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

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* and *Anopheles gambiae*, and 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 recently, but these require elaborate experimental pipelines for each species (Jiang et al., 2015; Hu and

33 Matsunami, 2018; Jones et al., 2019). The high-throughput methods also generate many false positives,
34 so it is recommended that the identified responses be further verified with targeted experiments
35 (Nishizumi and Sakano, 2015).

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

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

52

53 **Results**

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

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

63 sequence similarity and response distance, we surprisingly found no correlation at all (**Fig. 1**): the
64 Pearson's correlation coefficient was 0.0018 (P = 0.97, N = 276 pairs of receptors) in *D. melanogaster*,
65 and 0.026 (P = 0.35, N = 1225 pairs) in *A. gambiae*.

66

67 **Identification of response-determining positions**

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

80 We first confirmed that the subsequences of amino acids formed by these 20 positions in each OR were
81 indeed related to the responses. We found the correlation between the response distance and sequence
82 similarity (calculated using only the subsequences formed by the 20 positions) to be -0.56 (P= 1.7×10^{-24},
83 N = 276 pairs; **Fig. 2A1**) in *D. melanogaster*, and -0.39 (P = 1.7×10^{-45}, N = 1225 pairs; **Fig. 2A2**) in *A.*
84 *gambiae*. Compared to the result seen with the full sequence (**Fig. 1**), these correlations were
85 significantly negative. If we varied the number of top ranked positions used in the analysis, the
86 correlation coefficients did not change for numbers between 10 and 40 but gradually reduced for larger
87 numbers of positions (**Figs. 2B1,2**), suggesting that lower-ranked residues are not important for the
88 prediction of responses. Subsequently, we used the top 20 positions as the response-determining
89 positions in our analysis (**Figs. 2B1,2** shows that our results would not change if we chose some other
90 number between 50 % and 200 % of this number).

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

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

104

105 **Predicting responses using the identified positions**

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

112 We first applied this approach to predict the responses of the 24 and the 50 ORs in the *D. melanogaster*
113 and *A. gambiae* datasets, respectively. By turn, each OR was treated as a query OR and its response was
114 predicted using the responses of the remaining ORs in the same species (excluding the query OR). We
115 found that the predicted response profiles (i.e., the 110-length vectors of odor responses) of the ORs
116 were correlated with their actual response profiles (**Figs. 3A,B**); the correlations were positive and
117 statistically significant in 21 of the 24 ORs in *D. melanogaster*, and 29 of the 50 ORs in *A. gambiae*. Mean
118 of these correlation values was significantly greater than zero: 0.46 ($P = 2.53 \times 10^{-7}$, t-test; $P = 1.63 \times 10^{-5}$,
119 sign-rank test; N = 24 ORs) in *D. melanogaster,* and 0.25 ($P = 1.18 \times 10^{-4}$, t-test; $P = 1.19 \times 10^{-4}$, sign-rank
120 test; N = 50 ORs) in *A. gambiae*. In contrast, control predictions obtained by shuffling the response

121 matrices (see **Methods**) were not correlated with the actual responses (**Fig. 3C**): mean correlation values
122 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
123 *A. gambiae*.

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

143

144 **Test on an independent dataset**

145 Next, we collected the responses of 26 other ORs in *D. melanogaster* from 10 different studies (de
146 Bruyne et al., 1999, 2001, 2010; Dobritsa et al., 2003; Goldman et al., 2005; Kreher et al., 2005, 2008;
147 Marshall et al., 2010; Ronderos et al., 2014; Dweck et al., 2015) deposited in the DoOR database (Münch
148 and Galizia, 2016). Each of these ORs had responses available for some of the 110 odors used previously;
149 overall, 68 of the previously used odors were represented in the new dataset. These 26 ORs were
150 entirely different from the 24 ORs used previously in selecting the response-determining positions, and

151 therefore provided an independent test for our approach. Using the set of top 20 response-determining
152 positions as selected earlier from the original *D. melanogaster* dataset, we found that the responses
153 predicted for the 26 novel ORs were significantly better with our approach compared to the control
154 predictions (**Fig. 4A,B**): the average $D_{Pred,Actual}$ of an OR (37.01) was smaller than average $D_{Ctrl,Actual}$
155 (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
156 improvement of about 28 %. Overall, among the 506 OR-odor responses predicted (non-grey squares in
157 **Fig. 4A**), in 380 (75.1%) cases the predictions were better than the control predictions. These results
158 show that once the response-determining positions have been identified using known responses of
159 some ORs in a species, the responses of other conspecific ORs can be predicted reasonably using only
160 their sequences.

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

171

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

173 Our results show that the identified subset of sequence positions can allow reliable prediction of the
174 neural responses for new ORs. To check where these top 20 positions are located within the protein
175 structure, we visualized these positions in the secondary structure of two representative ORs in *D.*
176 *melanogaster* and *A. gambiae* using the Protter tool (Omasits et al., 2014). These residues were not
177 limited to any one region, but were instead found in various parts of the protein (**Fig. 5**). It is important
178 to note that our analysis does not imply a causal or a mechanistic role for each of these residues in
179 interacting with the odor molecules, but only suggests that these residues may be involved in some
180 aspect of determination of odor responses, perhaps indirectly. Notably, the region with the highest

density of the top 20 positions was the seventh transmembrane helix, including 4 positions in *D. melanogaster* and 5 in *A. gambiae* (**Fig. 5**; similar numbers were seen in other receptors). Moreover, of the three response-determining positions that were conserved between the two species, two were located in the seventh transmembrane helix. These observations provide independent support to the idea that the seventh transmembrane helix plays an important role in determining odor specificity (Hughes et al., 2014; Ray et al., 2014).

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

Given 20 response-determining positions in *D. melanogaster* and *A. gambiae* each, and with 3 being common between the two species, 37 unique positions in the multiple sequence alignment were marked as the top positions out of a total of 817. Thus, each of the 47 mutated positions in the curated experimental set had a 37/817 probability of matching a top position, and the expected number of matches by chance = 47 x 37/817 = 2.13. So, while only about 2 of the 47 mutated positions were expected to be among the top positions by chance, we found as many as 9 of them to be among the top positions ($P = 2.28 \times 10^{-4}$; binomial test), providing experimental support to our response-determining positions. These 9 experimentally studied mutations, which map to the top positions in our alignment, are listed in **Table 1** (further, **Table S1** shows all positions). Among these, residue 146 in OR85b of *D. melanogaster* is one of the residues expected to play a crucial role in the activation of the odorant (Nichols and Luetje, 2010). Residue 109 mutant in OR1 of *B. mori* was found to change the reversal potential and rectification index of the BmOR1-Orco complex (Nakagawa et al., 2012). Point mutation of residues 165 and 194 in BmorOR1 of *B. mori* led to significant reduction in response to bombykol pheromone (Xu and Leal, 2013). Similarly, odor responses in *A. gambiae* were found to change with mutation of residue 369 in OR15 (Hughes et al., 2014), and residues 405 and 406 in OR1 (Ray et al., 2014). Among the remaining 38 out of the 47 mutations in the curated set, 10 and 8 were found to be within a distance of 1 and 2 residues, respectively, from a top column. Thus, in total, 27 of the 47

212 experimentally confirmed important positions matched or were adjacent to our computationally
213 predicted response-determining positions.

214

215 **Discussion**

216 Sequence-based approaches to response prediction require a reasonably high homology among the
217 receptors. In vertebrates, receptors with less than 40 % sequence identity were found to have little
218 overlap in their responses (Li et al., 2015). The insects ORs presented a particularly difficult test case, as
219 they have only about 20 % homology among conspecific ORs; indeed, we found that there was no
220 correlation between the overall sequence similarity and the response similarity among the ORs. We
221 were able to overcome this difficulty by focusing on subsets of positions that may be important for
222 determining odor responses. Our approach was successful in predicting the responses for not only the
223 set of fly and mosquito receptors whose data was used while developing the method, but also for a
224 completely novel set of receptors whose responses were taken from 10 independent experimental
225 datasets. Further, we found that the predictions require only about 15 ORs with known responses (**Fig.**
226 **4**). This computational approach can be particularly useful for receptors expressed in neurons that are
227 difficult to access in electrophysiology experiments, or for receptors that do not express well in
228 heterologous expression systems (Ronderos et al., 2014).

229 Our method allows prediction of responses of novel ORs for already studied odors. Previously developed
230 methods allow prediction of responses of already studied ORs to novel odors (Schmuker et al., 2007;
231 Haddad et al., 2008; Boyle et al., 2013; Gabler et al., 2013; Saberi and Seyed-Allaei, 2016; Bushdid et al.,
232 2018; Kepchia et al., 2019). These two approaches are complementary and can be combined in future
233 work to enable the prediction of the responses of novel ORs to novel odors, which would help with
234 large-scale deorphanization. While applying these approaches, care must be taken regarding the
235 concentrations of the odors in the training data, as changes in odor concentration can affect neural
236 responses (Hallem and Carlson, 2006; Olsen et al., 2010) and behavior (Wright et al., 2005; Yarali et al.,
237 2009). Another caveat is that these approaches are currently limited to predicting the response
238 magnitudes, and ignore the temporal patterns of spikes observed in the olfactory receptor neurons
239 (Spors et al., 2006; Raman et al., 2010; Su et al., 2011; Grillet et al., 2016; Egea-Weiss et al., 2018).

240 Unlike vertebrate ORs which are G-protein coupled receptors, insect ORs are heteromeric ligand-gated
241 ion channels (Wicher et al., 2008), composed of a specific OR and a conserved co-receptor Orco (Larsson

242 et al., 2004). The structure of any insect receptor was not known until very recently when the structure
243 of Orco from fig wasp was studied using Cryo-EM (Butterwick et al., 2018). The approach of predicting
244 responses using the three-dimensional modeling of ORs, that has seen some success in vertebrates
245 (Bavan et al., 2014), is unlikely to be very productive in insects in the near future given the very low
246 sequence similarity among the ORs and with other proteins, and the paucity of available structures.
247 Therefore sequence-based computational methods will be important for receptors whose structures
248 and responses have not been determined experimentally. Moreover, the computational predictions can
249 be used to identify the candidate ligands to be tested experimentally. The method can also be extended
250 to other types of chemosensory receptors, such as the ionotropic receptors and the gustatory receptors.

251 We have used a simple algorithm in our approach, where we first identify the important positions in the
252 sequences and then use the similarity at those positions to make the predictions. It may be possible to
253 achieve good predictions without explicitly identifying the important positions but using machine
254 learning-based approach that uses all the residues and learns a non-linear mapping between the
255 sequences and the responses. However, we preferred the simpler deterministic approach over the
256 blackbox approaches typically used in machine learning because the deterministic approach provides a
257 transparent rationale for and information about the sequence positions that are used for making the
258 predictions. Expectedly, a majority of the identified positions were located in the extracellular loops or
259 transmembrane helices (particularly the seventh transmembrane, which has been shown to be
260 important for determining responses), but some were also found in intracellular loops. A previous study
261 has shown that mutation of intracellular residues can affect the electrophysiological responses of
262 neurons (Nakagawa et al., 2012). We emphasize that our approach only looks for correlations between
263 the residues and responses, and it is likely that some of the identified positions, particularly those found
264 in intracellular regions, affect the odor responses indirectly rather than through direct interactions with
265 the odor molecules. By comparing the 47 positions previously found to be important in mutagenesis
266 studies in various organisms (Nichols and Luetje, 2010; Nakagawa et al., 2012; Xu and Leal, 2013;
267 Hughes et al., 2014; Ray et al., 2014), we found that 9 of them (a statistically significant fraction, more
268 than 4 times larger than that expected by chance) mapped to the top positions identified in our analysis
269 (**Table 1**). This shows that our method is using reliable positions for making the predictions. The other
270 top positions that have not been experimentally studied so far are good candidates for future
271 mutagenesis studies, and could help in understanding the mechanisms of odor-OR interactions. As more
272 experimental evidence accumulates, the response predictions can be further improved by giving more
273 weightage to the experimentally verified positions among the top columns.

274 **Methods**

275 **Response datasets**

276 Responses of 24 *D. melanogaster* ORs were obtained from a previous study (Hallem and Carlson, 2006)
277 and responses of 50 *A. gambiae* ORs were obtained from another study (Carey et al., 2010), for a set of
278 110 odors. The response values indicate the odor-elicited spiking response of a neuron expressing the
279 receptor, minus the response for the diluent or the spontaneous firing rate. To serve as an independent
280 dataset, responses of 26 other *D. melanogaster* ORs were obtained from 10 studies (de Bruyne et al.,
281 1999, 2001, 2010; Dobritsa et al., 2003; Goldman et al., 2005; Kreher et al., 2005, 2008; Marshall et al.,
282 2010; Ronderos et al., 2014; Dweck et al., 2015) deposited in the Database of Odorant Responses
283 (DoOR) (Münch and Galizia, 2016), for 68 odors that were common with the earlier set of 110 odors. If
284 the response of a particular OR to an odor was reported in multiple studies, we used the average of
285 those values as the actual response.

286 **Sequence analysis**

287 The olfactory receptor sequences for *D. melanogaster* and *A. gambiae* were obtained from the Database
288 of Olfactory Receptors (Nagarathnam et al., 2014) and the (Hill et al., 2002) study, respectively. The
289 sequence similarity between any two receptor sequences was determined using the Needleman
290 Wunsch Algorithm using the BLOSUM62 scoring matrix (Needleman and Wunsch, 1970; Henikoff and
291 Henikoff, 1992). Multiple sequence alignment was performed for a total of 140 (62 in *D. melanogaster*
292 and 78 in *A. gambiae*) olfactory receptor sequences using MAFFT multiple sequence alignment program,
293 and additional sequences from other species (for validation) were added individually using the '-add'
294 option (Katoh and Frith, 2012).

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

298 **Ranking the response-determining positions**

299 We identified the response-determining sequence positions for *D. melanogaster* and *A. gambiae* using
300 the following approach, which quantifies at each position how well the differences in amino acids across
301 the receptors correlate with the differences in their responses. These positions corresponded to
302 columns in the multiple sequence alignment, and were same for all the receptors within a species

303 (allowing different positions for each receptor can provide better correlations for individual receptors
304 due to overfitting but will have poor generalizability for other receptors).

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

315 $$SS_{(r_1,r_2),c} = BLOSUM62\left(A_{r_1,c}, A_{r_2,c}\right) \forall\ (r_1, r_2)\ where\ A_{r_1,c} \neq \text{ gap } \&\ A_{r_2,c} \neq \text{ gap}$$

316 Similarly, we computed a response distance vector (DR_c^o) for the same pairs of ORs in the column, each
317 element ($DR^o_{(r_1,r_2),c}$) of which contained the absolute difference between the responses of receptors
318 $r_1\ and\ r_2$ to an odor o:

319 $$DR^o_{(r1,r2),c} = \left|N^o_{r_1} - N^o_{r_2}\right| \forall\ (r_1, r_2)\ where\ A_{r_1,c} \neq \text{ gap } \&\ A_{r_2,c} \neq \text{ gap},$$

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

328 $$S_c = -\sum_{\substack{o=1,\\ R_c^o<0}}^{110} \log P_c^o$$

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

331 **Response prediction**

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

338 $$\overline{SSS}_{r,query} = \frac{SSS_{r,query} - \min_i SSS_{r_i,query}}{\max_i SSS_{r_i,query} - \min_i SSS_{r_i,query}} \times 100$$

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

341 $$\widetilde{N}^o_{query} = \frac{\sum_{j=1}^{n} (\overline{SSS}_{r_j,query} \times N^o_{r_j})}{\sum_{j=1}^{n} \overline{SSS}_{r_j,query}}$$

342 where n is the number of receptors with known responses and $N^o_{r_j}$ denotes the neural response of
343 receptor r_j to odor o.

344 **Control Predictions**

345 Control predictions were obtained by shuffling the OR-odor response matrix and taking the shuffled
346 response as the control prediction for each OR-odor combination (this operation allowed generation of
347 random predictions while maintaining the overall statistical properties of the response values). The
348 shuffling was performed 50 times independently, and the results (correlations or $D_{Ctrl,Actual}$ values)
349 were averaged over the 50 shufflings.

350 **Statistical analysis and code availability**

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

356

357 **Acknowledgements**

358 This work was supported by the Wellcome Trust/DBT India Alliance Fellowship (grant number
359 IA/I/15/2/502091) awarded to N.G., and ISF-UGC joint research program in which R.H. was supported by
360 the ISF (grant No. 2307/15) and N.G. was supported by the UGC (F.No. 6-11/2016[IC]). We also thank
361 members of the N.G. laboratory for helpful comments.

362 **Declaration of interests**

363 The authors declare no competing financial interests.

364 **Author Contributions**

365 N.G. conceived the study. S.C., A.G., R.H., and N.G. developed the methodology; S.C. and A.G. developed
366 the software and performed analysis; N.G. and R.H. secured funding; S.C., A.G., R.H., and N.G. wrote the
367 manuscript.

368

369 **References**

370 Attardo GM et al. (2019) The Glossina Genome Cluster: Comparative Genomic Analysis of the Vectors of African
371 Trypanosomes. bioRxiv.
372 Bavan S, Sherman B, Luetje CW, Abaffy T (2014) Discovery of Novel Ligands for Mouse Olfactory Receptor MOR42-
373 3 Using an In Silico Screening Approach and In Vitro Validation. PLoS One 9.
374 Boyle SM, McInally S, Ray A (2013) Expanding the olfactory code by in silico decoding of odor-receptor chemical
375 space. Elife:1–17.
376 Bushdid C, De March CA, Fiorucci S, Matsunami H, Golebiowski J (2018) Agonists of G-Protein-Coupled Odorant
377 Receptors Are Predicted from Chemical Features. J Phys Chem Lett 9:2235–2240.
378 Butterwick JA, del Mármol J, Kim KH, Kahlson MA, Rogow JA, Walz T, Ruta V, Mármol J, Kim KH, Kahlson MA,
379 Rogow JA, Walz T, Ruta V (2018) Cryo-EM structure of the insect olfactory receptor Orco. Nature 560:447–
380 452.

381 Carey AF, Wang G, Su C-Y, Zwiebel LJ, Carlson JR (2010) Odorant reception in the malaria mosquito Anopheles
382 gambiae. Nature 464:66–71.
383 Chahda JS, Soni N, Sun JS, Ebrahim SAM, Weiss L, Carlson JR (2019) The molecular and cellular basis of olfactory
384 response to tsetse fly attractants. PLOS Genet:1–22.
385 de Bruyne M, Clyne PJ, Carlson JR (1999) Odor coding in a model olfactory organ: the Drosophila maxillary palp. J
386 Neurosci 19:4520–4532.
387 de Bruyne M, Foster K, Carlson JR (2001) Odor coding in the Drosophila antenna. Neuron 30:537–552.
388 de Bruyne M, Smart R, Zammit E, Warr CG (2010) Functional and molecular evolution of olfactory neurons and
389 receptors for aliphatic esters across the Drosophila genus. J Comp Physiol A 196:97–109.
390 de Fouchier A, Walker WB, Montagne N, Steiner C, Binyameen M, Schlyter F, Chertemps T, Maria A, Francois M-C,
391 Monsempes C, Anderson P, Hansson BS, Larsson MC, Jacquin-Joly E (2017) Functional evolution of
392 Lepidoptera olfactory receptors revealed by deorphanization of a moth repertoire. Nat Commun 8:1–11.
393 Dobritsa AA, Naters WVDG Van, Warr CG, Steinbrecht RA, Carlson JR, Haven N, Vic C (2003) Integrating the
394 Molecular and Cellular Basis of Odor Coding in the Drosophila Antenna. Neuron 37:827–841.
395 Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature
396 450:203–218.
397 Dweck HKM, Ebrahim SAM, Farhan A, Hansson BS, Stensmyr MC (2015) Olfactory Proxy Detection of Dietary
398 Antioxidants in Drosophila. Curr Biol 25:455–466.
399 Egea-Weiss A, Renner A, Kleineidam CJ, Szyszka P (2018) High Precision of Spike Timing across Olfactory Receptor
400 Neurons Allows Rapid Odor Coding in Drosophila. ISCIENCE 4:76–83.
401 Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE (2015) Hymenoptera Genome Database :
402 integrating genome annotations in HymenopteraMine. Nucleic Acids Res 44:793–800.
403 Gabler S, Soelter J, Hussain T, Sachse S, Schmuker M (2013) Physicochemical vs. vibrational descriptors for
404 prediction of odor receptor responses. Mol Inform 32:855–865.
405 Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG (2008) Drosophila chemoreceptor gene evolution :
406 selection , specialization and genome size. Mol Ecol 17:1648–1657.
407 Goldman AL, Van Der Goes Van Naters W, Lessing D, Warr CG, Carlson JR (2005) Coexpression of two functional
408 odor receptors in one neuron. Neuron 45:661–666.
409 Grillet M, Campagner D, Petersen R, Mccrohan C, Cobb M (2016) The peripheral olfactory code in Drosophila
410 larvae contains temporal information and is robust over multiple timescales. Proc R Soc B 283:1–8.
411 Haddad R, Carmel L, Sobel N, Harel D (2008) Predicting the Receptive Range of Olfactory Receptors. PLoS Comput
412 Biol 4:0001–0008.
413 Hallem EA, Carlson JR (2006) Coding of Odors by a Receptor Repertoire. Cell 125:143–160.
414 Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–
415 10919.

416 Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ (2002) G
417 protein-coupled receptors in Anopheles gambiae. Science (80-) 298:176–178.
418 Hu XS, Matsunami H (2018) High-Throughput Odorant Receptor Deorphanization Via Phospho-S6 Ribosomal
419 Protein Immunoprecipitation and mRNA Profiling BT - Olfactory Receptors: Methods and Protocols. In
420 (Simoes de Souza FM, Antunes G, eds), pp 95–112. New York, NY: Springer New York.
421 Hughes DT, Wang G, Zwiebel LJ, Luetje CW (2014) A determinant of odorant specificity is located at the
422 extracellular loop 2-transmembrane domain 4 interface of an anopheles gambiae odorant receptor subunit.
423 Chem Senses 39:761–769.
424 Jiang Y, Gong NN, Hu XS, Ni MJ, Pasi R, Matsunami H (2015) Molecular profiling of activated olfactory neurons
425 identifies odorant receptors for odors in vivo. Nat Neurosci 18:1446–1454.
426 Jones EM, Jajoo R, Cancilla D, Lubock NB, Wang J, Satyadi M, Cheung R, de March C, Bloom JS, Matsunami H, Kosuri
427 S (2019) A Scalable, Multiplexed Assay for Decoding GPCR-Ligand Interactions with RNA Sequencing. Cell Syst
428 8:254–260.e6.
429 Karpe SD, Dhingra S, Brockmann A, Sowdhamini R (2017) Computational genome-wide survey of odorant receptors
430 from two solitary bees Dufourea novaeangliae (Hymenoptera : Halictidae) and Habropoda laboriosa (
431 Hymenoptera : Apidae). Sci Rep 7:1–13.
432 Katoh K, Frith MC (2012) Adding unaligned sequences into an existing alignment using MAFFT and LAST.
433 Bioinformatics 28:3144–3146.
434 Kepchia D, Xu P, Terryn R, Castro A, Schürer SC, Leal WS, Luetje CW (2019) Use of machine learning to identify
435 novel , behaviorally active antagonists of the insect odorant receptor co-receptor (Orco) subunit. Sci Rep
436 9:1–18.
437 Kreher SA, Kwon JY, Carlson JR (2005) The Molecular Basis of Odor Coding in the Drosophila Larva. Neuron 46:445–
438 456.
439 Kreher SA, Mathew D, Kim J, Carlson JR (2008) Translation of Sensory Input into Behavioral Output via an Olfactory
440 System. Neuron 59:110–124.
441 Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB (2004) Or83b encodes a broadly
442 expressed odorant receptor essential for Drosophila olfaction. Neuron 43:703–714.
443 Leal WS, Choo Y, Xu P, Silva CSB, Ueira-vieira C (2013) Differential expression of olfactory genes in the southern
444 house mosquito and insights into unique odorant receptor gene isoforms. Proc Natl Acad Sci U S A
445 110:18704–18709.
446 Li H, You Y, Zhang L (2018) Single Sensillum Recordings for Locust Palp Sensilla Basiconica. JoVE:e57863.
447 Li J, Haddad R, Santos V, Bavan S, Luetje CW (2015) Receptive range analysis of a mouse odorant receptor
448 subfamily. J Neurochem 134:47–55.
449 Lin C, Potter CJ (2015) Re-Classification of Drosophila melanogaster Trichoid and Intermediate Sensilla Using
450 Fluorescence-Guided Single Sensillum Recording. PLoS One:1–14.

451 Lombardo F, Salvemini M, Fiorillo C, Nolan T, Zwiebel LJ, Ribeiro JM, Arcà B (2017) Deciphering the olfactory
452 repertoire of the tiger mosquito Aedes albopictus. BMC Genomics:1–23.
453 Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon H-W, Pitts RJ, van Loon JJA, Takken W, Carlson JR, Zwiebel LJ
454 (2007) Odor Coding in the Maxillary Palp of the Malaria Vector Mosquito Anopheles gambiae. Curr Biol
455 17:1533–1544.
456 Mainland JD, Li YR, Zhou T, Liu WL in. L, Matsunami H (2015) Human olfactory receptor responses to odorants. Sci
457 data 2:1–9.
458 Malnic B, Hirono J, Sato T, Buck LB, Hughes H (1999) Combinatorial Receptor Codes for Odors. Cell 96:713–723.
459 Marshall B, Warr CG, Bruyne M De (2010) Detection of Volatile Indicators of Illicit Substances by the Olfactory
460 Receptors of Drosophila melanogaster. Chem Senses 35:613–625.
461 Morel P (2018) Gramm : grammar of graphics plotting in Matlab. J open source Softw 3:1–4.
462 Münch D, Galizia CG (2016) DoOR 2 . 0 - Comprehensive mapping of Drosophila melanogaster odorant responses.
463 Sci Rep 6:1–9.
464 Nagarathnam B, Karpe SD, Harini K, Sankar K, Iftekhar M, Rajesh D, Giji S, Archunan G, Balakrishnan V, Gromiha
465 MM, Nemoto W, Fukui K, Sowdhamini R (2014) DOR - a Database of Olfactory Receptors - Integrated
466 Repository for Sequence and Secondary Structural Information of Olfactory Receptors in Selected Eukaryotic
467 Genomes. Bioinform Biol Insights 8:147–158.
468 Nakagawa T, Pellegrino M, Sato K, Vosshall LB, Touhara K (2012) Amino acid residues contributing to function of
469 the heteromeric insect olfactory receptor complex. PLoS One 7:1–9.
470 Neafsey DE, Waterhouse RM et al. (2015) Highly evolvable malaria vectors: The genomes of 16 Anopheles
471 mosquitoes. Science (80-) 347:1–8.
472 Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid
473 sequence of two proteins. J Mol Biol 48:443–453.
474 Nichols AS, Luetje CW (2010) Transmembrane segment 3 of Drosophila melanogaster odorant receptor subunit
475 85b contributes to ligand-receptor interactions. J Biol Chem 285:11854–11862.
476 Nishizumi H, Sakano H (2015) Decoding and deorphanizing an olfactory map. Nat Neurosci 18:1432–1433.
477 Oka Y, Katada S, Omura M, Suwa M, Yoshihara Y, Touhara K (2006) Odorant Receptor Map in the Mouse Olfactory
478 Bulb: In Vivo Sensitivity and Specificity of Receptor-Defined Glomeruli. Neuron 52:857–869.
479 Olsen SR, Bhandawat V, Wilson RI (2010) Divisive Normalization in Olfactory Population Codes. Neuron 66:287–
480 299.
481 Olsson SB, Hansson BS (2013) Electroantennogram and Single Sensillum Recording in Insect Antennae. In:
482 Pheromone Signaling: Methods and Protocols, pp 157–177. Totowa, NJ: Humana Press.
483 Omasits U, Ahrens CH, Müller S, Wollscheid B (2014) Protter: Interactive protein feature visualization and
484 integration with experimental proteomic data. Bioinformatics 30:884–886.
485 Raman B, Joseph J, Tang J, Stopfer M (2010) Temporally Diverse Firing Patterns in Olfactory Receptor Neurons

486 Underlie Spatiotemporal Neural Codes for Odors. J Neurosci 30:1994–2006.
487 Ramasamy S, Ometto L, Crava CM, Revadi S, Kaur R, Horner DS, Pisani D, Dekker T, Anfora G, Rota-stabelli O (2016)
488 The Evolution of Olfactory Gene Families in Drosophila and. Genome Biol Evol 8:2297–2311.
489 Ray A, van Naters WG oe., Carlson JR (2014) Molecular determinants of odorant receptor function in insects. J
490 Biosci 39:555–563.
491 Ronderos DS, Lin C-C, Potter CJ, Smith DP (2014) Farnesol-Detecting Olfactory Neurons in Drosophila. J Neurosci
492 34:3959–3968.
493 Saberi M, Seyed-Allaei H (2016) Odorant receptors of Drosophila are sensitive to the molecular volume of
494 odorants. Sci Rep 6:1–11.
495 Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD (2009) Odor coding by a Mammalian receptor repertoire. Sci
496 Signal 2:1–14.
497 Schmuker M, De Bruyne M, Hähnel M, Schneider G (2007) Predicting olfactory receptor neuron responses from
498 odorant structure. Chem Cent J 1:1–10.
499 Spors H, Wachowiak M, Cohen LB, Friedrich RW (2006) Temporal Dynamics and Latency Patterns of Receptor
500 Neuron Input to the Olfactory Bulb. J Neurosci 26:1247–1259.
501 Su C, Martelli C, Emonet T, Carlson JR (2011) Temporal coding of odor mixtures in an olfactory receptor neuron.
502 Proc Natl Acad Sci 108:1–6.
503 Touhara K, Sengoku S, Inaki K, Tsuboi A, Hirono J, Sato T, Salano H, Haga T (1999) Functional identification and
504 reconstitution of an odorant. Proc Natl Acad Sci USA 96:4040–4045.
505 Viklund H, Elofsson A (2008) OCTOPUS: Improving topology prediction by two-track ANN-based preference scores
506 and an extended topological grammar. Bioinformatics 24:1662–1668.
507 Wang G, Carey AF, Carlson JR, Zwiebel LJ (2010) Molecular basis of odor coding in the malaria vector mosquito
508 Anopheles gambiae. Proc Natl Acad Sci 107:4418–4423.
509 Wicher D, Schäfer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, Hansson BS (2008) Drosophila odorant
510 receptors are both ligand-gated and cyclic-nucleotide- activated cation channels. Nature 452:1007–1011.
511 Wright GA, Thomson MGA, Smith BH (2005) Odour concentration affects odour identity in honeybees. Proc R Soc B
512 272:2417–2422.
513 Xu P, Leal WS (2013) Probing insect odorant receptors with their cognate ligands: Insights into structural features.
514 Biochem Biophys Res Commun 435:477–482.
515 Yarali A, Ehser S, Hapil FZ, Huang J, Gerber B (2009) Odour intensity learning in fruit flies. Proc R Soc B 276:3413–
516 3420.
517

518 **Legends**

519 **Fig. 1: Lack of correlation between response distance and sequence similarity among pairs of ORs.**
520 Scatter plots of distances in response vectors versus the sequence similarity, among pairs of ORs, reveal
521 the lack of correlation, in both *D. melanogaster* (**A1**) and *A. gambiae* (**A2**). Distance is calculated as the
522 L1-distance between the two 110-length response vectors containing number of spikes for each of the
523 110 odors. Sequence similarity is measured as the Needleman-Wunsch alignment score. Each point
524 corresponds to a pair of conspecific ORs (n = 276 pairs in *D. melanogaster* and n = 1225 pairs in *A.*
525 *gambiae*).

526

527 **Fig. 2: Properties of the predicted top 20 response determining positions.**
528 **A1**, **A2**, Scatter plots of distances in 110-odor response vectors versus the sequence similarity measured
529 at the top 20 positions, among pairs of ORs, show significant negative correlation, in both *D.*
530 *melanogaster* (**A1**) and *A. gambiae* (**A2**). Each point corresponds to a pair of conspecific ORs (n = 276
531 pairs in *D. melanogaster* and n = 1225 pairs in *A. gambiae*).
532 **B1**, **B2**, Bars indicate the negative correlation between the distance in response vectors and the
533 sequence similarity measured using the top N positions, plotted as a function of N in the range of 10 to
534 100. Note that the correlation values are relatively stable in both the species in the range of 10-40 and
535 decrease gradually in magnitude for larger values of N.
536 **C1**, **C2**, Bars indicate the number of top 20 positions in *D. melanogaster* (**C1**) or *A. gambiae* (**C2**) retained
537 among the top 20 positions when the positions were recalculated after removing different numbers of
538 receptors from the dataset. For each number, all possible combinations of that many receptors were
539 removed and the average number of columns preserved among the top 20 is reported. Error bars
540 represent s.e.m. over all combinations for each number.

541

542 **Fig. 3: Reliability of responses predicted using the top 20 positions.**
543 **A,** Scatter plots of the actual and the predicted odor responses for three representative ORs each from
544 *D. melanogaster* and *A. gambiae*. The correlation coefficients between the two responses are indicated.
545 The responses were spread around the y=x line (dashed line, corresponding to ideal predictions).
546 **B1, B2,** Violin plots showing the correlation coefficient between the predicted and the actual response
547 vectors of ORs in *D. melanogaster* **(B1)** or *A. gambiae* (n = 50) **(B2)**. Each point corresponds to one OR (n
548 = 24 ORs in *D. melanogaster* and n = 50 ORs in *A. gambiae*). Note that the correlations are mostly

549 positive. Statistically significant positive correlations (with P less than 0.05) are shown with black
550 outlines.
551 **C1, C2,** Violin plots showing the correlation coefficient between the control predictions and the actual
552 response vectors of ORs in *D. melanogaster* **(C1)** or *A. gambiae* (n = 50) **(C2)**. Note that the correlations
553 are close to zero in most cases.
554 **D1, D2,** Color maps showing the prediction quality for each OR-odor pair, for the set of 110 odors, and
555 all 24 ORs in *D. melanogaster* **(D1)** or all 50 ORs in *A. gambiae* **(D2)**. The prediction quality is measured
556 as $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$, where $D_{Pred,Actual}$ is the absolute distance between the predicted and
557 actual response, and $D_{Ctrl,Actual}$ is the absolute distance between the control prediction and the actual
558 response. The abundance of negative values (red shades) indicates that the predictions are usually
559 better than the control. Column averages and row averages of the matrix values are shown on top and
560 right, respectively.
561 **E1, E2,** Violin plots shows the average (over odors) of $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$ for each of the 24 ORs in
562 *D. melanogaster* **(E1)** or 50 ORs in *A. gambiae* **(E2)**.
563 In violin plots, horizontal line and the error bar indicate the mean and the s.e.m., respectively, within a
564 species.
565
566 **Fig. 4: Response prediction on an independent set of ORs.**
567 **A,** Color map showing the prediction quality (measured as $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$) for each OR-odor
568 pair, for the set of 68 odors and 26 novel ORs. Pairs whose values were not available are shown in grey.
569 Note the high frequency of negative values (red) and very low frequency of positive values (blue),
570 indicating that the predictions were closer to the actual value than the control were in most cases.
571 **B**, Violin plots shows the average (over odors) of $D_{Pred,Actual}$ - $D_{Ctrl,Actual}$ for each of the 26 ORs.
572 Horizontal line and the error bar indicate the mean and the s.e.m., respectively.
573 **C,** $D_{Pred,Actual}$ is plotted as a function of the number of ORs with known responses used in making the
574 predictions. These different numbers of ORs were sampled from the original set of 24 ORs with known
575 responses. For each number, the sampling was performed 10 times; for each sampling, selection of
576 response-determining positions and prediction of responses was performed.
577 **D,** For each number of sampled ORs, we calculated the fraction of the 506 OR-odor pairs (non-grey
578 values in panel **A**) for which the predicted responses were closer to the actual values than the control
579 predictions were. Both methods of generating control predictions (see **Methods**) were compared.

580 In panels **C** and **D**, errors bars indicate s.e.m. over the 10 samplings for each number from 3 to 21; for
581 24, all available ORs were used.
582
583 **Fig. 5: Localization of the top 20 response-determining positions.**
584 Protter plots showing the localization of the top 20 response-determining residues in representative
585 ORs: OR85b of *D. melanogaster* **(A1)** and OR1a of *A. gambiae* **(A2)**. Residues in the top 20 columns in the
586 corresponding species are shown with colored backgrounds, while the residues that are among the top
587 20 columns in both the species are shown with a black outline. One of the top 20 columns in *A. gambiae*
588 mapped to a gap in OR1a in the multiple sequence alignment, so only 19 positions are marked with
589 color.
590
591 **Table 1: Positions identified in previous mutagenesis studies that match the top positions identified in**
592 **our analysis.**
593 Of the 47 mutations that we collected from the mutagenesis studies, the 9 that matched the top
594 columns in *D. melanogaster* or *A. gambiae* are listed. The complete set of the 47 mutations is shown in
595 **Table S1**. TM: transmembrane helix; EL: extracellular loop.
596
597
598 **Table S1, relates to Table 1: Complete multiple sequence alignment of all the ORs.**
599 The table shows the multiple sequence alignment of all the ORs that have been used in this study. Each
600 row corresponds to one OR. The first two rows indicate with colored background the columns that were
601 ranked in the two species; among these the ranks of the top 20 columns are mentioned. The 47 residues
602 that have been identified as important in previous mutagenesis studies are marked with yellow color in
603 the matrix.

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

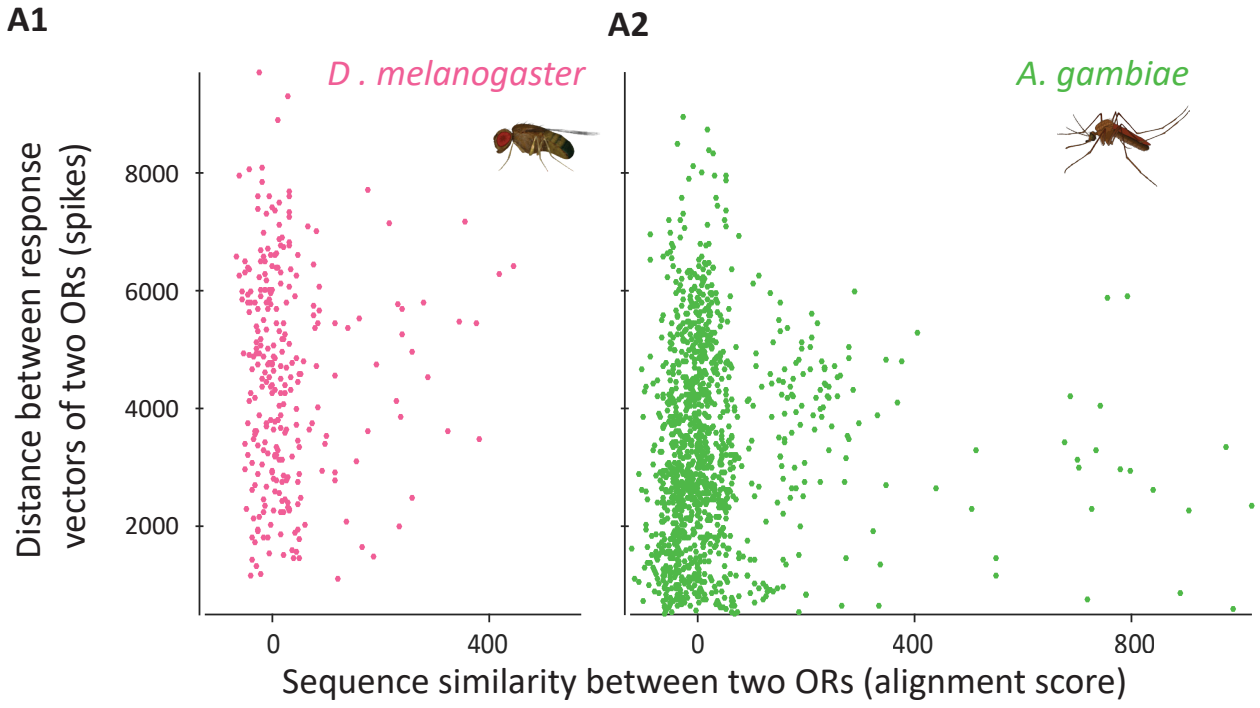

A1
A2
D . melanogaster
A. gambiae
Distance between response vectors of two ORs (spikes)
8000
6000
4000
2000
0
400
0
400
800
Sequence similarity between two ORs (alignment score)

Figure 2

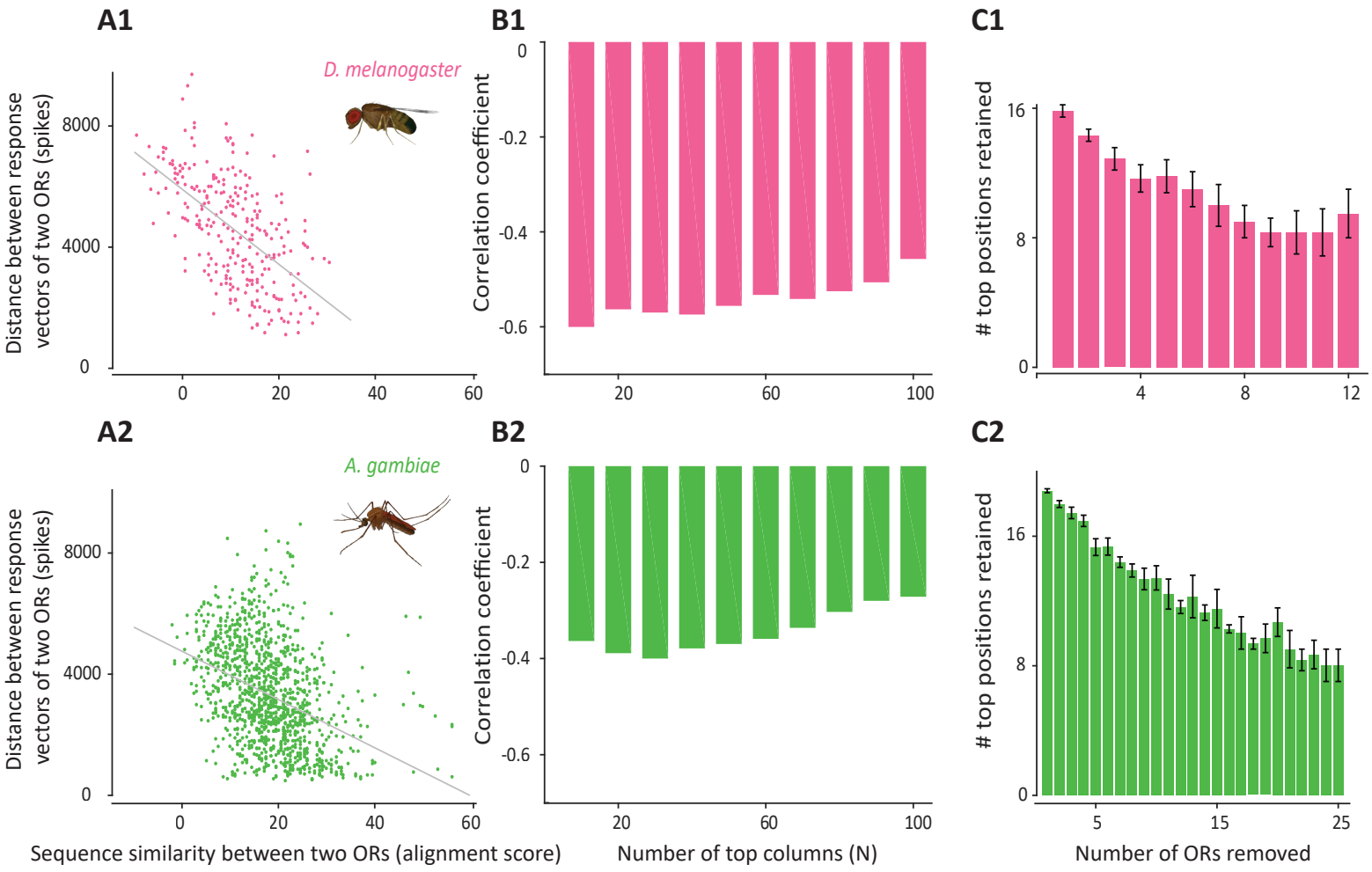

A1
D. melanogaster
Distance between response vectors of two ORs (spikes)
8000
4000
0
0
20
40
60
B1
Correlation coefficient
0
-0.2
-0.4
-0.6
20
60
100
C1
top positions retained
16
8
0
4
8
12
A2
A. gambiae
Distance between response vectors of two ORs (spikes)
8000
4000
0
0
20
40
60
Sequence similarity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

Figure 3

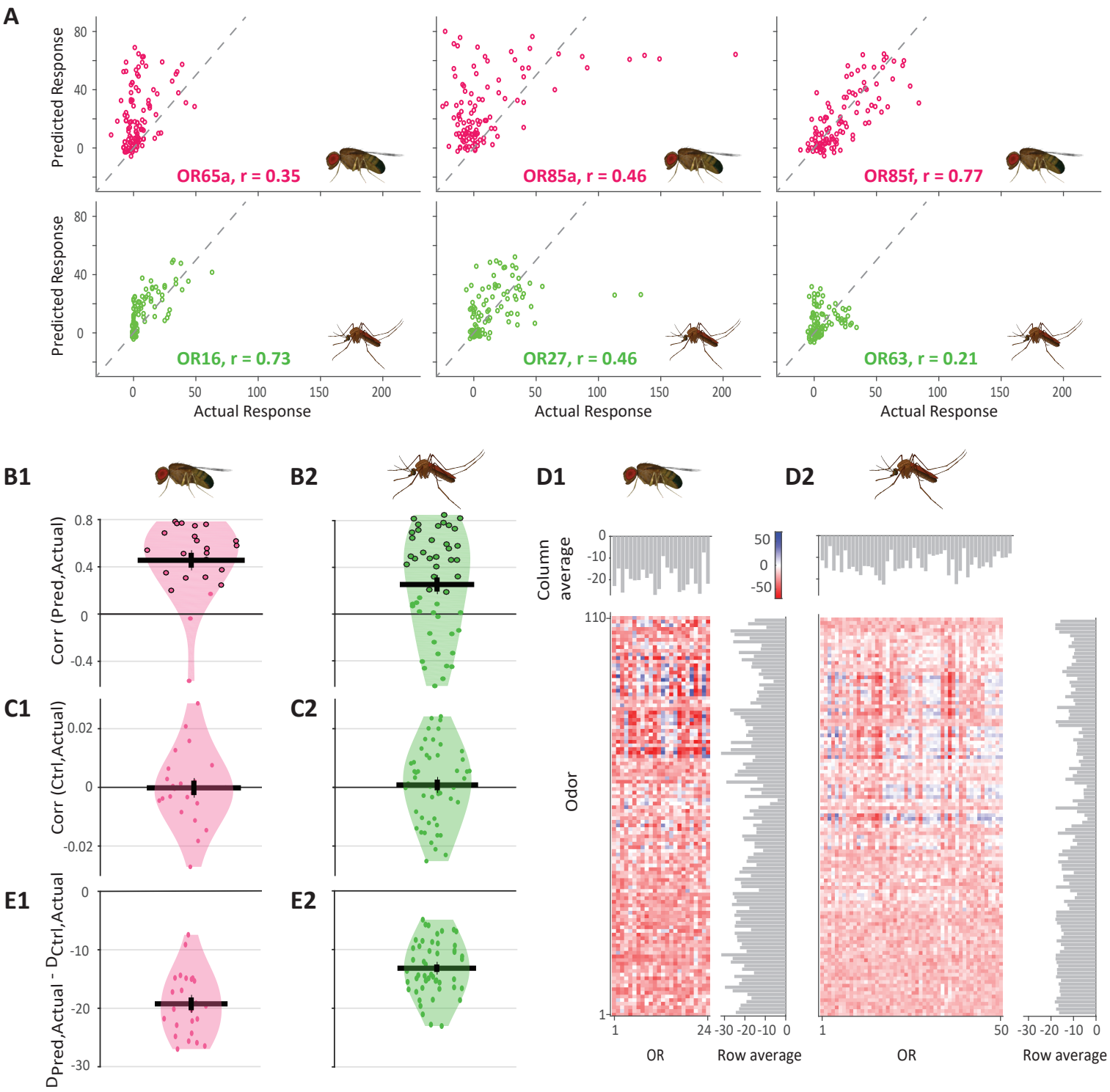

A
Predicted Response
Actual Response
OR65a, r = 0.35
OR85a, r = 0.46
OR85f, r = 0.77
OR16, r = 0.73
OR27, r = 0.46
OR63, r = 0.21
B1
B2
C1
C2
E1
E2
Corr (Pred,Actual)
Corr (Ctrl,Actual)
D1
D2
Column average
Odor
OR
Row average

Figure 4

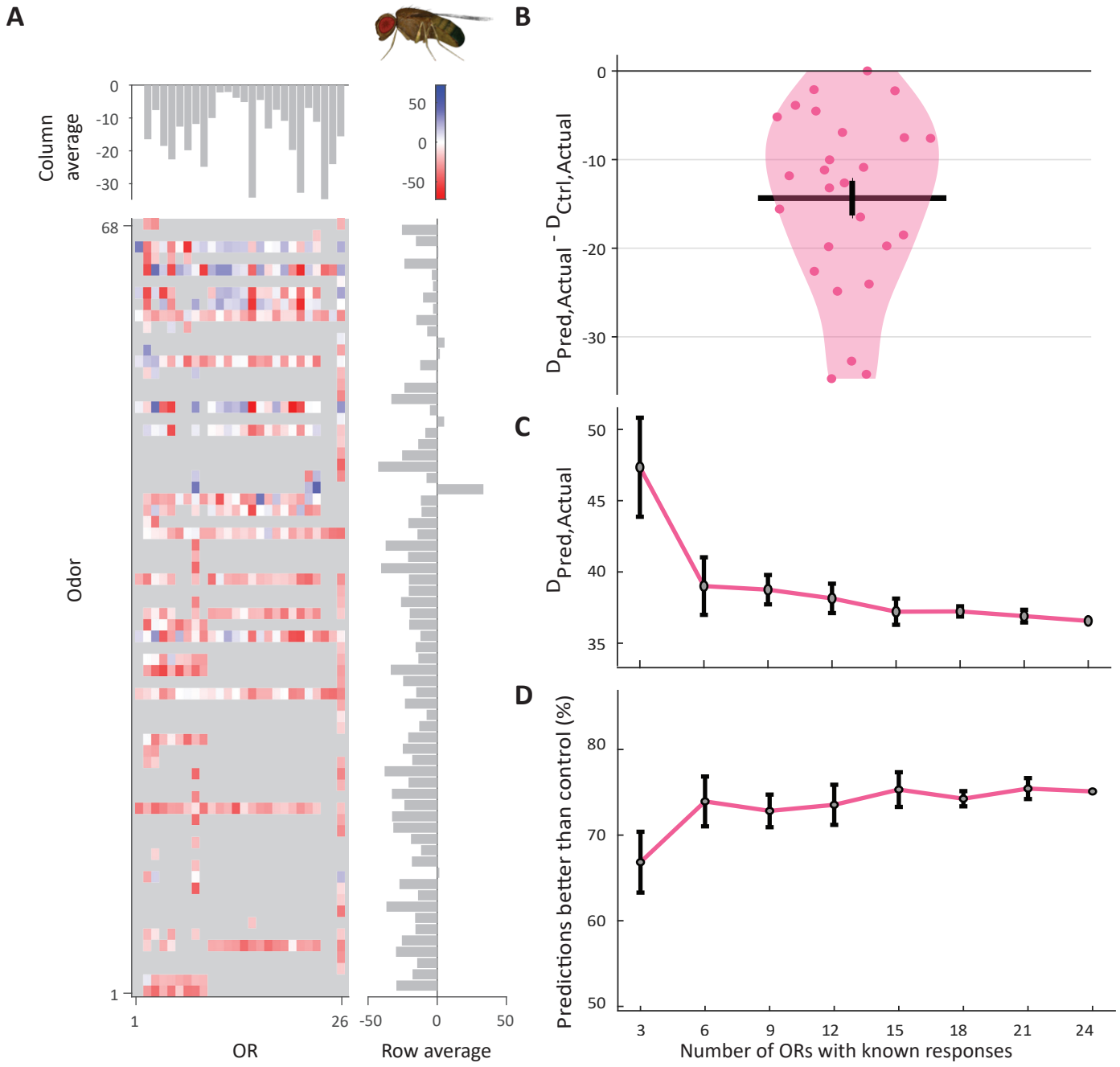
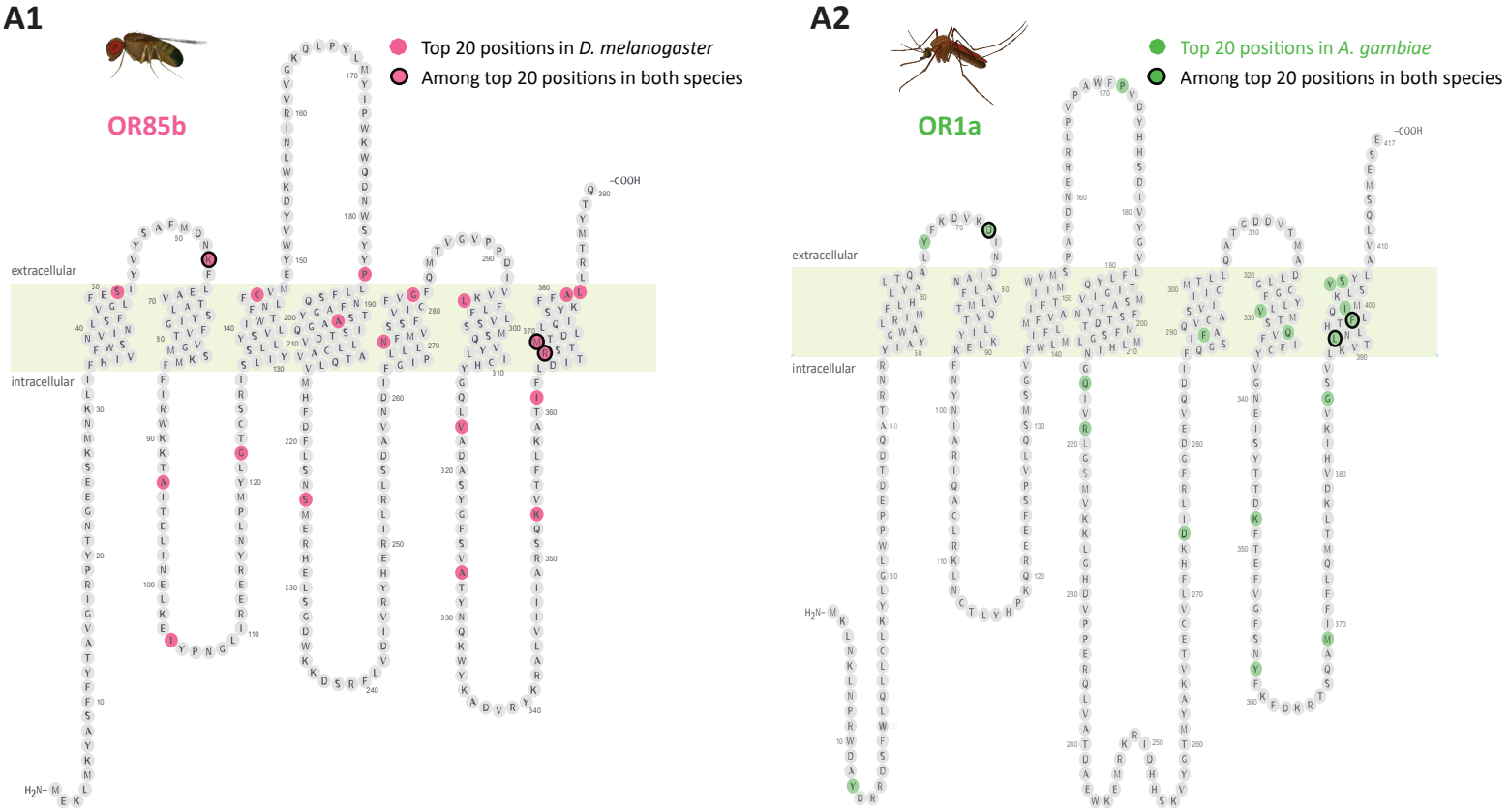

A
B
C
D
Column average
0
-10
-20
-30
50
0
-50
68
1
Odor
1
26
OR
-50
0
50
Row average
$D_{Pred,Actual} - D_{Ctrl,Actual}$
0
-10
-20
-30
$D_{Pred,Actual}$
50
45
40
35
Predictions better than control (%)
80
70
60
50
3
6
9
12
15
18
21
24
Number of ORs with known responses

Figure 5


A1
OR85b
Top 20 positions in *D. melanogaster*
Among top 20 positions in both species
extracellular
intracellular
-COOH
H2N-
A2
OR1a
Top 20 positions in *A. gambiae*
Among top 20 positions in both species
extracellular
intracellular
-COOH
H2N-