Analysis of Coding Principles in the Olfactory System and their Application in Cheminformatics

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

Vorgelegt beim Fachbereich 14 Biochemie, Chemie und Pharmazie

der Johann Wolfgang Goethe–Universität

in Frankfurt am Main

von

Michael Schmuker

aus Biberach an der Riß

Frankfurt 2007

vom Fachbereich 14 Biochemie, Chemie und Pharmazie der der Johann Wolfgang Goethe–Universität als Dissertation angenommen.

Dekan: Prof. Dr. Harald Schwalbe

Gutachter: Prof. Dr. Gisbert Schneider, Prof. Dr. Paul Wrede

Datum der Disputation: noch nicht bekannt

Erklärung

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung unterzogen habe.

Berlin, den 5. März 2007

Michael Schmuker

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt, dass ich die vorgelegte Dissertation über

Analysis of Coding Principles in the Olfactory System and their Application in Cheminformatics

selbständig angefertigt und mich anderer Hilfsmittel als der in der in ihr angegebenen nicht bedient habe, insbesondere, dass aus Schriften Entlehnungen, soweit sie in der Dissertation nicht ausdrücklich als solche mit Angabe der betreffenden Schrift bezeichnet sind, nicht stattgefunden haben.

Berlin, den 5. März 2007

Michael Schmuker

"All models are wrong, but some models are useful."

– George E. P. Box (1979)

Contents

Chapter 1

Introduction

When it comes to the analysis of sensory information, our own senses are still unmatched by most computational implementations. Moreover, it has turned out that engineers found similar solutions for efficient encoding of stimuli as they appear to be built into our brains.

For example, the wavelet-like encoding of visual information by the retina and subsequent visual processing areas, which has its counterpart in various image compression algorithms (Mallat, 1989). Another example is compression of audio information: The basilar membrane in the cochlea (the inner ear) is excited by different stimulus frequencies at different places, where similar frequencies excite nearby parts of the membrane. This phenomenon is called tonotopy, because of the topological projection of tones of different frequency (Nicholls et al., 2001). Hair cells in different parts of the basilar membrane thus respond to different audio frequencies, effectively providing a frequency decomposition of the original signal. Notably, analyses of the basilar membrane's coding characteristics have led to improvements in audio coding (Baumgarte, 2002).

The olfactory sense provides our perception of the chemical world. Through the course of evolution, its mechanisms to deal with complex chemical stimuli are likely to have evolved to cope optimally with this task. The analysis of

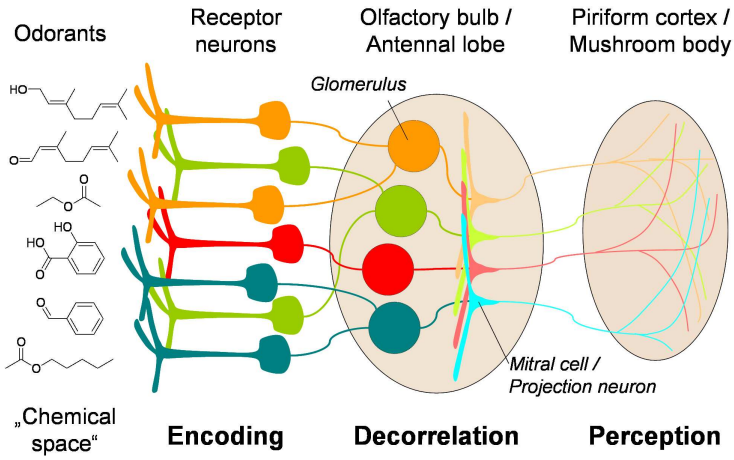

Figure 1.1: Overview of the architecture of olfactory systems (simplified, after Firestein (2001)).

this system promises to yield insight into efficient algorithms to encode and process chemical data, a task that is at the heart of cheminformatics.

1.1 Anatomy of the olfactory system

In order to understand the function of the olfactory system, it is essential to know its anatomy. This section can only serve as a "crash course" to olfaction, providing just enough information which is necessary in order to understand the scientific work we present here. More specific information is available in the original publications cited below.

One striking aspect of olfactory systems is its similar organization in a wide range of species (Hildebrand and Shepherd, 1997; Firestein, 2001). For example, the basic architecture is very similar in insects and in mammals. Figure 1.1 depicts this architecture.

The input is formed by the entirety of odorants ("chemical space" in Figure 1.1). Olfactory receptor neurons (ORNs) encode odorants to neural signals, forming the first stage of olfactory perception. The number of functional genes for olfactory receptors (ORs) has been estimated to about 60 in *Drosophila* (Vosshall, 2000), about 350 in humans (Glusman et al., 2001; Zozulya et al.,

2001), about 1000 in mice (Zhang and Firestein, 2002) and about 1200 in dogs (Olender et al., 2004). In either species, each ORN carries mostly one genotype of OR (depicted by neurons of different color in Figure 1.1), although exceptions to this rule exist (Mombaerts, 2004; Goldman et al., 2005). The regulation of this expression profile has recently been described in mice by Lomvardas et al. (2006).

The second stage in olfactory perception is embodied by the antennal lobe (in insects) resp. the olfactory bulb (in vertebrates). Axons of olfactory receptor neurons project onto so-called *glomeruli* in this structure. These glomeruli are sites of high synaptic connectivity between ORN axons and secondary neurons that project to higher processing areas. These secondary neurons are called "mitral cells" in mammals, and "projection neurons" in insects.

The pronounced connections between the secondary neurons via inhibitory interneurons inspired various hypotheses on the computational properties of this structure (see Cleland and Linster (2005) for a review). They have in common that it is involved in some form of *decorrelation* of the input.

Notably, a chemotopic arrangement of the glomeruli has been observed in vertebrates (Friedrich and Korsching, 1997; Uchida et al., 2000; Meister and Bonhoeffer, 2001) and insects (Sachse et al., 1999; Couto et al., 2005), in that similar odorants often activate neighboring glomeruli.

The axons of the secondary neurons finally project into the piriform cortex in mammals, and the mushroom body in insects. In both, these areas integrate inputs from a variety of sensory modalities (Heisenberg, 1998; Roesch et al., 2007), forming the ideal substrate for associative perception of scent.

1.2 Scope of this thesis

The conserved architecture of olfactory systems may indicate an optimum for processing chemical information. Understanding the organization and information processing concepts in the olfactory system promises to unveil effective

ways to encode and process chemical information. In this thesis, we aimed towards a better understanding of this system through modeling parts of the olfactory machinery, pursuing a highly interdisciplinary approach that connects chemistry with neurobiology and machine learning.

Our first goal was to describe the coding properties of ORNs, in terms of their preferred ligand characteristics. Assuming that activation of olfactory receptors (and in consequence the activation of ORNs) is the result of ligand-protein-interactions, it should depend on the molecular features of an odorant, and thus be predictable from the odorant's chemical structure. Based on this assumption, we derived Structure-Activity-Relationships for ORNs using vectorial descriptors of physicochemical molecular properties, and trained Artificial Neural Networks to predict ORN activation. We evaluated prediction accuracy through testing a novel set of odorants for ORN activation and comparing the results to our predictions. The outcome is presented in chapter 2.

The chemotopic arrangement of glomeruli on the secondary structure in the olfactory system provides insight in how chemical similarity is defined in nature. Hence, as a second goal for this thesis we wanted to investigate chemotopy in the insect antennal lobe. We aimed towards deriving regular projections of this three-dimensional structures on a regular grid to facilitate a systematic analysis. Self-Organizing Maps (SOMs) are particularly useful for this purpose, since they conserve local topology in the input space. In chapter 3, we describe SOMMER, a software that we developed to train and visualize SOMs with a variety of two- and three-dimensional topologies. Moreover, we show projections of *Drosophila*'s antennal lobes onto regular grids of different topologies, and demonstrate how these regular projections enable new ways to explore the antennal lobe's chemotopic organization.

Finally, our last goal was to investigate whether the coding and processing principles in the olfactory system can be applied to chemical information in general. To achieve this, we designed a simplified computational model that incorporates processing schemes which have been observed in the olfactory

system. This model allowed us to analyze the impact of olfactory processing strategies on retrospective screening of an odorant database and a collection of pharmaceutical compounds. The outcome of this analysis in presented in chapter 4.

Chapter 2

Functional characterization of olfactory receptors

2.1 Background

Olfactory Receptors (ORs) encode chemical stimuli in neuronal activity. The gene family of ORs consists of G-protein coupled receptors (GPCRs) and was first described for rats (Buck and Axel, 1991). In *Drosophila*, the organism we considered in this study, as well as in mammals and vertebrates in general, each Olfactory Receptor Neuron (ORN) carries one type of OR (Vosshall et al., 2000), such that the response of each ORN to a chemical substance is mainly determined by the receptor it expresses (Hallem et al., 2004).

The fact that there is no crystal structure available for any OR hampers structure-based approaches such as automated molecular docking to examine ligand binding characteristics. Although attempts have been made to use models based on homology to rhodopsin (Vaidehi et al., 2002; Floriano et al., 2004; Hall et al., 2004), these approaches suffered from the cumbersome creation of such a model and the remaining errors inherent to homology modeling (Becker et al., 2003; Kairys et al., 2006).

Araneda et al. (2000) pursued a ligand-based approach to characterize the rat's I7 OR. By testing a large number of ligands, they were able to establish a verbal characterization of preferred I7 ligands in terms of functional group, carbon chain length and rigidity. However, such an approach only provides qualitative data for a limited number of odorants. It does not describe ORN tuning in quantifiable parameters that can be determined for any chemical.

Here we present a method providing an objective way of predicting ORN responses to arbitrary odorants. We have developed a model that uses a distinct set of physicochemical parameters to describe the structure of odor molecules and predict their activity at *Drosophila* receptors.

We followed a classic approach to derive Structure-Activity-Relationships (SARs) by calculating molecular descriptors and training Artificial Neural Networks (ANNs), as it has been applied in other studies to characterize ligand affinity to specific receptors (Manallack et al., 1994; Schneider and Wrede, 1998; Winkler and Burden, 2002). Similar approaches were previously applied to model human psychophysical data, that is, odor and aroma characteristics (Tsantili-Kakoulidou and Kier, 1992; de Mello Castanho Amboni et al., 2000; Wailzer et al., 2001; Lavine et al., 2003). However, odor percepts are the result of a nonlinear transformation of ORN inputs in the brain and do not necessarily reflect OR properties (Sell, 2006). By contrast, we restricted our study to modeling receptor responses, because these are more likely to be dominated by physicochemical properties of the odorants, assuming OR activation is the result of ligand-receptor binding through intermolecular interactions.

In addition, we suggest that quantifying the molecular properties relevant for activating olfactory receptors reveals how chemical space is encoded by the receptor repertoire of a specific organism. One may assume that such an array of ORs has evolved to provide a useful representation of chemical space through an efficient coding scheme. Determining the actual properties of the chemical world that are detected by ORs may thus provide an efficient way to represent molecules in a computational framework in general.

2.2 Methods and data

2.2.1 Source data: odorants and ORN responses

We used the responses of *Drosophila* ORNs to 47 odorants that were measured through electrophysiological *in vivo* recordings by de Bruyne et al. (2001). These 47 odorants are depicted in Figure 2.1. Their names and the activity values (in spikes/s) are given in Table 2.1.

We prepared a database containing the molecular structures of each of those odorants and their activity (in spikes/s) on the neurons of the classes ab1D, ab2A, ab2B, ab3A, ab3B, ab5B and ab6A. The responses of these classes correspond to those of the OR10a, OR59b, OR85a, OR22a, OR85b, and OR47a receptors respectively (Hallem et al., 2004). No receptor has been identified yet for ab6A.

We chose these ORNs because interpretation of the response spectrum was not complicated by high responses to the solvent, and at least four molecules were active for these ORNs. This yielded a minimum ratio of active to inactive molecules of roughly 1 to 10, and allowed splitting of the data into a training and a validation set of the same size, with at least two instances of active molecules in each set (cf. section 2.2.4 “Neural Network Training”).

2.2.2 Definition of activity ranges

Compound activity was assessed by the magnitude of the neuronal response to odorant-enriched air, evaluated as the increase in action potential firing rate during a 500 ms stimulation with a 10-fold dilution of the headspace over the odorant diluted 1% in paraffin oil as described by de Bruyne et al. (2001). We transformed the continuous range of activity levels into all-or-none data by setting a lower and an upper threshold for each ORN. Molecules with activities below the lower threshold were considered inactive, while those with activities above the upper threshold were considered active. Active odorants are set in bold in the respective column in Table 2.1.

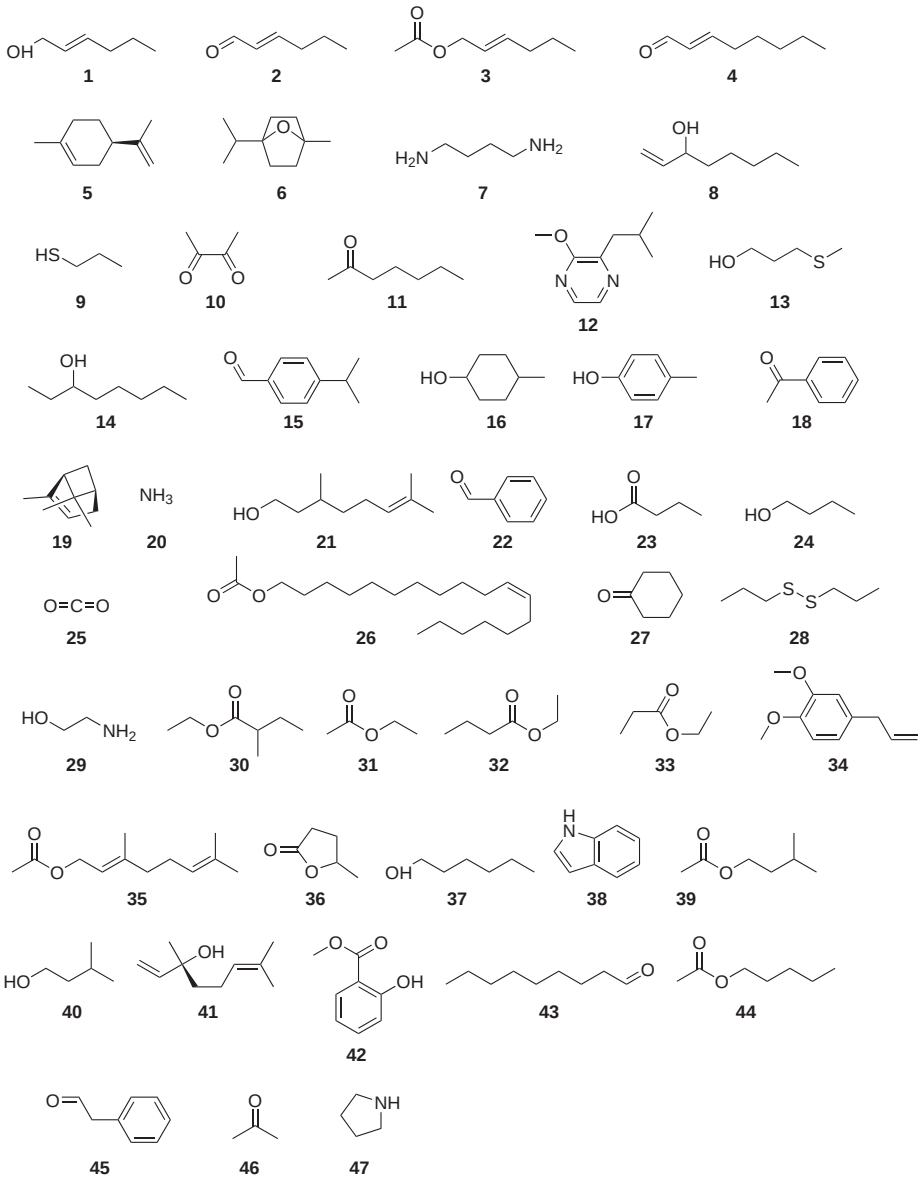

Figure 2.1: Odorant molecules tested by de Bruyne et al. (2001). Compound names are given in Table 2.1.

Table 2.1: Activity values (in spikes/s) and per-ORN thresholds. Spike rates set in bold indicate "active" odorants for the respective ORN. Compounds in brackets have uncertain activity (i.e. spike rates between the upper and lower threshold).

index	substance name	ab1D	ab2A	ab2B	ab3A	ab3B	ab5B	ab6A
1	(E)2-hexen-1-ol	-3	-3	**56**	2	16	0	**123**
2	(E)2-hexenal	1	-1	2	3	(24)	2	**85**
3	(E)2-hexenyl acetate	-3	8	0	**114**	-11	**22**	**78**
4	(E)2-octenal	1	-3	3	(20)	8	3	**71**
5	(R)-(+)-limonene	-2	-7	1	-4	-6	-8	-8
6	1,4-cineole	-4	-2	1	-1	3	0	-6
7	1,4-diaminobutane	-1	-11	-1	-1	19	(13)	3
8	1-octen-3-ol	0	-3	(14)	**49**	**39**	(13)	**175**
9	1-propanethiol	3	-11	2	5	(22)	7	-1
10	2,3-butanedione	2	**102**	1	(21)	**46**	-5	**42**
11	2-heptanone	0	5	1	**33**	**122**	**70**	**48**
12	2-isobutyl-3-methoxy-pyrazine	6	-4	-1	-6	2	-4	-3
13	3-(methylthio)-1-propanol	9	-7	4	2	10	**152**	14
14	3-octanol	2	-2	7	**57**	**112**	**27**	**162**
15	4-isopropylbenzaldehyde	13	-2	1	-7	16	-1	-14
16	4-methylcyclohexanol	7	0	3	(13)	4	-2	-15
17	4-methylphenol	(22)	-3	0	-1	(31)	1	-12
18	Acetophenone	**157**	-8	2	-7	1	1	-15
19	α-pinene	-4	-8	0	-3	13	-5	12
20	Ammonia	5	-10	-1	-1	17	5	-10
21	β-citronellol	-6	-7	3	-5	12	4	**53**
22	Benzaldehyde	**49**	-8	1	-5	3	-1	-13
23	Butanoic acid	1	-2	1	(20)	2	-4	**82**
24	Butanol	0	13	2	(11)	11	-5	-16
25	Carbon dioxide	0	5	1	4	14	1	10
26	*cis*-vaccenyl acetate	5	-6	1	(13)	-12	-1	-16
27	Cyclohexanone	-8	-2	-1	3	-1	-2	-15
28	Dipropyldisulphide	6	-10	-2	2	8	5	4
29	Ethanolamine	0	-9	5	4	9	8	-6
30	Ethyl 2-methylbutanoate	6	14	**23**	**141**	-7	2	2
31	Ethyl acetate	4	**156**	1	(14)	9	8	9
32	Ethyl butanoate	4	(23)	**73**	**145**	5	3	-2
33	Ethyl propionate	-1	**69**	(20)	**60**	-10	8	-12
34	Eugenol methyl ether	11	-1	1	-7	11	-1	-15
35	Geranyl acetate	2	5	0	1	(29)	-5	(26)
36	γ-valerolactone	**32**	-2	**23**	**32**	-2	-2	**47**
37	Hexanol	3	8	**67**	(20)	**87**	5	**134**
38	Indole	4	0	1	-1	10	-3	5
39	Iso-amyl acetate	**50**	7	9	**104**	8	**45**	(24)
40	Iso-amyl alcohol	1	1	4	(14)	6	1	-10
41	Linalool	-2	-7	-4	-1	14	-1	(36)
42	Methyl salicylate	**187**	-2	3	4	3	-3	5
43	Nonanal	1	-5	-1	-4	6	-4	-9
44	Pentyl acetate	5	(23)	2	**111**	(25)	**198**	**69**
45	Phenylacetaldehyde	**76**	-6	0	-3	12	-5	-19
46	Propanone	-3	**88**	1	1	2	1	(35)
47	Pyrrolidine	2	-5	6	-4	(24)	-1	(28)
	lower threshold	20	20	10	10	20	10	20
	upper threshold	30	30	20	30	35	20	40
	number of actives	6	4	6	10	5	6	13
	number of inactives	40	41	40	28	36	39	29

Odorants with an activity value between the two thresholds were excluded from the modeling process (bracketed in Table 2.1), because their activity cannot be determined with a high level of confidence. The dose-response curve of ORNs is a sigmoid, and small differences in odor delivery can in result in changes in the concentrations producing inconsistencies between the previously published results (de Bruyne et al., 2001) and the recordings in this study, particularly for these "borderline" odors.

To determine the two thresholds we used the following procedure (illustrated in Figure 2.2): Starting from activity histograms for each ORN, we estimated a lower threshold below which a molecule is considered inactive. Assuming that the activities of inactive compounds would be distributed around zero spikes/s (but without knowing the true distribution), we estimated the lower threshold to be where the first "gap" in the activity histogram distribution was located. Similarly, we estimated the upper threshold above which we considered molecules as being active.

In one case, additional data (Hallem et al., 2004) indicated that ethyl acetate, considered inactive at ab5B according to the threshold, may actually be a weak activator for the ab5B neuron. In consequence, we marked its activity as "unknown".

2.2.3 Descriptor calculation, selection and ranking

We calculated 203 physicochemical molecular descriptors using MOE (Chemical Computing Group, Montreal) for each odorant molecule, including calculated physical properties, subdivided surface areas, atom and bond counts, Kier & Hall connectivity and shape indices, adjacency and distance matrices, pharmacophore features, partial charge indices, potential energies, surface area, volume and shape indices and conformation dependent charge indices. Table A.1, starting on page 78 in the appendix provides a list of all descriptors that we used in this study, and how they are derived from the chemical structure.

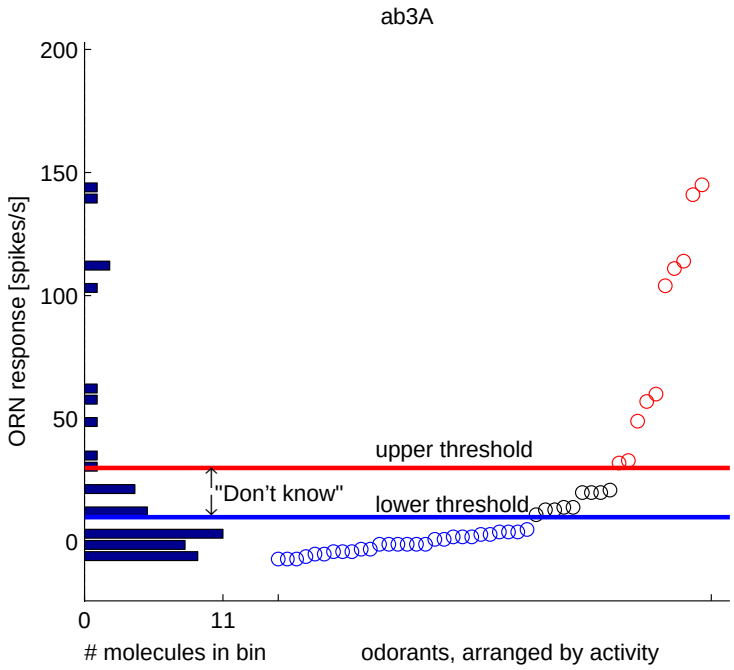

Figure 2.2: Binarizing activity using thresholds, here with respect to the ab3A ORN class. On the left, histogram representation of the activities, using 40 bins on the activity range. On the right, odorants arranged by activity (in spikes/s), with the highest activities most right. The two lines indicate the thresholds we determined. Odorants with activities below the lower threshold are considered "inactive" (blue circles), those above the upper threshold are considered "active" (red circles). The remaining odorants are excluded from the analysis (black circles).

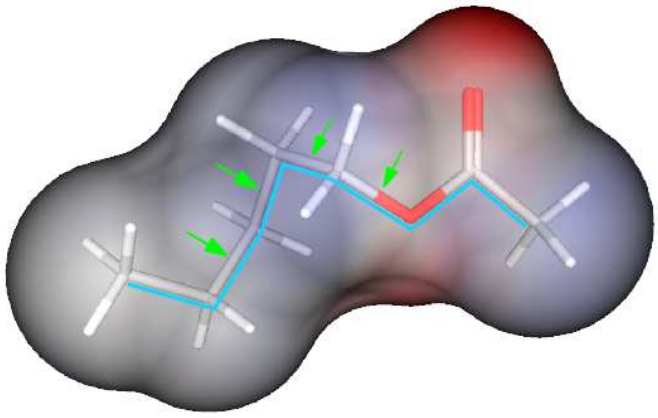

Figure 2.3: Some examples for molecular descriptors and their calculation from molecular structure, demonstrated for pentyl acetate. The molecular surface is colored by partial charge (blue=positive, red=negative partial charge, grey=neutral). The turquoise line denotes the longest chain in the molecule. The green arrows denote the rotatable single bonds (not counting the conjugated ester bond).

Figure 2.3 illustrates the meaning of various descriptors calculated for pentyl acetate. For example, the number of rotatable bonds (b_rotN in Table A.1) denotes the number of bonds in the molecule that have order 1, are not in a ring, and have at least two non-hydrogen neighbors. The double-bonded oxygen moiety of the ester group can function as a hydrogen bond acceptor, setting the number of H-bond acceptors to 1 (a_acc in Table A.1). The fractional negative surface area (e.g. PEOE_VSA_NEG in Table A.1) of pentyl acetate states the proportion of the molecular surface that has negative partial charge (indicated by red surface color), and has a value of 0.13. Finally, the longest chain (diameter in Table A.1 as defined by Petitjean (1992)) has a length of seven.

Prior to descriptor calculation, we generated heuristic 3D conformations with CORINA (Molecular Networks, Erlangen, Germany). At this stage, we used one conformation per molecule. Subsequently, those conformations were refined by energy minimization using MOE's MMFF94x force field, a modified version of the MMFF94s force field (Halgren, 1999). Minimization was stopped at a gradient of 10^{-5}.

Pruning unsuitable descriptors

Nine descriptors were discarded because they had zero variance across odor molecules. Some descriptors (e.g. the dipole moment) depend on the three-dimensional conformation of the molecule, which could lead to inconsistent modeling results for different conformations. Because we do not know which conformation of an odorant stimulates the ORN we sought to eliminate descriptors that vary strongly with 3D conformation.

To identify such strongly varying descriptors, we generated multiple conformers of all odorants using MOE's stochastic conformer generation functionality, using an energy cutoff of 5 kcal/mol. This resulted in a median nine conformers per molecule, with a maximum of 956 conformers for nonanal. For each descriptor the variance over all conformers of an odorant was calculated and scaled using the Fano Factor (Fano, 1947), $F_{\mathrm{D}} = \frac{\sigma_{\mathrm{D}}}{\mu_{\mathrm{D}}}$, with σ_{D} the variance and μ_{D} the mean of descriptor D over all conformations, without prior normalization. We calculated the mean F_{D} of each descriptor over all molecules and ranked the descriptors accordingly. Data from preliminary experiments (not shown) suggested a set of descriptors that particularly affected prediction quality through conformational variation. From those, the one with the smallest Fano Factor was the "dipole" descriptor with $F_{\mathrm{D}} = 0.03$. Therefore, we eliminated a total of 26 descriptors with a mean $F_{\mathrm{D}} \geq 0.03$.

Descriptor selection

Descriptors were ranked by their ability to separate active from inactive molecules. We quantified this ability using the Kolmogorov-Smirnov (KS) test (Manoukian, 1986). The KS-test compares the distribution of two series of data samples A and B by comparing, for each potential value x, the fraction of values from A less than x with the fraction of values from B less than x. The KS-value (k_{KS}) is the maximum difference over all x values. For each ORN, the descriptor values of all active odorants provided A, while B was provided

by the inactive odorants. Figure 2.4 illustrates this process for the b_1rotN descriptor at the ab3A ORN.

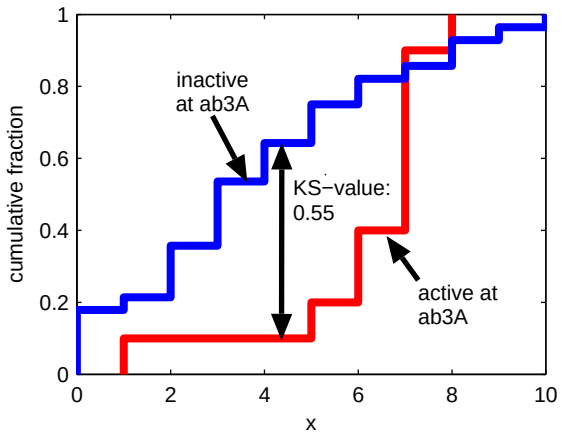

Figure 2.4: Calculation of the KS-value illustrated for the b_1rotN descriptor at the ab3A ORN. On the abscissa is the descriptor value, on the ordinate the cumulative fraction of odorants with a descriptor value equal to or less than the value on the abscissa. The KS-value is the maximum difference between the cumulative distribution functions of active (orange) and inactive (blue) compounds at ab3A.

The KS-test was performed using MATLAB R14 (The MathWorks, Natick, MA). We ranked the descriptors according to their p-value in the KS-test, that is, the probability that A and B stem from the same distribution. High KS-values result in low p-values. Descriptors with low p-values were ranked highest. Note that the ranking is specific and unique for each ORN. This is because for each ORN, different molecules constitute the active and inactive population, and in consequence the descriptor values for active and inactive molecules are differently distributed, which leads to different KS-values for the differences between distributions.

2.2.4 Artificial Neural Network training

We trained multilayer feed-forward Artificial Neural Networks (ANNs) to predict the activity of odorant molecules. Such networks have been described in detail elsewhere (Hertz et al., 1991; Zupan and Gasteiger, 1999). Briefly, a net-

work with k inputs, j neurons in the hidden layer, and i output neurons delivers the output O_i^μ in response to a pattern μ according to equation (2.1):

$$O_i^\mu = g\left(b_i + \sum_j W_{ij} \cdot g\left(b_j + \sum_k w_{jk} \cdot \xi_k^\mu\right)\right) \quad , \tag{2.1}$$

with $g(x)$ the transfer function of the output and hidden layer neurons respectively, b_i, b_j the bias of the neurons, W_{ij} the weight of the jth hidden neuron to the ith output neuron, w_{jk} the weight of kth input neuron to the jth hidden neuron, and ξ_k^μ the kth element of input pattern μ. We used a sigmoidal transfer function $g(x) = \frac{1}{1+e^{-x}}$, where x is the net input of a neuron.

The MATLAB Neural Network Toolbox was used for ANN modeling, employing backpropagation training with a gradient descent algorithm as implemented in MATLAB's `traingdx` function (Hertz et al., 1991).

Descriptor values were scaled to zero mean and unit standard deviation (autoscaling) prior to network training. We assigned a target value of 1 to active molecules and 0 to inactive molecules. We formed 250 pairs of equally sized training and validation data sets by random splitting, keeping the fraction of active to inactive molecules identical in both sets.

Network performance during training was assessed using the mean standard error (MSE, equation (2.2))

$$\mathrm{MSE}(O_{\mathrm{expect}}, O_{\mathrm{predict}}) = \frac{1}{S}\sum_{i=1}^{S}\left(O_{\mathrm{expect}} - O_{\mathrm{predict}}\right)^2 , \tag{2.2}$$

where O_{predict} was the output of the network and O_{expect} was given by the target values.

The MSE on the training data served as fitness function during training. ANN training was stopped when the MSE on the validation data did not decrease for 5,000 training epochs.

2.2.5 Model performance evaluation

Two factors greatly influence the outcome of ANN training: The ANN architecture (how many neurons to use in the hidden layer) and the number of inputs (molecular descriptors). More neurons in the hidden layer or a higher number of inputs to the ANN may allow for more complex description of the data, but the resulting model is also susceptible to overfitting, that is, modeling fine details without revealing the global data structure. Because these parameters are difficult to estimate in advance, we trained many networks with different combinations of parameters, varying the number of neurons in the hidden layer from one to four. In the special case of one hidden neuron, the ANN was reduced to a single neuron, which essentially is a Perceptron architecture (Hertz et al., 1991). To vary the number of descriptors, we cumulatively used the first $1, 2, \ldots 30$ descriptors from the ranked list, meaning we used the first descriptor, then the first two and so on until we used all 30 highest-ranked descriptors.

In total, we trained 30,000 ANN models per ORN (4 architectures$\times$30 input dimensionalities$\times$250 repetitions with different data splitting). We proceeded with selection of models with high predictive accuracy in cross-validation. We used the Matthews Correlation Coefficient MCC (Matthews, 1975) to assess prediction quality (eq. (2.3)):

$$\mathrm{MCC} = \frac{P \cdot N + O \cdot U}{\sqrt{(N+U) \cdot (N+O) \cdot (P+U) \cdot (P+O)}}, \tag{2.3}$$

where P is the number of true positives, that is, data instances that are active and have also been predicted active. N (true negatives) is the number of data instances that are inactive and have been predicted inactive. O denotes the number of overpredicted instances that were predicted active in spite of being inactive, and U is the number of underpredicted instances, that is, active instances predicted inactive. During each training run, we recorded the MCC on the training data as well as on the validation data for this run.

Model selection

A well-trained, well-generalizing model will have a high MCC both on the training and validation data. Hence, we selected ANNs with a training MCC equal or greater than their validation MCC, differing by no more than 0.1. From all ANNs fulfilling these criteria, we selected those with the maximum training MCC. If the selection resulted in more than one ANN, we used all selected ANNs and combined their prediction values by averaging.

For some ORNs, additional odorant activity data was available from other sources (Hallem et al., 2004; Stensmyr et al., 2003), providing an additional selection constraint on the models (see Table 2.2). Models failing to correctly predict the additional activity data were discarded. Of the additional compounds, Ethyl-3-hydroxybutyrate, a strong activator for ab3A according to (Hallem et al., 2004), was not tested in (de Bruyne et al., 2001), making it suitable as an additional validation point. Ethyl acetate was weakly active in (Hallem et al., 2004) at the ab5B ORN but inactive in the original data. Assuming that it truly is an activator of ab5B, we excluded it from network training and used it to validate the ANN predictions. The remaining compounds in Table 2.2 were originally excluded from training because their activity fell in between the upper and lower activity threshold and thus could not be derived with certainty. Since the additional sources suggest they are active, we used them as validation compounds for model selection.

2.2.6 Electrophysiology

We used the models to predict activity for a new set of odorants and tested the predictions in a new set of measurements from *Drosophila* ORNs in cooperation with Marien de Bruyne and Melanie Hähnel from the Freie Universität Berlin. Electrical activity was recorded extracellularly by inserting glass electrodes into individual sensilla on the antenna of *Drosophila melanogaster* males as previously described (de Bruyne et al., 2001; Dobritsa et al., 2003). Each

Table 2.2: Additional odorant activity data that was used for model selection. Sources: a: (Stensmyr et al., 2003), b: (Hallem et al., 2004).

ORN	Odorant Name	Source	Remarks
ab1D	furfural	a	-
ab2B	cyclohexanol	a	-
	(R)-ethyl-3-hydroxybutyrate	b	unknown stereoisomer, not tested by de Bruyne et al. (2001)
ab3A	butyl acetate	b	-
	ethyl acetate	b	unsure activity in b
	1-hexanol	b	unsure activity in b
ab3B	pentyl Acetate	b	unsure activity in b
	E2-hexenal	b	unsure activity in b
ab5B	ethyl acetate	b	considered inactive in b

sensillum houses several ORNs, either 4 (ab1 sensilla) or 2 (ab2, 3, 4, 5 and 6 sensilla). Neuronal excitation was measured as counts of spikes (action potentials) produced during a 500 ms stimulation period. Spike rates for each odorant were averaged from at least 9 (ab1 and ab2 sensilla), 7 (ab3 sensillum) or 3 individuals (ab5 and ab6 sensilla). It has previously been shown that spikes produced by the neurons in each of these sensilla can be reliably separated based on amplitude and shape differences (Clyne et al., 1997; de Bruyne et al., 2001; Stensmyr et al., 2003). The models were based on data generated with Tungsten electrodes but tested using saline filled glass electrodes. Both are standard methods that have been shown to produce similar results (Dobritsa et al., 2003). Most odorants were dissolved at 1% v/v in paraffin oil and air from a 5ml syringe, containing 10μl on a small piece of filter paper, was injected with a ca. 9-fold dilution factor (de Bruyne et al., 2001). Three odorants were tested at a 100 times lower concentration (see Table 2.4) because they were extremely potent activators for some ORNs.

2.2.7 Odorants

Odorants were obtained from Sigma, Aldrich or Fluka, of purity >99% or highest available, except for octanal (98%), salicylaldehyde (98%), ethyl 3-hydroxybutanoate (98%) and 2-octanone (98%). Except for (S)-(+)-carvone, all chiral odorants were applied as racemic mixtures.

2.3 Results and discussion

The goal of the first part of this thesis was twofold: First, we aimed at predicting ORN responses from molecular structure. Second, we wanted to describe structure-activity relationships between the odorant and the activated receptor.

To achieve the first aim, we trained artificial neural network models on an existing dataset of ORN responses, using selected subsets of chemical descriptors for odorant representation. We then recorded the responses of these same ORNs to a new set of chemicals to test whether the models we generated can be used to predict an odorants activity.

With the second aim in mind we analyzed the set of discriminative descriptors in order to characterize chemical properties that favor activation of each ORN.

2.3.1 Modeling ORN response and testing

We trained ANNs to model the activity of seven *Drosophila* ORNs in response to stimulation with odorant molecules. As training data we used ORN response data obtained in a previous study by *in vivo* electrophysiology (de Bruyne et al., 2001). We defined thresholds in activity such that a given compound can be classified as either "active", "inactive", or "uncertain", depending on the spike rate it elicits in the ORN. Compounds with uncertain activity were not used for training the ANNs for that specific receptor.

After selecting relevant descriptors for each ORN, we trained 30,000 ANN

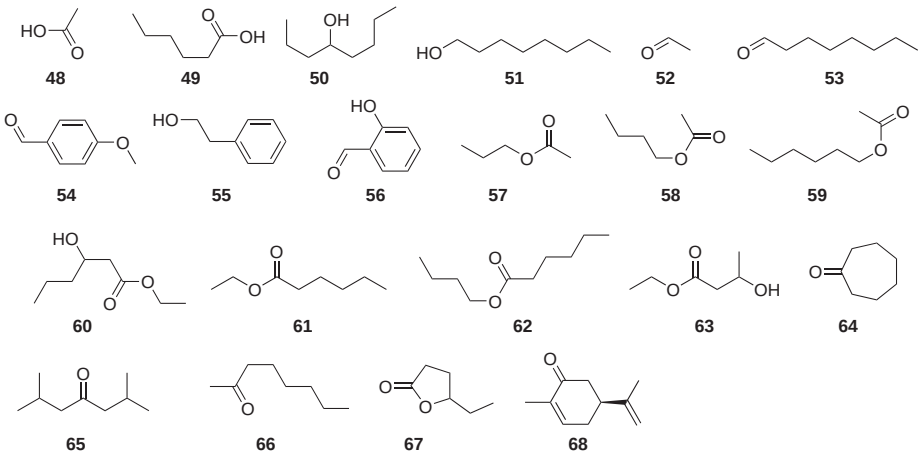

Figure 2.5: These odorants were screened to check prediction quality. Compounds names are given in Table 2.4.

models per ORN, selected those with the highest predictive power, and used them to predict ORN responses to 21 compounds, which were subsequently tested *in vivo*. These compounds, in the following referred to as "test data", are shown in Figure 2.5. We also assayed ten compounds that had already been tested by de Bruyne et al. (2001).

Spike rates in the test data were transformed into binary all-or-none data using the same thresholds as we used for the training data. Molecules with spike rates between the upper and lower threshold were excluded from the analysis, like in the training data.

We assessed prediction performance using the Matthews Correlation Coefficient for binary data (MCC, eq. 2.3). Table 2.3 shows the MCC for the training data and the test data. We excluded ethyl-3-hydroxybutyrate at ab2B and butyl acetate at ab3A from the calculation of the test set's MCC, since these molecules were used to select the best models (see section 2.2.5 on page 17). These compounds have entered the modeling process prior to testing and hence are not valid "test" compounds for those ORNs.

Five out of seven models succeeded in correctly predicting the training data. The training predictions for the ab3B and 6A neuron show imperfect performance, but still correlate with the activity in the training data.

The prediction of ORN response to novel molecules shows a mixed picture: For the ab3A ORN, the model achieved an MCC of 0.85, providing reliable prediction. For the ab1D, 2A, 5B and 6A ORNs, the MCCs range from 0.66 to 0.69, still indicating good performance. In contrast, the models showed only weak performance for the ab2B (MCC $= 0.17$) and 3B ORNs (MCC $= 0.34$).

The discrepancy between performance on the training data and the test data for some receptors may have several causes. First, although we used cross-validated training and, in some cases, additional activity data for model selection, due to the large number of models we built, it is possible that some models perfectly predict all training data, albeit by chance. Second, descriptor selection was performed on the whole data set instead of a cross-validated procedure, possibly "over-optimizing" descriptor space for the training data. However, because of the data splitting necessary for cross-validation, the number of data instances in one part of the data would have been too small for the statistical test we used to select descriptors. In both cases, the performance on the independent test set reveals the actual quality of prediction. This set contained only substances that did not enter the model creation at any point and is thus not affected by the above issues.

Table 2.4 gives detailed insight into the compounds we used for testing and the results of the screening, in comparison with the predictions. It should be noted that one compound (cyclohexanone) was inactive at ab3A in the training data (3 spikes/s), but active in screening (33 spikes/s). A similar observation made for 4-methylphenol at the ab1D neuron: its activity was uncertain in the training data (22 spikes/s), but it was inactive in screening (5 spikes/s).

Table 2.3: Matthew's Correlation Coefficient (MCC) for the training and the test data.

ORN		ab1D	ab2A	ab2B	ab3A	ab3B	ab5B	ab6A
MCC	training	1.00	1.00	1.00	1.00	0.77	1.00	0.86
	test	0.69	0.69	0.17	0.85	0.34	0.68	0.66

These differences may be a consequence of the effect that a slight variation in concentration may suffice to elicit a response (de Bruyne et al., 2001).

A possible source of error is that it is not always certain that the compound actually arriving at the receptor neuron did not undergo degradation, or that traces of other compounds contaminated the stimulus, for example as by-products from synthesis or as remnants after purification. These effects cannot be addressed by this study, but would require analysis of the air stream in parallel to the measurements, for example by gas chromatography (Vetter et al., 2006; Lin et al., 2005).

One point of discussion is the threshold setting for activity assignment, in that it followed no algorithmic procedure. However, these thresholds proved to be sensible choices, and appeared reasonable to us according to the data. First of all, the application of thresholds was necessary to simplify the data. As in any modeling study, simplifications have to be introduced in order to focus on the most relevant features, especially when the amount of data is limited. In this case, we chose to discard the quantitative activity data in favor of a binary active/inactive prediction. Although our threshold settings may have enhanced the aforementioned difference in activity assignment, these were more likely due to changes in the experimental setup, or variance in the *Drosophila* stock between the measurements of the training and test sets. Further, the models do not take into account the different vapor pressures of the compounds or effects of dose dependency of the responses, because the required data was not available for all compounds.

We also did not explicitly address any possible effects of modifiers of OR activity such as Olfactory Binding Proteins (OBPs). These proteins populate the aqueous lymph surrounding olfactory dendrites and have been shown to be involved in olfaction. *Drosophila* mutants devoid of the LUSH OBP have defects in avoiding high alcohol concentrations (Kim et al., 1998) and lack response to a pheromone (Xu et al., 2005). It has also been suggested that OBPs are involved in shuttling hydrophobic odorants through the lymph (Kaissling,

Table 2.4: Measured response (r, in spikes/s) and predicted activation (p, 1=predicted active, 0=predicted inactive). The upper ten compounds were also part of the training set. ORN set in bold responses are considered active, those in plain font inactive according to the thresholds in ORN response (see Table 2.1). Comparisons between predictions and measurements are marked up according to the following: — true negative, ++ true positive, o false positive (overpredicted) and u false negative (underpredicted). Responses to ethyl-3-hydroxybutyrate at ab2B and butyl acetate at ab3A have not been taken into account to assess prediction quality (see text).

		ab1D			ab2A			ab2B			ab3A			ab3B			ab5B			ab6A		
Index	substance name	r		p	r		p	r		p	r		p	r		p	r		p	r		p
23	butanoic acid	2	—	0	3	—	0	0	—	0	**19**	++	1	11	—	0	1	—	0	3	o	1
37	hexanol	5	—	0	6	—	0	**53**	++	1	(21)		1	**95**	++	1	6	—	0	**168**	++	1
14	3-octanol	0	—	0	1	—	0	2	—	0	**40**	++	1	**124**	++	1	13	o	1	**222**	++	1
22	benzaldehyde	**32**	++	1	3	—	0	2	—	0	8	—	0	14	—	0	0	—	0	11	—	0
17	4-methylphenol	5	o	1	1	—	0	-2	—	0	9	—	0	10	—	0	2	—	0	9	—	0
31	ethyl acetate	0	—	0	**175**	++	1	4	—	0	**38**	++	1	10	—	0	11	o	1	9	o	1
44	pentyl acetate	2	—	0	18	—	0	0	—	0	**153**	++	1	**64**	++	1	**180**	++	1	**166**	++	1
27	cyclohexanone	3	—	0	2	—	0	-1	—	0	**33**	u	0	(20)		0	2	—	0	12	—	0
10	2-heptanone	3	—	0	10	—	0	-2	—	0	**58**	++	1	**207**	++	1	**62**	++	1	**126**	++	1
35	geranyl acetate	2	—	0	(21)		0	-1	—	0	(19)		0	15	—	0	0	—	0	**65**	u	0
48	ethanoic acid	2	—	0	3	—	0	-2	—	0	9	o	1	16	—	0	1	—	0	12	o	1
49	hexanoic acid	0	—	0	3	—	0	1	o	1	**33**	++	1	11	o	1	3	o	1	4	o	1
50	4-octanol	1	—	0	0	—	0	4	—	0	**53**	++	1	8	o	1	(18)		0	**82**	++	1
51	octanol	7	—	0	6	—	0	**58**	u	0	(17)		1	18	o	1	6	—	0	**59**	u	0
52	acetaldehyde	2	—	0	1	o	1	0	—	0	9	—	0	10	—	0	1	—	0	0	—	0
53	octanal	3	—	0	2	—	0	2	—	0	9	—	0	(24)		1	4	—	0	(26)		0
54	4-methoxybenzaldehyde	(25)		1	4	—	0	1	—	0	6	—	0	18	—	0	1	—	0	(27)		0
55	2-phenylethanol	1	—	0	-3	—	0	4	—	0	(14)		0	12	—	0	2	—	0	11	—	0
56	salicylaldehyde	**46**	++	1	0	—	0	-1	—	0	7	—	0	16	—	0	1	—	0	17	—	0
57	propyl acetate	0	—	0	**35**	++	1	4	o	1	**112**	++	1	(29)		0	**40**	++	1	(27)		0
58	butyl acetate	4	o	1	17	—	0	-4	—	0	**135**		1	**68**	u	0	**116**	++	1	**101**	++	1
59	hexyl acetate	-1	—	0	9	—	0	-1	—	0	**72**	++	1	(33)		1	**61**	++	1	**106**	++	1
60	ethyl 3-hydroxyhexanoate*	1	—	0	5	—	0	9	o	1	**57**	++	1	17	—	0	**47**	u	0	(35)		0
61	ethyl hexanoate*	2	—	0	6	—	0	-1	—	0	**217**	++	1	3	o	1	1	—	0	3	—	0
62	butyl hexanoate	1	—	0	2	—	0	1	—	0	**63**	++	1	8	o	1	3	—	0	18	—	0
63	ethyl 3-hydroxybutyrate*	1	—	0	10	—	0	**204**		1	(14)		1	18	—	0	2	o	1	11	—	0
64	cycloheptanone	2	—	0	-2	—	0	-1	—	0	(27)		0	(20)		0	3	—	0	13	—	0
65	2,6-dimethyl-4-heptanone	2	—	0	1	—	0	**40**	u	0	(23)		1	14	—	0	4	—	0	12	—	1
66	2-octanone	1	—	0	3	—	0	3	—	0	**37**	++	1	**153**	++	1	**34**	++	1	**160**	++	1
67	gamma-hexalactone	4	—	0	6	—	0	-2	o	1	**46**	++	1	9	—	0	1	—	0	19	—	0
68	(S)-(+)-carvone	2	—	0	-5	—	0	-3	—	0	(11)		0	19	—	0	3	—	0	(29)		0

* These odorants were tested at a 100 times lower concentration

2001). The model, being trained on the activation data of ORNs in their "native surround" (i.e. the lymph), implicitly treats everything between the odorant and ORN activation as a "black box" and hence also contains effects of OBPs, if present.

2.3.2 Interpretation of descriptor selection

As stated above, we selected subsets of descriptors that are best suited for separating active from inactive compounds prior to ANN training. In addition to reducing the "noise" introduced into the data by unsuitable descriptors, the ranked list of descriptors can also give insight into the SAR of the ORNs. Since each descriptor represents a molecular feature, descriptors in the selected subset point to potentially preferred molecular features detected by an ORN. The sum of preferred features determines an ORN's "receptive field".

The descriptor rankings were produced using the p-value from a Kolmogorov-Smirnov test (KS-test) for significant difference between two data sets (inactive vs. active compounds), separately for each ORN. Descriptors with the lowest p-values were ranked highest. The ranked lists of descriptors including their associated p-values are given in table A.2 in the appendix.

We observed that the set of highest ranking descriptors is different for each ORN. This may correspond to a different SAR for each ORN, in that different chemotypes are recognized by different receptors.

In the following, we describe how the descriptor rankings relate to the SARs of the ORNs in this study. For the sake of compact display, we refer to individual descriptors by their abbreviations. More elaborate explanations of all descriptors that appear here and in the ranked lists are provided in table A.1 in the appendix.

ab1D

For ab1D, the highest ranked descriptor is std_dim3, a 3D shape descriptor, that describes the standard deviation along the principal component axis of the

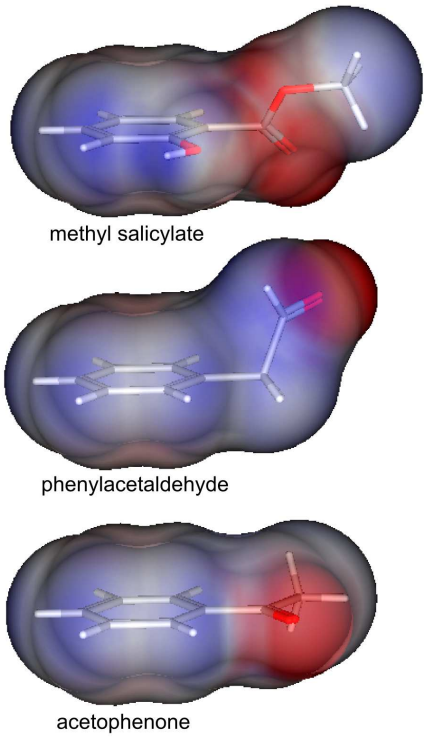

Figure 2.6: Three activators of ab1D, methyl salicylate (187 spikes/s), phenylacetaldehyde (76 spikes/s) and acetophenone (157 spikes/s) reveal their disk-like shape in surface representation. Red areas indicate negative partial charge, blue areas positive partial charge, and white indicates neutral (=no) charge.

atom coordinates. Typical activators of ab1D like methyl salicylate, acetophenone and phenylacetaldehyde have disk-like shape, which is due to their aromatic ring systems (see Figure 2.6). Hence, they will have small values for this descriptor, discriminating them from the other molecules in the data set. This descriptor does not feature strongly in the rankings of other ORNs that respond to aliphatic compounds. Furthermore, the high ranking of several descriptors for charge distribution on the molecular surface (such as PEOE_VSA_FPNEG, Q_VSA_FNEG, FCASA-) reflect the exposed carbonyl groups in most activators of ab1D, creating a focused negative partial charge distribution on the molecular surface (cf. Figure 2.6). Charge distribution descriptors feature high on the list of several ORNs.

ab2A

A strong effect of partial charge can also be observed for the activators of the ab2A ORN (ethyl acetate, 2,3-butanedione, propanone, ethyl propionate),

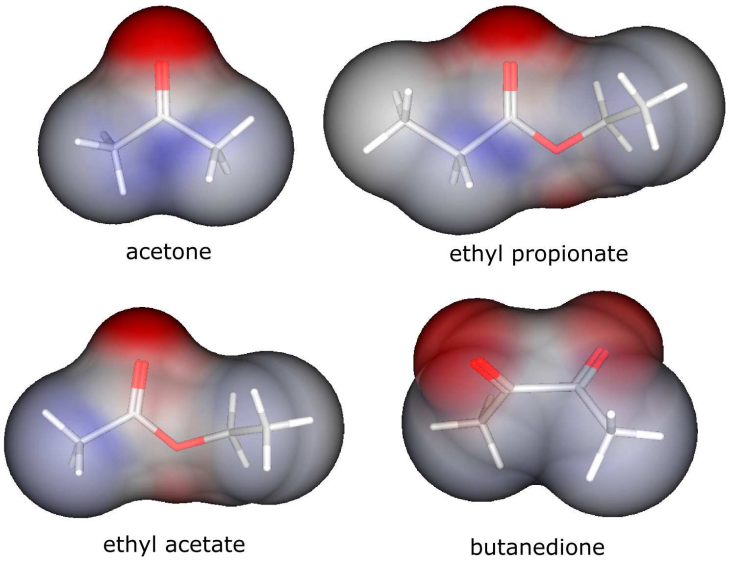

Figure 2.7: Conolly-surface representation for activators of the ab2A ORN. Color scheme is identical to Figure 2.6.

which are all comparably small and bear a focused negative partial charge on the molecular surface (cf. Figure 2.7). The focused charge is again represented in the highest scoring PEOE_VSA_FPNEG descriptor. The high rank of a_ICM can be related to the small molecule size. It describes the mean atom information content, which reflects the *entropy*, used by its information-theoretical meaning, in atom composition. For two equal-sized molecules, the one which is composed of more different atom types will have the higher entropy. Accordingly, for two molecules with the same number of different atom types, the smaller one will have higher entropy. Now the high scoring molecules incorporate only two atom types, namely O and C, as well as the majority of the remaining molecules in the data set. Thus, the smaller molecule size likely is the discriminating feature. Several connectivity descriptors (chi1v_C, chi1_C etc.) also reflect the importance of molecule size.

ab2B and ab3A

The AM1_HOMO descriptor, which is an index for "reactivity", yields a high rank for the ab3A neuron. Moreover, the MNDO_HF descriptor (heat of formation) correlates well with ab3A spike rate change (Pearson correlation coefficient: -0.55, $p < 10^{-4}$). Also, the ionization potential (reflected in the AM1_IP, PM3_IP and MNDO_IP descriptors) yields a high rank. All these descriptors relate to the reactivity of a molecule and are negatively correlated with activity. This seems evident if one considers that most activators of ab3A are esters, which are less reactive than for example aldehydes and primary alcohols, two groups to which many of the non-activators belong.

Similar observations can be made for ab2B, where four of the five activators of the ab2B ORN (ethyl butanoate, hexanol, γ-valerolactone, ethyl-2-methylbutanoate) have a slightly elevated ionization potential according to the AM1_IP descriptor, compared to non-activators (e.g. 3-methylthio-1-propanol, benzaldehyde or linalool), as well as a high ranking of the AM1_HOMO descriptor.

ab5B

For the ab5B ORN, the highest ranked descriptors are related to molecular shape, expressed by descriptors developed by Hall and Kier (1991) (KierA3, KierA1, KierA2, KierFlex, Kier2, Kier3). In combination with the high ranked b_1rotR descriptor (the relative number of rotatable bonds in the molecule), this reflects ab5B's preference for larger, flexible ligands, such as pentyl acetate, 2-heptanone and 3-octanol.

ab6A

Finally, for the ab6A ORN the Kier3 and Kier2 descriptors described by Hall and Kier (1991) rank highest. According to Todeschini and Consonni (2000), Kier2 encodes information about the "spatial density of atoms" in a molec-

ular graph, while Kier3 encodes the "centrality of branching"; Kier3 values are larger when branching is located at the extremities of the molecular graph or when no branching happens in the molecule, and they are smaller when branching is located near the center of the molecule. Interestingly, the single ANN model that was selected for prediction of ab6A activity only used these two descriptors. Considering that the descriptor values of activators all lie inside a very small range in which no non-activators are present (data not shown), and the fact that the selected ANN model has two hidden neurons, the network simply "cut out" the value range in which the activators of ab6A lie, a typical effect of overtraining. This may be a possible explanation for the rather poor predictive performance of the ab6A model. The ab6A ORN shows a somewhat broader selectivity characteristic: activators are not as easy to discriminate from non-activators as for the other ORNs, and our method of assigning binary activity values may not have been appropriate in this case. Here it is important to note that ab6A is the only ORN in this study for which the receptor gene could not yet be identified (Hallem et al., 2004).

General remarks

The results we present in this section should be taken as an example of how to extract knowledge from such an analysis. It is not justified to interpret an individual descriptor as the sole discriminating feature. Rather, the KS-statistics demonstrate that many features are suitable for classification. Descriptor selection is the result of a statistical procedure, and depends on the composition of the data set. Moreover, the ANN models combine the information obtained from the selected features to represent a more complex and nonlinear (except for Perceptron-type ANNs) relationship between molecular structure and activity than is suggested by the inherently linear descriptor ranking.

With these notes of caution, one might speculate that binding of odor molecules is achieved through different receptor-ligand interaction mechanisms at each OR. For example, our study suggests that ab2A is activated at least in part

by the polarity of small ligands, whereas ab5B appears to require the larger ligands with flexible side chains. While in the past the classification of chemical stimuli was based on functional group or chemical class, the use of physicochemical descriptors provides a different view on the molecular features that govern ORN activation.

A systematic analysis of ORN selectivity was complicated by the limited amount of ORN response data. Only recently, more comprehensive data on *Drosophila* ORN responses became available (Hallem and Carlson, 2006). Although the data was acquired using a different methodology (heterologous expression of OR genes in an "empty" ORN), it is possible that more data on these ORs will yield better results. This may be a fruitful task for a future study. It will be interesting to see if the abstract description of chemical entities as we used here can aid to reveal a logical structure in the selectivity of ORNs.

2.3.3 Using ORN responses to predict ORN responses

If ORN responses really span some sort of chemical space, it should as well be possible to use the spike rates as a descriptor. To assess this hypothesis, we tried to predict activity of one ORN using responses of the remaining ORNs. We used the logarithm of the spike rates, because principle component analysis showed that this transformation results in a more uniform distribution with less outliers (data not shown). ANN training and model selection followed the same protocol as above, except that only 150 pairs of test and training data were formed, and no additional validation data was available to prune networks that showed poor generalization. Since only six descriptors were available to train the ANNs, we did not apply KS-statistics for data reduction.

The results are given as correlation coefficients in Table 2.5. ORNs ab3A, ab3B, ab5B, and ab6A show moderate correlation (MCC between 0.47 to 0.66) on the test set, but prediction completely failed for ab1D, ab2A (MCC$= 0$, respectively) and ab2B (MCC$= -0.10$). This indicates that this approach indeed works, at least for four out of seven receptors. The failure at the remaining

Table 2.5: Matthew's Correlation Coefficient (MCC) for training and screening (test) using ORN responses as descriptor.

ORN		ab1D	ab2A	ab2B	ab3A	ab3B	ab5B	ab6A
MCC	training	0.55	0.0	0.76	1.00	0.77	0.80	0.72
	test	0.0	0.0	-0.10	0.47	0.66	0.54	0.54

three may results from the fact that for these receptors there are too few actives in the test set, namely one for each ab1D (salicylaldehyde) and ab2A (propyl acetate), and three for ab2B (octanol, ethyl 3-hydroxybutanoate and 2-octanone).

These results suggest that ORNs do not code in an "orthogonal" way, i.e. their responses are not uncorrelated. If this was the case, the above analysis must have failed. Rather, the properties each ORN class encodes seem to overlap between classes, providing a partly redundant coding scheme.

2.4 Conclusion

We have demonstrated that it is possible to predict *Drosophila* ORN responses from molecular structure. The approach performed well on the majority of receptors, considering that only few data was available for training. The features that were selected as being suitable for model training indicate that each ORN has different preferences regarding the physicochemical properties of its potential ligands. Finally, the ORN responses themselves can effectively be used as a descriptor to predict responses of other ORNs, providing evidence that ORNs indeed analyze chemical space in a way that can be exploited to predict receptor-ligand affinities. Moreover, it indicates that the encoding by olfactory receptors is partly redundant.

Acknowledgments

The electrophysiological part of this work has been carried out at the Freie Universität Berlin in cooperation with Marien de Bruyne and Melanie Hähnel.

Chapter 3

Modeling the insect antennal lobe with self-organizing maps

3.1 Background

The antennal lobe in insects (and the analogous structure in vertebrates, the olfactory bulb) are located at the second stage of olfactory processing. With several lines of evidence suggesting a chemotopic ordering in this neural structure, they provide a fascinating target to study how chemical similarity is defined in the olfactory system. Here, we show how self-organizing maps (SOMs) can be used to investigate this issue.

3.1.1 Self-organizing maps

SOMs were introduced as a feature extraction and data mapping approach by Kohonen (1982). Many variations of Kohonen's original concept have been conceived ever since (Kohonen, 2001). In the area of bioinformatics they have been primarily used for visualizing protein and DNA sequence and structure spaces (Arrigo et al., 1991; Ferrán and Ferrara, 1991; Schuchhardt et al., 1996; Hanke and Reich, 1996; Schneider et al., 1998; Aires-de-Sousa and Aires-de-

Sousa, 2003; Schneider and Fechner, 2004; Fankhauser and Mäser, 2005; Bensmail et al., 2005), drug design tasks (Schneider and Wrede, 1998; Polanski and Walczak, 2000; Givehchi et al., 2003; Schneider and Nettekoven, 2003; Teckentrup et al., 2004; Xiao et al., 2005), surface and property visualization and prediction (Gasteiger et al., 1994; Anzali et al., 1996; Hasegawa et al., 2002; Roche et al., 2002; Balakin et al., 2005), and binding site analysis (Stahl et al., 2000; Del Carpio-Muñoz et al., 2002) — often in conjunction with other clustering and pattern matching techniques. Typically, the use of SOMs has been restricted to two-dimensional (2D) projections of higher-dimensional data.

3.1.2 The SOMMER Application

We developed SOMMER, the **S**elf-**O**rganizing **M**ap **M**aker for **E**ducation and **R**esearch as a toolbox for the training and visualization of two- and three-dimensional (3D) unsupervised SOMs. The extra dimension in the SOM grid may allow for a better low-dimensional mapping of complex data manifolds. Moreover, the 3D grid allows for SOM topologies which are not available in 2D space, and SOMMER provides map topologies for planar rectangular, toroidal, cubic and spherical projections (see Figure 3.1).

The software was written in Java and makes use of the 3D visualization capabilities of the Java3D-package (https://java3d.dev.java.net/). By displaying the training process of the SOM, it illustratively demonstrates how SOM neurons self-organize to map the data distribution. This feature does not only facilitate the understanding of the training process (being particularly valuable for teaching), but can be used to assess the usefulness of a mapping solution. Integrated data processing tools provide means for data normalization. The SOM topology can also be set to a 2D scaffold. In this case, the user benefits from the 3D display of the data distribution, making eventual glitches in the 2D data mapping obvious. High-quality images can be saved for publication purposes.

The use of spherical lattices for self-organizing maps was first proposed

by Ritter (1998) as an example for the application of SOMs in non-Euclidian spaces. Sangole and Knopf (2003b) used deformations of spherical SOMs to create three-dimensional representations of numeric data sets that can be used for visual assessment of data set similarity and data classification. Wu and Takatsuka (2004) showed that spherical SOMs can converge to smaller quantization errors in less training epochs than SOMs with a rectangular, two-dimensional lattice. It has also been shown that spherical SOMs can yield low-dimensional feature maps of data distributed on the surface of a hypersphere with a lower embedding error than planar SOMs can (Nishio et al. (2004), Poster abstract for the Eighth Annual International Conference on Computational Molecular Biology (RECOMB), San Diego, USA). These observations are not surprising due to the fact that a SOM-projection is best if the dimensionality of the SOM is identical to the dimension of the data (Kohonen, 1982, 2001)—still, they demonstrate possible primary applications of spherical SOMs.

Suggesting an alternative application scenario for SOMs, Sangole and Knopf (2003a) showed how deformable spherical SOMs can be applied to create representations of freeform objects with the application to object recognition and shape registration, and demonstrated the capability of the spherical SOM to mimic object surfaces and reconstruct missing points.

Here, we apply SOMs to provide a regular projection of a three-dimensional object, namely the antennal lobe of *Drosophila*.

3.1.3 Chemotopy in Drosophila's antennal lobes

The antennal lobes of *Drosophila* are structures of neuropil in the olfactory system. Axons from primary receptor neurons converge onto this structure, forming *glomeruli*. These glomeruli are sites of high synaptic connectivity between primary receptor neurons and secondary projection neurons. Axons from receptor neurons bearing the same olfactory receptor converge on the same glomerulus (Couto et al., 2005; Fishilevich and Vosshall, 2005).

It is speculated that the arrangement of glomeruli follows a chemotopic

rule, such that similar odorants activate nearby glomeruli, as has been previously reported in the honeybee (Sachse et al., 1999), zebrafish (Friedrich and Korsching, 1997), mouse (Uchida et al., 2000) and rat (Meister and Bonhoeffer, 2001).

Currently, there is no clear evidence for or against a chemotopic ordering in the antennal lobe of *Drosophila*. While Fishilevich and Vosshall (2005) stated that chemotopy is not obvious in the antennal lobe, Couto et al. (2005) have reported that they did observe a clear chemotopic ordering, which was based on the chain length of aliphatic esters.

In a recent study, Hallem and Carlson (2006) provide the most comprehensive data set for *Drosophila* olfactory receptor responses available to date. Based on the mapping from receptor to glomerulus defined by Fishilevich and Vosshall (2005) and Couto et al. (2005), they derived glomerular activation from the activation of the "driving" receptor, i.e. the receptor expressed in the receptor neuron class that innervates the glomerulus. Although the authors could not establish a clear correspondence between chemical similarity and distance of the activated glomeruli, it is not entirely clear from their publication how they measured glomerulus distance. Since they state glomerular distance in μm, they presumably used the Euclidian distance in three-dimensional space.

However, the antennal lobes are spherical structures with the glomeruli mainly arranged on their surfaces. Hence, the Euclidian distance between two glomeruli may not be an appropriate measure for their separation, since this would ignore their spherical arrangement. The degree of separation, as used by Couto et al. (2005) is certainly a better measure to quantify the distance between two glomeruli, because it is based on the neighborhood structure and takes the local topology into account.

SOMs provide a topological mapping of the original data, which makes them particularly useful in this scenario. They enable straightforward, algorithmically well defined and reproducible projection of the three-dimensional arrangement of glomeruli onto the two-dimensional plane.

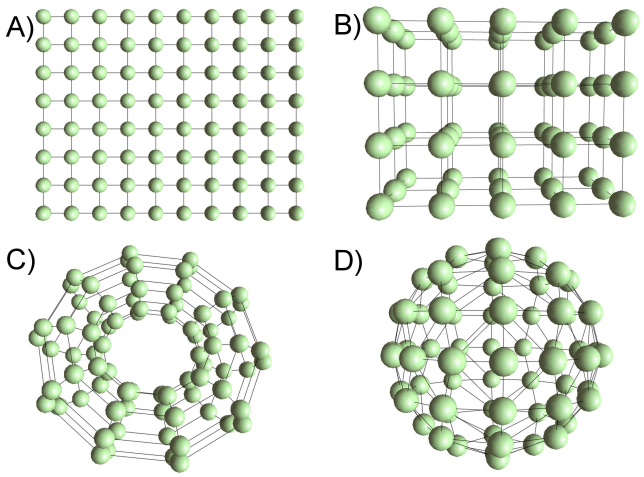

Figure 3.1: Topologies implemented in SOMMER. A) Rectangular 8×10, B) cubic $5 \times 4 \times 3$, C) toroidal 8×10, D) spherical with $f = 15$ (cf. equation 3.1).

3.2 Methods and data

3.2.1 Self-Organizing Maps

SOM Topologies

An SOM consists of a set of units which are linked by edges. These units are also called *neurons*, because the origin of the SOM was inspired by the various topographic mappings found in the brain (Kohonen, 1982). Figure 3.1 depicts the linkage topologies available in SOMMER. Currently, these are:

- Rectangular: a rectangular $X \times Y$ grid, containing $X \cdot Y$ neurons (Figure 3.1A).
- Cubic: a cubic $X \times Y \times Z$ grid, containing $X \cdot Y \cdot Z$ neurons (Figure 3.1B).
- Toroidal: rectangular topology with wrapped edges, resulting in a torus (Figure 3.1C).
- Spherical: a sphere with the neurons laid out regularly on its surface (Figure 3.1D).

Creating the spherical topology is handled differently than the other topologies, because distributing a number of points evenly on a spherical surface is a non-trivial task. Ritter (1998) used subdivisions of the icosahedron to tessellate the sphere. While that approach yields an almost regular tessellation, it offers a limited choice for the total number of neurons N, which obeys the formula $|N| = 10 \cdot f^2 + 2$, with f the subdivision frequency. Hence, for $f = 1, 2, 3, \ldots$ the number of neurons is quantized to $|N|$ =12, 42, 82, 162, 322, 642, 1282,... neurons.

We adopted the tetrahedron-based tessellation method from Java3D. The number of neurons obeys equation 3.1

$$|N| = \left(\frac{f + (3 - (f-1)\,\%\,4)}{2} \right)^2 + 2, \tag{3.1}$$

with % denoting modulo division. It leaves the choice between 6, 18, 38, 66, 102, 146, 198,... neurons. We chose this tessellation method over the icosahedron method because it allows for a more fine-grained tuning of the number of SOM neurons.

Training Algorithm

The algorithm implemented in SOMMER is based on the work of Loos and Fritzke (Loos HS, Fritzke B (1998), DemoGNG v.1.5 Manual). Since the grid-based distance metric used therein is not applicable to all available topologies, we generalized the distance between two neurons on the grid to a graph-based topological distance, such that $d_{\text{topo}}(n_1, n_2)$ is equal to the number of graph edges on the shortest path between the two neurons n_1 and n_2. With this slight modification, the algorithm is suitable for training an SOM with any topology, as long as some topological distance is defined.

The "winner neuron" $n_{\text{w}}(\xi)$ of a given data pattern ξ is defined as the neuron with minimal distance $d(\mathbf{n}, \xi)$, where $\mathbf{n}$ is the neuron's prototype vector (i.e. its coordinates in data space) and ξ the vector associated with the data pattern.

SOMMER implements Euclidean distance and the Manhattan (or city-block) metric to measure d.

In every training epoch t, the prototype vector $\mathbf{n}$ of each neuron is updated according to equation 3.2

$$\Delta\mathbf{n} = \lambda(t) \cdot \nu(n_\mathrm{w}, n, t) \cdot (\xi - \mathbf{n}), \tag{3.2}$$

with n_w the winner neuron, and the time-dependent learning rate $\lambda(t)$ defined by equation 3.3

$$\lambda(t) = \lambda_\mathrm{i} \left(\frac{\lambda_\mathrm{f}}{\lambda_\mathrm{i}}\right)^{\frac{t}{t_\mathrm{max}}}, \tag{3.3}$$

with λ_i the initial learning rate at $t = 0$ and λ_f the final rate at $t = t_\mathrm{max}$.

The neighborhood function ν determines how strongly a neuron n is adapted relative to the winner neuron n_w. We use a Gaussian neighborhood function (equation 3.4)

$$\nu(n_\mathrm{w}, n, t) = e^{\frac{-d_\mathrm{topo}(n_\mathrm{w}, n)^2}{2\sigma t^2}}, \text{ with } \sigma(t) = \sigma_\mathrm{i} \left(\frac{\sigma_\mathrm{f}}{\sigma_\mathrm{i}}\right)^{\frac{t}{t_\mathrm{max}}}, \tag{3.4}$$

where σ_i and σ_f refer to initial and final values of the neighborhood function.

The SOM training algorithm works as follows:

1. Initialize the prototype vector $\mathbf{n}$ of each neuron to a random vector.
2. Choose a data pattern ξ and determine the winner neuron $n_\mathrm{w}(\xi)$ with minimal distance $d(n_\mathrm{w}, \xi)$.
3. Adapt each neuron n according to equation 3.2.
4. Increase the training epoch $t = t + 1$.
5. If $t < t_\mathrm{max}$ continue with step 2, else terminate.

3.2.2 Three-dimensional models of the antennal lobes

Models of *Drosophila*'s antennal lobes are provided by the Flybrain database as VRML-models (Armstrong et al. (1995), http://www.flybrain.org, Accession Number AB00203). We used the model of "specimen 4". From the VRML-file, the surface coordinates of the glomeruli were extracted and converted to a format readable by SOMMER. In order to reduce the amount of surface data, we extracted 1000 representative surface points (see Figure 3.2B) from the available 5808 surface points by the usage of the MaxMin algorithm, using the Java-version of the application described by Schmuker et al. (2004).

3.2.3 Odorant data set

110 Odorants from the publication by Hallem and Carlson (2006) were converted into a database using ChemOffice 2002 (Cambridgesoft, Cambridge, MA). 184 Molecular descriptors were calculated with MOE (Molecular Computing Group, Montreal). We only calculated 2D-descriptors to minimize variations due to unknown conformation or stereo-configuration.

3.3 Results and Discussion

We used an SOM to project the spherical arrangement of glomeruli in *Drosophila*'s antennal lobes onto the two-dimensional plane. The projection in this plane allows for a more accessible visualization of activation patterns in response to odorant stimuli of the antennal lobe than it is possible using the original, three-dimensional structure.

3.3.1 SOM representations of the antennal lobe

Figure 3.2A shows a VRML-model of *Drosophila*'s antennal lobe from the Flybrain database (Armstrong et al., 1995). The glomeruli are colored according to their anatomical location: Blue for ventral, yellow for dorsal. Shadings of these

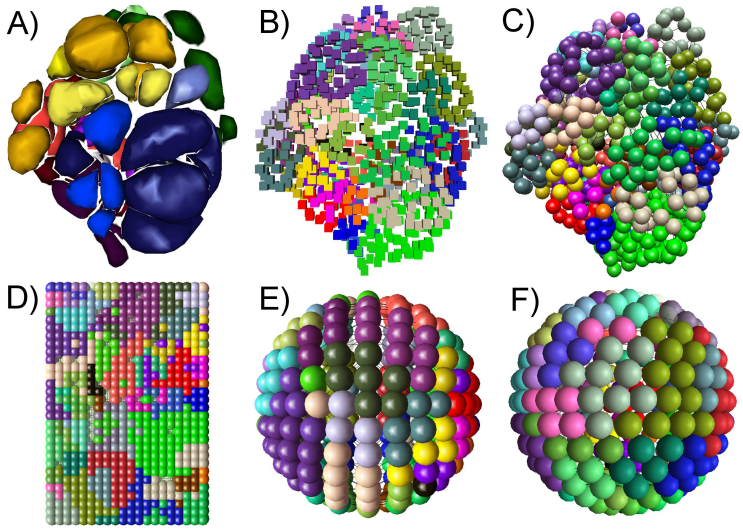

Figure 3.2: *Drosophila*'s antennal lobe and SOM representations thereof. A) The original model of the antennal lobe. Each bulky structure corresponds to one glomerulus. Colors are assigned based on the anatomical location of the glomeruli (see text). B) 1000 surface points extracted from the model. Each glomerulus has been assigned a unique color. C) An SOM after training; spheres represent SOM units. D) Two-dimensional projection by a rectangular 20×30 SOM. E) Representation by a spherical SOM with $f = 30$, front view, F) Back view.

colors indicate central, lateral or anterior positions.

A representative subset of surface points (Figure 3.2B) served as training data for SOM. We trained a planar rectangular SOM with 20×30 neurons and a spherical SOM with $f = 30$ (cf. equation 3.1). A trained SOM mimics the three-dimensional shape of the antennal lobe (Figure 3.2C). By "unfolding" the SOM, i.e. arranging it according to its inherent topology, a planar rectangular (Figure 3.2D) and a spherical representation (Figure 3.2E and F) were obtained.

Each unit of the SOM is colored according to which glomerulus the majority of surface points assigned to it belong to. Each surface point gets assigned to the unit closest to it (the winner neuron n_w, see Methods). Hence, each group of units bearing the same color identifies the same glomerulus.

The unfolded representations facilitate the inspection of neighborhood relationships between single glomeruli. The spherical representation is likely

to deliver more accurate results, due to the congruence between the original, spherical shape of the lobe and the topology of the SOM. On the other hand, to inspect the entire spherical SOM, at least two views are needed, as shown in Figure 3.2E and F. In contrast, the planar rectangular SOM delivers a projections that is likely to have some mapping errors, but can readily be visualized on two-dimensional media such as paper or a screen.

In addition, although the antennal lobe has spherical organization, its real shape rather resembles a semi-sphere. Only its front half (the part which is shown in Figure 3.2A) is populated with glomeruli. The spherical SOM however tries to map the structure to an entire sphere, which is prone to introduce mapping errors. Because of those disadvantages we decided to use the planar rectangular representation of the antennal lobe in the remainder of this study.

3.3.2 Two-dimensional projections of activation patterns

Of 49 glomeruli that have been described by Fishilevich and Vosshall (2005) and Couto et al. (2005), 21 can be mapped to a driving receptor which has been characterized by Hallem and Carlson (2006). Table 3.1 summarizes the mapping we derived from those publications. Olfactory receptors are identified by their genes, while the glomeruli are named by their position. We adopted the naming convention used in the Flybrain database, where "D" means dorsal, "V" ventral, "A" anterior, "L" lateral and "C" central. DA4_e and DL3_e are annotated as extra compartments in the Flybrain database.

There is disagreement on the targeting of the 47b OR: in Fishilevich and Vosshall (2005) VA1m is given as its target, while Couto et al. (2005) state VA1v/l. From the published imaging data, it is not clear to decide which assignment is true. Rather it is possible that slight variations in the experiment or variations between individual flies cause the observed targeting to fluctuate. In any case, the suffixes "m" and "v/l" described the medial and ventral/lateral subdomains of the VA1 glomerulus. We chose to assign it to the medial subdomain, which appears to be the best compromise based on the published data.

Table 3.1: Mapping of receptor genes to glomeruli as described by Couto et al. (2005) and Fishilevich and Vosshall (2005).

OR	Glom.	Remarks
2a	DA4	
7a	DL5	
9a	VM3	
10a	DL1	
19a	DC1	
22a	DM2	
23a	DA3	
35a	VC3l	
43a	DA4_e	extra compartment
43b	VM2	
47a	DM3	
47b	VA1m	VA1v/l in Couto et al. (2005)
49b	VA5	
59b	DM4	
65a	DL3_e	extra compartment
67a	DM6	
82a	VA6	
85a	DM5	
85f	DL4	
88a	VA1d	
98a	VM5	

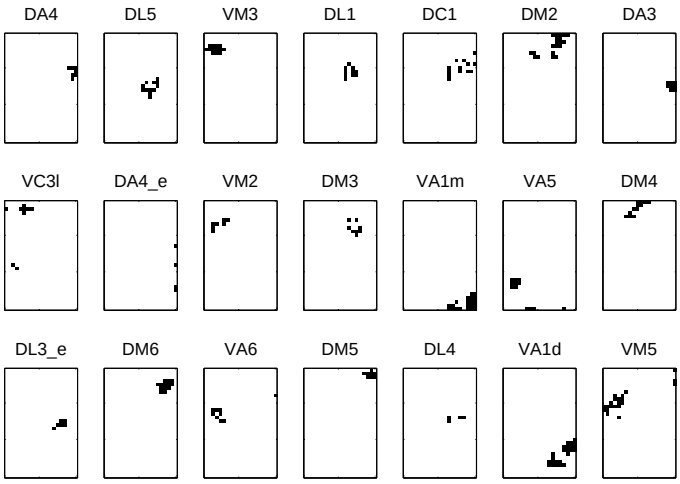

Figure 3.3: The position of 21 glomeruli for which receptors have been characterized by Hallem and Carlson (2006) in the two-dimensional SOM-projection of the antennal lobe.

Figure 3.3 shows the mapping of those 21 glomeruli in the two-dimensional projection provided by the SOM. Due to the two-dimensional projection some glomeruli have been fragmented. This is most prominent for the VC3l, DA4_e and VM5 glomeruli. Such distortions are a consequence of mapping high-dimensional structures to low-dimensional space.

Note that the anatomical arrangement has coarsely been conserved. For example, all dorsal glomeruli are projected inside the upper left quadrant of the map, while ventral glomeruli occupy the remaining regions. For the lateral, anterior and central locations such a regular mapping is more difficult to describe on the map.

3.3.3 Projected activity maps

Using the activity data provided by Hallem and Carlson (2006), and assuming that glomerular activation is identical to receptor activation, we derived two-dimensional projections of glomerular activation.

Figure 3.4 shows some examples for projected activation maps. While 1-hexanol evokes a distributed pattern with a wide range of spike rates (Figure

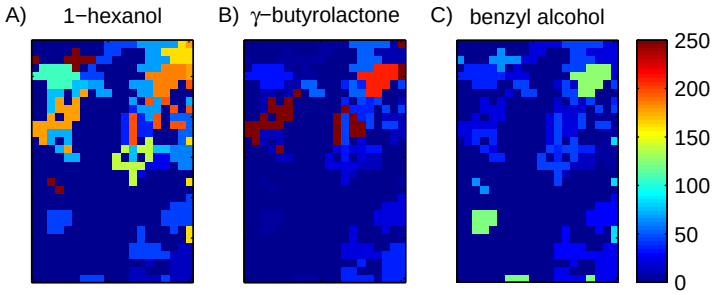

Figure 3.4: SOM-projections of glomerular activity patterns. Color indicates the activation in spikes/s of the respective olfactory receptor neuron innervating the glomerulus. A) 1-hexanol, B) γ-butyrolactone, C) benzyl alcohol.

3.4A), γ-butyrolactone evokes strong activation in four glomeruli (Figure 3.4B). Benzyl alcohol fails to evoke spike rates larger than 100 spikes/s, but activates glomeruli which are not activated by the other two odorants (Figure 3.4C).

These two-dimensional projections also may allow comparisons to activation patterns observed in vertebrates, where the olfactory bulb is more planar. These activation maps are frequently obtained by techniques that image neuronal activity, and the resulting images are two-dimensional (Friedrich and Korsching, 1997; Uchida et al., 2000). Moreover, the mapping of activation patterns with SOMs as presented here may enable comparisons between activation patterns in the secondary olfactory organs of different species.

3.3.4 Analysis of chemotopy

Chemotopy, in the sense that we analyze it here, is the arrangement of glomeruli such that the activation patterns evoked by odorants, when presented in sequence of monotonically changing values of one descriptor, exhibit some kind of directional shift. For example, if molecular weight is represented in the antennal lobe's topology, low-weight odorants would activate regions of the antennal lobe distant from those activated by high-weight compounds, and intermediate-weight compounds would activate glomeruli in between.

In order to visualize the topological representation of a certain descriptor,

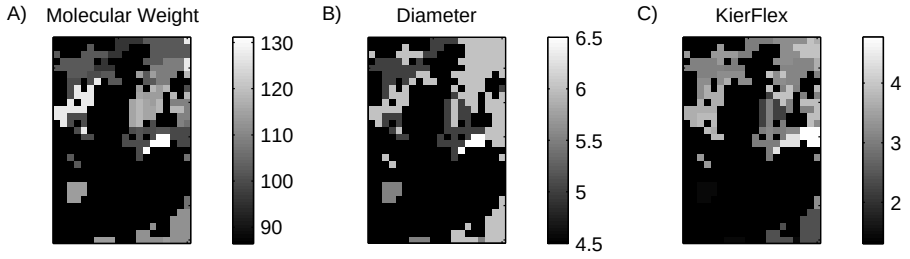

Figure 3.5: Distribution of selected descriptors in the antennal lobe. A) Molecular Weight, B) Diameter, C) KierFlex. Color indicates the median descriptor value for compounds that evoke spike rates above 50 spikes/s in the respective glomerulus. Inactive glomeruli appear in black.

we replaced the glomeruli in the activity map with the median descriptor value of molecules that evoke spike rates above 50 spikes/s in this glomerulus. It is important to note that we used the *absolute* spike rates instead of the baseline-corrected spike rates for activation assessment. Figure 3.5 shows the representation of three selected molecular descriptors in the antennal lobe.

Considering the upper right area of Figure 3.5A, there seems to be a trend for low-weight compounds being represented in the upper right corner, while heavier compounds tend to activate glomeruli more central in the map. However, this is most prominent in the upper half of the map. For example, the two glomeruli that are represented in the center of the map, namely VA6 and DL5 (cf. Figure 3.3) do not follow this trend. A possible explanation for this is that chemotopy on molecular weight may not be a global feature of the antennal lobe, but restricted to a subset of glomeruli.

Diameter describes the largest value in the distance matrix and corresponds to chain length (Petitjean, 1992). This feature also shows a coarse chemotopy (Figure 3.5B), partially supporting the findings described by Couto et al. (2005). Small diameter compounds particularly activate areas in the left half of the map, while large diameter compounds activate glomeruli on the right. The chemotopic arrangement is however by far not as clear as for molecular weight.

The KierFlex descriptor, an index for molecular flexibility defined by Hall

and Kier (1991), provides a third example. We could not detect a clear chemotopic representation in our maps (Figure 3.5C).

It must be noted that the distributions shown in Figure 3.5 vary with the threshold that is applied to determine glomerular activation. Too low a threshold causes the median descriptor values to fluctuate, because then compounds with low activation also contribute to the calculation of the median, possibly introducing noise in the calculation. Hence, the lower the threshold, the closer the descriptor value assigned to a glomerulus get to the real median of descriptors. Since it is difficult to estimate the correct threshold algorithmically, we chose the threshold manually, taking care not to introduce artifacts that may lead to an overestimation of the chemotopic effect.

The validity of the results is also determined by the diversity of the odorant set, in particular on the question if it is representative for *Drosophila*'s olfactory space. The odorant set we used here offers a large variety of chemical classes, but there is no objective means to quantify if it provides a representative sample of the fruit fly's olfactory space.

3.4 Conclusion

We developed SOMMER to train and visualize SOMs of arbitrary topology, and produced mappings of *Drosophila*'s antennal lobe on regular topologies. These mappings can be used to provide two-dimensional images of glomerular activation in response to an odorant. Moreover, the topological mapping of the antennal lobe onto the two-dimensional plane allowed us to observe that some chemical feature such as molecular weight and diameter appeared to be represented in a chemotopic way on parts of this neural structure.

Acknowledgment

Natalie Jäger and Joanna Wisniewska are thanked for careful preparation of a database with the odorant data set from the original publication by Hallem and Carlson (2006). André Brück, Alireza Givehchi, Evgeny Proschak, Kai Scheiffele, Florian Schwarte and Yusuf Tanrikulu are thanked for cooperation in the development of the SOMMER prototype. A manuscript on SOMMER which this chapter is partly based on has been published in the Journal of Molecular Modeling (Schmuker et al., 2007).

Chapter 4

A novel method for processing and classification of chemical data inspired by insect olfaction

The chemical sense of insects has evolved to encode and classify odorants. Thus, the neural circuits in their olfactory system are likely to implement an efficient method for coding, processing and classification of chemical information. In this chapter, we describe a method to process molecular descriptors and classify molecules which is based on neurocomputational principles observed in olfactory systems.

4.1 Background

The mechanisms that enable olfactory discrimination are remarkably similar across species, and even phyla (Hildebrand and Shepherd, 1997; Firestein, 2001).

Several principles of organization hold in insects as well as in vertebrates. One such principle is that each primary olfactory sensory neuron (OSN) specifically expresses one type of olfactory receptor (OR), as has been demonstrated e.g. in mice (Chess et al., 1994; Lomvardas et al., 2006) and in Drosophila (Couto et al., 2005; Fishilevich and Vosshall, 2005), although exceptions to this rule exist (Mombaerts, 2004; Goldman et al., 2005). Notably, ORs are seven-transmembrane G-protein coupled receptors (GPCRs), and state the largest genomic family of GPCRs (Buck and Axel, 1991; Mombaerts, 1999; Gao and Chess, 1999; Clyne et al., 1999).

Araneda et al. (2000) were the first to define ligand selectivity of an OR by the usage of medicinal chemistry techniques. Their findings for the rat I7 receptor have been supplemented by additional studies in vertebrates (Araneda et al., 2004; Mori et al., 2006) and insects (de Bruyne et al., 2001; Stensmyr et al., 2003; Hallem and Carlson, 2006), and even in human cells *in vitro* (Shirokova et al., 2005). A general result of those studies is that one odorant typically activates a number of different ORs, while each OR has rather broad ligand selectivity.

The results we present in chapter 2 also indicated that each receptor appears to analyze a specific part of chemical space, which can be described as a specific combination of features. More abstractly put, each receptor samples a certain region of chemical space. For example, the most potent ligands for the ab3A receptor shared an ester group and carbon chain of length inside a certain interval. In addition, the fact that we were able to develop a predictive model for some ORs using the responses of the other ORs suggested that the OR responses correlate to some extent.

Another characteristic of olfactory systems is that OSNs expressing a specific receptor make synaptic contacts with a defined subset of second-order projection neurons in the antennal lobe in *Drosophila* (Korsching, 2002; Keller and Vosshall, 2003), resp. mitral cells in the olfactory bulb in mice and zebrafish (Korsching, 2001). These connections are formed in spatially discrete areas, the

glomeruli.

It has long been speculated (and in part also shown) that the distribution of glomeruli in the insect antennal lobe and the vertebrate olfactory bulb is ordered such that receptors preferring ligands with similar chemical properties project to nearby glomeruli (Friedrich and Korsching, 1997; Uchida et al., 2000; Meister and Bonhoeffer, 2001; Couto et al., 2005; Johnson et al., 2005). There is some evidence that the distance of glomeruli correlates with the distance between their genomic sequences (Couto et al., 2005). Further, it has been demonstrated that receptor sequence similarity at least in some cases correlates with the chemical properties of preferred ligands (Schuffenhauer et al., 2003; Kratochwil et al., 2005; Keiser et al., 2007).

The chemotopic organization of the secondary structure can be exploited in computational processes. For example, the "contrast" between neighboring glomeruli could be enhanced by lateral inhibition (Cleland and Linster, 2005). In addition, results from a computational study by Linster et al. (2005) demonstrated that inhibition based on response correlation rather than spatial separation performs better at explaining the transformation that occurs in the antennal lobe.

From the secondary structures, olfactory information is passed on to higher brain areas. In mammals, mitral cells from the OB form highly overlapping projections in the piriform cortex (Zou et al., 2001). Similarly, projection neurons in Drosophila's AL send their axons to regions in the lateral horn and the mushroom bodies (Marin et al., 2002; Wong et al., 2002). The lateral horn and the mushroom bodies as well as the piriform cortex receive input from all sensory modalities (Heisenberg, 1998; Roesch et al., 2007). Thus, all information is present here to assign a perceptual quality to a chemical stimulus.

Upon these parallels in organization of neural connectivity, the question arises whether this architecture has properties that make it superior to other coding strategies for chemical information.

4.1.1 A simplified computational model

In a simplified approach, insect as well as vertebrate olfactory systems can be subdivided into three stages of functional organization: In the first stage, OSNs encode the stimulus features into neuronal signals. The second stage decorrelates these signals, optimizing stimulus representation. In the third stage these representations (or *patterns*) are associated with perceptual qualities.

Here, we present a computational model of information processing in the olfactory system that follows this design. By implementing the process of odor quality perception as a machine learning process, we analyze the impact of this three-step architecture on the accuracy of scent quality prediction from molecular structure.

4.2 Methods and data

4.2.1 Source data

The chemical space for this experiment was defined by a set of 836 odorants from the 2004 Sigma-Aldrich *Flavors and Fragrances* catalog (Sigma-Aldrich, 2004). In the "organoleptic properties" section of the catalog, each of the odorants therein is assigned a various number scent qualities, such as *'alliaceous'*, *'fruity'*, *'floral'*, including subclasses such as *'floral (Hyacinth)'*, *'fruity (Banana)'* and the like. One odorant can have more than one scent annotation, e.g. (1R)-(-)-Myrtenol is annotated as smelling *'campherous'*, *'medicinal'*, *'minty'* and *'woody'*, while ethyl 3-hydroxyhexanoate has notes of *'citrus'*, *'citrus (other)'*, *'fruity'*, *'fruity (Grape)'*, *'fruity (Pineapple)'* and *'smoky'*. After removing scents that occur less than five times in the data set, we yielded a total of 66 scent qualities.

4.2.2 Descriptor calculation

Molecules and their odor components were extracted from the Sigma-Aldrich *Flavors and Fragrances* catalog 2004. Using their accession numbers, all compounds were carefully checked for correctness with the machine-readable form of the Sigma catalog. Three-dimensional molecular models were obtained with CORINA (Molecular Networks, Erlangen, Germany), using one conformer per molecule. Partial charges were computed using MOE version 2005.06 (Chemical Computing Group, Montreal, Canada) using the MMFF94x force field (a modified version of MMFF94s (Halgren, 1999)). Prior to descriptor calculation, we performed an additional energy minimization using MOE and the MMFF94x force field, stopping at a gradient of 10^{-4}. Descriptors were calculated using MOE. We used all available two-dimensional (2D) descriptors, resulting in a 184-dimensional descriptor space.

Although only using 2D descriptors, we calculated the three-dimensional models because a molecules' conformation affects the distribution of partial charges, which is relevant for some 2D descriptors.

4.2.3 SOM training

We used SOMMER for SOM training (Schmuker et al., 2007, cf. chapter 3). The molecular descriptors were autoscaled (i.e. scaled to unit variance and zero mean) prior to SOM training. We trained toroidal SOMs with 12×15, 8×12, 5×7, 1×4 and 1×2 units, respectively. Note that the largest representation has approximately the same dimensionality (i.e. 180) as the original 184-dimensional descriptor set. Table 4.1 shows the parameters that we used for SOM training for all variants except the 1×2 SOM, for which we used maximal time $t_{\max} = 100$ and a final neighborhood value $\sigma_f = 0.5$. During training, the descriptor vectors were presented to the SOM in random sequence, one per time step. The training algorithm is described in detail in section 3.2.1, on page 37*f*.

Table 4.1: Parameters for SOM training.

Parameter	Value
Distance Function	Manhattan
$t_{\max}$	70000
σ_i	5.0
σ_f	0.1
λ_i	0.7
λ_f	0.01

4.2.4 Machine learning and performance assessment

We used the Naive Bayes classifier as implemented in the WEKA machine learning suite (Witten and Frank, 2005) for all classification experiments. The Naive Bayes classifier is a probabilistic classifier based on Bayes' theorem. Given a set of feature vectors F with known class adherence C, a conditional model for class adherence can be formulated using Bayes' theorem:

$$p(C|F) = \frac{p(C)p(F|C)}{p(F)}\,. \tag{4.1}$$

Assuming all n elements f_i, $i = 1, \ldots, n$ of the feature vector F are conditionally independent, eq. 4.1 can be rewritten as

$$p(C|F) = \frac{p(C)\prod_i p(f_i|C)}{p(F)}\,. \tag{4.2}$$

Probabilities were estimated assuming a normal distribution for the feature vectors. In practice, the denominator is omitted because it does not depend on C and hence is effectively constant.

Receiver-Operator Characteristic (ROC) curves were generated by arranging compounds by decreasing predicted probability of class adherence and cumulatively calculating rates of false and true positives.

In all classification experiments, the classifier was trained 50 times using 5-fold crossvalidation (leading to a 80/20 data split for training and test data), thus obtaining 50 probabilities for class adherence for each compound. Classi-

fier performance was assessed as the median AUC value of all 50 crossvalidation repetitions.

The assignment of scent is equivalent to a multi-label classification problem if the perceptual qualities (e.g. *'floral'* or *'fruity (Banana)'*) are treated as labels that can be assigned to any odorant. Hence, we trained the classifier separately for the 66 scent classes. In consequence, each of the 66 resulting classifiers would only distinguish between e.g. *'floral'* and not *'floral'*, or *'fruity (Banana)'* and not *'fruity (Banana)'*.

4.3 Results

In this study, we present a computational model mimicking the neurocomputational principles found in the olfactory system and analyze the impact of these principles on scent prediction from molecular structure.

4.3.1 Representing odorants as two-dimensional patterns

In the first step of the model the stimuli were encoded using "virtual receptors". Like olfactory receptors responding to ligands sharing similar properties (cf. chapter 2), a virtual receptor will respond to ligands that occupy the same region in chemical space. Figure 4.1A) illustrates the concept of the virtual receptor: the smaller the distance between an odorant and te virtual receptor in descriptor space, the higher the activation of this virtual receptor will be.

In our model, chemical space was defined by 184 molecular descriptors. Considering an array of n virtual receptors, each receptor has a position described by a coordinate vector p in the m-dimensional descriptor space. The response of a virtual receptor to an odorant should be the larger the smaller the distance between odorant and receptor is. Hence, we defined the response $\mathbf{r}_i$ of the ith receptor ($i = 1, 2, \ldots, n$) to an odorant s as

$$\mathbf{r}_i = 1 - \frac{d(s, p_i) - d_{\min}}{d_{\max} - d_{\min}}, \tag{4.3}$$

with p_i the coordinates of the ith receptor, $d(s, p_i)$ the Manhattan distance (sum of absolute coordinate differences) between s and p_i, $d_{\min}$ and $d_{\max}$ the minimal and maximal distance between any s and p_i. Thus, $\mathbf{r}_i = 0$ if $d(s, p_i)$ is maximal and $\mathbf{r}_i = 1$ if $d(s, p_i)$ is minimal.

The coordinates of the receptors should be chosen such that they cover all relevant parts of chemical space. We used a Self-Organizing Map (SOM) to arrange our virtual receptors in the 184-dimensional descriptor space. This space was defined by the Sigma-Aldrich *Flavors and Fragrances* catalog (Sigma-Aldrich, 2004), a collection of commercially available odorous compounds. The neighborhood-preserving topological organization of the SOM naturally leads to a chemotopic arrangement of its units, such that neighboring units are more similar in their ligand characteristics than units that are more separated in the SOM topology (Figure 4.1B). The pattern of activity can be arranged on a two-dimensional rectangular plane according to the projection that is defined by the SOM topology (Figure 4.1C).

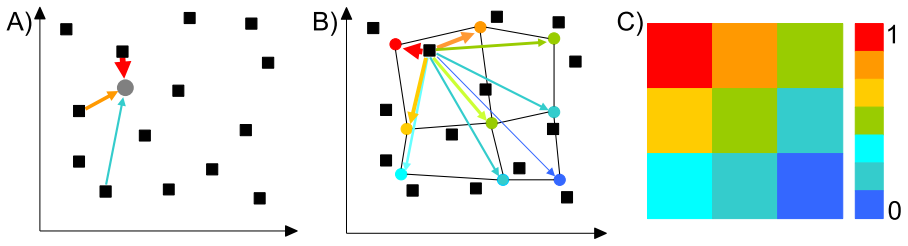

Figure 4.1: Creation of virtual activity patterns: schematic. A) A virtual receptor (gray disc) is defined as a point in chemical space. Arrow color indicates the amount of activation by an odorant (squares). B) Placement of virtual receptors through training of an SOM; lines connecting receptors symbolize neighborhood relationships in the SOM topology. C) Projection of the activity pattern evoked by one odorant to a two-dimensional rectangular plane according to the topological arrangement of the receptors in the SOM. Each rectangle corresponds to one receptor, color indicates amount of activation (see colorbar).

The SOMs we trained had toroidal architecture, and thus can be visualized as two-dimensional grids. Figure 4.2 depicts two odorants (A: butyl pheny-

lacetate, D: butyl levulinate) and the resulting activation patterns (Figure 4.2B and E) for a 12×15 SOM. Due to the toroidal grids, the upper and lower edges of the patterns are connected, as are left and the right edges.

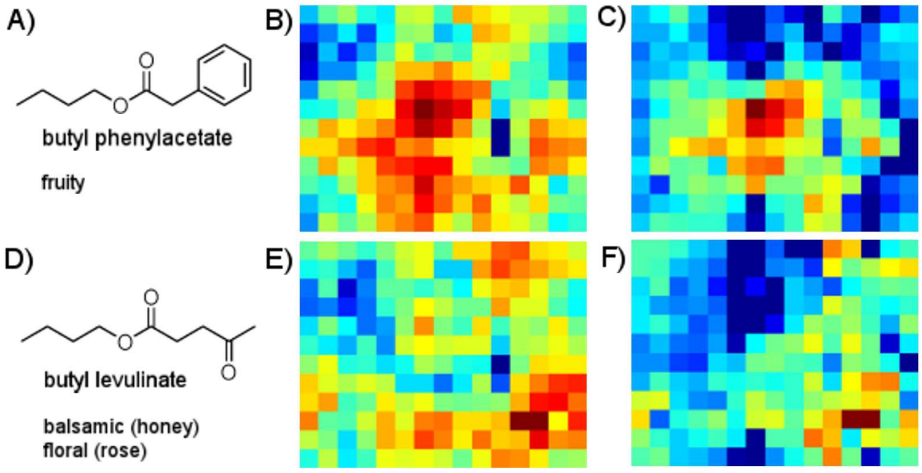

Figure 4.2: Two example odorants (A: butyl phenylacetate, D: butyl levulinate), their corresponding patterns before correlation-based filtering (B and E) and after the filtering (C and F). Red corresponds to maximal, blue to minimal activation.

Most of the patterns showed multiple 'islands' of high activation, thus most odorants activated several units that are not necessarily neighbors on the SOM grid. This reflects that the SOM corresponds to a manifold rather than a hyperplane in the descriptor space, i.e. it is 'folded' and not planar. In part, this is certainly due to the toroidal structure of the SOM, but may also be a consequence of the neighborhood structure in odorant space.

4.3.2 Transformation in the antennal lobe

In the scope of the model, the activation patterns correspond to activations of glomeruli in the antennal lobe. Linster et al. (2005) suggested that processing in the antennal lobe implements correlation-based lateral inhibition. That is, if two glomeruli are activated by a highly overlapping set of ligands, they will inhibit each other's response. This enforces a 'winner takes most'-situation, such

that the glomerulus with the stronger response will inhibit the response of the weaker glomerulus, effectively making their output more dissimilar. The more correlated the firing patterns of the two glomeruli are, the more pronounced this effect will be.

To account for this, we computed the post-lobal pattern vector $\mathbf{r}'$ from the pre-lobal input vector $\mathbf{r}$ (cf. eq. 4.3) by equation 4.4

$$\mathbf{r}' = \mathbf{r} - q\left(\frac{\mathbf{C}\cdot\mathbf{r}^T}{n}\right), \tag{4.4}$$

with n the number of virtual receptors, q an arbitrary weight and $\mathbf{C}$ a matrix where $\mathbf{C}_{i,j}$ contained Pearson's correlation coefficient between the responses of the ith and jth receptor. In addition, all negative elements as well as all elements on the diagonal of $\mathbf{C}$ were set to zero. Figure 4.2 C and F show the post-lobal response patterns.

The most salient difference between the patterns is that there is less overall activation. Also, the sites of highest activation remain unchanged (in the center in Figure 4.2C and in the lower left in Figure 4.2F), while large portions of the remaining pattern get sparser (i.e., show less activity).

In order to analyze the effect of receptor count, we trained SOMs of different sizes ranging from 2 to 180 units. Figure 4.3 shows patterns from different SOM sizes. With increasing resolution the peaks in the activation landscape become more distinct. While the higher-dimensional patterns may be visually more appealing, the question remains if they actually contain more information about the stimulus than their lower-dimensional counterparts. We address this issue in the next section.

4.3.3 Retrospective scent prediction from virtual receptor activation patterns

We performed a retrospective scent prediction experiment in order to examine the information content conveyed by the patterns. We used odor annotations to

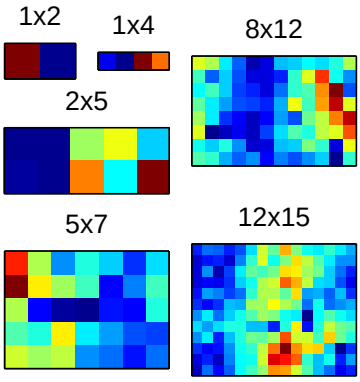

Figure 4.3: Activation patterns produced by (-)-Carvyl propionate for different SOM sizes. Red corresponds to maximal, blue to minimal activation.

836 odorants from the 2004 Sigma-Aldrich *Flavors and Fragrances* catalog (Sigma-Aldrich, 2004) as targets to train a Naive Bayes classifier.

After removing scents that occurred less than five times in the data set, we obtained a total of 66 scent qualities. We trained the classifier separately for each scent class, hence each of the 66 resulting classifiers would only distinguish between e.g. *'smoky'* and not *'smoky'*, or *'fruity (Banana)'* and not *'fruity (Banana)'*. Of all scents, the *'fruity'* annotation was most frequent, with 319 out of the 836 compounds bearing this attribute.

The model has two free parameters that both affected classification performance: the SOM size (i.e. receptor count) and q, the weight of correlational inhibition (cf. eq. 4.4). To illustrate the impact of q, we trained the classifier on *'fruity'* scents, using the 12×15 patterns as input and varied q between zero (i.e. no processing) and two. Figure 4.4A shows the distribution of ROC curves for classification of *'fruity'* scents from the 50 crossvalidation runs using the patterns generated with the 12×15 SOM layout. In both the best and the worst cases (Figure 4.4B and C) classification was best for $q = 2$ (best Area-Under-Curve (AUC) $= 0.82$, worst $= 0.75$, median $= 0.79$), followed by $q = 1$ (best AUC $= 0.78$, worst $= 0.72$, median $= 0.75$) and $q = 0$ (best AUC $= 0.75$, worst $= 0.68$, median $= 0.72$). Although the $q = 0$ and $q = 1$ patterns yielded similar classification power for low false-positive rates in the best case (Figure 4.4B), the filtered patterns yielded an overall better performance than the unfiltered representations. In the worst case (Figure 4.4C) $q = 2$ and $q = 1$ were

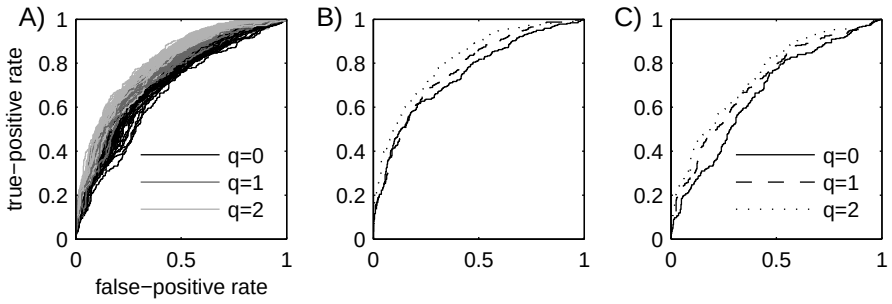

Figure 4.4: ROC curves for classification of the 'fruity' scent and 12×15 SOM layout for unprocessed patterns ($q = 0$), and patterns processed by correlational inhibition with $q = 1$ and $q = 2$. A) overlay of all ROC curves generated during cross-validation, B) best ROC, C) worst ROC.

on par for small positive rates, with $q = 2$ yielding the higher total AUC.

This trend is also apparent when comparing classification performance for all 66 scents against q and SOM architecture, as Figure 4.5 shows. We used the median AUC from all crossvalidation runs as performance indicator. Generally, patterns generated with $q = 2$ outperformed $q = 0$ and $q = 1$ for almost all architectures. For the 12×15 representations, the median AUC values were 0.68 ($q = 2$), 0.65 ($q = 1$) and 0.63 ($q = 0$). These differences are significant (Wilcoxon rank sum test, $p < 10^{-7}$).

Performance also gradually decreased with dimensionality, but only 2×5 and smaller representations significantly differ in their median AUC values from the larger representations (Wilcoxon rank sum test, $p < 0.05$). Hence, overall classification performance did not suffer from a reduction of dimensionality by a factor of five.

4.3.4 Correlation-based vs. distance-based inhibition

As Linster et al. (2005) discuss, another plausible mechanism for lateral processing in the antennal lobe is to organize interglomerular inhibition by distance. For comparison, we also analyzed the "classifyability" of scents (i.e. the performance in retrospective screening) when using distance-based inhibition.

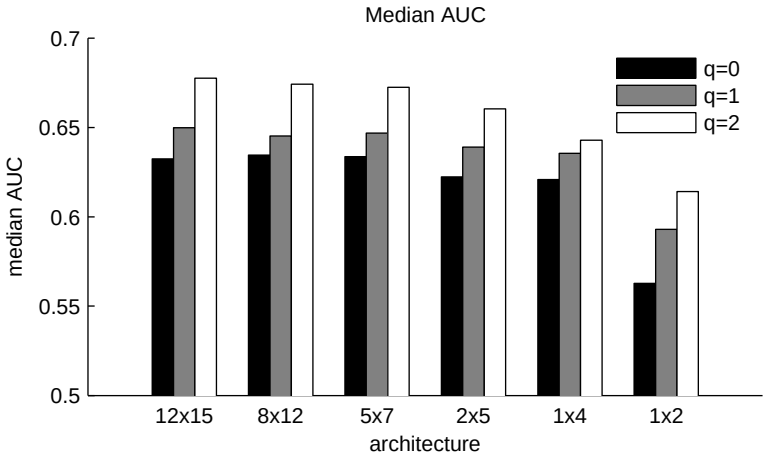

Figure 4.5: Median AUC values for scent prediction using unprocessed patterns ($q = 0$), and patterns processed by correlational inhibition with $q = 1$ and $q = 2$. Ordinate truncated to emphasize differences.

Similar to equation 4.4, we defined

$$\mathbf{r}' = \mathbf{r} - q\left(\frac{\mathbf{D} \cdot \mathbf{r}^T}{n}\right), \tag{4.5}$$

with $\mathbf{D}$ the relative distance matrix. $\mathbf{D}_{i,j}$ contains the relative distance between the ith and jth SOM unit, calculated according to eq. 4.6:

$$\mathbf{D}_{i,j} = -1 \cdot \frac{d_{\text{toro}}(i,j) - d_{\text{toro,min}}}{d_{\text{toro,max}} - d_{\text{toro,min}}}, \; i \neq j, \tag{4.6}$$

where d_{toro} denotes the euclidian distance on the toroidal surface on which the SOM units are arranged, $d_{\text{toro,min}}$ and $d_{\text{toro,max}}$ the minimum and maximum distance. Thus, $\mathbf{D}$ is 1 where distance is minimal, and 0 where distance is maximal. In addition, all elements on the diagonal of $\mathbf{D}$ were set to zero.

We compared median AUC values for scent prediction using patterns processed by distance-based inhibition with predictions using correlation-based inhibition for patterns. Figure 4.6 shows the results: there is virtually no increase in prediction performance by distance-based processing. This somewhat surprising result strongly argues for the hypothesis that correlational in-

hibition is a more effective mechanism to shape the input signals than distance-based inhibition (Linster et al., 2005).

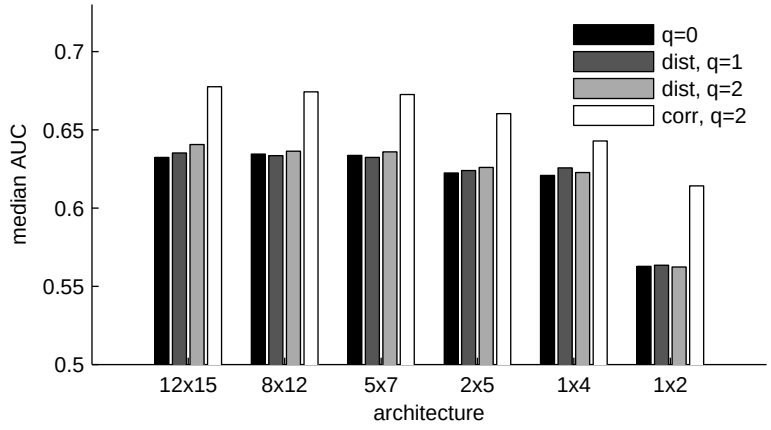

Figure 4.6: Median AUC values for scent prediction using patterns processed with distance-based inhibition, with q=1 and q=2. For comparison, AUC values for correlation-based inhibition (q=2) are also given.

4.3.5 Analysis of decorrelation

In the introduction to this thesis, we mentioned that the antennal lobe is said to decorrelate the receptor signals. In the present work, this decorrelation is achieved through mutual inhibition between projection neurons that innervate glomeruli with correlated response patterns. In order to quantify the amount of decorrelation we calculated the residual correlation between virtual receptors before and after correlational inhibition.

Figure 4.7 depicts correlation matrices indicating the residual correlation between all virtual receptors. While in the unprocessed receptor responses there was high residual correlation (Figure 4.7A, mean correlation $= 0.38$), it gradually decreased with increasing q; For $q = 1$ the mean residual correlation is 0.24 (Figure 4.7C), while it decreased to 0.02 for $q = 2$ (Figure 4.7E). The effect of decorrelation can be observed more clearly in the histograms that depict the distribution of correlation coefficients in the matrices (Figure 4.7B,D,F): The

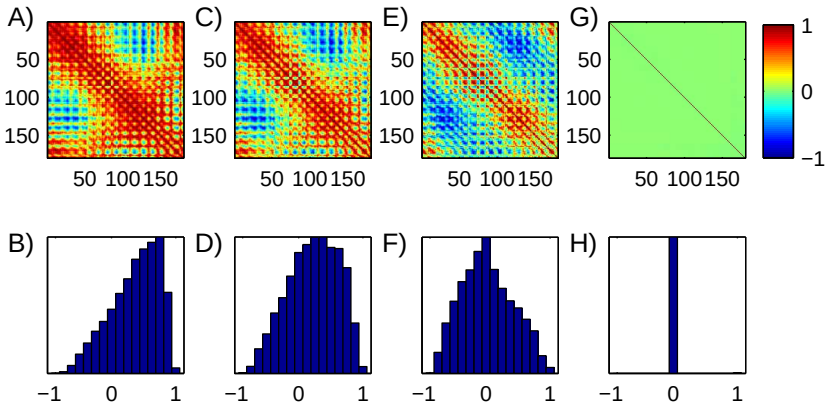

Figure 4.7: Correlation matrices (upper row) and their distribution histograms (lower row) for virtual receptor responses and PCA scores. A,B) $q = 0$, C,D) $q = 1$, E,F) $q = 2$, G,H) PCA. Histograms have been scaled to identical maximum count.

peak of the distribution shifts towards zero as q increases.

Figure 4.7G shows the residual correlation after applying Principal Component Analysis (PCA) to the original descriptor data. PCA is frequently used for dimensionality reduction prior to training machine-learning classifiers. The dimensions produced by PCA are orthogonal and have no residual correlation, which is illustrated by the single peak at zero for the correlation histogram in Figure 4.7H.

PCA thus achieves maximum decorrelation on the data, but it is not clear beforehand if the resulting patterns are also better suited for classification. To investigate whether maximum decorrelation also corresponds to maximal classification performance, we compared the classification performance of our method with the performance that can be achieved using PCA for dimensionality reduction on the original descriptor set. Figure 4.8 shows median AUC values from retrospective classification using patterns processed with $q = 2$ and using the first n principal components of the original dataset that explained most variance. We chose the dimensionality of the reduced data to match the dimensionality of the patterns.

The correlation-filtered patterns yielded higher AUC values for higher di-

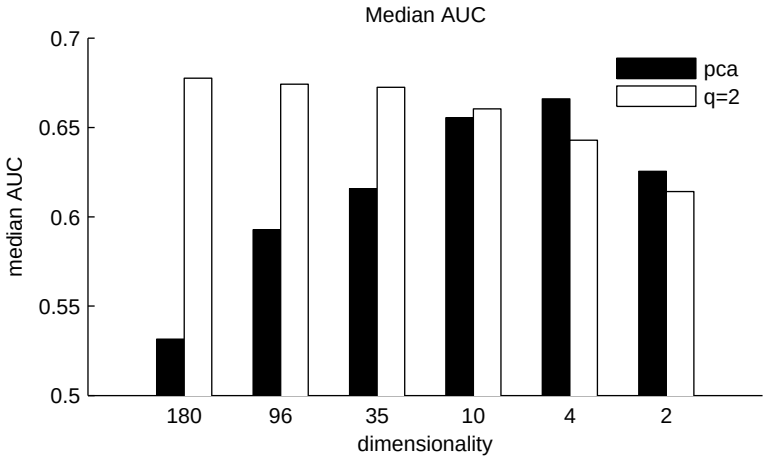

Figure 4.8: Median AUC values for scent prediction with correlation-filtered patterns and PCA-transformed representations of the original feature space.

mensionalities (max: 0.68 using 180 dimensions), while the principal components seemed to work best for low dimensionalities (max: 0.67 using 4 dimensions). The difference between the AUC values is significant ($p < 0.05$, Wilcoxon rank sum test). A possible explanation for this behavior is that when using PCA, the dimensions explaining less variance introduce noise. The Naive Bayes classifier has no means to distinguish "noise" variables from those that explain a high amount of variance, and thus produces inaccurate results.

Notably, when using the original 184-dimensional descriptor set without dimensionality reduction to train the classifier, we also obtained a median AUC value of 0.67 (data not shown). In this case, despite its high dimensionality, the classifyability of the data did not suffer.

In conclusion, maximum decorrelation by PCA does not necessarily increase classification performance. PCA may be the method of choice for data preprocessing if dimensionality reduction is important, e.g. if computational resources are limited, but care must be taken not to use too many principle components for data representation. In real brains, due to their highly parallel architecture data dimensionality may not be the limiting factor. Rather, robustness to noise and capacity of the code may be more important. The lat-

ter is provided by the higher dimensionality of the proposed coding scheme, while robustness is increased by the residual redundancy due to non-absolute decorrelation.

4.3.6 Application to pharmaceutical data

Pattern recognition and -classification on chemical data is not only important for studying olfaction, but also for virtual screening in pharmaceutical applications. In this process, regression models or machine-learning classifiers are trained on activity data for a certain pharmaceutical target in order to predict the activity of novel compounds.

We tested if the method we propose is also suited for pharmaceutical data. Chemical space was given by the COBRA database (Schneider and Schneider, 2003), version 6.1. In analogy to the procedure stated above, we placed virtual receptors by training SOMs of various dimensionality, using the same parameters as for the Flavors & Fragrances data (cf. Table 4.1). Virtual receptor patterns were derived following equation 4.3 and processed according to equation 4.4. Naive Bayes classifiers were trained on the patterns using pharmaceutical activity at 115 targets (e.g. Cyclooxygenase 2, Thrombin, mGluR5) and their superclasses (e.g. Enzyme, GPCRs, Ion Channels). We repeated 5-fold crossvalidation only 10 times (in contrast to 50 times above) in order to save computing time. For comparison, we also trained the classifiers on the original descriptors processed by PCA.

Figure 4.9 shows the median AUC values for the pharmaceutical data. The results slightly differ from those obtained for the *Flavors & Fragrances* catalog. First, overall performance on the COBRA data set was better than on the *Flavors & Fragrances* data. Second, principal components outperformed the virtual activation patterns in terms of classification performance. Third, the patterns processed with $q = 2$ were not always performing best.

The first two points may be a consequence of the fact that many of the original descriptors we employed have been optimized towards good modeling

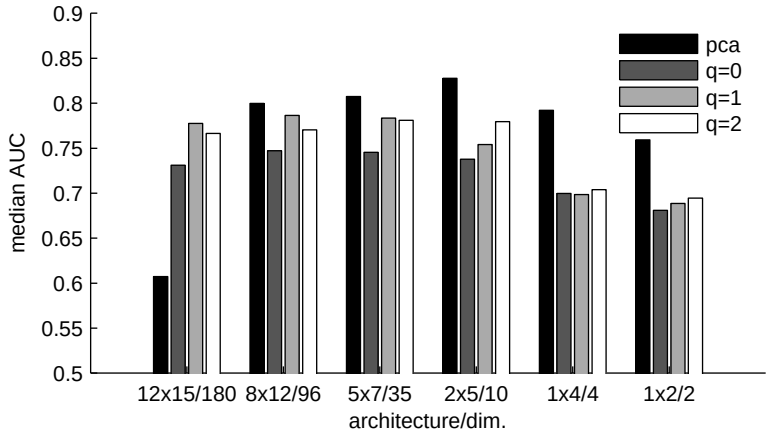

Figure 4.9: Median AUC values for the classification of pharmaceutical targets.

properties for pharmaceutics. Odorants are typically smaller than pharmaceutics and have considerably different chemical properties, their volatility being probably among the most obvious. Moreover, the chemistry behind both substance classes is fairly different (Grabowski and Schneider, 2007): Many pharmaceutical compounds have been tuned towards easy synthesizability on an industrial scale, while odorants typically are natural products, emerging as secondary metabolites or decay products. Hence, the gap in classification performance compared to odorant data is not surprising.

The fact that patterns processed with $q = 2$ were not always classified best requires a more thorough look at the results. Considering patterns with a dimensionality of up to 2×10, lower settings of q did not perform better than patterns processed with $q = 2$. Only for higher dimensionalities, the patterns processed with $q = 1$ performed better than those with $q = 2$. An analysis of the filtered patterns revealed that for $q = 2$ and higher dimensionalities many virtual receptor responses got set to zero, because the subtractive term in equation 4.4 became equal or larger than $\mathbf{r}_i$ (data not shown). Thus, only those virtual receptor signals with highest activation 'survived' the functional inhibition process, effectively replacing the soft winner-takes-most situation to a hard winner-takes-all one. Clearly, this result points out the need for q to be

adjusted in order to obtain best results. It also shows that there is not one optimal setting of q, but rather that this optimum depends on the application and the data.

4.4 Discussion

Scope of the model

The processing scheme we present here provides a very simplified model of neural computation in the olfactory system. Our focus was on providing a framework that enabled us to study certain aspects of computational principles, instead of trying to build a biologically accurate simulation of the olfactory system. We tried to keep the simulation overhead as small as possible, so that the essence of the processing strategies would stay obvious. More realistic models, in terms of biological plausibility, are particularly useful when one tries to answer more biological questions, like e.g. Huerta et al. (2004) and Nowotny et al. (2005) did.

Classifier choice

Although other classifiers, such as Artificial Neural Networks or Associative Memory classifiers may appear a more natural choice for modeling brain functions (Rolls and Treves, 1997; Haberly, 2001), the Naive Bayes classifier has the advantage that there are no free parameters that can affect the prediction quality. Despite the fact that the "naive" assumption of independence between features is often inaccurate, this classifier has proven to work well in real-world learning paradigms (Bender et al., 2004). Besides, it has been demonstrated how Bayesian classifiers can be implemented in neural structures (Barber et al., 2003).

Performance of scent prediction

The pure retrospective character of this study makes it difficult to assert how well scent prediction would work in "real life". For example, although cross-validation and multiple repetitions were used to assess prediction performance, prospective results may be worse than indicated by this study, especially when different scent annotation protocols were used for the training data. Predictions will be best for data from the same domain/source. For new sources, training data from that domain would be required to achieve best results.

The quality of the training data may also be an issue, since the protocol for scent assignment in our data source is not known (and therefore not reproducible). Only vague reports exist on how the labels were derived (Zarzo and Stanton, 2006). Further, there is also no guarantee for the reliability of the labeling, e.g. it is not clear if *all* scent notes are given for every odorant. Hence, it is possible that the data set contains a certain amount of false-negatives. Even using an ideal classifier, i.e. one that predicts the scent of any odorant with 100% accuracy, these odorants would show up as false-positives, since not all scent notes which the classifier (correctly!) predicts are annotated.

Another point that must be noted here is that we trained only one SOM per architecture. For proper crossvalidation, this stage should also be repeated on separate data folds. This would require the introduction of an additional cross-validation layer wrapped around the derivation of virtual activity patterns and classifier training. The high computational requirements for this task prohibited this systematic analysis.

In consequence, we use the prediction performance as a *relative* measure to compare different mechanisms for processing in the antennal lobe. The results should not be interpreted as providing an actual prediction method for scent. This may change with the availability of high-quality data, and means of objective testing of the predictions. Then, solid conclusions on the performance of scent prediction can be made.

Outlook

The q factor showed to have a large impact on the outcome of the correlational filtering step, with classifyability of the patterns improving for rising q, up to a certain level. Finding optimal values of q for a given data set may be a worthwhile topic for further research. Possible approaches include the use of meta-optimization techniques to derive q empirically, as demonstrated by Meissner et al. (2006) for the number of hidden neurons in an Artificial Neural Network, or its estimation from statistical properties of the data, like variance or cross-correlation.

Among the questions we did not address here are the effects of odorant concentration and combinations of odorants (mixtures). For both, we can suggest straightforward implementations: Odorant concentration could be implemented via "gain control" of the activation patterns, i.e. multiplication of the pattern with a concentration-dependent scalar, while odorant mixtures could be represented by additive or even nonlinear combinations of their activation patterns. The effects of processing in the virtual antennal lobe on those extensions as well as their impact on classification power provide a tantalizing prospect for future research.

4.5 Conclusion

We have presented a computational framework that implemented processing principles observed in olfactory systems. This method effectively captured relevant properties of the original data that allowed a machine-learning classifier to learn odorant classification. Besides reducing dimensionality of the original data, it also exhibited robustness against overdetermined representations, a situation where principal components of the original data failed. In addition, the application of this framework is not limited to the olfactory domain, but can also be used for virtual screening in a pharmaceutical compound database.

Acknowledgments

Volker Majczan is thanked for assistance in the preparation of the odorant database from the Sigma-Aldrich catalog. Natalie Jäger and Joanna Wisniewska have done preliminary research on odorant classification that was helpful in the design of the experiments described here.

Chapter 5

Conclusion and outlook

In this work, we aimed at analyzing the coding and processing principles in the olfactory system in order to gain knowledge on efficient processing of chemical information. The highly interdisciplinary character of this goal is reflected in the methods we employed to reach it, which combined cheminformatics with neurobiology and machine learning.

Functional characterization of olfactory receptors

Olfactory receptors work at the interface between the chemical world of volatile molecules and the perception of scent in the brain. Their main purpose is to translate chemical space into information that can be processed by neural circuits. Assuming that these receptors have evolved to cope with this task, the analysis of their coding strategy promises to yield valuable insight in how to encode chemical information in an efficient way.

In chapter 2, we analyzed olfactory coding by modeling responses of primary olfactory neurons to small molecules using a large set of physicochemical molecular descriptors and Artificial Neural Networks. We then tested these models by recording receptor neuron responses to a new set of odorants and successfully predicted the responses of five out of seven receptor neurons. Correlation coefficients ranged from 0.66 to 0.85, demonstrating the applicability

of our approach for the analysis of olfactory receptor activation data.

In addition, we demonstrated that the molecular descriptors which are best suited for response prediction vary for different receptor neurons, implying that each receptor neuron detects a different aspect of chemical space. The chemical meaning of these descriptors helps understand structure-response relationships for olfactory receptors and their "receptive fields". Finally, we demonstrated that receptor responses themselves can be used as descriptors in a predictive model of neuron activation, indicating that olfactory receptors encode chemical space in a way that can be exploited to predict receptor-ligand affinities. Moreover, this result suggests that coding at the receptor level is not decorrelated, but partly redundant.

Future research in this area will certainly benefit from a growing amount of olfactory receptor response data. The accuracy of both the preferred ligand features and activation predictions is likely to increase when receptor neuron response data for a greater variety of odorants becomes available. A greater data basis may also allow for quantitative models of receptor neuron activation.

Modeling the insect antennal lobe with self-organizing maps

One of the most intriguing features of the olfactory system lies in the stereotypic anatomical organization of the second processing stage, namely the antennal lobe in insects and the olfactory bulb in vertebrates. Several lines of evidence suggest that it implements a chemotopic spatial ordering, in that similar chemotypes activate nearby regions in this neural structure. This phenomenon provides an intriguing possibility to investigate how chemical similarity is defined in nature.

Our goal was to investigate the chemotopic organization of the antennal lobe with Self-Organizing Maps (SOMs). SOMs provide a topological mapping of the input data, which makes them particularly useful in this scenario. They enable straightforward, algorithmically well defined and reproducible

projection of the three-dimensional arrangement of glomeruli onto the two-dimensional plane.

For this purpose we developed SOMMER, the Self-Organizing Map Maker for Education and Research. SOMMER provides rectangular, toroidal, cubic and spherical SOM topologies and is a valuable, multi-purpose research tool to create SOMs from all kinds of data. In chapter 3 we provide an overview on the functionality that SOMMER provides and demonstrate the use of SOMs to produce two- and three-dimensional mappings of *Drosophila*'s antennal lobe.

We used the two-dimensional projections to derive maps of glomerular responses to odorants. In those activation maps, each glomerulus was assigned the spike rate response of the receptor neuron that states its main input. We produced such maps for receptor neuron responses to a set of 110 odorants.

These maps enabled us to discern the preferred chemical features that are represented in a glomerulus. We calculated a set of 184 physicochemical descriptors for each of the odorants and derived the median value of each descriptor in each glomerulus, taking into account all odorants that activate this glomerulus. Since each descriptor corresponds to a molecular property, we obtained a map that reflects the distribution of molecular features on the antennal lobe.

The analysis of this map revealed a clear trend for a chemotopic representation of molecular weight. A chemotopic ordering was also observed for molecular diameter (which is related to chain length), albeit to a lower extent. In contrast, no ordering was apparent for molecular flexibility.

The SOMMER application can provide the basis for several subsequent research projects. It is publicly available in binary and source code and can easily be adapted to current research needs. For example, arbitrary topologies can easily be implemented in SOMMER, making it an extremely versatile research tool. In this work we preferred the two-dimensional rectangular projection, but it will be interesting to investigate how SOMMER's spherical SOM topology performs in projections of the antennal lobe and other neural structures.

A novel method for processing and classification of chemical data inspired by insect olfaction

Chapter 4 provides a simple computational model for the entire olfactory system that builds upon the findings and the software described in chapters 2 and 3. Our goal in this part of the thesis was to design a method to process molecular descriptors and classify molecules, based on neurocomputational principles observed in the olfactory system.

In the framework we presented, we mimicked the three-stage architecture of the olfactory system and the processing schemes that are realized therein. Our main focus hereby was on providing a model that enabled us to study certain aspects of computational principles, rather than providing a biologically accurate simulation.

The first stage in olfactory processing is modeled by "virtual receptors", which are defined as discrete points in odorant space. The odorant space is set up by 184 physicochemical descriptors that reflect the chemical features of the odorants. The magnitude of the response of a virtual receptor to an odorant depends on their distance in the 184-dimensional odorant space: The closer the odorant is to the receptor, the higher the response.

The positions of virtual receptors were derived by training an SOM in odorant space. The coordinates of the virtual receptors were then obtained from the prototype vectors of the trained SOM. Since SOMs preserve the local topology of the input space, we obtained a chemotopic representation of the odorant space, much like the one observed in the antennal lobe or the olfactory bulb.

We implemented the decorrelation step in the antennal lobe as correlation-based lateral inhibition, where the response from one receptor is decreased by the weighted average of all other responses, where the weight is defined by the amount of their correlation. This creates a winner-takes-most situation, where the competition between two receptor signals is strongest if they correlate most. In addition, we introduced a scaling factor q which allowed us to regulate the overall weight of correlational inhibition.

The last step in our simplified model of olfactory computation consisted in assigning a perceptual quality to the input. We achieved this by training a Naive Bayes classifier on the odorant's annotated scents using the processed signals as input.

We tested the performance of our model by retrospective screening of an odorant database. Prediction accuracy was quantified by the Area Under the receiver-operating-characteristic Curve (AUC). The results showed that the representation of chemical information in our model is suitable to perform this task. We achieved median AUC values over all scent qualities of up to 0.72 for the unprocessed virtual activation patterns, depending on the number of virtual receptors.

Processing the patterns by correlational inhibition had a favorable impact on classification performance: For example, for the largest number of virtual receptors, the median AUC values increased to 0.75 for $q = 1$ and 0.79 for $q = 2$. The same trend could be observed for smaller numbers of receptors.

We demonstrated that this processing method effectively performs a "moderate decorrelation" of the input patterns, in that there remained residual correlation between some dimensions of the output. This is in contrast to the result one obtains from methods like principal component analysis (PCA), which produce uncorrelated output.

A comparison of classification performance between fully decorrelated patterns obtained with PCA and moderately decorrelated patterns generated by correlational inhibition revealed that both methods reach similar AUC values, although for different dimensionalities. Patterns transformed by PCA performed best for low dimensionalities, while the performance of those transformed with correlational inhibition reached their maximum performance for higher dimensionalities. Moreover, while the performance of the PCA-transformed patterns decreased quickly with increasing dimensionality, those processed with correlational inhibition seem to saturate in their performance when dimensionality was increased. This result indicates that this processing method

is more robust against overdetermined data sets, and particularly effective when data redundancy and robustness is preferred over low dimensionality, like for environments such as the brain where data is processed in a highly parallel manner.

Finally, we demonstrated that the application of this processing method is not limited to the olfactory domain by performing virtual screening in a pharmaceutical database.

Since the weight of correlational inhibition (expressed by the factor q) crucially influences the outcome of the processing scheme, finding an optimal setting for it may provide one starting point for future research. Another direction may be to incorporate odor concentration and mixtures of odors into the analysis.

Appendix

A.1 Molecular descriptors used for SAR

Table A.1 explains the meaning of the descriptors (adapted from the MOE user manual (Chemical Computing Group, Montreal, Canada)). Some descriptors occur in several variants, depending on the theory or algorithm underlying their calculation. For example, charge distribution for descriptors prefixed with Q was calculated using the MMFF94x force-field (Halgren, 1999), while those prefixed with PEOE are based on calculations with the Partial Equalization of Orbital Electronegativities (PEOE) method proposed by Gasteiger and Marsili (1980).

The following conventions are used in the table: n: the number of atoms (not counting hydrogens); m: the number of bonds (except bonds to hydrogen atoms); a: the sum of $(r_i/r_c - 1)$ where r_i is the covalent radius of atom i, and r_c is the covalent radius of a carbon atom; p_2: the number of paths of length 2 and p_3 the number of paths of length 3 in the molecular graph.

Table A.1: Molecular descriptors and their meaning.

Descriptor	Meaning
AM1_HOMO	Energy (eV) of the Highest Occupied Molecular Orbital calculated using the MOPAC AM1 Hamiltonian (Stewart, 1993).
AM1_IP	Ionization potential (kcal/mol) calculated using the AM1 Hamiltonian (Stewart, 1993).
AM1_LUMO	Energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the MOPAC AM1 Hamiltonian (Stewart, 1993).
E	Value of the potential energy.
E_str	Bond stretch potential energy.
FASA+	Fractional ASA+ calculated as ASA+ / ASA.
FASA-	Fractional ASA- calculated as ASA- / ASA.
FCASA+	Fractional CASA+ calculated as CASA+ / ASA.
FCASA-	Fractional CASA- calculated as CASA- / ASA.
Kier2	Second kappa shape index: $(n-1)^2/m^2$ (Hall and Kier, 1991).
Kier3	Third kappa shape index: $(n-1)\cdot(n-3)^2/p_3^2$ for odd n, and $(n-3)\cdot(n-2)^2/p_3^2$ for even n (Hall and Kier, 1991).
KierA1	First alpha modified shape index: $s\cdot(s-1)^2/m^2$ where $s=n+a$ (Hall and Kier, 1991).
KierA2	Second alpha modified shape index: $s\cdot(s-1)^2/m^2$ where $s=n+a$ (Hall and Kier, 1991).
KierA3	Third alpha modified shape index: $(n-1)\cdot(n-3)^2/p_3^2$ for odd n, and $(n-3)\cdot(n-2)^2/p_3^2$ for even n where $s=n+a$ (Hall and Kier, 1991).
KierFlex	Kier molecular flexibility index: (KierA1)·(KierA2)/n (Hall and Kier, 1991)
MNDO_HF	Heat of formation (kcal/mol) calculated using the MNDO Hamiltonian (Stewart, 1993).

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
MNDO_HOMO	Energy (eV) of the Highest Occupied Molecular Orbital calculated using the MNDO Hamiltonian (Stewart, 1993).
MNDO_IP	Ionization potential (kcal/mol) calculated using the MNDO Hamiltonian (Stewart, 1993).
PEOE_PC+	Total positive partial charge: the sum of the positive partial charges
{Q, PEOE}_RPC+	Relative positive partial charge: the largest positive q_i divided by the sum of the positive q_i
{Q, PEOE}_RPC-	Relative negative partial charge: the smallest negative q_i divided by the sum of the negative q_i
{Q, PEOE}_VSA+0	Sum of v_i where q_i is in the range [0.00,0.05)
{Q, PEOE}_VSA+5	Sum of per-atom van der Waals surface v_i where q_i is in the range [0.25,0.30)
{Q, PEOE}_VSA-1	Sum of v_i where q_i is in the range [-0.10,-0.05)
{Q, PEOE}_VSA_FHYD	Fractional hydrophobic van der Waals surface area
{Q, PEOE}_VSA_FNEG	Fractional negative van der Waals surface area
{Q, PEOE}_VSA_FPNEG	Fractional negative polar van der Waals surface area
{Q, PEOE}_VSA_FPOL	Fractional polar van der Waals surface area
{Q, PEOE}_VSA_FPOS	Fractional positive van der Waals surface area
{Q, PEOE}_VSA_FPPOS	Fractional positive polar van der Waals surface area
{Q, PEOE}_VSA_HYD	Total hydrophobic van der Waals surface area
{Q, PEOE}_VSA_NEG	Total negative van der Waals surface area
{Q, PEOE}_VSA_PNEG	Total negative polar van der Waals surface area
{Q, PEOE}_VSA_POL	Total polar van der Waals surface area
{Q, PEOE}_VSA_POS	Total positive van der Waals surface area
{Q, PEOE}_VSA_PPOS	Total positive polar van der Waals surface area
PM3_HOMO	Energy (eV) of the Highest Occupied Molecular Orbital calculated using the PM3 Hamiltonian (Stewart, 1993).

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
PM3_IP	Ionization potential (kcal/mol) calculated using the PM3 Hamiltonian (Stewart, 1993).
PM3_LUMO	Energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian (Stewart, 1993).
RPC+	Same as Q_RPC+
SMR	Molecular refractivity, calculated by an atomic contribution model (Wildman and Crippen, 1999)
SMR_VSA0	Sum of the approximate accessible van der Waals surface area v_i such that the contribution to Molar Refractivity for atom i (R_i) is in [0,0.11]
SMR_VSA5	Sum of v_i such that R_i is in (0.15,0.20]
SMR_VSA7	Sum of v_i such that $R_i > 0.56$
SlogP	Log of the octanol/water partition coefficient, calculated by an atomic contribution model (Wildman and Crippen, 1999)
SlogP_VSA1	Sum of v_i such that the contribution to logP(o/w) for atom i (L_i) is in (-0.4,-0.2]
SlogP_VSA2	Sum of v_i such that L_i is in (-0.2,0]
SlogP_VSA4	Sum of v_i such that L_i is in (0.1,0.15]
SlogP_VSA7	Sum of v_i such that L_i is in (0.25,0.30]
SlogP_VSA8	Sum of v_i such that L_i is in (0.30,0.40]
VDistEq	If m is the sum of the distance matrix entries then VdistEq is defined to be the sum of $\log_2 m - p_i \log_2 p_i / m$ where p_i is the number of distance matrix entries equal to i
VDistMa	If m is the sum of the distance matrix entries then VDistMa is defined to be the sum of $\log_2 m - D_{ij} \log_2 D_{ij} / m$ over all i and j.

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
a_ICM	Atom information content (mean). Let n_i be the number of occurrences of atomic number i in the molecule. Let $p_i = n_i/n$ where n is the sum of the n_i. The value of a_ICM is the negative of the sum over all i of $p_i \log p_i$.
a_IC	Atom information content (total). This is calculated to be a_ICM times n.
a_aro	Number of aromatic atoms
a_hyd	Number of hydrophobic atoms
a_nC	Number of carbon atoms
a_nH	Number of hydrogen atoms
a_nO	Number of oxygen atoms
apol	Sum of the atomic polarizabilities, with polarizabilities taken from (CRC, 1994)
b_1rotN	Number of rotatable single bonds (not including conjugated single bonds, such as petide and ester bonds)
b_1rotR	Fraction of rotatable single bonds
b_ar	Number of aromatic bonds
b_rotN	Number of rotatable bonds
b_rotR	Fraction of rotatable bonds
balabanJ	Balaban's connectivity topological index (Balaban, 1982)
bpol	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens) with polarizabilities taken from (CRC, 1994)
chi0_C	Carbon connectivity index (order 0). This is calculated as the sum of $1/\sqrt{d_i}$, with d_i the number of bonded non-hydrogen atoms, over all carbon atoms i with $d_i > 0$

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
chi0v_C	Carbon valence connectivity index (order 0). This is calculated as the sum of $1/\sqrt{v_i}$ over all carbon atoms i with $v_i > 0$, with $v_i = (p_i - h_i)/(Z_i - p_i - 1)$ where p_i is the number of s and p valence electrons and Z_i the atomic number of atom i.
chi1	Atomic connectivity index (order 1) from (Hall and Kier, 1991) and (Hall and Kier, 1977). This is calculated as the sum of $1/\sqrt{d_i d_j}$ over all bonds between heavy atoms i and j where $i < j$
chi1_C	Carbon connectivity index (order 1). This is calculated as the sum of $1/\sqrt{d_i d_j}$ over all bonds between carbon atoms i and j where $i < j$
chi1v	Atomic valence connectivity index (order 1) from (Hall and Kier, 1991) and (Hall and Kier, 1977). This is calculated as the sum of $1/\sqrt{v_i v_j}$ over all bonds between heavy atoms i and j where $i < j$
chi1v_C	Carbon valence connectivity index (order 1). This is calculated as the sum of $1/\sqrt{v_i v_j}$ over all bonds between carbon atoms i and j where $i < j$
dens	Mass density: molecular weight divided by van der Waals volume (calculated using a grid approximation with spacing 0.75 Å)
density	Molecular mass density: Weight divided by the van der Waals volume (calculated using a connection table approximation)
diameter	Largest value in the distance matrix (Petitjean, 1992).

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
glob	Globularity, or inverse condition number (smallest eigenvalue divided by the largest eigenvalue) of the covariance matrix of atomic coordinates. A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object.
logP(o/w)	Log of the octanol/water partition coefficient, calculated from a linear atom type model implemented in MOE
mr	Molecular refractivity, calculated from an 11 descriptor linear model implemented in MOE
petitjean	Value of (diameter-radius) / diameter, with diameter the largest value in the distance matrix and radius defined as follows: If r_i is the largest matrix entry in row i of the distance matrix D, then radius is defined as the smallest of the r_i (Petitjean, 1992)
petitjeanSC	Petitjean graph Shape Coeffecient as defined in (Petitjean, 1992): (diameter-radius) / radius
rgyr	Radius of gyration
std_dim1	Standard dimension 1: the square root of the largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis
std_dim2	Standard dimension 2: the square root of the second largest eigenvalue of the covariance matrix of the atomic coordinates
std_dim3	Standard dimension 3: the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates

Table A.1: Molecular descriptors and their meaning (*cont.*).

Descriptor	Meaning
vsa_acc	Approximation to the sum of VDW surface areas of pure hydrogen bond acceptors (not counting acidic atoms and atoms that are both hydrogen bond donors and acceptors such as -OH)
Zagreb	Zagreb index: the sum of d_i^2 over all heavy atoms i, with d_i the number of non-hydrogen atoms to which atom i is bonded

A.2 Descriptor ranks and p-values from KS-statistics

Table A.2: Descriptor ranks and p-values.

Rank	ab1D	p-value	ab2A	p-value
1	std_dim3	0.0015	PEOE_VSA_FPNEG	0.0035
2	PEOE_VSA_FPNEG	0.0083	a_ICM	0.0071
3	a_ICM	0.0124	PEOE_VSA_NEG	
4	dens		apol	
5	Q_VSA_FNEG	0.0141	SMR_VSA5	
6	Q_VSA_FPOS		chi1v_C	0.0099
7	FCASA-		chi1_C	
8	Q_VSA_POS	0.0161	PEOE_RPC+	
9	KierA3		PEOE_VSA_HYD	
10	Q_VSA_NEG	0.0206	a_hyd	
11	FASA-		SMR	
12	a_nH	0.0233	chi0v_C	0.0138
13	b_1rotR		chi1v	
14	b_rotR		PEOE_VSA_FHYD	
15	density	0.0263	PEOE_VSA_FPOL	
16	SlogP_VSA7	0.0297	E_str	
17	Kier3	0.0334	mr	
18	KierFlex		SlogP	
19	SlogP_VSA4		logP(o/w)	
20	glob		a_nC	0.0189
21	vsa_acc	0.0375	chi0_C	
22	a_aro	0.0420	chi1	
23	b_ar		PEOE_VSA-1	
24	b_rotN	0.0471	rgyr	
25	E		vsa_hyd	
26	bpol		weinerPath	0.0256
27	SMR_VSA3		b_count	
28	zagreb	0.0526	Q_VSA_HYD	
29	SMR_VSA0		SlogP_VSA1	
30	PEOE_VSA+0	0.0587	vdw_vol	

Table A.2: Descriptor ranks and p-values (*cont.*).

Rank	ab2B	p-value	ab3A	p-value
1	PEOE_VSA_FPPOS	0.0055	balabanJ	0.0001
2	AM1_HOMO	0.0109	MNDO_HF	0.0003
3	AM1_IP		PEOE_VSA_FPPOS	0.0004
4	PEOE_RPC+	0.0124	FASA+	0.0004
5	FASA+	0.0161	AM1_HOMO	0.0007
6	MNDO_HF	0.0233	PM3_HOMO	
7	chi1_C	0.0263	AM1_IP	
8	PEOE_VSA_FHYD		PM3_IP	
9	PEOE_VSA_FPOL		MNDO_HOMO	0.0010
10	Q_VSA_FPPOS		MNDO_IP	
11	MNDO_HOMO	0.0334	PEOE_VSA+5	0.0014
12	MNDO_IP		PEOE_VSA_POL	
13	PM3_HOMO		Q_RPC+	0.0016
14	PM3_IP		RPC+	
15	PEOE_VSA_PPOS	0.0375	b_1rotR	0.0029
16	glob		PEOE_VSA_PNEG	
17	PM3_LUMO	0.0471	PEOE_VSA_PPOS	
18	FCASA+		PEOE_VSA_POS	0.0039
19	PEOE_VSA_FNEG	0.0526	FCASA+	
20	PEOE_VSA_FPOS		SlogP_VSA2	0.0045
21	Q_RPC+	0.0653	SlogP_VSA1	0.0060
22	RPC+		SMR_VSA0	
23	AM1_LUMO	0.0727	PEOE_VSA+0	0.0091
24	PEOE_PC+		a_nO	0.0118
25	PEOE_PC-		PEOE_VSA_FHYD	
26	PEOE_VSA_POL		PEOE_VSA_FPOL	
27	E_str		vsa_other	
28	SMR_VSA6		b_1rotN	0.0135
29	std_dim2	0.0894	b_rotR	
30	Q_RPC-	0.0989	PEOE_VSA-5	

Table A.2: Descriptor ranks and p-values (*cont.*).

Rank	ab3B	p-value	ab5B	p-value
1	SlogP_VSA8	0.0323	KierA3	0.0010
2	std_dim2	0.0429	KierA1	0.0064
3	Density	0.0460	KierA2	
4	PEOE_VSA_POS	0.0604	KierFlex	
5	Dens	0.0645	Kier2	0.0097
6	balabanJ		Kier3	
7	PEOE_VSA+0		b_1rotR	0.0119
8	PEOE_RPC-	0.0836	balabanJ	0.0176
9	PM3_HOMO		rgyr	
10	PM3_IP		std_dim1	
11	Q_VSA_PPOS	0.0891	b_1rotN	0.0213
12	Kier2		chi1v	
13	Kier3		Q_VSA_POS	0.0257
14	AM1_HOMO	0.1139	b_rotN	0.0309
15	AM1_IP		MNDO_HF	
16	zagreb		VDistEq	0.0370
17	KierA2	0.1210	PEOE_VSA+0	
18	a_ICM	0.1617	std_dim2	0.0440
19	KierA3		a_IC	0.0618
20	KierFlex		b_rotR	
21	SMR_VSA5		bpol	
22	VDistMa	0.2016	glob	0.0728
23	b_1rotR		SMR_VSA7	0.0854
24	b_rotR		chi0v_C	0.0998
25	Q_VSA_FPPOS		FASA+	0.1161
26	MNDO_HOMO		PEOE_PC+	
27	MNDO_IP		vsa_hyd	
28	VDistEq	0.2127	PEOE_VSA_FPPOS	0.1346
29	b_1rotN		PEOE_VSA_HYD	
30	FCASA-	0.2242	Kier1	

Table A.2: Descriptor ranks and p-values (*cont.*).

Rank	ab6A	p-value	
1	Kier3	0.0003	
2	Kier2	0.0017	
3	balabanJ	0.0026	
4	VDistEq	0.0037	
5	SlogP_VSA8	0.0056	
6	KierA3	0.0066	
7	KierA2	0.0079	
8	std_dim2	0.0098	
9	b_rotR	0.0158	
10	b_1rotR	0.0226	
11	KierFlex	0.0304	
12	b_rotN	0.0351	
13	MNDO_HF	0.0368	
14	FASA+		
15	FCASA+		
16	petitjean	0.0424	
17	petitjeanSC		
18	MNDO_HOMO		
19	MNDO_IP		
20	diameter	0.0557	
21	E	0.0665	
22	PEOE_RPC+		
23	AM1_HOMO	0.0757	
24	AM1_IP		
25	Q_VSA_FPPOS		
26	PM3_HOMO		
27	PM3_IP		
28	SlogP_VSA4		
29	rgyr	0.0824	
30	zagreb	0.0859	

References

Aires-de-Sousa, J. and Aires-de-Sousa, L. (2003). Representation of DNA sequences with virtual potentials and their processing by (SEQREP) Kohonen self-organizing maps. *Bioinformatics*, 19(1):30–36.

Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J., and Polanski, J. (1996). The comparison of geometric and electronic properties of molecular surfaces by neural networks: application to the analysis of corticosteroid-binding globulin activity of steroids. *J Comput Aided Mol Des*, 10(6):521–534.

Araneda, R. C., Kini, A. D., and Firestein, S. (2000). The molecular receptive range of an odorant receptor. *Nat Neurosci*, 3(12):1248–1255.

Araneda, R. C., Peterlin, Z., Zhang, X., Chesler, A., and Firestein, S. (2004). A pharmacological profile of the aldehyde receptor repertoire in rat olfactory epithelium. *J Physiol*, 555(Pt 3):743–756.

Armstrong, J. D., Kaiser, K., Mller, A., Fischbach, K. F., Merchant, N., and Strausfeld, N. J. (1995). Flybrain, an on-line atlas and database of the drosophila nervous system. *Neuron*, 15(1):17–20.

Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A., and Damiani, G. (1991). Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *Comput Appl Biosci*, 7(3):353–357.

Balaban, A. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89:399–404.

Balakin, K. V., Ivanenkov, Y. A., Savchuk, N. P., Ivashchenko, A. A., and Ekins, S. (2005). Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discov Technol*, 2(2):99–113.

Barber, M. J., Clark, J. W., and Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Comput*, 15(8):1843–1864.

Baumgarte, F. (2002). Improved audio coding using a psychoacoustic model based on a cochlear filter bank. *Speech and Audio Processing, IEEE Transactions on*, 10(7):495–503.

Becker, O. M., Shacham, S., Marantz, Y., and Noiman, S. (2003). Modeling the 3D structure of GPCRs: advances and application to drug discovery. *Current Opinion in Drug Discovery and Development*, 6:353–361.

Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004). Similarity searching of chemical databases using atom environment descriptors (molprint 2d): evaluation of performance. *J Chem Inf Comput Sci*, 44(5):1708–1718.

Bensmail, H., Golek, J., Moody, M. M., Semmes, J. O., and Haoudi, A. (2005). A novel approach for clustering proteomics data using bayesian fast fourier transform. *Bioinformatics*, 21(10):2210–2224.

Box, G. (1979). *Robustness in Statistics*, chapter Robustness in the strategy of scientific model building. Academic Press, New York.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–187.

Chess, A., Simon, I., Cedar, H., and Axel, R. (1994). Allelic inactivation regulates olfactory receptor gene expression. *Cell*, 78(5):823–834.

Cleland, T. A. and Linster, C. (2005). Computation in the olfactory system. *Chem Senses*, 30(9):801–813.

Clyne, P., Grant, A., O'Connell, R., and Carlson, J. R. (1997). Odorant response of individual sensilla on the *drosophila* antenna. *Invert Neurosci*, 3(2-3):127–135.

Clyne, P. J., Warr, C. G., Freeman, M. R., Lessing, D., Kim, J., and Carlson, J. R. (1999). A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. *Neuron*, 22(2):327–338.

Couto, A., Alenius, M., and Dickson, B. J. (2005). Molecular, anatomical, and functional organization of the *Drosophila* olfactory system. *Curr Biol*, 15(17):1535–1547.

CRC (1994). *CRC Handbook of Chemistry and Physics*. CRC Press.

de Bruyne, M., Foster, K., and Carlson, J. R. (2001). Odor coding in the *Drosophila* antenna. *Neuron*, 30(2):537–552.

de Mello Castanho Amboni, R. D., da Silva Junkes, B., Yunes, R. A., and Heinzen, V. E. (2000). Quantitative structure-odor relationships of aliphatic esters using topological indices. *J Agric Food Chem*, 48(8):3517–3521.

Del Carpio-Muñoz, C. A., Ichiishi, E., Yoshimori, A., and Yoshikawa, T. (2002). MIAX: a new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases. *Proteins*, 48(4):696–732.

Dobritsa, A. A., van der Goes van Naters, W., Warr, C. G., Steinbrecht, R. A., and Carlson, J. R. (2003). Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. *Neuron*, 37(5):827–841.

Fankhauser, N. and Mäser, P. (2005). Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics*, 21(9):1846–1852.

Fano, U. (1947). Ionization yield of radiations. II. the fluctuations of the number of ions. *Physical Reviews*, 72:26–29.

Ferrán, E. A. and Ferrara, P. (1991). Topological maps of protein sequences. *Biol Cybern*, 65(6):451–458.

Firestein, S. (2001). How the olfactory system makes sense of scents. *Nature*, 413(6852):211–218.

Fishilevich, E. and Vosshall, L. B. (2005). Genetic and functional subdivision of the drosophila antennal lobe. *Curr Biol*, 15(17):1548–1553.

Floriano, W. B., Vaidehi, N., and Goddard, W. A. (2004). Making sense of olfaction through predictions of the 3-D structure and function of olfactory receptors. *Chem Senses*, 29(4):269–290.

Friedrich, R. W. and Korsching, S. I. (1997). Combinatorial and chemotopic odorant coding in the zebrafish olfactory bulb visualized by optical imaging. *Neuron*, 18(5):737–752.

Gao, Q. and Chess, A. (1999). Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. *Genomics*, 60(1):31–39.

Gasteiger, J., Li, X., and Uschold, A. (1994). The beauty of molecular surfaces as revealed by self-organizing neural networks. *J Mol Graph*, 12(2):90–97.

Gasteiger, J. and Marsili, M. (1980). Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, 36:3219.

Givehchi, A., Dietrich, A., Wrede, P., and Schneider, G. (2003). ChemSpaceShuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization. *QSAR & Combinatorial Science*, 22(5):549–559.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D. (2001). The complete human olfactory subgenome. *Genome Res*, 11(5):685–702.

Goldman, A. L., der Goes van Naters, W. V., Lessing, D., Warr, C. G., and Carlson, J. R. (2005). Coexpression of two functional odor receptors in one neuron. *Neuron*, 45(5):661–666.

Grabowski, K. and Schneider, G. (2007). Properties and architecture of drugs and natural products revisited. *Current Chemcial Biology*, 1:115–127.

Haberly, L. B. (2001). Parallel-distributed processing in olfactory cortex: new insights from morphological and physiological analysis of neuronal circuitry. *Chem Senses*, 26(5):551–576.

Halgren, T. A. (1999). MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry*, 20:720–729.

Hall, L. H. and Kier, L. B. (1977). The nature of structure-activity relationships and their relation to molecular connectivity. *European Journal of Medicinal Chemistry*, 12:307–312.

Hall, L. H. and Kier, L. B. (1991). *Reviews in Computational Chemistry*, volume 2, chapter The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, pages 367–422. Purdue University at Indianapolis, Indianapolis.

Hall, S. E., Floriano, W. B., Vaidehi, N., and Goddard, W. A. (2004). Predicted 3-D structures for mouse I7 and rat I7 olfactory receptors and comparison of predicted odor recognition profiles with experiment. *Chem Senses*, 29(7):595–616.

Hallem, E. A. and Carlson, J. R. (2006). Coding of odors by a receptor repertoire. *Cell*, 125(1):143–160.

Hallem, E. A., Ho, M. G., and Carlson, J. R. (2004). The molecular basis of odor coding in the *Drosophila* antenna. *Cell*, 117(7):965–979.

Hanke, J. and Reich, J. G. (1996). Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. *Comput Appl Biosci*, 12(6):447–454.

Hasegawa, K., Matsuoka, S., Arakawa, M., and Funatsu, K. (2002). New molecular surface-based 3D-QSAR method using kohonen neural network and 3-way PLS. *Comput Chem*, 26(6):583–589.

Heisenberg, M. (1998). What do the mushroom bodies do for the insect brain? an introduction. *Learn Mem*, 5(1-2):1–10.

Hertz, J., Palmer, R. G., and Krogh, A. S. (1991). *Introduction to the theory of neural computation*. Westview Press, Boulder, Colorado.

Hildebrand, J. G. and Shepherd, G. M. (1997). Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu Rev Neurosci*, 20:595–631.

Huerta, R., Nowotny, T., Garca-Sanchez, M., Abarbanel, H. D. I., and Rabinovich, M. I. (2004). Learning classification in the olfactory system of insects. *Neural Comput*, 16(8):1601–1640.

Johnson, B. A., Farahbod, H., Saber, S., and Leon, M. (2005). Effects of functional group position on spatial representations of aliphatic odorants in the rat olfactory bulb. *J Comp Neurol*, 483(2):192–204.

Kairys, V., Fernandes, M. X., and Gilson, M. K. (2006). Screening drug-like compounds by docking to homology models: a systematic study. *Journal of Chemical Information and Modeling*, 46:365–379.

Kaissling, K. E. (2001). Olfactory perireceptor and receptor events in moths: a kinetic model. *Chem Senses*, 26(2):125–150.

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*, 25(2):197–206.

Keller, A. and Vosshall, L. B. (2003). Decoding olfaction in *Drosophila*. *Curr Opin Neurobiol*, 13(1):103–110.

Kim, M. S., Repp, A., and Smith, D. P. (1998). LUSH odorant-binding protein mediates chemosensory responses to alcohols in *Drosophila* melanogaster. *Genetics*, 150(2):711–721.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, V43(1):59–69.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer, Berlin Heidelberg New York, 3rd edition.

Korsching, S. I. (2001). Odor maps in the brain: spatial aspects of odor representation in sensory surface and olfactory bulb. *Cell Mol Life Sci*, 58(4):520–530.

Korsching, S. I. (2002). Olfactory maps and odor images. *Curr Opin Neurobiol*, 12(4):387–392.

Kratochwil, N. A., Malherbe, P., Lindemann, L., Ebeling, M., Hoener, M. C., Mhlemann, A., Porter, R. H. P., Stahl, M., and Gerber, P. R. (2005). An automated system for the analysis of g protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. *J Chem Inf Model*, 45(5):1324–1336.

Lavine, B. K., Davidson, C. E., Breneman, C., and Katt, W. (2003). Electronic van der waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases. *J Chem Inf Comput Sci*, 43(6):1890–1905.

Lin, D. Y., Zhang, S., Block, E., and Katz, L. C. (2005). Encoding social signals in the mouse main olfactory bulb. *Nature*, 434:470–477.

Linster, C., Sachse, S., and Galizia, C. G. (2005). Computational modeling suggests that response properties rather than spatial position determine connectivity between olfactory glomeruli. *J Neurophysiol*, 93(6):3410–3417.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–413.

Mallat, S. G. (1989). Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:2091–2110.

Manallack, D. T., Ellis, D. D., and Livingstone, D. J. (1994). Analysis of linear and nonlinear QSAR data using neural networks. *Journal of Medicinal Chemistry*, 37:3758–3767.

Manoukian, E. B. (1986). *Mathematical Nonparametric Statistics*. Gordon and Breach Science Publishers, New York.

Marin, E. C., Jefferis, G. S. X. E., Komiyama, T., Zhu, H., and Luo, L. (2002). Representation of the glomerular olfactory map in the drosophila brain. *Cell*, 109(2):243–255.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Biophysica Acta*, 405:442–451.

Meissner, M., Schmuker, M., and Schneider, G. (2006). Optimized particle swarm optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics*, 7:125.

Meister, M. and Bonhoeffer, T. (2001). Tuning and topography in an odor map on the rat olfactory bulb. *J Neurosci*, 21(4):1351–1360.

Mombaerts, P. (1999). Seven-transmembrane proteins as odorant and chemosensory receptors. *Science*, 286(5440):707–711.

Mombaerts, P. (2004). Odorant receptor gene choice in olfactory sensory neurons: the one receptor-one neuron hypothesis revisited. *Curr Opin Neurobiol*, 14(1):31–36.

Mori, K., Takahashi, Y. K., Igarashi, K. M., and Yamaguchi, M. (2006). Maps of odorant molecular features in the mammalian olfactory bulb. *Physiol Rev*, 86(2):409–433.

Nicholls, J. G., Wallace, B. G., Martin, A. R., and Fuchs, P. A. (2001). *From Neuron to Brain*. Sinauer Associates, 4th edition.

Nowotny, T., Huerta, R., Abarbanel, H. D. I., and Rabinovich, M. I. (2005). Self-organization in the olfactory system: one shot odor recognition in insects. *Biol Cybern*, 93(6):436–446.

Olender, T., Fuchs, T., Linhart, C., Shamir, R., Adams, M., Kalush, F., Khen, M., and Lancet, D. (2004). The canine olfactory subgenome. *Genomics*, 83(3):361–372.

Petitjean, M. (1992). Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *Journal of Chemical Information and Computer Science*, 32:331–337.

Polanski, J. and Walczak, B. (2000). The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput Chem*, 24(5):615–625.

Ritter, H. (1998). Self-organizing maps in non-euclidean spaces. In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 97–108. Elsevier, Amsterdam.

Roche, O., Trube, G., Zuegge, J., Pflimlin, P., Alanine, A., and Schneider, G. (2002). A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *Chembiochem*, 3(5):455–459.

Roesch, M. R., Stalnaker, T. A., and Schoenbaum, G. (2007). Associative encoding in anterior piriform cortex versus orbitofrontal cortex during odor discrimination and reversal learning. *Cereb Cortex*, 17(3):643–652.

Rolls, E. and Treves, A. (1997). *Neural Networks and Brain Function*. Oxford University Press, Oxford, United Kingdom.

Sachse, S., Rappert, A., and Galizia, C. G. (1999). The spatial representation of chemical structures in the antennal lobe of honeybees: steps towards the olfactory code. *Eur J Neurosci*, 11(11):3970–3982.

Sangole, A. and Knopf, G. K. (2003a). Shape registration using deformable self-organizing feature maps. *International Journal of Smart Engineering System Design*, 5:439–454.

Sangole, A. and Knopf, G. K. (2003b). Visualization of randomly ordered numeric data sets using spherical self-organizing feature maps. *Computers and Graphics*, 27:963–976.

Schmuker, M., Givehchi, A., and Schneider, G. (2004). Impact of different software implementations on the performance of the maxmin method for diverse subset selection. *Mol Divers*, 8(4):421–425.

Schmuker, M., Schwarte, F., Brück, A., Proschak, E., Tanrikulu, Y., Givehchi, A., Scheiffele, K., and Schneider, G. (2007). Sommer: self-organising maps for education and research. *Journal of Molecular Modeling*, 13(1):225–228.

Schneider, G. and Fechner, U. (2004). Advances in the prediction of protein targeting signals. *Proteomics*, 4(6):1571–1580.

Schneider, G. and Nettekoven, M. (2003). Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. *J Comb Chem*, 5(3):233–237.

Schneider, G., Sjöling, S., Wallin, E., Wrede, P., Glaser, E., and von Heijne, G. (1998). Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins*, 30(1):49–60.

Schneider, G. and Wrede, P. (1998). Artificial neural networks for computer-based molecular design. *Progress in Biophysical Molecular Biology*, 70:175–222.

Schneider, P. and Schneider, G. (2003). Collection of bioactive reference compounds for focused library design. *QSAR & Combinatorial Science*, 22(7):713–718.

Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D., and Wrede, P. (1996).

Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng*, 9(10):833–842.

Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci*, 43(2):391–405.

Sell, C. S. (2006). On the unpredictability of odor. *Angew Chem Int Ed Engl*, 45(38):6254–6261.

Shirokova, E., Schmiedeberg, K., Bedner, P., Niessen, H., Willecke, K., Raguse, J.-D., Meyerhof, W., and Krautwurst, D. (2005). Identification of specific ligands for orphan olfactory receptors. g protein-dependent agonism and antagonism of odorants. *J Biol Chem*, 280(12):11807–11815.

Sigma-Aldrich (2004). *Flavors and Fragrances Catalog*. Sigma-Aldrich, Milwaukee, WI.

Stahl, M., Taroni, C., and Schneider, G. (2000). Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng*, 13(2):83–88.

Stensmyr, M. C., Giordano, E., Balloi, A., Angioy, A.-M., and Hansson, B. S. (2003). Novel natural ligands for drosophila olfactory receptor neurones. *J Exp Biol*, 206(Pt 4):715–724.

Stewart, J. J. P. (1993). *MOPAC Manual*, 7th edition.

Teckentrup, A., Briem, H., and Gasteiger, J. (2004). Mining high-throughput screening data of combinatorial libraries: development of a filter to distinguish hits from nonhits. *J Chem Inf Comput Sci*, 44(2):626–634.

Todeschini, R. and Consonni, V. (2000). *Handbook of molecular descriptors*. Wiley-VCH, Weinheim.

Tsantili-Kakoulidou, A. and Kier, L. B. (1992). A quantitative structure-activity relationship (QSAR) study of alkylpyrazine odor modalities. *Pharm Res*, 9(10):1321–1323.

Uchida, N., Takahashi, Y. K., Tanifuji, M., and Mori, K. (2000). Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features. *Nature Neuroscience*, 3:1035–1043.

Vaidehi, N., Floriano, W. B., Trabanino, R., Hall, S. E., Freddolino, P., Choi, E. J., Zamanakos, G., and Goddard, W. A. (2002). Prediction of structure and function of G protein-coupled receptors. *Proc Natl Acad Sci U S A*, 99(20):12622–12627.

Vetter, R. S., Sage, A. E., Justus, K. A., Cardé, R. T., and Galizia, C. G. (2006). Temporal integrity of an airborne odor stimulus is greatly affected by physical aspects of the odor delivery system. *Chem Senses*, 31(4):359–369.

Vosshall, L. B. (2000). Olfaction in drosophila. *Curr Opin Neurobiol*, 10(4):498–503.

Vosshall, L. B., Wong, A. M., and Axel, R. (2000). An olfactory sensory map in the fly brain. *Cell*, 102(2):147–159.

Wailzer, B., Klocker, J., Buchbauer, G., Ecker, G., and Wolschann, P. (2001). Prediction of the aroma quality and the threshold values of some pyrazines using artificial neural networks. *J Med Chem*, 44(17):2805–2813.

Wildman, S. and Crippen, G. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Modeling*, 39(5):868–873.

Winkler, D. A. and Burden, F. R. (2002). Application of neural networks to large dataset QSAR, virtual screening, and library design. *Methods in Molecular Biology*, 201:325–367.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Wong, A. M., Wang, J. W., and Axel, R. (2002). Spatial representation of the glomerular map in the drosophila protocerebrum. *Cell*, 109(2):229–241.

Wu, Y. and Takatsuka, M. (2004). The geodesic self-organizing map and its error analysis. In Estivill-Castro, V., editor, *Proceedings of the Twenty-Eighth Australasian Conference on Computer Science*, volume 38, pages 343–351, Darlinghurst. Australian Computer Society.

Xiao, Y.-D., Clauset, A., Harris, R., Bayram, E., Santago, P., and Schmitt, J. D. (2005). Supervised self-organizing maps in drug discovery. 1. Robust behavior with overdetermined data sets. *J Chem Inf Model*, 45(6):1749–1758.

Xu, P., Atkinson, R., Jones, D. N. M., and Smith, D. P. (2005). *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. *Neuron*, 45(2):193–200.

Zarzo, M. and Stanton, D. T. (2006). Identification of latent variables in a semantic odor profile database using principal component analysis. *Chem Senses*, 31(8):713–724.

Zhang, X. and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. *Nat Neurosci*, 5(2):124–133.

Zou, Z., Horowitz, L. F., Montmayeur, J. P., Snapper, S., and Buck, L. B. (2001). Genetic tracing reveals a stereotyped sensory map in the olfactory cortex. *Nature*, 414(6860):173–179.

Zozulya, S., Echeverri, F., and Nguyen, T. (2001). The human olfactory receptor repertoire. *Genome Biol*, 2(6):RESEARCH0018.

Zupan, J. and Gasteiger, J. (1999). *Neural Networks in Chemistry and Drug Design*. Wiley-VCH, Weinheim.

Zusammenfassung

Unser Geruchssinn vermittelt uns die Wahrnehmung der chemischen Welt. Im Laufe der Evolution haben sich in unserem olfaktorischen System Mechanismen entwickelt, die wahrscheinlich optimal auf die Erfüllung dieser Aufgabe angepasst sind. Die Analyse dieser Verarbeitungsstrategien verspricht Einblicke in effiziente Algorithmen für die Kodierung und Verarbeitung chemischer Information, deren Entwicklung und Anwendung dem Kern der Chemieinformatik entspricht.

In dieser Arbeit nähern wir uns der Entschlüsselung dieser Mechanismen durch die rechnerische Modellierung von funktionellen Einheiten des olfaktorischen Systems. Hierbei verfolgten wir einen interdisziplinären Ansatz, der die Gebiete der Chemie, der Neurobiologie und des maschinellen Lernens mit einbezieht.

Funktionelle Charakterisierung von olfaktorischen Rezeptorneuronen

Olfaktorische Rezeptoren arbeiten an der Schnittstelle zwischen dem chemischen Raum und der Geruchswahrnehmung im Gehirn. Sie kodieren Eigenschaften von Geruchsmolekülen in Information, die von neuronalen Schaltkreisen im Gehirn weiterverarbeitet werden kann. Im ersten Teil dieser Arbeit widmeten wir uns daher der Charakterisierung der Kodierungseigenschaften der olfaktorischen Rezeptoren.

Basierend auf publizierten Antworten von sieben Rezeptorneuronklassen der Fruchtfliege *Drosophila melanogaster* auf eine Auswahl von 47 Duftstoffen konstruierten wir ein Modell, das die Antwort eines Rezeptorneurons auf einen beliebigen olfaktorischen Stimulus ausgehend von dessen chemischer Struktur vorhersagen konnte. Hierzu repräsentierten wir die Duftstoffe durch vektorielle chemische Deskriptoren physikochemischer Eigenschaften. Jeder Duftstoff wurde somit durch eine Reihe von 203 Zahlen dargestellt. Mit diesen trainier-

ten wir Künstliche Neuronale Netze darauf, aktivierende Duftstoffe von nicht-aktivierenden zu unterscheiden. Dies geschah separat für jede der sieben Rezeptorneuronklassen.

Um die Vorhersagekraft der erhaltenen Modelle zu testen durchsuchten wir eine Duftstoff-Datenbank nach aktivierenden Molekülen. Anschließend wurden in Kooperation mit Marien de Bruyne und Melanie Hähnel von der Freien Universität Berlin *in vivo* Antworten von Rezeptorneuronen auf eine Auswahl der gefundenen Duftstoffe getestet und mit den Vorhersagen verglichen.

Für die Mehrzahl der untersuchten Rezeptorneuronklassen fanden wir eine Korrelation zwischen den Vorhersagen und den gemessenen Antworten. Der Matthews Korrelationskoeffizient lag bei 0.85 für die ab3A-Neuronklasse, 0.69 für ab1D und ab2A, 0.68 für ab5B und 0.66 für die ab6A-Neuronklasse, was wir als erfolgreiche Vorhersage werteten. Für die zwei verbleibenden Neuronklassen konnten wir jedoch keine verlässliche Vorhersage erstellen; die Korrelationskoeffizienten lagen bei 0.34 (ab3B) und 0.17 (ab2A).

Weiterhin konnten wir zeigen, dass für jede Neuronklasse eine andere Kombination chemischer Eigenschaften am besten zur Aktivierungsvorhersage geeignet ist. So können beispielsweise Aktivatoren von ab1D-Neuronen wie z.B. Methyl Salicylat, Phenylacetaldehyd und Acetophenon am besten durch ihre flache, scheibenähnliche Form und eine exponierte negative Partialladung beschrieben werden, wohingegen Neuronen der ab5B-Klasse größere Liganden mit flexiblen Seitenketten wie z.B. Pentyl Acetat, 2-Heptanon oder 3-Octanol zu bevorzugen scheinen.

Olfaktorische Rezeptorneuronen scheinen also unterschiedliche Kombinationen chemischer Eigenschaften zu kodieren, und ähneln darin wiederum den Deskriptoren, die wir ursprünglich zur Erstellung der Modelle benutzt haben. Wir konnten zeigen, dass die Rezeptorantworten selbst auch wieder als Deskriptoren zur Erstellung prädiktiver Modelle benutzt werden können, wenn auch mit geringerer Vorhersagegenauigkeit.

Modellierung des Antennallobus von Insekten mit selbstorganisierenden Karten

Eine der faszinierendsten Eigenschaften des olfaktorischen Systems liegt in der stereotypen Organisation der zweiten neuronalen Stufe, dem Antennallobus in Insekten bzw. dem *Bulbus olfactorius* in Vertebraten. In dieser Struktur konvergieren Axone von Rezeptorneuronen aus den nasalen Epithelia in so genannten Glomeruli, wo sie Synapsen mit sekundären Projektionsneuronen (in Insekten) bzw. Mithralzellen (in Vertebraten) bilden. Mehrere Indizien weisen darauf hin, dass diese Struktur eine chemotopische Organisation aufweist, d.h. dass ähnliche Duftstoffe nah beieinander liegende Areale dieser Struktur aktivieren. Dieses Phänomen ermöglicht es zu untersuchen wie chemische Ähnlichkeit in der Natur dargestellt wird.

Unser Ziel war es, die chemotopische Organisation des Antennallobus mit selbstorganisierenden Merkmalskarten (SOMs) zu erforschen. Mithilfe von SOMs können topologische Abbildungen der Eingabedaten erstellt werden, was sie in diesem Szenario besonders nützlich erscheinen läßt. Sie erlauben eine algorithmisch gut beschriebene, reproduzierbare Projektion der dreidimensionalen Anordnung der Glomeruli in die zweidimensionale Ebene.

Zu diesem Zweck entwickelten wir SOMMER, ein Programm zur Erstellung und Visualisierung zwei- und dreidimensionaler SOMs. SOMMER bietet rechteckige, toroidale, kubische und sphärische SOM-Topologien, die wir benutzten um zwei- und dreidimensionale Modelle des Antennallobus von *Drosophila* zu erstellen.

Wir benutzten die zweidimensionalen Modelle um darauf die Aktivierung von Glomeruli in Antwort auf Duftstimuli darzustellen. In diesen Aktivierungskarten wurde jedem Glomerulus die Feuerrate der Rezeptorneuronklasse die seinen hauptsächlichen Eingang stellt eingefärbt. Wir erstellten solche Karten für einen publizierten Datensatz von Neuronantworten auf 110 Duftstoffe.

Anhand dieser Aktivierungskarten untersuchten wir welche chemischen

Eigenschaften bevorzugt in den jeweiligen Glomeruli repräsentiert werden. Wir berechneten 184 physikochemische Deskriptoren für jeden der Duftstoffe. Um die Ausprägung eines bestimmten Deskriptors in einem Glomerulus zu quantifizieren berechneten wir anschließend den Median dieses Deskriptors, wobei wir nur Duftstoffe mit einbezogen die den jeweiligen Glomerulus zu aktivieren vermochten. Da jeder Deskriptor eine molekulare Eigenschaft wiedergibt, erhielten wir eine Karte die die Verteilung der chemischen Eigenschaften auf dem Antennallobus reflektiert.

Die Analyse dieser Karte offenbarte eine klare Tendenz für chemotopische Repräsentation des Molekulargewichts. Auch der Durchmesser eines Moleküls (der mit der Kettenlänge zusammenhängt) zeigte eine chemotopische Ordnung, wenn auch in geringerer Ausprägung. Im Gegensatz dazu konnten wir keine solche Ordnung für molekulare Flexibilität feststellen.

Eine neuartige Methode zur Verarbeitung und Klassifikation chemischer Daten, inspiriert durch den Geruchssinn der Insekten

Unser Ziel im letzten Teil dieser Arbeit war der Entwurf einer Methode zur Verarbeitung chemischer Deskriptoren und Klassifikation von Molekülen, die auf den im olfaktorischen System beobachteten Verarbeitungsstrategien basiert. Aufbauend auf den Ergebnissen unserer vorangegangenen Studien erstellten wir hierzu ein vereinfachtes rechnerisches Modell des gesamten olfaktorischen Systems.

Hierbei lag unser Augenmerk vor allem auf der Erstellung eines Modells, dass uns die Analyse bestimmter Aspekte rechnerischer Strategien ermöglichte, und weniger auf einer biologisch korrekten Simulation.

Die erste Stufe olfaktorischer Verarbeitung modellierten wir mit “virtuellen Rezeptoren”, definiert als diskrete Punkte im chemischen Duftraum. Dieser Duftraum wird aufgespannt durch 184 physikochemische Deskriptoren die

die chemischen Eigenschaften der Duftstoffe wiedergeben. Die Stärke der Antwort eines virtuellen Rezeptors auf einen Duftstimulus wird dabei durch deren Abstand im 184-dimensionalen Duftraum bestimmt: Je näher Rezeptor und Duftstoff beieinander liegen, desto größer die Antwort.

Die Platzierung der virtuellen Rezeptoren ermittelten wir durch Trainieren einer SOM im Duftraum, wobei die Koordinaten der Rezeptoren durch die Prototyp-Vektoren der trainierten SOM gegeben wurden. Da SOMs die lokale Topologie des Eingaberaumes bewahren, erhielten wir dadurch eine chemotopische Repräsentation des Duftraumes, ähnlich jener die man auf dem Antennallobus oder dem *Bulbus olfactorius* findet.

In der zweiten Stufe der olfaktorischen Verarbeitung werden im Antennallobus die Rezeptorsignale dekorreliert. Wir implementierten diesen Schritt durch korrelationsabhängige laterale Inhibition, wobei die Antwort eines Rezeptors um das nach Korrelation gewichtete Mittel aller anderen Rezeptoren vermindert wurde. Hierbei entsteht eine *Winner-Takes-Most*-Situation, wobei der Wettbewerb zwischen zwei Rezeptoren am stärksten ist wenn ihre Antwortspektren höchste Korrelation zeigen. Zusätzlich führten wir den skalaren Faktor q zur Gewichtung der gesamten Inhibition ein.

In der letzte Stufe unseres Modells ordneten wir den Eingabedaten eine Wahrnehmungsqualität (einen *Duft*) zu. Das erreichten wir durch Training eines Naiven Bayes-Klassifikators auf den annotierten Düften der Duftstoffe, wobei wir die verarbeiteten Signale als Eingangsdaten verwendeten.

Wir testeten die Leistungsfähigkeit unseres Modells mittels retrospektiven Screenings einer Duftstoffdatenbank. Die Vorhersagegenauigkeit wurde hierbei durch die Fläche unter der Receiver-Operating-Charateristic-Kurve (Area Under Curve, AUC) quantifiziert. Die Ergebnisse zeigten, dass die Repräsentation chemischer Information in unserem Modell für diese Aufgabe geeignet ist. Im Median über alle Düfte erreichten wir AUC-Werte von bis zu $0,72$ auf unverarbeiteten Aktivierungsmustern, abhängig von der Anzahl virtueller Rezeptoren.

Die Filterung der Aktivierungsmuster mit korrelationsabhängiger Inhibition verbesserte deren Klassifizierbarkeit: So stiegen die medianen AUC-Werte für die höchste Anzahl virtueller Rezeptoren auf $0,75$ für $q = 1$ und $0,79$ für $q = 2$. Denselben Trend konnten wir auch mit weniger Rezeptoren beobachten.

Wir konnten zeigen, dass diese Verarbeitungsmethode eine "moderate Dekorrelation" darstellt, d.h. dass eine restliche Korrelation zwischen den Ausgabedimensionen verbleibt. Dies steht im Gegensatz zu Datenanalysetechniken wie der Hauptkomponentenanalyse, bei der die Ausgabedimensionen unkorreliert sind.

Bei einem Vergleich der Klassifikationsleistung auf unkorrelierten Mustern (erhalten durch eine Hauptkomponentenanalyse der vektoriellen Deskriptoren) und der moderat dekorrelierten Muster wurden für beide ähnliche AUC-Werte erreicht, jedoch bei verschiedenen Eingabedimensionen. Unkorrelierte Muster zeigten die beste Klassifizierbarkeit bei niedriger Eingabedimensionalität, wobei die durch korrelationsbasierte Inhibition gefilterten Muster bei höherer Dimensionalität beste Ergebnisse lieferten. Allerdings verschlechterte sich die Klassifizierbarkeit unkorrelierter Muster rapide mit ansteigender Dimensionalität, wohingegen die moderat dekorrelierten Muster mit steigender Dimensionalität eine "Sättigung" erreichten, d.h. auf konstant hohem Niveau blieben.

Die Anwendung dieser Methode ist nicht auf olfaktorische Daten beschränkt. Dies konnten wir durch ein erfolgreiches virtuelles Screening einer pharmazeutischen Datenbank zeigen.

Danksagung

An dieser Stelle möchte ich mich herzlich bei Gisbert Schneider für die interessante und lehrreiche Zeit bedanken, die ich in seiner Arbeitsgruppe verbringen durfte. Seine fachliche Expertise und sein Optimismus haben mich immer wieder zu Höchstleistungen angespornt.

Der Arbeitsgruppe danke ich für die nette Arbeitsatmosphäre. Im besonderen gilt mein Dank Evgeny Byvatov, Steffen Renner, Uli Fechner, Andreas Schüller, Tobias Noeske, Alexander Böcker, Michael Meissner, Tina Grabowski, Yusuf Tanrikulu, Eugen Proschak, Matthias Rupp, Martin Weisel, Swetlana Derksen, Jan Hiss, Manuel Nietert, Lutz Franke und Alireza Givehchi für die durchweg gute Zusammenarbeit, auf wissenschaftlicher wie auch auf menschlicher Ebene.

Brigitte Scheidemantel-Geiß danke ich für ihr Organisationstalent und offenes Ohr bei Verwaltungsfragen. Auch Norbert Dichter bin ich zu tiefstem Dank verpflichtet für die hervorragende Rechneradministration, ohne die diese Arbeit nicht zustandegekommen wäre.

An dieser Stelle möchte ich auch jene Kooperationspartner und Studenten, die direkt in die Entstehung von Teilen dieser Arbeit involviert waren erwähnen: Vielen Dank an Marien de Bruyne und Melanie Hähnel von der FU Berlin (Kapitel 2, Florian Schwarte, André Brück, Kai Scheiffele (Kapitel 3), Volker Majczan (Kapitel 4) sowie Natalie Jäger und Joanna Wisniewska (Kapitel 3 und 4); Ihre Beiträge sind am Ende der jeweiligen Kapitel aufgeführt.

Mein ganz spezieller Dank geht an Andi, Anna, Christina, Felix, Henning, Inna, Iris, Julia, Leyla, Michi, Moritz, Rüdiger, Tina, Thorsten, Uli, Volker und allen anderen die zu den vielen schönen Erinnerungen beitrugen, die ich aus Frankfurt mitnehmen werde.

Am allermeisten möchte ich jedoch meiner Freundin Caroline sowie meinen Eltern und Geschwistern danken, die mit Ihrer andauernden Unterstützung großen Anteil am Gelingen dieser Arbeit hatten.

Curriculum vitae

Zur Person

Michael Schmuker

geboren am 26. 6. 1975

in Biberach an der Riß

Staatsangehörigkeit: deutsch

Familienstand: ledig

Schulische Ausbildung

1982 – 1986	Mittelberg-Grundschule, Biberach an der Riß
1986 – 1995	Wieland-Gymnasium, Biberach an der Riß
6/1995	Abitur mit Note 1,2

Hochschulausbildung

10/1996 – 9/2003	Studium der Biologie, Albert-Ludwigs-Universität Freiburg im Breisgau
9/1999 – 6/2000	ERASMUS-Stipendiat an der Université de Montpellier II, Frankreich, Studium der Computerwissenschaften und der Biologie
Sommer 2000 und Sommer 2001	Praktika bei Fa. Hoffmann-La Roche unter Anleitung von Dr. Gisbert Schneider

10/2002 – 9/2003	Diplomarbeit in der Abteilung von Prof. Dr. Ad Aertsen, Neurobiologie und Biophysik, Albert Ludwigs-Universität Freiburg: *"Modeling Homogeneity Detection in Primate Visual Cortex with Spiking Neurons"*, unter Anleitung von Dr. Marc-Oliver Gewaltig (Honda Research Institute Offenbach) und Dr. Thomas Wachtler (Uni Freiburg)
10/2003	Diplom der Biologie, mit den Prfungsfächern Neurobiologie, Biochemie, Bioinformatik (Note 1,4)
10/2003 – 12/2006	Promotion in der Arbeitsgruppe von Prof. Dr. Gisbert Schneider an der Johann Wolfgang Goethe–Universität Frankfurt am Main: *"Analysis of Coding Principles in the Olfactory System and their Application in Cheminformatics"*
seit 1/2007	Postdoktorale Arbeit innerhalb eines Projekts des Bernstein Center for Computational Neuroscience (BCCN) Berlin mit Prof. Randolf Menzel und Dr. Martin Nawrot

Preise und Auszeichnungen

Erster Preis für das Poster "A novel method for processing and classification of odorants inspired by insect olfaction", präsentiert auf der 2^{nd} *German Conference on Cheminformatics*, verliehen durch die Gesellschaft Deutscher Chemiker (GDCh), Goslar, November 2006.

Abiturpreis für die beste Leistung im Fach Chemie, gestiftet vom Fonds der chemischen Industrie, Biberach an der Riß, Juni 1995.

List of publications

Peer-reviewed articles

Schmuker, M., de Bruyne, M., Hhnel, M. and Schneider, G. (2007) Predicting olfactory receptor neuron responses from odorant structure. *Submitted for publication.*

Renner, S., Hechenberger, M., Noeske, T., Böcker, A., Jatzke, C., Schmuker, M., Parsons, C.G., Weil, T. and Schneider, G. (2007). Scaffold-hopping by 3D-pharmacophores and neural network ensembles. Accepted for publication in *Angewandte Chemie*.

Schmuker, M., Schwarte, F., Brück, A., Proschak, E., Tanrikulu, Y., Givehchi, A., Scheiffele, K. and Schneider, G. (2007). SOMMER: Self-Organizing Maps for Education and Research. *Journal of Molecular Modeling*, 13(1):225–228.

Meissner, M., Schmuker, M., and Schneider, G. (2006). Optimized particle swarm optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics*, 7:125.

Schmuker, M., Givehchi, A., and Schneider, G. (2004). Impact of different software implementations on the performance of the maxmin method for diverse subset selection. *Mol Divers*, 8(4):421–425.

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G.I. and Schneider, G. (2001). Deciphering apicoplast targeting signals - feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. *Gene* 280(1–2):19–26.

Diploma thesis

Schmuker, M. (2003). Modeling homogeneity detection with spiking neurons in primate visual cortex. Diploma Thesis, Albert-Ludwigs-Universität Freiburg

im Breisgau, Germany.

Talks

Schmuker, M.: Deciphering the olfactory code to enhance chemical pattern recognition. Seminar talk at the Bernstein Center for Computational Neuroscience Berlin, Germany, February (2007).

Schmuker, M.: Modeling olfactory coding of chemical information for pattern recognition. Seminar talk at the Department of Neural Information Processing, Universität Ulm, Germany, November (2006).

Schmuker, M.: Defining chemical space to predict olfactory receptor neuron responses from molecular structure. Seminar talk at the Max Planck Institute for Medical Optics, Heidelberg, Germany, May (2006).

Schmuker, M.: A model for surface detection in primate visual cortex. Seminar talk at the Institute for Theoretical Biology, Humboldt Universität zu Berlin, Germany, March (2003).

Conference contributions

Schmuker, M., and Schneider, G.: A novel method for processing and classification of odorants inspired by insect olfaction. Poster, 2^{nd} *German Conference on Cheminformatics*, Goslar, Germany, November (2006).

Schmuker, M., de Bruyne, M., Hähnel, M. and Schneider, G.: Predicting drosophila olfactory receptor responses from odorant molecular structure - a journey into chemical space. Poster, 5^{th} *Forum of European Neuroscience (FENS)*, Vienna, Austria, July (2006).

Schmuker, M., de Bruyne, M., Hähnel, M. and Schneider, G.: A journey into chemical space to predict *Drosophila*'s olfactory receptor neuron responses. Pos-

ter, *14th Spring School in Life Sciences "Sense of Smell"*, Jerusalem, Israel, April (2006).

Schmuker, M., de Bruyne, M. and Schneider, G.: Towards nature-derived coding of odor molecules. Poster, *1st German Conference on Cheminformatics*, Goslar, Germany, November (2005).

Schmuker, M., de Bruyne, M. and Schneider, G.: Towards predicting olfactory receptor responses. Poster, *30th Göttingen Neurobiology Conference*, February (2005).

Schmuker, M., Givehchi, A., Brück, A., Proschak, E., Scheiffele, K., Schwarte, F., Tanrikulu, Y. and Schneider G: The SOMMER project: 3D SOMs in research and teaching. Software presentation, *18th CIC Workshop*, Boppard, Germany, November (2004).

Schmuker, M., Koerner, U., Koerner, E., Gewaltig, M.O., Wachtler, T. and Aertsen, A.: A model of rapid surface detection in primate visual cortex. Poster, *29th Göttingen Neurobiology Conference*, June (2003).