Molecular basis of chemosensory perception

Cédric Bouysset

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES FONDAMENTALES ET APPLIQUÉES

THÈSE DE DOCTORAT

Bases moléculaires de la perception chimiosensorielle

Molecular basis of chemosensory perception

Cédric BOUYSSET

Institut de Chimie de Nice, UMR CNRS 7272

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

Dirigée par : Dr. Sébastien Fiorucci, Pr. Serge Antonczak

Soutenue le : 21 octobre 2021

Devant le jury, composé de :

Pr. Serge Antonczak, Université Côte d'Azur

Dr. Sébastien Buchoux, Evotec

Dr. Sébastien Fiorucci, Université Côte d'Azur

Pr. Esther Kellenberger, Université de Strasbourg

Pr. Matthieu Montes, Conservatoire National des Arts et Métiers

Pr. Jana Sopkovà-de Oliveira Santos, Université de Caen Normandie

Bases moléculaires de la perception chimiosensorielle

Jury :

Rapporteurs
Esther Kellenberger, Professeure des universités, Université de Strasbourg
Jana Sopkovà-de Oliveira Santos, Professeure des universités, Université de Caen Normandie

Examinateurs
Sébastien Buchoux, Docteur en Sciences chimiques (Chimie-Physique), Evotec
Matthieu Montes, Professeur des universités, Conservatoire National des Arts et Métiers

Directeurs
Sébastien Fiorucci, Maître de conférences HDR, Université Côte d'Azur
Serge Antonczak, Professeur des universités, Université Côte d'Azur

Abstract

Molecular basis of chemosensory perception

Smell and taste perception consists in a chemical stimulation of transmembrane receptors lying on the surface of sensory cells located in the nasal or oral cavity. The receptors involved in olfaction and the perception of bitter, sweet and umami taste all belong to the well-studied G protein-coupled receptors (GPCR) family, yet to date their exact tridimensional structure still eludes us. In this thesis, I study the molecular structures at the frontline of chemosensory perception, namely receptors and their ligands, through a computational lens. To begin with, I expose how quantitative structure-activity relationships (QSAR) lead us to the discovery of natural semiochemicals that effectively disrupt the destructive behavior of a pest to crop plants, *Spodoptera littoralis*, by targeting its olfactory receptors. I then apply a similar machine learning strategy to develop an online predictive platform that estimates the relative sweetness of molecules based on their structure, resulting in the discovery of a novel sweet-tasting lignan scaffold. Finally, I make use of molecular modeling and available mutagenesis data to provide relevant three-dimensional models of bitter taste receptors, and predict molecular switches involved in ligand-sensing and receptor activation. Besides, I design a Python library that encodes interactions in molecular complexes as fingerprints for an efficient analysis of molecular dynamics trajectories, docking results and experimental structures, and showcase it on a variety of scenarios involving GPCRs. Overall, this thesis illustrates the implementation of computational strategies to gain knowledge on chemosensory perception, from taste to olfaction, at the molecular level.

Keywords: Molecular modeling, machine learning, chemoinformatics, olfaction, taste, GPCR

Bases moléculaires de la perception chimiosensorielle

La perception olfactive et gustative provient d'une stimulation de nature chimique des récepteurs transmembranaires émergeants de la surface de cellules sensorielles situées dans la cavité nasale ou orale. Les récepteurs impliqués dans la perception des odeurs et des goûts amer, sucré et umami appartiennent à la famille des récepteurs couplés aux protéines G (RCPG). Malgré une connaissance approfondie de cette famille, une définition précise de la structure tridimensionnelle des RCPG chimiosensoriels nous échappe encore à ce jour. Dans cette thèse, je mets en lumière les structures moléculaires en première ligne de la perception chimiosensorielle, à savoir les récepteurs et leurs ligands, au travers d'un microscope computationnel. Dans un premier temps, je mets en avant un cas concret d'utilisation des relations quantitatives structure à activité (QSAR) par la découverte de composés sémiochimiques naturels interférant avec le comportement destructeur d'un insecte ravageur de cultures agricoles, *Spodoptera littoralis*, en ciblant ses récepteurs olfactifs. Par la suite, en me basant sur une méthode d'apprentissage automatique similaire, je conçois une plateforme en ligne permettant la prédiction du pouvoir sucrant de molécules en se basant sur leur structure, ce qui nous a permis de révéler un composé sucré innovant de la famille des lignanes. Pour finir, grâce aux outils de modélisation moléculaire et aux données de mutagénèse dirigée, je construits des modèles 3D de récepteurs au goût amer afin de prédire les interrupteurs moléculaires impliqués dans la détection des ligands et l'activation de ces récepteurs. En parallèle, je développe une librairie Python qui encode les interactions de complexes moléculaires sous forme d'empreinte numérique afin d'analyser des trajectoires de dynamique moléculaire, des structures issues d'amarrage moléculaire, ou des structures expérimentales, et je mets en valeur ce logiciel dans une multitude de scenarios impliquant des RCPG. Dans l'ensemble, ces travaux de thèse illustrent la mise en œuvre de méthodes numériques pour extraire des informations sur la perception chimiosensorielle, qu'elle soit olfactive ou gustative, à l'échelle moléculaire.

Mots clés : Modélisation moléculaire, apprentissage automatique, chémoinformatique, olfaction, gustation, RCPG

Acknowledgments

First, I would like to thank my supervisors, Dr. Sébastien Fiorucci and Pr. Serge Antonczak, for welcoming me in their team. While I'm not the most expressive person, I always greatly appreciated your support and guidance. More specifically to Seb, as you managed to give me freedom in my research, yet you were always present whenever I needed advice and feedback or for listening to me, and this has been immensely helpful.

I would also like to express my gratitude to my thesis jury, Pr. Esther Kellenberger and Pr. Jana Sopkovà-de Oliveira Santos, as well as Dr. Sébastien Buchoux and Pr. Matthieu Montes, for accepting to be referees and inspectors of my research. I also thank Dr. Nathanaël Guigo and Dr. Romain Gautier for being part of my thesis monitoring committee.

I am grateful to the French Ministry of Higher Education and Research, GIRACT and the Gen Foundation for the financial support given to me to conduct my research, as well as ECRO for their travel grant, and the Complex Systems Academy of Excellence and the SFCi for awarding me a prize.

A lot of the value of this thesis comes from experimental validation performed by our collaborators, so I would like to take the opportunity to thank Dr. Loïc Briand and Dr. Christine Belloir from the CSGA (Dijon), Dr. Gabriela Caballero-Vidal and Pr. Emmanuelle Jacquin-Joly from the iEES (Versailles), Peihua Jiang from the Monell Center (USA), and MeeRa Rhyu from the KFRI (Korea). I would have liked to learn more about site directed mutagenesis and *in vitro* testing directly from you, but unfortunately some virus decided otherwise.

On a more personal note, I would like to properly thank all the people who I've spent memorable times with during my PhD thesis. This includes the rest of the team: Jérôme for all the computer infrastructure that you built over the years, for pushing me to apply for grants, and for giving me the keys to a successful oral presentation. Jérémie, maybe not for your Python scripting skills (although I must confess I had fun helping you with matplotlib) or your old-school artistic choices (at some point you have to move past VMD and the greyscale), but definitely for your continuous guidance, your modeling expertise, your cooking skills and your silly jokes. Matej, who I guess will become the new Python expert, and Xhino, even if this pandemic has made it a bit harder to get to know each other. I'd also like to thank ex members

of the lab, especially Xiao for the great food, the various trips, the salsa lessons and, in general, all the joyful moments. I'd also like to thank Maxence for beta-testing some of my software projects and for the great work you've accomplished with my models. And last but not least, Jody, my "brollègue", who had no choice but to bear with me for 3 years (or is it the other way around?)! You may not shine with your cooking skills and your dubious taste for food, but I had great fun converting you from the evil side (R) to the good guys (Python), fixing your computer (if only there was a bugfix for the chair-to-keyboard interface problem), exchanging scientific ideas and spending those 3 years by your side.

I would also like to thank all my friends, from the ICN and elsewhere: Ben, Bertille, Simon, Manon, Aurel, Ana, Laurane, Johnny, Marie, Emilie "capt'n crabe", Emilie PEF, Conny, and Tammy, you've made these years unforgettable to me through your good humor and your friendliness, and I cannot thank you enough for it. A special mention for Boris, my Overwatch buddy, we may not be the best (definitely not), but it was a fun ride! Speaking of video games, I have a special thought for my "virtual" friends, the CUT team, especially Alu, God, Fox and Anod1 who have accompanied me from my bachelor years to the end of this academic journey, no matter which game we play it's always a pleasure to "fight" by your side.

I would also like to acknowledge the technical support from the MDAnalysis core developers, especially Irfan, Richard, Fiona, Oliver, and Lily, for providing me with their extremely valuable knowledge and experience in Python software development, thanks to your help my skills have vastly improved. I also want to thank random internet people who answered technical questions on forums or blog posts and indirectly helped me through that. I'm also grateful to the creators behind some of the software or websites that I used to generate the figures displayed in this thesis. This includes BioRender.com, which I used to generate some of the figures for biological pathways presented in this document, but also ChimeraX, PyMOL and VMD for 3D structures, as well as RDKit for 2D ones.

Last but not least, I would like to thank my family, especially my father, mother and brother, from the bottom of my heart. You have provided me with your endless support, your love, and your care, and even if I'm not very skilled when it comes to expressing my emotions, I hope that this paragraph shows my deep and undying love for you. I would also like to thank Mika, the "goodest" dog, for her amazing skill of making me smile just from her presence.

Table of contents

Acronyms

AdaBoost	Adaptive Boosting
Cryo-EM	Cryo-Electron Microscopy
EAG	Electroantennography
EC_{50}	Half maximal effective concentration
GPCR	G Protein-Coupled Receptor
kNN	k-Nearest Neighbors
logSw	Relative sweetness in logarithmic scale
LOO	Leave-One-Out
MD	Molecular Dynamics
ML	Machine Learning
MOE	Main Olfactory Epithelium
MSA	Multiple Sequence Alignment
OR	Olfactory Receptor
OSN	Olfactory Sensory Neuron
PCA	Principal Component Analysis
PDB	Protein Data Bank
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
R	Ramachandran number
RF	Random Forest
RMSD	Root Mean Square Deviation
SlitOR	*Spodoptera littoralis* Olfactory Receptor
SMILES	Simplified Molecular Input Line Entry Specification
SNP	Single-Nucleotide Polymorphism
SSR	Single Sensillum Recordings
SVM	Support Vector Machine
T1R *TAS1R*	Taste Receptor type 1
T2R *TAS2R*	Taste Receptor type 2 / Bitter taste receptor
TAAR	Trace Amine-Associated Receptor
TM	Transmembrane domain
TRC	Taste Receptor Cell
t-SNE	t-distributed Stochastic Neighbor Embedding
V1R *V2R*	Vomeronasal type-1 or type-2 Receptor
VNO	Vomeronasal Organ
WT	Wild type

Introduction

Chemosensory perception

Chemosensation allows multicellular organisms to evaluate the chemical composition of their surroundings and communicate with each other, a crucial ability for both survival and reproduction [1]. This type of primal sense has evolved in humans and most animals to be distinguished in two senses: olfaction and taste. Olfaction corresponds to the detection of volatile compounds by olfactory sensory neurons located in the nasal cavity. Depending on the route taken by said volatile compound, the resulting percept will be defined as an odor (through the nostrils i.e., the orthonasal pathway) or an aroma (from the oral cavity and through the pharynx i.e., the retronasal pathway).

Taste originates from the detection of sapid molecules by taste buds mostly located in lingual papillae on the tongue. The taste sensation is comprised of 5 basic taste modalities, namely saltiness, sourness, bitterness, sweetness, and umami [2], and should not be mistaken with chemesthesis. The latter, also termed trigeminal sense, corresponds to sensations detected by the somatosensory system and includes pungency, coolness, astringency, and metallicness [3]. Additionally, the characterization of fat taste, also referred to as oleogustus, as a basic taste modality is still under debate [4]. Finally, the flavor of food items is a multisensory modality that results from aroma (the perception of odorant compounds released during mastication by the retronasal pathway), in conjunction with taste and chemesthesis.

Anatomy of taste and smell

In vertebrates, the olfactory system is divided in two systems: the main olfactory epithelium (MOE) mainly responsible for odorant detection, and the vomeronasal organ (VNO) which mainly detects pheromones, although both organs can detect odorants and pheromones [5, 6] (Figure 1a). In insects, the olfactory sensory neurons (OSNs) are housed in sensilla (sensory hairs) that can be found on the maxillary palp and antennas [7], while the MOE is located below the cribriform plate in the nasal cavity of mammals (Figure 1b). These OSNs are, in both insect and vertebrates, bipolar neurons that extend a dendrite ending in ciliated projections, while the axon joins specialized olfactory structures in the brain [7, 8] (Figure 1ab).

In mammals, OSNs follow the one-receptor one-neuron paradigm where a single neuron only expresses a specific olfactory receptor (OR) [9]. It enables a combinatorial code where one odorant can activate multiple neurons each expressing a different OR, and one neuron can be activated by a diversity of odorants. With around 400 functional OR genes [10], thanks to such combinatorial code humans have been estimated to be able to discriminate more than 1 trillion odors [11], although these claims have been firmly disputed [12, 13].

Taste, on the other hand, is detected by specialized gustatory cells found in taste buds. The majority of taste buds are usually located on the tongue within papillae, but some exceptions have been found, like chickens for which they are mostly found on the palate and lower beak [14]. Additionally, gustatory papillae are also located on the palate, pharynx, larynx, and upper esophagus [15]. Taste buds are made of five types of cells, of which two are responsible for gustatory functions, namely receptor (type II) and presynaptic (type III) taste cells [16] while the other cells include glia-like (type I), basal (type IV) and marginal (type V) cells [15] (Figure 1c). Type II cells primarily expresses receptors responsible for sweet, umami and bitter taste perception, while type III cells respond to salty and sour stimuli. Since only type III cells have synaptic contact with nerve fibers, type II cells use ATP as a neurotransmitter to activate presynaptic cells or nerve fibers directly. While type II cells are tuned to a single taste modality since they primarily express a single class of taste receptor for either umami, bitter, or sweet, type III cells can respond more broadly to other tastants, especially bitter compounds, and can integrate the signal from neighboring receptor cells [16].

To explain taste coding i.e., how the afferent nerve fibers carry taste stimuli to the brain, two hypotheses have been proposed and are still under discussion: the “labelled line” and “across-fibre pattern” models [17, 18]. On the one hand, the “labelled line” model suggests that each afferent fiber is tuned to a specific taste, on the other hand, the “across-fiber” model states that the afferent fibers can transmit information for several taste modalities. Recent advances seem to favor a combination of both models, but also stress the importance of temporal coding since firing rates could play a role in encoding taste quality [19].

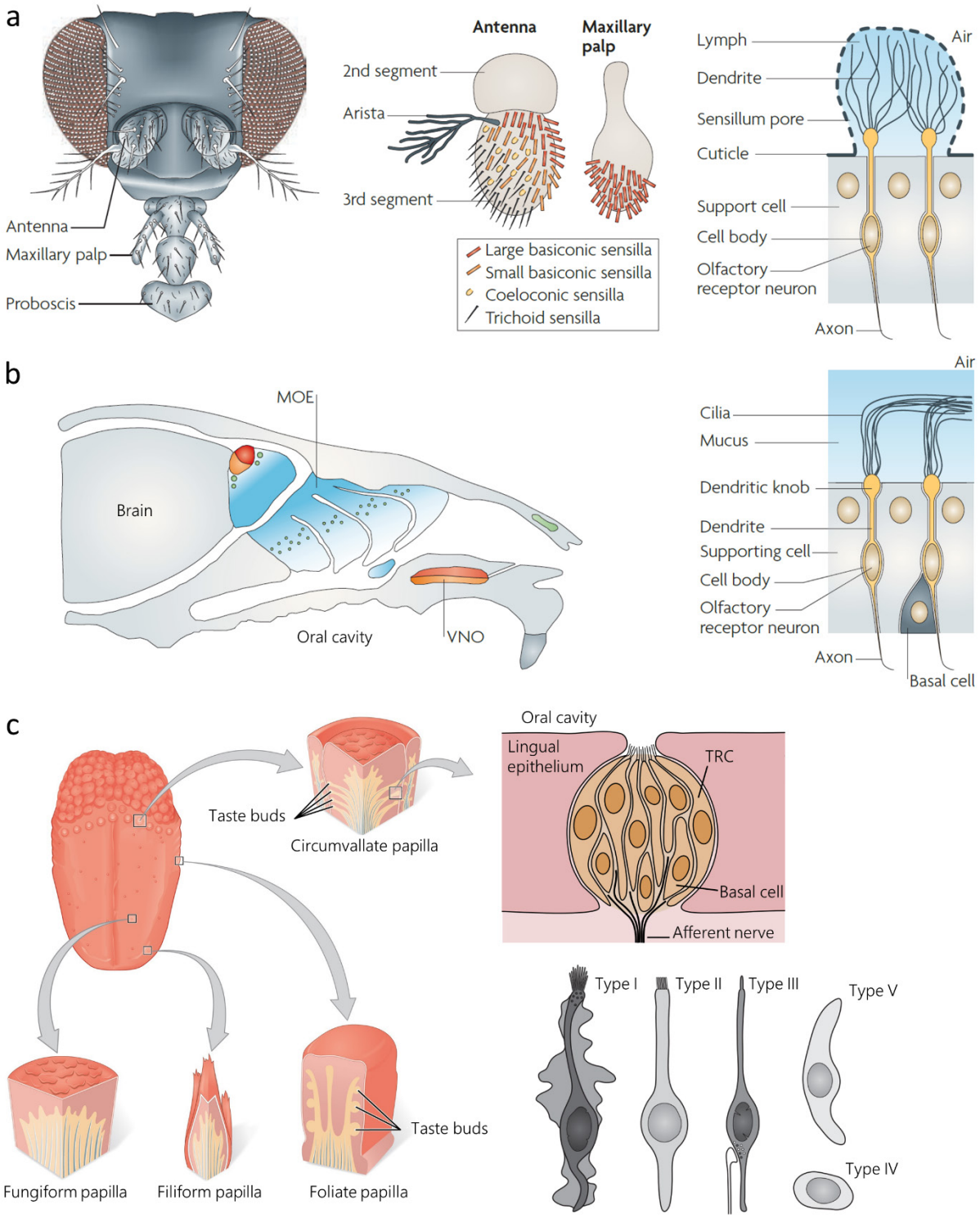


Figure 1: Anatomy of chemosensory perception. Structures responsible for smell perception in **a)** the fruit fly and **b)** mouse. MOE: main olfactory epithelium, VNO: vomeronasal organ. Adapted by permission from Springer Nature: Macmillan Publishers Ltd., Nat Rev Neurosci, "Olfactory signalling in vertebrates and insects: differences and commonalities", Kaupp, U. B., © (2010). **c)** Structures responsible for taste perception. TRC: taste receptor cell. Tongue and papillae adapted from OpenStax [CC BY 4.0], taste bud from NEUROtiker [GFDL, CC BY-SA 2.5], taste cells from Jonas Töle [CC0].

Chemosensory receptors are transmembrane receptors

Both chemical senses, smell and taste, rely on transmembrane receptors to detect odorant or tastants molecules in the extracellular domain. In vertebrates, these chemoreceptors belong to two families of transmembrane receptors: ion channels and G protein-coupled receptors (GPCRs) (Figure 2).

The first family, ion channels, participate in sour and salt taste perception. Several genes have previously been proposed to code for candidate sour receptors, including the polycystic kidney disease 2-like 1 (PKD2L1) receptor which has been demonstrated to be expressed in taste cells responsible for sour taste [20]. However, knockout of the PKD2L1 gene in mice had minor effects on sour perception, indicating only partial contributions from this receptor [21]. More recently, otopetrin 1 (OTOP1) was identified as a proton-selective ion channel that is highly expressed in taste cells [22], and was later confirmed by knockout experiments as the proper sour taste receptor [23, 24].

Regarding salty perception, an ion channel specific to sodium, the epithelial sodium channel (ENaC), has been identified as the main salt taste receptor of some vertebrates [25]. This ion channel, often called “amiloride-sensitive” due to amiloride being a known inhibitor, is responsible for the attractive behavior resulting from NaCl consumption at low concentrations. However, at higher concentrations of NaCl, another salt transduction pathway is used and leads to an aversive response, with the particularity of being insensitive to amiloride and less ion specific [26, 27]. The activation of this specific pathway could however depend on the chloride anion more than sodium cation [28], and the corresponding chloride receptor has yet to be discovered. Furthermore, the amiloride-insensitive response is mediated by type II or type III cells, depending on the type of papillae [28], while amiloride-sensitive response may rely on type I cells which were previously thought to only play a support role in taste buds [29].

Additionally, ion channels also play a role in olfaction as the functional odorant-sensing unit of insects is a ligand-gated ion channel [30]. Briefly, their olfactory receptor is a heteromer of unknown stoichiometry made of a highly conserved subunit named Orco, and an odorant-binding receptor [31]. Interestingly, these two subunits adopt a fold with seven transmembrane domains similar to G protein-coupled receptors (GPCRs), although with an inverted topology [32].

Alongside ion channels, another type of transmembrane receptors, the GPCR family, takes part in taste and odorant perception. This receptor family has a typical seven-helix fold, is

subdivided in six classes, and two of them are of interest for chemosensation. The class A family (rhodopsin-like) concerns the largest number of chemosensory receptors as it covers taste receptors type 2 (TAS2Rs) which are activated by bitter compounds, olfactory receptors (ORs), vomeronasal type-1 receptors (V1Rs), and trace amine-associated receptors (TAARs) [33, 34]. The class C family (glutamate), on the other hand, includes taste receptors type 1 (T1Rs) responsible for both sweet and umami perception, as well as vomeronasal type-2 receptors (V2Rs) [33, 35] and is structurally characterized by a very large extracellular domain that binds agonists.

Vertebrate olfaction rests upon ORs, TAARs, V1Rs and V2Rs. While ORs detect a variety of volatile molecules and their breadth of tuning ranges from narrow to broad, TAARs are specialized in binding biogenic amines and are never expressed in the same OSN as ORs [33], and vomeronasal receptors mainly bind pheromones (V1Rs) or peptides (V2Rs) [36]. Finally, the structural characteristics of insect ORs, T1Rs, and TAS2Rs are studied in more details in the corresponding chapters of this thesis.

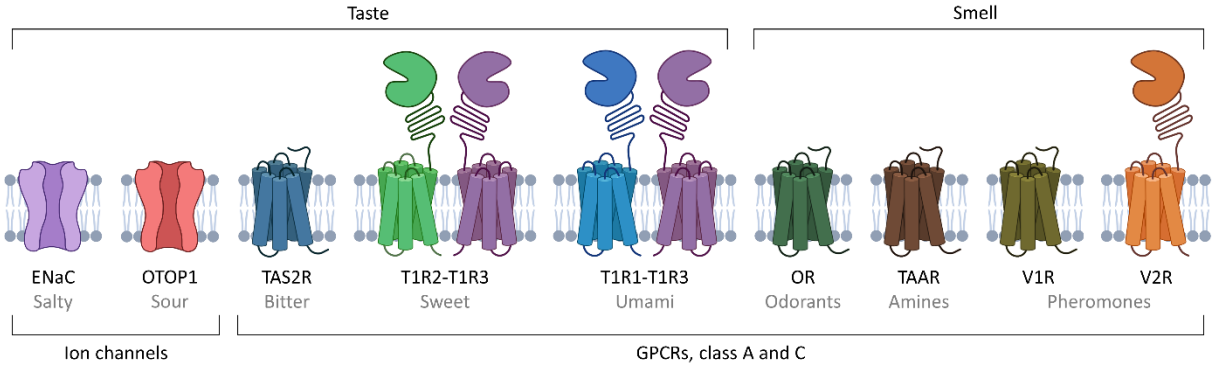


Figure 2: Schematic representation of chemosensory receptor structures in vertebrates, along with their function and family of transmembrane receptors.

Transduction of chemosensory stimuli

For the two types of transmembrane receptors, the signal transduction mechanism is different and can be either metabotropic for GPCRs, or ionotropic for ion channels. Although each chemosensory receptor may have specificities regarding the molecular structures involved in transduction, a general process is presented here for both classes, and further detailed in the next chapters.

In GPCRs, upon ligand binding, conformational changes occur in the receptor leading to its activation which promotes the binding of a G protein. The G protein is a heterotrimer made of a Gα subunit bound to guanosine diphosphate (GDP), and Gβ and Gγ subunits. When the G

protein binds to the GPCR, the GDP-binding site is destabilized which ultimately results in GDP dissociation, rapidly replaced by a guanosine triphosphate (GTP) taken from the cytosol. In turn, it promotes Gα conformational changes leading to the dissociation of the GTP-bound Gα from the Gβγ subunits [37]. Both subunits can then trigger a variety of signaling cascades involving different downstream effector proteins, leading to the release of neurotransmitters in the synaptic cleft made with an afferent neuron.

In ion channels, signal transduction is more straightforward as it does not typically involve secondary messengers. As the channel opens (triggered by ligand binding or other events), ions such as Ca^{2+}, Na^{+}, or K^{+} flow through the membrane, leading to a depolarization of the chemosensory cell thus stimulating an afferent neuron.

Perireceptor events that influence chemosensory perception

Several mechanical, biological, and biochemical events can modulate the perception of the environment through chemical senses. For example, sniffing is known to influence the intensity perceived by humans when smelling odorants [38], but similar mechanical optimizations occur in other species [8], such as moths using their wings to maximize the airflow passing through their sensilla [39]. Additionally, nasal anatomy may influence olfactory sensitivity as specific nasal features can contribute to create a vortex inside the nasal cavity which potentially maximizes residence time in the olfactory region [40].

Furthermore, changes in pH can greatly affect sweet taste perception, as shown by miraculin, a protein extracted from the red berry *Richadella dulcifica*. At neutral pH, miraculin inhibits the sweet taste of aspartame, cyclamate, and three sweet tasting proteins (neoculin, thaumatin and brazzein), but at acidic pH it becomes itself a highly potent sweetener [41].

In addition to salivary pH, salivary composition can influence flavor perception by interacting with food components. For instance, both mucins and α-amylase decrease the release of aroma compounds, and salivary enzymes are able to lower the concentrations of esters, thiols and aldehyde from mixtures of aroma components [42]. Salivary composition is also subject to inter-individual variability and differences in the proteome and metabolome of individuals can influence their sensitivity to food components like oleic acid [43].

Similarly to saliva, several proteins can contribute to olfactory perception before and after odorant binding. Odorant binding proteins (OBPs) are small and soluble proteins that participate in the transport of odorants through the aqueous nasal mucus or sensillary lymph towards ORs. They adopt the typical structure of lipocalins and interact with odorants mainly

through hydrophobic interactions, hence their role in facilitating the transit of lipophilic odorants [44]. Besides, xenobiotic metabolizing enzymes (XMEs) are involved in the degradation of potentially toxic compounds as well as odorants, and are highly expressed in the olfactory epithelium [44]. They participate in odorant clearance to maintain sensitivity but can also affect perception by transforming the initial odorant into another OR-binding metabolite. Overall, both OBPs and XMEs can influence the availability of odorants to ORs, and polymorphism on the corresponding genes could be in part responsible for the variability in odorant and aroma perception.

Polymorphism on chemosensory receptors

This last point raises a key part in chemosensory perception: how interindividual variability in taste and olfactory receptors affects perception and more. One of the most striking examples related to taste perception is the difference in phenylthiocarbamide (PTC) sensitivity related to haplotypes of the bitter taste receptor TAS2R38. Three positions can be subject to single-nucleotide polymorphisms (SNPs) on this receptor, A49P, V262A and I296V, and constitute two common haplotypes: AVI, the most common but recessive non-taster allele, and the taster PAV haplotype. The AVI/PAV heterozygotes are the most common in the population and can taste PTC, while AVI homozygotes are non-tasters and PAV ones are more sensitive to PTC (super-tasters) compared to the heterozygotes [45]. Similar effects are observed with another related compounds, 6-propyl-2-thiouracil (PROP), as both molecules contain a thioamide moiety. While PROP and PTC don't appear in food items, other thioamide or related compounds, namely goitrin and sinigrin, exist in several cruciferous vegetables such as Brussel sprouts, cabbage, and broccoli and were shown to be affected by the same polymorphism [46] (Figure 3). This could be part of the reasons that explain the avoidance of cruciferous vegetables by individuals, especially young children for whom PROP sensitivity was linked to a lower acceptance of raw broccoli [47].

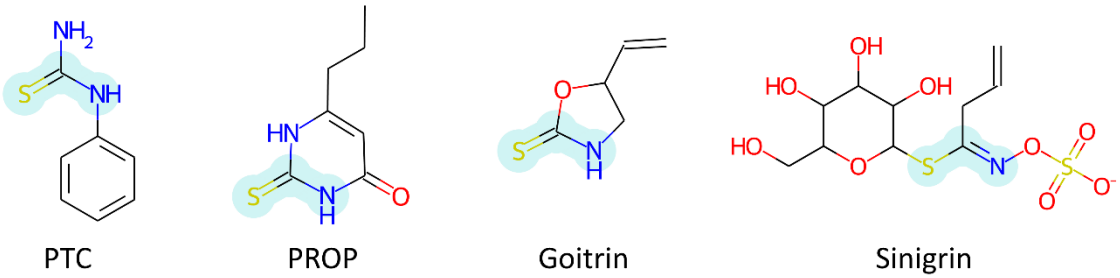


Figure 3: Structures of PTC, PROP, goitrin and sinigrin with the thioamide or equivalent moiety highlighted.

Odorant perception can also be greatly affected by genetic variations, as evidenced with androstenone, a pheromone in boars which is detected by the human olfactory receptor OR7D4. Its odor is described as unpleasant, urine- and sweat-like, or pleasant, vanilla- and sweet-like depending on individuals, and can taint the flavor of pork meat. Two SNPs exist for this receptor, R88W and T133M, leading to the most common and functional allele RT and the non-functional WM haplotype. The RT homozygotes are sensitive to androstenone odor, while the RT/WM and WM/WM genotypes are insensitive to this smell, and OR7D4 genotype influences the perception and liking of androstenone-tainted pork meat [48].

As shown in both examples, genotype shapes chemosensory perception, and because perception dictates food acceptance and dietary intake, genotype can also affect consumer habits. Hypersensitivity to repulsive odors and tastes due to genetic variations can thus lead to the rejection of food with positive health effects like fruits and vegetables containing phenols, triterpenes, and organosulfur compounds [49].

Relevance of chemosensory research

A better understanding of olfaction and taste at the molecular level can have direct applications in many industries. For example, perfumers and agri-food companies might be interested in the development of novel odorants or tastants with specific properties to comply with new safety regulations, fulfill consumer's expectations, or cover new market opportunities. The agricultural industry can also benefit from advances in the field as new odorant repellants for pest protection appear, and because some insects like mosquitoes also function as disease vectors, such repellants could also be of use in epidemiology. Scent also finds applications in marketing strategies through olfactory marketing, as it can help to build a brand's identity. It can also entice consumers to purchase products, as the sense of smell is often associated with an emotional response because of the close relationship between the olfactory cortex and the limbic system responsible for memory and emotional processing [50].

More surprisingly, chemosensory research has implications that go well beyond olfaction and taste. Genetic mutations on the bitter taste receptor TAS2R38 have been previously associated with decreased risks of obesity [51] and reduced odds of cigarette smoking [52], while an SNP on TAS2R16 was shown to decrease the risk of alcohol dependence [53]. Although the reasons for such associations are not known, clinical relevance of bitter taste receptors could be explained by their ectopic expression. TAS2Rs are indeed expressed in extra-oral tissues, but

their function in these diverse locations is not always understood. For instance, motile cilia found on epithelial cells of the human airway express bitter taste receptors, and their stimulation by bitter compounds increases ciliary beat frequency, potentially as a defense mechanism to propel noxious compounds out of the airway [54]. A haplotype of TAS2R9 has also been associated with altered glucose and insulin homeostasis [55], and associations of TAS2Rs with several other systems and diseases, including cardiac, vascular, testis, semen and cancerous cells and Parkinson disease, have been suggested [56]. Sweet and umami taste receptors also participate in extraoral functions, as showcased by the presence of T1R3 in the gastrointestinal tract to promote endocrine response through nutrient detection [57]. Extraoral roles of the sweet taste receptor has been more thoroughly reviewed by Laffitte *et al.* [58]. Similar diversified roles of ectopic olfactory receptors, notably in heart, lung, sperm, skin and cancerous tissues, have also been shown [59].

Computational strategies applied to chemical senses

To date, experimental structural information about chemosensory receptors is extremely scarce. For olfaction, some light has been shed on the olfactory co-receptor (Orco) of *Apocrypta bakeri* [60] thanks to cryo-electron microscopy (cryo-EM). Regarding taste, only ion channels have been fully resolved by cryo-EM recently, with structures for the zebrafish candidate sour receptor Otop1 [61] and the human salty taste receptor ENaC [62]. Additionally, part of the structure of the extracellular domain of the medaka fish's sweet taste receptor has been obtained by crystallography [63]. Thus, the transmembrane domains of all G protein-coupled chemosensory receptors, from taste to olfaction, have yet to be determined experimentally. Consequently, computational strategies have been previously used to predict the structure [35, 64] and study the dynamics [65, 66] of these receptors.

Concurrently, both structure-based [67, 68] and ligand-based [69–71] approaches have been applied to search for new active ligands, thereby accelerating the deorphanization of some chemosensory receptors or widening the explored chemical space.

In this way, this thesis was focused on getting a better picture of the molecular determinants of chemosensory perception through a computational lens. To reach this goal, two practical and fundamental objectives were set: i/ identifying small molecules that can modulate the activity of olfactory or taste receptors, and ii/ developing modeling protocols that can help us achieve a better understanding of the molecular processes occurring during taste perception. In order to

tackle these objectives, different computational methods were used. Quantitative structure-activity relationship (QSAR) models combine machine-learning with molecular features extracted from the structure of compounds to identify active substances, and such approach was applied to the rational discovery of novel odorants and tastants. For the second objective, I focused on generating 3D models of bitter taste receptors using homology modeling to identify molecular switches that play a role in ligand binding and signal transduction.

My contributions to chemosensory research are gathered in this thesis in three chapters, one for each olfactory and gustatory modality, where I assembled the scientific publications that I authored. In the first two chapters, machine-learning algorithms were implemented to identify candidate odorant or sapid molecules which were validated through *in vitro* or *in vivo* experiments performed by collaborators. The first chapter is focused on *Spodoptera littoralis*, a pest to crop plants, and our efforts in finding natural odorants that can disrupt the insect's behavior for applications in biocontrol strategies. The second chapter revolves around the search for natural intense sweeteners and the development of a web-based predictive platform. While QSAR models can successfully guide screening campaigns related to chemosensory problems, data is often scarce or poorly labelled which is far from ideal with machine-learning methods. Structure-based approaches thus appear as credible alternatives despite the other challenges they raise. In this direction, the last chapter is dedicated to the development of a molecular modeling protocol for reconstructing bitter taste receptors which recapitulate experimental mutagenesis data, followed by the identification of molecular switches that participate in ligand sensing and signal transduction.

In parallel to these main research questions, I was involved in side-projects, one of which could potentially be applied to decipher the allosteric activation mechanism of TAS2Rs. I developed a Python library, named ProLIF, which can extract interaction fingerprints from complexes (in MD trajectories, docking poses, or crystal structures) that combine ligand, protein, DNA or RNA molecules. The library was showcased on class A GPCRs to highlight key interactions that participate in ligand and G protein binding, as well as differences in inter-helical interactions between active and inactive structures. Such analysis could be of use to gain knowledge on structure-function relationships in bitter taste receptors.

Altogether, this thesis illustrates the implementation of computational strategies to gain knowledge on chemosensory perception at the molecular level.

References

1. Prasad BC, Reed RR (1999) Chemosensation: molecular mechanisms in worms and mammals. Trends in Genetics 15:150–153. https://doi.org/10.1016/S0168-9525(99)01695-9

2. Trivedi BP (2012) Gustatory system: The finer points of taste. Nature 486:S2–S3. https://doi.org/10.1038/486S2a

3. Roper SD (2014) TRPs in Taste and Chemesthesis. In: Nilius B, Flockerzi V (eds) Mammalian Transient Receptor Potential (TRP) Cation Channels. Springer International Publishing, Cham, pp 827–871

4. DiPatrizio NV (2014) Is fat taste ready for primetime? Physiology & Behavior 136:145–154. https://doi.org/10.1016/j.physbeh.2014.03.002

5. Herrada G, Dulac C (1997) A Novel Family of Putative Pheromone Receptors in Mammals with a Topographically Organized and Sexually Dimorphic Distribution. Cell 90:763–773. https://doi.org/10.1016/S0092-8674(00)80536-X

6. Baxi KN, Dorries KM, Eisthen HL (2006) Is the vomeronasal system really specialized for detecting pheromones? Trends in Neurosciences 29:1–7. https://doi.org/10.1016/j.tins.2005.10.002

7. Vosshall LB, Stocker RF (2007) Molecular Architecture of Smell and Taste in Drosophila. Annu Rev Neurosci 30:505–533. https://doi.org/10.1146/annurev.neuro.30.051606.094306

8. Ache BW, Young JM (2005) Olfaction: Diverse Species, Conserved Principles. Neuron 48:417–430. https://doi.org/10.1016/j.neuron.2005.10.022

9. Malnic B, Hirono J, Sato T, Buck LB (1999) Combinatorial Receptor Codes for Odors. Cell 96:713–723. https://doi.org/10.1016/S0092-8674(00)80581-4

10. Malnic B, Godfrey PA, Buck LB (2004) The human olfactory receptor gene family. Proceedings of the National Academy of Sciences 101:2584–2589. https://doi.org/10.1073/pnas.0307882100

11. Bushdid C, Magnasco MO, Vosshall LB, Keller A (2014) Humans Can Discriminate More than 1 Trillion Olfactory Stimuli. Science 343:1370–1372. https://doi.org/10.1126/science.1249168

12. Gerkin RC, Castro JB (2015) The number of olfactory stimuli that humans can discriminate is still unknown. eLife 4:e08127. https://doi.org/10.7554/eLife.08127

13. Meister M (2015) On the dimensionality of odor space. eLife 4:e07865. https://doi.org/10.7554/eLife.07865

14. Ganchrow D, Ganchrow JR (1985) Number and distribution of taste buds in the oral cavity of hatchling chicks. Physiology & Behavior 34:889–894. https://doi.org/10.1016/0031-9384(85)90009-5

15. Witt M (2019) Anatomy and development of the human taste system. In: Handbook of Clinical Neurology. Elsevier, pp 147–171

16. Tomchik SM, Berg S, Kim JW, et al (2007) Breadth of Tuning and Taste Coding in Mammalian Taste Buds. Journal of Neuroscience 27:10840–10848. https://doi.org/10.1523/JNEUROSCI.1863-07.2007

17. Chandrashekar J, Hoon MA, Ryba NJP, Zuker CS (2006) The receptors and cells for mammalian taste. Nature 444:288–294. https://doi.org/10.1038/nature05401

18. Simon SA, de Araujo IE, Gutierrez R, Nicolelis MAL (2006) The neural mechanisms of gustation: a distributed processing code. Nat Rev Neurosci 7:890–901. https://doi.org/10.1038/nrn2006

19. Ohla K, Yoshida R, Roper SD, et al (2019) Recognizing Taste: Coding Patterns Along the Neural Axis in Mammals. Chemical Senses 44:237–247. https://doi.org/10.1093/chemse/bjz013

20. Huang AL, Chen X, Hoon MA, et al (2006) The cells and logic for mammalian sour taste detection. Nature 442:934–938. https://doi.org/10.1038/nature05084

21. Horio N, Yoshida R, Yasumatsu K, et al (2011) Sour Taste Responses in Mice Lacking PKD Channels. PLoS ONE 6:e20007. https://doi.org/10.1371/journal.pone.0020007

22. Tu Y-H, Cooper AJ, Teng B, et al (2018) An evolutionarily conserved gene family encodes proton-selective ion channels. Science 359:1047–1050. https://doi.org/10.1126/science.aao3264

23. Zhang J, Jin H, Zhang W, et al (2019) Sour Sensing from the Tongue to the Brain. Cell 179:392-402.e15. https://doi.org/10.1016/j.cell.2019.08.031

24. Teng B, Wilson CE, Tu Y-H, et al (2019) Cellular and Neural Responses to Sour Stimuli Require the Proton Channel Otop1. Current Biology 29:3647-3656.e5. https://doi.org/10.1016/j.cub.2019.08.077

25. Chandrashekar J, Kuhn C, Oka Y, et al (2010) The cells and peripheral representation of sodium taste in mice. Nature 464:297–301. https://doi.org/10.1038/nature08783

26. Lewandowski BC, Sukumaran SK, Margolskee RF, Bachmanov AA (2016) Amiloride-Insensitive Salt Taste Is Mediated by Two Populations of Type III Taste Cells with Distinct Transduction Mechanisms. J Neurosci 36:1942–1953. https://doi.org/10.1523/JNEUROSCI.2947-15.2016

27. Lindemann B (2001) Receptors and transduction in taste. Nature 413:219–225. https://doi.org/10.1038/35093032

28. Roebber JK, Roper SD, Chaudhari N (2019) The Role of the Anion in Salt (NaCl) Detection by Mouse Taste Buds. J Neurosci 39:6224–6232. https://doi.org/10.1523/JNEUROSCI.2367-18.2019

29. Vandenbeuch A, Clapp TR, Kinnamon SC (2008) Amiloride-sensitive channels in type I fungiform taste cells in mouse. BMC Neurosci 9:1. https://doi.org/10.1186/1471-2202-9-1

30. Nakagawa T, Vosshall LB (2009) Controversy and consensus: noncanonical signaling mechanisms in the insect olfactory system. Current Opinion in Neurobiology 19:284–292. https://doi.org/10.1016/j.conb.2009.07.015

31. Carraher C, Dalziel J, Jordan MD, et al (2015) Towards an understanding of the structural basis for insect olfaction by odorant receptors. Insect Biochemistry and Molecular Biology 66:31–41. https://doi.org/10.1016/j.ibmb.2015.09.010

32. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical Membrane Topology and Heteromeric Function of Drosophila Odorant Receptors In Vivo. PLoS Biol 4:e20. https://doi.org/10.1371/journal.pbio.0040020

33. Spehr M, Munger SD (2009) Olfactory receptors: G protein-coupled receptors and beyond. Journal of Neurochemistry 109:1570–1583. https://doi.org/10.1111/j.1471-4159.2009.06085.x

34. Di Pizio A, Levit A, Slutzki M, et al (2016) Comparing Class A GPCRs to bitter taste receptors. In: Methods in Cell Biology. Elsevier Ltd, pp 401–427

35. Chéron J-B, Golebiowski J, Antonczak S, Fiorucci S (2017) The anatomy of mammalian sweet taste receptors: Modeling Sweet Taste Receptors. Proteins 85:332–341. https://doi.org/10.1002/prot.25228

36. Kaupp UB (2010) Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci 11:188–200. https://doi.org/10.1038/nrn2789

37. Hilger D, Masureel M, Kobilka BK (2018) Structure and dynamics of GPCR signaling complexes. Nat Struct Mol Biol 25:4–12. https://doi.org/10.1038/s41594-017-0011-7

38. Laing DG (1985) Optimum perception of odor intensity by humans. Physiology & Behavior 34:569–574. https://doi.org/10.1016/0031-9384(85)90050-2

39. Loudon C, Koehl MA (2000) Sniffing by a silkworm moth: wing fanning enhances air penetration through and pheromone interception by antennae. Journal of Experimental Biology 203:2977–2990. https://doi.org/10.1242/jeb.203.19.2977

40. Li C, Jiang J, Kim K, et al (2018) Nasal Structural and Aerodynamic Features That May Benefit Normal Olfactory Sensitivity. Chemical Senses 43:229–237. https://doi.org/10.1093/chemse/bjy013

41. Koizumi A, Tsuchiya A, Nakajima K -i., et al (2011) Human sweet taste receptor mediates acid-induced sweetness of miraculin. Proceedings of the National Academy of Sciences 108:16819–16824. https://doi.org/10.1073/pnas.1016644108

42. Mosca AC, Chen J (2017) Food-saliva interactions: Mechanisms and implications. Trends in Food Science & Technology 66:125–134. https://doi.org/10.1016/j.tifs.2017.06.005

43. Mounayar R, Morzel M, Brignot H, et al (2014) Salivary markers of taste sensitivity to oleic acid: a combined proteomics and metabolomics approach. Metabolomics 10:688–696. https://doi.org/10.1007/s11306-013-0602-1

44. Heydel J-M, Coelho A, Thiebaud N, et al (2013) Odorant-Binding Proteins and Xenobiotic Metabolizing Enzymes: Implications in Olfactory Perireceptor Events: Odorant-Binding Proteins and Metabolizing Enzymes. Anat Rec 296:1333–1345. https://doi.org/10.1002/ar.22735

45. Kim U, Jorgenson E, Coon H, et al (2003) Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. Science 299:1221–1225. https://doi.org/10.1126/science.1080190

46. Behrens M, Gunn HC, Ramos PCM, et al (2013) Genetic, Functional, and Phenotypic Diversity in TAS2R38-Mediated Bitter Taste Perception. Chemical Senses 38:475–484. https://doi.org/10.1093/chemse/bjt016

47. Keller KL, Steinmann L, Nurse RJ, Tepper BJ (2002) Genetic taste sensitivity to 6-n-propylthiouracil influences food preference and reported intake in preschool children. Appetite 38:3–12. https://doi.org/10.1006/appe.2001.0441

48. Lunde K, Egelandsdal B, Skuterud E, et al (2012) Genetic Variation of an Odorant Receptor OR7D4 and Sensory Perception of Cooked Meat Containing Androstenone. PLoS ONE 7:e35259. https://doi.org/10.1371/journal.pone.0035259

49. Reed DR, Tanaka T, McDaniel AH (2006) Diverse tastes: Genetics of sweet and bitter perception. Physiology & Behavior 88:215–226. https://doi.org/10.1016/j.physbeh.2006.05.033

50. RajMohan V, Mohandas E (2007) The limbic system. Indian J Psychiatry 49:132. https://doi.org/10.4103/0019-5545.33264

51. Ortega FJ, Agüera Z, Sabater M, et al (2016) Genetic variations of the bitter taste receptor TAS2R38 are associated with obesity and impact on single immune traits. Mol Nutr Food Res 60:1673–1683. https://doi.org/10.1002/mnfr.201500804

52. Cannon DS, Baker TB, Piper ME, et al (2005) Associations between phenylthiocarbamide gene polymorphisms and cigarette smoking. Nicotine Tob Res 7:853–858. https://doi.org/10.1080/14622200500330209

53. Hinrichs AL, Wang JC, Bufe B, et al (2006) Functional Variant in a Bitter-Taste Receptor (hTAS2R16) Influences Risk of Alcohol Dependence. The American Journal of Human Genetics 78:103–111. https://doi.org/10.1086/499253

54. Shah AS, Ben-Shahar Y, Moninger TO, et al (2009) Motile Cilia of Human Airway Epithelia Are Chemosensory. Science 325:1131–1134. https://doi.org/10.1126/science.1173869

55. Dotson CD, Zhang L, Xu H, et al (2008) Bitter Taste Receptors Influence Glucose Homeostasis. PLoS ONE 3:e3974. https://doi.org/10.1371/journal.pone.0003974

56. Jeruzal-Świątecka J, Fendler W, Pietruszewska W (2020) Clinical Role of Extraoral Bitter Taste Receptors. IJMS 21:5156. https://doi.org/10.3390/ijms21145156

57. Margolskee RF, Dyer J, Kokrashvili Z, et al (2007) T1R3 and gustducin in gut sense sugars to regulate expression of Na+-glucose cotransporter 1. Proceedings of the National Academy of Sciences 104:15075–15080. https://doi.org/10.1073/pnas.0706678104

58. Laffitte A, Neiers F, Briand L (2014) Functional roles of the sweet taste receptor in oral and extraoral tissues: Current Opinion in Clinical Nutrition and Metabolic Care 17:379–385. https://doi.org/10.1097/MCO.0000000000000058

59. Chen Z, Zhao H, Fu N, Chen L (2018) The diversified function and potential therapy of ectopic olfactory receptors in non-olfactory tissues. J Cell Physiol 233:2104–2115. https://doi.org/10.1002/jcp.25929

60. Butterwick JA, del Mármol J, Kim KH, et al (2018) Cryo-EM structure of the insect olfactory receptor Orco. Nature 560:447–452. https://doi.org/10.1038/s41586-018-0420-8

61. Saotome K, Teng B, Tsui CC (Alex), et al (2019) Structures of the otopetrin proton channels Otop1 and Otop3. Nat Struct Mol Biol 26:518–525. https://doi.org/10.1038/s41594-019-0235-9

62. Noreng S, Bharadwaj A, Posert R, et al (2018) Structure of the human epithelial sodium channel by cryo-electron microscopy. eLife 7:e39340. https://doi.org/10.7554/eLife.39340

63. Nuemket N, Yasui N, Kusakabe Y, et al (2017) Structural basis for perception of diverse chemical substances by T1r taste receptors. Nat Commun 8:15530. https://doi.org/10.1038/ncomms15530

64. de March CA, Kim S-K, Antonczak S, et al (2015) G protein-coupled odorant receptors: From sequence to structure: Odorant Receptors Sequence. Protein Science 24:1543–1548. https://doi.org/10.1002/pro.2717

65. Chéron J-B, Soohoo A, Wang Y, et al (2019) Conserved Residues Control the T1R3-Specific Allosteric Signaling Pathway of the Mammalian Sweet-Taste Receptor. Chemical Senses 44:303–310. https://doi.org/10.1093/chemse/bjz015

66. de March CA, Topin J, Bruguera E, et al (2018) Odorant Receptor 7D4 Activation Dynamics. Angew Chem 130:4644–4648. https://doi.org/10.1002/ange.201713065

67. Topin J, de March CA, Charlier L, et al (2014) Discrimination between Olfactory Receptor Agonists and Non-agonists. Chem Eur J 20:10227–10230. https://doi.org/10.1002/chem.201402486

68. Spaggiari G, Di Pizio A, Cozzini P (2020) Sweet, umami and bitter taste receptors: State of the art of in silico molecular modeling approaches. Trends in Food Science & Technology 96:21–29. https://doi.org/10.1016/j.tifs.2019.12.002

69. Bushdid C, de March CA, Fiorucci S, et al (2018) Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features. J Phys Chem Lett 9:2235–2240. https://doi.org/10.1021/acs.jpclett.8b00633

70. Chéron J-B, Casciuc I, Golebiowski J, et al (2017) Sweetness prediction of natural compounds. Food Chemistry 221:1421–1425. https://doi.org/10.1016/j.foodchem.2016.10.145

71. Dagan-Wiener A, Nissim I, Ben Abu N, et al (2017) Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Sci Rep 7:12074. https://doi.org/10.1038/s41598-017-12359-7

Chapter I

Reverse chemical ecology targeting ORs applied to pest control

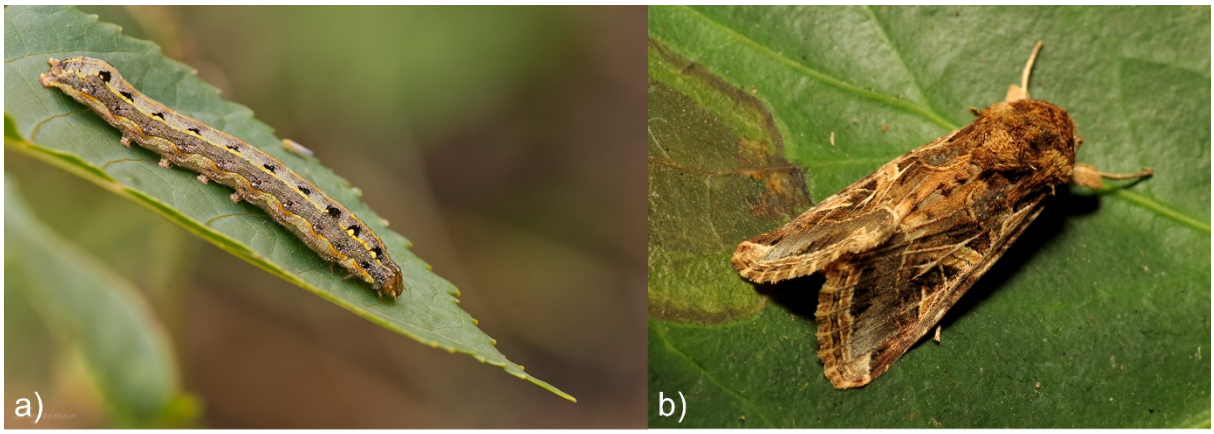


Figure 1: *Spodoptera littoralis* at different stages of its life. **a)** Caterpillar [© David Marquina Reyes, CC BY-NC-ND 2.0] **b)** Adult moth [© Katja Schulz, CC BY 2.0]

The cotton leafworm *Spodoptera littoralis* (Figure 1) is a polyphagous insect labelled as a quarantine pest by the European and Mediterranean Plant Protection Organization (EPPO) because of its potential economic impact [1]. Native of Africa, the noctuid moth is also found widely in Mediterranean Europe and parts of the Middle East [2, 3]. The widespread presence of this pest can be explained by its broad host range, with around 80 known host plants and crops [3, 4]. For these reasons, it is considered to be one of the most destructive pests among the *Lepidoptera* order [2].

Many of the damaging behaviors caused by insect pests, including *Spodoptera littoralis*, are closely related to olfaction as odorants convey information that take part in critical aspects of their lives such as reproduction, food, and oviposition [5]. This makes the olfactory system a promising target for biocontrol strategies using semiochemicals i.e., attractants or repellents, to better regulate pest behavior.

In insects, odorant stimuli are perceived by olfactory receptors (ORs) expressed at the membrane of olfactory sensory neurons (OSNs). These OSNs are found in sensilla (sensory hairs filled with lymph) located on the antenna and maxillary palp, and project directly to an olfactory glomerulus in the antennal lobe [6]. Monitoring the response of an insect to odorants is possible via electroantennography (EAG) and allows for OR deorphanization. In practice, a mutant *Drosophila* OSN that does not respond to odors, called the “empty neuron”, is used to generate constructs expressing any transgenic OR of interest [7]. Single sensillum recording is then used to monitor the electrophysiological response of a sensillum exposed to odorants, characterizing the effect of each ligand on the studied OR expressed in the OSNs.

By targeting ORs for pest control, the chances of disturbing other animal species are lower as insect and vertebrate ORs are known to be fundamentally different. Indeed, in insects the functional odorant-sensing unit is an heteromeric complex made of an Orco subunit (OR coreceptor, formerly Or83b) and one or more variable odorant-binding subunits (OR_x) as first discovered in *Drosophila Melanogaster* [8], while in vertebrates only the OR, a class A GPCR, is needed for detecting odorants [9]. This Orco subunit is highly conserved among insects and has homologs in distant insect species [10], contrasting with the high level of variability in OR_x within and across insect lineages likely related to the ecological niche of each specie [11]. Compared to mammalian ORs, insect ORs share the seven transmembrane domains structural arrangement of GPCRs, but they are characterized by an inverted membrane topology (Figure 2a), with an intracellular N- and extracellular C-termini [12].

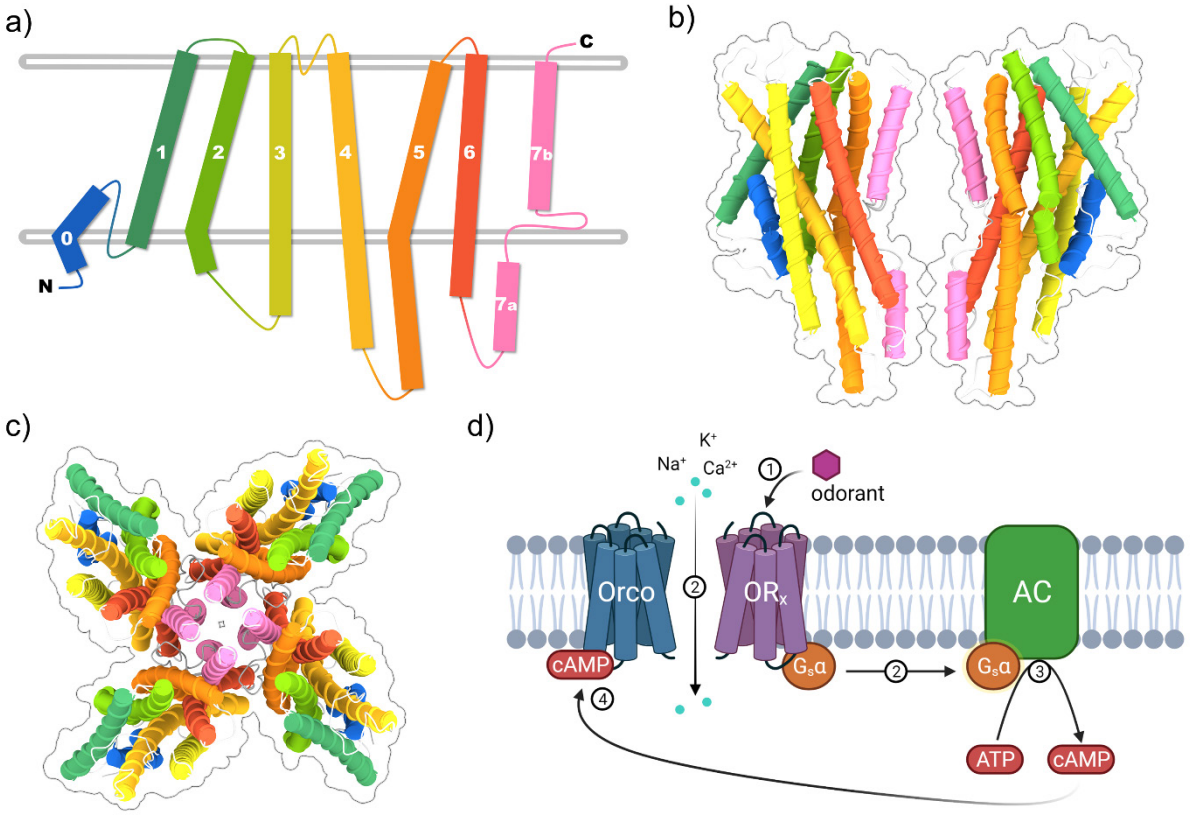


Figure 2: Structure and function of insect ORs. **a)** Orco topology and secondary structure. The structure of insect ORs is likely similar **b)** *Apocrypta bakeri* Orco tetramer viewed from the side (only 2 opposing subunits are shown for clarity, PDB id 6C70). **c)** Orco homotetramer viewed from the top. **d)** Signaling pathway of insect ORs. Upon odorant binding to OR_x, the complex becomes permeable to Na^+, K^+, and Ca^{2+} ions causing a short depolarization. Simultaneously, a G protein binds to the active OR_x, exchanges GDP for GTP with its α subunit ($G_s\alpha$) and then dissociates as an activated $G_s\alpha$ and a Gβγ dimer. The $G_s\alpha$ then binds to adenylyl cyclase (AC) which catalyzes the conversion of ATP to cAMP. Finally, cAMP binds to Orco and increases the activity of the ion channel.

Regarding signal transduction, the OR_x/Orco complex acts directly as a ligand-gated ion-channel, although the corresponding ion-conducting pore remains to be clearly identified as either formed by Orco only or the OR_x/Orco interface [13]. The recent cryo-EM elucidation of a wasp's Orco homotetramer (Figure 2b,c) illustrates that the TM segments that line the pore and subunit interfaces (S5 to S7, Figure 2a-c) tend to be more conserved than the other segments in both Orco and OR_x of various insects [14], suggesting a role in the stabilization of subunit interactions. It is thus possible that the functional OR_x/Orco complex structure could resemble that of the homotetramer (Figure 2c), where one or more Orco would be replaced by one or more OR_x to form an heterotetramer. While the characterization of insect ORs as odorant-gated ion channels is not disputed, other intracellular signaling cascades are known to affect their activity, giving them an unexpected metabotropic flavor. In fact, G proteins are known to be involved in the insect olfactory response [15] and odorant perception is altered by mutations affecting the cAMP signaling pathway [16]. The current consensus to explain this concurrent signaling pathway is that insect ORs are metabotropically-regulated ionotropic receptors: the immediate and short response from the odorant-gated ion channel activation is followed by a regulation of the ionotropic response by a slower G-protein-mediated pathway which sensitizes the receptor [11, 15, 17, 18] (Figure 2d). This confirms that insect ORs are distinct from vertebrates chemosensory GPCRs and as such suggests that odorants impacting insect behavior are less likely to simultaneously affect mammals, birds, fish, and amphibians, making the use of odorants a viable and valuable option for pest management.

In this chapter, the main goal was to develop a more environmentally friendly pest control strategy than insecticides, by identifying bio-olfacticides i.e., natural volatile molecules able to disrupt pest behavior through its olfactive functions. To achieve this, the moth *Spodoptera littoralis* was chosen as a model organism thanks to the recent deorphanization of part of its OR repertoire [19] (Figure 3). Two of the receptors, namely SlitOR24 and SlitOR25, were targeted as they are known to be expressed at the larval stage and to partake in caterpillar attraction when activated [20]. Because of the lack of structural data on the ORs (apart for Orco), we relied on a ligand-based *in silico* protocol to identify the new semiochemicals: this reverse chemical ecology approach relies on the link between OR activity and insect behavior to rationally design active ligands that can interfere with pest actions.

In the first publication presented in this chapter, a proof-of-concept machine learning model was developed to target SlitOR25 and used to virtually screen a large database of commercially available compounds.

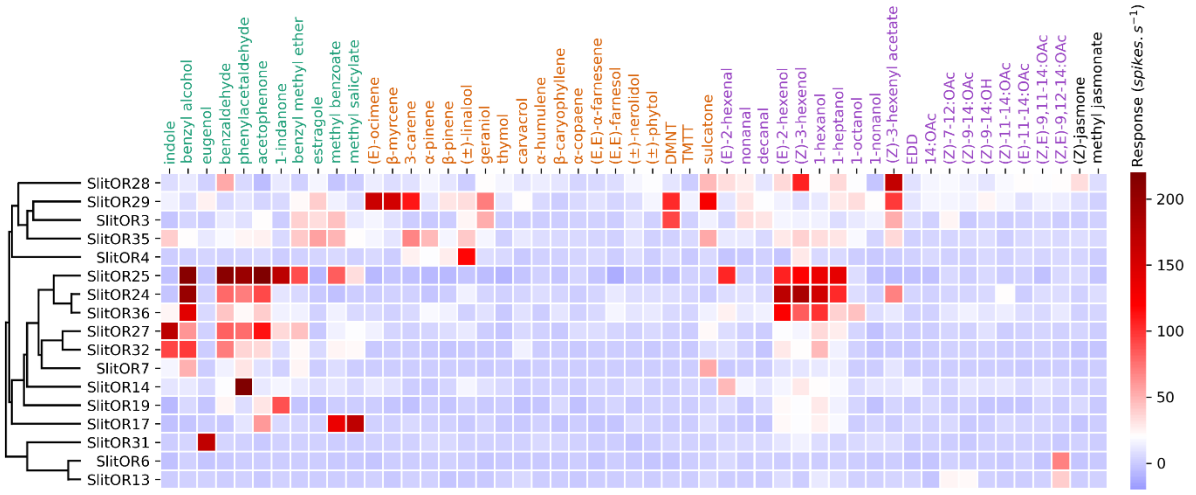


Figure 3: EAG screening of 17 *Spodoptera Littoralis* ORs (SlitORs) with 51 odorants at high dosage. Readapted from de Fouchier et al. [19]: SlitORs are classified based on a cluster analysis of response spectra and odorants are classified depending on their moieties (green: aromatic compounds, orange: terpenes, purple: aliphatics, black: unclassified).

Then, 32 candidate ligands were experimentally tested, both *in vitro* and *in vivo*, and revealed 9 novel agonists for the receptor. Building upon these encouraging results, new QSAR models were setup for both receptors in a second publication, while SlitOR25 benefited from a feedback loop as the initial dataset was augmented with the newly discovered active and inactive compounds. This time, the virtual screening was performed on an in-house focused library of natural volatile molecules to bias the search towards putative odorant molecules while considering food safety and environment protection concerns in the context of crop protection. New natural odorants were validated *in vivo* for both receptors and elicited an attractive response from the caterpillars in behavioral assays. This chapter thus provides an example of successful rational design of natural semiochemicals pertaining to pest control, driven by an *in silico* ligand-based approach.

Contributions

Publication 1

I analyzed the odorant chemical space of *S. littoralis* in comparison with *Drosophila Melanogaster*, trained the final QSAR model, and virtually screened a commercial database to propose compounds for experimental validation. Hubert Grunig performed preliminary

modeling experiments. Our collaborators performed the *in vitro* and *in vivo* assays. Gabriela Caballero-Vidal and I contributed equally as first authors.

Publication 2

I performed the modeling experiments (dataset preparation, machine-learning, applicability domain definition) and virtually screened the in-house database of natural compounds to identify odorant candidates. Our collaborators performed the *in vivo* and behavioral assays on those molecules. Gabriela Caballero-Vidal and I contributed equally as first authors.

References

1. Spodoptera littoralis (SPODLI)[Datasheet]| EPPO Global Database. https://gd.eppo.int/taxon/SPODLI/datasheet. Accessed 10 Jun 2021
2. Lopez-Vaamonde C (2010) Spodoptera littoralis (Boisduval, 1833) - African cotton leaf worm (Lepidoptera, Noctuidae). Chapter 14: Factsheets for 80 representative alien species. In: Alien terrestrial arthropods of Europe, Pensoft Publishers
3. Cocquempot C, Ramel J-M (2008) La noctuelle africaine du coton en voie de sédentarisation en France ? PHM Revue Horticole 506:33–36
4. Salama HS, Dimetry NZ, Salem SA (1971) On the Host Preference and Biology of the Cotton Leaf Worm Spodoptera littoralis Bois. Zeitschrift für Angewandte Entomologie 67:261–266. https://doi.org/10.1111/j.1439-0418.1971.tb02122.x
5. Dethier VG (1947) Chemical insect attractants and repellents, Blakiston
6. Vosshall LB, Stocker RF (2007) Molecular Architecture of Smell and Taste in Drosophila. Annu Rev Neurosci 30:505–533. https://doi.org/10.1146/annurev.neuro.30.051606.094306
7. Montagné N, Wanner K, Jacquin-Joly E (2021) Olfactory genomics within the Lepidoptera. In: Insect Pheromone Biochemistry and Molecular Biology. Elsevier, pp 469–505
8. Neuhaus EM, Gisselmann G, Zhang W, et al (2005) Odorant receptor heterodimerization in the olfactory system of Drosophila melanogaster. Nat Neurosci 8:15–17. https://doi.org/10.1038/nn1371
9. Kato A, Touhara K (2009) Mammalian olfactory receptors: pharmacology, G protein coupling and desensitization. Cell Mol Life Sci 66:3743–3753. https://doi.org/10.1007/s00018-009-0111-6
10. Larsson MC, Domingos AI, Jones WD, et al (2004) Or83b Encodes a Broadly Expressed Odorant Receptor Essential for Drosophila Olfaction. Neuron 43:703–714. https://doi.org/10.1016/j.neuron.2004.08.019
11. Wicher D, Miazzi F (2021) Functional properties of insect olfactory receptors: ionotropic receptors and odorant receptors. Cell Tissue Res 383:7–19. https://doi.org/10.1007/s00441-020-03363-x
12. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical Membrane Topology and Heteromeric Function of Drosophila Odorant Receptors In Vivo. PLoS Biol 4:e20. https://doi.org/10.1371/journal.pbio.0040020
13. Nakagawa T, Vosshall LB (2009) Controversy and consensus: noncanonical signaling mechanisms in the insect olfactory system. Current Opinion in Neurobiology 19:284–292. https://doi.org/10.1016/j.conb.2009.07.015

14. Butterwick JA, del Mármol J, Kim KH, et al (2018) Cryo-EM structure of the insect olfactory receptor Orco. Nature 560:447–452. https://doi.org/10.1038/s41586-018-0420-8
15. Deng Y, Zhang W, Farhat K, et al (2011) The Stimulatory Gαs Protein Is Involved in Olfactory Signal Transduction in Drosophila. PLoS ONE 6:e18605. https://doi.org/10.1371/journal.pone.0018605
16. Martín F, Charro M, Alcorta E (2001) Mutations affecting the cAMP transduction pathway modify olfaction in Drosophila. Journal of Comparative Physiology A: Sensory, Neural, and Behavioral Physiology 187:359–370. https://doi.org/10.1007/s003590100208
17. Carraher C, Dalziel J, Jordan MD, et al (2015) Towards an understanding of the structural basis for insect olfaction by odorant receptors. Insect Biochemistry and Molecular Biology 66:31–41. https://doi.org/10.1016/j.ibmb.2015.09.010
18. Wicher D (2013) Sensory receptors—design principles revisited. Front Cell Neurosci 7:. https://doi.org/10.3389/fncel.2013.00001
19. de Fouchier A, Walker WB, Montagné N, et al (2017) Functional evolution of Lepidoptera olfactory receptors revealed by deorphanization of a moth repertoire. Nat Commun 8:15709. https://doi.org/10.1038/ncomms15709
20. de Fouchier A, Sun X, Caballero-Vidal G, et al (2018) Behavioral Effect of Plant Volatiles Binding to Spodoptera littoralis Larval Odorant Receptors. Front Behav Neurosci 12:264. https://doi.org/10.3389/fnbeh.2018.00264

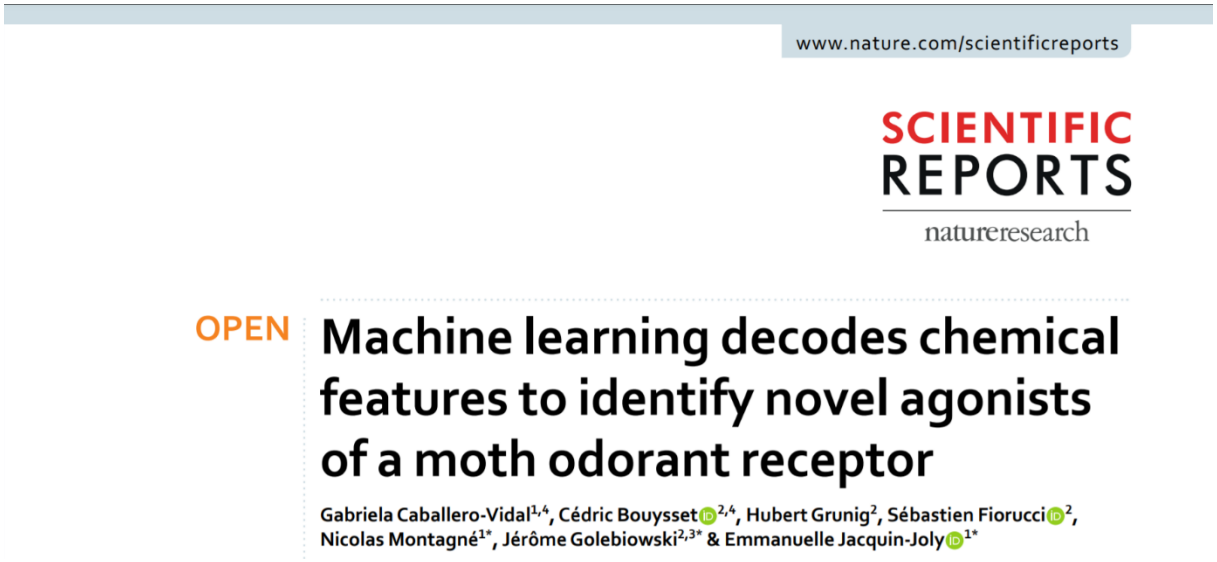
Publication 1

Machine learning decodes chemical features to identify novel agonists of a moth odorant receptor

Gabriela Caballero-Vidal†, Cédric Bouysset†, Hubert Grunig, Sébastien Fiorucci, Nicolas Montagné*, Jérôme Golebiowski*, & Emmanuelle Jacquin-Joly*

Scientific Reports 2020, *10* (1), 1655–1655.

doi.org/10.1038/s41598-020-58564-9



Abstract

Odorant receptors expressed at the peripheral olfactory organs are key proteins for animal volatile sensing. Although they determine the odor space of a given species, their functional characterization is a long process and remains limited. To date, machine learning virtual screening has been used to predict new ligands for such receptors in both mammals and insects, using chemical features of known ligands. In insects, such approach is yet limited to Diptera, whereas insect odorant receptors are known to be highly divergent between orders. Here, we extend this strategy to a Lepidoptera receptor, SlitOR25, involved in the recognition of attractive odorants in the crop pest *Spodoptera littoralis* larvae. Virtual screening of 3 million molecules predicted 32 purchasable ones whose function has been systematically tested on SlitOR25, revealing 11 novel agonists with a success rate of 28%.

Our results show that Support Vector Machine optimizes the discovery of novel agonists and expands the chemical space of a Lepidoptera OR. More, it opens up structure-function relationship analyses through a comparison of the agonist chemical structures. This proof-of-concept in a crop pest could ultimately enable the identification of OR agonists or antagonists, capable of modifying olfactory behaviors in a context of biocontrol.

Introduction

Animals are exposed in their environment to a plethora of odorant molecules from a variety of chemical structures. Some of these molecules contain valuable information to carry out essential activities such as the identification of food sources, oviposition sites, mating partners, conspecifics and predators. Animals detect odorants via olfactory sensory neurons (OSNs) housed in dedicated olfactory organs, and the mechanisms underlying this detection have been particularly well studied in insects and mammals[1]. In insects, the primary olfactory organs consist of the antennae and the maxillary palps, which are covered by olfactory sensilla that house the OSNs[2]. In mammals, OSNs are mainly localized within the olfactory epithelium of the nasal cavity. In both insects and mammals, large multigenic families of odorant receptor proteins (ORs) mediate odorant recognition, each OSN expressing a single receptor (plus ORco in insects, see below) that controls its detection spectrum. These ORs are seven-transmembrane (TM) domain receptors[3–5], yet mammalian and insect ORs belong to distinct unrelated families[6]. Mammalian ORs are members of the class A rhodopsin-like G protein–coupled receptors (GPCR)[7], whereas insect OR membrane topology is opposite to that of GPCRs, with

a cytoplasmic N-terminus and an extracellular C-terminus[8]. Furthermore, insect ORs form heteromers with a well conserved coreceptor named ORco[8–10], and these heteromers are gated directly by chemical stimuli[11].

Understanding how the OR repertoire of an animal contributes to odor sensing and adaptation to a specific environment relies on the capacity to identify natural ligands of these ORs, a process called deorphanization. Yet, the ligands of several mammalian and insect ORs have been identified using different expression systems[12–19]. However, the number of chemicals used to stimulate the ORs is limited due to practical handling and duration of the experimentation. Consequently, potential stimuli that are tested on ORs of a given species are generally only a small portion of the vast array of ecologically relevant odorants. In insects, such sets of potential stimuli consisted of up to 100 molecules used to challenge *Drosophila melanogaster*[19] (even up to 500 in one study but with only one replicate[20]) and *Anopheles gambiae* ORs[16,17], but only fifty have been used to stimulate the ORs of a moth, *Spodoptera littoralis*[18]. Given that the potential odor space for an animal is almost unlimited, it is likely that the main ligand(s) of some deorphanized ORs still remains unidentified. The problem of selecting the candidate molecules to be tested becomes even more critical when trying to identify agonists or antagonists of particular ORs that are not natural ligands but could have an impact on the behavior of pest and disease vector insects[21].

Several recent studies revealed that the application of machine learning in the context of virtual screening opens up the possibility to enlarge animal odor spaces. Machine learning based on odorant chemical descriptors allowed predicting receptor–odorant interactions in both insects[22–25] and mammals[26], although their ORs do not belong to the same protein families. Notably, quantitative structure-activity relationship (QSAR) is an *in silico* ligand-based method used to predict biological activity of untested chemicals, based on chemical features shared by active molecules[27]. In *D. melanogaster*, virtual screening of more than 240,000 chemical structures identified a large array of novel OR activators and inhibitors[25]. An *in silico* screening of 0.5 million compounds identified agonist or antagonist targeting the mosquito CO_2 receptor, leading to the discovery of new attractants and repellents for those harmful disease vectors[24]. More recently, antagonists for the insect coreceptor ORco have been identified by screening a library of 1280 odorant molecules[28]. In mammals, a more modest virtual screening of 258 chemicals anyhow identified new agonists of four human ORs[26]. Although efficient, this approach requires prior knowledge on the response spectrum of a given OR and its application has thus been restricted to model species with cumulative odorant-receptor functional data.

We have recently deorphanized a large array of ORs in the noctuid moth *Spodoptera littoralis* through heterologous expression in *Drosophila* OSNs[18]. This offers an unprecedented opportunity to test such a computational approach in a non-dipteran insect. *Spodoptera littoralis* is a polyphagous moth[29] present in Africa, the Middle East and Southern Europe[30]. At the larval stage, *S. littoralis* is responsible for extensive damage in a large number of crops of economic importance[29]. Establishing machine learning virtual screening efficiency in such an herbivorous pest species will open new routes for the identification of possible agonists and antagonists to be used in biocontrol strategies. In addition, screening structurally related molecules can bring crucial information to determine structure-function relationships. Here, we focused on *S. littoralis* OR25 (SlitOR25), an odorant receptor that is particularly suitable for this approach. Over a panel of ~52 volatile organic compounds, SlitOR25 is strongly activated by nine agonists and moderately activated by four.[18] Also, it is expressed at both larval and adult stages and its activation has been correlated with caterpillar attraction[31]. Based on properties of the previously identified SlitOR25 ligands, we carried out an *in silico* screening of a chemical space of more than three million chemicals, leading to the prediction of 90 potential agonists, of which 32 were commercially available. The activity of these 32 compounds was further functionally tested on SlitOR25 expressed in *Drosophila* OSNs. We revealed enrichment of SlitOR25 agonists, with a hit rate of 28%. With the current lack of any OR structure - apart that of ORco[32]- , this machine-learning protocol based on chemical molecular descriptors thus represents an efficient tool for addressing ligand structure-function relationship in addition to identifying novel unexpected ligands for moth ORs, extending their odor space outside the presupposed relevant odorants.

Results and discussion

In silico prediction of SlitOR25 agonists

First, the published SlitOR25 chemical space[18] was analyzed through calculation of its known ligand chemical descriptors and projection on the *Drosophila melanogaster* Database of Odorant Responses (DoOR v2.0)[33], considered as prototypical. Figure 1a simplifies this chemical space using a t-distributed stochastic neighbor embedding (t-SNE) algorithm in two dimensions. Agonists were split into two distinct clusters, suggesting that a machine learning model (Fig. 1b) should be able to identify rules to separate them from non-agonists (see Supplementary Tab. S1 for a list of the considered molecules). Then, the external dataset to be

screened was obtained by filtering ~90 million molecules from the PubChem database as described in the method section. More than three million molecules corresponding to organic potential volatile molecules were extracted and were evaluated by the optimized Support Vector Machine (SVM). After an additional filter associated with the applicability domain obtained by a similarity search with the known agonists, 90 molecules were predicted as agonists (Fig. 1c and Supplementary Tab. S2). The performance of the SVM is resumed in Tab. 1 and Supplementary Tab. S3.

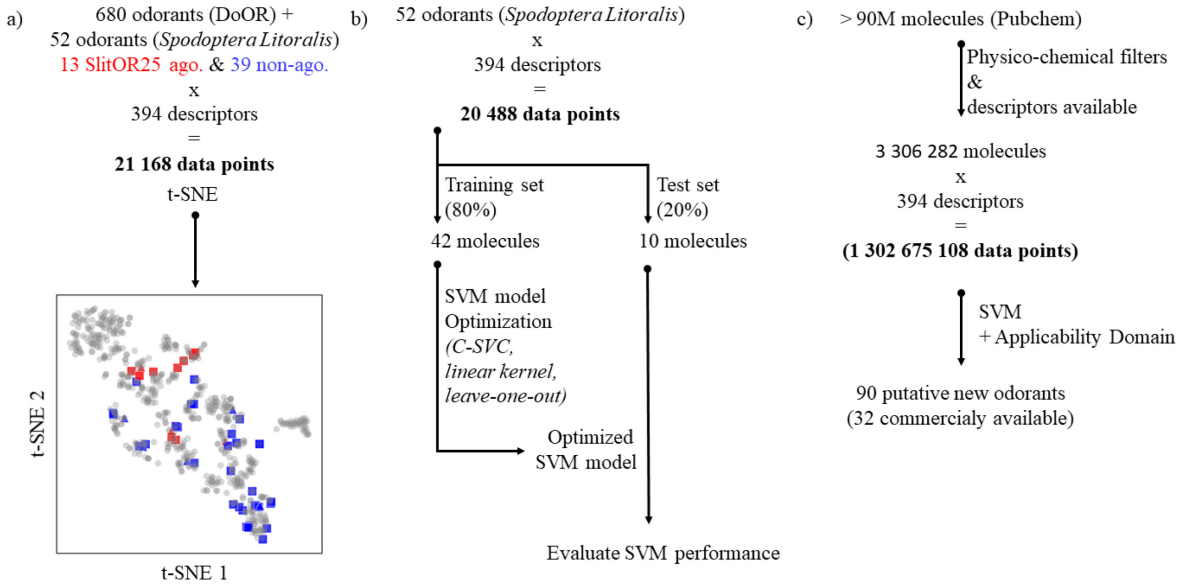


Figure 1: Analysis of insect odorant molecular space and protocol used for *Spodoptera littoralis* OR25 (SlitOR25) virtual screening. **(a)** Visualization of SlitOR25 and *Drosophila melanogaster* olfactory chemical spaces based on a t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction method. The agonists (ago) and non-agonists of SlitOR25 are shown in red and blue, respectively, and agonists of *D. melanogaster* are shown in gray. Chemicals of the training set are shown in squares while those of the test set are shown as triangles **(b)** Workflow of the Support Vector Machine (SVM) model based on an 80%/20% split of the initial database. Forty-two molecules constituting the training set were used to find optimized SVM parameters while 10 molecules were kept for a blind evaluation by the optimized SVM (Supplementary Tab. S1). C-SVC: C-Support Vector Classification. **(c)** Virtual screening of more than three million molecules extracted from the PubChem database resulted in 90 agonist candidates.

Table 1. 5-fold random split Support Vector Machine performance metrics. %CC: percentage of instances correctly classified, MCC: Matthews correlation coefficient.

Dataset	%CC	Precision	Recall	MCC
Training	0.90±0.03	0.77±0.05	0.84±0.08	0.77±0.07
Test	0.92±0.06	0.88±0.16	0.91±0.12	0.83±0.12

Table 2. Predicted agonists (this study) and known ligands[18] (in bold) tested on SlitOR25.

Compounds	CAS	Provider	Purity
1-Naphthaldehyde	66-77-3	Alfa Aesar	97%
2'-Fluoroacetophenone	445-27-2	Alfa Aesar	97%
Phenylglyoxal monohydrate	1074-12-0	Acros organics	97%
Terephthalaldehyde	623-27-8	Alfa Aesar	98%
Isophthalaldehyde	626-19-7	Alfa Aesar	98%
1,3-benzenedimethanol	626-18-6	Alfa Aesar	98%
2-Fluorobenzaldehyde	446-52-6	Alfa Aesar	97%
2-Fluorobenzyl alcohol	446-51-5	Alfa Aesar	98%
4-Fluorobenzaldehyde	459-57-4	Alfa Aesar	98%
4-Flourobenzyl alcohol	459-56-3	Alfa Aesar	97%
3,4-Difluorobenzaldehyde	34036-07- 2	Alfa Aesar	98%
3,4-Difluorobenzyl alcohol	85118-05- 4	Alfa Aesar	99%
2,3,4-Trifluorobenzyl alcohol	144284- 24-2	Alfa Aesar	97%
Salicylic acid	69-72-7	VWR chemicals	98%
3-Fluorobenzyl alcohol	456-47-3	Alfa Aesar	98%
3- Fluorobenzaldehyde	456-48-4	Alfa Aesar	97%
2,5-Difluorobenzaldehyde	2646-90-4	Alfa Aesar	98%
2,6-Difluorobenzaldehyde	437-81-0	Alfa Aesar	97%
3,5-Difluorobenzyl alcohol	79538-20- 8	Alfa Aesar	97%
3,5-Difluorobenzaldehyde	32085-88- 4	Alfa Aesar	97%
2,4-Difluorobenzyl alcohol	56456-47- 4	Alfa Aesar	98%
2,4-Difluorobenzaldehyde	1550-35-2	Alfa Aesar	98%
2,3-Difluorobenzaldehyde	2646-91-5	Alfa Aesar	98%
3,4,5-Trifluorobenzyl alcohol	220227- 37-2	Alfa Aesar	97%
2,4,5-Trifluorobenzyl alcohol	144284- 25-3	Alfa Aesar	98%
1,3-Indanedione	606-23-5	Alfa Aesar	97%
p-tolualdehyde	104-87-0	Alfa Aesar	98%
4'-Fluoroacetophenone	403-42-9	Alfa Aesar	99%
2',4'-Difluoroacetophenone	364-83-0	Alfa Aesar	98%
2-Methoxybenzoic acid	579-75-9	Alfa Aesar	98%
2,3-Difluorobenzyl alcohol	75853-18- 8	Alfa Aesar	97%
2,5-Difluorobenzyl alcohol	75853-20- 2	Alfa Aesar	98%
Benzaldehyde	100-52-7	Sigma-Aldrich	99,5%
Z-3-hexenol	928-96-1	Sigma-Aldrich	98%
Methyl salicylate	119-36-8	Sigma-Aldrich	99%
2-phenyl acetaldehyde	122-78-1	Sigma-Aldrich	98%
Benzyl methyl ether	538-86-3	Sigma-Aldrich	98%
Methyl benzoate	93-58-3	Acros organics	97%
Benzyl alcohol	100-51-6	Sigma-Aldrich	99%
Acetophenone	98-86-2	Acros organics	99%
E-2-hexenol	928-95-0	Sigma-Aldrich	96%
E-2-hexenal	6728-26-3	Sigma-Aldrich	98%
1-hexanol	111-27-3	Sigma-Aldrich	98%
1-heptanol	111-70-6	Sigma-Aldrich	99%

Effect of predicted agonists on SlitOR25 activity

Among the predicted potential novel agonists of SlitOR25, 32 molecules were commercially available at high purity (Tab. 2). These molecules were mainly fluorinated derivatives of known ligands (acetophenone, benzyl alcohol, benzaldehyde). To verify whether these were indeed agonists of SlitOR25, we performed single-sensillum recordings on *D. melanogaster* flies expressing SlitOR25 in ab3A OSNs instead of the endogenous receptor OR22a, a heterologous expression system known as the "empty neuron"[34]. A first screen with a high concentration of the 32 candidate agonists (10^{-2} dilution) revealed that nine of them elicited a significant response (<0.05, Fig. 2), representing a 28% success rate. For comparison, 30% of 138 *in silico* predicted odorants activated the mosquito CO_2 receptor in a first round[24]. Machine learning models based on ligand topology predicted 138 antagonists for mosquito ORco, out of which 45 were active (32%)[28]. In this last study, it has to be noticed that 58 active antagonists were used to feed the machine learning, a number that is much higher than the 13 ligands we used. In *Drosophila*, another study revealed that the success rate of an optimized QSAR greatly depends on the receptor (varying from 27% to 71%)[25] and that lowest rates were obtained for ORs tuned to aromatics (around 30%). Here, we add new evidences that machine learning is of great help to discover novel ligands for Lepidoptera ORs.

Looking in detail at the new ligands identified for SlitOR25, none presented a reverse agonist activity (reduction of spontaneous activity), whereas this has been observed for 13% of predicted ligands for *D. melanogaster* ORs when tested on OSNs[25]. This is likely attributed to the nature of the screened receptor, where reverse agonists would be part of a far-removed chemical space compared to agonists. However, with the current lack of any structure of an insect OR (apart that of ORco)[32], providing a mechanistic view on the way agonists work is extremely difficult.

Dose-response analyses

To compare the responses evoked on SlitOR25 by the nine newly identified agonists to those evoked by the previously known natural ligands, we conducted dose-response SSR experiments, using dilutions ranging from 10^{-7} to 10^{-2}, and effective doses 50 (ED50s) were calculated. Statistical analyses for the responses of all molecules tested in dose-response are detailed in Table 3.

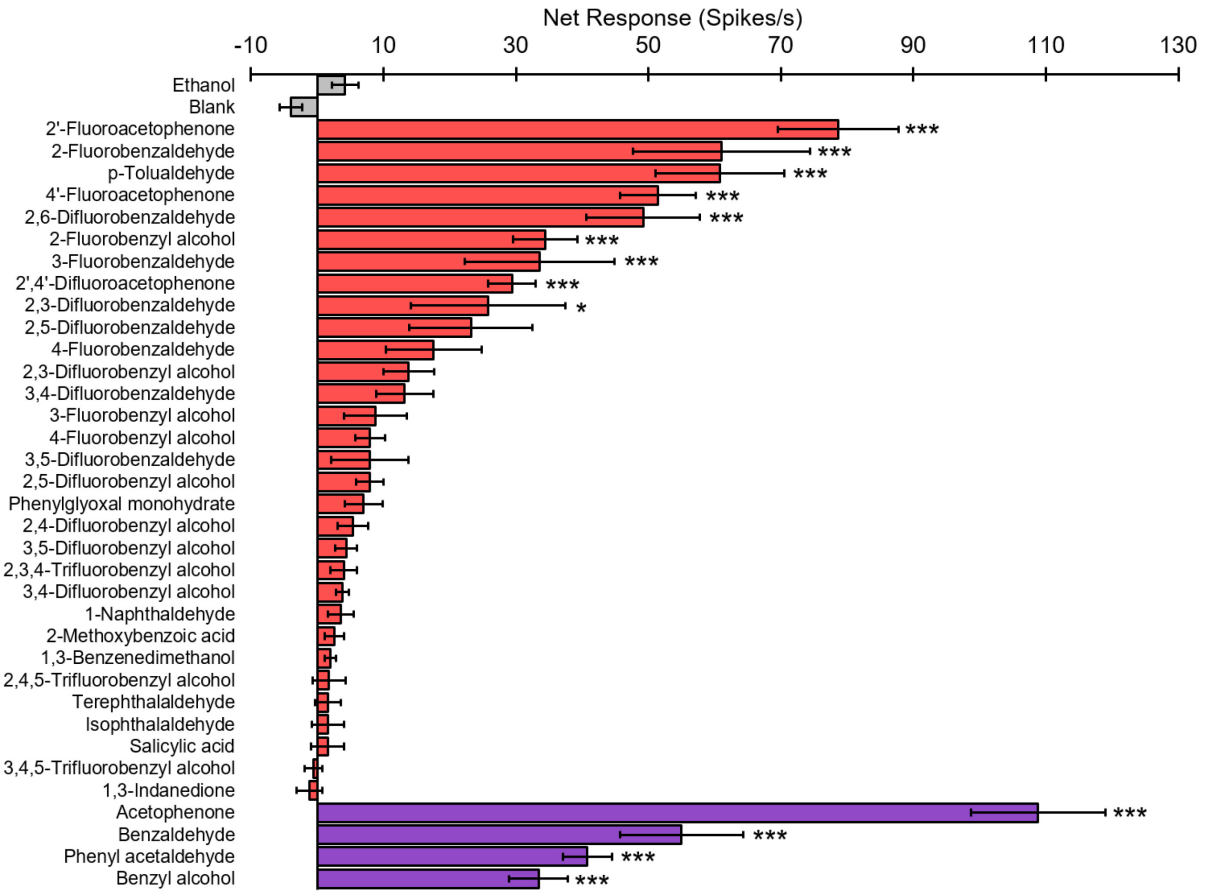


Figure 2. Response of *Drosophila* ab3A OSNs expressing SlitOR25 to 32 candidate ligands predicted via ligand-based QSAR approach. Responses are presented ± s.e.m. Grey bars: controls (ethanol solvent, blank). Red bars: predicted compounds tested in SSR at high doses (10^{-2}, ethanol dilution). Purple bars: known SlitOR25 ligands used as positive controls[18] (10^{-2}, ethanol dilution). Asterisks indicate statistically significant differences between responses to the odorant and to the solvent (Kruskal–Wallis test followed by a Dunnett multiple comparison test, * $p<0.05$, *** $p<0.001$, n=10).

For predicted molecules structurally related to the ligand acetophenone (Fig. 3a), statistically significant responses ($p<0.05$) were observed for all tested molecules from 10^{-6} dilution, but ED50s were higher than that of acetophenone although efficiencies were similar. For the newly predicted ligands structurally related to benzaldehyde (Fig. 3b), detection threshold started from 10^{-4} dilution. Interestingly, their ED50s were all lower than that of benzaldehyde (higher potency), although 2-fluorobenzaldehyde efficiency was much lower. The predicted agonist 2-fluorobenzyl alcohol exhibited a higher activation threshold than the structurally related-known ligand benzyl alcohol (Fig. 3c). Our results demonstrated that machine learning was very efficient in identifying new strong ligands for SlitOR25.

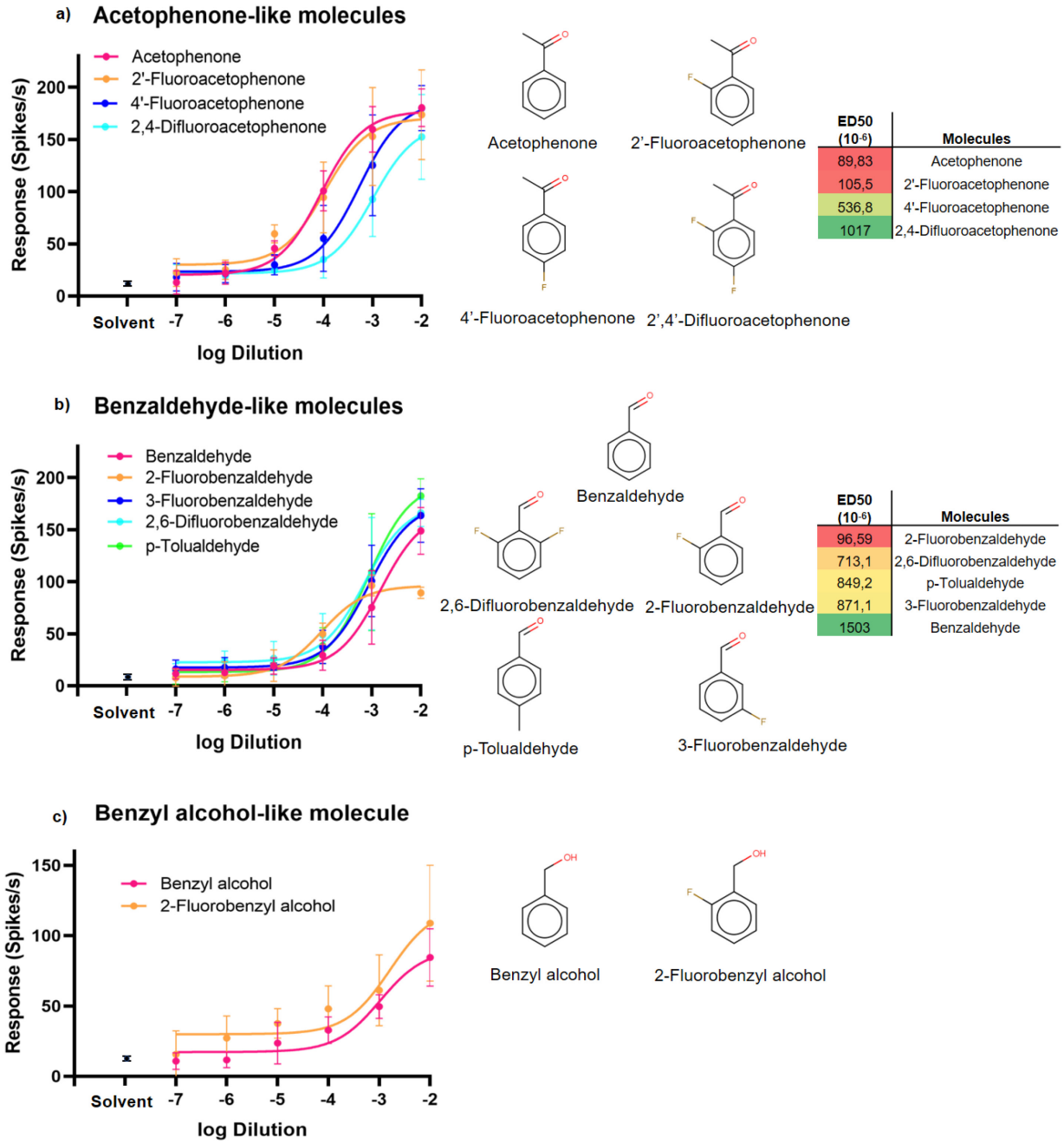


Figure 3. Dose-response activities (measured via SSR) of newly identified and three previously identified ligands on SlitOR25 expressed in *Drosophila* ab3A OSNs, structures and ED50 values. SSR responses are presented ± s.e.m. Only molecules with a significant activity in the screening tests ($p<0.05$, 10^{-2} dilution, Fig. 2) were tested in dose-response using dilutions from 10^{-7} to 10^{-2}. (a) Molecules structurally related to the known ligand acetophenone: 2'-fluoroacetophenone, 4'-fluoroacetophenone, 2',4'-difluoroacetophenone. (b) Molecules related to the ligand benzaldehyde: 2-fluorobenzaldehyde, 3-fluorobenzaldehyde, 2,6-difluorobenzaldehyde, p-tolualdehyde. (c) Molecules related to the ligand benzyl alcohol: 2-fluorobenzyl alcohol.

Table 3. Statistics for the responses of known (acetophenone, benzyl alcohol and benzaldehyde) and new SlitOR25 ligands tested in dose-response test. Solvent: ethanol.

Tested molecules	Dilutions					
	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
Acetophenone	NS	***	***	***	***	***
Benzyl alcohol	NS	NS	NS	***	***	***
Benzaldehyde	NS	***	***	***	***	***
2'-Fluoroacetophenone	NS	***	***	***	***	***
2-Fluorobenzaldehyde	NS	NS	NS	***	***	***
2-Fluorobenzyl alcohol	NS	NS	***	***	***	***
3-Fluorobenzaldehyde	NS	*	***	***	***	***
2,6-Difluorobenzaldehyde	***	***	***	***	***	***
p-Tolualdehyde	NS	NS	***	***	***	***
4'-Fluoroacetophenone	NS	***	***	***	***	***
2',4'-Difluoroacetophenone	***	***	***	***	***	***

Asterisks indicate statistically significant differences between responses to the odorant and to solvent (Kruskal–Wallis test followed by a Dunnett multiple comparison test, * $p<0.05$, *** $p<0.001$, n=5).

While acetophenone has been previously reported as the best ligand for SlitOR25[18], 2-fluorobenzaldehyde appeared as equivalent. The best agonists for SlitOR25 were acetophenone, 2'-fluoroacetophenone, and 2-fluorobenzaldehyde, with ED50 of ~100.10^{-6}. Other benzaldehyde derivatives appeared much less potent (ED50 in the range 500 to 1000.10^{-6}), followed by two benzyl alcohols (ED50 > 1000.10^{-6}. Independent of the pharmacophore approach and by visually inspecting the structures, the presence of a Fluor atom at the ortho position in the ring (position 2) maintains the agonist behaviour for the three chemical families (aldehydes, ketones, alcohol). Multiple fluorinations had either a weak beneficial effect on benzaldehyde derivatives or decreased or abolished the response in other series (Supplementary Fig. S1).

The predicted molecules we functionally tested present strongly intertwined chemical spaces. The functional assays we conducted revealed that some were strong agonists and other were non-agonists (Supplementary Fig.S1), allowing us to tentatively recapitulate the features required for being an agonist through a pharmacophore approach (Fig. S2). However, the model was not able to discriminate agonist from non-agonists based on the position of the Fluor atom on the aromatic cycle.

Alternatively, a statistical analysis of the descriptors able to discriminate between agonists and non-agonists revealed 105 descriptors out of the 394 processed initially. These descriptors can

either be constitutional, topologic, or electronic. They are hardly interpretable but could serve as a basis for a further screening protocol.

Conclusion

Machine learning widens the chemical space of a moth odorant receptor

In this study, we have used machine learning to predict novel agonists for SlitOR25, a broadly tuned receptor in the Lepidoptera *S. littoralis*. A Support Vector Machine was fed with 52 ligands for which the activity was already reported. After optimization, a database of more than 90 million chemicals was filtered and screened. Out of the three million of potentially useful molecules, 90 were predicted as agonists, of which 32 were commercially available. *In vivo* functional assays and dose-response analyses on these latter assessed nine novel molecules as moderate or strong agonists for the receptor.

Modeling has already been shown to provide accurate information and facilitate the selection of active molecules on odorant receptors. In insects, it has been applied only in two Diptera models, the fruit fly and the mosquito[24,25,28]. In this study, we reveal that a conventional machine learning approach is efficient for the identification of novel agonists for a moth receptor, whose amino acid sequence is unrelated to that of Diptera ORs.

It has to be noticed that none of the novel agonists discovered here has been previously described in the literature to be active on moth ORs and most are not described as plant emitted volatiles. Although they may not be encountered by insects in the wild, we have anyhow extended the chemical space of *S. littoralis* and the cumulated results open up ligand structure-function relationship analyses. More importantly, we have opened a closed-loop machine learning, where the new highly potent agonists discovered here could be used to train new models, further improving predictions in alternative and far removed chemical spaces.

Methods

Reagents

Reagents were purchased from various vendors (Tab. 2) at the highest available purity (ranging from 96 to 99% depending on the molecules) and were dissolved in ethanol. Ethanol: 96% purity, Carlo Erba reagents.

Quantitative Structure Activity Relationship

Softwares: Knime v3.2.1 was used to build the workflow, chemical descriptors were computed with Dragon v6.0.40 and the LibSVM v2.89 was used for the machine learning protocol[35].

Training and test set: The initial database of 52 volatiles (Supplementary Tab. S1) was obtained from[18]. The previously identified strong agonists of SlitOR25 were benzenoids (acetophenone, benzyl alcohol, benzaldehyde, phenyl acetaldehyde, 1-indanone) and short aliphatic alcohols and aldehydes (1-hexanol, 1-heptanol, (Z)3-hexenol, (E)2-hexenal)[18], which are compounds emitted mainly by flowers and leaves[36]. The receptor also responds to four other molecules (methyl salicylate, methyl benzoate, benzyl methyl ether, (E)2-hexenol), with weaker but still significant responses. The SlitOR25 database thus contains 13 agonists and 39 non-agonists. It was randomly split into a training set of 42 molecules and a test set of 10 molecules. Molecules of the training set were considered for the optimization of the model. Those of the test set were not used to build the model but to assess its performance.

External test set: 3 306 388 molecules out of 90 million were extracted from the Pubchem database[37] according to the following physico-chemical properties obtained directly on the website: each molecule has to contain a combination of C, H, O, N, F, S, or Cl elements with less than 20 heavy atoms, a molecular weight lower than 200 $g.mol^{-1}$, and a LogP in the range [0, 5].

Chemical space analysis: The Database of Odorant Response (DoOR v2.0)[33] was used to analyze the *S. littoralis* chemical space that has been used to train the machine learning model. Excluding salts from the analysis, DoOR contains 680 odorants that have been experimentally

tested on *D. melanogaster.* The t-SNE dimensionality reduction method was used to evaluate how our database of 52 ligands span a typical insect chemical space.

Molecular descriptors: For each dataset of the QSAR model (training, test and external sets) 4885 descriptors were computed using the *Dragon* software (version 6.0.38) based on 3D sdf files obtained directly from Pubchem. Constant or near-constant (variance lower than 0.005) descriptors were excluded from the database as well as descriptors with at least one missing value. Each descriptor of the final matrix was normalized using a min-max protocol (range [0,1]) before the split between training and test sets. Note that a normalization before or after the split did not affect the nature of the predicted agonists. Redundant descriptors were removed (absolute pair correlation greater than or equal to 0.95). The final SVM matrix contained 394 molecular descriptors. It was used for the t-SNE visualization of the database containing both the *S. littoralis* and *D. melanogaster* chemical spaces (see SI for details on t-SNE). The descriptors were computed on a machine with an intel Xeon with 32 GB of memory.

Setting up the QSAR model: Various numerical models, such as Random Forest or Perceptron (data not shown), were tested prior optimizing the chosen supervised machine learning method, *i.e.* Support Vector Machine (SVM). A brute force optimization was applied to assess the exhaustive parameter value combination. The C-SVC (C-Support Vector Classification) model with a linear kernel was finally used.

The C-SVC parameters were optimized in a two-step process. First a 5-fold-random split was performed with a cost ranging from 1 to 10 with a step of 1. Epsilon varied between 0.0001 and 0.1 with a step of 0.01. The model's accuracy remained identical for values in this range. Second a more precise 5-fold-random split sampling was performed, with a cost between 0.5 and 1.5 using a step of 0.1, and epsilon between 0.001 and 0.01 with a step of 0.001. Again, the accuracy was identical to that obtained with default settings (accuracy 0.9±0.09).

The optimized SVM parameters were accordingly set as follows: cost = 1.0, epsilon 0.001. The leave-one-out cross validation method was used. Each of the 13 agonists was given a score of 1 and the non-agonists were given a score of 0.

Applicability domain: A Tanimoto score, which measures the similarity between compounds (and varies between 0 and 1 whereby a value closer to 1 indicates greater similarity) was

calculated from Pubchem molecular fingerprints (881 Pubchem molecular descriptors obtained from the CDK module of Knime). The use of Pubchem fingerprints has already been shown to correctly capture biological activities[38]. Putative new odorants which has a Tanimoto index higher than 0.92 with respect to the Training set were considered belonging to the applicability domain. In our case, this corresponds to 90 molecules.

Single-sensillum recordings of Drosophila olfactory sensory neurons

Flies were reared on standard cornmeal-yeast-agar medium and kept in a climate and light-controlled environment (25 °C, 12 h light: 12 h dark cycle). SlitOR25-expressing flies were obtained by crossing the line *w;Δhalo/CyO;UAS-SlitOr25*[18] with the line *w; Δhalo/CyO;Or22a-Gal4*[34]. For each experiment, a 2- to 8-day-old fly was restrained in a pipette tip with only the head protruding. The tip was fixed on a microscope glass slide and one antenna was gently maintained using a glass capillary. The preparation was placed under a constant 1.5 $L.min^{-1}$ flux of charcoal-filtered and humidified air delivered through a glass tube of a 7 mm diameter, and observed with a light microscope (BX51WI, Olympus, Tokyo, Japan) equipped with a 100X magnification objective. Action potentials from ab3A OSNs were recorded using electrolytically sharpened tungsten electrodes (TW5-6, Science Products, Hofheim, Germany). The reference electrode was inserted into the eye and the recording electrode was inserted at the base of an ab3 sensillum using a motor-controlled PatchStar micromanipulator (Scientifica, Uckfield, United Kingdom). The electrical signal was amplified using an EX-1 amplifier (Dagan Corporation, Minneapolis, MN, USA), high-pass (1Hz) and low-pass (3 kHz) filtered and digitized (10 kHz) through a Digidata 1440A acquisition board (Molecular Devices, Sunnyvale, CA, USA) then recorded and analyzed using pCLAMP™ 10 (Molecular Devices).

The responses of ab3A OSNs were calculated by subtracting the spontaneous firing rate (in $spikes.s^{-1}$) from the firing rate during the odorant stimulation. The time windows used to measure these two firing rates lasted for 500 ms and were respectively placed 500 ms before and 100 ms after the onset of stimulation (to consider the time for the odorants to reach the antenna). Stimulus cartridges were built by placing a 1 cm^2 filter paper in the large opening end of a Pasteur pipette and loading 10 µl of the odorant solution onto the paper (10^{-2} dilution in ethanol), or 10 µL of ethanol as control. Evaporation time before using the cartridge was 10 minutes. Odorant stimulations were performed by inserting the tip of the pipette into a hole in the glass tube and generating a 500 ms air pulse (0.6 $L.min^{-1}$), which reached the permanent air flux while going through the stimulation cartridge.

The absence of the endogenous receptor OR22a in ab3A OSNs was verified using ethyl hexanoate (a strong ligand of OR22a) as a stimulus. Then, the SlitOR25 response spectrum was established using the panel of 32 predicted agonists (Tab. 2) and four already known ligands as controls. The stimulus cartridges were used at most two times per fly and a maximum of eight times in total. The entire panel of molecules was tested ten times on ten different flies expressing SlitOR25. Odorants were considered as active if the response they elicited was statistically different from the response elicited by the solvent alone (Kruskal–Wallis test followed by a Dunnett multiple comparison test, $p<0.05$).

For molecules that yielded a statistically significant response, dose-response experiments were conducted with odorant dilutions ranging from 10^{-2} down to 10^{-7}. Each dilution was tested in five different flies expressing SlitOR25. ED50 were calculated (except for benzyl alcohol and 2-fluorobenzyl alcohol) using GraphPad PRISM V.8.1.2 software.

SlitOR25 pharmacophore hypothesis

For the generation of the SlitOR25 pharmacophore, we considered a dataset of eleven odorants that are active on SlitOR25, as well as fourteen inactive compounds. All these molecules are derivatives of acetophenone described in this work. The pharmacophore was generated with up to four features, chosen between H-bond donors/acceptors, hydrophobic sites, and aromatic rings. Even considering several conformations for each molecule, the pharmacophore hypotheses generated by the software CATALYST (version 4.9.1, Accelrys Inc., San Diego, CA, August 2004) were identical, comprised of an aromatic ring and a H-bond acceptor. The addition of exclusion volumes did not improve the model and was thus discarded.

Acknowledgments

The authors thank Philippe Touton (iEES-Paris) for insect rearing and Xiaojing Cong (ICN, Nice) for help with the software PRISM V.8.1.2. This work has been funded by Inra, Sorbonne Université and the French National Research Agency (ANR-16-CE21-0002-02). GC-V received doctoral fellowships from BECAL and the National Council of Science and Technology of Paraguay. C.B. has been granted by GIRACT for a PhD bursary. We also benefited from funding from the French government, through the UCAJEDI “Investments in the Future” project managed by the ANR grant No. ANR-15-IDEX-01.

Author contributions

E.J-J., J.G., N.M. and S.F. conceived and designed the experiments. G.C-V and C.B designed and performed the experiments and analyzed the data. H.G. perform experiments. E.J-J., J.G., G.C-V., C.B., N.M. and S.F. wrote and revised the paper. All authors reviewed the manuscript.

References

1. Kaupp, U. B. Olfactory signalling in vertebrates and insects: differences and commonalities. *Nat Rev Neurosci* **11**, 188–200 (2010).

2. Leal, W. S. Odorant Reception in Insects: Roles of Receptors, Binding Proteins, and Degrading Enzymes. *Annu. Rev. Entomol.* **58**, 373–391 (2013).

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

4. Vosshall, L. B., Amrein, H., Morozov, P. S., Rzhetsky, A. & Axel, R. A Spatial Map of Olfactory Receptor Expression in the Drosophila Antenna. *Cell* **96**, 725–736 (1999).

5. Clyne, P. J. *et al.* A Novel Family of Divergent Seven-Transmembrane Proteins. *Neuron* **22**, 327–338 (1999).

6. Su, C.-Y., Menuz, K. & Carlson, J. R. Olfactory Perception: Receptors, Cells, and Circuits. *Cell* **139**, 45–59 (2009).

7. Mombaerts, P. Seven-Transmembrane Proteins as Odorant and Chemosensory Receptors. *Science* **286**, 707–711 (1999).

8. Benton, R., Sachse, S., Michnick, S. W. & Vosshall, L. B. Atypical Membrane Topology and Heteromeric Function of Drosophila Odorant Receptors In Vivo. *PLoS Biol* **4**, e20 (2006).

9. Larsson, M. C. *et al.* Or83b Encodes a Broadly Expressed Odorant Receptor Essential for Drosophila Olfaction. *Neuron* **43**, 703–714 (2004).

10. Vosshall, L. B. & Hansson, B. S. A Unified Nomenclature System for the Insect Olfactory Coreceptor. *Chemical Senses* **36**, 497–498 (2011).

11. Silbering, A. F. & Benton, R. Ionotropic and metabotropic mechanisms in chemoreception: 'chance or design'? *EMBO Rep* **11**, 173–179 (2010).

12. Peterlin, Z., Firestein, S. & Rogers, M. E. The state of the art of odorant receptor deorphanization: A report from the orphanage. *Journal of General Physiology* **143**, 527–542 (2014).

13. Montagné, N., de Fouchier, A., Newcomb, R. D. & Jacquin-Joly, E. Advances in the Identification and Characterization of Olfactory Receptors in Insects. in *Progress in Molecular Biology and Translational Science* vol. 130 55–80 (Elsevier, 2015).

14. Silva Teixeira, C. S., Cerqueira, N. M. F. S. A. & Silva Ferreira, A. C. Unravelling the Olfactory Sense: From the Gene to Odor Perception. *CHEMSE* bjv075 (2015) doi:10.1093/chemse/bjv075.

15. Wang, B., Liu, Y., He, K. & Wang, G. Comparison of research methods for functional characterization of insect olfactory receptors. *Sci Rep* **6**, 32806 (2016).

16. Wang, G., Carey, A. F., Carlson, J. R. & Zwiebel, L. J. Molecular basis of odor coding in the malaria vector mosquito Anopheles gambiae. *Proceedings of the National Academy of Sciences* **107**, 4418–4423 (2010).

17. Carey, A. F., Wang, G., Su, C.-Y., Zwiebel, L. J. & Carlson, J. R. Odorant reception in the malaria mosquito Anopheles gambiae. *Nature* **464**, 66–71 (2010).

18. de Fouchier, A. *et al.* Functional evolution of Lepidoptera olfactory receptors revealed by deorphanization of a moth repertoire. *Nat Commun* **8**, 15709 (2017).

19. Hallem, E. A. & Carlson, J. R. Coding of Odors by a Receptor Repertoire. *Cell* **125**, 143–160 (2006).

20. Mathew, D. *et al.* Functional diversity among sensory receptors in a *Drosophila* olfactory circuit. *Proc Natl Acad Sci USA* **110**, E2134–E2143 (2013).

21. Ray, A. Reception of odors and repellents in mosquitoes. *Current Opinion in Neurobiology* **34**, 158–164 (2015).

22. Katritzky, A. R. *et al.* Synthesis and bioassay of improved mosquito repellents predicted from chemical structure. *Proceedings of the National Academy of Sciences* **105**, 7359–7364 (2008).

23. Oliferenko, P. V. *et al.* Promising Aedes aegypti Repellent Chemotypes Identified through Integrated QSAR, Virtual Screening, Synthesis, and Bioassay. *PLoS ONE* **8**, e64547 (2013).

24. Tauxe, G. M., MacWilliam, D., Boyle, S. M., Guda, T. & Ray, A. Targeting a Dual Detector of Skin and CO2 to Modify Mosquito Host Seeking. *Cell* **155**, 1365–1379 (2013).

25. Boyle, S. M., McInally, S. & Ray, A. Expanding the olfactory code by in silico decoding of odor-receptor chemical space. *eLife* **2**, e01120 (2013).

26. 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).

27. Mansouri, K. & Judson, R. S. In Silico Study of In Vitro GPCR Assays by QSAR Modeling. in *In Silico Methods for Predicting Drug Toxicity* (ed. Benfenati, E.) vol. 1425 361–381 (Springer New York, 2016).

28. Kepchia, D. *et al.* Use of machine learning to identify novel, behaviorally active antagonists of the insect odorant receptor co-receptor (Orco) subunit. *Sci Rep* **9**, 4055 (2019).

29. Salama, H. S., Dimetry, N. Z. & Salem, S. A. On the Host Preference and Biology of the Cotton Leaf Worm Spodoptera littoralis Bois. *Zeitschrift für Angewandte Entomologie* **67**, 261–266 (1971).

30. EFSA Panel on Plant Health. Scientific Opinion on the pest categorisation of Spodoptera littoralis. *EFSA Journal* doi:10.2903/j.efsa.2015.3987.

31. de Fouchier, A. *et al.* Behavioral Effect of Plant Volatiles Binding to Spodoptera littoralis Larval Odorant Receptors. *Front. Behav. Neurosci.* **12**, 264 (2018).

32. Butterwick, J. A. *et al.* Cryo-EM structure of the insect olfactory receptor Orco. *Nature* **560**, 447–452 (2018).

33. Münch, D. & Galizia, C. G. DoOR 2.0 - Comprehensive Mapping of Drosophila melanogaster Odorant Responses. *Sci Rep* **6**, 21841 (2016).

34. Dobritsa, A. A., van der Goes van Naters, W., Warr, C. G., Steinbrecht, R. A. & Carlson, J. R. Integrating the Molecular and Cellular Basis of Odor Coding in the Drosophila Antenna. *Neuron* **37**, 827–841 (2003).

35. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).

36. Knudsen, J. T., Eriksson, R., Gershenzon, J. & Ståhl, B. Diversity and Distribution of Floral Scent. *The Botanical Review* **72**, 1–120 (2006).

37. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**, D1102–D1109 (2019).

38. Hao, M., Wang, Y. & Bryant, S. H. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica Chimica Acta* **806**, 117–127 (2014).

Supporting information

Supplementary Table S1. Database of molecules used to train the machine learning model. Agonists marked with an * were not considered as strong agonists in the work by de Fouchier et al 2017[18], but the receptor response was still significantly different from solvent. They were thus included in our agonist list.

NAME	CAS	Classification	Training / Test set
benzaldehyde	100-52-7	agonist	training set
phenylacetaldehyde	122-78-1	agonist	training set
(E)2-hexenal	6728-26-3	agonist	training set
(E)2-hexenol	928-95-0	agonist*	training set
(Z)3-hexenol	928-96-1	agonist	training set
1-hexanol	111-27-3	agonist	training set
1-heptanol	111-70-6	agonist	test set
benzyl alcohol	100-51-6	agonist	test set
acetophenone	98-86-2	agonist	training set
1-indanone	83-33-0	agonist	training set
methyl salicylate	119-36-8	agonist*	training set
methyl benzoate	93-58-3	agonist*	training set
benzyl methyl ether	538-86-3	agonist*	test set
1-octanol	111-87-5	non agonist	training set
(Z)9-14: OH	35153-15-2	non agonist	training set
(Z)7-12:OAc	14959-86-5	non agonist	test set
(Z,E)-9,12-14:OAc	30507-70-1	non agonist	training set
(Z)-jasmone	488-10-8	non agonist	training set
α-copaene	3856-25-5	non agonist	training set
nonanal	124-19-6	non agonist	training set
sulcatone	110-93-0	non agonist	test set
α-humulene	6753-98-6	non agonist	training set
(E)11-14:OAc	33189-72-9	non agonist	training set
TMTT	62235-06-7	non agonist	training set
(Z)11-14:OAc	20711-10-8	non agonist	training set
decanal	112-31-2	non agonist	training set
(Z)9-14:OAc	16725-53-4	non agonist	test set
methyl jasmonate	39924-52-2	non agonist	training set
(E,E)-α-farnesene	502-61-4	non agonist	training set
(±)-linalool	78-70-6	non agonist	training set
(±)-phytol	7541-49-3	non agonist	training set
carvacrol	499-75-2	non agonist	training set
eugenol	97-53-0	non agonist	test set
β-myrcene	123-35-3	non agonist	training set
(±)-nerolidol	7212-44-4	non agonist	training set
hexane	110-54-3	non agonist	training set
β-caryophyllene	87-44-5	non agonist	test set

DMNT	19945-61-0	non agonist	training set
1-nonanol	143-08-8	non agonist	training set
EDD	3025-30-7	non agonist	training set
3-carene	13466-78-9	non agonist	test set
14:OAc	638-59-5	non agonist	training set
indole	120-72-9	non agonist	training set
(Z,E)-9,11-14:OAc	50767-79-8	non agonist	training set
geraniol	106-24-1	non agonist	training set
(Z)3-hexenyl acetate	3681-71-8	non agonist	training set
β-pinene	127-91-3	non agonist	training set
(E)-ocimene	3779-61-1	non agonist	test set
α-pinene	80-56-8	non agonist	training set
estragole	140-67-0	non agonist	training set
thymol	89-83-8	non agonist	training set
(E,E)-farnesol	106-28-5	non agonist	training set

Supplementary Table S2. Panel of 90 predicted agonist molecules for SlitOR25.

NAME	CID	CAS (if available)
Salicylic acid	338	69-72-7
P-Tolualdehyde	7725	104-87-0
4'-Fluoroacetophenone	9828	403-42-9
2-Fluoroacetophenone	9947	450-95-3
2-methoxybenzoic acid	11370	579-75-9
1,3-Indanedione	11815	606-23-5
terephthalaldehyde	12173	623-27-8
Pent-2-enal	12993	764-39-6
2-oxo-2-phenylacetaldehyde	14090	1074-12-0
2-Penten-1-ol	15306	1576-95-0
Isophthalaldehyde	34777	626-19-7
2',4'-Difluoroacetophenone	67770	364-83-0
2-Fluorobenzyl alcohol	67969	446-51-5
2-Fluorobenzaldehyde	67970	446-52-6
3-Fluorobenzyl alcohol	68008	456-47-3
3-Fluorobenzaldehyde	68009	456-48-4
4-Fluorobenzyl alcohol	68022	459-56-3
4-Fluorobenzaldehyde	68023	459-57-4
1,3-Benzenedimethanol	69374	626-18-6
2,4-Difluorobenzaldehyde	73770	1550-35-2
(2,6-difluorophenyl)methanol	87921	19064-18-7
2,4-Difluorobenzyl alcohol	91867	56456-47-4
2'-Fluoroacetophenone	96744	445-27-2
3H-indene-1,2-dione	123358	16214-27-0
2,6-difluorobenzaldehyde	136284	437-81-0
2,5-Difluorobenzaldehyde	137663	2646-90-4
2,3-Difluorobenzaldehyde	137664	2646-91-5
Benzocyclobutenone	137953	3469-06-5
Hydroperoxy(phenyl)methanol	286896	
2,3-Difluorobenzyl alcohol	447153	75853-18-8
2,5-Difluorobenzyl alcohol	522599	75853-20-2
3,5-Difluorobenzyl alcohol	522721	79538-20-8
3,4-Difluorobenzyl alcohol	522833	85118-05-4
hex-3-ene-1,6-diol	549321	67077-43-4
3,4-Difluorobenzaldehyde	588088	34036-07-2
3,5-Difluorobenzaldehyde	588160	32085-88-4
4H-naphthalen-1-one	2754230	19369-49-4
(2,3,4-trifluorophenyl)methanol	2777027	144284-24-2
2,4,5-Trifluorobenzyl alcohol	2777035	144284-25-3
3,4,5-Trifluorobenzyl alcohol	2777040	220227-37-2
2-phenylmalonaldehyde	3672296	26591-66-2
2-fluorohexan-1-ol	10441694	1786-48-7
2-Fluoroindan-1-one	11029998	700-76-5
2-(4-fluorophenyl)acetaldehyde	11126322	1736-67-0
1H-inden-1-one	11815384	480-90-0
Naphthalenone	12446728	57392-28-6
2-fluoro-2-phenylacetaldehyde	12602096	13344-76-8
8-methylidenebicyclo[4.2.0]octa-1,3,5-trien-7-one	13167180	88180-40-9

6aH-cyclopropa[a]inden-6-one	15732192	
2-(3-fluorophenyl)acetaldehyde	15811999	75321-89-0
2-(2-Fluorophenyl)Acetaldehyde	17770161	75321-85-6
3-fluorohexan-1-ol	19105682	
bicyclo[2.2.2]octa-1,3,5-trien-8-one	19743341	
3-oxo-2-phenylprop-2-enal	21258278	
4-(fluoromethyl)benzaldehyde	21407901	64747-66-6
3-(fluoromethyl)benzaldehyde	23080897	96258-62-7
hydroxy(phenyl)methanolate	23517413	
2-(2-oxoethenyl)benzaldehyde	45083582	89002-82-4
hexa-2,5-dien-1-ol	53752206	28465-64-7
2,5-Hexadienal	53799150	24058-41-1
5-fluoroinden-1-one	55266475	
3-methylidene-6-(oxomethylidene)cyclohexa-1,4-diene-1-carbaldehyde	56633662	
2-fluoropent-3-en-1-ol	57051182	
fluoro-(4-fluorophenyl)methanol	57224117	
(E)-3-Oxo-2-phenylprop-1-en-1-olate	59895713	
fluoro-(2-fluorophenyl)methanol	66718278	
bicyclo[3.2.2]nona-1(7),5,8-trien-4-one	67715125	
[3-(fluoromethyl)phenyl]methanol	68528076	
fluoro-(3-fluorophenyl)methanol	69304374	
(2,3-difluorophenyl)-fluoromethanol	70187444	
Bicyclo[4.1.0]hepta-1,3,5-triene-7-carboxaldehyde	71332736	102073-01-8
5-fluorohexan-1-ol	72823953	
4-fluoro-3H-indene-1,2-dione	83069838	
3,3-difluoro-2H-inden-1-one	83669798	
bicyclo[3.3.1]nona-1,3,5(9)-trien-6-one	87233327	
4-fluorohexan-1-ol	87401947	
4-formylbenzoyl fluoride	90160302	
2-(2,3-difluorophenyl)-2-fluoroacetaldehyde	90375715	
3-fluoro-2,3-dihydroinden-1-one	91882489	
naphthalene-1-carbaldehyde	101170232	
oxidooxy(phenyl)methanol	101334094	
5-fluoro-2-methylidene-3H-inden-1-one	101875887	
(3S)-3-(fluoromethyl)-2,3-dihydroinden-1-one	102233594	
2-(oxomethylidene)indene-1,3-dione	102578882	
2-(2,4-difluorophenyl)-2-fluoroacetaldehyde	105435719	
1-(2-ethenyl-4-fluorophenyl)ethanone	108327546	
2,4-difluoro-2,3-dihydroinden-1-one	117942772	
2-fluoro-3-methyl-2,3-dihydroinden-1-one	118515426	
1H-Inden-1-one	119092183	67864-38-4
3-fluoro-3-methyl-2H-inden-1-one	122380797	

Supplementary Table S3. Five-fold random split Support Vector Machine performance metrics. TP: true positives, TN: true negatives, FP: false positives, FN: false negatives, %CC: percentage of instances correctly classified, MCC: Matthews correlation coefficient.

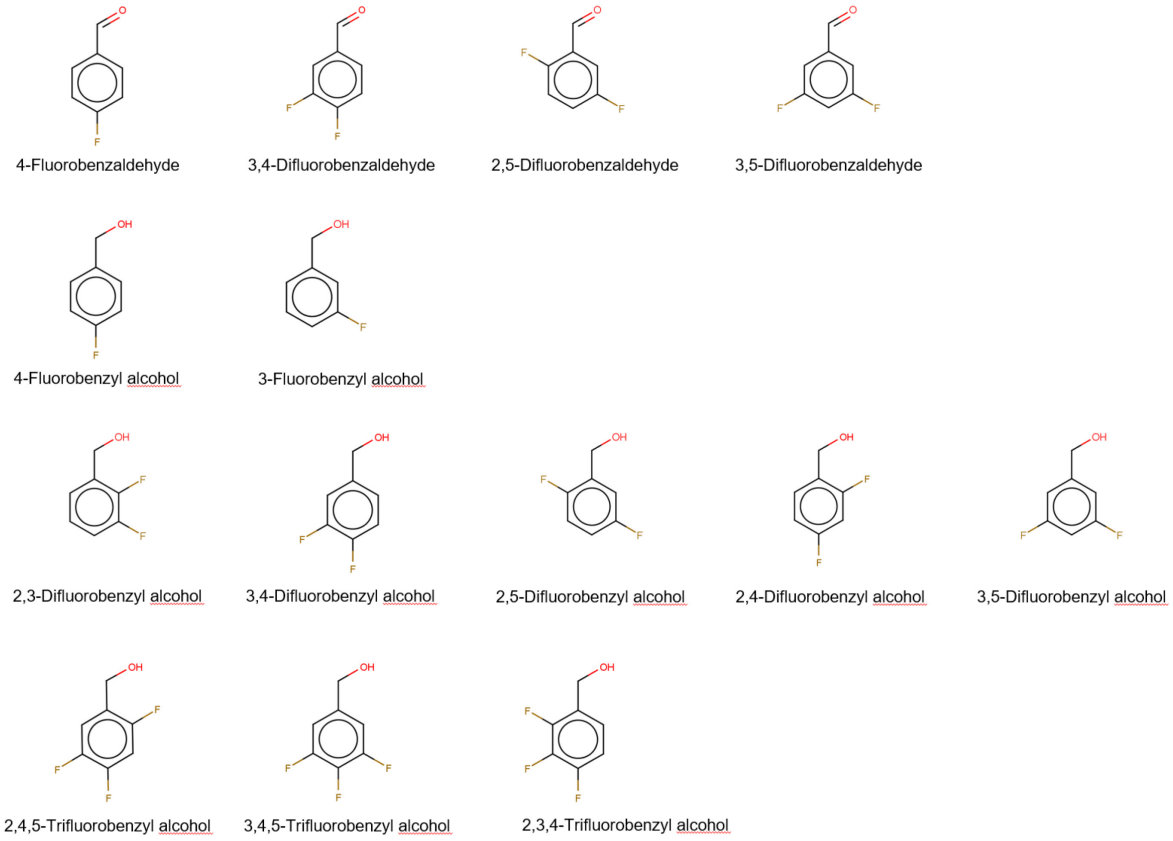
Dataset	TP	TN	FP	FN	%CC	Precision	Recall	MCC
Training	8.17±1.12	29.83±1.35	2.50±0.71	1.50±0.71	0.90±0.03	0.77±0.05	0.84±0.08	0.77±0.07
Test	3.00±0.76	6.17±0.83	0.50±0.71	0.33±0.44	0.92±0.06	0.88±0.16	0.91±0.12	0.83±0.12

The Mathews correlation coefficient (MCC) is obtained as follows:

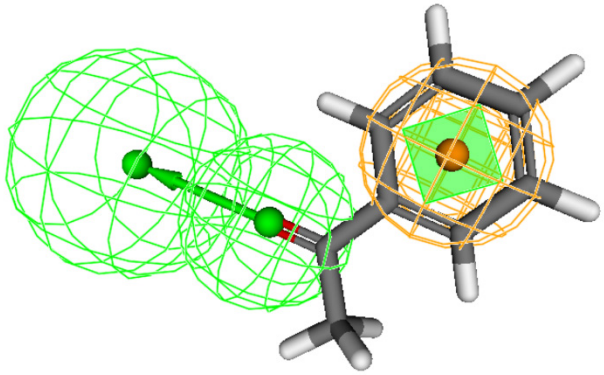
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Precision = \frac{TP}{TP + FP}$$

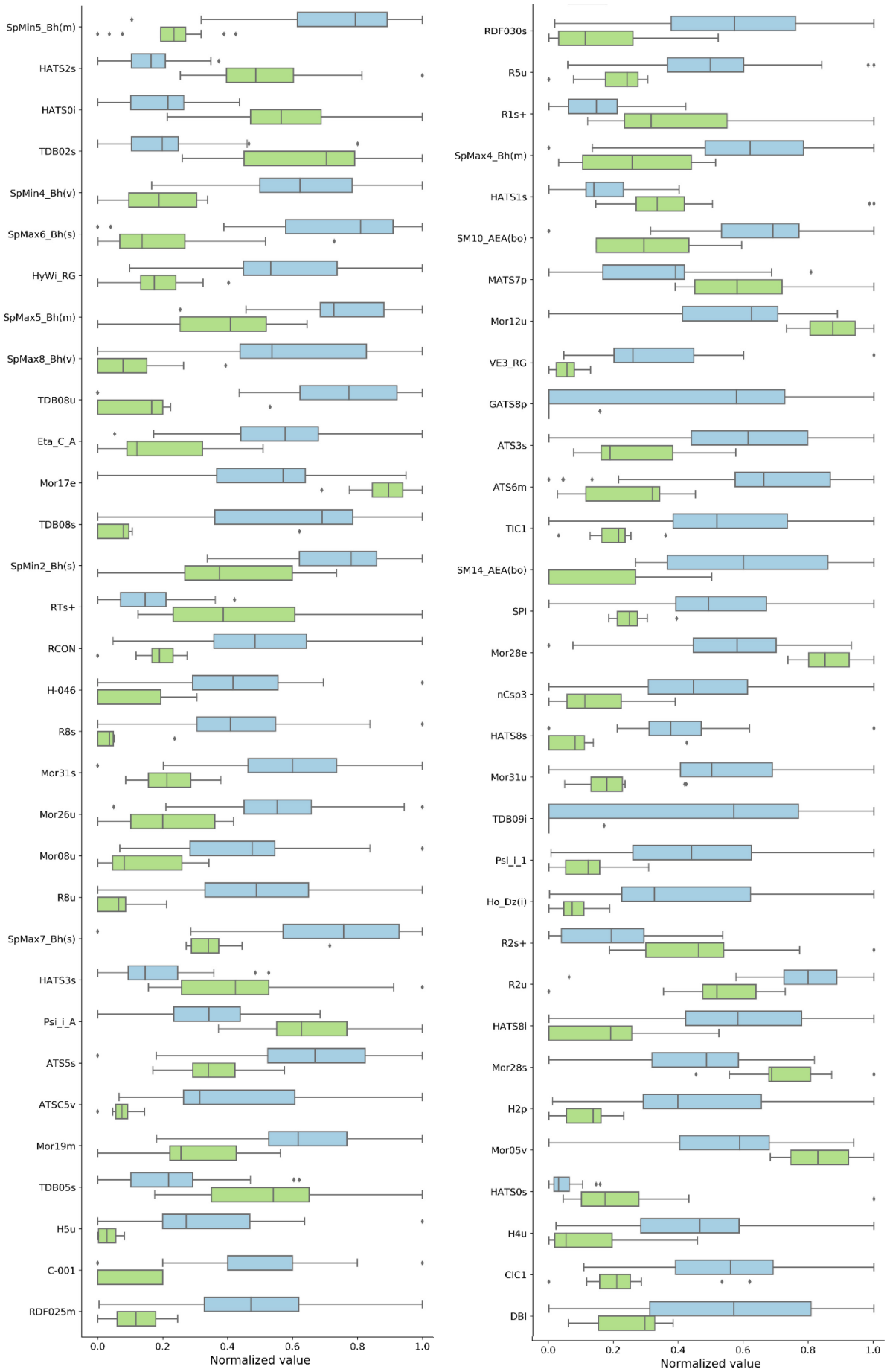
$$Recall = \frac{TP}{TP + FN}$$

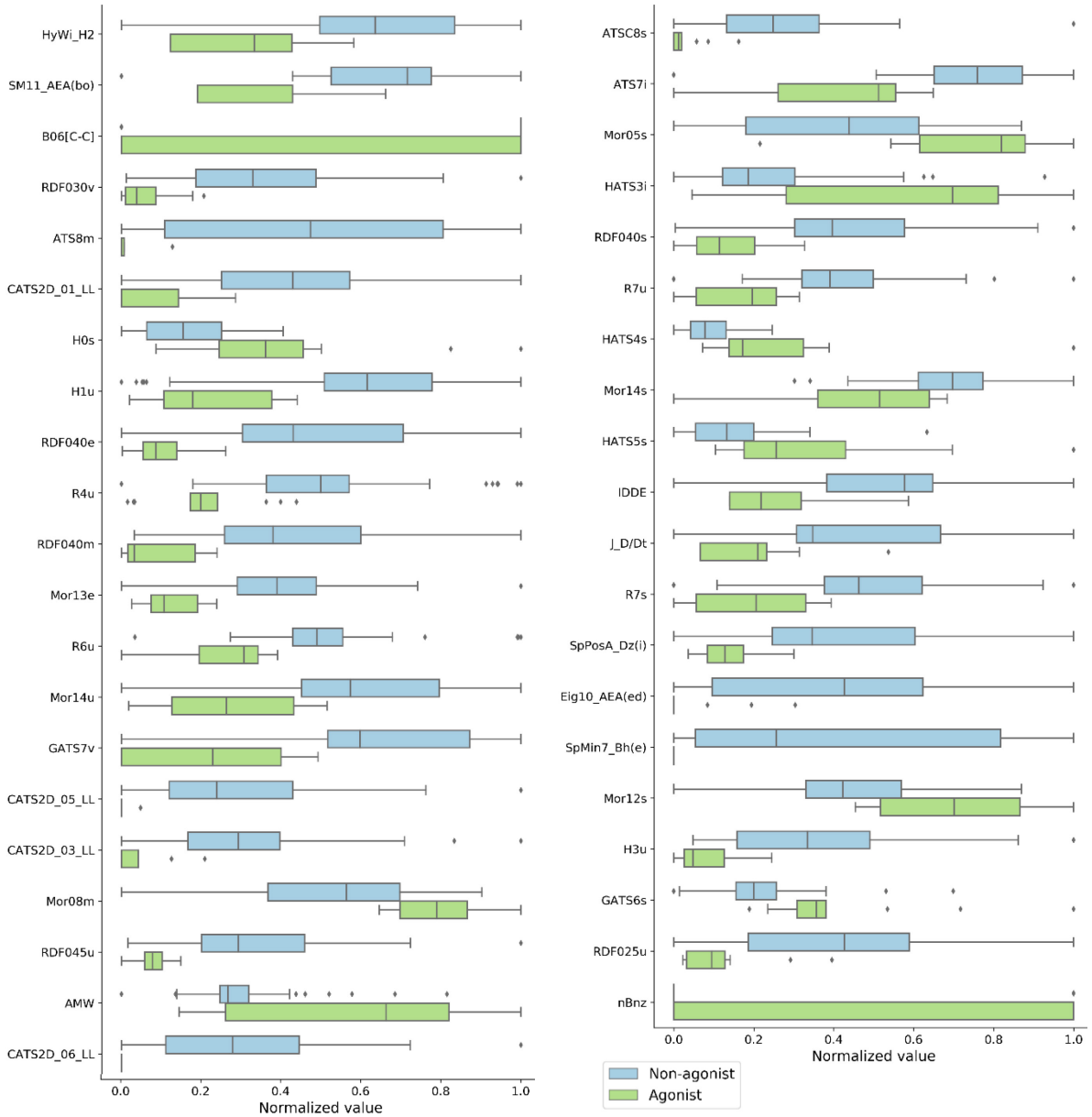


Supplementary Figure S1. Chemical structure of predicted but non-active ligands for SlitOR25.



Supplementary Figure S2. SlitOR25 pharmacophore hypothesis. The pharmacophore bears an aromatic cycle (orange sphere) and a hydrogen bond acceptor (green spheres). Acetophenone perfectly fits into this pharmacophore. Note that non-agonists also fit into the pharmacophore model, emphasizing that the model does not accurately discriminate agonists from non-agonists.


SpMin5_Bh(m)
HATS2s
HATS0i
TDB02s
SpMin4_Bh(v)
SpMax6_Bh(s)
HyWi_RG
SpMax5_Bh(m)
SpMax8_Bh(v)
TDB08u
Eta_C_A
Mor17e
TDB08s
SpMin2_Bh(s)
RTs+
RCON
H-046
R8s
Mor31s
Mor26u
Mor08u
R8u
SpMax7_Bh(s)
HATS3s
Psi_i_A
ATS5s
ATSC5v
Mor19m
TDB05s
H5u
C-001
RDF025m
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value
RDF030s
R5u
R1s+
SpMax4_Bh(m)
HATS1s
SM10_AEA(bo)
MATS7p
Mor12u
VE3_RG
GATS8p
ATS3s
ATS6m
TIC1
SM14_AEA(bo)
SPI
Mor28e
nCsp3
HATS8s
Mor31u
TDB09i
Psi_i_1
Ho_Dz(i)
R2s+
R2u
HATS8i
Mor28s
H2p
Mor05v
HATS0s
H4u
CIC1
DBI
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value



Supplementary Figure S3. Descriptors statistically significant (amongst the 394 in total) to discriminate agonists from non-agonists (t-test) are represented as boxplots. The two samples (agonists and non-agonists) were considered as independent. The Benjamini-Hochberg procedure was used to consider the false discovery rate (FDR). Using a FDR of 0.1%, the resulting critical p-value was set to 2.5E-04, which resulted in the identification of the 105 descriptors shown here. Each descriptor values were normalized between 0 and 1 for easier visualization.

t-SNE calculation:

The t-SNE implementation of the Python package scikit-learn was used with the following parameters: embedding initialization through principal component analysis (PCA) instead of random, learning rate of 300, early exaggeration of 15, perplexity of 30, and 1000 iterations. Different parameters close to the ones recommended in the package documentation were tested until compounds which are structurally similar were plotted close to each other, and dissimilar molecules were plotted distant from one another.

Publication 2

Reverse chemical ecology in a moth: machine learning on odorant receptors identifies new behaviorally active agonists

Gabriela Caballero-Vidal†, Cédric Bouysset†, Jérémy Gévar, Hayat Mbouzid, Céline Nara, Julie Delaroche, Jérôme Golebiowski, Nicolas Montagné*, Sébastien Fiorucci*, & Emmanuelle Jacquin-Joly*

Under revision

Abstract

The concept of reverse chemical ecology (exploitation of molecular knowledge for chemical ecology) has recently emerged in conservation biology and human health. Here, we extend this concept to crop protection. Targeting odorant receptors from a crop pest insect, the noctuid moth *Spodoptera littoralis*, we demonstrate that reverse chemical ecology has the potential to accelerate the discovery of novel crop pest insect attractants and repellents. Using machine learning, we first predicted novel natural ligands for two odorant receptors, SlitOR24 and 25. Then, electrophysiological validation proved *in silico* predictions to be highly sensitive, as 93% and 67% of predicted agonists triggered a response in *Drosophila* olfactory neurons expressing SlitOR24 and SlitOR25, respectively, despite a lack of specificity. Last, when tested in Y-maze behavioral assays, the most active novel ligands of the receptors were attractive to caterpillars. This work provides a template for rational design of new eco-friendly semiochemicals to manage crop pest populations.

Keywords

Semiochemicals, insects, *Spodoptera littoralis*, behavior, crop protection

Abbreviations

AUROC: Area Under the Receiver Operating Characteristics curve

kNN: k-nearest neighbors

LOO: leave-one-out

MCC: Matthews correlation coefficient

OR: odorant receptor

OSN: olfactory sensory neuron

QSAR: quantitative-structure-activity relationship

SSR: single sensillum recordings

SVC: Support Vector Classifier

SVM: Support Vector Machine

Introduction

Insects detect and use odorant information from the external environment to make important decisions, such as selecting a mating partner, a food source or an oviposition site [1]. Depending on the species ecology, odorant signals can repel or attract insects, or do nothing. Among behaviorally relevant molecules, one can cite the moth sex pheromones that attract males from some distance away. Because of such olfactory-triggered behaviors, odorant molecules have been exploited to develop control strategies against insect pests and disease vector populations [2-4] that are integrated in combination with other strategies in Integrated Pest Management. For instance, synthetic moth sex pheromones have been used for decades for population monitoring or mating disruption [2], aggregation pheromones and/or host plant volatiles are used for mass trapping, and non-host or toxic odorants are used as repellents. However, the identification of such active molecules is usually difficult, because it mainly relies on bioassay-guided approaches, including fastidious behavioral assays on multiple individuals.

In this context, reverse chemical ecology has recently emerged as a powerful alternative to identify relevant signals for a given species. This approach proposes to screen olfactory proteins linked to a particular behavior in order to identify putative behaviorally active semiochemicals [5]. It has been promoted by the recent advances in our understanding of the molecular basis of insect olfaction in the last two decades, especially the discovery of their odorant receptors (ORs)[6-8]. These ORs are transmembrane proteins primarily responsible for odorant detection. They are expressed in olfactory sensory neurons (OSNs) housed in olfactory sensilla, located mainly on the antennae. ORs form ion channels together with a subunit called Orco (OR coreceptor) that is highly conserved across insect species [9-11]. Odorants activate the corresponding OR-Orco complex that transforms the chemical signal into an electrical signal that is transmitted to the brain, leading to the behavioral response [12]. Identifying molecules that will be active on target ORs remains difficult [4], but ligand-based *in silico* strategies relying on the chemical structures of active compounds have proven quite effective for virtual screening of ORs. Quantitative-structure-activity relationship (QSAR) models, which have been widely used in medical chemistry [13, 14], have been applied with success to predict the activity of semiochemicals on ORs from model insects such as *Drosophila melanogaster* [15] and the mosquitoes *Aedes aegypti* and *Anopheles gambiae* [16-19].

In the present study, we used QSAR models to predict ligands for ORs from a non-model insect species, the crop pest moth *Spodoptera littoralis*, revealing it is possible to use machine learning to identify OR agonists outside Diptera [20]. We have previously identified ligands for a large

number of *S. littoralis* ORs (hereafter SlitOR) using heterologous expression in the empty neuron system [21]. Moreover, behavioral assays have shown that *S. littoralis* caterpillars are attracted by plant volatiles that activate SlitOR24 and SlitOR25 [22]. With the final aim of identifying new attractive semiochemicals for *S. littoralis* larvae, we thus prioritized these two ORs that presented a large overlapping receptive range, including aromatic compounds and green leaf volatiles. We virtually screened a judiciously selected natural product library to identify novel ligands. This led to success rates of 67% and – even more impressively – 93% active molecules on SlitOR25 and SlitOR24, respectively.Finally, we conducted behavioral experiments to investigate the activity of the most potent agonists of SlitOR24 and SlitOR25. This work, combining machine learning, electrophysiological analyses and behavioral assays, not only expands the list of natural SlitOR ligands but also successfully identifies new larval attractants that can potentially be implemented in eco-friendly control strategies. Whereas the concept of reverse chemical ecology has been successfully applied in conservation biology (targeting endangered species [23]) and human health (targeting disease vectors [5]), our work now demonstrates its great potential in agriculture.

Materials and Methods

Insects

S. littoralis larvae were reared on a semi-artificial diet [24] under the following conditions: 22°C, 60% relative humidity and 16:8-h light: dark cycle. Fourth-instar larvae (L4) were used for behavioral assays.

Transgenic *D. melanogaster* flies expressing SlitOR24 and 25 were obtained by crossing the lines *w;Δhalo/CyO;UAS-SlitOR24* and *w;Δhalo/CyO;UAS-SlitOR25* [21] with the line *w;Δhalo/CyO;OR22a-Gal4* [25]. Flies were reared on standard nutrient medium made of cornmeal, yeast and agar. Flies were kept at 25 °C, under a 12:12-h light: dark cycle.

Modeling

Datasets

The SlitOR24 QSAR model was built using the dataset of 51 experimentally tested molecules (10 actives, 41 inactives) from [21]. The SlitOR25 model was built using the same dataset enriched with 32 molecules experimentally tested in [20], resulting in a dataset of 83 molecules labelled as active (25 molecules) or inactive (58 molecules) against SlitOR25. An in-house

library of 158 plant volatile organic compounds (Online Resource 1) was screened by the two numerical models. All molecules were collected as SMILES strings, the major tautomers at pH 7.0 were retrieved with cxcalc (Calculator Plugins, Marvin 18.3.0, 2018, ChemAxon), and the resulting molecules were standardized with the standardizer python package v0.1.7 (for salt removal and structure normalization). Molecular descriptors were computed directly from the standardized SMILES using Dragon v6.0.38. Feature exclusion was performed within the software based on the following criteria: constant or near-constant descriptors, descriptors with at least one missing value and highly correlated descriptors (absolute pair correlation greater than or equal to 0.95 for SlitOR25 and 0.9 for SlitOR24) were excluded. This resulted in libraries of 288 and 493 descriptors for SlitOR24 and SlitOR25, respectively.

The SlitOR24 and SlitOR25 datasets (Online Resource 2) were split in training and test sets using a common clustering method, the sphere-exclusion approach, which can select a diverse subset of compounds in a dataset. For both sets, descriptors were normalized between 0 and 1, and the split was initialized by putting in the test set the compound closest to the center of the normalized dataset. At each iteration the new compound to be added to the test set was selected using a MinMax procedure, the dissimilarity radius to exclude compounds from the test set was set to 4.8 for SlitOR24 and 4.0 for SlitOR25, and the algorithm was stopped once the test set reached 24% of the size of the original dataset. This resulted in training sets of 39 molecules (8 actives, 31 inactives) and 64 molecules (18 actives, 46 inactives), and test sets of 12 molecules (2 actives, 10 inactives) and 19 molecules (7 actives, 12 inactives) for SlitOR24 and SlitOR25, respectively. For both datasets, each descriptor was then denormalized and normalized only based on the training set min and max values. To quantify the uncertainty of prediction resulting from the initial choice of compounds in the training and test sets, five alternative splits were generated using the same strategy. The same sphere-exclusion approach was used to define the new training/test sets with initial compounds chosen randomly and not at the center of the normalized distribution as for the final model. Due to imbalanced data (less active than inactive compounds) and to facilitate comparison, only the first five splits that had the same activity distribution (active/inactive) as in the split used for the final model were investigated.

Machine-Learning

QSAR models were trained and evaluated using Weka v3.8.2 [26]. Several classification algorithms were optimized in “leave-one-out” (LOO) cross-validation loops: C-SVC (LibSVM v1.0.10) (SVC: Support Vector Classifier; SVM: Support Vector Machine), k-nearest neighbors

(kNN), RandomTree, DecisionTree, and RandomForest. Cost-sensitive models were also tested without providing a significant improvement in performance. Once the optimal algorithm and hyperparameters were identified for each OR based on Matthews correlation coefficient (MCC) (Table 1), the final models were trained on the full training set and parametrized as follow: for SlitOR24 a RandomForest was chosen and trained with 100 trees, unlimited maximum depth for each tree, and no feature randomly chosen; for SlitOR25 a kNN classifier (IBk) was chosen with 9 neighbors, weighted by the inverse of the Euclidean distance, and a brute force neighbor search. Finally, the performances of the SlitOR24 and SlitOR25 models were assessed on the test sets.

Applicability Domain

A similarity distance approach [27] was used to estimate the applicability domain of the two selected models. A distance cutoff is defined as $Dc = \langle D \rangle + Z\sigma$ where $\langle D \rangle$ and σ are the mean and standard deviation of Euclidean distances of each training set compound with their nearest neighbor in the descriptor space, and Z is an empirical parameter. The parameter Z was incremented until all training set compounds had their distance with their kNN lower or equal to Dc. For SlitOR25, we kept the same number of neighbors as in the model (k=9) and for SlitOR24, we used k=6 based on our benchmark of different learners during the training phase. For each external compound, its distance with the kNN was measured and a reliability score was estimated as $reliability = 1 + \frac{D-Dc}{Dc}$.

Single sensillum recordings on neurons expressing SlitOR24 and SlitOR25

Single sensillum recordings were performed on *Drosophila* ab3A neurons expressing SlitOR24 or SlitOR25, using fly lines previously generated [21]. A 2 to 8-day-old fly was immobilized in a pipette tip, only the head sticking out. The fly was placed on a microscope glass slide under a constant 1.5 $L.min^{-1}$ flux of charcoal-filtered and humidified air delivered through a glass tube of a 7 mm diameter. The experiments were monitored using a light microscope (Olympus BX51WI, Tokyo, Japan) equipped with a 100X magnification objective. Action potentials from ab3A OSNs were recorded using electrolytically sharpened tungsten electrodes (TW5-6, Science Products, Hofheim, Germany). One reference electrode was inserted into the eye and the recording electrode was inserted at the base of an ab3 sensillum using a motor-controlled PatchStar micromanipulator (Scientifica, Uckfield, United Kingdom). Odorants were purchased from Sigma-Aldrich (Saint-Louis, MO, USA). Stimulus cartridges were built by

placing a 1 cm² filter paper in a Pasteur pipette and loading 10 µl of the odorant solution onto the paper (10^{-2} dilution in paraffin oil), or 10 µL of paraffin oil or a paper without any odorant as controls. Odorant stimulations were performed by inserting the tip of the pipette into a hole in the glass tube and generating a 500 ms air pulse (0.6 $L.min^{-1}$). The responses of ab3A OSNs were calculated by subtracting the spontaneous firing rate (in $spikes.s^{-1}$) from the firing rate during the odorant stimulation.

The stimulation panel consisted, for each SlitOR, of an already known agonist [21] used as positive control (benzyl alcohol for SlitOR24 and acetophenone for SlitOR25), paraffin oil as a negative control, 34 predicted agonists and 5 molecules randomly chosen among the predicted non-agonists for both ORs (Online Resource 3). Each stimulus cartridge was used at maximum eight times in total. The panel of molecules was tested on five (for predicted non-agonists) to eight-ten (for predicted agonists) different flies expressing SlitOR24 or SlitOR25. Odorants were considered as active if the response was statistically different from the response elicited by the solvent alone (Kruskal–Wallis test followed by a Dunnett multiple comparison test, $p<0.05$).

Larvae behavior in Y-tube olfactometer

Behavioral experiments were performed in a Y-tube olfactometer. The olfactometer consisted of a 2.1 cm inner diameter glass Y-tube, the main segment was 13 cm long, and each of the two arms was 9.5 cm long. L4 larvae were used and starved overnight (16 to 20 hours starvation) prior to the experiments. All experiments were performed under red light, to avoid biases due to visual cues. Charcoal-purified air was delivered into each arm of the olfactometer at a flow rate of 0.5 $L.min^{-1}$, stabilized using a flowmeter (Key Instruments, Trevose, PA, USA) to ensure that equal air streams entered each arm. The temperature of the room was maintained at 24 ° C during all tests. The experimental set-up was first tested with different controls: i) paraffin oil in each arm, a configuration expected to induce no larval choice, ii) a 10^{-2} dilution of benzyl alcohol, an odorant known to induce larvae attraction [22], in one arm and paraffin oil in the other arm (larval choice expected) and iii) a 10^{-2} dilution of (*E*)-ocimene, a molecule inactive on larval behavior [22], versus paraffin oil (no larval choice expected). Seven of the strongest agonists of both SlitOR24 and 25 were tested for behavioral activity. Odorants were diluted in paraffin oil (dilutions 10^{-2} and 10^{-3}). Ten µl of diluted odorants or control (paraffin oil) were loaded on a filter paper. A paper with solvent alone was placed in one arm and a filter paper with the odorant dilution in the other arm. One larva at a time was placed in the main arm of

the olfactometer and the behavior was recorded during 10 minutes with a digital camera located above the device. Each larva was tested only once. To avoid any bias during the test, the olfactometer was switched from one side to the other between each test and up to three times, before washing the olfactometer with TFD4 detergent (Franklab, Montigny-le-Bretonneux, France) diluted at 3% for 30 minutes, then rinsing with distilled water and 95% ethanol. Once dry, all glass parts were put in an oven at 200 °C overnight. We analyzed two different parameters: 1) the choice made by the caterpillar and 2) the time spent in each arm. We considered that the caterpillar made a choice when three quarters of its body length entered an arm. Larvae that did not make a choice within ten minutes were not included in the statistical analysis. This explains the variable numbers of replicates for each test, ranging from 27 to 34. All behavioral assays were carried out within a 4 h time interval during larvae photophase.

Statistics

Single sensillum recording data were analyzed using a Kruskal–Wallis test followed by nonparametric multiple comparisons using 'nparcomp' R package (type: Dunnett). For behavioral data, a Chi-squared test for given probabilities was used to verify the significance of caterpillars' choice and a paired Student t-test was used to compare the time spent by larvae in each arm of the Y-tube olfactometer.

Results

Virtual screening of SlitOR24 and SlitOR25

Model performance

Each SlitOR model was parameterized with a LOO strategy, re-trained on the full training set once the best parameters were identified, and validated using the independent test set (Table 1). For SlitOR24, due to the limited number of active molecules in the training set, the model appeared to be mostly tuned to classify correctly the non-agonists. For SlitOR25, the model came with the benefit of an expanded applicability domain. However, the decrease in performance on the test set, compared to a previous preliminary model we conducted on this OR [20], is probably linked to the increased chemical diversity and thus to the complexity of the problem. Overall, both current SlitOR24 and SlitOR25 models had satisfying predictive abilities with MCC $\geq$ 0.4, and AUROC (Area Under the Receiver Operating Characteristics curve) $\geq$ 0.8, and were suitable to prioritize compounds for experimental testing.

Table 1. Performance evaluation of the SlitOR24 and SlitOR25 QSAR models using different metrics. LOO: performance of the best model using a leave-one-out cross-validation strategy, TP: true positives, TN: true negatives, FP: false positives, FN: false negatives, FPR: false positive rate, MCC: Matthews correlation coefficient, AUROC: area under the receiver operating characteristics curve.

	Dataset	TP	TN	FP	FN	Accuracy	Precision	Recall	FPR	MCC	AUROC
SlitOR24	LOO	4	29	2	4	0.85	0.67	0.50	0.06	0.49	0.83
	Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	0.99
	Test	1	9	1	1	0.83	0.50	0.50	0.10	0.40	0.80
SlitOR25	LOO	15	34	12	3	0.77	0.56	0.83	0.26	0.52	0.84
	Training	18	46	0	0	1.00	1.00	1.00	0.00	1.00	1.00
	Test	5	10	2	2	0.79	0.71	0.71	0.17	0.55	0.89

To estimate the generalization error, five similar models were generated using alternative splits for preparing the training and test sets (Online Resources 4 and 5). When changing the distribution of compounds in the training and test sets, the overall performance of the predictive models remained similar compared to the final model. In details, the MCC was still above 0.4 and the AUROC ranged from 0.7 to 0.9 for both SlitOR24 and 25 models except for two alternate SlitOR24 models. For these two, no true positive compound was identified, mostly due to the small size of the dataset. One has to note that the false positive rate (i.e. the number of false positive prediction over the total number of inactive compounds) was higher for SlitOR25 models (0.17-0.50) than for SlitOR24 ones (0.00-0.10) and may lead to incorrectly classify non agonists and overestimate the number of compounds to be experimentally tested.

The current SlitOR25 and SlitOR24 machine learning models were used to virtually screen an in-house library of 158 natural volatile organic compounds. 28 and 67 molecules were predicted as agonists and within the applicability domain of SlitOR24 and SlitOR25 models, respectively, with 27 molecules in common (Online Resource 3). The 67 molecules predicted as agonists for SlitOR25 were re-screened by our previously published model [20] and 20 of them were predicted as agonists by both SlitOR25 models (Online Resource 3).

Electrophysiological responses of SlitOR24 and SlitOR25 to the predicted agonists and non-agonists

To validate *in silico* predictions, we performed single sensillum recordings on *Drosophila* OSNs expressing SlitOR24 or SlitOR25 with a stimulus panel containing the 27 molecules predicted as agonists for both SlitOR24 and SlitOR25, 6 molecules predicted as agonists only for SlitOR25 by both the current and the published SlitOR25 models, and one molecule predicted as an agonist only for SlitOR24. We also tested five molecules predicted as non-agonists for both receptors (Online Resource 3) and one already known agonist for each OR as control [21]. In total, we tested 39 molecules on both receptors (28+11 and 33+6 predicted agonists+non-agonists for SlitOR24 and SlitOR25, respectively). As expected, both ORs responded to their respective positive control (Fig. 1).

For SlitOR24, 26 predicted agonists out of 28 were active (Fig. 1A), representing a 93% success rate of prediction. Among the six agonists predicted only for SlitOR25, four were active on SlitOR24 although they were not predicted as agonists by the model. Six molecules from the panel triggered responses above 100 spikes.s^{-1} (1-pentanol, (*Z*)-2-hexenol, 2-hexanol, (*E*)-3-hexenol, 2-heptanol and 2-phenylethanol, the latter eliciting the highest response), thus being as active as the previously identified agonist benzyl alcohol. Eight agonists triggered responses between 50 and 100 spikes.s^{-1} (1-hexen-3-ol, 2-hexanone, benzyl cyanide, 3-heptanone, 2-heptanone, furfuryl alcohol , 4-methyl-2-pentanol and heptanal).

For SlitOR25, 22 out of the selected 33 predicted agonists were active, representing a 67% success rate (Fig. 1B). As expected, the agonist predicted only for SlitOR24 did not elicit any SlitOR25 response. Two molecules from the panel triggered responses above 100 spikes.s^{-1} (2-heptanol and benzyl cyanide) and were more active than the previously identified agonist acetophenone, and six triggered responses between 50 and 100 spikes.s^{-1} (2-phenylethanol, (*E*)-3-hexenol, heptanal, 1-hexen-3-ol, 3-heptanone, 2-heptanone). None of the six non-agonists predicted for SlitOR25 elicited a significant response. In short, with a recall of 0.84 vs 1.00, both models were highly sensitive, even if the SlitOR25 model was less precise (0.93 vs 0.67) and specific (0.75 vs 0.35) than the SlitOR24 one, as expected by the evaluation metrics of the trained models (Online Resource 6).

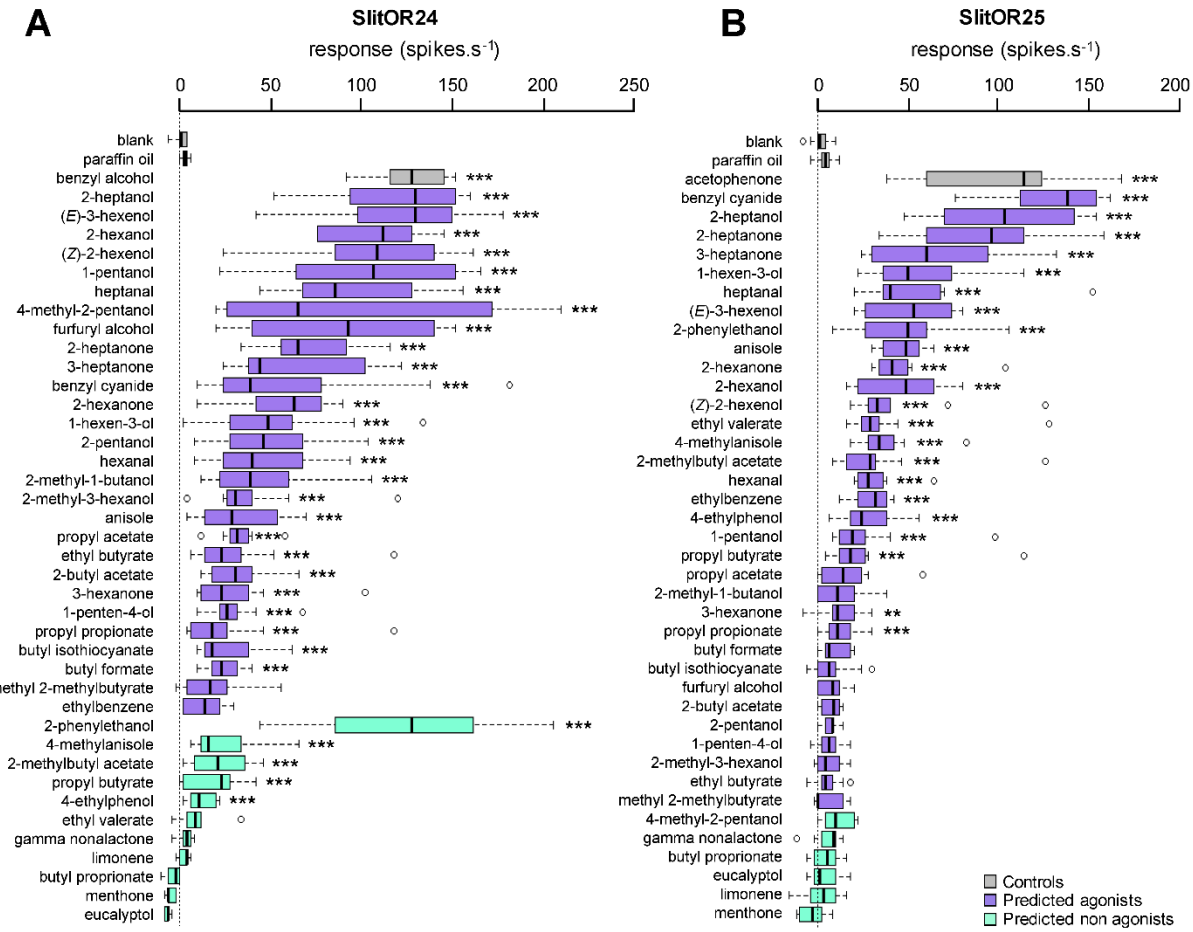


Fig. 1 Responses of SlitOR24 and SlitOR25 to predicted ligands. Single-sensillum recording (SSR) responses (spikes.s^{-1}) of *Drosophila* ab3A neurons expressing SlitOR24 (**A**) and SlitOR25 (**B**) during stimulation with QSAR model-predicted ligands. Grey bars represent negative controls (solvent and filter paper without odorant) and positive controls (known ligands for the respective OR [21]). Purple bars represent predicted agonists. Turquoise bars represent predicted non-agonists. All molecules were tested at a 10^{-2} dilution in paraffin oil. Box plots show the median (line), 25–75% percentiles (box), 10–90% percentiles (whisker), and outliers (dots). Asterisks indicate statistically significant differences between responses to the odorant and to the solvent alone (Kruskal–Wallis non parametric ANOVA followed by a Dunnett's multiple comparison test, ** $p<0.01$, *** $p<0.001$, n=8-10 for predicted agonists, n=5 for non-agonists)

Behavioral effect of newly identified agonists

The newly identified OR agonists were then tested for their effect on larvae behavior. In all behavioral experiments, larvae were starved for 16 to 20 hours since previous experiments have shown that starved larvae are more motivated to orientate toward odor sources than satiated larvae and that such starvation has no impact in larval survival or mobility [28]. Before testing

the effect of SlitOR24 and SlitOR25 ligands on the larval behavior, the experimental setup was first validated using different controls: paraffin oil (solvent), benzyl alcohol (known attractant) and (*E*)-ocimene (neutral) at dilution 10^{-2} [22]. As expected, larvae did not make any choice when exposed to both arms loaded with solvent. Larvae were statistically more attracted to the arm containing benzyl alcohol than to the control arm whereas no choice was observed using (*E*)-ocimene (Fig. 2). Then, seven of the molecules that elicited the highest neuronal responses in flies expressing SlitOR24 and SlitOR25 [(*Z*)-2-hexenol, (*E*)-3-hexenol, 2-phenylethanol, benzyl cyanide, 2-heptanol, anisole and 2-hexanone] were used in the same behavioral assay at two different dilutions (10^{-2} and 10^{-3}).

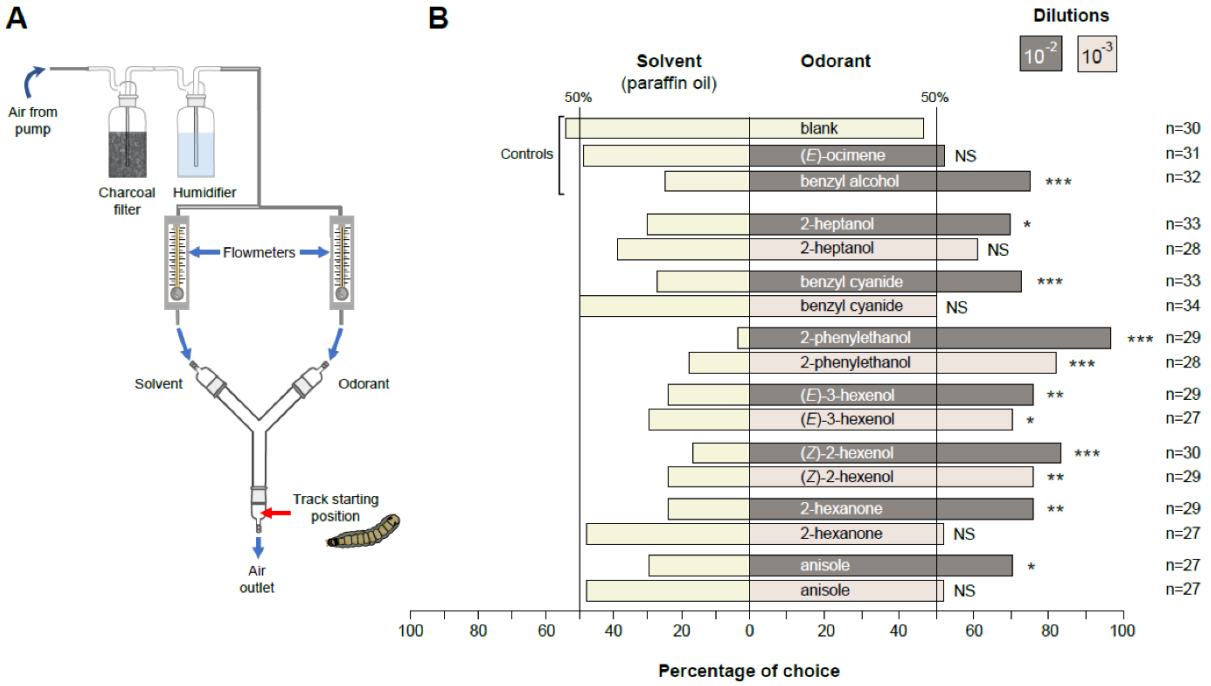


Fig. 2 Behavioral responses (percentage of choice) of *S. littoralis* larvae to predicted ligands shown to be active on SlitOR24 and SlitOR25. (A) Experimental setup used to study caterpillar's behavior. In this device, there is an air inlet, which circulates through two filters (active carbon and water bubbles), from where it passes to two flowmeters, to finally reach the Y-tube olfactometer. At the base of the olfactometer, the air outlet and the starting point for the larva are indicated. **(B)** Percentage of larval choice to (left/right): blank/blank (paraffin oil), neutral control (paraffin oil/ocimene), positive control (paraffin oil/benzyl alcohol), active ligands on ORs (paraffin oil/compounds). Dark grey bars at right represent the caterpillar's choice at 10^{-2} dilution, and light grey bars represent caterpillar's choice at 10^{-3} dilution. Asterisks indicate statistically significant preferences of larvae for the odorant side (Chi-squared test for given probabilities, * $p<0.05$, ** $p<0.01$, *** $p<0.001$, NS: not significant). Numbers of replicates (n) are indicated on the right.

Results showed that for the 10^{-2} dilution, all molecules tested were attractive to the larvae (Fig. 2), with percentages of choice between 69.6% (2-heptanol) and 96.5% (2-phenylethanol). At the 10^{-3} dilution, larvae retained preference for three compounds: (*Z*)-2-hexenol, (*E*)-3-hexenol and 2-phenylethanol. Regarding the time spent in each arm (Fig. 3), larvae spent significantly more time in the arm containing five out of the seven molecules when tested at the 10^{-2} dilution: benzyl cyanide, (*Z*)-2-hexenol, 2-phenylethanol, anisole and 2-hexanone. At the 10^{-3} dilution, larvae spent more time in the arm containing three molecules: (*E*)-3-hexenol, 2-phenylethanol and 2-hexanone. Strikingly, the time spent by larvae on the arm containing (*E*)-3-hexenol was higher at the lowest dilution.

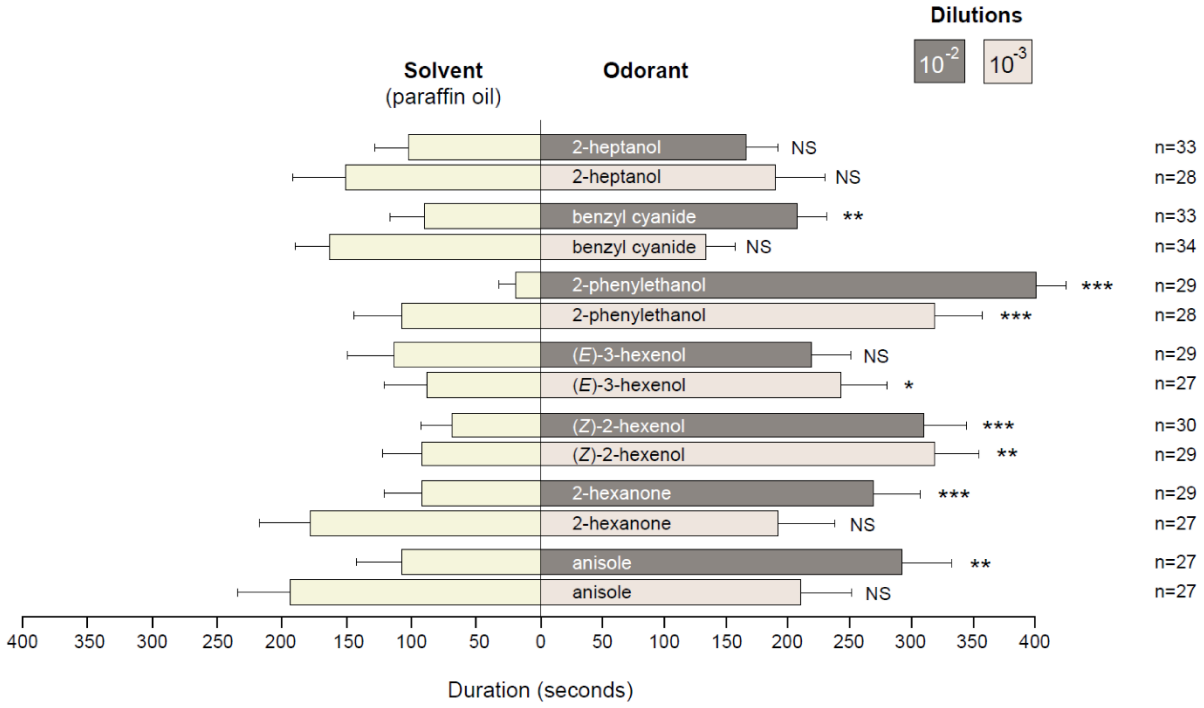


Fig. 3 Behavioral responses (time in each arm) of *S. littoralis* larvae to predicted ligands shown to be active on SlitOR24 and SlitOR25. Time (in seconds) spent by the larvae in each arm on the Y-tube olfactometer. Bars at left represent the time spent in the arm containing the solvent (paraffin oil). Bars at right: Dark grey bars represent the time spent in the arm containing the odorant at 10^{-2} dilution, and light grey bars represent the time spent in the arm containing the odorant at 10^{-3} dilution. Asterisks indicate statistically significant differences between the time spent by larvae in each arm (Paired Student t-test, * $p<0.05$, ** $p<0.01$, *** $p<0.001$, NS: not significant). Numbers of replicates are indicated on the right and error bars indicate SEM.

Discussion

Reverse chemical ecology has recently appeared as a promising approach to identify behaviorally active semiochemicals that could be used for pest control strategies. In *Helicoverpa armigera* caterpillars, a combination of transcriptomic analyses, functional characterization of ORs and behavioral assays led to the identification of OR ligands that are behaviorally active (attractive and repulsive) for first-instar larvae [29]. A link between the activation of some ORs and attraction was also demonstrated in another species of pest caterpillars, the cotton leafworm *S. littoralis* [22]. These works thus showed that caterpillar ORs have a great potential as targets in reverse chemical ecology, yet the chances to identify behaviorally active molecules remain limited by the number of molecules tested on the target ORs. The incorporation of *in silico* modeling to the functional studies could fill this gap, since it has proven efficient when applied to the identification of new mosquito repellents [5, 18, 19]. Recently, we have published a proof-of-concept that revealed that such an approach can be extended to crop pest ORs [20]. Focusing on a single *S. littoralis* OR, SlitOR25, we could predict new agonists via machine learning that were indeed active on this OR, with a reasonable success rate of 28%. However, we did not investigate their behavioral activity. Anyhow, the chemical structures of the newly identified SlitOR25 agonists precluded their use for pest control, as most agonists were fluorinated compounds that cannot be used in the field [20].

In the present work, the objective was threefold. The first one was to improve our machine learning model for the prediction of agonists. The second objective was to predict natural, plant derivate, non-toxic and affordable agonists that would be compatible with pest control. The last objective was to investigate the behavioral activity of predicted agonists. To reach these objectives, we focused on the broadly tuned receptors SlitOR24 and SlitOR25 [21], whose activation has been linked to larvae attraction [22] and that were thus highly relevant for a reverse chemical ecology strategy. More, SlitOR25 has been used to establish the machine learning proof-of-concept on Lepidoptera ORs [20], and the data acquired (additional ligands) are perfectly suited to be used for model improvement.

First, we revealed that the QSAR models are highly precise since 67% and 93% of predicted agonists triggered a response in *Drosophila* olfactory neurons expressing SlitOR25 and SlitOR24, respectively. Even if the models lack specificity, notably for SlitOR25, they were sufficiently accurate to predict many new agonists. The SlitOR24 success rate was notably higher than what has been reported previously for Diptera. In *Drosophila*, more than 240,000 compounds were first screened *in silico* to find new OR ligands [15]. OR-optimized descriptors

allowed to rank the untested molecules, identifying the top 500 hits for each OR. Predicted compounds were experimentally tested on nine ORs, showing 71% of success rate compared to only 10% when using non-predicted odors [15]. In mosquitoes, Tauxe et al. 2013 obtained a ~30% success rate when trying to identify CO_2 receptor activators by using molecular descriptors [18]. In a second round of *in silico* prediction, they increased prediction accuracy through a SVM-based approach, yielding an improved success rate of 74%. In the present work, while the SlitOR24 model appeared exceptionally precise to identify true agonists (93%), it has to be noticed that it did miss some of them. Some of the molecules predicted as agonists only for SlitOR25 appeared to be agonists for SlitOR24 (false negative rate of 18%). Reversely, the SlitOR25 model was less efficient to identify true agonists (precision of 67% and a false positive rate of 65%), but was highly sensitive and succeeded in predicting all the non-agonists. More, combining the SlitOR25 model with the previously published one [20] (Online Resource 3) guided us to prioritize the most promising compounds. As already reported in mosquitoes [18], such results suggest that model combination, in addition to cumulative experimental data to feed models, offer a way to improve insect OR ligand identification.

One has to keep in mind that our models are based on experimental data obtained from ORs expressed in the empty neuron system of *Drosophila*, which lacks perireceptor proteins such as odorant-binding proteins and odorant-degrading enzymes [12]. We cannot rule out that response spectra of caterpillar ORs expressed in a fly neuron may somehow differ from the response of the corresponding caterpillar neurons, leading to a potential confounding effect on the modeling. However, we have previously shown that, when expressed in the empty neuron system, SlitOR24 and SlitOR25 exhibit exactly the same response spectrum than the two corresponding olfactory neurons from *S. littoralis* adult antennae (see Supplementary Figure S3 in [21]). Thus, we can be confident in the use of models based on empty neuron SSR data for identifying molecules active on *S. littoralis* caterpillars.

To reach the second objective, the QSARs have been used here to screen an in-house virtual library of plant compounds, while our previous efforts focused on a large subset of the Pubchem database selected on physico-chemical properties that led to the identification of structurally related, mainly fluorinated, predicted ligands [20]. Through this approach, we have identified new agonists for SlitOR24 and SlitOR25 (more or equally active as previously identified ligands), greatly extending their initially described response spectra [21]. Both ORs presented a large overlapping receptive range, including aliphatic alcohols, aromatic compounds and green leaf volatiles. Interestingly, a large majority (74%) of predicted ligands for SlitOR25

were also active on SlitOR24. This suggests that the binding pocket of both ORs would be quite similar and opens up further studies on structure-function relationships. The tridimensional structure of an insect (*Machilis hrabei*) OR has been recently elucidated [30] and even if the sequence identity between MhraOR5 and SlitOR24 or 25 is low (<20%), one can try to extrapolate the corresponding binding pockets. Interestingly, in line with the experimental data, a multiple sequence alignment suggests that the residues from the putative odorant-binding sites of SlitOR24 and 25 are highly conserved (Online Resource 7).

The behavioral effects of the new ligands that elicited high neuronal responses were investigated on larvae, and all proved to be attractive. These data not only confirmed the former hypothesis that SlitOR24 and OR25 activation is linked to larval attraction [22], but also demonstrated that reverse chemical ecology is efficient in predicting behaviorally active odorants. Interestingly, many of these new attractants for *S. littoralis* larvae have never been reported to be relevant cues for adults or larvae on this species. Among the new ligands for SlitOR25, benzyl cyanide (a nitrogenous aromatic compound) induced the highest OSN firing rate and a high attraction rate. It has been shown previously that benzyl cyanide is a herbivore-induced volatile emitted by diverse plants, like the black poplar *Populus nigra* and Brussels sprouts *Brassicae oleracea* [31, 32]. One the one hand, such signal indicates actual presence of herbivores, and thus the possible presence of adequate food for larvae. On the other hand, benzyl cyanide has also been reported to be attractive to different parasitoid species that use this cue to detect the presence of host larvae [31, 32]. Benzyl cyanide is also naturally emitted by some insect species, and is notably known as a male anti-aphrodisiac pheromone in the desert locust [33] as well as in the butterfly *Pieris brassicae.* In this latter species, it is transferred to the females while mating, making them less attractive to conspecific males [34]. In turn, this anti-aphrodisiac is exploited by parasitoid wasps such as *Trichogramma brassicae* to detect laid eggs for further parasitization [35]. The most potent attractant for *S. littoralis* larvae at both doses tested was 2-phenylethanol, an aromatic compound that induced the highest firing rate in OSNs expressing SlitOR24. 2-phenylethanol is released by flowers, fruits or vegetative tissues of a large array of plants from a multitude of families [36] and it may be important for caterpillar foraging behavior. It is documented as one of the most attractive compounds - together with phenylacetaldehyde - for *H. armigera* adults [37, 38] and elicited high neuronal responses in *Heliothis virescens* females [39].

Although we propose here a probable role in caterpillar foraging behavior, the potential ecological significance of these *S. littoralis* larval attractants remains to be determined, as well

as their behavioral effects on adults, in which SlitOR24 and 25 are also expressed in antennae [40, 41]. Anyhow, our work shows that reverse chemical ecology can be applied efficiently to identify behaviorally-active volatiles that could ultimately implement semiochemical-based control strategies against agricultural pests. Improved membrane protein tridimensional structure resolution [30, 42] and prediction [43, 44] will give access to structural details of the odorant-binding pocket of insect ORs then contributing to expand the chemical space to be explored by structure-based virtual screening.

References

1. Dethier VG (1947) Chemical Insect Attractants and Repellents. Philadelphia

2. Witzgall P, Kirsch P, Cork A (2010) Sex pheromones and their impact on pest management. J Chem Ecol 36:80-100. doi: 10.1007/s10886-009-9737-y

3. Nyasembe VO, Torto B (2014) Volatile phytochemicals as mosquito semiochemicals. Phytochem Lett 8:196-201. doi: 10.1016/j.phytol.2013.10.003

4. Ray A (2015) Reception of odors and repellents in mosquitoes. Curr Opin Neurobiol 34:158-64. doi: 10.1016/j.conb.2015.06.014

5. Choo YM, Xu P, Hwang JK, Zeng F, Tan K, Bhagavathy G, et al. (2018) Reverse chemical ecology approach for the identification of an oviposition attractant for *Culex quinquefasciatus*. Proc Natl Acad Sci U S A 115:714-9. doi: 10.1073/pnas.1718284115

6. Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, Axel R (1999) A spatial map of olfactory receptor expression in the *Drosophila* antenna. Cell 96:725-36

7. Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim J, Carlson JR (1999) A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. Neuron 22:327-38

8. Gao Q, Chess A (1999) Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. Genomics 60:31-9

9. Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB (2004) Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. Neuron 43:703-14

10. Sato K, Pellegrino M, Nakagawa T, Vosshall LB, Touhara K. Insect olfactory receptors are heteromeric ligand-gated ion channels (2008) Nature 452:1002-6. doi: 10.1038/nature06850

11. Wicher D, Schafer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, et al. (2008) *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. Nature 452:1007-11. doi: 10.1038/nature06861

12. Leal WS (2013) Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. Annu Rev Entomol 58:373-91. doi: 10.1146/annurev-ento-120811-153635

13. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin, II, Cronin M, et al. (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57:4977-5010. doi: 10.1021/jm4004285

14. Mansouri K, Judson RS (2016) In Silico Study of In Vitro GPCR Assays by QSAR Modeling. Methods Mol Biol 1425:361-81. doi: 10.1007/978-1-4939-3609-0_16

15. Boyle SM, McInally S, Ray A (2013) Expanding the olfactory code by in silico decoding of odor-receptor chemical space. Elife 2:e01120. doi: 10.7554/eLife.01120

16. Katritzky AR, Wang Z, Slavov S, Tsikolia M, Dobchev D, Akhmedov NG, et al. (2008) Synthesis and bioassay of improved mosquito repellents predicted from chemical structure. Proc Natl Acad Sci U S A 105:7359-64. doi: 10.1073/pnas.0800571105

17. Oliferenko PV, Oliferenko AA, Poda GI, Osolodkin DI, Pillai GG, Bernier UR, et al. (2013) Promising *Aedes aegypti* repellent chemotypes identified through integrated QSAR, virtual screening, synthesis, and bioassay. PLoS One 8:e64547. doi: 10.1371/journal.pone.0064547

18. Tauxe GM, MacWilliam D, Boyle SM, Guda T, Ray A (2013) Targeting a dual detector of skin and CO2 to modify mosquito host seeking. Cell 155:1365-79. doi: 10.1016/j.cell.2013.11.013

19. Kepchia D, Xu P, Terryn R, Castro A, Schurer SC, Leal WS, et al. (2019) Use of machine learning to identify novel, behaviorally active antagonists of the insect odorant receptor co-receptor (Orco) subunit. Sci Rep 9:4055. doi: 10.1038/s41598-019-40640-4

20. Caballero-Vidal G, Bouysset C, Grunig H, Fiorucci S, Montagne N, Golebiowski J, et al. (2020) Machine learning decodes chemical features to identify novel agonists of a moth odorant receptor. Sci Rep 10:1655. doi: 10.1038/s41598-020-58564-9

21. de Fouchier A, Walker WB, 3rd, Montagne N, Steiner C, Binyameen M, Schlyter F, et al. (2017) Functional evolution of Lepidoptera olfactory receptors revealed by deorphanization of a moth repertoire. Nat Commun 8:15709. doi: 10.1038/ncomms15709

22. de Fouchier A, Sun X, Caballero-Vidal G, Travaillard S, Jacquin-Joly E, Montagne N (2018) Behavioral Effect of Plant Volatiles Binding to *Spodoptera littoralis* Larval Odorant Receptors. Frontiers in behavioral neuroscience 12:264. doi: 10.3389/fnbeh.2018.00264

23. Zhu J, Arena S, Spinelli S, Liu D, Zhang G, Wei R, et al. (2017) Reverse chemical ecology: Olfactory proteins from the giant panda and their interactions with putative pheromones and bamboo volatiles. Proc Natl Acad Sci U S A 114:E9802-E10. doi: 10.1073/pnas.1711437114

24. Poitout S, Buès R (1974) Elevage de chenilles de vingt-huit espèces de Lépidoptères Noctuidae et de deux espèces d'Arctiidae sur milieu artificiel simple. Particularités de l'élevage selon les espèces. Ann Zool Ecol anim. 6:431-41

25. Dobritsa AA, van der Goes van Naters W, Warr CG, Steinbrecht RA, Carlson JR (2003) Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. Neuron 37:827-41

26. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition ed. CA, USA

27. Tropsha A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science 22:69-77. doi: doi.org/10.1002/qsar.200390007

28. Poivet E, Rharrabe K, Monsempes C, Glaser N, Rochat D, Renou M, et al. (2012) The use of the sex pheromone as an evolutionary solution to food source selection in caterpillars. Nat Commun 3:1047. doi: 10.1038/ncomms2050

29. Di C, Ning C, Huang LQ, Wang CZ (2017) Design of larval chemical attractants based on odorant response spectra of odorant receptors in the cotton bollworm. Insect Biochem Mol Biol 84:48-62. doi: 10.1016/j.ibmb.2017.03.007

30. del Marmol J, Yedlin M, Ruta V (2021) The structural basis of odorant recognition in insect olfactory receptors. Preprint at https://doi.org/10.1101/2021.01.24.427933

31. Clavijo McCormick A, Boeckler GA, Kollner TG, Gershenzon J, Unsicker SB (2014) The timing of herbivore-induced volatile emission in black poplar (*Populus nigra*) and the influence

of herbivore age and identity affect the value of individual volatiles as cues for herbivore enemies. BMC Plant Biol 14:304. doi: 10.1186/s12870-014-0304-5

32. Smid HM, Van Loon JJA, Posthumus MA, Vet LEM (2002) GC-EAG-analysis of volatiles from Brussels sprouts plants damaged by two species of *Pieris caterpillars*: Olfactory receptive range of a specialist and a generalist parasitoid wasp species. Chemoecology 12:169–76

33. Ferenz H-J, Seidelmann K (2003) Pheromones in relation to aggregation and reproduction in desert locusts. Physiological Entomology 28:11-8. doi: 10.1046/j.1365-3032.2003.00318.x

34. Andersson J, Borg-Karlson AK, Wiklund C (2003) Antiaphrodisiacs in pierid butterflies: a theme with variation! J Chem Ecol 29:1489-99. doi: 10.1023/a:1024277823101

35. Fatouros NE, Huigens ME, van Loon JJ, Dicke M, Hilker M. (2005) Chemical communication: butterfly anti-aphrodisiac lures parasitic wasps. Nature 433:704. doi: 10.1038/433704a

36. Knudsen JT, Eriksson R, Gershenzon J, Ståhl B (2006) Diversity and distribution of floral scent. Botanical Review 72:1-120

37. Gregg PC, Del Socorro AP, Henderson GS (2010) Development of a synthetic plant volatile-based attracticide for female noctuid moths. II. Bioassays of synthetic plant volatiles as attractants for the adults of the cotton bollworm, *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae). Australian Journal of Entomology 49:21-30. doi: 10.1111/j.1440-6055.2009.00734.x

38. Guo M, Du L, Chen Q, Feng Y, Zhang J, Zhang X, et al. (2021) Odorant Receptors for Detecting Flowering Plant Cues Are Functionally Conserved across Moths and Butterflies. Mol Biol Evol 38:1413-27. doi: 10.1093/molbev/msaa300

39. Rostelien T, Stranden M, Borg-Karlson AK, Mustaparta H (2005) Olfactory receptor neurons in two Heliothine moth species responding selectively to aliphatic green leaf volatiles, aromatic compounds, monoterpenes and sesquiterpenes of plant origin. Chem Senses 30:443-61. doi: 10.1093/chemse/bji039

40. Poivet E, Gallot A, Montagne N, Glaser N, Legeai F, Jacquin-Joly E. (2013) A comparison of the olfactory gene repertoires of adults and larvae in the noctuid moth *Spodoptera littoralis*. PLoS One 8:e60263. doi: 10.1371/journal.pone.0060263

41. Walker WB, 3rd, Roy A, Anderson P, Schlyter F, Hansson BS, Larsson MC (2019) Transcriptome Analysis of Gene Families Involved in Chemosensory Function in *Spodoptera littoralis* (Lepidoptera: Noctuidae). BMC Genomics 20:428. doi: 10.1186/s12864-019-5815-x

42. Butterwick JA, Del Marmol J, Kim KH, Kahlson MA, Rogow JA, Walz T, et al. (2018) Cryo-EM structure of the insect olfactory receptor Orco. Nature 560:447-52. doi: 10.1038/s41586-018-0420-8

43. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020) Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117:1496-503. doi: 10.1073/pnas.1914677117

44. Benton R, Dessimoz C, Moi D (2020) A putative origin of the insect chemosensory receptor superfamily in the last common eukaryotic ancestor. Elife 9. doi: 10.7554/eLife.62507

Declarations

Funding

This work has been funded by INRAE, Sorbonne Université and the French National Research Agency (ANR-16-CE21-0002). We thank BECAL and the National Council of Science and

Technology of Paraguay for the doctoral fellowships awarded to G.C-V. We also benefited from funding from the French government, through the UCAJEDI "Investments in the Future" project managed by the ANR grant No. ANR-15-IDEX-01 (to S.F. and J.Go.) and by GIRACT (Geneva, Switzerland) [9th European PhD in Flavor Research Bursaries for first year students to C.B.] and the Gen Foundation (Registered UK Charity No. 1071026) [a charitable trust which principally provides grants to students/researchers in natural sciences, in particular food sciences/technology to C.B.]. Computation for the work described in this paper was supported by the Université Côte d'Azur's Center for High-Performance Computing.

Conflicts of interest/Competing interests

The authors declare no conflict of interest nor competing interest.

Availability of data and material

All in silico data and weka model files have been deposited on GitHub (https://github.com/chemosim-lab/SlitOR_data). Transformant flies are available on request to Emmanuelle Jacquin-Joly (emmanuelle.joly@inrae.fr) or Nicolas Montagné (nicolas.montagne@sorbonne-universite.fr).

Code Availability

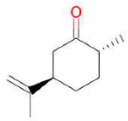
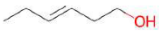
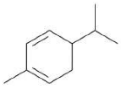
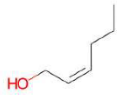
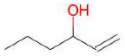
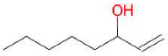
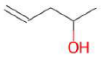
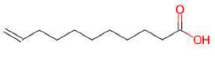
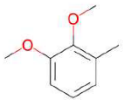
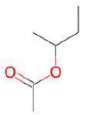
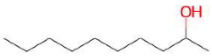
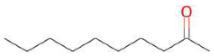
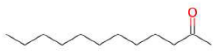
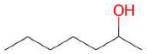
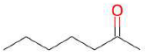
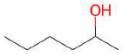
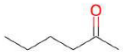
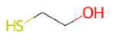
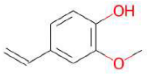
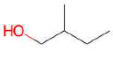
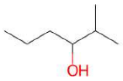
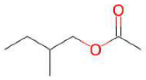
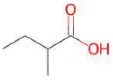
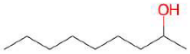
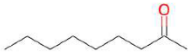
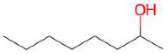
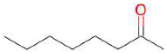
All the binary files necessary to screen the database with the machine-learning models, the code used to define the applicability domain of the models, a description of the content of each file and how to use them are made available on GitHub (https://github.com/chemosim-lab/SlitOR_data).

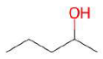
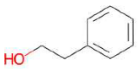
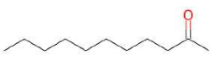
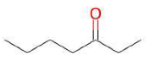
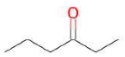
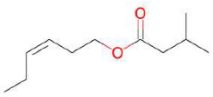
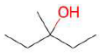
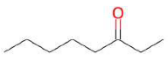
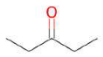
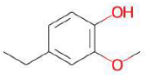
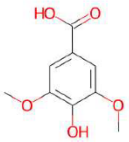
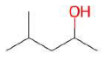
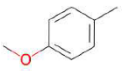
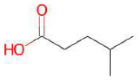
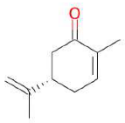
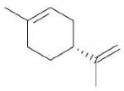
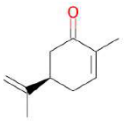
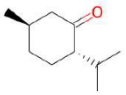
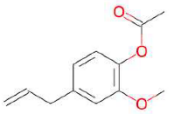
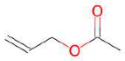
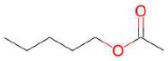
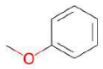
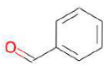
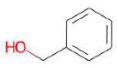
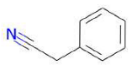
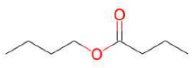
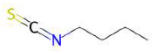
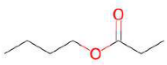
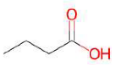
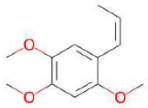
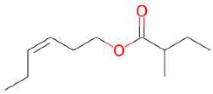
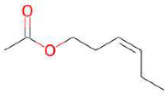
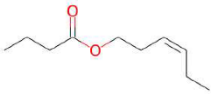
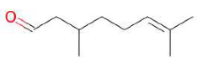
Authors' contributions

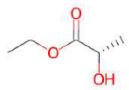
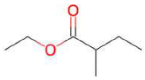
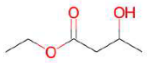
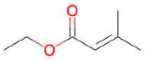
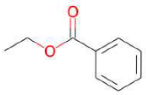
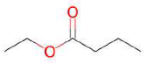
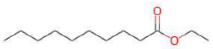
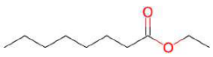
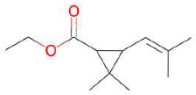
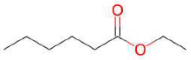
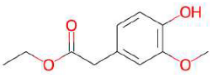
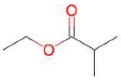
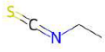
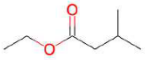
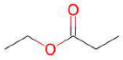
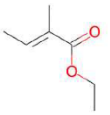
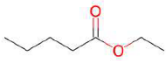
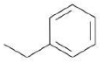
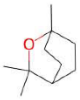
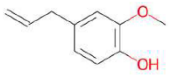
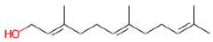
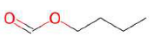
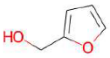
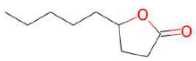
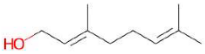
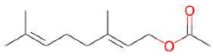
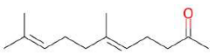
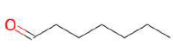
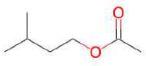
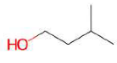
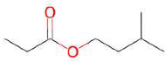
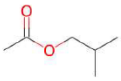
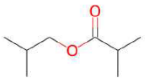
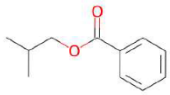
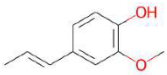
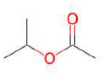
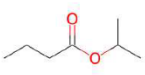
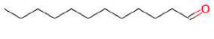
E.J-J., J.Go., N.M. and S.F. conceived and designed the experiments. G.C-V. and C.B. designed and performed the experiments and analyzed the data. J.Gé. designed the behavioral experiments. H.M, C.N. and J.D. performed the behavioral experiments. E.J-J., J.Gé., G.C-V., C.B., N.M. and S.F. wrote and revised the paper.

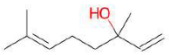
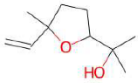
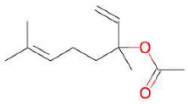
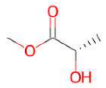
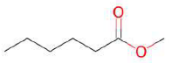
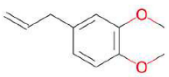
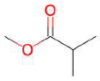
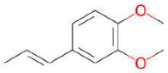
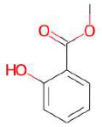
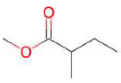
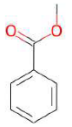
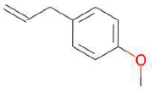
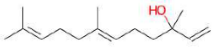
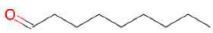
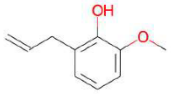
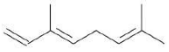
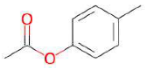
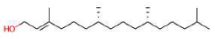
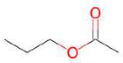
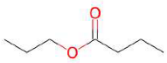
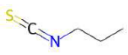
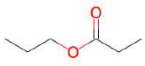
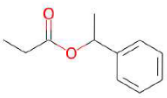
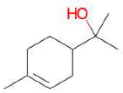
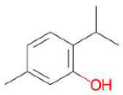
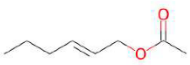
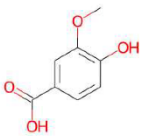
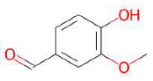
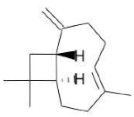
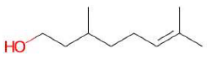
Supporting information

Online Resource 1. In-house library of plant volatile organic compounds

 (+)-dihydrocarvone 5524-05-0	 (E)-3-hexen-1-ol 928-97-2	 (R)-α-phellandrene 4221-98-1	 (Z)-2-hexen-1-ol 928-94-9	 (methylthio)methane 75-18-3
1- tridecanol 112-70-9	 1-butanol 71-36-3	 1-decanol 112-30-1	 1-heptanol 111-70-6	 1-hexanol 111-27-3
 1-hexen-3-ol 4798-44-1	 1-nonanol 143-08-8	 1-octanol 111-87-5	 1-octen-3-ol	 1-pentanol 71-41-0
 1-penten-4-ol 625-31-0	 1-undecanol 112-42-5	 1-undecenoic acid 112-38-9	 2,3-butanediol 513-85-9	 2,3-dimethoxytoluene 4463-33-6
 2-butanol 78-92-2	 2-butanone 78-93-3	 2-butyl acetate 105-45-4	 2-decanol 1120-06-5	 2-decanone 693-54-9
 2-dodecanone 6175-49-1	 2-heptanol 543-49-7	 2-heptanone 110-43-0	 2-hexanol 626-93-7	 2-hexanone 591-78-6
 2-mercaptoethanol	 2-methoxy-4-vinylphenol 7786-61-0	 2-methyl-1-butanol 137-32-6	 2-methyl-3-hexanol 617-29-8	 2-methylbutyl acetate 624-41-9
 2-methylbutyric acid 116-53-0	 2-nonanol 628-99-9	 2-nonanone 821-55-6	 2-octanol	 2-octanone 111-13-7

 2-pentanol 6032-29-7	 2-pentanone 107-87-9	 2-phenyl ethanol 60-12-8	 2-undecanone 112-12-9	 3-heptanone 106-35-4
 3-hexanone 589-38-8	 3-hexenyl 3-methylbutanoate 35154-45-1	 3-methyl-2-butanol 598-75-4	 3-methyl-3-pentanol 77-74-7	 3-octanone 106-68-3
 3-pentanone 96-22-0	 4-butyrolactone 96-48-0	 4-ethylguaiacol	 4-ethylphenol 123-07-9	 4-hydroxy 3,5 benzoic acid 530-57-4
 4-methyl-2-pentanol 108-11-2	 4-methylanisole 100-66-3	 4-methylvaleric acid 646-07-1	 D-carvone 2244-16-8	 D-limonene 5989-27-5
 L-carvone 6485-40-1	 L-menthone 14073-97-3	 acetyleugenol	 allyl acetate 591-87-7	 amyl acetate 628-63-7
 anisole 100-66-3	 benzaldehyde 100-52-7	 benzyl alcohol 100-51-6	 benzyl cyanide 140-29-4	 butyl butyrate 109-21-7
 butyl isothiocyanate 592-82-5	 butyl propionate 590-01-2	 butyric acid 107-92-6	 caproaldehyde (hexanal) 66-25-1	 cis-2,4,5-trimethoxy-1-propenylbenzene
 cis-3-hexenyl 2-methylbutanoate 53398-85-9	 cis-3-hexenyl acetate 3681-71-8	 cis-3-hexenyl butyrate 16491-36-4	 citronellal 106-23-0	dodecane

 ethyl (S)-(-)-lactate 687-47-8	 ethyl 2-methylbutyrate 7452-79-1	 ethyl 3-hydroxybutyrate 5405-41-4	 ethyl 3-methylcrotonate 638-10-8	 ethyl benzoate 93-89-0
 ethyl butyrate 105-54-4	 ethyl caprate 110-38-3	 ethyl caprylate	 ethyl chrysanthemumate 97-41-6	 ethyl hexanoate
 ethyl homovanillate	 ethyl isobutyrate 97-62-1	 ethyl isothiocyanate 542-85-8	 ethyl isovalerate 108-64-5	 ethyl propionate 105-37-3
 ethyl tiglate 5837-78-5	 ethyl valerate 539-82-2	 ethylbenzene 100-41-4	 eucalyptol 470-82-6	 eugenol
 farnesol	 formic acid butyl ester 592-84-7	 furfuryl alcohol 98-00-0	 gamma-nonanoic lactone 104-61-0	 geraniol 106-24-1
 geranyl acetate 16409-44-2	 geranylacetone 3796-70-1	 heptaldehyde 111-71-7	 indole 120-72-9	 isoamyl acetate 123-92-2
 isoamyl alcohol	 isoamyl propionate 105-68-0	 isobutyl acetate 110-19-0	 isobutyl isobutyrate 97-85-8	 isobutylbenzoate
 isobutyric acid 79-31-2	 isoeugenol	 isopropyl acetate 108-21-4	 isopropyl butyrate 638-11-9	 laurinaldehyde 112-54-9

 linalool 78-70-6	 linalool oxide 60047-17-8	 linalyl acetate 115-95-7	 methionol 505-10-2	 methyl (S)-(-)-lactate 27871-49-4
 methyl acetate	 methyl caproate 106-70-7	 methyl eugenol	 methyl isobutyrate 547-63-7	 methyl isoeugenol
 methyl salicylate	 methyl-2-methylbutyrate 868-57-5	 methylbenzoate	 methylchavicol 140-67-0	 nerolidol
 nonanal 124-19-6	 o-eugenol	 ocimene 13877-91-3	 octanal	octane 111-65-9
 p-tolyl acetate 140-39-6	 p-xylene	 phytol	 propyl acetate 109-60-4	 propyl butyrate 105-66-8
 propyl isothiocyanate 628-30-8	 propyl propionate 106-36-5	 styrallyl propionate 120-45-6	 terpineol anhydre 8000-41-7	tetradecane 629-59-4
 thymol	 trans-2-hexenyl acetate 2497-18-9	tridecane	 vanilic acid	 vanillin
 α-pinene 80-56-8	 β-caryophyllene 87-44-5	 β-citronellol 106-22-9		

Online Resource 2. Dataset used for training and validating the QSAR model.

https://github.com/chemosim-lab/SlitOR_data/blob/main/supp%20data%202%20SlitOR24%2B25_dataset.xlsx

Online Resource 3. Molecules predicted as agonists (A) or non-agonists (N) within the applicability domain of SlitOR24 and SlitOR25 (previous model from [20]) models. Molecules that were further experimentally tested on both ORs are indicated (Yes/No) as well as their activity on the corresponding OR (Yes/No).

Molecules	CAS	SlitOR24 prediction	SlitOR25 prediction	Previous SlitOR25 prediction	Molecules tested in SSR	Activity on SlitOR24	Activity on SlitOR25
3-heptanone	106-35-4	A	A	A	Yes	Yes	Yes
anisole	100-66-3	A	A	A	Yes	Yes	Yes
1-hexen-3-ol	4798-44-1	A	A	A	Yes	Yes	Yes
ethyl benzene	100-41-4	A	A	A	Yes	No	Yes
(E)-3-hexen-1-ol	928-97-2	A	A	A	Yes	Yes	Yes
(Z)-2-hexen-1-ol	928-94-9	A	A	A	Yes	Yes	Yes
2-heptanol	543-49-7	A	A	A	Yes	Yes	Yes
1-pentanol	71-41-0	A	A	A	Yes	Yes	Yes
2-heptanone	110-43-0	A	A	A	Yes	Yes	Yes
2-hexanone	591-78-6	A	A	A	Yes	Yes	Yes
2-hexanol	626-93-7	A	A	A	Yes	Yes	Yes
hexanal	66-25-1	A	A	A	Yes	Yes	Yes
benzyl cyanide	140-29-4	A	A	A	Yes	Yes	Yes
heptaldehyde	111-71-7	A	A	A	Yes	Yes	Yes
4-methylanisole	104-93-8		A	A	Yes	Yes	Yes
2-phenyl ethanol	60-12-8		A	A	Yes	Yes	Yes
ethyl valerate	539-82-2		A	A	Yes	No	Yes
4-ethylphenol	123-07-9		A	A	Yes	Yes	Yes
2-methyl butyl acetate	624-41-9		A	A	Yes	Yes	Yes
propyl butyrate	105-66-8		A	A	Yes	No	Yes
2-butyl acetate	105-45-4	A	A		Yes	Yes	No
furfuryl alcohol	98-00-0	A	A		Yes	Yes	No
formic acid butyl ester	592-84-7	A	A		Yes	Yes	No
methyl-2-methylbutyrate	868-57-5	A	A		Yes	No	No
1-penten-4-ol	625-31-0	A	A		Yes	Yes	No
3-hexanone	589-38-8	A	A		Yes	Yes	Yes
ethyl butyrate	105-54-4	A	A		Yes	Yes	No
2-methyl-1-butanol	137-32-6	A	A		Yes	Yes	No
2-pentanol	6032-29-7	A	A		Yes	Yes	No
2-methyl-3-hexanol	617-29-8	A	A		Yes	Yes	No
propyl acetate	109-60-4	A	A		Yes	Yes	No
butyl isothiocyanate	592-82-5	A	A		Yes	Yes	No
propyl propionate	106-36-5	A	A		Yes	Yes	Yes
4-methyl-2-pentanol	108-11-2	A			Yes	Yes	No
2-methoxy-4-vinylphenol	7786-61-0		A		No		
isopropyl acetate	108-21-4		A		No		
1-butanol	71-36-3		A		No		
allyl acetate	591-87-7		A		No		
L-carvone	6485-40-1		A		No		

Molecules	CAS	SlitOR24 prediction	SlitOR25 prediction	Previous SlitOR25 prediction	Molecules tested in SSR	Activity on SlitOR24	Activity on SlitOR25
D-carvone	2244-16-8		A		No		
isoamyl acetate	123-92-2		A		No		
isobutyl acetate	110-19-0		A		No		
ethyl (S)-(-)-lactate	687-47-8		A		No		
4-butyrolactone	96-48-0		A		No		
2,3-butanediol	513-85-9		A		No		
ethyl 3-hydroxybutyrate	5405-41-4		A		No		
2-pentanone	107-87-9		A		No		
p-xylene	106-42-3		A		No		
methionol	505-10-2		A		No		
ethyl isovalerate	108-64-5		A		No		
methyl (S)-(-)-lactate	27871-49-4		A		No		
3-pentanone	96-22-0		A		No		
ethyl 3-methylcrotonate	638-10-8		A		No		
ethyl propionate	105-37-3		A		No		
isopropyl butyrate	638-11-9		A		No		
ethyl tiglate	5837-78-5		A		No		
4-methylvaleric acid	646-07-1		A		No		
isobutyl isobutyrate	97-85-8		A		No		
propyl isothiocyanate	628-30-8		A		No		
2-methylbutyric acid	116-53-0		A		No		
ethyl isobutyrate	97-62-1		A		No		
p-Tolyl acetate	140-39-6		A		No		
ethyl 2-methylbutyrate	7452-79-1		A		No		
3-methyl-3-pentanol	77-74-7		A		No		
ethyl benzoate	93-89-0		A		No		
methyl isobutyrate	547-63-7		A		No		
2,3-dimethoxytoluene	4463-33-6		A		No		
3-methyl-2-butanol	598-75-4		A		No		
D-limonene	5989-27-5	N	N		Yes	No	No
gamma-nonanoic lactone	104-61-0	N	N		Yes	No	No
(-)-menthone	14073-97-3	N	N		Yes	No	No
butyl propionate	590-01-2	N	N		Yes	No	No
eucalyptol	470-82-6	N	N		Yes	No	No

Online Resource 4. Performance of the QSAR models when changing the initial test compound used in the sphere-exclusion algorithm to obtain the training and test datasets. Only the first five splits that had the same activity distribution as in the split used for the final model were investigated.

	Split #	Dataset	TP	TN	FP	FN	Accuracy	Precision	Recall	FPR	MCC	AUROC
SlitOR24	1	LOO	4	29	2	4	0.85	0.67	0.50	0.06	0.49	0.93
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	1.00
		Test	0	10	0	2	0.83	NA	0.00	0.00	NA	0.50
	2	LOO	3	29	2	5	0.82	0.60	0.38	0.06	0.38	0.80
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	1.00
		Test	1	10	0	1	0.92	1.00	0.50	0.00	0.67	0.95
	3	LOO	3	29	2	5	0.82	0.60	0.38	0.06	0.38	0.80
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	1.00
		Test	1	10	0	1	0.92	1.00	0.50	0.00	0.67	0.95
	4	LOO	4	29	2	4	0.85	0.67	0.50	0.06	0.49	0.83
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	1.00
		Test	0	10	0	2	0.83	NA	0.00	0.00	NA	0.85
	5	LOO	3	29	2	5	0.82	0.60	0.38	0.06	0.38	0.80
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	1.00
		Test	1	10	0	1	0.92	1.00	0.50	0.00	0.67	0.95
	Model	LOO	4	29	2	4	0.85	0.67	0.50	0.06	0.49	0.83
		Training	7	31	0	1	0.97	1.00	0.88	0.00	0.92	0.99
		Test	1	9	1	1	0.83	0.50	0.50	0.10	0.40	0.80
SlitOR25	1	LOO	14	38	8	4	0.81	0.64	0.78	0.17	0.57	0.87
		Training	18	46	0	0	1.00	1.00	1.00	0.00	1.00	1.00
		Test	6	8	4	1	0.74	0.60	0.86	0.33	0.51	0.77
	2	LOO	15	36	10	3	0.80	0.60	0.83	0.22	0.57	0.89
		Training	18	46	0	0	1.00	1.00	1.00	0.00	1.00	1.00
		Test	6	9	3	1	0.79	0.67	0.86	0.25	0.59	0.81
	3	LOO	15	36	10	3	0.80	0.60	0.83	0.22	0.57	0.89
		Training	18	46	0	0	1.00	1.00	1.00	0.00	1.00	1.00
		Test	6	9	3	1	0.79	0.67	0.86	0.25	0.59	0.81
	4	LOO	11	39	7	7	0.78	0.61	0.61	0.15	0.46	0.89
		Training	16	41	5	2	0.89	0.76	0.89	0.11	0.75	0.96
		Test	6	6	6	1	0.63	0.50	0.86	0.50	0.36	0.69
	5	LOO	15	38	8	3	0.83	0.65	0.83	0.17	0.62	0.89
		Training	16	39	7	2	0.86	0.70	0.89	0.15	0.69	0.94
		Test	7	6	6	0	0.68	0.54	1.00	0.50	0.52	0.85
	Model	LOO	15	34	12	3	0.77	0.56	0.83	0.26	0.52	0.84
		Training	18	46	0	0	1.00	1.00	1.00	0.00	1.00	1.00
		Test	5	10	2	2	0.79	0.71	0.71	0.17	0.55	0.89

Online Resource 5. Range of metrics values (min and max) for all splits investigated in Online Resource 4.

Target	Dataset	Accuracy	Precision	Recall	FPR	MCC	AUROC
SlitOR24	LOO	0.82–0.85	0.60–0.67	0.38–0.50	0.06	0.38–0.49	0.80–0.93
	Training	0.97	1.00	0.88	0.00	0.92	1.00
	Test	0.83–0.92	NA–1.00	0.00–0.50	0.00–0.10	NA–0.67	0.50–0.95
SlitOR25	LOO	0.77–0.83	0.56–0.65	0.61–0.83	0.10–0.26	0.46–0.62	0.84–0.89
	Training	0.86–1.00	0.70–1.00	0.89–1.00	0.00–0.15	0.69–1.00	0.94–1.00
	Test	0.63–0.79	0.50–0.71	0.71–1.00	0.17–0.50	0.36–0.59	0.69–0.89

Online Resource 6. Performance of the QSAR models on the new experimental data.

Target	TP	TN	FP	FN	Accuracy	Precision	Recall	FPR	MCC
SlitOR24	26	6	2	5	0.82	0.93	0.84	0.25	0.53
SlitOR25	22	6	11	0	0.72	0.67	1.00	0.65	0.49

Online Resource 7. List of putative SlitOR24 and 25 residues corresponding to MhOR5 odorant binding site according to ClustalO and MAFFT multiple sequence alignments (MSA). The MSA have been performed on the EMBL-EBI webserver.

MhOR5	**SlitOR24**		**SlitOR25**	
	ClustalO	MAFFT	ClustalO	MAFFT
V88	V88	V88	V88	V88
Y91	I91	I91	L91	L91
F92	H92	H92	Q92	Q92
S151	T153	T153	T153	T153
G154	V156	V156	A156	A156
W158	Y160	Y160	F160	F160
M209	I195	F197	I195	F197
I213	Y199	S201	Y199	S201
Y380	Y322	Y322	F322	F322
Y383	Y325	Y325	Y325	Y325

Chapter II

Do computers have a sweet tooth? Machine learning for natural sweeteners

Sweet taste is universally and innately perceived as pleasant [1, 2]. This mostly makes sense from an evolutionary point of view as it rewards the consumption of caloric food and because parts of our body, mostly the brain, require sugars as input to function properly [3]. While this made sense for early vertebrates, nowadays the reward mechanism for sugar consumption is overly stimulated due to the abundance of free sugars in processed foods, up to the point where its excessive consumption can foster addictive behaviors greater than or equal to drugs [4–6]. Interestingly, sugars and sweeteners both trigger a pleasant sensation through the brain reward mechanism, but sweeteners don't necessarily foster satiety [7]. Additionally, sugar-sweetened beverages (the largest source of added sugar intake in the US) have been proven to promote excess weight gain, type II diabetes and cardiovascular diseases [8] as well as dental caries [9]. The concern for public health has nurtured proposals to regulate added sugars similarly to tobacco or alcohol, which are the two other main risk factors in non-communicable diseases, mainly through taxes or imposing age limits on their purchase [10]. Low-calorie sweeteners thus appear as an ideal alternative that could satisfy both consumers, food-processing industry, and public health agencies. However, sugar consumption is only part of the equation as obesity, diabetes, and cardiovascular diseases also depend on other external factors such as saturated fats consumption and physical activity. Unfortunately, among the plethora of sweeteners available, none of them can reproduce the sensory profile of table sugar, sucrose [11]. Indeed, some sweeteners, like saccharin, suffer from a bitter aftertaste, some, like stevia, can be perceived with both bitter [12] and menthol aftertaste, while others, like aspartame, suffer from sweetness lag i.e., the sweetness is delayed [13]. Other factors also play a role in both consumer and industry acceptance, like solubility and thermal stability for food and beverage preparations, as well as cost, safety, and patentability [14]. For these reasons, the search for the ideal sweetener is still open. The strategy often used by the food-processing industry to circumvent these limitations is to combine different sugars and sweeteners to mitigate the downsides of each individual sweet additive and come closer to the sweetness profile of sucrose with less calories [13].

One way to design new sweeteners could be to take advantage of the current knowledge on the molecular structures responsible for sweet taste perception. Sweet tastants are detected by the sweet taste receptor found in type II taste cells [15]. It belongs to the GPCR family and is structured as a heterodimer made of two taste receptors type 1 subunits (T1R): T1R2 and T1R3, although the T1R3 homodimer can also be functional and responsive to saccharides albeit at higher concentrations [16]. Interestingly, T1R2 is pseudogenized in several felines including cats, tigers, cheetahs and lions explaining their indifference to sweet taste stimuli [17, 18].

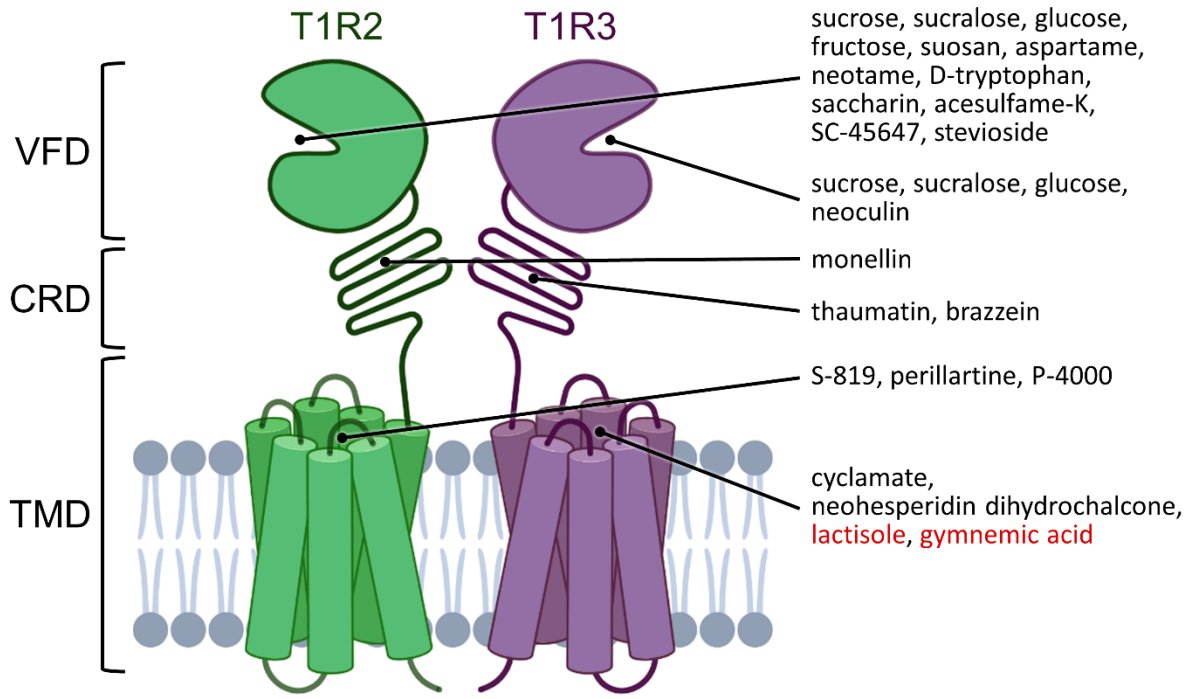


Figure 1: Structure of the sweet taste receptor, labelled with ligands (inhibitors in red) and their binding site (adapted and updated from [19–22]). VFD: Venus flytrap domain, TMD: transmembrane domain, CRD: cysteine-rich domain.

The T1R2 and T1R3 subunits belong to the class C GPCR family which notably comprises metabotropic glutamate receptors (mGluRs) and gamma-amino-butyric acid (GABA) type B receptors. The class C receptor structures are arranged in 3 distinct domains: a large extracellular N-terminal domain called the Venus Flytrap Domain (VFD), the typical 7 helix transmembrane domain (TMD) characteristic of GPCRs, and a cysteine-rich domain (CRD) that connects those two domains [23]. For the sweet taste receptor, several binding pockets have been identified, one in each of these domains (Figure 1) [19, 24]. In the case of the sweet taste receptor, the VFD binds natural sugars which tend to be polar, the CRD is a quite rigid structure [25] that can bind sweet tasting proteins, and the TMD binds sweeteners and, for T1R3, negative allosteric modulators. However, applying structure-based approaches to the sweet taste receptor for the discovery of new sweeteners is challenging for two reasons. Firstly, there is no known structure of the receptor, except for the VFD of T1R2 and T1R3 of the medaka fish [26], although class C structures are available and could help building homology models as was done previously [23]. Secondly, multiple binding sites are known and predicting which one to choose for each ligand is not necessarily straightforward.

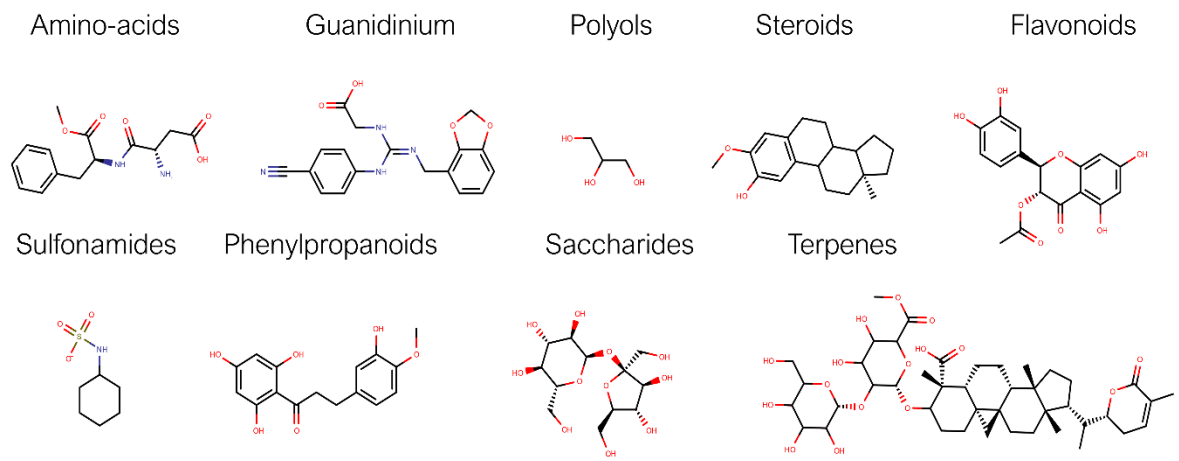


Figure 2: Diverse structures extracted from the SweetenersDB, labelled by chemical classes.

To date, more than 300 small-molecule sugars and sweeteners are known, and a description of their sweet taste intensity, called relative sweetness, is available [27]. Relative sweetness is calculated as the concentration ratio between a solution of sweetener and a solution of sucrose perceived with the same intensity. Hence, sucrose has a relative sweetness of 1, and the most intense sweetener, lugduname, has a relative sweetness of approximately 225 000. A striking chemical diversity can be found among the list of known sugars and sweeteners, with not only a large variety of saccharides, but also of polyols, polyphenols, amino-acid derivatives, terpenes, and phenylpropanoids, among others (Figure 2). Such disparity can be explained by the multitude of binding sites upon which those sweet molecules can bind and opens the question of whether or not there are more sweet scaffolds to discover. Considering all the small-molecule data available, ligand-based methods appear as credible alternatives to search for novel sweet compounds. These approaches have been investigated in the past, starting with pharmacophore models as early as 1914 [28], later followed by machine-learning models that either classify molecules as sweet or non-sweet, or models that predict the relative sweetness [27].

In this chapter, I focus on the design of an online QSPR platform that can predict the relative sweetness of compounds based on their structure in order to identify novel intense natural sweeteners. Starting by updating and curating the database of sugars and sweeteners previously established by the group (SweetenersDB) [27], I then used state-of-the-art machine-learning protocols to train and validate a model based on open-source descriptors (Figure 3). I also enforced a thorough evaluation of the applicability domains of the model to quantitatively estimate the quality (in terms of applicability, reliability, and certainty) of each prediction.

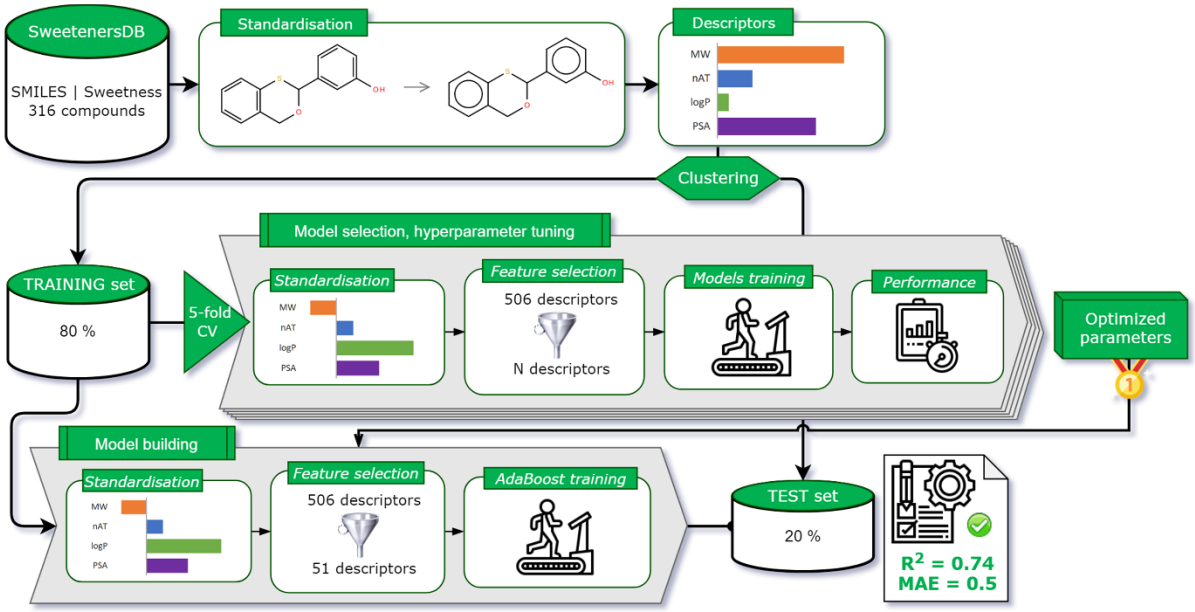


Figure 3: Machine-learning workflow applied for the development of the QSPR model of sweetness prediction.

The model was then implemented on a freely accessible webserver (PrediSweet, http://chemosimserver.unice.fr/predisweet) and used to screen a dataset of natural compounds. Three compounds were prioritized and one was validated by in vitro assays, corresponding to a novel sweet scaffold belonging to the lignan family.

Contributions

Publication 3

I updated and curated the database of sugars and sweeteners (SweetenersDB), trained and validated the machine-learning model for sweetness prediction completed with the definition of the applicability domains, set up the webserver and deployed the model there (PrediSweet), and screened the dataset of natural compounds to identify putative sweeteners and prioritized three compounds. Our collaborators tested these compounds *in vitro*.

Oral and poster presentations

This work was presented as a poster during the 2nd UCA Complex Days meeting (2019) for which I received a "best poster award", and the 26th PACA Chemistry Day (2019). It was also presented orally during the 21st GGMM congress (2019), and the 9th meeting of the French chemoinformatics society (SFCi, 2019) for which I was awarded the "best oral communication".

References

1. Ventura AK, Mennella JA (2011) Innate and learned preferences for sweet taste during childhood: Current Opinion in Clinical Nutrition and Metabolic Care 14:379–384. https://doi.org/10.1097/MCO.0b013e328346df65
2. Reed DR, Knaapila A (2010) Genetics of Taste and Smell. In: Progress in Molecular Biology and Translational Science. Elsevier, pp 213–240
3. Mergenthaler P, Lindauer U, Dienel GA, Meisel A (2013) Sugar for the brain: the role of glucose in physiological and pathological brain function. Trends in Neurosciences 36:587–597. https://doi.org/10.1016/j.tins.2013.07.001
4. Lenoir M, Serre F, Cantin L, Ahmed SH (2007) Intense Sweetness Surpasses Cocaine Reward. PLoS ONE 2:e698. https://doi.org/10.1371/journal.pone.0000698
5. Madsen HB, Ahmed SH (2015) Drug versus sweet reward: greater attraction to and preference for sweet versus drug cues: Drug versus sweet reward. Addiction Biology 20:433–444. https://doi.org/10.1111/adb.12134
6. Avena NM, Rada P, Hoebel BG (2008) Evidence for sugar addiction: Behavioral and neurochemical effects of intermittent, excessive sugar intake. Neuroscience & Biobehavioral Reviews 32:20–39. https://doi.org/10.1016/j.neubiorev.2007.04.019
7. Murray S, Tulloch A, Criscitelli K, Avena NM (2016) Recent studies of the effects of sugars on brain systems involved in energy balance and reward: Relevance to low calorie sweeteners. Physiology & Behavior 164:504–508. https://doi.org/10.1016/j.physbeh.2016.04.004
8. Malik VS, Hu FB (2012) Sweeteners and Risk of Obesity and Type 2 Diabetes: The Role of Sugar-Sweetened Beverages. Curr Diab Rep 12:195–203. https://doi.org/10.1007/s11892-012-0259-6
9. Touger-Decker R, van Loveren C (2003) Sugars and dental caries. The American Journal of Clinical Nutrition 78:881S-892S. https://doi.org/10.1093/ajcn/78.4.881S
10. Lustig RH, Schmidt LA, Brindis CD (2012) The toxic truth about sugar. Nature 482:27–29. https://doi.org/10.1038/482027a
11. Chéron J-B, Marchal A, Fiorucci S (2019) Natural Sweeteners. In: Encyclopedia of Food Chemistry. Elsevier, pp 189–195
12. Hellfritsch C, Brockhoff A, Stähler F, et al (2012) Human Psychometric and Taste Receptor Responses to Steviol Glycosides. J Agric Food Chem 60:6782–6793. https://doi.org/10.1021/jf301297n
13. O'Brien-Nabors L (2001) Alternative sweeteners, 3rd ed., rev.expanded. M. Dekker, New York
14. DuBois GE, Prakash I (2012) Non-Caloric Sweeteners, Sweetness Modulators, and Sweetener Enhancers. Annu Rev Food Sci Technol 3:353–380. https://doi.org/10.1146/annurev-food-022811-101236
15. Clapp TR, Yang R, Stoick CL, et al (2004) Morphologic characterization of rat taste receptor cells that express components of the phospholipase C signaling pathway. J Comp Neurol 468:311–321. https://doi.org/10.1002/cne.10963
16. Nelson G, Hoon MA, Chandrashekar J, et al (2001) Mammalian Sweet Taste Receptors. Cell 106:381–390. https://doi.org/10.1016/S0092-8674(01)00451-2
17. Li X, Li W, Wang H, et al (2006) Cats Lack a Sweet Taste Receptor. The Journal of Nutrition 136:1932S-1934S. https://doi.org/10.1093/jn/136.7.1932S

18. Li X, Glaser D, Li W, et al (2009) Analyses of Sweet Receptor Gene (Tas1r2) and Preference for Sweet Stimuli in Species of Carnivora. Journal of Heredity 100:S90–S100. https://doi.org/10.1093/jhered/esp015
19. DuBois GE (2016) Molecular mechanism of sweetness sensation. Physiology & Behavior 164:453–463. https://doi.org/10.1016/j.physbeh.2016.03.015
20. Sigoillot M, Brockhoff A, Meyerhof W, Briand L (2012) Sweet-taste-suppressing compounds: current knowledge and perspectives of application. Appl Microbiol Biotechnol 96:619–630. https://doi.org/10.1007/s00253-012-4387-3
21. Kim S-K, Chen Y, Abrol R, et al (2017) Activation mechanism of the G protein-coupled sweet receptor heterodimer with sweeteners and allosteric agonists. Proc Natl Acad Sci USA 114:2568–2573. https://doi.org/10.1073/pnas.1700001114
22. Belloir C, Neiers F, Briand L (2017) Sweeteners and sweetness enhancers. Current Opinion in Clinical Nutrition and Metabolic Care 20:279–285. https://doi.org/10.1097/MCO.0000000000000377
23. Chéron J-B, Golebiowski J, Antonczak S, Fiorucci S (2017) The anatomy of mammalian sweet taste receptors: Modeling Sweet Taste Receptors. Proteins 85:332–341. https://doi.org/10.1002/prot.25228
24. Xu H, Staszewski L, Tang H, et al (2004) Different functional roles of T1R subunits in the heteromeric taste receptors. Proceedings of the National Academy of Sciences 101:14258–14263. https://doi.org/10.1073/pnas.0404384101
25. Wellendorph P, Bräuner-Osborne H (2009) Molecular basis for amino acid sensing by family C G-protein-coupled receptors: The family C of G-protein-coupled receptors. British Journal of Pharmacology 156:869–884. https://doi.org/10.1111/j.1476-5381.2008.00078.x
26. Nuemket N, Yasui N, Kusakabe Y, et al (2017) Structural basis for perception of diverse chemical substances by T1r taste receptors. Nat Commun 8:15530. https://doi.org/10.1038/ncomms15530
27. Chéron J-B, Casciuc I, Golebiowski J, et al (2017) Sweetness prediction of natural compounds. Food Chemistry 221:1421–1425. https://doi.org/10.1016/j.foodchem.2016.10.145
28. DuBois GE (2011) Validity of early indirect models of taste active sites and advances in new taste technologies enabled by improved models: Taste receptor models and breakthrough applications. Flavour Fragr J 26:239–253. https://doi.org/10.1002/ffj.2042

Publication 3

Novel Scaffold of Natural Compound Eliciting Sweet Taste Revealed by Machine Learning

Cédric Bouysset, Christine Belloir, Serge Antonczak, Loïc Briand, & Sébastien Fiorucci*

Food Chemistry 2020, *324*, 126864.

doi.org/10.1016/j.foodchem.2020.126864

Food Chemistry 324 (2020) 126864

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Short communication

Novel scaffold of natural compound eliciting sweet taste revealed by machine learning

Cédric Bouysset[a], Christine Belloir[b], Serge Antonczak[a], Loïc Briand[b], Sébastien Fiorucci[a,*]

Abstract

Sugar replacement is still an active issue in the food industry. The use of structure-taste relationships remains one of the most rational strategy to expand the chemical space associated to sweet taste. A new machine learning model has been setup based on an update of the SweetenersDB and on open-source molecular features. It has been implemented on a freely accessible webserver. Cellular functional assays show that the sweet taste receptor is activated *in vitro* by a new scaffold of natural compounds identified by the *in silico* protocol. The newly identified sweetener belongs to the lignan chemical family and opens a new chemical space to explore.

Keywords

Sweet taste, machine learning, natural compounds, sweetener, sweet taste receptor

Introduction

Consumer interest in natural high potency sweeteners has grown spectacularly in recent years, fueled by concerns about sugar overconsumption and the use of artificial additives in foods. There are three main strategies to reduce sugar intake: an abrupt reduction of sugar without substitution, the use of flavor materials to modify sweet taste perception and the use of alternative sweeteners. Though many low-calorie sweeteners are known, only few of them are used by the food industry (Belloir, Neiers, & Briand, 2017). The search of novel intense sweeteners, possessing the same chemosensory profile as sucrose, remains open and challenging.

All sweet tasting compounds are detected by a single heterodimeric G protein-coupled receptor composed of T1R2 and T1R3 subunits expressed at the surface of taste buds (Li et al., 2002; Nelson et al., 2001). However, no experimental 3D-structure of the T1R2/T1R3 sweet taste receptor is available and ligand-based approaches such as Structure Activity Relationship (SAR), are relevant to establish a link between the structure of a compound and its sweet taste. From original studies of Edna W. Deutsch & Corwin Hansch (Deutsch & Hansch, 1966), followed a year later by Robert S. Shallenberger & Terry E. Acree (Shallenberger & Acree, 1967) to recent structure-taste relationship models (Achary, Toropova, & Toropov, 2019; Arnoldi, Bassoli, Merlini, & Ragg, 1991; Barker, Hattotuwagama, & Drew, 2002; Bassoli et al., 2001; Chéron, Casciuc, Golebiowski, Antonczak, & Fiorucci, 2017; Drew et al., 1998;

Rojas, Tripaldi, & Duchowicz, 2016; Spillane & McGlinchey, 1981; Spillane et al., 2000, 1996; Spillane, McGlinchey, Muircheartaigh, & Benson, 1983; Spillane & Sheahan, 1989; Tuwani, Wadhwa, & Bagler, 2019; Van Der Heijden, Brussel, & Peer, 1979; Vepuri, Tawari, & Degani, 2007; Walters, 2006; Zheng, Chang, Xu, Xu, & Lin, 2019), the quest to understand the molecular features underlying sweet taste perception is still active.

In this study, we present the first online tool able to predict sweet taste perception based on a machine learning protocol. We have updated and curated the previous database of 316 sweet compounds (SweetenersDB) and added new applicability domain metrics to assess the robustness of the predictions. A novel scaffold of natural sweetener, belonging to the lignan chemical family, that have never been annotated as sweet have been identified and experimentally validated.

Materials and Methods

Data preparation

Based on our previous work (Chéron et al., 2017), the database of sugars and sweeteners (Figure S1), named SweetenersDB, was curated and updated with missing compounds (Ruiz-Aceituno, Hernandez-Hernandez, Kolida, Moreno, & Methven, 2018). Each compound was labelled with a relative sweetness value, corresponding to a measure of the sweet taste intensity relative to sucrose. Relative sweetness is defined as the concentration ratio between a sucrose solution and a solution of sweetener perceived with the same intensity. The relative sweetness of each compound was transformed in logarithmic scale for easier manipulation, and it will be later referred to as logSw. For compounds that were already present in the database, we updated the SMILES (Simplified Molecular Input Line Entry System) to isomeric SMILES in order to differentiate stereoisomers. When the information on stereocenters was not available, we either regrouped the stereoisomers in a single entry with their average logSw value if the logSw difference was lower than 0.2, or we discarded both compounds. The resulting dataset consisted of 316 compounds in SweetenersDB (Table S1). The machine learning protocol was applied to two datasets of interest: 4796 natural compounds (Table S2) extracted from the SuperNatural II database and the phyproof catalogue from PhytoLab, already pre-screened by our previous model (Chéron et al., 2017).

Every compound in the datasets were collected as SMILES strings and sanitized with RDKit (Landrum et al., 2018). To assess the importance of predicting protonation states, the major

microspecies of each compound was also determined with ChemAxon cxcalc tool (ChemAxon, 2018) at physiological salivary pH (pH=6.5). Structures were then standardized using the "standardizer" (EMBL-EBI, 2017) Python package: salts are removed from the structure, and a set of around 30 structure-normalization rules are applied to each molecular graph to cover most of tautomerization reactions. 0D, 1D and 2D descriptors were computed using Dragon v6.0.38 (Talete srl, 2014), RDKit (Landrum et al., 2018), Mordred (Moriwaki, Tian, Kawashita, & Takagi, 2018), and ChemoPy (Cao, Xu, Hu, & Liang, 2013). Descriptors from the three latter packages were regrouped as "open-source" descriptors. For each of these two descriptors sets, the initial number of features was reduced by removing those that could not be calculated for a molecule, as well as near-constant features (two or less unique values), features with a standard deviation below 0.001, and features with a correlation greater than 0.95. The resulting datasets consisted of 635 descriptors for the Dragon dataset, and 506 features for the "open-source" dataset. To avoid any model bias due to overfitting, the number of features used by the model is a hyperparameter that has been optimized.

The updated SweetenersDB was split in training and test sets using a Sphere Exclusion clustering algorithm. Dragon descriptors were chosen for this procedure: they were normalized between 0 and 1, and the clustering was initiated from the compound that is closest to the center of the dataset in the descriptor hyperspace. 64 diverse compounds (20.3%) were selected for the test set, leaving 252 compounds in the training set (Figure 1, Table S1). The chemical space was mapped using a t-distributed Stochastic Neighbor Embedding (t-SNE) analysis. t-SNE was performed with the scikit-learn python package (v0.20.2) (Pedregosa et al., 2011) using default parameters (perplexity of 30, early exaggeration of 12, learning rate of 200 and 1000 iterations) except for the embedding initialization which was done with principal component analysis.

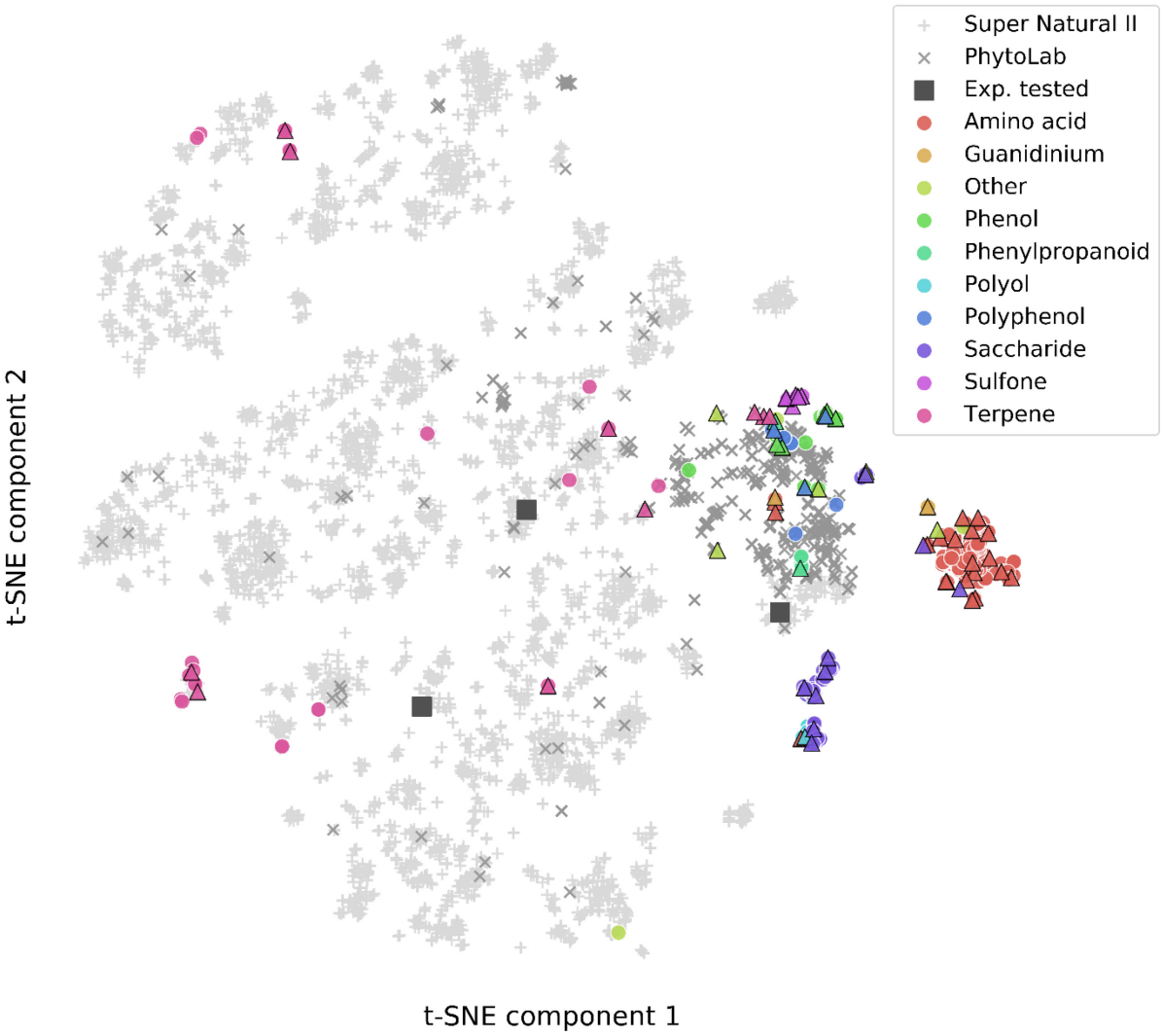


Figure 1: Representation of the SweetenersDB chemical space based on a t-SNE dimensionality reduction method. Known sweet chemical families in the training and test set are represented by circles and triangles, respectively. Light and dark grey data points represent natural compounds that were predicted as intensely sweet (logSw ≥ 2) by both our previous and current models (Table S2). Grey squares represent natural molecules experimentally tested in the present study.

Machine-learning model for sweetness prediction

Several regression algorithms from the python package scikit-learn were evaluated: Random Forest, Support Vector Machine (SVM), Adaptative Boosting with a Decision Tree base estimator (AdaBoost Tree), and k-Nearest Neighbors. Five-fold cross validation was performed with hyperparameter tuning using a grid search. The workflow for each cross-validation fold was as follow: standardization of descriptors, feature selection, and model training. Selection of descriptors was done by keeping a given percentile of the highest ranked descriptors based

on their Mutual Information with our endpoint. The optimal percentile of features was tuned as a parameter of the Grid Search.

Once optimal hyperparameters were found for each model, final models were trained using the full training dataset. Their predictive performance was evaluated based on criteria previously defined by Golbraikh and Tropsha (Golbraikh & Tropsha, 2002). For the "Dragon" models, only the SVM model did not pass all criteria, and for the "open source" model, only the AdaBoost Tree passed all criteria. In both cases, the AdaBoost Tree model was selected as the best performing model, using 32 descriptors for the "Dragon" model, and 51 descriptors for the "open source" model (Figure S2 and Table S4). A summary of their performances is reported in the results section (Table 1) and detailed in supporting information (Table S3).

In addition to training and validating several models for sweetness prediction, a web server implementing the "open-source" model was developed and is freely available at the following address: http://chemosimserver.unice.fr/predisweet/

Other chemoinformatics solutions are available but none of them has been implemented on a webserver. For instance, the e-Sweet platform (Zheng et al., 2019) is based on a consensus model of various machine learning protocols. The database used to train and test their model is very similar to the database used to setup Predisweet and e-Sweet performs as well as our model (R^2 on the test set is in the same range [0.75-0.78] for both solutions). Recently a new functionality to predict sweetness has been implemented on the BitterSweet webserver (Tuwani et al., 2019). The performance of BitterSweet is comparable to e-Sweet and Predisweet (R^2 of 0.72 on our test set) but the protocol is still unpublished, and seven molecules of the test set has not been considered as sweet.

Webserver interface

The user is asked for one or several molecules which can either be drawn directly on the chemical structure editor Ketcher or inputted as a simple text query or file in the SMILES format. The workflow (Figure 2) followed by query compounds is the same as used during model development. First, a molecule is generated from the SMILES string with RDKit to assess its sanity. The structure is then standardized using the "standardizer" Python module. The 51 molecular descriptors selected during model development are computed and standardized based on the training set transformations. The descriptors are passed to the AdaBoost Tree model in order to predict the logSw. Finally, the quality of each prediction is assessed based on three metrics, namely the applicability, reliability, and decidability domains

(Hanser, Barber, Marchaland, & Werner, 2016). The applicability domain indicates if the compound is within the descriptor range of the training set and its score is computed using a convex hull approach. The reliability domain highlights the density of information around the compound. The reliability score is calculated by counting the number of molecules from the training set that are inside a sphere centered on the query. The decidability domain shows the confidence in the prediction that was made. The decidability score is based on the weights of each decision tree that compose the AdaBoost model. It is computed by summing the weights of decision trees that made a prediction close to the model prediction and dividing it by the sum of all weights.

Each molecule is indexed in the database with its InChIKey, which avoids making predictions for the same molecule twice. For a seamless user experience, the name of each molecule is retrieved by querying PubChem with the pubchempy Python package, and a 2D representation of the compound is generated with RDKit.

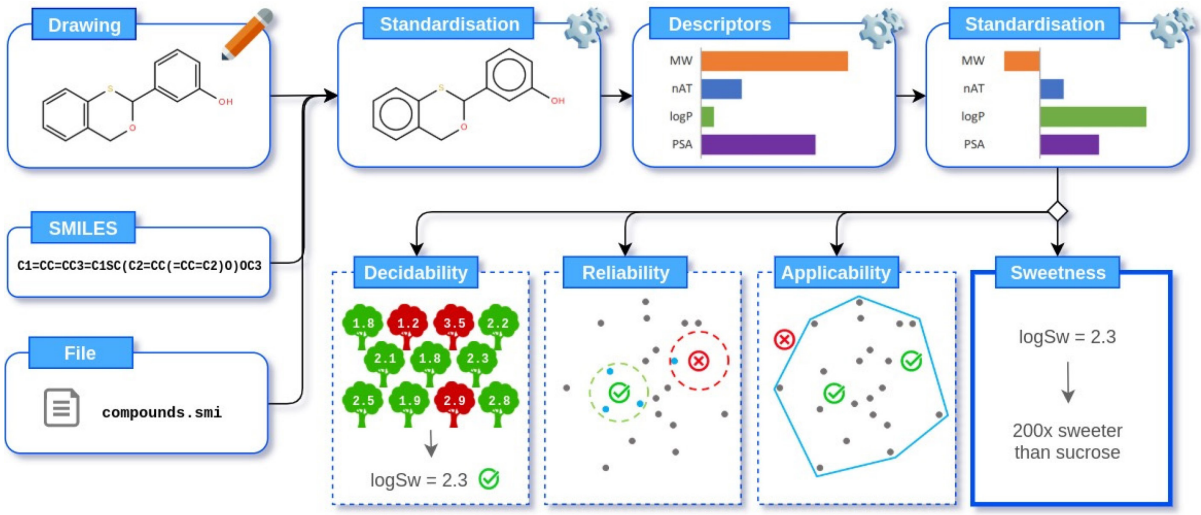


Figure 2: Workflow followed by each molecule submitted to the webserver.

Functional expression of the human sweet taste receptor

In order to validate the sweetness of the three natural compounds, we employed a cell-based expression system for the human T1R2/T1R3 sweet taste receptor as previously described (Poirier et al., 2012; Sigoillot et al., 2018). Briefly, the cDNAs coding human T1R2 and T1R3 subunits were cloned into pcDNA3 and pcDNA4 expression plasmids, respectively. HEK293T cells stably expressing Gα16gust44 and T1R3 were seeded at a density of 0.4 $\times 10^6$ cells per

well into 96-well black walled, clear bottom microtiter plates (Falcon) in high-glucose DMEM supplemented with 2 mM GlutaMAX, 10% dialyzed foetal bovine serum, penicillin/streptomycin, G418 (400 µg/mL) and zeocin (250 µg/mL) at 37 °C and 6.3% CO_2, in a humidified atmosphere. Twenty-four hours later, HEK293T-Gα16gust44-T1R3 cells were transiently transfected with pcDNA3-T1R2 (120ng/well) with Lipofectamine 2000. Calcium signal of mock-transfected cells (HEK293T Gα16gust44 cells stably expressing T1R3 transfected with pcDNA3 empty vector) were always measured in parallel and compared. Twenty-four hours after transfection, the cells were loaded for 1 hour at 37°C with the calcium indicator Fluo4-AM (Molecular Probes) diluted in C1 buffer (130 mM NaCl, 5 mM KCl, 10 mM Hepes pH 7.4, 2 mM $CaCl_2$) in the presence of pluronic acid (0.025%, w/v) and probenecid (2.5 mM). After washing with C1 buffer, cells were stimulated with a range of sweet tasting compounds. The fluorescence intensity was measured for 90 seconds (excitation 488 nm, emission 510 nm) into an automated fluorimetric FlexStation®3 Multi-Mode microplate reader. The change in fluorescence upon stimulus application were averaged, mock-substracted and baseline-corrected. The EC_{50} values were calculated using SigmaPlot software by nonlinear regression using the function:

$$f(x) = min + \frac{max - min}{1 + \left(\frac{x}{EC_{50}}\right)^{-\text{Hillslope}}}$$

Chemicals

All tested compounds (arctiin, ginsenoside Rd and jujuboside A, Figure 3) were purchased from Phytolab GmbH & Co. KG, with the exception of sucralose obtained from Sigma-Aldrich. All the compounds were dissolved first in DMSO (100 mM in 100% DMSO), and then diluted with the C1 buffer solution; except for sucralose, which was dissolved in the C1 buffer solution directly.

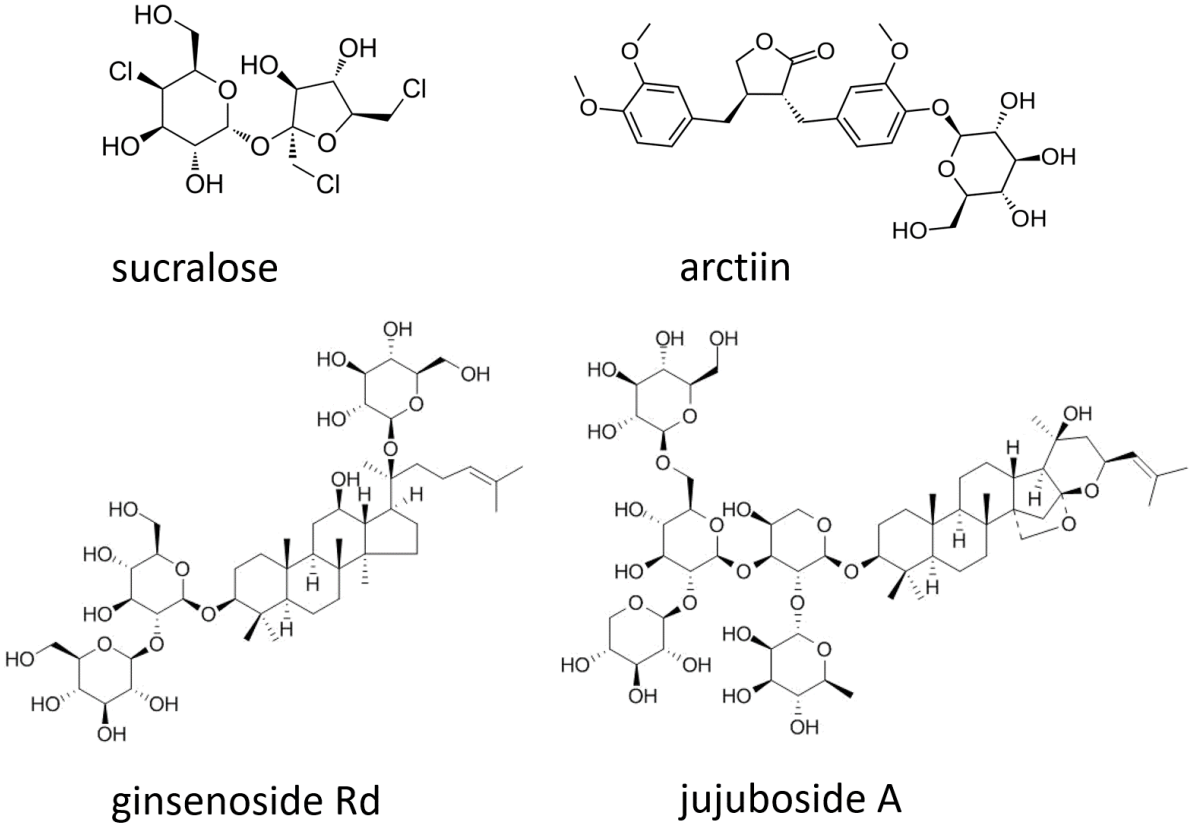


Figure 3: Structure of the tested compounds

Results and discussion

New machine-learning model based on open-source features

The performance of the Open-source and Dragon models has been compared. Both models show good predictivity on the test set according to state-of-the-art QSAR rules (Table 1). Slightly more than 90% of the test set are predicted with an absolute error lower than a log unit (Figure S3). The models are less accurate for high sweetness values since they have been trained with less information for highly potent sweeteners. Improving the quality of the machine learning model would then requires i) expanding the chemical diversity of sweet compounds and ii) a larger database of *in vivo* and *in vitro* experiments. A threshold of LogSw larger than 2 has then been chosen to minimize false positive predictions prior *in vitro* validation. Since similar performance have been obtained for both models, the open-source version has been implemented on a webserver, freely accessible at the following address: http://chemosimserver.unice.fr/predisweet/. Another model has been set up with descriptors

calculated at salivary pH to assess the effect of the protonation state on the model performance. Even though more than a quarter of the molecules had different descriptor values between the default and the salivary pH dataset, there was no significant difference in terms of performance. The protonation assessment step thus has been skipped in the final protocol. We emphasize that the model has not been trained to predict bitter taste and we envision to include this feature in a future work. Additionally, any QSAR model has a field of application that clearly defines the boundaries within which the model should be used, usually referred to as the applicability domain. We have implemented three different metrics to explicitly inform the user whether the model and its prediction can be trusted for a particular query molecule.

Table 1: Performance of the models according to Golbraikh and Tropsha rules. (Golbraikh & Tropsha, 2002)

Rules	Open-source model	Dragon model
$R^2 > 0.6$	0.74	0.75
$Q^2 > 0.5$	0.84	0.79
$\|R^2 - R_0^2\|/R^2 < 0.1$	0.02	0.05
$0.85 <= k <= 1.15$	0.93	0.90
$\|R_0^2 - R'_0^2\| < 0.3$	0.07	0.12

Identification of a new sweet scaffold

A large database of natural compounds has been virtually screened to identify new putative sweeteners. The analysis of the resulting sweet chemical space of ~4800 natural compounds shows that it does not fully overlap the chemical space of known sweeteners (Figure 1). It suggests that a large part of the natural chemical space remains unexplored. We have finally selected three natural compounds that have been tested for their ability to activate the human sweet taste receptor T1R2/T1R3 expressed in HEK cells, as previously reported (Poirier et al., 2012). As a negative control, HEK293T Gα16gust44 cells stably expressing T1R3 were mock-transfected with the empty expression vector to control for T1R2-independent non-specific signals. In addition to a LogSw value higher than 2, the price and the commercial availability

were two important criteria in the compound choice. Two of them, Jujuboside A and Ginsenisode Rd, belong to the triterpene chemical family. The third one, arctiin, possesses a lignan scaffold. As shown in Figure 4b, application of arctiin on T1R2/T1R3-expressing cells evoked calcium responses in a dose-dependent manner, while no fluorescence signals were observed with mock transfected cells. The half-maximal effective concentrations (EC_{50}) of arctiin was 2.5 ± 0.4 mM. As a control, we determined the concentration-response curve for the high-intensity sucralose (Figure 4a) leading to an EC_{50} value of 87 ± 13 µM, in agreement with reported values (Assadi-Porter et al., 2010; Masuda et al., 2012; Servant et al., 2010). In contrast, jujuboside A and ginsenisode Rd showed detectable activity on the T1R2/T1R3 receptor, but only at the highest tested concentration (Figure 4c and d) precluding establishment of complete dose–response curve and calculation of EC_{50} values. This concentration used was the maximum one that did not induce any side effects on mock transfected cells.

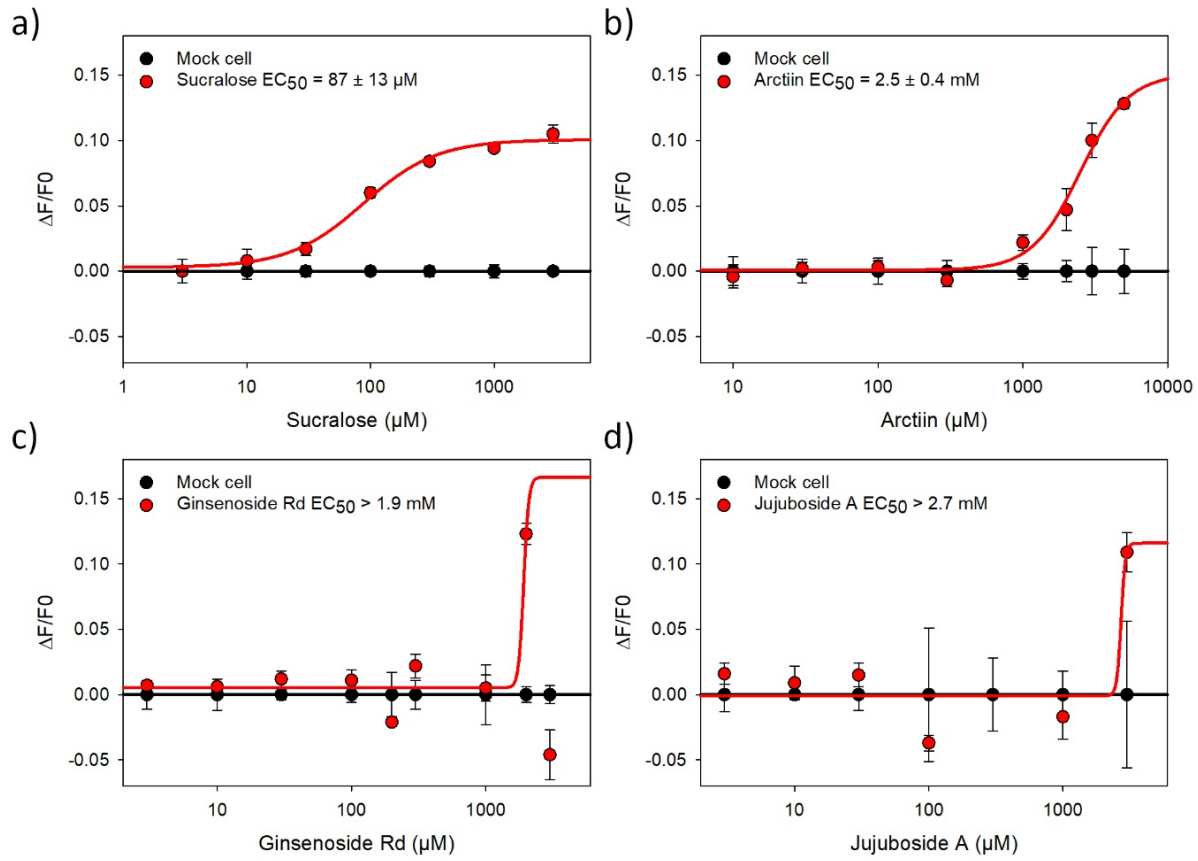


Figure 4: Response of the human sweet taste receptor to the three natural compounds identified by the machine learning protocol and sucralose used as a control. Dose-response curves of T1R2/T1R3-expressing cells (red curve) and mock-transfected cells (black curve). All concentrations were measured in triplicate and each experiment was repeated at least 2 times.

Conclusion

In this study we have used machine learning to predict novel agonists of the sweet taste receptor. An AdaBoost Tree model was setup based on open-source chemical features optimized on a curated database of 316 known sweet agents (SweetenersDB) and implemented on a freely available webserver. The virtual screening of a large database of natural compounds identified thousands of putative sweeteners, of which three were selected for *in vitro* functional assays of the human sweet taste receptor and dose-response analyses. Among them, we identified arctiin as a novel agonist of the T1R2/T1R3 sweet taste receptor with an EC_{50} value of 2.5±0.4mM. It belongs to the lignan chemical family, polyphenols found in plants, of which epi-lyoniresinol has already been annotated as slightly sweet by sensory analyses (Cretin et al., 2015; Marchal, Cretin, Sindt, Waffo-Téguo, & Dubourdieu, 2015). As numerous natural sweeteners, arctiin might also possess bitter taste but it would require additional experiments out of the scope of the present study to assess its aftertaste. Nevertheless, our results confirm that the lignan chemical family opens a new chemical space for the search of new sweet agents and machine learning is a fruitful approach in this context.

Acknowledgements

This work was supported by the French Ministry of Higher Education and Research [PhD Fellowship], by GIRACT (Geneva, Switzerland) [9th European PhD in Flavor Research Bursaries for first year students] and the Gen Foundation (Registered UK Charity No. 1071026) [a charitable trust which principally provides grants to students/researchers in natural sciences, in particular food sciences/technology]. We also benefited from funding from the French government, through the UCAJEDI “Investments in the Future” project managed by the ANR grant No. ANR-15-IDEX-01.

Author contributions

Cédric Bouysset: Data curation, Software, Investigation, Writing - review & editing. Christine Belloir: Investigation, Writing - review & editing. Serge Antonczak: Conceptualization, Writing - review & editing. Loïc Briand: Funding acquisition, Conceptualization, Methodology, Writing - review & editing, Validation. Sébastien Fiorucci: Funding acquisition, Conceptualization, Methodology, Writing - review & editing, Validation.

References

Achary, P. G. R., Toropova, A. P., & Toropov, A. A. (2019). Combinations of graph invariants and attributes of simplified molecular input-line entry system (SMILES) to build up models for sweetness. *Food Research International*, *122*, 40–46. https://doi.org/10.1016/j.foodres.2019.03.067

Arnoldi, A., Bassoli, A., Merlini, L., & Ragg, E. (1991). Isovanillyl sweeteners. Synthesis, conformational analysis, and structure–activity relationship of some sweet oxygen heterocycles. *J. Chem. Soc., Perkin Trans. 2*, (9), 1399–1406. https://doi.org/10.1039/P29910001399

Assadi-Porter, F. M., Maillet, E. L., Radek, J. T., Quijada, J., Markley, J. L., & Max, M. (2010). Key Amino Acid Residues Involved in Multi-Point Binding Interactions between Brazzein, a Sweet Protein, and the T1R2-T1R3 Human Sweet Receptor. *Journal of Molecular Biology*, *398*(4), 584–599. https://doi.org/10.1016/j.jmb.2010.03.017

Barker, J. S., Hattotuwagama, C. K., & Drew, M. G. B. (2002). Computational studies of sweet-tasting molecules. *Pure and Applied Chemistry*, *74*(7), 1207–1217. https://doi.org/10.1351/pac200274071207

Bassoli, A., Drew, M. G. B., Hattotuwagama, C. K., Merlini, L., Morini, G., & Wilden, G. R. H. (2001). Quantitative Structure-Activity Relationships of Sweet Isovanillyl Derivatives. *Quantitative Structure-Activity Relationship*, *20*(1), 3–16. https://doi.org/10.1002/1521-3838(200105)20:1<3::AID-QSAR3>3.0.CO;2-H

Belloir, C., Neiers, F., & Briand, L. (2017). Sweeteners and sweetness enhancers. *Current Opinion in Clinical Nutrition and Metabolic Care*, *20*(4), 279–285. https://doi.org/10.1097/MCO.0000000000000377

Cao, D. S., Xu, Q. S., Hu, Q. N., & Liang, Y. Z. (2013). ChemoPy: Freely available python package for computational biology and chemoinformatics. *Bioinformatics*, *29*(8), 1092–1094. https://doi.org/10.1093/bioinformatics/btt105

ChemAxon. (2018). *Calculator Plugins*. Retrieved from http://www.chemaxon.com

Chéron, J. B., Casciuc, I., Golebiowski, J., Antonczak, S., & Fiorucci, S. (2017). Sweetness prediction of natural compounds. *Food Chemistry*, *221*, 1421–1425. https://doi.org/10.1016/j.foodchem.2016.10.145

Cretin, B. N., Sallembien, Q., Sindt, L., Daugey, N., Buffeteau, T., Waffo-Teguo, P., … Marchal, A. (2015). How stereochemistry influences the taste of wine: Isolation, characterization and sensory evaluation of lyoniresinol stereoisomers. *Analytica Chimica Acta*, *888*, 191–198. https://doi.org/10.1016/j.aca.2015.06.061

Deutsch, E. W., & Hansch, C. (1966). Dependence of relative sweetness on hydrophobic bonding [22]. *Nature*, Vol. 211, p. 75. https://doi.org/10.1038/211075a0

Drew, M. G. B., Wilden, G. R. H., Spillane, W. J., Walsh, R. M., Ryder, C. A., & Simmie, J. M. (1998). Quantitative Structure−Activity Relationship Studies of Sulfamates RNHSO $_3$ Na: Distinction between Sweet, Sweet-Bitter, and Bitter Molecules. *Journal of Agricultural and Food Chemistry*, *46*(8), 3016–3026. https://doi.org/10.1021/jf980095c

EMBL-EBI. (2017). *standardiser*. Retrieved from https://github.com/flatkinson/standardiser

Golbraikh, A., & Tropsha, A. (2002). Beware of q^2! *Journal of Molecular Graphics and Modelling*, *20*(4), 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1

Hanser, T., Barber, C., Marchaland, J. F., & Werner, S. (2016). Applicability domain: towards a more formal definition. *SAR and QSAR in Environmental Research*, *27*(11), 893–909. https://doi.org/10.1080/1062936X.2016.1250229

Landrum, G., Kelley, B., Tosco, P., sriniker, gedeck, NadineSchneider, … Avery, P. (2018, April 20). *rdkit/rdkit: 2018_03_1 (Q1 2018) Release*. https://doi.org/https://doi.org/10.5281/zenodo.1222070

Li, X., Staszewski, L., Xu, H., Durick, K., Zoller, M., & Adler, E. (2002). Human receptors for sweet and umami taste. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(7), 4692–4696. https://doi.org/10.1073/pnas.072090199

Marchal, A., Cretin, B. N., Sindt, L., Waffo-Téguo, P., & Dubourdieu, D. (2015). Contribution of oak lignans to wine taste: Chemical identification, sensory characterization and quantification. *Tetrahedron*, *71*(20), 3148–3156. https://doi.org/10.1016/j.tet.2014.07.090

Masuda, K., Koizumi, A., Nakajima, K., Tanaka, T., Abe, K., Misaka, T., & Ishiguro, M. (2012). Characterization of the Modes of Binding between Human Sweet Taste Receptor and Low-Molecular-Weight Sweet Compounds. *PLoS ONE*, *7*(4), e35380. https://doi.org/10.1371/journal.pone.0035380

Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1). https://doi.org/10.1186/s13321-018-0258-y

Nelson, G., Hoon, M. A., Chandrashekar, J., Zhang, Y., Ryba, N. J. P., & Zuker, C. S. (2001). Mammalian sweet taste receptors. *Cell*, *106*(3), 381–390. https://doi.org/10.1016/S0092-8674(01)00451-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poirier, N., Roudnitzky, N., Brockhoff, A., Belloir, C., Maison, M., Thomas-Danguin, T., … Briand, L. (2012). Efficient Production and Characterization of the Sweet-Tasting Brazzein Secreted by the Yeast Pichia pastoris. *Journal of Agricultural and Food Chemistry*, *60*(39), 9807–9814. https://doi.org/10.1021/jf301600m

Rojas, C., Tripaldi, P., & Duchowicz, P. R. (2016). A New QSPR Study on Relative Sweetness. *International Journal of Quantitative Structure-Property Relationships*, *1*(1), 78–93. https://doi.org/10.4018/ijqspr.2016010104

Ruiz-Aceituno, L., Hernandez-Hernandez, O., Kolida, S., Moreno, F. J., & Methven, L. (2018). Sweetness and sensory properties of commercial and novel oligosaccharides of prebiotic potential. *Lwt*, *97*(April), 476–482. https://doi.org/10.1016/j.lwt.2018.07.038

Servant, G., Tachdjian, C., Tang, X. Q., Werner, S., Zhang, F., Li, X., … Karanewsky, D. S. (2010). Positive allosteric modulators of the human sweet taste receptor enhance sweet taste. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4746–4751. https://doi.org/10.1073/pnas.0911670107

Shallenberger, R. S., & Acree, T. E. (1967). Molecular theory of sweet taste [16]. *Nature*, Vol. 216, pp. 480–482. https://doi.org/10.1038/216480a0

Sigoillot, M., Brockhoff, A., Neiers, F., Poirier, N., Belloir, C., Legrand, P., … Briand, L. (2018). The Crystal Structure of Gurmarin, a Sweet Taste–Suppressing Protein: Identification of the Amino Acid Residues Essential for Inhibition. *Chemical Senses*, *43*(8), 635–643. https://doi.org/10.1093/chemse/bjy054

Spillane, W. J., & McGlinchey, G. (1981). Structure—activity studies on sulfamate sweeteners II: Semiquantitative structure-taste relationship for sulfamate (rnhso 3−) sweeteners—the role of R. *Journal of Pharmaceutical Sciences*, *70*(8), 933–935. https://doi.org/10.1002/jps.2600700826

Spillane, W. J., McGlinchey, G., Muircheartaigh, I., & Benson, G. A. (1983). Structure–activity

studies on sulfamate sweetners III: Structure–taste relationships for heterosulfamates. *Journal of Pharmaceutical Sciences*, *72*(8), 852–856. https://doi.org/10.1002/jps.2600720804

Spillane, W. J., Ryder, C. A., Curran, P. J., Wall, S. N., Kelly, L. M., Feeney, B. G., & Newell, J. (2000). Development of structure–taste relationships for sweet and non-sweet heterosulfamates†. *Journal of the Chemical Society, Perkin Transactions 2*, (7), 1369–1374. https://doi.org/10.1039/b0024821

Spillane, W. J., Ryder, C. A., Walsh, M. R., Curran, P. J., Concagh, D. G., & Wall, S. N. (1996). Sulfamate sweeteners. *Food Chemistry*, *56*(3), 255–261. https://doi.org/10.1016/0308-8146(96)00022-2

Spillane, W. J., & Sheahan, M. B. (1989). Semi-quantitative and quantitative structure–taste relationships for carboand hetero-sulphamate ($RNHSO_3^-$) sweeteners. *J. Chem. Soc., Perkin Trans. 2*, (7), 741–746. https://doi.org/10.1039/P29890000741

Talete srl. (2014). *Dragon (Software for Molecular Descriptor Calculation)*.

Tuwani, R., Wadhwa, S., & Bagler, G. (2019). BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-43664-y

Van Der Heijden, A., Brussel, L. B. P., & Peer, H. G. (1979). Quantitative structure-activity relationships (QSAR) in sweet aspartyl dipeptide methyl esters. *Chemical Senses*, *4*(2), 141–152. https://doi.org/10.1093/chemse/4.2.141

Vepuri, S. B., Tawari, N. R., & Degani, M. S. (2007). Quantitative structure-activity relationship study of some aspartic acid analogues to correlate and predict their sweetness potency. *QSAR and Combinatorial Science*, *26*(2), 204–214. https://doi.org/10.1002/qsar.200530191

Walters, D. E. (2006). Analysing and predicting properties of sweet-tasting compounds. In *Optimising Sweet Taste in Foods* (pp. 283–291). https://doi.org/10.1533/9781845691646.3.283

Zheng, S., Chang, W., Xu, W., Xu, Y., & Lin, F. (2019). e-Sweet: A Machine-Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness. *Frontiers in Chemistry*, *7*(JAN), 35. https://doi.org/10.3389/fchem.2019.00035

Supporting information

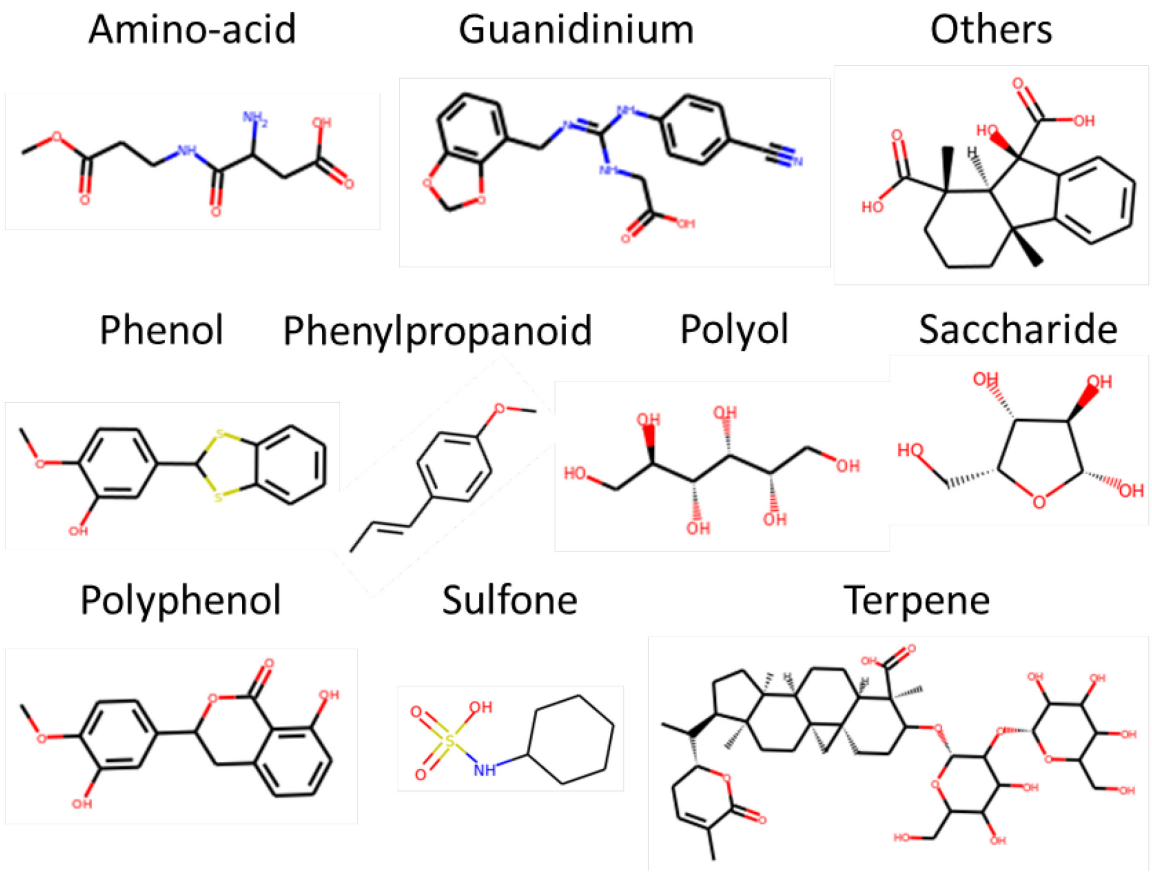


Figure S1: Chemical families present in the database.

Table S1: list of molecules in SweetnersDB v2.0

See https://ars.els-cdn.com/content/image/1-s2.0-S0308814620307263-mmc2.xlsx

Table S2: list of natural compounds used to map the chemical space of sweeteners (Figure 1)

See https://ars.els-cdn.com/content/image/1-s2.0-S0308814620307263-mmc4.xlsx

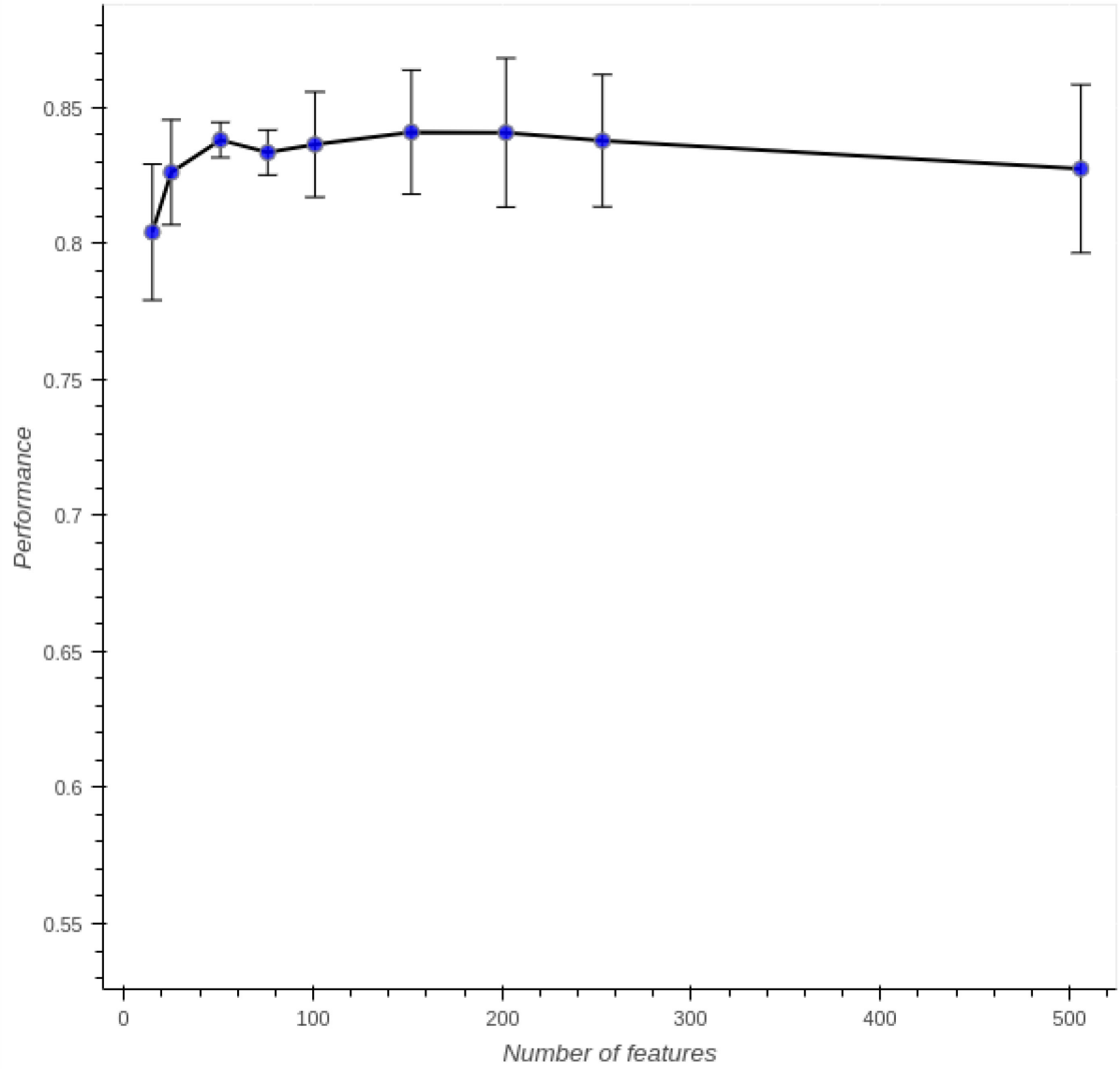


Figure S2: Performance of the "open-source" AdaBoost Tree model for different number of features obtained during cross-validation on the training set (where Performance is the coefficient of determination R^2).

Table S3: Evaluation metrics for the performance of the "open-source" and "Dragon" AdaBoost Tree models.

Model	Dataset	R	R^2	$\frac{\lvert R^2 - R_0^2 \rvert}{R^2}$	k	R_0^2	$\frac{\lvert R^2 - R'^2_0 \rvert}{R^2}$	k'	R'^2_0	RMSE	MAE
Open-source	Training	0.997	0.995	0.000	1.001	0.995	0.000	0.997	0.995	0.084	0.050
	LOO	0.914	0.835	0.041	1.003	0.801	0.014	0.934	0.824	0.467	0.344
	Test	0.858	0.737	0.104	0.989	0.660	0.015	0.925	0.725	0.666	0.496
Dragon	Training	0.998	0.996	0.000	0.999	0.996	0.000	0.999	0.996	0.072	0.024
	LOO	0.888	0.789	0.058	1.000	0.743	0.020	0.919	0.773	0.529	0.376
	Test	0.867	0.749	0.212	1.022	0.590	0.052	0.900	0.709	0.651	0.498

LOO: Leave-One-Out, R: correlation coefficient, R^2: coefficient of determination, k and k': slopes of the regression lines through the origin for the observed vs. predicted and predicted vs. observed values respectively, R_0^2 and R'^2_0: corresponding coefficients of determination, RMSE: root mean squared error, MAE: mean absolute error

Table S4: Descriptors used by each model

Model	Number of descriptors	Descriptors
Open-source	51	rdkit_BertzCT rdkit_EState_VSA10 rdkit_HallKierAlpha rdkit_MaxAbsEStateIndex rdkit_MaxPartialCharge rdkit_MinEStateIndex mordred_ATS0Z mordred_AATS4d mordred_AATS0p mordred_AATS1p mordred_AATS5p mordred_ATSC2c mordred_ATSC3c mordred_ATSC1dv mordred_ATSC2s mordred_AATSC2c mordred_AATSC3c mordred_AATSC1dv mordred_AATSC2dv mordred_AATSC2s mordred_AATSC3s mordred_AATSC1Z mordred_AATSC0v mordred_AATSC0p mordred_AATSC0i mordred_AATSC2i mordred_MATS1c mordred_MATS1s mordred_GATS1dv mordred_GATS1s mordred_GATS1se mordred_GATS1p mordred_GATS2p mordred_GATS2i mordred_BCUTc-1h mordred_AXp-1d mordred_AXp-2d mordred_AETA_alpha mordred_ETA_dAlpha_B mordred_ETA_dEpsilon_D mordred_ETA_psi_1 mordred_AMID_O mordred_RotRatio chemopy_GATSp2 chemopy_IC1 chemopy_MATSm2 chemopy_MATSm5 chemopy_MATSp2 chemopy_bcute1 chemopy_bcute2 chemopy_bcutm2
Dragon	32	Mp C% MAXDP piPC05 piPC08 piPC10 piID X1A X4A ChiA_Dt AVS_B(m) SpMaxA_B(m) AVS_B(v) SpPosA_B(v) SpDiam_B(v) MATS1m MATS2e MATS3e GATS1e GATS2p GATS2i GATS1s GATS2s SpMax2_Bh(m) SpMax2_Bh(v) P_VSA_v_3 P_VSA_e_2 P_VSA_i_2 Eta_alpha_A SM12_AEA(ri) CATS2D_03_AL PDI

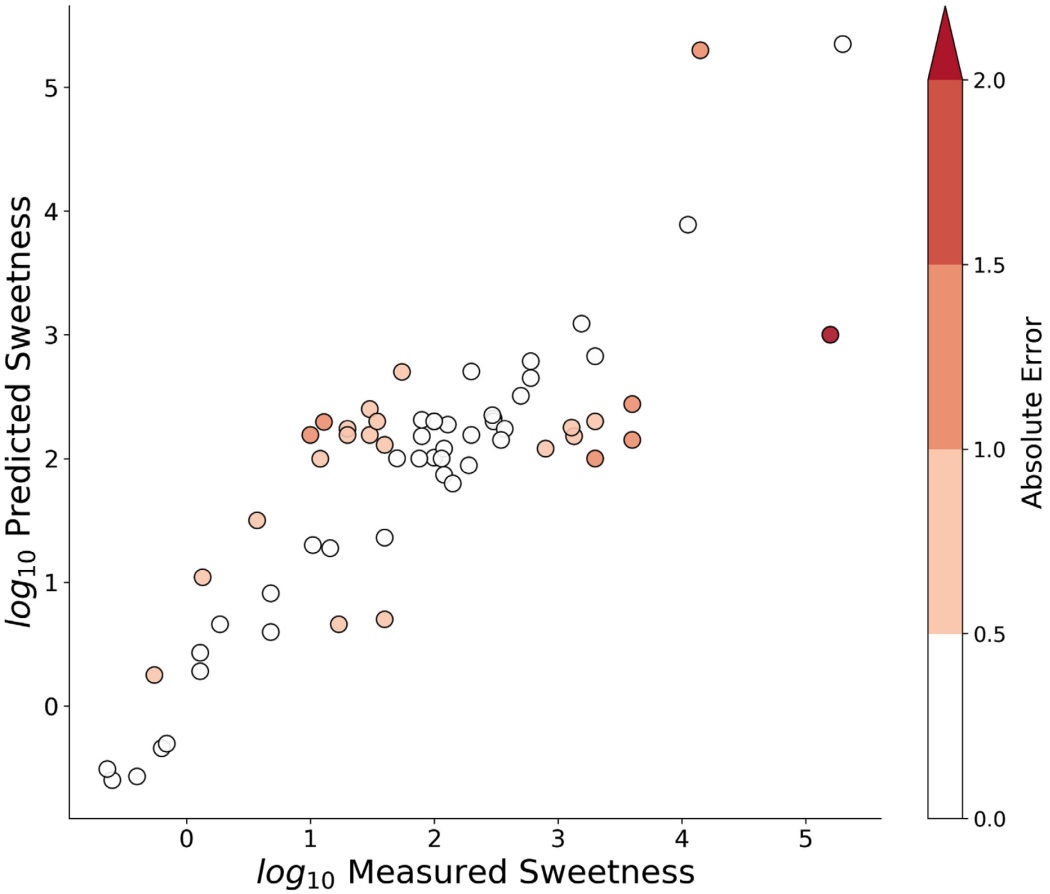


Figure S3: Predicted and observed logSw values on the test set (predictions from the "open-source" AdaBoost Tree model).

Table S5: Experimental EC_{50} of sweet taste receptor ligands and predicted sweetness values using various machine learning models.

Molecule	**Measured EC_{50}**	**Predicted logSw**		
		PrediSweet[b]	**e-Sweet[c]**	**BitterSweet[d]**
Sucralose (control)	87 ± 13 µM	2.78	2.69	2.28
Arctiin	2.5 ± 0.4 mM	2.78	1.81	bitter
Ginsenoside Rd	> 1.9 mM[a]	2.51	2.34	2.30
Jujuboside A	> 2.7 mM[a]	2.57	2.30	neither sweet nor bitter

a: maximum concentration (showing detectable activity) that did not induce any side effects on mock transfected cells. b: using the "open-source" model described in the present study. c: using the e-Sweet software (Zheng et al., 2019) and consensus model CM01. d: using the BitterSweet webserver (Tuwani et al., 2019). The results from the BitterSweet webserver must be considered with care. Tuwani et al., 2019 refers to the "bitter *vs* sweet" classification model and not to the sweetness regression model. The latter has not been published to date.

Chapter III

Structure-function relationships for bitter taste receptors

Taste perception allows humans, and in general most vertebrates, to appraise the nutritive value of food through five basic taste modalities: salty, sour, sweet, umami, and bitter. From an evolutionary point of view, bitter taste is believed to have had a role in protecting living beings from ingesting toxic compounds present in food items by triggering an aversive behavior [1]. In fact, more than 60% of bitter compounds reported in the BitterDB [2], a database of bitter molecules, are toxic [3]. However, the story is not as simple as "toxic compounds are bitter". Bitter taste sensitivity depends on the occurrence of bitter and toxic compounds in an animal's diet. For instance, carnivores can afford lower bitter taste sensitivity thresholds than herbivores since comparatively, they are not often in contact with potentially toxic compounds which mostly come from plants, but a more sensitive bitter taste would be too restrictive on an herbivore's diet and would become a handicap [1, 4]. In contrast with the idea of bitter taste as a warning system, a cross-cultural tendency to seek bitter medicine when signs of illness appear is observed, as bitter substances can also suggest pharmacological activity since many drugs are bitter [5]. This is even the case for animals since chimpanzees [6], mice [7] and ruminants [8] will actively search for bitter-tasting plants or solutions when infected by a parasite, and more surprisingly, when they are healthy. Bitter taste could then be not exclusively a marker of toxicity but of pharmacological activity i.e., in a beneficial or harmful way, and help to control the intake amount.

The receptors responsible for bitter taste perception, the taste receptors type 2 (TAS2Rs), are expressed in type II cells present in taste buds. These receptors belong to the GPCR family [9, 10], just as taste receptors type 1 (T1Rs) which are responsible for sweet and umami taste perception. Type II cells can also express T1R receptors, but most type II cells will only express one class of taste GPCR, either T1R or TAS2R [11]. While it has been shown that TAS2Rs can oligomerize as both homo or heterodimers, no functional consequence was found, neither as agonists that would bind specifically to the heteromer, nor by having a different localization on the plasma membrane, nor by displaying varied pharmacological properties [12]. TAS2Rs also appear to be glycoproteins as glycosylation of an asparagine of the second extracellular loop (ECL2), which is conserved in the entire mammalian repertoire, is important for protein maturation and membrane insertion but not for their function as it can be rescued by other means [13]. Despite the existence of several TAS2Rs and a subsequent combinatorial code, this is not sufficient to be able to discriminate the bitter stimuli generated by different ligands [14]. Indeed, while type II taste cells can coexpress different combinations of TAS2Rs [9], and different ligands may activate different subsets of taste cells [15], the differentiated signals converge downstream in the gustatory pathway [16, 17], leading to a single bitter sensation. The human

genome contains 25 functional genes coding for TAS2Rs and 11 pseudogenes [18], while there are 33 functional and 3 pseudo genes in the mouse genome [19]. In general, birds have a smaller functional TAS2R repertoire (none in penguin, 1 in pigeon, 3 in chicken) [20, 21] likely because of their diet, as explained previously. However, a lower number of TAS2Rs isn't necessarily associated with a smaller receptive range for bitter taste, as it may be compensated by a broader tuning width of TAS2Rs to detect more ligands [22]. In humans, there are 4 broadly tuned, 6 narrowly tuned, 3 specific, and 4 orphaned TAS2Rs, while the remaining 8 have an intermediate receptive range [23–27]. Among this last category, we have shown that TAS2R7 can detect metal ions [28] (see Appendix) in addition to organic compounds. While TAS2Rs are undoubtedly GPCRs, their sub-classification is more complicated. Historically, they were first thought to be distantly related to pheromone receptors expressed in the vomeronasal organ [9, 10], then classified with class F GPCRs due to three similar motifs in their consensus sequence [29], but more recent work tags them as related to class A GPCRs [30, 31], or even as their own class T family in the GPCR database (GPCRdb) [32].

Like all GPCRs, TAS2Rs are metabotropic receptors that rely on secondary messengers to convert the chemical signal (binding of a bitter tastant) to an action potential that will be carried to the nervous system (Figure 1).

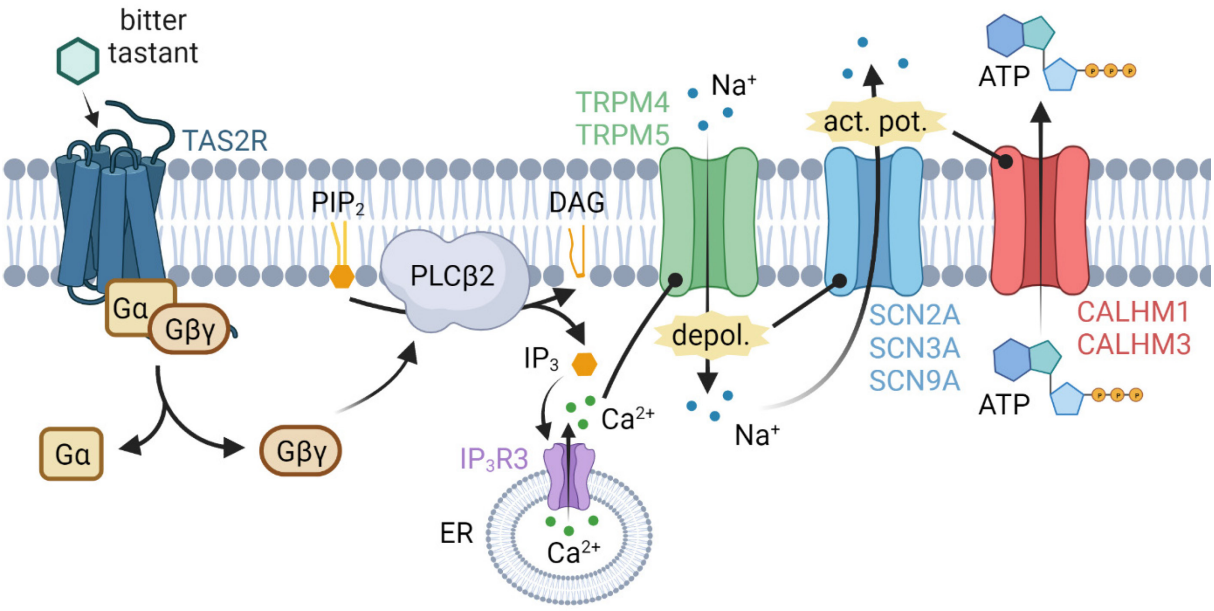


Figure 1: Signal transduction for bitter taste. PIP_2: phosphatidylinositol 4,5-bisphosphate, PLCβ2: phospholipase C beta 2, DAG: diacylglycerol, IP_3: inositol trisphosphate, ER: endoplasmic reticulum, IP_3R3: inositol triphosphate receptor, TRPM4/5: transient receptor potential melastatin, depol.: depolarization, SCN2A/3A/9A: voltage-gated sodium channels, act. pot.: action potential, ATP: adenosine 5'-triphosphate, CALHM1/3: calcium homeostasis modulators.

Briefly, upon ligand binding on a TAS2R, an heterotrimeric G protein dissociates into a Gα-gustducin subunit and a Gβγ complex (Gβ3 and Gγ13) which then activates a cascade of reactions involving a phospholipase C (PLCβ2) and an inositol triphosphate receptor (IP_3R3) to liberate calcium ions in the cytoplasm [33]. The increase in Ca^{2+} level activates members of the transient receptor potential melastatin (TRPM) family which in turn triggers the depolarization of the cell followed by the generation of action potentials by voltage-gated sodium channels (SCN) [34]. This culminates with the release of ATP in the extracellular medium by calcium homeostasis modulators (CALHM), where ATP acts as a neurotransmitter for sensory nerve fibers through the purinergic signaling pathway [35].

While the signal transduction has been extensively described, very little is known about structure-function relationships in these receptors because of the absence of experimental structures. The molecular switches involved in ligand-sensing and G protein signaling mostly remain speculative, yet a deeper understanding could provide guidance for the rational design of pharmacologically active compounds, either as selective agonists targeting ectopically expressed TAS2Rs or as antagonists acting as bitter-taste blockers, using structure-based drug-design approaches. However, the problematic lack of tridimensional structure of TAS2Rs could be addressed using different modeling approaches. The method of choice used in most cases of structure prediction is homology modeling. It allows building a 3D model of a target protein by using another related protein for which a structure is available, called a template, based on a sequence alignment (Figure 2, see Appendix for details). In the case of TAS2Rs, all previous homology modeling attempts used class A GPCRs as templates [36] with a sequence identity below 15% for the transmembrane domains and ranging from 13% to 29% considering all class A GPCRs without structures [31].

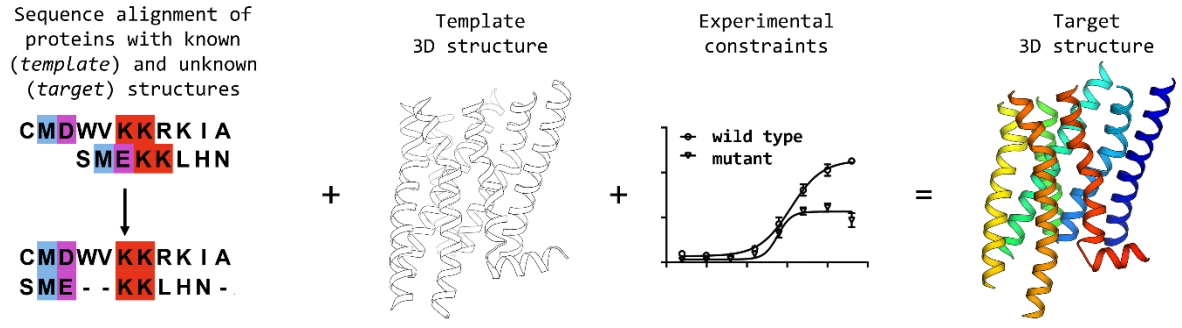


Figure 2: Homology modeling workflow. Single-point mutagenesis data can be used to iteratively improve the produced models by ensuring positions that are involved in binding in a ligand-dependent manner are oriented towards the cavity of the binding pocket.

The main difference between the currently published TAS2R models lies in divergent multiple sequence alignments between targets and templates, highlighting that this process is not straightforward. Alternatively, *ab initio* methods such as GPCR-I-TASSER [37] can readily generate 3D models from the amino-acid sequence of any GPCR by reconstructing transmembrane segments (TMs) individually using replica-exchange Monte Carlo (REMC) simulations. The TMs are then assembled in a second set of REMC simulations using a reduced representation of the TM segments (Cα atoms and sidechain center of mass only) to more efficiently sample the conformational space [38]. However, this *ab initio* framework is only used if homology modeling is not possible because of the absence of a suitable template. More recently, deep-learning-based methods have emerged at the forefront of protein folding. DeepMind, a Google subsidiary, won the 13th CASP, a blind structure prediction challenge, using their AlphaFold model [39] which relies on coevolution to predict distance distributions between all pairs of Cβ atoms in a protein to derive a potential that is minimized by gradient descent to generate a 3D structure [40]. For the following 14th CASP challenge, their AlphaFold 2 model reached a scientific breakthrough with results comparable in some cases to experimental methods using a completely different implementation [41]. While the results are likely less promising for transmembrane proteins since they are underrepresented in the PDB, the technological leap provided by DeepMind will hopefully inspire further progress to complement and improve template-based modeling.

In this chapter, our focus was on providing an integrative protocol that combines homology modeling with single-point mutagenesis data to deliver 3D models that describe the binding pocket of TAS2Rs as accurately as possible. As a follow-up of our modeling efforts of TAS2R7 [28] (see Appendix), our sequence alignment was adjusted by including the entire mammalian TAS2R repertoire to more precisely identify conserved positions, as well as olfactory receptors (which are a subcategory of class A GPCRs) to guide the alignment of ambiguous sequence segments. We then extrapolated the molecular switches of class A GPCRs to this family of taste receptors and validated them with *in vitro* functional assays. This work paves the way for future analysis on the molecular recognition and signal transduction of bitter taste receptors.

Contributions

Appendix publication A1

My role was to predict how the human TAS2R7 interacts with metal ions using a 3D structure of the receptor that I generated by comparative modeling. Site-directed mutagenesis was

performed by our collaborators on four charged or polar residues close to a region with a negatively charged electrostatic potential. Two residues that are critical for the recognition of metal ions by TAS2R7 were identified this way.

Publication 4

Jérémie Topin setup, updated, and curated the database of single-point mutagenesis data, analyzed the results, and supervised the study. I updated and curated the database, built the 3D models of TAS2Rs, designed a custom scoring function to select ideal 3D models, and analyzed the results. Jody Pacalon analyzed the volume and hydrophobicity of TAS2Rs binding pocket. Our collaborators performed *in vitro* assays and analyzed the results. Jérémie Topin and I contributed equally as first authors.

Poster presentations

I presented this work during poster sessions at the European Chemoreception Research Organization (ECRO, 2019) and Groupement de Recherche Odorant Odeur Olfaction (GDR-O3, 2019) annual meetings, and the International Symposium on Olfaction and Taste (ISOT, 2020) and Weurman symposium (2021).

References

1. Meyerhof W (2005) Elucidation of mammalian bitter taste. In: Amara SG, Bamberg E, Grinstein S, et al (eds) Reviews of Physiology, Biochemistry and Pharmacology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 37–72

2. Wiener A, Shudler M, Levit A, Niv MY (2012) BitterDB: a database of bitter compounds. Nucleic Acids Research 40:D413–D419. https://doi.org/10.1093/nar/gkr755

3. Nissim I, Dagan-Wiener A, Niv MY (2017) The taste of toxicity: A quantitative analysis of bitter and toxic molecules: THE TASTE OF TOXICITY. IUBMB Life 69:938–946. https://doi.org/10.1002/iub.1694

4. Glendinning JI (1994) Is the bitter rejection response always adaptive? Physiology & Behavior 56:1217–1227. https://doi.org/10.1016/0031-9384(94)90369-7

5. Mennella JA, Beauchamp GK (2008) Optimizing oral medications for children. Clinical Therapeutics 30:2120–2132. https://doi.org/10.1016/j.clinthera.2008.11.018

6. Koshimizu K, Ohigashi H, Huffman MA (1994) Use of Vernonia amygdalina by wild chimpanzee: Possible roles of its bitter and related constituents. Physiology & Behavior 56:1209–1216. https://doi.org/10.1016/0031-9384(94)90368-9

7. Vitazkova SK, Long E, Paul A, Glendinning JI (2001) Mice suppress malaria infection by sampling a 'bitter' chemotherapy agent. Animal Behaviour 61:887–894. https://doi.org/10.1006/anbe.2000.1677

8. Villalba JJ, Miller J, Ungar ED, et al (2014) Ruminant self-medication against gastrointestinal nematodes: evidence, mechanism, and origins. Parasite 21:31. https://doi.org/10.1051/parasite/2014032

9. Adler E, Hoon MA, Mueller KL, et al (2000) A Novel Family of Mammalian Taste Receptors. Cell 100:693–702. https://doi.org/10.1016/S0092-8674(00)80705-9

10. Matsunami H, Montmayeur J-P, Buck LB (2000) A family of candidate taste receptors in human and mouse. Nature 404:601–604. https://doi.org/10.1038/35007072

11. Roper SD, Chaudhari N (2017) Taste buds: cells, signals and synapses. Nat Rev Neurosci 18:485–497. https://doi.org/10.1038/nrn.2017.68

12. Kuhn C, Bufe B, Batram C, Meyerhof W (2010) Oligomerization of TAS2R Bitter Taste Receptors. Chemical Senses 35:395–406. https://doi.org/10.1093/chemse/bjq027

13. Reichling C, Meyerhof W, Behrens M (2008) Functions of human bitter taste receptors depend on N-glycosylation. Journal of Neurochemistry 106:1138–1148. https://doi.org/10.1111/j.1471-4159.2008.05453.x

14. Spector AC, Kopka SL (2002) Rats Fail to Discriminate Quinine from Denatonium: Implications for the Neural Coding of Bitter-Tasting Compounds. J Neurosci 22:1937–1941. https://doi.org/10.1523/JNEUROSCI.22-05-01937.2002

15. Caicedo A, Roper SD (2001) Taste Receptor Cells That Discriminate Between Bitter Stimuli. Science 291:1557–1560. https://doi.org/10.1126/science.1056670

16. Geran LC, Travers SP (2009) Bitter-Responsive Gustatory Neurons in the Rat Parabrachial Nucleus. Journal of Neurophysiology 101:1598–1612. https://doi.org/10.1152/jn.91168.2008

17. Spector AC, Glendinning JI (2009) Linking peripheral taste processes to behavior. Current Opinion in Neurobiology 19:370–377. https://doi.org/10.1016/j.conb.2009.07.014

18. Go Y, Satta Y, Takenaka O, Takahata N (2005) Lineage-Specific Loss of Function of Bitter Taste Receptor Genes in Humans and Nonhuman Primates. Genetics 170:313–326. https://doi.org/10.1534/genetics.104.037523

19. Shi P, Zhang J, Yang H, Zhang Y (2003) Adaptive Diversification of Bitter Taste Receptor Genes in Mammalian Evolution. Molecular Biology and Evolution 20:805–814. https://doi.org/10.1093/molbev/msg083

20. Zhao H, Li J, Zhang J (2015) Molecular evidence for the loss of three basic tastes in penguins. Current Biology 25:R141–R142. https://doi.org/10.1016/j.cub.2015.01.026

21. Wang K, Zhao H (2015) Birds Generally Carry a Small Repertoire of Bitter Taste Receptor Genes. Genome Biol Evol 7:2705–2715. https://doi.org/10.1093/gbe/evv180

22. Behrens M, Korsching SI, Meyerhof W (2014) Tuning Properties of Avian and Frog Bitter Taste Receptors Dynamically Fit Gene Repertoire sizes. Molecular Biology and Evolution 31:3216–3227. https://doi.org/10.1093/molbev/msu254

23. Meyerhof W, Batram C, Kuhn C, et al (2010) The Molecular Receptive Ranges of Human TAS2R Bitter Taste Receptors. Chemical Senses 35:157–170. https://doi.org/10.1093/chemse/bjp092

24. Thalmann S, Behrens M, Meyerhof W (2013) Major haplotypes of the human bitter taste receptor TAS2R41 encode functional receptors for chloramphenicol. Biochemical and Biophysical Research Communications 435:267–273. https://doi.org/10.1016/j.bbrc.2013.04.066

25. Dotson CD, Zhang L, Xu H, et al (2008) Bitter Taste Receptors Influence Glucose Homeostasis. PLoS ONE 3:e3974. https://doi.org/10.1371/journal.pone.0003974

26. Roland WSU, van Buren L, Gruppen H, et al (2013) Bitter Taste Receptor Activation by Flavonoids and Isoflavonoids: Modeled Structural Requirements for Activation of hTAS2R14 and hTAS2R39. J Agric Food Chem 61:10454–10466. https://doi.org/10.1021/jf403387p

27. Ueno Y, Sakurai T, Okada S, et al (2011) Human Bitter Taste Receptors hTAS2R8 and hTAS2R39 with Differential Functions to Recognize Bitter Peptides. Bioscience, Biotechnology, and Biochemistry 75:1188–1190. https://doi.org/10.1271/bbb.100893

28. Wang Y, Soohoo AL, Lei W, et al (2019) Metal ions activate the human taste receptor TAS2R7. Chemical Senses 44:339–347. https://doi.org/10.1093/chemse/bjz024

29. Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB (2003) The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. Mol Pharmacol 63:1256–1272. https://doi.org/10.1124/mol.63.6.1256

30. Nordstrom KJV, Sallman Almen M, Edstam MM, et al (2011) Independent HHsearch, Needleman-Wunsch-Based, and Motif Analyses Reveal the Overall Hierarchy for Most of the G Protein-Coupled Receptor Families. Molecular Biology and Evolution 28:2471–2480. https://doi.org/10.1093/molbev/msr061

31. Di Pizio A, Levit A, Slutzki M, et al (2016) Comparing Class A GPCRs to bitter taste receptors. In: Methods in Cell Biology. Elsevier Ltd, pp 401–427

32. Isberg V, Mordalski S, Munk C, et al (2016) GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Res 44:D356–D364. https://doi.org/10.1093/nar/gkv1178

33. Behrens M, Ziegler F (2020) Structure-Function Analyses of Human Bitter Taste Receptors—Where Do We Stand? Molecules 25:4423. https://doi.org/10.3390/molecules25194423

34. Kinnamon SC, Finger TE (2019) Recent advances in taste transduction and signaling. F1000Res 8:2117. https://doi.org/10.12688/f1000research.21099.1

35. Finger TE, Danilova V, Barrows J, et al (2005) ATP signaling is crucial for communication from taste buds to gustatory nerves. Science 310:1495–1499. https://doi.org/10.1126/science.1118435

36. Spaggiari G, Di Pizio A, Cozzini P (2020) Sweet, umami and bitter taste receptors: State of the art of in silico molecular modeling approaches. Trends in Food Science & Technology 96:21–29. https://doi.org/10.1016/j.tifs.2019.12.002

37. Zhang J, Yang J, Jang R, Zhang Y (2015) GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome. Structure 23:1538–1549. https://doi.org/10.1016/j.str.2015.06.007

38. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738. https://doi.org/10.1038/nprot.2010.5

39. Senior AW, Evans R, Jumper J, et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710. https://doi.org/10.1038/s41586-019-1923-7

40. Senior AW, Evans R, Jumper J, et al (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins 87:1141–1148. https://doi.org/10.1002/prot.25834

41. Callaway E (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 588:203–204. https://doi.org/10.1038/d41586-020-03348-4

Publication 4

Functional Molecular Switches of Mammalian G Protein-Coupled Bitter-Taste Receptors

Jérémie Topin†*, Cédric Bouysset†, Jody Pacalon, Yiseul Kim, MeeRa Rhyu, Sébastien Fiorucci*, & Jérôme Golebiowski

Under review

doi.org/10.1101/2020.10.23.348706

Abstract

Bitter taste receptors (TAS2Rs) are a poorly understood subgroup of G protein-coupled receptors (GPCRs). The experimental structure of these receptors has yet to be determined, and key-residues controlling their function remain mostly unknown. We designed an integrative approach to improve comparative modeling of TAS2Rs. Using current knowledge on class A GPCRs and existing experimental data in the literature as constraints, we pinpointed conserved motifs to entirely re-align the amino-acid sequences of TAS2Rs. We constructed accurate homology models of human TAS2Rs. As a test case, we examined the accuracy of the TAS2R16 model with site-directed mutagenesis and *in vitro* functional assays. This combination of *in silico* and *in vitro* results clarify sequence-function relationships and identify the functional molecular switches that encode agonist sensing and downstream signaling mechanisms within mammalian TAS2Rs sequences.

Keywords

Bitter taste receptor, GPCR, structure-function relationships, integrative structural biology

Introduction

Bitterness is one of the basic taste modalities detected by the gustatory system. It is generally considered to be a warning against the intake of noxious compounds [1] and, as such, is often associated with disgust and food avoidance [2]. At the molecular level, this perception is initiated by the activation of bitter taste receptors. In humans, 25 genes functionally express these so-called type 2 taste receptors (TAS2Rs), which provide the capacity to detect a wide array of bitter chemicals [3]. Further, TAS2Rs are also ectopically expressed in non-chemosensory tissues, making them important emerging pharmacological targets [4-6].

TAS2Rs are G protein-coupled receptors [7] (GPCRs) classified as distantly related to class A GPCRs. They were previously classified with class F GPCRs [8] and more recently as a separate sixth class evolved from class A [9, 10]. The sequence similarity between TAS2Rs and class A GPCRs is in the range of 14%-29% [11]. Structure-based sequence alignment has placed TAS2Rs in the class A family, which contains the olfactory chemosensory receptors sub-family [12]. TAS2Rs have been recently labelled as class T in the GPCR database (GPCRdb) (Fig. 1a) [13].

Structurally, GPCRs are made up of seven transmembrane (TM) helices named TM1 to TM7 that form a bundle across the cell membrane. How GPCRs achieve specific robust signaling and how these functions are encoded in their sequences are pending fundamental questions. GPCR activation relies on so-called molecular switches, which allosterically connect the ligand binding pocket to the intracellular G protein coupling site in order to trigger downstream signaling [14]. In class A GPCRs (including olfactory receptors, ORs), these molecular switches consist of conserved sequence motifs (Fig. 1c). The “toggle/transmission switch” $CWxP^{TM6}$ (or $FYGx^{TM6}$ in ORs) senses agonist binding. The other motifs, which propagate the signal, include the “hydrophobic connector” $PIF^{TM3\text{-}5\text{-}6}$, the $NPxxY^{TM7}$, the “ionic lock” DRY^{TM3}, and a hydrophobic barrier between the last two [15-18].

To date, experimental structures have not been determined for any TAS2Rs, but the following hallmark motifs have been defined based on sequence conservation: $NGFI^{TM1}$, $LAxSR^{TM2}$, $KIANFS^{TM3}$, LLG^{TM4}, PF^{TM5}, $HxKALKT^{TM6}$, YFL^{TM6}, and $PxxHSFIL^{TM7}$ [7]. These conserved motifs are highly dissimilar between TAS2Rs and class A GPCRs (Fig. 1b,d and Table 1), leading to different sequence alignments. The main discrepancies occur in TM3, TM4, TM6, and TM7 [11, 19-30], making it difficult to infer TAS2R functional molecular switches.

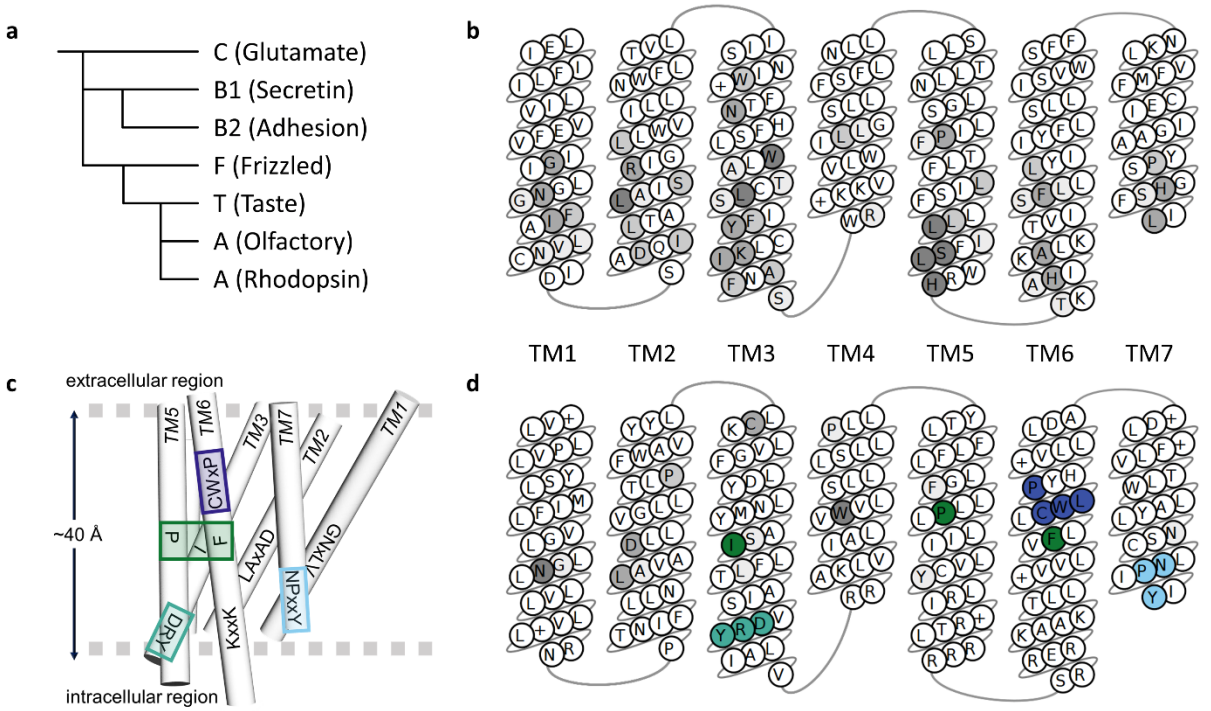


Fig. 1 a Schematic phylogenetic tree of GPCR classes according to Cvicek et al. [12]. **b** Snake plot representation of transmembrane segments (TM) of mammalian TAS2Rs consensus sequences, colored in grey scale according to sequence conservation. **c** Non-olfactory class A GPCR sequence hallmarks (transmission switch in blue, hydrophobic connector in green, ionic lock in sea green, hydrophobic barrier in light blue). **d** Snake plot representation of non-olfactory class A GPCR consensus sequences.

Table 1 Key residues and consensus motifs. Superscripts refer to the Ballesteros–Weinstein numbering scheme.

TM	class A GPCR	OR	TAS2R
1	$GN^{1.50}xLV$	$GN^{1.50}LLI$	$N^{1.50}GFI$
2	$LAxAD^{2.50}$	$LSFxD^{2.50}$	$LAxSR^{2.50}$
3	$L^{3.43}$ $DR^{3.50}Y$	$L^{3.43}$ $MAYDR^{3.50}YVAIC$	$L^{3.43}$ $K^{3.50}IANFS$
4	$W^{4.50}$	$W^{4.50}$	$L^{4.50}LG$
5	$P^{5.50}$ $Y^{5.58}$	$P^{5.50}F$ $Y^{5.58}$	$P^{5.50}F$ $F^{5.58}$
6	$K^{6.32}xxK$ $CW^{6.48}LP$	$RxK^{6.32}AFSTC$ $FY^{6.48}G$	$HxK^{6.32}ALKT$ $YF^{6.48}L$
7	$SxxNP^{7.50}xxY$	$PxxNP^{7.50}xIY$	$PxxHS^{7.50}FIL$

These discrepancies remain a central issue in understanding the complex allosteric TAS2R machinery. The present study aims to identify the molecular switches that control TAS2R functions. We present an integrative protocol that advances comparative modeling of TAS2Rs. Case studies of site-directed mutagenesis followed by *in vitro* functional assays on human TAS2R16 then evaluated the roles of the predicted molecular switches in TAS2Rs.

Methods

Sequence alignment

Automatic multiple sequence alignment (MSA) of TAS2Rs was performed with class A and class F templates (labelled *ClustalO* and *classF,* respectively*)* using ClustalO [31] with default settings in the Jalview interface (v2.11.0) [32]. These MSAs were not modified. Another MSA, labelled *Chemosim,* was completed using class A templates, 339 class II ORs and TAS2Rs. The *Chemosim* alignment was then manually refined using constraints from functional assays in the literature (as described in the results section). We specifically focused on the 339 class II ORs because they contain relevant motifs for TAS2Rs alignment and because TM sequence conservation is higher than in a mixture of class I and class II human ORs. TM segments were predicted by the PPM webserver [33]. The final *Chemosim* MSA is provided as a supplementary information file (TAS2R-OR-templates.pir).

Template selection for comparative modeling of bitter taste receptors

Class A GPCR templates were selected by submitting each of the 25 human TAS2Rs UniprotKB accession numbers to the Swiss-Model modeling server [34]. From the proposed templates for human TAS2Rs, 46 with at least 10% sequence identity were kept. Templates were then grouped by protein name and sorted by resolution and average sequence identity with TAS2Rs. The highest resolution template from each group was retained, resulting in 19 templates. Finally, six GPCR class A templates were selected to maximize structural diversity. As TAS2Rs have been suggested to be part of the same family as the frizzled receptors [35], 3 class F GPCR templates were also considered: the human FZD4 receptor [36] and 2 structures of the human SMO receptor [37]. The PDB code for the six class A templates were as follows: rhodopsin (6FUF) [38], β1-adrenergic (4BVN) [39], β2-adrenergic receptor (5JQH) [40], angiotensin II type 1 (4YAY) [41], chemokine receptor CXCR4 (3ODU) [42], serotonin receptor 5-HT2C (6BQG) [43].

Integrative structural modeling of TAS2R

Using the protocols described above (*Chemosim*, *Gomodo*, *ClustalO*, *GPCRdb*, *BitterDB*, and *classF*), we built a large number of 3D models and evaluated and ranked them using a meta-score defined as the average of the pocket and helicity score (Fig. 2). This score provides a unique descriptor that accounts for both GPCR structural requirements and TAS2R experimental constraints.

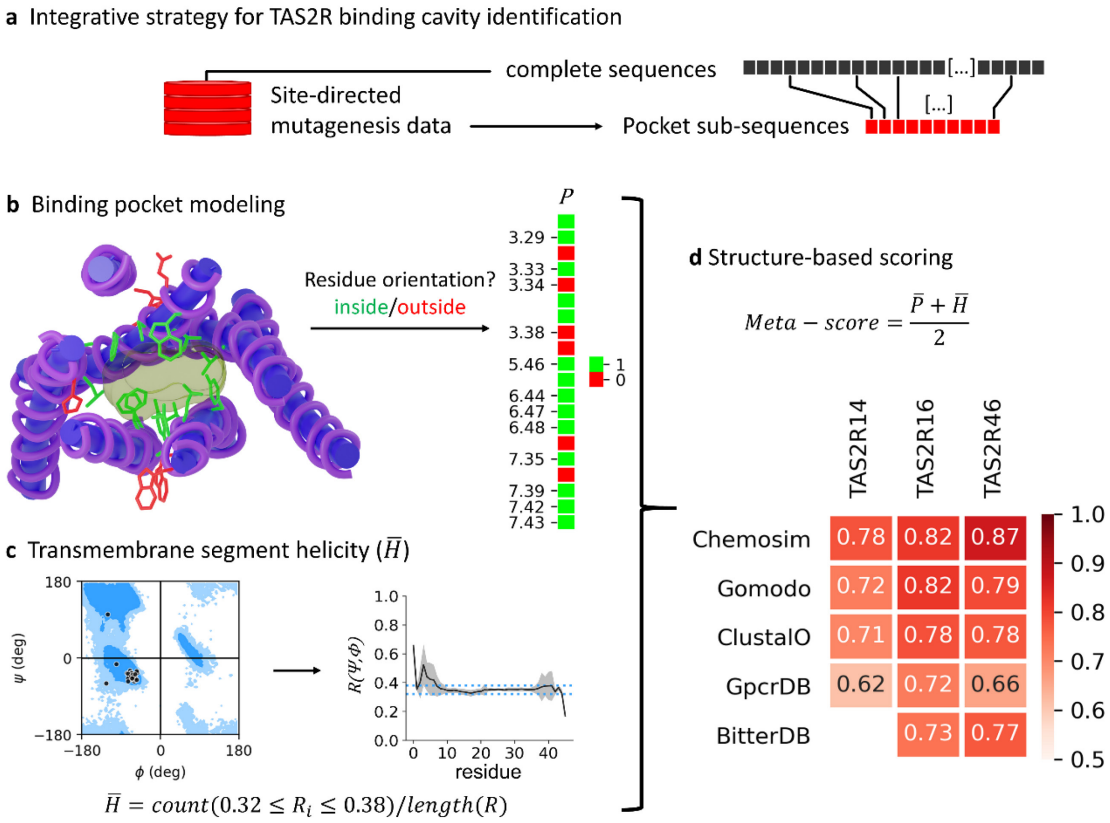


Fig. 2 a An integrative approach to identify the TAS2R binding pocket that is used as a constraint in comparative modeling with the Chemosim protocol. **b** A pocket fingerprint was extracted based on the positions of binding residues in the 3D model. The light brown surface represents the binding pocket. **c** The helicity of the TM segment was analyzed and **d** combined with the pocket fingerprint to calculate a structure-based normalized meta-score. The meta-scores of the best 3D models of TAS2R14, 16 and 46 structures generated by the different comparative modeling protocols are shown in panel d.

For each alignment (*ClustalO*, *Chemosim,* and *classF*) and each template, we generated 1000 homology models using Modeller v9.21 [44] with a maximum of 300 conjugate gradient minimization steps and refinement by molecular dynamics with simulated annealing ("md_level"=slow). The remaining parameters were set to default from the "automodel" class. The BitterDB and GPCRdb webservers provided additional 3D models of each TAS2R.

The GOMoDo [45] webserver was also used to automatically generate models of TAS2Rs based only on the sequence (labelled *Gomodo* in the analysis). Default options were used, excepting the number of models which was set to the maximum (99 models).

Evaluation of the model pocket score: To identify residues oriented toward the binding pocket, the following protocol was implemented in Python: i/ For each of the 25 human TAS2Rs, a reference 3D model was selected from the *Chemosim* models. All reference models were then structurally aligned to the TAS2R16 reference. ii/ A unique grid of points broadly covering the binding site of class A GPCRs was generated and aligned to the coordinates of the TAS2R16 reference. iii/ Each TAS2R model was aligned to its reference based on the alpha carbons of the TM residues. iv/ Residues whose sidechain center of mass (SCM) was within 8.0 angstroms of any grid point, and whose angle between the SCM, the alpha carbon, and any grid point was lower or equal to 30 degrees, were considered as oriented towards the pocket. Only residues annotated as involved in ligand binding were kept (see supplementary file TAS2R-msa_annotated.xlsx). v/ The pocket score was calculated as the fraction of residues oriented towards the pocket for each TM, averaged across all TMs. 3D structure alignment was performed with MDAnalysis v1.0.0 [46], and distance and angle calculations were performed with scipy v1.5.0 [47] and numpy v1.19.0 [48].

Evaluation of TM helicity: The Ramachandran number [49] (R) was used to check the structural quality of the TM domains of each model produced. R, which is based on the ϕ and ψ dihedral angles, can be seen as a short numerical form of the Ramachandran plot. First, we analyzed the helicity of 358 class A GPCR X-ray structures to set the experimental range and found an average value of 0.35. Thus, a residue was considered in an alpha-helix conformation if its R value fell between 0.32 and 0.38.

To discard misshapen 3D models having severe kinks in the middle of TM domains, we introduced a function based on R. We defined the function $f(r) = count(|r_i - R_{ref}| \leq \sigma)$, where r is a moving subset of six consecutive R values that are shifted forward until all R values

for a given TM helix have been sampled; $R_{ref} = 0.35$ is the average R value based on X-ray structures; and σ=0.07 is a parameter that was optimized to exclude misfolded TM proteins while keeping X-ray structures. If at any point the result of $f(r)$ was lower than 4 for any TM residue, the model was discarded. A helicity score ($\overline{H}$) was then calculated as the fraction of TM residues satisfying the condition: $\overline{H} = count(0.32 \leq R_i \leq 0.38)/length(R)$. Among all considered X-ray structures, the minimum $\overline{H}$ value obtained was 0.789. This threshold was used to filter out irrelevant models.

Assessing meta-score accuracy: The meta-score was defined as the average of the pocket and helicity scores. The relevance of the meta-score was assessed by building a homology model of the human smoothened receptor (class F) from a β2-adrenoceptor template (class A, with a low shared sequence identity [9%] with class F, PDB 5JQH). Using the experimental structure of a human smoothened receptor (PDB 4JKV), the RMSD of the best model was then calculated from the meta-score or from the scores available in Modeller or the QMEANBrane [50] webserver. As shown in Fig. S2, the meta-score outperformed classical metrics when ranking GPCR models based on distantly related GPCR templates.

Cell culture and transfection

Plasmids encoding TAS2R16 and G16αgust44 were constructed as previously described [51]. $G16_{\alpha gust44}$ and TAS2R16 were cloned into a CMV promoter-based vector and expressed constitutively. Point mutations on the TAS2R16 clone were obtained from a commercial service (Macrogen Inc., Seoul, Republic of Korea), which also performed DNA sequencings of the mutant genes. The TAS2R16 and $G16_{\alpha gust44}$ expression plasmids were co-transfected (4:1) into HEK293T cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA). Cellular responses were measured 18–24 h after transfection. Cells were cultured at 37°C in a humidified atmosphere of 5% CO_2. The culture medium was Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS), 100 IU/ml penicillin G, 100 µg/ml streptomycin, 2 mM L-glutamine, and 1 mM sodium pyruvate (Invitrogen).

Quantitative measurement of intracellular Ca^{2+} in bitter taste receptors upon stimulation with salicin

The compound-induced changes in cytosolic Ca^{2+} concentrations were measured using a FlexStation III microplate reader (Molecular Devices, Sunnyvale, CA, USA). Cells transfected with TAS2R16 were seeded onto 96-well black-wall CellBind surface plates (Corning, NY, USA). After 18–24 h seeding, the cells were washed with assay buffer (130 mM NaCl, 10 mM glucose, 5 mM KCl, 2 mM $CaCl_2$, 1.2 mM $MgCl_2$, and 100 mM HEPES; pH 7.4) and incubated in the dark, first at 37°C for 30 min, and then at 27°C for 15 min in assay buffer consisting of Calcium-4 (FLIPR Calcium 4 Assay Kit, Molecular Devices). After the samples were treated, the cell fluorescence intensity (excitation, 486 nm; emission, 525 nm) was measured. The results were plotted with $\Delta F/F^0$ on the y-axis, where ΔF is the change in Calcium-4 fluorescence intensity at each time point, and F^0 is the initial fluorescence intensity. The responses from at least three wells ($n = 3$) with the same stimulus were averaged.

Results and discussion

Matching conserved motifs between Class A GPCRs and TAS2Rs

The prediction of TAS2Rs tertiary structure based on sequence similarity remains challenging due to discrepancy in the published alignment [11, 19-30]. We have already shown that refining the sequence alignment of ORs with non-olfactory class A GPCRs by including site-directed mutagenesis produces relevant three-dimensional models of chemosensory receptors. These models have been supported by a large amount of experimental data [16, 18, 52, 53]. We thus apply a similar integrative strategy to TAS2Rs. To overcome the lack of sequence similarity between TAS2Rs and GPCRs with known structures, we inserted 339 human class II OR sequences in the alignment. Subsequent manual data curation involved integration of site-directed mutagenesis data from the literature for 136 amino-acids positions, i.e. 45% of the entire TAS2Rs sequence (see ESI TAS2R-msa_annotated.xlsx). Our alignment (Fig. S1) highlights the key residues and consensus motifs in all human TAS2Rs, which correspond to the functional molecular switches in ORs and non-olfactory class A GPCRs (Fig. 1b,d). They are detailed above and summarized in Table 1.

TM1, 2 and 4 did not contain motifs involved in downstream signaling. In TM1, the $NGFI^{TM1\text{-}TAS2R}$ motif corresponds to $GNLLI^{TM1\text{-}OR}$ in OR and $GNxLV^{TM1\text{-}classA}$ in non-olfactory GPCR templates (see Fig. S1). In TM2, $R^{2.50\text{-}TAS2R}$ in the $LAxSR^{TM2\text{-}TAS2R}$ motif aligns with

$D^{2.50\text{-OR/classA}}$, which in class A GPCRs constitutes a sodium ion binding site that stabilizes inactive receptor conformations [54]. Position 2.50 in TAS2Rs is positively-charged and unlikely to be involved in sodium binding. Rather, it is hypothesized to stabilize the structure of TAS2Rs. [21] The sequence alignment of TM4 was not straightforward, as it lacks the canonical $W^{4.50\text{-OR/classA}}$. The highly conserved leucine $L^{4.50}$ of the $LLG^{TM4\text{-}TAS2R}$ motif aligns with the most conserved $W^{4.50\text{-OR/class A}}$.

TM3, 5, 6, and 7 contained functional molecular switches which have been identified in class A GPCR experimental structures [14].

In TM3, $K^{3.50}$ in the $KIANFS^{TM3\text{-}TAS2R}$ motif matches $R^{3.50}$ of the $DRY^{TM3\text{-}classA}$ and $MAYDRYVAIC^{TM3\text{-}OR}$ motifs. The DRY motif constitutes the ionic lock in ORs and non-olfactory class A GPCRs. This also aligns the highly conserved $L^{3.43}$, with a leucine found at position 3.43 in both non olfactory class A GPCRs and OR (Table 1).

In TM5, the conserved $P^{5.50}$ of the $PF^{TM5\text{-}TAS2R}$ motif corresponds to the $PF^{TM5\text{-}OR}$ and $P^{TM5\text{-}classA}$ motifs/residue involved in the so-called "hydrophobic connector" ($P^{5.50}I^{3.40}F^{6.44}$ in class A GPCRs). Another conserved aromatic residue that is found in 52% of TAS2Rs, $F^{5.58}$, consistently aligns with the conserved $Y^{5.58}$ known to be important for GPCR activation [18, 55].

In TM6, the $HxKALKT^{TM6\text{-}TAS2R}$ motif matches both a comparable motif in non-olfactory class A GPCRs and the typical OR motif $RxKAFST^{TM6\text{-}OR}$. The "toggle/transmission switch" ($CW^{6.48}LP^{classA}$ and $FY^{6.48}G^{OR}$) aligns with the $YF^{6.48}L$ motif in TAS2Rs. The location of this $YF^{6.48}L$ motif at the bottom of the pocket is consistent with site-directed mutagenesis results, suggesting a ligand-sensing role, as is the case for class A GPCRs [16, 56].

The extracellular part of TM7 is well-documented to belong to the ligand binding pocket in TAS2Rs and other GPCRs [20, 24, 56]. This is consistent with its high sequence variability (see Fig. S1). TM7 intracellular residues show higher conservation, as they are involved in GPCR signaling [16, 56]. These conserved motifs, however, show little similarity between TAS2Rs and other GPCRs. Here, the comparison with ORs is highly instructive: from the $P^{7.46}xLNP^{7.50}xIY^{TM7\text{-}OR}$ motif found in ORs, $P^{7.46}$ is shared with TAS2Rs, and $NP^{7.50}xxY$ is found in other class A GPCRs. $P^{7.46}$ and $P^{7.50}$ are conserved in 76% and 28% of human TAS2Rs, respectively. The $PxxHSFIL^{TM7\text{-}TAS2R}$ motif is consequently aligned with $PxLNPxIY^{TM7\text{-}OR}$, which itself matches the highly conserved $xxxNPxxY^{TM7\text{-}classA}$ motif [20].

Predicted tertiary structure of TAS2Rs

Based on this refined alignment, we tested various protocols and structural templates to build accurate 3D homology models of TAS2Rs. Among the TAS2Rs, receptors TAS2R14, 16, and 46 were selected to evaluate the approach, as previous work on these receptors involving site-directed mutagenesis provides data to determine the residues within their binding pocket. According to our meta-score, the best models of these three receptors were obtained using the *Chemosim* approach and a single template, either the β2-adrenoceptor (PDB 5JQH) or the β1-adrenoceptor (PDB 4BVN) structure (Fig. 2 & S3). The performance of each protocol is compared in Fig. S3 and S4. *Gomodo* and *ClustalO* approaches led to comparable models, with slight improvement over *BitterDB* and, in most cases, substantial improvement over *GPCRdb*. The use of class F templates systematically led to models with misfolded helices (Fig. S4).

These models and analysis were then extrapolated to the full human TAS2Rs repertoire. Even if limited experimental data is available, we were able to define a consensus TAS2R cavity based on the positions identified simultaneously in TAS2R14, 16 and 46. We also extended the definition of a specific TAS2R cavity to residues identified by site-directed mutagenesis. The best models for the entire TAS2R family were obtained using GPCR templates in their closed conformation (Fig. S6), with the exception of TAS2R38, for which the open-conformation 5-HT_{2C} receptor (PDB 6BQG) was best. On average, the templates 5JQH, 4BVN all of which correspond to adrenergic receptors, performed best. In this study, we found no relationship between the performance of the protocols and the percentage sequence identity of the templates used to build the models. At 10–15%, the sequence identity between TAS2Rs and class A templates is too low to be a discriminating criterion.

The best *Chemosim* model obtained for each human TAS2R is provided as a PDB file in the supplementary information. Projecting TAS2Rs sequence conservation onto the 3D structure showed that the models retain the structural characteristics of the GPCR (Fig. S5). The most conserved residues were located in the intracellular region of the receptor that binds the G protein, while the greatest variability was found in the extracellular ligand-binding pocket. Analysis of the binding cavity (Fig. S7) revealed high diversity within the hTAS2Rs family. The pocket volume ranged up to 400 $Å^3$ and 700 $Å^3$ for hTAS2R13 and hTAS2R39, respectively, corresponding to the structural features of a GPCR [57]. Although no obvious structure-function relationship was revealed by the analysis of the cavity volume, the hydrophobicity partially correlated with the receptor range of response. The binding cavities of TAS2Rs with broad ligand spectrums tended to be more hydrophobic than those of narrow-

spectrum receptors (Fig. S7), consistent with previous studies showing a correlation between hydrophobicity and GPCR promiscuity [58, 59].

Evaluating the function role of molecular switches

To evaluate the functional role of the predicted molecular switches, twelve residue positions on TAS2R16 were subjected to site-directed mutagenesis followed by *in vitro* functional assays with salicin (Fig. 3 and Table S2). The residues mostly belonged to TM3 and TM6, which, in GPCRs, are well-known to be involved in agonist sensing and activation [14].

Using our model as a basis, we investigated residues found in the ligand binding pocket ($90^{3.35}$, $91^{3.36}$, and $185^{5.47}$) and at or around the predicted molecular switches ($45^{2.39}$, $97^{3.41}$, $221^{6.29}$, $222^{6.30}$, $236^{6.44}$, and $239^{6.47}$). Residues 42^{ICL1}, 43^{ICL1}, and $100^{3.44}$ were predicted to be far from the molecular switches. All mutants showed a specific, dose-dependent response to salicin (Fig. 3), confirming that they are expressed and functional at the cell surface.

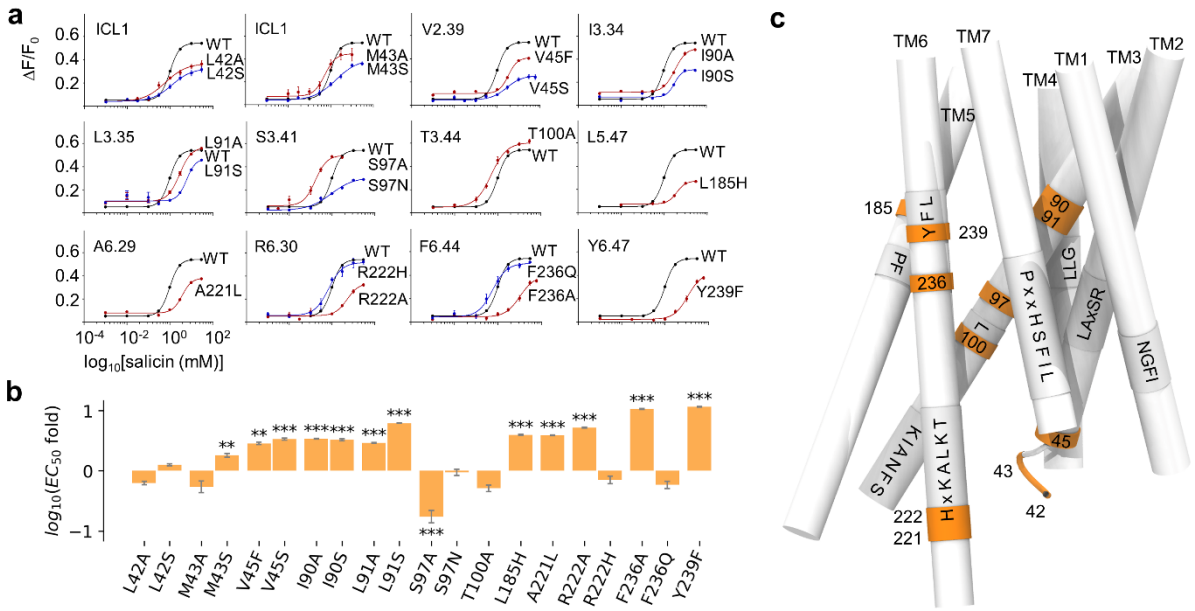


Fig. 3 a *In vitro* functional assays of wild-type (WT) TAS2R16 and single-point mutants stimulated by salicin. **b** EC_{50} fold (compared to WT) expressed as $\log(EC_{50}(MUT)/EC_{50}(WT))$ for the twenty TAS2R16 mutants considered in this study. Positive values indicate a reduced response to salicin in the mutated receptor compared to the WT. *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$ versus the WT group (one-way ANOVA followed by Dunnett's test). **c** Representative structure of TAS2R16 highlighting the location of the mutated residues. The TM domains are presented as sticks. The positions of mutated residues are colored in orange, and the molecular switches revealed by the sequence alignment are indicated on the structure.

The L42^{ICL1}A/S, M43^{ICL1}A, and T100$^{3.44}$A mutations served as negative controls (Table S2) and generally did not statistically affect salicin potency (Fig. 3 and Table S3). Only mutation of position 43 to a serine induced a weak decrease of salicin-dependent response in TAS2R16 compared to WT.

The TASR216 I90A/S$^{3.35}$, L91A/S$^{3.36}$, and L185H$^{5.47}$ mutants showed a reduced response to salicin, consistent with their orientation toward the interior of the receptor bundle (Fig. 3 and Table S3). Positions 3.35 and 5.47 have been previously reported to directly interact with ligands [26, 30, 60].

Position 239$^{6.47}$ is conserved as Y (64%) and F (8%) in human TAS2Rs (Fig. 4a). In mammals, an aromatic residue (F, Y or H) is also found in 85% of the sequences. Conservation of an aromatic residue also occurs in ORs [16]. The Y239F$^{6.47}$ mutation decreased the potency of salicin by a factor of 11, confirming its importance in receptor activation (Fig. 3). Position Y239$^{6.47}$ corresponded to Y239 and Y241 in TAS2R10 and TAS2R46, respectively. For both of these receptors, the tyrosine to phenylalanine mutation is reported to lead to a significant reduction in ligand responsiveness [20, 61]. Born et al. also observed a complete loss of response to agonists with the Y239A$^{6.47}$ TAS2R10 construction [61]. Further, we found that the introduction of an alanine at this position eliminated any response to salicin (data not shown). Altogether, these observations highlight the functional equivalence of the Y$^{6.47}$FLx motif in TAS2Rs with the F$^{6.47}$YGx in ORs [16] and the C$^{6.47}$WLP [14] in non-olfactory class A GPCRs [9]. This motif is particularly important as it forms part of the cradle of the binding pocket and senses the presence of agonists [56]. Adjacent to Y239$^{6.47}$, the aromatic residue F240$^{6.48}$ is conserved as aromatic in 72% of human TAS2Rs and in 67% of mammalian TAS2Rs. As the toggle-switch residue, its nature and function in agonist sensing is similar in ORs (conserved as F$^{6.48}$) [16] and non-olfactory GPCRs (conserved as W$^{6.48}$) [14]. F240$^{6.48}$ has previously been reported to affect TAS2R16 agonist response. Sakurai et al. showed that mutation of F240$^{6.48}$ to a leucine residue in TAS2R16 drastically alters the function of the receptor, while mutation to aromatic residues (Y and W) leads to moderate changes in the EC50 [19]. Further, the potencies of various other agonists were affected in the same manner, highlighting the critical role this residue plays in signal initiation, as is the case for numerous class A GPCRs [14-16].

The hydrophobic connector molecular switch involved in class A GPCRs activation [15] was conserved as P$^{5.50}$I$^{3.40}$F$^{6.44}$ [14, 15, 17]. Similarly to other TAS2Rs, a P$^{5.50}$A$^{3.40}$F$^{6.44}$ motif (Fig. 4b) was located at the core of TAS2R16, close to the cradle of the binding pocket. In class A

GPCRs, this motif, together with NPxxY^{TM7}, holds a central role in receptor signaling, ligand-independent constitutive activation, and β-arrestin signaling in the β2-adrenoceptor [17]. It is plausible that this motif has similar functions in TAS2Rs [62], as suggested by the modulated response to salicin we found in our mutants (Fig. 3). F236$^{6.44}$, conserved in 75% of mammalian TAS2Rs as Y/F (Fig. 4b), is predicted to be part of the hydrophobic connector molecular switch. The F236A$^{6.44}$ TAS2R16 mutant consistently showed a significantly weaker response to salicin, while no difference in response was found for the F236Q$^{6.44}$ mutant. In a previous study, Thomas et al. found that a F236Y$^{6.44}$ mutation prevented agonist-dependent signaling [26]. In TAS2R14, an alanine residue occupies position 6.44, and mutation to a leucine leads to a decrease in receptor sensitivity to numerous ligands [60].

Adjacent to position 3.40, S97$^{3.41}$ does not belong to the binding pocket and points toward the membrane. In accordance with a previous report showing its importance for TAS2R16 trafficking [26], the S97A$^{3.41}$ mutation altered receptor response (gain of function).

Our model predicted that V45$^{2.39}$ is part of a hydrophobic cluster in the intracellular part of TM2 and is conserved as a hydrophobic residue in 72% of TAS2Rs. This hydrophobic area occurs near the highly conserved L229$^{7.53}$ (96% and 93% in humans and mammals, respectively) and the HSFIL^{TM7} motifs and likely forms part of the hydrophobic barrier that prevents flooding of the intracellular region. Mutating V45$^{2.39}$ into a hydrophilic residue (S) strongly altered salicin activation both in this work and in the literature [26]; substitution with a bulkier hydrophobic residue (F) was better tolerated.

In TM6, position 6.29 and adjacent residues have been documented to control G protein selectivity in class A GPCRs [63]. A221$^{6.29}$ and H222$^{6.30}$ are conserved in 60% and 92% of human TAS2Rs, respectively, and in 70% and 94% of mammalian TAS2Rs (Fig. 4c). Position 222$^{6.30}$ is an arginine in TAS2R16. Salicin induced reduced responses in the A221L$^{6.29}$ and R222A$^{6.30}$ mutants, whereas the response of the R222H$^{6.30}$ mutant was not statistically different from the WT. In TASR2R4, the H233A$^{6.30}$ mutation inhibited the response to quinine [64]. Altogether, these findings highlight the need for a positive charge at position 6.30 for G protein-coupling and selectivity.

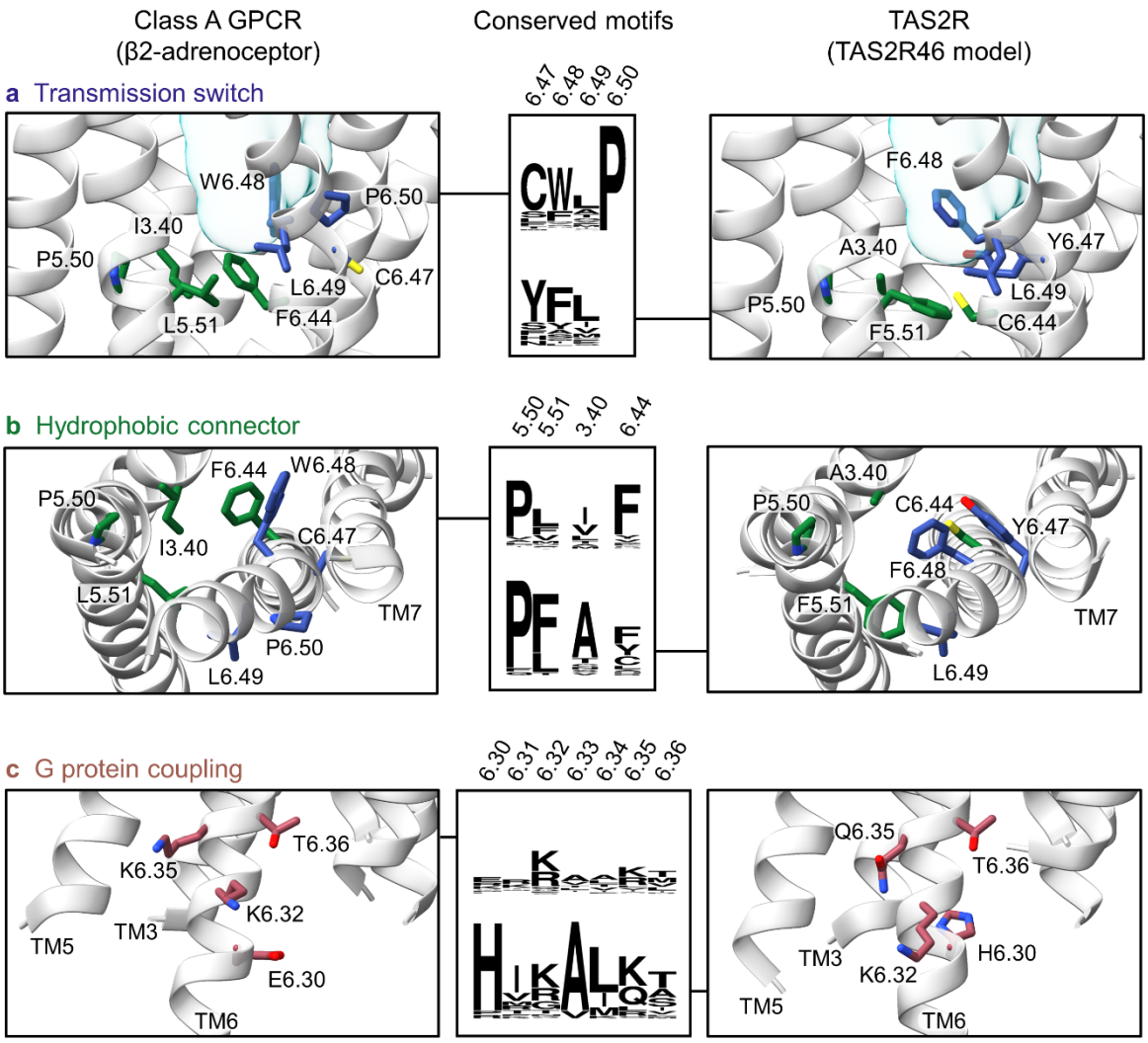


Fig. 4 Sequence logos and molecular details of conserved motifs involved in the activation mechanism of class A GPCRs and TAS2Rs, i.e. **a** the transmission switch (colored in blue), **b** the hydrophobic connector (in green), and **c** the G protein-coupling region (in red). The binding pocket is depicted as a pale blue surface. The structure of the β2-adrenoceptor is taken from PDB code 5JQH.

Conclusions

This study elucidates key residues and consensus functional motifs of bitter taste receptors (TAS2Rs) using a combination of bioinformatics, molecular modeling, and *in vitro* assays. The consensus sequence motifs match well-known ones in class A GPCRs. Further, we performed sequence alignment of human TAS2Rs with olfactory and non-olfactory class A GPCRs, including residue conservation and experimental data as constraints. Using site-directed mutagenesis, we then evaluated the functional roles of these motifs in TAS2R16 as a case study.

In addition to the residues lining the binding pocket, we identified the "toggle/transmission switch" (the $YF^{6.48}L$ motif in TM6) and the "hydrophobic connector" ($P^{5.50}A^{3.40}F^{6.44}$) for agonist sensing. Other molecular switches were identified in the intracellular regions of TM6 and TM7 that are suggested to be involved in G protein selectivity or in receptor activation. These molecular switches extends to mammalian TAS2Rs (see supplementary files). The approach, templates, and 3D model provided in this study serve as a foundation for rational design of specific TAS2Rs agonists and antagonists and for decoding sequence-structure-function relationships in these receptors.

Code and data availability

The scripts used to generate and analyze the models as well as PDB files of TAS2Rs 3D models with the highest meta-score have been deposited on GitHub. (https://github.com/chemosim-lab/TAS2R_data)

Author contributions

JT†, CB†, and JP performed numerical modeling. YK and MR conducted functional assays. JT, CB, and SF performed data curation. JT, CB, JP, YK, and MR conducted formal analyses. JT, SF, and JG supervised and managed the study and wrote the paper. MR and JG provided resources for this study.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

The authors thank Dr. Xiaojing Cong for fruitful discussion and critical review of the manuscript. This work was funded by the French Ministry of Higher Education and Research [PhD Fellowship to CB], by the National Research Foundation of Korea (NRF) [grant number NRF2020R1A2C2004661], by GIRACT (Geneva, Switzerland) [9th European PhD in Flavor Research Bursaries for first year students to CB], and the Gen Foundation (Registered UK Charity No. 1071026), a charitable trust that primarily funds research in natural sciences, particularly food sciences/technology [grant to CB and JT]. We also benefited from funding by the French government through the UCAJEDI "Investments in the Future" project, managed

by the ANR [grant No. ANR-15-IDEX-01 to SF and JG]. Computation for the work described in this paper was supported by the Université Côte d'Azur's Center for High-Performance Computing.

References

[1] B. Lindemann, Receptors and transduction in taste, Nature, 413 (2001) 219-225.

[2] W. Meyerhof, Elucidation of mammalian bitter taste, in: Reviews of physiology, biochemistry and pharmacology, Springer, 2005, pp. 37-72.

[3] K.L. Mueller, M.A. Hoon, I. Erlenbach, J. Chandrashekar, C.S. Zuker, N.J. Ryba, The receptors and coding logic for bitter taste, Nature, 434 (2005) 225-229.

[4] S.-J. Lee, I. Depoortere, H. Hatt, Therapeutic potential of ectopic olfactory and taste receptors, Nature Reviews Drug Discovery, 18 (2019) 116-138.

[5] S. Foster, K. Blank, L. See Hoe, M. Behrens, W. Meyerhof, J. Peart, W. Thomas, Novel bitter taste receptor ligands elicit G protein-dependent negative inotropic effects in mouse heart (LB572), The FASEB Journal, 28 (2014) LB572.

[6] 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 (2015) 533-545.

[7] E. Adler, M.A. Hoon, K.L. Mueller, J. Chandrashekar, N.J. Ryba, C.S. Zuker, A novel family of mammalian taste receptors, Cell, 100 (2000) 693-702.

[8] R. Fredriksson, M.C. Lagerström, L.-G. Lundin, H.B. Schiöth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints, Molecular pharmacology, 63 (2003) 1256-1272.

[9] K.J. Nordström, M. Sällman Almén, M.M. Edstam, R. Fredriksson, H.B. Schiöth, Independent HHsearch, Needleman–Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families, Molecular biology and evolution, 28 (2011) 2471-2480.

[10] A. Krishnan, M.S. Almén, R. Fredriksson, H.B. Schiöth, The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi, PloS one, 7 (2012) e29817.

[11] A. Di Pizio, A. Levit, M. Slutzki, M. Behrens, R. Karaman, M.Y. Niv, Comparing Class A GPCRs to bitter taste receptors: Structural motifs, ligand interactions and agonist-to-antagonist ratios, in: Methods in cell biology, vol. 132, Elsevier, 2016, pp. 401-427.

[12] V. Cvicek, W.A. Goddard III, R. Abrol, Structure-based sequence alignment of the transmembrane domains of all human GPCRs: phylogenetic, structural and functional implications, PLoS computational biology, 12 (2016) e1004805.

[13] C. Munk, V. Isberg, S. Mordalski, K. Harpsøe, K. Rataj, A. Hauser, P. Kolb, A. Bojarski, G. Vriend, D. Gloriam, GPCRdb: the G protein-coupled receptor database–an introduction, British Journal of Pharmacology, 173 (2016) 2195-2207.

[14] X. Deupi, J. Standfuss, Structural insights into agonist-induced activation of G-protein-coupled receptors, Current opinion in structural biology, 21 (2011) 541-551.

[15] Q. Zhou, D. Yang, M. Wu, Y. Guo, W. Guo, L. Zhong, X. Cai, A. Dai, W. Jang, E.I. Shakhnovich, Common activation mechanism of class A GPCRs, eLife, 8 (2019).

[16] 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, Journal of the American Chemical Society, 137 (2015) 8611-8616.

[17] A.-M. Schönegge, J. Gallion, L.-P. Picard, A.D. Wilkins, C. Le Gouill, M. Audet, W. Stallaert, M.J. Lohse, M. Kimmel, O. Lichtarge, Evolutionary action and structural basis of the allosteric switch controlling β 2 AR functional selectivity, Nature communications, 8 (2017) 1-12.

[18] C.A. de March, J. Topin, E. Bruguera, G. Novikov, K. Ikegami, H. Matsunami, J. Golebiowski, Odorant receptor 7D4 activation dynamics, Angewandte Chemie, 130 (2018) 4644-4648.

[19] T. Sakurai, T. Misaka, M. Ishiguro, K. Masuda, T. Sugawara, K. Ito, T. Kobayashi, S. Matsuo, Y. Ishimaru, T. Asakura, Characterization of the β-d-glucopyranoside binding site of the human bitter taste receptor hTAS2R16, Journal of Biological Chemistry, 285 (2010) 28373-28378.

[20] A. Brockhoff, M. Behrens, M.Y. Niv, W. Meyerhof, Structural requirements of bitter taste receptor activation, Proceedings of the National Academy of Sciences, 107 (2010) 11110-11115.

[21] N. Singh, S.P. Pydi, J. Upadhyaya, P. Chelikani, Structural basis of activation of bitter taste receptor T2R1 and comparison with Class A G-protein-coupled receptors (GPCRs), Journal of Biological Chemistry, 286 (2011) 36032-36041.

[22] R. Karaman, S. Nowak, A. Di Pizio, H. Kitaneh, A. Abu-Jaish, W. Meyerhof, M.Y. Niv, M. Behrens, Probing the binding pocket of the broadly tuned human bitter taste receptor TAS2R14 by chemical modification of cognate agonists, Chemical biology & drug design, 88 (2016) 66-75.

[23] F. Fierro, A. Giorgetti, P. Carloni, W. Meyerhof, M. Alfonso-Prieto, Dual binding mode of "bitter sugars" to their human bitter taste receptor target, Scientific reports, 9 (2019) 1-16.

[24] 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, Journal of chemical theory and computation, 11 (2015) 4439-4449.

[25] A. Di Pizio, L.-M. Kruetzfeldt, S. Cheled-Shoval, W. Meyerhof, M. Behrens, M.Y. Niv, Ligand binding modes from low resolution GPCR models and mutagenesis: chicken bitter taste receptor as a test-case, Scientific reports, 7 (2017) 1-11.

[26] A. Thomas, C. Sulli, E. Davidson, E. Berdougo, M. Phillips, B.A. Puffer, C. Paes, B.J. Doranz, J.B. Rucker, The Bitter Taste Receptor TAS2R16 Achieves High Specificity and Accommodates Diverse Glycoside Ligands by using a Two-faced Binding Pocket, Scientific reports, 7 (2017) 7753.

[27] X. Biarnés, A. Marchiori, A. Giorgetti, C. Lanzara, P. Gasparini, P. Carloni, S. Born, A. Brockhoff, M. Behrens, W. Meyerhof, Insights into the binding of Phenyltiocarbamide (PTC) agonist to its target human TAS2R38 bitter receptor, PloS one, 5 (2010).

[28] S. Prasad Pydi, J. Upadhyaya, N. Singh, R. Pal Bhullar, P. Chelikani, Recent advances in structure and function studies on human bitter taste receptors, Current Protein and Peptide Science, 13 (2012) 501-508.

[29] J.P. Slack, A. Brockhoff, C. Batram, S. Menzel, C. Sonnabend, S. Born, M.M. Galindo, S. Kohl, S. Thalmann, L. Ostopovici-Halip, Modulation of bitter taste perception by a small molecule hTAS2R antagonist, Current Biology, 20 (2010) 1104-1109.

[30] Y. Wang, A.L. Zajac, W. Lei, C.M. Christensen, R.F. Margolskee, C. Bouysset, J. Golebiowski, H. Zhao, S. Fiorucci, P. Jiang, Metal ions activate the human taste receptor TAS2R7, Chemical senses, 44 (2019) 339-347.

[31] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular systems biology, 7 (2011).

[32] 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 (2009) 1189-1191.

[33] 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 research, 40 (2012) D370-D376.

[34] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F.T. Heer, T.A.P. de Beer, C. Rempfer, L. Bordoli, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic acids research, 46 (2018) W296-W303.

[35] R. Fredriksson, H.B. Schiöth, The repertoire of G-protein–coupled receptors in fully sequenced genomes, Molecular pharmacology, 67 (2005) 1414-1425.

[36] S. Yang, Y. Wu, T.-H. Xu, P.W. de Waal, Y. He, M. Pu, Y. Chen, Z.J. DeBruine, B. Zhang, S.A. Zaidi, Crystal structure of the Frizzled 4 receptor in a ligand-free state, Nature, 560 (2018) 666-670.

[37] X. Zhang, F. Zhao, Y. Wu, J. Yang, G.W. Han, S. Zhao, A. Ishchenko, L. Ye, X. Lin, K. Ding, Crystal structure of a multi-domain human smoothened receptor in complex with a super stabilizing ligand, Nature communications, 8 (2017) 1-10.

[38] C.-J. Tsai, F. Pamula, R. Nehmé, J. Mühle, T. Weinert, T. Flock, P. Nogly, P.C. Edwards, B. Carpenter, T. Gruhl, Crystal structure of rhodopsin in complex with a mini-Go sheds light on the principles of G protein selectivity, Science advances, 4 (2018) eaat7052.

[39] J.L. Miller-Gallacher, R. Nehme, T. Warne, P.C. Edwards, G.F. Schertler, A.G. Leslie, C.G. Tate, The 2.1 Å resolution structure of cyanopindolol-bound β1-adrenoceptor identifies an intramembrane Na+ ion that stabilises the ligand-free receptor, PloS one, 9 (2014).

[40] D.P. Staus, R.T. Strachan, A. Manglik, B. Pani, A.W. Kahsai, T.H. Kim, L.M. Wingler, S. Ahn, A. Chatterjee, A. Masoudi, Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation, Nature, 535 (2016) 448-452.

[41] H. Zhang, H. Unal, C. Gati, G.W. Han, W. Liu, N.A. Zatsepin, D. James, D. Wang, G. Nelson, U. Weierstall, Structure of the angiotensin receptor revealed by serial femtosecond crystallography, Cell, 161 (2015) 833-844.

[42] B. Wu, E.Y. Chien, C.D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F.C. Bi, Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists, Science, 330 (2010) 1066-1071.

[43] Y. Peng, J.D. McCorvy, K. Harpsøe, K. Lansu, S. Yuan, P. Popov, L. Qu, M. Pu, T. Che, L.F. Nikolajsen, 5-HT2C receptor structures reveal the structural basis of GPCR polypharmacology, Cell, 172 (2018) 719-730. e714.

[44] B. Webb, A. Sali, Comparative Protein Structure Modeling Using MODELLER, Current Protocols in Bioinformatics, 54 (2016) 5.6.1-5.6.37.

[45] M. Sandal, T.P. Duy, M. Cona, H. Zung, P. Carloni, F. Musiani, A. Giorgetti, GOMoDo: a GPCRs online modeling and docking webserver, PloS one, 8 (2013) e74092.

[46] R.J. Gowers, M. Linke, J. Barnoud, T.J.E. Reddy, M.N. Melo, S.L. Seyler, J. Domanski, D.L. Dotson, S. Buchoux, I.M. Kenney, MDAnalysis: a Python package for the rapid analysis

of molecular dynamics simulations, in, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

[47] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature methods, 17 (2020) 261-272.

[48] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, Array programming with NumPy, Nature, 585 (2020) 357-362.

[49] R.V. Mannige, J. Kundu, S. Whitelam, The Ramachandran number: an order parameter for protein geometry, PloS one, 11 (2016) e0160023.

[50] G. Studer, M. Biasini, T. Schwede, Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane), Bioinformatics, 30 (2014) i505-i511.

[51] T. Ueda, S. Ugawa, H. Yamamura, Y. Imaizumi, S. Shimada, Functional interaction between T2R taste receptors and G-protein α subunits expressed in taste receptor cells, Journal of Neuroscience, 23 (2003) 7376-7380.

[52] L. Charlier, J. Topin, C. Ronin, S.-K. Kim, W.A. Goddard, R. Efremov, J. Golebiowski, How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case, Cellular and molecular life sciences, 69 (2012) 4205-4213.

[53] C. Bushdid, A. Claire, J. Topin, M. Do, H. Matsunami, J. Golebiowski, Mammalian class I odorant receptors exhibit a conserved vestibular-binding pocket, Cellular and molecular life sciences, 76 (2019) 995-1004.

[54] V. Katritch, G. Fenalti, E.E. Abola, B.L. Roth, V. Cherezov, R.C. Stevens, Allosteric sodium in class A GPCR signaling, Trends in biochemical sciences, 39 (2014) 233-244.

[55] Y. Miao, S.E. Nichols, P.M. Gasper, V.T. Metzger, J.A. McCammon, Activation and dynamic network of the M2 muscarinic receptor, Proceedings of the National Academy of Sciences, 110 (2013) 10982-10987.

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

[57] J.A. Dalton, I. Lans, J. Giraldo, Quantifying conformational changes in GPCRs: glimpse of a common functional mechanism, BMC bioinformatics, 16 (2015) 1-15.

[58] A. Levit, T. Beuming, G. Krilov, W. Sherman, M.Y. Niv, Predicting GPCR promiscuity using binding site features, Journal of chemical information and modeling, 54 (2014) 184-194.

[59] A. Di Pizio, M.Y. Niv, Promiscuity and selectivity of bitter molecules and their receptors, Bioorganic & medicinal chemistry, 23 (2015) 4082-4091.

[60] S. Nowak, A. Di Pizio, A. Levit, M.Y. Niv, W. Meyerhof, M. Behrens, Reengineering the ligand sensitivity of the broadly tuned human bitter taste receptor TAS2R14, Biochimica et Biophysica Acta (BBA)-General Subjects, 1862 (2018) 2162-2173.

[61] S. Born, A. Levit, M.Y. Niv, W. Meyerhof, M. Behrens, The human bitter taste receptor TAS2R10 is tailored to accommodate numerous diverse ligands, Journal of Neuroscience, 33 (2013) 201-213.

[62] D. Kim, S. Cho, M.A. Castaño, R.A. Panettieri, J.A. Woo, S.B. Liggett, Biased TAS2R bronchodilators inhibit airway smooth muscle growth by downregulating phosphorylated extracellular signal–regulated kinase 1/2, American journal of respiratory cell and molecular biology, 60 (2019) 532-540.

[63] T. Flock, A.S. Hauser, N. Lund, D.E. Gloriam, S. Balaji, M.M. Babu, Selectivity determinants of GPCR–G-protein binding, Nature, 545 (2017) 317-322.

[64] S.P. Pydi, N. Singh, J. Upadhyaya, R.P. Bhullar, P. Chelikani, The third intracellular loop plays a critical role in bitter taste receptor activation, Biochimica et Biophysica Acta (BBA)-Biomembranes, 1838 (2014) 231-236.

Supplementary information

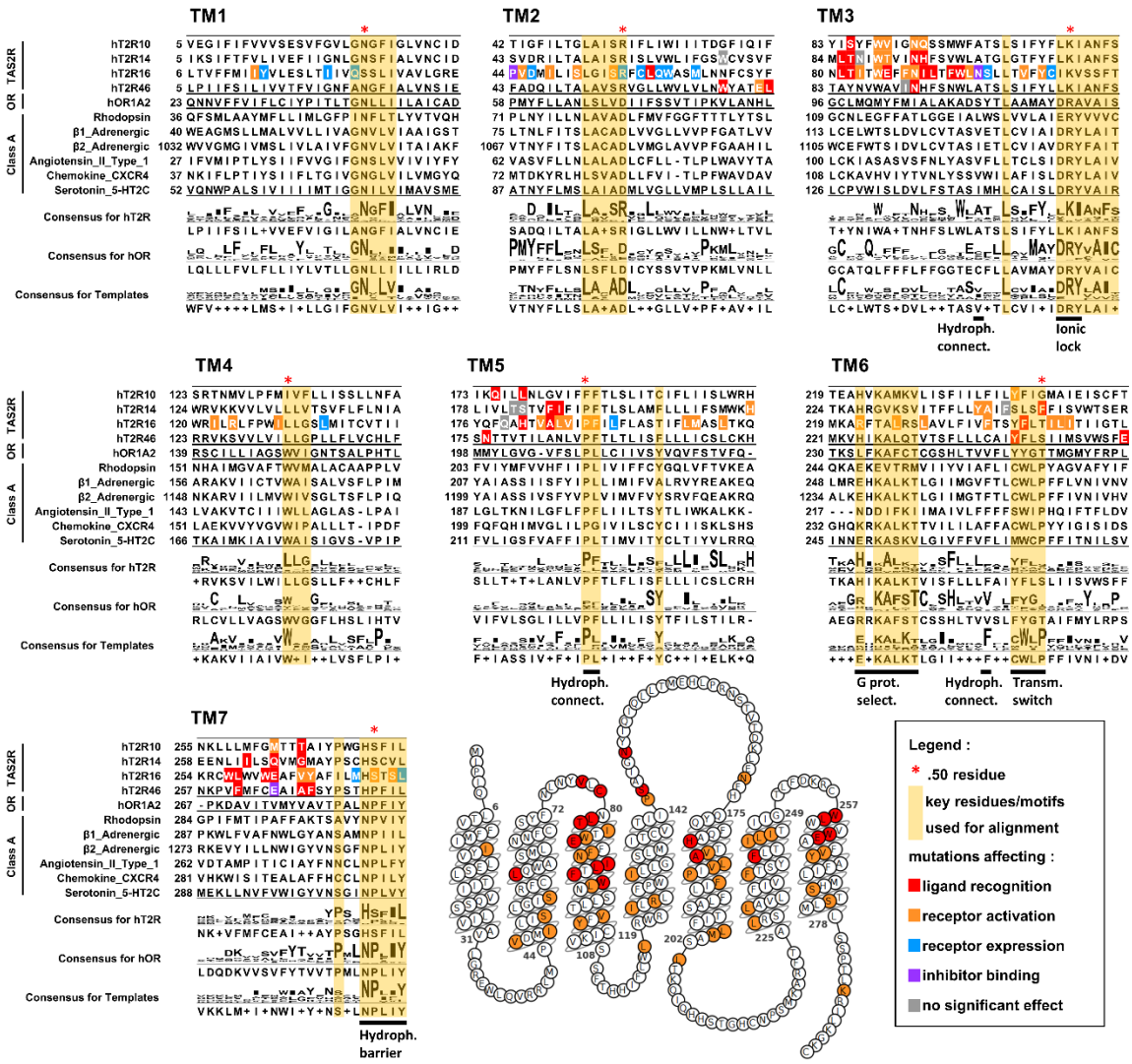


Figure S1. Alignment and results from site directed mutagenesis.

Selected human type 2 taste receptors (TAS2R), human Olfactory Receptors (OR) and non-OR class A G protein-coupled receptors (GPCRs). In the sequence alignment, residues are colored by their roles as reported in the literature (see legend). The aligned motifs are highlighted in yellow. Consensus sequences for TAS2Rs, ORs and Templates contained 25 human TAS2Rs, 339 human class II ORs, and 6 class A GPCRs, respectively. Functional molecular switches (transmission switch, hydrophobic connector, ionic lock, and hydrophobic barrier) and residues involved in G-protein coupling are indicated under the alignment.

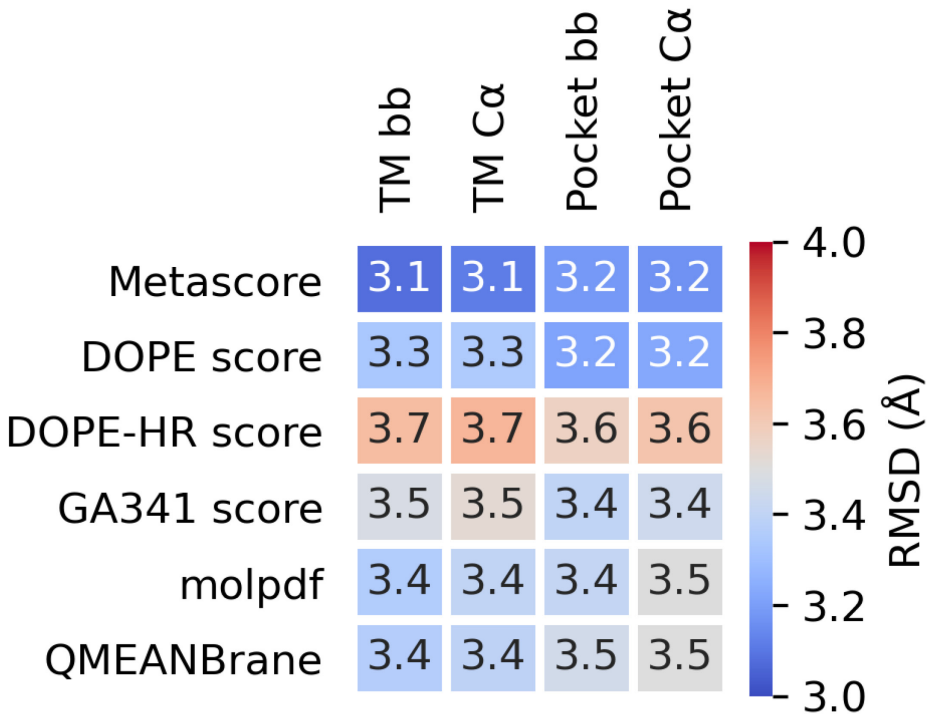


Figure S2. RMSD of class F models built using a class A template

The human smoothened receptor (class F) models were built by homology modeling with a class A template (β2-adrenoceptor, PDB 5JQH [1]). The sequence alignment was taken from the GPCRdb [2] and manually refined with UCSF Chimera's structure-based sequence alignment tool (v1.14) [3] based on the 5JQH template and a structure of the smoothened receptor (PDB 4JKV [4]). The same Modeller [5] protocol detailed in the manuscript was used to generate 1000 models of the smoothened receptor. The models were structurally aligned to the 4JKV reference based on the trans membrane (TM) domains and ranked by their meta-scores. Finally, the RMSDs between the reference and each best model were calculated based on the TM domain backbone (TM bb), the TM alpha carbons (TM Ca), the pocket residue backbone (Pocket bb), and the pocket residue alpha carbons (Pocket Ca). The pocket residues were identified by visual inspection of four class F X-ray structures in complex with a ligand (PDB codes 6O3C [6], 4JKV [4], 4QIM [7], and 4N4W [7]).

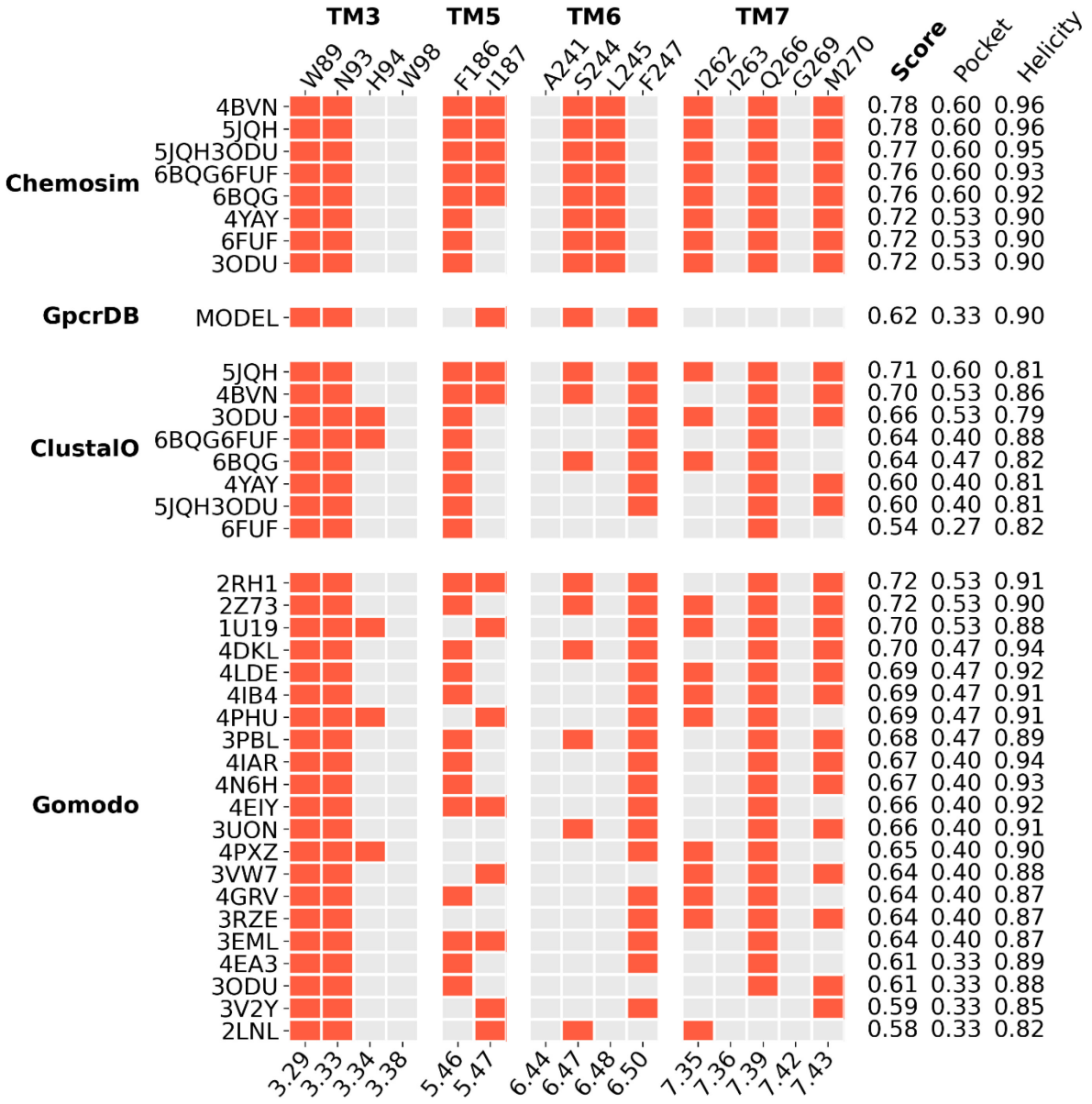


Figure S3.a. Detailed analysis of TAS2R14 binding pocket residues

Meta-scores of top models for each protocol and template. Best models following the Gomodo [8] and ClustalO [9] protocols were selected based on their DOPE score [10]. For BitterDB [11], the only available model did not satisfy our structure quality criteria. The x-axis labels correspond to the Ballesteros-Weinstein numbering of each residue [12]. The left y-axis provides the PDB code of each template except for GPCRdb, where the model was retrieved directly from their website. The right y-axis shows the meta-score, pocket score, and helicity score for each selected model.

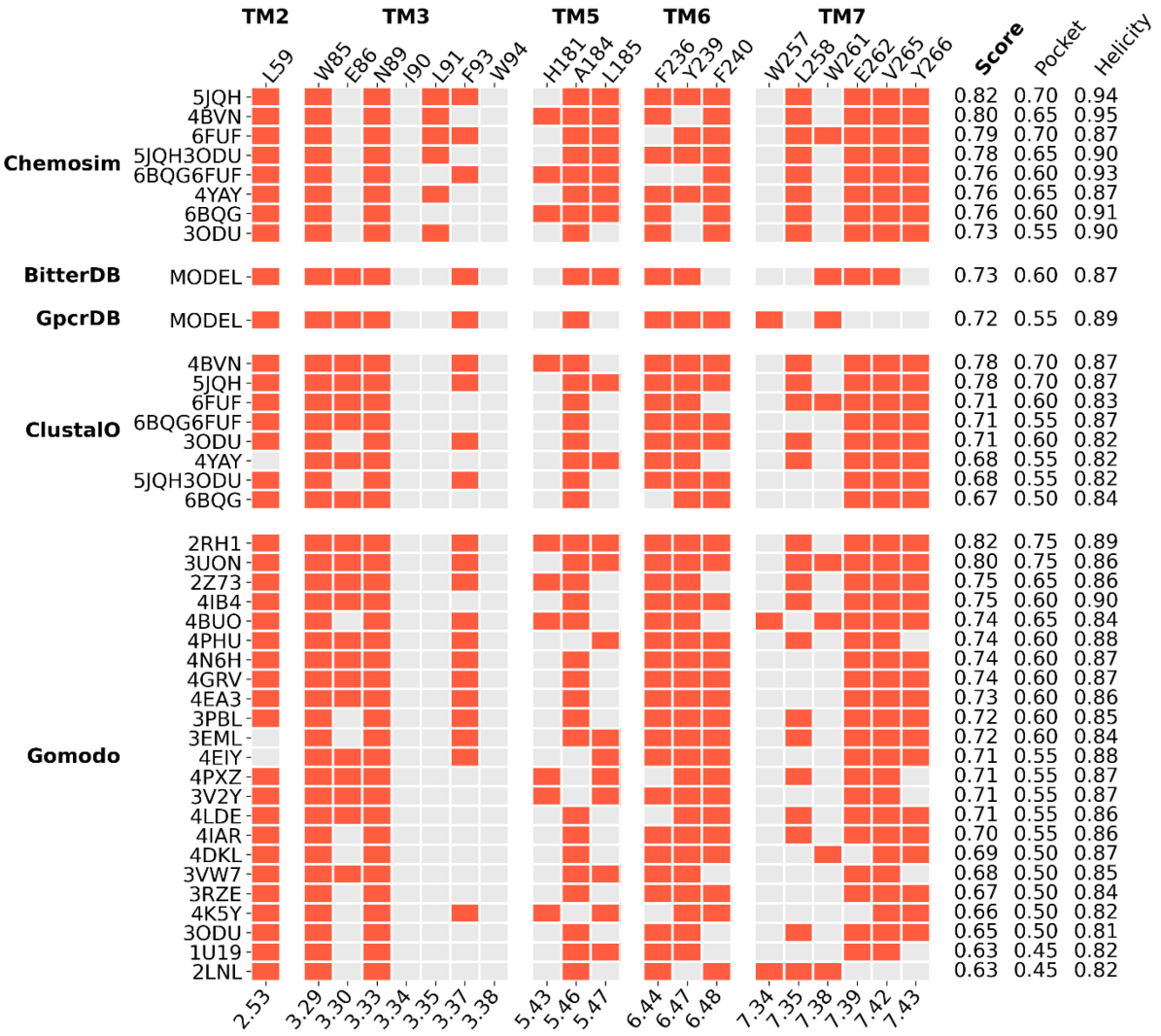


Figure S3.b. Detailed analysis of TAS2R16 binding pocket residues

Meta-scores of top models for each protocol and template. Best models following the Gomodo and ClustalO protocols were selected based on their DOPE score. The x-axis labels correspond to the Ballesteros-Weinstein numbering of each residue. The left y-axis provides the PDB code of each template except for BitterDB and GPCRdb, where the model was retrieved directly from their website. The right y-axis shows the meta-score, pocket score, and helicity score for each selected model.

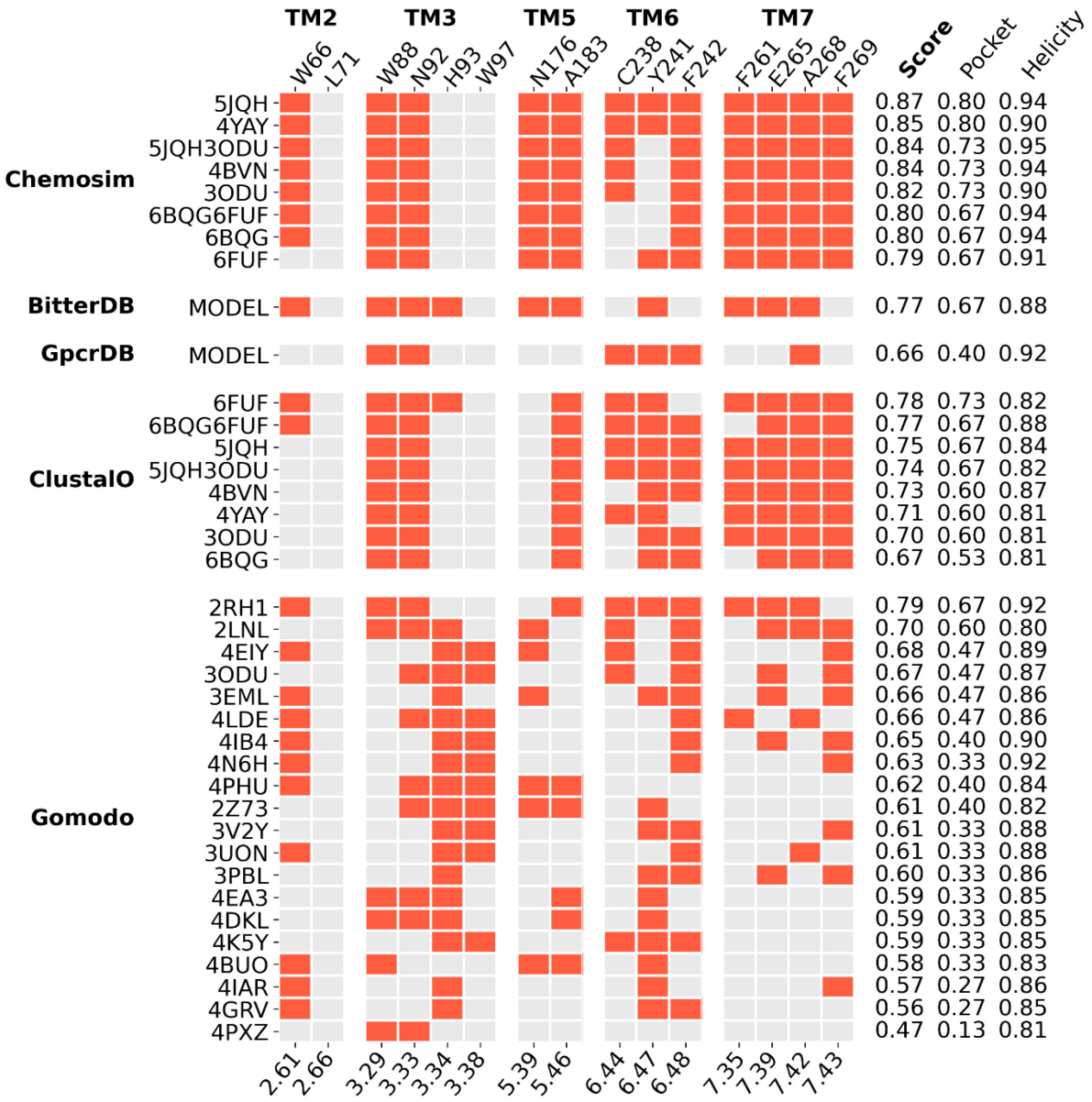


Figure S3.c. Detailed analysis of TAS2R46 binding pocket residues

Meta-scores of top models for each protocol and template. Best models following the Gomodo and ClustalO protocols were selected based on their DOPE score. The x-axis labels correspond to the Ballesteros-Weinstein numbering of each residue. The left y-axis provides the PDB code of each template except for BitterDB and GPCRdb, where the model was retrieved directly from their website. The right y-axis shows the meta-score, pocket score, and helicity score for each selected model.

Figure S4.a. Analysis of TAS2R14 transmembrane helicity

Ramachandran number (R) plot of each residue, numbered by their Ballesteros-Weinstein (BW) position, for the models produced by the best template for each protocol. Standard deviation is represented by the shaded area, and the green zone corresponds to R values typically found in alpha helices of crystallographic GPCR structures (0.32 to 0.38).

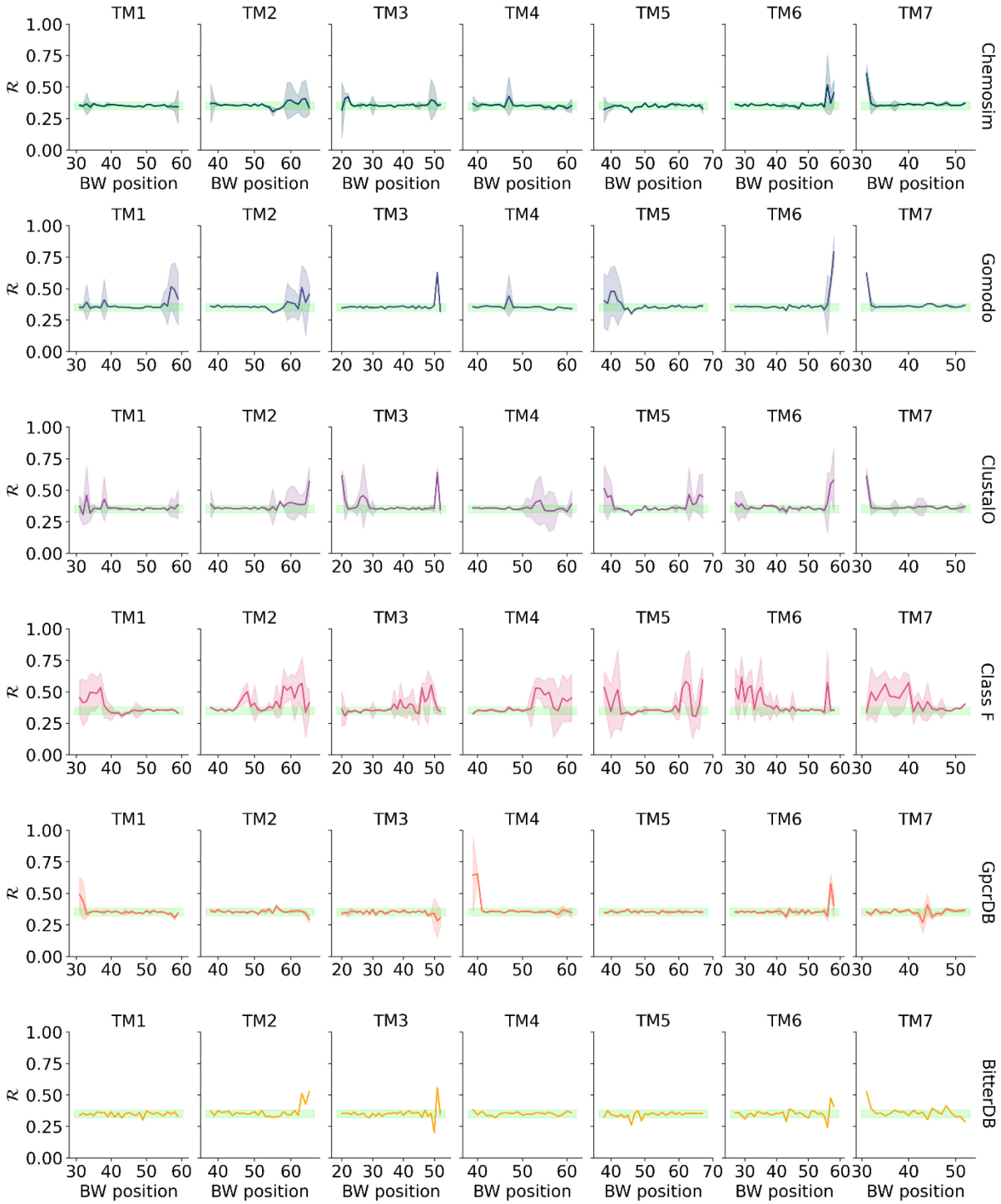


Figure S4.b. Analysis of TAS2R16 transmembrane helicity

See figure caption S4.a.

Figure S4.c. Analysis of TAS2R46 transmembrane helicity

See figure caption S4.a.

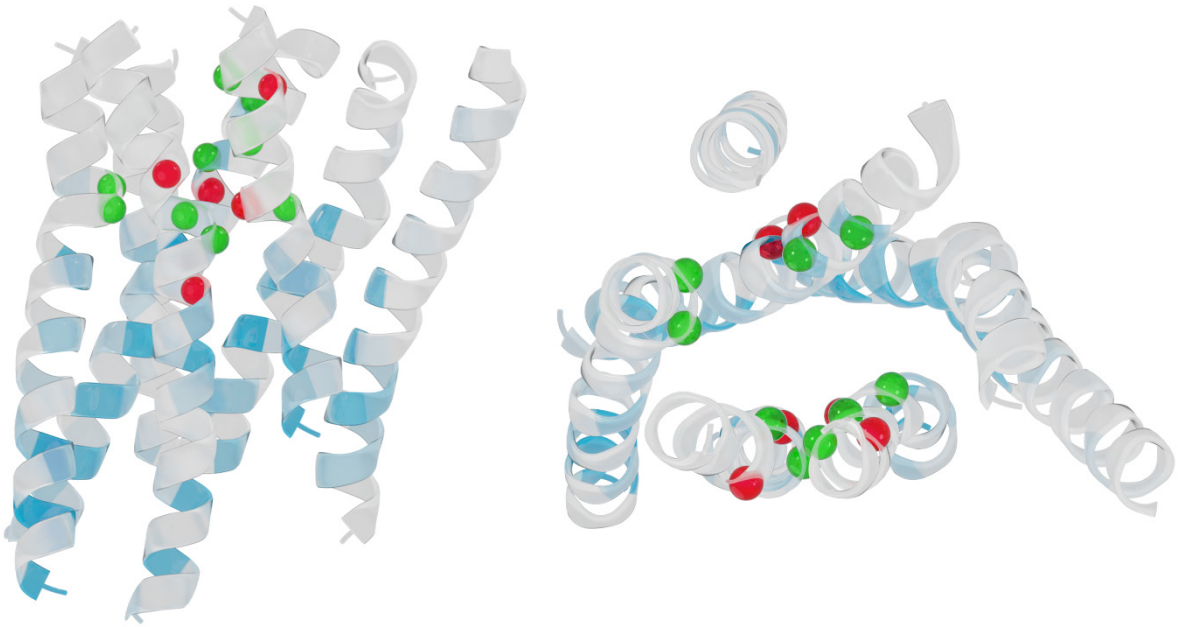


Figure S5.a. Structure of the TAS2R14 model with the highest meta-score

Structure of the best *Chemosim* model obtained from the present study. The residues defining the binding pocket are shown as spheres if their side chains are oriented outward (red) or inward (green) from the pocket and follow from the results shown in Fig S3. Positions of the highly conserved residues in the human TAS2R family are indicated by a color scale, from 50% or less conservation (white) to 100% (blue).

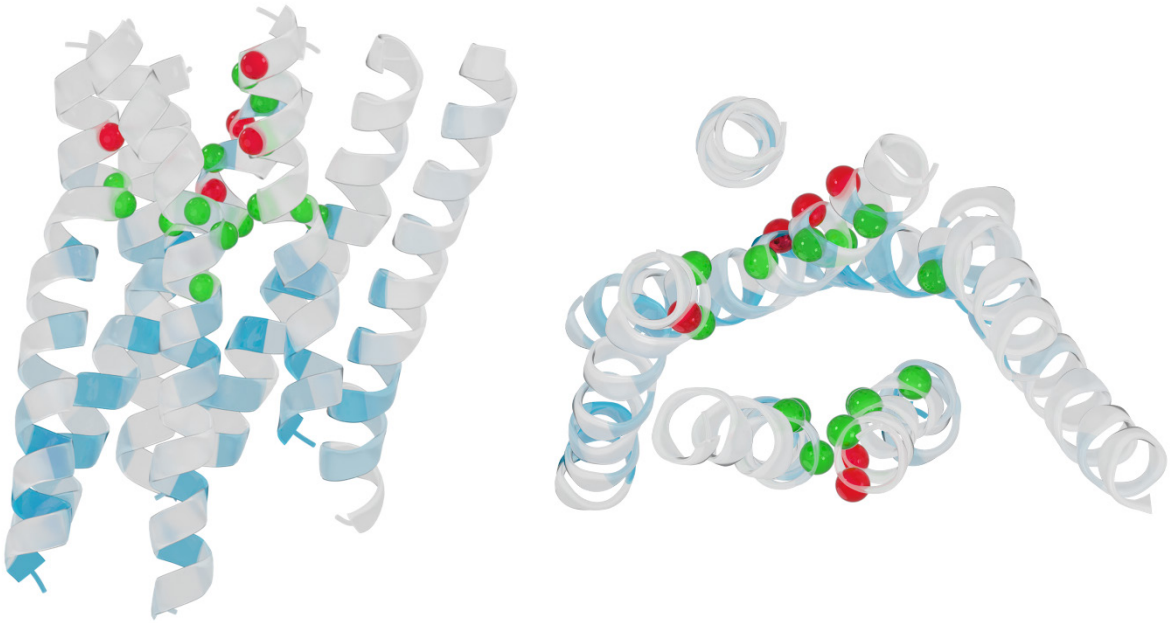


Figure S5.b. Structure of the TAS2R16 model with the highest meta score

See figure caption S5.a.

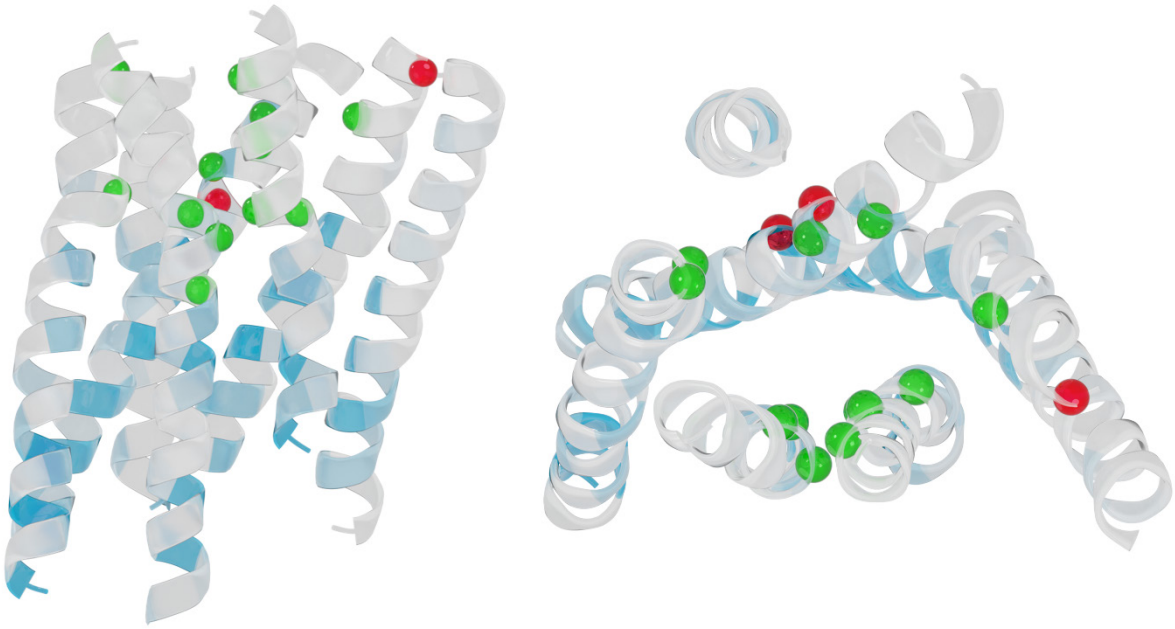


Figure S5.c. Structure of the TAS2R46 model with the highest meta score

See figure caption S5.a.

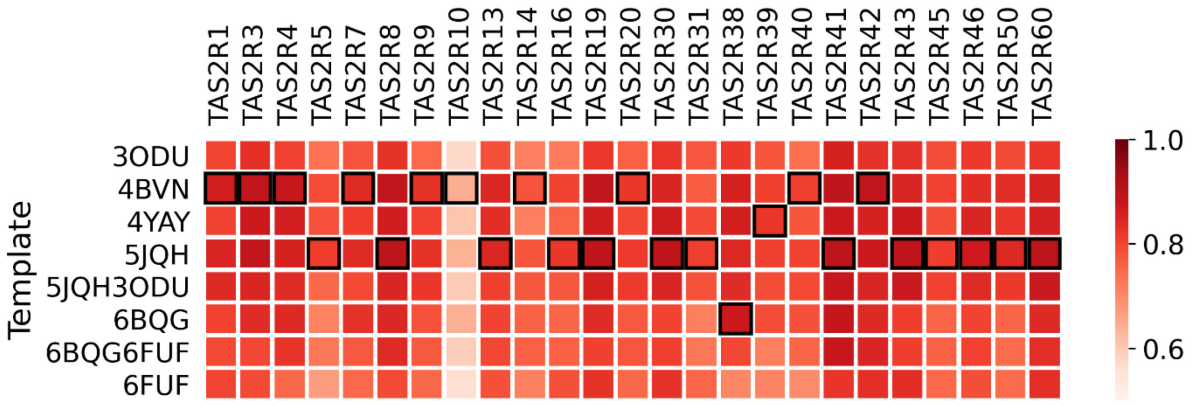


Figure S6. Selection of TAS2R models according to various class A templates

TAS2R models were built following the *Chemosim* protocol. The best models are shown with black boxes and were selected according to the highest meta-score. For all receptors, a consensus TAS2R cavity was used for the detection of residues oriented in the binding pocket. This consensus cavity was composed of residues 3.29, 3.33, 3.34, 3.38, 5.46, 6.44, 6.47, 6.48, 7.35, 7.39, 7.42, and 7.43 and was completed by receptor-specific cavity residues highlighted in the annotated TAS2Rs MSA that is provided in the supplementary files (TAS2R-msa-annotated.xlsx).

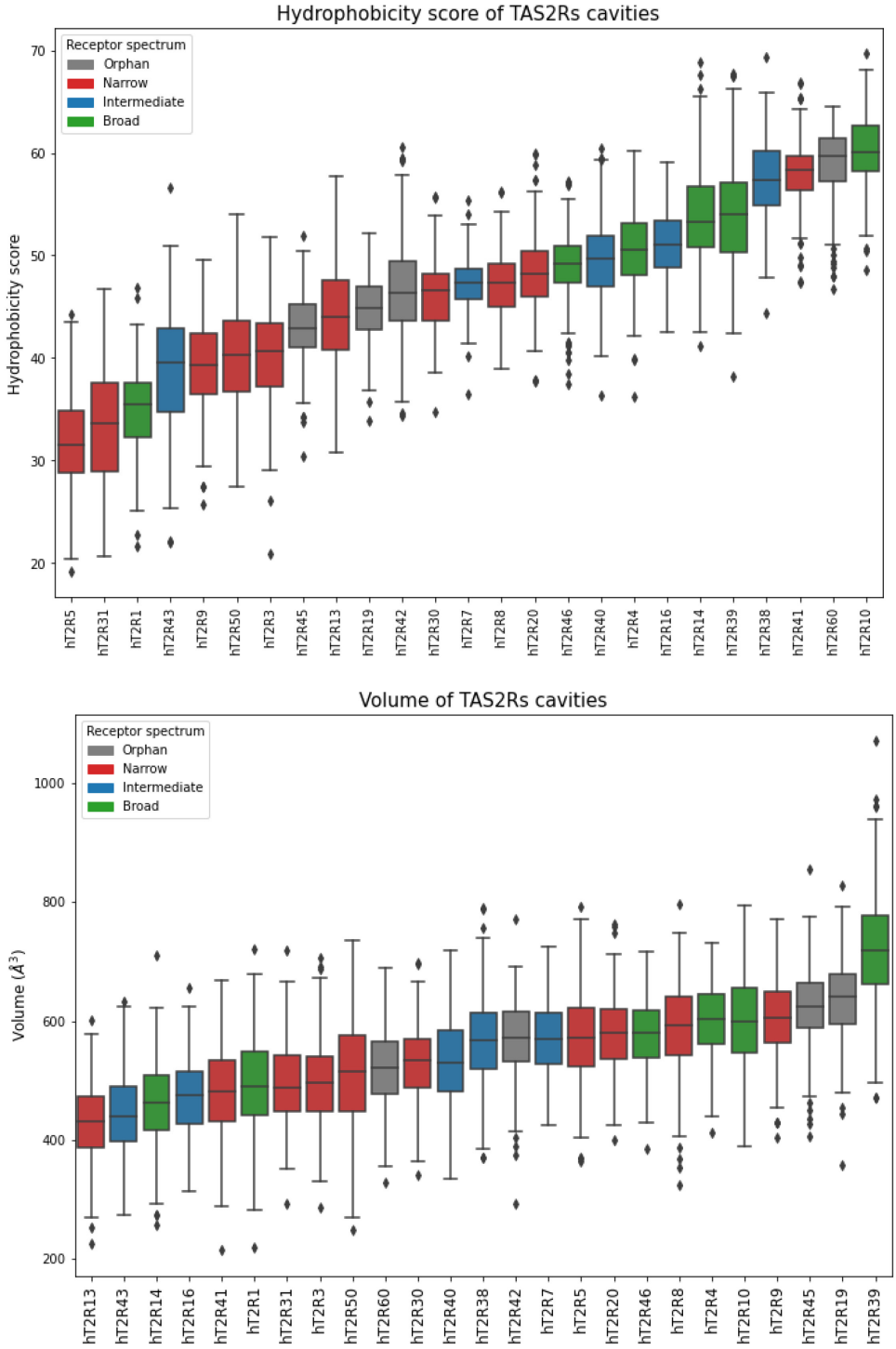


Figure S7. Structural analysis of TAS2R binding pocket

Box-plot of hydrophobicity and volume of TAS2Rs binding pocket. The box extends from the lower to upper quartile values of the data, with a line at the median and outliers plotted in diamonds. The top 250 models for each TAS2R produced by the *Chemosim* protocol and selected templates as shown in Figure S6 were analyzed by MDpocket [13] and colored according to the receptive range (broad, intermediate/specific, narrow, and orphan receptors in green, blue, red, and grey, respectively). A positive hydrophobicity score means that the cavity is mainly hydrophobic.

Table S1. Summary of the most conserved TAS2R amino acids

The most conserved TAS2R residues (above 80% sequence identity) and those involved in TAS2R hallmarks (in yellow/bold) used for multiple sequence alignment with OR and class A templates.

	ClassA motif	OR motif	TAS2R Motif	TAS2R Consensus	Conservation	BW numbering	TAS2R14	TAS2R16	TAS2R46
TM1				G	88%	1.46	G20	I21	G20
	GNxLV	GNLLI	NGFI	**N**	**92%**	**1.50**	N24	S25	N24
				G	**72%**	**1.51**	S25	S26	G25
				F	**84%**	**1.52**	F26	L27	F26
				I	**92%**	**1.53**	I27	I28	I27
ICL1				W	80%		W35	W36	W35
TM2				D	84%	2.40	D45	D46	D45
				I	84%	2.42	I47	I48	I47
				L	80%	2.43	L48	L49	L48
	LAxAD	LSxxD	LAxSR	**L**	**100%**	**2.46**	L51	L52	L51
				A	**64%**	**2.47**	A52	G53	A52
				S	**84%**	**2.48**	S54	S55	S54
				R	**96%**	**2.49**	R55	R56	R55
				L	92%	2.53	L58	L59	L58
TM3				W	84%	3.29	W89	W85	W88
				N	84%	3.33	N93	N89	N92
				W	100%	3.38	W98	W94	W97
	L	L	L	**L**	**96%**	**3.43**	L103	L99	L102
				F	80%	3.46	F106	F102	F105
				Y	92%	3.47	Y107	Y103	Y106
	DRY	MAYDRYVAIC	KIANFS	**K**	**92%**	**3.50**	K110	K106	K109
				I	**88%**	**3.51**	I111	V107	I110
				A	**76%**	**3.52**	A112	S108	A111
				N	**64%**	**3.53**	N113	S109	N112
				F	**84%**	**3.54**	F114	F110	F113
				S	**64%**	**3.55**	S115	T111	S114
ICL2				F	88%		F119	F115	F118
				L	88%		L122	L118	L121
TM4				K	84%	4.39	K123	R119	K122
	W	W	LLG	**L**	**88%**	**4.50**	L134	L130	L133
				L	**80%**	**4.51**	L135	L131	L134
				G	**72%**	**4.52**	V136	G132	G135
ECL2				N	100%		N162	N163	N161
				T	96%		T164	T165	T163
TM5	P	PF	PF	**P**	**92%**	**5.50**	P190	P188	P187
				F	**72%**	**5.51**	F191	F189	F188
				L	80%	5.55	L195	L193	L192
	Y	Y	F	**F**	**52%**	**5.58**	F198	T196	F195
				L	100%	5.61	L201	L199	L198
				S	100%	5.64	L204	S202	L201
				L	96%	5.65	M205	L203	L202
ICL3				H	96%		H208	Q206	H205
				G	84%		I218	G213	I215
				D	84%		D221	N216	D218
				P	80%		A222	P217	P219
TM6	KxxK	RxKAFSTC	HxKALKT	**H**	**92%**	**6.30**	H227	R222	H224
				K/R	**60%**	**6.32**	G229	T224	K226
				A	**92%**	**6.33**	V230	A225	A227
				L	**64%**	**6.34**	K231	L226	L228
				K/Q	**88%**	**6.35**	S232	R227	Q229
				T/S	**60%**	**6.36**	V233	S228	T230
				F	96%	6.40	F237	L232	F234
				L	80%	6.43	Y240	V235	L237
	CWLP	FYG	YFL	**Y**	**64%**	**6.47**	S244	Y239	Y241
				F	**60%**	**6.48**	L245	F240	F242
				L/I/V	**76%**	**6.49**	S246	L241	L243
TM7	NPxxY	PxxNPxIY	PxxHSFIL	**P**	**76%**	**7.46**	P273	I269	P272
				H	**96%**	**7.49**	H276	H272	H275
				S	**68%**	**7.50**	S277	S273	P276
				F	**60%**	**7.51**	C278	T274	F277
				I	**76%**	**7.52**	V279	S275	I278
				L	**96%**	**7.53**	L280	L276	L279
				I	92%	7.54	I281	M277	I280
				N	80%		N284	S280	N283
				L	96%		L287	L283	L286

Table S2. Mutations tested *in vitro* to assess the 3D model

Mutations	TAS2R motifs	Location/role
I90A/S$^{3.35}$, L91A/S$^{3.36}$, L185A$^{5.47}$	n.a.	inside pocket
T100A$^{3.44}$	Negative control	outside pocket
S97A/N$^{3.41}$	n.a.	receptor surface/ receptor trafficking
F236A/Q$^{6.44}$	P$^{5.50}$A$^{3.40}$F$^{6.44}$	pocket cradle/ hydrophobic connector, agonist sensing
Y239F$^{6.47}$	YF$^{6.48}$L	pocket cradle/ transmission switch, agonist sensing
V45S/F$^{2.39}$	Next to PxxHS$^{7.50}$FIL	intracellular part/ hydrophobic barrier
L42A/S^{ICL1}, M43A/S^{ICL1}	Negative control	intracellular part
A221L$^{6.29}$, R222A/H$^{6.30}$	HxK$^{6.32}$ALKT	G protein binding site/ G protein selectivity

Table S3. Salicin-induced *in vitro* response in wild-type and mutant TAS2R16

	Mutations	EC_{50} † (mM)	Maximal Response ($\Delta F/F_0$)
WT		0.98 ± 0.01	0.55
I90	I90A	3.34 ± 0.03***	0.50
	I90S	3.20 ± 0.11***	0.31
L91	L91A	2.85 ± 0.03***	0.57
	L91S	6.05 ± 0.03***	0.47
L42	L42A	0.61 ± 0.04	0.37
	L42S	1.23 ± 0.05	0.33
M43	M43A	0.53 ± 0.12	0.45
	M43S	1.77 ± 0.13**	0.40
V45	V45S	3.30 ± 0.12***	0.41
	V45F	2.79 ± 0.12**	0.26
S97	S97A	0.17 ± 0.04***	0.50
	S97N	0.92 ± 0.11	0.31
T100	T100A	0.50 ± 0.06	0.61
L185	L185H	3.87 ± 0.05***	0.27
A221	A221L	3.78 ± 0.04***	0.38
R222	R222A	5.10 ± 0.08***	0.34
	R222H	0.69 ± 0.10	0.52
F236	F236A	10.38 ± 0.11***	0.39
	F236Q	0.57 ± 0.08	0.50
Y239	Y239F	11.30 ± 0.08***	0.42

† Values are means ± SEM; Statistical significance is indicated by *** $P < 0.001$, ** $P < 0.01$, and * $< P$ 0.05 vs. the WT group (one-way ANOVA followed by Dunnett's test)

Other supplementary information files

The MSA of human TAS2Rs and a selection of ORs and class A templates (TAS2R-OR-templates.pir); the MSA of reviewed mammalian TAS2R sequences obtained from Uniprot (mammalian-TAS2R.pir); and an annotated MSA of human TAS2Rs (TAS2R-msa-annotated.xlsx).

References

[1] D.P. Staus, R.T. Strachan, A. Manglik, B. Pani, A.W. Kahsai, T.H. Kim, L.M. Wingler, S. Ahn, A. Chatterjee, A. Masoudi, Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation, Nature, 535 (2016) 448-452.

[2] A.J. Kooistra, S. Mordalski, G. Pándy-Szekeres, M. Esguerra, A. Mamyrbekov, C. Munk, G.M. Keserű, D.E. Gloriam, GPCRdb in 2021: integrating GPCR sequence, structure and function, Nucleic Acids Research, 49 (2021) D335-D343.

[3] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera—a visualization system for exploratory research and analysis, Journal of computational chemistry, 25 (2004) 1605-1612.

[4] C. Wang, H. Wu, V. Katritch, G.W. Han, X.-P. Huang, W. Liu, F.Y. Siu, B.L. Roth, V. Cherezov, R.C. Stevens, Structure of the human smoothened receptor bound to an antitumour agent, Nature, 497 (2013) 338-343.

[5] B. Webb, A. Sali, Comparative Protein Structure Modeling Using MODELLER, Current Protocols in Bioinformatics, 54 (2016) 5.6.1-5.6.37.

[6] I. Deshpande, J. Liang, D. Hedeen, K.J. Roberts, Y. Zhang, B. Ha, N.R. Latorraca, B. Faust, R.O. Dror, P.A. Beachy, Smoothened stimulation by membrane sterols drives Hedgehog pathway activity, Nature, 571 (2019) 284-288.

[7] C. Wang, H. Wu, T. Evron, E. Vardy, G.W. Han, X.-P. Huang, S.J. Hufeisen, T.J. Mangano, D.J. Urban, V. Katritch, Structural basis for Smoothened receptor modulation and chemoresistance to anticancer drugs, Nature communications, 5 (2014) 1-11.

[8] M. Sandal, T.P. Duy, M. Cona, H. Zung, P. Carloni, F. Musiani, A. Giorgetti, GOMoDo: a GPCRs online modeling and docking webserver, PloS one, 8 (2013) e74092.

[9] F. Sievers, D.G. Higgins, Clustal Omega for making accurate alignments of many protein sequences, Protein Science, 27 (2018) 135-145.

[10] M.y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures, Protein science, 15 (2006) 2507-2524.

[11] A. Wiener, M. Shudler, A. Levit, M.Y. Niv, BitterDB: a database of bitter compounds, Nucleic acids research, 40 (2012) D413-D419.

[12] J.A. Ballesteros, H. Weinstein, [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors, in: Methods in neurosciences, vol. 25, Elsevier, 1995, pp. 366-428.

[13] P. Schmidtke, A. Bidon-Chanal, F.J. Luque, X. Barril, MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories, Bioinformatics, 27 (2011) 3276-3285.

Discussion

Data quality

In the first two chapters, I have demonstrated applications of a ligand-based method, namely machine-learning (ML), to tackle chemosensory problems. In the first part, the models were trained to classify compounds as agonists or non-agonists of a pest insect olfactory receptors, and then used to prioritize volatile molecules that had a behavioral effect on said insect. In the second chapter, the ML model was developed to estimate the sweetness of compounds relative to sucrose and served to discover a novel scaffold of sweetener. In both cases, the models were trained on data obtained experimentally i.e., someone bought each compound before carrying out an experiment to measure the chemosensory endpoint. This raises a question on the impact of the quality of experimental data on the predictions made by ML models. It is usually well accepted that the output of a model reflects the data it was trained on, meaning that a model trained on noisy data will produce faulty results, which is perfectly illustrated by the GIGO concept: garbage in, garbage out.

In the present cases, the first subset of errors can come from an inexact description of the molecules used in the experiment. Such errors can arise from the identifier of the molecule, such as a name that corresponds to more than one molecule, or a name with an incomplete stereochemistry, or an incorrect CAS number. Alternatively, inaccuracies can also emerge from the structure i.e., when using non-isomeric SMILES although the stereoisomer used in the study is well defined, or because of the absence of structural standardization steps that enforce ambiguous moieties such as aromatic rings, nitro groups and tautomers to be represented in a canonical way, which may impact descriptor calculation. These errors can be time-consuming to repair and are typically resolved during the data curation step of the modeling pipeline. The second set of errors comes directly from the measurement of the endpoint and mostly includes disparities in experimental procedures and measurement errors inherent to fluctuations in the readings of the apparatus.

For example, in the second chapter the model was trained to predict the relative sweetness of compounds which is a property that is not measured by an instrument but perceived by humans. To obtain this value, a panel is used to evaluate a solution of sweetener and a solution of sucrose until both are perceived isointense. Depending on the experimental setup, the reference

concentration might be at the detection threshold or at high concentration, and the relative sweetness might then be calculated as the concentration ratio or the mass ratio between both solutions. The sensory panel might be composed of highly trained individuals, or regular panelists, and the distribution of sex and age might differ, and most importantly their taste receptors might carry single-nucleotide polymorphisms (SNPs) which will affect how each compound is perceived. All these differences in the protocol add noise in the collected data which is then passed to the ML algorithm. In our case, the resulting model had a mean absolute error of 0.5 on the $\log_{10}$ of relative sweetness in the test case, corresponding to a 3-fold error on the relative sweetness. While this may seem considerable at first, to some extent it reflects the large error bar that accompanies the experimental data.

Conversely, in the *Spodoptera littoralis* dataset, the electroantennography results are directly obtained from the olfactory sensory neuron (OSN) of each fly exposed to an odorant, so that there is no interpretation of the olfactory signal by the brain. These neurons express the OR of interest thus SNPs may not affect ligand binding and signal transduction, although peri-receptor events peculiar to each fly might slightly modulate the response. Also, all the measurements were performed by researchers from the same lab using the same protocol. Overall, this experimental setup is less likely to produce noisy data in comparison with the sugars and sweeteners dataset.

Extracting knowledge from machine-learning models

QSAR models are often described as black boxes that can easily and often successfully predict properties, albeit without conveying any meaningful information to explain the decision process. This holds true in the cases studied in the first two chapters, as the algorithms used either reproduce parts of the training data without learning anything from it (k-Nearest Neighbors) or are too complex to extract the process leading to the final decision in understandable terms (Support Vector Machine, Random Forest, AdaBoost Trees) as opposed to more simple models like a decision tree or a multilinear regressor. However, some recent efforts towards the interpretability of machine-learning models have come through [1], such as LIME (local interpretable model-agnostic explanations) [2] or SHAP (Shapley additive explanations) [3]. These are post-hoc interpretability analysis methods that can be applied to any kind of model and features.

For example, in LIME, surrogate interpretable models are trained on top of the black box model to explain its predictions. To do so, it creates small variations of an instance that is being predicted to create a new dataset, labels it with predictions from the black box, and then trains an interpretable model on this data. The surrogate model is thus a local approximation of the black box model. Unfortunately, LIME and other similar methods do not consider the correlation between features when generating the local dataset from the instance being predicted, although these correlations are present in molecular descriptors i.e., changing the fraction of oxygen atoms should also change the molecular weight and other properties. Since this is not the case with such method, the molecules corresponding to those incorrect descriptors would also be invalid, which is a limiting factor for the application of such explainable AI methods to ligand-based problems.

Currently, the most straightforward way towards interpretability appears to be the restriction to both interpretable descriptors and ML algorithms during the development of the model, which often leads to a decrease in performance. Recent advances in deep learning applied to drug discovery also came with their share of methodological developments for interpretability, some of which were reviewed by Jiménez-Luna *et al.* [4] among other explainable AI methods. For example, Preuer *et al.* [5] used integrated gradients to extract atom-wise contributions to the final decision of their deep feed-forward neural network, and map these contributions on the structure. They also proposed a method to extract the general molecular substructures that are learned by convolutional neural networks, instead of limiting interpretability to individual predictions. Tang *et al.* [6] designed a message-passing neural network based on self-attention, an architecture which allows them to formulate attention weights as individual atom contributions for a given prediction. All of these atomic contribution approaches also benefit from powerful visualization techniques such as similarity maps [7] that display each atom's weight as a colored contour map directly onto the molecular structure.

Overall, model interpretability is currently highly dependent on which chemical features and ML algorithm were used, and the choice of whether or not explaining the prediction is important should thus be considered at the early stages of model development rather than after virtual screening.

Applicability domain

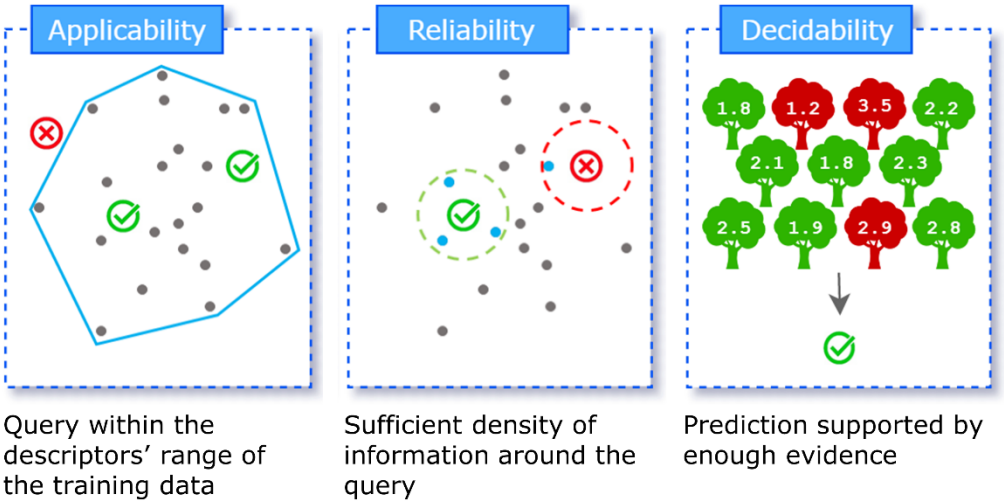


Figure 1: Concepts covered by different implementations of an applicability domain.

While I have already described the concept of the applicability domain(s) in more details in the Publication 3 section, this term is usually reserved to QSAR models. Briefly, an applicability domain determines if a model should be used and trusted to predict the activity of a given molecule (Figure 1).

However, structural models, even more so homology models, also have scenarios where it is acceptable to use them and some where it is not. For example, the homology models of TAS2Rs presented in the last chapter are optimized to recapitulate as best as possible what we know on the binding pocket while maintaining the typical GPCR structure with 7 alpha-helical domains. By mutating a position in the sequence and studying its effect on the response by testing multiple known ligands, it is possible to uncover if the mutated position interacts directly with the ligand and thus if the position is part of the binding pocket. If, for the same mutant, some wild-type active ligands respond and others do not, then it is reasonable to assume that this specific position is in contact with the ligand and that it belongs to the pocket. By following this methodology, we tuned our homology models to reflect the information gathered from single-point mutagenesis data available in the literature. While no model could satisfy all these constraints, our best ones could position from 60 to 85% of the identified residues inside the pocket except for TAS2R10 (33%). This makes most of our models suitable for a more thorough analysis of the properties of the binding pocket and for structure-based virtual screening approaches such as docking.

However, the other parts of the receptor, especially loops and the G protein coupling site, might not be as accurate as the binding pocket as there is little to no information available for these regions, hence the need to clearly define the applicability domain of the model, in our case the binding pocket, where the model can be used with confidence. This does not mean, however, that the model cannot be used outside of this scope but rather that the information that will be extracted from it will be less reliable. For example, once our TAS2R16 model was generated, we then used it to hypothesize which residues are involved in signal transduction and tested them *in vitro* with single-point mutagenesis for validation. This clearly fell outside of the applicability domain (except for the "toggle switch" residue which senses the ligand), yet the experiments validated that these positions are relevant during receptor activation, thus expanding the scope of use of the model.

Conversely, using any sort of numerical model (ligand-based or structural) within its applicability domain does not imply that the outcome should be considered as a proof, but rather as a hypothesis pending experimental validation. Thanks to the work done by our collaborators in Dijon, Versailles, and South Korea, all the hypotheses that I generated computationally were to some extent validated experimentally.

Virtual screening of TAS2Rs

Subsequently to the generation of the homology models of the TAS2R repertoire, we started investigating their usefulness in finding agonists for two orphan receptors (TAS2R42 and TAS2R60) and a narrowly tuned one (TAS2R20, 3 known ligands) through virtual screening (VS). This work was carried out by Maxence Lalis, an intern student who I co-supervised.

He started by updating the BitterDB with recently published experimentally tested molecules (both active and inactive). The next step was to establish a docking protocol benchmarked on three receptors with sufficient mutagenesis data to ensure the structural quality of the homology models, and different receptive ranges: TAS2R14 (broadly tuned with 171 agonists), TAS2R16 (β-D-glucopyranoside specialist with 19 agonists) and TAS2R46 (intermediate with 71 agonists).

For each receptor, three models were selected through clustering based on the sidechain of residues involved in ligand binding according to mutagenesis data, followed by the selection of the best model in each cluster according to our metascore defined in Publication 4. Several strategies were investigated to optimize the discrimination of active from inactive molecules.

As shown by de Graaf *et al.* [8] and Jaiteh *et al.* [9], the second extracellular loop (ECL2) of class A GPCR homology models can greatly affect ligand enrichment, yet in cases where experimental data on the loop is scarce and the accuracy of its modeling is uncertain, it is best to remove ECL2 from the models for VS. In our case, both full and loopless models were considered during the optimization of the VS protocol. Other optimizations included the size of the grid box used by AutoDock Vina [10] and different rescoring strategies were examined such as taking the minimum, average or median score of the generated poses, as well as the CorrScore [11], Ligand Efficiency (LE), LESA and LEln [12]. From our benchmark study, removing the ECL2 from the models, fitting the grid box to accommodate all pocket residues involved in ligand binding according to mutagenesis data, and scoring each ligand by their consensus best score (taking, for each homology model, the minimum Vina scores among all poses and then averaging it) produced the best results, as evaluated by the area under the receiver operating characteristics curve (AUROC).

The next step was to virtually screen a database of commercially available compounds, Sigma Aldrich, to identify putative agonists prior to experimental validation. The dataset, comprised of 233,097 molecules, was screened with the above protocol applied to both benchmarked receptors (TAS2R14, TAS2R16 and TAS2R46) and orphan/narrowly-tuned receptors (TAS2R20, TAS2R42, TAS2R60). Several filtering strategies were included to reduce the list of prioritized compounds (Figure 2). First, the top 4% of the best scored ligands were kept. Second, we used a filter based on the Euclidean distance between the center of mass of known agonists and that of the ligand, as the poses of some inactive molecules tended to drift towards the corners of the grid box during the benchmark. This way, the first half of ligands with their center of mass closest to that of known agonists were kept. Next, a filter based on bitter tastants physicochemical properties, namely the molecular weight, atomic logP, number of rings and hydrogen bond donors and acceptors, was added. Molecules that fell outside of the ranges explored by known bitter molecules from the BitterDB were excluded. Additionally, an interaction fingerprint (IFP) between each docking pose and their receptor was calculated and compared to a reference fingerprint constructed from the mutagenesis data. Ligands with a Tanimoto similarity to the reference below 0.2 were discarded. Finally, to further reduce the list of compounds, a last filtering step was performed using the Bemis-Murcko scaffold [13] of the remaining hits. Each compound was clustered according to its scaffold, and among clusters that contained enough molecules, the one with the best Vina consensus score was selected.

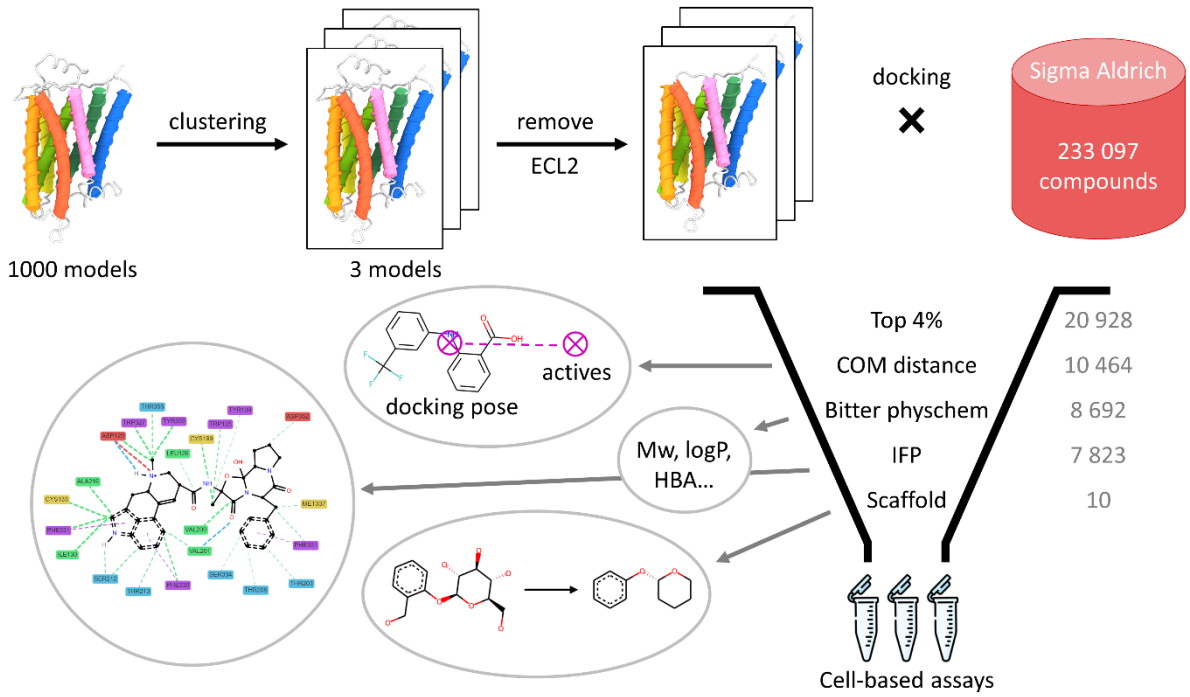


Figure 2: TAS2R virtual screening workflow. Numbers on the right side correspond to the number of hits remaining after each filtering step for TAS2R14. COM distance: distance between the center of mass of known actives and the docking pose. Bitter physchem: Physicochemical properties of known bitter molecules. IFP: interaction fingerprint. Scaffold: Murcko scaffold clustering.

The cutoff on the number of compounds was set to 50 for TAS2R14 and TAS2R16, and 30 for the others. Interestingly, some of the selected scaffolds match with known agonists for TAS2R14, TAS2R46 and TAS2R20. Amidst the molecules that were not selected by the clustering procedure, those that were described as tasting bitter in the literature, without being assigned to a specific TAS2R, were added to the pool of selected compounds.

This protocol resulted in the selection of 20 compounds for TAS2R14 (Figure 3), half by clustering and the other half from the literature search that labelled them as bitter. Because 6 of the compounds were either not available commercially or were not described as soluble in DMSO, 14 compounds were sent to our collaborators for *in vitro* testing. Pending validation of these compounds, the hits selected for the other TAS2Rs will be experimentally tested as well. As shown in Figure 3, compound N, which corresponds to isorhamnetin, is a known agonist for both TAS2R14 and TAS2R39 [14] and will serve as a positive control. For the other compounds, the Tanimoto coefficient with their closest known TAS2R agonist is quite low (0.34 ± 0.13), suggesting that the VS approach can explore a broad part of the chemical space compared to ligand-based methods which tend to be more strictly limited by their applicability domain.

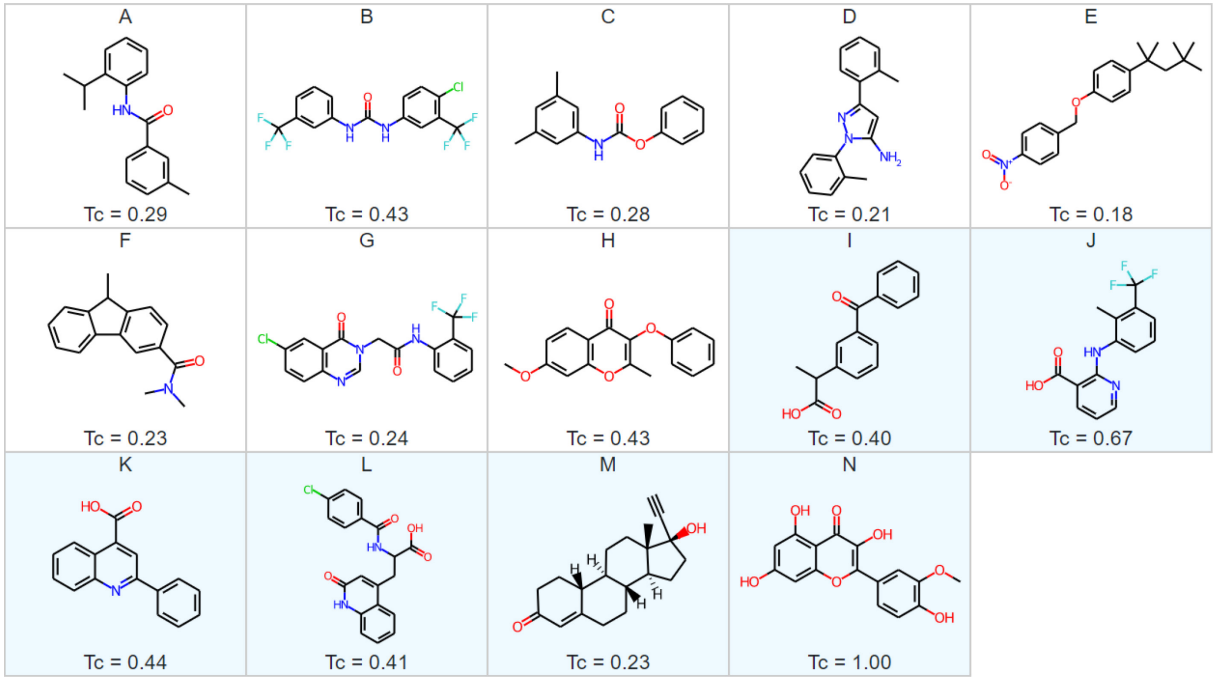


Figure 3: Potential hits selected for cell-based assays. Structures described as "bitter" in the literature are highlighted in pale blue. Tc: Tanimoto coefficient to the closest known TAS2R agonist.

Towards a better understanding of class A GPCRs

Because molecular interactions are at the root of ligand sensing and receptor activation, providing tools to automatize the analysis of interactions made by biological complexes can prove useful in the computational toolkit of chemists. One such analysis, called an interaction fingerprint (IFP), encodes 3D interactions between a ligand and a protein as a bitvector, where each bit represents the presence or absence of a specific type of interaction (hydrogen bond, π-stacking…etc.) between the ligand and a residue. As a side-project, I have extended the IFP concept to work with molecular complexes formed of any combination of ligand, protein, DNA, or RNA molecules, extracted from MD trajectories, docking poses, or crystallographic structures, by developing a Python library called ProLIF. The library also lets users define their own interactions, comes with several tutorials, and integrates seamlessly within the Python ecosystem as special attention was given to the interoperability between ProLIF and other packages, whether for data analysis or visualization. A publication describing the implementation and showcasing some examples of analysis on both ligand-protein and protein-protein interactions of class A GPCRs is currently under revision and is provided in the Appendix.

While the tridimensional structure of most chemosensory receptors is unknown, a reasonable number of crystallographic structures are available for class A GPCRs, including some which are co-crystallized with a ligand in the orthosteric binding pocket. However, despite binding to the same site, these ligands can exhibit agonism or antagonism, the reason for which is unknown. We hypothesized that a ligand's differentiation as agonist or antagonist is dictated by its interactions with specific residues inside the pocket. Using the GPCRdb [15], we collected 205 PDB files of class A GPCRs in complex with a ligand in the orthosteric site. Each entry was labelled with their receptor conformational state (active or inactive) and ligand activity (agonist, antagonist, or inverse agonist). After renumbering the residues according to the protein's UniProt sequence and protonating the structures with PDB2PQR [16], ProLIF was used to extract the IFP of all complexes. Since some ligand-GPCR pairs have been crystallized more than once (especially in the opsin family), the corresponding entries were regrouped and their IFP were merged. The residue numbering was then converted to the generic structure-based GPCRdb numbering for class A GPCRs [15] to be able to compare the different structures. This resulted in 148 unique pairs of ligand-GPCR complexes and their consensus IFP, which we used to identify putative differences in the modes of interaction between agonists, and antagonists, as well as between active and inactive structures. The different types of interactions (hydrophobic, H-bond donor…etc.) available for each position were regrouped as a single ligand-residue contact and converted to a contingency table for three groups: agonist *vs* antagonist, agonist *vs* others, and active *vs* inactive state. Other possible groups were not included in the study because not enough data was available. For each group, only positions that had at least 10 members in one of the classes were kept. From there, a two-tailed Boschloo statistical test [17] with a 95% confidence interval was run on each table, and to minimize type I errors due to multiple testing, the p-values were subsequently corrected using the Benjamini-Hochberg procedure [18] with a 5% false discovery rate.

As shown in Figure 4, multiple positions seem to govern ligand differentiation in TM2, TM3 and TM4, and more surprisingly in ECL1. Residue 3.40, which is a conserved hydrophobic position, interact significantly more with agonists, and both 2.64 and 3.40 are often involved when the receptor is in an active state. The residue in the most conserved position of the first extracellular loop, ECL1.50, which is conserved as an aromatic residue, seem to be involved in the negative modulation of the receptor as it is more often interacting with antagonists and inverse agonists than with agonists, and the inactive conformation of the receptor is more often present when this residue takes part in ligand-protein interactions. Another highly crucial

position for the distinction between agonist and antagonist or inverse agonist is the residue 2.60 which is not conserved and can be either hydrophobic or polar.

However, at position 2.59 the orexin family is responsible for 90% of interactions with antagonists, meaning that our preliminary analysis can be biased by imbalanced data and should thus be considered with precautions. Additionally, more data would be needed to conclude on potential family-specific binding modes as the most populated family, adrenoceptors, only has 24 unique complexes experimentally resolved (with 10 agonists, 8 antagonists, 6 inverse agonists).

While we used X-ray structures in this study, such IFPs could also be generated from docking poses and fed to an ML classifier to specifically search for antagonists or inverse agonists in a virtual screening process, as investigated on the β2-adrenergic receptor by Jiménez-Rosés *et al.* [19]. Provided the validation of our TAS2R docking protocol, one could also imagine applying the same IFP approach for the design of specific TAS2R antagonists, notably pertaining to their ectopic expression as potential drug targets.

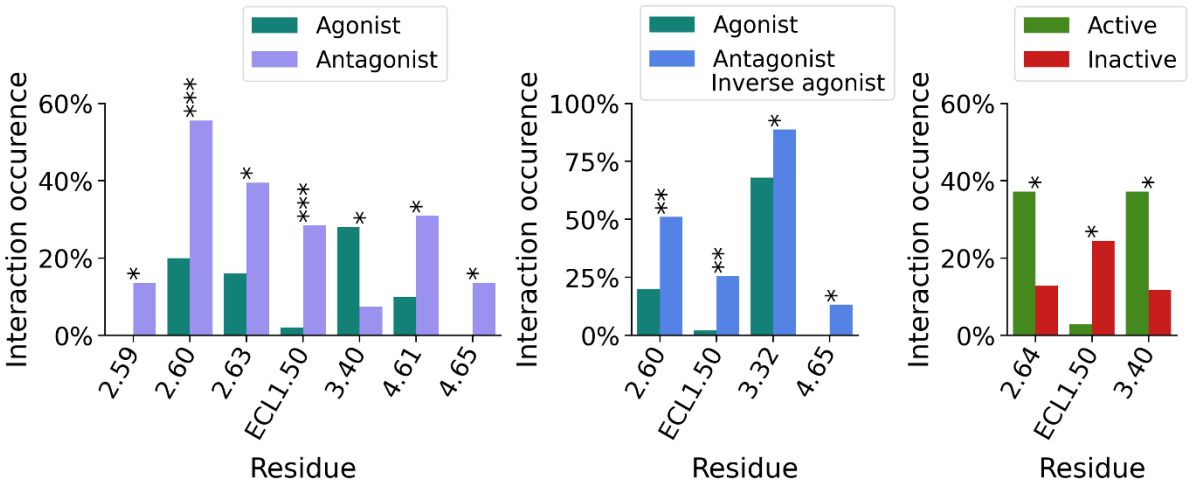


Figure 4: Class A GPCR positions involved in ligand differentiation. Each percentage represents the number of complexes in which the residue is interacting with the ligand, divided by the total number of complexes in each class: 50 agonists, 81 antagonists, 98 antagonists & inverse agonists, 35 actives, and 94 inactives. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

References

1. Molnar C (2019) Interpretable machine learning: a guide for making Black Box Models interpretable. Lulu, Morrisville, North Carolina

2. Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:160204938

3. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc.

4. Jiménez-Luna J, Grisoni F, Schneider G (2020) Drug discovery with explainable artificial intelligence. Nat Mach Intell 2:573–584. https://doi.org/10.1038/s42256-020-00236-4

5. Preuer K, Klambauer G, Rippmann F, et al (2019) Interpretable Deep Learning in Drug Discovery. In: Samek W, Montavon G, Vedaldi A, et al (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing, Cham, pp 331–345

6. Tang B, Kramer ST, Fang M, et al (2020) A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. J Cheminform 12:15. https://doi.org/10.1186/s13321-020-0414-z

7. Riniker S, Landrum GA (2013) Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. J Cheminform 5:43. https://doi.org/10.1186/1758-2946-5-43

8. de Graaf C, Foata N, Engkvist O, Rognan D (2008) Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening. Proteins 71:599–620. https://doi.org/10.1002/prot.21724

9. Jaiteh M, Rodríguez-Espigares I, Selent J, Carlsson J (2020) Performance of virtual screening against GPCR homology models: Impact of template selection and treatment of binding site plasticity. PLoS Computational Biology 16:1–25. https://doi.org/10.1371/journal.pcbi.1007680

10. Oleg T, Arthur J. O (2010) AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. Journal of computational chemistry 31:455–461. https://doi.org/10.1002/jcc

11. Carta G, Knox AJS, Lloyd DG (2007) Unbiasing Scoring Functions: A New Normalization and Rescoring Strategy. J Chem Inf Model 47:1564–1571. https://doi.org/10.1021/ci600471m

12. Ben Shoshan-Galeczki Y, Niv MY (2020) Structure-based screening for discovery of sweet compounds. Food Chemistry 315:126286. https://doi.org/10.1016/j.foodchem.2020.126286

13. Bemis GW, Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928

14. Roland WSU, van Buren L, Gruppen H, et al (2013) Bitter Taste Receptor Activation by Flavonoids and Isoflavonoids: Modeled Structural Requirements for Activation of hTAS2R14 and hTAS2R39. J Agric Food Chem 61:10454–10466. https://doi.org/10.1021/jf403387p

15. Isberg V, Mordalski S, Munk C, et al (2016) GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Res 44:D356–D364. https://doi.org/10.1093/nar/gkv1178

16. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Research 32:W665–W667. https://doi.org/10.1093/nar/gkh381

17. Boschloo RD (1970) Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. Statistica Neerland 24:1–9. https://doi.org/10.1111/j.1467-9574.1970.tb00104.x

18. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

19. Jiménez-Rosés M, Morgan BA, Sigstad MJ, et al (2021) Prediction of ligand-receptor pharmacological activities using a combined docking and machine learning approach. https://doi.org/10.1101/2021.03.18.434755

Conclusion

Smell and taste perception originates from the detection of small molecules by chemosensory receptors expressed in the nasal or oral cavity. Most of these receptors belong to the GPCR family, a well-studied group targeted by more than 30% of modern therapeutic drugs with several structures experimentally solved. Yet, the exact tridimensional structure of these G protein-coupled chemosensory receptors still eludes us and limits our understanding of their structure-function relationships. How do they bind ligands, how do they transmit this information to intracellular secondary messengers during signal transduction, and can we rationally design active molecules targeted at these receptors, are all open questions.

The aim of my PhD thesis was to study the molecular structures at the frontline of chemosensory perception, namely receptors and their ligands, through a computational lens. Accordingly, two objectives were set: the first one to connect machine-learning (ML) algorithms with taste and odor properties of small compounds, and the second one to decipher the molecular basis of taste perception.

To reach the first objective, I focused on two different subjects that were both ideal candidates for machine-learning but also challenging. In the first chapter, I described a 2-step approach for the rational design of natural semiochemicals that can disrupt the destructive behavior of a pest insect, *Spodoptera littoralis*. Those compounds were identified for their ability to interact with specific olfactory receptors of the noctuid moth, by screening a focused library of natural volatile molecules with ML models. Those models were built on the premise of a previous proof-of-concept model that scanned a more diverse chemical space of commercially available compounds. The difficulty here was to train accurate models with few and imbalanced data, but thanks to a feedback loop between *in vivo* data and *in silico* methods, the models' performance at predicting active compounds improved over time. This led to 2 publications:

- Caballero-Vidal, G.; Bouysset, C.; Grunig, H.; Fiorucci, S.; Montagné, N.; Golebiowski, J.; Jacquin-Joly, E. Machine Learning Decodes Chemical Features to Identify Novel Agonists of a Moth Odorant Receptor. *Scientific Reports* **2020**, *10* (1), 1655–1655. https://doi.org/10.1038/s41598-020-58564-9.
- Caballero-Vidal, G.; Bouysset, C.; Gévar J.; Mbouzid H.; Nara C.; Delaroche J.; Golebiowski, J.; Montagné, N.; Fiorucci, S.; Jacquin-Joly, E. Reverse chemical ecology in a moth: machine learning on odorant receptors identifies new behaviorally active agonists. *Under revision.*

In a second chapter, I described the development of an online QSPR platform designed to predict the relative sweetness of compounds, and its subsequent use in the search of intense natural sweeteners. The more challenging part here was to work with noisy data which ultimately impaired the performance of the ML model, as well as the development of a web platform from scratch. It resulted in the discovery of a novel sweet-tasting scaffold belonging to the lignan family, described in a publication:

- Bouysset, C.; Belloir, C.; Antonczak, S.; Briand, L.; Fiorucci, S. Novel Scaffold of Natural Compound Eliciting Sweet Taste Revealed by Machine Learning. *Food Chemistry* **2020**, *324*, 126864–126864. https://doi.org/10.1016/j.foodchem.2020.126864.

Both chapters concluded in the usefulness of ML to drive the search for chemosensory-active molecules, albeit without providing knowledge on the reasons for their activity. It also requires known active compounds to start with, which hampers its application to the deorphanization of chemosensory receptors, although proteochemometrics i.e., including descriptors from both ligands and receptors, could help to bridge that gap.

For the second objective, a focus was made on bitter taste receptors as their function does not require oligomerization, which simplifies the modeling approach compared to sweet and umami taste receptors. In the associated chapter, I described our approach for reconstructing tridimensional models of TAS2Rs from their sequence. The integrative protocol combined homology modeling and single-point mutagenesis data to provide models of each receptor that accurately describe their binding pocket. This protocol was first used on TAS2R7 to identify which residues are critical for the recognition of metal ions by this specific receptor and was then improved before reconstructing the entire human TAS2R repertoire. During the comparative modeling with class A GPCRs, we identified analogous motifs to the molecular switches that govern ligand sensing and signal transduction and validated them experimentally with *in vitro* functional assays. The most demanding task in the project was the definition of an empirical score that can distinguish models that are faithful to the mutagenesis data while discarding those that are structurally deformed. The corresponding work was described in 2 publications:

- Wang, Y.; Soohoo, A. L.; Lei, W.; Christensen, C.; Margolskee, R. F.; Bouysset, C.; Golebiowski, J.; Zhao, H.; Fiorucci, S.; Jiang, P. Metal Ions Activate the Human Taste Receptor TAS2R7. *Chemical Senses* **2019**, *44,* 339–347. https://doi.org/10.1093/chemse/bjz024.
- Topin, J.; Bouysset, C.; Pacalon, J.; Kim, Y.; Rhyu, M.; Fiorucci, S.; Golebiowski, J. Functional Molecular Switches of Mammalian G Protein-Coupled Bitter-Taste Receptors. **2020**. https://doi.org/10.1101/2020.10.23.348706. *Under review.*

We subsequently started searching for agonists of one narrowly tuned and two orphan TAS2Rs by virtual screening, and while this task is still pending experimental validation, the TAS2R models that I generated will hopefully bring insights on bitter tastants recognition and receptor activation in the near future.

Aside from my main research topic, I was involved in two distinct projects, one related to the development of software for the analysis of interactions in biomolecular complexes, and another on the alteration of chemosensory perception during the Covid-19 pandemic. In the former project, I designed a Python library that can automatically detect a variety of interactions in complexes involving ligand, protein, DNA or RNA molecules obtained from molecular dynamics (MD) trajectories, docking poses, and experimental structures. These interactions are encoded as a binary fingerprint which can then be employed in a series of tasks such as machine-learning or rescoring docking results. While we showcased the usefulness of the library on class A GPCRs in the publication below, it could be applied to any receptor, including TAS2Rs:

- Bouysset, C.; Fiorucci, S. ProLIF: a library to encode molecular interactions as fingerprints. *Under revision.*

I also participated in worldwide efforts to better understand chemosensory loss (anosmia and ageusia) often occurring after a SARS-CoV-2 infection. I brought my experience in web development to collaboratively design and maintain the website of the Global Consortium for Chemosensory Research (GCCR, https://gcchemosensr.org/) where we shared several studies related to the loss of smell and taste, in more than thirty languages. The data collected from the dissemination of these questionnaires allowed us to publish 2 articles where we concluded that Covid-19 impairs not only smell but also taste and chemesthesis, and that smell loss is the best predictor of Covid-19 for people with symptoms of a respiratory illness:

- Parma, V.; Ohla, K.; Veldhuizen, M. G.; Niv, M. Y.; Kelly, C. E.; Bakke, A. J.; Cooper, K. W.; Bouysset, C.; […]; Hayes, J. E. More Than Smell—COVID-19 Is Associated With Severe Impairment of Smell, Taste, and Chemesthesis. *Chemical Senses* **2020**, *45*, 609–622. https://doi.org/10.1093/chemse/bjaa041.
- Gerkin, R. C.; Ohla, K.; Veldhuizen, M. G.; Joseph, P. V.; Kelly, C. E.; Bakke, A. J.; Steele, K. E.; Farruggia, M. C.; Pellegrino, R.; Pepino, M. Y.; Bouysset, C.; […]; Parma, V. Recent Smell Loss Is the Best Predictor of COVID-19 Among Individuals With Recent Respiratory Symptoms. *Chemical Senses* **2021**, *46*, 1–12. https://doi.org/10.1093/chemse/bjaa081.

The work presented in this thesis layouts the foundations for employing computational models to either predict properties related to chemical senses i.e., taste and smell, or to unravel the molecular mechanisms behind chemosensory receptors' activation. To further advance our knowledge on structure-function relationships, the ideal scenario would be to get insights from experimental structures, but since these are not available yet, one could start from MD simulations based on the homology models generated here. Furthermore, to expose the interaction networks that govern receptor activation, one could use the interaction fingerprint library developed here and apply it to such MD trajectories to automatically extract key interactions. Finally, future chemosensory ML models will likely benefit from having more data, although cleaner data would be more influential, and could be improved by seeking interpretability more than performance.

Appendix

Methods

Quantitative structure-activity relationships (QSAR)

QSAR tries to establish a link between molecules and an endpoint i.e., their activity or any other biological, physical, or chemical property. It relies on the assumption that the molecular structure contains features (moieties, electronic properties…*etc.*) that are related to the property of interest to be able to derive a relationship between these features and the endpoint.

Molecular structures

The first step for building a QSAR model is to collect and curate a dataset of molecules with their activity. In the simplest cases, a tabular data format is used, where each row corresponds to a molecular structure, typically a SMILES string [1], alongside an activity/property value and some additional metadata.

Once the data has been collected, each molecular representation is standardized. This step ensures that the input molecules, usually gathered from different sources, are all consistent with one another and follow a unified representation. The standardization procedure is typically comprised of several steps: stripping salts and solvents, adding or removing hydrogen atoms, neutralizing charges, and forcing a canonical representation for aromatic rings or tautomers. The order in which those steps are performed is important since each action relies on the output of the previous one, and changing the order might change the final output structure. This step might also reveal some unreadable structures which must be corrected manually. Open-source tools [2] and commercial software solutions (ChemAxon Standardizer) exist to simplify and automatize the standardization procedure.

Once this is done, the next step is to curate the database. This includes detecting duplicates and defining how to appropriately handle them. This action is facilitated by the previous step and can be performed by simply searching for text duplications in the InChiKey obtained from the standardized structure. In the case of a textual endpoint property (e.g., an odor description) some additional corrections might be applicable, such as manipulating the string to be all

lowercase. The exact workflow applied for curating the database is in most cases specific to the problem being solved.

Molecular features

Once the dataset has been cleaned, the next step consists in calculating molecular features which will accurately and numerically describe each compound. The aim of molecular features is to define the chemical space that is relevant for a given problem i.e., to find appropriate representations of the molecules with regard to the property being modeled. Two types of features can be used (and are sometimes combined): fingerprints and molecular descriptors.

Molecular fingerprints typically decompose each chemical structure in a set of moieties and accounts for the presence (or sometimes the count) of each specific moiety, resulting in a bitstring representation of the molecule. Those moieties can be predefined (e.g., the MACCS keys fingerprint) or extracted automatically from each structure (e.g., circular topological fingerprints like ECFP [3]). The fingerprint is typically folded to fit in a fixed number of bits to reduce memory usage and computational cost in the framework of a similarity search, although it introduces bit collisions which hampers interpretability and adds noise. In the use case of QSAR models, it is best not to use folded fingerprints [4].

Another type of features that are often used in QSAR are molecular descriptors. As defined by Todeschini and Consonni, “the molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.” [5]. Descriptors can be constitutional (atom counts, molecular weight…*etc.*) or topological (based on the molecular graph) among others, but they can also depend on the conformation of a molecule (orbital energies, radius of gyration…*etc.*). This last point raises the open question on how to generate relevant 3D conformations for each molecule, as the biologically active conformation might be very different from the most stable conformation in the gas phase.

Regardless of the kind of molecular features used, it is quite common that some of the features are redundant or unhelpful, and keeping them would only hinder model training. Thus, an additional curation step can be performed to limit the number of molecular features available by discarding the heavily correlated or constant descriptors, as well as excluding the ones that could not be calculated for a given compound.

Machine learning (ML)

The final stage of QSAR modeling is to train and validate a machine-learning model. Firstly, the dataset of molecular features and their corresponding activity must be split into a training set and an external validation set. Several strategies exist to perform this split in a rational way such as the sphere-exclusion algorithm, but a random split can also lead to acceptable results [6].

Because several ML algorithms depend on distance calculations between data points during model training, it can help to scale features in the dataset to make sure that the algorithm is not biased towards features with highly variable ranges. Two feature scaling techniques are often used: standardization and normalization. In standardization, the features in the training data are scaled so that they exhibit the same properties as a standard normal distribution (average of 0 and standard deviation of 1), according to the following equation:

$$x' = \frac{x - \mu}{\sigma}$$

where x is the input feature, and μ and σ are the mean and standard deviation calculated on the training set for a given feature, respectively. In normalization (or min-max scaling), the features are scaled between 0 and 1 as follow:

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

where Xmin and Xmax are the minimum and maximum values in the training set for a given feature, respectively. There is no rule-of-thumb to determine *a priori* which scaling procedure should be used. Once feature scaling is settled, the same transformations that were used on the training set must be applied on the external test set. Additionally, scaling must be performed after splitting the dataset in training and test sets, as otherwise it could leak information from the training data into the test data, thus skewing the final evaluation of the model.

Subsequently, another splitting is used to optimize hyperparameters for each trained ML algorithm. Since hyperparameter tuning is the step that is most prone to overfitting, several splits are usually investigated in parallel to ensure that the chosen parameters are stable across different repartitions of the training set compounds. It can be achieved in a few different ways, but the most common is k-fold cross-validation (CV). This method randomly subdivides the dataset in k subsets of equal size, and each subset is used once as an internal validation set while the others are used for training the model. Several derived forms exist, such as leave-one-out where k corresponds to the number of compounds in the dataset, or stratified k-fold CV which

ensures an equivalent distribution of the endpoint property between each partition. Nested CV, which applies the principle of CV to both external and internal splits, is also often used in order to limit the potential bias resulting from the initial splitting strategy and measure the ability of the best performing model to generalize to unseen samples.

Inside the internal CV loop, the hyperparameters are usually fine-tuned using a brute-force grid search i.e., all possible combinations of the ML algorithm's parameters are investigated one after the other, and the optimal algorithm and parameters are selected according to the performance on the internal test set. The final model is then trained on the complete training set and its performance on the external test set are exposed.

This last point about performance raises the question of which metric is used to optimize the model. Selecting this metric, named loss function, obviously depends on the type of task (regression or classification) but also on the aspects of the model that should be strengthened. For example, in a classification task one might want to penalize false positives more than false negatives, but in another task, both should be balanced. In the former case, an appropriate loss function would be based on the true negative rate (TNR, also called specificity), while in the latter case, balanced accuracy (BA) might be more suitable:

$$TNR = \frac{TN}{TN+TP} \qquad BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Additionally, a variety of metrics exist to properly report different aspects of the final model and expose its strengths and possible shortcomings. For regression tasks, Golbraikh and Tropsha have proposed a list of criteria to assess the predictive ability of QSAR models [7]. For classification, since class imbalance is not rare and can make typical metrics like precision or accuracy severely misleading, metrics like the F1 score or the Matthews correlation coefficient are usually preferred.

Finally, the model should be accompanied with a definition of its applicability domain to inform end-users about the confidence in a given prediction [8]. Ideally, authors of QSAR models should also be concerned about making their modeling protocol FAIR (findable, accessible, interoperable, reusable) to maximize the beneficial impact for the science community, and Artrith *et al.* [9] have recently shared a set of practical guidelines for this matter.

Homology Modeling

Homology modeling uses existing experimental structures of proteins to rebuild a tridimensional model of a similar yet unresolved protein.

Template search

The first step in comparative modeling is to search for suitable templates that exhibit sufficient sequence similarity with the target protein. One can use BLAST [10] to perform such query on sequences of proteins available on the Protein Data Bank (PDB), which will search for identical sequence fragments between the target and putative template, and later use this result to build a pairwise alignments between both. Briefly, a pairwise alignment tries to maximize the matching between identical or similar residues in both sequences and introduces gaps when the substitution score, as obtained from a substitution matrix (such as BLOSUM62 [11]), is not favorable.

Alternatively, in chapter 3 we decided to directly take the templates used by an automatic model building webserver, SWISS-MODEL [12], when building models for the human bitter-proteome.

The final templates are usually selected based on sequence identity and X-ray resolution but depending on the purpose of the model other criteria can also influence the selection, such as the presence of a ligand in the binding pocket or the activation state of the protein.

Multiple sequence alignment (MSA)

Once template structures have been selected, the next step is to align the complete set of sequences (both targets and templates) together. This step is most of the time performed automatically using a heuristic method, such as Clustal Omega [13] which we used in chapter 3, because finding the optimal solution to an MSA can be too computationally demanding for more than a few sequences.

In ClustalO, the algorithm starts by generating a distance matrix for a subset of sequences using a modified mBed method [14]. Briefly, mBed identifies a small number of reference sequences and computes their distance to the complete set of sequences to create vector embeddings of each sequence. These distances are calculated using the k-tuple method [15]. This step is followed by clustering using k-means based on these sequence embeddings. Next, within each cluster the full distance matrix is approximated from the sequence embeddings with mBed and

a dendrogram is calculated with the UPGMA hierarchical clustering method as implemented in MUSCLE [16]. The dendrograms are then joined to generate a guide tree based on the barycenter of each subcluster. The combination of mBed with clustering is what allows ClustalO to be computationally efficient even with a large number of sequences, as the complete pairwise distance matrix that is usually required to obtain the guide tree is actually never computed.

The final step implies the computation of the MSA using HHAlign [17], which relies on using Hidden Markov Model (HMM) profiles to construct a progressive alignment based on the guide tree.

While such MSA algorithm typically offers a good starting point, the alignment obtained is rarely the optimal solution thus manual refinement is often required and can be aided by taking advantage of available mutagenesis and structural data. For instance, when applying homology modeling to G protein-coupled receptors, once the transmembrane (TM) domains of the target receptor has been identified, the MSA should be corrected to minimize the presence of gaps in the TMs as it often leads to misshaped α-helices.

Model generation

With the target-template alignment done, the models can be generated by comparative modeling. In chapter 3, we used the Modeller [18] software which models proteins by satisfying spatial restraints, but other methods (rigid-body assembly, segment matching) exist [19].

The first step is to derive the restraints from the alignment based on the assumption that the distance between aligned residues should be similar in both target and template. From an analysis of families of proteins with resolved structures, tables of correlations between a variety of spatial characteristics (α-carbon distances, dihedral angles…*etc.*) were obtained. These tables are used to create the homology-derived restraints from the input target-template alignment and expressed as probability density functions. Next, these restraints are combined with the CHARMM22 force-field terms [20] (which constrains bond lengths, dihedrals and non-bonded interactions) to formulate an objective function. The final step is to minimize this objective function using a combination of conjugate gradients and simulated-annealing molecular dynamics, which allows for an efficient sampling of the conformational space. By slightly varying the initial coordinates of the structure, Modeller can produce an ensemble of models for the target protein.

Model selection

Once models are built, we can finally prioritize one or more tridimensional structures. The most common approach for models built with Modeller is to use the DOPE score [21] which is directly available in the software. DOPE is a statistical potential derived from crystallographic structures based on inter-atomic distances. However, it was trained exclusively on globular proteins, which limits its practical use when working with transmembrane receptors. For this reason, we decided to create our own scoring function to select models in chapter 3.

To ensure that the models have reasonable geometry, several structure validation solutions such as MolProbity [22] exist. These methods typically check for Ramachandran outliers, atomic clashes and bad rotamers. Visual inspection can also be used to select the final model based on target-specific expert knowledge.

Finally, if several models are needed, one can combine the above-mentioned methods with clustering based on residues of interest to prioritize a representative ensemble of structures. This could be used to select a subset of models with variable binding-pocket residues' sidechain orientation prior to docking, which can serve as a good compromise (in terms of computing time) between rigid docking on a single structure and flexible docking.

References

1. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model 28:31–36. https://doi.org/10.1021/ci00057a005

2. Bento AP, Hersey A, Félix E, et al (2020) An open source chemical structure curation pipeline using RDKit. J Cheminform 12:51. https://doi.org/10.1186/s13321-020-00456-1

3. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t

4. Gütlein M, Kramer S (2016) Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. J Cheminform 8:60. https://doi.org/10.1186/s13321-016-0173-z

5. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors, 1st ed. Wiley

6. Martin TM, Harten P, Young DM, et al (2012) Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? J Chem Inf Model 52:2570–2578. https://doi.org/10.1021/ci300338w

7. Golbraikh A, Tropsha A (2002) Beware of q^2! Journal of Molecular Graphics and Modelling 20:269–276. https://doi.org/10.1016/S1093-3263(01)00123-1

8. Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. SAR and QSAR in environmental research 27:893–909. https://doi.org/10.1080/1062936X.2016.1250229

9. Artrith N, Butler KT, Coudert F-X, et al (2021) Best practices in machine learning for chemistry. Nat Chem 13:505–508. https://doi.org/10.1038/s41557-021-00716-z

10. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

11. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89:10915–10919. https://doi.org/10.1073/pnas.89.22.10915

12. Schwede T (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31:3381–3385. https://doi.org/10.1093/nar/gkg520

13. Sievers F, Wilm A, Dineen D, et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75

14. Blackshields G, Sievers F, Shi W, et al (2010) Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol Biol 5:21. https://doi.org/10.1186/1748-7188-5-21

15. Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proceedings of the National Academy of Sciences 80:726–730. https://doi.org/10.1073/pnas.80.3.726

16. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113

17. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125

18. Šali A, Blundell TL (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 234:779–815. https://doi.org/10.1006/jmbi.1993.1626

19. Webb B, Sali A (2014) Comparative protein structure modeling using MODELLER

20. MacKerell AD, Bashford D, Bellott M, et al (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. J Phys Chem B 102:3586–3616. https://doi.org/10.1021/jp973084f

21. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524. https://doi.org/10.1110/ps.062416606

22. Chen VB, Arendall WB, Headd JJ, et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21. https://doi.org/10.1107/S0907444909042073

Other publications

The following section includes two publications that are relevant to this thesis:

- Publication A1, in which I participated but not as a first author
- Publication A2, which is not directly related to the main research topic of this thesis

Publication A1

Metal ions activate the human taste receptor TAS2R7

Yi Wang†, Amanda L. Zajac†, Weiwei Lei, Carol M. Christensen, Robert F. Margolskee, Cédric Bouysset, Jérôme Golebiowski, Huabin Zhao, Sébastien Fiorucci, & Peihua Jiang*

Chemical Senses 2019, *44*, 339–347.

doi.org/10.1093/chemse/bjz024

Chemical Senses, 2019, Vol 44, 339–347
doi:10.1093/chemse/bjz024
Original Article
Advance Access publication 23 April 2019

OXFORD

Original Article

Metal Ions Activate the Human Taste Receptor TAS2R7

Yi Wang[1,2,*], Amanda L. Zajac[1,*], Weiwei Lei[1], Carol M. Christensen[1], Robert F. Margolskee[1], Cédric Bouysset[3], Jérôme Golebiowski[3,4], Huabin Zhao[2,©], Sébastien Fiorucci[3,©] and Peihua Jiang[1]

Abstract

Divalent and trivalent salts exhibit a complex taste profile. They are perceived as being astringent/drying, sour, bitter, and metallic. We hypothesized that human bitter taste receptors may mediate some taste attributes of these salts. Using a cell-based functional assay, we found that TAS2R7 responds to a broad range of divalent and trivalent salts, including zinc, calcium, magnesium, copper, manganese, and aluminum, but not to potassium, suggesting TAS2R7 may act as a metal cation receptor mediating bitterness of divalent and trivalent salts. Molecular modeling and mutagenesis analysis identified 2 residues, $H94^{3.37}$ and $E264^{7.32}$, in TAS2R7 that appear to be responsible for the interaction of TAS2R7 with metallic ions. Taste receptors are found in both oral and extraoral tissues. The responsiveness of TAS2R7 to various mineral salts suggests it may act as a broad sensor, similar to the calcium-sensing receptor, for biologically relevant metal cations in both oral and extraoral tissues.

Keywords

TAS2R7, metal ions, bitter taste, metallic taste

Introduction

Divalent salts evoke a complex taste profile, described as metallic, bitter, and astringent (Lim and Lawless 2005). Despite recent progress in the identification of the taste receptor repertoire for sweet and bitter compounds, the molecular mechanisms underlying the complex sensory attributes of divalent salts are largely unknown (Bachmanov and Beauchamp 2007). Using rodent models, Riera *et al.* (Riera *et al.* 2009) showed that sensory attributes of complex-tasting divalent salts are mediated at least partially by transient receptor potential cation channel subfamily M member 5 (Trpm5) and transient receptor potential vanilloid-1 (Trpv1) channels. Direct activation of Trpv1 by divalent ions may explain the astringency sensation of divalent ions (Riera *et al.* 2009). Trpm5 is a shared signaling element for sweet, umami, and bitter taste transduction (Perez *et al.* 2002; Zhang *et al.* 2003). The involvement of Trpm5 for the taste of divalent salts indicates it may be mediated in part by transduction mechanisms similar to that for sweet, bitter, and umami tastes. Interestingly, the sweet and umami receptor subunit T1R3 is reported to be involved in the taste of calcium and magnesium (Tordoff *et al.* 2008). However, calcium- and magnesium-containing salts are primarily perceived as bitter tasting

(Lim and Lawless 2005; Yang and Lawless 2005). Yet how bitterness of these metallic ions is detected is unclear.

Bitter taste is mediated by type 2 taste receptors (TAS2Rs) that are expressed in a subset of taste bud cells (Chandrashekar *et al.* 2000; Matsunami *et al.* 2000). TAS2Rs are G protein-coupled receptors (GPCRs) within the rhodopsin family (Chandrashekar *et al.* 2000; Matsunami *et al.* 2000). Humans possess 25 functional TAS2Rs. However, the numbers of TAS2R genes vary greatly among mammalian species, ranging from 0 to 54 in amphibian, presumably correlating with the specific ecological niche of a species (Feng *et al.* 2014; Go *et al.* 2005; Jiang *et al.* 2012; Jiao *et al.* 2018; Liman 2006; Shi and Zhang 2006; Wang and Zhao 2015). Most human TAS2Rs have been deorphanized, and their receptive ranges are heterogeneous (Meyerhof *et al.* 2010). Some receptors such as TAS2R14 and TAS2R10 are broadly tuned, responding to a wide range of structurally diverse bitter compounds, whereas some others such as TAS2R38 and TAS2R16 are more specialized, responding to relatively few compounds with specific chemical motifs (Bufe *et al.* 2005; Kim *et al.* 2003; Meyerhof *et al.* 2010). This combinatorial TAS2R coding scheme may explain why a relatively limited number of receptors can detect a broad range of structurally diverse bitter compounds.

Given the bitter-taste attribute of multiple divalent salts, we hypothesized that divalent salts may activate one or more TAS2Rs, therefore producing a bitter sensation, contributing to the complex taste attributes of metal ions. To test this hypothesis, we examined which bitter receptor(s) are responsive to divalent salts and found that TAS2R7 responded to all divalent salts tested. In addition, TAS2R7 responded to trivalent salts such as aluminum sulfate. In contrast, potassium chloride, a monovalent salt, does not activate TAS2R7, indicating its specificity. Further structural and functional analyses and molecular modeling revealed H94 and E264 of TAS2R7 as two key residues for the receptor's interaction with metallic ions.

Materials and Methods

Preparation of human TAS2R constructs and site-directed mutants

The coding sequences of human TAS2Rs were amplified from human genomic DNA, then subcloned into pcDNA3.1(+) vector, with the herpes simplex virus glycoprotein D epitope (HSV) at the C-terminal and a signal peptide consisting of the first 45 amino acid residues of the rat somatostatin receptor 3 at the N-terminal, essentially as described previously (Bufe *et al.* 2002). Point mutations in human TAS2R7 (NCBI Reference Sequence: NP_076408.1) were

constructed by site-directed mutagenesis. All the constructs were confirmed by Sanger sequencing.

Chemicals

All tested compounds were purchased from Sigma-Aldrich, with the exception of diphenidol hydrochloride (Reagent World) and L-praziquantel (manufactured by Shaoxing Pharmaceutical Co. Ltd.). All the metal ions were dissolved in the assay solution (130 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, and 10 mM glucose; pH 7.4) unless specified otherwise, and the bitter compounds (diphenidol, quinine and chlorphenamine) were dissolved first in DMSO as stock solution and then diluted with the assay solution; the final DMSO concentration was below 0.5%, with the exception of cromolyn, which is dissolved in the assay solution directly. For our initial screening (Table 1), Hanks' balanced salt solution (HBSS, ThermoFisher, Cat#: 14025134) supplemented with 10 mM HEPES were used as the assay buffer. Because HEPES and other buffering agents partially precipitated certain metal ions, the assay solution without buffering agents as described above was used for further characterization of TAS2R7.

Functional assays of human TAS2Rs

Human embryonic kidney 293 (PEAKrapid, ATCC # CRL-2828) cells were cultured in Opti-MEM medium with 4% fetal bovine serum. One day before transfection, cells were seeded on a 96-well plate at a density of 25,000 per well. Cells were then transiently transfected with a TAS2R construct (0.1 μg/well) along with a G protein Gα16-gust44 (0.1 μg/well) construct by Lipofectamine 2000 (0.5 μl/well). For controls, only Gα16-gust44 was used (mock transfection). Twenty-four hours after transfection, cells were washed with HBSS including 10 mM HEPES and loaded with Fluo-4 in the dark for 1 hour. After incubation, cells were washed two times with HBSS (including 10 mM HEPES), incubated in the dark for another 30 minutes, and then washed with assay solution once more before running the assay using a FlexStation III reader. Relative fluorescence units (excitation at 494 nm, emission at 516 nm, and auto cutoff at 515 nm) were read every 2 seconds for 2 minutes. Calcium mobilization traces were recorded.

Immunostaining

Cells were seeded onto poly-lysine coated coverslips in 24-well plates and transfected with a wild-type or mutant TAS2R7 receptor construct (0.25 μg/well), along with Gα16-gust44 (0.25 μg/well) by lipofectamine (2.5 μl/well). 24hr post transfection, cells were fixed with 4%

paraformaldehyde in phosphate buffered saline (PBS) for 30 min. Cells were then washed with 3 exchanges of PBS, and incubated with blocking buffer (2% donkey serum, 0.3% Triton x-100, in SuperBlock (PBS) buffer (ThermoFisher, Cat #37515)) for 1 hr at RT. An anti-HSV antibody (Millipore, Cat# MAC123, 1:1000) was applied overnight. An Alexa Fluor 488-labled Donkey anti-mouse secondary antibody (Abcam, Cat#: ab150105, 1:1000) were used for fluorescence visualization.

Data analysis

Calcium mobilization traces were raw data obtained from single wells. Changes in fluorescence (ΔF) were calculated as the peak fluorescence minus baseline fluorescence (Lei *et al.* 2015). The calcium mobilization was quantified as the percentage of change (ΔF) relative to baseline (F). Each data point for bar graphs and dose-dependent responses was averaged from triplicates (mean ± SD). Calcium mobilization traces and bar graphs along with dose-dependent plots were all generated by GraphPad Prism 7. Analysis of variance with Dunnett's multiple comparisons test were used for statistical analysis. * indicates $P < 0.05$.

Molecular modeling of TAS2R7

The 3D structure of the human TAS2R7 was obtained by comparative modeling using Modeller 9.19 (Sali and Blundell 1993) based on the crystal structure of the 5-HT_{2C} serotonin receptor, PDB identifier 6BQG (Peng *et al.* 2018) (**Fig. S1**). The best homology model according to the DOPE score has been energy minimized using AMBER (Case *et al.* 2005) and the AMBER ff14SB force field (Maier *et al.* 2015) parameter prior to structural validation with PROCHECK. Electrostatic potential was calculated with the APBS program (Baker *et al.* 2001). To obtain accurate electrostatic properties, we used the two-step focusing technique and a grid spacing lower than 0.5 Å in each space dimension. The molecular surface was generated using a water probe with a radius of 1.4 Å. The dielectric constant of the protein and the solvent was fixed to 2 and 80, respectively. The protonation states of titratable residues were predicted at pH 6.5 through the H++ server (Gordon *et al.* 2005). Cromolyn was docked within the TAS2R7 binding cavity using Autodock Vina (Trott and Olson 2010). The Zn^{2+} cation was manually docked into the TAS2R7 model. The cation-receptor complex was energy minimized with the AMBER software using 500 steps of the steepest descent optimization followed by 1,000 steps of conjugate gradient optimization with positional restraints of 50 $kcal \cdot mol^{-1} \cdot Å^{-2}$ on backbone heavy atoms.

Results

Identification of TAS2Rs for metal ions

To determine whether a TAS2R responds to metal ions, we expressed all 25 human bitter receptors individually in HEK293 cells (PEAKrapid) by transient transfection of a TAS2R along with a coupling chimeric G protein, Gα16-gust44. All TAS2Rs were cloned from human genomic DNA. Activation of human TAS2Rs was monitored by the calcium mobilization assay (Lei *et al.* 2015). We tested these receptors individually for their responses toward metal ions: $ZnSO_4$ (20 mM), $CuSO_4$ (20 mM), and $MgCl_2$ (20 mM) (**Table 1, Fig. 1A**). No receptors showed responsiveness to these metal ions, with the exception of TAS2R7, which consistently showed robust responses toward all three divalent salts. To determine the breadth of tuning of TAS2R7 toward metal ions, we also tested $MnCl_2$ (20 mM), $Al_2(SO_4)_3$ (20 mM), and $CaCl_2$ (20 mM) (**Fig. 1B**). All divalent and trivalent ions activated the receptor, albeit with variable degrees of efficacy (**Fig. 1A, B**). $ZnSO_4$ solution is acidic (pH ~5) at the concentration we tested, as is $Al_2(SO_4)_3$ solution (pH ~3). To determine if pH affected the activity of TAS2R7, we tested the responsiveness to TAS2R7 to 1 mM citric acid (pH ~3) (**Fig. 1C**). No specific response was detected. Therefore, the responses of TAS2R7 toward metal ions were specific. In contrast to divalent and trivalent cations, the monovalent salt KCl did not activate the receptor, suggesting that TAS2R7 is specifically tuned to divalent and trivalent salts (**Fig. 1C, D)**. To determine whether anions might affect the potency and efficacy of cations, we compared the responses of TAS2R7 toward $ZnSO_4$ and $ZnCl_2$. No obvious differences were found between two types of anions (EC_{50} of $ZnSO_4$: 3.21 mM, $ZnCl_2$: 3.42 mM) (**Fig. S2**).To our knowledge, aside from CaSR, TAS2R7 is the only GPCR that can be activated by multiple metal ions (Brown *et al.* 1993; McGehee *et al.* 1997; Saidak *et al.* 2009).

Table 1. Responses of all 25 human TAS2Rs to metallic ions.

		TAS2R																								
Substance	(mM)	1	3	4	5	7	8	9	10	13	14	16	19	20	30	31	38	39	40	41	42	43	45	46	50	60
$ZnSO_4$	20	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
$CuSO_4$	20	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
$MgCl_2$	20	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

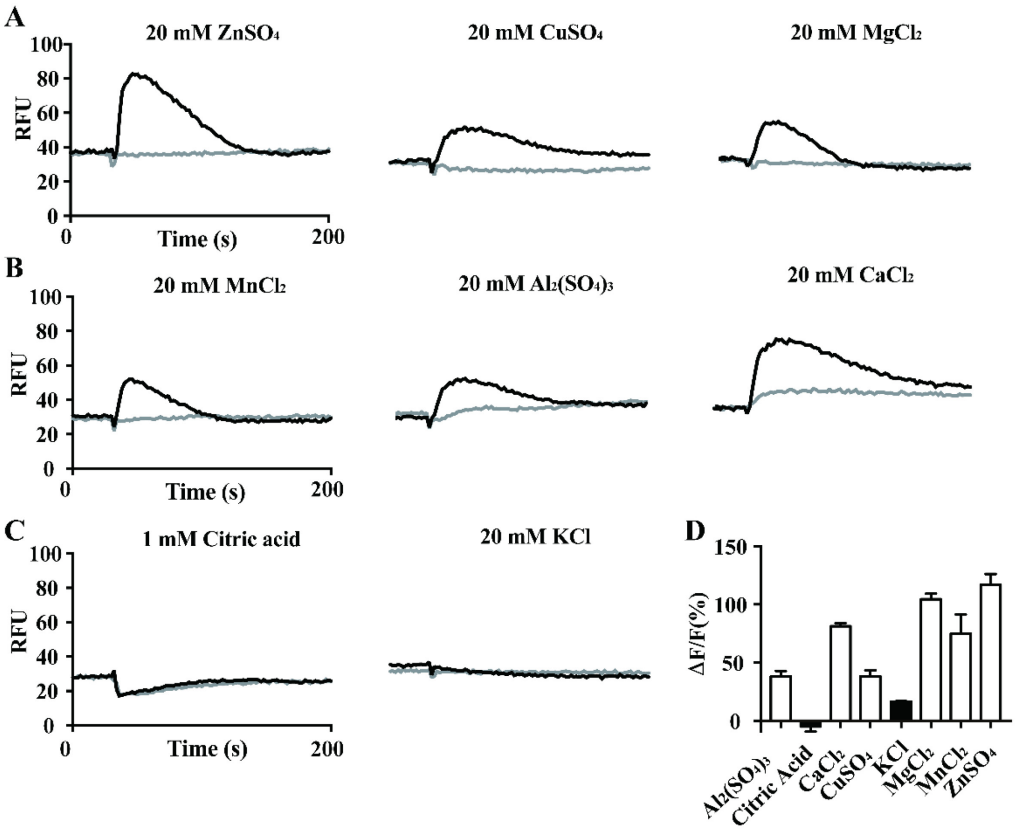


Figure 1. Metal ions activate TAS2R7. (A–C) HEK293 cells transfected with human TAS2R7 with Gα16-gust44 were assayed for their responses to metal ions and citric acid. Black traces, representative calcium mobilization traces of TAS2R7 to compounds; gray traces, mock-transfected cells used as control. RFU, relative fluorescence unit. (D) Quantitative analysis of responses of TAS2R7 to metallic ions and citric acid. Data are percentage change (mean ± SD) in fluorescence (peak RFU – baseline RFU, denoted ΔF) from baseline fluorescence (denoted F) averaged from triplicates. Experiments were replicated three times.

TAS2R7 responds to metal ions in a dose-dependent manner

To determine the sensitivity of TAS2R7 toward metal ions, we generated concentration-response functions (**Fig. 2**). TAS2R7 responded to all metal ions we tested in a dose-dependent manner (Fig. 2A), while mock-transfected cells showed no responses to metal ions at any concentration we tested (Fig. 2B). Nevertheless, the efficacy differs among different cations. The receptor appears to be most sensitive toward aluminum sulfate (EC_{50}, 39±15 μM), followed by $CuSO_4$ (EC_{50},1.04±0.36 mM) , $ZnSO_4$ (EC_{50}, 33.36±0.14 mM), $MgCl_2$ (EC_{50}, 6.07±1.07mM), $CaCl_2$ (EC_{50}, 5.27±0.50 mM), and $MnCl_2$ (EC_{50}, 6.59±1.73 mM). Mock-transfected cells showed no responses to any concentration of $Al_2(SO_4)_3$ tested.

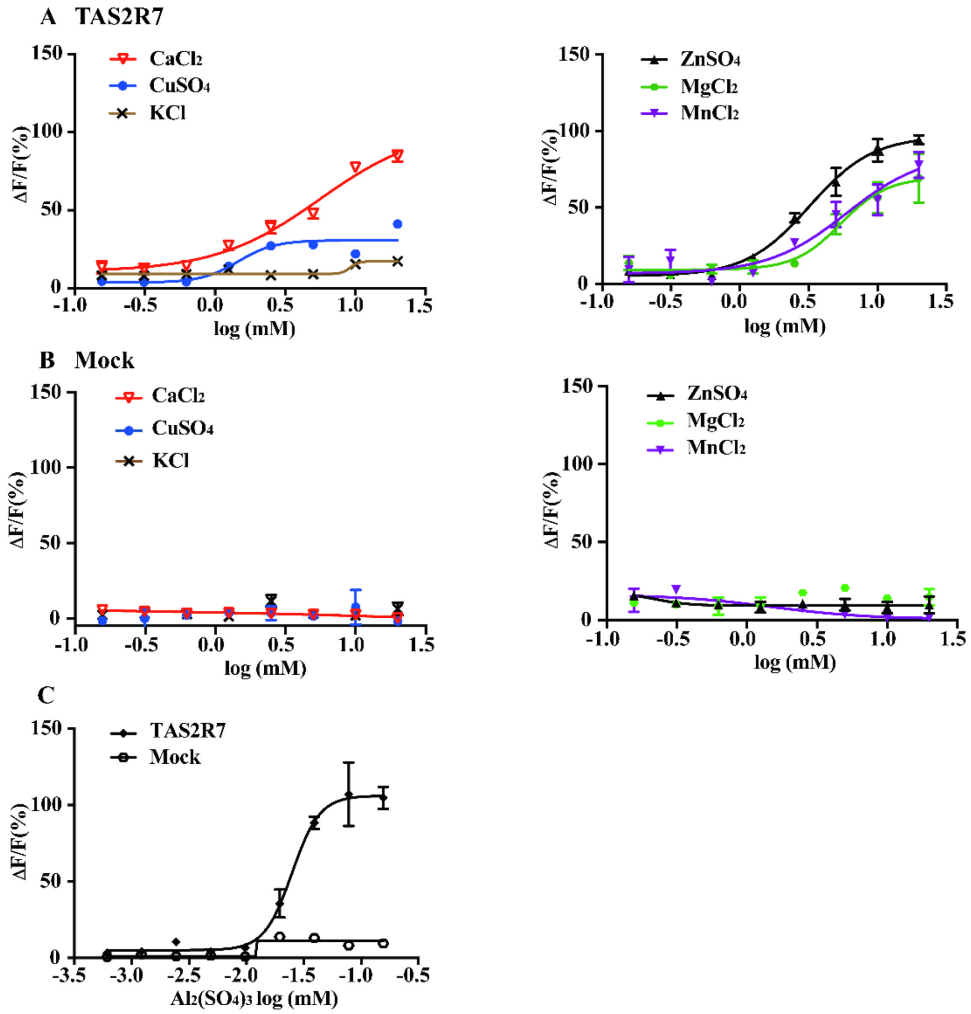


Figure 2. TAS2R7 responds to metal ions dose dependently. HEK293 cells transiently transfected with human TAS2R7 with Gα16-gust44 showed dose-dependent responses to metal ions: $CaCl_2$, $CuSO_4$, $ZnSO_4$, $MgCl_2$, $MnCl_2$, and $Al_2(SO_4)_3$ (A, C). KCl does not activate TAS2R7 at any concentrations tested (A, left panel). Mock-transfected cells (Gα16-gust44 only, Mock) were used as controls for cell transfected with TAS2R7 in response to metal ions (B, C). GraphPad Prism 7 was used to fit the curve (sigmoidal). Experiments were replicated three times.

As expected, the receptor was also not responsive to any concentration of KCl. Thus, TAS2R7 interacts differently with different ions.

Our assay solution contains 2 mM calcium ion, which supports optimal assay condition for the calcium mobilization assay, yet TAS2R7 responds to calcium. Therefore, to determine if the presence of calcium affects the responses of TAS2R7 to metal ions, we performed calcium mobilization assays using assay solution containing no calcium (130 mM NaCl, 5 mM KCl, and 10 mM glucose; pH 7.4). All the tested compounds were dissolved in the same assay

solution. As expected, TAS2R7 showed robust responses to all six metal ions tested under this condition (**Fig. 3A**). Concentration-dependent curves were similar in the presence and absence of calcium in the assay solution.

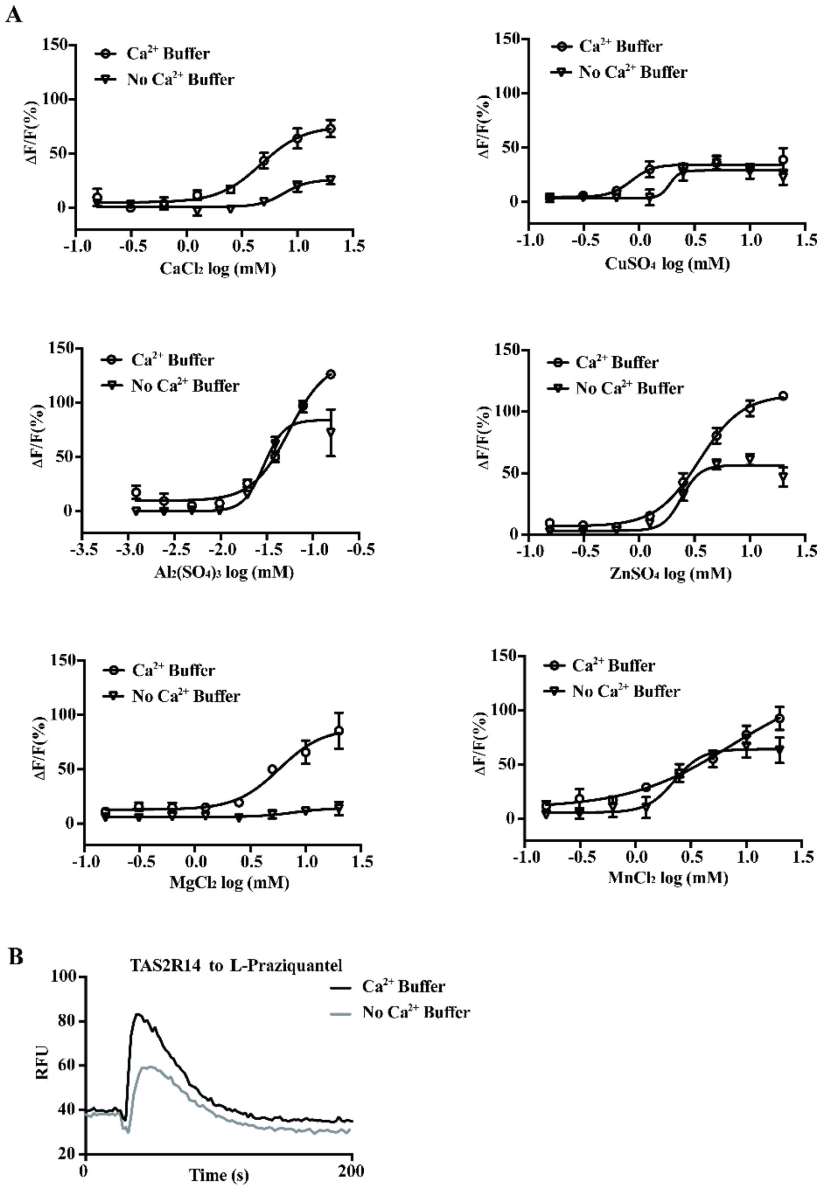


Figure 3. Responses of TAS2R7 to metal ions in the absence and presence of calcium in the assay solution. (A) Responses of HEK293 cells transiently transfected with human TAS2R7 with Gα16-gust44 to six metal ions in the presence and absence of calcium in the assay solution, including $CaCl_2$, $CuSO_4$, $ZnSO_4$, $MgCl_2$, $MnCl_2$, and $Al_2(SO_4)_3$ respectively . GraphPad Prism 7 was used to draw the dose-dependent curves. (B) TAS2R14 was expressed along with Gα16-gust44 in the HEK293 cells, and the responses to 0.5 mM L-praziquantel were assayed with the presence and absence of calcium. Black traces, calcium mobilization with the presence of calcium; gray traces, with the absence of calcium. Experiments were replicated twice.

The EC_{50} of six metal ions in the absence of calcium is as follows, $CaCl_2$, 4.70 mM; $CuSO_4$, 0.85 mM; $ZnSO_4$, 3.49 mM; $MgCl_2$, 5.78 mM; $MnCl_2$, 7.19 mM; $Al_2(SO_4)_3$, 55 μM , respectively, similar to the EC_{50}s in the presence of calcium ($CaCl_2$, 7.56 mM; $CuSO_4$, 1.89 mM; $ZnSO_4$, 2.41 mM; $MgCl_2$, 7.84 mM; $MnCl_2$, 2.24 mM; $Al_2(SO_4)_3$, 29 μM). However, the maximal responses to all metal ions were smaller in the absence than in the presence of calcium, especially towards $MgCl_2$. This appears to be a general phenomenon for this type of assay, as shown by reduced response amplitude for other GPCRs as well (e.g., TAS2R14 to L-praziquantel, **Fig. 3B**). All the dose-dependent curves were replicated at least twice.

TAS2R7 is a narrowly tuned receptor

TAS2R7 has been reported to respond to certain bitter compounds, including diphenidol, quinine, cromolyn, and chlorphenamine (Meyerhof *et al.* 2010). To further determine the tuning properties of TAS2R7, we examined its responsiveness to bitter compounds that were previously shown to activate the receptor (**Fig. 4A**) (Meyerhof *et al.* 2010). At the concentrations reported previously, none of the compounds we tested (diphenidol, quinine, cromolyn, and chlorphenamine) triggered detectable responses in cells transiently transfected with TAS2R7 in our hands (Meyerhof *et al.* 2010). However, cromolyn at a higher dose (10 mM) did elicit a robust response in cells specifically transfected with TAS2R7 but not in mock-transfected cells. We further confirmed the requirement of high doses of cromolyn to activate the receptor by dose-response analysis (EC_{50}, 5.9 mM) (**Fig. 4B**). For other compounds, even higher doses produced no responses (**Fig. 4A**). Thus, our data indicate that TAS2R7 selectively responds to metal ions and cromolyn. We also performed cell-based assay with the presence and absence of calcium for cromolyn (**Fig. 4C**). As expected, the maximal response is smaller using the assay solution containing no calcium than the assay solution containing calcium, while the EC_{50}s are comparable (with calcium: 6.67 mM; without calcium: 5.22 mM). Therefore, we used assay solution containing calcium for our further analysis of the receptor to have a better readout.

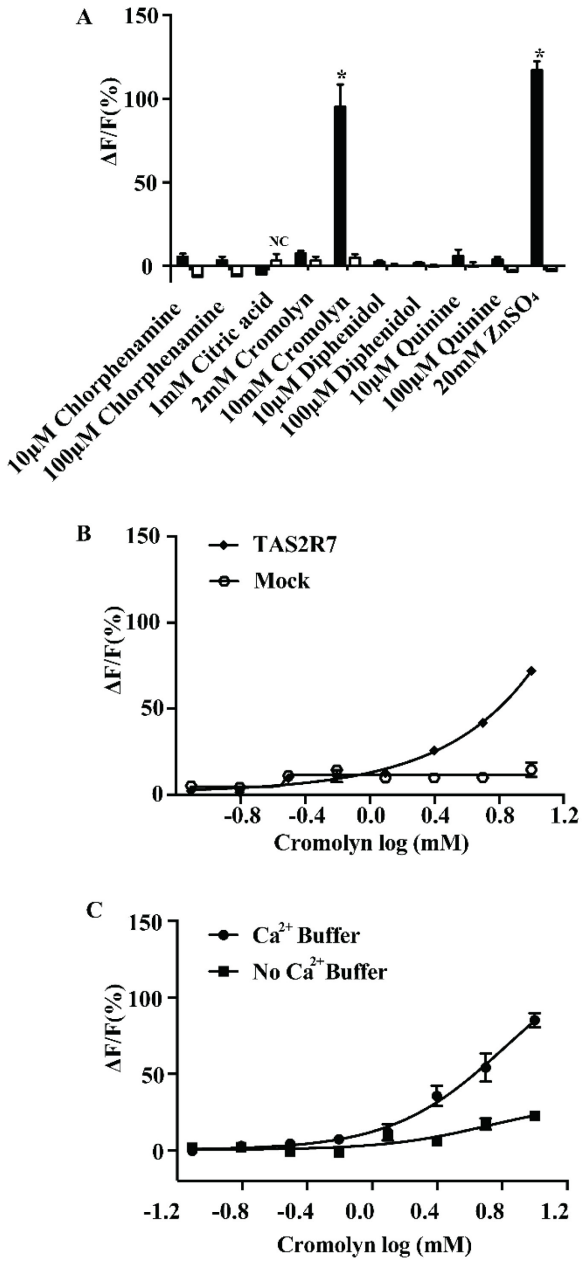


Figure 4. TAS2R7 is a narrowly tuned receptor. HEK293 cells were transiently transfected with human TAS2R7, coupled with Gα16-gust44, and their responses assayed to previously reported bitter ligands. Two-tailed t-tests were used to determine whether there is a significantly difference between the TAS2R7-transfected cells and mock-transfected (Gα16-gust44 only) cells. (A) Responses to $ZnSO_4$ and citric acid were chosen as positive control and negative control (NC), respectively. Bitter compounds that stimulate significant responses are indicated with an asterisk (*). ($p < 0.05$) (B) Cromolyn activates TAS2R7 in a dose-dependent manner. Experiments were replicated three times. (C) Dose-dependent curves of TAS2R7 toward cromolyn with the presence and absence of calcium. Experiments were replicated twice.

Molecular modeling and site-directed mutagenesis identify two residues of TAS2R7 critical for the recognition of metal ions

To predict how TAS2R7 interacts with metal ions, a homology model of TAS2R7 was built based on the crystal structure of the 5-HT_{2C} serotonin receptor (Peng *et al.* 2018). We first automatically docked cromolyn into the GPCR binding cavity formed by helices 2, 3, 5, 6, and 7, because metal ions are too small for initial docking simulations (**Fig. 5C**). The results of docking simulations identified a pocket similar to that defined by Liu *et al.* (Liu *et al.* 2018). All amino acids involved in contact with the ligand are part of the typical TAS2R binding pocket (**Fig. S3**).

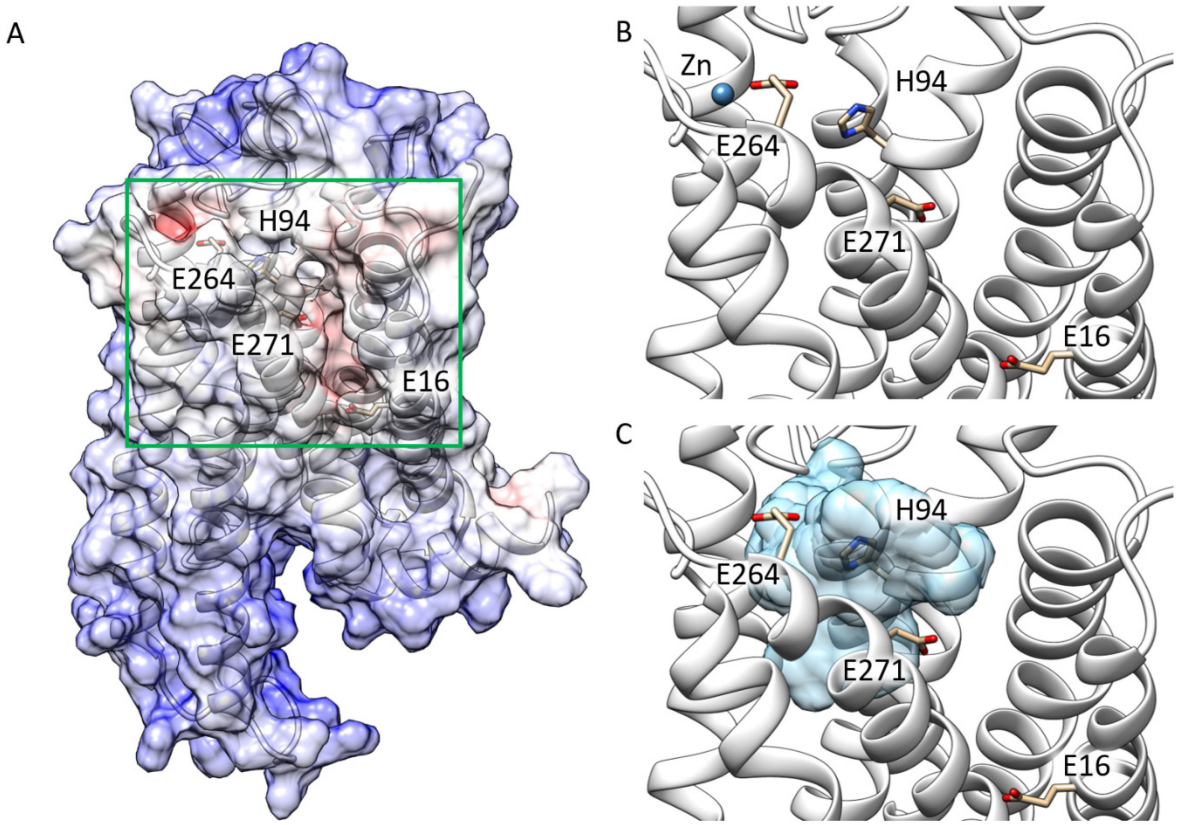


Figure 5. Molecular model of TAS2R7's binding pocket with docked ligands. (A) Electrostatic potential (±10 kbT/e) mapped onto the molecular surface of the protein. Red and blue colors indicate negatively and positively charged regions, respectively. The most attractive cavity for cation binding is delimited by the green box. (B) Minimized structure of TAS2R7 interacting with Zn^{2+}. (C) Binding cavity of TAS2R7 (in light blue) explored by cromolyn in the docking simulations.

The electrostatic potential computed on the TAS2R7 model shows a negatively charged region (**Fig. 5A**) suitable for attracting cations. Accordingly, negatively charged or polar residues in this area, $E16^{1.42}$, $H94^{3.37}$, $E264^{7.32}$, and $E271^{7.39}$ (the superscripts refer to the Ballesteros-Weinstein notation (Ballesteros and Weinstein 1995)) are considered to interact with metal ions through strong electrostatic interactions (**Fig. 5B**).

To assess the importance of these residues, we performed site-directed mutagenesis. We mutated the negatively charged residues $E16^{1.42}$, $E264^{7.32}$, and $E271^{7.39}$ to Q (glutamine), K (lysine) or L (leucine). The facing $H94^{3.37}$ was mutated to F (phenylalanine). HEK293 cells that expressed mutant receptors along with Gα16-gust44 were examined for their responses to metal ions (20 mM for all except 0.16 mM $Al_2(SO_4)_3$) and cromolyn (10 mM, as a positive control) to assess receptor's function. To determine the expression level of each receptor, we stained the wildtype or mutant receptor-transfected cells using an anti-HSV antibody since all the receptors are tagged with HSV at c-terminal. There was no obvious difference in the intensity of the staining among mutants and wildtype receptors (**Fig. S4**). Compared to the wild-type receptor (**Fig. 6A**), two classes of mutants were noted: those showing significantly diminished responses to only a subset of metal ions (**Fig. 6B**), and those showing either normal or reduced responses to both metal ions and cromolyn (**Fig. 6C**).

For example, $H94^{3.37}F$ showed diminished responses specifically $MnCl_2$ (**Fig. 6B**). In contrast, $E264^{7.32}K$ showed specific loss of responses to $ZnSO_4$ and $Al_2(SO_4)_3$, and $E264^{7.32}L$ responded to $ZnSO_4$ but not to $Al_2(SO_4)_3$. With both $E264^{7.32}K$ and $E264^{7.32}L$, the overall responses of mutant receptors to metal ions and cromolyn were reduced. Similarly, substitution of glutamate with glutamine ($E264^{7.32}Q$) led to a mutant receptor showing reduced responses to metal ions and cromolyn but did not specifically affect the receptor response to a particular metal ion. Substitution of glutamate at $E16^{1.42}$ with other residues showed no specific effects on the activity of metallic ions. However, with the exception of $E16^{1.42}L$, all other mutations led to relatively smaller responses to both metal ions and cromolyn compared with wild-type TAS2R7. Substitution of $E271^{7.39}$ with either glutamine or leucine led to a mutant receptor showing slightly reduced responses to all metal ions and cromolyn *in vitro*. Together, our mutagenesis data suggest the involvement of $H94^{3.37}$ and $E264^{7.32}$ in interacting with metal ions.

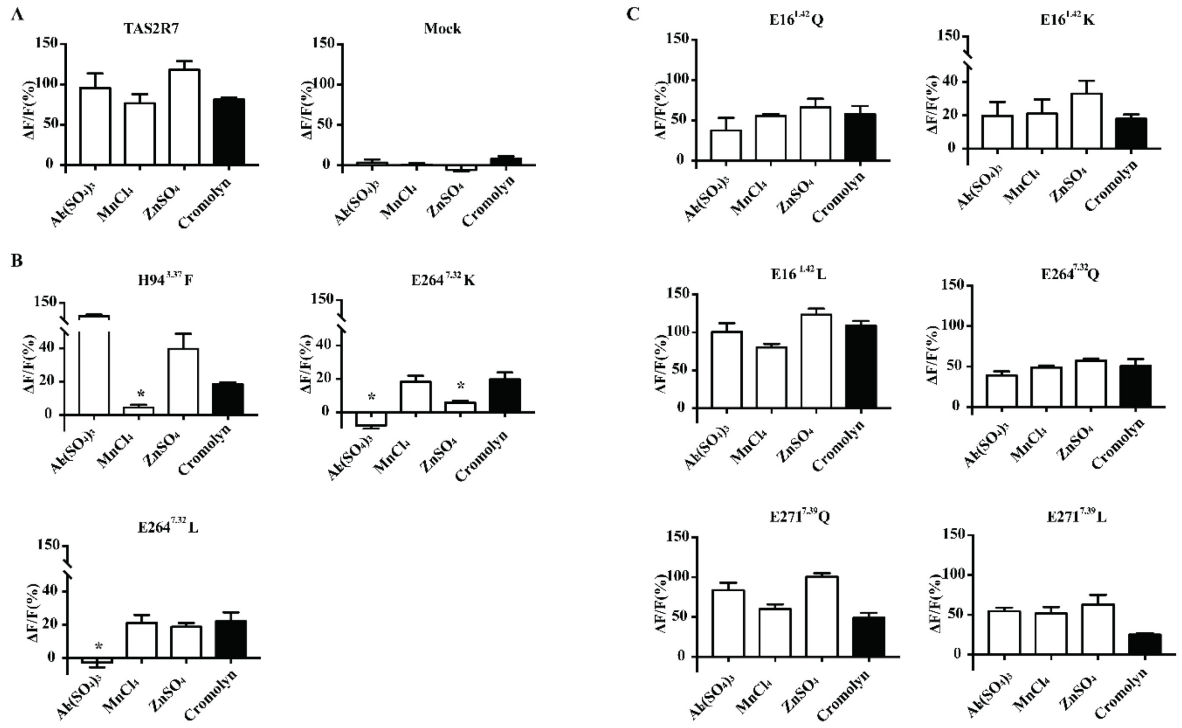


Figure 6. Mutagenesis analysis of the predicted binding pocket for metal ions. Wild-type (A) and mutant receptors (B & C) were expressed along with Ga16-gust44 in HEK293 cells, and their responses to metal ions and cromolyn were examined. Panel B includes mutant receptors showing selectively reduced responses to certain metal ions. Panel C includes mutant receptors showing no specific reduction in responses to metal ions. Dunnett's multiple comparisons test was performed to determine when the responses to metallic ions of mutants were significantly decreasing from that of cromolyn, indicated with an asterisk (*). ($p < 0.05$). Experiments were replicated three times.

Discussion

TAS2R7 as a metal ion detector

By systematically assaying all the human bitter receptors for their responsiveness to metal ions, we found that TAS2R7 acts as a receptor for divalent and trivalent cations. To our knowledge, only CaSR and GPR39 have been previously shown to be metal-sensing receptors (Brown *et al.* 1993; Holst *et al.* 2007; McGehee *et al.* 1997; Saidak *et al.* 2009). Identification of TAS2R7 as a metal-ion-sensing receptor broadens our understanding how metal ions are sensed.

TAS2Rs evolved to detect bitter substances (which are potentially harmful or toxic) in diets. Activation of these receptors would then induce aversive behavior as a defense mechanism

(Bachmanov and Beauchamp 2007). Most natural compounds that taste bitter are plant derived. Some plants are known to be rich in minerals. Vegetable bitterness is shown to be related to calcium content (Tordoff and Sandell 2009). Thus, activation of TAS2R7 may contribute to bitterness associated with calcium-rich (or mineral-rich) vegetables. Future work is warranted to determine if blocking TAS2R7 (e.g., inhibitors of TAS2R7) can reduce bitterness or metallic taste of metal ions or mineral-rich foods.

Taste disturbance is a widely reported side effect for cancer patients who receive chemotherapy or radiotherapy (Comeau *et al.* 2001). Often, they complain about bitter taste or metallic taste (Comeau *et al.* 2001). It is conceivable that such treatments may alter bitter receptor gene expression, such as upregulation of TAS2R7 that is normally expressed at a low level. Metal ions in the blood may activate the receptor, leading to bitter/metallic taste perception in pathological conditions. Blocking TAS2R7 activity may provide a therapeutic strategy for alleviating chemotherapy- or radiotherapy-induced taste disturbance.

Interaction of metal ions and TAS2R7

Our structure-function analysis of TAS2R7 showed differential requirements of H94 in helix 3 and E264 in helix 7 for their interaction with different metal ions. Substitution of the histidine residue at position 94 ($H^{3.37}$) with phenylalanine diminished responsiveness of the receptor toward $MnCl_2$ more than toward $Al_2(SO_4)_3$, $ZnSO_4$, and cromolyn *in vitro*. Conversely, substitution of the negative-charged glutamate residue at position 264 ($E^{7.32}$) with positive-charged lysine rendered the receptor insensitive to $Al_2(SO_4)_3$ and $ZnSO_4$ but still responsive to $MnCl_2$. Similarly, substitution with the neutral but slightly bulkier leucine residue also rendered the receptor insensitive to $Al_2(SO_4)_3$. Altogether, our demonstration of the contribution of H94 and E264 to a binding pocket for metal ions is supported by both mutagenesis analysis and molecular modeling. Additionally, we showed that these ions interact distinctively with residues lining this binding pocket. Especially, the presence or absence of calcium in the assay solution appears to influence the responses of TAS2R7 distinctly for different metal ions. We don't know the reason but speculate that calcium may work cooperatively with certain ions (e.g., $ZnSO_4$, $MgCl_2$) than with others (e.g., $CuSO_4$). Future detailed structure-function analysis of interactions of the receptor and metal ions will provide further insights into how metal ions activate the receptor.

Potential extraoral function of TAS2R7

Recently, TAS2Rs have been shown to be expressed not only in the oral cavity but also in many other tissues in the body (Behrens and Meyerhof 2011). However, the endogenous cognate ligands for these extra-oral receptors are largely unknown. Compared with other TAS2Rs, TAS2R7 is reported to be weakly expressed in taste bud cells (Behrens *et al.* 2007). Using immunostaining and RT-PCR, it has been shown that TAS2R7 is also expressed in pancreatic islet cells (Chen *et al.* 2007).

Zinc is known to be an important regulator of islet function. Pancreatic β cells contain high concentrations of zinc in the secretory granules (Wijesekara *et al.* 2009). Upon excitation of β cells, Zn^{2+} is coreleased at high concentrations with insulin into the extracellular space of the islet. Given the presumptive expression of TAS2R7 in a subset of islet cells, it is tempting to speculate that the released Zn^{2+} may act on TAS2R7-expressing cells to regulate glucose homeostasis. Indeed, using human genetic approaches, Dotson *et al.* (Dotson *et al.* 2008) showed that a nonsynonymous coding SNP in TAS2R7 is associated with type 2 diabetes mellitus. However, we found no significant difference in the responsiveness of TAS2R7 having isoleucine residue at the position 304 and the receptor carrying M304 toward divalent and trivalent metals (data not shown).

There is compelling evidence supporting that extracellular Al^{3+} at micromolar concentrations activates a GPCR-like signaling pathway in certain cells (Spurney *et al.* 1999). Aluminum has been shown to be a weak agonist for CaSR (Spurney *et al.* 1999). Given the efficacious response of TAS2R7 toward Al^{3+}, it is possible that TAS2R7 mediates certain biological responses elicited by aluminum ions. Indeed, Al^{3+} administered systemically can reach 50 μM in serum in animal studies and stimulates osteoblast-mediated de novo bone formation in vivo and osteoblast proliferation in vitro (Lau *et al.* 1991). This is within the sensitivity of TAS2R7 to Al^{3+} (**Fig. 2**).

Another study performed by Velazquez-Fernandez *et al.* (Velazquez-Fernandez *et al.* 2006) showed that TAS2R7 is upregulated in parathyroid adenoma samples compared to parathyroid hyperplasia samples, suggesting a potential link between TAS2R7 and regulation of calcium homeostasis. However, CaSR acts as a principal regulator of calcium homeostasis.

CaSR is known to respond to a variety of divalent and trivalent ions (18-20). Despite similarity in the responses to divalent and trivalent ions of CaSR and TAS2R7, differences between these two receptors are notable. For example, TAS2R7 responds to zinc ions, and CaSR does not. Thus, in terms of specificity for metal ions, TAS2R7 appears to be more broadly tuned. The

physiological role of TAS2R7 in extraoral tissues and the possibility of metal ions as its endogenous ligands warrant future investigation.

Acknowledgments

We thank members of the Jiang laboratory for discussion.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

Funding

This work was supported in part by a grant from the Bill and Melinda Gates Foundation (OPP1159241). Calcium assays were performed at the Monell Chemosensory Receptor Signaling Core, which was supported in part by NIH–National Institute on Deafness and Other Communication Disorders Core Grant DC011735 (R.F.M.). Y.W. was partially supported by the China Scholarship Council, and H.Z. was supported by the National Natural Science Foundation of China (31722051). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Y.W., A.L.Z., C.C., R.F.M, and P.J. designed research; Y.W., A.L.Z., and W.L. performed research; J.G., C.B. and S.F. performed molecular modeling; Y.W., A.L.Z. J.G., H.Z., S.F., and P.J. analyzed data; Y.W., S.F., and P.J. wrote the paper.

References

Bachmanov AA, Beauchamp GK. 2007. Taste receptor genes. Annu Rev Nutr 27: 389-414.

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98: 10037-10041.

Ballesteros JA, Weinstein H. 1995. [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods in Neurosciences 25: 366-428.

Behrens M, Foerster S, Staehler F, Raguse JD, Meyerhof W. 2007. Gustatory expression pattern of the human TAS2R bitter receptor gene family reveals a heterogenous population of bitter responsive taste receptor cells. J Neurosci 27: 12630-12640.

Behrens M, Meyerhof W. 2011. Gustatory and extragustatory functions of mammalian taste receptors. Physiol Behav 105: 4-13.

Brown EM, Gamba G, Riccardi D, Lombardi M, Butters R, Kifor O, Sun A, Hediger MA, Lytton J, Hebert SC. 1993. Cloning and characterization of an extracellular Ca(2+)-sensing receptor from bovine parathyroid. Nature 366: 575-580.

Bufe B, Breslin PA, Kuhn C, Reed DR, Tharp CD, Slack JP, Kim UK, Drayna D, Meyerhof W. 2005. The molecular basis of individual differences in phenylthiocarbamide and propylthiouracil bitterness perception. Curr Biol 15: 322-327.

Bufe B, Hofmann T, Krautwurst D, Raguse JD, Meyerhof W. 2002. The human TAS2R16 receptor mediates bitter taste in response to beta-glucopyranosides. Nat Genet 32: 397-401.

Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ. 2005. The Amber biomolecular simulation programs. J Comput Chem 26: 1668-1688.

Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, Zuker CS, Ryba NJ. 2000. T2Rs function as bitter taste receptors. Cell 100: 703-711.

Chen J, Yan P, Wang J, Cai Y, Cao L, Cao Y. 2007. The localization of TAS2R7 in artery and pancreas. AChemS XXXIX Abstract. P138.

Comeau TB, Epstein JB, Migas C. 2001. Taste and smell dysfunction in patients receiving chemotherapy: a review of current knowledge. Support Care Cancer 9: 575-580.

Dotson CD, Zhang L, Xu H, Shin YK, Vigues S, Ott SH, Elson AE, Choi HJ, Shaw H, Egan JM, Mitchell BD, Li X, Steinle NI, Munger SD. 2008. Bitter taste receptors influence glucose homeostasis. PLoS One 3: e3974.

Feng P, Zheng J, Rossiter SJ, Wang D, Zhao H. 2014. Massive losses of taste receptor genes in toothed and baleen whales. Genome Biol Evol 6: 1254-1265.

Go Y, Satta Y, Takenaka O, Takahata N. 2005. Lineage-specific loss of function of bitter taste receptor genes in humans and nonhuman primates. Genetics 170: 313-326.

Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. 2005. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. Nucleic Acids Res 33: W368-371.

Holst B, Egerod KL, Schild E, Vickers SP, Cheetham S, Gerlach LO, Storjohann L, Stidsen CE, Jones R, Beck-Sickinger AG, Schwartz TW. 2007. GPR39 signaling is stimulated by zinc ions but not by obestatin. Endocrinology 148: 13-20.

Jiang P, Josue J, Li X, Glaser D, Li W, Brand JG, Margolskee RF, Reed DR, Beauchamp GK. 2012. Major taste loss in carnivorous mammals. Proc Natl Acad Sci U S A 109: 4956-4961.

Jiao H, Wang Y, Zhang L, Jiang P, Zhao H. 2018. Lineage-specific duplication and adaptive evolution of bitter taste receptor genes in bats. Mol Ecol 27: 4475-4488.

Kim UK, Jorgenson E, Coon H, Leppert M, Risch N, Drayna D. 2003. Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. Science 299: 1221-1225.

Lau KH, Yoo A, Wang SP. 1991. Aluminum stimulates the proliferation and differentiation of osteoblasts in vitro by a mechanism that is different from fluoride. Mol Cell Biochem 105: 93-105.

Lei W, Ravoninjohary A, Li X, Margolskee RF, Reed DR, Beauchamp GK, Jiang P. 2015. Functional Analyses of Bitter Taste Receptors in Domestic Cats (Felis catus). PLoS One 10: e0139670.

Lim J, Lawless HT. 2005. Qualitative differences of divalent salts: multidimensional scaling and cluster analysis. Chem Senses 30: 719-726.

Liman ER. 2006. Use it or lose it: molecular evolution of sensory signaling in primates. Pflugers Arch 453: 125-131.

Liu K, Jaggupilli A, Premnath D, Chelikani P. 2018. Plasticity of the ligand binding pocket in the bitter taste receptor T2R7. Biochim Biophys Acta Biomembr 1860: 991-999.

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 11: 3696-3713.

Matsunami H, Montmayeur JP, Buck LB. 2000. A family of candidate taste receptors in human and mouse. Nature 404: 601-604.

McGehee DS, Aldersberg M, Liu KP, Hsuing S, Heath MJ, Tamir H. 1997. Mechanism of extracellular Ca2+ receptor-stimulated hormone release from sheep thyroid parafollicular cells. J Physiol 502 (Pt 1): 31-44.

Meyerhof W, Batram C, Kuhn C, Brockhoff A, Chudoba E, Bufe B, Appendino G, Behrens M. 2010. The molecular receptive ranges of human TAS2R bitter taste receptors. Chem Senses 35: 157-170.

Peng Y, McCorvy JD, Harpsoe K, Lansu K, Yuan S, Popov P, Qu L, Pu M, Che T, Nikolajsen LF, Huang XP, Wu Y, Shen L, Bjorn-Yoshimoto WE, Ding K, Wacker D, Han GW, Cheng J, Katritch V, Jensen AA, Hanson MA, Zhao S, Gloriam DE, Roth BL, Stevens RC, Liu ZJ. 2018. 5-HT2C Receptor Structures Reveal the Structural Basis of GPCR Polypharmacology. Cell 172: 719-730 e714.

Perez CA, Huang L, Rong M, Kozak JA, Preuss AK, Zhang H, Max M, Margolskee RF. 2002. A transient receptor potential channel expressed in taste receptor cells. Nat Neurosci 5: 1169-1176.

Riera CE, Vogel H, Simon SA, Damak S, le Coutre J. 2009. Sensory attributes of complex tasting divalent salts are mediated by TRPM5 and TRPV1 channels. J Neurosci 29: 2654-2662.

Saidak Z, Brazier M, Kamel S, Mentaverri R. 2009. Agonists and allosteric modulators of the calcium-sensing receptor and their therapeutic applications. Mol Pharmacol 76: 1131-1144.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815.

Shi P, Zhang J. 2006. Contrasting modes of evolution between vertebrate sweet/umami receptor genes and bitter receptor genes. Mol Biol Evol 23: 292-300.

Spurney RF, Pi M, Flannery P, Quarles LD. 1999. Aluminum is a weak agonist for the calcium-sensing receptor. Kidney Int 55: 1750-1758.

Tordoff MG, Sandell MA. 2009. Vegetable bitterness is related to calcium content. Appetite 52: 498-504.

Tordoff MG, Shao H, Alarcon LK, Margolskee RF, Mosinger B, Bachmanov AA, Reed DR, McCaughey S. 2008. Involvement of T1R3 in calcium-magnesium taste. Physiol Genomics 34: 338-348.

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

Velazquez-Fernandez D, Laurell C, Saqui-Salces M, Pantoja JP, Candanedo-Gonzalez F, Reza-Albarran A, Gamboa-Dominguez A, Herrera MF. 2006. Differential RNA expression profile by cDNA microarray in sporadic primary hyperparathyroidism (pHPT): primary parathyroid hyperplasia versus adenoma. World J Surg 30: 705-713.

Wang K, Zhao H. 2015. Birds Generally Carry a Small Repertoire of Bitter Taste Receptor Genes. Genome Biol Evol 7: 2705-2715.

Wijesekara N, Chimienti F, Wheeler MB. 2009. Zinc, a regulator of islet function and glucose homeostasis. Diabetes Obes Metab 11 Suppl 4: 202-214.

Yang HH, Lawless HT. 2005. Descriptive Analysis of Divalent Salts. J Sens Stud 20: 97-113.

Zhang Y, Hoon MA, Chandrashekar J, Mueller KL, Cook B, Wu D, Zuker CS, Ryba NJ. 2003. Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. Cell 112: 293-301.

Supporting information

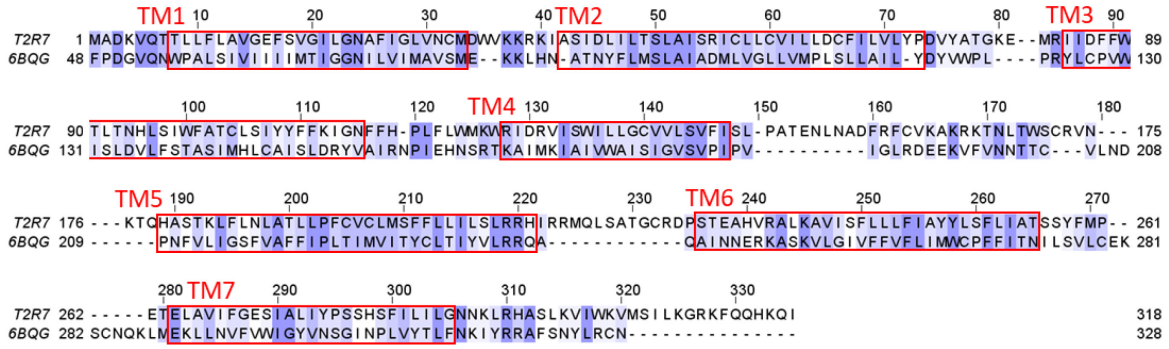


Figure S1. Alignment of TAS2R7 and 5-HT$_{2C}$ serotonin receptor (PDB 6BQG) sequences.

Transmembrane helices are delimited by red boxes. Conserved residues are shown in dark blue. Aligned residues with a positive Blosum62 score are shown in light blue.

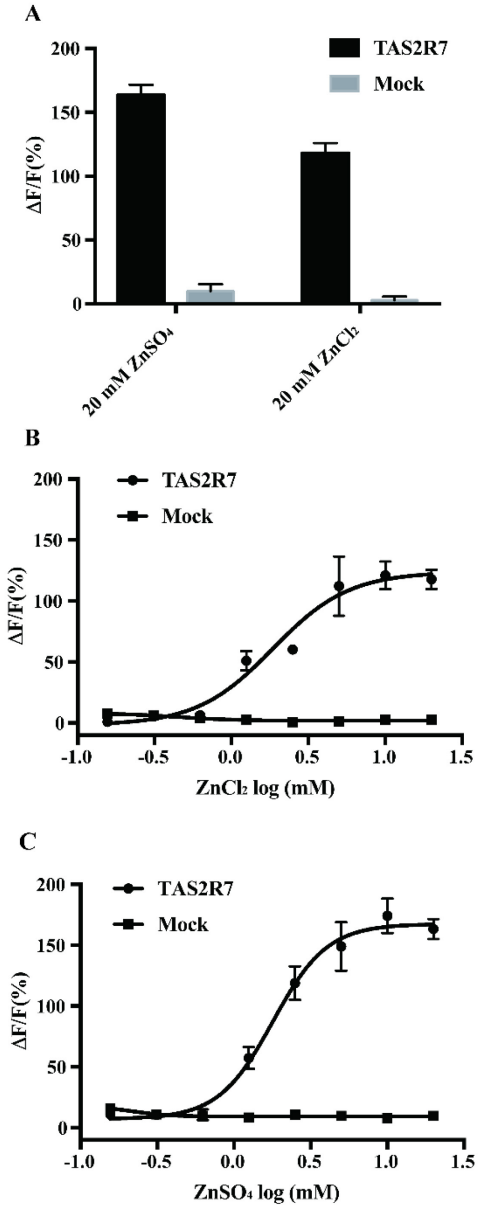


Figure S2. Responses of TAS2R7 toward $ZnCl_2$ and $ZnSO_4$.

Responses of HEK293 cells transiently transfected with human TAS2R7 with Gα16-gust44 to $ZnSO_4$ and $ZnCl_2$, respectively. A) Quantitative analysis of responses of TAS2R7 to 20 Mm $ZnSO_4$ and $ZnCl_2$. Data are percentage change (mean ± SD) in fluorescence (peak RFU – baseline RFU, denoted ΔF) from baseline fluorescence (denoted F). Experiments were replicated three times. B, C) Dose-dependent curves of TAS2R7 toward $ZnSO_4$ and $ZnCl_2$, Experiments were replicated twice.

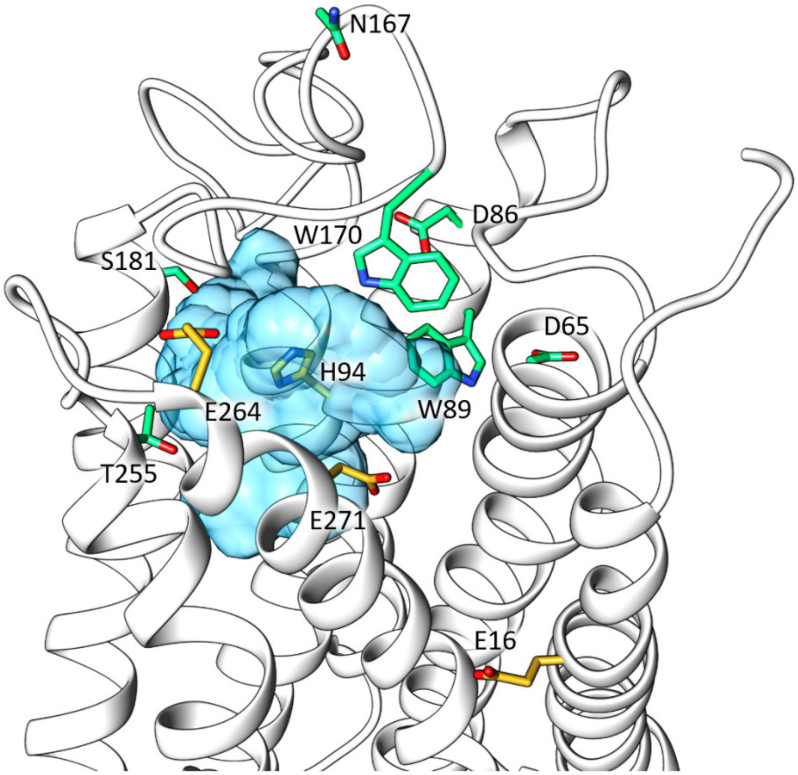


Figure S3. Comparison of our TAS2R7 homology model with a previously published model. The binding cavity of TAS2R7 (in blue) was explored with cromolyn during the docking simulations. Residues proposed by Liu *et al.* (21) to be part of the binding pocket are shown in green: D65, D86, W89, N167, W170, S181, T255, E271. Residues affecting metallic interaction suggested by the present study are shown in yellow: E16, H94, E264, E271.

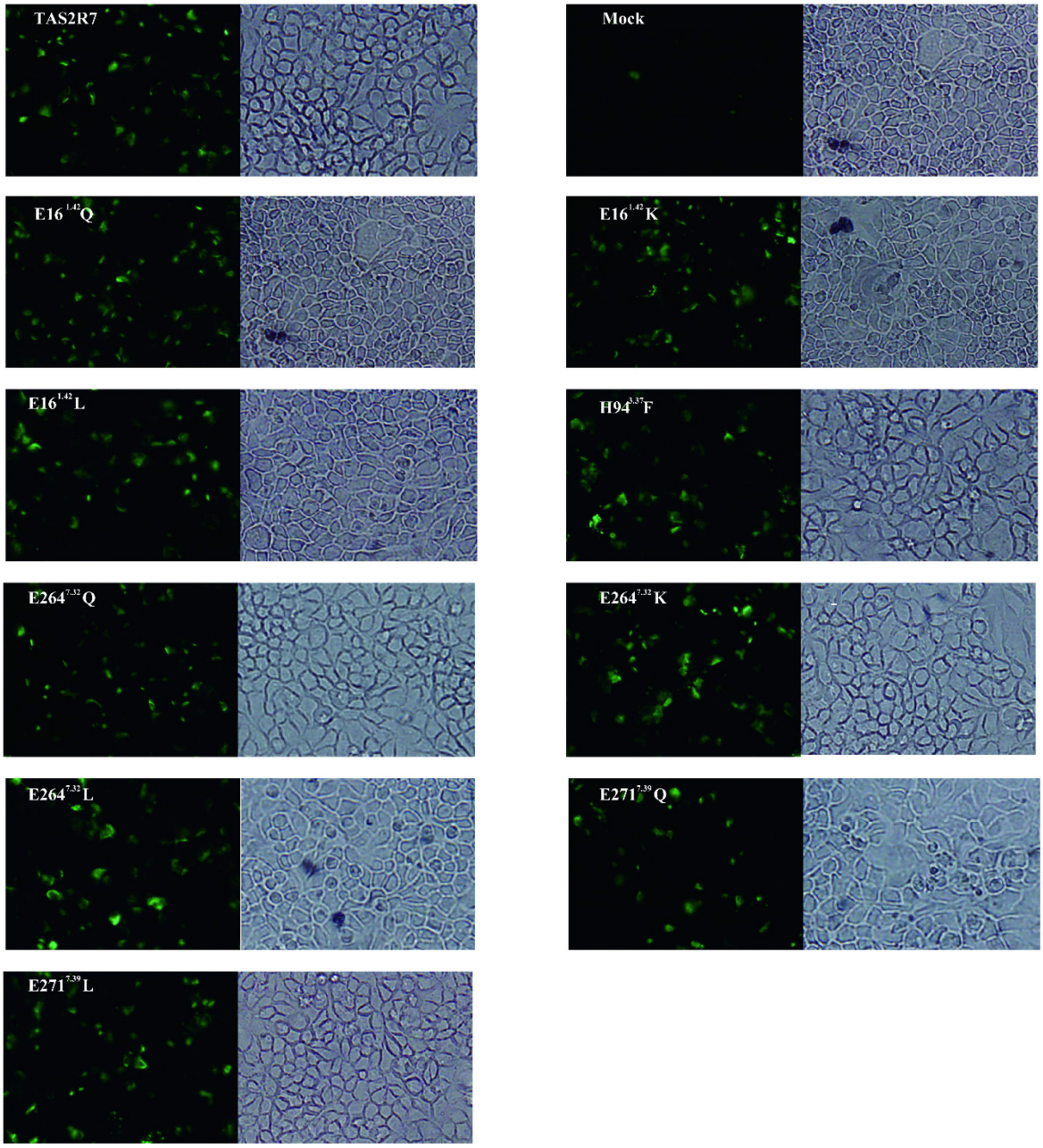


Figure S4. **Immunostaining of cells transfected with mutants and wildtype TAS2R7 receptors.** HEK293 cells expressing TAS2R7 or its mutants were immunostained with an anti-HSV antibody. An Alexa Fluor 488-labled Donkey anti-mouse secondary antibody was used for fluorescence visualization (Green). A brightfield image of the same field was shown next to the fluoresent image. Images were taken with the same exposure time and the same setting.

Appendix

Publication A2

ProLIF: a library to encode molecular interactions as fingerprints

Cédric Bouysset*, Sébastien Fiorucci*

Under revision

Abstract

Interaction fingerprints are vector representations that summarize the three-dimensional nature of interactions in molecular complexes, typically formed between a protein and a ligand. This kind of encoding has found many applications in drug-discovery projects, from structure-based virtual-screening to machine-learning. Here, we present ProLIF, a Python library designed to generate interaction fingerprints for molecular complexes extracted from molecular dynamics trajectories, experimental structures, and docking simulations. It can handle complexes formed of any combination of ligand, protein, DNA, or RNA molecules. The available interaction types can be fully reparametrized or extended by user-defined ones. Several tutorials that cover typical use-case scenarios are available, and the documentation is accompanied with code snippets showcasing the integration with other data-analysis libraries for a more seamless user-experience. The library can be freely installed from our GitHub repository (https://github.com/chemosim-lab/ProLIF).

Keywords

Interaction fingerprint, structural biology, molecular dynamics, docking, virtual screening, Python

Introduction

Interactions between and within molecular structures are the driving force behind biological processes, from protein folding to molecular recognition. The decomposition of interactions by residues in biomolecular complexes can provide insights into structure-function relationships, and characterizing the nature of each of these interactions can guide medicinal chemists in structure-based drug discovery projects [1]. Approaches to encode the interactions observed in 3D structural data in the form of a binary fingerprint have been developed in the past [2–6] and applied successfully to a variety of projects. For example, de Graaf et al. [7] used the Tanimoto similarity between the interaction fingerprint (IFP) of a crystallographic reference and the IFP of docking poses to rescore virtual screening results on a G protein-coupled receptor (GPCR). Rodríguez-Pérez et al. [8] showed that IFPs can achieve superior predictive performance than ligand fingerprints (ECFP4) for the classification of kinase inhibitor binding modes with machine-learning models. Finally, Mpamhanga et al. [9] showed that one can use the IFP for clustering, and then shortlist a reasonable number of binding modes prior to visual inspection.

More recently, the approach was also implemented for molecular dynamics (MD) simulations to study ligand unbinding [10]. While the typical IFP usually encodes pre-established interactions (hydrogen bond, π-stacking…*etc.*) on a per-residue basis, other implementations exist. Sato et al. [11] developed a pharmacophore-based IFP which relies on the pharmacophoric features of the ligand atoms in contact with the protein and the distance between each of these pharmacophores to generate a bitvector. Da et al. [12] developed an IFP that relies on the atomic environment of both the protein and ligand interacting atoms to set the positions of a bit in the fingerprint, rather than relying on protein residues and predefined interactions, which has the advantage of implicitly encoding every possible type of interaction. This protocol was later reimplemented in Python by Wójcikowski et al. [13], but other more classical Python-based IFP implementations exist [14–18]. In this paper, we introduce a new Python library, ProLIF, that overcomes several limitations encountered by these programs, namely working exclusively with the output of specific docking programs, not being compatible with the analysis of MD trajectories, being restricted to a specific kind of complex (usually protein-ligand complexes), depending on residue or atom type naming conventions, or not being extensible or configurable regarding interactions.

Implementation

ProLIF can deal with RDKit [19] molecules or MDAnalysis [20] Universe objects as input, which allows supporting most 3D molecular formats, from docking to MD simulations. While most MD topology files do not keep explicit information about bond orders and formal charges, MDAnalysis is able to infer this information if all hydrogen atoms are explicit in the structure while converting the structure to an RDKit molecule. The RDKit parent molecule is then automatically fragmented in child residue molecules based on residues name, number, and chain to make it easier to work on a per-residue basis when encoding the interactions.

When calculating an interaction fingerprint, each interaction is typically defined as two groups of atoms that satisfy geometrical constraints based on distances and/or angles (Table 1). Here the selection of atoms is made using SMARTS queries (Table 2), which is more precise than relying on elements or atomic weights and is also more universal than relying on force-field-specific atom types.

Table 1: Interactions currently available in ProLIF

Interaction	Ligand*	Protein*	Distance (Å)	Angle (deg)
Anionic Cationic	Anion Cation	Cation Anion	≤ 4.5	
CationPi PiCation	Cation Aromatic	Aromatic Cation	(+)-ctd ≤ 4.5	$\langle \vec{n}, \overrightarrow{ctd \cdots (+)} \rangle \in [0, 30]$
PiStacking	Aromatic	Aromatic	ctd-ctd ≤ 6.0 min ≤ 3.8	$\langle \vec{n}, \vec{n} \rangle \in [0, 90]$
EdgeToFace	Aromatic	Aromatic	ctd-ctd ≤ 6.0 min ≤ 3.8	$\langle \vec{n}, \vec{n} \rangle \in [50, 90]$
FaceToFace	Aromatic	Aromatic	ctd-ctd ≤ 4.5 min ≤ 3.8	$\langle \vec{n}, \vec{n} \rangle \in [0, 40]$
HBAcceptor HBDonor	HBAcceptor HBDonor	HBDonor HBAcceptor	D-A ≤ 3.5	$\langle \overrightarrow{HD}, \overrightarrow{HA} \rangle \in [130, 180]$
XBAcceptor XBDonor	XBAcceptor XBDonor	XBDonor XBAcceptor	X-A ≤ 3.5	$\langle \overrightarrow{XD}, \overrightarrow{XA} \rangle \in [130, 180]$ $\langle \overrightarrow{AX}, \overrightarrow{AR} \rangle \in [80, 140]$
MetalAcceptor MetalDonor	Ligand Metal	Metal Ligand	≤ 2.8	
Hydrophobic	Hydrophobic	Hydrophobic	≤ 4.5	

*Although "ligand" and "protein" are used here, all the listed interactions can be applied to any molecular complex (protein-protein, DNA-protein…etc.). (-): anion, (+): cation, ctd: centroid of the aromatic ring, min: minimum value in the distance matrix between both aromatic rings, n: normal to the aromatic ring plane, D: hydrogen/halogen bond donor, A: hydrogen/halogen bond acceptor, H: hydrogen atom, X: halogen atom, R: atom linked to a halogen bond acceptor.

Table 2: SMARTS patterns used in the definition of interactions.

Name	SMARTS pattern(s)
Anion	[-{1-}]
Cation	[+{1-}]
Aromatic	a1:a:a:a:a:a:1 a1:a:a:a:a:1
HBAcceptor	[N,O,F,-{1-};!+{1-}]
HBDonor	[#7,#8,#16][H]
XBAcceptor	[#7,#8,P,S,Se,Te,a;!+{1-}][*]
XBDonor	[#6,#7,Si,F,Cl,Br,I]-[Cl,Br,I,At]
Metal	[Ca,Cd,Co,Cu,Fe,Mg,Mn,Ni,Zn]
Ligand	[O,N,-{1-};!+{1-}]
Hydrophobic	[#6,#16,F,Cl,Br,I,At;+0]

The library is designed so that users can easily modify existing interactions, as there is usually no consensus on the empirical thresholds (distance, angles) that should be used. For example, the hydrogen bond DH…A can be defined as a distance between H and A lower or equal to 3.0 Å [9] or as a distance between D and A lower or equal to 3.5 [4, 14, 21] or 4.1 Å [15, 22], and the angles constraints can also vary. ProLIF is also designed to let users define custom interactions.

Each interaction is written as a Python class that implements a “detect” method which takes two RDKit molecules as input, typically a ligand and a protein residue, and outputs a Boolean (True if the interaction is present, else False) as well as the indices of atoms responsible for the interaction. All interaction classes are then gathered inside a “Fingerprint” class that can generate a bitvector from two RDKit molecules, and optionally return the atom indices. By default, the Fingerprint class is configured to generate a bitvector with the following interactions: hydrophobic, π-stacking, π-cation and cation-π, anionic and cationic, and H-bond donor and acceptor, although more specific interactions are available (see Table 1). This Fingerprint class is designed with two scenarios in mind, post-processing MD trajectories or

docking results, thus it provides user-friendly functions to generate the complete array of interactions for each pair of interacting residues.

Finally, the interaction is stored inside the Fingerprint class as a mapping between a pair of "ligand" and "protein" residues, and the corresponding interaction bitvector. For easier post-processing, the interaction fingerprint can then be converted to a pandas DataFrame object [23], which facilitates the search for specific interactions and the aggregation of results.

Results and discussion

By relying on the interoperability with popular open-source libraries (MDAnalysis and RDKit), it can support a wide range of molecular formats typically found in docking experiments and MD simulations. Because it directly relies on SMARTS patterns to define the chemical moieties that partake in interactions, it is also compatible with any kind of molecular complex, including complexes made of ligands, proteins, DNA or RNA molecules. Interoperability also allows for data analysis to be substantially easier: as mentioned in the Implementation section the IFP can be directly exported to a pandas DataFrame (one of Python's most popular data analysis library), and the documentation contains tutorials on how to visualize the interactions as graphs or how to display them on the 3D structure of the complex.

Analysis of an MD trajectory of a GPCR in complex with a ligand

The code to run ProLIF on an MD trajectory can be as simple as follow:

```
import prolif as plf
import MDAnalysis as mda
# load trajectory with MDAnalysis
u = mda.Universe("topology.pdb", "traj.xtc")
# create selections for the ligand and protein
lig = u.atoms.select_atoms("resname LIG")
prot = u.atoms.select_atoms("protein")
# generate interaction fingerprint
fp = plf.Fingerprint()
fp.run(u.trajectory, lig, prot)
```

Here, we showcase an analysis based on the fingerprint obtained from a 500ns MD simulation of the 5-HT1B receptor (class A aminergic GPCR) in complex with ergotamine retrieved from the GPCRmd webserver (id 90) [24]. In class A GPCRs, each position is annotated in superscript notation according to the Ballesteros-Weinstein numbering scheme [25], a generic residue numbering denoting both the helix and position relative to the most conserved residue labelled as number 50.

Exporting the fingerprint to a DataFrame allows to easily address common questions like which residues are involved in a specific type of interaction, which interactions does a specific residue do, which are the most frequent types of interactions, or which are the residues most frequently interacting with the ligand. In this MD trajectory, there is constantly at least one hydrophobic, H-bond donor and cationic interaction, while H-bond acceptor and π-stacking interactions occur respectively in 92% and 85% of the analyzed frames (see analysis notebook in supplementary data). F331$^{6.52}$ is responsible for half of the π-stacking interactions occurring during the simulation, and the ten residues that interact with the ligand the most frequently are (in descending order): D129$^{3.32}$, I130$^{3.33}$, F330$^{6.51}$, V201$^{ECL2.52}$, F331$^{6.52}$, S212$^{5.42}$, W327$^{6.48}$, V200$^{ECL2.51}$, C133$^{3.36}$ and F351$^{7.35}$ which are all in contact with the ligand in at least 97% of frames. This is in agreement with the known interactions available from experimental structures as listed on the GPCRdb webpage [26] for the human 5-HT1B receptor, except for S212$^{5.42}$ which isn't reported to make H-bond interactions with ligands. The difference is likely due to the fact that this analysis is based on an MD trajectory while GPCRdb gathers interactions from experimental structures. However, GPCRdb also lists mutational data for S212$^{5.42}$, and mutating this position to an alanine does not affect the binding affinity to ergotamine [27] which can coincide with the MD simulation since the ligand makes a hydrogen bond with the backbone and not the sidechain. Mutating S212$^{5.43}$ to a bulkier residue could potentially affect this interaction and decrease the binding affinity.

Because ProLIF keeps track of the atom indices responsible for interactions, it is possible to display detailed 2D or 3D interaction plots. Examples of scripts to generate such plots are given in the documentation. An exception is made for the ligand interaction network diagram which has been directly included in the source code of ProLIF under the LigNetwork class. This LigNetwork diagram (Figure 1) is interactive and allows repositioning the residues but also hiding specific residue types or interactions by clicking the legend. It can show the interaction diagram at a precise frame or aggregate the results and only display interactions that appear frequently, controlled by a frequency threshold. In the latter case, to keep the plot readable for each ligand-protein-interaction group only the most frequent ligand atom is shown, as it might differ between frames.

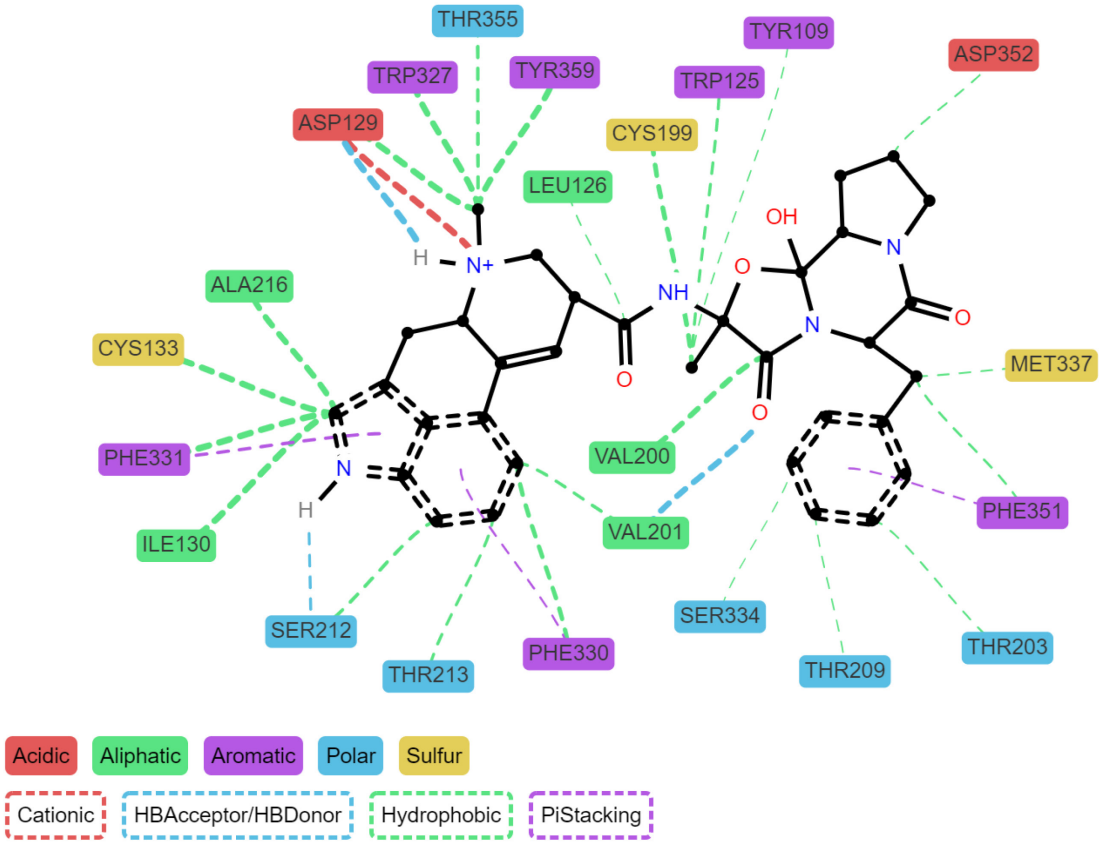


Figure 1: Ligand interaction network for the ergotamine agonist bound to the 5-HT1B receptor. Each interaction is shown as a dashed line between the residue and the ligand, and the width of the line is linked to the frequency of the interaction in the simulation. Only interactions occurring in at least 30% of frames are shown here.

The fingerprint can also be converted to an RDKit bitvector to make use of the similarity/distance metric functions implemented. This allows to investigate the presence of different binding modes in the simulation. In Figure 2, we show the Tanimoto similarity matrix between each interaction fingerprint during the MD simulation. Two clusters are visible (from frame 400 to 1400, and from frame 1400 to 2100) which reveals changes in the interactions between ergotamine and 5-HT1B. Indeed, in the second cluster the phenyl ring of ergotamine gets closer to the indole moiety, which disrupts hydrophobic contacts with W125$^{3.28}$, H-bonding with S212$^{5.42}$ and π-stacking with F351$^{7.35}$ to create new hydrophobic interactions with T203^{ECL2}, T209$^{5.39}$, S334$^{6.55}$ and D352$^{7.36}$.

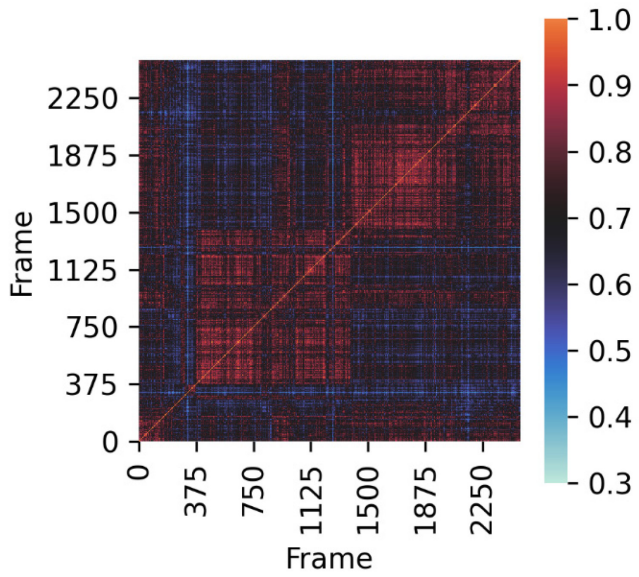


Figure 2: Tanimoto similarity matrix of ligand-protein interactions between each frame of the MD trajectory.

Analyzing protein-protein interactions (PPI)

The analysis of intra- and inter-molecular interactions can also be applied to investigate protein dynamics and function with ProLIF. Because ProLIF requires explicit hydrogen atoms, we preprocess PDB files of X-ray structures in the current section with the PDB2PQR [28] webserver as follows: AMBER force-field and naming scheme, protonation states assignments with PROPKA at pH 7.0, H-bond network optimization and removal of water molecules.

In this first example, we focus on the activation mechanism of a class A GPCR and show how ProLIF can help pinpoint intramolecular structural modifications upon receptor activation. GPCRs are membrane-embedded receptors arranged in seven helical transmembrane domains (labelled TM1 to TM7) followed by a shorter helix (H8) that lies at the interface between the membrane and the cytosol. This family shares conserved key motifs in each TM domain, and some of the motifs are part of molecular switches that mediate ligand binding or receptor activation. Among them, the DRY motif in TM3 and the NPxxY motif in TM7 have been reported to be part of the allosteric mechanism [29]. Briefly, upon ligand binding, the signal propagates from the binding pocket to the ionic lock (comprised of the DRY motif) through a network of hydrophobic residues. The ionic lock maintaining the receptor in its inactive form is disrupted, leading to an increase of the inter-helix distances (notably TM3-TM6). At the same time, the hydrophobic barrier cannot prevent anymore the flooding of the intracellular part of the receptor thereby creating an intracellular crevice required for G protein coupling. $R^{3.50}$ of

the DRY motif is known to stabilize the inactive form of the rhodopsin receptor through a salt-bridge with $D^{6.30}$ known as the "ionic-lock" [29]. This position can also interact with $Y^{5.58}$ through an H-bond, and is reported to be critical for the formation of the active state in the β2 adrenergic receptor [30]. For the NPxxY motif, the mutation of $Y^{7.53}$ disrupts interactions with $N^{2.40}$ in the β2 adrenergic receptor [31], and $Y^{7.53}$ is also reported to have an aromatic interaction with $F^{8.50}$ which stabilizes the inactive conformation of the rhodopsin receptor [32].

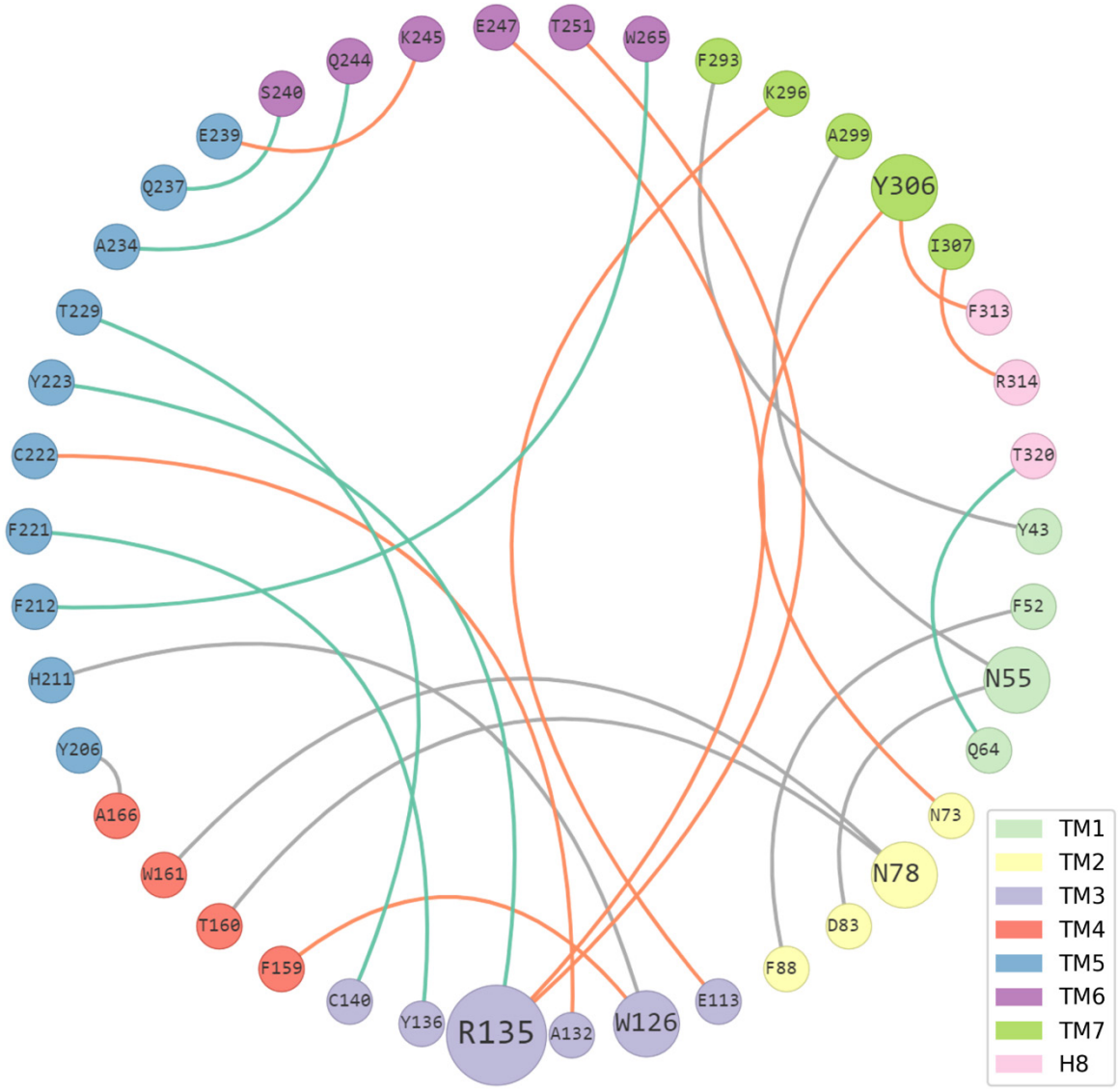


Figure 3: Residue interaction network for the bovine rhodopsin. Residues are colored by transmembrane domain (TM). Interactions that only appear in the active (PDB 6FK6) or inactive (PDB 1U19) state of the receptor are respectively shown in green or orange, and the ones that appear in both are in grey. Each residue node is scaled based on its number of interactions. For clarity, interactions that occur within the same TM (as labelled by GPCRdb) and interactions between residues that are less than 3 residues apart are not shown, as well as hydrophobic interactions (as defined in the implementation) and residues that did not participate in any interaction.

As an example, the residue interaction network of the bovine rhodopsin in both active (PDB 6FK6) and inactive (PDB 1U19) states is studied to reveal the structural changes involving these two motifs. As seen in Figure 3, the ionic lock between R135$^{3.50}$ and E247$^{6.30}$ is only visible in the inactive form of the receptor, while the interaction between R135$^{3.50}$ and Y223$^{5.58}$ was only detected in the active form. Y306$^{7.53}$, which is part of the NPxxY motif in TM7, takes part in both key interactions that stabilize the inactive form of the receptor previously described: an H-bond interaction with N73$^{2.40}$ and a π-stacking interaction with F313$^{8.50}$. Finally, in rhodopsin, the salt-bridge between K296$^{7.43}$ and E113$^{3.28}$ is known to be crucial in the activation cycle of the receptor and is only disrupted when K296$^{7.43}$ transiently bounds to retinal [33], which is in agreement with the interactions reported here.

The final step in GPCR signal transduction being an intermolecular process between the GPCR and a G-protein, ProLIF can also be used in this case to highlight positions that dictate the coupling specificity in a series of GPCR-G-protein complexes. Here, we reproduce the analysis of interactions between the β2 adrenoceptor and the Gαs/Gβ1 complex by Flock et al. [34] where the authors used a "van der Waals contact" interaction based on Venkatakrishnan et al. [35] which considers two residues as interacting if any interatomic distance is below or equal to their van der Waals interaction distance (the sum of their van der Waals radii plus a tolerance factor of 0.6 Å). We reimplemented this in ProLIF (see analysis notebook in supplementary data) and applied it to the same structure (PDB 3SN6) to obtain the PPI network shown in Figure 4. The interaction network remains mostly the same as with Figure S6 of the original study [34] and highlights the importance of positions I135$^{3.54}$, P138$^{34.50}$, F139$^{34.51}$, Q229$^{5.68}$, R239^{ICL3} and T274$^{6.36}$ for GPCR-G protein coupling. Using the default ProLIF implementation would help clarifying the types of interactions involved (H-bond, ionic…etc.) for a better understanding of coupling specificity when several GPCR-G protein complexes are investigated.

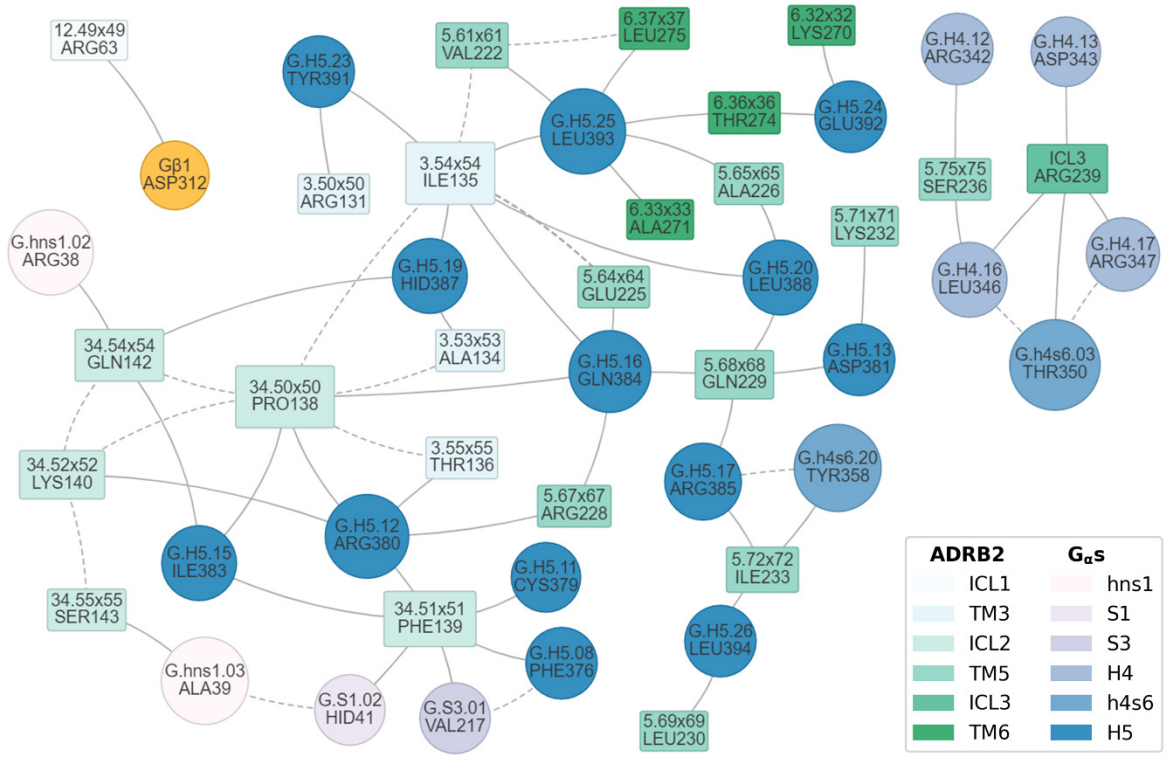


Figure 4: Interaction network between the β2 adrenoceptor (ADRB2) and G protein complex (Gαs and Gβ1). ADRB2 residues are shown as rectangles in shades of green, and G protein residues are shown as ellipses in shades of blue for Gαs and in yellow for Gβ1. For ADRB2, ICL denotes the intracellular loops while TM corresponds to the transmembrane domains. For Gαs, the common Gα numbering (CGN) system is used [36]. Each node is scaled by its number of interactions. Inter and intra protein interactions are respectively shown as plain and dashed lines. Residues that do not participate in GPCR-G protein interactions are not shown, and interactions between covalently bonded residues or residues of the same helix (as labelled by GPCRdb) are hidden.

Conclusions

ProLIF is a new Python library that overcomes limitations encountered by other freely available IFP programs. One of the main differences is the support of MD trajectories, while still being compatible with other molecular structure files like docking and experimental structures. By design, it is also not restricted to a particular kind of molecular complex but supports any combination of ligand, protein, DNA, or RNA molecules, thanks to its absence of dependency to force-field specificities such as atom types or residue naming convention. It also has a user-friendly API, comes with several tutorials, and allows creating custom interactions or reconfiguring existing ones. Finally, it focuses on the integration with typical data-analysis

packages and visualization tools for a seamless user experience within the Python ecosystem. Possible improvements include the addition of more interactions types, but also more types of fingerprints such as the pharmacophoric [11] or circular [12, 13] fingerprints. Adding a command-line interface would also extend the userbase to researchers inexperienced in Python. Another point of interest could be the extension to other popular visualization libraries for a more streamlined data analysis experience for users.

Availability and requirements

Project name: ProLIF

Project home page: https://github.com/chemosim-lab/ProLIF

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3.6 or higher, and several open-source Python packages listed in the project's documentation

License: Apache License 2.0

Declaration

Availability of data and materials

The datasets and code supporting the conclusions of this article are available in the supplementary materials and in the GitHub repository, https://github.com/chemosim-lab/ProLIF-paper, or through the Zenodo archive, https://doi.org/10.5281/zenodo.4945869.

Competing interests

The authors declare no competing financial interest.

Funding

This work was funded by the French Ministry of Higher Education and Research [PhD Fellowship to CB], by GIRACT (Geneva, Switzerland) [9th European PhD in Flavor Research Bursaries for first year students to CB] and the Gen Foundation (Registered UK Charity No. 1071026) [a charitable trust which principally provides grants to students/researchers in natural sciences, in particular food sciences/technology to CB].

Authors' contributions

CB designed the software and analyzed the data. SF managed the study. CB and SF interpreted the data and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Jody Pacalon and Matej Hladiš for their helpful comments on the manuscript. The authors would also like to thank users who reported issues with the code and documentation which ultimately led to improvements

Authors' information

CB received a stipend from Google for contributing code to MDAnalysis as part of a Google Summer of Code program. The contributed code, which allows for interoperability between MDAnalysis and RDKit, is part of the dependencies that the software described in this manuscript relies on.

References

1. Fischer A, Smieško M, Sellner M, Lill MA (2021) Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. J Med Chem 64:2489–2500. https://doi.org/10.1021/acs.jmedchem.0c02227

2. Deng Z, Chuaqui C, Singh J (2004) Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. Journal of Medicinal Chemistry 47:337–344. https://doi.org/10.1021/jm030331x

3. Kelly MD, Mancera RL (2004) Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. Journal of Chemical Information and Computer Sciences 44:1942–1951. https://doi.org/10.1021/ci049870g

4. Marcou G, Rognan D (2007) Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. Journal of Chemical Information and Modeling 47:195–207. https://doi.org/10.1021/ci600342e

5. Perez-Nueno VI, Rabal O, Borrell JI, Teixido J (2009) APIF: A new interaction fingerprint based on atom pairs and Its application to virtual screening. Journal of Chemical Information and Modeling 49:1245–1260. https://doi.org/10.1021/ci900043r

6. Jasper JB, Humbeck L, Brinkjost T, Koch O (2018) A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening. Journal of Cheminformatics 10:1–13. https://doi.org/10.1186/s13321-018-0264-0

7. de Graaf C, Kooistra AJ, Vischer HF, et al (2011) Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H 1 Receptor. J Med Chem 54:8195–8206. https://doi.org/10.1021/jm2011589

8. Rodríguez-Pérez R, Miljković F, Bajorath J (2020) Assessing the information content of structural and protein–ligand interaction representations for the classification of kinase inhibitor binding modes via machine learning and active learning. J Cheminform 12:36. https://doi.org/10.1186/s13321-020-00434-7

9. Mpamhanga CP, Chen B, McLay IM, Willett P (2006) Knowledge-based interaction fingerprint scoring: A simple method for improving the effectiveness of fast scoring functions. Journal of Chemical Information and Modeling 46:686–698. https://doi.org/10.1021/ci050420d

10. Kokh DB, Doser B, Richter S, et al (2020) A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. Journal of Chemical Physics 153:. https://doi.org/10.1063/5.0019088

11. Sato T, Honma T, Yokoyama S (2010) Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. Journal of Chemical Information and Modeling 50:170–185. https://doi.org/10.1021/ci900382e

12. Da C, Kireev D (2014) Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. Journal of Chemical Information and Modeling 54:2555–2561. https://doi.org/10.1021/ci500319f

13. Wójcikowski M, Kukiełka M, Stepniewska-Dziubinska MM, Siedlecki P (2019) Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. Bioinformatics 35:1334–1341. https://doi.org/10.1093/bioinformatics/bty757

14. Radifar M, Yuniarti N, Istyastono EP (2013) PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting. Bioinformation 9:325–328. https://doi.org/10.6026/97320630009325

15. Salentin S, Schreiber S, Haupt VJ, et al (2015) PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res 43:W443–W447. https://doi.org/10.1093/nar/gkv315

16. Jubb HC, Higueruelo AP, Ochoa-Montaño B, et al (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. Journal of Molecular Biology 429:365–371. https://doi.org/10.1016/j.jmb.2016.12.004

17. Istyastono EP, Radifar M, Yuniarti N, et al (2020) PyPLIF HIPPOS: A Molecular Interaction Fingerprinting Tool for Docking Results of AutoDock Vina and PLANTS. J Chem Inf Model 60:3697–3702. https://doi.org/10.1021/acs.jcim.0c00305

18. Adasme MF, Linnemann KL, Bolz SN, et al (2021) PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Research gkab294. https://doi.org/10.1093/nar/gkab294

19. Landrum G, Tosco P, Kelley B, et al (2021) rdkit/rdkit: 2021_03_2 (Q1 2021) Release. Zenodo

20. Gowers R, Linke M, Barnoud J, et al (2016) MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Austin, Texas, pp 98–105

21. Wójcikowski M, Zielenkiewicz P, Siedlecki P (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. J Cheminform 7:26. https://doi.org/10.1186/s13321-015-0078-2

22. Hajiebrahimi A, Ghasemi Y, Sakhteman A (2017) FLIP: An assisting software in structure based drug design using fingerprint of protein-ligand interaction profiles. Journal of Molecular Graphics and Modelling 78:234–244. https://doi.org/10.1016/j.jmgm.2017.10.021

23. Reback J, McKinney W, Jbrockmendel, et al (2021) pandas-dev/pandas: Pandas 1.2.4. Zenodo

24. Rodríguez-Espigares I, Torrens-Fontanals M, Tiemann JKS, et al (2020) GPCRmd uncovers the dynamics of the 3D-GPCRome. Nat Methods 17:777–787. https://doi.org/10.1038/s41592-020-0884-y

25. Ballesteros JA, Weinstein H (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods in Neurosciences 25:366–428. https://doi.org/10.1016/S1043-9471(05)80049-7

26. Kooistra AJ, Mordalski S, Pándy-Szekeres G, et al (2021) GPCRdb in 2021: integrating GPCR sequence, structure and function. Nucleic Acids Research 49:D335–D343. https://doi.org/10.1093/nar/gkaa1080

27. Wang C, Jiang Y, Ma J, et al (2013) Structural Basis for Molecular Recognition at Serotonin Receptors. Science 340:610–614. https://doi.org/10.1126/science.1232807

28. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Research 32:W665–W667. https://doi.org/10.1093/nar/gkh381

29. Katritch V, Cherezov V, Stevens RC (2013) Structure-Function of the G Protein–Coupled Receptor Superfamily. Annu Rev Pharmacol Toxicol 53:531–556. https://doi.org/10.1146/annurev-pharmtox-032112-135923

30. Weis WI, Kobilka BK (2018) The Molecular Basis of G Protein–Coupled Receptor Activation. Annu Rev Biochem 87:897–919. https://doi.org/10.1146/annurev-biochem-060614-033910

31. Han DS, Wang SX, Weinstein H (2008) Active State-like Conformational Elements in the β_2 -AR and a Photoactivated Intermediate of Rhodopsin Identified by Dynamic Properties of GPCRs. Biochemistry 47:7317–7321. https://doi.org/10.1021/bi800442g

32. Fritze O, Filipek S, Kuksa V, et al (2003) Role of the conserved NPxxY(x)5,6F motif in the rhodopsin ground state and during activation. Proceedings of the National Academy of Sciences 100:2290–2295. https://doi.org/10.1073/pnas.0435715100

33. Robinson PR, Cohen GB, Zhukovsky EA, Oprian DD (1992) Constitutively active mutants of rhodopsin. Neuron 9:719–725. https://doi.org/10.1016/0896-6273(92)90034-B

34. Flock T, Hauser AS, Lund N, et al (2017) Selectivity determinants of GPCR–G-protein binding. Nature 545:317–322. https://doi.org/10.1038/nature22070

35. Venkatakrishnan AJ, Deupi X, Lebon G, et al (2013) Molecular signatures of G-protein-coupled receptors. Nature 494:185–194. https://doi.org/10.1038/nature11896

36. Flock T, Ravarani CNJ, Sun D, et al (2015) Universal allosteric mechanism for Gα activation by GPCRs. Nature 524:173–179. https://doi.org/10.1038/nature14663