Sequence-based prediction of olfactory receptor responses

Shashank Chepurwar[1], Abhishek Gupta[1,2], Rafi Haddad[3], Nitin Gupta[1,*]

[1] Department of Biological Sciences and Bioengineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh 208016, India

[2] Department of Chemistry, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh 208016, India

[3] The Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat-Gan, 5290002, Israel

* Corresponding author, email address: guptan@iitk.ac.in

Keywords: olfaction, olfactory, anopheles, drosophila, receptors

Summary

Computational prediction of how strongly an olfactory receptor responds to various odors can help in bridging the widening gap between the large number of receptors that have been sequenced and the small number of experiments measuring their responses. Previous efforts in this area have predicted the responses of a receptor to some odors, using the known responses of the same receptor to other odors. Here we present a method to predict the responses of a receptor without any known responses, by using available data about the responses of other conspecific receptors and their sequences. We applied this method to olfactory receptors in insects *Drosophila melanogaster* (both adult and larva) and *Anopheles gambiae*, and to mouse and human olfactory receptors. We found the predictions to be in significant agreement with the experimental measurements. The method also provides clues about the response-determining positions within the receptor sequences.

Introduction

Odor sensing begins with the binding of odor molecules to olfactory receptors (ORs) expressed on the membranes of olfactory sensory neurons. An organism often expresses a repertoire of tens or hundreds of types of receptors, which combinatorially can detect a very large number of odor stimuli (Malnic et al., 1999). The amino acid sequences of these receptors – and consequently their tuning profiles to odors – vary from species to species, in accordance with their environmental niches.

Identification of the odors that a given OR responds to, also known as deorphanization, is fundamental to understanding olfactory processing. In cases where the sensory neurons expressing a particular OR are known and easy to target with electrodes, the cognate odors can be identified using extracellular recordings from those neurons (Lu et al., 2007; Olsson and Hansson, 2013; Li et al., 2018). An alternative approach is to express the OR in an easy-to-target cell using heterologous expression systems — including cell lines, *Xenopus* oocytes, or the empty neuron system; these have been used in different species including humans (Touhara et al., 1999; Mainland et al., 2015), mice (Oka et al., 2006; Saito et al., 2009), fruit flies (de Bruyne et al., 2001; Hallem and Carlson, 2006; Lin and Potter, 2015), mosquitoes (Carey et al., 2010a; Wang et al., 2010), moths (de Fouchier et al., 2017), and tsetse flies (Chahda et al., 2019). However, such experiments need to be performed separately for each receptor and are time-consuming; moreover, some receptors do not express well in the heterologous systems (Ronderos et al., 2014). High-throughput methods for identifying OR-odor interactions have also been developed

33 recently, but these require elaborate experimental pipelines for each species (Jiang et al., 2015; Hu and
34 Matsunami, 2018; Jones et al., 2019). The high-throughput methods also generate many false positives,
35 so it is recommended that the identified responses be further verified with targeted experiments
36 (Nishizumi and Sakano, 2015).

37 With lowering costs of sequencing, the sequences of ORs are rapidly becoming available for various
38 organisms, including multiple species of fruit flies (Drosophila 12 Genomes Consortium, 2007; Gardiner
39 et al., 2008; Ramasamy et al., 2016), mosquitoes (Leal et al., 2013; Neafsey and Waterhouse, 2015;
40 Lombardo et al., 2017), bees (Elsik et al., 2015; Karpe et al., 2017) and tsetse flies (Attardo et al., 2019).
41 However experimental data about odor responses are available for relatively few ORs. Computational
42 predictions can potentially bridge this gap. Indeed, computational methods have been developed to
43 predict the response of an OR to new odors based on its known responses to other odors, by analyzing
44 the chemical structure (Schmuker et al., 2007), molecular volume (Saberi and Seyed-Allaei, 2016), or
45 other physicochemical parameters of odor molecules ((Haddad et al., 2008; Boyle et al., 2013; Gabler et
46 al., 2013; Bushdid et al., 2018; Kepchia et al., 2019)). However, these methods can only be used for ORs
47 whose responses to some odors are known.

48 Here, we present a computational approach for predicting the responses of an OR even when no odor
49 responses are available for that OR; our method instead uses the known responses of other conspecific
50 ORs and the sequence similarities among them. We demonstrate the effectiveness of this approach by
51 comparing our predictions with the experimentally measured responses recordings in insects. Our
52 method also provides insights into the response-determining positions in the receptor sequences.

53

54 **Results**

55 **Lack of correlation between OR sequences and responses**

56 As odor responses of an OR depend on its three-dimensional structure, which in turn depends on the
57 amino-acid sequence, we expected that the sequence similarity between a pair of ORs will relate to the
58 similarity in their responses. To test this idea, we used large scale datasets of odor responses of 24 ORs
59 in *D. melanogaster* (Hallem and Carlson, 2006) and 50 ORs in *A. gambiae* (Carey et al., 2010) to sets of
60 110 odors; these odor sets were partially different between the two species but identical for all ORs
61 within a species. For each pair of ORs within a species, we calculated the distance between their 110-
62 length response vectors (**see Methods**). We quantified amino-acid sequence similarity for each pair of

63 ORs within a species using global sequence alignment. While we expected to see a negative correlation
64 between sequence similarity and response distance, we surprisingly found no correlation at all (**Fig. 1**):
65 the Pearson's correlation coefficient was 0.0018 (P = 0.97, N = 276 pairs of receptors) in *D.*
66 *melanogaster*, and 0.026 (P = 0.35, N = 1225 pairs) in *A. gambiae*.

67

68 **Identification of response-determining positions**

69 The task of sequence-based predictions is made challenging by the very low level of conservation among
70 the ORs: the percentage amino-acid identity between pairs of ORs was only 20.23 ± 4.48 % in *D.*
71 *melanogaster* and 20.60 ± 8.17 % in *A. gambiae* (mean ± s.d. calculated over all pairs within a species).
72 The lack of correlation between the sequences and the responses is partly because only a small fraction
73 of residues in the sequence are involved in determining odor-specificity, and the similarity of these
74 residues is not represented well by the overall sequence similarity (Man et al., 2007). However, as there
75 is very little structural information available for ORs, the positions of these response-determining
76 residues are not known. We took an empirical approach to rank each position in the multiple sequence
77 alignment of all receptor sequences according to its importance in determining the odor responses (see
78 **Methods**). Within a species, the ranking was done jointly over all ORs to avoid overfitting. Using this
79 ranked list of positions for each species, we selected the top 20 positions that should help in predicting
80 the odor responses of new ORs.

81 We first confirmed that the subsequences of amino acids formed by these 20 positions in each OR were
82 indeed related to the responses. We found the correlation between the response distance and sequence
83 similarity (calculated using only the subsequences formed by the 20 positions) to be -0.56 (P= 1.7×10^{-24},
84 N = 276 pairs; **Fig. 2A1**) in *D. melanogaster*, and -0.39 (P = 1.7×10^{-45}, N = 1225 pairs; **Fig. 2A2**) in *A.*
85 *gambiae*. Compared to the result seen with the full sequence (**Fig. 1**), these correlations were
86 significantly negative. Similar negative correlations were observed even if we used only half of the odors
87 for calculating the top 20 positions and the other half for calculating the response similarity
88 (**Supplementary Fig. S1**). If we varied the number of top ranked positions used in the analysis, the
89 correlation coefficients did not change for numbers between 10 and 40 but gradually reduced for larger
90 numbers of positions (**Figs. 2B1,2**), suggesting that lower-ranked residues are not important for the
91 prediction of responses. Subsequently, we used the top 20 positions as the response-determining

92 positions in our analysis (**Figs. 2B1,2** shows that our results would not change if we chose some other
93 number between 50 % and 200 % of this number).

94 We then checked the robustness of the top 20 positions by reducing the number of receptors used for
95 ranking the positions. When we removed one receptor (this was done by removing each receptor, one
96 at a time, to get the averages and error bars), we found that on average ~16 of the 20 positions in *D.*
97 *melanogaster* and ~19 of the 20 positions in *A. gambiae* were retained in the top 20. The number of
98 retained positions dropped gradually as more receptors were left out (again, all possible combinations
99 were tested for each number). But even with half of the receptors removed, at least 8 of the original 20
100 positions remained in top 20 in both the species (**Figs. 2C1,2**). This analysis indicated that the selected
101 positions were robust to minor perturbations in the datasets.

102 Although the top 20 alignment positions were chosen independently in *D. melanogaster* and *A.*
103 *gambiae*, we found that 3 of the top 20 positions were conserved between the species. This number
104 was three times the number of conserved positions expected by chance (20 x 20/369 = 1.08, where 369
105 is the number of alignment positions that were ranked in both the species), further increasing our
106 confidence in the reliability of the identified positions.

107

108 **Predicting responses using the identified positions**

109 Having identified reasonable candidates for response-determining positions, common to all ORs within a
110 species, we asked if it is possible to reliably predict the responses of a query OR to a panel of odors using
111 only the sequence information about the amino acids present at the identified positions. We tested this
112 idea using a simple and intuitive approach, in which the response of a query OR was estimated as the
113 mean of the responses of known ORs, weighted by their similarity to the query OR at the response-
114 determining positions using the BLOSUM62 amino-acid substitution matrix (see **Methods**).

115 We first applied this approach to predict the responses of the 24 and the 50 ORs in the *D. melanogaster*
116 and *A. gambiae* datasets, respectively. By turn, each OR was treated as a query OR and its response was
117 predicted using the responses of the remaining ORs in the same species (excluding the query OR). We
118 found that the predicted response profiles (i.e., the 110-length vectors of odor responses) of the ORs
119 were correlated with their actual response profiles (**Figs. 3A,B**); the correlations were positive and
120 statistically significant in 21 of the 24 ORs in *D. melanogaster*, and 29 of the 50 ORs in *A. gambiae*. Mean

121 of these correlation values was significantly greater than zero: 0.46 (P = 2.53 x 10^{-7}, t-test; P = 1.63 x 10^{-5},
122 sign-rank test; N = 24 ORs) in *D. melanogaster,* and 0.25 (P = 1.18 x 10^{-4}, t-test; P = 1.19 x 10^{-4}, sign-rank
123 test; N = 50 ORs) in *A. gambiae*. In contrast, control predictions obtained by shuffling the response
124 matrices (see **Methods**) were not correlated with the actual responses (**Fig. 3C**): mean correlation values
125 were 0.0002 (P = 0.95, t-test; N = 24 ORs) in *D. melanogaster,* and 0.0008 (P = 0.65, t-test; N = 50 ORs) in
126 *A. gambiae*.

127 The correlation coefficient between the actual and the predicted odor response vectors tells whether
128 these responses have the same relative magnitude across different odors, but does not tell how close
129 the two responses are in absolute terms (a predicted response that is several times larger than the
130 actual response for every odor will have a perfect correlation). We therefore also used a distance-based
131 metric and checked whether our predictions were closer to the actual responses than the *control*
132 predictions were. We define $D_{Pred,Actual}$ as the absolute difference between the predicted response
133 and the actual response of an OR to an odor, and $D_{Ctrl,Actual}$ as the absolute difference between the
134 control prediction and the actual response. Thus, these two terms indicate the error in the actual and
135 the control predictions. If our predicted response is better (closer to the actual response) than the
136 control prediction, the value of $D_{Pred,Actual} - D_{Ctrl,Actual}$ should be negative. The heat maps in **Fig. 3D**
137 show the comparisons for all the predictions in *D. melanogaster* and *A. gambiae* and reveal a very high
138 abundance of negative values. In *D. melanogaster*, the average $D_{Pred,Actual}$ of an OR (28.60) was
139 smaller than average $D_{Ctrl,Actual}$ (47.84) by 19.24 spikes (P = 7.30 x 10^{-15}, paired t-test; N = 24 ORs; **Fig.**
140 **3E1**), an improvement of more than 40 %. In *A. gambiae*, the average $D_{Pred,Actual}$ (20.11) was smaller
141 than average $D_{Ctrl,Actual}$ (33.29) by 13.18 spikes (P = 8.46 x 10^{-26}, paired t-test; N = 50 ORs; **Fig. 3E2**),
142 again an improvement of about 40 %. Thus, our predicted responses using the sequence similarity at the
143 response-determining positions were significantly closer to the actual responses than random
144 predictions were. These improvements were substantial considering the absolute values of odor
145 responses in *D. melanogaster* (mean ± s.d. = 32.43 ± 49.29 spikes, N = 24 x 110 OR-odor combinations)
146 and *A. gambiae* (mean ± s.d. = 20.31 ± 37.35 spikes, N = 50 x 110).

147

148 **Test on an independent dataset**

149 Next, we collected the responses of 26 other ORs in *D. melanogaster* from 10 different studies (de
150 Bruyne et al., 1999, 2001, 2010; Dobritsa et al., 2003; Goldman et al., 2005; Kreher et al., 2005, 2008;

151 Marshall et al., 2010; Ronderos et al., 2014; Dweck et al., 2015) deposited in the DoOR database (Münch
152 and Galizia, 2016). Each of these ORs had responses available for some of the 110 odors used previously;
153 overall, 68 of the previously used odors were represented in the new dataset. These 26 ORs were
154 entirely different from the 24 ORs used previously in selecting the response-determining positions, and
155 therefore provided an independent test for our approach. Using the set of top 20 response-determining
156 positions as selected earlier from the original *D. melanogaster* dataset, we found that the responses
157 predicted for the 26 novel ORs were significantly better with our approach compared to the control
158 predictions (**Fig. 4A,B**): the average $D_{Pred,Actual}$ of an OR (37.01) was smaller than average $D_{Ctrl,Actual}$
159 (51.36) by 14.35 spikes (P = 1.13 x 10^{-7}, paired t-test; P = 5.96 x 10^{-8}, sign-rank test; N = 26 ORs), an
160 improvement of about 28 %. Overall, among the 506 OR-odor responses predicted (non-grey squares in
161 **Fig. 4A**), in 380 (75.1%) cases the predictions were better than the control predictions. These results
162 show that once the response-determining positions have been identified using known responses of
163 some ORs in a species, the responses of other conspecific ORs can be predicted reasonably using only
164 their sequences.

165 While predicting the responses of the new ORs above, we had relied on the known responses of 24 ORs.
166 Can similar predictions be made if responses of fewer ORs are available? To check this, we sampled
167 subsets of ORs from the original dataset of 24 ORs, and used these samples to select response-
168 determining positions and make the response predictions. The sizes of these samples were varied from
169 3 to 21 in intervals of 3, and for each size the sampling was performed 10 times. We found that the
170 difference between the predicted and the actual response ($D_{Pred,Actual}$) decreased gradually as the
171 number of sampled ORs increased but saturated for 15 or more ORs (**Fig. 4C**). A similar trend was
172 observed when we checked how the number of sampled ORs affected the fraction of predictions that
173 were better than control (**Fig. 4D**). These results suggest that a relatively small set of ORs with known
174 responses may be enough to make reasonable predictions for new ORs.

175 **Predicting responses in larval *Drosophila***

176 Although the OR responses in larval *Drosophila melanogaster* differ from the responses in adult
177 *Drosophila* (Kreher et al., 2008), the ORs themselves are the same at the molecular level. Therefore, we
178 checked if the top positions identified in adult *Drosophila* can be used to predict the larval responses.
179 We tested this idea on a dataset of 21 ORs and 26 odors (each available at two different concentrations,
180 10^{-2} and 10^{-4}) from larval *Drosophila* (Kreher et al., 2008). For each concentration, we found that the
181 predictions were significantly better than control predictions (**Supplementary Fig. S2**): at 10^{-2}

182 concentration, the average $D_{Pred,Actual}$ (38.01) was smaller than average $D_{Ctrl,Actual}$ (54.72) by 16.71
183 spikes (P = 6.32 x 10^{-12}, paired t-test; N = 21 ORs); at 10^{-4} concentration, the average $D_{Pred,Actual}$ (12.38)
184 was smaller than average $D_{Ctrl,Actual}$ (16.7) by 4.32 spikes (P = 1.1 x 10^{-8}, paired t-test; N = 21 ORs).

185

186 **Properties of the response-determining positions**

187 Our results show that the identified subset of sequence positions can allow reliable prediction of the
188 neural responses for new ORs. To check where these top 20 positions are located within the protein
189 structure, we visualized these positions in the secondary structure of two representative ORs in *D.*
190 *melanogaster* and *A. gambiae* using the Protter tool (Omasits et al., 2014). These residues were not
191 limited to any one region, but were instead found in various parts of the protein (**Fig. 5**). It is important
192 to note that our analysis does not imply a causal or a mechanistic role for each of these residues in
193 interacting with the odor molecules, but only suggests that these residues may be involved in some
194 aspect of determination of odor responses, perhaps indirectly. Notably, the region with the highest
195 density of the top 20 positions was the seventh transmembrane helix, including 4 positions in *D.*
196 *melanogaster* and 5 in *A. gambiae* (**Fig. 5**; similar numbers were seen in other receptors). Moreover, of
197 the three response-determining positions that were conserved between the two species, two were
198 located in the seventh transmembrane helix. These observations are in agreement with previous studies
199 suggesting that the seventh transmembrane helix plays an important role in determining odor specificity
200 (Hughes et al., 2014; Ray et al., 2014).

201 Next, we collected data from various experimental studies that have mutated individual residues in
202 various ORs and have checked their effect on odor responses. We collected a total of 47 such mutated
203 positions from 5 studies (Nichols and Luetje, 2010; Nakagawa et al., 2012; Xu and Leal, 2013; Hughes et
204 al., 2014; Ray et al., 2014) and compared them with our identified set of response-determining
205 positions. In addition to *D. melanogaster* and *A. gambiae*, these studies also included data from *Bombyx*
206 *mori* and *Culex quinquefasciatus.* Sequences from these other species were added to our existing
207 multiple sequence alignment (see **Methods**), so that the mutated positions in all these species can be
208 compared with our alignment positions.

209 Given 20 response-determining positions in *D. melanogaster* and *A. gambiae* each, and with 3 being
210 common between the two species, 37 unique positions in the multiple sequence alignment were
211 marked as the top positions out of a total of 817. Thus, each of the 47 mutated positions in the curated

212 experimental set had a 37/817 probability of matching a top position, and the expected number of
213 matches by chance = 47 x 37/817 = 2.13. So, while only about 2 of the 47 mutated positions were
214 expected to be among the top positions by chance, we found as many as 9 of them to be among the top
215 positions ($P = 2.28 \times 10^{-4}$; binomial test), providing experimental support to our response-determining
216 positions. These 9 experimentally studied mutations, which map to the top positions in our alignment,
217 are listed in **Table 1** (further, **Table S1** shows all positions). Among these, residue 146 in OR85b of *D.*
218 *melanogaster* is one of the residues expected to play a crucial role in the activation of the odorant
219 (Nichols and Luetje, 2010). Residue 109 mutant in OR1 of *B. mori* was found to change the reversal
220 potential and rectification index of the BmOR1-Orco complex (Nakagawa et al., 2012). Point mutation of
221 residues 165 and 194 in BmorOR1 of *B. mori* led to significant reduction in response to bombykol
222 pheromone (Xu and Leal, 2013). Similarly, odor responses in *A. gambiae* were found to change with
223 mutation of residue 369 in OR15 (Hughes et al., 2014), and residues 405 and 406 in OR1 (Ray et al.,
224 2014). Among the remaining 38 out of the 47 mutations in the curated set, 10 and 8 were found to be
225 within a distance of 1 and 2 residues, respectively, from a top column. Thus, in total, 27 of the 47
226 experimentally confirmed important positions matched or were adjacent to our computationally
227 predicted response-determining positions.

228

229 **Prediction of vertebrate olfactory responses**

230 As our method does not make any assumption about the composition of the receptor sequences, it
231 should be applicable to any family of chemosensory receptors. We next sought to check its utility in
232 predicting the responses of the vertebrate ORs. Saito et al. (Saito et al., 2009) have provided a dataset of
233 EC50 values of 52 mouse and 10 human ORs that responded to at least one of the 63 odors tested. We
234 extracted a binary response matrix (1 for response, 0 for no response) from this dataset: among the
235 mouse responses (52 x 63 = 3276 values), this matrix included 262 (approximately 8%) 1s and rest were
236 0s. Using the multiple sequence alignment provided in the same study (Saito et al., 2009) and mouse
237 data, we calculated the top 20 response determining positions. We then calculated the response of each
238 mouse receptor using the weighted average of the responses of the remaining receptors, as done for
239 insect receptors (see **Methods**); for comparison with actual responses, this value was also converted to
240 a binary response using a threshold, set to keep the total fraction of 1s in the predicted responses also
241 at 8%. The predictions thus obtained had a false alarm rate of about 7% and a hit rate of about 21%,
242 resulting in sensitivity index d' = 0.68 ($P < 0.001$, estimated by comparing the d' value with those

243 obtained for 1000 shuffled control predictions). We next asked if human OR responses could also be
244 predicted using the mouse training data. Using the top positions, the responses, and the threshold
245 identified from the mouse data, we predicted the binary responses of the 10 human ORs. Among the 10
246 x 63 values in the original human dataset, 78 (~12%) were 1s. Our prediction yielded 581 0's and 49 1s,
247 with a false alarm rate of about 7% and a hit rate of about 13% (d' = 0.34, P = 0.047 based on 1000
248 shuffled predictions). In summary, these results show that our method can also be used to predict
249 vertebrate OR responses with better than chance accuracy.

250

251 **Discussion**

252 Sequence-based approaches to response prediction require a reasonably high homology among the
253 receptors. In vertebrates, receptors with less than 40 % sequence identity were found to have little
254 overlap in their responses (Li et al., 2015). The insects ORs presented a particularly difficult test case, as
255 they have only about 20 % homology among conspecific ORs; indeed, we found that there was no
256 correlation between the overall sequence similarity and the response similarity among the ORs. We
257 were able to overcome this difficulty by focusing on subsets of positions that may be important for
258 determining odor responses. Our approach was successful in predicting the responses for not only the
259 set of fly and mosquito receptors whose data was used while developing the method, but also for a
260 completely novel set of receptors whose responses were taken from 10 independent experimental
261 datasets. Further, we found that the predictions require only about 15 ORs with known responses (**Fig.
262 4**). This computational approach can be particularly useful for receptors expressed in neurons that are
263 difficult to access in electrophysiology experiments, or for receptors that do not express well in
264 heterologous expression systems (Ronderos et al., 2014). Finally, we also showed that our method can
265 also be used for vertebrate receptors.

266 Our method allows prediction of responses of novel ORs for already studied odors. Previously developed
267 methods allow prediction of responses of already studied ORs to novel odors (Schmuker et al., 2007;
268 Haddad et al., 2008; Boyle et al., 2013; Gabler et al., 2013; Saberi and Seyed-Allaei, 2016; Bushdid et al.,
269 2018; Kepchia et al., 2019). These two approaches are complementary and can be combined in future
270 work to enable the prediction of the responses of novel ORs to novel odors, which would help with
271 large-scale deorphanization. While applying these approaches, care must be taken regarding the
272 concentrations of the odors in the training data, as changes in odor concentration can affect neural

273 responses (Hallem and Carlson, 2006; Olsen et al., 2010) and behavior (Wright et al., 2005; Yarali et al.,
274 2009). Another caveat is that these approaches are currently limited to predicting the response
275 magnitudes, and ignore the temporal patterns of spikes observed in the olfactory receptor neurons
276 (Spors et al., 2006; Raman et al., 2010; Montague et al., 2011; Su et al., 2011; Grillet et al., 2016; Egea-
277 Weiss et al., 2018).

278 Unlike vertebrate ORs which are G-protein coupled receptors, insect ORs are heteromeric ligand-gated
279 ion channels (Wicher et al., 2008), composed of a specific OR and a conserved co-receptor Orco (Larsson
280 et al., 2004). The structure of any insect receptor was not known until very recently when the structure
281 of Orco from fig wasp was studied using Cryo-EM (Butterwick et al., 2018). The approach of predicting
282 responses using the three-dimensional modeling of ORs, that has seen some success in vertebrates
283 (Bavan et al., 2014), is unlikely to be very productive in insects in the near future given the very low
284 sequence similarity among the ORs and with other proteins, and the paucity of available structures.
285 Therefore sequence-based computational methods will be important for receptors whose structures
286 and responses have not been determined experimentally. The sequence-based predictions are expected
287 to fare better when the responses being predicted are in the same range as the responses in the training
288 dataset, but may not fare as well in predicting uniquely strong ligands for a novel receptor (for the latter
289 goal, three-dimensional modeling may be necessary). Our computational predictions can be used to
290 identify the candidate ligands to be tested experimentally. The method can also be extended to other
291 types of chemosensory receptors, such as the ionotropic receptors and the gustatory receptors.

292 We have used a simple algorithm in our approach, where we first identify the important positions in the
293 sequences and then use the similarity at those positions to make the predictions. It may be possible to
294 achieve good predictions without explicitly identifying the important positions and instead using a
295 machine learning-based approach that uses all the residues and learns a non-linear mapping between
296 the sequences and the responses. However, we preferred the simpler deterministic approach over the
297 blackbox approaches typically used in machine learning because the deterministic approach provides a
298 transparent rationale for and information about the sequence positions that are used for making the
299 predictions. Expectedly, a majority of the identified positions were located in the extracellular loops or
300 transmembrane helices (particularly the seventh transmembrane, which has been experimentally shown
301 to be important for determining responses), but some were also found in intracellular loops. A previous
302 study has shown that mutation of intracellular residues can affect the electrophysiological responses of
303 neurons (Nakagawa et al., 2012). We emphasize that our approach only looks for correlations between

304 the residues and responses, and it is likely that some of the identified positions, particularly those found
305 in intracellular regions, affect the odor responses indirectly rather than through direct interactions with
306 the odor molecules. By comparing the 47 positions previously found to be important in mutagenesis
307 studies in various organisms (Nichols and Luetje, 2010; Nakagawa et al., 2012; Xu and Leal, 2013;
308 Hughes et al., 2014; Ray et al., 2014), we found that 9 of them (a statistically significant fraction, more
309 than 4 times larger than that expected by chance) mapped to the top positions identified in our analysis
310 (**Table 1**). This shows that our method is using reliable positions for making the predictions. The other
311 top positions that have not been experimentally studied so far are good candidates for future
312 mutagenesis studies, and could help in understanding the mechanisms of odor-OR interactions. As more
313 experimental evidence accumulates, the response predictions can be further improved by giving more
314 weightage to the experimentally verified positions among the top positions.

315 **Methods**

316 **Response datasets**

317 Responses of 24 *D. melanogaster* ORs were obtained from a previous study (Hallem and Carlson, 2006)
318 and responses of 50 *A. gambiae* ORs were obtained from another study (Carey et al., 2010), for a set of
319 110 odors. The response values indicate the odor-elicited spiking response of a neuron expressing the
320 receptor, minus the response for the diluent or the spontaneous firing rate. To serve as an independent
321 dataset, responses of 26 other *D. melanogaster* ORs were obtained from 10 studies (de Bruyne et al.,
322 1999, 2001, 2010; Dobritsa et al., 2003; Goldman et al., 2005; Kreher et al., 2005, 2008; Marshall et al.,
323 2010; Ronderos et al., 2014; Dweck et al., 2015) deposited in the Database of Odorant Responses
324 (DoOR) (Münch and Galizia, 2016), for 68 odors that were common with the earlier set of 110 odors. If
325 the response of a particular OR to an odor was reported in multiple studies, we used the average of
326 those values as the actual response.

327 **Sequence analysis**

328 The olfactory receptor sequences for *D. melanogaster* and *A. gambiae* were obtained from the Database
329 of Olfactory Receptors (Nagarathnam et al., 2014) and the (Hill et al., 2002) study, respectively. The
330 sequence similarity between any two receptor sequences was determined using the Needleman
331 Wunsch Algorithm using the BLOSUM62 scoring matrix (Needleman and Wunsch, 1970; Henikoff and
332 Henikoff, 1992). Multiple sequence alignment was performed for a total of 140 (62 in *D. melanogaster*
333 and 78 in *A. gambiae*) olfactory receptor sequences using MAFFT multiple sequence alignment program,

334 and additional sequences from other species (for validation) were added individually using the '-add'
335 option (Katoh and Frith, 2012).

336 The OCTOPUS tool (Viklund and Elofsson, 2008) was used to predict the membrane topology of the
337 receptors and the Protter tool was used to visualize the 2-dimensional topology of the receptors
338 (Omasits et al., 2014).

339 **Ranking the response-determining positions**

340 We identified the response-determining sequence positions for *D. melanogaster* and *A. gambiae* using
341 the following approach, which quantifies at each position how well the differences in amino acids across
342 the receptors correlate with the differences in their responses. These positions corresponded to
343 columns in the multiple sequence alignment, and were same for all the receptors within a species
344 (allowing different positions for each receptor can provide better correlations for individual receptors
345 due to overfitting but will have poor generalizability for other receptors).

346 From the total of 817 alignment columns in the multiple sequence alignment, we first eliminated the
347 columns which were either fully conserved or had gaps in more than half of the sequences, leaving 378
348 columns to be ranked in *D. melanogaster* and 392 columns to be ranked in *A. gambiae*, of which 369
349 were common to both species (see **Table S1**). We denote an element of the alignment by $A_{r,c}$ where c
350 indicates the alignment column and $r \in [1, 24]$ or $[1, 50]$ in *D. melanogaster* and *A. gambiae*,
351 respectively, indicates the row corresponding to a receptor with known responses. The following steps
352 were carried out separately for each species. To calculate the rank for each of the columns, we first
353 calculated a sequence similarity vector (SS_c) for each column, each element ($SS_{(r_1,r_2),c}$) of which
354 contained the pairwise amino acid similarity scores, calculated using BLOSUM62 matrix, between all
355 pairs of receptors (r_1, r_2) that did not have a gap in the alignment column c:

356 $$SS_{(r_1,r_2),c} = BLOSUM62\left(A_{r_1,c}, A_{r_2,c}\right) \forall\ (r_1, r_2)\ where\ A_{r_1,c} \neq \text{ gap } \&\ A_{r_2,c} \neq \text{ gap}$$

357 Similarly, we computed a response distance vector (DR_c^o) for the same pairs of ORs in the column, each
358 element ($DR^o_{(r_1,r_2),c}$) of which contained the absolute difference between the responses of receptors
359 r_1 and r_2 to an odor o:

360 $$DR^o_{(r1,r2),c} = \left|N^o_{r_1} - N^o_{r_2}\right| \forall\ (r_1, r_2)\ where\ A_{r_1,c} \neq \text{ gap } \&\ A_{r_2,c} \neq \text{ gap},$$

361 where N_r^o denotes the neural response of receptor r to odor o. We used the L1-norm instead of the
362 Euclidean distance to avoid giving extra weightage to larger values. Note that SS_c and DR_c^o are both
363 vectors of the same length; this length is equal to the number of all possible pairs of receptors that did
364 not have a gap in column c. We then calculated the Pearson's correlation coefficient (R_c^o) between the
365 sequence similarity (SS_c) and response difference (DR_c^o) vectors for each column c and each odor o. The
366 P-value of this correlation, denoted by P_c^o, was also noted. A negative value of R_c^o indicates that the
367 receptors that have high sequence similarity at the alignment position c also have high response
368 similarity (small distance) for odor o. Finally, we calculated the total score of a column as

371 $$S_c = -\sum_{\substack{o=1,\\ R_c^o<0}}^{110} \log P_c^o$$

369 Use of P-values allowed us to give more weightage to highly significant correlations. Columns with high
370 S_c scores were used subsequently as the response-determining positions.

372 **Response prediction**

373 We first generated the subsequences using the amino acids at the response-determining positions; this
374 was done for the receptors with known responses and also for the receptor whose response is being
375 predicted (*query*). Subsequence similarity ($SSS_{r,query}$) between the query receptor and each of the
376 known receptors was calculated by taking the average of BLOSUM62 score for all positions in the two
377 subsequences (excluding positions with gaps). The similarity values were then linearly scaled (denoted
378 by $\overline{SSS}_{r,query}$) to a range of 0 to 100 to avoid negative weights in the subsequent steps:

379 $$\overline{SSS}_{r,query} = \frac{SSS_{r,query} - \min_i SSS_{r_i,query}}{\max_i SSS_{r_i,query} - \min_i SSS_{r_i,query}} \times 100$$

380 The predicted response ($\widetilde{N}_{query}^o$) of the query receptor to an odor o was calculated as a weighted
381 average of the known responses:

382 $$\widetilde{N}_{query}^o = \frac{\sum_{j=1}^n (\overline{SSS}_{r_j,query} \times N_{r_j}^o)}{\sum_{j=1}^n \overline{SSS}_{r_j,query}}$$

383 where n is the number of receptors with known responses and $N_{r_j}^{o}$ denotes the neural response of
384 receptor r_j to odor o.

385 **Control Predictions**

386 Control predictions were obtained by shuffling the OR-odor response matrix and taking the shuffled
387 response as the control prediction for each OR-odor combination (this operation allowed generation of
388 random predictions while maintaining the overall statistical properties of the response values). The
389 shuffling was not limited to shifting of rows or columns, but included the shuffling of all elements in the
390 matrix. The shuffling was performed 50 times independently, and the results (correlations or $D_{Ctrl,Actual}$
391 values) were averaged over the 50 shufflings.

392 **Statistical analysis and code availability**

393 All analyses were performed in MATLAB. A modified version of Gramm plotting toolbox (Morel, 2018)
394 was used to draw the plots. We used Pearson correlation to measure the correlations, and t-tests and
395 Wilcoxon signed-rank tests to compare the means. All tests were two-tailed. The number of sample
396 points is shown for each test in the results. The code developed in this study can be accessed from
397 https://github.com/neuralsystems/OR_response_prediction.

398

399 **Acknowledgements**

400 We thank members of the N.G. laboratory for helpful comments.

401 **Funding**

402 This work was supported by the Wellcome Trust/DBT India Alliance Fellowship (grant number
403 IA/I/15/2/502091) awarded to N.G., and ISF-UGC joint research program in which R.H. was supported by
404 the ISF (grant No. 2307/15) and N.G. was supported by the UGC (F.No. 6-11/2016[IC]).

405 **Declaration of interests**

406 The authors declare no competing financial interests.

407 **Author Contributions**

408 N.G. conceived the study. S.C., A.G., R.H., and N.G. developed the methodology; S.C. and A.G. developed
409 the software and performed analysis; N.G. and R.H. secured funding; S.C., A.G., R.H., and N.G. wrote the
410 manuscript.

411

412

413 **References**

414 Attardo GM et al. (2019) The Glossina Genome Cluster: Comparative Genomic Analysis of the Vectors of African
415 Trypanosomes. bioRxiv.
416 Bavan S, Sherman B, Luetje CW, Abaffy T (2014) Discovery of Novel Ligands for Mouse Olfactory Receptor MOR42-
417 3 Using an In Silico Screening Approach and In Vitro Validation. PLoS One 9.
418 Boyle SM, McInally S, Ray A (2013) Expanding the olfactory code by in silico decoding of odor-receptor chemical
419 space. Elife:1–17.
420 Bushdid C, De March CA, Fiorucci S, Matsunami H, Golebiowski J (2018) Agonists of G-Protein-Coupled Odorant
421 Receptors Are Predicted from Chemical Features. J Phys Chem Lett 9:2235–2240.
422 Butterwick JA, del Mármol J, Kim KH, Kahlson MA, Rogow JA, Walz T, Ruta V, Mármol J, Kim KH, Kahlson MA,
423 Rogow JA, Walz T, Ruta V (2018) Cryo-EM structure of the insect olfactory receptor Orco. Nature 560:447–
424 452.
425 Carey AF, Wang G, Su C-Y, Zwiebel LJ, Carlson JR (2010) Odorant reception in the malaria mosquito Anopheles
426 gambiae. Nature 464:66–71.
427 Chahda JS, Soni N, Sun JS, Ebrahim SAM, Weiss L, Carlson JR (2019) The molecular and cellular basis of olfactory
428 response to tsetse fly attractants. PLOS Genet:1–22.
429 de Bruyne M, Clyne PJ, Carlson JR (1999) Odor coding in a model olfactory organ: the Drosophila maxillary palp. J
430 Neurosci 19:4520–4532.
431 de Bruyne M, Foster K, Carlson JR (2001) Odor coding in the Drosophila antenna. Neuron 30:537–552.
432 de Bruyne M, Smart R, Zammit E, Warr CG (2010) Functional and molecular evolution of olfactory neurons and
433 receptors for aliphatic esters across the Drosophila genus. J Comp Physiol A 196:97–109.
434 de Fouchier A, Walker WB, Montagne N, Steiner C, Binyameen M, Schlyter F, Chertemps T, Maria A, Francois M-C,
435 Monsempes C, Anderson P, Hansson BS, Larsson MC, Jacquin-Joly E (2017) Functional evolution of
436 Lepidoptera olfactory receptors revealed by deorphanization of a moth repertoire. Nat Commun 8:1–11.
437 Dobritsa AA, Naters WVDG Van, Warr CG, Steinbrecht RA, Carlson JR, Haven N, Vic C (2003) Integrating the
438 Molecular and Cellular Basis of Odor Coding in the Drosophila Antenna. Neuron 37:827–841.
439 Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature
440 450:203–218.
441 Dweck HKM, Ebrahim SAM, Farhan A, Hansson BS, Stensmyr MC (2015) Olfactory Proxy Detection of Dietary

442 Antioxidants in Drosophila. Curr Biol 25:455–466.
443 Egea-Weiss A, Renner A, Kleineidam CJ, Szyszka P (2018) High Precision of Spike Timing across Olfactory Receptor
444 Neurons Allows Rapid Odor Coding in Drosophila. ISCIENCE 4:76–83.
445 Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE (2015) Hymenoptera Genome Database :
446 integrating genome annotations in HymenopteraMine. Nucleic Acids Res 44:793–800.
447 Gabler S, Soelter J, Hussain T, Sachse S, Schmuker M (2013) Physicochemical vs. vibrational descriptors for
448 prediction of odor receptor responses. Mol Inform 32:855–865.
449 Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG (2008) Drosophila chemoreceptor gene evolution :
450 selection , specialization and genome size. Mol Ecol 17:1648–1657.
451 Goldman AL, Van Der Goes Van Naters W, Lessing D, Warr CG, Carlson JR (2005) Coexpression of two functional
452 odor receptors in one neuron. Neuron 45:661–666.
453 Grillet M, Campagner D, Petersen R, Mccrohan C, Cobb M (2016) The peripheral olfactory code in Drosophila
454 larvae contains temporal information and is robust over multiple timescales. Proc R Soc B 283:1–8.
455 Haddad R, Carmel L, Sobel N, Harel D (2008) Predicting the Receptive Range of Olfactory Receptors. PLoS Comput
456 Biol 4:0001–0008.
457 Hallem EA, Carlson JR (2006) Coding of Odors by a Receptor Repertoire. Cell 125:143–160.
458 Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–
459 10919.
460 Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ (2002) G
461 protein-coupled receptors in Anopheles gambiae. Science (80-) 298:176–178.
462 Hu XS, Matsunami H (2018) High-Throughput Odorant Receptor Deorphanization Via Phospho-S6 Ribosomal
463 Protein Immunoprecipitation and mRNA Profiling BT - Olfactory Receptors: Methods and Protocols. In
464 (Simoes de Souza FM, Antunes G, eds), pp 95–112. New York, NY: Springer New York.
465 Hughes DT, Wang G, Zwiebel LJ, Luetje CW (2014) A determinant of odorant specificity is located at the
466 extracellular loop 2-transmembrane domain 4 interface of an anopheles gambiae odorant receptor subunit.
467 Chem Senses 39:761–769.
468 Jiang Y, Gong NN, Hu XS, Ni MJ, Pasi R, Matsunami H (2015) Molecular profiling of activated olfactory neurons
469 identifies odorant receptors for odors in vivo. Nat Neurosci 18:1446–1454.
470 Jones EM, Jajoo R, Cancilla D, Lubock NB, Wang J, Satyadi M, Cheung R, de March C, Bloom JS, Matsunami H, Kosuri
471 S (2019) A Scalable, Multiplexed Assay for Decoding GPCR-Ligand Interactions with RNA Sequencing. Cell Syst
472 8:254–260.e6.
473 Karpe SD, Dhingra S, Brockmann A, Sowdhamini R (2017) Computational genome-wide survey of odorant receptors
474 from two solitary bees Dufourea novaeangliae (Hymenoptera : Halictidae) and Habropoda laboriosa (
475 Hymenoptera : Apidae). Sci Rep 7:1–13.
476 Katoh K, Frith MC (2012) Adding unaligned sequences into an existing alignment using MAFFT and LAST.

477 Bioinformatics 28:3144–3146.
478 Kepchia D, Xu P, Terryn R, Castro A, Schürer SC, Leal WS, Luetje CW (2019) Use of machine learning to identify
479 novel , behaviorally active antagonists of the insect odorant receptor co-receptor (Orco) subunit. Sci Rep
480 9:1–18.
481 Kreher SA, Kwon JY, Carlson JR (2005) The Molecular Basis of Odor Coding in the Drosophila Larva. Neuron 46:445–
482 456.
483 Kreher SA, Mathew D, Kim J, Carlson JR (2008) Translation of Sensory Input into Behavioral Output via an Olfactory
484 System. Neuron 59:110–124.
485 Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB (2004) Or83b encodes a broadly
486 expressed odorant receptor essential for Drosophila olfaction. Neuron 43:703–714.
487 Leal WS, Choo Y, Xu P, Silva CSB, Ueira-vieira C (2013) Differential expression of olfactory genes in the southern
488 house mosquito and insights into unique odorant receptor gene isoforms. Proc Natl Acad Sci U S A
489 110:18704–18709.
490 Li H, You Y, Zhang L (2018) Single Sensillum Recordings for Locust Palp Sensilla Basiconica. JoVE:e57863.
491 Li J, Haddad R, Santos V, Bavan S, Luetje CW (2015) Receptive range analysis of a mouse odorant receptor
492 subfamily. J Neurochem 134:47–55.
493 Lin C, Potter CJ (2015) Re-Classification of Drosophila melanogaster Trichoid and Intermediate Sensilla Using
494 Fluorescence-Guided Single Sensillum Recording. PLoS One:1–14.
495 Lombardo F, Salvemini M, Fiorillo C, Nolan T, Zwiebel LJ, Ribeiro JM, Arcà B (2017) Deciphering the olfactory
496 repertoire of the tiger mosquito Aedes albopictus. BMC Genomics:1–23.
497 Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon H-W, Pitts RJ, van Loon JJA, Takken W, Carlson JR, Zwiebel LJ
498 (2007) Odor Coding in the Maxillary Palp of the Malaria Vector Mosquito Anopheles gambiae. Curr Biol
499 17:1533–1544.
500 Mainland JD, Li YR, Zhou T, Liu WL in. L, Matsunami H (2015) Human olfactory receptor responses to odorants. Sci
501 data 2:1–9.
502 Malnic B, Hirono J, Sato T, Buck LB, Hughes H (1999) Combinatorial Receptor Codes for Odors. Cell 96:713–723.
503 Man O, Willhite DC, Crasto CJ, Shepherd GM, Gilad Y (2007) A Framework for Exploring Functional Variability in
504 Olfactory Receptor Genes. PLoS One 2:e682.
505 Marshall B, Warr CG, Bruyne M De (2010) Detection of Volatile Indicators of Illicit Substances by the Olfactory
506 Receptors of Drosophila melanogaster. Chem Senses 35:613–625.
507 Montague SA, Mathew D, Carlson JR (2011) Similar Odorants Elicit Different Behavioral and Physiological
508 Responses, Some Supersustained. J Neurosci 31:7891–7899.
509 Morel P (2018) Gramm : grammar of graphics plotting in Matlab. J open source Softw 3:1–4.
510 Münch D, Galizia CG (2016) DoOR 2 . 0 - Comprehensive mapping of Drosophila melanogaster odorant responses.
511 Sci Rep 6:1–9.

512 Nagarathnam B, Karpe SD, Harini K, Sankar K, Iftekhar M, Rajesh D, Giji S, Archunan G, Balakrishnan V, Gromiha
513 MM, Nemoto W, Fukui K, Sowdhamini R (2014) DOR - a Database of Olfactory Receptors - Integrated
514 Repository for Sequence and Secondary Structural Information of Olfactory Receptors in Selected Eukaryotic
515 Genomes. Bioinform Biol Insights 8:147–158.
516 Nakagawa T, Pellegrino M, Sato K, Vosshall LB, Touhara K (2012) Amino acid residues contributing to function of
517 the heteromeric insect olfactory receptor complex. PLoS One 7:1–9.
518 Neafsey DE, Waterhouse RM et al. (2015) Highly evolvable malaria vectors: The genomes of 16 Anopheles
519 mosquitoes. Science (80-) 347:1–8.
520 Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid
521 sequence of two proteins. J Mol Biol 48:443–453.
522 Nichols AS, Luetje CW (2010) Transmembrane segment 3 of Drosophila melanogaster odorant receptor subunit
523 85b contributes to ligand-receptor interactions. J Biol Chem 285:11854–11862.
524 Nishizumi H, Sakano H (2015) Decoding and deorphanizing an olfactory map. Nat Neurosci 18:1432–1433.
525 Oka Y, Katada S, Omura M, Suwa M, Yoshihara Y, Touhara K (2006) Odorant Receptor Map in the Mouse Olfactory
526 Bulb: In Vivo Sensitivity and Specificity of Receptor-Defined Glomeruli. Neuron 52:857–869.
527 Olsen SR, Bhandawat V, Wilson RI (2010) Divisive Normalization in Olfactory Population Codes. Neuron 66:287–
528 299.
529 Olsson SB, Hansson BS (2013) Electroantennogram and Single Sensillum Recording in Insect Antennae. In:
530 Pheromone Signaling: Methods and Protocols, pp 157–177. Totowa, NJ: Humana Press.
531 Omasits U, Ahrens CH, Müller S, Wollscheid B (2014) Protter: Interactive protein feature visualization and
532 integration with experimental proteomic data. Bioinformatics 30:884–886.
533 Raman B, Joseph J, Tang J, Stopfer M (2010) Temporally Diverse Firing Patterns in Olfactory Receptor Neurons
534 Underlie Spatiotemporal Neural Codes for Odors. J Neurosci 30:1994–2006.
535 Ramasamy S, Ometto L, Crava CM, Revadi S, Kaur R, Horner DS, Pisani D, Dekker T, Anfora G, Rota-stabelli O (2016)
536 The Evolution of Olfactory Gene Families in Drosophila and the Genomic Basis of chemical-Ecological
537 Adaptation in Drosophila suzukii. Genome Biol Evol 8:2297–2311.
538 Ray A, van Naters WG oe., Carlson JR (2014) Molecular determinants of odorant receptor function in insects. J
539 Biosci 39:555–563.
540 Ronderos DS, Lin C-C, Potter CJ, Smith DP (2014) Farnesol-Detecting Olfactory Neurons in Drosophila. J Neurosci
541 34:3959–3968.
542 Saberi M, Seyed-Allaei H (2016) Odorant receptors of Drosophila are sensitive to the molecular volume of
543 odorants. Sci Rep 6:1–11.
544 Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD (2009) Odor coding by a Mammalian receptor repertoire. Sci
545 Signal 2:1–14.
546 Schmuker M, De Bruyne M, Hähnel M, Schneider G (2007) Predicting olfactory receptor neuron responses from

547 odorant structure. Chem Cent J 1:1–10.
548 Spors H, Wachowiak M, Cohen LB, Friedrich RW (2006) Temporal Dynamics and Latency Patterns of Receptor
549 Neuron Input to the Olfactory Bulb. J Neurosci 26:1247–1259.
550 Su C, Martelli C, Emonet T, Carlson JR (2011) Temporal coding of odor mixtures in an olfactory receptor neuron.
551 Proc Natl Acad Sci 108:1–6.
552 Touhara K, Sengoku S, Inaki K, Tsuboi A, Hirono J, Sato T, Salano H, Haga T (1999) Functional identification and
553 reconstitution of an odorant. Proc Natl Acad Sci USA 96:4040–4045.
554 Viklund H, Elofsson A (2008) OCTOPUS: Improving topology prediction by two-track ANN-based preference scores
555 and an extended topological grammar. Bioinformatics 24:1662–1668.
556 Wang G, Carey AF, Carlson JR, Zwiebel LJ (2010) Molecular basis of odor coding in the malaria vector mosquito
557 Anopheles gambiae. Proc Natl Acad Sci 107:4418–4423.
558 Wicher D, Schäfer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, Hansson BS (2008) Drosophila odorant
559 receptors are both ligand-gated and cyclic-nucleotide- activated cation channels. Nature 452:1007–1011.
560 Wright GA, Thomson MGA, Smith BH (2005) Odour concentration affects odour identity in honeybees. Proc R Soc B
561 272:2417–2422.
562 Xu P, Leal WS (2013) Probing insect odorant receptors with their cognate ligands: Insights into structural features.
563 Biochem Biophys Res Commun 435:477–482.
564 Yarali A, Ehser S, Hapil FZ, Huang J, Gerber B (2009) Odour intensity learning in fruit flies. Proc R Soc B 276:3413–
565 3420.
566

567 **Legends**

568 **Fig. 1: Lack of correlation between response distance and sequence similarity among pairs of ORs.**
569 Scatter plots of distances in response vectors versus the amino-acid sequence similarity (a larger value
570 indicates more similarity), among pairs of ORs, reveal the lack of correlation, in both *D. melanogaster*
571 (**A1**) and *A. gambiae* (**A2**). Distance is calculated as the L1-distance between the two 110-length
572 response vectors containing number of spikes for each of the 110 odors. Sequence similarity is
573 measured as the Needleman-Wunsch alignment score. Each point corresponds to a pair of conspecific
574 ORs (n = 276 pairs in *D. melanogaster* and n = 1225 pairs in *A. gambiae*).

575

576 **Fig. 2: Analysis with response determining positions.**
577 **A1**, **A2**, Scatter plots of distances in 110-odor response vectors versus the sequence similarity measured
578 at the top 20 positions, among pairs of ORs, show significant negative correlation, in both *D.*
579 *melanogaster* (**A1**) and *A. gambiae* (**A2**). Each point corresponds to a pair of conspecific ORs (n = 276
580 pairs in *D. melanogaster* and n = 1225 pairs in *A. gambiae*).
581 **B1**, **B2**, Bars indicate the negative correlation between the distance in response vectors and the
582 sequence similarity measured using the top N positions, plotted as a function of N in the range of 10 to
583 100. Note that the correlation values are relatively stable in both the species in the range of 10-40 and
584 decrease gradually in magnitude for larger values of N.
585 **C1**, **C2**, Bars indicate the number of top 20 positions in *D. melanogaster* (**C1**) or *A. gambiae* (**C2**) retained
586 among the top 20 positions when the positions were recalculated after removing different numbers of
587 receptors from the dataset. For each number, all possible combinations of that many receptors were
588 removed and the average number of positions preserved among the top 20 is reported. Error bars
589 represent s.e.m. over all combinations for each number.

590

591 **Fig. 3: Reliability of responses predicted using the top 20 positions.**
592 **A,** Scatter plots of the actual and the predicted odor responses for three representative ORs each from
593 *D. melanogaster* and *A. gambiae*. The correlation coefficients between the two responses are indicated.
594 The responses were spread around the y=x line (dashed line, corresponding to ideal predictions).
595 **B1, B2,** Violin plots showing the correlation coefficient between the predicted and the actual response
596 vectors of ORs in *D. melanogaster* **(B1)** or *A. gambiae* (n = 50) **(B2)**. Each point corresponds to one OR (n
597 = 24 ORs in *D. melanogaster* and n = 50 ORs in *A. gambiae*). Note that the correlations are mostly

598 positive. Statistically significant positive correlations (with P less than 0.05) are shown with black
599 outlines.
600 **C1, C2,** Violin plots showing the correlation coefficient between the control predictions and the actual
601 response vectors of ORs in *D. melanogaster* **(C1)** or *A. gambiae* (n = 50) **(C2)**. Note that the correlations
602 are close to zero in most cases.
603 **D1, D2,** Color maps showing the relative prediction error for each OR-odor pair, for the set of 110 odors,
604 and all 24 ORs in *D. melanogaster* **(D1)** or all 50 ORs in *A. gambiae* **(D2)**. The relative prediction error is
605 measured as $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$, where $D_{Pred,Actual}$ is the absolute distance between the
606 predicted and actual response, and $D_{Ctrl,Actual}$ is the absolute distance between the control prediction
607 and the actual response. The abundance of negative values (red shades) indicates that the predictions
608 are usually better than the control. Column averages and row averages of the matrix values are shown
609 on top and right, respectively.
610 **E1, E2,** Violin plots shows the average (over odors) of $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$ for each of the 24 ORs in
611 *D. melanogaster* **(E1)** or 50 ORs in *A. gambiae* **(E2)**.
612 In violin plots, horizontal line and the error bar indicate the mean and the s.e.m., respectively, within a
613 species.
614
615 **Fig. 4: Response prediction on an independent set of ORs.**
616 **A,** Color map showing the relative prediction error (measured as $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$) for each OR-
617 odor pair, for the set of 68 odors and 26 novel ORs. Pairs whose values were not available are shown in
618 grey. Note the high frequency of negative values (red) and very low frequency of positive values (blue),
619 indicating that the predictions were closer to the actual value than the control were in most cases.
620 **B**, Violin plots shows the average (over odors) of $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$ for each of the 26 ORs.
621 Horizontal line and the error bar indicate the mean and the s.e.m., respectively.
622 **C,** $D_{Pred,Actual}$ is plotted as a function of the number of ORs with known responses used in making the
623 predictions. These different numbers of ORs were sampled from the original set of 24 ORs with known
624 responses. For each number, the sampling was performed 10 times; for each sampling, selection of
625 response-determining positions and prediction of responses was performed.
626 **D,** For each number of sampled ORs, we calculated the fraction of the 506 OR-odor pairs (non-grey
627 values in panel **A**) for which the predicted responses were closer to the actual values than the control
628 predictions were. Both methods of generating control predictions (see **Methods**) were compared.

629 In panels **C** and **D**, errors bars indicate s.e.m. over the 10 samplings for each number from 3 to 21; for
630 24, all available ORs were used.
631
632 **Fig. 5: Localization of the top 20 response-determining positions.**
633 Protter plots showing the localization of the top 20 response-determining residues in representative
634 ORs: OR85b of *D. melanogaster* **(A1)** and OR1 of *A. gambiae* **(A2)**. Residues in the top 20 positions in the
635 corresponding species are shown with shaded backgrounds, while the residues that are among the top
636 20 positions in both the species are shown with a black outline. One of the top 20 positions in *A.*
637 *gambiae* mapped to a gap in OR1 in the multiple sequence alignment, so only 19 positions are shaded.
638
639 **Table 1: Positions identified in previous mutagenesis studies that match the top positions identified in**
640 **our analysis.**
641 Of the 47 mutations that we collected from the mutagenesis studies, the 9 that matched the top
642 positions in *D. melanogaster* or *A. gambiae* are listed. The complete set of the 47 mutations is shown in
643 **Table S1**. TM: transmembrane helix; EL: extracellular loop.
644
645 **Supplementary Table S1: Complete multiple sequence alignment of all the ORs.**
646 The table shows the multiple sequence alignment of all the ORs that have been used in this study. Each
647 row corresponds to one OR. The first two rows indicate with colored background the positions that were
648 ranked in the two species; among these the ranks of the top 20 positions are mentioned. The 47
649 residues that have been identified as important in previous mutagenesis studies are marked with yellow
650 color in the matrix.
651
652

Table 1

Study	Mutated receptor	Mutated residue	Position in multiple sequence alignment	Rank among top positions	Topological position
Hughes et al, 2014	*A. gambiae* OR15	F369L	783	2 in *D. melanogaster* & 17 in *A. gambiae*	TM 7
Nakagawa et al 2012	*B. mori* OR1	Y109A	234	14 in *D. melanogaster*	TM 3
Nichols and Luetje, 2010	*D. melanogaster* OR85b	C146S	335	20 in *D. melanogaster*	TM 3
Ray et al, 2014	*A. gambiae* OR1	Y405H	791	6 in *A. gambiae*	TM 7
Ray et al, 2014	*A. gambiae* OR1	S406A	792	20 in *A. gambiae*	TM 7
Xu and Leal, 2013	*B. mori* OR1	P165A	335	20 in *D. melanogaster*	TM 3
Xu and Leal, 2013	*B. mori* OR1	P194A	398	9 in *A. gambiae*	EL 2
Xu and Leal, 2013	*A. gambiae* OR10	P154A	398	9 in *A. gambiae*	EL 2
Xu and Leal, 2013	*C. quinquefasciatus* OR10	P157A	398	9 in *A. gambiae*	EL 2

Figure 1

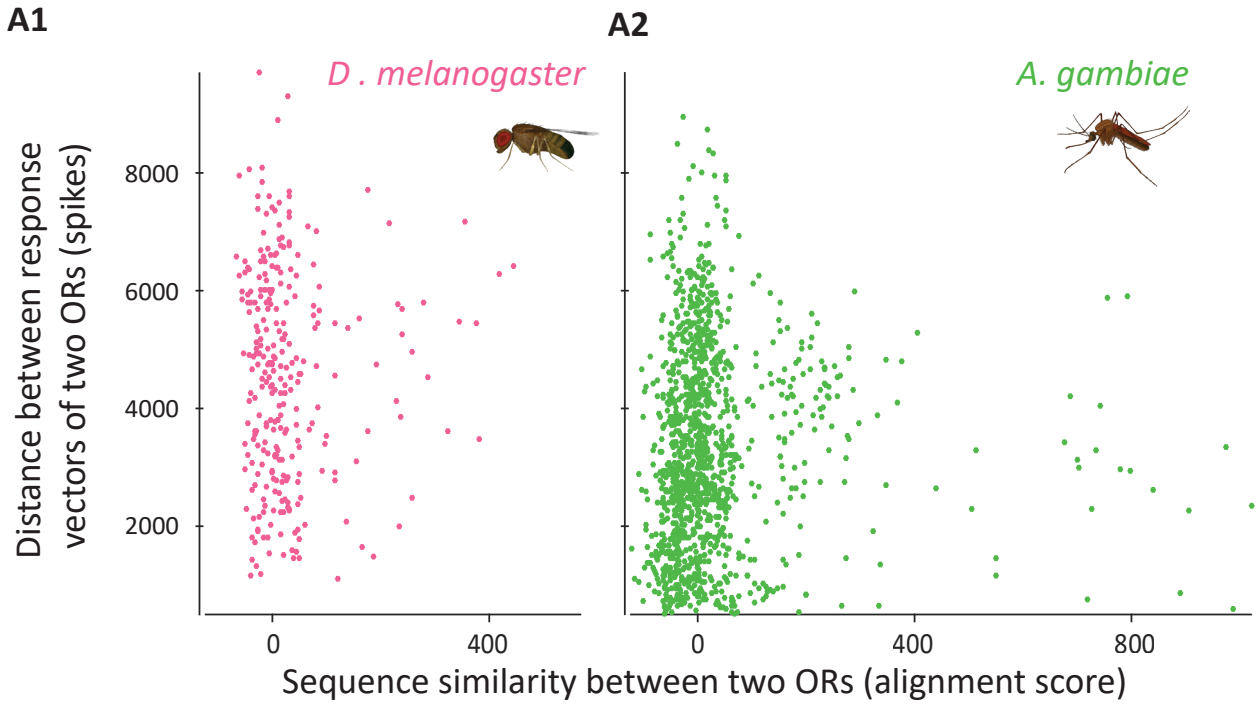

A1
A2
D . melanogaster
A. gambiae
Distance between response vectors of two ORs (spikes)
8000
6000
4000
2000
0
400
0
400
800
Sequence similarity between two ORs (alignment score)

Figure 2

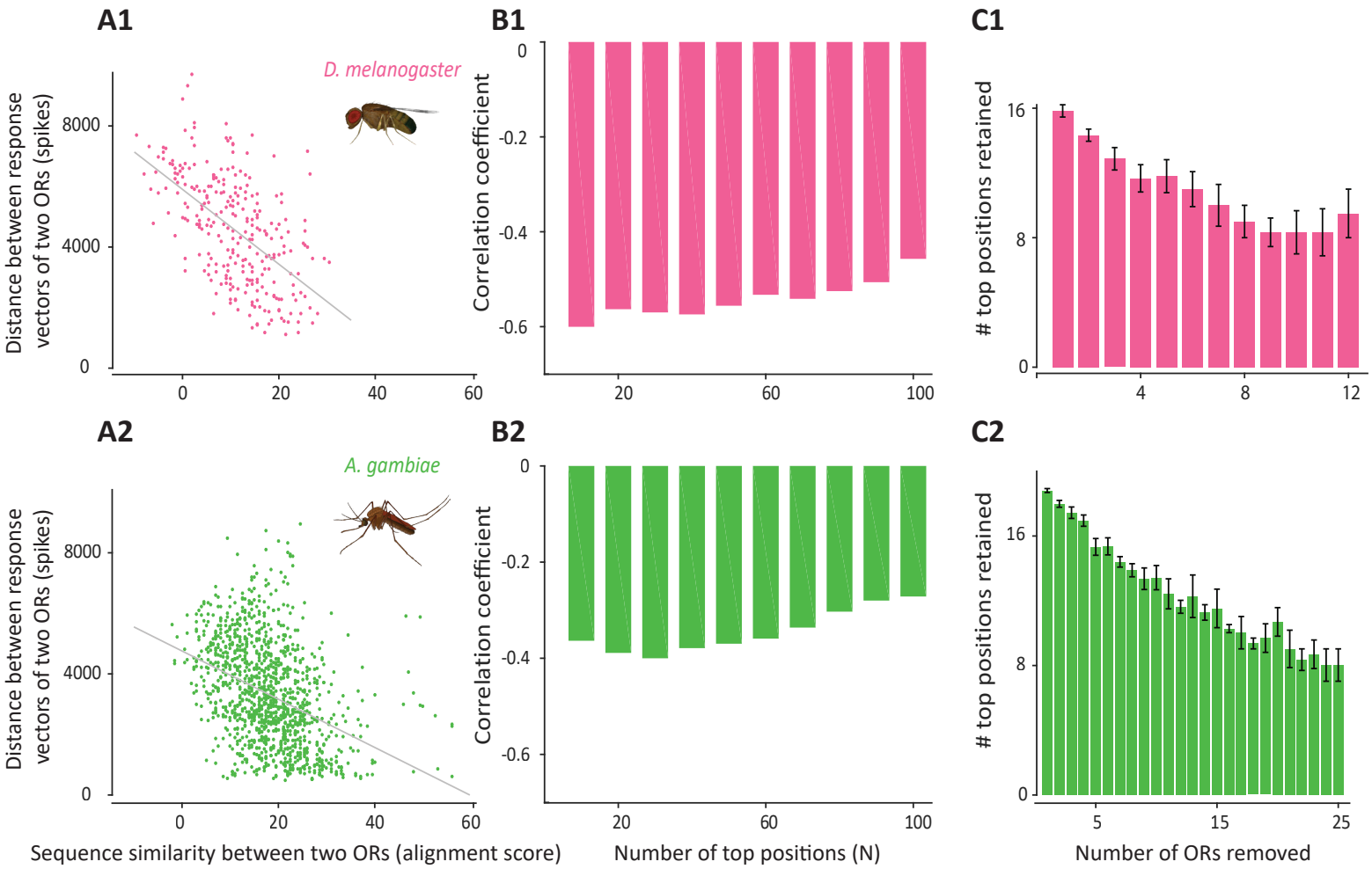

A1
D. melanogaster
Distance between response vectors of two ORs (spikes)
8000
4000
0
0
20
40
60
B1
Correlation coefficient
0
-0.2
-0.4
-0.6
20
60
100
C1
top positions retained
16
8
0
4
8
12
A2
A. gambiae
Distance between response vectors of two ORs (spikes)
8000
4000
0
0
20
40
60
Sequence similarity between two ORs (alignment score)
B2
Correlation coefficient
0
-0.2
-0.4
-0.6
20
60
100
Number of top positions (N)
C2
top positions retained
16
8
0
5
15
25
Number of ORs removed

Figure 3

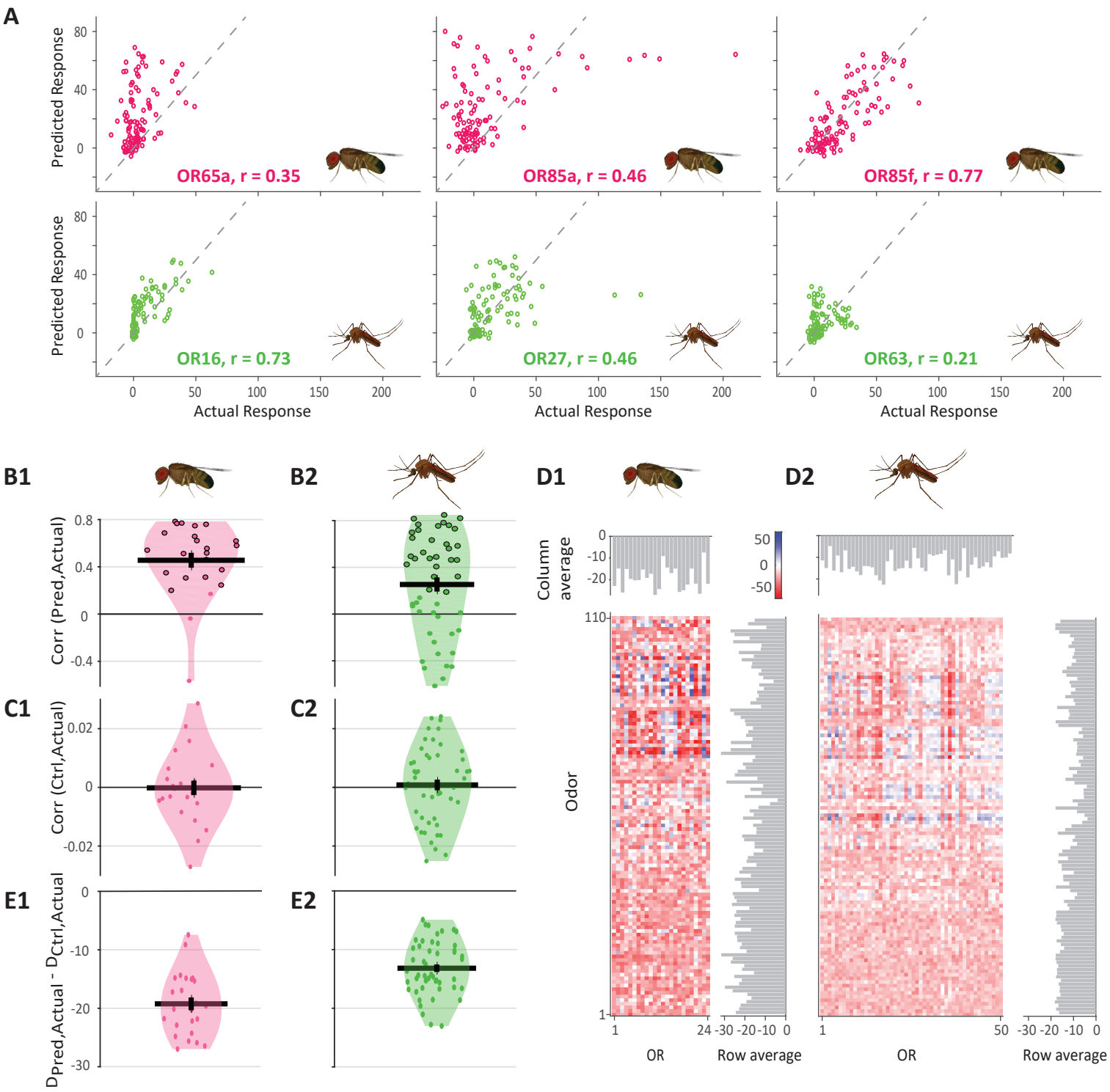

A
Predicted Response
Actual Response
OR65a, r = 0.35
OR85a, r = 0.46
OR85f, r = 0.77
OR16, r = 0.73
OR27, r = 0.46
OR63, r = 0.21
B1
B2
C1
C2
E1
E2
Corr (Pred,Actual)
Corr (Ctrl,Actual)
D1
D2
Column average
Odor
OR
Row average

Figure 4

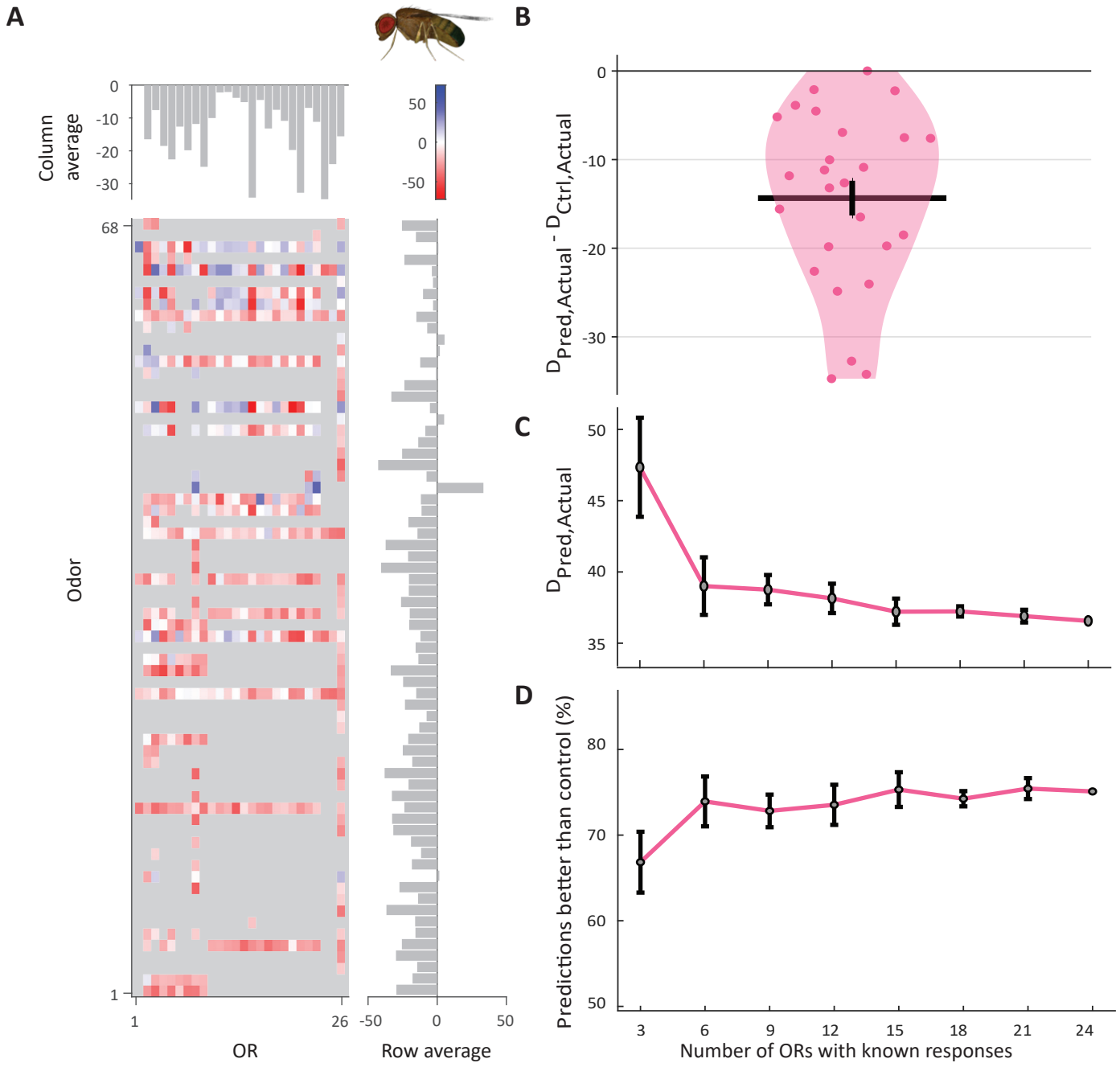
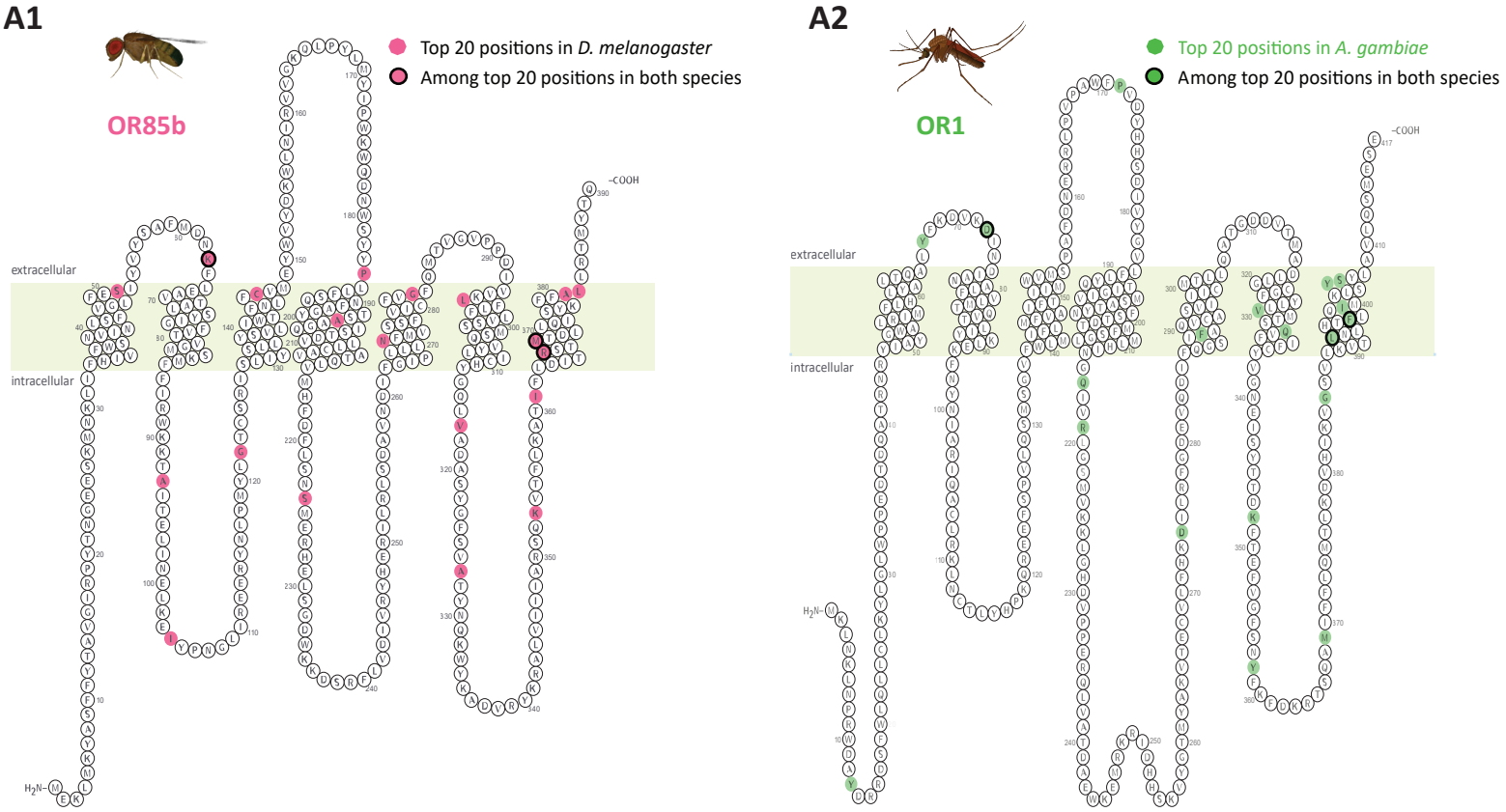
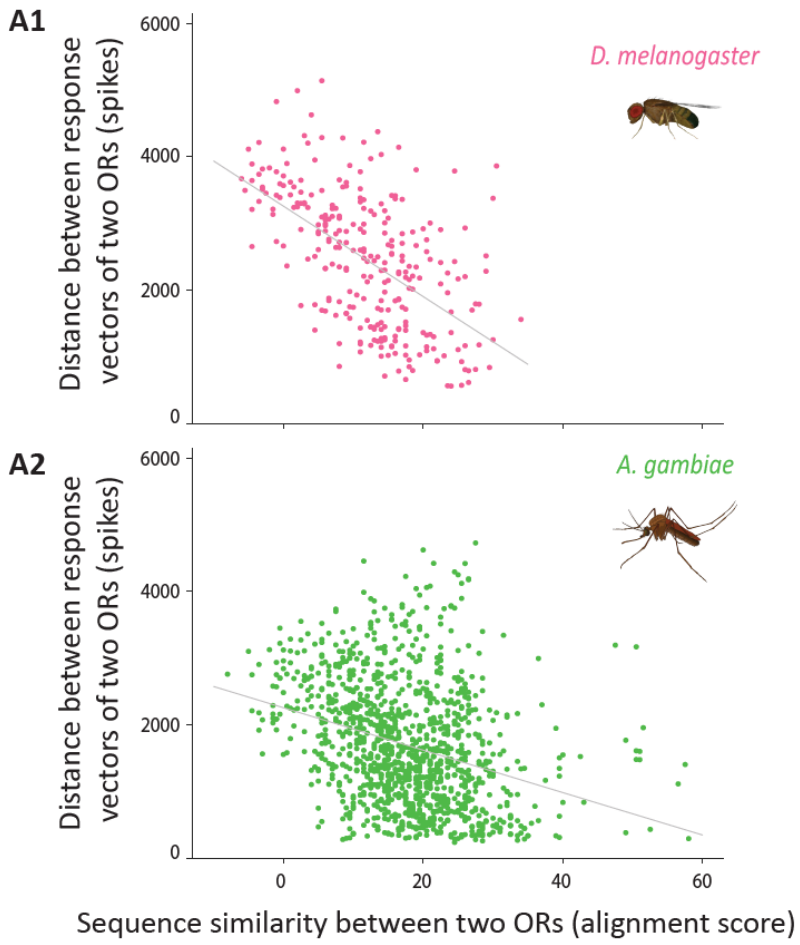

A
B
C
D
Column average
0
-10
-20
-30
50
0
-50
68
1
Odor
1
26
OR
-50
0
50
Row average
$D_{Pred,Actual} - D_{Ctrl,Actual}$
0
-10
-20
-30
$D_{Pred,Actual}$
50
45
40
35
Predictions better than control (%)
80
70
60
50
3
6
9
12
15
18
21
24
Number of ORs with known responses

Figure 5


A1
OR85b
Top 20 positions in *D. melanogaster*
Among top 20 positions in both species
extracellular
intracellular
-COOH
H2N-
A2
OR1
Top 20 positions in *A. gambiae*
Among top 20 positions in both species
extracellular
intracellular
-COOH
H2N-

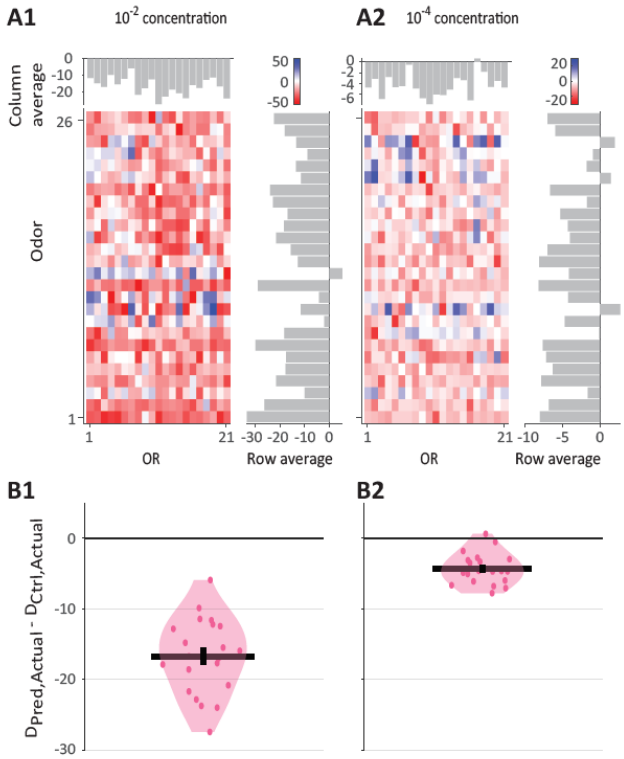
Supplementary Figure S1



Supplementary Fig. S1: Cross-validation of correlation between response similarity and sequence similarity at the top 20 positions.

Scatter plots of distances in response vectors (using half of the odors) versus the sequence similarity measured at the top 20 positions, among pairs of ORs. This figure is similar to Fig. 2A1,2, except that here only half of the odors (in odd-numbered rows of the original dataset) were used to calculate the top 20 positions, and the remaining half were used to calculate the response similarity, thus providing a cross-validation. The observed Pearson correlations are -0.56 in both *D. melanogaster* (**A1**), and -0.31 in *A. gambiae* (**A2**). Each point in the plots corresponds to a pair of conspecific ORs (n = 276 pairs in *D. melanogaster* and n = 1225 pairs in *A. gambiae*).

Supplementary Figure S2



Supplementary Fig. S2: Predictions in larval *Drosophila*.

A1, A2 Color maps showing the relative prediction error for each OR-odor pair, for 21 ORs in *D. melanogaster* larvae for the set of 26 odors, at the dilution of 10^{-2} **(A1)** or 10^{-4} (**A2**). The relative prediction error is measured as $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$, as described in **Fig. 3**.

B1, B2 Violin plots shows the average (over odors) of $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$ for each of the 21 ORs in panels **A1** and **A2,** respectively. Horizontal line and the error bar indicate the mean and the s.e.m., respectively.