Supporting information for

Agonists of G protein-coupled odorant receptors are predicted from chemical features

C. Bushdid[#], C. A. de March[#], S. Fiorucci, H. Matsunami*, J. Golebiowski*

Experimental section

Luciferase assay in Hana3A cells

Dual-Glo Luciferase assay (Promega, Madison, USA) was used to determine the activities of firefly luciferase (Luc) and Renilla luciferase (Rluc) in Hana3A cells as previously described.[1] Luc luminescence, driven by a cAMP response element promoter (CRE-Luc; Stratagene California, California, USA), was used to determine the cell activation level. For each well of a 96-well plate, 5 ng SV40-RL, 10 ng CRE-Luc, 5 ng human RTP1s,[2] 2.5 ng M3 receptor,[3] and 5 ng of Rho-tagged odorant receptor plasmid DNA were transfected 24 h before the monitoring. After transfection, the odorants were injected into each well at a given concentration and left for 3.5 h. The luminescence of Luc and Rluc were then monitored. The normalized activity for each well was further calculated as (Luc-400)/(Rluc-400). The basal activity of the ORs of interest was averaged from four wells in the absence of odorants. For each receptor, odorant dilution was chosen so that it could allow a comparison with the control agonists while preventing cytotoxicity. The concentrations were set to 100 µM for OR51E1, 150 µM for OR1A1, and 300 µM for OR2W1 and MOR256-3. Odorant-induced activity was averaged from four wells and further corrected by subtracting the basal activity of that receptor. The response of the empty vector was also monitored as a control of the specificity of the odorant receptor response in four wells for each odorant and each concentration.
A molecule was considered agonist if it triggered an OR response higher or equal to 10% of that of the strongest tested agonist.

Efficacy of each tested compound was evaluated for different mutant ORs, whereby an efficacy value was identified as significantly different compared to the efficacy of OR51E1 *wt*; this was assessed using a one-way ANOVA and post hoc Dunnett's tests, and a significant difference was defined as $p < 0.01$.

Chemical space analysis

To examine the chemical space of all the ligands tested on OR51E1 prior to this study, we first calculated 4884 chemical and topological descriptors of 2577 commercially available odorants.[4] Molecular descriptors are mathematical values that describe the structure or shape of molecules and can be used to predict their activity and properties. To reduce the dimensionality of these descriptors, (obtained using Dragon software) [5] we performed a principal component analysis (PCA) using Knime.[6]
Principal component analysis (PCA) is a well-established method for dimensionality reduction that takes N points in an M-dimensional space and generates an orthogonal basis whereby these N points

are projected into a new M-dimensional space, but in which each successive dimension explains the maximal possible variance.

Because Dragon generated a very large number of descriptors, each was normalized to prevent descriptors with larger ranges from artificially dominating the dimensionality of the descriptor space. We further filtered for variance (0.05 cut-off) and for correlation (correlation filter set to 0.95) and obtained 66 descriptors. PCA was performed to reduce dimensionality; accordingly, the molecular features space that could be explained by each of the first 10 PCs accounted for ~80% of the variance. The effective dimensionality of the odorant space profile was much smaller than 66, with the first two PCs accounting for ~48% of the total variance and the first four for ~62% of the total variance. The full weight of all descriptors can be found in the Supporting Information (Figure S1).

Eight chemical features which are of interest due to their pharmacological importance, were selected to build a radar plot. These descriptors were estimated for the agonist and non-agonist groups. Namely, the descriptors estimated were Molecular Weight, Moriguchi octanol-water partition coefficient (LogP), number of donor atoms for H-bonds (#H donor), number of acceptor atoms for H-bonds (#H acceptor), total surface area from P_VSA-like descriptors, hydrophilic factor, surface area of acceptor atoms from P_VSA-like descriptors, and surface area of donor atoms from P_VSA-like descriptors.

Support Vector Machine model

Our numerical protocol comprised three steps, as follows: first we removed molecules in our library that had already been tested on OR51E1 according to the previous literature. From the remaining reduced library, we also excluded compounds that do not belong to the applicability domain of the model. Second, the remaining data were virtually screened and split into agonists and non-agonists using a supervised learning algorithm, *i.e.* Support Vector Machine (SVM). Third, assessment of the predictions was made by *in vitro* functional assays. Further details on these three steps will now be discussed.

Agonist vs. non-agonist spaces balance

To our knowledge, OR51E1 has been tested against 127 molecules, [4, 7-13] 7 of which were eliminated because of conflicting evidence as to whether the ligand was an agonist or a non-agonist. Of these 120 tested molecules, twenty-four were considered as agonists and 96 as non-agonists.

To avoid overfitting the model with non-agonist features, we created a set that was made up of a balanced number of agonists and non-agonists; we thus selected 24 non-agonists from the total 96. For this, the 96 known non-agonists were reduced to 24 representative molecules using a PCA followed by a k-medoids clustering approach. These 24 molecules were selected for the rest of the model building to span the chemical space of OR51E1. The final balanced test set included a total of 48 molecules (24

agonists and 24 non-agonists), and a model was built using a supervised machine learning method (Support Vector Machine, SVM).

Parameters such as splitting of the dataset were set to random and the proportion of molecules in the test set and the dataset were modified over several iterations. In our final model, a splitting proportion of 70:30 was chosen where 70% of the dataset (33 molecules) were allocated to the learning set and 30% (15 molecules) were allocated to the test set. Information about the splitting and molecules used to build the model are provided in a separate file (File S1). A C-SVC SVM model with a linear kernel was used. The SVM parameters were as follows: Cost (C)= 1 and Epsilon= 0.001. The kernel parameters were left to their default settings: degree= 3; gamma= 0; and coef0= 0.

Virtual screening and applicability domain

Our initial library containing 258 chemicals available for *in vitro* testing was filtered to exclude molecules that had already been tested on the target receptor by previous studies. This resulted in a total of 176 untested molecules that were retained for the virtual screening of OR51E1 (see File S2).

The SVM model was constructed on a randomly selected set of compounds. The mathematical model consequently learned from their molecular properties. The applicability domain of the model is the chemical space associated with the learning set. Indeed, the model cannot be expected to reliably predict the activity of molecules that are too different from the ones it has learned from.

We calculated *Pubchem* molecular fingerprints of the learning set and compared them to those of compounds in our library using a Tanimoto score, which measures the similarity between compounds and varies between 0 and 1, whereby a value closer to 1 indicates greater similarity. The molecules which has a Tanimoto index higher than 0.85 with respect to the learning set are considered as belonging to the applicability domain. They were therefore virtually screened by our model.

OR51E1 3D modeling

The 3D model of OR51E1 was built according to the protocol previously published [14]. Briefly, all 396 human OR sequences were aligned to the sequence of GPCRs for which the experimental structure is known. Manual adjustments were performed to be consistent with 123 mutational data of the literature. A homology model was obtained using the crystal structures of bovine rhodopsin receptor (PDB id: 1U19), CXCR4 chemokine receptor (3ODU), human adenosine A2A receptor (2YDV), and human chemokine CXCR1 receptor (2LNL) as structural templates using Modeller.[15]

The N-terminal structure was omitted to avoid perturbing the modeling protocol. Five models were obtained and the one consistent with the *in vitro* data and several structural constraints (no large folded structure in extra-cellular loops should be observed, all trans-membranes helices (TMs) folded as α-helices, and a tiny α-helix structure between TM3 and TM4) was kept for the ligand-docking step.

Agonists' structures and parameters were prepared with the antechamber module of AMBER with AM1-BCC charges. They were docked into the receptor cavity, using flexible docking parameters on residues His108$^{3.33}$, Tyr254$^{6.48}$ and Phe257$^{6.51}$ with Autodock Vina.[16] The structure associated with the best pose of each ligand was further subjected to a 10,000 step energy minimization process (5000 steps of steep descent). The resulting structure was considered for structural analysis.

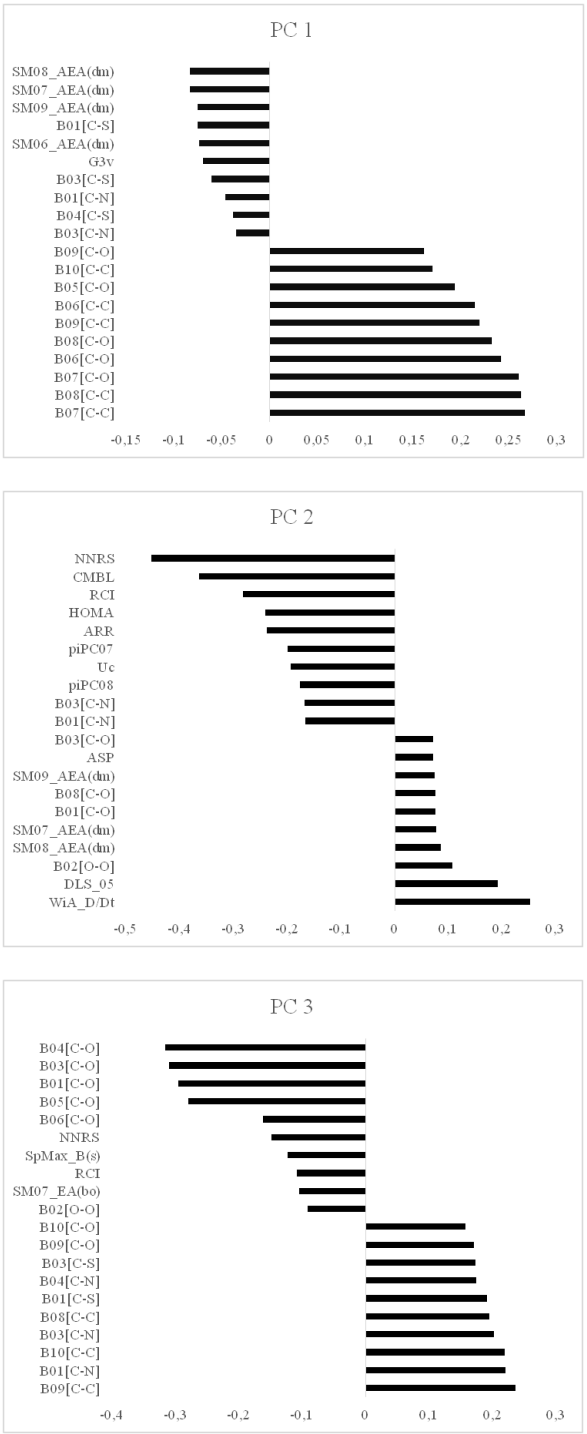

Figure S1. Descriptors having the largest weight in the 3 first principal components of the chemical space analysis. The descriptors meanings can be found at: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf

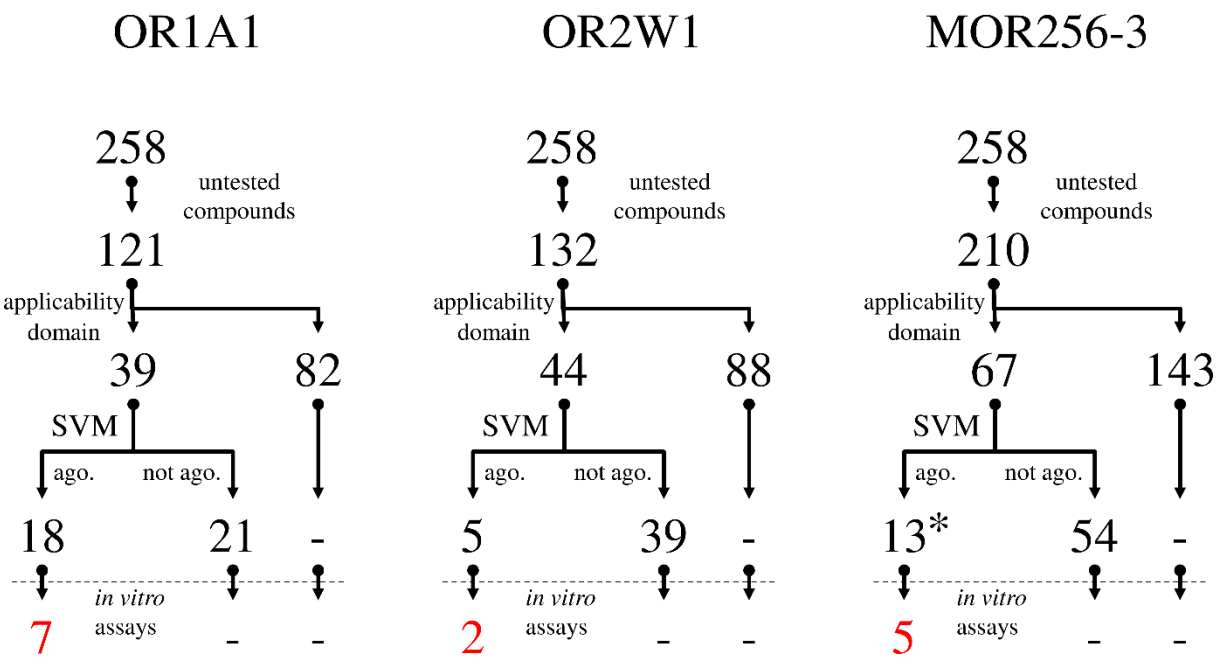

Figure S2: Workflow used for each additionally tested receptor (OR1A1, OR2W1, MOR256-3). In each case the compounds which had not been tested on the receptor were excluded, then the applicability domain of the model was defined, and predicted agonists were identified and tested in vitro. * means that in this case, the 10 compounds with the highest Tanimoto score were tested in vitro.

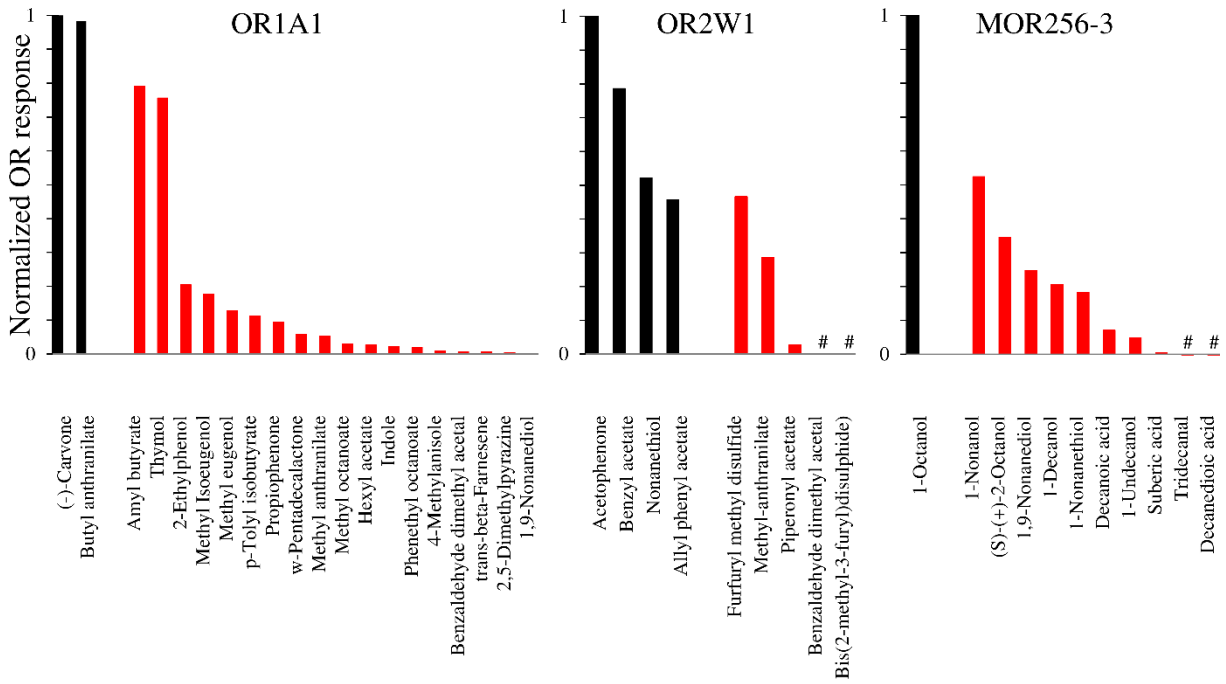

Figure S3: in vitro screening of agonists predicted (in red) by each model for OR1A1, MOR256-3 and OR2W1 and comparison with controls (in black). Seven novel agonists (triggering an OR response c.a. 10% of the strongest control) are identified for OR1A1, 2 for OR2W1and five for MOR256-3. # means that the recorded response was below zero. For OR2W1, the values of benzaldehyde dimethyl acetal and bis(2-methyl-3-furyldisulphide) are -0.1 and -1.48, respectively. A negative response does not mean antagonist activity but rather cell toxicity. For tridecanal and decanedioic acid in MOR256-3, the values are -0.01 and -0.03, respectively.

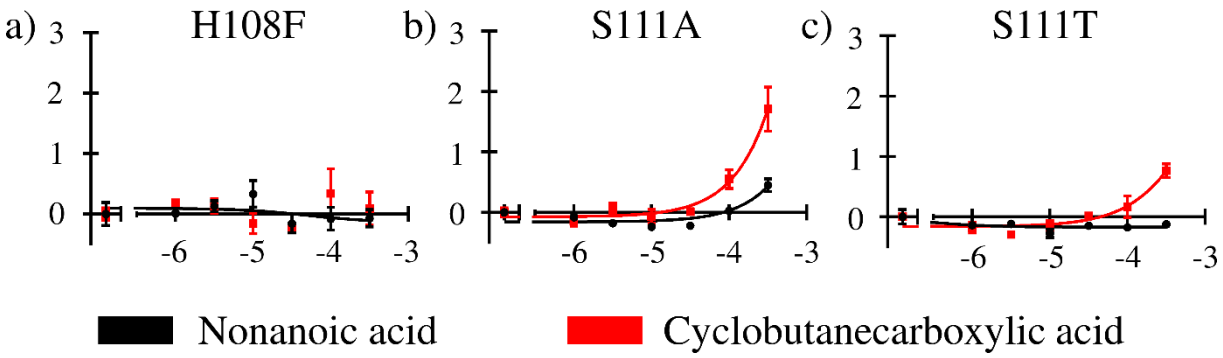

Figure S4: Dose response curves of OR51E1 a) H108F mutant, b) S111A mutant, c) S111T mutant.

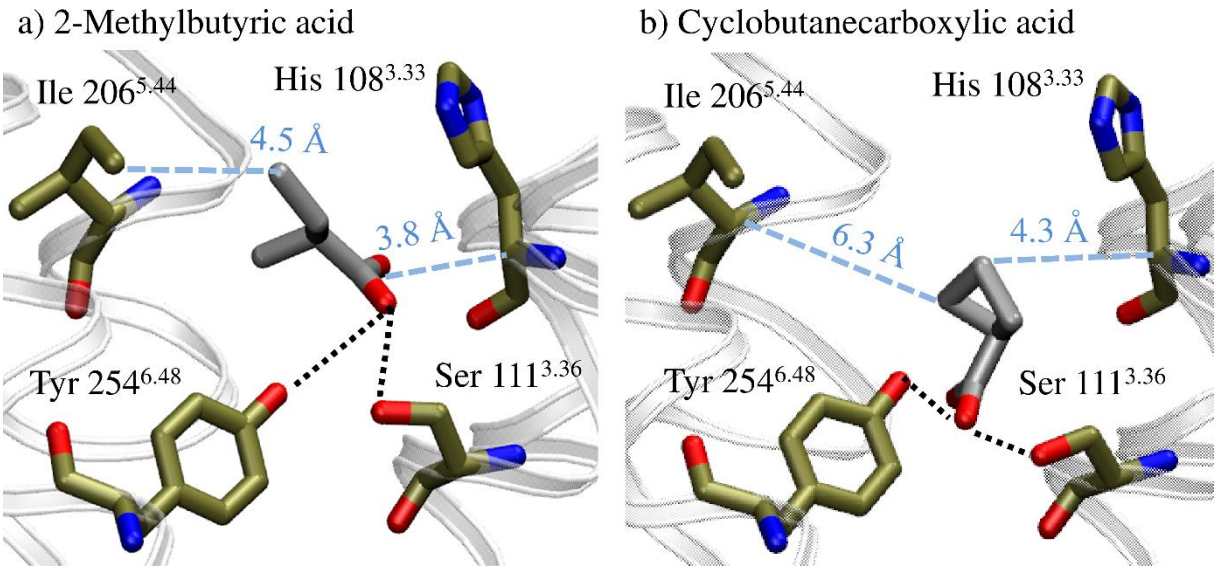

Figure S5: 2-methylbutyric acid shows closer contacts with H108 and I206 with respect to cyclobutanecarboxylic acid. a) Binding mode of 2-methylbutyric acid in the binding cavity of OR51E1. Carbon atoms are shown in gold (or gray in the case of the ligand), oxygen atoms are in red, and nitrogen atoms in blue. 2-methylbutyric acid is closer to the residues at the top of the binding cavity than cyclobutanecarboxylic acid. The closest distances between each acid and the receptor are shown in light blue. b) Binding mode of cyclobutanecarboxylic acid in the binding cavity.

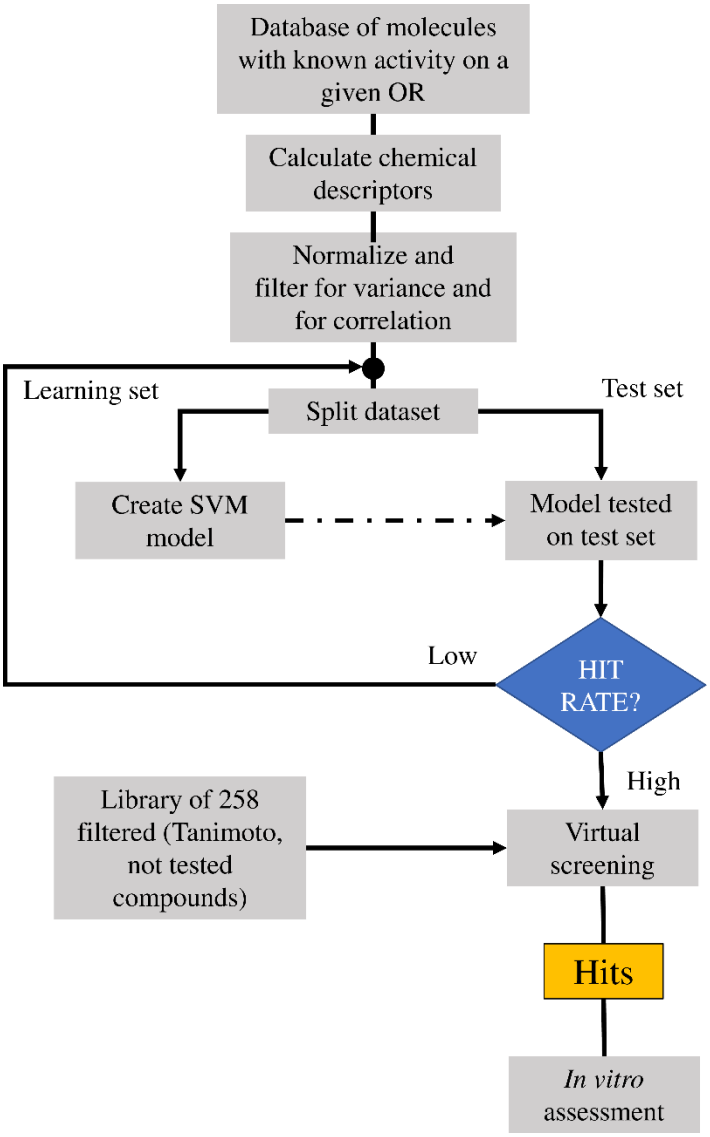

Figure S6. QSAR workflow.

Table S1. Molecules predicted by the SVM model on OR51E1 and their Tanimoto similarity scores.

CAS	Name	CID	Agonist of OR51E1	Tanimoto score
3658-80-8	Dimethyl trisulfide	19310	NO	1
116-53-0	2-Methylbutyric acid	8314	YES	1
6169-06-8	(S)-(+)-2-Octanol	2723888	NO	0,93
3721-95-7	Cyclobutanecarboxylic acid	19494	YES	0,88

File S1: All molecules which were tested on the studied receptors. For each OR, agonists and non-agonists were obtained from references [4, 7-12, 17-21].

File S2: 258 molecules available in the laboratory forming the virtual screening library.

References

1. Zhuang, H.; Matsunami, H. Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. *Nat. Protoc.* **2008,** *3* (9), 1402-13.
2. Keller, A.; Zhuang, H.; Chi, Q.; Vosshall, L. B.; Matsunami, H. Genetic variation in a human odorant receptor alters odour perception. *Nature* **2007,** *449* (7161), 468-72.
3. Li, Y. R.; Matsunami, H. Activation state of the M3 muscarinic acetylcholine receptor modulates mammalian odorant receptor signaling. *Sci. Signal.* **2011,** *4* (155), ra1.
4. Mainland, J. D.; Keller, A.; Li, Y. R.; Zhou, T.; Trimmer, C.; Snyder, L. L.; Moberly, A. H.; Adipietro, K. A.; Liu, W. L.; Zhuang, H.; Zhan, S.; Lee, S. S.; Lin, A.; Matsunami, H. The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **2014,** *17* (1), 114-20.
5. Dragon, T. s. *Software for Molecular Descriptor Calculation*, 6.0; 2014.
6. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner*. Springer: Berlin Heidelberg, 2007.
7. Saito, H.; Chi, Q.; Zhuang, H.; Matsunami, H.; Mainland, J. D. Odor coding by a Mammalian receptor repertoire. *Sci. Signal.* **2009,** *2* (60), ra9.
8. Audouze, K.; Tromelin, A.; Le Bon, A. M.; Belloir, C.; Petersen, R. K.; Kristiansen, K.; Brunak, S.; Taboureau, O. Identification of odorant-receptor interactions by global mapping of the human odorome. *PLoS One* **2014,** *9* (4), e93037.
9. Adipietro, K. A.; Mainland, J. D.; Matsunami, H. Functional evolution of mammalian odorant receptors. *PLoS Genet.* **2012,** *8* (7), e1002821.
10. Fujita, Y.; Takahashi, T.; Suzuki, A.; Kawashima, K.; Nara, F.; Koishi, R. Deorphanization of Dresden G protein-coupled receptor for an odorant receptor. *J. Recept. Signal Transduct.* **2007,** *27* (4), 323-34.
11. Geithe, C.; Andersen, G.; Malki, A.; Krautwurst, D. A Butter Aroma Recombinate Activates Human Class-I Odorant Receptors. *J. Agric. Food Chem.* **2015,** *63* (43), 9410-20.
12. Jovancevic, N.; Dendorfer, A.; Matzkies, M.; Kovarova, M.; Heckmann, J. C.; Osterloh, M.; Boehm, M.; Weber, L.; Nguemo, F.; Semmler, J.; Hescheler, J.; Milting, H.; Schleicher, E.; Gelis, L.; Hatt, H. Medium-chain fatty acids modulate myocardial function via a cardiac odorant receptor. *Basic Res. Cardiol.* **2017,** *112* (2), 13.
13. Veithen, A.; Philippeau, M.; Chatelain, P. High-Throughput Receptor Screening Assay. In *Springer Handbook of Odor*, Buettner, A., Ed. Springer: 2017; pp 505-525.
14. de March, C. A.; Kim, S. K.; Antonczak, S.; Goddard, W. A., 3rd; Golebiowski, J. G protein-coupled odorant receptors: From sequence to structure. *Protein Sci.* **2015,** *24* (9), 1543-8.
15. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **2006,** *Chapter 5*, Unit 5 6.
16. Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010,** *31* (2), 455-61.
17. Geithe, C.; Noe, F.; Kreissl, J.; Krautwurst, D. The Broadly Tuned Odorant Receptor OR1A1 is Highly Selective for 3-Methyl-2,4-nonanedione, a Key Food Odorant in Aged Wines, Tea, and Other Foods. *Chem. Senses* **2017,** *42* (3), 181-193.
18. Schmiedeberg, K.; Shirokova, E.; Weber, H. P.; Schilling, B.; Meyerhof, W.; Krautwurst, D. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. *J. Struct. Biol.* **2007,** *159* (3), 400-12.
19. Block, E.; Jang, S.; Matsunami, H.; Sekharan, S.; Dethier, B.; Ertem, M. Z.; Gundala, S.; Pan, Y.; Li, S.; Li, Z.; Lodge, S. N.; Ozbil, M.; Jiang, H.; Penalba, S. F.; Batista, V. S.; Zhuang, H. Implausibility of the vibrational theory of olfaction. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (21), E2766-74.
20. Braun, T.; Voland, P.; Kunz, L.; Prinz, C.; Gratzl, M. Enterochromaffin cells of the human gut: sensors for spices and odorants. *Gastroenterology* **2007,** *132* (5), 1890-901.

21. Yu, Y.; de March, C. A.; Ni, M. J.; Adipietro, K. A.; Golebiowski, J.; Matsunami, H.; Ma, M. Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (48), 14966-71.