: Top and side views of the 3D model of the fish's field in a coronal slice, for two curved bends; color indicates electric potential, hence the boundaries between solid colored regions are equipotential lines. Contrast is enhanced to make the differences in the field between the two postures more visible.

My original goal was to solve for the fish's field in the entire 3D mesh, but 3D implementation of the field model in GetDP proved problematic: while reasonable field predictions could be made within a single mesh, the solver seemed to encounter local minima that made comparison of results between meshes difficult. While further work could probably resolve this, the 3D solution was also very time-consuming to compute, especially given the large number of simulations needed to study interactions of multiple bends. (For example, testing pairs of bends of $\pm 20^\circ$ among four bend locations with two-degree resolution requires 2400 simulations.) I therefore reduced the original 3D mesh to a 2D model, by taking a coronal slice through the midline of the fish mesh, and solving for the field in this slice. This method neglects some features of the fish's shape which likely impact the fish's field, such as the narrowness of the tail compared to the body, however for a qualitative estimate of the effects of body bends, I will assume it is sufficient.

3.4 Results

3.4.1 Proprioceptive mossy fiber tuning curves are diverse

The posture dependence of the granule cell basis, and thus the negative image they form, is determined by the mossy fiber representation of the tail manipulation. Mossy fiber encoding of posture is not well understood: some cells vary their firing rate monotonically with tail position, while others seem to have preferred tail angles, or to encode tail velocity. In addition, proprioceptive mossy fibers are often tuned to a particular bend location; because the tail is being moved by an external manipulator, it is unclear exactly which joints are being bent, and at what angles.

Instead of trying to construct model mossy fibers with known tuning properties, I extracted tail angle-dependent tuning curves from recorded responses of a population of mossy fibers during tail manipulation. The downside of this setup is that I don't know exactly what the mossy fiber input to the granule cell basis is encoding. But because the mossy fiber recordings are derived from the same setup as the measured negative images I will study, I do know that they are the inputs that shape the granule cell basis during negative image formation. To extract tuning curves, I divided tail angles into 16 bins, and found the average tonic firing rate of each recorded mossy fiber as a function of the tail's location. 60 out of 80 recorded mossy fiber firing rates were monotonic

functions of tail position. The remaining 20 were non-monotonic, exhibiting clear peaks or dips in firing rate at intermediate tail angles. Most fibers' tuning curves were also direction dependent, although this could be an artifact of how the fish was restrained during recording (for example, ipsi→contra bends could have affected joints in a different pattern than contra→ ipsi bends.) Because an equal number of mossy fibers preferred ipsi- vs contra-directed bends, I arbitrarily chose to use ipsi-directed tuning curves as input to model granule cells.

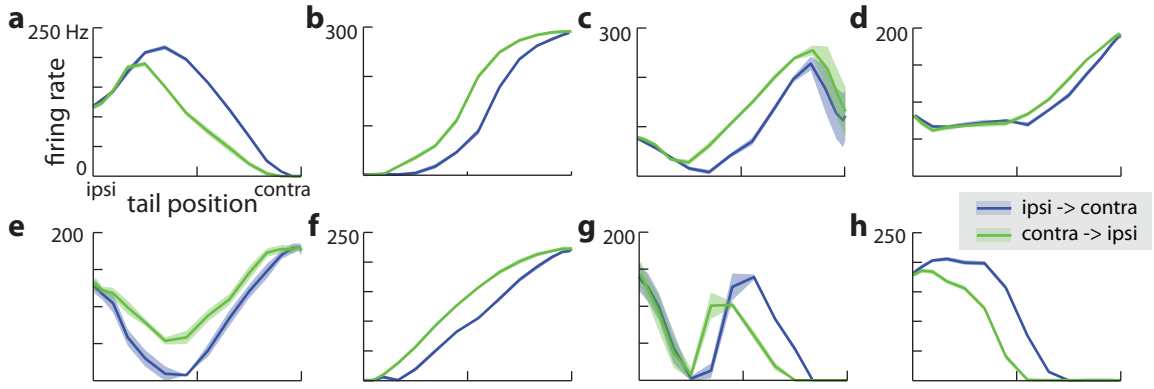


Figure 3.15: Example tuning curves of recorded mossy fibers, showing firing rate as a function of tail position. Cells in panels **b**, **d**, **f**, and **h** are (essentially) monotonic, while cells in panels **a**, **c**, **e**, and **g** are non-monotonic.

3.4.2 The model granule cell basis can form diverse proprioceptive negative images

I first tested the ability of the model granule cell basis to recreate experimentally measured negative images. Using the experimental setup outlined in Figure 3.5, Tim Requarth studied cancellation of artificial sensory consequences of posture. One advantage of this setup is that the relationship between tail angle and field strength is controlled experimentally, so it can be changed to take any form. In the previous chapter, we saw that the shape of the negative image learned to cancel an artificial sensory signal reflects the shape of the granule cell basis. Therefore by pairing unnatural sensory signals with posture, Tim sought to test whether posture-specific negative images could reveal limitations of the granule cell basis.

Strikingly, the fish was able to learn a diverse set of tail angle/field strength relationships. As

shown in Figure 3.4c, moving the tail contra to ipsi of a receptor near the head normally causes the EOD amplitude to increase roughly linearly with tail angle. However the fish could also learn the reverse of this relationship, as well as nonmonotonic relationships between tail angle and field strength. Moving the tail only on one side of the fish's body during pairing, or over a restricted range of bend angles, resulted in negative images that smoothly generalized over bend regions not seen during training.

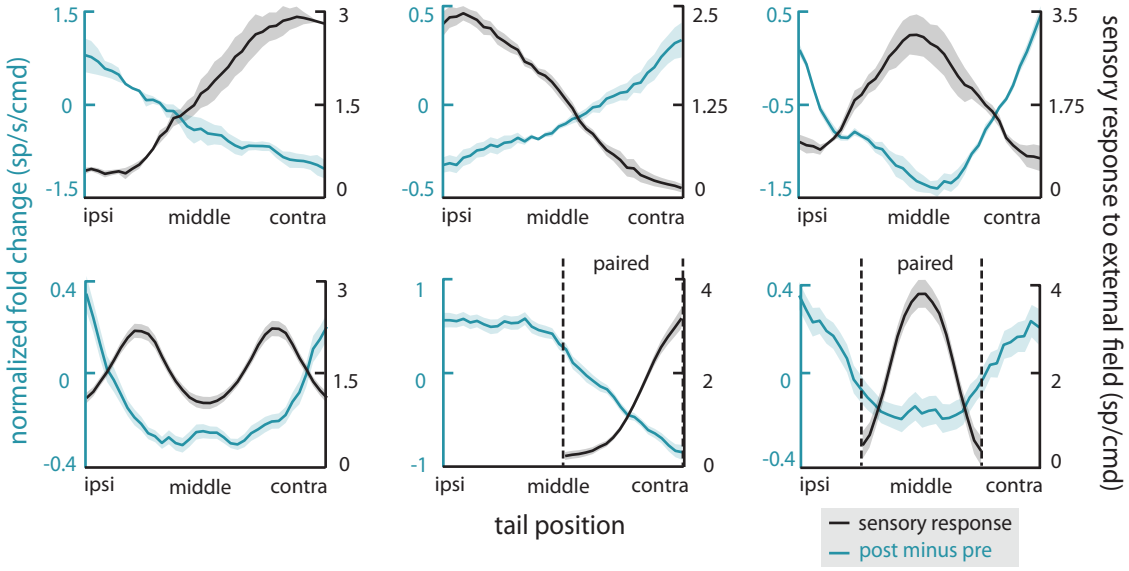


Figure 3.16: Externally applied tail angle/EOD amplitude relationships (black) and the resulting negative image learned by the fish (blue), as functions of tail angle. The upper-middle plot is the natural EOD relationship: the EOD amplitude is stronger when the tail is ipsilateral to the recorded cell's receptive field, and weaker when the tail is contralateral to the cell's receptive field. The fish is able to learn a surprising range of negative images with reasonable accuracy (although it failed to fully learn the W-shaped relationship on the bottom left.) The bottom center and right plots are generalization experiments, in which the fish's tail was only moved through the indicated region during learning.

I simulated negative image formation in a model efferent cell receiving proprioceptively-modulated input from a set of model granule cells. The model granule cells received proprioceptive input from the set of experimentally-recorded mossy fibers discussed in the previous section; synaptic

weights, inputs per granule cell, and spiking threshold were selected as described in the previous chapter. The negative images formed by the model were a good match for those in the data, with the exception of the W-shaped function and the V-shaped generalization function (bottom left and bottom right of Figures 3.16 and 3.18), in which the model learned the extremes of the negative image more slowly than was observed in the data.

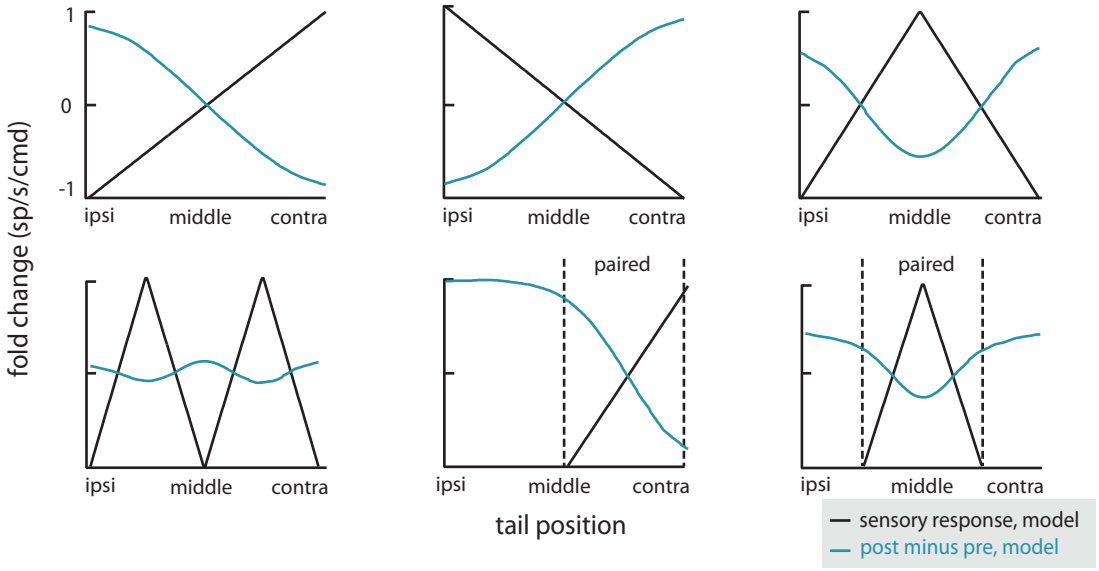


Figure 3.17: Simulated tail angle/EOD amplitude relationships (black) and the negative image learned by the model granule cell basis (blue), as functions of tail angle. Notable differences from the experimentally measured negative images are the W-shaped plot on the bottom left, and the V-shaped generalization experiment on the bottom right, both of which formed only shallow negative images at extreme bends.

The disparity between model and recorded negative images could simply mean that I am not using the correct mossy fiber basis as input to my model granule cells. The mossy fiber recordings used in the model were collected several years before the pairing study, and used a different tail manipulation setup- therefore small differences between mossy fiber input to the model granule cells and the actual mossy fiber responses during pairing could contribute to the model's inability to match the experiment. Alternatively, increasing the granule cells' thresholds decreases the number of cells active at central tail positions, and decreases the amplitude of the central peak in the W-

shaped curve; however making this change also causes the remaining negative images to look less like the data.

Interestingly, I found that the W-shaped negative image was particularly sensitive to the way the tail was moved during learning. In Figure 3.18, I simulated a sawtooth movement of the tail during learning. Modest changes in this movement resulted in substantially different negative images of the W-shaped pairing; no other pairing showed this extreme sensitivity.

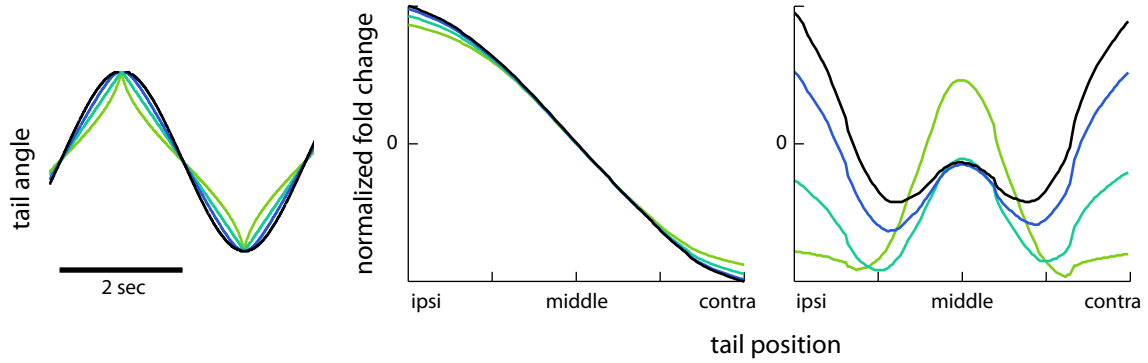


Figure 3.18: Negative images of the upper left and bottom left pairings from Figure 3.18, for different tail manipulations. Changing the tail movement from a sawtooth to a sinusoid increased the proportion of learning trials in which the tail was at extreme positions, and thus increased the magnitude of the negative image at the two bend extremes for the W pairing. Changing the tail movement to spend less time at extreme positions (light green line) drove a stronger negative image at small tail angles. All other pairings were much less affected by these changes, as seen from the linear pairing in the center plot.

To make sense of this result, I look at the matrix governing the dynamics of learning for this system, described in Section 3.2.3. This section derives the matrix $\mathbf{M}(p, p')$ that determines how a voltage perturbation at posture p affects the negative image at posture p'. Figure 3.19 the first five eigenvectors of M, and the first 25 eigenvalues. eigenvectors with large eigenvalues represent posture/field relationships that the system can learn quickly. The W-shaped pairing is strongly correlated with the 5th eigenvector, and weakly correlated with the first and third eigenvectors. Because the eigenvalue of the 5th eigenvector is very small, it does not have a strong effect on learning. Instead, depending on how the tail is manipulated during training, either the first or

third eigenvector will dominate the negative image of the W-shaped pairing, leading to a negative image that is peaked either at the two extremes (if the first mode dominates) or in the middle (if the third dominates).

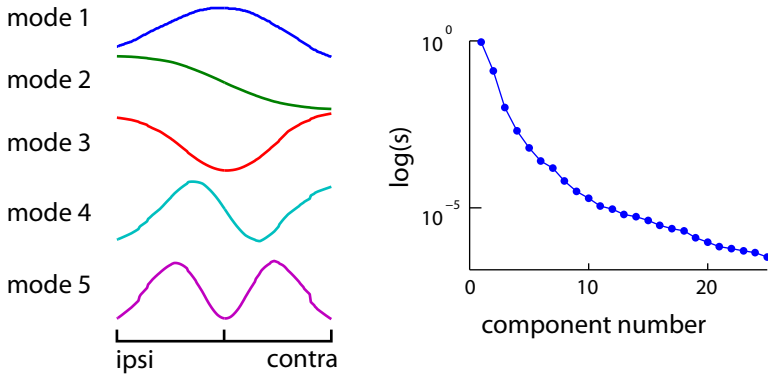


Figure 3.19: **Left:** First five eigenvectors of the learning matrix **M**, reflecting the five tail angle/field strength relationships learned most quickly by the system. **Right:** First 25 eigenvalues of **M**; the magnitude of an eigenvalue determines how quickly its corresponding eigenvector is learned during negative image formation.

3.4.3 Single- and double-joint bends and their effect on the fish's field

Having confirmed that model granule cells could form posture-specific negative images, I set out to study the effect of posture on the fish's field more systematically, using the electrostatic model described in the methods.

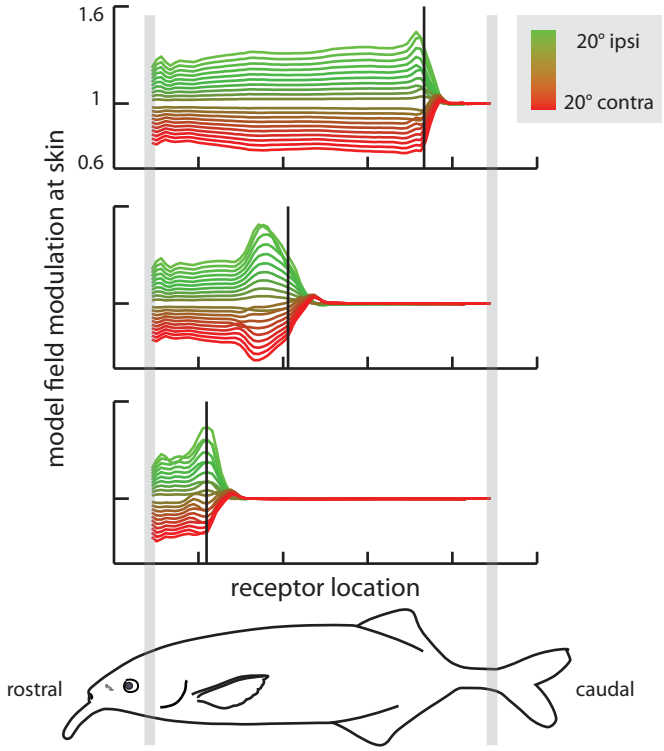


Figure 3.20: Effects of three example bends on field strength at the fish's skin, from the 2D mesh model. In each plot, I bent the fish mesh at the location indicated by the vertical black line, and measured the modulation in the EOD-generated field at the skin on one side of the model fish. Green lines show the modulation from bends 20° ipsi to the measurement site, while red lines show modulation from bends 20° contra, and the x axis is aligned with the fish's body as indicated below the plots. All bends induced strong modulation rostral of the bend location. Aside from magnifying effects near the location of the bend, which may depend on the detailed geometry of the mesh, fold modulation of the fish's field was roughly constant from the bend to the head. The field at skin caudal to the bend location was unaffected by the bend—this makes sense, as the distance from the electric organ to the skin does not change at these locations.

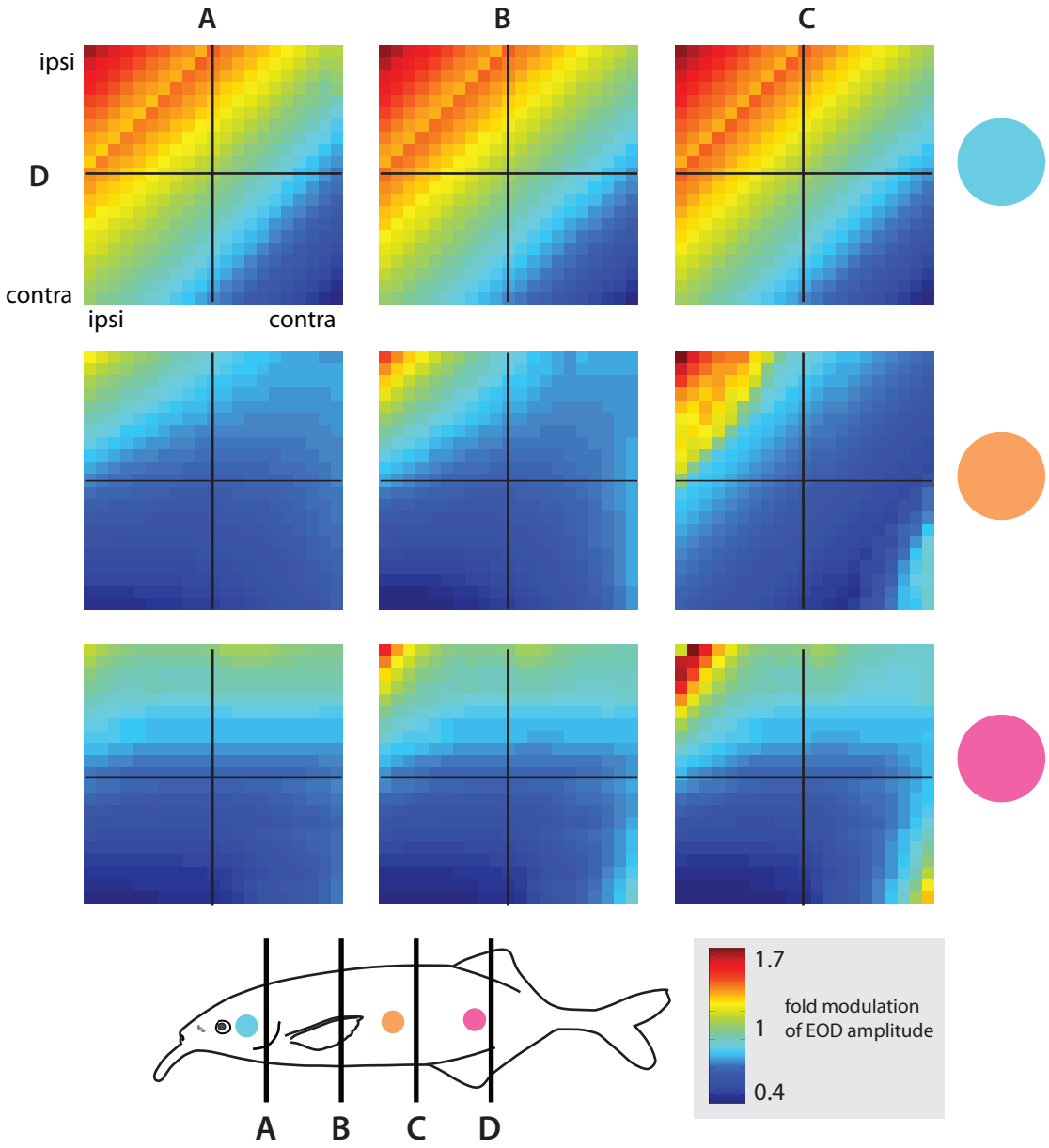


Figure 3.21: Interaction of pairs of bends, at joint angles up to ±20°, measured at three locations along the fish's body; bend and measurement locations are indicated on the fish schematic. I focused on the interaction of a bend near the fish's tail with bends further up the body. The top row shows effect of bend pairs on the field measured at the blue dot. For all three locations paired with the tail bend, the two bends summed completely linearly. By contrast, at the pink dot, modulation of the field was dominated by tail bends, with bends near the head having little effect (similar to the results from single bends.) At extreme bends, when the two joints are both bent ipsi or both contra, the strength of the measured field increases.

I observed three types of effects in pairs of bends, depending on the relative location of the two bends, and the location of the receptor (see Figure 3.21). First, bends caudal of the receptor tended to combine linearly: the field modulation at the receptor was proportional to the sum of the angles of all tail-ward bends. Second, bends rostral of the receptor only weakly modulated the strength of the field, as was the case in the single-bend experiment. Finally, for receptors in the caudal half of the fish, the strength of the measured field increased when both joints were bent in the same direction. This is most likely caused by the combined effect of the two bends bringing the fish's head closer to the skin: because the EOD field propagates along the length of the fish's body, the fish's head is more positively charged than its environment. Thus for extreme bends, and at receptors near the middle to caudal regions of the fish's skin, the fish's own head starts to register as an object to the electrosensory system.

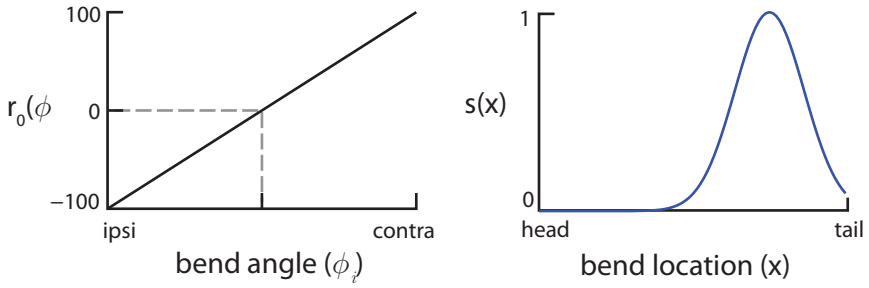
3.4.4 Forming negative images over families of postures

Using my results from simulations of the fish's field, I looked at how a model granule cell basis could cancel effects of pairs of bends at different sensor locations. Preliminary recording of proprioceptive mossy fibers indicates that they are tuned to a particular bend location along the length of the fish's body [Nate Sawtell, unpublished observations]. For bends near a mossy fiber's preferred bend location, the fiber's tonic firing rate is strongly modulated by bend angle, while for bends away from the preferred location, the fiber's tonic firing rate is unaffected or only weakly modulated. Recorded mossy fibers tended to be broadly tuned for bend location.

Based on these observations, I built a simple model of mossy fiber tuning as a function of bend angle and location. Given a set of bends at locations x_i along the fish's body, with bend angles ϕ_i, I set the response of the model mossy fiber to be the sum of the modulation induced by each bend individually, thresholded at 0 and with a maximum firing rate of 200 Hz. Mossy fiber responses were determined by a function $s(x)$ controlling the spatial tuning of each fiber, and a function $r_0(\phi)$ of mossy fiber firing rate modulation as a function of bend angle. Bend modulations of mossy fiber firing rate were relative to a baseline firing rate of the fiber when all joints were straight, f_0.

$$r(\mathbf{x}, \phi) = \Big[f_0 + \sum_{i=1}^{N_{\text{joints}}} s(x_i) r_0(\phi_i) \Big] +$$

I set $s(x)$ to be a Gaussian, and I defined $r_0(\phi)$ to increase linearly with bend angle, and picked a preferred body side (ipsi vs contra) at random with equal probability.



I tested negative image formation at the three receptor locations shown in Figure 3.21, for the three depicted combinations of tail and body bends, under the assumption that all bend angles were encountered with equal frequency. I first looked at the rate of negative image formation as a function of the number of mossy fiber inputs to model granule cells. For simplicity, I assumed that every granule cell received the same number of mossy fiber inputs, and that mossy fiber tuning curves spanned approximately a quarter of the fish's length. I found that learning rates were highly dependent on the location of the model efferent cell's receptive field, but that learning at all receptor locations tested was fastest when the number of mossy fiber inputs to granule cells was small, between 1 and 2 per cell. Interestingly, granule cell populations receiving only single mossy fiber inputs learned slower than those receiving two inputs in some conditions—this seems to be because the granule cell basis is slow to learn nonlinear interactions of bends with only single mossy fiber inputs.

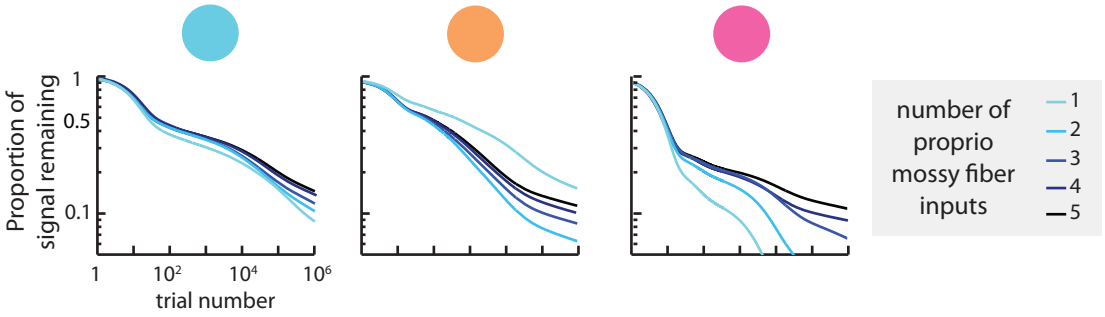


Figure 3.22: Negative image formation by a model efferent cell receiving input from a set of model granule cells, plotted as a function of the number of mossy fiber inputs granule cells received. The three plots show rate of negative image formation for different efferent cell receptive fields, from rostral (left) to caudal (right); precise locations are indicated by colored dots, which correspond to the locations used in Figure 3.21.

Note that a trial in this example corresponds to a single EOD. When at rest, the fish typically discharges its electric organ at a rate of about 5 Hz. Thus 1000 trials in Figures 3.22 and 3.23 corresponds to a little over three minutes of pairing- while 10^6 trials is on the order of two days.

I next looked at dependence of negative image formation on tuning width of the mossy fibers, assuming two proprioceptive mossy fiber inputs per granule cell. As with the number of inputs, learning rates varied as a function of receptor locations. For model efferent cells with receptive fields near the head, narrowly tuned mossy fibers initially learned at rates comparable to broader tuned mossy fibers, but their learning rate dropped off with time. This may be due to slower generalization by these cells, due to sharper tuning of granule cells. Conversely, for efferent cells with receptive fields near the tail, granule cell bases with narrowly tuned mossy fiber inputs learned fastest, perhaps due to the greater complexity of negative images formed near the tail.

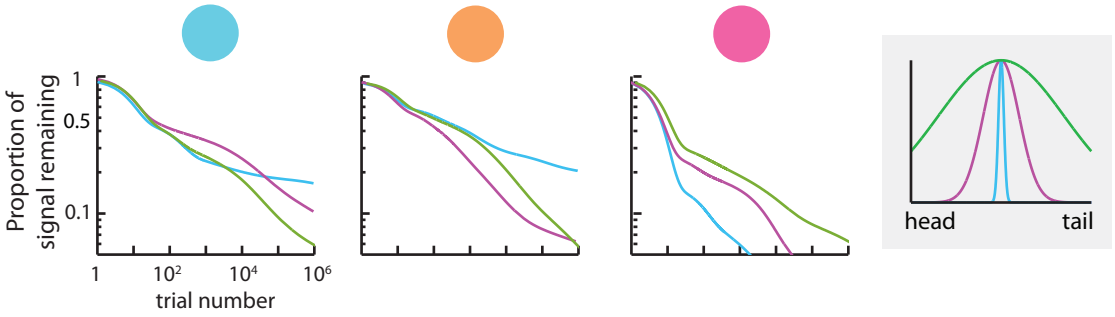


Figure 3.23: Negative image formation as a function of the tuning width of the mossy fiber basis; receptive field locations are again from Figure 3.21. The learning rate was highest for narrowly-tuned mossy fibers in model efferent cells with caudal receptive fields, while efferent cells with rostral receptive fields learned fastest with broadly tuned mossy fibers.

3.5 Discussion

In this chapter, I developed a model for proprioceptive negative image formation in the active electrosensory system, and studied the sensory consequences of posture in a field model of the fish. By modulating granule cell activation, tonically active proprioceptive mossy fibers create a granule cell basis that responds to the combined effects of posture and the EOD.

The performance of the granule cell basis is sensitive to the precise representation of posture in mossy fibers. In Section 3.4.4, I looked at one such factor, the tuning width of mossy fibers. I assumed the simplest possible mossy fiber tuning, in which effects of separate bends were scaled according to the fiber's tuning for bend location and summed linearly, however the actual responses of mossy fibers could be much more complex. Proprioceptive mossy fibers originate in the spinal cord, and as was discussed in Section 3.4.1, their response functions are diverse. In addition to encoding posture, some proprioceptive cells appear to be tuned to specific bend angles, or the velocity of tail movement.

Incorporating additional mossy fiber tuning properties into the model from Section 3.4.4 could yield further insight into how the mossy fiber basis can be tuned to increase the rate of negative image formation. Because the effect of posture varies with the location of an efferent cell's receptive

field, negative image formation may be aided by a body map in the electrosensory lobe, in which the response properties of mossy fibers innervating the ELL are matched to the receptive fields of efferent cells. Unlike the model of the previous chapter, which dealt with one-dimensional negative images in the paralyzed fish, the active system in the behaving fish must be able to form a diverse set of negative images simultaneously, and respond to changes in these negative images on biologically relevant timescales (recall that 10^6 trials in the plots of Section 3.4.4 corresponds to roughly 2 days of pairing.)

Using the models developed in this chapter, I hope to identify additional ways in which the mossy fiber basis can be tuned to solve the particular computational problems faced by efferent cells with different receptive fields. These results should hint at ways in which mossy fiber innervation of the ELL could be wired to optimize the learning rate of posture-specific negative images. Ideally, experimental investigation of these predictions could then search for an anatomically-constructed internal model among mossy fibers, by looking for patterns of mossy fiber tuning that are matched to the receptive fields of their target efferent cells.

Chapter 4

Odor representation in the *Drosophila* mushroom body

The *Drosophila* olfactory periphery closely resembles the mammalian olfactory system. Odorants bind to a set of olfactory receptor neurons in the antenna, which converge on a set of glomeruli in the antenna lobe, as in the mammalian olfactory bulb. The responses of each glomerulus are relayed by a set of projection neurons, the equivalent of mitral cells in vertebrates, to two higher brain regions: the lateral horn and the mushroom body. The lateral horn is associated with innate behavioral responses to odors, such as sex-specific responses to pheromones during courtship [Datta et al 2008], whereas the mushroom body is required for learned responses to odors. For instance, the mushroom body is required in classical conditioning, in which flies learn to avoid or approach otherwise neutral odors via the pairing of an odor with shock or reward [de Belle and Heisenberg, 1994; Dubnau, Grady, Kitamoto and Tully, 2001; McGuire, Le, and Davis, 2001].

The anatomy of the mushroom body bears strong resemblance to the circuitry of the cerebellum: signals from the sensory periphery are expanded into a sparse, high-dimensional representation. A small set of Purkinje-like neurons read out odor representations via a set of synapses that undergo plasticity during learning. In this and the following chapter, I construct a model of the *Drosophila* olfactory system and examine its performance in odor learning tasks. In this chapter I develop the model, expanding on previous work to more closely match experimental data from the olfactory

periphery and mushroom body. In the next chapter, I will discuss the circuitry thought to underly odor learning, and present two hypotheses by which that circuitry could mediate the acquisition of odor-specific behaviors.

4.1 Anatomy of the fly olfactory system

This section will review olfactory anatomy discussed in the introduction, and introduce terminology and abbreviations used in the rest of the chapter.

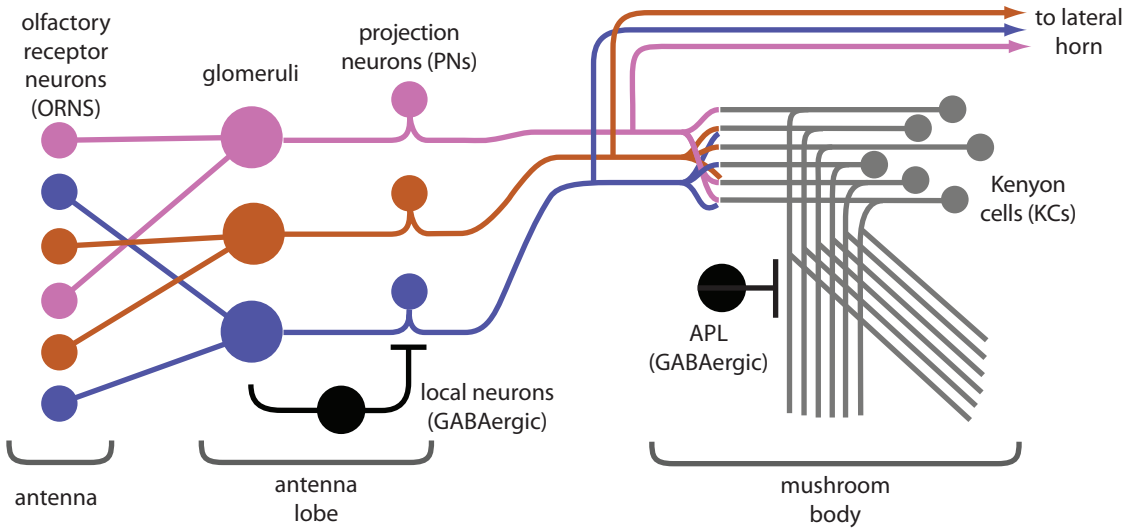


Figure 4.1: The olfactory processing stream in *Drosophila*, with all elements of the model marked.

Olfactory receptor neurons (ORNs) on the antenna and maxillary palps of the fly each express one of a family of roughly 50 odorant receptors. Most odors activate several classes of odor receptors and interaction can either increase or decrease ORN firing rates (ORNs are spontaneously active). In my model, I take ORN firing rates from a large dataset measuring the responses of 23 ORN types to a panel of 110 odors [Hallem and Carlson, 2006]. Roughly 30% of odor-ORN interactions in the dataset are inhibitory, and four odors elicit a purely inhibitory response. Responses of ORNs to odors are not uniform: some ORNs are characteristically excited by odors, while others are predominantly inhibited. In addition, odors which strongly drive one type of ORN often also strongly drive several others. Odors which strongly drive many/most classes of ORNs are called

"public" odors, and are often fruity odors like isopentyl acetate (banana) or food-related odors like yeast. Some other odors only activate a single type of ORN at typical tested concentrations; these are referred to as "private" odors. Example private odors are geranyl acetate (rose) and methyl salicylate (wintergreen).

ORN axons project to the antenna lobes, which are situated in a similar position to the olfactory bulbs in vertebrates. The antenna lobes are composed of a set of roughly 50 glomeruli, one for each ORN class, and all ORNs of the same class converge exclusively onto their designated glomerulus. Each glomerulus is read out by a small number of projection neurons (PNs; roughly five per glomerulus). Within the antenna lobes are additional local interneurons (LNs) that receive odor-evoked input, presumably from ORNs, and feed back onto glomeruli. There appear to be several classes of interneurons, with different patterns of glomerular innervation; LNs are predominantly inhibitory and GABAergic, though excitatory cholinergic LNs exist as well [Wilson and Laurent 2005; Olsen, Bhandawat, and Wilson, 2007; Shang et al 2007]. Like ORNs, PNs are spontaneously active and can be either excited or inhibited by odors; however unlike ORNs, odor-mediated inhibition of PNs can come from two sources: inhibition of the ORN class innervating a PN's glomerulus (called the cognate ORN), or lateral inhibition between glomeruli mediated by LNs. As will be discussed in the model, lateral inhibition is thought to play an important role in normalizing odor representations in the PNs so that effects of odor intensity are removed from the PN responses, while information about odor identity is preserved.

Each PN axon bifurcates outside the antenna lobe to innervate two regions, the mushroom body and the lateral horn (like the antenna lobe, both structures appear bilaterally in the fly brain.) Relatively little is known about the anatomy of the lateral horn, though the hypothesis that it drives odor-evoked innate behaviors has some experimental support [Datta et al 2008]. The mushroom body contains approximately 2000 Kenyon cells (KCs), small granule cell-like neurons. PN axons form excitatory connections with KC dendrites in the mushroom body calyx. Each KC has around 7 dendritic claws, each of which forms a single synapse with a single PN axon. PN-KC connectivity is random, in the sense that one KC claw forming a connection with a given PN does not alter the probability of the KC's other claws from synapsing with any other PN [Caron, Ruta, Abbott, and Axel, 2013; Murthy, Fiete, and Laurent, 2008; Honegger, Campbell, and Turner, 2011]. However

the distribution of connections is not uniform: some PNs send out more axonal projections and are more likely to synapse with KCs than others.

All 2000 KCs extend their axons in two structures, the medial and vertical lobes of the mushroom body. There are three subclasses of KC, distinguished by the order in which they emerge developmentally: α/β, α'/β', and γ KCs. The α/β and α'/β' KCs have bifurcating axons which send one branch to the medial lobe (forming the β and β' lobes) and one to the vertical lobe (forming α and α' lobes). γ KCs have a single axon, which extends only to the medial lobe. There is no difference in PN-KC connectivity across the three KC classes, and I will treat them interchangeably in this study. But some differences do exists between KC classes. Physiologically, α'/β' KCs are slightly more responsive to odors and are the only KCs which are spontaneously active (albeit at only around 0.1 Hz) [Turner, Bazhenov, and Laurent 2008]. And in memory tasks, one study found that neurotransmission from α'/β' KCs is required for acquisition and consolidation of odor memory, but not for retrieval, while α/β KCs are required only for retrieval [Krashes et al 2006].

Unlike the ORNs and PNs, KC responses to odors are very sparse, with a typical odor evoking spikes in 5-10% of KCs, and any given KC responding to 5-10% of odors on average (although narrowness of tuning varies across cells in the KC population, as will be discussed in the results portion of this chapter.) Contributing to KC sparseness is a fairly high spiking threshold in KCs, as well as recurrent inhibition of the KC population by an unusual cell called the Anterior Paired Lateral neuron (APL). Each mushroom body (in the left and right hemispheres of the fly brain) is innervated by a single massive APL neuron, which sends processes to the mushroom body calyx as well as the vertical and medial lobes. While it is tricky to distinguish axons from dendrites in flies, it appears that the APL receives input from KCs in the vertical and medial lobes, and inhibits KCs at their dendritic terminals in the calyx. Rather than fire action potentials, APL releases GABA at a rate proportional to its level of depolarization [Papadoupulou, Cassenaer, Nowotny, and Laurent, 2011]. APL appears to play a role in normalizing KC responses to odors, similar to Golgi cells in the cerebellum, although some studies have hypothesized a role of APL in the process of learning itself [Pitman et al 2011].

Thus, dense odor codes in the olfactory periphery are transformed to sparse, high-dimensional codes in the mushroom body. Previous work has established a basic model of odor representation in the

mushroom body, indicating that KC connectivity and sparseness are tuned to create a maximally high-dimensional representation of odors among KCs [Caron, Ruta, Abbott, and Axel, 2013; Luo, Axel and Abbott, 2010; Turner, Bazhenov, and Laurent, 2008]. Here, I refine previous models to more closely match observed responses in the olfactory periphery and mushroom body, and look at some properties of odor representation in the new model. In the next chapter, I will discuss evidence for the role of the mushroom body in odor learning, recent findings regarding neural readout of odor representations by KCs, and how this readout underlies the acquisition of novel associative odor memories in the fly.

4.2 Objectives

The goal of this chapter is to develop a dynamic firing rate/spiking model of the fly olfactory system, from ORNs up through the mushroom body. This work will expand upon a previous model built using time-averaged responses of olfactory neurons [Luo, Axel, and Abbott 2010]. I will focus on two key areas for improvement: first, the earlier model did not capture the dynamics of PN responses, which are shaped by slow lateral inhibition. As described in following sections, the inhibitory LNs mediating divisive normalization in PNs act on the order of hundreds of milliseconds; models that disregard this and capture just the time-averaged effect of lateral inhibition overestimate its effect on the excitatory input relayed by PNs to the mushroom body. And second, I will develop a more biologically realistic model for APL-mediated inhibition of KCs. In the original model, APL was effectively implemented by subtracting off the first principal component of PN responses to odors; as a result, odor representations in KCs were very well decorrelated, and odor learning in the model was straightforward. However this model of APL leads to several inconsistencies with recorded KC responses, such as a drop in active cells as odor concentration increases, which is the opposite of what is seen in vivo. Explicitly modeling APL responses to odors leads to a drop in the dimensionality of KC responses, as KCs become more correlated across odors, but it also leads to a distribution of KC tuning curves that is more clearly in line with what is observed experimentally. In the next chapter, I will discuss learning rules by which the mushroom body may overcome problematic correlations between odor representations.

4.3 Methods

4.3.1 Construction of the dynamic mushroom body model

4.3.1.1 Obtaining ORN responses to odors

Previously published data [Hallem and Carlson, 2006] records the response of 23 classes of ORNs to a panel of 110 different odors, as well as responses to nine fruit odors and ten chemical odors at each of four concentrations. Each odor was presented for six 500 ms trials, and ORN responses are reported in spikes per second relative to the baseline spontaneous firing rate of the ORN (also reported). While most odor-evoked responses were excitatory, approximately one quarter of odor/receptor combinations showed a drop in ORN firing rate in response to the odor, and four of the tested odors had purely suppressive effects on ORN firing. Electrophysiological recordings in ORNs show that odor-evoked spiking undergoes little adaptation [Bhandawat et al 2007], thus ORN firing rates in the dynamic model were modeled as a simple step response, filtered by the cell's synaptic membrane filter (unless mentioned otherwise, all model cells were assumed to have a membrane time constant of 10 ms):

$$\tau_m \frac{dORN}{dt} = (ORN_{\text{spont}} - ORN(t)) + sI(t)$$

where s is the odor-dependent ORN activation from the Hallem and Carlson data, and $I(t)$ is the stimulus timecourse, usually a 500 ms step.

In some instances, I generated additional synthetic odors from the Hallem and Carlson dataset. To study representations of odor mixtures, I assumed for simplicity that ORN activation by the set of mixed odors was a linear sum of activation by single odors. If odor concentration is not high enough that odorants are competing for ORN receptors, this assumption should be reasonably accurate. I also used synthetic odors to study dimensionality of odor representations in different stages of olfactory processing; I typically generated 4890 synthetic odors, to increase the total number of odors in the dataset to 5000. For each synthetic odor, I randomly drew a response for each ORN from its set of responses to the 110 real odors,

Unlike previous work by Luo et al, I chose not to generate synthetic responses for the 18 olfactory glomeruli not recorded by Hallem and Carlson. While this could diminish the model's performance

on odor discrimination tasks, it avoids artificially increasing the dimensionality of odor representations in the ORNs and subsequent levels of processing. It therefore provides a reasonable baseline for learning: additional glomerular responses could only improve performance.

4.3.1.2 Modeling projection neuron responses to odors

In previous work, Olsen, Bhandawat and Wilson fit a PN model to two features of the ORN-PN transformation: first, PNs have a nonlinear input-ouput transformation of their cognate ORNs, in which the PN amplifies weak ORN responses and saturates at stronger ORN responses [Olsen, Bhandawat, and Wilson, 2010]. And second, a population of local inhibitory interneurons (LNs) divisively normalize PN responses by an amount proportional to the total ORN input to the antenna lobe. LN normalization effectively stretches the input-output function of PNs, such that the more glomeruli a given odor recruits, the more strongly it must activate a given glomerulus to drive up the firing rate of its cognate PN. Together, the two effects reduce the effect of stimulus intensity on the PN population response, and equalize PN input to the mushroom body across odors.

The model exhibited two shortcomings: first, it gave only the average firing rate of the PNs over a 500 ms stimulus presentation, whereas recorded PN spiking exhibits multiphasic structure which is likely to impact Kenyon cell spiking. And second, the Olsen model was fit using only glomeruli which were strongly excited by odors, and did not allow PN firing rates to drop below their spontaneous levels. Like ORNs, PNs are spontaneously active in the absence of odor, and are inhibited during odor presentation via two mechanisms: either through odor-evoked suppression of their cognate ORNs, or through lateral inhibition from other glomeruli [Wilson and Laurent 2005].

4.3.1.3 Adding odor-evoked inhibition of projection neurons

I addressed the second shortcoming by building a model that permits odor-evoked inhibition of PNs. Temporarily ignoring lateral inhibition, odor-evoked PN responses in the Olsen model are given by:

$$PN - PN_{\text{spont}} = R_{\max}\frac{(ORN - ORN_{\text{spont}})^{1.5}}{\sigma^{1.5} + (ORN - ORN_{\text{spont}})^{1.5}}$$

where PN_{spont} and ORN_{spont} are the spontaneous firing rates of a given PN and ORN (ie for a single glomerulus), and $R_{\max}$ and σ are parameters fit to the data. The 1.5 exponent was also fit to the shape of the ORN-PN nonlinearity.

To allow PN firing rates to drop below their spontaneous values, I fit this model with a thresholded tanh function:

$$PN = \Big[R_{\max} \tanh\big(g(ORN - ORN_{\text{spont}} + c)\big) + PN_{\text{spont}}\Big]_+$$

where $R_{\max}$ is defined as before. g and c are new parameters of the tanh model, which were fit using nonlinear least squares to minimize the cost function formed from the difference of the two models:

$$C = \int_0^{x_{\max}} dx \Big(R_{\max} \frac{x^{1.5}}{\sigma^{1.5} + x^{1.5}} - R_{\max} \tanh\big(g(x + c)\big)\Big)^2$$

which was evaluated using a rectangle approximation. Note that x replaces $ORN - ORN_{\text{spont}}$; thus $x_{\max}$ was set to the largest experimentally-observed ORN response.

4.3.1.4 Adding lateral inhibition to the revised model

Lateral inhibition between glomeruli is mediated by a population of GABAergic local neurons (LNs) in the antenna lobe. LNs are broadly tuned to odors, but their tuning curves are heterogeneous, as is the spatial extent of their innervation of glomeruli. However, Olsen et al found that the net effect of the LN population is to divisively normalize PN responses by the total ORN input to the antenna lobe:

$$PN - PN_{\text{spont}} = R_{\max} \frac{(ORN - ORN_{\text{spont}})^{1.5}}{\sigma^{1.5} + b^{1.5} + (ORN - ORN_{\text{spont}})^{1.5}}$$

where $b = m \sum_i (ORN_i - ORN_{i,\,\text{spont}})$. I kept this approach in the tanh model, introducing a divisive inhibition term with additional parameters k_1 and k_2:

$$PN = \Big[R_{\max} \tanh\Big(g(ORN - ORN_{\text{spont}} + c) \frac{k_1}{b + k_2}\Big) + PN_{\text{spont}}\Big]_+ \tag{4.1}$$

k_1 and k_2 were again fit with nonlinear least squares, by finding values to minimize

$$C = \int_0^{b_{\max}} db \int_0^{x_{\max}} dx \Big(R_{\max} \frac{x^{1.5}}{\sigma^{1.5} + b^{1.5} x^{1.5}} - R_{\max} \tanh\big(g(x + c) \frac{k_1}{b + k_2}\big)\Big)^2$$

where $b_{\max}$ was the maximum value of b computed from the Hallem and Carlson dataset.

4.3.1.5 Spontaneous projection neuron activity

For the PN's threshold (the value of $ORN - ORN_{\text{spont}}$ at which $PN > 0$) to be independent of b in this model, PN_{spont} should scale like $k_1/(b + k_2)$; I therefore set

$$PN_{\text{spont}} = ORN_{\text{spont}} \frac{k1}{b + k2}$$

this gives a mean spontaneous firing rate of 14.5 Hz among modeled PNs, compared to 13.3 Hz in ORNs in the Hallem and Carlson dataset. Complete data on the spontaneous firing rates of PNs is unavailable, but they are known to show spontaneous activity of several Hz; the value used here can easily be adjusted in future models if additional information about PN responses becomes available.

4.3.1.6 Building the dynamic projection neuron model

Experimental evidence suggests that PN responses are shaped by slow lateral inhibition mediated by LNs. Both LNs and PNs are inhibited by GABA: IPSPs in LNs have a time constant of 100 ms and are completely abolished by picrotoxin, suggesting they are mediated by GABA_A. IPSPs in PNs are a sum of a fast ($\tau = 100$ ms) picrotoxin-abolished component and a slower component with $\tau = 400$ ms which was abolished by the GABA_B antagonist CGP54626, suggesting a combination of GABA_A and GABA_B- mediated inhibition in PNs [Wilson and Laurent, 2005].

I modeled the LN population as a single unit receiving excitatory input from all ORNs, and assumed threshold-linear GABA release as a function of the LN firing rate. I further assumed that GABA inhibition of LNs was divisive. With these assumptions, I found that superlinear recruitment of LNs by the ORNs was needed to match the Olsen model; additional experimental investigation would help to further constrain the LN model and determine whether this feature is reasonable.

Defining temporary variables I_{PN} and I_{LN} to be the amount of GABA-mediated inhibition of PNs

and LNs, dynamics of GABA receptor activation and LN and PN responses are given by:

$$\begin{aligned}
\tau_{\mathrm{GABA_{A/B}}}\frac{d\mathrm{GABA_{A/B}}}{dt} &= -\mathrm{GABA_{A/B}}(t) + [LN(t)]_+ \\
I_{PN} &= w_A\mathrm{GABA_A} + w_B\mathrm{GABA_B} \\
I_{LN} &= \mathrm{GABA_A} \\
\tau_m\frac{dPN}{dt} &= \Big[-PN(t) + R_{\max}\tanh\Big(g(ORN - ORN_{\mathrm{spont}} + c)\frac{k_1}{I_{PN}+k_2}\Big) + PN_{\mathrm{spont}}\Big]_+ \\
\tau_m\frac{dLN}{dt} &= \Big[-LN(t) + m\big(\sum_i ORN_i(t)\big)^3\frac{k_1'}{I_{LN}+k_2'}\Big]
\end{aligned}$$

$\tau_{\mathrm{GABA_A}} = 100$ ms, $\tau_{\mathrm{GABA_B}} = 400$ ms, $w_A = \frac{1}{4}$ and $w_B = \frac{3}{4}$ were fit to IPSPs recorded in [Wilson Laurent 2005]. Parameters of the PN equation are taken from Eq. 4.1, with the exception of $R_{\max}$. In the Olsen model, $R_{\max}$ was the maximum number of spikes fired during a one second stimulus presentation. In the dynamic model the PN response may transiently exceed this value at odor onset; I therefore increased $R_{\max}$ to 200Hz, which is near the peak firing rate observed in PNs. In the LN model, k_1' and k_2' did not strongly affect PN performance, and for simplicity were chosen to be similar to k_1 and k_2. m and the cubic exponential were fit by hand to the Olsen model.

The time-averaged firing rates of PNs in the dynamic model are a close fit to those from the Olsen model for all odors tested (Figure 4.2).

4.3.1.7 Spiking model

To generate spikes from the dynamic model, PN firing rates were fed into a Poisson process with a 3 ms refractory period; the refractory period was fit to match the mean/variance relationship of PN spiking reported in [Olsen Bhandawat and Wilson 2010]. Five PNs were modeled per glomerulus, each with the same underlying firing rate but a different realization of the Poisson process.

For experiments in which PN variability was unimportant, the firing rates of PNs from the dynamic model were used directly as input to model KCs.

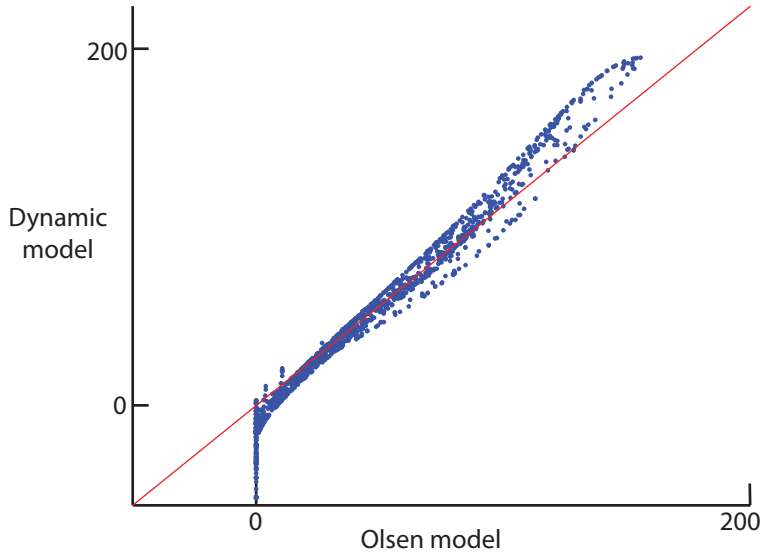


Figure 4.2: Fit of the dynamic PN model to the Olsen model. Each point is the response of one PN for one of 110 odors, for the dynamic model vs the Olsen model. To match the data to which the Olsen model was fit, the response of the dynamic model is computed as the average firing rate over a 500 ms odor presentation, minus the average spontaneous firing rate in a 500 ms window prior to odor presentation. The dynamic model is a good fit to the output of the Olsen model over all tested odors (note that the Olsen model can't produce PN responses below baseline firing rates, giving rise to the vertical excursion at x = 0.)

4.3.1.8 Projection neuron → Kenyon cell connectivity

PN-KC connectivity was derived from the experimental data of Caron et al [Caron, Ruta, Abbott and Axel 2013], who experimentally measured connection probability between PNs and KCs. Each KC has a small number of dendritic claws, which synapse with a single PN; the authors verified that PN-KC connectivity was random, meaning that one KC claw forming a synapse with a given PN did not affect connection probabilities of the remaining claws– however, some PNs were much more likely to synapse with KCs than others.

For each model KC, the number of dendritic claws was randomly drawn from a set of 200 experimentally-measured claw counts (mean 6.8 claws, with max of 11 and min of 2.) PN inputs at each claw were randomly and independently drawn from a set of 342 measured connections between glomeruli and

KCs; glomeruli in the set averaged 13.8 connections to KCs, with a max of 33. One glomerulus, da3, had no observed connections with KCs; other than that, the lowest number of observed connections was 2. After selecting a glomerulus, one of the five PNs innervating that glomerulus was randomly selected as input to the given KC claw. PN-KC synaptic weights were assumed to be binary, and normalized by the number of claws on each KC.

4.3.1.9 Kenyon cell threshold

Unlike PNs, KCs show little to no spontaneous activity: α/β and γ KCs are silent in the absence of odor, while α'/β' KCs have a spontaneous firing rate of around 0.1 Hz [Turner, Bazhenov, and Laurent 2007]. To match this in the model, I set KC thresholds relative to the time-averaged spontaneous PN input they received. I typically set all KC thresholds to a fixed amount above this value, although an alternative approach which may also work would be to set the threshold as a fixed number of standard deviations of the spontaneous PN input to each cell.

The sparsity of odor-evoked responses in model KCs is determined by both the KC spiking threshold and the strength of recurrent inhibition of KCs by APL (see below). A recent paper indicates that the number of KCs responding to an odor roughly doubles when APL is silenced [Lin et al 2014], therefore I set the KC threshold to achieve twice the desired final sparsity across a panel of odors, then tuned APL inhibition until this response was halved.

4.3.1.10 APL

APL is a single giant interneuron which innervates the entire mushroom body, and releases GABA in a graded manner proportional to the number of KCs activated by an odor [Papadopoulou et al 2011]. It extends a putitive dendrite-like process to the lobes of the mushroom body, where KC axons are found, and an axon-like process to the calyx, where PN axons synapse with KC dendritic claws. It is therefore hypothesized that APL receives excitatory input from KC axons, and recurrently inhibits KCs either presynaptically (yielding divisive inhibition) or postsynaptically (subtractive inhibition) at their claws in the calyx.

The optimum role for APL-mediated inhibition of KCs is unclear. In the Marr-Albus model of the

cerebellum, recurrent inhibition of the KC-equivalent granule cells normalizes stimulus representations, but in the fly olfactory system a stage of normalization already occurs in the transformation from ORNs to PNs, making the need for APL-mediated normalization questionable. In a previous model of the mushroom body [Luo, Axel, and Abbott 2010], APL was used to remove the first principal component of PN responses in the 110-odor ensemble; while this approach substantially improved mushroom body performance on odor learning tasks, it resulted in patterns of KC activity which did not match well with available data (see Results).

In most models discussed here, I assume that APL receives equal input from all KCs, and inhibits all KCs with equal weight; inhibition is indicated as either divisive or subtractive. In other instances I test the effects of plasticity in KC→APL or APL→KC connections, either as a means of altering the distribution of KCs responding to odors, or as a part of odor learning.

In all these conditions, I assume APL release of GABA is a linear function of its membrane potential, which evolves as

$$\tau_m \frac{dAPL}{dt} = -APL + \sum_i KC_i$$

4.3.2 Metrics used in model analysis

4.3.2.1 Lifetime sparseness of model cells and odors

In Figure 4.7, I compute lifetime sparseness of KC tuning as in [Perez-Orive et al 2002] using data received from Glenn Turner; this metric was taken from [Willmore and Tolhurst 2001, Rolls and Tovee 1995]. Lifetime sparseness of the i^{th} model KC across an ensemble of odors is given by:

$$S_p(i) = \frac{1}{1-\frac{1}{N}}\left(1 - \frac{\sum_{j=1}^{N}(r_j^2/N)}{\sum_{j=1}^{N}(r_j/N)^2}\right) \tag{4.2}$$

where r_j is the response of KC i to odor j, and N is the total number of tested odors. Because the shape of the distribution depends on the number of odors tested, I matched the data from Turner and computed sparseness over a set of 25 odors (18 odors for PNs). Unfortunately several of the odors used by Turner were not in the Hallem and Carlson dataset, so rather than match the odors directly, I selected random subsets of 25 (or 18) odors from the Hallem and Carlson dataset, and averaged over multiple odor sets to obtain the data presented in Results.

The same formulation was used to compute the sparseness of population representations of odors. For odor i, population sparseness $S_p(i)$ is again given by Eq. 4.2; in this case r_j is the response of KC j to odor i, and N is the total number of simulated KCs. Again, the shape of the distribution depends on the number of cells used, so I computed population sparseness in a subset of 109 KCs or 7 PNs to match the sample size of the experimental data. I repeated this computation across multiple sets of cells and averaged the results to obtain the data in Results.

4.3.2.2 Representation angles

The activity of a population of N neurons can be represented as an N-dimensional vector $\mathbf{r} = (r_1, r_2, \ldots, r_N)$, where r_i is the firing rate of the i^{th} neuron – either at a given time t, or averaged over some time window. To get a sense of how similar two patterns of neural activity are, I compute the angle between the two $\mathbf{r}$ vectors. For example, if $\mathbf{r}_A$ and $\mathbf{r}_B$ are the responses of two N-dimensional neural populations:

$$\theta_{AB} = \arccos\Big(\frac{\mathbf{r}_A \cdot \mathbf{r}_B}{|\mathbf{r}_A||\mathbf{r}_B|}\Big)$$

$\theta_{AB} = 0$ when $\mathbf{r}_A = \mathbf{r}_B$, and $\theta_{AB} = \pi/2$ when $\mathbf{r}_A \perp \mathbf{r}_B$ (note that for large N, θ_{AB} quickly approaches $\pi/2$ for two random vectors $\mathbf{r}_A$ and $\mathbf{r}_B$.) I typically use representation angles to compare two similar signals, such as the response of ORNs vs PNs to the same odor.

4.3.2.3 Dimensionality of odor representations

To study the suitability of the mushroom body as a basis for learned odor responses, I analyzed the dimension of KC responses across a panel of odors, computed as follows. The time-averaged responses of a set of N neurons to a set of P odors can be summarized by an $N \times P$ matrix M. While multiple metrics for dimensionality of a matrix exist, I calculated dimension in terms of the singular values $\{s_i\}$ of M:

$$D = \Big(\sum_{i=1}^{N} s_i\Big)^2 \Big(\sum_{i=1}^{N} s_i^2\Big)^{-1}$$

Dimensionality reflects the number of patterns of activity needed to account for the population response across all stimuli. It is a good predictor of network performance on binary odor classification tasks, in which a linear readout must learn random mappings of odors into categories of salient (appetitive/aversive) vs neutral. If a neural population's responses are highly correlated across odors, the number of mappings a linear readout will be able to learn will be limited, and correspondingly, dimensionality of matrix M will be low. Conversely, if population responses to odors are perfectly decorrelated, a linear readout will be able to learn a response to one subset of odors without altering its response to the other, allowing arbitrary mapping of odors into categories. With decorrelated population responses, the neural activity pattern representing odor A will be linearly independent from all other odor representations, therefore the dimension of M will be high.

4.4 Response properties of the mushroom body model

To get a better understanding of the model, I compared model performance to data from previous studies of PNs and KCs. The results help to verify that the model is a reasonable representation of the fly olfactory system, and highlight some features of the transformation of odor representation from the olfactory periphery to the mushroom body.

4.4.1 PN response dynamics are shaped by slow lateral inhibition

PN responses to odors are scaled by divisive inhibition mediated by a population of GABAergic interneurons called local neurons (LNs) that receive input from ORNs and are broadly tuned across odors. PNs express both $GABA_A$ and $GABA_B$ receptors, which have intrinsic timescales of 100 ms for $GABA_A$ vs 400 ms for $GABA_B$; LNs also recurrently inhibit themselves via $GABA_A$ receptors [Wilson Laurent 2005]. The slow timescale of $GABA_B$ inhibition has been cited to explain the temporal diversity of PN responses: PNs can be inhibited by odors for periods spanning from 100 ms to several seconds, a property not seen in their ORN inputs. PN responses to odors can also be multiphasic, showing a period of excitation followed by inhibition, or vice versa. Because KCs in the mushroom body receive input from only a few PNs (7 on average), the dynamics of individual

PN responses likely have a strong effect on odor representation in their downstream KCs.

I implemented LN-mediated lateral inhibition in my dynamic model of PNs, and adjusted inhibitory weights to match the divisive inhibition model of Olsen et al (see Methods). Because the timescale of LN-mediated inhibition is slow, it lags the initial excitatory input from ORNs to PNs, giving rise to multiphasic responses of PNs to odor inputs. Figure 4.3 shows responses of the model PN population to 500 ms presentation of several example odors, as well as example traces from seven recorded PNs for comparison. Typical model PNs have a transient peak in firing rate shortly after odor onset, which decays as lateral inhibition grows. After odor offset, there is a transient dip in PN activity, as inhibition decays back to baseline. Lateral inhibition is stronger for public odors such as isopentyl acetate that activate many glomeruli, whereas private odors that only activate a single glomerulus strongly drive less inhibition, and evoke more sustained PN responses.

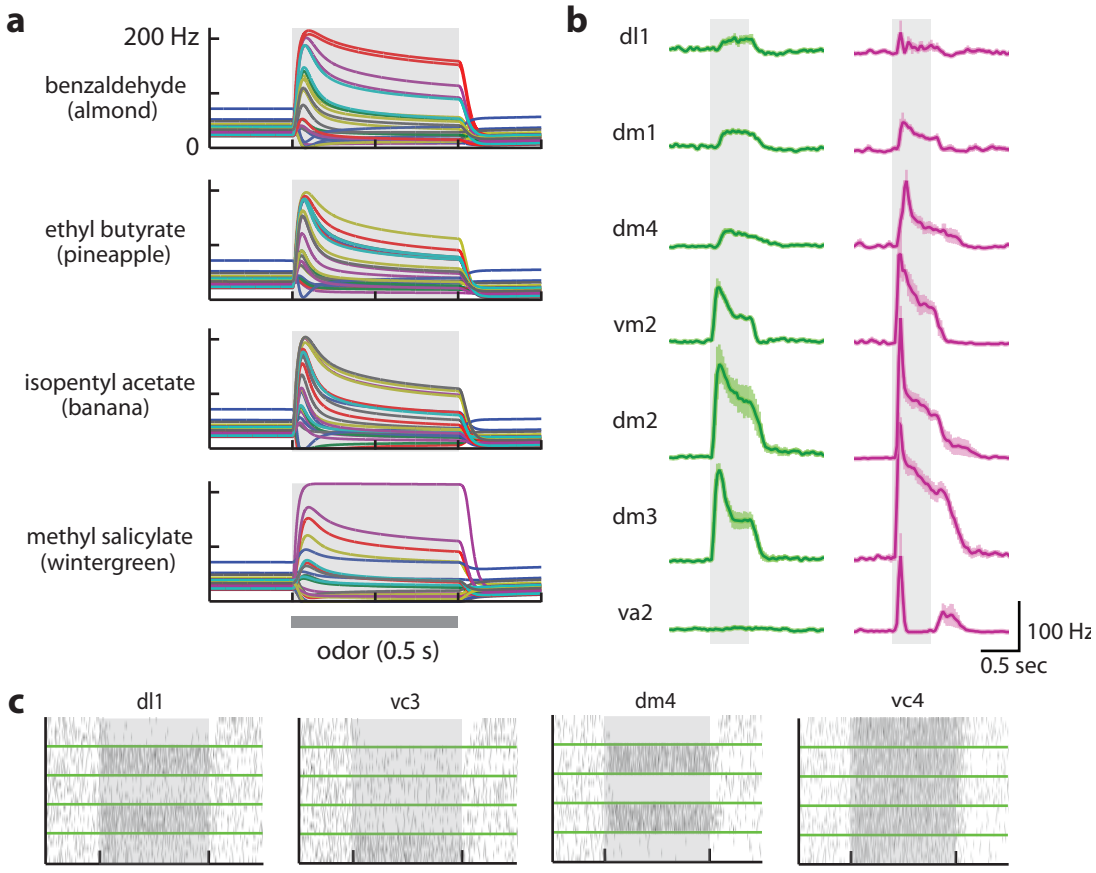


Figure 4.3: **Odor representations by model PNs evolve over time. a.** Firing rates of the 23 PNs in the dynamic model, in response to four sample odors (each colored line is a different PN.) Gray regions mark the time of odor presentation, which has duration of 500 ms unless otherwise noted. Strong public odors like isopentyl acetate, which activates many glomeruli, show a pronounced onset transient followed by a drop in firing rate due to lateral inhibition. Private odors like methyl salicylate, which only activates one glomerulus strongly, drive less lateral inhibition, allowing sustained responses in a small number of glomeruli. **b.** Extracellularly-recorded firing rates of seven example ORNs (green) and their cognate PNs (pink) in response to a 500 ms presentation of isopentyl acetate, reproduced from Bhandawat et al [Bhandawat et al 2007]. Note odor-evoked inhibition in PNs innervating glomeruli dl1 and va2, as well as transient onset responses in dm1-dm4. **c.** Example spike rasters generated from model PNs innervating glomeruli dl1, vc3, dm4 and vc4, responding to a panel of five odors. Each row shows the single-trial spiking response of a single PN, with the odor presentation window marked in gray and responses to different odors delineated by green lines.

The model PNs show similar responses to recorded PNs, though there is clearly additional temporal structure in the recorded PN responses that slow lateral inhibition alone cannot account for. Some of this structure is inherited from ORNs: in the model, ORN input to PNs was a simple step input, whereas recorded ORNs clearly show a transient spike in firing rate at odor onset, and a drop in firing rate as time increases, possibly due to adaptation at the receptor. Some ORNs also had higher affinity for odorants than others, showing sustained firing after odor offset. In addition, the nonlinear firing rate transformation from ORNs to PNs is known to vary across glomeruli, as can be seen in Figure 4.3b: PNs innervating glomeruli vm2 and dm3 have different response magnitudes despite receiving nearly the same excitatory input. While there is not enough data available to model the full range of effects that shape PN dynamics, incorporating LNs into the model is at least sufficient to match the general shape of PN responses in a way which is also consistent with previous models of divisive normalization of PNs.

4.4.2 Lateral inhibition determines PN representation of odors in mixtures

I also verified that slow LN-mediated inhibition in my model could reproduce previous experimental findings on the role of lateral inhibition in shaping PN responses to odor mixtures. Recordings from PNs responding to mixtures of public and private odors show that mixing a public odor with a private odor makes the response to the private odor become more transient, as shown in Figure 4.4 [Olsen Bhandawat and Wilson 2010]. Applying GABA receptor antagonists to the antenna lobe blocks the increase in transience, suggesting that it is mediated by lateral inhibition in the antenna lobe rather than odorant competition at the level of the ORNs.

To recreate this study in the dynamic PN model, I presented the model with a private odor (2,3-butanedione) at low or high concentrations, either by itself or mixed with low or high concentrations of a public odor (isopentyl acetate), and looked at the response to the private odor (measured in the glomerulus activated most strongly by that odor.) Because odor concentrations were low, I assumed linear mixing of odor representations at the level of the ORNs. As in the experiments, mixing the public odor with the private caused the response to the private odor to become more transient, due to increased recruitment of lateral inhibition by the public odor. Increasing the concentration of the private odor also caused the PN response to become more transient, though

the overall increase in sustained firing for mixtures of strong private with weak/strong public odors was marginal in the model compared to the data.

The effect of this interaction is to boost representation of public odors in the PNs, and subsequently in the mushroom body and lateral horn. Public odors, which strongly activate many glomeruli in the fly, are often innately appetitive fruit odors associated with food sources. Reducing responses to private odors in mixtures could therefore be an adaptation to enhance the representation of behaviorally relevant stimuli.

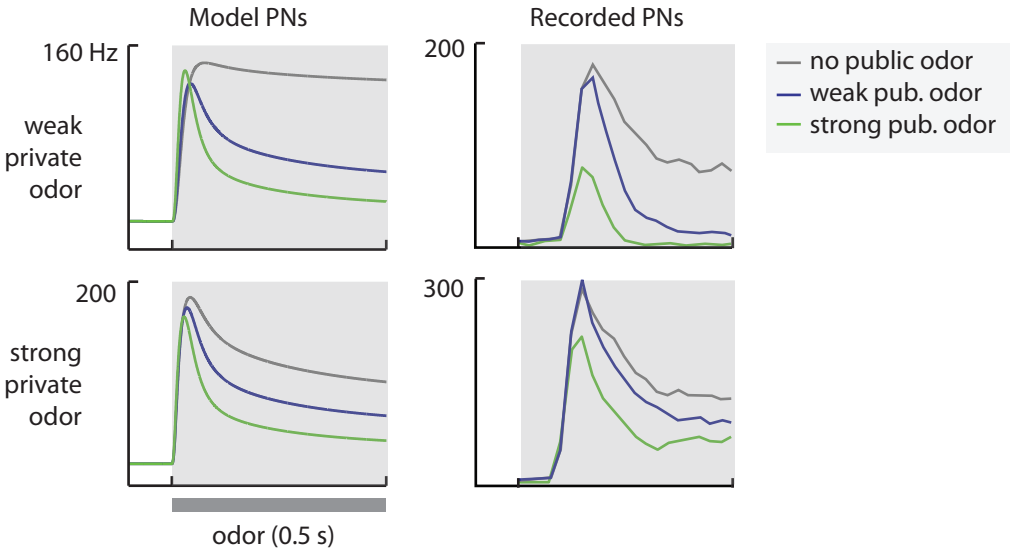


Figure 4.4: **Temporal effects of odor mixtures. Right:** extracellularly-recorded firing rate response of a PN to 2-butanone, a private odor that only activates one glomerulus strongly; 2-butanone was presented either alone or mixed with the public odor isopentyl acetate, and each odor diluted to either 10^{-5} (weak) or $10^{-}3$ (strong). The PN shows sustained firing in response to the private odor alone, but mixing with a public odor attenuates the sustained portion of the response. Increasing the concentration of the private odor reduces the strength of this effect. **Left:** recreation of the response transience effect by model PNs. I modeled mixtures of the private odor 2,3-butanedione with public odor isopentyl acetate, with the two odors at dilutions of either 10^{-4} (weak) or 10^{-2} (strong), using concentration-dependent ORN *(continued on following page)*

Figure 4.4: *(continued)* responses to the two odors taken from the Hallem and Carlson dataset. As in the experiment, mixing the private odor with a public odor made the response to the private odor more transient, while increasing the concentration of the private odor somewhat reduced the effect of mixing.

4.4.3 The PN representation of odors evolves over time

From a computational perspective, the scaling of PN responses by lateral inhibition is thought to normalize PN population activity so that the total PN input to the mushroom body is roughly equalized across odors, removing effects of odor intensity while preserving representation of odor identity [Olsen Bhandawat and Wilson 2010]. However as noted above, the timescale of lateral inhibition is slow, and odor presentation drives a transient spike in PN firing rates before lateral inhibition has had time to kick in. This peak in firing rate can reach up to 300 Hz in experimental data, and is sufficient to evoke spiking responses in downstream KCs. Divisively normalized odor representations among PNs may therefore not be representative of the actual stimulus driving KC activity, at least for the 500 ms pulses of odors studied experimentally.

To get a better understanding of how much PN normalization shapes the input to KCs, I compared output of the dynamic PN model to that of two alternative models: one in which lateral inhibition between PNs was instantaneous, and one with no lateral inhibition. At each timestep, I computed the angle between the odor representation in the dynamic PNs with that of the two models, as outlined in the Methods. As shown in Figure 4.5, the model with no inhibition is more strongly correlated with the output of the dynamic model during the first several hundred milliseconds after odor onset, as indicated by the smaller angle between the two model PN populations. After a few hundred milliseconds, the effects of lateral inhibition begin to affect the PN response, and the instantaneous inhibition model becomes the better fit to the dynamic model.

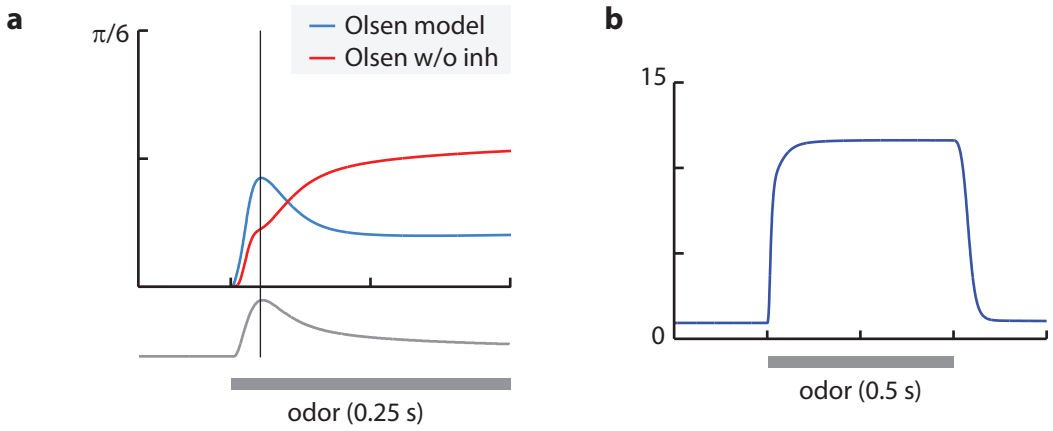


Figure 4.5: **a.** Representation angle between the dynamic PN population and two alternative models: one in which lateral inhibition is instantaneous (Olsen model), and one with no lateral inhibition (Olsen w/o inh); plotted below the time axis is the average PN response across cells and odors. Two PN populations with responses different only by a scale factor will have an angle of zero between them, whereas the angle approaches $\pi/2$ for orthogonal PN representations. In the first several hundred milliseconds, the dynamic model is more similar to the model without any lateral inhibition. Further into the stimulus period, inhibition begins to be recruited in the dynamic model, and the representation changes to be more like the Olsen divisive normalization model. **b.** I also measured the dimensionality of the dynamic PN response over the course of a stimulus presentation, using the metric described in the Methods. While PNs in the dynamic model show complex temporal structure in their responses, the dimensionality of their representation is stable over the course of stimulus presentation, and reaches its maximum value before lateral inhibition has fully kicked in.

While normalization changes the representation of odors by PNs, it does not alter the dimensionality of input to the mushroom body. I computed dimension of the PN response over a set of 1000 real and synthetic odors, and found that it was unaffected by the dynamics of PN responses during odor presentation. Because of this effect, as well as the slow timescale of PN inhibition, it seems possible that normalization among the PNs might not be required for associative learning of odors in the mushroom body, and could instead provide some more general role in sensory adaptation.

One unexplained feature of the fly olfactory system is that it seems to have two sequential stages

in which odor representations are normalization via recurrent inhibition: LNs pool activity across glomeruli to scale PN responses in the antenna lobe, and the giant inhibitory neuron APL pools activity across KCs to scale KC activation in the mushroom body. While the dynamics of APL-mediated inhibition have not been studied, it could be that the two stages of normalization act on two different timescales: APL could facilitate learning by provide rapid normalization of odor representations in the mushroom body, while LNs help the olfactory periphery adapt its responses to odor fluctuations on longer timescales.

4.4.4 Odor representations are sparse but not uniform in model Kenyon cells

I next characterized the odor-evoked activity of Kenyon cells in my mushroom body model. In experimental data, 5-10% of KCs respond reliably to a given odor, and individual KCs are narrowly tuned, responding to 5-10% of odors on average. To tune the responses of KCs in my model, I therefore adjusted KC thresholds and APL-mediated recurrent inhibition as outlined in the Methods, such that an average of either 5% or 10% of KCs responded to odors in the Hallem and Carlson dataset. For simplicity, I will neglect PN spiking variability in this section, and focus on results from the rate model.

Population responses to each odor are shown on the left in Figure 4.6, while individual KC tuning widths (the fraction of odors to which each cell responded) are shown on the right. With the model tuned for 10% mean sparsity, only one odor (glycerol) consistently failed to evoke a response in the model KC population, which may be attributed to glycerol's very weak activation of ORNs. When sparsity was decreased to 5% (fewer KCs responding), three odors (glycerol, g-decalactone, and a-humulene) and occasional others failed to evoke a response.

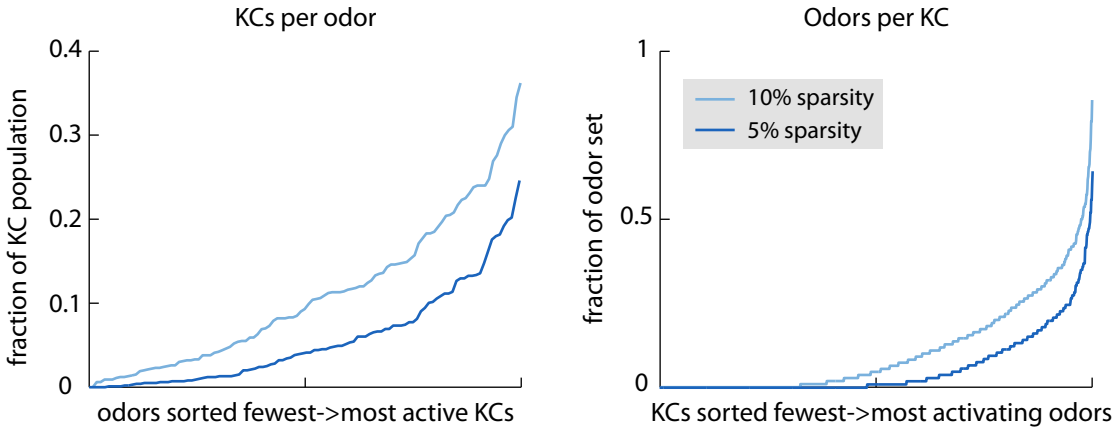


Figure 4.6: Heavy tails in KC representation of odors, and in odor tuning of KCs. **Left.** Fraction of the KC population responding to odors in the Hallem and Carlson dataset, sorted from smallest population to largest population. The model was tuned so that an average of either 5% or 10% of KCs responded to each odor. **Right.** Fraction of the tested odor ensemble to which each KC responded, sorted from least to most responsive KC. A substantial portion of KCs were silent for all odors: 32% of cells in the model with 10% sparsity, and 48% of cells in the model with 5% sparsity.

The tuning width of model KCs was far from uniform, with many KCs responding to none of the presented odors, and a few generalist KCs that responded to over half of the 110 tested odors. In the model with 10% sparsity, 32% of KCs failed to respond to any of the presented odors, while 1.7% of KCs responded to over half of the odors presented. In the model with 5% sparsity, 48% of KCs failed to respond, while 4% responded to over half of the odors presented. The nonuniformity of KC responses has important implications for the capacity of the mushroom body to form odor-specific associative memories, which is discussed below.

It is important to note that the nonuniformity of KC responses depends strongly on the model setup used, in which all KCs received the same recurrent inhibition from APL, and had the same threshold relative to their baseline input. I am able to reduce the number of silent and over-responsive KCs by either tuning KC thresholds or adjusting the weight of APL inhibition onto individual KCs, producing models in which nearly every KC responds to the target value of 5% or 10% of simulated

odors. Both adjustments could be viewed as homeostatic effects in KCs: cells that never spike, or spike too often, scale their synaptic weights to adjust their firing rate.

But comparison with experimental data suggests that real KCs exhibit nonuniform tuning similar to that of the model KCs in Figure 4.6. In Figure 4.7, I compare sparsity in model KCs to imaging data from Glenn Turner, using the lifetime sparseness metric from [Perez-Orive et al 2002], after [Willmore and Tolhurst 2001, Rolls and Tovee 1995]. KC results in this figure come from the model tuned to 10% sparsity. To confirm my model of the olfactory periphery was reasonable, I also computed the lifetime sparseness of the 23 simulated PNs, comparing my result to the measured sparseness in a set of PNs from 7 glomeruli. In a second experiment on a smaller panel of six odors, a substantial proportion of imaged KCs did not respond to any of the odors presented– on the order of 50-75% (Turner, private communication). These findings suggest that fine-tuning of KC thresholds and APL recurrence is not required to create a sparse representation that matches that of the mushroom body, and that KC responses are predominantly determined by the PN inputs they receive.

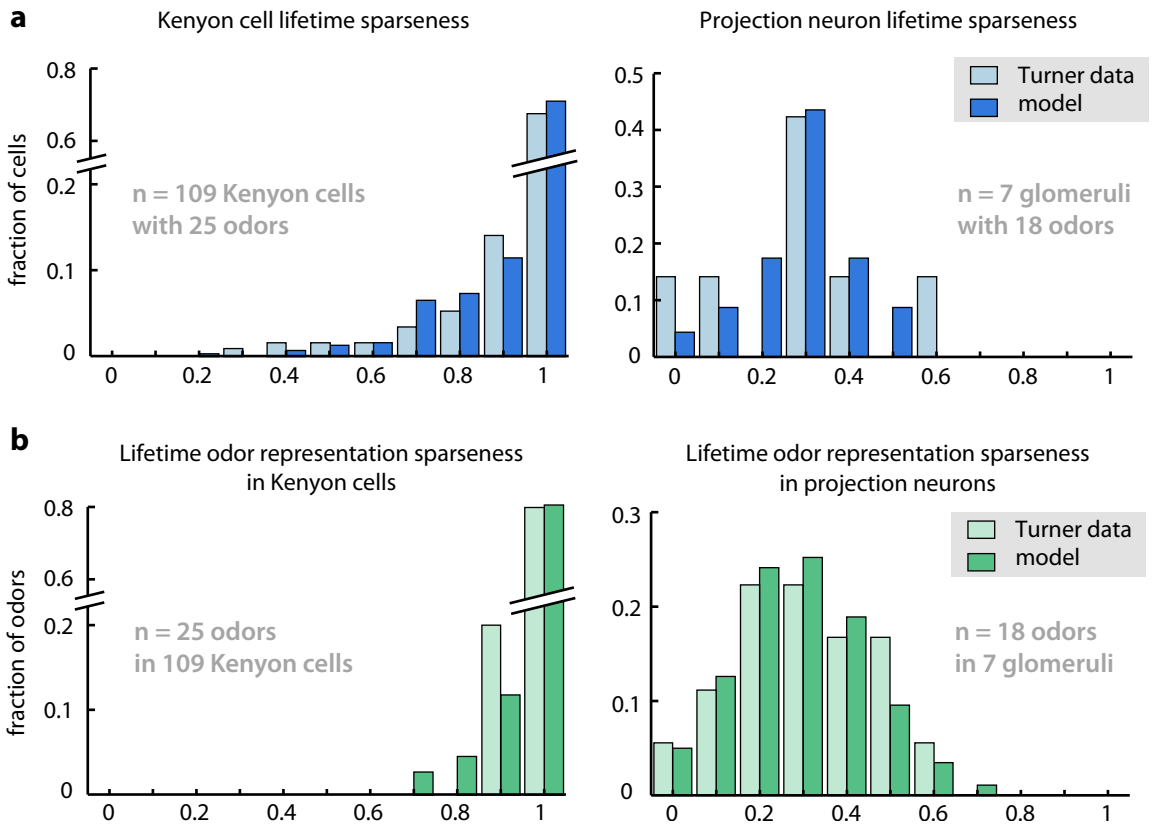


Figure 4.7: **a.** Lifetime sparseness of the model KC and PNs, ie the proportion of odors which evoke a response in a given cell. Because response sparseness depends on the set of odors used for testing, I chose to match the experimental data and compute sparseness in sets of 25 odors for KCs and 18 odors for PNs; I then averaged across subsets to construct the histogram. Plotted for comparison are experimentally measured lifetime sparseness of a set of 109 KCs and 7 glomeruli [Turner]. The model is tuned such that an average of 10% of KCs respond to a given odor. **b.** Lifetime sparseness of odor representations by the model KC and PN populations, ie the proportion of cells which respond to an odor. As above, sparseness depends on the size of the cell population, thus to match the measured values to the Turner data I computed sparseness on random subsets of 109 KCs or 7 PNs, and averaged across sampled subsets to construct the histogram. For comparison are experimentally measured population sparseness of odor representations on a set of 25 odors for KCs and 18 odors for PNs.

4.4.5 Kenyon cell responses scale with odor concentration

Calcium imaging in the mushroom body has shown that the number of KCs responding to an odor increases with odor concentration [Peter Wang, unpublished observations]. ORN population activity increases with odor concentration, but the representation of the odor also changes: in addition to having firing rates that are nonlinear with odor concentration, individual ORNs have different binding affinities for odors, so odors that activate a single ORN at low concentrations could activate several at higher concentrations. Due to the two stages of normalization in the olfactory system model (local interneuron-mediated lateral inhibition in PNs and APL-mediated recurrent inhibition in KCs), it is not clear that increasing ORN recruitment by odors will lead to an increase in the number of active KCs. Indeed, in the model by Luo et al. [Luo, Axel, and Abbott, 2010] the number of active KCs dropped with increasing odor concentration, because recruitment of APL and lateral inhibition between glomeruli together increased more sharply with odor concentration than did the activation of individual ORNs.

I studied concentration-dependent responses of KCs to ten monomolecular odors at concentrations of 10^{-8}, 10^{-6}, 10^{-4}, and 10^{-2}, as well as nine fruit odors at concentrations of 10^{-6}, 10^{-4}, 10^{-2}, and undiluted, using concentration-dependence data from the Hallem and Carlson dataset. While the number of active KCs in the model does increase with odor concentration, as observed experimentally, the firing rate of active KCs drops on average (Figure 4.8c). This can be attributed to two effects. First, at higher odor concentration the transient response of PNs becomes larger relative to the sustained response. I therefore expect there to be an increase in the number of KCs that are only transiently active, which means that the average firing rate of active KCs will drop. But in addition, increased recruitment of KCs at odor onset drives stronger APL-mediated inhibition of KCs, causing KCs that showed sustained firing at low odor concentrations to have transient responses at high odor concentrations. This effect is evidenced by the drop in KC input to APL during the second half of the stimulus period in Figure 4.8e, despite PN input to KCs remaining high during this interval (seen in Figure 4.8d). To confirm that the effect was due to APL inhibition, I blocked APL output in the model and found that the drop in KC input to APL was eliminated.

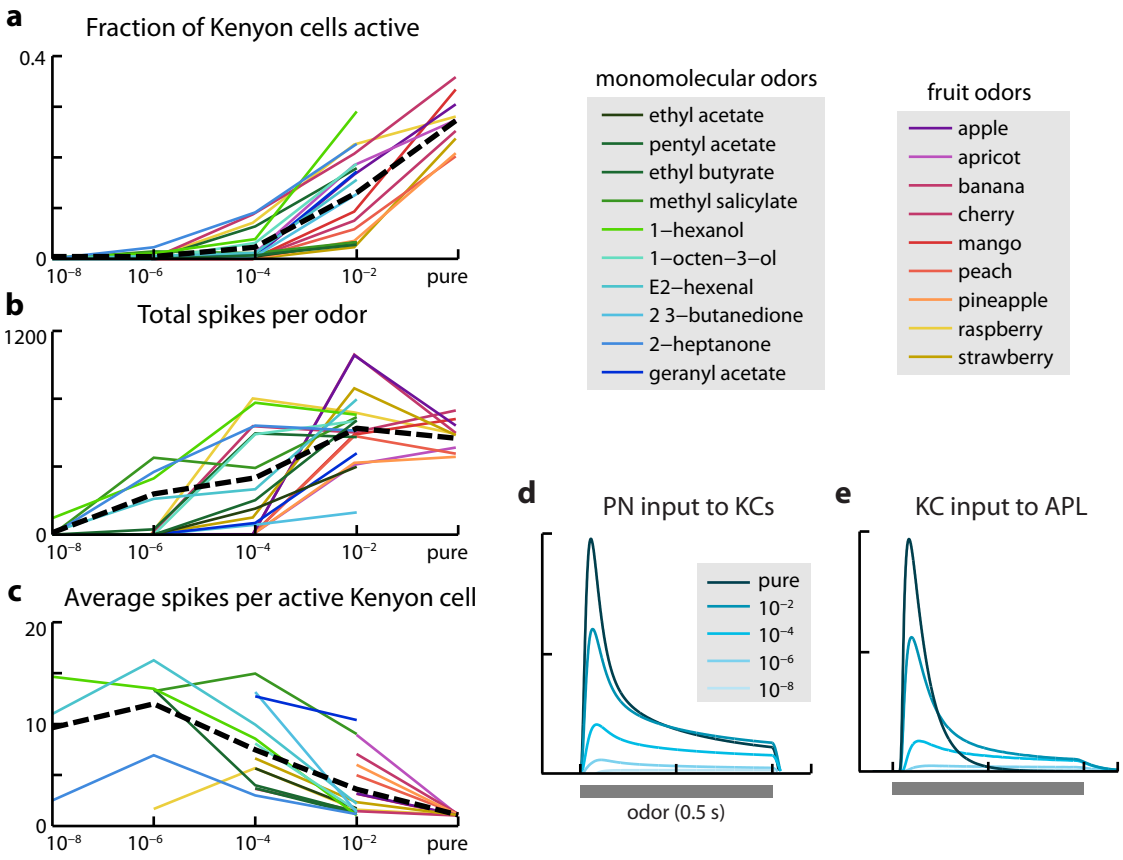


Figure 4.8: Concentration-dependent activity of the model KC population to ten monomolecular odors and nine fruit odors. Dashed black line is average across tested odors. Only monomolecular odors were tested at concentration 10^{-8}, and only fruit odors were tested undiluted; at other concentrations all odors were used. **a.** The number of active KCs increased with concentration for all odors, as is observed experimentally. Some odors failed to evoke any response in the model at low concentrations (10^{-8} or 10^{-6}), but this is not out of line with observed odor sensitivity in flies. **b.** Population response to odors can also be measured as the total number of evoked spikes in the population of 2000 KCs. This number also increased with odor concentration, though not as sharply as the count of active KCs. **c.** The average number of odor-evoked spikes in active KCs dropped at higher odor concentrations. (Concentrations which did not activate any KCs are not plotted.) **d.** Total excitatory input from PNs to the KC population, averaged across the 19 tested odors. **e.** Average KC input to APL across the 19 odors, used as a proxy for KC population spiking. (Legend is same as in panel d.)

The observed drop in KC firing rates at high odor concentrations could be a useful feature for preserving learned behaviors across odor concentrations. Without this effect, a higher-order neuron tuned to respond to odor A would respond non-selectively to other odors at high concentrations, simply because those odors would elicit a stronger total response. Decreasing the firing rates of KCs with concentration compensates for the increase in the number of active KCs, to make the total population firing rate of the KCs more concentration invariant (Figure 4.8b). At the same time, KCs in the model which were active at low concentrations of an odor almost always remained active at higher concentrations: only 5.5% of KCs that fired more than one spike in response to an odor at 10^{-4} did not respond to that odor at 10^{-2}, and these cells typically turned off due to odor-driven inhibition of their ORN inputs at the higher concentration, rather than the effects of APL inhibition. The dynamic model therefore equalizes odor representations across concentrations, while keeping the KC representation of that odor largely intact, two features which should contribute to a stable readout of odor identity across concentrations.

The drop in KC firing rates could be checked easily experimentally either by recording odor-evoked responses of individual KCs or by measuring the total KC input to APL over multiple concentrations. The experiment simulated above, blocking inhibition of KCs and measuring the input to APL, would also be informative to recreate as it would help to disentangle the effects of PN normalization (via lateral inhibition) and KC normalization (via APL) on odor representations in the mushroom body. APL-mediated normalization of KC responses is one of the least constrained aspects of the model, due to the lack of detailed knowledge of its connectivity and response to odors. Thus failure to reproduce the concentration-dependence of KC firing rates experimentally could suggest useful modifications of the APL model, and lead to a better understanding of the role of APL in shaping KC representations of odors.

4.4.6 Kenyon cells preserve distance of representations in odor mixtures

The sparse, high-dimensional representation of odors by Kenyon cells is useful for learning odor-specific behavior: neglecting effects of noise, the dimensionality expansion from PNs to KCs ensures that the representation of a learned odor will have very limited overlap with representations of other odors However in the case of odor mixtures, it becomes useful for similarity between stimuli to be

preserved: for example, a learned association with a pure odor should be recalled if that odor is presented in a mixture. In a previous imaging study, Campbell, Honegger et al. examined both correlation between KC representations of odors in different mixture ratios, and the discriminability of pairs of mixtures in flies [Campbell, Honegger, et al 2013]. Comparing a 0:100 to a 100:0 mixture, 30:70 mixture to a 70:30 mixture, and a 40:60 mixture to a 60:40 mixture, the authors found that the correlation between KC responses increased as mixtures became more similar. Mixtures of 40:60 and 60:40 could be discriminated just barely above chance by flies, and at chance levels by a linear classifier trained on imaged KC responses. These results suggest some preservation of distance in odor representations from ORNs to KCs.

First, to check that the model was performing reasonably, I compared representation of pairs of odors to the representation of their 50:50 mixture. Previous imaging studies in the mushroom body report that most KCs respond sublinearly to odor mixtures, meaning the response of a KC to the mixture of odors A and B is less than the sum of its responses to odors A and B alone. For simplicity I assumed that ORN responses were weighted sums of the responses to the two odors in the mixture. While there is not enough data to make a quantitative comparison, model responses to mixtures were predominantly sublinear, in line with the data. KCs that responded to both of the mixed odors individually were almost always active in response to the mixture, suggesting that the representation of the odor mixture maintained some similarity to the representation of the two mixed odors.

(I'll have a couple plots of Peter's data here for comparison, plus a few more mixtures below, but you get the idea)

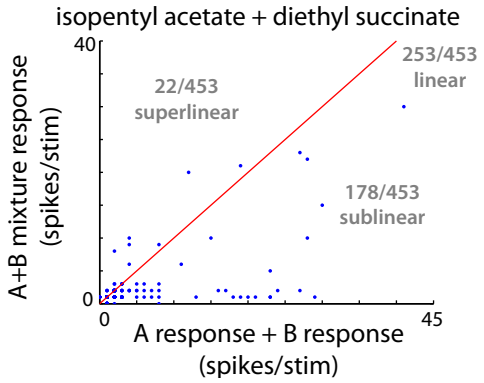


Figure 4.9: Plot of KC response to 50:50 odor mixtures vs response to each odor on its own. Each point is a KC, red line is y=x. Model KCs are predominantly sublinear or linear, and all but two KCs that fired more than one spike in response to the two individual odors were also active for the mixture of both odors.

To more thoroughly examine the effect of odor mixing on KC representations, I simulated responses to 500 random pairs of odors in mixing ratios of $A:B = \{100{:}0,\ 90{:}10,\ 80{:}20,\ 70{:}30,\ 60{:}40,\ 50{:}50,\ 40{:}60,\ 30{:}70,\ 20{:}80,\ 10{:}90,$ and $0{:}100\}$. For each mixture I computed the $N \times 1$ vector R of the mean-subtracted firing rates of N KCs ($N = 2000$ in the model), then calculated the correlation $C_{i,j} = \frac{R_i \cdot R_j}{|R_i||R_j|}$ between each pair of mixtures (i, j). To get a sense of the dependence of representation distance on mixture similarity, I then computed the average of each diagonal band of C:

$$D(x) = \left(\frac{1}{M - |x|}\right) \sum_i C_{i,i+x}$$

where M is the number of mixtures simulated. Hence $D(0)$ is the average of the diagonal of C (so $D(0) = 1$), while $D(\pm 10)$ is the average correlation of the 100:0 mix with the 0:100 mix, ie the average correlation between pure odors in the dataset.

As seen in Figure 4.10a-b, odors evoking strongly correlated activity in ORNs, such as two similar mixing ratios of an odor pair, also evoke correlated patterns of activity in KCs. Correlation in KCs falls off much faster than in ORNs or PNs, which gives the correlation matrices in the bottom row of Figure 4.10a a somewhat block-diagonal structure. Rather than correlate with public or private

odor mixings, the shape of the blocks in individual correlation matrices seem to be most closely tied to the binding affinity of the two tested odors for ORNs: for example, ethyl 3-hydroxybutyrate and ethylcinnamate both have very weak affinity for all ORNs relative to the two odors they are mixed with. As a result, the model predicts that odors mixed with either of these two should be identifiable in mixtures even at low mixing ratios.

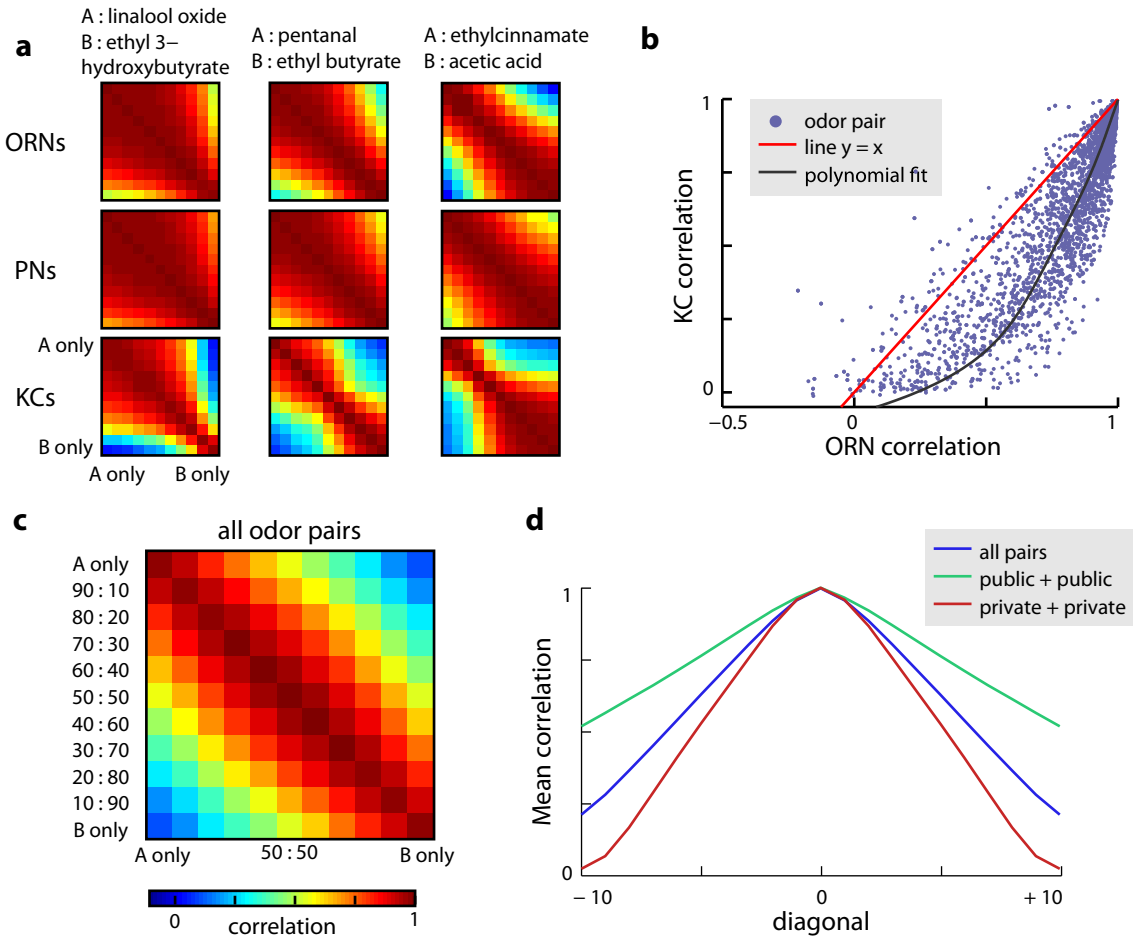


Figure 4.10: Correlation of odor mixture representations, computed as described in the text. **a.** Correlation matrices of three example odor mixtures, computed for ORNs, PNs, and KCs; refer to panel c for full axis labels and color bar. Note that in KCs the transition from mixtures resembling A to mixtures resembling B can be quite sharp, and occurs at different ratios for different odors. The point of transition seems to be determined by the difference in total ORN activation between the two odors. **b.** Plot of mixture correlation in ORNs vs KCs for all 11 mixtures of all 500 odor pairs (points downsampled for display purposes.) While KC representations are less correlated than those of ORNs, pairs of mixtures which are highly correlated in ORNs typically remain so in KCs. **c.** Average KC correlation matrix over 500 odor pairs, testing each pair with 11 mixing ratios. **d.** Plot of $D(x)$ (see text), the average of terms on the x^{th} diagonal of the covariance matrix. Public odors are more strongly correlated with each other than the average odor pair, while private odors are less strongly correlated.

4.4.7 Dimensionality of odor representations is maximized by the mushroom body

Finally, I examined the relationship between properties of PN-KC connectivity and the dimensionality of odor representations, as described in the Methods. The dimensionality is a measure of the diversity of the KC population's responses across odors. If KC responses to odors are strongly correlated, the number of mappings from odor to behavior that a linear readout of KC activity will be able to form is small, and the dimensionality will be low. This results in poor performance on odor learning tasks, for which it should be possible to set the response of a linear readout from KCs to arbitrary values for any given set of odors. If KC responses to different odors are decorrelated, the dimensionality will be high, and the response of a linear readout to any given odor can be changed arbitrarily without impacting its response to other odors.

I computed dimensionality of KC responses over a panel of 5000 synthesized odors, using the metric outlined in the Methods. I examined how three features of PN-KC connectivity contribute to KC dimensionality: the number of PN inputs to a KC, the amount of structure in PN-KC connections, and the sparseness of KC responses. To add structure to PN-KC connections, I arbitrarily defined five groups of glomeruli in the model, and allowed each KC to receive input from only one group. To vary the degree of structure, I made KC group conformity probabilistic, such that a given PN-KC connection conformed to the assigned group with probability p, or receive random input from any glomerulus with probability $1 - p$, giving connectivity ranging from completely random ($p = 0$) to completely structured ($p = 1$). I found that dimensionality was highest in the completely random network, and dropped as the degree of structure increased.

I then varied the number of PN inputs to each KC, and varied the sparseness of KC representations by adjusting the spiking threshold of all KCs uniformly, keeping APL inhibition tuned to roughly halve the number of active KCs, as described in the Methods. The maximal dimensionality was observed when KC responses were sparse (10-20%), and with roughly 5-10 inputs per KC. However, I found that dimensionality continued to slowly rise when the number of inputs was increased to 25 or more. This second, slower increase in dimensionality could be because at high numbers of claws, all KCs receive roughly the same input, and the representation of odors depends on which subset of KCs crosses threshold first, activating APL and inhibiting the remainder of the population. Such a

system could decorrelate KC representations across odors, but it should also be more vulnerable to noise in the PN inputs to KCs. Measuring the dimensionality of odor representations using spiking PNs should clarify this question.

Overall, this model shows that the values which maximize dimensionality fall within the ranges found experimentally in the fly olfactory circuit: PN-KC connectivity is unstructured, odors activate roughly 10% of KCs, and each KC receives an average of 6.8 PN inputs. These findings suggest that the KC representation of odors is structured to allow the fly to learn arbitrary and specific associations of odors with behaviors.

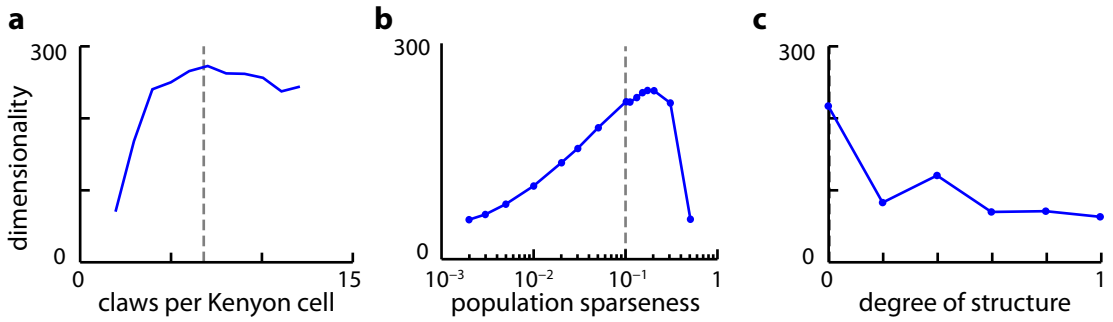


Figure 4.11: Dimensionality of the KC representation of odors is related to several parameters of the model; each point is an average over two instantiations of the model. Experimentally determined values of each parameter are indicated by the gray line. **a.** Dimensionality as a function of the number of PN-KC connections peaks at around 5-10 connections per KC, then drops off before gradually increasing again. **b.** Dimensionality as a function of KC population sparseness, varied by adjusting KC spiking thresholds and keeping APL inhibition tuned as described in the methods. **c.** Dimensionality obtained when a variable degree of structure is imposed on PN-KC connectivity. Structure was set by a parameter p, ranging from fully unstructured ($p = 0$) to fully structured ($p = 1$), by restricting PN-KC connections to assigned groups of glomeruli with probability p.

4.5 Discussion

The mushroom body lies at the interface of sensory processing and behavioral control in *Drosophila*. Its inputs transform odor representations from a dense code of ORN firing rates to a sparse,

high-dimensional pattern of Kenyon cell activity. In the following chapter, we will see how this architecture allows a representation of odor identity to be translated by the mushroom body to a representation of the odor's behavioral salience, which in subsequent layers of processing determines the fly's response to the odor. Unlike the lateral horn pathway, the mushroom body is special because Kenyon cell representations impose no particular meaning on odors; instead, meaning is determined by how those representations are read out. While the neural circuitry that reads out odor representations from the mushroom body is beginning to be identified, the mechanism by which these structures modify their responses during learning is still unknown.

Learning behavioral responses to odors is simple in the abstract: by projecting a stimulus up to a high-dimensional space, a simple linear readout can be trained to map the stimulus to a behavioral response. But correlations in odor representations degrade the capacity of the readout to form arbitrary mappings, constraining the capacity of the system to form multiple associations. In order to get a sense of how big a problem these correlations are for the fly olfactory system, I had to first build a model which reasonably approximated KC responses. For each piece of the model, I used a collection of experimental observations to constrain model parameters, and to verify that the odor-driven activity of model cells was in line with biology. While the model could of course be refined with additional levels of biological detail, the goal of the model is not to capture every detail of Kenyon cell responses, but to provide a reasonable estimate of the problems the fly's learning circuitry would need to solve in order to keep behavioral responses odor-specific and diverse. In the next chapter, I will review the challenges to odor learning introduced in the current mushroom body model, and devise possible solutions by which they can be overcome. These solutions make testable predictions about the activity of the mushroom body readout circuitry, and can be used to clarify the mechanism of odor learning from Kenyon cell representations.

Chapter 5

A circuit mechanism for associative learning in the mushroom body

While not needed for innate behavioral responses to odors, the mushroom body is required for the formation and retrieval of odor memories in associative learning [de Bell and Heisenberg, 1994; Dubnau, Grady, Kitamoto and Tully 2001; Keene, Leung, Armstrong and Waddell 2007]. Associative learning allows organisms to adapt their response to sensory cues based on prior experience, by combining a representation of the sensory environment with an internal code of valence based on the animal's experiences. During learning, consistent association of a sensory stimulus with a salient experience forms a stable associative memory. Subsequent presentation of the trained stimulus retrieves the stored memory, allowing the animal to predict the valence of the stimulus and produce an appropriate behavioral response.

In the previous chapter, I described how odor representations are transformed from low-dimensional, anatomically preserved codes in the olfactory periphery to sparse, high-dimensional, and random patterns of Kenyon cell activation in the mushroom body. In this chapter, I will review the results of a recent experimental effort that has fully characterized the set of cells peripheral to the mushroom body, and identified a population of only 34 neurons that form the mushroom body's only output to the remainder of the brain. Some of the identified cells have previously been implicated in retrieval of associative memories [Séjourné et al 2011] and the circuit architecture of the identified cells with

the mushroom body bears strong resemblance to the canonical cerebellar circuit for associative learning [Heisenberg, 2003; Farris 2011; Ohyama, Nores, Murphy and Mauk, 2003; Christian and Thompson 2003].

To examine this theory, I construct a basic model of associative learning using the observed circuitry of the mushroom body, and study its capacity for learning of odor-specific responses generated by my mushroom body model. While odor representations in my mushroom body model are less suited to associative learning than an idealized random representation, I find that a simple modification of the associative learning rule allows large numbers of associative memories to be stored. Using two variations on this modification, I make predictions of the responses of neurons peripheral to the mushroom body over the course of learning, and show some evidence suggesting that a similar mechanism is at work in the mushroom body.

5.1 Learning circuitry of the mushroom body

A recent investigation fueled by the powerful genetic tools available in fly has uncovered the complete population of neurons innervating the mushroom body. The KC representation of odors is read out by a population of 34 output neurons, and learning in these output neurons is modulated by a set of roughly 130 dopaminergic neurons. The connectivity of these neurons readily suggests a basic circuit for acquisition of conditioned responses in the output neurons, in which dopaminergic neurons play a key role in driving memory acquisition.

5.1.1 Identification of mushroom body output neurons using split-GAL4 lines and photoactivatable GFP

As described in the previous chapter, the roughly 2000 Kenyon cells of the mushroom body extend their axons to form two structures called the medial and vertical lobes. KC axons either bifurcate to both lobes (α/β and α'/β' KCs) or go only to the medial lobe (γ KCs); axons in each lobe are segregated by KC type, thus defining a total of five distinct lobes (α, α', β, β', and γ) [Tanaka, Tanimoto and Ito 2008].

KCs themselves do not project outside of the mushroom body, and there has been substantial interest in determining how the KC representation of odors is read out by the fly. In a heroic recent study by Aso et al, thousands of *Drosophila* split-GAL4 lines were screened for neurons innervating the lobes of the mushroom body. The GAL4-UAS system is a powerful biochemical method for studying gene expression in flies, in which the yeast transcription factor GAL4 is expressed in genetically distinct neural subpopulations; crossing with a line linking a reporter gene to GAL4's target enhancer sequence, UAS, labels all cells expressing GAL4 [Brand and Perrimon 1993]. Split-GAL4 operates via the same mechanism, but the gene coding for the GAL4 protein is broken into two pieces that can be linked to separate promoters, causing only cells expressing both pieces to be labeled [Pfeiffer, Ngo et al 2010; Jenett, Rubin et al 2012].

Aso et al screened 7,000 GAL4 lines and 2,500 split-GAL4 lines, and found over 400 split-GAL4 lines that strongly labeled single cells or small groups of cells innervating the mushroom body. They verified the GAL4 results in a followup study using photoactivatable GFP, which labels all neurons innervating an illuminated area. By targeting illumination over the mushroom body, the authors identified two innervating neurons missed by the original study, which has since been identified genetically- but otherwise found that the split-GAL4 screen had successfully identified all neurons innervating the mushroom body [Aso et al, in preparation].

5.1.2 Mushroom body output neurons tile the mushroom body lobes

The innervation pattern of the identified neurons shows remarkable structure. A first class of neurons, termed mushroom body output neurons (MBONs), have dendrites that characteristically innervate particular lobes of the mushroom body. The MBON dendrites cleanly divide the two lobes and the base of the pedunculus into a small number of serial, nonoverlapping compartments, within which the dendrite densely contacts either all KCs or an anatomically distinct subpopulation (KC axons in the α'/β' lobes are clustered into anterior, middle, and posterior groups, in the α/β lobes into posterior, surface, and core groups, and in the γ lobe into main and dorsal groups.) The MBONs innervating each compartment are genetically distinct, and each projects characteristically to other regions of the brain.

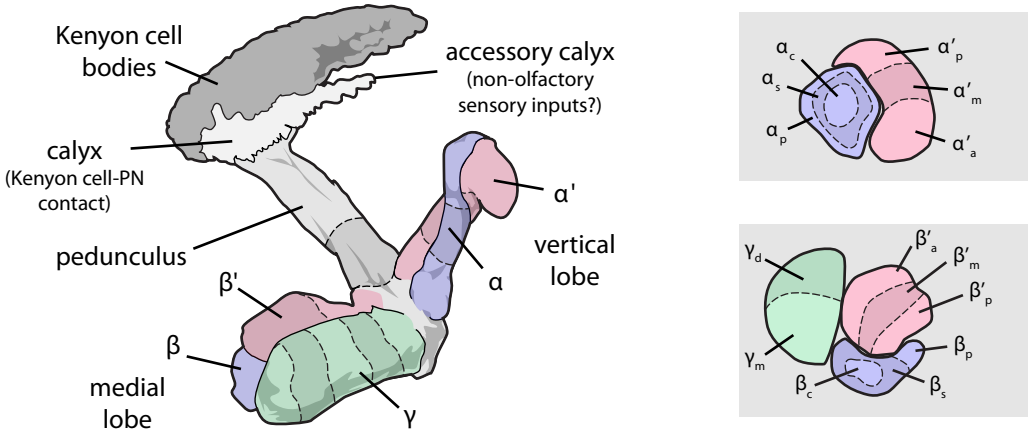


Figure 5.1: Gross anatomy of the mushroom body, reproduced from Tanaka et al [Tanaka, Tanimoto, and Ito 2008] with addition of compartments and γ-lobe subdivisions from Aso et al [Aso et al, in preparation]. Dashed lines reflect the compartments defined by MBON dendrites (some lines obscured). The α and α' lobes are each divided into three compartments, β and β' are each divided into two, γ is divided into five; there is an additional compartment of α/β Kenyon cell axons in the pedunculus. On the right are cross-sectional views of the two lobes, showing anterior, middle, and posterior (a, m, and p) regions of α'/β' lobes, posterior, surface, and core (p, s, and c) regions of α/β lobes, and main and dorsal regions of the γ lobe.

MBONs project from the mushroom body to several regions. A small number of excitatory neurons innervating the α or α' lobes project to lateral horn, a region implicated in control of innate odor-related behaviors. Other neurons, most likely inhibitory, send their axons to other compartments of the mushroom body, suggesting they might gate the response of MBONs in those compartments. And every MBON projects to four neuropils just outside the mushroom body: the crepine (CRE), the superior medial protocerebrum (SMP) the superior intermediate protocerebrum (SIP) and the superior lateral protocerebrum (SLP). MBON axon terminals exhibit distinct innervation patterns within each neuropil, suggesting the presence of stereotyped circuitry.

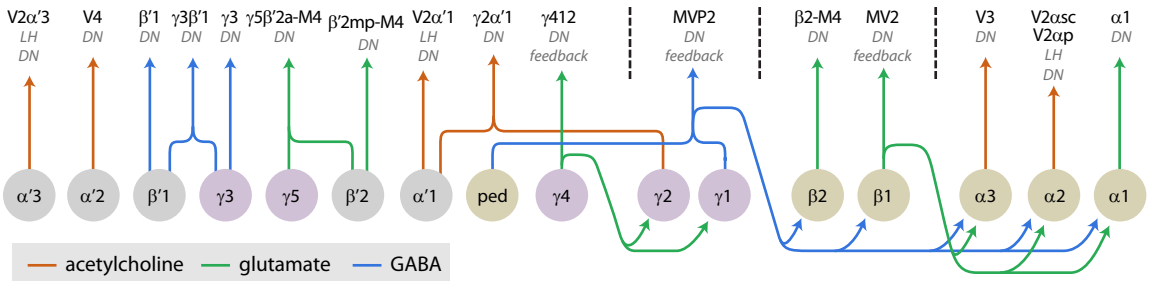


Figure 5.2: MBON innervation of mushroom body compartments, organized into four groups in order of increasing complexity of inputs; reproduced from Aso et al [Aso et al, in preparation]. Compartments are color coded by the type of Kenyon cells they include (yellow = α/β, gray = α'/β', purple = γ), and labeled based on their location in the lobes (lower numbers are more proximal to the pedunculus). MBONs project from the mushroom body to the lateral horn (LH), to the dopaminergic neuropils CRE, SMP, SIP, and SLP (labeled collectively as DN), and to other mushroom body compartments; the targets of each MBON are shown below its name in gray. The leftmost group of nine compartments are the simplest, having only feedforward input. The MBON γ412 projects from γ4 to γ1+2 to form the second group. And MVP2 feeds back from γ1 onto all α/β lobe compartments aside from the pedunculus, and MV2 from β1 additionally feeds back onto all three α lobe compartments, forming the third and fourth groups. Note that acetylcholine is an excitatory neurotransmitter in fly, while glutamate can be either excitatory or inhibitory, but appears to be inhibitory in the mushroom body.

5.1.3 Dopamine controls learning in the mushroom body

In addition to the MBONs, each compartment of the mushroom body is innervated by the axons of a small number of dopaminergic neuron types. Distinct classes of dopaminergic neurons innervate the same specific compartments in the mushroom body as the MBONs– most target a single specific mushroom body compartment, though there are several exceptions. 90% of the dendritic arbors of the dopaminergic neurons reside in the four neuropils targeted by MBON axons: CRE, SMP, SIP, and SLP, where they appear to make contact with MBON axons.

Previous studies have identified two groups of dopaminergic neurons, PPL1 and PAM, that play a role in aversive and appetitive odor learning, respectively. Thus far there does not appear to be a

clear segregation of PPL1 and PAM dendrites among the four neuropils peripheral to the mushroom body, although this will be clarified in further investigation. There appears to be some stereotyped connectivity between MBONs and dopaminergic neurons within the peripheral neuropils, although the full extent of this connectivity is still being determined. The full wiring diagram of MBONs and dopaminergic neurons is shown below.

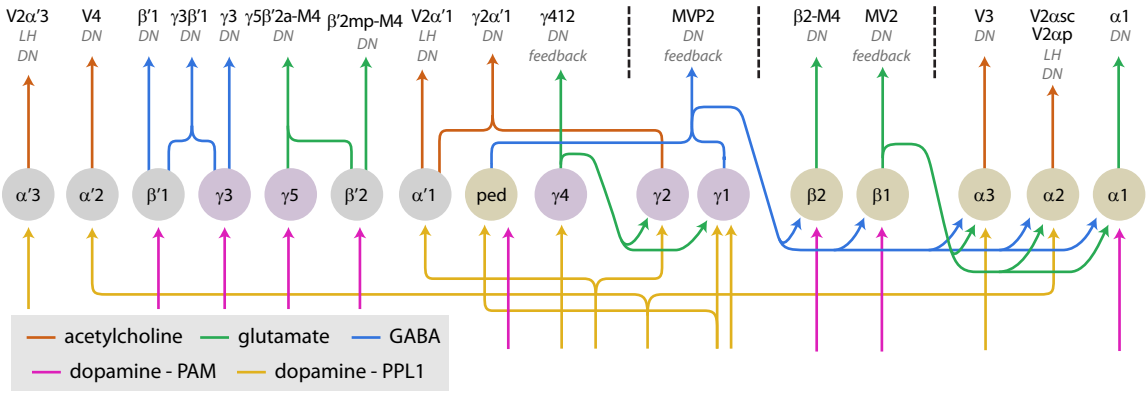


Figure 5.3: MBON compartments showing innervation by dopaminergic neurons [Aso et al, in preparation]. Differences between PAM and PPL1 neurons are discussed below.

As seen in Figure 5.3 above, dopaminergic neurons have extensive recurrent connections with the mushroom body. Dopamine in flies was recently found to play a role in learning and memory similar to its role in vertebrates. Cells in the mushroom body express two forms of dopamine receptors, DAMB and dDA1, the latter of which has been shown to be required for both appetitive and aversive memory [Kim, Lee, and Han 2007]. Different subclasses of dopaminergic neurons seem to signal different kinds of learning signals: the PAM cluster of neurons responds selectively to sugar rewards, and has been shown to be required for appetitive reinforcement learning [Liu et al 2012], while PPL1 neurons are sufficient for aversive memory formation [Claridge-Chang et al 2009]. Dopaminergic neurons have also been found to be modulated by motivational state [Krashes et al 2009]. A subset of PPL1 neurons express receptors for neuropeptide F, a molecule released during food deprivation that drives prolonged feeding behavior [Wu et al 2003]. Blocking of PPL1 neurons to simulate the effect of neuropeptide F allows appetitive conditioning in satiated flies (normally appetitive associations can only be formed when the fly is starved), while stimulating the same neurons suppresses appetitive conditioning in hungry flies [Krashes et al 2009]. This

suggests that the fly's motivational state can gate associative learning by suppressing or driving certain dopaminergic neuron subpopulations.

The mechanism for dopamine's role in odor learning was hinted at by a recent study in locust [Cassenaer and Laurent 2012]. The study focuses on octopamine, rather than dopamine, although at least in flies octopamine is known to act on the mushroom body indirectly, by activating PAM dopaminergic neurons [Burke et al 2012; Waddell 2013]. The locust study finds that the synapses between Kenyon cells and MBONs normally obey a Hebbian spike-timing-dependent plasticity (STDP) rule- however, when KC and MBON spiking is followed by a pulse of octopamine, the shape of this learning rule is changed to be purely depressive. This suggests that dopamine could drive odor learning in MBONs by modulating the synaptic weights of their KC inputs.

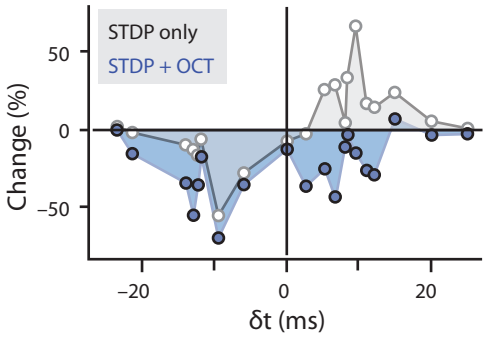


Figure 5.4: KC-MBON STDP learning rule, reproduced from Cassenaer and Laurent [Cassenaer and Laurent 2012]. In gray, the normal STDP rule in KC-MBON synapses, where δt is the time of the postsynaptic spike minus the presynaptic spike, and the y axis shows the percent change in KC-evoked EPSP size in MBONs following five trials in which pre- and postsynaptic spikes were paired at the given δt. In blue, the STDP rule observed when octopamine is injected 1s after pairing.

5.1.4 Outline of a learning circuit in the mushroom body

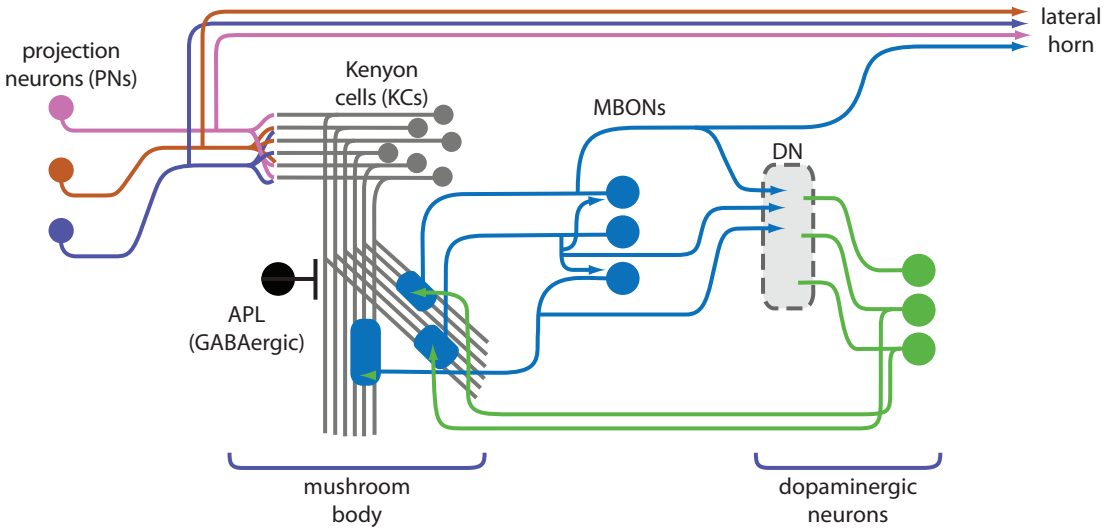


Figure 5.5: Summary of circuit architecture of the mushroom body, showing innervation of the mushroom body lobes by MBONs, and feedback via dopaminergic neurons driven by MBON input.

The architecture of the mushroom body and its peripheral neurons bears strong resemblance to that of the cerebellum, reviewed in Figure 5.6 [Farris 2011; Marr, 1969; Albus, 1971; Ito 1984; Medina and Mauk 2000]. Intrinsic neurons (Kenyon cells/ granule cells) form a small number of synapses with incoming fibers (projection neurons/mossy fibers), and transform the densely, ordered representation in their inputs to sparse, random, and high dimensional patterns of activation. A population of readout neurons (MBONs/Purkinje cells) receive vastly convergent input (up to half of Kenyon cells/200,000 parallel fibers from granule cells), and project their axons to a downstream nucleus (lateral horn/the deep cerebellar nucleus) where they are recombined with projection neuron/mossy fiber inputs. Another set of "teacher" cells (dopaminergic neurons/climbing fibers) gates plasticity from intrinsic neurons to readout neurons. The teacher cells respond to reward/punishment signals, such as food or shocks, and also receive negative feedback from the output of the system, either directly in the case of MBONs, or indirectly via disinhibition of the deep cerebellar nucleus in the case of the cerebellum.

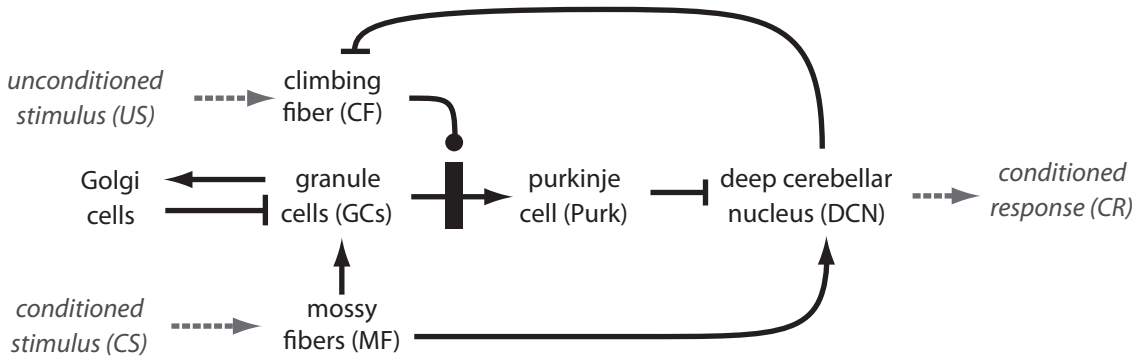


Figure 5.6: The basic circuit architecture underlying cerebellar learning, derived from the model of Medina et al [Medina et al, 2000; Medina, Nores, Ohyama and Mauk, 2000]. The mechanism by which this circuit drives associative learning is reviewed in depth in the introduction to this thesis; but in brief: climbing fiber activity evokes dendritic action potentials called complex spikes in Purkinje cells, triggering synaptic plasticity at granule cell to Purkinje cell synapses. During learning, an unconditioned stimulus (eg a shock) modulates climbing fiber spiking, driving either LTD or LTP in synapses with granule cells activated by the conditioned stimulus (eg a tone or an odor). In the case of an increase in climbing fiber activity, complex spikes evoke LTD at granule cell synapses, causing a temporally-specific decrease in the Purkinje cell response to the conditioned stimulus upon future encounters. The drop in the Purkinje cell response to the conditioned stimulus disinhibits the DCN, which drives downstream motor centers to elicit the conditioned response. Increased DCN activity also inhibits the climbing fiber, balancing the input to the climbing fiber evoked by the unconditioned stimulus and restoring the climbing fiber to its baseline firing rate.

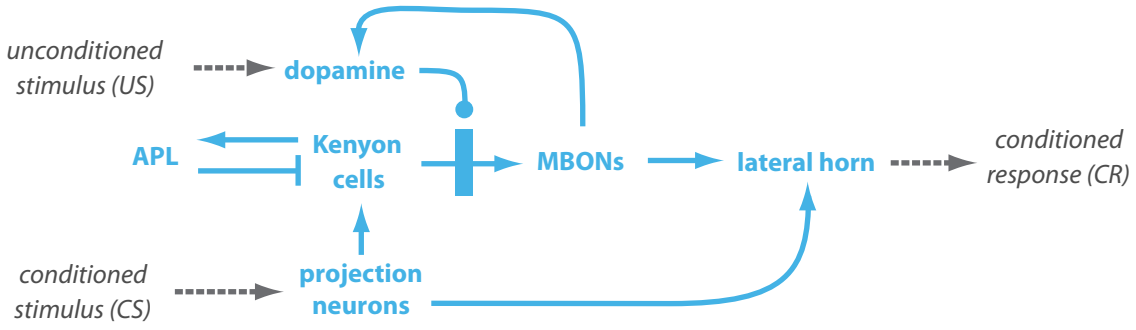


Figure 5.7: The architecture of the mushroom body, arranged to show parallels with the cerebellum in Figure 5.6. Dopaminergic neurons take the place of climbing fibers in gating plasticity from KCs to MBONs, which show modified responses to odors following learning of conditioned avoidance [Séjourné et al 2011]. The lateral horn contains stereotyped circuits involved in driving innate behaviors [Jefferis et al 2007; Datta et al 2008] therefore changing MBON input to the lateral horn could activate or inactivate different motivational states in the fly, or trigger specific behaviors.

The mushroom body both resembles the cerebellar learning circuit physically, and is known to play a similar role in associative learning, suggesting a common circuit mechanism for associative learning in the brain [Heisenberg 2003; Farris 2011]. In this chapter, I construct a simple model of associative memory formation and retrieval by a single MBON, inspired by a classical model of cerebellar learning [Medina et al, 2000], and discuss the circuit mechanisms by which this model can form large numbers of associative memories. I then summarize a few possible alternative models for memory formation, and the predictions made by each. As the interaction of MBONs with dopaminergic neurons and with the lateral horn becomes better understood, the model developed here can be expanded to fit the constraints of the observed circuitry, and provide a more complete picture of the sites of learning in the mushroom body. The complexity of learning uncovered in the cerebellum [Boyden, Katoh, and Raymond 2004] warns that plasticity on the level of individual MBONs may be only part of the story of learning in the mushroom body. In particular, the multilayered interactions between compartments of the mushroom body suggests some form of gating or hierarchical processing of learned behaviors, and has no direct equivalent in models of cerebellar learning. But the powerful ensemble of genetic tools available in fly, as well as the

relatively small number of neurons making up the mushroom body's inputs and outputs, offer hope that the complete circuit mechanism underlying the acquisition, storage, and retrieval of associative memories in the mushroom body can be uncovered.

5.2 Modeling dopamine-mediated odor learning in a mushroom body output neuron

Having previously developed a model of odor representations in the mushroom body, I asked whether dopamine-modulated plasticity in a model MBON receiving input from my model KCs was sufficient to drive acquisition of a conditioned response. MBONs are broadly tuned across odors; following learning of a conditioned response to an odor, their response to the conditioned odor is specifically altered, while responses to other odors are unaffected [Séjourné et al 2011]. I asked whether a dopamine-modulated learning rule could drive a change in the odor-evoked firing rate of a model MBON that is specific to the conditioned odor. Because the fly should be able to form multiple associative memories in its lifetime, I also investigated whether dopamine-modulated learning in KC-MBON synapses could drive learning of multiple odors, without altering MBON responses to odors not in the training set. For simplicity, I will ignore the effects of feedback interactions between mushroom body compartments. I will first treat dopamine as an external signal conveying information about the unconditioned stimulus. I will then examine the effects of feedback connections from the mushroom body, in which dopamine is modulated both by the unconditioned stimulus and by the odor-evoked response of the model MBON.

5.2.1 Reward-modulated learning framework

In behaving flies, an odor and its associated punishment or reward are often separated in time—for instance, odor cues signaling nearby food are experience before the reward associated with finding the food occurs—meaning a mechanism for associative learning should have some capacity to combine information across time. There is evidence for such a learning mechanism in Cassenaer and Laurent's study of octopamine-modulated STDP in locusts [Cassenaer and Laurent 2012], in which local administration of octopamine one second after spike pairing in KCs and MBONs was

sufficient to alter the shape of the STDP learning rule at the KC-MBON synapse.

A general model for learning with delays between stimulus and reward is reward-modulated spike-timing dependent plasticity (R-STDP) [Fremaux, Sprekeler and Gerstner 2010]. In this framework, a candidate change in synaptic weights derived from an unsupervised learning rule such as STDP is effective only if it is followed by a reward. Focusing on learning in a single MBON, for a synapse from the i^{th} KC to the MBON, R-STDP is formulated as:

$$\begin{aligned} \tau_s \frac{ds_i}{dt} &= -s_i(t) + \eta UL_i(t) \\ \frac{dw_i}{dt} &= R(t)s_i(t) \end{aligned}$$

where R = reward and s_i is a synaptic eligibility trace that stores candidate weight changes from the unsupervised learning rule UL_i, scaled by an arbitrary parameter η (In following sections I will typically neglect the subscript i for simplicity.) Biologically, s could correspond to elevated calcium concentrations in active synapses, or another molecular indicator of recent synaptic activity. The reward signal $R(t)$ locks in s_i as an actual change in the synaptic weight w_i (which I constrain to be nonnegative). In keeping with the small number of dopaminergic neurons innervating the mushroom body, I assumed R to be a global reward signal across all synapses on a single MBON, rather than being synapse-specific.

5.2.1.1 Choice of UL

In selecting UL, I chose to ignore timing components of KC and MBON responses, and focus on learning a change in the time-averaged odor-evoked firing rate of MBONs. I used a simple unsupervised learning rule with two terms reflecting nonassociative (α) and associative (β) components of learning:

$$UL_i = \alpha r_{\mathrm{KC}_i} - \beta r_{\mathrm{KC}_i} r_{\mathrm{ON}}$$

where r_{KC_i} is the total odor-evoked activity of the i^{th} KC, and r_{ON} is the odor-driven firing rate of the MBON. Ignoring the effect of the reward term, the unsupervised learning rule has a fixed

point ($UL_i = 0$) at

$$r^*_{\text{ON}} = \frac{\alpha}{\beta}$$

that is stable when ($\alpha > 0$, $\beta > 0$), and unstable when ($\alpha < 0$, $\beta < 0$).

5.2.2 Using dopamine to encode odor valence

The most straightforward role of dopaminergic neurons is to encode the valence of odor stimuli—ie whether it is followed by a reward or shock. I refer to odors paired with an odor or shock "trained" odors, and odors that are not paired "untrained" odors, and set the dopamine signal $R(t)$ as:

$$R(t) \quad = \quad \begin{array}{|ll} 0 & \text{untrained odor} \\ 1 & \text{trained odor} \end{array}$$

Under these conditions, MBON responses to trained odors will converge to the fixed point of the unsupervised learning rule, $r^*_{\text{ON}} = \alpha/\beta$.

To test performance of the learning rule, I simulated learning of random sets of odors from the Hallem and Carlson dataset [Hallem and Carlson 2006]. I initialized MBON responses to all odors to be equal, then divided the odors into two sets: of trained odors, A, and untrained odors, B. I drove the mushroom body model with an alternating train of odor pulses, drawing the first from set A, second from set B, and so on, selecting the stimulus odor from each set at random each time (I excluded five odors from the dataset that activated fewer than 0.5% of KCs.) Odors were followed after 2 seconds by a pulse of dopamine, of amplitude determined by $R(t)$ (ie 1 for odors in set A, and zero for odors in set B). I set synaptic eligibility traces to zero before each odor pulse, to simulate the effect of long delays between odor presentations.

After training, I set a threshold equal to the minimum change in MBON response to the trained odors in set A, and counted the fraction of odors not in set A for which the readout neuron response crossed the threshold. These are false positives, or cases of overgeneralization, in which training the MBON to respond preferentially to some odors causes it to also respond to odors not in the training set.

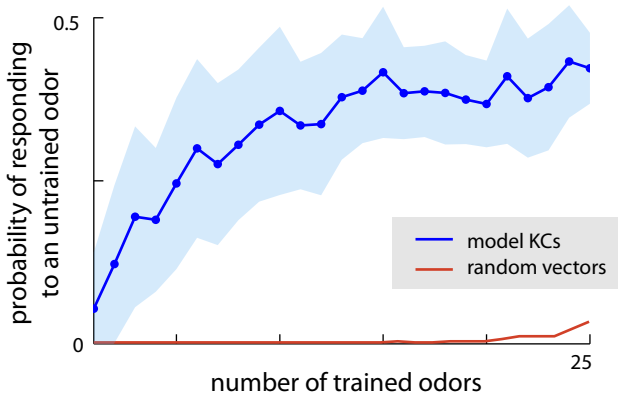


Figure 5.8: Probability of responding to untrained odors, as a function of the number of trained odors. While the learning rule performs well on sparse random vectors (red line), actual KC representations of odors have substantial overlap, and training on a small set of odor causes substantial overgeneralization to untrained odors.

The valence-based learning rule quickly overgeneralized as the number of trained odors increased (Figure 5.8): when the size of the training set reached only five odors, the MBON responded incorrectly to 25% of untrained odors. This is mainly due to overlap in KC representation of odors: if KC representation of two odors is very similar, changing the MBON response to the first odor will significantly modify its response to the second, potentially leading to the second odor evoking a false positive response. To verify that overlapping representations were at fault, and not the number or sparseness of KCs, I replaced KC responses with random binary vectors of the same mean sparsity; I found that with the random KC responses, MBON performance in the associative learning task was vastly improved (Figure 5.8, red line).

Thus, correlations between odor representations in the KCs seems to impair learning of odor-specific responses. There are two ways this problem could be solved. First, I could adjust my model of the olfactory periphery. In particular, I had little information about the response properties and connectivity of the normalizing GABAergic neuron APL. It could be that APL connectivity with Kenyon cells is finely tuned to decorrelate representations of odors prior to learning. Alternatively, a more carefully tuned learning rule could reduce the extent of overgeneralization by the MBONs.

Adjusting the learning rule is preferable to trying to decorrelate odor representations completely. While correlated odor representations interfere with learning of odor-specific responses, there are cases in which they may also be helpful for the fly. For example, odors in nature are not usually encountered at consistent concentrations and mixing ratios. In the previous chapter, I showed that my model KCs preserve similarity between different mixing ratios in binary odor mixtures, as has been previously observed experimentally [Glenn Turner, private communication]. This preservation of similarity means that associative memories formed from one mixture containing an odor are more likely to also be recalled by other, similar mixtures with that odor. More stringent decorrelation of odors will disrupt this preservation of odor similarity, impairing generalization in cases where it is useful to the fly. Instead, I use a more flexible learning rule that allows MBON responses to familiar, non-salient odors to be kept distinct from responses to behaviorally meaningful odors.

5.2.3 Adding a second fixed point of learning reduces overgeneralization

Overgeneralization occurs because modifying MBON responses to some odors alters their responses to other odors. This can be avoided by designing a learning rule that controls the MBON response to both trained and untrained odors. To achieve this in the odor valence model, I added a second term to the learning rule:

$$\begin{aligned} \tau_s \frac{ds}{dt} &= -s(t) + \alpha r_{\mathrm{KC}}(t) - \beta r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t) \\ \frac{dw}{dt} &= \big(R(t)s(t)\big) + \big(\gamma r_{\mathrm{KC}}(t) - \delta r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t)\big) \\ R(t) &= \{0 \text{ for untrained odors}, 1 \text{ for trained odors}\} \end{aligned}$$

For untrained odors, the first term of dw/dt is zero and the learning rule has a single fixed point governed by the second term; this fixed point is stable when $\gamma > 0$, $\delta > 0$, and unstable otherwise. But for odors paired with reward, both terms of learning are nonzero, and the fixed point of learning moves to a new value. The learning rule therefore has two effective fixed points for trained and untrained odors:

$$r^*_{\mathrm{ON}} = \begin{cases} \gamma/\delta & \text{untrained} \\ (\gamma + e^{-\Delta t/\tau_s}\alpha)/(\delta + e^{-\Delta t/\tau_s}\beta) & \text{trained} \end{cases}$$

where Δt is the time between odor presentation and reward, so that $e^{-\Delta t/\tau_s}$ is the decay of the synaptic eligibility trace at the time the reward is presented. (If $\alpha/\beta = \gamma/\delta$ the two fixed points are the same, but this case is easily avoided.) If the fly is trained to respond to odor A, its ongoing exposure to other odors in its environment will cause this learning rule to selectively strengthen readout from KCs that are highly specific to odor A, leaving the response to other odors unaffected. I tested this in a toy example with two odors A and B, and seven KCs—one specific to odor A, one to odor B, and five responding to both odors. I use learning rule parameters of $\alpha = 0$, $\beta = -4$, $\gamma = 5$, $\delta = 1$—so that learning decreased the MBON response to the trained odor. In preliminary investigation of MBONs during learning, some cells increased their response to odors after learning, while others decreased their response [Daisuke Hattori, unpublished observations]; the different directions of firing rate change following learning could reflect different roles of MBONs in modulating downstream circuitry, but so far not enough is known about downstream targets of MBONs to draw conclusions.

As seen below, the model MBON is able to learn a specific response to odor A while preserving its baseline response for odor B. The synaptic weights from KCs that respond to both odors (mixed neurons) are not strongly affected by learning: presentation of odor B drives a decrease in synaptic weight that counteracts the increase evoked by the pairing of odor A with dopamine.

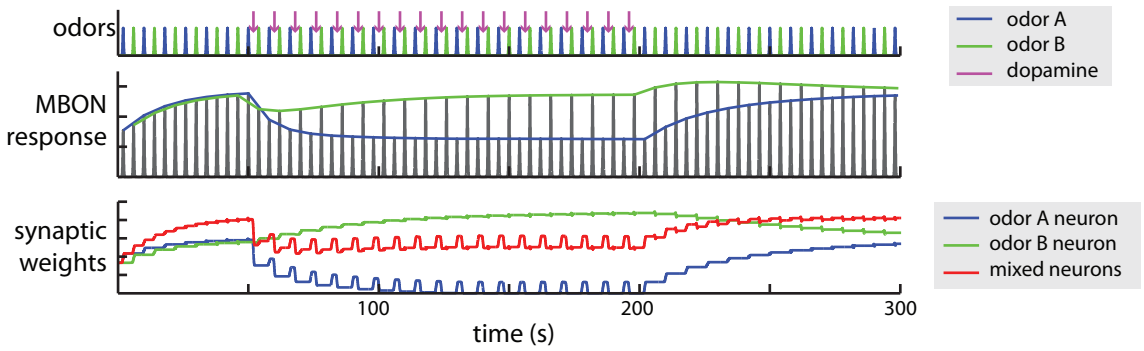


Figure 5.9: Learning rule performance in a toy model. **Top**: timing of odor stimuli. Initially, odors A and B are presented to the mushroom body model in alternating pulses. After 50 seconds of stimulation, presentation of odor A is followed after 2 seconds by presentation of dopamine. At 200 seconds, dopamine signaling is turned off. **Middle**: firing rate of the model MBON; responses to odors A and B are connected by blue and green lines, respectively. In the first 50 seconds, the MBON response to both odors adapts to the fixed point γ/δ of untrained odors. Upon pairing of odor A with dopamine, the response of the MBON to odor A drops; the response to odor B is transiently affected, but is quickly restored to the untrained odor fixed point. When dopamine is turned off at 200 s, the trained response to odor A is extinguished, and the MBON response returns to the untrained fixed point for both odors. **Bottom**: synaptic weights from KCs to MBONs: in the toy example, one neuron responded to odor A, one to odor B, and five to both odors. During training, the synaptic weight from the odor A neuron drops, causing the MBON firing rate to decrease to the trained fixed point. Mixed neurons are also affected by pairing, but to a lesser degree, while the odor B neuron increases its synaptic weight to compensate in the drop in mixed neuron synaptic weights, and bring the MBON response to odor B back to the untrained fixed point.

5.2.3.1 Exposure to multiple untrained odors increases specificity of learned responses

I next tested odor learning as in the previous section, by simulating learning of random sets of taken from 105 odors in the Hallem and Carlson dataset [Hallem and Carlson 2006]. As before, I define a set of trained odors, A, and of untrained odors, B, and drove the mushroom body model

with alternating pulses of odors drawn from the two sets. I found that if I divided the entire dataset between groups A and B, the MBON responded to trained odors without any instances of overgeneralization. When I decreased the size of set B the probability of overgeneralization rose, though it did so gradually, such that encountering a limited example set of untrained odors decreased the probability of overgeneralizing even for odors not encountered during learning. For example, given set A = 25 odors and set B = 50 odors, approximately 17% of odors not encountered during learning evoked false positive responses, as opposed to 40% in the original learning formulation.

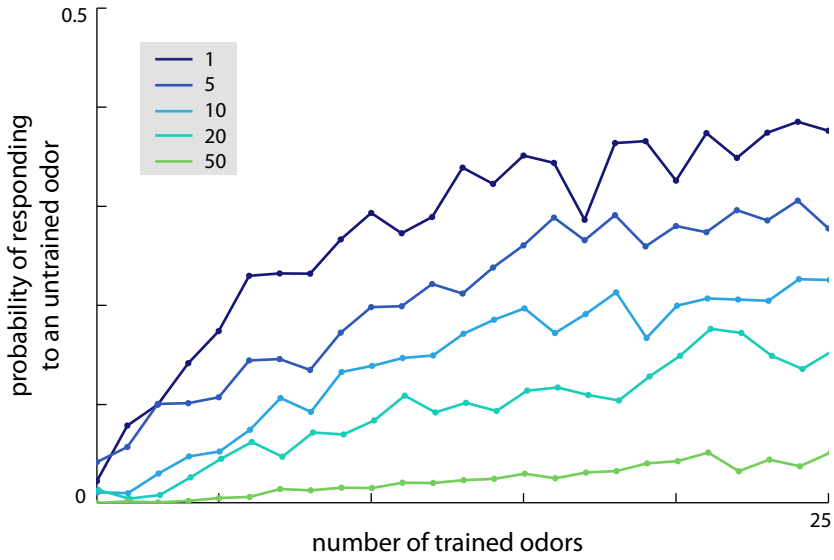


Figure 5.10: Probability of overgeneralization, ie responding to odors not in the set A of trained odors, as a function of the size of set A, plotted for different sizes of set B (legend). I included both odors in set B and odors not encountered during training when measuring the probability of overgeneralization. The probability of overgeneralization dropped as the number of odors in set B increased; I found that the model never overgeneralized to odors in set B, but also that increasing the size of set B decreased the probability that the model would overgeneralize to odors not encountered during training.

5.2.3.2 Mushroom body output neurons show evidence for slow adaptation

As seen in the toy example, this learning rule causes a slow adaptation of the MBON to untrained odors. Upon first presentation of an odor the MBON response is small, given the randomly initialized synaptic weights from activated KCs; in subsequent trials, the response increases to the fixed point γ/δ. Interestingly, a similar slow adaptation is observed in MBON responses to odors presented without training [Daisuke Hattori, unpublished observations]. In the figure below, the fly was presented with stimulus blocks of four odors; each odor was presented once in a block, as a train of ten one-second pulses of odor. The block of four odors was then repeated several times—thus for any particular odor, there was a delay of a few minutes from its presentation in block n to its presentation in block $n + 1$, during which other odors were being presented. From the first to the fourth such block, the average MBON response to each odor increased, suggesting a slow adaptation to odors on a timescale of minutes. (KC representations of odors show no such adaptation under these conditions.)

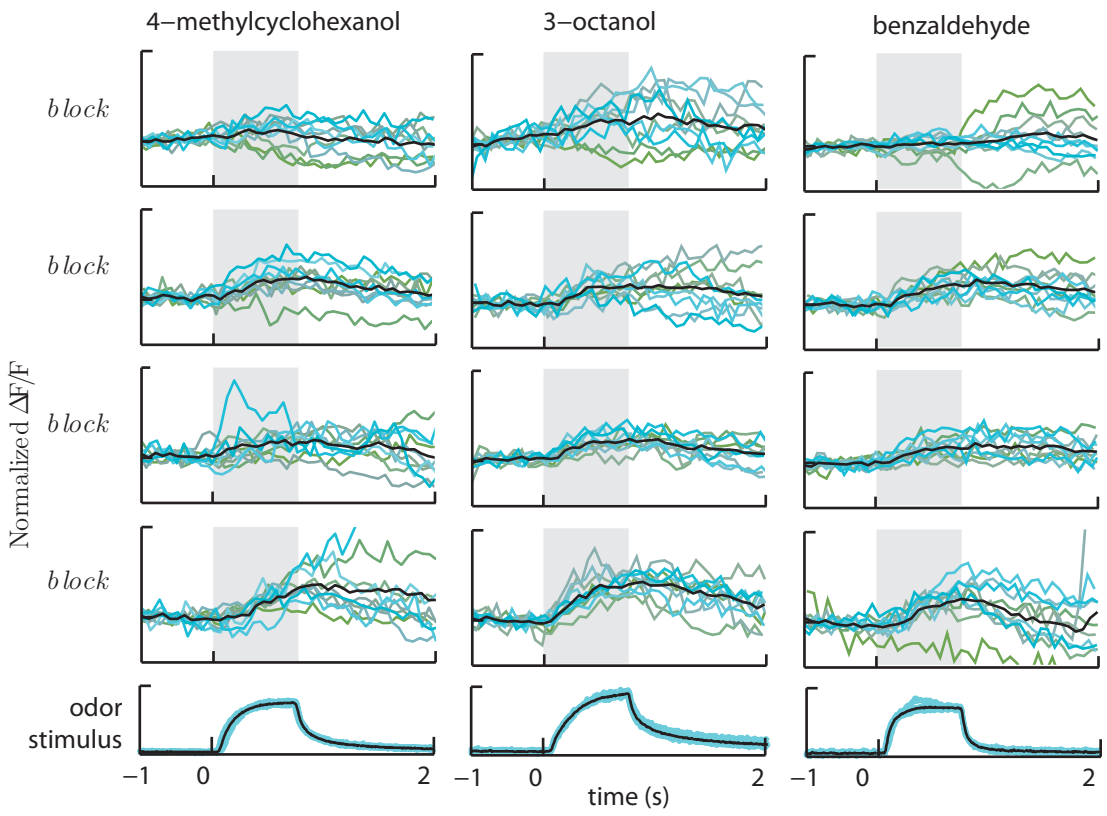


Figure 5.11: Calcium imaging of the response of an MBON to three odors (a fourth stimulus of plain air was included, but is not shown here; MBON responses to the air were negligible in all blocks). This particular MBON showed an increase in response to all three odors tested, while other MBONs showed a decrease in response on a similar timescale [Daisuke Hattori, unpublished observations].

In the valence model I have presented in this section, this adaptation could be a signature of the MBON evolving to a fixed firing rate for untrained odors, as a means of preserving specificity of learned responses.

5.2.4 Predictions about learning from the modified valence model

5.2.4.1 Dopamine changes the shape of the KC-MBON learning rule

In this model, the value of $R(t)$ reflects the presence or absence of dopamine in the system. Consistent with the findings of Cassenaer and Laurent in locust, dopamine effectively changes the KC-MBON learning rule, from $\gamma r_{\mathrm{KC}}(t) - \delta r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t)$ in the absence of dopamine to $(\gamma + e^{-\Delta t/\tau_s}\alpha) r_{\mathrm{KC}}(t) - (\delta + + e^{-\Delta t/\tau_s}\beta) r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t)$ in the presence of dopamine.

5.2.4.2 The learned MBON response depends on the timing of the dopamine signal

The response of the MBON after learning, $r^*_{\mathrm{ON}} = (\gamma + e^{-\Delta t/\tau_s}\alpha)/(\delta + e^{-\Delta t/\tau_s}\beta)$, depends on the value of Δt, the time between odor presentation and dopamine response. The model assumes that the synaptic eligibility trace decays exponentially following odor presentation, which is not necessarily the case in the mushroom body– but independent of the exact form of decay, the learned MBON response should have some dependence on Δt. This could be tested experimentally by repeating the experiment of Cassenaer and Laurent with different delays between spike pairing and application of dopamine: my model predicts that the shape of the learning rule should smoothly evolve as Δt is increased, converging in the limit of large Δt at the original learning rule shape observed in the absence of dopamine.

5.2.4.3 Persistent presentation of dopamine is required to maintain the learned response

The fixed point of r^*_{ON} is controlled by the presence or absence of dopamine. As a result, trained odors should continue to evoke a dopamine response, even after the trained response has been learned. Additionally, blocking activity of dopaminergic neurons should lead to forgetting of the learned response, because the fixed point of the learning rule will return to its untrained value. This prediction conflicts with the results of Berry et al, in which blocking dopamine lead to reduced forgetting of learned responses in flies [Berry, Cervantes-Sandoval, Nicholas, and Davis 2012]. I will

discuss a possible solution to this conflict in the following section.

5.2.4.4 Learned changes in MBON responses cannot be bidirectional

The valence model predicts that individual MBONs will either always increase or always decrease their responses to learned odors. The fixed point for learned odors is determined by the timing of dopamine Δt and by the learning rule parameters α, β, γ, and δ, which are not odor-dependent—therefore it is impossible to have $r^*_{\mathrm{ON}} > \gamma/\delta$ for learned response A and $r^*_{ON} < \gamma/\delta$ for learned response B. MBON learning has not been studied using large panels of trained odors—it will be informative to see whether the prediction of this model holds, or whether the effect of learning on MBON firing rates is more flexible than this learning rule permits.

5.3 A possible biological mechanism for valence learning without forgetting

The valence model presented above fails to replicate one previous experimental result, which finds that neither learning or forgetting can occur in the mushroom bodies in the absence of dopamine [Berry, Cervantes-Sandoval, Nicholas, and Davis 2012]. Because the valence model depends on he presence of dopamine to keep r_{ON} at the fixed point for learned odors, blocking dopamine will lead to forgetting of learned associative memories, as the MBON response decays back to the untrained fixed point for all odors.

To fix this contradiction, I made an adjustment to the valence model based on a hypothesis presented by Berry et al, that learning and forgetting are mediated by separate dopamine receptors in the mushroom body. The mushroom body expresses two types of dopamine receptor: DAMB, which has a high affinity for dopamine and thus a low threshold for activation [Han, Millar, Grotewiel, and Davis 1996], and dDA1, which has a low affinity for dopamine, thus a high threshold for activation [Kim, Lee, and Han 2007]. I designed a modified learning rule incorporating these two receptor

classes:

$$\begin{aligned}
\tau_s \frac{ds_{\mathrm{dDA1}}}{dt} &= -s_{\mathrm{dDA1}}(t) + \alpha r_{\mathrm{KC}}(t) - \beta r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t) \\
\tau_s \frac{ds_{\mathrm{DAMB}}}{dt} &= -s_{\mathrm{DAMB}}(t) + \gamma r_{\mathrm{KC}}(t) - \delta r_{\mathrm{KC}}(t) r_{\mathrm{ON}}(t) \\
\frac{dw}{dt} &= DAMB(t) \cdot s_{\mathrm{DAMB}}(t) + dDA1(t) \cdot s_{\mathrm{dDA1}}(t)
\end{aligned}$$

where $DAMB(t)$ and $dDA1(t)$ are activation levels of DAMB and dDA1 receptors. The learning rule works similarly to the previous model—at low concentrations of dopamine only the DAMB receptor is activated, and learning evolves to the fixed point of the first unsupervised learning rule, $r^*_{ON} = \gamma/\delta$. At high concentrations of dopamine both receptors are activated, and learning evolves to the fixed point of the combined learning rules, $r^*_{ON} = (\gamma + k\alpha)/(\delta + k\beta)$, where k depends on the strength of activation of the two receptors. I gave both receptor classes sigmoidal activation functions,

$$\begin{aligned}
DAMB(t) &= (1 + e^{-m_1(\theta_{\mathrm{DAMB}} - R(t))})^{-1} \\
dDA1(t) &= (1 + e^{-m_2(\theta_{\mathrm{dDA1}} - R(t))})^{-1}
\end{aligned}$$

parameterized by activation thresholds θ_{DAMB} and θ_{dDA1} that determine the value of R at which receptors are 50% activated, and slopes m_1 and m_2 that determine rate of activation around that threshold. For learning to be successful, I found that the linear portions of the dDA1 and DAMB activation functions had to be well separated, so that each was saturated near the other's activation threshold, as in the following example.

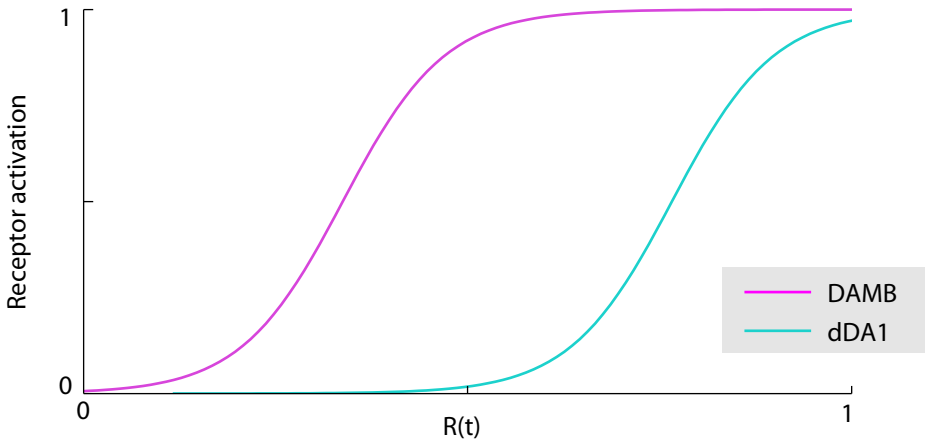


Figure 5.12: Example activation functions for the DAMB and dDA1 dopamine receptors; parameters here are set to $\theta_{\text{DAMB}} = 0.35$, $\theta_{\text{dDA1}} = 0.75$, and $m_1 = m_2 = 15$.

To reproduce the results of the previous valence model, I set $R(t)$ according to:

$$\begin{aligned} R(t) &\approx \theta_{\text{DAMB}} && \text{untrained} \\ &> \theta_{\text{dDA1}} && \text{trained} \end{aligned}$$

This formulation works exactly like the previous valence model, except $R(t)$ is now some small, nonzero value for untrained odors. Setting $R(t) = 0$ to recreate the effect of blocking dopamine causes activation of both $dDA1(t)$ and $DAMB(t)$ to approximately zero, thus blocking all changes in synaptic weights in the absence of dopamine as observed in Berry et al [Berry, Cervantes-Sandoval, Nicholas, and Davis 2012].

5.3.0.5 MBON feedback onto dopaminergic neurons

I have so far neglected MBON feedback onto dopaminergic neurons, and set $R(t)$ using the valence of the odor stimulus. To determine whether the dDA1/DAMB model could still work when dopaminergic neurons were driven by MBONs, I modified $R(t)$ to include a component driven by the MBON response:

$$\begin{aligned} R(t) &= c \cdot r_{\text{ON}} + \text{A} && \text{untrained} \\ &= c \cdot r_{\text{ON}} + B && \text{trained} \end{aligned}$$

where c is a scalar weight, A is the externally-driven activation of dopaminergic neurons for untrained odors, and B is the externally-driven activation of dopaminergic neurons for trained odors.

Momentarily ignoring A and B, the dynamics of dw/dt are driven by $R(t)$, and have a critical point at the value R_{crit}, where

$$\left|DAMB(t) \cdot s_{\text{DAMB}}(t)\right| = \left|dDA1(t) \cdot s_{\text{dDA1}}(t)\right|$$

(recall that DAMB(t) and dDA(t) are both functions of $R(t)$.) For $R(t) < R_{\text{crit}}$, the effects of DAMB dominate and r_{ON} is driven to γ/δ, while for $R(t) > R_{\text{crit}}$, effects of dDA1 dominate and r_{ON} is driven to $(\gamma+\alpha)/(\delta+\beta)$. When $R(t)$ is itself a function of r_{ON}, this leads to positive or negative feedback in learning, depending on the sign of c and the two fixed points of dw/dt:

$$\left.\begin{array}{ll} c>0; & \gamma/\delta < (\gamma+\alpha)/(\delta+\beta) \\ c<0; & \gamma/\delta > (\gamma+\alpha)/(\delta+\beta) \end{array}\right| \text{ positive feedback}$$

$$\left.\begin{array}{ll} c>0; & \gamma/\delta > (\gamma+\alpha)/(\delta+\beta) \\ c<0; & \gamma/\delta < (\gamma+\alpha)/(\delta+\beta) \end{array}\right| \text{ negative feedback}$$

Positive feedback

In the case of positive feedback, the R_{crit} point is unstable. Consider the case where $c > 0$, $\gamma/\delta < (\gamma+\alpha)/(\delta+\beta)$. If $R(t) < R_{\text{crit}}$, learning drives r_{ON} down towards γ/δ; the decrease in r_{ON} causes a further drop in $R(t)$, leading to a further decrease in r_{ON}. When $R(t) > R_{\text{crit}}$, learning drives r_{ON} up towards $(\gamma+\alpha)/(\delta+\beta)$, an increase that causes a further increase in $R(t)$ above R_{crit}.

For the system to learn in the presence of positive feedback, I set A and B such that $c \cdot r_{\text{ON}} + A < R_{\text{crit}}$ for untrained odors, and $c \cdot r_{\text{ON}} + B > R_{\text{crit}}$ for trained odors. If MBON feedback is strong, ie learning alters r_{ON} to the extent that $c \cdot r_{\text{ON}} + A > R_{\text{crit}}$, the learned response will persist in the absence of the reinforcing signal. This formulation is also susceptible to false positives: if a random untrained odor drives the MBON enough to push $c \cdot r_{\text{ON}} + A$ above R_{crit}, positive feedback will drive the MBON response to the trained fixed point upon repeat presentations of the odor. Alternatively, MBON feedback could be kept weak enough that $R(t)$ can only exceed R_{crit} when an odor is paired with reward/shock– that is, $c \cdot r_{\text{ON}} + A < R_{\text{crit}}$ for all attainable values of r_{ON}.

This setup is more robust against false positives, but learned responses will not be preserved if an odor ceases to be paired with the reinforcing signal.

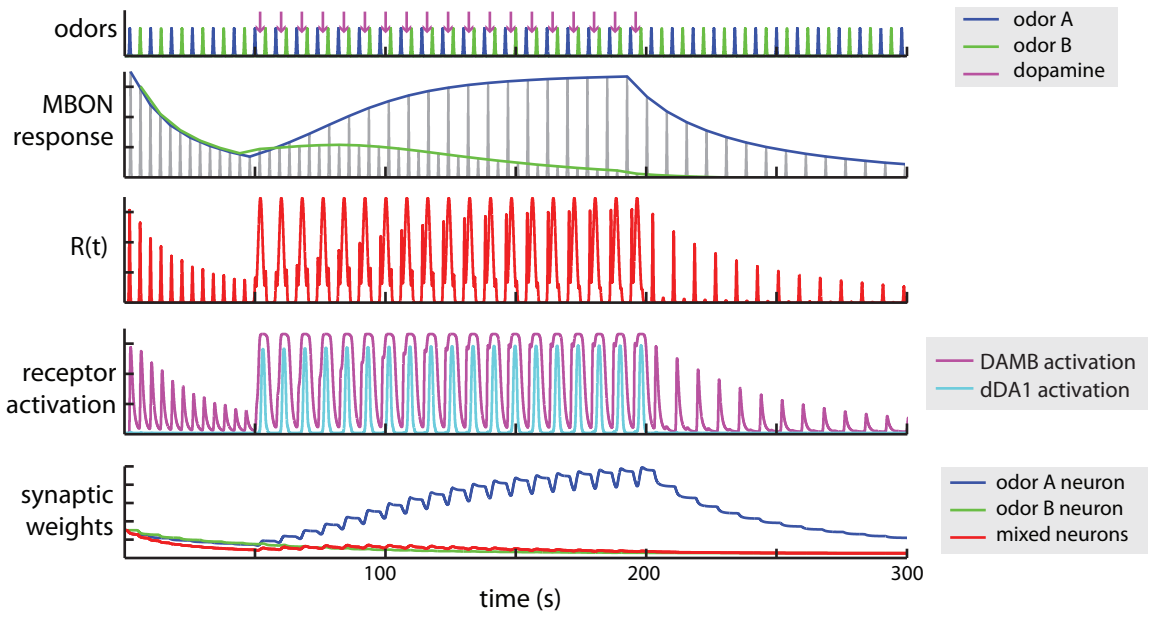


Figure 5.13: Example of learning in the DAMB/dDA1 model, using the toy model described in Figure 5.9; in this case, the learning rule is set to drive a decreased MBON response to untrained odors (odor B), and an increased response to trained odors (odor A). Odor, MBON response, and synaptic weights plots are as in Figure 5.9. The middle plot shows $R(t)$, which is a weighted sum of the MBON firing rate and the external valence signal. Beneath this is a plot showing the activation of the two dopamine receptors by $R(t)$. Low values of $R(t)$ only activate DAMB, which drives a decrease in synaptic weights, while high values of $R(t)$ are strong enough to activate dDA1, which drives an increase in synaptic weights. The change in MBON response to the trained odor is not strong enough to drive $R(t)$ past R_{crit}, thus after pairing is turned off at 200 seconds, the response of the MBON to odor A decays back to the untrained firing rate.

Negative feedback

With negative feedback, the critical point at R_{crit} is stable, and changes in $R(t)$ driven by valence cues are opposed by the resulting changes in r_{ON}. For learning to create significant changes in r_{ON}, the effect of valence cues on $R(t)$ must be large compared to the effects of r_{ON}. At the same time, the model's ability to adjust synaptic weights is limited, because the fixed point R_{crit} usually occurs

on the linear portion of the dDA1 activation function, at which point DAMB is not sensitive to changes in $R(t)$ (see activation functions in Figure 5.12). I found learning more difficult to control as a result, although further study could turn up parameter settings in which the negative feedback regime offers better control of the MBON response.

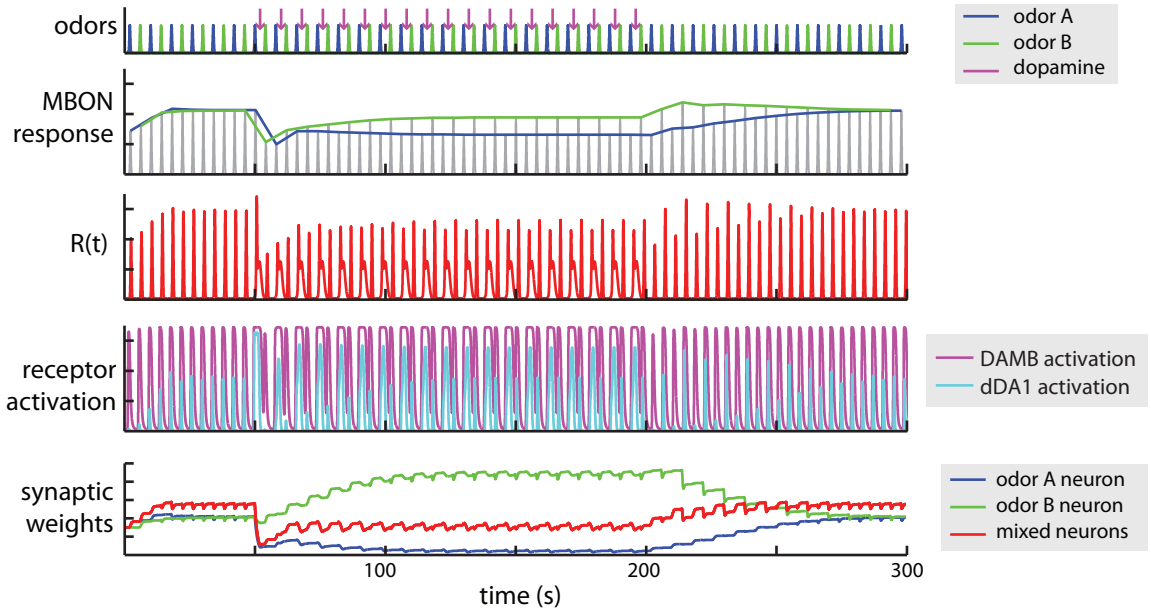


Figure 5.14: Example of learning in the negative feedback regime. Prior to pairing with dopamine, the critical point R_{crit} is stable, and MBON firing rates converge to a value at which $\left|DAMB(t) \cdot s_{\text{DAMB}}(t)\right| = \left|dDA1(t) \cdot s_{\text{dDA1}}(t)\right|$. Upon pairing with dopamine, increased recruitment of dDA1 relative to DAMB drives down the MBON response to odor A. Because changes in r_{ON} counteract changes in $R(t)$, it is difficult to drive large changes in the MBON response to the trained odor.

5.3.0.6 Increased overgeneralization of learning in DAMB mutants

An interesting prediction from this model is the effect of mutations in dDA1 and DAMB receptors on learning. I focused on the positive feedback model for $R(t)$, though these effects should hold in models without MBON feedback onto dopaminergic neurons as well.

dDA1 mutant flies have previously been found to be incapable of both aversive and appetitive learning [Kim, Lee and Han 2007]. Using learning parameters from the toy learning example of

Figure 5.13, I created a dDA1 mutant by setting the dDA1 activation function to zero for all values of $R(t)$. The MBON response decays to the fixed point of the DAMB-mediated learning rule, and is not affected by external input to $R(t)$:

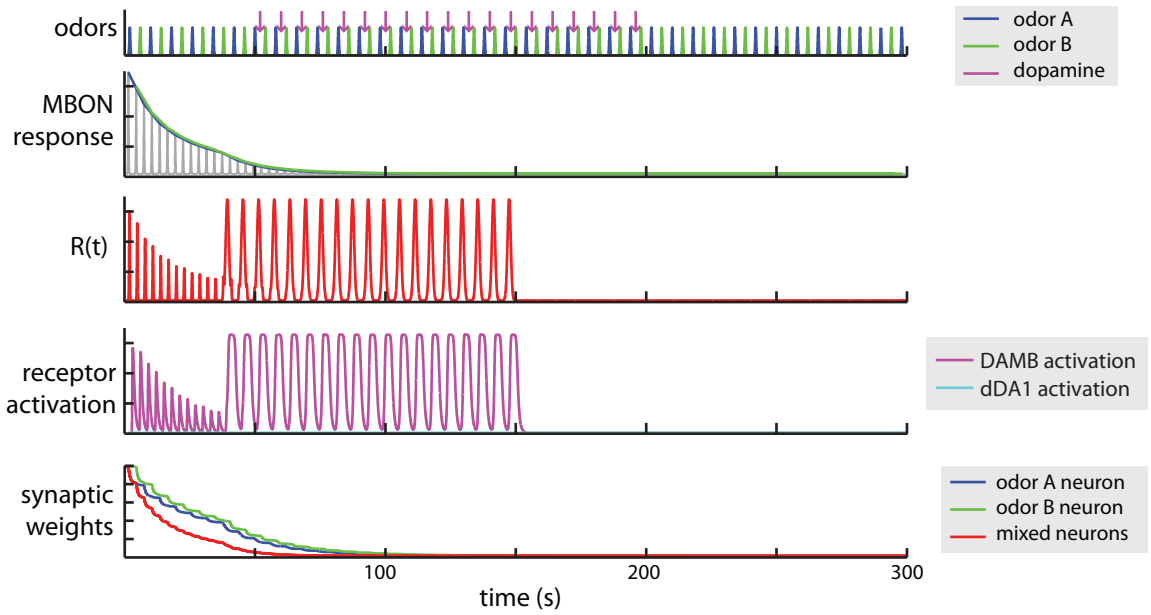


Figure 5.15: dDA1 mutant model: MBON response to all odors decays to the fixed point of untrained odors, γ/δ, which in this case was zero.

Similarly, I created DAMB mutant flies by setting the DAMB activation function to zero for all values of $R(t)$. These models were able to learn in the presence of dopamine, however without the dDA1-mediated adaptation to untrained odors, they were much more prone to generalizing to other odors. It would be interesting to test this prediction experimentally, to determine whether DAMB does indeed play a role in preserving specificity of learned behavior in flies.

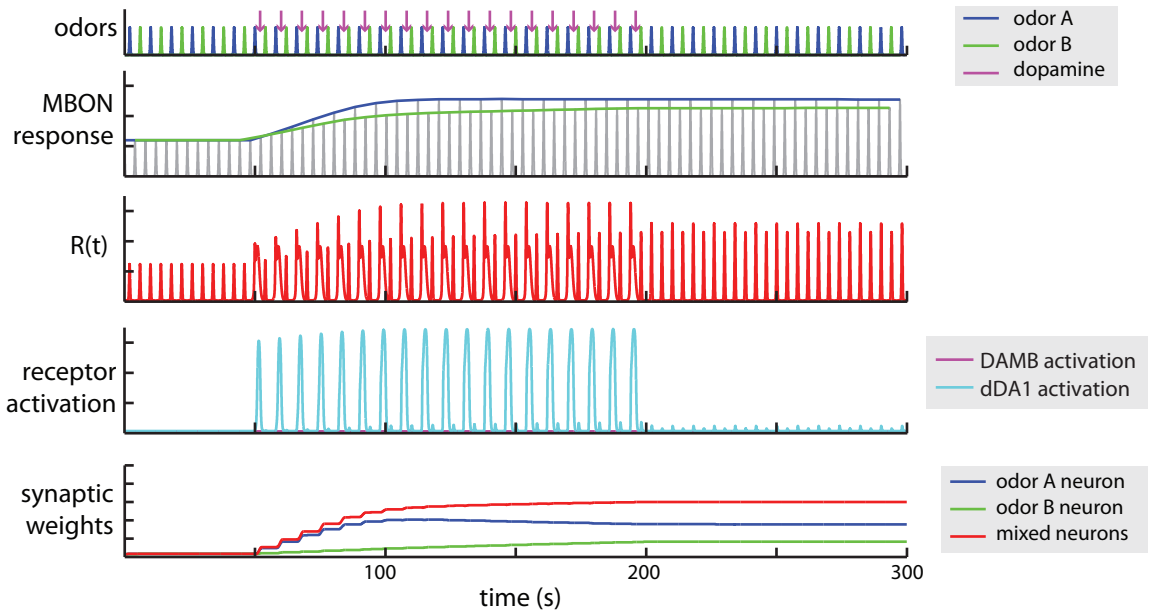


Figure 5.16: DAMB mutant model: the MBON response to the trained odor increases to the fixed point of the dDA1-mediated learning rule. Because a population of neurons in the toy model respond to both odor A and odor B, changing the MBON response to odor A also alters the MBON response to odor B. Without DAMB-mediated adaptation, this model predicts that the MBON will rapidly overgeneralize as was observed in Figure 5.8.

5.4 Using $R(t)$ to encode readout error

In classical models of cerebellar learning, climbing fibers are used to encode error signals [Marr, 1969; Albus, 1971; Ito 1972; Miles and Lisberger 1981]. While this theory has its holes—for example, cerebellar learning can occur even in training conditions that do not modulate climbing fiber activity [Ke, Guo and Raymond 2009]—using $R(t)$ as an error signal also fixes the overgeneralization problem encountered in the original odor valence formulation. As with the modified valence formulation, this approach creates two fixed points for trained and untrained odors.

In the error signal approach, the firing rate of dopaminergic neurons encodes the difference between the odor-evoked MBON response and some target firing rate:

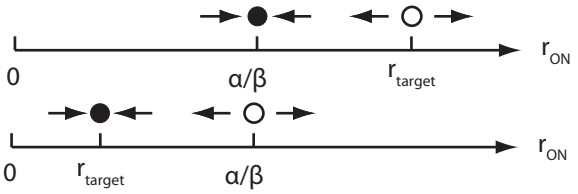
$$R(t) = r_{\text{target}} - r_{\text{ON}}$$

where r_{target} takes one value by default, and switches to a different value when an odor is paired with reward or shock. The full reward-modulated learning rule is thus

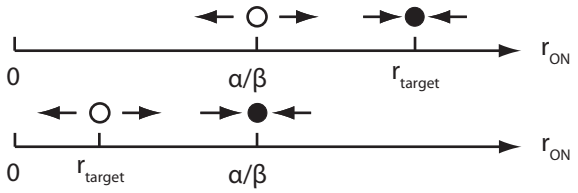
$$\begin{aligned} \tau_s \frac{ds}{dt} &= -s(t) + \eta U L \\ \frac{dw}{dt} &= \big(r_{\text{target}} - r_{\text{ON}}\big) s(t) \end{aligned}$$

which has fixed points $r_{\text{ON}} = \{r_{\text{target}}, \alpha/\beta\}$. Stability of the two fixed points of learning depend on their relative placement, and on the sign of α and β:

1. α > 0 , β > 0



2. α < 0 , β < 0



where filled circles mark stable fixed points, and open circles unstable fixed points. (Learning rules in which the signs of α and β do not match have only the r_{target} fixed point greater than zero, with stability determined by which of α or β is positive.) By having r_{target} be nonzero for untrained (not paired with reward or shock) odors, the MBON response to untrained odors evolves according to the effect of the unsupervised learning rule, either to the fixed point of the learning rule or to zero/infinity. If r_{target} is set as follows, r_{ON} will evolve towards α/β over repeat presentations of an odor; when that odor is paired with a reward, r_{ON} will adjust up/down to r_{target} for the paired odor.

1) $\alpha > 0$ $\beta > 0$	untrained:	$r_{\text{target}} > \alpha/\beta$
	trained:	$r_{\text{target}} < \alpha/\beta$
2) $\alpha < 0$ $\beta < 0$	untrained:	$r_{\text{target}} < \alpha/\beta$
	trained:	$r_{\text{target}} > \alpha/\beta$

5.4.1 Predictions about learning

5.4.1.1 Dopamine signal should go to zero once a conditioned response has been learned

In the error signal model of learning, dopaminergic neurons encode a distance of the MBON response from a target value. Thus for trained odors, the dopamine response will go to zero once the response has been learned (that is, once $r_{\text{ON}} = r_{\text{target}}$). For untrained odors, the dopamine response should not quite reach zero, as learning drives $r_{\text{ON}} \to \frac{\alpha}{\beta}$, and the output neuron never precisely reaches r_{target}. This unusual feature could be avoided by simply setting $r_{\text{target}} = \frac{\alpha}{\beta}$ for untrained odors. This is a clear point of different with the valence model, in which dopamine signals must be maintained in order to preserve the learned MBON response.

5.4.1.2 Blocking dopamine should prevent all changes in synaptic weights

Because all synaptic weight changes are gated by $R(t)$, blocking dopamine (setting $R(t)$ to zero) should block synaptic plasticity, consistent with the results of Berry et al [Berry, Cerbantes-Sandoval, Nicholas, and Davis 2012].

5.4.1.3 Learned changes in MBON responses cannot be bidirectional

As with the valence model, this model imposes some limitations on the fixed points of learning: if $\alpha > 0$, $\beta > 0$, the value r_{target} for learned odors must always be less than the fixed point α/β to ensure its stability; similarly if $\alpha < 0$, $\beta < 0$, r_{target} for learned odors must always be greater than α/β. In either case, changes in MBON firing rate due to learning cannot be bidirectional: an MBON should either always decrease or always increase its response to trained odors.

5.4.1.4 Dopamine must be able to modulate synaptic weights bidirectionally

Error-based learning requires the dopamine signal to convey both positive and negative values. A mechanism for encoding negative values has been described in previous models of cerebellar learning

[Medina et al 2000], in which climbing fibers are tonically active. In these models, climbing fiber activity controls the sign of bidirectional plasticity at the granule cell-Purkinje cell synapse: an increase in climbing fiber firing rates induces LTD at granule cell-Purkinje cell synapses, while a drop in climbing fiber firing rate below baseline induces LTP, thus allowing the climbing fiber signal to flip the sign of changes in synaptic weight [Hirano 1990; Sakurai 1986; Salin, Malenka, and Nicoll 1996].

The climbing fiber error model has often been challenged in the cerebellum literature—not only because it is unclear that climbing fibers control learning [Ke, Guo, and Raymond 2009], but because tonic firing rates of climbing fibers are very low (around 1Hz), making precise encoding of error signals questionable. Similar problems face the mushroom body: if we consider $R(t)$ as dopamine concentration relative to baseline, the baseline response of dopaminergic neurons to odors must still be high enough for negative values of $R(t)$ to be encoded.

The bidirectionality of learning is also key to this formulation: below-baseline dopamine levels must flip the direction of the unsupervised learning rule. The two dopamine receptors DAMB and dDA1 could provide a mechanism for bidirectionality, although their concentration dependence may make error-based learning more difficult to implement.

5.4.1.5 Dopaminergic neurons need a memory trace to compute the error signal

Finally, it is important to note that the error signal is a difference between the MBON response to an odor and the observed reward, two events that can be separated in time by several seconds. For the error signal to be computed, either MBON responses to odors must be sustained, or a memory trace of recent inputs must be stored in the dopaminergic neurons. Recorded responses of dopaminergic neurons and MBONs both seem to be time locked to the odor stimulus [Daisuke Hattori, private communication], but little is known about the dynamics of dopaminergic neuron responses during odor learning.

5.4.1.6 Dopamine changes the shape of the KC-MBON learning rule

This result does not emerge as naturally from the error model as it does from the valence model, but it can be produced. In the error model $R(t)$ encodes $r_{\text{target}} - r_{\text{ON}}$, so an injection of dopamine following odor presentation is equivalent to setting a high value of r_{target}. If the unsupervised learning rule satisfies $\alpha < 0, \beta < 0$, setting $r_{\text{target}} > \alpha/\beta$ flips the sign of the unsupervised learning rule. This might account for the changed in learning rule shape observed experimentally by Cassenaer and Laurent [Cassenaer and Laurent, 2012].

5.5 Discussion

In this section, I studied formation of multiple associative memories in a single MBON and found that, with an appropriate choice of learning rule, many associative memories can be stored. The actual associative memory capacity of the fly is extremely difficult to measure experimentally, as it would require many training and test sessions on large sets of odors. However by assuming a high memory capacity is important and looking at the conditions this imposes on the form of associative learning, I can still make predictions about activity patterns of MBONs and dopaminergic neurons during learning, as well as the effect of mutations in the dopamine receptor machinery.

There are a few other learning alternatives I did not explore in this chapter. First, I have not discussed the role of the giant inhibitory neuron APL, which previous studies have found to be required for associative learning [Pitman et al 2011]. Another similarly large neuron, the serotonergic DPM, is gap-junction coupled with APL and extensively innervates the mushroom body [Tanaka, Tanimoto, and Ito, 2008]. It has also been suggested as a mediator of learning, on its own [Krashes et al, 2007] or in conjunction with APL [Wu et al 2011]. I found that APL-KC plasticity could help to make KC representations of learned odors less variable, by weakening inhibition of KCs responding to learned odors and making them more likely to respond, providing a form of pattern completion. However I have not systematically investigated the memory capacity of this system, or its susceptibility to overgeneralization.

Finally, models of cerebellar learning suggest another alternative site of associative plasticity, in

the lateral horn. As was shown in Figure 5.7, projection neurons extend processes to both the mushroom body and the lateral horn. Similarly in Figure 5.6, mossy fibers project to both granule cells and the deep cerebellar nucleus (DCN). Studies of cerebellar learning have found that mossy fiber synapses with the DCN are also plastic, and that their plasticity is gated by Purkinje cells [Raymond, Lisberger and Mauk 1996; Mauk 1997]; this plasticity has been hypothesized to underly effects such as savings, in which a conditioned response that is trained, extinguished, and then trained again is learned faster the second time [Medina, Garcia and Mauk, 2001]. The interaction of MBONs and projection neurons in the lateral horn has not been studied, but a similar two-stage learning process could also be at work in flies. In this case, the overgeneralization problem I observe in my mushroom body model could be overcome by the gradual re-writing of acquired memories from the mushroom body to the lateral horn. In this case, changes in MBON responses to odors might only last for a short period until a more robust and specific memory can be written to the lateral horn. Further study of the downstream targets of MBONs will be needed for this hypothesis to be investigated experimentally.

Chapter 6

General discussion

Associative learning makes the implicit assumption that an experience from the past provides predictive information about the future that can be applied to the present. This assumption has a physical underpinning: when a sensory stimulus is linked to a salient experience, the readout of a neural population encoding that stimulus is altered via synaptic plasticity, producing a stable, long-lasting change in the stimulus-evoked response of downstream circuitry. For learning to be useful, it must incorporate additional assumptions about how to relate previous experiences to the present context—for example, if odor A has been predictive of reward in the past, what does it mean if odor A is encountered in a mixture with odor B?

With the right neural representation of sensory stimuli, a simple learning rule can produce a learned response that generalizes in a behaviorally useful way. In the passive and active systems of the electric fish, the set of granule cell basis functions available to efferent cells determines the shape of the negative images efferent cells can form. We found that the precise shape of the basis in the passive system is tailored to the family of signals the fish encounters in nature, allowing negative images of these stimuli to form more rapidly. Future experimental work should determine whether similar adaptations exist in the active system, where rapid formation of negative images is important due to the large set of posture-specific sensory effects that must be canceled.

In fly, the mushroom body does not maximally decorrelate odors, but instead preserves similarities of odor representations that emerge in the olfactory receptors (Figure 4.8). Thus a learned response

to odor A is likely to also be evoked by the mixture of odor A and B, due to overlapping odor representations in the mushroom body. Learning rules that drive output neuron responses to separate fixed points for trained and untrained odors can ensure that this useful property is preserved while preventing overgeneralization to unrelated odors.

Cerebellum-like structures are a special case of learning, in which individual readout neurons play identifiable roles in mapping sensory stimuli to behavioral responses. But sensory representations shape the learning capacity of systems in a more general context. Deep belief networks with many layers of nonlinear transformations vastly outperform comparably-sized neural networks with fewer layers [Bengio 2009]. After training on object recognition tasks, units in the uppermost layers of deep networks show tuning for remarkably complex features, such as faces and parts of vehicles [Sermanet et al 2014]. No one layer of deep belief nets is essential for their success; rather, their multilayered structure converts sensory input into a representation of the most relevant features for object recognition, allowing for more robust performance. The visual processing stream works in a similar way, gradually transforming small features of the visual scene into abstract and condition-invariant representations of objects; learned associations that would be impossibly complex in primary visual cortex become far easier to implement using representations in higher areas.

All of the models in this thesis have focused on learning in single neurons, either MG cells, efferent cells, or MBONs. In the previous chapter, I discussed a hypothesis by which plasticity at KC-to-MBON synapses may lead to more lasting and stable modifications of the projection neuron input to the lateral horn. This hypothesis is based on a similar circuit in cerebellum, in which Purkinje cells gate plasticity from mossy fibers to the deep cerebellar nucleus. Likewise, there is a two-stage learning process in the ELL, in which MG1, MG2, LG, and LF cells are all recurrently connected, but only LG and LF cells have processes that leave the electrosensory lobe. Both MG cells and LF/LG cells form negative images from their granule cell inputs, although learning is usually easier to drive in MG cells than LF/LG cells [Nate Sawtell, private communication]. An exciting future challenge will be to use the models established here to look at how learning occurs on the population level, and how multiple sites of plasticity benefit adaptive control of behavior.

Bibliography

[1] LM Aitkin and J Boyd. Acoustic input to the lateral pontine nuclei. *Hearing research*, 1(1):67–77, 1978.

[2] James S Albus. A theory of cerebellar function. *Mathematical Biosciences*, 10(1):25–61, 1971.

[3] Christopher Assad. *Electric field maps and boundary element simulations of electrolocation in weakly electric fish*. PhD thesis, California Institute of Technology, 1997.

[4] David Babineau, André Longtin, and John E Lewis. Modeling the electric field of weakly electric fish. *Journal of experimental biology*, 209(18):3636–3651, 2006.

[5] Clare VH Baker, Melinda S Modrell, and J Andrew Gillis. The evolution and development of vertebrate lateral line electroreceptors. *The Journal of experimental biology*, 216(13):2515–2522, 2013.

[6] C Bell and G von der Emde. Electric organ corollary discharge pathways in mormyrid fish. *Journal of Comparative Physiology A*, 177(4):463–479, 1995.

[7] Curtis C Bell. An efference copy which is modified by reafferent input. *Science*, 214(4519):450–453, 1981.

[8] Curtis C Bell. Properties of a modifiable efference copy in an electric fish. *J Neurophysiol*, 47(6):1043–1056, 1982.

[9] Curtis C Bell, Angel Caputi, Kirsty Grant, and Jacques Serrier. Storage of a sensory pattern by anti-hebbian synaptic plasticity in an electric fish. *Proceedings of the National Academy of Sciences*, 90(10):4650–4654, 1993.

[10] Curtis C Bell, Kirsty Grant, and Jacques Serrier. Sensory processing and corollary discharge effects in the mormyromast regions of the mormyrid electrosensory lobe. i. field potentials, cellular activity in associated structures. *Journal of neurophysiology*, 68:843–843, 1992.

[11] Curtis C Bell, Victor Han, and Nathaniel B Sawtell. Cerebellum-like structures and their implications for cerebellar function. *Annu. Rev. Neurosci.*, 31:1–24, 2008.

[12] Curtis C Bell, Victor Z Han, Yoshiko Sugawara, and Kirsty Grant. Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature*, 387(6630):278–281, 1997.

[13] Curtis C Bell and Charles J Russell. Termination of electroreceptor and mechanical lateral line afferents in the mormyrid acousticolateral area. *Journal of Comparative Neurology*, 182(3):367–382, 1978.

[14] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[15] Jacob A Berry, Isaac Cervantes-Sandoval, Eric P Nicholas, and Ronald L Davis. Dopamine is required for learning and forgetting in¡ i¿ drosophila¡/i¿. *Neuron*, 74(3):530–542, 2012.

[16] Vikas Bhandawat, Shawn R Olsen, Nathan W Gouwens, Michelle L Schlief, and Rachel I Wilson. Sensory processing in the drosophila antennal lobe increases reliability and separability of ensemble odor representations. *Nature neuroscience*, 10(11):1474–1482, 2007.

[17] David Bodznick and John C Montgomery. The physiology of low-frequency electrosensory systems. In *Electroreception*, pages 132–153. Springer, 2005.

[18] David Bodznick, John C Montgomery, and David J Bradley. Suppression of common mode signals within the electrosensory system of the little skate raja erinacea. *Journal of experimental biology*, 171(1):107–125, 1992.

[19] James M Bower. Control of sensory data acquisition. *International review of neurobiology*, 41:489–513, 1997.

[20] Edward S Boyden, Akira Katoh, and Jennifer L Raymond. Cerebellum-dependent learning: the role of multiple plasticity mechanisms. *Neuroscience*, 27, 2004.

[21] Andrea H Brand and Norbert Perrimon. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *development*, 118(2):401–415, 1993.

[22] Christopher J Burke, Wolf Huetteroth, David Owald, Emmanuel Perisse, Michael J Krashes, Gaurav Das, Daryl Gohl, Marion Silies, Sarah Certel, and Scott Waddell. Layered reward signalling through octopamine and dopamine in drosophila. *Nature*, 2012.

[23] Holly R Campbell, Johannes Meek, Jianmei Zhang, and Curtis C Bell. Anatomy of the posterior caudal lobe of the cerebellum and the eminentia granularis posterior in a mormyrid fish. *Journal of Comparative Neurology*, 502(5):714–735, 2007.

[24] Robert AA Campbell, Kyle S Honegger, Hongtao Qin, Wanhe Li, Ebru Demir, and Glenn C Turner. Imaging a population code for odor identity in the drosophila mushroom body. *The Journal of Neuroscience*, 33(25):10568–10581, 2013.

[25] AA Caputi, R Budelli, K Grant, and CC Bell. The electric image in weakly electric fish:

physical images of resistive objects in gnathonemus petersii. *Journal of experimental biology*, 201(14):2115–2128, 1998.

[26] Sophie JC Caron, Vanessa Ruta, LF Abbott, and Richard Axel. Random convergence of olfactory inputs in the drosophila mushroom body. *Nature*, 497(7447):113–117, 2013.

[27] Stijn Cassenaer and Gilles Laurent. Conditional modulation of spike-timing-dependent plasticity for olfactory learning. *Nature*, 482(7383):47–52, 2012.

[28] Ling Chen, Jonathan L House, Rüdiger Krahe, and Mark E Nelson. Modeling signal and background components of electrosensory scenes. *Journal of Comparative Physiology A*, 191(4):331–345, 2005.

[29] Kimberly M Christian and Richard F Thompson. Neural substrates of eyeblink conditioning: acquisition and retention. *Learning & memory*, 10(6):427–455, 2003.

[30] Adam Claridge-Chang, Robert D Roorda, Eleftheria Vrontou, Lucas Sjulson, Haiyan Li, Jay Hirsh, and Gero Miesenböck. Writing memories with light-addressable reinforcement circuitry. *Cell*, 139(2):405–415, 2009.

[31] Shaun P Collin and Darryl Whitehead. The functional roles of passive electroreception in non-electric shes. *Animal Biology*, 54(1):1–25, 2004.

[32] Blender Online Community. *Blender*. Blender Foundation, 2014.

[33] Trinity B Crapse and Marc A Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008.

[34] Sandeep Robert Datta, Maria Luisa Vasconcelos, Vanessa Ruta, Sean Luo, Allan Wong, Ebru Demir, Jorge Flores, Karen Balonze, Barry J Dickson, and Richard Axel. The drosophila pheromone cva activates a sexually dimorphic neural circuit. *Nature*, 452(7186):473–477, 2008.

[35] Jackson John David. Classical electrodynamics, 1975.

[36] Ronald L Davis. Olfactory memory formation in drosophila: from molecular to systems neuroscience. *Annu. Rev. Neurosci.*, 28:275–302, 2005.

[37] J Steven de Belle and Martin Heisenberg. Associative odor learning in drosophila abolished by chemical ablation of mushroom bodies. *Science*, 263(5147):692–695, 1994.

[38] Paul Dean, John Porrill, Carl-Fredrik Ekerot, and Henrik Jörntell. The cerebellar microcircuit as an adaptive filter: experimental and computational evidence. *Nature Reviews Neuroscience*, 11(1):30–43, 2009.

[39] Nina Deisig, Harald Lachnit, Martin Giurfa, and Frank Hellstern. Configural olfactory learn-

ing in honeybees: negative and positive patterning discrimination. *Learning & Memory*, 8(2):70–78, 2001.

[40] Nina Deisig, Harald Lachnit, Jean-Christophe Sandoz, Klaus Lober, and Martin Giurfa. A modified version of the unique cue theory accounts for olfactory compound processing in honeybees. *Learning & Memory*, 10(3):199–208, 2003.

[41] Marco A Diana, Yo Otsu, Gilliane Maton, Thibault Collin, Mireille Chat, and Stéphane Dieudonné. T-type and l-type ca2+ conductances define and encode the bimodal firing pattern of vestibulocerebellar unipolar brush cells. *The Journal of neuroscience*, 27(14):3823–3838, 2007.

[42] Josh Dubnau, Lori Grady, Toshi Kitamoto, and Tim Tully. Disruption of neurotransmission in drosophila mushroom body blocks retrieval but not acquisition of memory. *Nature*, 411(6836):476–480, 2001.

[43] Patrick Dular, Christophe Geuzaine, F Henrotte, and Willy Legros. A general environment for the treatment of discrete problems and its application to the finite element method. *Magnetics, IEEE Transactions on*, 34(5):3395–3398, 1998.

[44] John C Eccles, M Ito, and J Szentagothai. The cerebellum as a neuronal machine, 1967. *New York*, page 272.

[45] Jacob Engelmann, João Bacelo, Michael Metzen, Roland Pusch, Beatrice Bouton, Adriana Migliaro, Angel Caputi, Ruben Budelli, Kirsty Grant, and Gerhard Von Der Emde. Electric imaging through active electrolocation: implication for the analysis of complex scenes. *Biological cybernetics*, 98(6):519–539, 2008.

[46] Jacob Engelmann, E van den Burg, J Bacelo, M de Ruijters, S Kuwana, Y Sugawara, and K Grant. Dendritic backpropagation and synaptic plasticity in the mormyrid electrosensory lobe. *Journal of Physiology-Paris*, 102(4):233–245, 2008.

[47] Jörg-Peter Ewert, Harald Burghagen, and Evelyn Schürg-Pfeiffer. Neuroethological analysis of the innate releasing mechanism for prey-catching behavior in toads. In *Advances in vertebrate neuroethology*, pages 413–475. Springer, 1983.

[48] Sarah M Farris. Are mushroom bodies cerebellum-like structures? *Arthropod structure & development*, 40(4):368–379, 2011.

[49] GS Fraenkel and DL Gunn. L., 1961: The orientation of animals, 1940.

[50] Nicolas Frémaux, Henning Sprekeler, and Wulfram Gerstner. Functional requirements for reward-modulated spike-timing-dependent plasticity. *The Journal of Neuroscience*, 30(40):13326–13337, 2010.

[51] Karl von Frisch. About the color vision of fish and bees. 1919.

[52] Bertram Gerber and Juliane Ullrich. No evidence for olfactory blocking in honeybee classical conditioning. *Journal of experimental biology*, 202(13):1839–1854, 1999.

[53] Christophe Geuzaine and Jean-François Remacle. Gmsh: A 3-d finite element mesh generator with built-in pre-and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[54] Martin Giurfa and Jean-Christophe Sandoz. Invertebrate learning and memory: fifty years of olfactory conditioning of the proboscis extension response in honeybees. *Learning & Memory*, 19(2):54–66, 2012.

[55] I Gormezano, Neil Schneiderman, Edward Deaux, and Isreal Fuentes. Nictitating membrane: classical conditioning and extinction in the albino rabbit. *Science*, 138(3536):33–34, 1962.

[56] Fernando Guerrieri, Harald Lachnit, Bertram Gerber, and Martin Giurfa. Olfactory blocking and odorant similarity in the honeybee. *Learning & Memory*, 12(2):86–95, 2005b.

[57] Fernando Guerrieri, Marco Schubert, Jean-Christophe Sandoz, and Martin Giurfa. Perceptual and neural olfactory similarity in honeybees. *PLoS biology*, 3(4):e60, 2005a.

[58] Elissa A Hallem and John R Carlson. Coding of odors by a receptor repertoire. *Cell*, 125(1):143–160, 2006.

[59] Kyung-An Han, Neil S Millar, Michael S Grotewiel, and Ronald L Davis. Damb, a novel dopamine receptor expressed specifically in drosophila mushroom bodies. *Neuron*, 16(6):1127–1135, 1996.

[60] Martin Heisenberg. Mushroom body memoir: from maps to models. *Nature Reviews Neuroscience*, 4(4):266–275, 2003.

[61] T Hirano. Depression and potentiation of the synaptic transmission between a granule cell and a purkinje cell in rat cerebellar culture. *Neuroscience letters*, 119(2):141–144, 1990.

[62] Kyle S Honegger, Robert AA Campbell, and Glenn C Turner. Cellular-resolution population imaging reveals robust sparse coding in the drosophila mushroom body. *The Journal of Neuroscience*, 31(33):11772–11785, 2011.

[63] Sung-Tae Hong, Sunhoe Bang, Seogang Hyun, Jongkyun Kang, Kyunghwa Jeong, Donggi Paik, Jongkyeong Chung, and Jaeseob Kim. camp signalling in mushroom bodies modulates temperature preference behaviour in drosophila. *Nature*, 454(7205):771–775, 2008.

[64] Carl D Hopkins. Evolution of electric communication channels of mormyrids. *Behavioral Ecology and Sociobiology*, 7(1):1–13, 1980.

[65] CD Hopkins, K-T Shieh, DW McBride Jr, and M Winslow. A quantitative analysis of passive

electrolocation behavior in electric fish; pp. 45–59. *Brain, behavior and evolution*, 50(Suppl. 1):45–59, 1997.

[66] Masao Ito. Neural design of the cerebellar motor control system. *Brain research*, 40(1):81–84, 1972.

[67] Vladimir Itskov and LF Abbott. Pattern capacity of a perceptron for sparse discrimination. *Physical review letters*, 101(1):018101, 2008.

[68] Richard B Ivry. The representation of temporal information in perception and motor control. *Current opinion in neurobiology*, 6(6):851–857, 1996.

[69] Gregory SXE Jefferis, Christopher J Potter, Alexander M Chan, Elizabeth C Marin, Torsten Rohlfing, Calvin R Maurer Jr, and Liqun Luo. Comprehensive maps of¡ i¿ drosophila¡/i¿ higher olfactory centers: Spatially segregated fruit and pheromone representation. *Cell*, 128(6):1187–1203, 2007.

[70] John Allen Jellies. Associative olfactory conditioning in drosophila melanogaster and memory retention through metamorphosis. *Illinois State University, Normal, IL*, 1981.

[71] Arnim Jenett, Gerald M Rubin, Teri-TB Ngo, David Shepherd, Christine Murphy, Heather Dionne, Barret D Pfeiffer, Amanda Cavallaro, Donald Hall, Jennifer Jeter, et al. A gal4-driver line resource for¡ i¿ drosophila¡/i¿ neurobiology. *Cell reports*, 2(4):991–1001, 2012.

[72] Ad J Kalmijn. The detection of electric fields from inanimate and animate sources other than electric organs. In *Electroreceptors and Other Specialized Receptors in Lower Vertrebrates*, pages 147–200. Springer, 1974.

[73] Masashi Kawasaki. Evolution of time-coding systems in weakly electric fishes. *Zoological science*, 26(9):587–599, 2009.

[74] Michael C Ke, Cong C Guo, and Jennifer L Raymond. Elimination of climbing fiber instructive signals during motor learning. *Nature neuroscience*, 12(9):1171–1179, 2009.

[75] Alex C Keene, Markus Stratmann, Andreas Keller, Paola N Perrat, Leslie B Vosshall, and Scott Waddell. Diverse odor-conditioned memories require uniquely timed dorsal paired medial neuron output. *Neuron*, 44(3):521–533, 2004.

[76] Young-Cho Kim, Hyun-Gwan Lee, and Kyung-An Han. D1 dopamine receptor dda1 is required in the mushroom body neurons for aversive and appetitive learning in drosophila. *The Journal of neuroscience*, 27(29):7640–7647, 2007.

[77] Toshihiro Kitamoto. Conditional modification of behavior in drosophila by targeted expression of a temperature-sensitive shibire allele in defined neurons. *Journal of neurobiology*, 47(2):81–92, 2001.

[78] Michael J Krashes, Shamik DasGupta, Andrew Vreede, Benjamin White, J Douglas Armstrong, and Scott Waddell. A neural circuit mechanism integrating motivational state with memory expression in¡ i¿ drosophila¡/i¿. *Cell*, 139(2):416–427, 2009.

[79] Michael J Krashes, Alex C Keene, Benjamin Leung, J Douglas Armstrong, and Scott Waddell. Sequential use of mushroom body neuron subsets during¡ i¿ drosophila¡/i¿ odor memory processing. *Neuron*, 53(1):103–115, 2007.

[80] Gilles Laurent and MOHAMMAD Naraghi. Odorant-induced oscillations in the mushroom bodies of the locust. *The Journal of neuroscience*, 14(5):2993–3004, 1994.

[81] HW Lissmann and KE Machin. The mechanism of object location in gymnarchus niloticus and similar fish. *Journal of Experimental Biology*, 35(2):451–486, 1958.

[82] Chang Liu, Pierre-Yves Plaçais, Nobuhiro Yamagata, Barret D Pfeiffer, Yoshinori Aso, Anja B Friedrich, Igor Siwanowicz, Gerald M Rubin, Thomas Preat, and Hiromu Tanimoto. A subset of dopamine neurons signals reward for odour memory in drosophila. *Nature*, 2012.

[83] Rodolfo Llinás and John P Welsh. On the cerebellum and motor learning. *Current opinion in neurobiology*, 3(6):958–965, 1993.

[84] Francesca Locatelli, Luisa Bottà, Francesca Prestori, Sergio Masetto, and Egidio D'Angelo. Late-onset bursts evoked by mossy fibre bundle stimulation in unipolar brush cells: evidence for the involvement of h-and trp-currents. *The Journal of physiology*, 591(4):899–918, 2013.

[85] Sean X Luo, Richard Axel, and LF Abbott. Generating sparse and selective third-order responses in the olfactory system of the fly. *Proceedings of the National Academy of Sciences*, 107(23):10713–10718, 2010.

[86] David Marr. A theory of cerebellar cortex. *The Journal of physiology*, 202(2):437–470, 1969.

[87] Michael D Mauk. Roles of cerebellar cortex and nuclei in motor learning: contradictions or clues? *Neuron*, 18(3):343–346, 1997.

[88] Michael D Mauk, Joseph E Steinmetz, and Richard F Thompson. Classical conditioning using stimulation of the inferior olive as the unconditioned stimulus. *Proceedings of the National Academy of Sciences*, 83(14):5349–5353, 1986.

[89] Sean MJ McBride, Giovanna Giuliani, Catherine Choi, Paul Krause, Dana Correale, Karli Watson, Glenn Baker, and Kathleen K Siwicki. Mushroom body ablation impairs short-term memory and long-term memory of courtship conditioning in¡ i¿ drosophila melanogaster¡/i¿. *Neuron*, 24(4):967–977, 1999.

[90] David A McCormick, Gregory A Clark, David G Lavond, and Richard F Thompson. Initial localization of the memory trace for a basic form of learning. *Proceedings of the National Academy of Sciences*, 79(8):2731–2735, 1982.

[91] David A McCormick, Joseph E Steinmetz, and Richard F Thompson. Lesions of the inferior olivary complex cause extinction of the classically conditioned eyeblink response. *Brain research*, 359(1):120–130, 1985.

[92] David A McCormick and Richard F Thompson. Cerebellum: essential involvement in the classically conditioned eyelid response. *Science*, 223(4633):296–299, 1984.

[93] Sean E McGuire, Phuong T Le, and Ronald L Davis. The role of drosophila mushroom body signaling in olfactory memory. *Science*, 293(5533):1330–1333, 2001.

[94] Javier F Medina, Keith S Garcia, and Michael D Mauk. A mechanism for savings in the cerebellum. *The Journal of Neuroscience*, 21(11):4081–4089, 2001.

[95] Javier F Medina, Keith S Garcia, William L Nores, Nichole M Taylor, and Michael D Mauk. Timing mechanisms in the cerebellum: testing predictions of a large-scale computer simulation. *The Journal of Neuroscience*, 20(14):5516–5525, 2000.

[96] Javier F Medina and Michael D Mauk. Computer simulation of cerebellar information processing. *nature neuroscience*, 3:1205–1211, 2000.

[97] Javier F Medina, William L Nores, Tatsuya Ohyama, and Michael D Mauk. Mechanisms of cerebellar learning suggested by eyelid conditioning. *Current opinion in neurobiology*, 10(6):717–724, 2000.

[98] J Meek, K Grant, and C Bell. Structural organization of the mormyrid electrosensory lateral line lobe. *Journal of experimental biology*, 202(10):1291–1300, 1999.

[99] J Meek, Kirsty Grant, Y Sugawara, TGM Hafmans, M Veron, and JP Denizot. Interneurons of the ganglionic layer in the mormyrid electrosensory lateral line lobe: morphology, immunohistochemistry, and synaptology. *Journal of Comparative Neurology*, 375(1):43–65, 1996.

[100] Randolf Menzel, J Erber, and Th Masuhr. Learning and memory in the honeybee. In *Experimental analysis of insect behaviour*, pages 195–217. Springer, 1974.

[101] FA Miles and SG Lisberger. Plasticity in the vestibulo-ocular reflex: a new hypothesis. *Annual review of neuroscience*, 4(1):273–299, 1981.

[102] Makoto Mizunami, Masayuki Iwasaki, Michiko Nishikawa, and Ryuichi Okada. Modular structures in the mushroom body of the cockroach. *Neuroscience letters*, 229(3):153–156, 1997.

[103] Makoto Mizunami, Josette M Weibrecht, and Nicholas J Strausfeld. Mushroom bodies of the cockroach: their participation in place memory. *Journal of Comparative Neurology*, 402(4):520–537, 1998.

[104] Enrico Mugnaini, Gabriella Sekerková, and Marco Martina. The unipolar brush cell: a remarkable neuron finally receiving deserved attention. *Brain research reviews*, 66(1):220–245, 2011.

[105] Uli Müller. Learning in honeybees: from molecules to behaviour. *Zoology*, 105(4):313–320, 2002.

[106] Mala Murthy, Ila Fiete, and Gilles Laurent. Testing odor response stereotypy in the drosophila mushroom body. *Neuron*, 59(6):1009–1023, 2008.

[107] Donata Oertel and Eric D Young. What's a cerebellar circuit doing in the auditory system? *Trends in neurosciences*, 27(2):104–110, 2004.

[108] Tatsuya Ohyama, William L Nores, Matthew Murphy, and Michael D Mauk. What the cerebellum computes. *Trends in neurosciences*, 26(4):222–227, 2003.

[109] Shawn R Olsen, Vikas Bhandawat, and Rachel I Wilson. Excitatory interactions between olfactory processing channels in the drosophila antennal lobe. *Neuron*, 54(1):89–103, 2007.

[110] Shawn R Olsen, Vikas Bhandawat, and Rachel Irene Wilson. Divisive normalization in olfactory population codes. *Neuron*, 66(2):287, 2010.

[111] Maria Papadopoulou, Stijn Cassenaer, Thomas Nowotny, and Gilles Laurent. Normalization for sparse encoding of odors by a wide-field interneuron. *Science*, 332(6030):721–725, 2011.

[112] John M Pearce. A model for stimulus generalization in pavlovian conditioning. *Psychological review*, 94(1):61, 1987.

[113] Javier Perez-Orive, Ofer Mazor, Glenn C Turner, Stijn Cassenaer, Rachel I Wilson, and Gilles Laurent. Oscillations and sparsening of odor representations in the mushroom body. *Science*, 297(5580):359–365, 2002.

[114] Barret D Pfeiffer, Teri-T B Ngo, Karen L Hibbard, Christine Murphy, Arnim Jenett, James W Truman, and Gerald M Rubin. Refinement of tools for targeted gene expression in drosophila. *Genetics*, 186(2):735–755, 2010.

[115] Jena L Pitman, Wolf Huetteroth, Christopher J Burke, Michael J Krashes, Sen-Lin Lai, Tzumin Lee, and Scott Waddell. A pair of inhibitory neurons are required to sustain labile memory in the¡ i¿ drosophila¡/i¿ mushroom body. *Current Biology*, 21(10):855–861, 2011.

[116] Tim Requarth and Nathaniel B Sawtell. Plastic corollary discharge predicts sensory consequences of movements in a cerebellum-like circuit. *Neuron*, 82(4):896–907, 2014.

[117] Robert A Rescorla, Allan R Wagner, et al. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99, 1972.

[118] Patrick D Roberts and Curtis C Bell. Computational consequences of temporally asymmetric learning rules: Ii. sensory image cancellation. *Journal of computational neuroscience*, 9(1):67–83, 2000.

[119] Edmund T Rolls and Martin J Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2):713–726, 1995.

[120] David J Rossi, Simon Alford, Enrico Mugnaini, and N Traverse Slater. Properties of transmission at a giant glutamatergic synapse in cerebellum: the mossy fiber-unipolar brush cell synapse. *Journal of neurophysiology*, 74:24–24, 1995.

[121] Charly V Rousseau, Guillaume P Dugué, Andréa Dumoulin, Enrico Mugnaini, Stéphane Dieudonné, and Marco A Diana. Mixed inhibitory synaptic balance correlates with glutamatergic synaptic phenotype in cerebellar unipolar brush cells. *The Journal of Neuroscience*, 32(13):4632–4644, 2012.

[122] Masaki Sakurai. Synaptic modification of parallel fibre-purkinje cell transmission in in vitro guinea-pig cerebellar slices. *The Journal of Physiology*, 394(1):463–480, 1987.

[123] Paul A Salin, Robert C Malenka, and Roger A Nicoll. Cyclic amp mediates a presynaptic form of ltp at cerebellar parallel fiber synapses. *Neuron*, 16(4):797–803, 1996.

[124] Nathaniel B Sawtell, Claudia Mohr, and Curtis C Bell. Recurrent feedback in the mormyrid electrosensory system: cells of the preeminential and lateral toral nuclei. *Journal of neurophysiology*, 93(4):2090–2103, 2005.

[125] Nathaniel B Sawtell and Alan Williams. Transformations of electrosensory encoding associated with an adaptive filter. *The Journal of Neuroscience*, 28(7):1598–1612, 2008.

[126] Eric J Schwartz, Jason S Rothman, Guillaume P Dugué, Marco Diana, Charly Rousseau, R Angus Silver, and Stéphane Dieudonné. Nmda receptors with incomplete mg2+ block enable low-frequency transmission through the cerebellar cortex. *The Journal of Neuroscience*, 32(20):6878–6893, 2012.

[127] Lonnie L Sears and Joseph E Steinmetz. Dorsal accessory inferior olive activity diminishes during acquisition of the rabbit classically conditioned eyelid response. *Brain research*, 545(1):114–122, 1991.

[128] Julien Séjourné, Pierre-Yves Plaçais, Yoshinori Aso, Igor Siwanowicz, Séverine Trannoy, Vladimiros Thoma, Stevanus R Tedjakumala, Gerald M Rubin, Paul Tchénio, Kei Ito, et al. Mushroom body efferent neurons responsible for aversive olfactory memory retrieval in drosophila. *Nature neuroscience*, 14(7):903–910, 2011.

[129] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun.

Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[130] Yuhua Shang, Adam Claridge-Chang, Lucas Sjulson, Marc Pypaert, and Gero Miesenböck. Excitatory local circuits and their implications for olfactory processing in the fly antennal lobe. *Cell*, 128(3):601–612, 2007.

[131] Brian H Smith and Susan Cobey. The olfactory memory of the honeybee apis mellifera. ii. blocking between odorants in binary mixtures. *Journal of Experimental Biology*, 195(1):91–108, 1994.

[132] Roger W Sperry. Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of comparative and physiological psychology*, 43(6):482, 1950.

[133] Joseph E Steinmetz, David G Lavond, and Richard F Thompson. Classical conditioning of the rabbit eyelid response with mossy fiber stimulation as the conditioned stimulus. *Bulletin of the Psychonomic Society*, 23(3):245–248, 1985.

[134] K Takeda. Classical conditioned response in the honey bee. *Journal of Insect Physiology*, 6(3):168–179, 1961.

[135] Nobuaki K Tanaka, Hiromu Tanimoto, and Kei Ito. Neuronal assemblies of the drosophila mushroom body. *Journal of Comparative Neurology*, 508(5):711–755, 2008.

[136] Tim Tully and William G Quinn. Classical conditioning and retention in normal and mutantdrosophila melanogaster. *Journal of Comparative Physiology A*, 157(2):263–277, 1985.

[137] Glenn C Turner, Maxim Bazhenov, and Gilles Laurent. Olfactory representations by drosophila mushroom body neurons. *Journal of neurophysiology*, 99(2):734–746, 2008.

[138] Toby Tyrrell and David Willshaw. Cerebellar cortex: its simulation and the relevance of marr's theory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1277):239–257, 1992.

[139] Gerhard von der Emde. Discrimination of objects through electrolocation in the weakly electric fish, gnathonemus petersii. *Journal of Comparative Physiology A*, 167(3):413–421, 1990.

[140] Gerhard Von der Emde. Active electrolocation of objects in weakly electric fish. *Journal of experimental biology*, 202(10):1205–1215, 1999.

[141] Gerhard von der Emde. Non-visual environmental imaging and object detection through active electrolocation in weakly electric fish. *Journal of Comparative Physiology A*, 192(6):601–612, 2006.

[142] Gerhard von der Emde, Monique Amey, Jacob Engelmann, Steffen Fetz, Caroline Folde,

Michael Hollmann, Michael Metzen, and Roland Pusch. Active electrolocation in gnathonemus petersii: Behaviour, sensory performance, and receptor systems. *Journal of Physiology-Paris*, 102(46):279 – 290, 2008. Electrosensory Systems.

[143] Gerhard von der Emde and Curtis C Bell. Nucleus preeminentialis of mormyrid fish, a center electrosensory feedback. i. electrosensory and coro discharge responses. *Nucleus*, 76(3), 1996.

[144] Gerhard Von Der Emde, Stephan Schwarz, Leonel Gomez, Ruben Budelli, and Kirsty Grant. Electric fish measure distance in the dark. *Nature*, 395(6705):890–894, 1998.

[145] E Von Holst and H Mittelstaedt. The principle of reafference. *Naturwissenschaften*, 37:464–476, 1950.

[146] Scott Waddell. Reinforcement signalling in¡ i¿ drosophila¡/i¿; dopamine does it all after all. *Current opinion in neurobiology*, 23(3):324–329, 2013.

[147] Alanna J Watt, Hermann Cuntz, Masahiro Mori, Zoltan Nusser, P Jesper Sjöström, and Michael Häusser. Traveling waves in developing cerebellar cortex mediated by asymmetrical purkinje cell connectivity. *Nature neuroscience*, 12(4):463–473, 2009.

[148] Lon A Wilkens and Michael H Hofmann. Behavior of animals with passive, low-frequency electrosensory systems. In *Electroreception*, pages 229–263. Springer, 2005.

[149] Alan Williams, Patrick D Roberts, and Todd K Leen. Stability of negative-image equilibria in spike-timing-dependent plasticity. *Physical Review E*, 68(2):021923, 2003.

[150] Benjamin Willmore and David J Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255–270, 2001.

[151] Rachel I Wilson and Gilles Laurent. Role of gabaergic inhibition in shaping odor-evoked spatiotemporal patterns in the drosophila antennal lobe. *The Journal of neuroscience*, 25(40):9069–9079, 2005.

[152] Chia-Lin Wu, Meng-Fu Maxwell Shih, Jason Sih-Yu Lai, Hsun-Ti Yang, Glenn C Turner, Linyi Chen, and Ann-Shyn Chiang. Heterotypic gap junctions between two neurons in the¡ i¿ drosophila¡/i¿ brain are critical for memory. *Current Biology*, 21(10):848–854, 2011.

[153] Qi Wu, Tieqiao Wen, Gyunghee Lee, Jae H Park, Haini N Cai, and Ping Shen. Developmental control of foraging and social behavior by the¡ i¿ drosophila¡/i¿ neuropeptide y-like system. *Neuron*, 39(1):147–161, 2003.

[154] Zheng Wu, Anita E Autry, Joseph F Bergan, Mitsuko Watabe-Uchida, and Catherine G Dulac. Galanin neurons in the medial preoptic area govern parental behaviour. *Nature*, 509(7500):325–330, 2014.

[155] Melis Yilmaz and Markus Meister. Rapid innate defensive responses of mice to looming visual stimuli. *Current Biology*, 23(20):2011–2015, 2013.