Numerical modeling of olfaction

Caroline Bushdid

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES FONDAMENTALES ET APPLIQUÉES

THÈSE DE DOCTORAT

Numerical Modeling of Olfaction

Modèles numériques des mécanismes de l'olfaction

Caroline BUSHDID

Institut de Chimie de Nice

Présentée en vue de l'obtention du grade de docteur en Chimie **de l'**Université Côte d'Azur

Dirigée par : Pr. Jérôme Golebiowski
Soutenue le : 6 novembre 2018

Devant le jury, composé de :
Dr. Loïc Briand, Rapporteur
CSGA, Dijon
Dr. Laurent Chaloin, Rapporteur
IRIM, Montpellier
Pr. Andrea Büttner, Examinatrice
Fraunhofer Institute, Erlangen
Dr. Isabelle André, Examinatrice
LISBP, Toulouse
Pr. Uwe Meierhenrich, Examinateur
ICN, Nice
Pr. Jérôme Golebiowski, Directeur de thèse
ICN, Nice

Fondation Edmond Roudnitska

UNIVERSITE COTE D'AZUR

Ecole doctorale Sciences Fondamentales et Appliquées

Institut de Chimie de Nice

THESE DE DOCTORAT

Présentée en vue de l'obtention du
grade de docteur en Chimie de
l'Université Côte d'Azur par

CAROLINE BUSHDID

NUMERICAL MODELING OF OLFACTION

MODELES NUMERIQUES DES MECANISMES DE L'OLFACTION

Thèse dirigée par Prof. Jérôme Golebiowski
Soutenue le 6 novembre 2018

Jury :

Dr. Loïc Briand	CSGA, Dijon	Rapporteur
Dr. Laurent Chaloin	IRIM, Montpellier	Rapporteur
Pr. Andrea Büttner	Fraunhofer Institute, Erlangen	Examinatrice
Dr. Isabelle André	LISBP, Toulouse	Examinatrice
Pr. Uwe Meierhenrich	ICN, Nice	Examinateur
Pr. Jérôme Golebiowski	ICN, Nice	Directeur de thèse

Abstract

Humans have ~400 genes encoding odorant receptors (ORs) that get differentially activated by a virtually infinite space of small organic molecules. The combinatorial code resulting from this activation could allow the human nose to discriminate more than one trillion different olfactory stimuli. But how is the percept encoded in the structure of a molecule?
To understand how our nose decrypts the structure of molecules, numerical models were used to study the main protagonists of olfaction: ORs and odorants. These approaches included machine-learning methods to explore and exploit existing data on ORs, and molecular modeling to understand the mechanisms behind molecular recognition.
In this thesis I first review the structure-odor relationships from a chemist's point of view. Then, I explain how I developed a machine learning protocol which was validated by predicting new ligands for four ORs. In addition, molecular modeling was used to understand how molecular recognition takes place in ORs. In particular, a conserved vestibular binding site in a class of human ORs was discovered, and the role of the orthosteric binding cavity was studied. The application of these techniques allows upgrading computer aided deorphanization of ORs. My thesis also establishes the basis for testing computationally the combinatorial code of smell perception. Finally, it lays the groundwork for predicting the physiological response triggered upon odorant stimulation. Altogether, this work anchors the structure-odor relationship in the post-genomic era, and highlights the possibility to combine different computational approaches to study smell.

Keywords: Odorant receptors, olfaction, deorphanization, machine-learning, molecular modeling, virtual screening.

Résumé

L'Homme possède ~400 gènes codant pour des récepteurs aux odorants (ROs) qui sont différentiellement activés par un espace virtuellement infini de molécules. Le code combinatoire qui résulte de cette activation permettrait au nez humain de discriminer plus de mille milliards de stimuli olfactifs différents. Mais comment le percept est-il encodé dans la structure d'une molécule ?
Pour comprendre comment notre nez décrypte la structure des molécules odorantes, des modèles numériques ont été utilisés pour étudier les principaux protagonistes de l'olfaction : les ROs et les odorants. Ici, l'apprentissage automatique est utilisé pour explorer et exploiter les données déjà existantes sur les ROs. D'autre part, la modélisation moléculaire est employée pour comprendre les mécanismes qui sous-tendent la reconnaissance moléculaire.
Dans cette thèse j'ai passé en revue les relations structure-odeur du point de vue d'un chimiste. J'ai ensuite développé un protocole d'apprentissage automatique, qui a été validé pour prédire de nouveaux ligands pour quatre ROs. La modélisation moléculaire a été utilisée pour comprendre la reconnaissance moléculaire des ROs. Notamment, l'existence d'un site vestibulaire conservé dans une classe de ROs a été mis en évidence et le rôle de la cavité de liaison orthostérique dans les ROs a été étudiée. L'application de ces techniques permet de moderniser la déorphanisation guidée par ordinateur. Dans sa globalité, mes travaux ont aussi permis de préparer le terrain pour tester de façon virtuelle le code combinatoire des odeurs, et pour prédire la réponse physiologique déclenchée par ces molécules. Dans son ensemble, ce travail ancre la relation structure-odeur dans l'ère post-génomique, et souligne la possibilité de combiner différentes approches computationnelles pour étudier l'olfaction.

Mots clés : Récepteurs aux odorants, olfaction, déorphanisation, apprentissage automatique, modélisation moléculaire, criblage virtuel
.

Acknowledgments

First of all, I would like to express my most sincere gratitude towards my thesis supervisor, Prof. Jérôme Golebiowski. I will always remember his continuous support and guidance. It was a great pleasure to work in his team and learn how to conduct research. His stabilizing confidence, his patience while explaining scientific concepts, and his encouragements to pursue activities unrelated to research (such as teaching, giving broad audience talks, and participating in contests) contributed to build the researcher and person that I am today. My valuable PhD experience is now summarized in this thesis manuscript.

In addition to my thesis supervisor, I would like to thank my thesis committee Dr. Loïc Briand, Dr. Laurent Chaloin, Prof. Andrea Büttner, Dr. Isabelle André, and Prof. Uwe Meierhenrich for accepting to judge the work I conducted over the last three years.

A special thanks to my ICN colleagues whom I had the pleasure to meet, and notably to the ChemoSim team who accompanied me on a daily basis during this journey. In particular thank you to Hubert Grunig (and his python scripts), Jean-Baptiste Chéron (and his helpful computer skills), Sébastien Fiorucci (and his Snickers), Cédric Bouysset (and his insights on cheminformatics), Xiaojing Cong (and her delicious Chinese food), Jérémie Topin (and his artistic director talents) and Serge Antonczak (and his help with the administrative riddles). They were all immensely supportive throughout the thesis process, and provided valued perspective and advice. Their good humour contributed to build the trademark friendly and benevolent atmosphere in the lab which I will miss. I would like to particularly highlight my gratitude towards Jérémie Topin as we worked closely together during these past years. I appreciated his unceasing help in the analysis, production, and treatment of my results, but also his expertise and help for making the images which can be seen in this manuscript.

On an equally important note, I would like to express my gratitude towards all our collaborators outside the ChemoSim lab. Without them, my PhD wouldn't be as rich as it is today. Claire de March, and Hiroaki Matsunami, at Duke University provided uncountable *in vitro* experiments to test my hypotheses; Graham Hughes and Emma Teeling at University College Dublin contributed by performing phylogenetic analysis of OR sequences, Cheil Moon's team at the DGIST in South Korea brought an interesting "ectopic" perspective to our

work, and Moustafa Bensafi and his team in Lyon allowed me to share my molecular chemist expertise. I also would like to acknowledge the generosity of Loïc Briand and Christine Belloir, from the CSGA in Dijon, who both had a formative role. During the two weeks they hosted me in their lab, they taught me everything I know about *in vitro* testing, site directed mutagenesis and cell culture.

Furthermore, I am grateful towards all the organisms who granted me with financial aid to endorse my research during these past years. These include the Roudnitska Foundation (in particular, thank you to Michel Roudnitska and to Gilles Sicard), the Giract, the Gen Foundation, the Prix Pierre Laffitte committee and the Royal Society of Chemistry.

Last but not least, I am profoundly grateful for my family and in particular for my parents who have always shown unconditional support. They have always trusted my projects and had my back. Words are not enough to describe my endless thankfulness. I hope this little paragraph can contribute to show my profound appreciation. I would like to also thank my friends - those I met in Nice and those from abroad - for their moral support, their good humor and their positivity. Last but not least, thank you to Simo, for bearing with me - specially through the last months - for all his support and love.

Contents

Introduction (EN)

The olfactory system governs our sense of smell and plays a leading role in the perception of flavors. Without the sense of smell, food loses its most elementary flavor quality. But odors also play a central role in the avoidance of danger and in the establishment of interindividual non-verbal communications (1).

Chemists have tried establishing relationships between the structure of a molecule and its odorant properties since the XXth century. Such relationships have been identified for some particular odorant families and sometimes for chemical functions but their overall success has been limited (2). Recently, a crowd-funded challenge called the DREAM olfaction prediction challenge resulted in the development of several models which used machine learning algorithms that could accurately predict the pleasantness and intensity of molecules, as well as 8 out of 19 semantic descriptors (i.e. « garlic », « fish », « sweet », « fruit », « burnt », « spices », « flower », « sour ») with an average correlation of predictions across all models above 0.5 (chance giving a correlation of 0.05) (3). **This study showed the possibility to establish a predictive relationship between certain olfactory qualities and molecular numerical descriptors of odorants**. It is however likely that to go further, it is necessary to consider the molecular protagonists of our sense of smell.

The sense of smell begins when an odorant is recognized by one or more odorant receptors (ORs) expressed at the surface of olfactory sensory neurons (OSN). ORs are members of the G protein-coupled receptor (GPCR) family of membrane proteins. These types of receptors are the largest class of signaling proteins in the genome, and eminent environmental sensors. They are responsible for transducing physical or chemical signals into cellular responses. In ORs, the interaction which results from an odorant binding initiates a cascade of reactions inside the cell. Ultimately, the membrane of the neuron is depolarized and the olfactory message is transmitted to higher regions of the brain. Overall, the message encoded into the chemical structure is transformed into a neural activity code that gives rise to the perception of an odor (4).

The olfactory system evolved to successfully detect and discriminate amongst an extremely large number of volatile compounds present in our environment. To detect this wide array of structurally diverse chemicals, it is commonly accepted that the nose developed the so-called combinatorial code of olfaction (5). The combinatorial code of olfaction hypothesis relies on

the fact that an odorant can bind and activate one or more ORs and activate, and conversely, an OR can recognize several odorants. Thus, the nose uses particular sets of ORs to detect and distinguish far more airborne chemicals than there are individual ORs. ORs can therefore be considered as the gatekeepers to odorant perception. **Understanding how ORs are activated by odorant molecules is essential to structure-odor relationships. Potentially their activation could even be connected with behavioral and emotional output!**

The dimensionality of odorant space (millions of chemicals, hence billions of possible mixtures) and of receptor space (~400 functional ORs in humans) highlights the necessity to use computational approaches to better understand olfaction. Indeed, despite still being debated, the nose can discriminate a much larger number of olfactory stimuli than what was considered in the XXth century (6) and the number of molecules that theoretically have an odor is estimated to be greater than 27 billion (7).

In my PhD research project, **I focused on using numerical models to broaden our understanding of olfaction**. I mainly used two computational approaches: machine learning to mine data related to olfaction, and molecular modeling to understand the mechanisms underlying molecular recognition by ORs. This thesis presents a compilation of my research articles tracing the research I performed hand-in-hand with our collaborators. All of these research articles are therefore highly transdisciplinary and attempt to provide answers to practical as well as fundamental questions. The practical questions I aim to answer are: Is it possible to predict an OR activation simply by looking at a chemical structure? Does this chemical structure encode an emotional response? And more fundamental questions I sought to tackle were: How do differences in the amino-acids inside the ORs binding cavities affect molecular recognition? What is the role of conserved residues in the odorant receptor sub-genome?

First, an article covering a broad presentation of what is known about the sense of smell is presented. I review the state-of-the-art on structure-odor relationships, as viewed from a chemist's perspective. I discuss the role that ORs play in smell perception and explain the combinatorial code. I finally show how mutations can alter our sense of smell and wrap up the review with the role of ORs outside the olfactory epithelium. Originally written in French, the article is referenced as:

- Bushdid C., de March C.A., Topin J., Antonczak S., Bensafi M., Golebiowski J. 'Ces molécules qui nous mènent par le bout du nez. Le codage moléculaire de la perception des odeurs' *Actualité Chimique* 406, 21-30, **2016**

Then, journal articles presenting how machine learning can be used to better understand olfaction are compiled.

First, a methodological paper presents a protocol that can be used to mine the odorant space associated with ORs. I describe in detail the protocol, from the special care that should be put into the collection and curation of the data, to the establishment of a model. I emphasize on the necessity of assessing the applicability domain of the model, as well as on how to balance out the training set in order to avoid over-fitting. This article is referenced as:

- Bushdid C., de March C.A., Matsunami H., Golebiowski J. 'Numerical models and *in vitro* assays to study Odorant Receptors' *Methods in Molecular Biology* 1820, 77-93, **2018**

Second, an article showing how this method can be applied to explore the chemical spaces of ORs is provided. Through this article I show that the chemical space associated with already deorphanized ORs can be explored and broadened. I also focus on understanding how the differential response to an odorant is encoded into the receptors sequence. This article is referenced as:

- Bushdid C., de March C.A., Fiorucci S., Matsunami H., Golebiowski J. 'Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features' *J. Phys. Chem. Lett.* 9, 2235-2240, **2018**

Next, a focus is brought to the insights molecular modeling can bring to olfaction.

I first report the discovery of a conserved vestibular binding site in class I human OR genome. This article highlights the discovery of a new interaction site in these mammalian ORs that is specific to this sub-genome and could be responsible for the recognition of different chemical spaces by these ORs. This article is referenced as:

- Bushdid C., de March C.A., Topin J., Do M., Matsunami H., Golebiowski J. 'Mammalian class I odorant receptors exhibit a conserved vestibular binding pocket' ***currently under revision***

Second, I focused on understanding the specific role of the orthosteric binding cavity of an OR in its recognition profile. I showed how the binding cavity regulates both the number of chemicals an OR can be activated by, and also its responsiveness. This article is under preparation at the time of the writing of the document. It will be referenced as:

- Bushdid C., de March C.A., Yu Y.., Cong X., Ma M., Matsunami H., Golebiowski J. 'Molecular recognition profile of olfactory receptors is governed by their binding pocket' ***in preparation***

To conclude, two perspectives of the application of the research I performed will be discussed. The first one regards deploying molecular modeling on all human ORs to predict

odorant/OR interaction. The second concerns applying data-mining methods to predict psychophysiological responses upon olfactory stimulation in humans.

My contribution in this thesis can be summarized in these main points. I first reviewed the state-of-the-art knowledge about olfaction and odorant chemicals. I then showed how machine learning approaches will help chemists predict novel ligands for membrane receptors. I established the first steps suggesting that, in the foreseeable future, these types of methods might even be used to predict psychophysiological responses upon odor stimulation. I also demonstrated that evolutionary pressure led to the conservation of specific interaction sites in the mammalian OR genome. I explored the role the orthosteric binding pocket plays in the molecular recognition and activation process in ORs. Lastly, I led the groundwork for a genome-wide exploration of olfactory receptors. All of my results will be of use to deepen our understanding of the mechanisms and operations of the sense of smell.

References

1. R. J. Stevenson, An initial evaluation of the functions of human olfaction. Chem. Senses 35, 3-20 (2010).
2. C. S. Sell, On the unpredictability of odor. Angew. Chem. Int. Ed. 45, 6254-6261 (2006).
3. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820-826 (2017).
4. S. Firestein, How the olfactory system makes sense of scents. Nature 413, 211-218 (2001).
5. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).
6. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. Science 343, 1370-1372 (2014).
7. C. W. Yu, K. A. Prokop-Prigge, L. A. Warrenburg, J. D. Mainland, Drawing the Borders of Olfactory Space. Chem. Senses 40, 565-565 (2015).

Introduction (FR)

Le système olfactif régit notre odorat et joue un rôle de premier plan dans la perception des saveurs. Sans l'odorat, la nourriture perd sa qualité de sensorielle la plus centrale. Mais les odeurs jouent également un rôle crucial dans la prévention de dangers et dans l'établissement de communications non verbales interindividuelles (1).

Pour les chimistes l'établissement d'une relation entre la structure d'une molécule et ses propriétés odorantes reste un défi depuis le début XXe siècle. De telles relations ont été établies pour certaines familles odorantes et parfois pour des fonctions chimiques, mais leur succès global est resté limité (2). Récemment, le défi de prédiction olfactive DREAM a abouti au développement de plusieurs modèles qui utilisaient des algorithmes d'apprentissage automatique pour prédire des qualités olfactives. Il est à présent possible de prédire la valence et l'intensité perçue de molécules odorantes, ainsi que 8 descripteurs sémantiques sur 19 testés, (i.e. « odeur de poisson », « sucrée », « fruitée », « brûlée », « épicée », « fleurie », « acide ») avec une corrélation moyenne des prédictions - entre tous les modèles - supérieure à 0,5 (le hasard donnant une corrélation de 0,05) (3). **Cette étude montre la possibilité d'établir une relation prédictive entre certaines qualités olfactives et les descripteurs numériques moléculaires des substances odorantes**. Il est cependant probable que pour aller plus loin, il soit nécessaire de prendre en compte les protagonistes moléculaires entrant en jeu dans la perception des odeurs.

L'odorat commence par la reconnaissance moléculaire d'un odorant par un ou plusieurs récepteurs aux odorants (ROs) exprimés dans les neurones sensoriels olfactifs. Les ROs appartiennent à la famille de récepteurs membranaires couplés à la protéine G (RCPG). Ce type de récepteurs constituent la plus grande classe de protéines de signalisation dans le génome et ce sont d'importants capteurs environnementaux. Ils sont responsables pour la transduction des signaux physiques ou chimiques en réponse cellulaire. Dans les OR, l'interaction résultant d'une liaison avec un odorante initie odorant initie une cascade de transduction à l'intérieur de la cellule. Il en résulte une dépolarisation de la membrane du neurone et la transmission du message olfactif à d'autres régions du cerveau. Dans l'ensemble, le message codé dans la structure chimique est transformé en un code d'activité neuronale qui donne lieu à la perception d'une odeur (4).

Le système olfactif a évolué pour détecter et discriminer avec succès un très grand nombre de composés volatils présents dans notre environnement. Pour détecter ce large éventail de composés structurellement divers, il est communément admis que le nez adopte ce que l'on appelle le code combinatoire de l'olfaction (5). Dans ce code combinatoire, un odorant peut se lier à un ou plusieurs ROs et le(s) activer et, inversement, un RO peut reconnaître un ou plusieurs odorants. Ainsi, le nez utilise des sous-ensembles particuliers de ROs pour détecter et discriminer plus de molécules odorantes qu'il n'y a de ROs uniques. Etant donné que les ROs transforment le signal chimique en une série de pics de potentiel d'action interprétables par le cerveau, ils peuvent être considérés comme la pierre angulaire de la perception des odeurs. Le décryptage des mécanismes qui sous-tendent la reconnaissance moléculaire est fondamental pour déchiffrer le code de la perception olfactive. Comprendre comment les RO sont activés par des molécules odorantes est essentiel pour les relations structure-odeur. Potentiellement, leur activation pourrait même être liée à des mesures comportementales et émotionnelles !

La dimensionnalité de l'espace odorant (des millions de molécules, donc des milliards de mélanges possibles) et de l'espace récepteur mettent en évidence la nécessité d'utiliser des approches numériques pour mieux comprendre l'olfaction. En effet, malgré un débat sur le chiffre exact au sein de la communauté, le nez peut distinguer un nombre beaucoup plus important de stimuli olfactifs que ce qui a été considéré au XXe siècle (6). D’autant plus, le nombre de molécules ayant une odeur est théoriquement supérieur à 27 milliards (7).

Dans mon projet de recherche de doctorat, je me suis concentrée sur l'utilisation de modèles numériques pour approfondir nos connaissances sur l’odorat. J'ai principalement utilisé deux approches informatiques : l'apprentissage automatique pour extraire des informations à partir de données existantes sur l'olfaction et la modélisation moléculaire pour comprendre les mécanismes mis-en-jeu lors de la reconnaissance moléculaire par les ROs. Cette thèse présente une compilation d’articles de recherche retraçant les recherches que j'ai effectuées en partenariat avec nos collaborateurs. Tous ces articles sont hautement transdisciplinaires et tentent de fournir des réponses à des questions aussi bien pratiques que fondamentales. Les questions pratiques auxquelles je souhaite répondre sont les suivantes : est-il possible de prédire l’activation d’un RO en examinant simplement une structure chimique ? Cette structure chimique encode-t-elle une réponse émotionnelle ? Les questions plus fondamentales que j’aborderai sont les suivantes : Comment les différences dans les acides

aminés à l'intérieur des cavités de liaison des ROs affectent-elles la reconnaissance moléculaire ? Quel est le rôle des résidus conservés dans le sous-génome des ROs ?

Premièrement, un article couvrant une présentation large de ce que l'on sait sur l'odorat est présenté. Je passe en revue l'état de l'art sur les relations structure-odeur, du point de vue d'un chimiste. Je discute du rôle que jouent les ROs dans la perception des odeurs et explique le code combinatoire. Je montre enfin comment les mutations peuvent modifier la perception olfactive et mentionne le rôle des ROs en dehors du nez. Cet article a été initialement écrit en français et est référencé :

- Bushdid C., de March C.A., Topin J., Antonczak S., Bensafi M., Golebiowski J. 'Ces molécules qui nous mènent par le bout du nez. Le codage moléculaire de la perception des odeurs' *Actualité Chimique* 406, 21-30, **2016**

Ensuite, des articles décrivant la manière dont l'apprentissage automatique peut être utilisé pour mieux comprendre l'olfaction sont présentés.
D'abord un article méthodologique décrit un protocole pouvant être utilisé pour explorer l'espace chimique associé aux ROs. J'y souligne l'attention particulière qui doit être apportée à la collecte et au traitement des données, ainsi qu'à l'établissement d'un modèle. J'insiste sur la nécessité d'évaluer le domaine d'applicabilité du modèle, ainsi que la manière d'équilibrer les données utilisées pour l'apprentissage afin d'éviter un ajustement excessif. Cet article est référencé sous :

- Bushdid C., de March C.A., Matsunami H., Golebiowski J. 'Numerical models and *in vitro* assays to study Odorant Receptors' *Methods in Molecular Biology* 1820, 77-93, **2018**

Ensuite, un article montrant l'application de cette méthode pour explorer les espaces chimiques des ROs est fourni. À travers cet article, je montre que l'espace chimique associé aux ROs déjà déorphanisés peut être exploré et élargi. Je me concentre également sur la compréhension de la façon dont la réponse différentielle à une substance odorante est codée dans la séquence des récepteurs. Cet article est référencé comme :

- Bushdid C., de March C.A., Fiorucci S., Matsunami H., Golebiowski J. 'Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features' *J. Phys. Chem. Lett.* 9, 2235-2240, **2018**

Dans un troisième temps, l'utilisation de la modélisation moléculaire pour étudier l'olfaction est présentée.

Premièrement, je décris la découverte d'un site de liaison vestibulaire conservé dans une sous-partie du génome de ROs humains. Cet article met en évidence un site d'interaction qui est conservé spécifiquement dans le sous-génome des récepteurs de la classe I et qui pourrait être responsable de la reconnaissance de différents espaces chimiques par ces ROs. Cet article est référencé sous :

- Bushdid C., de March C.A., Topin J., Do M., Matsunami H., Golebiowski J. 'Mammalian class I odorant receptors exhibit a conserved vestibular binding pocket' ***actuellement en révision***

Deuxièmement, je me suis concentrée sur la compréhension du rôle spécifique que la cavité de liaison orthostérique joue dans le profil de reconnaissance d'une famille de ROs. J'ai montré comment la cavité de liaison régule à la fois le spectre de reconnaissance, mais aussi que la force de sa réponse est encodé par des résidus en dehors de la cavité. Cet article est en cours de rédaction au moment de l'écriture de ce manuscrit. Il sera référencé comme :

- Bushdid C., de March C.A., Yu Y., Cong X., Ma M., Matsunami H., Golebiowski J. 'Molecular recognition profile of olfactory receptors is governed by their binding pocket' ***en préparation***

Pour conclure, deux perspectives de l'application de la recherche que j'ai effectuée seront discutées. La première concerne le déploiement de la modélisation moléculaire pour créer des modèles atomiques de tous les ROs humains. Ceci permettrait de prédire les interactions odorant-RO à l'origine du code combinatoire. La seconde concerne l'application de méthodes d'apprentissage automatique pour prédire les réponses psychophysiologiques mesurées lors de stimulations olfactives chez l'homme.

Ma contribution dans cette thèse peut être résumée en ces principaux points : J'ai d'abord passé en revue les connaissances sur l'olfaction et les molécules odorantes. J'ai ensuite montré comment les méthodes d'apprentissage automatique aideront les chimistes à prédire de nouveaux ligands pour ces récepteurs. J'ai établi les bases pour permettre la prédiction de la réponse psychophysiologique lors d'une stimulation olfactive. J'ai également montré que la pression évolutive a conduit à la conservation de sites d'interaction spécifiques dans le génome olfactif des mammifères. J'ai exploré le rôle que joue la cavité de liaison orthostérique dans le processus de reconnaissance moléculaire et d'activation dans les ROs. Finalement, j'ai mené les travaux préparatoires pour permettre l'exploration à l'échelle du

génome de ces récepteurs. Tous mes résultats seront utiles pour approfondir notre compréhension des mécanismes et du fonctionnement de notre nez.

References

1. R. J. Stevenson, An initial evaluation of the functions of human olfaction. Chem. Senses 35, 3-20 (2010).
2. C. S. Sell, On the unpredictability of odor. Angew. Chem. Int. Ed. 45, 6254-6261 (2006).
3. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820-826 (2017).
4. S. Firestein, How the olfactory system makes sense of scents. Nature 413, 211-218 (2001).
5. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).
6. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. Science 343, 1370-1372 (2014).
7. C. W. Yu, K. A. Prokop-Prigge, L. A. Warrenburg, J. D. Mainland, Drawing the Borders of Olfactory Space. Chem. Senses 40, 565-565 (2015).

Part I
State-of-the-art on structure-odor relationships

"The scent organ was playing a delightfully refreshing Herbal Capriccio."

Aldous Huxley (Brave New World, 1932)

Publication 1

Those molecules that lead us by the nose: The molecular coding of odor perception

CAROLINE BUSHDID, CLAIRE A. DE MARCH, JÉRÉMIE TOPIN, SERGE ANTONCZAK, MOUSTAFA BENSAFI, JÉRÔME GOLEBIOWSKI

IN ACTUALITÉ CHIMIQUE 406, 21-31, **2016**

"*Those who control odors control the hearts of men*" – P. Süskind (Le Parfum, 1985, 263p)

Abstract

The sense of smell is a “chemical” sense as it allows us to perceive volatile molecules present in our environment. The chemical information extracted from them eventually modifies our behaviors and influences our relationships with others. But smells remain highly elusive since we have deeply personal connections to them. In this context, the chemists’ primary concerns are the molecular features of odorant compounds which will be translated into a perceivable odor. Despite many efforts, the establishment of a reliable structure – odor relationship remains highly challenging. To unravel such a link, interdisciplinary research combining chemistry, biology, behavioral and social sciences will surely be required. This article focuses on the structure – odor relationship, with a particular emphasis on the physiological stages leading to smell perception. It notably highlights the fundamental role of receptors expressed by our olfactory neurons.

Introduction

10% of people in France suffer from an impairment of smell ranging from very slight hyposmia (partial loss of smell) to total anosmia (total loss of smell) (see http://www.olfaction.cnrs.fr). Yet every day we are confronted with odorous stimulations. Beyond the natural smells that surround us, we use perfumes, modify our body odors by the use of deodorants or shampoos and our clothes are impregnated by the fragrance of detergents or fabric softeners. Our daily lives are driven by actions involving odors and we agree on the use of arbitrary descriptors such as the smell of “clean”, “fresh” or “food”. These adjectives are rooted in our cultures due to a repeated exposure to the smells that social or marketing laws have imposed on us. We can clearly see the intimate link that exists between the molecular message and the cultural aspect of odors. As in any form of art, the perfumer or flavorist apprentices are in quest of “beautiful” and “good” scents, even before the marketing concepts are defined. Although the creation of a perfume is still not considered a work of art, organizations such as the E. Roudnitska Foundation or the International Society of Perfumers-Creators are fighting for this community to be recognized as being comprised of artists.

While in ancient times, smells were controlled, valued and sublimated, (1) other epochs have been darker for perfumes. In the middle ages, smells were mostly recognized through the suffering sensations they elicited. The pestilence of the cities and the lack of hygiene of the population, mixed with the great modesty due to religion, led to the association of odors with

the world of animality and witchcraft. Even later, for a philosopher like Kant, the subjective character of smells brings man back to his animality and is described as the most vulgar of the five senses deserving only contempt (2). It was not until the end of the 19th century that the perception of smells began to be rehabilitated by philosophers such as Nietzsche, who valued odors, their evocative powers, as well as their perception (3).

Thus, at the turn of the last century, the notion that the modern, civilized human being can do without the sense of smell became obsolete: its hidden meaning, without words, previously considered superfluous, has become an object of studies and therefore new knowledge. Simultaneously, society became more hedonistic and thus more mindful on the emotional effects elicited by smells. Odors are present in our food, source of both pleasure and social interactions; they influence our sexuality, our relationships with others in general and particularly with our children. They contribute to our emotional balance and our well-being, since the loss of smell has deleterious consequences on this equilibrium (4, 5). To sum up, odors play a fundamental role in behaviors such as food intake or the detection of environmental hazards (6). Generally speaking, just like appearance, eloquence or purchasing power, they have become social markers.

But the perception of smells is first of all "chemical". Its function is above all to quantify and qualify the volatile molecules present in our environment. Chemists plays a leading role in the development and characterization of these molecules that lead us through the nose. And by chemists, we mean the great family of molecular scientists, which extends from the physical chemists to the molecular biologist as well as organic chemists or the neuro-chemists. All of these disciplines contribute to understanding the tremendous complexity of the mechanisms involved from the inhalation of a molecule to the perception of its smell, or even the modification of our behavior. This interdisciplinarity will allow us to remove the locks to crack the molecular code of odorant perception. Since the discovery of the genes encoding olfactory receptors, smell perception research was projected into the post-genomic era, which can be defined as the period after the completion of human genome sequencing. An epoch which is "dominated by transdisciplinarity, speed and the centrality of information technologies that mark contemporary life sciences" (7). In this context, this article supplements a previous update published in 2005 (8).

The olfactory space

The auditory and visual spaces are relatively well defined. Their perception can be predicted on the basis of the physical properties of stimuli, such as wavelength or frequency. Our visual system distinguishes millions of colors ranging from wavelengths between 390 and 700 nm. Similarly, in terms of audition, we distinguish about 340,000 different tones ranging from 20 to 20,000 Hz.

The olfactory space remains to be defined because it seems to be much subtler. Despite numerous studies aiming to establish a link between the physicochemical characteristics of odorant molecules and the olfactory percept, no universal rule governing a relationship between molecular structure and odor has been established. The dimensionality of the odorant chemical space – which is virtually infinite – and especially the lack of data analysis methods as complex as the vocabulary related to odors are major obstacles to this definition. The description of an odor is much more approximate than the characterization of a color or a sound. It remains both highly variable and too subjective from one individual to another because it is influenced by culture (especially between trained and untrained subjects) (9).

The difficulty of measuring, characterizing or categorizing the odorant chemical space is illustrated by the fact that the resolution of the human smell has only recently been established. It seems that we would be able to discriminate not less than one trillion olfactory stimuli (10). This new estimate far surpasses the previous ones which limited our detection capacity to 10,000 smells! Although this estimate remains controversial, it illustrates that from a physiological point of view, we are far from being deprived in olfaction. But to discriminate is one thing, to characterize and verbalize is something else.

The ability of humans to identify odorants in complex mixtures is also known. We are able to identify compounds when they are present in simple mixtures containing only a few different molecules (2 to 5). Intriguingly, the odor of mixtures of at least 30 molecules associated with different odorous notes converges towards a single olfactory note. It is referred to as the "olfactory white" to echo its similarity to visual or auditory "whites" (which are mixtures of all colors or mixtures all the sound frequencies) (11).

Smell, culture and emotion

One can't help but wonder, could it be that the odorant message is not the finality, but merely an intermediate in the hedonic link that connects us to those molecules? Other descriptors

which are directly associated with our valuation are used well before the olfactory descriptors. Some descriptors are even non-verbal as they can directly be measure from the response of our bodies (vide infra).

From a perceptual point of view, the simplest dimension is the hedonic character, which represents the pleasantness or unpleasantness of a smell. In psychology it is called "valence". The preference for some olfactory notes over others is strongly related to our environment and our past. This cultural effect seems to be begin in utero, since certain food preferences (for carrots, vanilla, broccoli, and even anise) are associated with the consumption of this types of foods by the mother during pregnancy and breastfeeding (12, 13). These dietary preferences can even be retained until adulthood. They tend to come from our regions of origin!

But how can these psychological effects be linked to physical and chemical characteristics? Research conducted more than 40 years ago had suggested a relationship between the molecular weight of odorants and their hedonic valence (the lightest odorants being the most unpleasant). Nevertheless, the relationship between molecular structure and hedonic perception cannot be reduced to such a simple chemical characteristic. Intriguingly, although the physiological basis could not be established, structurally "simple" odorants (as opposed to those with an embranched chemical structure – considered as "complex") are perceived as having a more negative valence (14).

Although no universal rule has so far been established, some empirical rules, centered on chemical or odorous families, have allowed chemists to attempt to categorize odors based on their physicochemical properties.

After the publication of this review article, a study was published by Keller *et al.* (15). It described how a crowd-funded project called the "DREAM Olfaction Prediction Challenge" used machine learning to predict human olfactory perceptions from a set of unpublished perceptual data. Briefly, perceptual data was gathered from 49 individuals who rated 476 different molecules by intensity and pleasantness as well as by 19 other semantic descriptors. Once the data was gathered, 21 teams competed to produce the best models using machine learning. This study showed that the models outperformed previous models in the prediction of pleasantness. They could also, to a certain extent, predict the intensity of a given molecule as well as 8 of the 19 semantic descriptors.

The models, which used chemical features as an input, only need a small fraction of the calculated features to achieve optimal prediction. By analyzing the chemical features associated with perception, Keller *et al.* showed that sulfuruous molecules were rated as smelling like "burnt" and "garlic" and that pleasantness correlated with the molecular size and

with the structural similarity to certain molecules (paclitaxel and citronellyl phenyl acetate). Additionally, the predictive models allowed reverse-engineering a desired perceptual profile, opening the way to the establishment of more powerful predictive tools using different perceptual measurements.

Even though these results are encouraging, the models were limited due to the interindividual variability observed when rating an odor, but also by the use of semantic descriptors which might not be the best choice for predicting a smell.

Olf-Active molecules

Ancient beliefs attributed to smells the powers of triggering physical reactions.

More prosaically, in aromatherapy, essential oils are thought to have properties such as anti-stress, energizing and even anti-depressive. Robust scientific bases remain to be established. However, it is easy to envision an effect of odors on our well-being since the limbic system – closely related to our mood, our memory and our sexual desire – is directly and strongly recruited by our olfactory system (16). There are few studies on the evaluation of a so-called "psycho-physiological" effect upon odor stimulation. This effect can be assessed by measuring physiological parameters under the control of the autonomic nervous system. These could be body temperature, sweating, heart beat rate, breathing rate or muscle contraction. For example, the essential oils of ylang-ylang, peppermint or the linalool molecule (which has a smell of lavender and bergamot) lower the body temperature of the studied subjects, suggesting that they have relaxing properties (17). Cis-3-hexenol (a molecular carrying a characteristic cut grass odor) and trans hexen-2-al (eliciting a green apple odor) significantly reduce stress and anxiety in rodents, suggesting a potential relaxing effect of the so-called "green" odors in humans (18). For the moment, our understanding of the effects of odors on our body remains largely unexplored. The identification of a rational mechanism linking the chemical properties of odorant molecules to their psycho-physiological effects – and not only to their odor – remains to be established.

Odorant molecules

The olfactory sensation is conditioned by various factors that depend on the molecular structure of odorants. In addition to a certain hydrophobicity, the molecule that codes for an odor must be sufficiently volatile to be transported by the air we breathe. This term, although

intuitive, is rather delicate to define in perfumery. Molecular weight, vapor pressure (which directly relates to the quantity of molecules present in the gas phase) or logP (water / octanol partition coefficient) are typically good indicators of this volatility. (8) This definition, focused on the molecule, has its flaws. Oxygen, nitrogen or methane are perfect counterexamples. None of them has a smell, although their physical and chemical characteristics (they are volatile and hydrophobic) correspond to the criteria stated above. Among these highly hydrophobic molecules the absence of stimulation of olfactory receptors could be at the core of the absence of odor. But then, how to explain the tenacious smell of ozone (O_3, which comes from the Greek *ozô*, which means "to exhale an odor"), causing the characteristic smell of photocopy rooms? Concentration can also influence the smell of a molecule. Again, the following example highlights the difficulty to establish a structure-odor relationship: 4-mercapto-4-methylpentan-2-one molecule (called "cat ketone") has an odor of cat urine at high concentration while its dilution gives it a note of "blackcurrant" or "cabernet-sauvignon".

One of the challenges a perfumer is confronted to, beyond the realization of the perfume itself, is to consider the matrix that will deliver the perfume (for example a body cream, a detergent or a shampoo). Although the previous descriptors are rational indicators for predicting the behavior of molecules, their interactions with complex matrices make predictions extremely delicate (19).

In perfumery, the terms "substantivity" and "retention" of an ingredient are preferred. These terms reflect the behavior over time of an ingredient, usually on a particular matrix such as hair, skin or clothes. These parameters are of course equally important in fine perfumery as in functional perfumery, here the fragrances are incorporated into bases such as shampoos, creams or detergents.

Musk or musks

The natural musky odor comes from muscone (Figure 1), produced in the anal glands of musk deers (*Moschus moschiferus*). Historically, the musky smells were achieved by drying those glands and then infusing them to obtain a perfumery ingredient. Nowadays, these sources of supply have become obsolete for both ethical and economic reasons (20).

Due to the macrocyclic structure of natural musky smelling compounds, their synthesis and production has long been a challenge for organic synthesis. The difficulty lies in the entropy of the system: during the cyclization of a long linear structure, the probability of an

intermolecular reaction occurring is greater than that of an intramolecular reaction. This synthesis is now possible but its experimental constraints are not adapted to an industrial production. The use of musky odor compounds with a simpler chemical structure is therefore preferred in perfumery.

For example, the discovery of Musk Baur (Figure 1) has been of major importance for the perfume industry. In 1888, Baur, a chemist wishing to optimize explosives, fortuitously developed a "nitromusc" by modifying the structure of trinitrotoluene (TNT)! The derivatives of this compound were widely used as an alternative for musky notes until polycyclic musks were discovered in the 1950s (Figure 1).

Note that other animals can also produce molecules with musky odors, such as the "civet" cat (*Viverra civetta*) that produces civetone (Figure 2). Today, even though the glands are not used in perfumery anymore, the animals are still exploited to produce Kopi Luwak coffee that has a particular musky flavor. The animals eat the coffee fruit but are unable to digest the seeds, which are excreted after maceration in the intestinal tract of the cat. This is where they acquire their particular taste that makes this coffee the most expensive beverage in the world! The kilo of coffee costs several hundred US dollars.

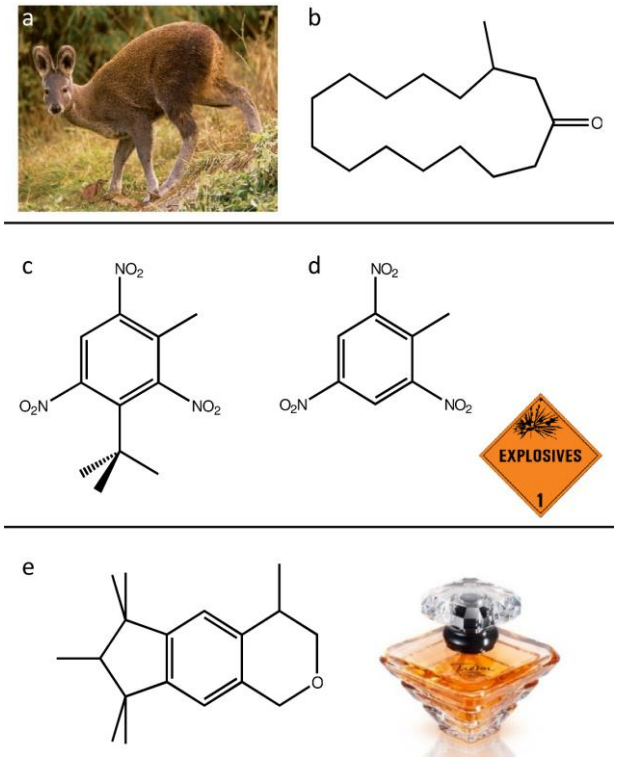


Figure 1. a) Musk Deer, **b)** A musky odor molecule of animal origin: muscone. **c)** A derivative of TNT with a musky odor: Baur musk, and **d)** TNT **e)** A polycyclic musk: the Galaxolide® musk that gives off the musky odor of many detergents or the Trésor perfume by Lancôme.

Musk and smell of clean

Musky odor molecules are also now widely associated with cleanliness because of their widespread use in functional perfumery. They are used both for their perfuming qualities and as fixatives. Their large size and high boiling point makes them ideal for retaining more volatile molecules and thus optimize the kinetics associated with perfuming properties. In order to obtain a laundry detergent with a "substantive" smell, the odorant compounds must effectively be delivered on clothes and it must survive rinsing and drying. Musks fulfill this function perfectly in laundry detergents and their repeated use eventually resulted in the association of their smell with that of clean linen.

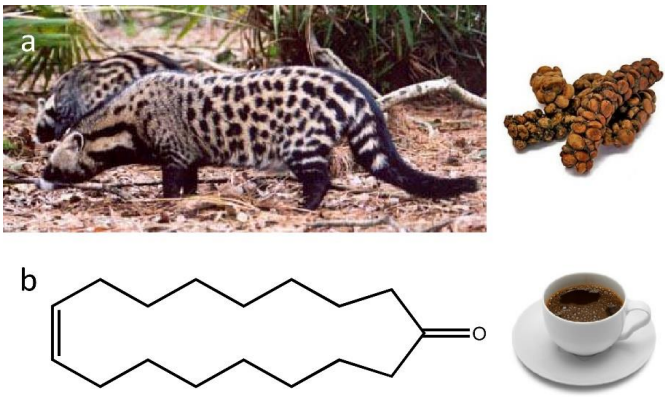


Figure 2. a) Civet cat: the excrement produced by this animal when eating coffee fruit will be the source of Kopi Luwak, **b)** Civetone: a molecule with a musky smell of animal origin.

Structure-odor relationships

The introduction of synthetic products as ingredients in perfumery – in particular with the famous aldehyde overdose of Chanel N° 5 – marks the beginning of the interest of the industry for these compounds. The rational design of molecules associated with previously established odors has become one of the fantasies of the perfume industry.

The efforts of chemists to establish such structure-odor relationships have mainly focused on olfactory families of interest for the field of perfumery (woody, green, floral or musky scents) (21, 22). And although there is no absolute rule, some chemical functions appear to be associated with a characteristic odor.

The ester function is known in the world of perfume for its fruity smell. Many former graduates of Terminale S recall their chemistry Practical Work on the synthesis of isoamyl acetate and its characteristic banana odor. In the same way, the organic chemist recognizes the

fruity odor of ethyl acetate, a solvent widely used in synthesis. But even if the presence of this functional group is a good indicator of the fruity smell of a molecule, its precise olfactory quality remains hard to define (coconut, pear or peach for example). In addition, less than half of the molecules associated with fruit odor do indeed carry an ester function. The other functions responsible for this odor belong to the family of ketones, aldehydes or lactones. (21) By striving to connect a structure to an odor, the perfume industry has been inspired by pharmacological approaches that compile the physical and chemical characteristics of molecules bearing the same olfactory note in order to deduce the most general possible rules. These "olfactophores" (from *olfacto* = "odor" and *phorós* = "carry" in ancient Greek) are models that gather the structural information of compounds belonging to the same olfactory family and that have the same odor. The characteristics most often encountered are steric and hydrophobic interactions, as well as specific polar, acidic or basic features (Figure 3).

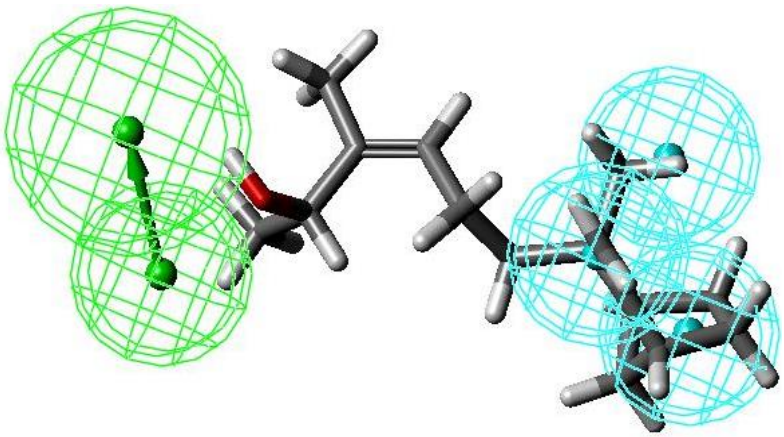


Figure 3. An olfactophore for sandalwood odor with optimal superposition of a characteristic compound carrying this odor. In the model, three hydrophobic features (sky blue) are separated in space by a hydrogen bonding donor group feature (green).

In general, the extreme subtlety of the links between structure and odor makes the rational design of odorant compound always delicate. Often even subtle changes (change of the position of a double bond, chirality or addition of a simple methyl group) can lead to changes in the quality and odor intensity of an odorant, or even the loss of its smell. (23) The two enantiomers of limonene are a typical example: these two molecules have distinct odors. The R-enantiomer has an orange odor whereas its mirror image, the S-enantiomer, has an odor similar to that of lemon (Figure 4). In many other cases, we are unable to tell the difference between enantiomers. Conversely, molecules of very different structures can belong to the same olfactory family. For example, in the case of the camphor-smelling molecules, no consensus of chemical functionality can be established (Figure 4).

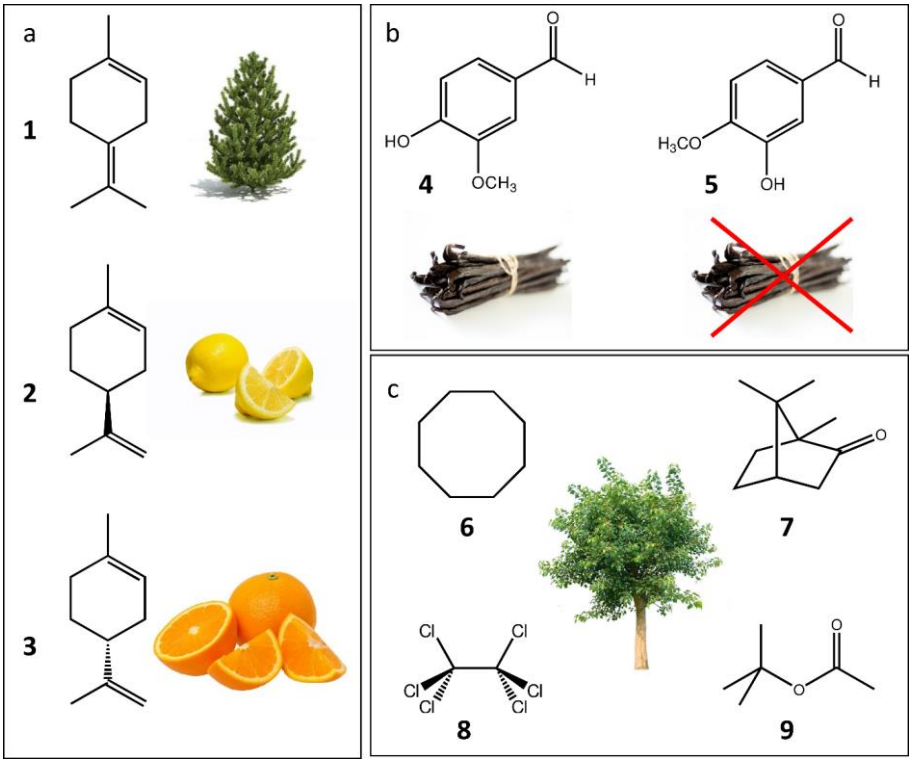


Figure 4. Similar molecular structures can have different olfactory characteristics, whereas different molecular structures can have a similar smell. **a)** Terpineol (1) has a pine odor, S-limonene (2) has an odor associated with that of lemon, while R-limonene (3) has an orange odor. **b)** Despite their structural similarity, isovanillin (5) does not have the same characteristic vanilla odor as vanillin (4) at all. **c)** Cyclooctane (6), camphor (7), hexachloroethane (8) and ter-butyl acetate (9) all have, despite their very diverse structures, a camphorous odor.

Ultimately, the lack of a direct link between chemical structure and odor is an obstacle to the rational design of odorant compounds. Currently, by reproducing a pharmacological approach, the discovery of new compounds of interest is based on the design of olfactophores.

However, because of the complexity of the olfactory system, it remains in many cases due to serendipity (discovery due to chance). This olfactory anarchy originates from the complexity of the mechanisms of our perception of odors already mentioned in the early 1980's by Maurice Chastrette (24). It seems obvious that the chemist will have to go beyond the limits of his discipline to establish a universal structure-odor relationship, forcing him to consider the biology of the olfactory system and the neural encoding of our perception of odors. The major stages of perception are briefly described in Figure 5.

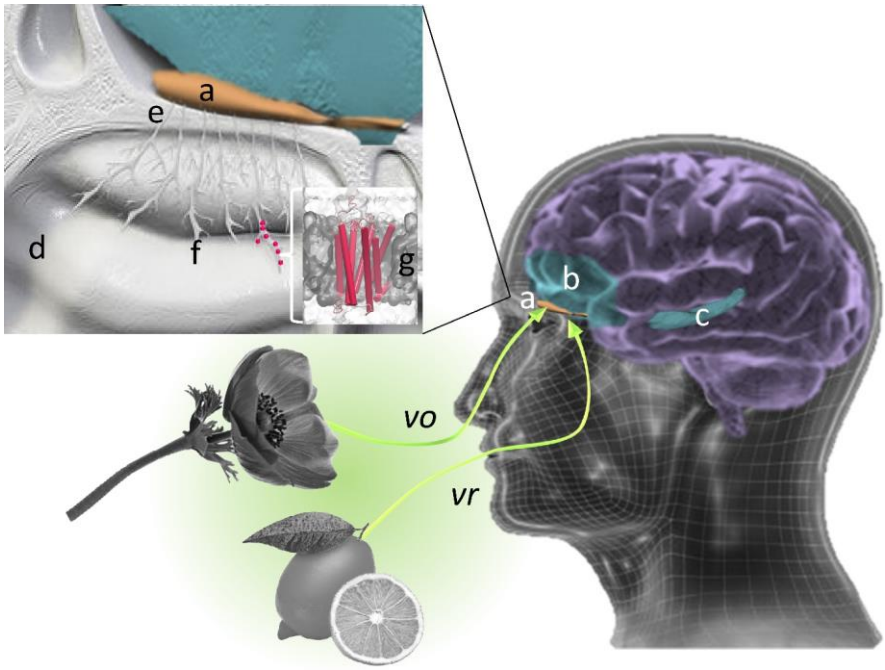


Figure 5. When we inhale (orthonasal route – vo) or ingest (retronasal pathway – vr) odorant molecules, part of our brain decodes their molecular message: it is our olfactory bulb (orange – a). The signal induced by these molecules is then transmitted to different zones such as the orbito-frontal cortex (blue – b) and the hippocampus (blue-c) which allow us to recognize its smell, to appreciate its valence and to trigger a possible emotional connection. But the olfactory bulb is in our cranial vault and is not really in direct contact with the odorous compounds. It is separated by the screened plate of the ethmoid bone (e). At the molecular level, odorous compounds are inhaled via the nasal cavity (d). The molecules then interact with our olfactory neurons (f), which express their olfactory receptors (g) in their membrane. The chemical message carried by the odorant molecules is transformed into a neuronal influx, which crosses the screened plate to be treated by our olfactory bulb.

Olfactory receptors, the cornerstones of odorant perception

In 1991, Linda Buck and Richard Axel highlight the existence of the gene family encoding Olfactory Receptors (ORs) (25). This discovery will be awarded by the 2004 Nobel Prize for Medicine. ORs belong to the superfamily of G Protein-Coupled Receptors (GPCRs) that

includes a large number of the proteins found in mammals. GPCRs play a vital role in the interaction of cells with their environments: they are associated with the recognition of a wide variety of extracellular stimuli such as amino acids, lipids, neurotransmitters, hormones and of course, scented molecules. Genes coding for olfactory receptors represent more than 4% of our genome and more than 3% of our proteome, making it the second gene family after those of the immune system.

In humans, there are about 800 olfactory receptor genes of which 396 are functional. In comparison, the dog has about 800 functional genes and the rat 1200. Our sense of smell might have lost its importance through evolution, thus explaining the ~400 non-functional olfactory receptor genes (a.k.a. pseudo-genes). There is a possible correlation between the acquisition of the trichromatic vision - more important for the survival of the species - and the degeneration of our olfactory receptor repertoire. Moreover, this degeneration does not seem to be about to stop because no plateau has yet been reached in the decrease of the number of functional genes.

During evolution, therefore, we may continue to lose our capacity to odor detect and discriminate odors (26). However, the relationship between the number of receptor genes and detection capacity is debated: our cognitive power being greater compared to that of the rat or dog, it could compensate our smaller repertoire of olfactory receptors (27).

When inhaled, a molecule comes into contact with the entire repertoire of receptors that we express in the olfactory epithelium. Depending on its physical and chemical properties, the molecule will differentially activate each of the ORs expressed in the neurons and it is this combinatorial code of activation of olfactory neurons that will be interpreted as an odor by our brain

Olfactory receptors are GPCRs

GPCRs (G Protein-Coupled Receptors) are key proteins in cellular communication and more generally in the perception of the environment. They allow us to see, smell, taste, make us shudder, activate our reward system, control our sleep, help with memory and are involved in many other mechanisms involving hormones and neurotransmitters. These receptors have a common three-dimensional structure composed of amino acids organized into seven helices connected by loops (Figure 6). These helices are embedded in the cell membrane. Each type of GPCR is more or less specific to a chemical space (β-adrenergic receptors mainly bind

catecholamines, the muscarinic receptor binds acetylcholine and olfactory receptors, or ORs, bind odor molecules).

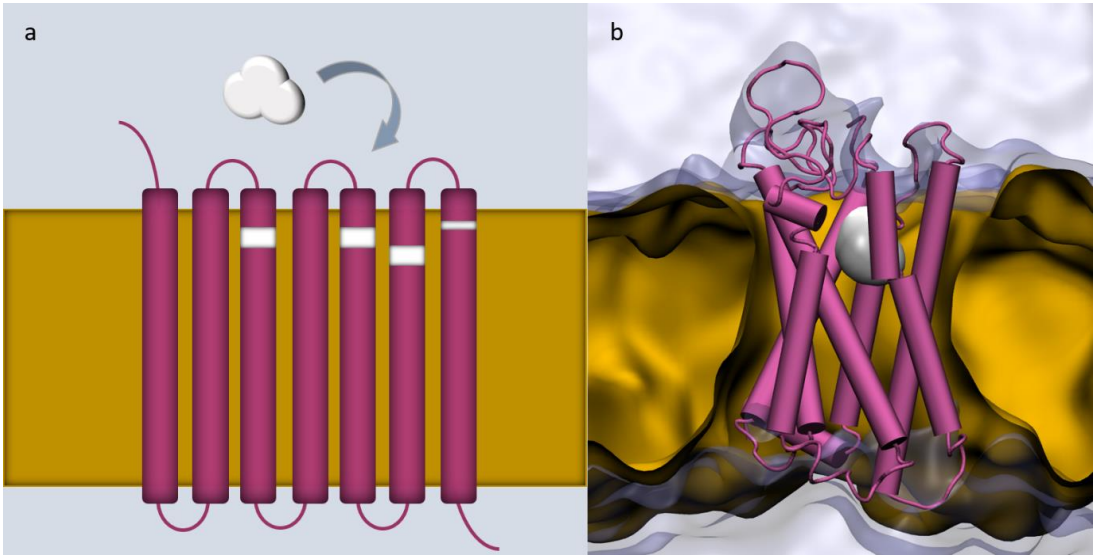


Figure 6. a) Schematic representation of an olfactory receptor. The blank areas represent the amino acids of the receptor that will be in contact with the ligand. **b)** Olfactory receptor in complex with an odorant. The GPCR is represented in purple, the membrane in yellow and the odorant in white.

In the case of olfactory receptors, the mechanism of receptor activation by the ligand is associated with a "molecular switch" located in the cradle of the binding site. Once an agonist is bound to the receptor cavity, the receptor becomes active through a conformational change. The conformational change of the receptor allows the binding to a protein called G protein. A cascade of biochemical reactions is then triggered and results in the opening of an ion channel leading to a calcium influx in the cell which is associated with a cell membrane depolarization. Finally, ORs transform the chemical message into a neuronal influx, which is then interpreted as an olfactory perception.

The combinatorial code of odor perception

Assuming that an OR can recognize several types of odorant molecules and that an odorant molecule can be recognized by several ORs (28), the combinatorial combination of these two partners (ORs and odorous) is virtually infinite. The commonly accepted hypothesis is that olfactory system decrypts odors through this "combinatorial code" which attributes to each molecule its own identity card of activated olfactory neurons (and thus of ORs). In principle, the perception of smells is analogous to the composition of a musical accord on a piano, except that it would have 396 keys, as shown schematically in Figure 7.

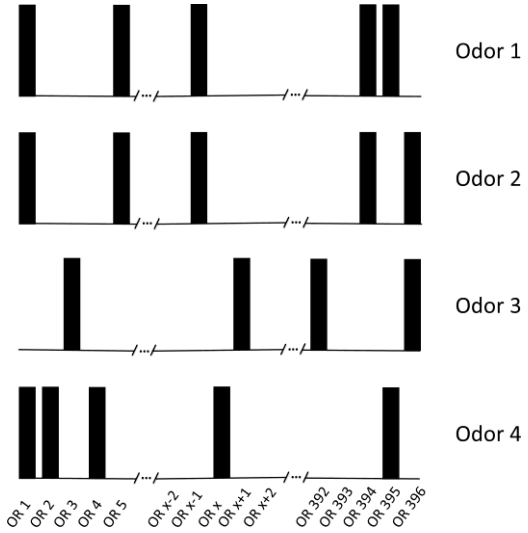


Figure 7. The differential activation of various ORs is at the origin of the combinatorial code of odors. The response of our repertoire of olfactory receptors is schematized by black bars. Depending on the structure of the odorant, the combinatorial activation code of our 396 receptors will be variable, thus justifying our ability to detect and discriminate an extraordinarily large number of volatile compounds.

Cracking this combinatorial code will in principle make it possible to attribute an odor to an odorant compound, based on the activation of our olfactory receptor repertoire. Understanding the mechanism of activation of these proteins becomes crucial in establishing structure-odor relationships. On this point, a debate has long raged between a classical pharmacological mechanism and the so-called "vibrational theory" (vide infra).

Peri-receptor phenomena also play an important role and add subtlety to an already extraordinarily complex system. These phenomena involve at least two types of biological protagonists found in our olfactory mucus. The first, odorant binding proteins (OBP), could help in the transportation of odorant molecules to the receptors. Second, degradation enzymes play a role of detoxification of the olfactory mucus. These enzymes chemically modify the compounds to eliminate them. However, some of these metabolites may still be able to

activate olfactory receptors. The combinatorial code would be even more complex since the nature of the chemical message would come, not from the pure molecule but from a bouquet formed by its metabolites. The role of these metabolites in perception has been tested in the presence of an inhibitor of these degradation enzymes. The presence of this blocker actually changed the perceived smell (29). This observation supports the hypothesis that the biotransformation of odorous molecules would have a significant impact on olfactory perception

Vibrational theory vs. pharmacological approach

The mechanisms leading to the perception of a smell have long been subject to theories and speculations. Two mechanisms of the recognition of molecules by ORs are historically most discussed: the vibrational theory which proposes an alternative to the more conventional pharmacological approach of the interaction between molecules and biological receptors.

The pharmacological approach, more widely accepted now, is based on the discovery describing the enzyme-substrate interaction. This is based on molecular complementarity of the "lock and key" type between the ligand and its receptor. In the 1960s, this concept was popularized by Amoore who developed his stereochemical theory. This theory postulates that the smell of a molecule is determined by its structure and more particularly by its shape and size (the reader will refer to article (8) for further information on this topic). This theory then evolved to integrate the existence of new parameters such as functional groups, hydrophobicity, polarity, acidity and basicity of the molecules, ultimately leading to the birth of olfactophores previously discussed (Figure 3).

Alternatively, the idea that the sense of smell works on the same principle as vision or hearing, defined as "spectral" senses as opposed to "chemical senses", is a notion that has hit the headlines. According to vibrational theory, often revisited, olfactory receptors mainly recognize the vibrations of odorant molecules and not their structure.

The most recent version attributes the activation of a receptor to the transfer of an electron through the OR via the odorant through an electron tunneling mechanism (30).

This theory was tested by comparing the odor of two isotopes, acetophenone and its deuterated analogue (which have different vibration spectra). In this case, the difference in olfactory quality of the two molecules was described as obvious. Later, these results were refuted by a test showing that humans did not have the ability to distinguish the odors of these two molecules. Nevertheless, other living organisms, like flies or bees would be able to

differentiate these two compounds. After lengthy debates and exchanges of articles supporting or refuting this theory, it seems that the debate is now settled. A trans-disciplinary approach combining molecular biology, organic chemistry and theoretical chemistry concluded on the "highly improbable" character of vibrational theory (31, 32).

It is therefore the conventional approach that seems to be adopted, involving the modulation of the structure of the receptor during its interaction with the odor molecule. To date, less than 30 protein structures belonging to the large family of GPCRs have been elucidated, but no olfactory receptors are found among these experimental structures. Nevertheless, the construction of a theoretical three-dimensional structure is possible thanks to molecular modeling. It consists of taking advantage of the known structures of GPCRs to deduce the nature of our ORs and their mechanisms of interactions with odorants (33, 34). This type of approach, at the frontier between computational chemistry and structural bio-informatics, can be considered as a computational microscope, in our case focused on our olfactory receptors (35, 36). By applying the laws of physics to all the atoms of an odorant - olfactory receptor complex, it is possible to describe and observe the interactions at the atomic level that allow our olfactory neurons to decode the chemical message carried by an odorant molecule (Figure 8). One can then imagine the development of a biologically inspired virtual nose that would redefine the concept of structure-odor relationships by explicitly considering the biological protagonists of odor perception.

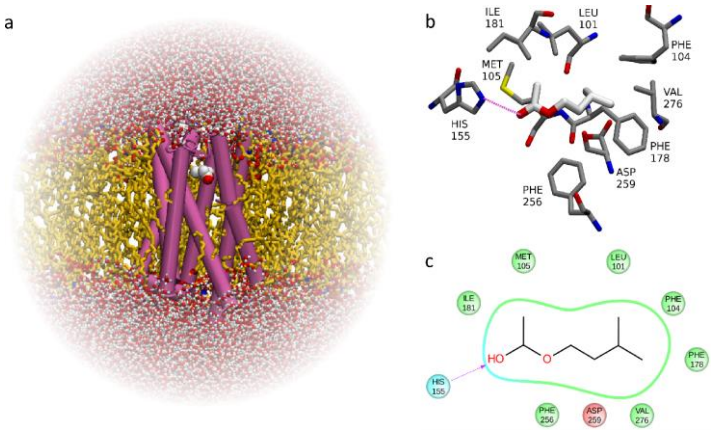


Figure 8. Computational microscope pointed at the olfactory receptors. **a)** In order to simulate realistic physiological conditions, the olfactory receptor (violet) containing an odorant (white) is inserted into a lipid membrane (yellow). The intra- and extracellular media are modeled by water molecules. **b)** Focusing on the receptor binding cavity. In this example, isoamyl acetate (with its carbon atoms in white and oxygen in red) is in contact with ten amino acids (carbon atom in gray, nitrogen in blue and oxygen in red). The odorant-receptor interactions can be of the hydrogen bond type (dotted lines) but are mostly non-polar. **c)** Scheme of the ligand-receptor interaction. In this representation of the active site, isoamyl acetate is placed in the center. The complementarity between the receiver cavity and the odorant allows the decoding of its chemical structure.

Structure - odor relationships in the post - genomic era

Since sequencing the complete human genome, the establishment of correlations between pathological or behavioral characteristics and the expression of one or more genes is possible. In this perspective, some intriguing relationships between our perception of pure compounds and the expression of olfactory receptor genes have been established. Functional genes encoding olfactory receptors are quite variable between individuals. While some genes are crucial for our survival, others related to the perception of odors can be modified over generations without affecting the viability of an individual. Specific mutations have been associated with differences in olfactory perception that can affect our behavior (Figure 9, ref (37) and internal references).

The so-called “green” odor (reminiscent of freshly cut grass) is typically associated with cis-3-hexenol. We do not all equally perceive this molecule. Its detection threshold (concentration at which an individual perceives its odor) is correlated with a slight variation in the sequence of the olfactory receptor 2J3 gene. A change in two amino acids on the 300 that make up our 2J3 receptor is enough to cause a modulation of the detection threshold of the individual carrying this mutation (38).

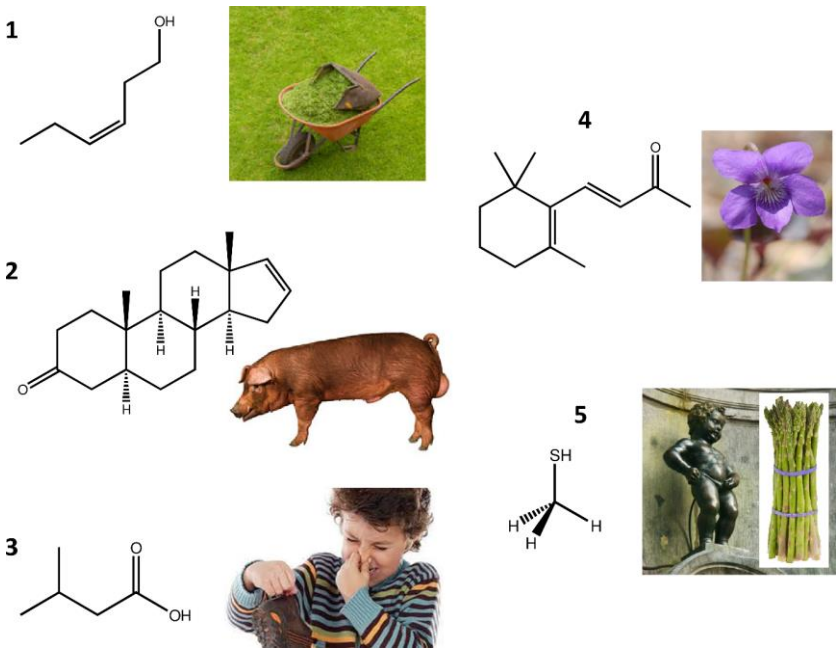


Figure 9. Molecules for which a chemo-genomic link has been established. 1: cis-3-hexenol, 2: androstenone, 3: isovaleric acid, 4: β-ionone, 5: mercaptan.

The perception of androstenone is even more documented. About ~ 25% of the population is almost anosmic to this compound. This odorant is incidentally the sexual pheromone of the pig, but it is also secreted by men in their axillary sweat. Although the majority of individuals

who are able to detect it describe it as having an “animal” and “urinous” type of smell, others, expressing a mutation affecting only two amino acids of their 7D4 receptor sequence (two out of 300 amino acids), describe the smell as being pleasant, reminiscent of honey and vanilla. Finally, mutants on a single amino acid position in the sequence of this receptor, have a much lower detection threshold than the average smeller, making them “super detectors” of androstenone.

These genetic differences have a direct consequence on the appreciation of meat from castrated (with a low amount of androstenone) or non-castrated (with a higher amount of androstenone) pig. Non-castrated pork has a much more potent and animalic taste than its castrated pork analog with low androstenone content. A correlation can even be established between the proportion of anosmic people to androstenone in France, Spain and Britain and the percentage of castrated pigs in these countries. This example clearly illustrates how our olfactory genome and our eating habits – as well as breeding methods – are intertwined.

To cite other examples, the differential sensitivity to isovaleric acid (described as having a body odor smell, of feet) among various populations could be partly due to a polymorphism of the gene expressing the 11H7P receptor. Detection of β-ionone (a floral and woody odor molecule present in both perfumes and food) would also be affected by a polymorphism in the gene encoding the 5A1 receptor.

A statistical analysis on about 30,000 genomes of people whose hedonic perception of coriander was documented established that the appreciation of dishes containing coriander is correlated with a sequence variation in an olfactory receptor gene (6A2). This mutation would affect the sensitivity to aldehydes (soap scents) present in the leaves, but this relationship remains to be clearly established.

The perception of the characteristic smell of urine after ingesting asparagus, related in particular to the presence of methanethiol, has been partially correlated with variations in 7M2 and 14C36 receptor genes in Caucasians (but not in African populations).

The examples are few in number, but emphasize interindividual differences, showing that the cultural effect is not the only one responsible for behavioral variations.

Olfactory receptors in other places than in the olfactory epithelium!

The genes that code for our olfactory receptors represent a non-negligible fraction of our genome (4%). These receptors play a much larger role in our bodies since they are not exclusively expressed in our olfactory epithelium. Recent research establishes the presence of

“olfactory” receptors in tissues unrelated to the perception of odors. Therefore, we now start to prefer the term “odorant receptors” to “olfactory” because the latter suggests a role exclusively related to the perception of odors.

Let’s consider the recently identified example of the odorant receptor in the trachea. This receptor specifically responds to the conjugate base of lactic acid that is produced during a lack of oxygenation. The increase of lactate ions during the lack of oxygen triggers an increase in the respiratory rate. This crucial phenomenon is controlled by a single olfactory receptor that “senses” the lactate ion (39). This receptor is also expressed in the kidneys and responds to short chain fatty acids. It thus controls the secretion of renin which is at the origin of the increase in blood volume and thirst (40).

OR51E2 receptor, also known as the Prostate Specific G protein-coupled receptor (PSGR), is overexpressed in prostate cancer cells. The presence of one of its agonists on tumors, β-ionone, with its floral and woody odor, activates cell proliferation (41).

Olfactory receptors are therefore involved in a multitude of living mechanisms whose list is only beginning to be revealed through genomic and proteomic studies.

The research efforts that are being made to understand the mechanisms of odor perception will open up opportunities in disciplinary fields that range from the chemistry of flavors and fragrances to biomedical sciences and neuroscience. This area of research is definitely a great showcase to demonstrate the strength of the synergy of interdisciplinary studies.

Conclusion

Odors, long misunderstood or despised, now fascinate and intrigue. Smell allows us to detect the molecules present in our environment which are of very variable nature. We are thus able to detect a danger or, on the contrary, to identify a beneficial source thanks to our sense of smell. Like all living beings, humans developed a strategy to discern and even discriminate the 27 billion potentially odorous molecules in the universe (42). This function is provided by our millions of olfactory neurons that selectively express each of the 396 types of olfactory receptors. The complexity that leads to the systems’ formidable accuracy highlights the difficulty of establishing a simple metric to link a chemical structure to an olfactory percept.

The establishment of a universal structure – odor relationship would represent a dramatic breakthrough from a fundamental point of view and would also be of great interest to companies in the perfume industry. The chemists of aromas and perfumes dream of being able to guide their research of new odorant molecules with targeted olfactory qualities.

Although this step is theoretically feasible, it will need to describe the steps that connect a molecule to the code associated with its odor. Research on the molecular mechanisms of odor perception is becoming more active and looks promising. The process undertaken to understand these mechanisms will also be of fundamental interest in the chemistry of drug design as these “olfactory” receptors prove to regulate physiological or pathological mechanisms in organs that have nothing to do with olfaction.

References

1. J. P. Brun, X. Fernandez, **Parfums antiques - De l'Archéologue au Chimiste**, Cinisello Balsamo, Milano, IT (2015).
2. C. Classen, D. Howes, A. Synnott, **Aroma - The Cultural History of Smell,** London, UK (1994).
3. A. Le Guérer, La réhabilitation de l’odorat. Le pouvoir des odeurs. In: R. Salesse, R. Gervais, Eds., **Odorat et Goût - De la neurobiologie des sens chimique aux applications,** 11-14 (2012).
4. A. Keller, D. Malaspina, Hidden consequences of olfactory dysfunction: a patient report series. BMC Ear Nose Throat Disord. 13(1):8 (2013).
5. I. Croy, S. Nordin, T. Hummel, Olfactory Disorders and Quality of Life—An Updated Review. Chem. Senses 39, 185-194 (2014).
6. R. J. Stevenson, An initial evaluation of the functions of human olfaction. Chem. Senses 35, 3-20 (2010).
7. S. Richardson, **What is Postgenomics?** In: 4S Annual Meeting Crowne Plaza Cleveland City Center Hotel, Clevelant, OH, 25-11-2014.
8. U. J. Meierhenrich, J. Golebiowski, X. Fernandez, D. Cabrol-Bass, De la molécule à l’odeur. Les bases moléculaires des premières étapes de l'olfaction. L'Actualité Chimique 289, 29-40 (2005).
9. C. Sezille, A. Fournel, C. Rouby, F. Rinck, M. Bensafi, Hedonic appreciation and verbal description of pleasant and unpleasant odors in untrained, trainee cooks, flavorists and perfumers. Front. Psychol. 5:12, (2014).
10. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. Science 343, 1370-1372 (2014).
11. T. Weiss, K. Snitz, A. Yablonka, R. M. Khan, D. Gafsou, E. Schneidman, N. Sobel, Perceptual convergence of multi-component mixtures in olfaction implies an olfactory white. Proc. Natl. Acad. Sci. U. S. A. 109, 19959-19964 (2012).
12. S. Nicklaus, S. Issanchou, Gustation, olfaction et préférences alimentaires chez l’enfant. In: R. Salesse, R. Gervais, Eds., **Odorat et Goût - De la neurobiologie des sens chimique aux applications**, 359-371 (2012).
13. B. Schaal, L. Marlier, R. Soussignan, Human foetuses learn odours from their pregnant mother's diet. Chem. Senses 25, 729-737 (2000).
14. F. Kermen, A. Chakirian, C. Sezille, P. Joussain, G. Le Goff, A. Ziessel, M. Chastrette, N. Mandairon, A. Didier, C. Rouby, M. Bensafi, Molecular complexity determines the number of olfactory notes and the pleasantness of smells. Sci. Rep. 1, 206 (2011).
15. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820-826 (2017).
16. E. A. Krusemark, L. R. Novak, D. R. Gitelman, W. Li, When the sense of smell meets emotion: anxiety-state-dependent olfactory processing and neural circuitry adaptation. J. Neurosci. 33, 15324-15332 (2013).

17. Y. Sugawara, A. Shigetho, M. Yoneda, T. Tuchiya, T. Matumura, M. Hirano, Relationship between mood change, odour and its physiological effects in humans while inhaling the fragrances of essential oils as well as linalool and its enantiomers. Molecules 18, 3312-3338 (2013).
18. Y. Nikaido, S. Miyata, T. Nakashima, Mixture of cis-3-hexenol and trans-2-hexenal attenuates behavioral and stress responses induced by 2,5-dihydro-2,4,5-trimethylthiazoline and electric footshock stress in rats. Physiol. Behav. 103, 547-556 (2011).
19. K. D. Perring, Volatility and Substantivity. In: C.S. Sell, Ed., **The Chemistry of Fragrances - From Perfumer to Consumer**, 199-213 (2006).
20. C. S. Sell, Ingredients for the Modern Perfumery Industry. In: C.S. Sell, Ed., **The Chemistry of Fragrances - From Perfumer to Consumer**, 52-131 (2006).
21. K. J. Rossiter, Structure-Odor Relationships. Chem. Rev. 96, 3201-3240 (1996).
22. P. Kraft, J. A. Bajgrowicz, C. Denis, G. Frater, Odds and trends: Recent developments in the chemistry of odorants. Angew. Chem. Int. Ed. 39, 2981-3010 (2000).
23. C. S. Sell, On the unpredictability of odor. Angew. Chem. Int. Ed. 45, 6254-6261 (2006).
24. M. Chastrette, D. Zakarya, in The Human Sense of Smell, D. G. Laing, R. L. Doty, W. Breipohl, Eds. (Springer-Verlag, Berlin Heidelberg, 1991), chap. 4, pp. 77-95.
25. L. Buck, R. Axel, A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175-187 (1991).
26. D. Pierron, N. G. Cortes, T. Letellier, L. I. Grossman, Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. Mol. Phylogenet. Evol. 66, 558-564 (2013).
27. S. Robin, P. Quignon, F. Galibert, Génétique et évolution des récepteurs olfactifs chez les Vertébrés. In: R. Salesse, R. Gervais, Eds., **Odorat et Goût - De la neurobiologie des sens chimiques aux applications**, 291-309 (2012).
28. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).
29. B. Schilling, R. Kaiser, A. Natsch, M. Gautschi, Investigation of odors in the fragrance industry. Chemoecology 20, 135-147 (2010).
30. L. Turin, A spectroscopic mechanism for primary olfactory reception. Chem. Senses 21, 773-791 (1996).
31. E. Block, S. Jang, H. Matsunami, S. Sekharan, B. Dethier, M. Z. Ertem, S. Gundala, Y. Pan, S. Li, Z. Li, S. N. Lodge, M. Ozbil, H. Jiang, S. F. Penalba, V. S. Batista, H. Zhuang, Implausibility of the vibrational theory of olfaction. Proc. Natl. Acad. Sci. U. S. A. 112, E2766-2774 (2015).
32. L. B. Vosshall, Laying a controversial smell theory to rest. Proc. Natl. Acad. Sci. U. S. A. 112, 6525-6526 (2015).
33. C. A. de March, Y. Yu, M. J. Ni, K. A. Adipietro, H. Matsunami, M. Ma, J. Golebiowski, Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. J. Am. Chem. Soc. 137, 8611-8616 (2015).
34. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U. S. A. 112, 14966-14971 (2015).
35. C. A. de March, J. Golebiowski, A computational microscope focused on the sense of smell. Biochimie 107 Pt A, 3-10 (2014).
36. J. Topin, C. A. de March, L. Charlier, C. Ronin, S. Antonczak, J. Golebiowski, Discrimination between olfactory receptor agonists and non-agonists. Chemistry 20, 10227-10230 (2014).
37. C. A. de March, S. Ryu, G. Sicard, C. Moon, J. Golebiowski, Structure-odour relationships reviewed in the postgenomic era. Flavour Fragr. J. 30, 342-361 (2015).
38. J. F. McRae, J. D. Mainland, S. R. Jaeger, K. A. Adipietro, H. Matsunami, R. D. Newcomb, Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, cis-3-hexen-1-ol. Chem. Senses 37, 585-593 (2012).
39. A. J. Chang, F. E. Ortega, J. Riegler, D. V. Madison, M. A. Krasnow, Oxygen regulation of breathing through an olfactory receptor activated by lactate. Nature 527, 240-244 (2015).
40. J. L. Pluznick, R. J. Protzko, H. Gevorgyan, Z. Peterlin, A. Sipos, J. Han, I. Brunet, L. X. Wan, F. Rey, T. Wang, S. J. Firestein, M. Yanagisawa, J. I. Gordon, A. Eichmann, J. Peti-Peterdi, M.

J. Caplan, Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation. Proc. Natl. Acad. Sci. U. S. A. 110, 4410-4415 (2013).
41. G. Sanz, I. Leray, A. Dewaele, J. Sobilo, S. Lerondel, S. Bouet, D. Grebert, R. Monnerie, E. Pajot-Augy, L. M. Mir, Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. PLoS One 9, e85110 (2014).
42. C. W. Yu, K. A. Prokop-Prigge, L. A. Warrenburg, J. D. Mainland, Drawing the Borders of Olfactory Space. Chem. Senses 40, 565-565 (2015).

Part II
Mapping odorant and receptor spaces using machine learning

"All models are wrong, but some are useful." – George E.P. Box

Machine learning to study smell

Odorants form a multidimensional stimulus space which is challenging to understand. In humans, about 400 receptors endow us with an extraordinary discriminative power. But how can we match a receptors' response to the properties of odorant molecules? **In other words, is the ligand-driven response of a G protein coupled odorant receptor a chemical property?**

Matching an odorant to an odorant receptor (OR) is a disconcerting task. In this context, high throughput screening – even through a robust mean to probe how the biological system weights given chemical relationships – remains time expensive and costly. **Virtual screening can thus be considered as a suitable alternative** to help exploring the chemical space associated with an OR and eventually, expanding it.

Virtual screening is the process of sifting through a compound database for molecules which will be submitted to experimental testing for a given property. These types of methods are designed to computationally screen large datasets, and to select a smaller subset of chemical entities with sought-after properties.

Virtual screening applied to olfaction: state-of-the art

Molecular modeling: pharmacophores and docking

From a chemists' perspective, the activity of a compound is related to its physical and chemical properties. A method widely used in the pharmaceutical industry consists in building pharmacophores. Pharmacophores are the compilation of chemicals known to act on a therapeutic target. The structures of ligands are superimposed and allow identifying the chemical features necessary at certain position in space to trigger a given pharmacological activity. These can be for example hydrophobic features, steric hindrances, or presence of polar atoms at certain key positions. This concept relies on the assumption that molecules having similar structures and physicochemical properties will have similar effects. In olfaction this technique was applied to build a pharmacophore for human OR1G1(1) using *in vitro* data from a previous study (2).

An alternative to computationally perform deorphanization is to build three-dimensional models of ORs. Here, docking of candidate ligands is performed. The benchmark used is a

function modeling the free energy of interaction. This method has proven successful in the identification of novel ligands for mouse OR42-3 (3) and human OR51E2 (4).

Both of these approaches rely on the "fitting" the ligand into i/ a pharmacophore or ii/ into a receptor cavity and estimating the cost of this fitting. They are first principle techniques as they rely on optimizing the energy of interaction, and eventually on conformational sampling. The strength of these techniques is that they allow a better understanding of the structural complementarity between agonists and their biological target. They can however be time consuming and rely on the quality of the model of the receptor, or on the conformational sampling in the case of the pharmacophore approach.

Data mining

Virtual screening techniques also include alternative compound classification methods such as clustering and partitioning, or machine learning approaches. Here, the goal is to predict compound class labels (i.e. active vs. non-active – using classification algorithms) or a specific numerical quantity (such as EC50 – using regression algorithms) on the basis of models which use certain types of descriptors as input.

In insect olfaction, such techniques were already applied on Drosophila. In 2013, Boyle *et al.* successfully identified OR-odorant interactions (5) by using the 'similarity property principle' (where it is reasoned that similar structures will elicit similar properties). And in the same species, Schmucker *et al.* (6) used artificial neural networks to predict OR neuron responses upon chemical stimulation.

Machine learning to study olfaction: predicting OR activation in mammals

In the midst of the fourth industrial revolution, there is much excitement about the potential of artificial intelligence (AI) to further pharmaceutical research. The fourth industrial revolution is characterized by its velocity and system impact, all driven by AI (7). The advantage of having an intelligent model is mainly that it demonstrates an ability to solve problems, it can learn from experience, and deal with new situations. As machine learning is a particular approach to AI, it is easy to access a "virtuous circle". Here, the first model which is created using an initial data set, can help obtaining new data which will in term be used to refine the model in an iterative process. (Fig.1) The advantage of doing this, is that the model is expected to be improved with each iteration.

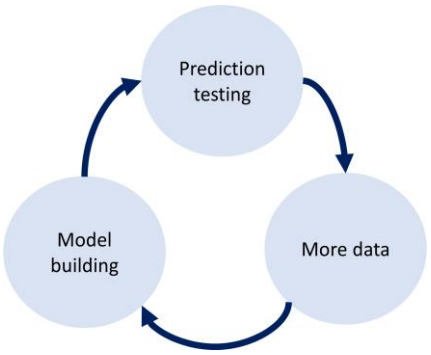


Figure 1: Virtuous circle in machine learning. First a model is built using available data, then predictions of the model are tested. Results of these prediction allow to gain access to more data which is in turn used to improve the initial model.

Protocol

During my PhD I set up a protocol able to predict mammalian OR response *in vitro* (humans and mice) using machine learning. This protocol should help exploring the chemical space related to a given OR, and eventually expanding it.

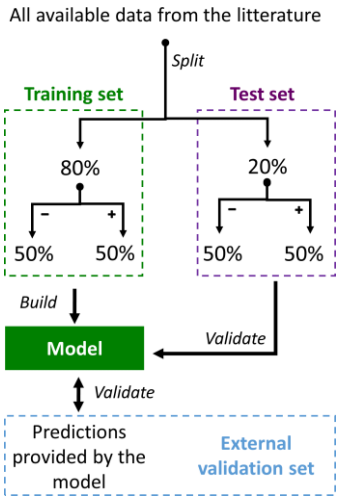


Figure 2: Workflow to build a machine learning model. The available data is split into a training and a test set in proportions usually ranging from 60:40 to 80:20. Ideally, each set contains a balanced number of molecules with either an enhancing activity (or positive effect "+") or a decreasing activity (negative effect "-"). The data in the training set is used to build a model while the data in the test set validates it. Once the model accurately predicts the test set, predictions are made and tested on an external validation set, which are then tested *in vitro*.

Briefly, all available data is gathered from the literature and labeled in a binary fashion as: activating (1) or not activating (0). Molecular descriptors are calculated for each molecule using Dragon software (8). Upon calculation of chemical descriptors, the initial dataset is split into a training and a test set by ensuring a balance in the number of agonists and non-agonists to avoid overfitting the model (Fig. 2). Several models, which tentatively associate the

receptor response with the chemical properties of the input ligand are built iteratively until an acceptable performance is obtained. Then, a library of compounds, having similar properties to the ones of the molecules used in the training set is screened on the model. Finally, the predictions are tested *in vitro* by our partners to assess the accuracy and allow discovering new ligands. They provide a feedback on the performance of our model.

The limits of machine learning

Chemical similarity and applicability domain

Since the QSAR model was developed using a certain set of compounds, it can only make prediction on compounds which are similar to the ones it was developed on. This is the so-called "applicability domain" of the model. Compounds which are too different from the ones in the training set cannot be expected to be accurately predicted. Accordingly, as the model is fitted exclusively on a certain chemical space – bearing specific physico-chemical and structural characteristics – the rest of it is largely left unexplored. Consequently, this hinders the possibility of discoveries of hits outside the known chemical space.

Need for consistent and high confidence data

A major problem when feeding the model is associated with discrepancies between studies coming from various data in the literature. Results can differ as a function of the functional assay that is used (Xenopus eggs, HeLa, HEK or HANA cells), and even EC50 values can vary between studies. As machine learning is required to predict certain characteristic, it can be biased if the data is not both consistent and of high quality. In practical terms, caution was necessary to identify an agonist or a non-agonist. If a molecule was considered to be an agonist by one study but was regarded to have an opposite effect in another one, the concentration at which the odorant was tested was first assessed. Then, if discrepancies remained between studies, the molecule was not used in the process of model building or testing. Consistency of data should be less problematic when a database is provided by only one group or person using a given method.

Furthermore, machine learning relies on the existence of data. It is therefore not a suitable technique for deorphanization of a receptor. Molecular modeling techniques such as docking or receptor-based pharmacophores should not be over-looked as alternatives. Given that ~80% of human ORs remain orphan, a combination of machine learning and molecular modeling techniques could be considered.

Contribution

In this article, the step-by-step procedure of the machine learning protocol I developed is described. Briefly, the protocol consisted in merging a linear classification algorithm (using Support Vector Machine) with a similarity measure relying on molecular fingerprints. This step is important because it helps optimizing the use of the model to a restricted set of data which represent the applicability domain of the model. Here, I developed and optimized the workflow which can be used in KNIME for creating a QSAR model as well as the protocol for comparing molecular similarity. The *in vitro* section of this protocol article was written by our collaborators in Duke University.

References

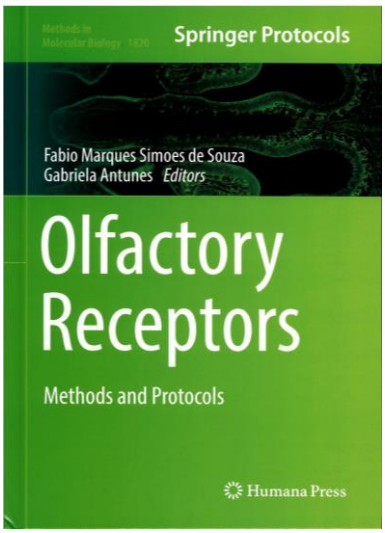
1. G. Sanz, T. Thomas-Danguin, H. Hamdani el, C. Le Poupon, L. Briand, J. C. Pernollet, E. Guichard, A. Tromelin, Relationships between molecular structure and perceived odor quality of ligands for a human olfactory receptor. Chem. Senses 33, 639-653 (2008).
2. G. Sanz, C. Schlegel, J. C. Pernollet, L. Briand, Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism. Chem. Senses 30, 69-80 (2005).
3. S. Bavan, B. Sherman, C. W. Luetje, T. Abaffy, Discovery of novel ligands for mouse olfactory receptor MOR42-3 using an in silico screening approach and in vitro validation. PloS One 9, e92064 (2014).
4. T. Abaffy, J. R. Bain, M. J. Muehlbauer, I. Spasojevic, S. Lodha, E. Bruguera, S. K. O'Neal, S. Y. Kim, H. Matsunami, A Testosterone Metabolite 19-Hydroxyandrostenedione Induces Neuroendocrine Trans-Differentiation of Prostate Cancer Cells via an Ectopic Olfactory Receptor. Front. Ocol. 8, 162 (2018).
5. S. M. Boyle, S. McInally, A. Ray, Expanding the olfactory code by in silico decoding of odor-receptor chemical space. eLife 2, e01120 (2013).
6. M. Schmuker, M. de Bruyne, M. Hahnel, G. Schneider, Predicting olfactory receptor neuron responses from odorant structure. Chem. Cent. J. 4, 1-11 (2007).
7. K. Schwab, **The Fourth Industrial Revolution**, World Economic Forum (2016).
8. TALETE srl. Dragon *Software for Molecular Descriptor Calculation* (2014).

Publication 2

Numerical models and in vitro assays to study Odorant Receptors

CAROLINE BUSHDID, CLAIRE A. DE MARCH, HIROAKI MATSUNAMI, JÉRÔME GOLEBIOWSKI

IN METHODS IN MOLECULAR BIOLOGY 1820, 77-93, **2018**



Abstract

Unravelling the sense of smell relies on understanding how odorant receptors recognize odorant molecules. Given the vastness of the odorant chemical space and the complexity of the odorant receptor space, computational methods are in line to propose rules connecting them. We hereby propose an *in silico* and an *in vitro* approach, which, when combined are extremely useful for assessing chemo-genomic links. In this chapter we mostly focus on the mining of already existing data through machine learning methods. This approach allows establishing predictions that map the chemical space and the receptor space. Then, we describe the method for assessing the activation of odorant receptors and their mutants through luciferase reporter gene functional assays.

1. Introduction

Odorant receptors (ORs) belong to the G-protein coupled receptor (GPCR) family that is the largest family of cell surface seven transmembrane receptors. This type of receptors mediates many important physiological functions and in the case of olfaction, OR function mostly results from the interaction of an OR with an odorant.

The human genome contains approximately 800 genes coding for ORs, of which only about 400 are intact. It is commonly accepted that the code of olfactory perception is a combinatorial one: each OR gets differentially activated when interacting with an odorant [1]. The signal resulting from this interaction is the first of a signaling cascade that will transform a chemical message into a series of neural impulses. This message will result in an olfactory perception. However, the role of ORs extends far beyond the field of olfaction, as ORs are ectopically expressed in many non-olfactory tissues, where they could play an important role.

In order to understand the role of ORs in olfaction, functional data that pairs odorants and odorant receptors is required. Despite enormous efforts in the community to deorphanize ORs *in vitro* [2-5], such methods have only seen limited success: indeed, less than 20% of human odorant receptors have been paired to at least one known ligand, and the rest of them remain orphan.

We hereby describe two numerical approaches, which can be synergized with *in vitro* experiments to have reliable assessments.

The first one is ligand based and useful for expanding the known agonists' chemical space (i.e. the type of chemicals) of an OR. In this case a computational model will learn the

physico-chemical properties necessary to activate a given receptor. A library of interesting molecules can then be screened and later identified hits can be tested *in vitro*. The second can be used to identify new odorant – OR pairs by using 3D modeling and docking. Here 3D homology modeling is performed because no experimental structures of ORs are available to date. Docking experiments provide insights to the affinity of ligands for the binding cavity and results obtained in this manner can also be confirmed by site directed mutagenesis prior to further analysis [6,7]. Alternatively, machine learning can be typically used to extend the chemical space associated with a given receptor [8].

2. Materials

2.1 Equipment

1. Computer: high-performance computer for the computational modeling approach.
2. Disk space for installing and storing the required files and software.
3. *In vitro* approach equipment: Silica gel columns (Qiagen minelute PCR purification kit), Thermal cycler (Biometra TProfessional TRIO Thermocycler), luminometer (BMG Labtech POLARstar OPTIMA), cell culturing fumehood and incubator.

2.2 Software List

1. KNIME (software for data analysis)
2. The Human OlfactoRy Data Explorer HORDE (webserver to download protein sequences)
3. Protein BLAST (webserver to find regions of similarity between biological sequences)
4. Jalview (freeware for sequence alignment)
5. Modeller (freeware for homology or comparative modeling of 3D protein structures)
6. VMD, Chimera or Pymol (3D visualization softwares)
7. PROPKA or H++ (softwares for optimal protonation)
8. AutoDock Vina (docking software)
9. Dragon (software for the calculation of chemical descriptors)

2.3 Data files for the computational model

1. Assemble all available experimental data regarding an odorant receptor and its cognate ligands.
2. 3D conformers (in SDF format) of the ligands are gathered using the Pubchem website.
3. Calculate the Molecular descriptors using Dragon [9].

4. The SDF 3D conformers of the database are screened and their related Dragon descriptors are calculated in the same manner.

2.4 Quantitative Structure Activity Relationships

1. Install the KNIME software (https://www.knime.com/) [10].

2.5 Sequence alignment and comparison:

1. OR sequences are downloaded using The Human OlfactoRy Data Explorer HORDE website (https://genome.weizmann.ac.il/horde/)
2. Use Protein BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) to find a class A GPCR with a high similarity to the query sequence for which the structure has already been experimentally resolved.
3. Use an alignment software such as Jalview (File>input alignment). Here the Mafft (Web services>alignment>Mafft with defaults) program is used to do a first alignment which should be further tuned by manual inspection (see Sequence alignment).

2.6 3D structure building and refinement:

1. Install Modeller (https://salilab.org/modeller/download_installation.html). This software is used for homology or comparative modeling of three-dimensional protein structures [11].
2. Preferred 3D visualization software (VMD, Chimera or Pymol).
3. A protonation webserver (H++, PROPKA).

2.7 Ligand Docking:

1. Install AutoDock Vina docking software.
2. Files containing the 3D structure of the odorants are in pdb format.

2.8. Reagents:

1. 5x Phusion HF buffer: F-518 5x Phusion HF reaction Buffer from Thermo Fisher Scientific.
2. 2mM dNTP: dilution of dNTP at 10mM (PCR Nucleotide Mix, REF: 04638956001) from Sigma Aldrich
3. Phusion pol : F-549 Phusion Hot Start II DNA polymerase 2U/µl from Thermo Fisher Scientific
4. 5µM Primer forward: from IDT

5. 5µM Primer reverse: from IDT

6. 1ng/µl Template plasmid: dilution of 100ng/µl plasmid from H. Matsunami lab

7. 5µM 5' primer: design CTC CAC AGG TGT CCA CTC from IDT

8. 5µM 3' primer: design CAC TGC ATT CTA GTT GTG from IDT

9. Buffer 3: B7003S NEBuffer 3 from New England Biolabs

10. MluI: R0198S 10,000 U/ml from New England Biolabs

11. NotI: R0189L 10,000 U/ml from New England Biolabs

12. BSA: B9001S Purified BSA 100x 10mg/ml from New England Biolabs

13. dW: 15230-147 Distilled Water from Gibco

14. Rho-pCI vector: pCI Mammalian expression vector from Promega with Rho tag inserted between NheI and EcoRI sites.

15. T4 ligase: M0202L T4 DNA ligase 400,000 U/ml from New England Biolabs

16. Buffer: B0202S 10x Buffer for T4 DNA ligase with 10mM ATP from New England Biolabs

17. 96-well plate: 3843 Assay plate, 96 well, with LE lid white with clear bottom Poly-D-lysine coated Polystyrene from Corning

18. Minimum Essential Medium (MEM) : 10-010-CV MEM 1x Minimum Essential Medium Eagle with Earle's salts & L-glutamine from Corning cellgro

19. FBS: 16000-044 Fetal Bovine Serum from Gibco

20. PSF (M10PSF): Penicillin/Streptomycin + Amphotericin

21. Penicillin/Streptomycin: P4333 Penicillin-Streptomycin solution stabilized with 10,000 U of penicillin and 10mg streptomycin from Sigma Aldrich

22. Amphotericin: 15290-018 Amphotericin B 250 µg/ml from Gibco

24. CRE-Luc: from Stratagene

25. RTP1S: 100ng/µl plasmid from H. Matsunami lab

26. SV40-RL: E6911 Promoter-Driven Control Renilla Luciferase Vectors pGL4.73[hRluc/SV40]

27. M3: muscarinic acetylcholine receptor M3 100ng/µl plasmid from Matsunami lab

28. Lipofectamine 2000: 11668-019 Lipofectamine 2000 Reagent 1mg/ml from Invitrogen

29. CD293: 11913-019 CD Medium for Suspension cultures from Gibco

30. Glutamine: 25030 L-Glutamine 200mM 100x from Gibco

31. CuCl2: C-6641 Cupric Chloride dehydrate from Sigma Aldrich

32. DMSO: D2650 Dimethyl Sulfoxide Hybri-Max, sterile filtered Bioreagent from Sigma Aldrich

33. Dual-Glo Luciferase buffer: in E2980 Dual-Glo Luciferase Assay System from Promega
34. Stop-and-Glo buffer: in E2980 Dual-Glo Luciferase Assay System from Promega
35. Firefly luciferase substrate: in E2980 Dual-Glo® Luciferase Substrate (lyophilized)
36. Renilla luciferase substrate: in E2980 Dual-Glo® Stop & Glo® Substrate

3.Methods

3.1 File preparation for machine learning

3.1.1 Information about the agonist/non-agonist activity of a ligand

The file preparation is a crucial step and probably the lengthiest one.

1. First, gather all available information on a given OR and its deorphanization status.
2. A file can be constituted containing i/ in the first column the SDF identifier or CID (see below) of the ligand and ii/ in the second a binary code: label 1 if the ligand is considered by the authors to be an agonist, and label 0 for a non-agonist. (see Note 1)

3.1.2 SDF file preparation and Dragon descriptor calculation

1. Download the SDF 3D conformers informations in Pubchem. Usually the file contains a header with a unique identification number (CID).
2. Once all the relevant SDFs have been compiled, Dragon calculates chemical descriptors for each molecule.
3. After calculation, a file containing the ID of the ligand (here the SDF header) and ~ 4000 descriptors is obtained.
4. Before continuing, the file should be checked and cleaned (exclude descriptors with errors such as ‘NaN’- extending for non-numeric-numbers).

3.2 Machine Learning Workflow setup

3.2.1 Dataset preparation

1. Each file is loaded into a KNIME workflow using the File Reader node, and both files are joined to contain all the descriptors and information about the activity of the ligand.

2. Descriptors are then normalized before being filtered out for low variance and for high correlation (Fig. 1).

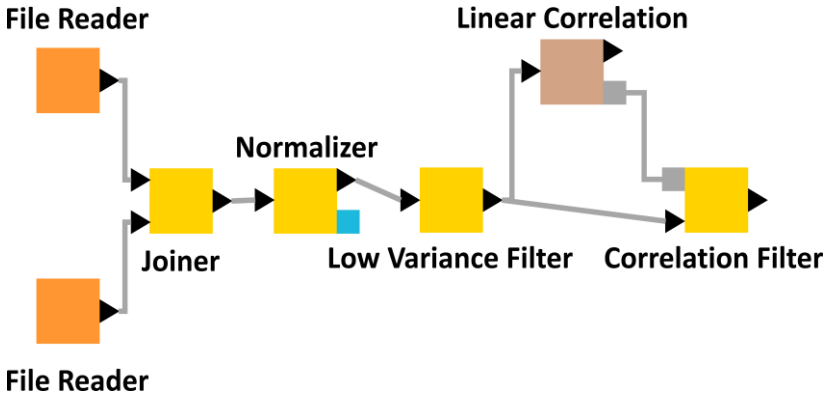


Figure 1. Dataset preparation nodes in KNIME.

3. Optional step: If the data set contains much more non-agonists than agonists, the data can be filtered to select only representative non-agonists, thus avoiding over-learning the criteria for a ligand to be a non-agonist, and obtaining a balanced dataset (Fig. 2).

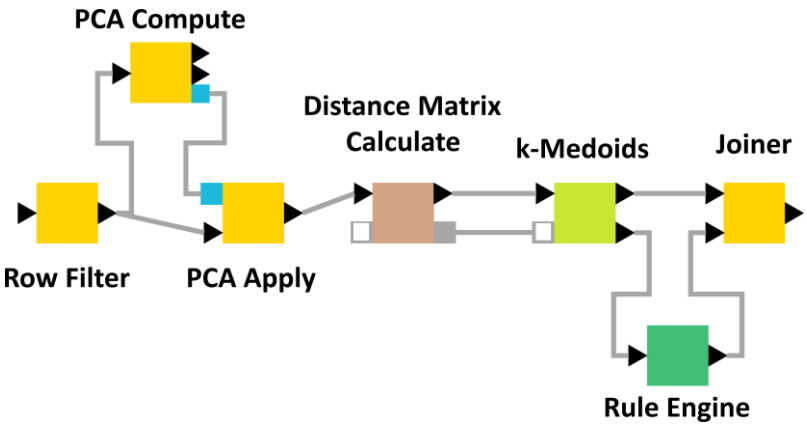


Figure 2. Filtration of ligands in case of unbalanced database.

4. Here add the Row Filter node and select only the non-agonists (0).

5. Perform a principal component analysis (PCA), add a Distance Matrix Calculate node, followed by a k-medoids node where the partition count should be set to the number of molecules that are needed obtain a balanced dataset. The output file created by the execution of this node k-Medoids contains the representative molecules which will be used for the model generation. The Rule Engine node allows filtering the non-agonists to ensure that there

are comparable amounts of agonists and non-agonists and that the non-agonists span the tested chemical space.

6. Concatenate the resulting file with the file containing only the agonists (obtained using a Row Filter from the initial dataset.)

3.2.2 Machine learning model generation

1. Link the dataset information to a Partitioning node to split the dataset into learning set and test set.

2. The learner chosen downstream will gain information to build the model from the learning set. Once the model is established, its performance will be evaluated thanks to the test set.

3. The learning set is then connected to an X-partitioner node which is linked to a learner (support vector machine (LIBSVM) or RandomForest) and then the Predictor node.

4. The loop is terminated by an X-Aggregator (Fig. 3). The X-partitioner and the X-aggregator nodes are necessary to perform cross validation. (see Note 2)

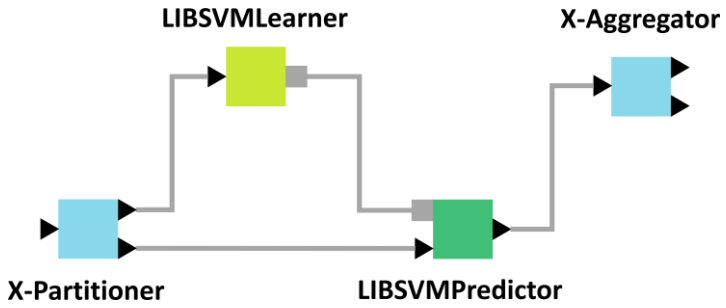


Figure 3. Workflow allowing the LIBSVM nodes to learn over several iterations using the leave-one-out method.

3.2.3 Assessment of the model performance

1. The output generated by the model is connected to a Rule Engine node to assess the number of true and false positives and negative.

2. The MathFormula node later allows assessing the performance of the model. Matthew's coefficient formula is informative, but any other preferred measure can be performed here.

3. The resulting model can be used as an input for another Predictor node that has as an input the second partitioning set (the test set).

4. The same workflow as above allows here to calculate a Matthew's coefficient score (MCS). It allows evaluating the performance of a model with newly tested molecules. Empirically, a good MCS is of ~0.3 and above, +1 representing a perfect learning of the SVM model. (see Note 3)

3.2.4 Agonists prediction and similarity score filtering

1. The file containing all the descriptors of the database to be tested has to be connected to a Predictor node. The output predicts potential new agonists and non-agonists.

2. It is important to note that molecules to be screened on the model should belong to the same chemical space as the learning set. Therefore, prior to screening a library it is highly desirable to assess the similarity using a *Tanimoto* score for example. A *Tanimoto* score above 0.85 is typically a good indicator that molecules belong to the same chemical space. To do so, SDF 3D descriptors of the molecules of interest (i.e. molecules used in the learning set and molecules predicted by the model as being agonists) must be loaded into two different SDF Reader nodes which are in turn each linked to a Fingerprints node.

3. Finally, a Similarity Search node allows calculating the similarity between the molecules predicted by the model and the molecules contained in the learning set (Fig. 4).

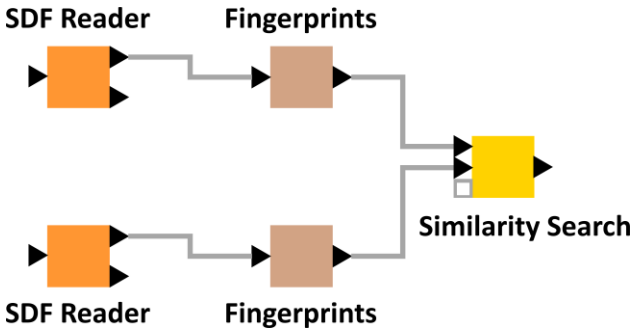


Figure 4. Assessment of applicability domain: Comparison of the chemical space of the molecules present in the library and on the molecules on which the model was built.

4. Once this crucial step is completed, Dragon descriptors for the molecules belonging to the same chemical space as the learning set can be calculated. These molecules will therefore be in line with what the model learned and can be screened on it through the Predictor node.

3.3 3D computational modeling

A chapter dedicated to the Molecular Modeling of Odorant/Olfactory Receptor Complexes can be found in the 2013 edition of Olfactory Receptors Methods and Protocols. [12] This is a short update of this chapter.

3.3.1 Sequence Alignment

1. Homology modeling requires accurate sequence alignment to existing structures. Since no experimental OR structure is available to date, sequence alignment of the target OR and at least one other class A GPCR for which an experimental structure has been solved should be performed.

2. Since ORs and other class A GPCRs show relatively small sequence identities (about 20 % to 30%) alignments between ORs and template GPCRs should be done carefully.

3. Since packing of the transmembrane helices of class A GPCRs with experimental structures is well conserved, it is safe to assume that ORs would share the packing. Therefore, gaps should be avoided within transmembrane helices and conserved motifs inside transmembrane regions should be used as anchoring points [13]. These include:

-GN pattern in transmembrane domain 1 (TM1)

-LAxAD motif in TM2 (LSxxD in ORs)

-A cysteine residue in the extracellular side of TM3 and DRY motif in the intracellular side of TM3

-A tryptophan residue in TM4

-A tyrosine residue in the intracellular part of TM5 (usually CY for non-ORs and SY for ORs)

-KA residues in TM6, the highly conserved CWLP motif in non-ORs is to be aligned with the FYG motif in ORs.

-NPxxY pattern in TM7.

The cysteine residue on top of TM3 and another highly conserved one in extracellular loop 2 (ECL2) form a cysteine bridge.

4. Some larger motifs which are conserved in ORs exclusively can be used if several OR sequences are aligned to produce several models, these include:

-LHxPMYFFLxxLSxxD in TM2

-MAYDRYVAICxPLxY in TM3

-KAFSTCxSH in TM6

-PxLNPxIYxLRN in TM7

A second cysteine pair is conserved in OR sequences and should be used as an anchoring point in the sequence of the extracellular loop 2.

5. Paste the sequences of ORs and GPCRs templates in FASTA format. Open in Jalview and execute the Mafft program. Manually ensure that the conserved regions cited above are correctly aligned. Further guidance on GPCR alignment can be found in refs [13,14].

3.3.2 From Sequence to Structure

1. Obtain experimental structures for the GPCRs used as templates via the Protein Data Bank website. Each PDB file should be cleaned from water molecules, lipids, beta-factors, ligands, nanoparticles, etc.

2. Modeller input files are prepared specifying the residues forming a disulfide bridge between the cysteine on top of TM3 and the one in ECL2.

3.3.3 3D model analysis, validation and preparation

In addition to the 3D model analysis and validation discussed in the 2013 edition these steps can be considered to estimate the quality of built OR 3D structures [12].

1. Model selection can be assessed using a DOPE score. This score is used to assess the energy of the protein model generated by Modeller.

2. The best model should also be visually checked. It should include structured and overall clean helices.

3. An important criterium in any computational biophysics experiment is to determine the protonation state of the protein. For this H++ and PropKa softwares are useful.

3.3.4 Building Protein-Ligand complexes

1. Dock the ligand inside the binding cavity of the protein.

2. Choose the docking conformation involving residues found to be important in site-directed mutagenesis experiments (vide infra).

3.4 In vitro approach

Two types of *in vitro* experiments are ideally combined with *in silico* studies: 1/ preliminary validation of the theoretical model, 2/ test of the hypothesis brought by the model. Both steps can be performed by monitoring OR activation, either through an odorant stimulation or via a spontaneous (basal) activity. Inserting mutations at position revealed to be of interest by the model will assess their importance.

3.4.1 Site directed mutagenesis

Follow sections 3.4.2 to 3.4.4 to accomplish the site directed mutagenesis. Figure 5 recapitulates the typical protocol of site-directed mutagenesis.

3.4.2 Design of primers

1. The design of the primer is the first step to introduce a mutation in a nucleotide sequence. The primers typically have a length of 18 to 25 nucleotides and an estimated annealing temperature between 56°C and 58°C.

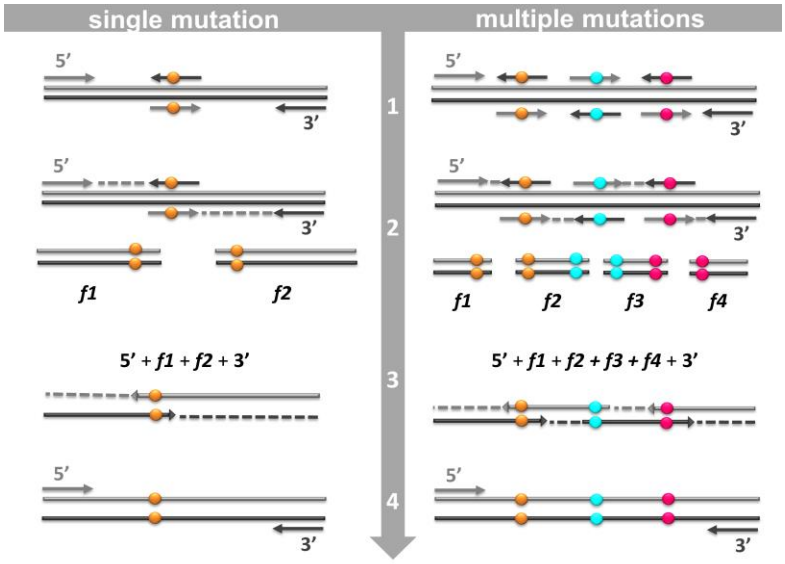


Figure 5. Principle of PCR-mediated site directed mutagenesis. Step 1: Design of the reverse and forward primers containing the desired mutation(s) (colored in orange, blue or pink) in comparison to the reference sequence (gray). Step 2: Application of the first PCR round to the primers and the reference to obtain fragments including the mutation(s). Step 3: Application of the second PCR round to create the full sequence of the desired mutant. Step 4: Final sequence of the mutant with all mutations inserted. It will further be amplified and purified prior insertion into a Rho-pCI vector.

2. To calculate the annealing temperature of a primer, A and T nucleotides account for 2°C while G and C account for 4°C. The mutated site should be located in the middle of the primer. We include a -4°C penalty per nucleotide mismatch. An example is provided in Figure 6.

- Amino acid sequence of the receptor, the mutation desired is L/V

[...VYTYGFLNSVIQT...]

- Corresponding nucleotide sequence

[… GTG TAC ACC TAC GGC TTC CTG AAC AGC GTG ATC CAG ACC …]

- Design of the primers

By mutating CTG in GTG, the L residue will be mutated in F.
C TAC GGC TTC GTG AAC AGC GT
The annealing temperature is then estimated as follows:
4 224 444 224 -424 224 244 42 = 58°C

Finally, the corresponding reverse primer is designed:

primer-L/V-forward: C TAC GGC TTC GTG AAC AGC GT

primer-L/V-reverse: AC GCT GTT CAC GAA GCC GTA G

Figure 6. Primer design protocol.

3.4.3 Site directed mutagenesis with Phusion polymerase

1. Two successive PCR rounds are required to reach the final mutant sequence (Fig. 5, steps 2-3). We use Phusion DNA polymerase (NEB) which allows robust and consistent amplification while minimizing error rate.

2. The PCR first round produces different fragments including the mutation based on the template. Coming back to the example of F, here we have two fragments: 1 and 2. The fragment 1 is delimited by the 5' (forward) and primer-L/V-reverse and the fragment 2 by primer-L/V-forward and 3' (reverse). Each of them is produced following the protocol of Table 1.

Table 1. Reagents mix (left) and Phusion PCR thermo-cycler conditions (right) for the first PCR round

Reagents	V(µL) for one sample
5x Phusion HF buffer	2
2mM dNTP	1
Phusion pol	0.1
5µM Primer forward	1
5µM Primer reverse	1
1ng/µl Template plasmid	1
Distilled water	5

Phusion PCR	
98°C	30sec
98°C	5sec
55°C	15sec
72°C	1min / kb
For 25 cycles	
72°C	5min
10°C	Pause

3. The production of the fragments is confirmed by running an agarose gel electrophoresis.

4. Dilute the PCR products ten times in distilled water for the second round PCR (Table 2). This second PCR aim is to assemble the fragments in order to obtain the final nucleotide sequence of the mutant (Fig. 5, step 3).

Table 2. Reagents mix (left) and Phusion PCR thermo-cycler conditions (right) for the second PCR round.

Reagents	V(µL) for one sample
5x Phusion HF buffer	5
2mM dNTP	2.5
Phusion pol	0.25
5µM 5' primer	1
5µM 3' primer	1
Fragment **1** 1/10	1.25
Fragment **2** 1/10	1.25
Distilled water	11

Phusion PCR	
98°C	30sec
98°C	5sec
55°C	15sec
72°C	1min / kb
For 25 cycles	
72°C	5min
10°C	Pause

5. The products are again confirmed by running an agarose gel electrophoresis.

3.4.4 Purification and insertion in Rho-pCI vector

1. Adding N-terminal extracellular domain sequences of rhodopsin (Rho-tag) to ORs has been shown to both improve the OR expression and convenient to evaluate its expression [15,16]. Our Rho-pCI vector has the insertion of the first 20 amino acids of human rhodopsin in the pCI vector (Promega) between the NheI and EcoRI sites. Here, insert the OR sequence between the MluI and NotI sites.

2. Purify the PCR products on silica gel columns (Qiagen minelute PCR purification kit) and digest with the restriction enzymes (Table 3). For this digestion, add the 100 ng/µL Rho-pCI vector to the samples to use it in the ligation step.

Table 3. Reagents mix for the digestion with restriction enzymes

	V (µL) for one sample
DNA	9
Buffer 3	2
MluI	0.5
NotI	0.5
BSA	0.2
dW	8

3. Digest at 37°C for 2 hours. The product is then purified on gel and on column.

4. The purified samples are now inserted in the Rho-pCI vector (produced previously) in a ligation step as described in Table 4.

Table 4. Reagents mix for the ligation

	V (µL) for one sample
Insert	1
Rho-pCI vector	0.5
T4 ligase	0.5
Buffer	0.5
dW	2.5

5. Ligation is done at room temperature for at least 1h. Use the ligation product for transformation by adding 2.5 µL of the ligation product to 40 µL of competent cells (DH5α).

6. Leave the mix on ice for 10min and undergo a heat shock for 30sec at 37-42°C and then plate it on LB-amp plate.

7. Inclubate plates at 37°C overnight and pick up colonies for a miniprep and sequence verification.

3.4.5 Dual Luciferase reporter gene assay

Several methods exist to experimentally monitor OR activation by an odorant. They can be divided into two families, either cAMP or calcium release reporting assays. Here we focus on cAMP reporter assays. The two protocols mostly used to monitor OR response by cAMP are the Glosensor [17] and the one presented in this chapter, the luciferase reporter gene assays, depicted in Figure 7 [18]. Follow the sections 3.4.6 to 3.4.8 to monitor OR activation.

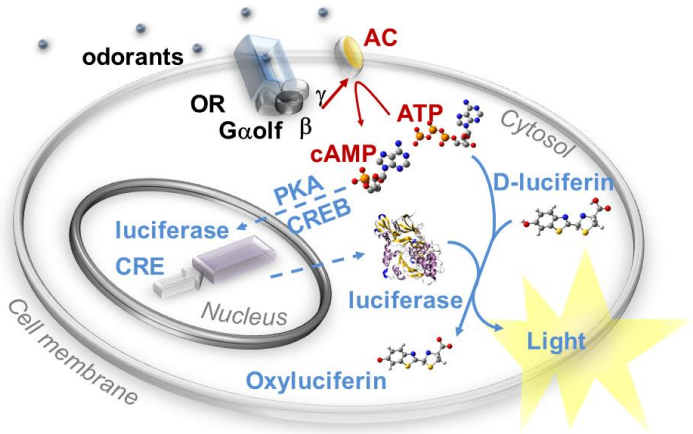


Figure 7. Principle of OR activation monitoring mediated by Firefly luciferase reporter gene. OR activation (black labels) is triggered by the binding of an odorant. Upon activation by an agonist, the α subunit of the Golf initiates the signaling pathway (red labels) by activating the Adenylate cyclase (AC) which produces cAMP from ATP. The cAMP release is monitored indirectly by firefly luciferase production through the activation of the reporter gene (blue labels). It is combined with the Renilla luciferase reporter which is constitutively active with the SV40 promoter to monitor the transfection efficiency and cell viability (internal control reporter).

3.4.6 Cell culture

Hana3A cells stably express Golf which couples with the activated OR to trigger the cAMP release. The following protocol is valid for one 96-well plate coated with poly D lysine.

1. The day before transfection, 1/10 of a 100% confluence 100mm plate of Hana3A is suspended in 6mL of Minimum Essential Medium (MEM) containing 10% FBS medium and PSF (M10PSF).

2. Add 50µL of the suspended cells in each well.

3. Incubate the 96 wells plate overnight.

3.4.7 Transfection

In addition to the receptor, several plasmids are added to the transfection step;

1. Add RTP1S. This is important to promote OR expression [2,19].

2. Add Muscarinic acetylcholine M3 receptor. It modulates OR signaling [16] and CRE-Luc (Firefly luciferase) and SV40-RL (Renilla luciferase) are necessary for the assay.

3. Add the empty vector Rho-pCI to the transfection plan as a control.

4. The mix of the plasmids is done in Mix1 (Table 5) and the Lipofectamine reagent is diluted in MEM in Mix2 (Table 6).

Table 5. Mix1 – mix and dilution of the plasmids

MEM	500µL
CRE-Luc	10ng
RTP1S	5ng
SV40-RL	5ng
M3	2.5ng

Table 6. Mix2 - dilution of the Lipofectamine reagent

MEM	500µL
Lipofectamine 2000	20µL

5. Add 5ng of OR plasmid (proportionally divide Mix1 and OR plasmid quantity if you have several receptors expressed in one plate)

6. Add Mix2 (proportionally divide if necessary) to Mix1 and incubate for 15 minutes.

7. Add 5000µL of M10 for one plate and replace the M10PSF of the 96 wells plate by 50µL of the mix per well.

8. The plate is incubated overnight.

3.4.8 Monitoring

1. Prepare odorant solution in CD293 with glutamine and supplemented with 30µM CuCl2.

2. Odorants are typically at 1M concentration in DMSO and kept at -20°C. Depending on their solubility, we may use lower concentrations or alternative solvents such as ethanol.
3. Screening of multiple odorants at one concentration are typically performed at 150µM and dose response curves are usually performed with concentrations of 0, 1, 3.16, 10, 31.6, 100, and 316µM. Cells are stimulated with 25µL of odorant solutions for 3.5 hours.
4. Add 10µL of Dual-Glo Luciferase buffer containing the Firefly Luciferase substrate per well and shake 5-10 min before measuring the Firefly luciferase activity using a luminometer.
5. Add 9.5µL of Stop-and-Glo buffer containing 1/100 volume of Renilla Luciferase substrate. 6. After 5-10min shaking, monitor the Renilla luciferase response with a luminometer.
7. To analyze the results, subtract the value of luminescence of an empty cell to each Firefly and Renilla luciferase response and divide the Firefly by the Renilla values:
8. Here, 400 is background reading with our luminometer. To ensure the specificity of the response, it is possible to normalize the OR by the Rho-pCI responses. Normalization to the basal activity of each receptor can be helpful when comparing the response of several ORs.

4. Notes

1. It can be useful to include a couple of additional column containing information about the type of assay and the concentration at which the odorant was tested, the source of the data, and EC50 values if applicable. This additional information can be helpful when contradicting data is available between two studies. It should be noted at this point that authors don't systematically distinguish between enantiomers and often use enantiomer mixtures. The establishment of this file is therefore delicate and particular attention should be brought to specifications in each publication.
2. Reproducibility is a major issue in machine learning methods, thus, a sorter node can be added between the dataset nodes to order ligands according to their activity, or CID number. In the Partitioning node, if the chosen sampling manner is set to random, remember to save the molecules present in the test or in the learning set for further reference.
3. Another criterion should be tested before saving the model: i.e. how well does the model filter out ligands? Is it too permissive considering most of your dataset to have a certain activity or is it able to largely discriminate the set? Several iterations and modifications of the splitting parameters in the partitioning node can be necessary before finding a model which captures the complexity of the chemical space.

References

1. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).
2. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Sci. Signal. 2, ra9 (2009).
3. C. Geithe, G. Andersen, A. Malki, D. Krautwurst, A Butter Aroma Recombinate Activates Human Class-I Odorant Receptors. J. Agric. Food Chem. 63, 9410-9420 (2015).
4. J. D. Mainland, A. Keller, Y. R. Li, T. Zhou, C. Trimmer, L. L. Snyder, A. H. Moberly, K. A. Adipietro, W. L. Liu, H. Zhuang, S. Zhan, S. S. Lee, A. Lin, H. Matsunami, The missense of smell: functional variability in the human odorant receptor repertoire. Nat. Neurosci. 17, 114-120 (2014).
5. K. A. Adipietro, J. D. Mainland, H. Matsunami, Functional evolution of mammalian odorant receptors. PLoS Genet. 8, e1002821 (2012).
6. C. A. de March, Y. Yu, M. J. Ni, K. A. Adipietro, H. Matsunami, M. Ma, J. Golebiowski, Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. J. Am. Chem. Soc. 137, 8611-8616 (2015).
7. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U. S. A. 112, 14966-14971 (2015).
8. C. Bushdid, C. A. de March, S. Fiorucci, H. Matsunami, J. Golebiowski, Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features. J. Phys. Chem. Lett. 9, 2235-2240 (2018)
9. TALETE srl. Dragon *Software for Molecular Descriptor Calculation* (2014).
10. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, **Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization**. Springer, Berlin Heidelberg, (2008).
11. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinformatics Chapter 5, Unit 5-6 (2006).
12. L. Charlier, J. Topin, C. A. de March, P. C. Lai, C. J. Crasto, J. Golebiowski, Molecular Modelling of Odorant/Olfactory Receptor Complexes, Methods Mol. Biol.,1003, 53-65 (2013).
13. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: from sequence to structure. Protein Sci. 24, 1543-1548 (2015).
14. V. Cvicek, W. A. Goddard, 3rd, R. Abrol, Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. PLoS Comput. Biol. 12, e1004805 (2016).
15. D. Krautwurst, K. W. Yau, R. R. Reed, Identification of ligands for olfactory receptors by functional expression of a receptor library. Cell 95, 917-926 (1998).
16. Y. R. Li, H. Matsunami, Activation state of the M3 muscarinic acetylcholine receptor modulates mammalian odorant receptor signaling. Sci. Signal. 4, ra1 (2011).
17. B. F. Binkowski, F. Fan, K. V. Wood, Luminescent biosensors for real-time monitoring of intracellular cAMP. Methods Mol. Biol. 756, 263-271 (2011).
18. H. Zhuang, H. Matsunami, Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nat. Protoc. 3, 1402-1413 (2008).
19. H. Zhuang, H. Matsunami, Synergism of accessory factors in functional expression of mammalian odorant receptors. J. Biol. Chem. 282, 15284-15293 (2007).

Linking molecular structure to in vitro activity

The olfactory signal generates when the interaction between an odorant molecule and at least one odorant receptors occurs. The molecule will trigger a response of the receptor if it is able to interact with the receptor and trigger the conformational changes necessary to the coupling of the G protein.

Knowing the physical and chemical properties a molecule must have to be an agonist or a non-agonist of a receptor is at the core of cracking the combinatorial code of olfaction. Indeed, linking the genetic sequence to the functional profile of the receptors should lead to the decipherment of the olfactory code.

In several studies using medicinal chemistry based approaches, a precise understanding of the chemical determinants which are responsible for activation or blocking a receptor is undertaken. For example, the molecular receptive range of rat receptor i7 was investigated thoroughly, and showed that this receptor responded to aliphatic aldehydes bearing a specific side chain length (1). Other factors influencing a receptors' response include features such as the topological surface area (2), the relative position of a functional group (3) or molecular conformation (4). The idea of these studies is to determine the « rules » that the ligands need to follow to agonize or antagonize a receptor.

However, in general, when looking at the reported agonists and non-agonists of most deorphanized ORs outside of these medicinal chemistry frameworks, **no obvious link between the structural motif of the ligands and their activity can be established**.

The amount of chemogenomic data generated by experimentalists working in the field of olfaction has exploded since the development of a protocol to express ORs in heterologous expression system (5). Notably, large scale deorphanization studies using these types of methods have accounted for largely deorphanizing human ORs (6-8). Of course, other studies, which are specific of a given OR (9-13) also contribute greatly to the enrichment of our olfactory knowledge in terms of odorant-OR pairs.

As a whole, all these types of studies provide **a wealth of data which is prone to be mined by computational approaches**. To what extent can we mine the data which has been generated by experimentalists to predict novel ligands?

Another interesting question regarding agonists often addressed in a medicinal chemistry framework regards the strength of the receptors' response. How can we build a model to understand how the strength of the agonists is encoded in the receptors' sequence?

In this article, I addressed these questions by using the previously established Machine Learning protocol, which uses Support Vector Machine (SVM) to predict the activity of a ligand by virtual screening. The results showed that it is indeed possible to predict, with a reasonable accuracy, the activity of a molecule using only molecular descriptors. I identified new agonists for four ORs: OR51E1, OR1A1, OR2W1 and MOR256-3. Furthermore, I created a three-dimensional structure of the main studied receptor, OR51E1, to understand how particular residues in the cavity of the receptor affect the efficacy of the ligand.

References

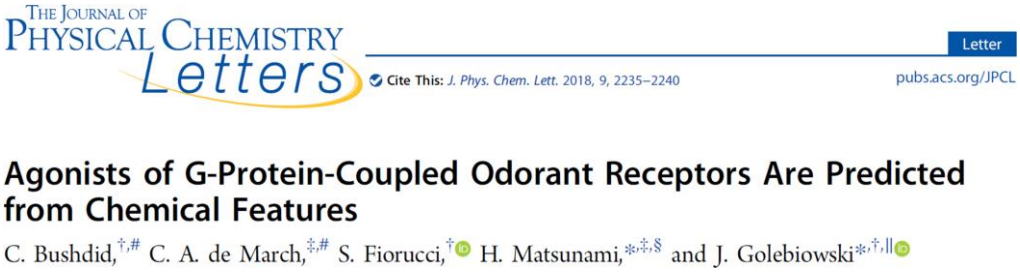
1. R. C. Araneda, A. D. Kini, S. Firestein, The molecular receptive range of an odorant receptor. Nat. Neurosci. 3, 1248-1255 (2000).
2. E. Poivet, Z. Peterlin, N. Tahirova, L. Xu, C. Altomare, A. Paria, D. J. Zou, S. Firestein, Applying medicinal chemistry strategies to understand odorant discrimination. Nat. Commun. 7:11157 (2016).
3. E. Poivet, N. Tahirova, Z. Peterlin, L. Xu, D. J. Zou, T. Acree, S. Firestein, Functional odor classification through a medicinal chemistry approach. Sci. Adv. 4, eaao6086 (2018).
4. Z. Peterlin, Y. Li, G. Sun, R. Shah, S. Firestein, K. Ryan, The importance of odorant conformation to the binding and activation of a representative olfactory receptor. Chemistry & biology 15, 1317-1327 (2008).
5. H. Zhuang, H. Matsunami, Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nat. Protoc. 3, 1402-1413 (2008).
6. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Sci. Signal. 2, ra9 (2009).
7. K. A. Adipietro, J. D. Mainland, H. Matsunami, Functional evolution of mammalian odorant receptors. PLoS Genet 8, e1002821 (2012).
8. J. D. Mainland, A. Keller, Y. R. Li, T. Zhou, C. Trimmer, L. L. Snyder, A. H. Moberly, K. A. Adipietro, W. L. Liu, H. Zhuang, S. Zhan, S. S. Lee, A. Lin, H. Matsunami, The missense of smell: functional variability in the human odorant receptor repertoire. Nat. Neurosci. 17, 114-120 (2014).
9. G. Sanz, C. Schlegel, J. C. Pernollet, L. Briand, Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism. Chem. Senses 30, 69-80 (2005).
10. S. Li, L. Ahmed, R. Zhang, Y. Pan, H. Matsunami, J. L. Burger, E. Block, V. S. Batista, H. Zhuang, Smelling Sulfur: Copper and Silver Regulate the Response of Human Odorant Receptor OR2T11 to Low-Molecular-Weight Thiols. J. Am. Chem. Soc. 138, 13281-13288 (2016).
11. C. Geithe, F. Noe, J. Kreissl, D. Krautwurst, 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 42, 181-193 (2017).
12. F. Noe, J. Polster, C. Geithe, M. Kotthoff, P. Schieberle, D. Krautwurst, OR2M3: A Highly Specific and Narrowly Tuned Human Odorant Receptor for the Sensitive Detection of Onion Key Food Odorant 3-Mercapto-2-methylpentan-1-ol. Chem. Senses 42, 195-210 (2017).
13. L. Ahmed, Y. Zhang, E. Block, M. Buehl, M. J. Corr, R. A. Cormanich, S. Gundala, H. Matsunami, D. O'Hagan, M. Ozbil, Y. Pan, S. Sekharan, N. Ten, M. Wang, M. Yang, Q. Zhang, R. Zhang, V. S. Batista, H. Zhuang, Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. Proc. Natl. Acad. Sci. U. S. A. 115, E3950-E3958 (2018).

Publication 3

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

CAROLINE BUSHDID, CLAIRE A. DE MARCH, SEBASTIEN FIORUCCI, HIROAKI MATSUNAMI, JEROME GOLEBIOWSKI

IN THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS 9, 2235-2240, **2018**



Abstract

Predicting the activity of chemicals for a given odorant receptor is a longstanding challenge. Here, the activity of 258 chemicals on the human G protein-coupled odorant receptor (OR)51E1, also known as prostate-specific G protein-coupled receptor 2 (PSGR2) was virtually screened by machine learning using 4884 chemical descriptors as input. A systematic control by functional *in vitro* assays revealed that a support vector machine algorithm accurately predicted the activity of a screened library. It allowed identifying two novel agonists *in vitro* for OR51E1. The transferability of the protocol was assessed on OR1A1, OR2W1, and MOR256-3 odorant receptors and in each case, novel agonists were identified with a hit rate of 39% to 50%. We further show how ligands' efficacy is encoded into residues within OR51E1 cavity using a molecular modeling protocol. Our approach allows widening the chemical spaces associated with odorant receptors. This machine learning protocol based on chemical features thus represents an efficient tool for screening ligands for G protein-coupled odorant receptors that modulate non-olfactory functions, or upon combinatorial activation, give rise to our sense of smell.

Introduction

Odorant receptor (OR) genes represent more than 4% of our proteome. (1) ORs belong to the class A G protein-coupled receptor (GPCR) family of proteins, which are responsible for transmitting signals across cell membranes. ORs play a crucial and central role in our sense of smell, endowing us with an extraordinary olfactory discrimination ability. (2) They are also ectopically expressed in various organs and modulate non-olfactory functions. (3) Therefore, understanding the chemical space ORs respond to is not only relevant to understand olfaction, but may also have a pharmacological impact.

Upon smelling, odorant compounds reach ORs present at the surface of olfactory sensory neurons. At the molecular level, odorants combinatorially activate a fraction of our 397 types of ORs, (1, 4) and the pattern of activation codes for a specific olfactory percept. Thus, ORs are very subtle molecular machines and small modifications in their amino acid sequence can drastically affect their response to chemicals. (5) Accordingly, the relationship between the ligand chemical space and the receptor space is complex and subtle; however, our understanding of how OR characteristics match ligand features remains limited in the field of odor perception, as well as more generally in research on GPCRs.

From a pharmacological perspective, odorants are no more than ligands for G protein-coupled odorant receptors. GPCR ligands can be classified as agonists, non-agonists, or antagonists. However, the defining features of GPCR agonists are not well characterized. It is very likely that molecular features of candidate ligands encode their activity, and the challenge for researchers lies in identifying rules that link chemical descriptors to receptor activation. For agonists, the strength of the receptor response (efficacy) is then determined by a finely tuned interaction between the ligand and the receptor, which controls the downstream signaling. (6)

Machine learning may be a useful general approach to explore the chemical space of mammalian ORs and to identify general chemogenomic rules. Indeed, machine learning approaches have been shown to predict olfactory perception from chemical features of odorants. (7) Focusing on GPCRs drug design, machine learning is considered sufficiently promising to increase the success rate of discovery of novel ligands in the coming years. (8) To date, machine learning mostly processed ligand similarity (9) or few types of ligand chemical descriptors or protocols. (10-13) Most GPCR studies focused on already known data and discussed the accuracy of machine learning models. However, the use of an external dataset using novel candidates is crucial to assess the real performance of the model, as recently performed on bitter taste receptors. (14)

In this work, we designed a synergistic approach that combines machine learning, *in vitro* luciferase assays, and homology modeling, (15) to i) identify novel ligands for ORs and ii) better understand the receptor features that govern agonist efficacy. We chose OR51E1 as the receptor of interest because it is a highly characterized OR. It also has ubiquitous functions in olfactory and non-olfactory tissues. (16) OR51E1 is also known as prostate-specific G protein-coupled receptor 2 (PSGR2) and is of particular interest as it is not only involved in odor recognition but also in human prostate or lung cancer cell proliferation. Furthermore, this receptor is postulated to be a marker for neuroendocrinic carcinoma cells (17-18).

Ligand activity is predicted by a support vector machine algorithm. All available data regarding OR51E1 were gathered (File S1). To our knowledge, 24 molecules are reported as agonists of OR51E1 while 96 are reported as non-agonists. Agonist and non-agonist spaces cannot easily be discriminated based on simple chemical descriptors (Fig. 1a). The 24 known agonists of this receptor have a wide array of chemical functions and structures. About 58% of these are carboxylic acids, and the remaining 42% belong to various chemical families (aldehydes, aromatic cycles with alcohols, esters, and ethers, as well as nitrogen- and sulfur-containing molecules). Of the 14 agonists within the carboxylic acid family, only one has a cyclic structure (5-norbornene-2-carboxylic acid) while the other 13 present aliphatic side

chains. The carbon chains range from 4 to 14 carbon atoms (butyric acid to tetradecanoic acid). To make sense of such a complex chemical space, we used a machine learning algorithm to capture the chemical characteristics necessary to identify a molecule as an agonist.

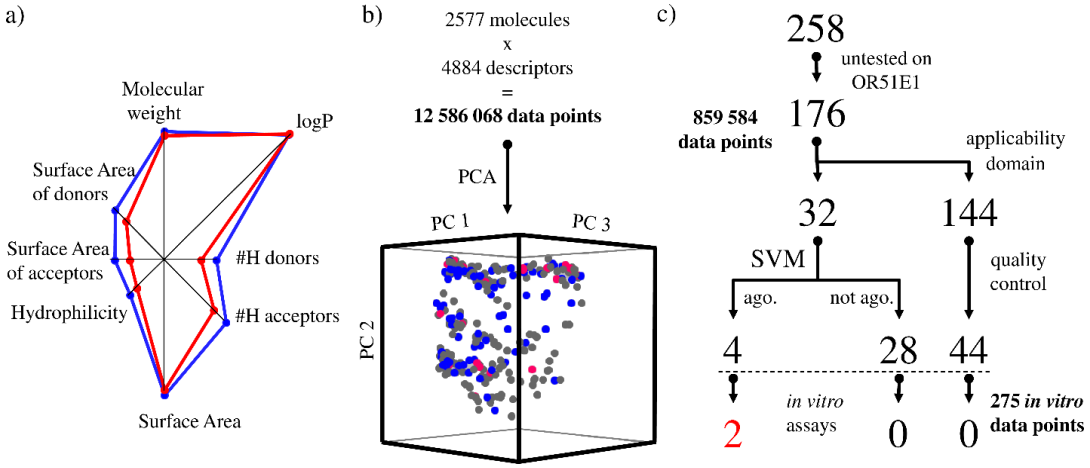


Figure 1. Odorant molecular space and protocol used for OR51E1 virtual screening. a) Spider plot representing the weight of eight attributes (molecular weight, logP, number of hydrogen-bond donor and acceptors, total surface area, hydrophobicity, and surface areas of acceptor and donor atoms) to define OR51E1 agonists (red) and non-agonists (blue) as identified prior to this study. b) Projection of our library of 176 original odorants, shown in gray, into the three first principal components of the odorant space computed on the basis of chemical descriptors of 2577 odorant compounds. Known agonists and non-agonists are shown in red and blue, respectively. c) Of our library of 258 odorants, 176 had been untested on OR51E1, and 32 belonged to the same chemical space (applicability domain) of our model. The virtual screening predicted 4 agonists, two of which elicited a receptor response *in vitro*. All predicted non-agonists were tested *in vitro* for quality control purposes, as were 44 randomly selected compounds initially excluded from the library; none triggered a receptor response *in vitro*.

We further defined the odorant molecular space by computing 4884 chemical, topological, and electronic descriptors of 2577 commonly used odorants. (19) Odorants belonging to our library, as well as odorants whose activity on OR51E1 has been previously reported (File S1 and S2), were projected onto this subspace formed by the first three principal components (PCs) which account for 56,6 % of variance to examine to what extent they covered the odorant space (Fig. 1b and Fig. S1). The first and the third molecular PCs were weighted by factors that are reasonable descriptors for the size and composition of a molecule (size of the molecule and presence/absence of oxygen, sulfur, or nitrogen atoms). The second PC was weighted by factors describing either the embranchment and/or complexity of the molecule (see the Supporting Information). In this case, agonists and non-agonists spaces were found to be strongly intertwined and highlight the need to mine them using non-linear algorithms. A

Quantitative Structure Activity Relationship approach featuring a machine learning protocol which uses a support vector machine (SVM) was performed.

As shown in Fig. 1c, the initial library was made up of 258 compounds, 176 of which had never been tested before on OR51E1. A first reduction of the testable compounds was performed by assessing the applicability domain of the model. The applicability domain encompasses molecules that are considered similar to the training set. This means that the SVM model "learned" from particular structures of the training set and only similar structures from the untested library can be reasonably screened by the model. Here, we used molecular fingerprints and a *Tanimoto* score to compare molecules from the untested library to the training set (see SI). In our untested library, 144 molecules that were structurally too dissimilar to those of the learning set were accordingly excluded from virtual screening. The remaining 32 compounds belonging to the applicability domain were virtually screened using the SVM algorithm.

These were split into 4 potential agonists and 28 non-agonists. The predicted agonists were 2-methyl-butyric acid, cyclobutanecarboxylic acid, dimethyl-trisulfur, and (S)-(+)-2-octanol. Functional luciferase *in vitro* assays were used to assess the activity of these 32 compounds as well as 44 compounds initially excluded from the applicability domain for quality control reasons (Fig. 1c). A total of 76 molecules were screened at 100 µM on OR51E1 and compared to a set of three positive controls and three negative controls (Fig. 2a). None of the molecules outside of the applicability domain nor those predicted as non-agonists triggered a receptor response. However, of the four potential agonists, two activated the receptor and two did not. Cyclobutanecarboxylic acid and 2-methyl butyric acid exhibited an agonist dose-dependent behavior *in vitro* while the two other predicted agonists (dimethyl-trisulfur and (S)-(+)-2-octanol) were false positives (Fig. 2b and Table S1). Thus, our model accurately captured the activity of 30 compounds out of the 32 belonging to the applicability domain. It shows an *in vitro* hit rate of 50% (two true agonists out of four predicted agonists) and a reliability of 94% (30 correct predictions out of 32).

The transferability of the protocol was assessed by applying it to two other human receptors (OR1A1, OR2W1) and a mouse receptor (MOR256-3), for which many agonists and non-agonists are reported (see File S1). For each system, the SVM algorithm processed the part of the untested library belonging to the applicability domain. Then, *in vitro* assays were used to compare their ability to trigger OR response to those of control agonists. Table 1 reports *in vitro* hit rates for the four systems studied. Novel agonists were predicted with hit rate in the range of [39%, 50%]. For OR1A1, OR2W1, and MOR256-3, *in vitro* data assessed 7, 2, and 5

novel agonists, respectively (Fig. S2, S3 and File S1). This suggests that machine learning approaches are useful to explore the wide chemical space associated with ORs, and potentially other GPCRs if a sufficient number of agonists and non-agonists are known. Here, the smallest database used for training a supervised model concerned OR51E1, consisting of 24 agonists and 24 non-agonists.

Focusing on OR51E1, agonists show a large range of activation strength. Compared with nonanoic acid, a known agonist of OR51E1, (20-21) the two novel agonists showed comparable potencies (see dose-response curves in Fig. 2b) but different efficacies. Namely, cyclobutanecarboxylic acid triggered a response 230% higher than nonanoic acid, while 2-methyl-butyric acid had an efficacy that was 50% lower (Fig. 2b). These variable strengths of the receptor responses suggest differential interactions and affinities between ligands and the receptor cavity. We further investigated the influence of the receptor cavity features on ligand activity using a 3D model.

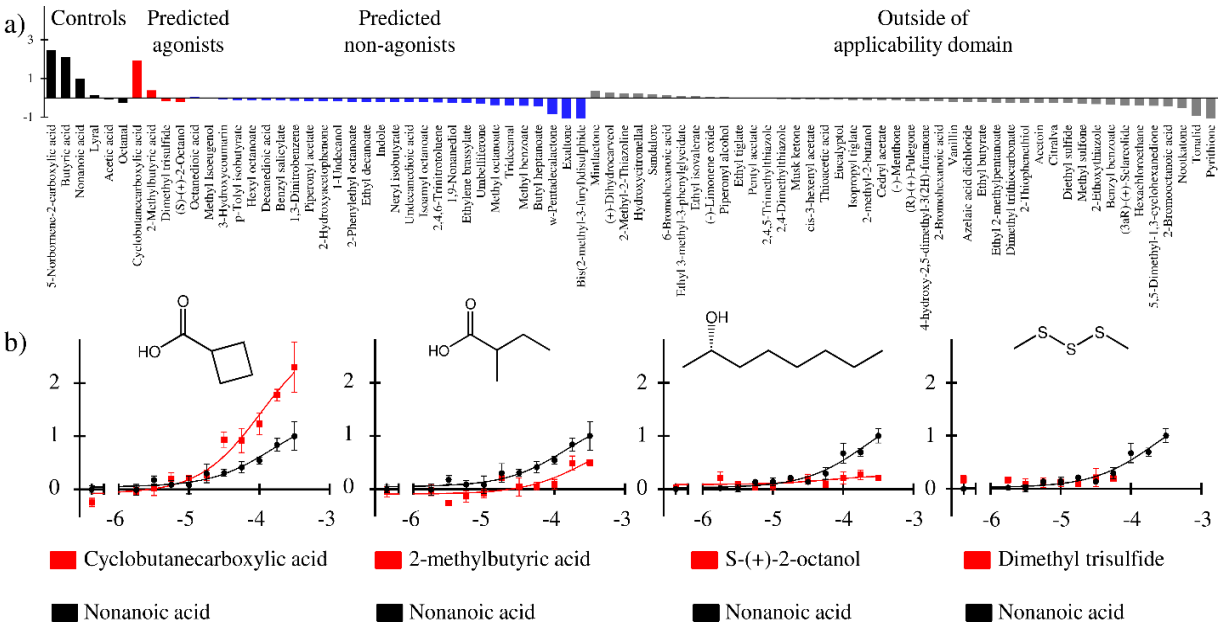


Figure 2. *In vitro* screening of OR51E1. a) Response of OR51E1 to 6 controls (3 non-agonists and 3 agonists – in black), four predicted agonists by the SVM model (in red), 28 predicted non-agonists by the SVM model (in blue), and 44 molecules not within the applicability domain (in gray). Odorants were injected at 100 µM and each response of OR51E1 was corrected by that of the empty vector (pCI) response and by basal activity of receptors. It should be noted that negative responses do not reflect inhibition of the response but rather cell toxicity. b) Dose-response curves for the four potential agonists as identified by the SVM model. Cyclobutanecarboxylic acid and 2-methylbutyric acid were found to activate the receptor while S-(+)-octanol and dimethyl trisulfide did not. All responses were normalized to that of nonanoic acid.

Table 1. Performance and transferability of the SVM model. Numbers of *in vitro*-validated and SVM-predicted agonists and corresponding hit rate for the four systems studied.

	OR51E1	OR1A1	OR2W1	MOR256-3
in vitro / SVM	2/4	7/18	2/5	5/10
Hit rate	50%	39%	40%	50%

Agonist efficacy is affected by the receptor cavity. From a structural point of view, ORs, like class A GPCRs, form a bundle within the cell membrane, and are composed of seven helices named TM1 to TM7. The canonical ligand binding site is located 10 Å below the extracellular side of the receptor. The largest sequence variability between ORs lies at this binding site, to potentially endow our olfactory system with a large discrimination of the chemical space. (22-23) The OR51E1 3D model was built based on X-ray class A GPCRs templates and showed a binding cavity with a solvent accessible surface area of ~300 $Å^3$, ~70% of which is apolar. Consistently, the whole set of 26 agonists of OR51E1 encompasses ligands with variable bulkiness (ranging from 90 to 360 $Å^3$) and a relatively high lipophilicity (logP ranging from -0.8 to 4.7).

In the present study, nonanoic acid and cyclobutanecarboxylic acid, whose efficacies were found to differ by two orders of magnitude, shared the same binding mode. They were both predicted to bind the receptor through their carboxylic moiety at the cradle of the cavity (Fig. 3a and b). The acidic function is in contact with the so-called toggle switch involved in receptor activation, $Y254^{6.48}$ (6.48 refers to the Ballesteros Weinstein notation (24)). The ligands bind between TM3 ($S111^{3.36}$, $H108^{3.33}$), TM5 ($I206^{5.44}$), and TM6 ($Y254^{6.48}$), which is in line with typical binding modes observed in X-ray structures of agonists bound to β2-adrenergic and opsin receptors. (25-26)

More specifically, residues H108 and I206 were found to be differentially in contact with the two agonists, while S111 and Y254 were found to interact with the carboxylic acid moiety. This suggests that H108 and I206 control the bulkiness of the agonist. Mutant ORs at these positions consistently showed differential efficacy modulation *in vitro* (Fig. 3c). The efficacy of nonanoic acid increased by more than two orders of magnitude when the receptor was reprogrammed with smaller residues at the top of its cavity (H108A and I206A in Fig. 3d). These mutants both showed a non-significant effect on the much smaller cyclobutanecarboxylic acid.

When H108 was mutated into a ~25% bulkier and less hydrophilic residue (H108F), the response of the receptor was totally abolished for both agonists, which is likely attributed to a complete blockage of the binding cavity (Fig. S4a). We show that these residues are located at

the upper part of the cavity and modulate its size. They control the accessibility of agonists within the binding cavity by sensing the size of their hydrophobic moiety. This accessibility then modulates the efficacy of the receptor response.

On the other hand, mutations of residue S111 similarly modulated the response to the two different agonists; this is suggestive of a role of S111 in acid moiety recognition, which triggers the activation mechanism once the ligand is bound (Fig. S4b and c).

Consistent with its low efficacy, the energy-minimized docked structure of 2-methylbutyric acid suggested that the ligand is closer to H108 and I206 compared to cyclobutanecarboxylic acid (Fig. S5). As for nonanoic acid, such tighter contacts with the top of the binding cavity would prevent these agonists from having a large efficacy.

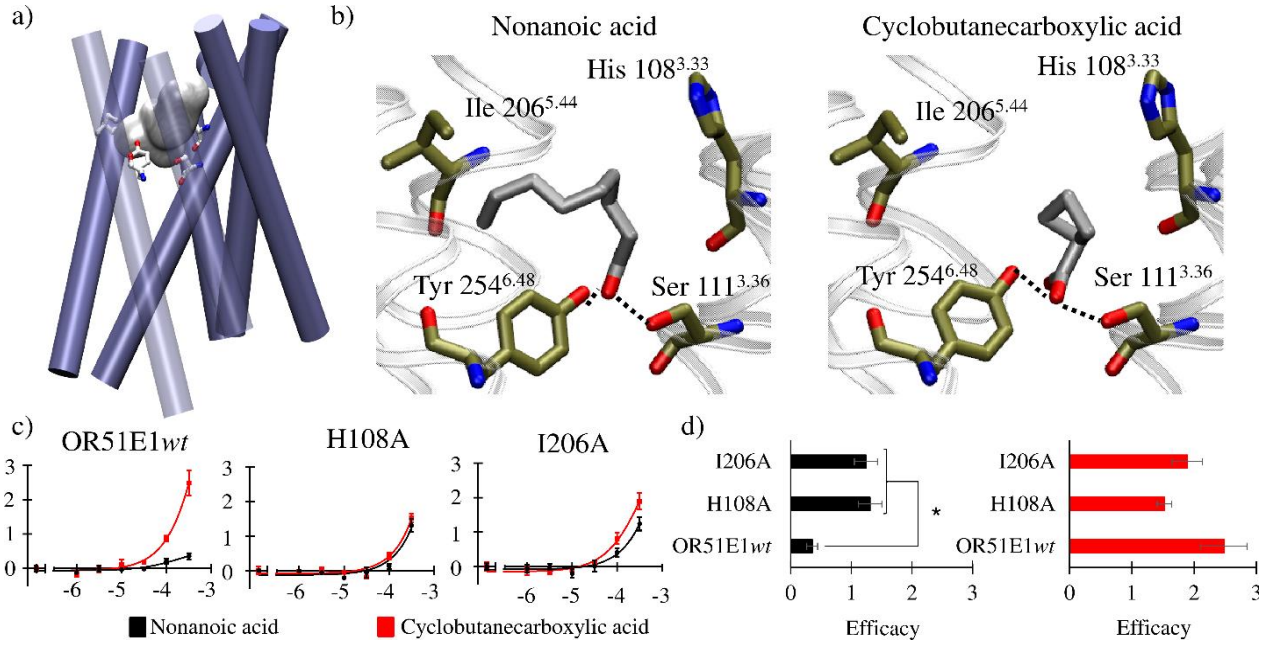


Figure 3. Some OR51E1 cavity residues control the efficacy of bound agonists. a) Overview of OR51E1 wild type (OR51E1 *wt*) bundle structure and binding cavity (white volume). Unfolded loops are omitted for clarity b) Binding mode of nonanoic acid (left) and cyclobutanecarboxylic acid (right) in the binding cradle of OR51E1 *wt*. Carbon atoms are shown in gold (or gray in case of the ligand), oxygen atoms in red, and nitrogen atoms in blue. Hydrogen bond contacts are shown as dashed lines. c) Dose-response curves of OR51E1 *wt* and OR51E1 mutants to nonanoic acid (black) and cyclobutanecarboxylic acid (Red). d) Efficacies of OR51E1 *wt* and OR51E1 mutants to nonanoic acid (black) and cyclobutanecarboxylic acid (red). Error bars indicate the SEM. (*$p < 0.01$)

In conclusion, we used supervised learning algorithm (SVM) to expand the chemical space of OR51E1, whereby we virtually screened a database of 176 compounds. After assessment by *in vitro* luciferase assays, we predicted 30 out of 32 ligand activities and revealed two novel agonists of OR51E1 *in vitro*, which were carboxylic acids. It is interesting to note that the

structures of the acid agonists reported for this receptor are very different; some are short-chained aliphatic acids (such as butyric acid), while others exhibit cyclic structures (such as 5-norbornene-2-carboxylic acid). However, OR51E1 cannot be considered simply as a carboxylic acid sensing receptor. Of the 45 acids that were tested on OR51E1, 64.5% were non-agonists. Our machine learning model was able to accurately learn from this complex set of known agonist and non-agonist structures and successfully predicted two novel ligands for this receptor. The applicability to other ORs was assessed through the use of a similar protocol on OR1A1, OR2W1, and MOR256-3. The models predicted the *in vitro* activity with a hit rate ranging from 39% to 50% and identified novel agonists in all cases. This suggests that machine learning is capable of predicting the activity of candidate compounds using their chemical features. Note that these machine learning models are limited to their applicability domain, which naturally restricts the chemical space research area. Other agonists present outside this domain could be discovered using molecular modeling to predict ligand activity based on receptor activation dynamics (27).

Further, to understand the link between the strength of a response and receptor features, a 3D model of OR51E1 was built. We showed how some residues in the cavity can selectively affect efficacy. Based on the 3D model, we successfully reprogrammed OR51E1 *wt* which confirmed that residues at the top of the orthosteric binding cavity control agonist efficacy. For these two mutants, nonanoic acid was specifically transformed into a highly efficacious agonist compared to the OR51E1 *wt*. This emphasizes the crucial role of some residues in modulating the strength of a receptor response. These residues belonging to TM3 and TM5 are well documented as being involved in ligand recognition in ORs (5).

The use of ligand-based machine learning approaches can be used to enhance our knowledge of the chemogenomic links between the vast space of odorants, made up of millions of molecules and that of receptors, comprising 397 functional genes in humans. Olfactory sensing is still in its infancy and such approaches will help widening the chemical space associated with ORs, for which few agonists and non-agonists are reported. By integrating descriptors for the receptor features, one can envision to produce numerical SVM models able to reliably predict not only the activity of candidate ligands but also their efficacy.

References

1. Y. Niimura, Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Curr. Genomics 13, 103-114 (2012).
2. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. Science 343, 1370-1372 (2014).
3. N. Kang, J. Koo, Olfactory receptors in non-chemosensory tissues. BMB Rep. 45, 612-622 (2012).
4. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).
5. C. A. de March, Y. Yu, M. J. Ni, K. A. Adipietro, H. Matsunami, M. Ma, J. Golebiowski, Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. J. Am. Chem. Soc. 137, 8611-8616 (2015).
6. X. J. Yao, G. Velez Ruiz, M. R. Whorton, S. G. Rasmussen, B. T. DeVree, X. Deupi, R. K. Sunahara, B. Kobilka, The effect of ligand efficacy on the formation and stability of a GPCR-G protein complex. Proc. Natl. Acad. Sci. U. S. A. 106, 9501-9506 (2009).
7. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820-826 (2017).
8. D. Wacker, R. C. Stevens, B. L. Roth, How Ligands Illuminate GPCR Molecular Pharmacology. Cell 170, 414-427 (2017).
9. S. M. Boyle, S. McInally, A. Ray, Expanding the olfactory code by in silico decoding of odor-receptor chemical space. eLife 2, e01120 (2013).
10. K. Mansouri, R. S. Judson, In Silico Study of In Vitro GPCR Assays by QSAR Modeling. Methods Mol. Biol. 1425, 361-381 (2016).
11. R. Shemesh, A. Toporik, Z. Levine, I. Hecht, G. Rotman, A. Wool, D. Dahary, E. Gofer, Y. Kliger, M. A. Soffer, A. Rosenberg, D. Eshel, Y. Cohen, Discovery and validation of novel peptide agonists for G-protein-coupled receptors. J. Biol. Chem. 283, 34643-34649 (2008).
12. J. Wu, Q. Zhang, W. Wu, T. Pang, H. Hu, W. K. B. Chan, X. Ke, Y. Zhang, J. Wren, WDL-RF: Predicting Bioactivities of Ligand Molecules Acting with G Protein-coupled Receptors by Combining Weighted Deep Learning and Random Forest. Bioinformatics, 34, 2271-2282 (2018).
13. M. Schmuker, M. de Bruyne, M. Hahnel, G. Schneider, Predicting olfactory receptor neuron responses from odorant structure. Chem. Cent. J. 1, 11 (2007).
14. W. Huang, Q. Shen, X. Su, M. Ji, X. Liu, Y. Chen, S. Lu, H. Zhuang, J. Zhang, BitterX: a tool for understanding bitter taste in humans. Sci. Rep. 6, 23450 (2016).
15. C. Bushdid, C. A. de March, H. Matsunami, J. Golebiowski, Numerical models and in vitro assays to study Odorant Receptors. Methods Mol. Biol. 1820, 77-93 (2018).
16. C. Flegel, S. Manteniotis, S. Osthold, H. Hatt, G. Gisselmann, Expression profile of ectopic olfactory receptors determined by deep sequencing. PloS One 8, e55368 (2013).
17. J. Leja, A. Essaghir, M. Essand, K. Wester, K. Oberg, T. H. Totterman, R. Lloyd, G. Vasmatzis, J. B. Demoulin, V. Giandomenico, Novel markers for enterochromaffin cells and gastrointestinal neuroendocrine carcinomas. Mod. Pathol. 22, 261-272 (2009).
18. J. Weng, J. Wang, X. Hu, F. Wang, M. Ittmann, M. Liu, PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer. Int. J. Cancer 118, 1471-1480 (2006).
19. J. D. Mainland, A. Keller, Y. R. Li, T. Zhou, C. Trimmer, L. L. Snyder, A. H. Moberly, K. A. Adipietro, W. L. Liu, H. Zhuang, S. Zhan, S. S. Lee, A. Lin, H. Matsunami, The missense of smell: functional variability in the human odorant receptor repertoire. Nat. Neurosci. 17, 114-120 (2014).
20. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Sci. Signal. 2, ra9 (2009).
21. K. Audouze, A. Tromelin, A. M. Le Bon, C. Belloir, R. K. Petersen, K. Kristiansen, S. Brunak, O. Taboureau, Identification of odorant-receptor interactions by global mapping of the human odorome. PloS One 9, e93037 (2014).
22. O. Man, Y. Gilad, D. Lancet, Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein Sci. 13, 240-254 (2004).

23. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: from sequence to structure. Protein Sci. 24, 1543-1548 (2015).
24. J. A. Ballesteros, H. Weinstein, Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: C. S. Stuart, Ed., **Methods Neurosci**., 25, 366-428 (1995).
25. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler, M. M. Babu, Molecular signatures of G-protein-coupled receptors. Nature 494, 185-194 (2013).
26. J. H. Park, T. Morizumi, Y. Li, J. E. Hong, E. F. Pai, K. P. Hofmann, H.-W. Choe, O. P. Ernst, Opsin, a Structural Model for Olfactory Receptors? Angew. Chem. Int. Ed. 52, 11021-11024 (2013).
27. C. de March, J. Topin, E. Bruguera, G. Novikov, K. Ikegami, H. Matsunami, J. Golebiowski, Odorant Receptor 7D4 Activation Dynamics. Angew. Chem. Int. Ed., 57, 4554-4558 (2018).

Supporting information

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. *For further references regarding sequence alignment please refer to Part IV of this manuscript, in the Methods section of "Genome-wide analysis of odorant receptors"*. 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.

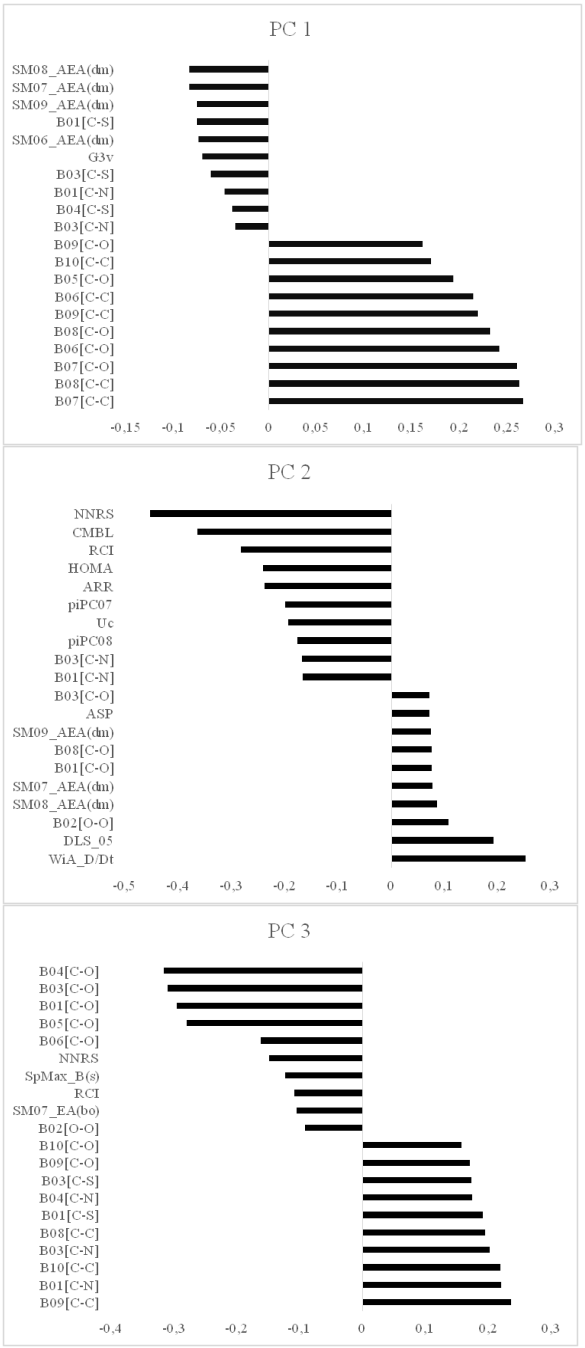


Figure S1. Descriptors having the largest weight in the 3 first principal components of the chemical space analysis. The descriptors meanings can be found at: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf

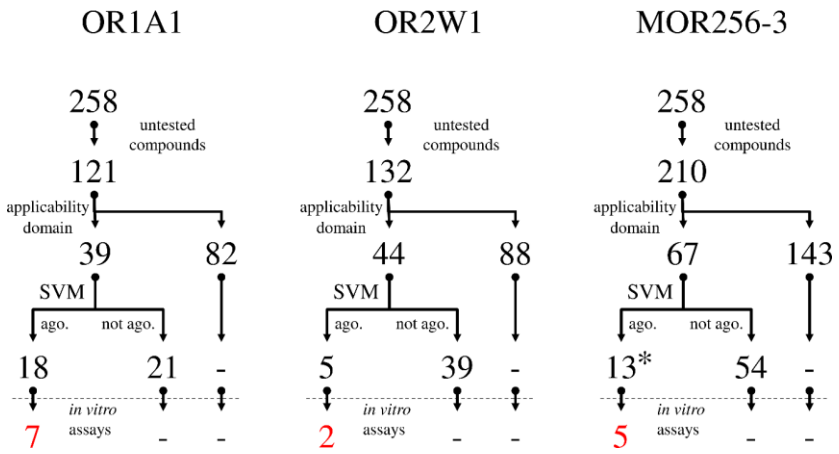


Figure S2. Workflow used for each additionally tested receptor (OR1A1, OR2W1, MOR256-3). In each case the compounds which had not been tested on the receptor were excluded, then the applicability domain of the model was defined, and predicted agonists were identified and tested *in vitro*. * means that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

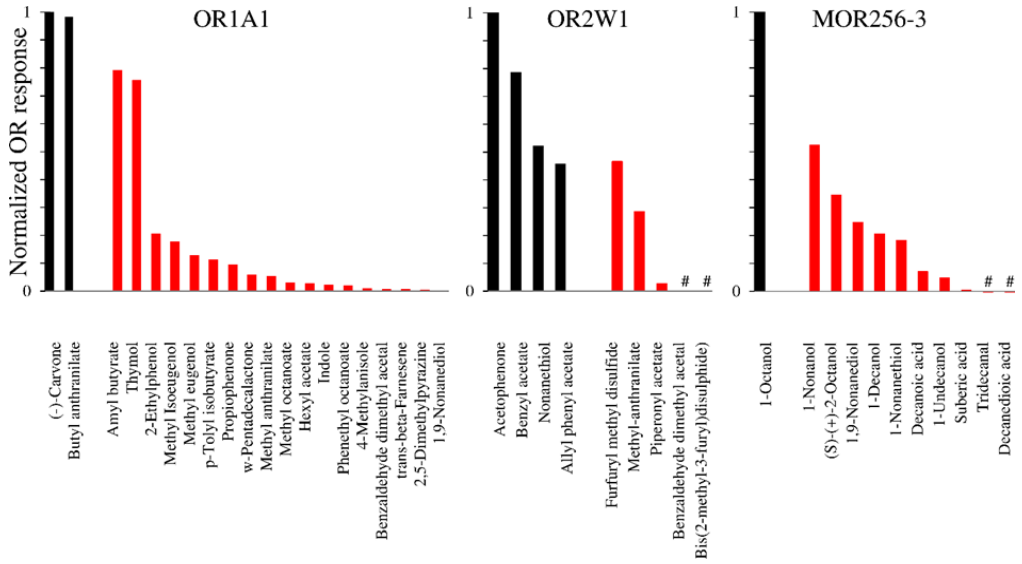


Figure S3. *In vitro* screening of agonists predicted (in red) by each model for OR1A1, MOR256-3 and OR2W1 and comparison with controls (in black). Seven novel agonists (triggering an OR response *ca*. 10% of the strongest control) are identified for OR1A1, 2 for OR2W1 and five for MOR256-3. # means 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.

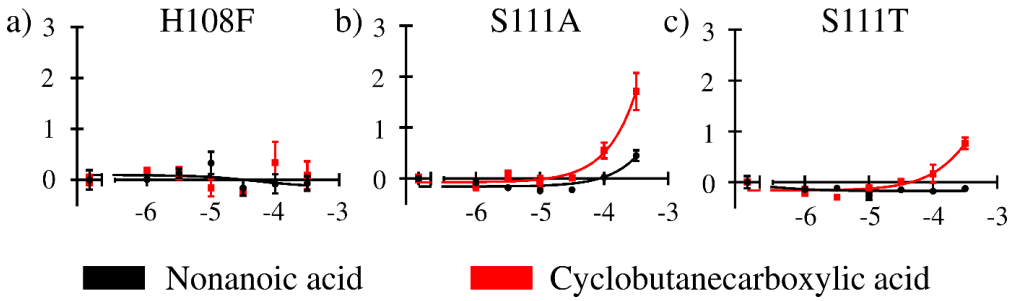


Figure S4. Dose response curves of OR51E1 a) H108F mutant, b) S111A mutant, c) S111T mutant.

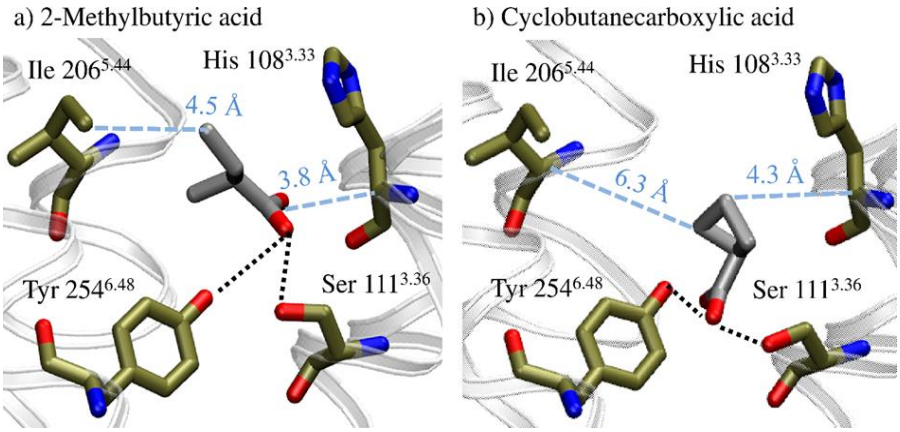


Figure S5. 2-methylbutyric acid shows closer contacts with H108 and I206 with respect to cyclobutanecarboxylic acid. a) Binding mode of 2-methylbutyric acid in the binding cavity of OR51E1. Carbon atoms are shown in gold (or gray in the case of the ligand), oxygen atoms are in red, and nitrogen atoms 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.

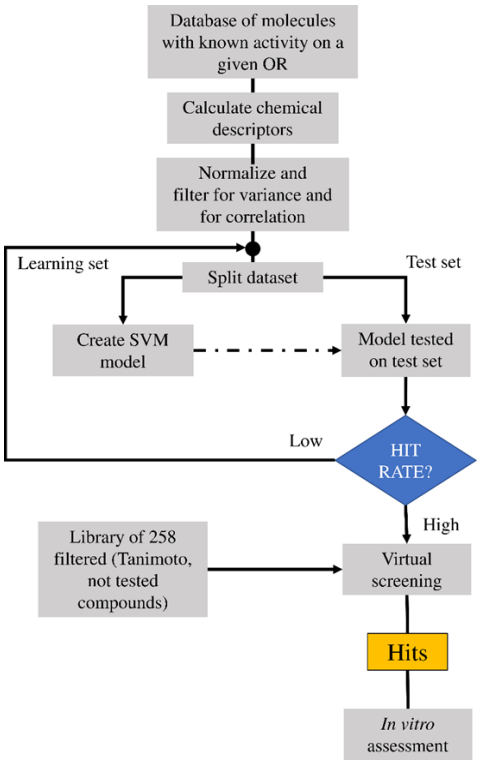


Figure S6. QSAR workflow.

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.

Files S1 and S2 can be found at: https://pubs.acs.org/doi/suppl/10.1021/acs.jpclett.8b00633

References

1. H. Zhuang, H. Matsunami, Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nat. Protoc. 3, 1402-1413 (2008).
2. A. Keller, H. Zhuang, Q. Chi, L. B. Vosshall, H. Matsunami, Genetic variation in a human odorant receptor alters odour perception. Nature 449, 468-472 (2007).
3. Y. R. Li, H. Matsunami, Activation state of the M3 muscarinic acetylcholine receptor modulates mammalian odorant receptor signaling. Sci. Signal. 4, ra1 (2011).
4. J. D. Mainland, A. Keller, Y. R. Li, T. Zhou, C. Trimmer, L. L. Snyder, A. H. Moberly, K. A. Adipietro, W. L. Liu, H. Zhuang, S. Zhan, S. S. Lee, A. Lin, H. Matsunami, The missense of smell: functional variability in the human odorant receptor repertoire. Nat. Neurosci. 17, 114-120 (2014).

5. TALETE srl. Dragon *Software for Molecular Descriptor Calculation* (2014).
6. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, **Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization**. Springer, Berlin Heidelberg, (2008).
7. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Sci. Signal. 2, ra9 (2009).
8. K. Audouze, A. Tromelin, A. M. Le Bon, C. Belloir, R. K. Petersen, K. Kristiansen, S. Brunak, O. Taboureau, Identification of odorant-receptor interactions by global mapping of the human odorome. PloS One 9, e93037 (2014).
9. K. A. Adipietro, J. D. Mainland, H. Matsunami, Functional evolution of mammalian odorant receptors. PLoS Genet. 8, e1002821 (2012).
10. Y. Fujita, T. Takahashi, A. Suzuki, K. Kawashima, F. Nara, R. Koishi, Deorphanization of Dresden G protein-coupled receptor for an odorant receptor. J. Recept. Signal Transduct. 27, 323-334 (2007).
11. C. Geithe, G. Andersen, A. Malki, D. Krautwurst, A Butter Aroma Recombinate Activates Human Class-I Odorant Receptors. J. Agric. Food Chem. 63, 9410-9420 (2015).
12. N. Jovancevic, A. Dendorfer, M. Matzkies, M. Kovarova, J. C. Heckmann, M. Osterloh, M. Boehm, L. Weber, F. Nguemo, J. Semmler, J. Hescheler, H. Milting, E. Schleicher, L. Gelis, H. Hatt, Medium-chain fatty acids modulate myocardial function via a cardiac odorant receptor. Basic Res. Cardiol. 112, 13 (2017).
13. A. Veithen, M. Philippeau, P. Chatelain, High-Throughput Receptor Screening Assay. In: A. Buettner, Ed., **Springer Handbook of Odor**, 505-525 (2017).
14. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: from sequence to structure. Protein Sci. 24, 1543-1548 (2015).
15. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinformatics Chapter 5, Unit 5-6 (2006).
16. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 31, 455-461 (2010).
17. C. Geithe, F. Noe, J. Kreissl, D. Krautwurst, 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 42, 181-193 (2017).
18. K. Schmiedeberg, E. Shirokova, H. P. Weber, B. Schilling, W. Meyerhof, D. Krautwurst, Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. J. Struct. Biol. 159, 400-412 (2007).
19. E. Block, S. Jang, H. Matsunami, S. Sekharan, B. Dethier, M. Z. Ertem, S. Gundala, Y. Pan, S. Li, Z. Li, S. N. Lodge, M. Ozbil, H. Jiang, S. F. Penalba, V. S. Batista, H. Zhuang, Implausibility of the vibrational theory of olfaction. Proc. Natl. Acad. Sci. U. S. A. 112, E2766-2774 (2015).
20. T. Braun, P. Voland, L. Kunz, C. Prinz, M. Gratzl, Enterochromaffin cells of the human gut: sensors for spices and odorants. Gastroenterology 132, 1890-1901 (2007).
21. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U. S. A. 112, 14966-14971 (2015).

Part III
Receptor-based understanding of olfaction using molecular modeling

"Corpora non agunt nisi fixata." – Paul Ehrlich

[A substance cannot act unless it is bound]

Studying odorant-OR interactions

Understanding the (bio)molecular basis of smell perception requires considering odorant receptors (ORs), but knowledge on the way they function remains scarce. Ligand-based understanding of olfaction can be enhanced by a more thorough model of the interactions that occur between odorants and receptors. In this part of the document, **I shift from a ligand-based paradigm to a receptor-based one**. To which extent can a 3D model of an OR predict the interactions with ligands? Are these models able to capture specific structural features which could be used to rationally design ligands? Such questions can hardly be answered by ligand-based approaches such as machine learning. In the following chapters, I will present the tools used during my PhD to study how ORs interact with odorants at the atomistic level.

Odorant receptors are G Protein Coupled Receptors

Odorant receptors belong to the family of class A G protein-coupled receptors (GPCRs). These types of receptors represent the most important family of membrane receptors and play a fundamental role in the communication of cells with their environment (1). Understanding the structural properties of these receptors is at the core of understanding the sense of smell and on getting a grasp on how molecular recognition functions.

The G protein-coupled receptor superfamily is very diverse in structure and function (2). Up to date, more than 800 human GPCR sequences have been identified. It is possible to further divide this superfamily into five main families: class A (or rhodopsin), class B (or secretin), class C (or glutamate), class F (or Frizzled/Taste) and Adhesion family (3). Most GPCRs belong to class A and this family comprises ORs.

From a structural point of view, GPCRs share a common architecture made up by seven trans-membrane (TM) helices which span the membrane and are connected by three intracellular and three extracellular loops (2). The N-terminal end is located at the extracellular side of the receptor, and the C-terminal end is at the intracellular side. (Fig. 1A) In GPCRs, three distinct functional regions are defined: 1) the extracellular part of the receptors, where ligand recognition takes place and where the largest sequence variability is observed (4) (Fig.1B – region in green), 2) the trans-membrane domain which experiences the largest conformational changes upon ligand binding and represents the structural hub of the protein, 3) the

intracellular section, which couples to signaling partners inside the cell, namely the G protein but also other partners involved in cell signaling such as β-arrestin (Fig.1B - depicted in blue).

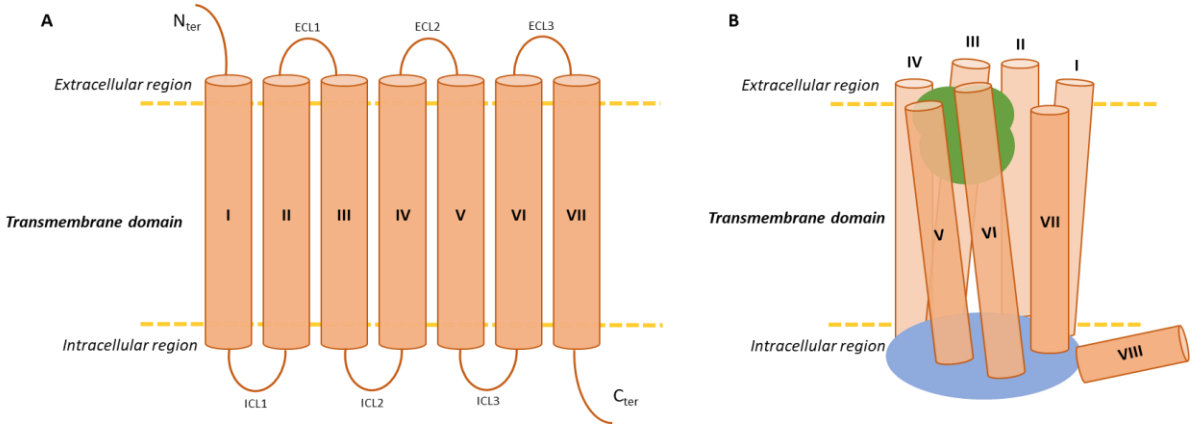


Figure 1: A. Topological representation of a GPCR. Its N-terminal extremity is located in the extracellular region, its C-terminal region in the intracellular region. The transmembrane (TM) domain are alpha helices which are connected by intracellular or extracellular loops (ICL or ECL, respectively). **B.** GPCRs form a structure similar to a barrel when folded. The C-terminal region often adopts a helical form which is termed “helix 8”. The orthosteric ligand binding site is depicted in green. In ORs, it is located between TMs III, IV, V and VI. G proteins and other intracellular partners bind the GPCR at the region depicted in blue.

The first crystallographic structure of a GPCR was made available in 2000 (5). Since then hundreds of structures of these proteins have been resolved with different types of agonists, antagonists, and even in complex with its intracellular partners: the G-protein (6) and β-arrestin (7). All structures exhibit a very conserved general fold and this enrichment in terms of availability of structure paves the way to building homology models of unresolved GPCR structures.

Molecular modeling of odorant receptors

No crystal structure is available for any OR. Molecular modeling is particularly useful to model their three-dimensional structures, based on available templates. Then, molecular dynamics describes at the atomistic level the dynamical behavior of proteins. Typically, molecular dynamics simulations produce data covering the microsecond timescale. Time-evolved snapshot of atomic coordinates, combined sequentially into sets called trajectories, result in detailed « movies » of how biomolecules behave over time under a variety of tunable conditions. Their dynamic behavior can thus be monitored and, in our case, gives insight into the mechanisms governing odorant recognition by ORs.

Linking the primary amino acid sequence of a protein to its structure is still challenging. Comparative modeling is certainly a good approach for getting an accurate model of unknown proteins. Applied to ORs, it becomes possible to obtain insight into the nature of the OR binding site and to make hypotheses on how information transmission takes place in these proteins. Briefly, all 397 human functional OR sequences are aligned to the sequence of class A GPCRs for which the experimental structure is known. *For further references regarding sequence alignment please refer to Part IV of this manuscript, in the Methods section of "Genome-wide analysis of odorant receptors"*. We selected bovine rhodopsin receptor (PDB id: 1U19), CXCR4 chemokine receptor (3ODU), human adenosine A2A receptor (2YDV), and human chemokine CXCR1 receptor (2LNL). Manual adjustments were performed to be consistent with more than 140 site-directed mutagenesis data points of the literature (8). Modeller (9) allows obtaining several homology models for one receptor which are all inspected individually. The best model is selected by ensuring its consistency with *in vitro* data and with several structural constraints.

These constraints are:

-no large folded structures in extra-cellular loops,

-all trans-membranes helices (TMs) should be α-helices,

-a small α-helix structure should be present in the C-terminal part of the receptor.

The best model is used for structural analysis, docking and molecular dynamics simulations.

Then, hypotheses on the role of specific residues, or specific motifs can be evaluated by the 3D model. Through collaborations with experimentalist partners, I had the opportunity to assess the conclusions made by my models through *in vitro* functional assays and site-directed mutagenesis.

In this part of my thesis, I focused on understanding how molecular recognition operates in ORs.

Binding sites in GPCRs

Broadly speaking, the orthosteric ligand binding is relatively well conserved amongst class A GPCRs (4). In most crystallographic structures, ligands are identified to contact TMs III, VI and VII (Fig.2). In addition to these consensus contacts, additional ones can be established with other TMs, as a function of the type of ligands. In PAR1 receptor for example, TM IV establishes ligand contacts (10). In the case of ORs, residues in TM V have also been shown to interact with the ligand (8, 11). This is also the case for rhodopsin, and β2-adrenergic receptor amongst others.

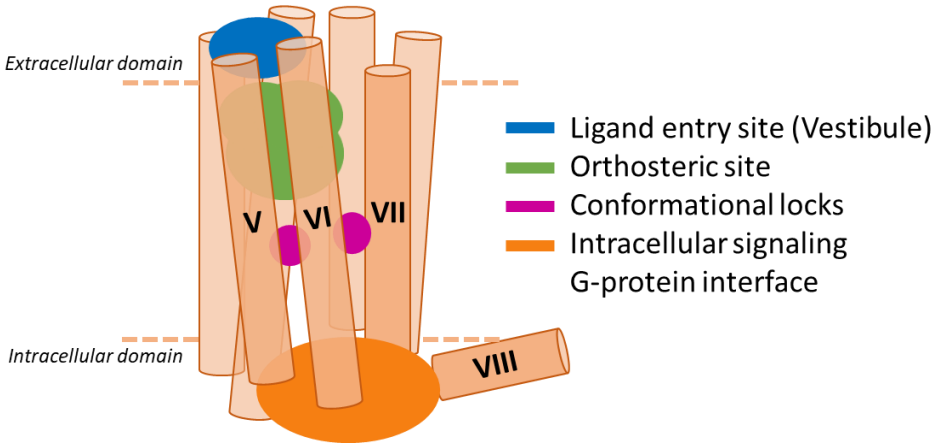


Figure 2: Schematic representation of different binding sites in a GPCR where allosteric interactions can occur.

In addition to this orthosteric site, allosteric binding sites have also been identified in class A GPCRs. These sites are topographically distinct from the orthosteric site and usually lead to a change in receptor conformation. Allosteric interaction can occur at the ligand entry site, at the conformational locks or in the intracellular protein interface (12).

Contributions

This part of the thesis gathers two articles. Article 4 is currently under revision while article 5 is in preparation.

Article 4 is based on a combination of genome-wide sequence analysis, molecular modeling, *in vitro* functional assays, and site-directed mutagenesis. It reports the discovery of a conserved and functional motifs in the extracellular domain of a sub-class of mammalian ORs. My contribution was to perform sequence analysis and a molecular dynamics simulation which showed that, when present, these motifs act as a vestibular binding site for the ligand.

In vitro results from our collaborators confirmed the functional role of residues forming this motif. I also analyzed the chemical space associated with these two sub-classes of ORs: I showed that ORs exhibiting this conserved vestibular site bind more hydrophilic chemicals. The presence of such a vestibular binding site might help explain why the properties of the ligands that are recognized by both classes of human ORs slightly differ. The *in vitro* mutational data was performed by our collaborators at Duke University

Article 5 aims to study the role of OR cavity and activation mechanism in ligand specificity. Despite a high sequence identity, a family of ORs exhibit different recognition spectra and different strengths of responses. To study the role of the cavity and that of the activation mechanism, we rationally designed the cavity of a narrowly tuned OR into the cavity of a broadly tuned OR. Our results suggest that the cavity composition is responsible for the recognition spectrum of an OR and for the basal activity. However, the strength of the response is mostly controlled by other mechanisms mostly excluding the role of the binding cavity. More generally, we show how affinity is not synonymous with strong response of an OR.

My contribution was to define the residues pointing into the binding cavity of each OR using sequence analysis and molecular modeling. I designed the mutations to be performed for fully reprogramming of the cavity. Our collaborators at UPenn produced the mutants while those at Duke University performed *in vitro* screenings of the wild-type and mutated ORs.

References

1. K. L. Pierce, R. T. Premont, R. J. Lefkowitz, Seven-transmembrane receptors. Nat. Rev. Mol. Cell Biol. 3, 639-650 (2002).

2. D. M. Rosenbaum, S. G. Rasmussen, B. K. Kobilka, The structure and function of G-protein-coupled receptors. Nature 459, 356-363 (2009).

3. R. Fredriksson, M. C. Lagerstrom, L. G. Lundin, H. B. Schioth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol. Pharmacol. 63, 1256-1272 (2003).

4. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler, M. M. Babu, Molecular signatures of G-protein-coupled receptors. Nature 494, 185-194 (2013).

5. K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, M. Miyano, Crystal structure of rhodopsin: A G protein-coupled receptor. Science 289, 739-745 (2000).

6. S. G. Rasmussen, B. T. DeVree, Y. Zou, A. C. Kruse, K. Y. Chung, T. S. Kobilka, F. S. Thian, P. S. Chae, E. Pardon, D. Calinski, J. M. Mathiesen, S. T. Shah, J. A. Lyons, M. Caffrey, S. H. Gellman, J. Steyaert, G. Skiniotis, W. I. Weis, R. K. Sunahara, B. K. Kobilka, Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 477, 549-555 (2011).

7. Y. Kang, X. E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T. A. White, O. Yefanov, G. W. Han, Q. Xu, P. W. de Waal, J. Ke, M. H. Tan, C. Zhang, A. Moeller, G. M. West, B.

D. Pascal, N. Van Eps, L. N. Caro, S. A. Vishnivetskiy, R. J. Lee, K. M. Suino-Powell, X. Gu, K. Pal, J. Ma, X. Zhi, S. Boutet, G. J. Williams, M. Messerschmidt, C. Gati, N. A. Zatsepin, D. Wang, D. James, S. Basu, S. Roy-Chowdhury, C. E. Conrad, J. Coe, H. Liu, S. Lisova, C. Kupitz, I. Grotjohann, R. Fromme, Y. Jiang, M. Tan, H. Yang, J. Li, M. Wang, Z. Zheng, D. Li, N. Howe, Y. Zhao, J. Standfuss, K. Diederichs, Y. Dong, C. S. Potter, B. Carragher, M. Caffrey, H. Jiang, H. N. Chapman, J. C. Spence, P. Fromme, U. Weierstall, O. P. Ernst, V. Katritch, V. V. Gurevich, P. R. Griffin, W. L. Hubbell, R. C. Stevens, V. Cherezov, K. Melcher, H. E. Xu, Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature 523, 561-567 (2015).

8. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: From sequence to structure. Protein Sci. 24, 1543-1548 (2015).

9. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinformatics Chapter 5, Unit 5-6 (2006).

10. C. Zhang, Y. Srinivasan, D. H. Arlow, J. J. Fung, D. Palmer, Y. Zheng, H. F. Green, A. Pandey, R. O. Dror, D. E. Shaw, W. I. Weis, S. R. Coughlin, B. K. Kobilka, High-resolution crystal structure of human protease-activated receptor 1. Nature 492, 387-392 (2012).

11. C. Bushdid, C. A. de March, S. Fiorucci, H. Matsunami, J. Golebiowski, Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features. J. Phys. Chem. Lett. 9, 2235-2240 (2018).

12. M. Congreve, C. Oswald, F. H. Marshall, Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs. Trends Pharmacol. Sci. 38, 837-847 (2017).

13. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U. S. A. 112, 14966-14971 (2015).

Publication 4

Mammalian class I odorant receptors exhibit a conserved vestibular binding pocket

CAROLINE BUSHDID, CLAIRE A. DE MARCH, JÉRÉMIE TOPIN, MATTHEW DO, HIROAKI MATSUNAMI, JÉRÔME GOLEBIOWSKI

UNDER REVISION

Abstract

Odorant receptors represent the largest family of mammalian G protein-coupled receptors. Phylogenetically, they are split into two classes (I and II). By analyzing the entire subclass I odorant receptors sequences, we identified two class I-specific and highly conserved motifs. These are predicted to face each other at the extracellular portion of the transmembrane domain, forming a vestibular site at the entrance to the orthosteric binding cavity. Molecular dynamics simulation combined with site-directed mutagenesis and *in vitro* functional assays confirm the functional role of this vestibular site in ligand-driven activation. Mutations at this part of the receptor differentially affect the receptor response to four agonists. Since this vestibular site is involved in ligand recognition, it could serve ligand design that target specifically this sub-genome of mammalian odorant receptors.

Introduction

Mammals sense their chemical environment when volatile molecules activate the odorant receptors (ORs) embedded in the membrane of their olfactory sensory neurons [1]. ORs belong to the class A G protein-coupled receptors (GPCRs) and represent more than 3% of all protein-coding genes [2]. Structurally, class A GPCRs are made up of seven transmembrane helices (TM1−TM7) forming a bundle within the cell membrane, connected by intra- and extra-cellular loops. Class A GPCRs show a remarkably conserved fold and, from a sequence point of view, they also share typical hallmark motifs that are distributed at various locations within their sequences [3].

In the absence of any OR experimental structure, molecular modeling has been the tool of choice for studying structure-function relationships in ORs. Comparative modeling benefits from the high level of conservation in the structural fold of class A GPCRs. Combined with *in vitro* site-directed mutagenesis and functional assays, it has proven extremely powerful for predicting various OR features, such as ligand binding and selectivity, and selective activation by agonist bound within their orthosteric binding site [4-9].

In ORs, the orthosteric binding cavity exhibits the greatest variability,[10] endowing mammals with an extraordinarily broad chemical discriminatory power. Phylogenetic classification of the mammalian olfactory receptor genome, when compared to those found in Xenopus, splits this family of receptors into two distinct classes [11-13]. In humans, the 58 class I ORs represent ~15% of the 397 receptors that are considered functional. Class I ORs

are also referred to as “fish-like” since they resemble the family that was first identified in aquatic animals such as fish and frog. They were considered to be ancestral ORs that were maintained in mammals [12]. The second class of ORs (class II or “tetrapod”) are thought to have appeared during tetrapod terrestrial evolution. However, the diversity of mammalian ORs remains much smaller than that found in fishes, where many more classes have been identified [14]. Interestingly, during evolution, class I ORs were maintained in a single cluster as opposed to class II OR genes which spread over most chromosomes [15]. Class I odorant receptor genes also share a specific and conserved genetic mechanism regulating the allelic exclusion during expression [16].

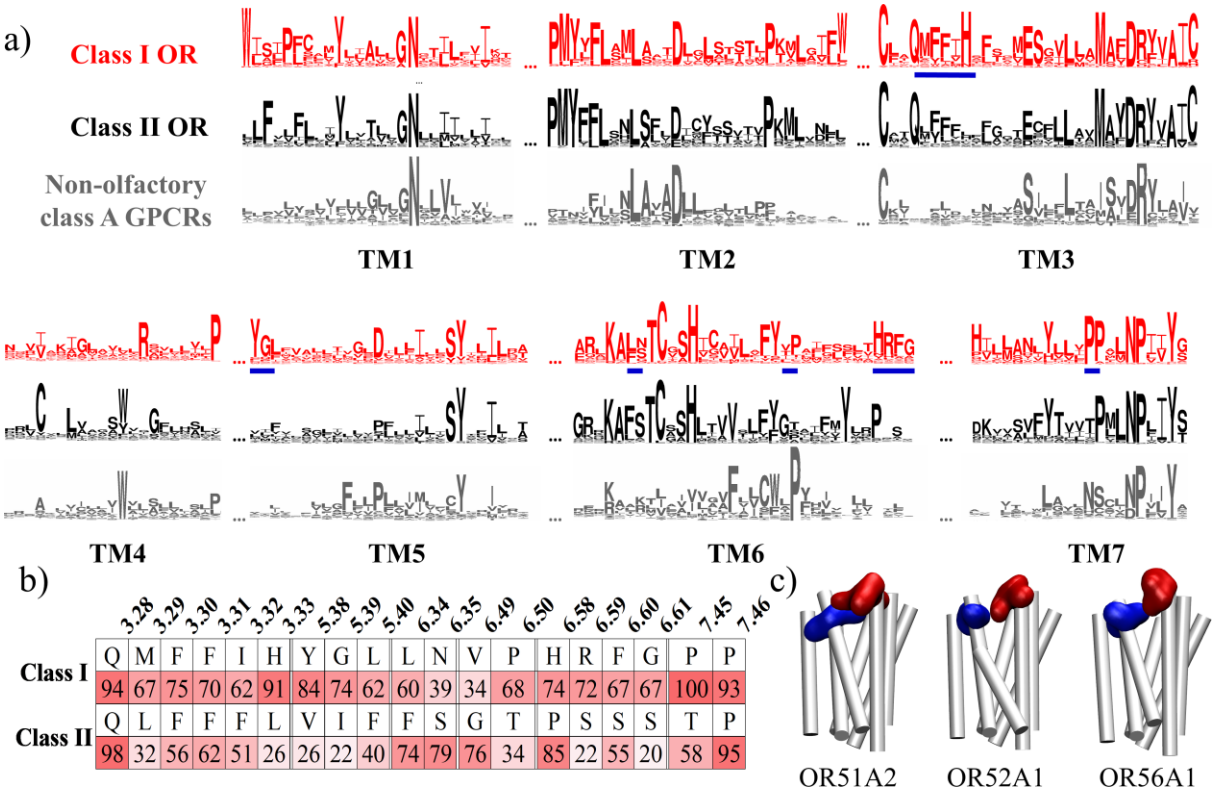
In mammals, class I and class II, despite sharing all the typical OR sequence-specific hallmarks, can be clearly distinguished by some characteristic features that are highly conserved within their sequences [13].The extracellular part of TM6 has been suggested to play an role in the differential ligand recognition between class I and class II ORs [17], although the associated structural features are yet to be uncovered. Between the two classes, no difference could be identified concerning the breadth of tuning, the number of agonists, or the sensitivity [18].

Extracellular allosteric sites have been identified in numerous class A GPCRs [19], such as the purinoceptor 1 (P2Y1 [20]), sphingosine-1-phosphate receptor 1 (S1P1[21]), β2-adrenergic receptor [22], and the muscarinic acetylcholine receptors [23]. More specifically, a so-called “vestibular binding site” has been identified at the top of the transmembrane domain in the muscarinic acetylcholine receptors [23], in the β2 adrenergic receptor [22] experimental structures, as well as in molecular models of the δ-opioid receptor. [24] Focusing on chemosensory receptors, such a site was reported in the bitter taste receptor TAS2R46[25] and in the trace amine associated receptor TAAR13c.[26] These discoveries laid the groundwork for the design of highly selective allosteric modulators. This type of vestibular binding site has never been identified in any OR.

In this article, we report the identification of two class I OR-specific motifs that face each other at the extracellular extremities of TM5 and TM6. These motifs are conserved in class I ORs but not in the class II. Molecular dynamics simulations showed that these motifs form a vestibular site at the entrance to the orthosteric binding cavity. Site-directed mutations at the vestibular binding site affected the receptors’ response to odorant stimulation *in vitro*, confirming the functional role of this vestibule.

Results

ORs show distinct specific sequence motifs. Figure 1 reports the conservation analysis of human GPCR sequences. Here, 397 human ORs were compared with 204 non-olfactory class A GPCRs. As observed previously in mice [13], both OR classes show identical conserved regions such as the PMYxFL motif in TM2, MAYDRYVAIC in TM3, SY residues in TM5, RxKAxxTCxSH and FY in TM6. Other motifs considered to be the hallmarks of the GPCR class A family (GN in TM1, DRY in TM3, KA in TM6, and NPxxY in TM7) can also be identified (Fig. 1a).



	3.28	3.29	3.30	3.31	3.32	3.33	5.38	5.39	5.40	6.34	6.35	6.49	6.50	6.58	6.59	6.60	6.61	7.45	7.46
Class I	Q	M	F	F	I	H	Y	G	L	L	N	V	P	H	R	F	G	P	P
	94	67	75	70	62	91	84	74	62	60	39	34	68	74	72	67	67	100	93
Class II	Q	L	F	F	F	L	V	I	F	F	S	G	T	P	S	S	S	T	P
	98	32	56	62	51	26	26	22	40	74	79	76	34	85	22	55	20	58	95

Figure 1. Conserved motifs in class I and class II human ORs and comparison with non-olfactory class A GPCRs. a) The logos summarize the conservation amongst 397 human OR sequences (58 are class I ORs and 339 are class II) and 204 non-olfactory class A GPCRs [10,27]. Major differences are underlined in blue. b) Highest conservation of class-specific motifs in TM3, TM5, TM6, and TM7. c) Three class I human OR structures (representative of each class I subgroup) showing the position of conserved motifs in YGL^{TM5} (blue) and HRFG^{TM6} (in red). Both motifs are predicted to face each other, in these representations intra- and extracellular loops are omitted for image clarity.

Once the typical class A GPCRs motifs are excluded, the hallmarks of class I and class II ORs become quite distinct (Files S3 and S4). Specific conserved motifs can be distinguished in both classes (Fig. 1b and c). As a general rule, TM3 of class A GPCRs is a structural and

functional hub [3]. Class I ORs show a highly conserved QM$^{3.29}$FFxH$^{3.33}$ motif (superscripts refer to the Ballesteros-Weinstein notation)[28] where the polar methionine and histidine residues are predicted to point towards the inside of the binding cavity of class I ORs [10,29]. These two conserved and hydrophilic residues (M$^{3.29}$ and H$^{3.33}$ respectively conserved at 67% and 91%, Fig. 1b) are not conserved in class II ORs (where only a conserved QxFFxx signature is found).

In TM5, a YGL$^{5.40}$ motif is specific to the extracellular side of class I ORs conserved at more than 60% for each residue, whereas in class II ORs the conservation rate at each position drops to less than 40% (Fig. 1b). TM6 also presents notable differences in conservation in the extracellular side of the receptors between both classes. A HRFG$^{6.61}$ motif, where each residue is conserved at about 70%, is specific to class I, but this motif is replaced by a conserved PxSx motif in class II.

Focusing on the middle of TM6, the highly conserved FY$^{6.48}$XX motif is shown to act as a toggle-switch in ORs [6] and is conserved as a FYxP motif in class I ORs. At the N-terminal part of TM6 (RxKAFSTCxSH motif), a highly conserved FS motif in class-II ORs is replaced by an LN motif in class I. Note that this motif was shown to be crucial in receptor activation and dynamics [4,30]. Finally, as previously reported in mice [13], three conserved proline residues in TM7 are present in class I ORs while only two can be found in class II. TM4 and TM5 are connected by the extracellular loop 2 (ECL2). Although this structure surely plays an active role in ligand recognition,[31] it may not be specific to class I or class II ORs given its poor conservation amongst mammalian ORs.

The conserved motifs YGL^{TM5} and HRFG^{TM6} form a vestibular site

In human, class I ORs are phylogenetically split into three sub-groups, OR51, OR52, and OR56.[12] YGL^{TM5} and HRFG^{TM6} are conserved for the two ORs that are representative of the two first groups (OR51A2 and OR52A1). In the third group (OR56 sub-family comprises 6 ORs), the typical OR56A1 receptor shows slightly different motifs in TM5 and TM6 (YQF^{TM5} and NLAR^{TM6}, see File S1), which nonetheless face each other as well. Above all, this highlights that the structural feature made up of these facing motifs is observed amongst the whole class I OR sub-genome (Fig. 1c). We focused specifically on understanding the functional role of such conserved and class-specific motifs in TM5 and TM6 extracellular regions. OR51E1 was considered prototypical because its sequence shows all the signatures of class I ORs and it has already been thoroughly studied, both *in vitro* and *in silico* [29,32,18]. The model was built using a previous protocol, where homology modeling was

confirmed to fulfill constraints obtained by functional *in vitro* assays [29]. In total, 141 mutated positions on mammalian ORs where gathered [4], covering 35% of the whole sequence of the receptor. As expected, YGL^{TM5} and HRFG^{TM6} form the last extracellular helical turn of TM5 and TM6, respectively, and face each other.

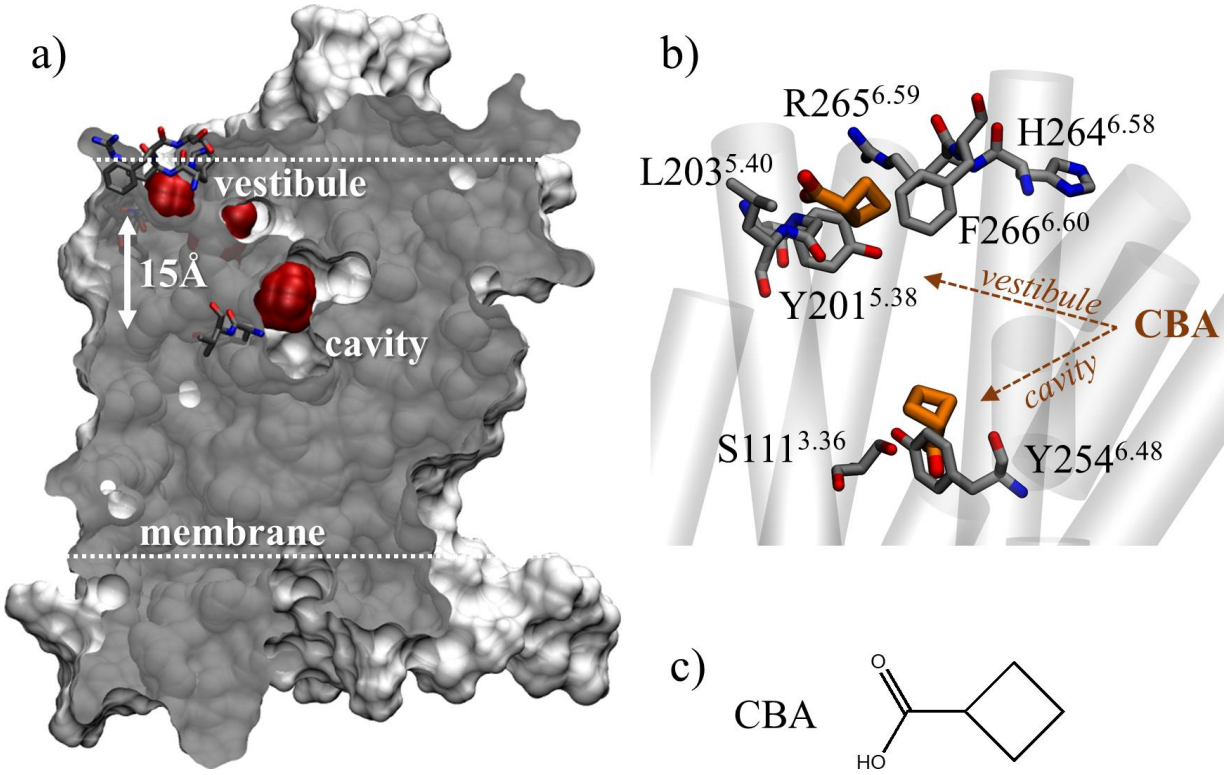


Figure 2. **OR51E1 conserved motifs form a vestibular binding site visited by ligands.** a) Cross section of OR51E1 van der Waals volume (gray). Extracellular loops were modeled but are omitted for image clarity. A vestibular binding site and the orthosteric binding cavity (red) are detected by a cavity detector. Y254$^{6.48}$ and V255$^{6.49}$ at the orthosteric cavity and HRFG^{TM5} and YGL^{TM6} forming the vestibule are shown in licorice. b) Superpositions of typical positions of CBA (shown in brown) at the cradle of the orthosteric binding cavity (S111$^{3.36}$ and Y254$^{6.48}$) and at the vestibular site. c) Chemical structure of CBA.

The two motifs are located at the solvent/membrane interface. The vestibule is predicted to be 2.5 helical turns (approximately 15 Å) above the orthosteric binding pocket (Fig. 2a). Four known agonists of OR51E1, cyclobutanoic acid (CBA), butyric acid (BA), isovaleric acid (IVA) and nonanoic acid (NA) were docked into the orthosteric binding site and the vestibule. The four acids with varied chemical propertied showed comparable affinities for both sites (Table S1 and Fig. 2b), suggesting that the vestibule is likely to interact with agonists.

At the orthosteric cavity, the agonists established contacts with S111$^{3.36}$, I206$^{5.44}$ and H108$^{3.33}$, as we have previously shown [29]. The ligands locations are consistent with the binding poses predicted by several ligand-OR interaction studies [29,4,10,8,7,9,33].

The dynamics of the four systems was investigated through molecular dynamics (MD) simulations. In three independent MD simulations, CBA visited both sites during the same run (Fig. 2, 3, and S1). The same behavior is observed for the three other agonists (Fig. S1), confirming the connection between the vestibular site and the orthosteric cavity.

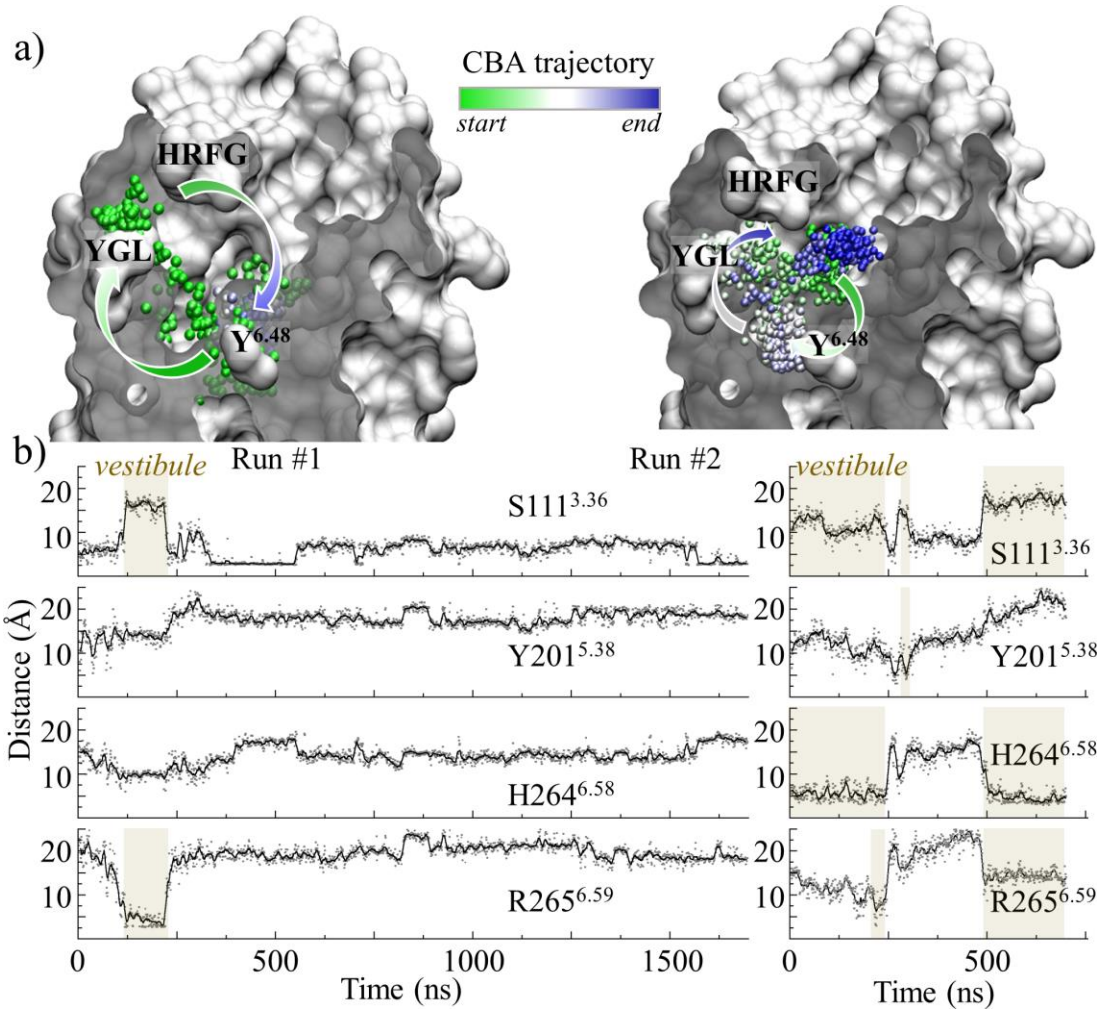


Figure 3. The vestibule is connected to the orthosteric binding cavity. a) Locations of CBA (as shown by a sphere corresponding to its carboxylic atom) during two independent MD trajectories initiated either at the orthosteric cavity (Run #1, left) or at the vestibule (Run #2, right). Green to blue colors indicate simulation time evolution. b) Distance analysis between a CBA oxygen atom and donor/acceptor groups of S111 (bottom of the cavity), and Y201, H264, and R265 (forming the vestibule). In both simulations, CBA engages opportunistic hydrogen bonds with residues forming the vestibule. Periods during which CBA visits the vestibule are highlighted.

We have previously found that S111$^{3.36}$ was involved in agonist binding to OR51E1 [29]. When the agonist are hydrogen-bonded with S111$^{3.36}$, they are close to the conserved motif (FY$^{6.48}$VP in class I or FY$^{6.48}$GT in class II ORs) acting as the activation toggle-switch of the receptor [6]. Although residues 255$^{6.49}$ and 256$^{6.50}$ do not strongly interact with the ligands

(Fig. S1), structural perturbations at these positions affect the general ability of the receptor to sense agonists and to activate [6,5]. The V255$^{6.49}$G and P256$^{6.50}$A mutations consistently modified the receptor's basal activity and decreased its efficacy (Fig. 4a,d).

In general, mutating the residues identified either at the orthosteric cavity or at the vestibule affected the basal activity (Fig. 4) without apparent impact on the surface expression (Fig. S2). This highlights the role of the vestibular site in the activation mechanism.

During the course of the MD simulations, the agonists formed transitory hydrophilic or hydrophobic contacts with residues belonging to the orthosteric cavity and the vestibule, whatever the starting point of the ligand (Fig. 3 and Fig. S1). This suggests an active role of these residues in ligand recognition. In the vestibule, transitory hydrogen bonds could be observed between the agonists and R265$^{6.59}$, H264$^{6.58}$, or Y201$^{5.38}$, as illustrated for CBA in Figure 3.

Mutations at the vestibular site differentially affect agonists recognition *in vitro*

The opportunistic and transitory agonist-vestibule interactions observed during the MD simulations underline the role of the vestibule in agonist recognition. R265A$^{6.59}$ as well as Y201$^{5.38}$A mutant ORs were no longer responsive to the agonists *in vitro* (Fig. 4a), confirming the crucial roles of these residues in ligand-driven receptor activation.

The EC50 confidence intervals for each ligand and each mutant (Tables S2 and S3) are compared in Fig. 4c to examine the changes in potency upon mutation. The V255$^{6.49}$G mutation has a negligible effect on the potency of all the agonists, consistent with the orientation of V255$^{6.49}$ towards the membrane. However, different modulations are observed when mutating residues located at the vestibule, namely L203$^{5.40}$, H264$^{6.58}$, and F266$^{6.60}$. All mutant ORs are less responsive to agonists but each mutation affects each ligand in a different manner. This suggests that the vestibular residues do interact with the ligands. Each residue of the vestibule thus plays a role in the ligand recognition process, which naturally depends on the chemical property of the stimulating odorant. (Fig. S3 and Table S2)

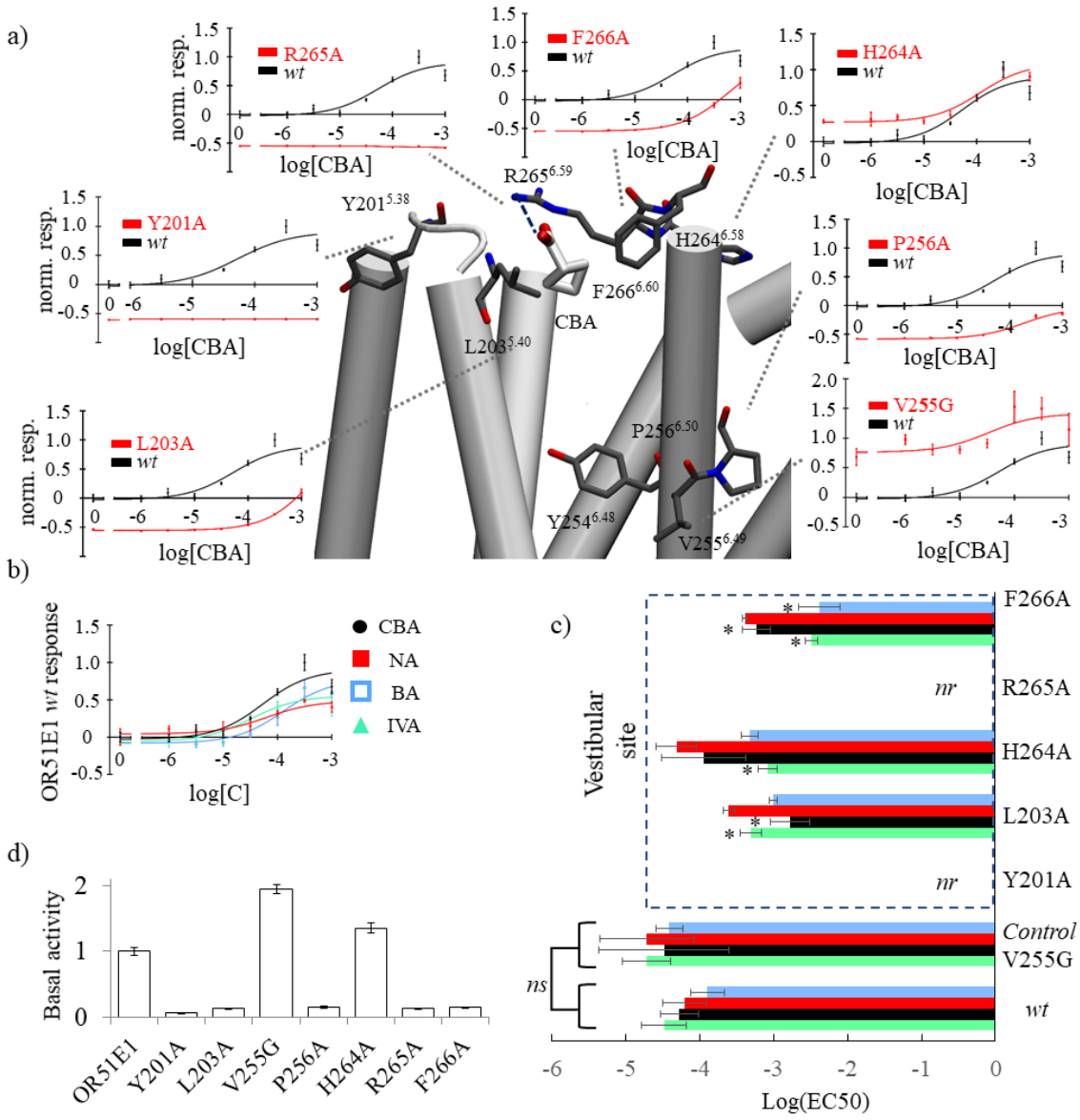


Figure 4. *In vitro* assessment of the functional role of the vestibule and the binding cavity. a) a snapshot where CBA bound at the vestibular site can form a hydrogen bond with R265$^{6.59}$ (blue dashed line) and hydrophobic contacts with L203$^{5.40}$ and F266$^{6.60}$. In this representation extracellular loops are omitted for image clarity. For each residue near the agonist, normalized dose-response curves of the response of OR51E1 mutants (red) in comparison with the *wt* (black) to CBA stimulation is shown. The y-axis represents the firefly luminescence normalized to Renilla firefly and normalized to the *wt* response. b) Normalized dose-response curves for OR51E1 *wt* to nonanoic acid (NA), isovaleric acid (IVA), and butyric acid (BA). c) Log(EC50) values for the four agonists. Comparison between the *wt* and the mutant ORs at the position V255$^{6.49}$ and at the vestibular site. '*nr*' indicates non-responsive. Mutation V255$^{6.49}$G, which does not interact with agonists has no statistically significant (ns) effect on the EC50 for any agonist. Mutations at the vestibular site differentially affect potency. A '*' indicates that the 95% EC50 confidence intervals do not intersect (Table S3). The H264$^{6.58}$A mutation only affects IVA binding while the F266$^{6.60}$A mutation affects all ligands except NA. The L203A$^{5.40}$ mutation mostly decreases both CBA and IVA potencies d) Normalized basal activity for all mutant ORs normalized to *wt*.

Discussion.

Class I odorant receptor sub-genome exhibit a vestibular cavity. Sequence analysis of human class I and class II ORs showed that both classes have distinct signatures in their sequences, as previously observed in mice [13]. Motifs within the ECL3 were considered to play an active role in ligand recognition by class I ORs [17]. Interestingly, we identified two motifs in the extracellular domain of class I ORs that face towards each other and are not conserved amongst class II ORs. Site-directed mutagenesis and functional assays determined that these motifs form a functional allosteric extracellular binding cavity, 15 Å removed from the canonical cavity of ORs. This structural feature is likely to be observed within the whole class I sub-genome.

Allosteric binding sites have been identified in many GPCRs from distinct classes [19]. In ORs notably, some residues located in helix 8 were reported to play a functional role in response to ligand binding [34]. In the case of extracellular allosteric sites and more specifically in muscarinic acetylcholine receptors, the vestibular site has been extensively studied since these receptors have strictly conserved binding sites, while the vestibule is an allosteric pocket that acts as a potential subtype selectivity filter [35]. Class I ORs show the opposite mechanism, where the vestibule appears to be highly conserved while the binding pocket (either in class I or class II) is strongly divergent. Coming back to the vestibular binding site, class II ORs, do not show any conservation at these positions. This could be connected with observed differences in ligand properties within the two classes.

From a sequence point of view, the most comparable vestibular motifs to those reported herein are found in the β2-adrenergic receptor (β2AR) vestibular entry site [22]. In the β2AR, the OR-equivalent motif of $YGL^{TM5\text{-}OR}$ is $YAI^{TM5\text{-}\beta2AR}$, while the $HRFG^{TM6\text{-}OR}$ equivalent is $HVIQ^{TM6\text{-}\beta2AR}$. These residues, as well as others located in the extracellular domain, interact with the agonists in the very first steps of the binding of alprenolol to the receptor [22]. Namely, $A200^{5.39}$, $H296^{6.58}$ and $V267^{6.59}$ interact with the ligand through hydrophobic interaction. An analysis of the structures of class II ORs studied by us previously did not reveal the presence of any vestibular cavity [33,36,4]. Hence, the presence of a vestibule in the class I olfactory sub-genome suggests that the function of class I ORs lies between class A GPCRs (where vestibules can be identified) and class II ORs.

Molecular dynamics simulations of OR51E1, a prototypical class I OR, showed that the motifs facing each other in the three-dimensional structure constitute a similar vestibular binding site as those found in the β2AR. In the simulation the ligands transiently visited the

vestibule highlighting the connection that exists between the vestibule and the orthosteric binding site. Docking experiments suggested that the vestibular binding site had a comparable binding affinity for the agonist with respect to the orthosteric binding cavity. The functionality of the vestibule was assessed through *in vitro* experiments. Here we show that mutation of the residues identified as being in contact with the ligand during molecular dynamics simulations strongly affected the response of the receptor. The residues constituting this vestibular pocket play an important role in agonists recognition.

Class I receptors have already been shown to bind more hydrophilic compounds than class II ORs [18]. We have revisited and updated all ligands considered to be agonists for any human OR (File S3). A statistical analysis of the agonist chemical properties shows that class I and class II ORs agonists have noticeably different properties (Fig. S4 and S5, Tables S4 and S5) [18,17,37]. Typically, class I OR agonists are more hydrophilic than the class II. Counterintuitively, an analysis of the residues lining the orthosteric binding cavities of class I ORs shows that they are less hydrophilic than those of class II ORs (Fig. S6). The vestibular site could act as a molecular sieve in class I ORs to favor hydrophilic ligands.

Beyond providing insight into the binding process of class I OR ligands, the identification of these TM5 and TM6 conserved motifs lays the groundwork for the rational design of allosteric modulators for the entire olfactory class I OR sub-genome. Such modulators will be of broad interest beyond olfaction since this class of receptors is expressed not only in the nose. They are found in numerous non-olfactory tissues, although their functional roles and expression levels remain to be established in several cases. More specifically, six class I ORs were shown to be the amongst the 40 most highly expressed ORs outside of the nose. These include, OR51E1 and OR51E2, OR52N4, OR52B6, OR52D1 and OR51B5 [38]. In addition, 53 class I human ORs transcripts were found to be expressed in polymorphonuclear leukocytes, B and T cells, whereas class II ORs were not detected in these cell types [39]. This wide-spread expression and evidence for important regulatory roles in various diseases makes class I ORs interesting and novel pharmacological targets. The design of ligands with specific interactions with this allosteric cavity would thus be extremely useful.

Materials and Methods

Chemical space analysis

To examine the characteristics of the molecules that class I and class II ORs respond to, all the data concerning deorphanization of human ORs was gathered. All references regarding OR deorphanization data and used in this study can be found in the Supplementary Material (File S3). Only molecules eliciting agonist activity on ORs were considered for further analysis: 92 were identified as agonists for class I ORs and 189 as agonists for class II ORs. We calculated 4884 chemical and topological descriptors using Dragon Software [40] on this set of molecule. We extracted eight descriptors that are of interest due to their pharmacological importance (they relate to the well-known *Lipinski's* rule of five) (File S4). The descriptors were normalized and averaged to create the spider plot. An average was identified as significantly different with a one-way ANOVA test, using $p < 0.05$.

Sequence logos

Multiple sequence alignment was performed using Jalview [41]. 397 sequences of human ORs were gathered and aligned as well as 204 non-olfactory class A GPCR sequences. The alignment was then split into three: 58 sequences were used for generating the class I sequence logo, 339 sequences were used to generate class II sequence logo, and 204 sequences for other class A GPCRs. Phylogenetic trees and secondary structure prediction were performed using the Jalview built-in Web Service application. Logos were generated using WebLogo. Sequence alignments for class I and class II ORs are provided in separate *fasta* files (Files S1 and S2).

Conservation analysis

Information about the conservation rate of each amino acid making up underlined motifs is given in Figure 1. These conservation rates were obtained by using Jalview [41] and the most conserved residue and its associated value are reported.

Human OR51E1 3D model

The protocol follows a previously published method [10]. Briefly, all 397 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 data from the 141 mutants previously described in 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 [42]. The N-terminal structure was omitted to avoid perturbing the modeling protocol. Five models were obtained and the one that was consistent with the *in vitro* data and several structural constraints (no large folded

structure in extra-cellular loops, all trans-membranes helices (TMs) folded as α-helices, and a tiny α-helix 8 at the C-terminal extremity) was kept.

Cyclobutanoic acid (CBA), nonanoic acid (NA), isovaleric acid (IVA), and butyric acid (BA) structures 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 $H108^{3.33}$, $Y254^{6.48}$ and $F257^{6.51}$ with Autodock Vina for the docking in the orthosteric binding cavity and $Y201^{5.38}$, $H264^{6.58}$ and $R265^{6.59}$ for docking in the vestibular binding site [43]. In each site, all docking poses were similar, and we considered the one with the lowest binding free energy for simulations.

Ligands binding poses considered for the MD simulations are given in Table S6. The cavity volumes were analyzed with MDPocket [44].

Molecular dynamics

The OR51E1 model was embedded into a model membrane. Its orientation in the membrane was determined using OPM server [45]. The simulation box is made of POPC lipids solvated using TIP3P water molecules in Maestro [46]. The total system is made up of ~36,000 atoms in a periodic bow of 74*59*89 $Å^3$. Molecular dynamics simulations were performed with sander and pmemd.cuda modules of AMBER16 with the ff14SB force-field for the protein, and the lipid14 for the membrane, and the gaff2 force-field for the ligands. Bonds involving hydrogen atoms are constrained using the SHAKE algorithm and long-range electrostatic interactions are handled with Particle Mesh Ewald (PME). The cut-off for non-bonded interactions is set to 8 Å. With CBA, an alternative run (Run # 3) with a cutoff set to 10 Å was also run for comparison purposes. Similar findings were obtained, i.e. a visit of both the vestibule and the orthosteric cavity (CBA run #3 behavior is shown in Fig. S1). MD simulations were stopped once the ligand sampled both the orthosteric cavity and the vestibule.

Temperature in kept constant in the system using a Langevin thermostat with a collision frequency of 2 ps^{-1}. In addition, a weak anisotropic algorithm with a relaxation time of 1 ps^{-1} is applied to keep a constant pressure. Snapshots are saved every 20 ps. The workflow used for energy minimization, thermalization, equilibration, and production of molecular dynamics simulations is detailed in Fig. S7. The total simulation time for this study is 5.6 µs. The RMS deviations of OR51E1 during all molecular dynamics simulations are shown in Fig. S8. The distance analysis was performed considering heteroatoms of each agonist and the closest H-bond donor/acceptor heteroatom of residue S111, Y201, H264, and R265.

Site directed mutagenesis

The coding sequence of OR51E1 was cloned into a pCI vector (Promega) and tagged at the N-terminal with the 20 first amino acids of rhodopsin. Site-directed mutagenesis was performed by Phusion DNA polymerase (NEB) [29]. The sequence of all plasmids was validated using the BigDye Terminator Sequencing Kit (Applied Biosystem).

Dual luciferase reporter gene assay

The Dual luciferase reported gene assay was used to evaluate the functionality of wild-type and mutant clones of OR51E1 in an *in vitro* system [29,47]. Hana3A cells have been cultured and plated the day before transfection at 1/10 of a 100% confluence 100mm plate into 96-well plates coated with poly D lysine. After overnight incubation, the required genes were transfected using, for each plate, 5 ng SV40-RL, 10 ng CRE-Luc, 5 ng human RTP1S [48], 2.5 ng M3 receptor [49] and 5 ng of receptor (OR51E1 wt or mutant) plasmid. After around 18h of transfection, cells were stimulated during 3.5 hours by 25 µL of odorant diluted in CD293 + glutamine + CuCl2. The luminescence of Firefly (Luc) and Renilla (Rluc) luciferase were then sequentially monitored by injecting the corresponding substrate. The activity in each well was normalized as (Luc-400)/(Rluc-400). The response of a receptor was also normalized to its basal activity as (NLX/NL0)-1 where NL0 is the normalized luminescence value at 0 µM of odorant and NLX the value at X µM.

Cell surface expression

Fluorescent Activated Cell Sorting (FACS) was conducted to evaluate cell surface expression of OR51E1 and mutants. Hana3A were seeded in a 35mm dish (Corning) in Minimum Essential Medium containing 10% FBS (M10). Lipofectamine2000 (Invitrogen) was used for transfection of plasmid OR and RTP1s DNA. At the time of transfection, green fluorescent protein (GFP) expression vector and RTP1s were co-transfected to monitor and improve the transfection efficiency. About 24 hrs post-transfection, cells were incubated 30min with PBS containing anti Rho-tag antibody 4D2 (gift from R. Molday), 15mM NaN3, and 2% FBS and then washed and incubated 30min with phycoerythrin (PE)-conjugated donkey anti-mouse IgG (Jackson Immunologicals). 7-amino-actinomycin D (7-AAD; Calbiochem), a fluorescent, cell-impermeant DNA binding agent, was added before flow cytometry to eliminate dead cells from analysis as 7-AAD selectively stains dead cells. The intensity of PE signal among the GFP-positive population was measured and plotted to evaluate the OR expression to the plasma membrane.

References

1. L. Buck, R. Axel, A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175-187 (1991).
2. Y. Niimura, Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Curr. Genomics 13, 103-114 (2012).
3. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler, M. M. Babu, Molecular signatures of G-protein-coupled receptors. Nature 494, 185-194 (2013).
4. C. A. de March, J. Topin, E. Bruguera, G. Novikov, K. Ikegami, H. Matsunami, J. Golebiowski, Odorant Receptor 7D4 Activation Dynamics. Angew. Chem. Int. Ed. 57, 4554-4558 (2018).
5. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Nat. Acad. Sci. U. S. A. 112, 14966-14971 (2015).
6. C. A. de March, Y. Yu, M. J. Ni, K. A. Adipietro, H. Matsunami, M. Ma, J. Golebiowski, Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. J. Am. Chem. Soc. 137, 8611-8616 (2015).
7. C. Geithe, J. Protze, F. Kreuchwig, G. Krause, D. Krautwurst, Structural determinants of a conserved enantiomer-selective carvone binding pocket in the human odorant receptor OR1A1. Cell. Mol. Life Sci. 74, 4209-4229 (2017).
8. L. Ahmed, Y. Zhang, E. Block, M. Buehl, M. J. Corr, R. A. Cormanich, S. Gundala, H. Matsunami, D. O'Hagan, M. Ozbil, Y. Pan, S. Sekharan, N. Ten, M. Wang, M. Yang, Q. Zhang, R. Zhang, V. S. Batista, H. Zhuang, Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. Proc. Nat. Acad. Sci. U. S. A. 115, E3950-E3958 (2018).
9. L. Gelis, S. Wolf, H. Hatt, E. M. Neuhaus, K. Gerwert, Prediction of a ligand-binding niche within a human olfactory receptor by combining site-directed mutagenesis with dynamic homology modeling. Angew. Chem. Int. Ed. 51, 1274-1278 (2012).
10. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: From sequence to structure. Protein Sci 24, 1543-1548 (2015).
11. J. Freitag, J. Krieger, J. Strotmann, H. Breer, Two classes of olfactory receptors in Xenopus laevis. Neuron 15, 1383-1392 (1995).
12. G. Glusman, A. Bahar, D. Sharon, Y. Pilpel, J. White, D. Lancet, The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm Genome 11, 1016-1023 (2000).
13. X. Zhang, S. Firestein, The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. 5, 124-133 (2002).
14. Y. Niimura, M. Nei, Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc. Nat. Acad. Sci. U. S. A. 102, 6039-6044 (2005).
15. Y. Niimura, M. Nei, Evolution of olfactory receptor genes in the human genome. Proc. Nat. Acad. Sci. U. S. A. 100, 12235-12240 (2003).
16. T. Iwata, Y. Niimura, C. Kobayashi, D. Shirakawa, H. Suzuki, T. Enomoto, K. Touhara, Y. Yoshihara, J. Hirota, A long-range cis-regulatory element for class I odorant receptor genes. Nature Commun. 8, 885 (2017).
17. J. Freitag, G. Ludwig, I. Andreini, P. Rossler, H. Breer, Olfactory receptors in aquatic and terrestrial vertebrates. J. Comp. Physiol. A 183, 635-650 (1998).
18. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Science signaling 2, ra9 (2009).
19. D. M. Thal, A. Glukhova, P. M. Sexton, A. Christopoulos, Structural insights into G-protein-coupled receptor allostery. Nature 559, 45-53 (2018).
20. X. Yuan, S. Raniolo, V. Limongelli, Y. Xu, The Molecular Mechanism Underlying Ligand Binding to the Membrane-Embedded Site of a G-Protein-Coupled Receptor. J. Chem. Theory Comput. 14, 2761-2770 (2018).
21. N. Stanley, L. Pardo, G. D. Fabritiis, The pathway of ligand entry from the membrane bilayer to a lipid G protein-coupled receptor. Sci. Rep. 6, 22639 (2016).

22. R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, D. E. Shaw, Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc. Nat. Acad. Sci. U. S. A. 108, 13118-13123 (2011).
23. A. C. Kruse, J. Hu, A. C. Pan, D. H. Arlow, D. M. Rosenbaum, E. Rosemond, H. F. Green, T. Liu, P. S. Chae, R. O. Dror, D. E. Shaw, W. I. Weis, J. Wess, B. K. Kobilka, Structure and dynamics of the M3 muscarinic acetylcholine receptor. Nature 482, 552-556 (2012).
24. D. Provasi, A. Bortolato, M. Filizola, Exploring molecular mechanisms of ligand recognition by opioid receptors with metadynamics. Biochemistry 48, 10020-10029 (2009).
25. M. Sandal, M. Behrens, A. Brockhoff, F. Musiani, A. Giorgetti, P. Carloni, W. Meyerhof, Evidence for a Transient Additional Ligand Binding Site in the TAS2R46 Bitter Taste Receptor. J. Chem. Theory Comput. 11, 4439-4449 (2015).
26. K. Sharma, S. Balfanz, A. Baumann, S. Korsching, Full rescue of an inactive olfactory receptor mutant by elimination of an allosteric ligand-gating site. Sci. Rep. 8, 9631 (2018).
27. V. Cvicek, W. A. Goddard, 3rd, R. Abrol, Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. PLoS Comput Biol 12, e1004805 (2016).
28. J. A. Ballesteros, H. Weinstein, Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: C. S. Stuart, Ed., **Methods Neurosci.**, 25, 366-428 (1995).
29. C. Bushdid, C. A. de March, H. Matsunami, J. Golebiowski, Numerical models and in vitro assays to study Odorant Receptors. Methods Mol. Biol. 1820, 77-93 (2018).
30. S. Katada, T. Hirokawa, Y. Oka, M. Suwa, K. Touhara, Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the odorant-binding site. J. Neurosci. 25, 1806-1815 (2005).
31. M. Wheatley, D. Wootten, M. T. Conner, J. Simms, R. Kendrick, R. T. Logan, D. R. Poyner, J. Barwell, Lifting the lid on GPCRs: the role of extracellular loops. Br. J. Pharmacol. 165, 1688-1703 (2012).
32. K. Audouze, A. Tromelin, A. M. Le Bon, C. Belloir, R. K. Petersen, K. Kristiansen, S. Brunak, O. Taboureau, Identification of odorant-receptor interactions by global mapping of the human odorome. PloS one 9, e93037 (2014).
33. L. Charlier, J. Topin, C. Ronin, S. K. Kim, W. A. Goddard, 3rd, R. Efremov, J. Golebiowski, How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case. Cell. Mol. Life Sci. 69, 4205-4213 (2012).
34. T. Sato, T. Kawasaki, S. Mine, H. Matsumura, Functional Role of the C-Terminal Amphipathic Helix 8 of Olfactory Receptors and Other G Protein-Coupled Receptors. Int. J. Mol. Sci. 17, 1930 (2016).
35. S. Granier, B. Kobilka, A new era of GPCR structural and chemical biology. Nat. Chem. Biol. 8, 670-673 (2012).
36. J. Topin, C. A. de March, L. Charlier, C. Ronin, S. Antonczak, J. Golebiowski, Discrimination between olfactory receptor agonists and non-agonists. Chemistry 20, 10227-10230 (2014).
37. A. Dunkel, M. Steinhaus, M. Kotthoff, B. Nowak, D. Krautwurst, P. Schieberle, T. Hofmann, Nature's chemical signatures in human olfaction: a foodborne perspective for future biotechnology. Angew. Chem. Int. Ed. 53, 7124-7143 (2014).
38. D. Massberg, H. Hatt, Human Olfactory Receptors: Novel Cellular Functions Outside of the Nose. Physiol. Rev. 98, 1739-1763 (2018).
39. A. Malki, J. Fiedler, K. Fricke, I. Ballweg, M. W. Pfaffl, D. Krautwurst, Class I odorant receptors, TAS1R and TAS2R taste receptors, are markers for subpopulations of circulating leukocytes. Journal of leukocyte biology 97, 533-545 (2015).
40. TALETE srl. Dragon *Software for Molecular Descriptor Calculation* (2014).
41. A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, G. J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191 (2009).
42. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit 5 6 (2006).

43. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31, 455-461 (2010).
44. P. Schmidtke, A. Bidon-Chanal, F. J. Luque, X. Barril, MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 27, 3276-3285 (2011).
45. M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, A. L. Lomize, OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 40, D370-376 (2012).
46. S. Schrödinger Release 2018-1: Maestro, LLC, New York, NY, 2018.
47. H. Zhuang, H. Matsunami, Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nature Protoc. 3, 1402-1413 (2008).
48. H. Zhuang, H. Matsunami, Synergism of accessory factors in functional expression of mammalian odorant receptors. J. Biol. Chem. 282, 15284-15293 (2007).
49. Y. R. Li, H. Matsunami, Activation state of the M3 muscarinic acetylcholine receptor modulates mammalian odorant receptor signaling. Sci. signal. 4, ra1 (2011).

Supporting Information

Table S1. Autodock VINA docking score between OR51E1 and its agonists in the orthosteric site and in the vestibular binding site (in kcal.mol^{-1}).

	CBA	NA	BA	IVA
Orthosteric site	-4.0	-4.9	-3.7	-4.0
Vestibular site	-4.5	-5.4	-3.8	-4.5

Table S2. Log EC50 values their standard deviations values in parenthesis.

	CBA	NA	BA	IVA
OR51E1*wt*	-4.276(0.301)	-4.203(0.259)	-3.893(0.228)	-4.483(0.305)
V255G	-4.481(0.638)	-4.717(0.881)	-4.412(0.184)	-4.721(0.320)
Y201A	-	-	-	-
L203A	-2.777(0.074)	-3.607(0.259)	-2.999(0.055)	-3.31(0.145)
H264A	-3.94(0.270)	-4.316(0.567)	-3.322(0.117)	-3.075(0.123)
R265A	-	-	-	-
F266A	-3.237(0.037)	-3.391(0.190)	-2.383(0.285)	-2.488(0.077)
P256A	-3.72(0.137)	-3.323(0.348)	-3.281(0.732)	-3.104(0.116)

Table S3. EC50 95% Confidence Intervals (asymptotic, computed from Figure S3 and Figure 4a). * means that the intervals do not intersect. 'ns' for not significant.

	CBA			NA			IVA			BA		
wt	***8.9E-06***	***3,2E-04***		***1E-05***	***3E-04***		***5E-06***	***2E-04***		***3E-05***	***5E-04***	
V255G	7.6E-07	1,4E-03	ns	1E-07	4E-03	ns	3E-06	1E-04	ns	1E-05	1E-04	ns
Y201A	-			-			-			-		
L203A	1.1E-03	2.6E-03	*	5E-05	1E-03	ns	2E-04	1E-03	*	7E-04	1E-03	*
H264A	2.3E-05	5.7E-04	ns	2E-06	1E-03	ns	4E-04	2E-03	*	2E-04	1E-03	ns
R265A	-			-			-			-		
F266A	4.6E-04	7.2E-04	*	1E-04	1E-03	ns	2E-03	5E-03	*	8E-04	2E-02	*
P256A	8.5E-05	4.3E-04	ns	6E-05	4E-03	ns	4E-04	2E-03	*	3E-04	8E-04	ns

Table S4. Summary of the variance analysis: R^2 (coefficient of determination). Results of the Fisher test: F. Pr is the risk to conclude that the null hypothesis (no significant difference between the two classes) is true.

	MW	NNRS	Pol	Hy	MLOGP	SAtot	SAacc	SAdon
R^2	0.001	0.021	0.002	0.072	0.053	0.000	0.134	0.117
F	0.144	5.862	0.497	21.674	15.492	0.105	43.162	36.769
Pr > F	0.704	0.016	0.482	< 0.0001	0.000	0.746	< 0.0001	< 0.0001

Table S5. Summary of the observations for the two receptors classes (normalized values). Tukey HSD (Honestly Significantly Different) test difference analysis between the observable with a 95 % confidence.

	MW	NNRS	Pol	Hy	MLOGP	SAtot	SAacc	SAdon
ClassI	0.334 a	0.356 b	0.248 a	0.138 a	0.631 b	0.351 a	0.165 a	0.096 a
ClassII	0.326 a	0.503 a	0.232 a	0.065 b	0.700 a	0.358 a	0.088 b	0.028 b
Pr > F	0.704	0.016	0.482	0.000	0.000	0.746	0.000	0.000
Significant	no	yes	no	yes	yes	no	yes	yes

Table S6. Starting positions and MD simulations times.

ligand		Orthosteric cavity	Vestibule
CBA	Run #1, 1.7 µs	X	
	Run #2, 0.7 µs		X
	Run #3, 0.7 µs	X	
NA	0.4 µs	X	
IVA	1.8 µs		X
BA	0.3 µs		X

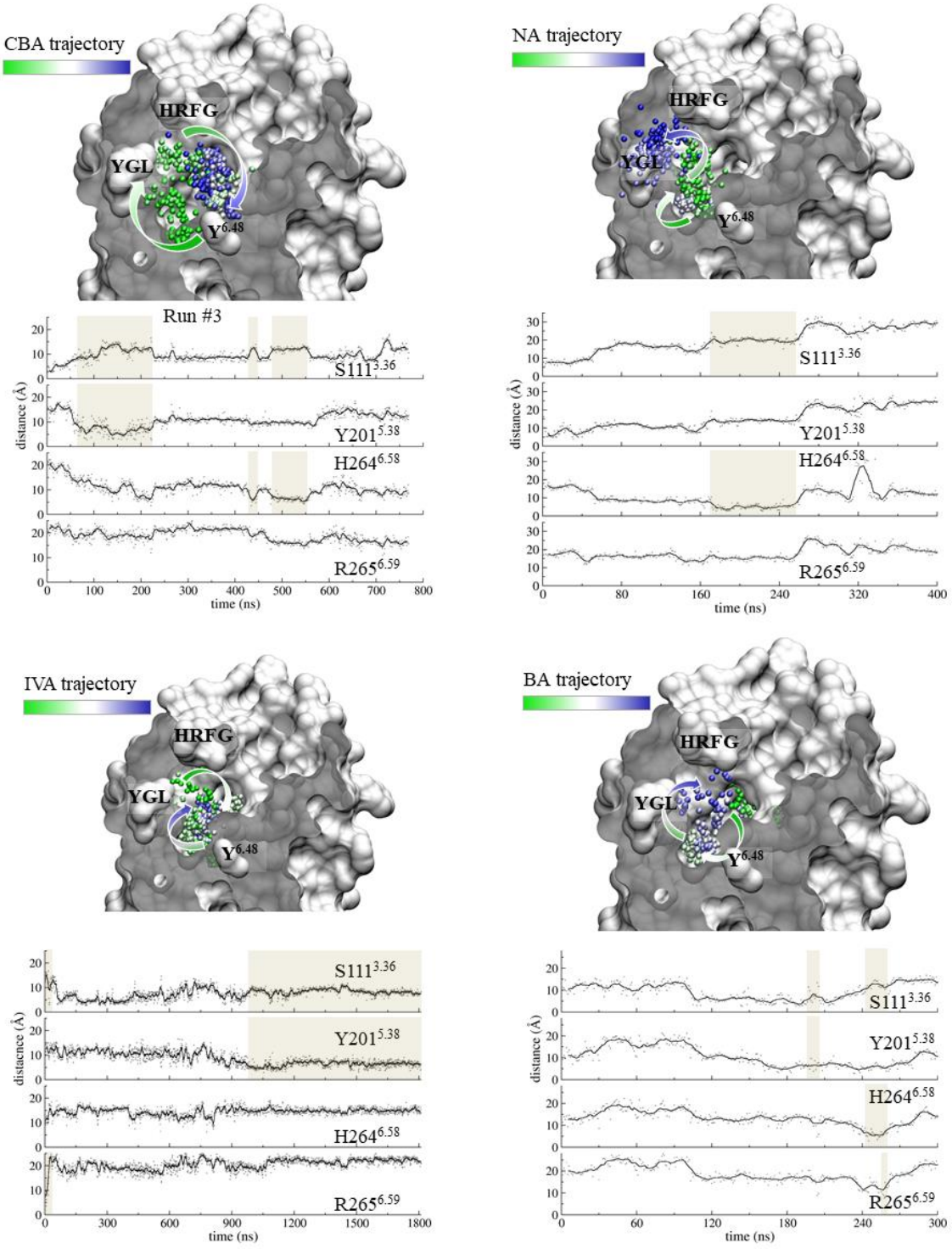


Figure S1. Molecular surface of OR51E1 and position of CBA (Run #3), NA, IVA, and BA represented by the carbon atom of their acid moiety (trajectory from green to blue). Hydrogen-bond analysis between the agonist oxygen atoms and donor/acceptor groups of S111 (bottom of the cavity), and Y201, H264, and R265 (forming the vestibule). Periods during which the agonist visits the vestibule are highlighted.

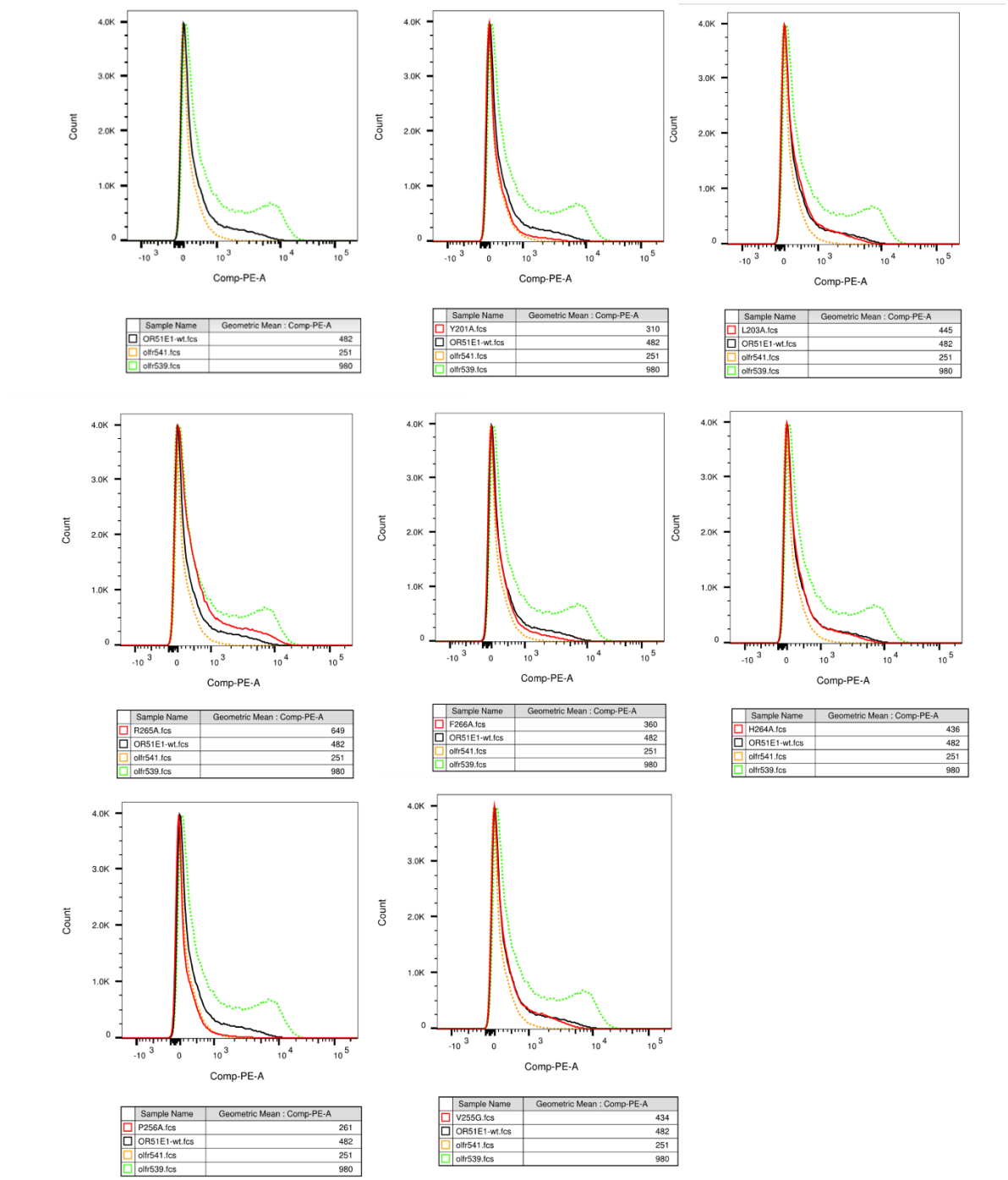


Figure S2. Cell surface expression levels of OR51E1 wt and mutants assessed by Fluorescent Activated Cell Sorting (FACS).

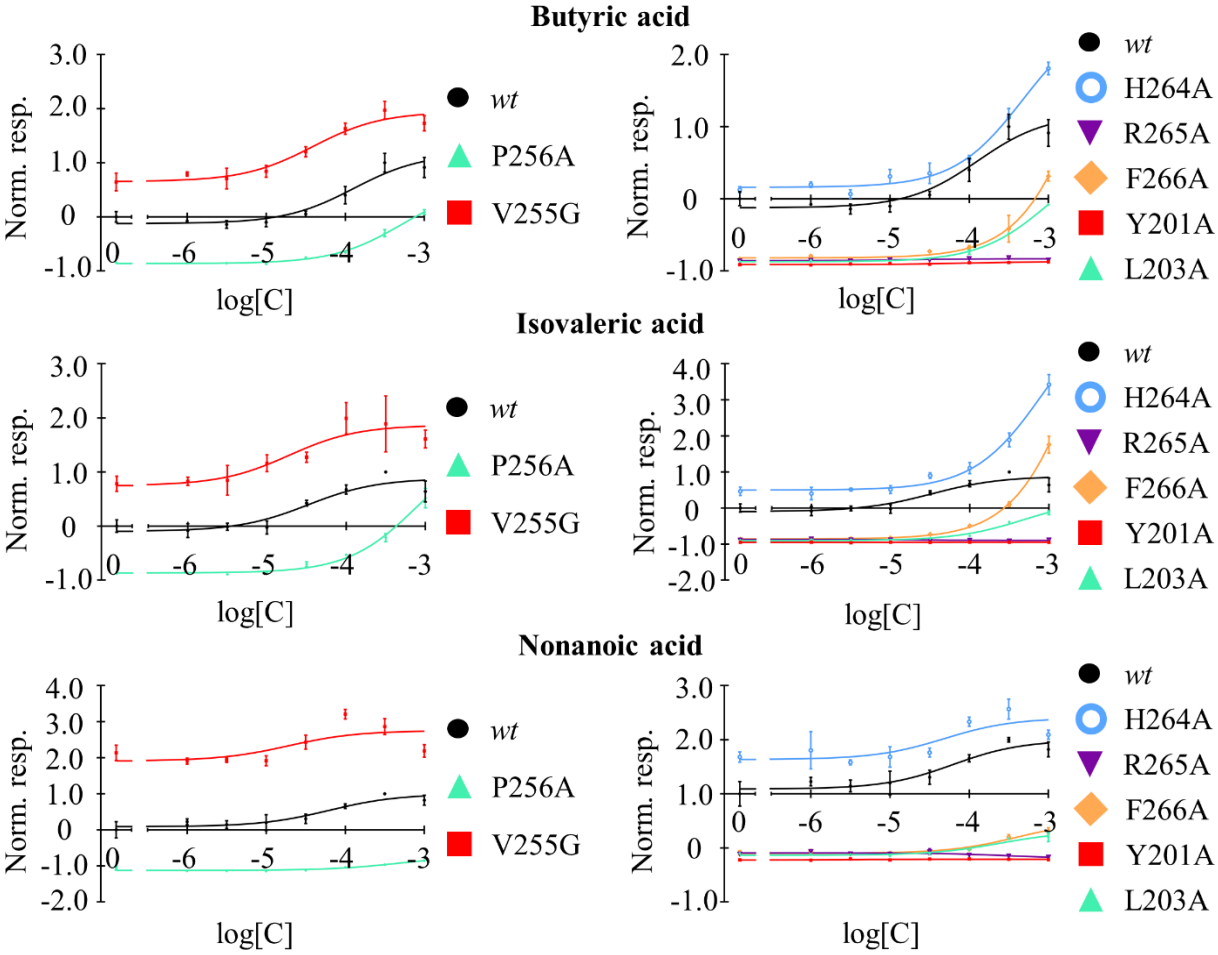


Figure S3. Normalized (to *wt*) dose-response curves of OR51E1 and its cavity and vestibule mutants to butyric acid, isovaleric acid and nonanoic acid.

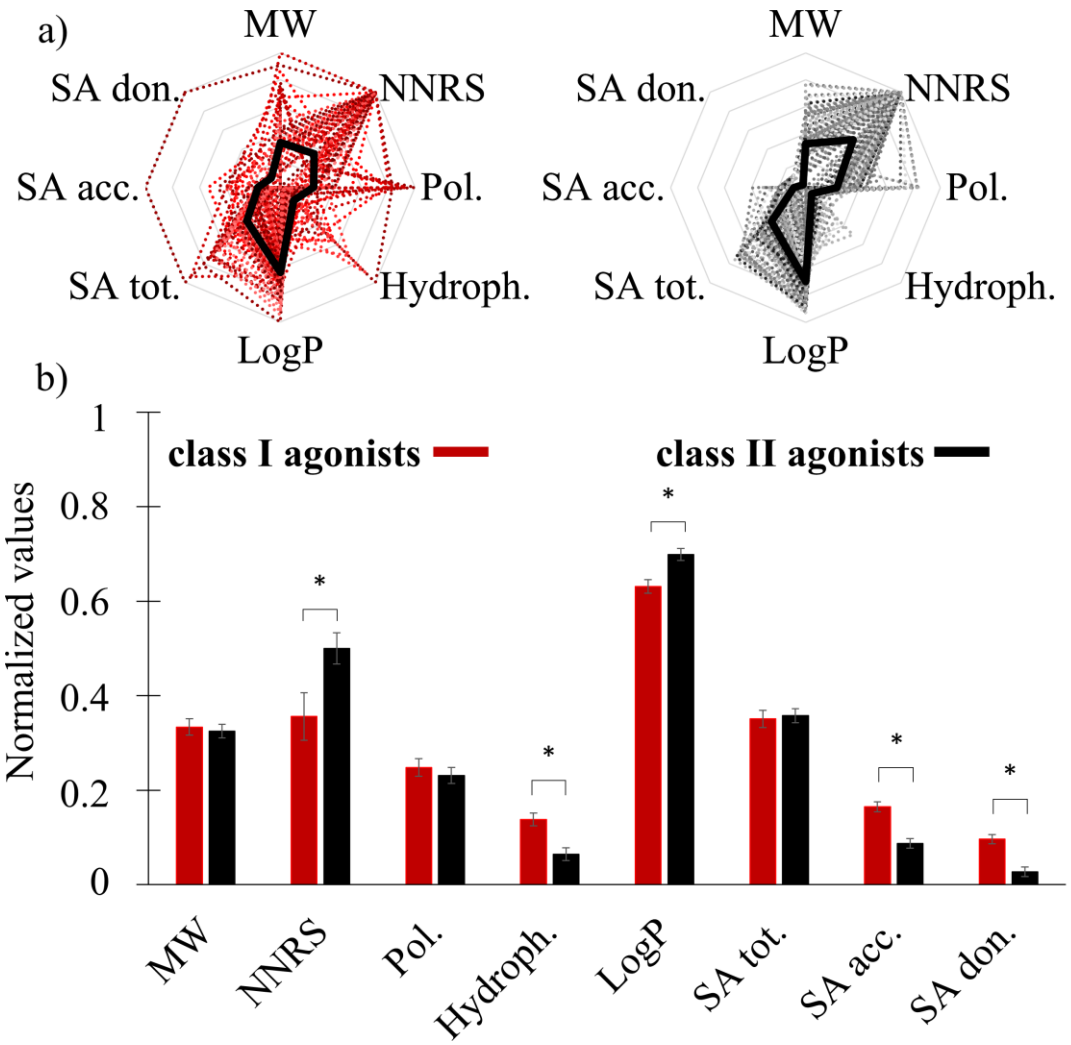


Figure S4. Chemical space properties of class I and class II human ORs. a) Spider plot representing chemical descriptor data from 92 agonists of class I ORs (left) and 189 agonists of class II (right). The values shown here have been averaged and normalized to the largest value found for all the descriptors (MW = Molecular weight, NNRS = Normalized number of ring systems, Pol. = Polarity number, Hydroph. = Hydrophylic factor, LogP = Moriguchi octanol/water partition coefficient, SA tot. = Total surface area, SA don. = Donor surface area, SA acc. = Acceptor surface area). B) Histogram representing the averages and standard error means in each class. Descriptors marked with a star are significantly different between class I and class II OR agonists (One-way ANOVA, statistical significance 0.05 Fischer test, Table S2 and S3).

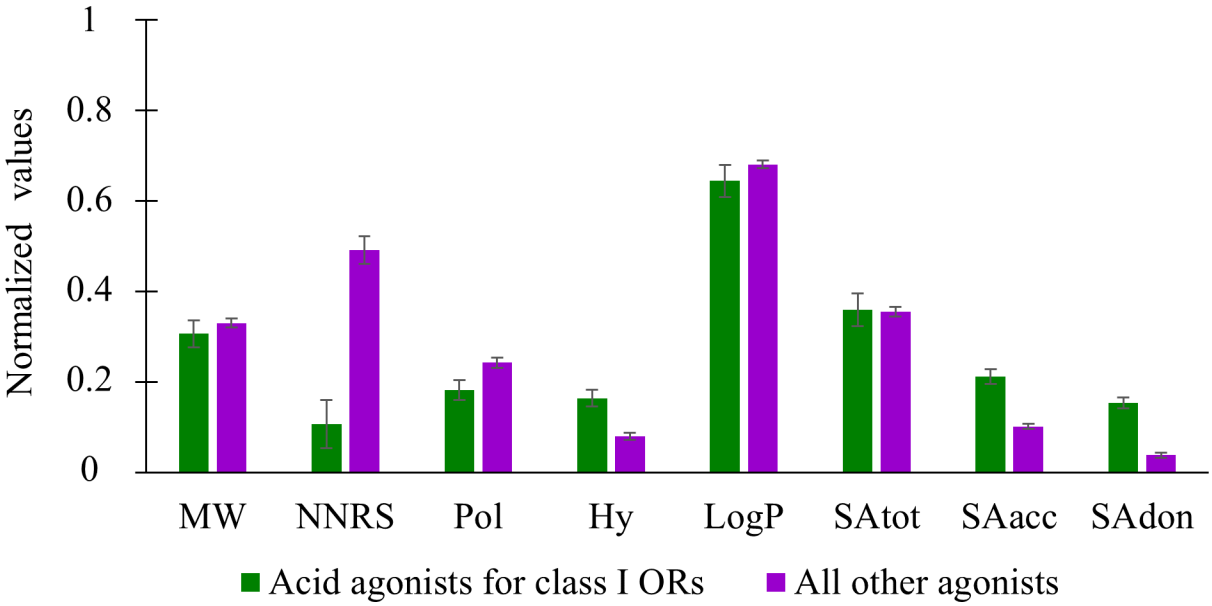


Figure S5. Chemical properties mean values as well as standard error means for the 28 acid agonists of class I ORs, and comparison with the 253 agonists reported for all human ORs (comprising two acids agonizing class II ORs). Small value of the number of cycle, and high values for SA acceptor and for SA donor are characteristics of the acid chemical family.

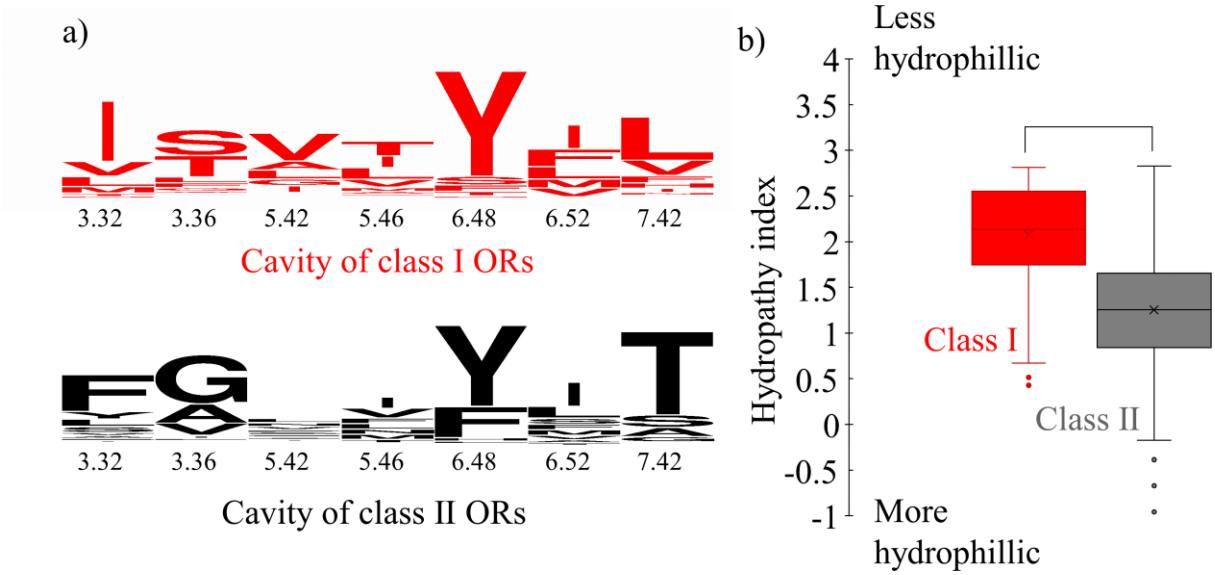


Figure S6. Hydrophobicity calculations of the orthosteric binding pocket. Residues 3.32, 3.36, 5.42, 5.46, 6.52 and 7.42 were extracted to obtain a seven amino acid peptide sequence mimicking all human OR cavities. The average hydrophopathy of each peptide was calculated using the GRAVY Calculator online server (www.gravy-calculator.de). The hydropathy index of an amino acid is a number representing the hydrophilic properties of the side-chain.[1] The smaller the number, the more hydrophilic the amino acid. a) Sequence alignment of the seven residues pointing towards the orthosteric binding cavity. b) Box-plot showing the average hydropathy score distribution in class I (red) and class II (black) ORs.

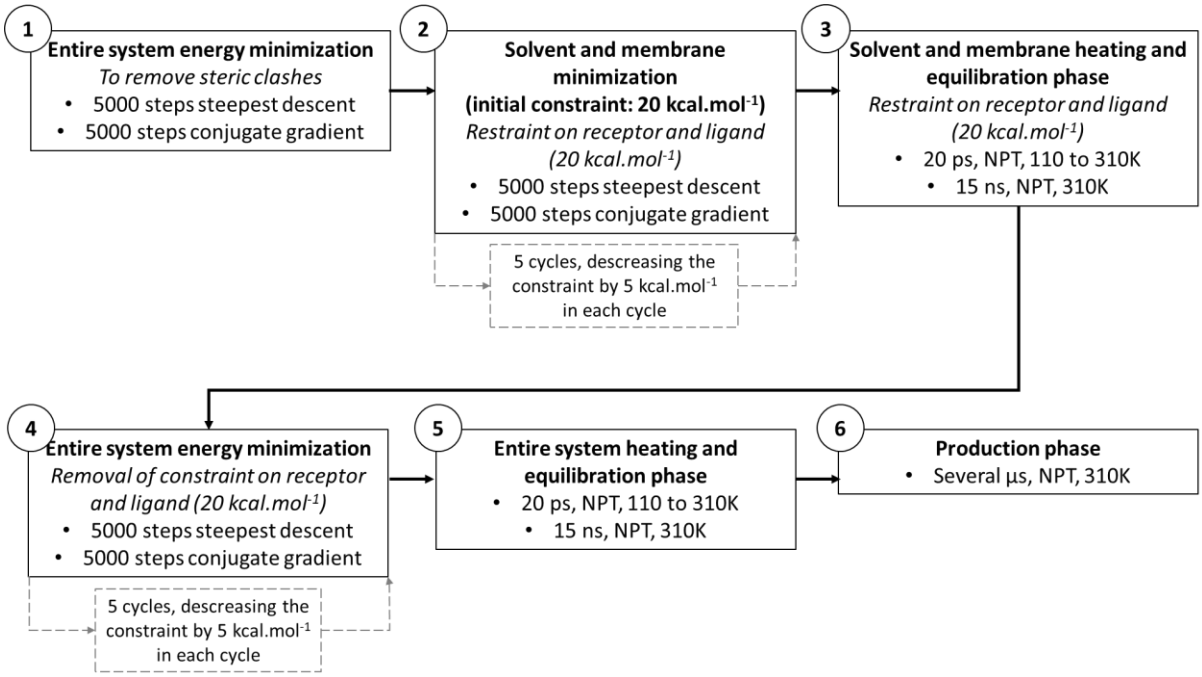


Figure S7. Workflow of molecular dynamics simulations.

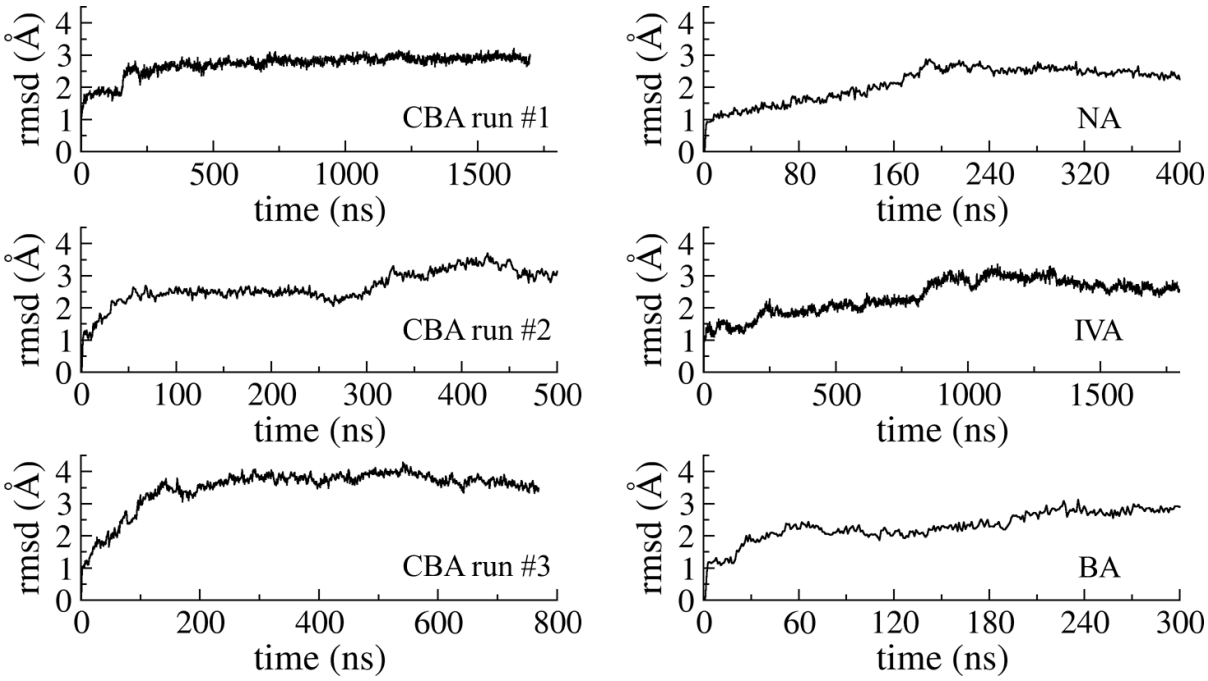


Figure S8. OR51E1 Transmembrane domain root mean square deviation (RMSD) during the various molecular dynamics simulation. The RMSD was computed on the Cα atoms with respect to the first frame.

The supplementary Files S1 to S4 can be found at:

http://chemosim.unice.fr/index.php/downloads/

Publication 5

Molecular recognition profile of odorant receptors is governed by their binding pocket

CAROLINE BUSHDID, CLAIRE A. DE MARCH, YIQUN YU, XIAOJING CONG, HIROAKI MATSUNAMI, JÉRÔME GOLEBIOWSKI

IN PREPARATION

Introduction

Mammalian olfaction begins when an odorant molecule interacts with an odorant receptor (OR) (1). ORs belong to the superfamily of G protein coupled receptors which are membrane proteins. The function of these receptors is to "recognize" odorant compounds and to transduce the signal into a cellular response. To detect the vast amount of odorant compounds potentially present in the environment, the olfactory system is thought to rely on a combinatorial code, where ORs play a central role. In this framework, a given odorant receptor (OR) detects multiple odorants, and a given odorant is detected by multiple ORs (2). Most ORs respond to a large chemical space (3) but some ORs only respond to a very specific set of odorants. Two terms are defined to study molecular recognition in ORs: i/molecular receptive range and ii/tuning breadth (4). The molecular receptive range (MRR) of an OR corresponds to the set of compounds that activate the OR. The tuning breadth, on the other hand, is the distribution of these compounds across chemical space. In this respect, a receptor can be defined as being broadly tuned (responding to a wide array of chemical structures) or on the contrary narrowly tuned (responding only to a certain class of chemical structures).

G protein-coupled odorant receptors, as all GPCRs, are allosteric machines. Ligand binding and G protein coupling sites are removed by more than 20 Å. The signaling process triggered by ligand recognition can be modeled as a two-step process, associated with specific equilibriums. They can be attributed to i/ the interaction at the receptor binding cavity, and ii/ the activation mechanism through an allosteric process from the cavity to the G protein coupling site. In the first step the ligand must reach a certain affinity to bind the receptor (eq. 1). In the second, the bound complex in its inactive form is in an equilibrium with a so-called active form, which controls the signaling process (eq. 2).

$$\mathrm{L} + \mathrm{OR} = (\mathrm{L/OR})_{bound} \qquad \mathrm{K}_1;\ \Delta\mathrm{G}_1 \qquad (1)$$

$$(\mathrm{L/OR})_{bound} = (\mathrm{L/OR})_{activated} \qquad \mathrm{K}_2;\ \Delta\mathrm{G}_2 \qquad (2)$$

Intuitively, the first step seems crucial in the ligand recognition process of the receptor, but we have shown that the activation mechanism *per se* also modulates the tuning of ORs (5).
In a well-studied sub-family of mouse ORs (MORs), (4-9) receptors sharing more than 50% amino-acid identity show very different tuning breadths. In particular MOR256-22, is a narrowly tuned receptor (4), while MOR256-31 is a broadly tuned one. Identifying and

quantifying what makes these receptors broadly and narrowly tuned is important for a better understanding how ORs "read" the chemical space.

In a joint approach using molecular modeling, site-directed mutagenesis and functional assays, we decomposed the role of the orthosteric binding cavity in the tuning breadth of an OR. A narrowly tuned OR binding cavity was totally mutated into that of a broadly tuned one. We first identified the nature of the residues lining the orthosteric binding cavity in the narrowly tuned mouse OR256-22. Then three-dimensional atomic models of these ORs confirmed the location of the residues. In silico and in vitro mutations were performed to progressively transform the cavity of MOR256-22 into that of the broadly tuned MOR256-31. Eight mutations were considered necessary to transform the binding pocket of MOR256-22 into that of MOR256-31 (the octa-mutant will be referred to by "22mut31"). Upon mutating the residues lining the orthosteric binding pocket, the tuning breadth, the molecular receptive range and the strength of the response are impacted. We show that the orthosteric binding pocket is mainly responsible for the recognition spectrum while the potency of the ligand for a given receptor is mostly encoded into the allosteric activation mechanism of the GPCR.

Results

The molecular receptive range of odorant receptors is controlled by the binding cavity. To identify key residues that underlie differential (i.e. broad or narrow) responsiveness between MOR256-22 and MOR256-31 we aligned their protein sequences. These ORs belong to the same subfamily, and share > 50% amino acid identity on their overall sequence. However, in the transmembrane (TM) region, the sequence identity of MOR256-22 with MOR256-31 was of 54.9% (Table 1). The conservation of each TM varied, with TM2 sharing the highest percentage identity (82%) followed by TMs 7 and 6 (76% and 73%, respectively). The least conserved TMs were TM4 and TM5 with ~35% sequence identities.

Table 1: Percentage of identity in the transmembrane region of MOR256-31 as compared to MOR256-22.

Domain	Percentage of Identity	Identical residues/Total residues
TM 1	37.0	10/27
TM 2	82.1	23/28
TM 3	44.1	15/34
TM 4	34.8	8/23
TM 5	35.7	10/28
TM 6	73.3	22/30
TM 7	76.0	19/25
All TMs	54.9	107/195

From crystallographic data for class A GPCRs it is established that the binding cavity is made up of residues belonging to TM3, TM6 and TM7. Notably, residues at positions corresponding to the Ballesteros-Weinstein notation 3.32, 3.33, 3.36, 6.48, 6.51, and 7.39 make consensus contacts with diverse ligands across these protein (10). In the case of ORs and in other class A GPCR, structures as well as mutants in diverse studies show that other residues belonging to TM5 also play an important role in ligand selectivity (11). These residues served as a starting point to identify those responsible for the diverging responsiveness profiles and to understand the molecular basis for the recognition spectrum of this receptor (Fig. 1A). 10 residues were identified as making up the binding cavity, of which 8 were distinct amongst both receptors. The cavities were hence considered to be very different.

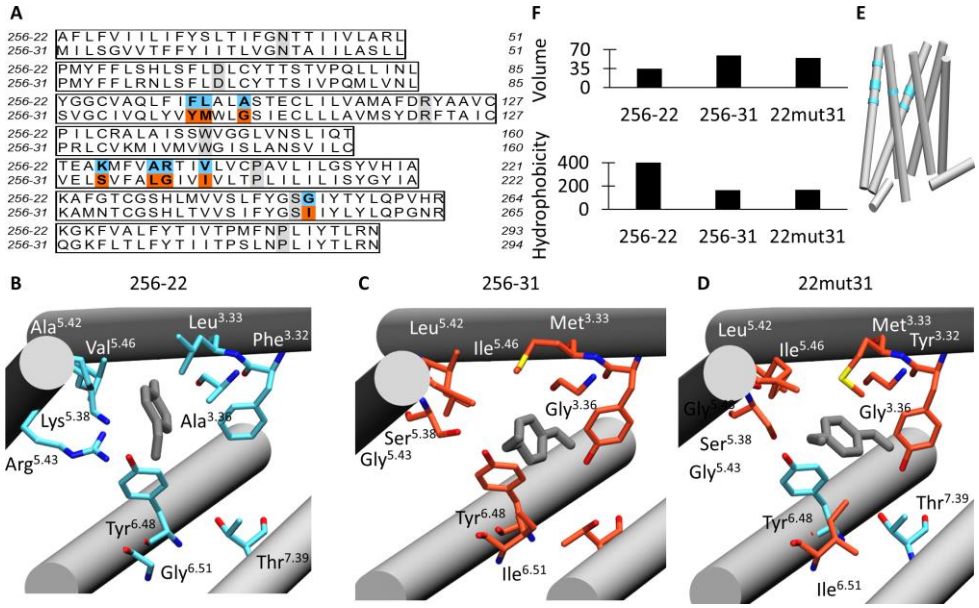


Figure 1. Binding cavity transformation of MOR256-22 A. Sequence alignment of MOR256-22 with MOR256-31. Only transmembrane domains are shown, highlighted residues concern those mutated during this study. **B.** Three-dimensional model of MOR256-22. Helices are shown in white, residues in contact with the ligand (anisaldehyde) are shown in cyan. Anisaldehyde is shown in grey. **C.** Three-dimensional model of MOR256-31. Helices are shown in white, residues in contact with the ligand (anisaldehyde) are shown in orange. **D.** Transformation of the cavity of MOR256-22 into MOR256-31 (22mut31). Mutated residues are shown in orange, other residues in contact with the ligand are shown in cyan. **E.** 3D model of MOR256-22, highlighted in cyan are the position of the residues in contact with the ligand. Loops are not shown for image clarity. **F.** Evolution of the binding pocket volume (in $Å^3$) of MOR256-22, MOR256-31 and 22mut31, and of their hydrophobicity score of MOR256-22, MOR256-31 and 22mut31.

Each residue at the positions identified in MOR256-22 as being in contact with the ligands were mutated *in silico* to corresponding residues in the target MOR256-31 (Fig. 1 B to E). To reprogram MOR256-22 into MOR256-31, eight positions were considered (F104Y, L205M,

A108G, K197S, A201L, R202G, V205I and G254I, see Fig. 1 D). More importantly, an octa-mutant of the binding cavity of MOR256-22, including all these positions was predicted to mimic the cavity of MOR256-31 (this octa-mutant will be referred to as 22mut31). Computational assessment of the characteristics of the binding cavity showed that wild-type MOR256-22 and MOR256-31 had very different pocket volume and hydrophobicity (Fig. 1E). After the reprogramming, 22mut31 nicely recovered the features of MOR256-31 cavity. Another control was also performed through the analysis of the position and orientation of anisaldehyde (a known agonist of MOR256-22) within the cavities. Again, the ligand was predicted to have a similar position in both 22mut31 and MOR256-31 (Fig. 1 B to D).

In vitro assays quantified the molecular receptive range of the receptors. Most of the single mutants corresponding to the mutations that reprogram MOR256-22 into MOR256-31 (i.e. F104Y, L105M, K197S, A201L, R202G and V205I) were unsuccessful at changing the MRR of MOR256-22 (Fig. 2A). For all these mutants, strong responses to trans-cinnamaldehyde and anisaldehyde were recorded as is the case for MOR256-22. On the other hand, the quadruple mutant, where only four mutations were performed out of the eight positions (MOR256-22 L105M/K197S/R202G/G254I – hereafter referred to as "22half31") exhibited a shift in its response profile when compared to that of MOR256-22: it responded to 2-heptanone, D-limonene, benzyl alcohol, acetophenone, coumarin and (-)-carvone which were not canonical ligands for this receptor. Finally, 22mut31 exhibited a much larger response profile, similar to that of MOR256-31, where sensitivity to trans-cinnamaldehyde was lost, but where recognition of more than 10 new ligands was gained. A general analysis of the MRR (Fig. 2B) reported that MOR256-22 (which only recognized anisaldehyde and trans-cinnamaldehyde) became a broadly tuned OR recognizing 16 ligands when its cavity was fully reprogrammed into that of MOR256-31.

A constituively active mutant of MOR256-22 is a broadly tuned receptor. In addition to the mutants conceived to redesign the cavity of MOR256-22 a mutation was performed (A108L) which rendered the receptor constitutively active, as already shown on MOR256-3 (7). The receptor exhibited a basal activity 4.2 times higher than that of MOR256-22 (Fig. 2C). As expected based on a previous study (5), the receptor responded to a wider MRR than MOR256-22 (Fig. 2 A and B).

The binding cavity affects the basal activity of the receptor. MOR256-22 showed a low basal activity (Fig.2C). On the other hand, all receptors which exhibited responses to a wide array of chemical structures, had extremely high basal activities. MOR256-31, 22mut31 and 22half 31 had respectively basal activities 14.9, 4.8 and 9.8 times higher than MOR256-22. A correlation was observed between the basal activity and the MRR of a given GPCOR (Fig. 2D).

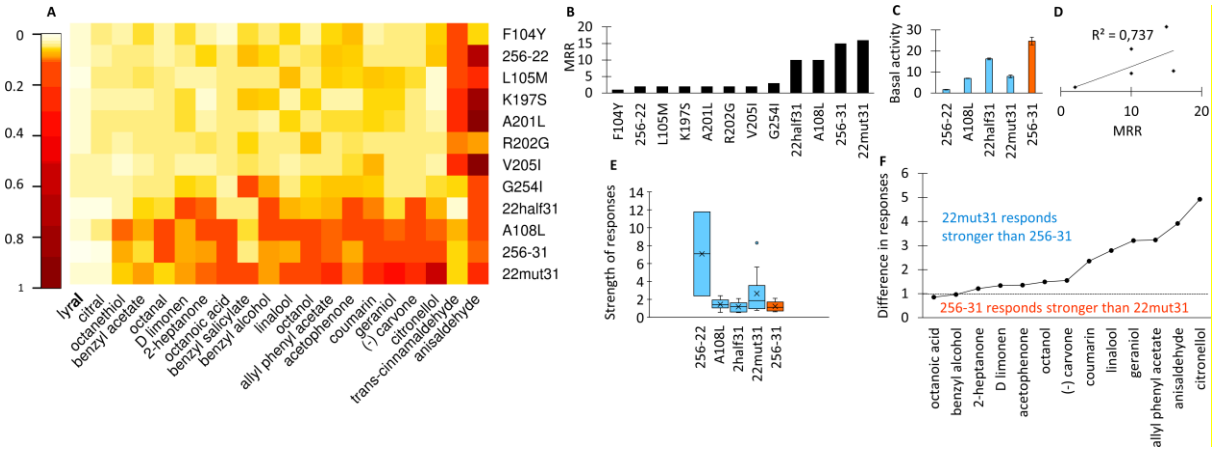


Figure 2. The response profile of MOR256-22 were shifted as a function of the residues lining its orthosteric binding pocket A. Heatmap representing the response profile for all the mutants allowing the transformation of MOR256-22 into MOR256-31 as well as the wild types, to a set of 21 previously identified odorants (*5*). All responses were normalized to the responses of all mutants and all odorants. They were ordered according to the molecular receptive range (MRR) of each receptor. The darker the shade of red, the stronger the receptor responded to a given ligand. **B.** Evolution of the molecular receptive range of MOR256-22 after each mutation. **C**. Basal activity of MOR256-22, MOR256-22 A108L, 22half31, 22mut31 and MOR256-31. **D.** Correlation between the basal activity and the MRR of wild-type and mutant ORs. **E.** Box plot showing the distribution of the response strengths to agonists of MOR256-22, 22half31, 22mut31 and MOR256-31 to the molecules in their molecular receptive range. Only agonists of the receptors were considered to construct this box plot. **F.** Ligand by ligand analysis on the difference in the strength of the response between MOR256-31 and 22mut31. Dotted line represents an equivalent response of MOR256-31 and of 22mut31 for a same ligand.

The binding cavity affects the strength of the response. MOR256-22 responded to only two agonists, but the recorded responses were very strong (Fig. 2E). Docking scores of anisaldehyde into the ligand binding cavities showed a greater affinity for MOR256-22 (-5.1 kcal.mol$^{-1)}$ than for MOR256-31 (-3.2 kcal.mol^{-1}) or 22mut31 (-3.7 kcal.mol^{-1}). This confirm that the cavity of MOR256-31 has a worse affinity for anisaldehyde than MOR256-22. Consistently, MOR256-31 and 22mut31 did not respond to anisaldehyde while MOR256-22 responded strongly to this ligand (Fig. 2A). Generally speaking, MOR256-31 responded

weakly to a wide array of chemical structures (Fig. B and E) while 22mut31 responded more strongly than MOR256-31 (Fig. 2 E and F) and to a wider array of structure than MOR256-22 (Fig. 2 A and B).

MOR256-31 and 22mut31 both responded to 13 ligands, Figure 2F shows a ligand by ligand comparison of the modification in the strength of the response between both receptors. 22mut31 responded 0.8 times weaker to octanoic acid than MOR256-31, but responded as strong as or stronger than MOR256-31 to all other compounds. On average, 22mut31 responds 2.3 times stronger to odorants than MOR256-31.

Discussion

In the current study, we investigated the molecular and structural features underlying the role of the orthosteric binding cavity in G protein-coupled odorant receptor ligand recognition. We identified eight residues which all point into the orthosteric binding cavity of two members of the MOR256 sub-family. These residues differ between so-called broadly and a narrowly tuned ORs. Upon reprogramming the cavity of a narrowly-tuned into a broadly-tuned OR, we measured how the molecular receptive range but also basal activity, were impacted. The strength of the response seemed to be dependent on the activation mechanism of the receptor.

Herein, a straightforward case was identified where the ligand binding cavity was responsible for the molecular receptive range of an OR (Fig. 2A). MOR256-22 and MOR256-31 showed different response profiles in addition to different tuning breadths. MOR256-22 had a low basal activity and responded strongly to 2 agonists. In contrast, MOR256-31 had a high basal activity, and responded weakly to 15 compounds. Interestingly, the cavity-chimeric mutant 22mut31 exhibited a basal activity 4.6 times higher than MOR256-22, and a MRR which was 8 times larger (Fig. 2 B and C). In addition, it responded on average 2.3 times stronger to its ligands than MOR256-31 (Fig 2E and F).

Our results suggested that the role of the binding cavity is two-fold: it affects the MRR (and tuning breadth) of the receptor, but it also enhances its basal activity. MRR and basal activity are thus intertwined which is in line with precedent observations by our group (5,7).

On the other hand, the strength of the response seems to be an intrinsic property of the receptor. Indeed, the mutant receptor 22mut31, even though exhibiting a shifted MRR and basal activity, retained its capacity to respond more strongly to odorants than MOR256-31 (Fig. 2F). As the cavity does not control the strength of response of the OR, we believe that the affinity of a ligand is not exclusively dependent on the binding pocket composition.

This comes as no surprise since GPCRs are intrinsically allosteric machines: when a ligand binds to the orthosteric binding pocket, the signal is transduced to the G-protein binding site through reversible conformational changes in the protein (12). We speculate that the weak response of a broadly tuned OR counter to a strong response of a narrowly tuned OR avoids a broadly tuned receptor from "dominating" the combinatorial code.

In conclusion, we studied how the binding cavity of a receptor was responsible for its MRR, tuning breadth and basal activity. The activation mechanism seems to control the strength of the response. Indeed, the mutant receptor, 22mut31, responded on average 2.3 times stronger to odorants than MOR256-31. This suggests that regions besides the orthosteric binding pocket are also involved in molecular recognition since they dictate the strength at which the receptor will respond. These could be molecular locks or other allosteric binding sites located within the protein itself.

Methods

3D model building

The protocol follows a previously published method (11). Briefly, all 397 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 data from the 153 mutants previously described in the literature. *For further references regarding sequence alignment please refer to Part IV of this manuscript, in the Methods section of "Genome-wide analysis of olfaction"*. 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 (13). The N-terminal structure was omitted to avoid perturbing the modeling protocol. Five models were obtained and the one that was consistent with the *in vitro* data and several structural constraints (no large folded structure in extra-cellular loops, all trans-membranes helices (TMs) folded as α-helices, and a tiny α-helix 8 at the C-terminal extremity) was kept. Visual inspection of all the residues pointing into the binding cavity were additionally performed to ensure the cavities were comparable (i.e. same orientation of the side-chains).

Anisaldehyde structure and parameters were prepared with the antechamber module of AMBER with AM1-BCC charges. It was docked into the receptor cavity, using rigid docking

parameters to ensure that the binding cavities remained comparable, with Autodock Vina (14).

Site directed mutagenesis and in vitro testing

The protocol follows the same steps as those presented in Part II, Publication 2. Here, the response of a receptor was tested in a single dose screening assays, and all experiments were performed in triplicate. All ORs were tested at the same concentrations and a positive response was determined at 300 μM, a near-saturating concentration. The results showing the strength of the response are normalized to that of the basal activity of the receptor.

References

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

2. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. Cell 96, 713-723 (1999).

3. K. Touhara, L. B. Vosshall, Sensing odorants and pheromones with chemosensory receptors. Annu. Rev. Physiol. 71, 307-332 (2009).

4. D. Kepchia, B. Sherman, R. Haddad, C. W. Luetje, Mammalian odorant receptor tuning breadth persists across distinct odorant panels. PloS One 12, e0185329 (2017).

5. Y. Yu, C. A. de March, M. J. Ni, K. A. Adipietro, J. Golebiowski, H. Matsunami, M. Ma, Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U. S. A. 112, 14966-14971 (2015).

6. X. Grosmaitre, S. H. Fuss, A. C. Lee, K. A. Adipietro, H. Matsunami, P. Mombaerts, M. Ma, SR1, a mouse odorant receptor with an unusually broad response profile. J. Neurosci. 29, 14545-14552 (2009).

7. C. A. de March, Y. Yu, M. J. Ni, K. A. Adipietro, H. Matsunami, M. Ma, J. Golebiowski, Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. J. Am. Chem. Soc. 137, 8611-8616 (2015).

8. J. Li, R. Haddad, S. Chen, V. Santos, C. W. Luetje, A broadly tuned mouse odorant receptor that detects nitrotoluenes. J. Neurochem. 121, 881-890 (2012).

9. K. Nara, L. R. Saraiva, X. Ye, L. B. Buck, A large-scale analysis of odor coding in the olfactory epithelium. J. Neurosci. 31, 9179-9191 (2011).

10. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler, M. M. Babu, Molecular signatures of G-protein-coupled receptors. Nature 494, 185-194 (2013).

11. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: From sequence to structure. Protein Sci 24, 1543-1548 (2015).

12. J. P. Changeux, A. Christopoulos, Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation. Cell 166, 1084-1102 (2016).

13. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Current protocols in bioinformatics Chapter 5, Unit 5-6 (2006).

14. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comp. Chem. 31, 455-461 (2010).

Part IV
Perspectives

"From somewhere at the bottom of the passage the smell of roasting coffee came floating into the street […] Then a door banged, seeming to cut off the smell as abruptly as though it had been a sound."

George Orwell (1984, 1949)

Genome-wide analysis of odorant receptors

Introduction

Cataloging the odorant/OR interaction is a challenging problem. Despite numerous and intensive efforts most ORs remain orphan. Unsurprisingly, deorphanization has happened at a low rate since in a classical molecular biology framework, it would be required to test all possible odorant/OR pairs, which of course is not feasible given the dimensionality of odorant space. Computational approaches such as bio-informatic analysis of OR sequences or generation of 3D models will allow a genome-wide analysis of the potential odorant/OR interactions.

The appeal of tagging a genetic sequence to a function is understandable. It was first expected that genetically identified families would be related to functional families. If that was the case, by looking at a gene sequence it would be possible to predict the chemical space an OR responds to. However, **phylogenetic organization based on full-length sequences does not reflect ligand recognition profile** (Fig.1)**.**

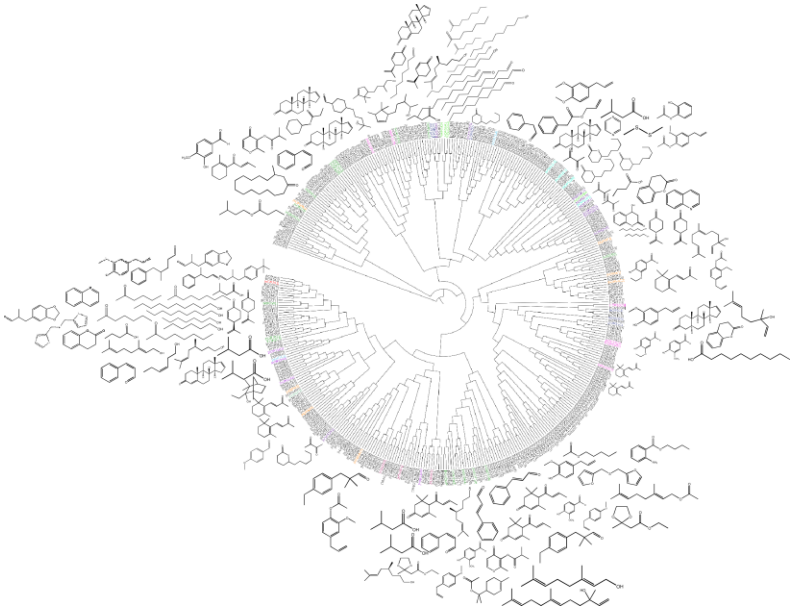


Figure 1: Phylogenetic tree of complete human OR sequences and their agonists. Deorphanized receptors are colored and their agonists are shown. No obvious relationship can be found in terms of "recognition spectrum" for a certain family of ORs.

OR genes are classified using the nomenclature 'OR*n*X*m*'. Here, *n* is a number that represents the family the OR belongs to. OR sequences sharing at least 40% sequence identity fall into the same family (i.e. OR1A1 and OR1B2 have at least 40% amino acids in common). X is a letter that represents the sub-family they belong to, in this context, two OR sequences belong

to the same sub-family if they have at least 60% sequence identity (i.e. OR1A1 and OR1A2 share at least 60% amino acids in common). Finally, *m* represents the rank of the OR inside a given family (1). This classification reflects whole sequence identity, but as many regions of the sequences are highly conserved – they have evolutionarily-conserved functional significance – the differences are expected to lie at their binding cavity. The cavity in turn is made up by 7 residues, which represent ~2% of the full sequence. It is thus likely that some receptors which do not belong to the same family based on their whole sequence, might nonetheless have very similar binding cavities and could recognize related chemicals. I defined such couples "neo-orthologs" and "neo-homologs".

A homolog is a gene related to a second gene by descent from a *common ancestor*. Orthologs are a sub-category of homologs. Here two genes in *different species* evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. We here asked whether it is possible to identify "neo-homologs" (our newly coined term) based only similar cavity composition and properties. "Neo-homologs" and "neo-orthologs" are genes that produce ORs with the same binding cavities but which do not share an overall sequence similarity. "Neo-orthologs" are a sub-category within "neo-homologs" where the same cavity is found between receptors of different species. They are expected to have similar tuning breadths and chemical spaces.

Two hypotheses can be established supposing that **similar binding sites will have similar recognition profiles**. First, we hypothesize that if two receptors have similar amino acids in their binding cavity (thus a similar "binding cavity sequence") then they will recognize the same types of ligands. Second, and in a subtler manner, we hypothesize that if two 3D models of receptors have similar 3D orthosteric binding pocket characteristics (hydrophobicity, size, etc.) then they will recognize similar ligands.

To identify potential similarities in the binding cavity from a sequence point of view, I tentatively built pseudo-peptides made-up with the residues forming the binding cavities of the olfactory sub-genome. Colleagues at the University College Dublin in Ireland helped building phylogenetic trees of these pseudo-peptides to identify neo-orthologs.

Additionally, I led the basis for building of a database containing all the 3D structures of human and mouse ORs with the aim of analyzing their binding cavities. The optimization process of the protocol will be detailed here. The idea is to identify "neo-orthologs" and "neo-homologs" through three-dimensional structures of ORs on the basis of the physical and chemical properties of their cavities.

Methods

Sequence alignment and dimensionality reduction

A typical odorant receptor sequence has 310 residues. Of these 310 residues, 7 can be considered as being in contact with odorants (2-5). These are residues $104^{3.32}$, $108^{3.36}$, $202^{5.42}$, $206^{5.46}$, $252^{6.48}$, $256^{6.52}$ and $278^{7.42}$ (2-6). Figure 2 shows an alignment of 15 ORs with the four class A G Protein-Coupled Receptors. A total of 153 mutations performed on 14 different ORs were gathered from the literature. These mutations cover ~32% of the entire sequence of a typical 310 amino acid sequence of an OR.

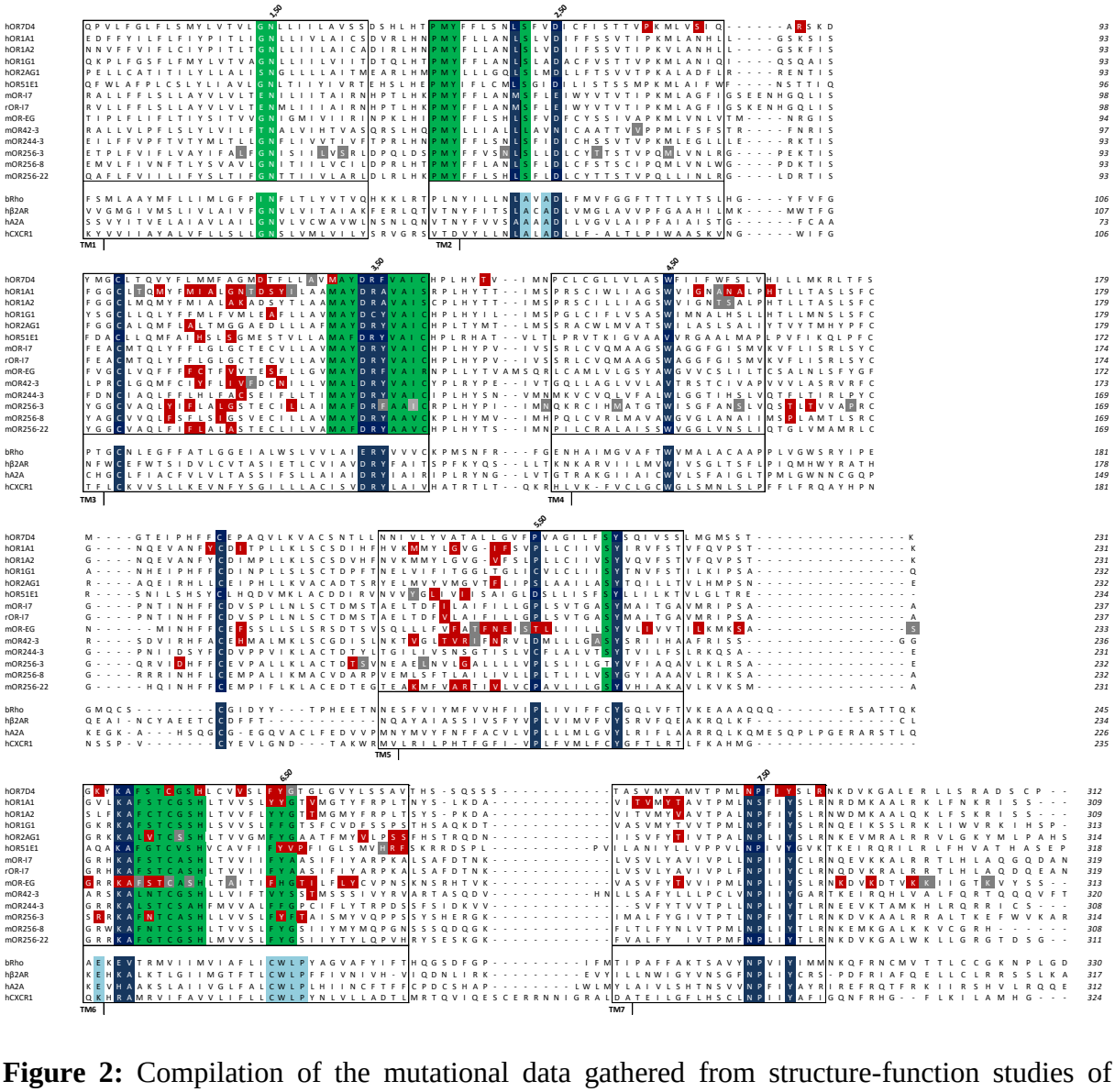


Figure 2: Compilation of the mutational data gathered from structure-function studies of ORs. Conserved motifs amongst class A GPRCs are highlighted in blue, those specific to ORs are highlighted in green, and those specific to non-olfactory class A GPCRs are marked in light blue. Mutations affecting ligand binding are marked in red, while mutations not affecting the receptors response are shown in grey.

The seven residues involved in ligand binding are largely variable. It could be possible that two receptors with diverging complete sequences might share similar binding cavity residues. To perform dimensionality reduction, I first aligned 397 human ORs with 1100 mouse ORs. Then, the ~310 amino acid sequences of ORs were reduced to the 7 residues lining the binding cavity, i.e. (Fig. 3).

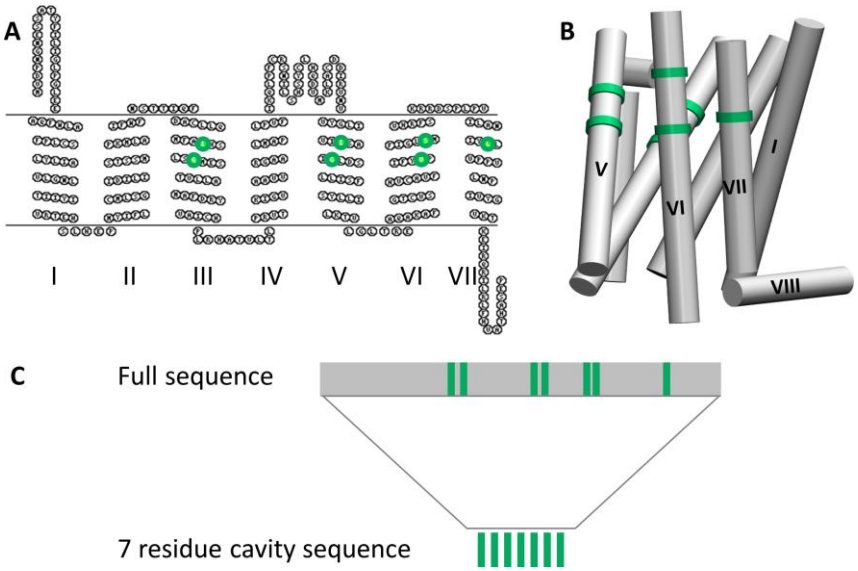


Figure 3: A. Snakeplot of the OR sequence with the seven consensus residues involved in odorant contact highlighted in green. **B.** Position of the residues involved in ligand contact on the three-dimensional structure of the receptor **C.** Scheme representing the position on the sequence of the residues involved in ligand contact, which make up the peptide reduction.

3D model building

A previously published protocol was used to build human OR models (2, 7). Briefly, all OR sequences were aligned to the sequence of class A GPCRs for which the experimental structure is known. Manual adjustments were performed to be consistent with more than 150 mutational data points available from the literature. Homology models were obtained for each receptor 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 (8). The N-terminal and C-terminal structures were omitted to avoid perturbing the modeling protocol.

Replica exchange molecular dynamics simulations for geometry optimization

Replica exchange molecular dynamics (REMD) is a simulation method which aims to improve the sampling of a system using different state coordinates, these coordinates were dictated by temperature in our case. The idea here is to run several copies of one system

randomly initialized at different temperatures and then exchange them to make the configuration at high temperatures available at low temperatures and vice versa. This allows conformational sampling of high and low energy configurations. In this way, energy barrier on the potential energy surface can be overcome, allowing the exploration of distinct conformational spaces than the ones accessible from the initial structure.

The REMD simulations was set up with the help from Dr. Xiaojing Cong. They were realized with 16 replicas, starting from one initial structure obtained using Modeller with an implicit membrane. To obtain the best structure, 10 models were produced using Modeller and the best one was selected using the provided DOPE score. To sample the conformational space of ORs, REMD simulations were performed on the systems by restricting the backbone of the initial model but allowing the side chains to be flexible. The replica temperatures ranged from 310 K to 500 K, following a distribution calculated with the Patriksson-van der Spoel approach (9). Exchange between replicas was attempted every 1000 simulation steps. This setup resulted in an average exchange probability of ~20%. 16 replicas MD in the NPT ensemble for each system were performed at variable time scales (vide infra).

Comparative modeling geometry optimization

Modeller is a comparative modeling software which performs modeling by satisfying spatial restraints. The spatial restraints are obtained through the provided sequence alignment with templates and supplemented by stereochemical restraints obtained from a force field. Once a model is obtained, geometry optimization is performed employing methods of conjugate gradient and molecular dynamics with simulated annealing. Much like in REMD, simulated annealing uses heat to overcome energetic barriers. Here the system is first heated and then slowly cooled in an iterative process with the aim to find a global minimum. In this manner, energy barriers on the potential energy surface can be overcome, and an optimal energy minimum can be found.

When sampling conformational space using comparative modeling, the 1000 models of ORs needed to be aligned because they all had slightly different orientations. Because loops are mostly unstructured in the models and are not predicted to be part of the cavity only the Cα atoms of the stable part of the protein (the bundle) were aligned using cpptraj (10).

Orthosteric pocket detection

To establish the region corresponding to the binding pocket of ORs, 1000 models of a receptor containing only Glycine residues were produced. Such models were chosen to ensure

a complete coverage of the possible orthosteric binding cavity. In this model, the positions corresponding to the residues in contact with the ligand were used as boundaries for the space to be searched by MDpocket (11). MDpocket is a pocket detection algorithm (4) that was used to detect and analyze the orthosteric binding. Briefly, MDpocket's geometry-base cavity detection algorithm aims to identify and characterize binding sites from the analysis of a conformational ensemble. To track the composition in amino acids of the binding pocket, the algorithm identifies the atoms of residues which are in contact with at least one alpha sphere of the binding pocket. An alpha sphere is a sphere that contacts four atoms on its boundary and contains no internal atom. The results are returned as a cumulative list. Additional binding pocket descriptors such as pocket volume, polar surface area or hydrophobicity are estimated based on the information of: the receptor atoms, the alpha spheres contained in a given pocket, or on the residues lining it.

Phylogenetic tree building

Phylogenetic analysis was conducted by Dr. Graham Hughes at University College Dublin. The trees were built using the seven amino acid and the full sequence alignments provided by me. These alignments were built after a critical analysis of our in-house sequence alignment and by considering mutational data from the literature (Fig. 2). Dr. Hughes' role was to build the trees and assess their similarity.

Results

Potential "neo-orthologs" were identified

A	3.32	3.36	5.42	5.46	6.48	6.52	7.42
hOR1I1	F	G	F	V	Y	F	G
Olfr1357	F	G	F	M	Y	F	G
Olfr15	F	G	V	F	Y	I	S
hOR2C1	F	G	V	F	Y	S	S
Olfr223	F	S	G	F	F	M	C
Olfr1367	F	G	V	F	Y	I	C
hOR2B11	F	G	L	F	Y	I	S
Olfr222	F	G	L	F	Y	I	S
Olfr1370	F	G	V	Y	Y	I	S

B	3.32	3.36	5.42	5.46	6.48	6.52	7.42
hOR1C1	F	V	V	L	Y	I	S
hOR5V1	F	V	T	I	Y	I	S
Olfr110	F	V	I	I	Y	I	S
Olfr111	F	V	I	I	Y	I	S
Olfr290	L	A	I	C	Y	I	G
Olfr291	L	A	I	F	Y	I	G

C	3.32	3.36	5.42	5.46	6.48	6.52	7.42
Olfr1320	F	G	L	V	Y	I	T
hOR11A1	F	A	L	C	Y	M	T
Olfr96	F	A	L	F	Y	M	T
Olfr279	F	A	I	L	Y	I	T
Olfr266	F	A	T	V	Y	I	T

Figure 4: Three groups of neo-orthologs (A, B and C) were identified. Boxed receptors represent ortholog pairs (which share more than 80% sequence identity). Receptors highlighted in green are deorphanized with at least one ligand.

Our results show that three groups of ORs could form neo-ortholog pairs. They are candidates for being tested experimentally (at least one of the ORs in the group has been deorphanized)

(Fig.4). The deorphanized ORs have similar residues in their binding cavities than ORs for which no ligands have yet been identified. These theoretical results give working hypotheses to be tested *in vitro* for further validation. For example, it would be interesting to test ligands of Olfr15 (12), hOR2C1 (12) and hOR2B11 (13) on hOR1I1, Olfr1357/223/1367/222 and 1370 (Fig. 4A). Interestingly, the former (Olfr15, hOR2C1 and hOR2B11) are all deorphanized and they all respond to at least one short chained alkane (between 4 and 9 carbon in chain length) bearing a thiol function. This suggests that the "neo-orthologs" will also respond to these types of compounds.

In a past study, attempts were made to predict ligands for a novel OR using sequence (12). The success rate of predicting ligands for a novel OR was reported to be just above chance. We expect that almost 10 years later, this guided deorphanization attempt will yield better results. However, to gain more insight into the physical and chemical properties of ORs, three-dimensional models of ORs were considered. After this sequence-based approach, I elaborated a structure-based building of the human olfactory genome.

Structure-based analysis - 3D models of all human ORs

A good model aims to reflect the physical world. The main drawback of homology modeling lies in the variability of the models associated with a single sequence. Depending on the protocol or the choice of the templates used for the comparative modeling, the same target could be modeled differently. For example, in 1000 structures of human OR51E1 models, the smallest pocket volume measured using MDpocket is of ~460 $Å^3$, while the largest is of 1325 $Å^3$. Similarly, the proportion of polar atoms found in the pocket ranged from ~23 polar atoms to ~37 polar atoms. This is of course an issue if the goal is to compare precisely all receptors. Intuitively, receptors known to be similar (similar sequences) have to be similar from the modeling point of view.

I developed a protocol to make sure that similar receptors are indeed similar in terms of structure. To do this, I optimized my protocol considering four "markers" used as benchmarks. Two couples of mice and human orthologs as well as two mutants of human ORs and their wild-type were considered, as shown in Table 1. These four couples each share the same cavity sequence. They should therefore have very similar computed cavity 3D properties. It is now possible to obtain comparable representations of the olfactory genome in three-dimensions.

Table 1. Name of reference OR and their corresponding marker. The comparison will be performed on a murin ortholog in the case of hOR2W1 and hOR51E1, and between mutants of hOR2J3 an hOR7D4 (* means according to www.genecards.org).

Reference OR	Mutant or ortholog	Cavity identity	Sequence identity
Human OR2W1	Mouse Olfr263	100%	85.5%*
Human OR51E1	Mouse Olfr558	100%	87.9%*
Human OR2J3	Human OR2J3_AQ	100%	99.4%
Human OR7D4	Human OR7D4_WM	100%	99.4%

The selection of a single homology structure of an OR would be too restrictive to estimate the position of the side chains in the binding cavity. This in turn affects the binding pocket detection. Therefore, I chose to perform a thorough sampling of the conformational ensemble for each structure prior to comparing the binding pocket amongst the benchmarks. The hypothesis is that a large sampling will allow the convergence of the cavity properties between two similar receptors. This is typically the ergodic hypothesis stated by Boltzmann.

Sampling conformational space using replica exchange molecular dynamics

In first attempts to build models of all ORs, it was clear that the orientation of the residues in the binding pocket were not similar from one model to another. Here, only one model was selected for each receptor (see Methods). The markers had very dissimilar binding cavity characteristics (Fig. 5) and thus comparison and further analysis of the models could not be accurate.

Figure 5A and B report the convergence analysis. Despite running REMD simulations on different time-scales, the difference in the parameters measured between references and markers did not converge. Alternatively, sampling of the conformational space using Modeller was performed.

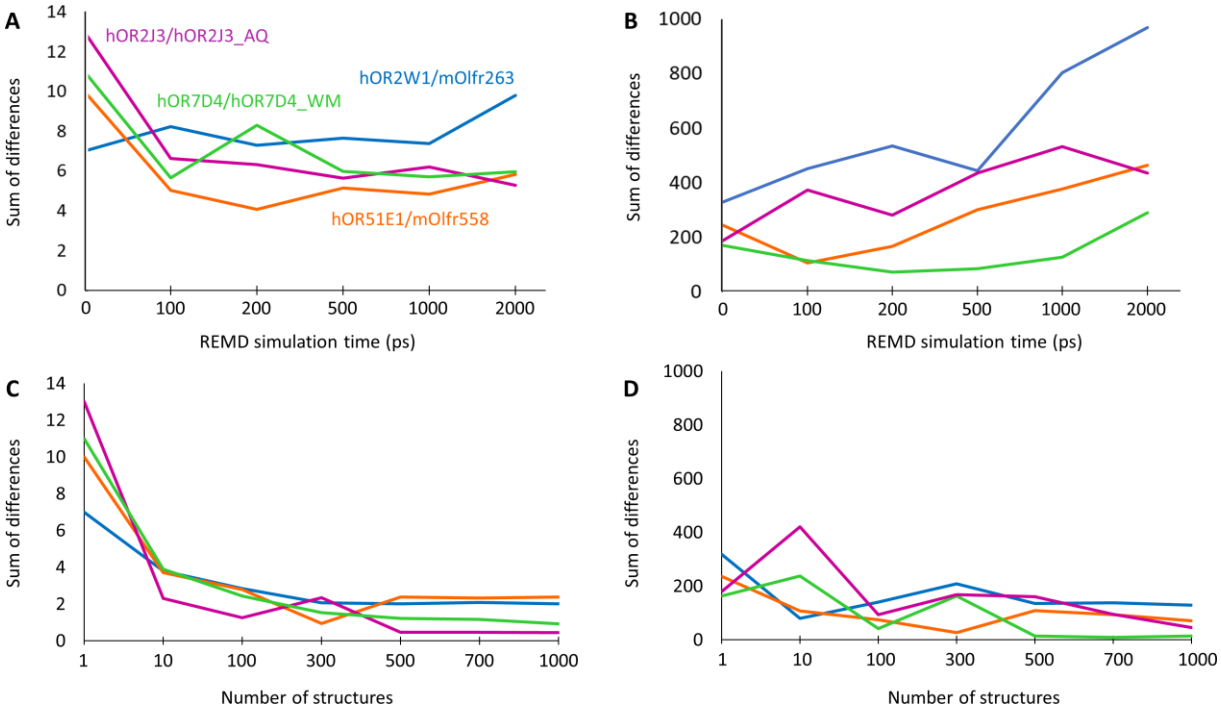


Figure 5: OR cavity sampling convergence analysis. Differences between a reference OR and its marker were measured. The larger the difference, the larger the divergence between models. A difference of zero would imply a perfect convergence. **A.** Sum of the differences in amino acids identified in the pocket as a function of REMD simulation time. **B.** Sum of the differences in the 3D descriptors as a function of REMD simulation time. **C.** Sum of differences in amino acids converges as a function of the number of structures generated by Modeller. **D.** Sum of the differences in 3D descriptors converge as a function of the number of structures generated by Modeller.

Sampling conformational space using Modeller

As can be seen on Figure 5C and D, if only one model is analyzed, the binding pocket features are very dissimilar between markers. However, the more structures are created, the smaller the differences become, signifying that each group of receptors has on average similar binding cavities. It is important to note that larger differences are expected to be observed between reference ORs and their orthologs as compared to reference ORs and their two-point mutant versions. For example, the cavities of hOR2J3 and its mutant are identical since the two amino acids in which they differ in are located on TM3 underneath the cradle of the orthosteric binding site and on intra-cellular loop 3. Consistently, the conformational sampling provided by Modeller shows both receptors have very similar binding cavities (magenta line in Figs. 5C and D). On the other hand, hOR2W1 and its mouse ortholog olfr263 share ~86% sequence similarity (Table 1). As expected, the binding cavities of these two receptors are comparable but not identical (blue line in Figs. 5C and D).

Based on this analysis, it appears that a sampling with one-thousand models ensures a correct examination of the binding cavities.

Why did Modeller perform better than REMD?

REMD in this case was performed by restraining the backbone of the protein thus sampling the movement of the side chains. Although this method is more elegant to sample conformational space than producing a large quantity of models, it proved to be extremely time-expensive. Due to this hindrance, we probably did not allow the system to converge since the measured time-scales are all rather short (below 2000ps of sampling time). On the other hand, Modeller, allowed to achieve sampling conformational space in a less time-costly manner and was thus designated as the method of choice to build our OR database. Additionally, Modeller's 1000 structure are all optimized at the minimum energy level, while REMD gives a trajectory sampling high and low energy levels.

Fig. 5A shows that REMD starts converging but this convergence is slower than the one observed in Fig. 5C which used Modeller. Restraining of the backbone atoms during the REMD conformational sampling could be in part responsible for this. In the Modeller protocol, it is possible to obtain models with slightly different backbones which also account for a larger binding pocket flexibility. It is likely that in the REMD protocol which involved a restrained backbone the side chains had less possibilities to accommodate in the cavity, which in turn affected accurate sampling.

Conclusion

Linking odorants to ORs is at the core of deciphering the combinatorial code of olfaction. To do this, it would be helpful to **tag a genetic sequence (or a protein) to a chemical space**.

In this part of the manuscript I proposed an alternative way to represent the olfactory genome (or proteome). My aim here was to re-classify ORs by similar binding cavities characteristics because similar binding cavities could mean similar ligand recognition profiles.

In collaboration with our Irish partners, phylogenetic trees of OR cavities were built. By representing **only the residues involved in ligand binding** I hope to identify new phylogenetic relationships between ORs. This sequence-based approach allowed obtaining groups of receptors which could possibly respond to the same chemicals.

To level up the sequence-based approach I used molecular modeling. In this structure-based approached, I sought to **represent the binding cavity of ORs in the most reliable way**. To do so, conformational sampling was necessary. The most efficient way to perform

conformational sampling was through the generation of 1000 3D models of ORs using Modeller. When outstretched on the entire olfactory genome, this process will allow identifying ORs with similar binding cavities, thus potentially similar chemical spaces. If deployed on human and mice ORs we might identify "neo-orthologs". Within a species we might discover "neo-homologs". These "neo-homologs" and "neo-orthologs" will be receptors which respond to the same ligands.

Outlook and future directions: building the olfactome

Besides being useful for linking receptor binding cavity properties to ligand properties, 3D models of ORs should **help probing computationally the combinatorial code of olfaction**.

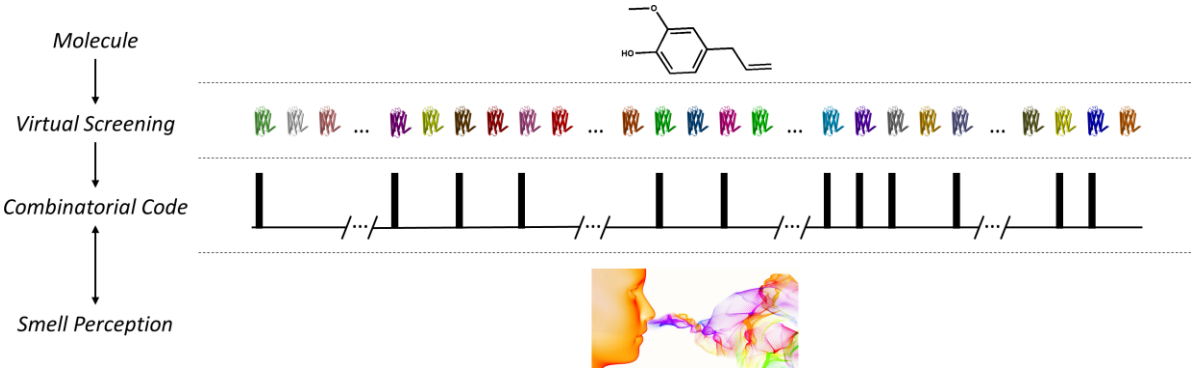


Figure 6: The olfactome. One molecule is screened on the entire 397,000 structures. Here one receptor represents the average of the 1000 Modeller structures. By virtually screening a chemical structure into these "pseudo-trajectories", it becomes possible to estimate the energy of interaction of the ligand with the OR. In turn, prediction of the activation of the receptors becomes possible. This activation pattern obtained on all ORs is the combinatorial code of olfaction. We speculate that similar combinatorial codes will be synonymous for similar smell perceptions.

The 1000 Modeller structures can be considered as a pseudo-trajectory. Docking a ligand into them will allow accounting for the intrinsic flexibility of the protein. By obtaining the average energy of interaction it will be possible to estimate the affinity of the ligand for the receptor. Deployment of this odorant/OR docking scheme on all structures should provide hypotheses to experimentally test the combinatorial code of olfaction (Fig. 6). The olfactome is thus a biologically inspired computational nose.

References

1. G. Glusman, A. Bahar, D. Sharon, Y. Pilpel, J. White, D. Lancet, The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm. Genome 11, 1016-1023 (2000).
2. C. A. de March, S. K. Kim, S. Antonczak, W. A. Goddard, 3rd, J. Golebiowski, G protein-coupled odorant receptors: from sequence to structure. Protein Sci. 24, 1543-1548 (2015).
3. C. A. de March, J. Topin, E. Bruguera, G. Novikov, K. Ikegami, H. Matsunami, J. Golebiowski, Odorant Receptor 7D4 Activation Dynamics. Angew. Chem. Int. Ed. 57, 4554-4558 (2018).
4. C. Geithe, J. Protze, F. Kreuchwig, G. Krause, D. Krautwurst, Structural determinants of a conserved enantiomer-selective carvone binding pocket in the human odorant receptor OR1A1. Cell. Mol. Life Sci. 74, 4209-4229 (2017).
5. C. Bushdid, C. A. de March, S. Fiorucci, H. Matsunami, J. Golebiowski, Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features. The journal of physical chemistry letters 9, 2235-2240 (2018).
6. A. J. Venkatakrishnan, X. Deupi, G. Lebon, C. G. Tate, G. F. Schertler, M. M. Babu, Molecular signatures of G-protein-coupled receptors. Nature 494, 185-194 (2013).
7. L. Charlier, J. Topin, C. A. de March, P. C. Lai, C. J. Crasto, J. Golebiowski, Molecular Modelling of odorant/olfactory receptor complexes. Methods Mol. Biol. 1003, 53-65 (2013).
8. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit 5-6 (2006).
9. A. Patriksson, D. van der Spoel, A temperature predictor for parallel tempering simulations. Phys. Chem. Chem. Phys. 10, 2073-2077 (2008).
10. D. R. Roe, T. E. Cheatham, 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. 9, 3084-3095 (2013).
11. P. Schmidtke, A. Bidon-Chanal, F. J. Luque, X. Barril, MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 27, 3276-3285 (2011).
12. H. Saito, Q. Chi, H. Zhuang, H. Matsunami, J. D. Mainland, Odor coding by a Mammalian receptor repertoire. Science signaling 2, ra9 (2009).
13. J. D. Mainland, A. Keller, Y. R. Li, T. Zhou, C. Trimmer, L. L. Snyder, A. H. Moberly, K. A. Adipietro, W. L. Liu, H. Zhuang, S. Zhan, S. S. Lee, A. Lin, H. Matsunami, The missense of smell: functional variability in the human odorant receptor repertoire. Nat. Neurosci. 17, 114-120 (2014).

Is emotion a chemical property?

Introduction

To achieve a better understanding of olfaction, in addition to decrypting the olfactory code, it is also necessary to identify its evolutionary function. **But what is the evolutionary function of olfaction?**

Olfaction is the sense that allows us to *perceive* odors, which are chemical stimuli present in our environment. In turn, the role of perception is to *collect information* about the environment and *guide* adaptive behaviors (1). Indeed, collecting accurate information by itself has no adaptative value while collecting accurate information and adapting the behavior provides an additional benefit for the organism.

In this part of my thesis, **I set the groundwork for studies aiming to predict emotional responses elicited by smells based on chemical features**. Such an approach will be of interest because it should allow to design and optimize odorants or perfumes to positively affect our emotional state. Such approaches are extremely appealing as they are non-pharmacologic. Using odor as a positive-effect inducer, it was shown that olfactory emotions reduced the stress responses in humans and improved the subjects' mood (2). This highlights the possibility to use smell to induce and optimize well-being. For example, it could be interesting to treat people showing loss of autonomy such as the elderly, and particularly subjects suffering from Alzheimer's disease. This population often shows a mood deterioration (i.e. high levels of anxiety) or an impoverishment of motivated behaviors (such as eating). It would be useful to establish and optimize non-invasive ways to maintain and extend these people's autonomy.

Recently, machine learning models were developed to predict human olfactory perception qualities from chemical feature to some extent (3). Odor intensity and pleasantness were shown to be quite accurately predicted but less than half of the 19 semantic descriptors were correctly estimated. The difficulty to predict these descriptors highlights a complication in olfaction: the semantic descriptors attached to odorant compounds are often not optimized to describe a smell but are often borrowed from other senses (i.e. "sweet" smell). Furthermore, our olfactory vocabulary is not standardized well enough to allow a universal classification of odorants which could be slowing down attempts to establish structure-odor relationships (4).

Also, these semantic descriptors imply a conscious categorization of odors, while most of the time we breathe in the molecules surrounding us without even being conscious of them or their importance to us.

Smells affect us on a physical, psychological and social level, and most of the time unconsciously. A protocol was developed in the laboratory by Dr. Jérémie Topin to measure the physiological changes arising upon odorant stimulation. An overview will be given here on the developed protocol, and on how machine learning can be used to predict the responses of our bodies. Here, my role was to adapt and optimize the previously presented machine learning protocol to tentatively predict a panel-averaged physiological response using chemical features as input. The measured responses are the heart beat rate variation (HBR), the skin conductance, the body temperature, and the breathing rate. Such responses are under the control of our autonomous nervous system and translate the autonomous response of our body upon a given stimulation. They can be associated with our mood or some basic emotions, such as anxiety, fear, or well-being (5).

Methods

Protocol to measure physiological response upon odorant stimulation

The following experiments were performed by Dr. Jérémie Topin. I will nevertheless shortly describe the protocol to illustrate the way he collected the data I used to build machine learning models.

In the lab, a panel of 16 volunteers was screened to measure their physiological response upon odorant stimulation. To do so, a volunteer sat on a chair and was asked to keep his eyes closed through the entire experiment. Volunteers were stimulated twice with five odorants and water was used as a control. The session was repeated another day using a randomized order of odorants. Physiological parameters were recorded continuously during the whole session. The procedure is described in Figure 1.

After a resting period of two minutes, odorants were presented to the subject every minute by placing the flask approximately five centimeters under the nose. Stimulation was synchronized with the beginning of the inhalation and extended for three full breathings.

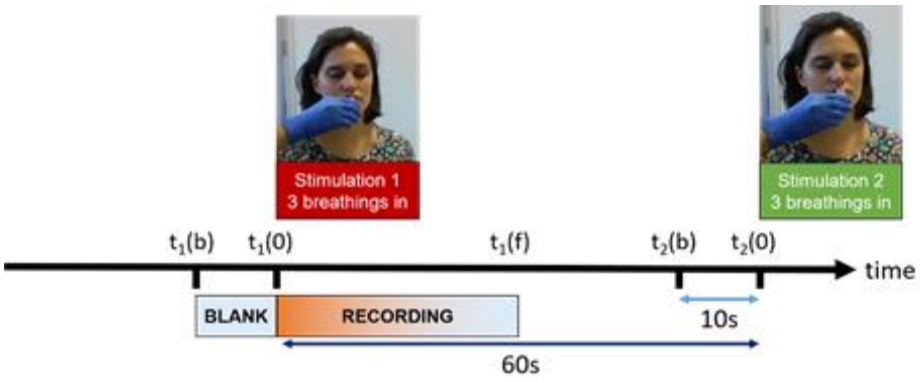


Figure 1: Olfactory stimulation protocol. The "blank" period is considered 10s prior to the beginning of an odorant stimulation (from $t_1(b)$ to $t_1(0)$). At $t_1(0)$, the first odorant is presented to the subject 5cm under his/her nose. Throughout three complete breathings, the subject inhales the odorant and physiological parameters are recorded. After having completed the three breathing cycles ($t_1(f)$), the odorant is removed from under the subjects' nose until 60s are reached after the beginning of the stimulation. The last 10s (from $t_2(b)$ to $t_2(0)$) of the 60s experiment are then used as a blank for the next stimulation which starts at $t_2(0)$.

A minimum of two sessions were considered for each subject (one male and one female volunteer participated to four sessions). The odorants delivery order was different between the two sessions. 40 stimulations were performed for each odorant.

The percentage of variation between blank and stimulation was defined to normalize the effect of odorant inhalation on heart beat rate, skin conductance, respiration rate and temperature. (6, 7) The values from all parameters acquired during the blank (10 seconds prior the stimulation) and those acquired during the stimulation period (from 0 to 25 sec) were separately averaged. The percent signal change (%variation) for each odorant was then calculated.

For each odorant, the percentage of variation compared to that obtained with water stimulation (considered as a blank), was computed. The data thus represents the response with respect to a blank odor.

Building models to predict physiological responses

Forty-four compounds were tested for physiological effects and 4884 chemical, topological and three-dimensional molecular descriptors calculated for each molecule using Dragon software (5). Descriptors describing the three-dimensional features of molecules were excluded because the stereochemistry of certain tested compounds was not precisely known. The best SVM model was obtained using classification (classifying odorants into enhancing HBR and diminishing HBR). Of the 44 compounds, 35 used in the training set, and nine were

left out to be tested on the obtained model. (Fig. 2A) In the training set, 19 molecules diminished the HBR while 16 enhanced the HBR, it was thus considered to be “balanced”.

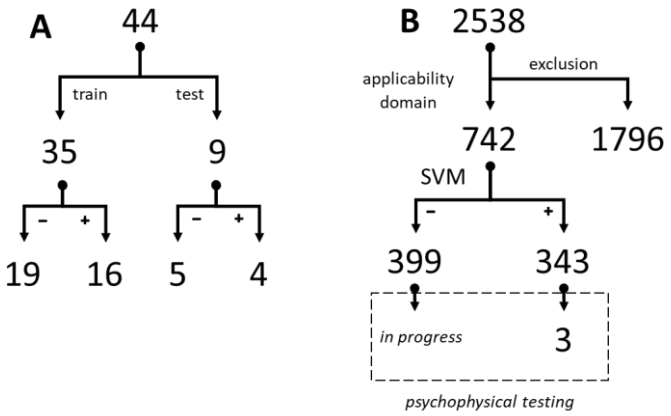


Figure 2: A. 44 compounds were used as input to create and validate the models on heart beat rate (HBR) evolution. Of these, 9 were used in the test set and 35 in the training. Both test and training set were balanced in both cases, i.e.~50% of molecules enhanced HBR and ~50% diminished it. **B.** Workflow for screening novel compounds on the model. On a total of 2538 molecules within a commercial library, 742 molecules belonged to the same applicability domain as the training set, of these, 343 were predicted as enhancing HBR (i.e. 399 with an opposite effect). I predicted the psychophysical response on 3 molecules at the time of writing of this manuscript.

The used algorithm was a support vector classification (ν-SVC), and the linear kernel (other kernels did not yield better results). SVM parameters (such as ν and ε) were iteratively optimized until an acceptable Matthew’s coefficient was obtained on the test set. The parameter ν approximates the fraction of training errors and support vectors, for example, if set to 0.05, 5% of the training set is allowed to be misclassified, and at least 5% of the training examples will be support vectors. The parameter ε is the tolerance of termination criterion, it represents the maximum number of iterations that will be reached in the optimization process of the model.

The best model had a Matthew’s coefficient of 0.35 on the test set, and 0.25 on the training set. Regression models using SVM were also tested, but the results were not satisfying compared to classification models. Matthew’s coefficient (MCS) allows evaluating the performance of a model with newly tested molecules. Empirically, a good score is of ~0.3 and above, +1 representing a perfect learning of the SVM model. It follows the formula:

$$MCS = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + TP)(TN + FN)}}$$

Where TP is the number of true positives, TN is the number of true negative, FP is the number of false positives, and FN is the number of false negatives.

Alternatively, models were also built to predict skin conductance. However, they all yielded bad results. Indeed, the training and test sets were not balanced in terms of enhancing and diminishing molecules. In particular here, skin conductance is enhanced in 88% of the molecules in the training (Table 1). The models were thus biased towards predicting enhancement and were considered unfit for further predictions.

Table 1. Physiological parameters of the 35 compounds used to generate models.

	Heart Beat Rate	Skin Conductance	Respiration Rate	Temperature
Enhancing	16	31	18	13
Diminishing	19	4	17	22

Results

Predicting heart beat rate from chemical features.

Molecular descriptors for a library of 2538 compounds for which physiological parameters had never been tested but which have a potentially odorant character (6) were calculated. Of these 2538 compounds, 742 had a *Tanimoto* similarity score larger than 0.8 when compared to the training set. They were thus considered to be sufficiently similar to the training set to be virtually screened on the model. The model predicted 399 compounds to cause a decrease in HBR while 343 compounds were predicted to act as HBR activators (Fig.2b). Ten compounds were selected to be used as an external validation set (5 with increasing HBR activities, and 5 with decreasing HBR activities), but only three could be tested before the writing of this manuscript. All the three compounds had an enhancing activity as predicted by the model. (Fig. 3)

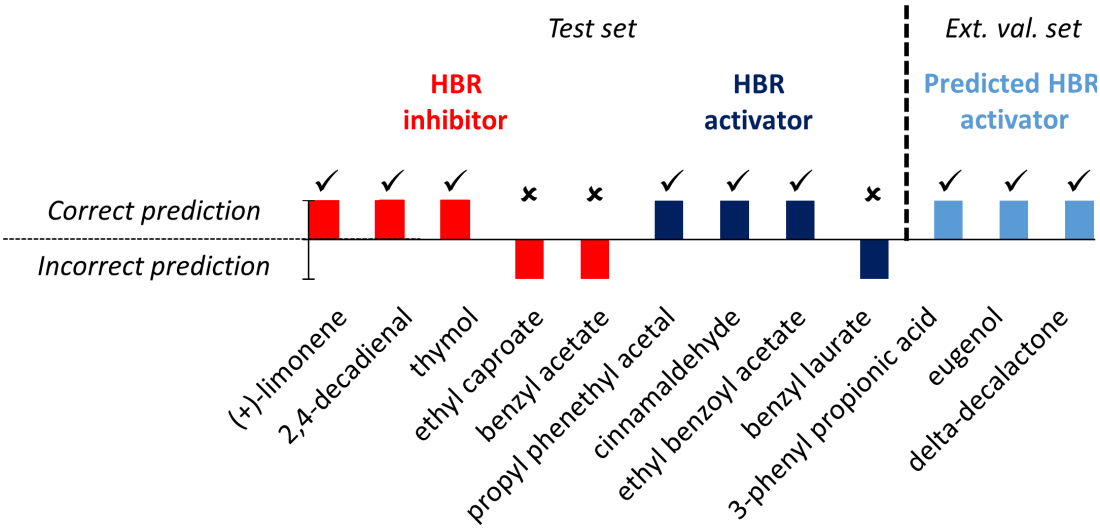


Figure 3: Correct and incorrect prediction of the model on the test set and the external validation set. Three molecules out of five are correctly predicted to be HBR inhibitors, and 3 out of 4 are correctly predicted to be activators. In the external validation set, three molecules predicted to activate HBR, were correct predictions. Further prediction results are pending.

Outlook and future directions

The model predicting HBR variation had a Matthew's coefficient on the test set of 0.35 and a global error rate of 33% (3 incorrectly predicted molecules out of nine, Fig. 3). The results of the prediction were nevertheless particularly encouraging since the predictions on three molecules in the external validation set were further assessed by experiment. The model suggests that HBR is partly encoded in the chemical structure. It seems promising to continue enriching the models with more data and experimenting with other machine learning methods. Caution needs however to be taken: even though our results are stimulating, the error rate on the test set remains rather high. Furthermore, our models could be limited by intra- and inter-individual differences.

An inherent limit of psychophysical data is the intra-individual difference (variation between the responses of a single individual upon identical stimulations at different times). Keller *et al*. (1) already reported a plateau in terms of the performance of their models which is due to the intra-individual differences. These were measured and allowed establishing the theoretical limit of their models. Our measurements could inherit of these limits of "inconsistent" data from individuals depending on the time of the day, their internal emotional state, and other subjective factors.

In addition to this intra-individual difference, inter-individual differences also present a limit to the model. It can be speculated that individual olfactory perception reflects the olfactory receptor subtype repertoire as well as cultural aspects. From this point of view, to adapt the models to populations, volunteers could be genotyped to make groups of people having a similar olfactory receptor repertoire. It is however important to note, that we expect that through our protocol, responses of evolutionary relevant molecules should be the same despite these genetic variations. Indeed, if smell is considered in its evolutionary framework, potentially toxic or unpleasant smells (such as rotten or diseased smells) are expected to have negative emotional responses. On the contrary, smells presenting an evolutionary advantage are expected to elicit a positive emotional response.

References

1. A. Keller, **Philosophy of Olfactory Perception**, Astoria, NY, USA (2016).

2. J. Lehrner, C. Eckersberger, P. Walla, G. Potsch, L. Deecke, Ambient odor of orange in a dental office reduces anxiety and improves mood in female patients. Physiol. Behav. 71, 83-86 (2000).

3. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. Science 355, 820-826 (2017).

4. A. Keller, L. B. Vosshall, Olfactory perception of chemically diverse molecules. BMC Neurosci. 17, 55 (2016).

5. E. A. Krusemark, L. R. Novak, D. R. Gitelman, W. Li, When the sense of smell meets emotion: anxiety-state-dependent olfactory processing and neural circuitry adaptation. J. Neurosci. 33, 15324-15332 (2013).

6. M. L. Loggia, M. Juneau, M. C. Bushnell, Autonomic responses to heat pain: Heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity. Pain 152, 592-598 (2011).

7. E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, J.-H. Sohn, Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. J. Physiol. Anthropol. 34, 25 (2015).

Summary and conclusion (EN)

Smell perception is triggered by the interaction of odorant receptors (ORs) and odorants. Using about 400 ORs and employing a combinatorial code, the human olfactory system detects hundreds of thousands of diverse chemical stimuli and achieves to make nuanced discriminations between them. A detailed understanding of ORs is necessary to fully decrypt the sense of smell.

The aim of this PhD thesis was to use numerical models to deepen our knowledge on olfaction. I focused on studying its principal protagonists: odorants and ORs.

Numerical methods such as machine learning and cheminformatics were used to predict receptor activation and even psychophysiological responses. I notably developed a protocol to predict the activity of an OR based on the features of chemical compounds. This allowed me to virtually screen a database of compounds and predict with a high reliability agonists for some receptors. This type of procedure helps exploring the chemical space associated with four ORs and expanding it. Additionally, the same type of protocol was used to predict human physiological changes measured upon odorant stimulation, paving the way towards the establishment of a structure-emotion relationship.

Moreover, I used molecular modeling and bio-informatic sequences analysis to investigate molecular recognition. In particular, I explored the role of the ligand binding cavity in ORs. A conserved vestibular binding pocket in class I ORs was thus identified. This type of ligand binding site has been observed in crystal structures as well as in other chemosensory receptors (bitter taste receptors and trace amine associated receptors). This discovery describes early stages of OR/ligand recognition and the possible role of allosteric modulators in their activation. Furthermore, these techniques were used to examine the role that the orthosteric binding cavity plays in ligand recognition and in the receptors' response. Identification of residues pointing towards the binding cavity, and subsequent mutation allowed transforming a narrowly tuned receptor into a broadly tuned one. Interestingly, molecular recognition and basal activity are governed by the cavity, while the strength of the response is mostly controlled by other mechanisms. By applying these techniques on the entire genome, OR deorphanization could be improved. Through sequence analysis, identification of so-called "neo-homologs" pairs should allow identifying new odorant/OR pairs. These "neo-homologs"

are receptors having different overall sequences but similar binding sites. In addition, structural analysis and optimization made it possible to compare models of ORs. This provides a possibility to deorphanize ORs by linking the physical and chemical properties of their cavities to the physical and chemical features of molecules forming odorant space. In the near future, these models will also allow probing the combinatorial code of olfaction computationally.

In addition to the research carried out in this thesis, my research activities prior to enrolling in my PhD concerned sensory analysis to infer the number of olfactory stimuli the nose is able to discriminate.

- Bushdid C., Magnasco M., Vosshall L.B., Keller A. 'Humans Can Discriminate More than 1 Trillion Olfactory Stimuli' *Science* 343, 1370-1372, **2014**

During my PhD and in collaboration with the DGIST in the Republic of Korea, I participated in the discovery and characterization of an ectopically expressed odorant receptor (OR10J5). I built a model of this receptor and calculated the energy of interaction between the ligand and the protein.

- Tong T., Ryu S.E., Min Y., de March C.A., Bushdid C., Golebiowski J., Moon T., Park T. 'Olfactory receptor 10J5 responding to alpha-cedrene regulates hepatic steatosis via the cAMP-PKA pathway' *Sci. Rep.* 7, 9471, **2017**

In another collaboration with the Centre de Recherche en Neurosciences in Lyon, my expertise as a molecular chemist was put forward. I was responsible for interpreting the rules that govern structure-odor relationships which were discovered using datamining.

- Licon C.C., Bosc G., Sabri M., Mantel M., Fournel A., Bushdid C., Golebiowski J., Robardet C., Plantevit M., Kaytoue M., Bensafi M. 'Chemical features mining provide new descriptive structure-odor relationships' ***submitted***

Finally, a review in French was published in ***Pollution Atmosphérique*** where the molecular origin of smell nuisances is reported.

- Bushdid C., Topin J., Golebiowski J. 'When the atmosphere smells like sulfure. Perception of malodors.' *Pollut. Atmos.* 234, 3, **2017**

My main research activities allowed establishing a solid groundwork for future studies concerning prediction and research in olfaction using computational methods. The work presented in this manuscript should particularly benefit computer aided deorphanization of ORs, but also virtual screening to test the combinatorial code of olfaction. In the near future machine learning and molecular modeling could be combined. Possibly, sequence-based

descriptors as well as 3D cavity descriptors will be linked to chemical descriptors (i.e. through proteochemometrics) and to activities or even psychophysiological measurements.

Given the complexity of odorant space and the involvement of a combinatorial code, many fundamental questions remain. Some may be successfully addressed by applying numerical approaches as established in this manuscript. These undeniably powerful methods will help to achieve a better understanding of the sense of smell.

Résumé et conclusion (FR)

La perception d'une odeur est déclenchée par l'interaction des molécules odorantes avec les récepteurs odorants (ROs). Grâce à un code combinatoire et en mobilisant plusieurs centaines de ROs, le système olfactif humain est capable de détecter et de discriminer de façon nuancée entre des milliers de stimulus différents. Une compréhension détaillée de ces protéines est donc nécessaire pour décrypter notre odorat.

L'objectif de cette thèse était d'utiliser des modèles numériques pour approfondir nos connaissances sur l'olfaction. Pour ce faire, je me suis concentrée sur l'étude de ses principaux protagonistes : les odorants et les ROs, en appliquant des modèles numériques.

Ici, des **méthodes numériques telles que l'apprentissage automatique et la chemo-informatique** ont été utilisées pour prédire l'activation de RO ainsi que des réponses psychophysiologiques. J'ai notamment développé un protocole pour prédire l'activation d'un RO en « apprenant » des caractéristiques des composés chimiques qui l'activent. Cela m'a permis de prédire de nouveaux agonistes avec une haute fiabilité pour quatre récepteurs. Ce type de procédure permet l'exploration voire l'expansion de l'espace chimique associé à un RO. Le même type de protocole a été utilisé pour prédire les changements physiologiques mesurés lors d'une stimulation odorante, ceci a permis d'ouvrir la voie pour l'établissement d'une relation structure-émotion.

Au cours de ma thèse j'ai aussi utilisé la **modélisation moléculaire et l'analyse bio-informatique de séquences** pour étudier la reconnaissance moléculaire dans les ROs. En particulier, j'ai exploré le rôle de la cavité dans les ROs. Une cavité de liaison vestibulaire conservée dans les ROs de classe I a ainsi été identifiée. Ce type de site de liaison a aussi été observé dans les structures cristallines de RCPGs classe A ainsi que dans d'autres récepteurs chimio-sensoriels (comme les récepteurs au goût amer et les récepteurs associés aux amines à l'état de trace). Cette découverte permettra l'approfondissement de nos connaissances sur le mécanisme moléculaire mis-en-jeu lors des premières étapes de la reconnaissance des ligands et d'étudier le rôle des modulateurs allostériques dans l'activation des ROs. Ces techniques ont aussi été utilisées pour examiner le rôle que joue la cavité de liaison orthostérique dans la reconnaissance des ligands et dans la réponse des récepteurs. Ici, l'identification des résidus composant la cavité de liaison et ses mutations *in vitro* ont permis de transformer un récepteur

à spectre de reconnaissance étroit en un récepteur à spectre de reconnaissance large. Il est intéressant de noter que la reconnaissance moléculaire et l'activité basale d'un RO dépendent de la composition de la cavité, alors que la force de la réponse d'un RO semble être encodée dans d'autres parties du récepteur.

Enfin, l'application de ces techniques sur l'ensemble du génome olfactif devrait faciliter la déorphanisation de ROs. En effet, grâce à l'analyse bio-informatique de séquence, des paires de « néo-homologues » permettront d'identifier de nouvelles paires odorant/ROs. Ces « néo-homologues » sont des récepteurs ayant des séquences différentes mais des cavités de liaison similaires. De plus, l'optimisation de la structure des modèles permettra de comparer les ROs. Cela offre la possibilité de déorphaniser les ROs en reliant les propriétés physiques et chimiques de leurs cavités aux caractéristiques physiques et chimiques des molécules candidates. Dans un avenir proche, ces modèles permettront aussi de tester le code combinatoire de l'olfaction computationnellement.

En plus des recherches effectuées dans cette thèse, mes activités de recherche avant mon doctorat ont porté sur l'utilisation de l'analyse sensorielle pour déduire le nombre de stimuli olfactifs que le nez est capable de discriminer.

- Bushdid C., Magnasco M., Vosshall L.B., Keller A. 'Humans Can Discriminate More than 1 Trillion Olfactory Stimuli' *Science* 343, 1370-1372, **2014**

Au cours de ma thèse et en collaboration avec la DGIST en Corée du Sud, j'ai participé à la découverte et à la caractérisation d'un récepteur olfactif ectopique (OR10J5). Ici, j'ai construit un modèle de ce récepteur et calculé l'énergie d'interaction entre ligands et récepteurs.

- Tong T., Ryu S.E., Min Y., de March C.A., Bushdid C., Golebiowski J., Moon T., Park T. 'Olfactory receptor 10J5 responding to alpha-cedrene regulates hepatic steatosis via the cAMP-PKA pathway' *Sci. Rep.* 7, 9471, **2017**

Dans le cadre d'une collaboration avec le Centre de Recherche en Neurosciences de Lyon, mon expertise en chimie moléculaire a été mise en avant pour interpréter les règles régissant les relations structure-odeur découvertes à l'aide de datamining.

- Licon C.C., Bosc G., Sabri M., Mantel M., Fournel A., Bushdid C., Golebiowski J., Robardet C., Plantevit M., Kaytoue M., Bensafi M. 'Chemical features mining provide new descriptive structure-odor relationships' ***soumis***

Enfin, une revue en français a été publiée dans ***Pollution Atmosphérique*** où l'origine moléculaire des nuisances olfactives est passée en revue

- Bushdid C., Topin J., Golebiowski J. 'When the atmosphere smells like sulfure. Perception of malodors.' *Pollution Atmosphérique*, 234, 3, **2017**

Mes principales activités de recherche ont permis d'établir une base solide pour de futures études concernant la prédiction et la recherche en olfaction à l'aide de méthodes numériques. Les travaux présentés dans ce manuscrit devraient particulièrement bénéficier à la déorphanisation assistée par ordinateur des ROs, mais également au criblage virtuel pour tester le code combinatoire de l'olfaction. Il est possible que l'apprentissage automatique et la modélisation moléculaire puissent être combinés dans le futur. Eventuellement, les descripteurs basés sur les séquences seront liés à des descripteurs chimiques (par des méthodes de protéochémométrie par exemple) ; mais on peut également envisager que des descripteurs de l'espace de cavités soient liés à des activités voire à des réponses psychophysiologiques.

Compte tenu de la complexité de l'espace odorant et de l'implication d'un code combinatoire, de nombreuses questions fondamentales subsistent. Celles-ci peuvent être traitées avec succès en appliquant des approches numériques comme cela est établi dans ce manuscrit. Ce sont des méthodes indéniablement puissantes qui faciliteront grandement la compréhension de l'odorat.