Master's Degree in Biomedical Engineering

Molecular mechanics-driven comparative analysis of protein-ligand binding pockets to investigate the relationship between taste perception and metabolic pathways

Supervisor
Prof. Marco Agostino Deriu

Candidate
Xhesika Hada

Co-supervisor
Eric Adriano Zizzi

Academic year 2020/2021

Abstract

The sense of taste in mammals, including humans, is a complex natural mechanism that acts as a sentinel system, allowing for a quick recognition of chemicals that enter the oral cavity and the discrimination between healthy and nutritious food and substances potentially toxic or dangerous to health. This crucial ability is enabled by the key players of the sense of taste, i.e. taste receptors: these are highly specialized proteins that play the role of molecular switches for specific cellular signalling pathways, ultimately resulting in the perception of the five basic tastes: sweet, bitter, umami, salty and sour. Moreover, several studies have shown that the localization of taste receptors is not limited to the oral cavity only, but is rather widespread throughout the human body, including the gastrointestinal tract and the central nervous system, where a direct sensation of taste is not evoked. In fact, in such tissues they are thought to serve a different set of functions revolving around nutrition and food absorption. In this context, where the interaction between taste receptors and tastants is not exclusive to the gustatory system alone, the present work is focused on a high-level analysis, starting from the molecular-level characterization of the binding site of the human sweet taste receptor, a class-C GPCR, in complex with its tastant agonist sucrose, to perform a similarity search to scan for the conserved tastant-binding site residues in the currently solved proteome, with the goal of shedding light on the putative function of food molecules within domains and pathways that are external to the gustatory system. In order to identify the proteins of greatest interest and significance, two successive filtering steps were performed, the first one relying on Solvent-Accessible Surface Area (SASA) calculations to extract proteins with an exposed-surface binding site with a shape similar to that present in the sweet receptor; the second step consisted in Molecular Docking calculations of sweet ligands to the proteins extracted from the first step, to restrict the analysis to those with binding affinity values above a fixed threshold. Such proteins were analyzed from a proteomic perspective, and it was observed that most of them are involved in enzymatic activities, e.g. hydrolases, oxidoreductases, or transferases, as well as in various biological processes that range from nucleocytoplasmic transport, biosynthesis of nitrogenous bases, up to post-translational modifications that control protein activity. The results obtained regarding the conservation of the binding site, the nature and role of the candidate proteins discovered, and the methodologies and platforms adopted in the present work can be used as a basis for future studies focused on the use of tastants and corresponding receptors as models for the engineering of drugs for the treatment of food-related diseases and disorders.

1. Introduction

The current chapter is a general introduction to the master's thesis work, framing it within its biological background and describing the rationale, objectives and organization of the present research.

Mammals are dependent on food as sources of energy and biochemical building blocks, and two of our major senses, taste, and olfaction, are involved in or even specifically devoted to distinguishing beneficial food from potentially harmful substances such as toxins [1]. Our sense of taste allows us to instantly evaluate the nutrional value of food before ingesting it, and it appears entirely reasonable that this ability has been crucial for the survival of our species, in a context where nourishing food sources were sparse[2]. However, in present-day society, several aspects revolving around the relationship between food and its taste are worth discussing. For instance, those who live in a condition of food scarcity or in contexts where food safety cannot be guaranteed must often times indeed rely on taste to identify nutritious food to eat and discriminate between safe and unsafe substances; conversely, those who live in an environment of abundant, palatable and mostly safe food are driven by taste to overconsume calorically dense foods, which may result in overnutrition-related diseases, such as diabetes and obesity [2,3]. For this reason, taste perception continues to play an important role in human health even at the present day, for diverse reasons, from avoiding spoilt food or toxins all the way to the prevention and treatment of conditions stemming from overnutrition [2]. An example in today's context, which also underlines the commercial importance of the role of taste, can be found in the deliberate addition of bitter substances in toxic products normally used in cleaning such as detergents or degreasers. This is done to prevent and discourage the ingestion of these substances by children who are more likely to prefer sweet foods than adults.

The mechanisms of nutrient sensing in the oral cavity have been well characterized and involve lingual taste receptors, which are chemosensors capable of detecting chemical stimuli from food such as sugars, amino acids, poisons, acids, and minerals and building up to the final taste perception. In general, sweet, salty, umami, sour and bitter are basic taste qualities liked to the underlying chemical nature of tastants: sweet is generally tied to carbohydrates, salty is a characteristic of minerals and ions, bitter is instinctively associated to harmful compounds, sour to spoiled food, and umami is the taste of protein and amino acid content. Additionally, there is emerging evidence that lipids can be detected by fatty acid receptors on taste cells, leading to the development of a sixth taste quality [4,5] . As can be easily understood, taste is the evolutionary

adaptation to the need of ingesting substances with specific nutritional qualities, while at the same time avoiding dangerous ones. Several studies have revealed that just as the tongue, the gastrointestinal system is also equipped with taste receptors and similar signaling pathways that give it the ability to sense nutrients and toxins. Furthermore, intestinal chemosensation is important as it evokes neuronal and hormonal responses following the ingestion of food that guide food metabolism, particularly satiety and energy homeostasis [5].

For all the aforementioned reasons, given the importance of taste receptors in the ingestion and processing of food, this work focuses on investigating the role of taste chemosensors holistically, starting from a molecular-level characterization of the binding site of the receptors involved in the sense of taste, followed by a similarity search of conserved binding-site residues in the currently solved proteome, carried out to elucidate the putative role of tastants across a variety of biochemical pathways even beyond the gustatory system.

With this overarching intent, the main purpose of this work is to investigate the molecular characteristics of the binding site of a human sweet taste receptor models and the sweet taste molecules that bind to it, as several studies show how sweet taste pathways are also present in the intestine and central nervous system (CNS).

The work is organized as follows:

Chapter 1 is the present introduction.

Chapter 2 provides a biological background on taste receptors and their characteristics, with a greater focus on the sweet receptor. Then their role in nutrition and metabolism is described and discussed.

Chapter 3 describes the methods used in the present work. An initial description of Molecular Modeling is presented followed by a theoretical discussion of Molecular Mechanics. The Solvent Accessible Surface Area (SASA) Calculation and the Molecular Docking algorithms are then presented and described in detail.

Chapter 4 describes the proposed pipeline for carrying out the similarity search of the sweet receptor binding site across all currently solved protein structures. After discussing the state of the art of similarity search algorithms and an overview of their underlying methodologies, the workflow implemented in the present work is described. Next, the final results are presented, of which an analysis from a proteomic perspective is mentioned, discussing the conservation of the binding site

and the nature and role of the discovered candidate proteins. Finally, possible future perspectives of the methodology developed and described in the present work are proposed.

2 Biological Background

The sense of taste plays a pivotal role in nutrition, as it is a means for humans and other mammals to evaluate the quality of food, to promote the consumption of foods of high nutritional content. Appetitive and pleasureable taste stimuli increase the consumption of nutrients while stimuli of repulsion discourage the ingestion of potential toxins. The five basic taste qualities reflect just this role:

- sweet taste allows for identifying sugary foods that are thought to have been essential for human survival
- umami taste allows to identify of foods containing proteins and therefore promotes the intake of amino acids that are important for the whole organism (e.g. essential amino acids)
- bitter taste allows for limiting the number of toxins consumed because it is an adaptive taste and at high intensity gives an aversive response, typically linked to nausea
- salty taste identifies ions and minerals, mainly sodium, which is involved in numerous functions including maintaining blood volume or membrane potential in cells
- sour taste indicates the presence of acids, and like the bitter taste it produces aversive responses at high intensities and appetitive responses at moderate-low intensities. Sour taste has been linked throughout evolution to spoiled food

Taking the nutritional context of today's industrialized nations into consideration, one of the biggest challenges is represented by the shift from undernutrition to overnutrition, due to the easy accessibility of highly caloric foods. This condition in many cases leads to the development of overnutrition-related diseases, such as obesity, type 2 diabetes mellitus, or fatty liver disease. This is one of the reasons why sweet taste plays a crucial role in human health[3].

2.1 Gustatory system: anatomy and function

The five basic taste sensations that humans can perceive are divided into sweet, umami, bitter, salty and sour. The perception of taste is a consequence of the interaction at the molecular level between food-related chemicals and the taste buds, which are sensorial organs dedicated to the sense of taste and are found on the tongue. Taste buds contain taste receptor cells (TRCs), also known as gustatory cells.

From an anatomical point of view, the gustatory systems presents a hierarchical structure, with the papillae , small protuberances on the upper part of the tongue that can be seen with the naked eye, at

the highest level; the papillae in turn contain the taste buds which, at the smaller scale, are composed of the Taste Receptor Cells.

Lingual papillae possess both a sensorial and a mechanical function, as their specific conformation maximizes the surface area of the tongue to increment the contact area between food and tongue and therefore the friction between the two.

The papillae are not present in equal numbers on the surface of the tongue and are arranged in a specific topological pattern, which is highlighted in *Figure 1* .

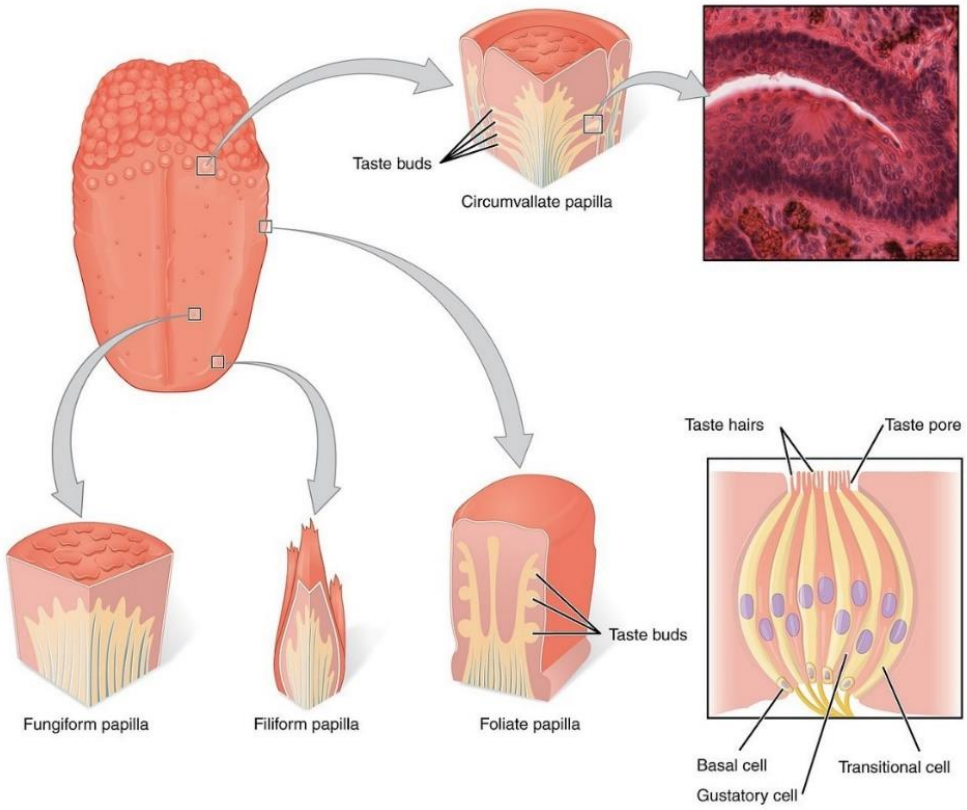


Figure 1: Localization of lingual papillae. Circumvallate, back of the tongue; foliate, side of the tongue, fungiform, middle and front of the tongue, and filiform, front of the tongue. A single taste bud is represented in the lower right corner.[6]

From a morphological perspective, the lingual papillae are classified as circumvallate, fungiform, foliate and filiform. All types contain taste buds, except for the filiform papillae, which have a more mechanical function, acting as grips to increase friction on the tongue.

Circumvallate papillae (or vallate papillae) are present in the smallest number, which can vary between 10 and 14, and are found in the posterior part of the oral tract of the tongue, particularly just in front of the terminal sulcus of the tongue, and are arranged in an arch shape. They are larger than the other lingual papillae, have a ring-like structure, contain many taste buds located mostly in the lateral surface of the papilla, and are innervated by a cranial nerve (*Figure 2, C*).

Foliate papillae are slightly higher in number than the circumvallate papillae, located along the lateral edges of the tongue, in the posterior part. They consist of parallel rows of ridges and grooves. Although human foliate papillae are less developed than the foliate papillae of other animals, they contain a high number of taste buds. The morphology and arrangement of human foliate papillae differ from individual to individual and change with age (*Figure 2*, *B*).

Fungiform papillae occupy the anterior part of the tongue and are mainly concentrated in the lateral areas and median sulcus. In addition, their distribution varies greatly with age [7], and they contain fewer taste buds than the foliate and circumvallate papillae. As the name suggests, they have an overall mushroom-like structure, with a narrow base and a bloated top, but can vary in appearance. The fungiform papillae can be seen with the naked eye as the typical red dots found on the tongue, a characteristic deriving from the fact that they are highly perfused (*Figure 2, A*).

Filiform papillae, the most abundant type, have a conical morphology and are distributed over the entire dorsal surface of the tongue. They are non-gustatory papillae, not containing any taste buds, and have two main functions: one mechanical, giving roughness to the tongue to enhance the retention of food, and the other sensorial, providing information about pain and temperature (*Figure 2 , A*).

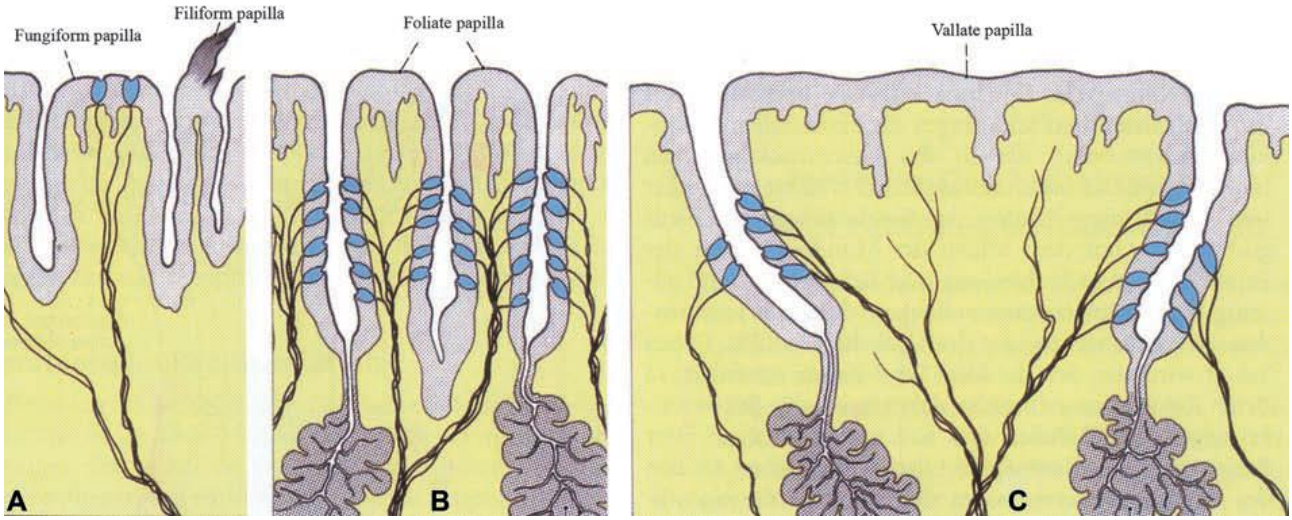


Figure 2: Schematic drawings of lingual papillae. (A) Fungiform papilla on the middle and filiform Papilla on the right. (B) Foliate papilla, (C) Vallate papilla. In figure B and C the taste buds are in the lateral epithelium wall of the papilla.[8]

Three of the four papillae, i.e., circumvallate, foliate, and fungiform, are covered by non-keratinized stratified squamous epithelium, while the filiform papillae differ in that they have partially keratinized stratified squamous epithelium which is what confers the typical rough surface of the tongue.

The circumvallate papillae receive innervation from the N IX cranial nerve, the glossopharyngeal nerve. The remaining anterior two-thirds of the tongue, where the other papillae are located, receive innervation from the chorda tympani nerve, a branch of the large facial N VII nerve [9].

2.2 Taste buds: anatomy and physiology

Inside the papillae, under the keratinous layer, reside the taste buds, neuroepithelial receptor cells which initiate taste signaling. In the human oral cavity there are about 5000 taste buds, mainly located on the tongue and palate, but also present to a lesser extent on the epiglottis, pharynx, and larynx [9].

The structure of taste buds can be described as garlic-bulb-like shaped, with a taste pore, i.e., a small epithelial hole at the end that communicates with the external environment. The basic units in this structure are the TRCs which are assembled in groups of about 80-100 cells. They are electrically active epithelial cells that can depolarize and release neurotransmitters. The latter are the chemical messengers through which taste receptor cells communicate with neighbouring neurons [1,5]. Each taste bud in adults is innervated by sensory ganglion neurons, from neuronal axons that branch and penetrate the taste bud.

TRCs are classified into four different types: type I, II, III cells, and taste cell precursors (sometimes referred to as type IV cells). Each taste bud contains TRCs belonging to the four categories regardless of their anatomical location. The different types of cells and the taste bud are shown in *Figure 3* and *Figure 4*.

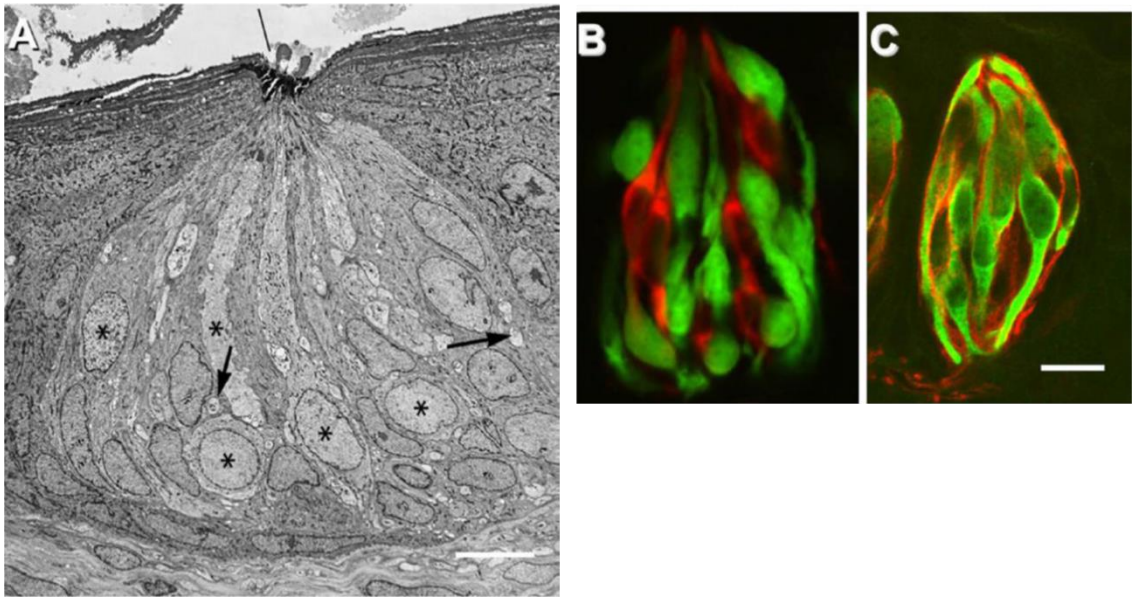


Figure 3 : Cell types and synapses in the taste bud. (A) Electron micrograph of a rabbit taste bud where the arrows indicate nerve profiles. Asterisks indicate Type II cells. (B), A taste bud from a mouse. The tissue is from a transgenic mouse expressing GFP only in Receptor (Type II) cells (green). The red cells represent presynaptic cells (immunostained) (C) Image of taste bud obtained by immunostaining, where the Type I cells are visible. [9]

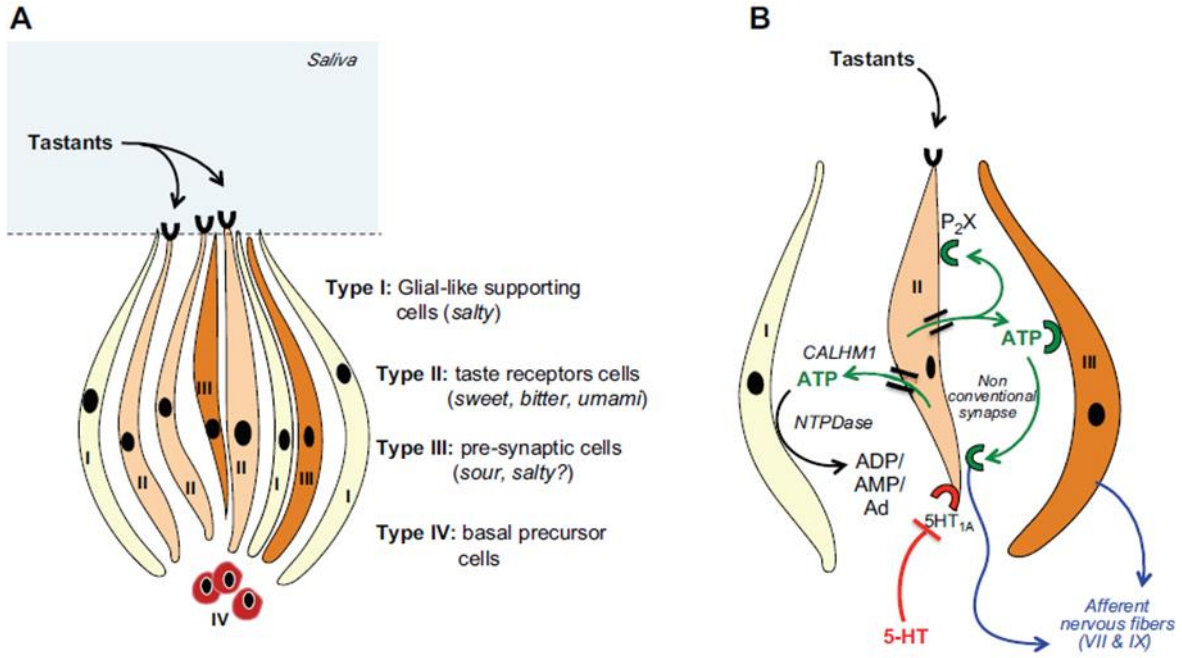


Figure 4: Representation of the different types of cells inside the taste buds. A) Shows the roles of the different cell types. B) Shows the role of ATP, as neurotransmitter, in the communication between the taste bud cells.[10]

TRCs are continuously replaced, on average every 10 days to compensate for mechanical, thermal, or toxin-induced damage to the gustatory epithelia[10]. In addition, the entire gustatory papilla as well as the gustatory epithelium, i.e. the taste organ, have the ability of regenerating completely upon destruction or removal, making them one of the very few organs in humans with a complete regenerative capability.

Type I cells are the most abundant type, and are believed to have glial-like functions, such as promoting synaptic transmission and limiting the spread of transmitters, mainly ATP and glutamate. They have also been suggested to transduce salty tasty through epithelial sodium channels (ENaC), although the precise mechanism is still not clear [1]. Type I cells have an irregularly shaped nucleus and they are usually characterized by cytoplasmatic lamellae that enwrap other taste cells within the taste bud. Moreover, they express NTPDase2, a nucleotidase that degrades ATP released by other taste cells via hydrolyzation, and GLAST, a transporter for glutamate . They also express ROMK, a potassium channel, capable of eliminating the K+ that accumulates in the extracellular space after the propagation of the action potential. Accumulation of K+ can affect the excitability of type II and III cells and decrease it [11].

Type II cells represent more than thirty percent of the gustatory system cells. Compared to type I cells, they have a more regular shape and are also called receptor cells, because they express G-protein coupled receptors (GPCR) that bind sweet, bitter, and umami taste compounds. Each cell responds mainly to a single tastants, i.e. they can react to only bitter or only sweet stimuli, not to both [12]. Type II cells form synapses with afferent sensory fibres: the binding of tastants starts the secretion and the release of ATP through pannexin 1 (Panx1) hemichannels and the consequent sensing transduction on the afferent fibres. ATP receptors also mediate the communication between type II and type III cells.

Type III cells are the least numerous type compared to the first two subclasses. They are presynaptic cells and are the only ones in the taste buds to form conventional neuronal synapses, i.e., synaptic junctions, with nerve terminals. The main neurotransmitters produced are serotonin (5-HT), acetylcholine, γ-aminobutyric acid (GABA), and noradrenaline (NE). When cells are depolarized, they release neurotransmitters through vesicles [1]. They don't express GPCRs but are involved in the perception of sour taste through the expression of polycystic kidney disease 2-like 1 protein (PKD2L1) and polycystic kidney disease 1-like 3 protein (PKD1L3) channels [1] and they respond also to carbonation [2].

Taste cell precursors have a spherical or ovoidal shape They constitute a heterogeneous group of cells that can be quiescent precursor cells or immature taste cells present at the base of the taste buds, and they are not directly involved in taste transduction.

To summarize, type I cells appear to be supportive glial-like cells. Type II cells play the role of sensory cells for sweet, bitter, or umami tastes via GPCRs. The binding to the tastants activates the receptor cells that secrete ATP which in turn excites the afferent sensory fibres and adjacent type III

cells. The latter can be stimulated either directly by the acid/sour taste molecules or indirectly by the sweet, bitter, and umami tastants. Therefore more gustatory stimuli are integrated into the papillae, or to be more specific in taste buds. Furthermore, the presynaptic cells release serotonin, noradrenaline, and GABA which allow the taste signal to be transmitted to the brain stem along cranial nerves VII and IX [1,5,9]. Serotonin and GABA transmitters also inhibit type II cells. The ATP itself secreted by type II cells exerts autocrine feedback on the cells themselves. The lack of this positive feedback was observed in-vivo on mice in which some purinoreceptors were absent, leading to a reduction in the taste-evoked secretion of ATP and cessation of the normal operation of taste buds [9].

2.3 *Signalling mechanism*

The transduction pathway of sweet, L-glutamate (umami), and bitter stimuli operated through type II taste receptor cells are represented schematically in *Figure 5*.

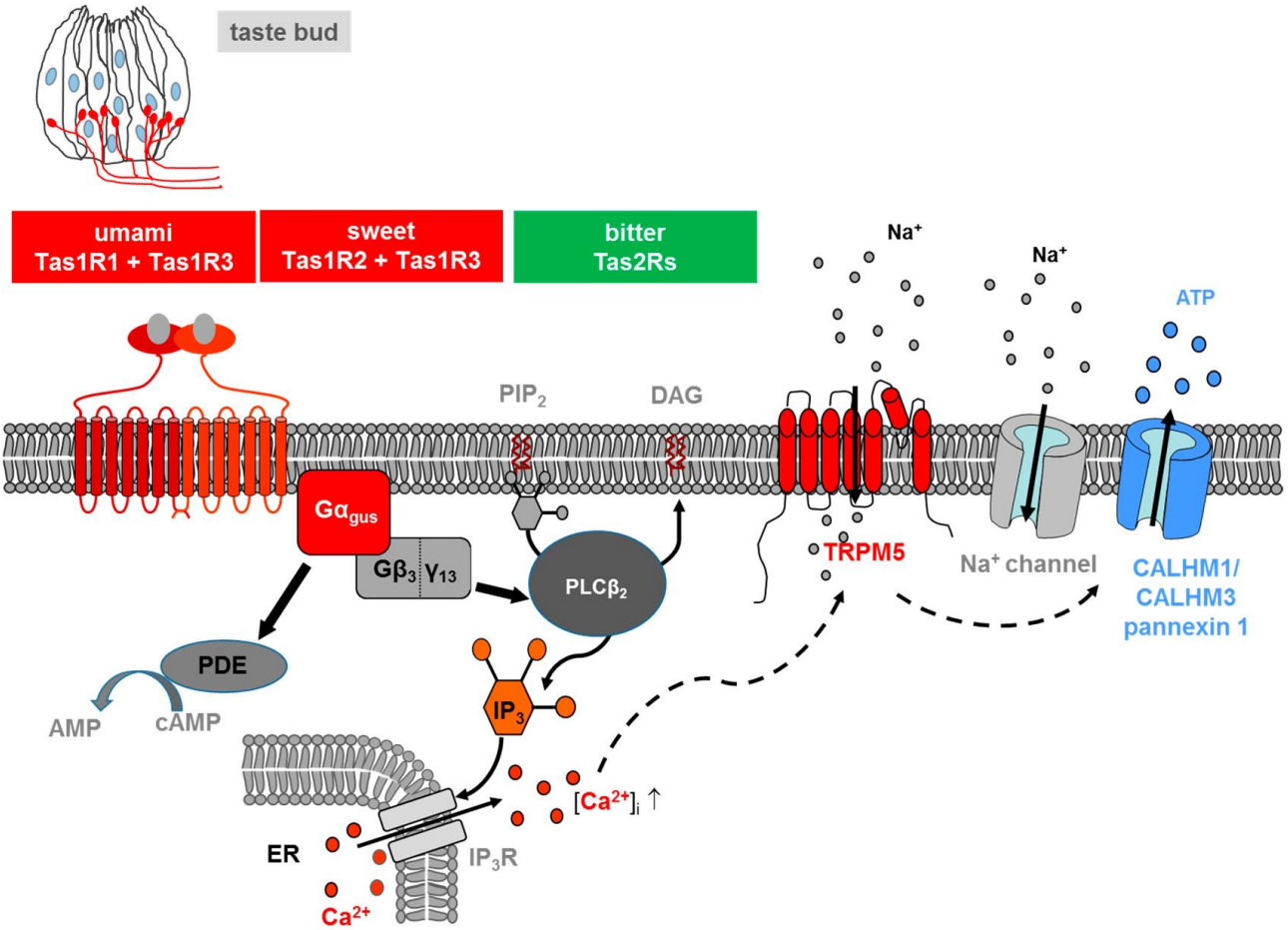


Figure 5: Signalling pathway on type II cells[13]

First, sweet and umami tastants are detected by GPCRs of the C family (T1Rs) that assemble into heterodimeric complexes, i.e., T1R2+T1R3 (sweet) or T1R1+T1R3 (umami), while bitter tastants are detected by A-family GPCRs (T2Rs) [14]. The binding to the tastants activates taste GPCRs which in turn stimulate a series of downstream enzymes and effectors in receptor cells. In all these taste stimuli, sweet, umami or bitter, the starting point of the transduction pathway is the activation of a trimeric G protein formed by α-gustducin (Gαgus) and a complex of Gβ3 and Gγ13 (Gβ3/γ13).

The complex Gβ3/γ13 then activates a membrane-bound enzyme, phospholipase C isoform β2 (PLCβ2) that catalyzes the digestion of plasma membrane phospholipids into inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG). IP3 is a second messenger, because by spreading in the intracellular space and binding to the IP3 receptors (IP3R), it allows the release of Ca^{2+} in the cytosol, as the IP3Rs are ion channels present in intracellular stores of calcium[13,14]. This is the canonical PLC/IP3 Ca^{2+} release signalling pathway[9].

The release of intracellular Ca^{2+} (dashed lines in *Figure 5*) has two consequences:

- the cell membrane is depolarized because calcium ions bind and activate the Transient Receptor Potential Melastatin 5 (TRPM5) ion channels, allowing Na^{+} ions to enter through voltage-gated sodium channels
- at the same time, the calcium ions bind to the pannexin1 gap junction hemichannels and the calcium homeostasis modulator channel (CALHM), formed by CALHM1 and CALHM3, and activates them (here the depolarization of the membrane also comes into play)

The result is a release of ATP from the cytosol to the extracellular environment.

As for the Gαgus (left of Figure 5), its activity is linked to the PDE-3 phosphodiesterase whose task is to degrade the second messenger cyclic-AMP (cAMP) to AMP. This cyclic nucleotide can inhibit the PLCβ2/IP3 pathway, with its degradation therefore this ability is blocked.

In summary, two pathways coexist in type II cells and are activated simultaneously, which ultimately results in the release of ATP in the taste bud.

2.4 Five basic taste qualities

As mentioned in the previous sub-chapter, sweet, bitter and umami receptors belong to the GPCR superfamily, while sour and salty tastants are identified by ion channels.

GPCRs share two common features: the first is the presence of seven transmembrane (TM) α-helices, and the second is the coupling to heterotrimeric G-proteins of the intracellular domain of the receptor. The seven α-helices are joined by six loops, three of them are extracellular and the remaining three are intracellular. The superfamily is classified into seven classes from A to F based on physiological and structural features[15]. The ones of interest for taste transduction are class A and class C as they are respectively the classes to which the sweet and umami receptors and the bitter receptor belong:

- Class A (Rhodopsin-like) is the largest family of GPCRs, including hormone, neurotransmitter, and light receptors. Receptors of class A have a small extracellular domain and they work as monomers.
- Class C (Metabotropic glutamate receptors) feature seven transmembrane α-helices with a large extracellular N-terminal domain (ATD), where ligands bind. Another important aspect of this class is the ability to form constitutive dimers, with homo- and heterodimerization, i.e. the association between two identical or non-identical proteins, respectively, depending on the members, which is mandatory for functionality. In addition to the sweet (T1R2+T1R3) and umami (T1R1+T1R3) receptors, there are eight metabotropic glutamate (mGlu) receptors, two heterodimeric-aminobutyric acid ($GABA_B$) receptors, a calcium-sensing receptor (CaS), a promiscuous L-α-amino acid receptor (GPRC6A), and some orphan receptors (*Figure 6*)[16–18].

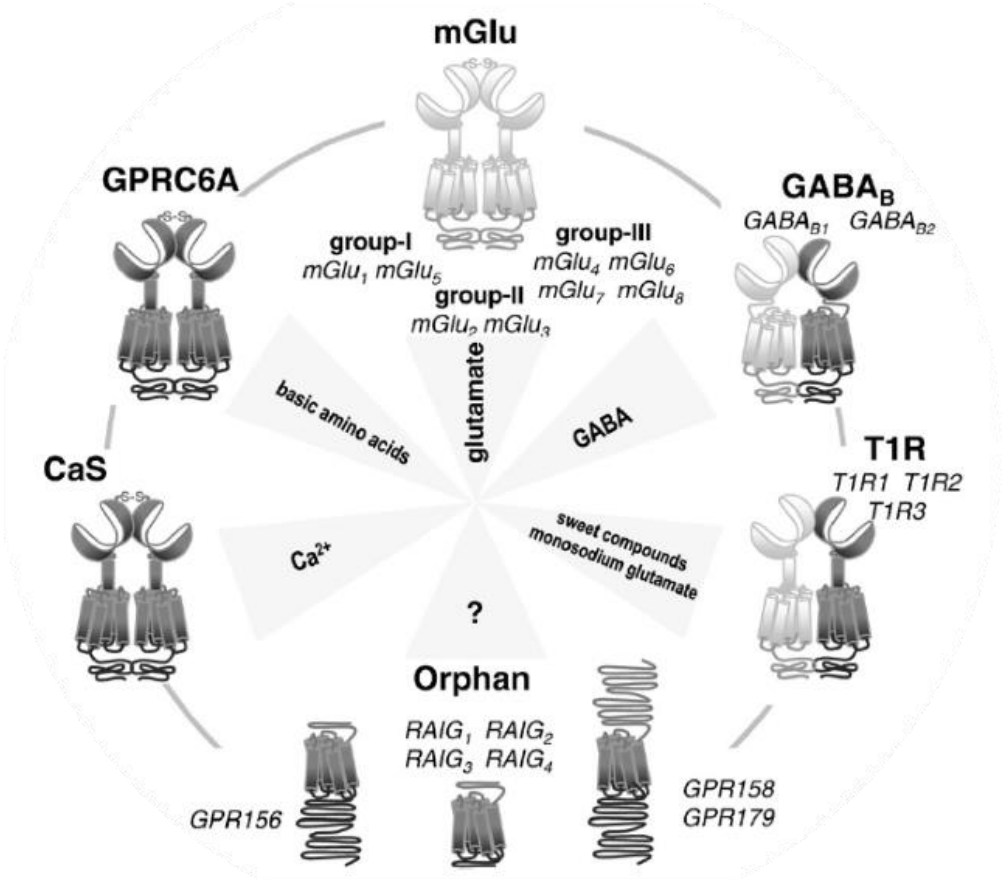


Figure 6: Schematic representation of Class C GPCR family. Their corresponding agonist is shown at the centre of the image. Heterodimers are represented with light grey and dark grey protomers[16]

On the other hand, the signalling mechanism of the ion channels involved in the perception of salty and sour taste is still unclear. The perception of sour taste is known to be related to the presence of H^+ ions. As mentioned before, the perception of sour tastants is thought to be mediated by PKD1L3 and PKD2L1 ion channels, while the perception of salt tastants is thought to be mediated by the ENaC ion channel.

Figure 7 contains a schematic summary of the receptors for the five basic qualities.

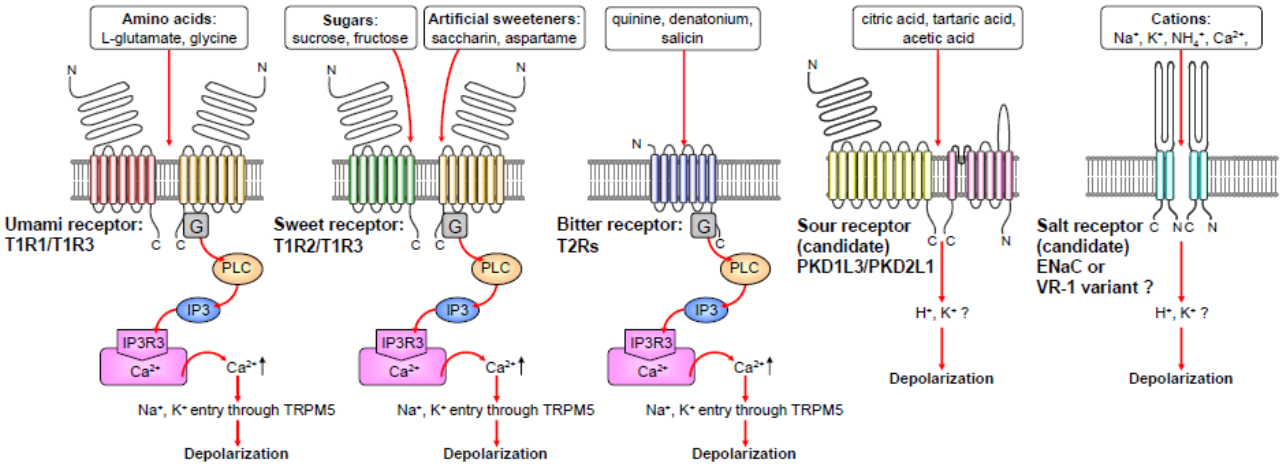


Figure 7: Schematic representation of the receptors of the five basic taste senses, from left to right umami, sweet, bitter, sour and salty, and pathway of stimulus transduction [19]

2.4.1 Sweet

The receptor that binds sweet tastants belongs to the GPCRs of class C. It is a heterodimer composed of T1R2 and T1R3. It is a transmembrane receptor, composed of an intracellular domain, a transmembrane part that consists of seven transmembrane helices (TM), and an extracellular N-terminus domain, consisting of a Venus flytrap (VFT) and cysteine-rich domains (CR) (*Figure 8*).

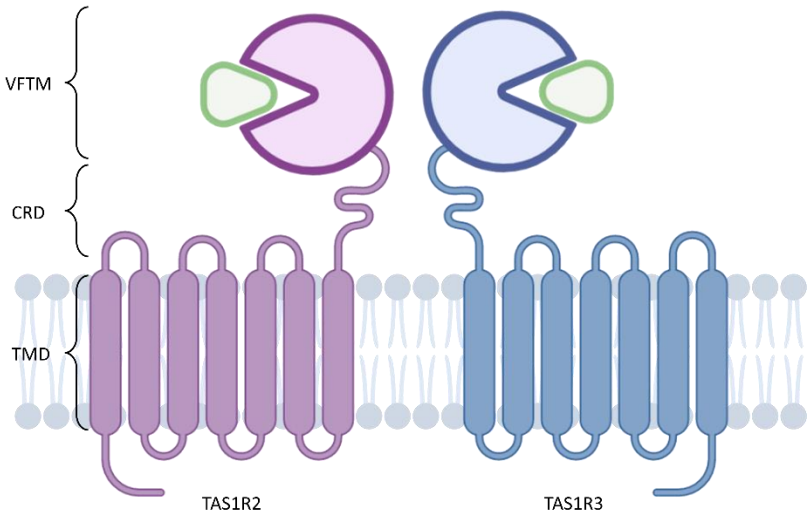


Figure 8: Sweet taste receptor schematic structure and orthosteric binding sitess in the Venus flytrap module (VFTM).[20]

The main active site of the sweet receptor, referred to as the orthosteric binding site, is found in the VFT domain, which can assume an open or closed configuration, depending on whether the ligand is unbound or bound, and remains in equilibrium between the two conformations in the absence of the ligand.

The conformation in which the VFT domain is found affects the functionality of the receptor. Since the receptor is made up of two subunits, both VFT domains are considered in defining the conformation: in the *open/open* conformation the receptor is inactive, while in the *closed/open* conformation the receptor is active. The ligand binds mainly in the open VFT and promotes the closed-form as the interactions between the two lobes present in the VFT domain stabilize this form. The active form of the receptor is also characterized by a rearrangement of the VFT dimer in which the relative orientation of the two domains changes[16].

The orthosteric binding site can recognize a series of compounds belonging to different classes and with different molecular weights:

- natural small sugars. such as glucose, sucrose, fructose, and sugar alcohols
- glycosides, such as stevioside and glycyrrhizin
- D-amino acids, such as D-tryptophan and D-phenylalanine
- artificial chemical compounds, such as sucralose, aspartame, neotame, saccharin, and cyclamate

The binding of sweeteners to the VFT domain leads to the subsequent activation of the TM domain and the intracellular region, and this leads to the activation of the whole receptor[16,17,21]. It should be noted that in taste receptors, only one subunit (T1R1 or T1R2) in the VFT dimer is responsible for binding the sugar compound[16]. Furthermore, as some studies have shown, some sweeteners such as stevioside and aspartame only bind to the VFT domain of the T1R2 subunit, while glucose and sucrose bind to the VFT domain of both subunits[21]. In addition to the orthosteric binding site, in the transmembrane region of both receptor subunits, there are allosteric sites to which positive allosteric modulators (PAMs) can bind, i.e. taste-free ligands with high binding selectivity to the sites in question. Their function is to increase the activity of agonists, i.e. sweet molecules that bind to the orthosteric site, by affecting the spatial conformation of the receptor.

2.4.2 Umami

Umami stands for delicious or savoury taste, it is a Japanese word used to describe the savoury taste of amino acids or oligopeptides. It was first used at the beginning of the 1900s by Dr Kikunae Ikeda, a Japanese chemist who discovered this taste sensation evoked by glutamic acid[22]. As the name suggests these receptors can enhance the palatability of food, to stimulate in humans the ingestion of safe foods, that is, foods containing fundamental nutritional resources.

Like sweet receptors, the receptors that can identify umami tastants are heterodimers, composed of T1R1 and T1R3 subunits belonging to class C GPCRs. The structure is also the same: large extracellular VFT domains and TM domains linked to the previous domains via CR domains (*Figure 9*) [23,24].

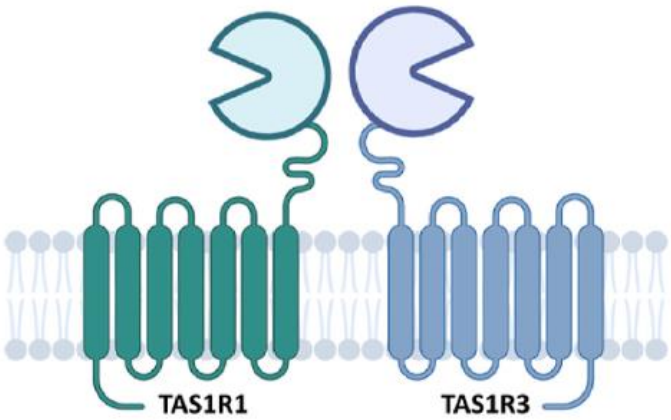


Figure 9: Umami taste receptor schematic structure.[20]

In humans, generally, the umami taste receptor is activated by monosodium L-glutamate (MSG), but can also be activated by amino acids like aspartate or even by organic acids like lactic, succinic, and propionic acids. The taste sensation can be increased by the binding to the orthosteric binding site of esters such as guanosine 5'monophosphate (GMP) and inosine 5'-monophosphate (IMP).

2.4.3 Bitter

The receptors that can identify bitter tastants are part of the T2Rs family of GPCRs. They have a similar function to that of class-A GPCRs, however, they don't share any remarkable sequence similarity. If compared to the T1Rs, the T2Rs family is quite numerous, containing close to 25 different receptors in humans, compared to just 3 for the T1Rs[25]. Interestingly, the number of receptors belonging to the T2Rs family varies from species to species, reasonably reflecting the differences in taste perception, nutritional needs and hazardous substances in different mammals.

T2Rs consist of a short extracellular N-terminus and intracellular C-terminus and seven transmembrane helices (TM). The binding site is located in the transmembrane part of the receptors (*Figure 10*) and the bitter compounds that can bind to it can be structurally very different, due to the different recognition mechanisms exhibited by these receptors.

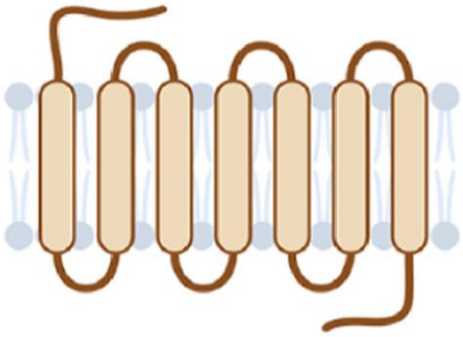


Figure 10: Bitter taste receptor schematic structure[20]

2.4.4 Salty

One of the transduction mechanisms proposed for the perception of salty taste is epithelial sodium channels (ENaC), which are highly selective Na^+ ion channels. Other mechanisms proposed for salty transduction include the transient receptor-potential channel, subfamily V, member 1 (TRPV1) and the transient receptor-potential cation channel, mucolipin subfamily, member 3 (TRPML3).

From a structural point of view, ENaC is made up of three homologous subunits α, β and γ, of which the α subunit is involved in the activity of the channel while the β and γ subunits are involved in the expression of the channel on the cell surface. There are also extracellular loops, that are needed for the channel function, and intracellular N- and C- termini. The C-terminus, in particular, is rich in proline. (*Figure 11*)

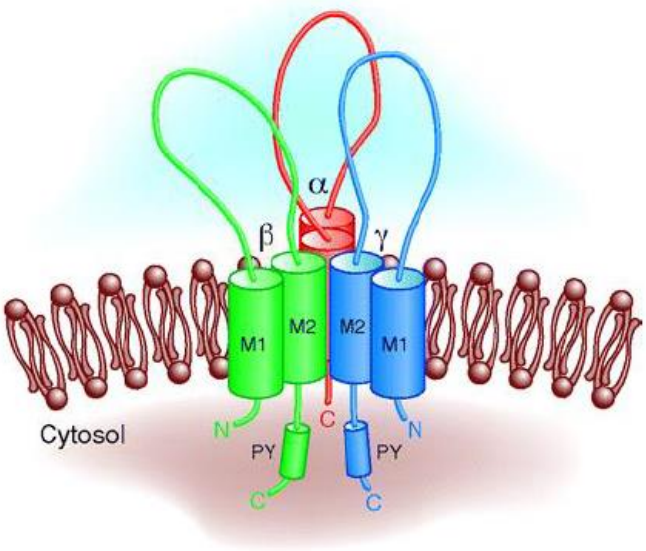


Figure 11: Epithelial sodium channel (ENaC) schematic subunits structure: α ENaC (red), β ENaC (green), and γ ENaC[26]

In the perception of salty taste, two outcomes may arise based on the concentration of the salty tastant. At high concentrations, the response obtained is repulsion, i.e. negative, while at lower concentrations, the response obtained is pleasure, i.e. positive.

2.4.5 Sour

The receptors that can recognize sour tastants are thought to be polycystic kidney disease-like (PKD) channels, in particular, PKD2L1 and PKD1L3. These are ion channels expressed in type III taste receptor cells of all populations of papillae. PKD2s are part of the transient receptor potential (TPR) superfamily of ion channels. The proteins belonging to the PKD1 family are composed of a long N-terminal extracellular domain and eleven transmembrane domains with another transmembrane domain at the C-terminus[27]. In *Figure 12* the taste PDK model is represented.

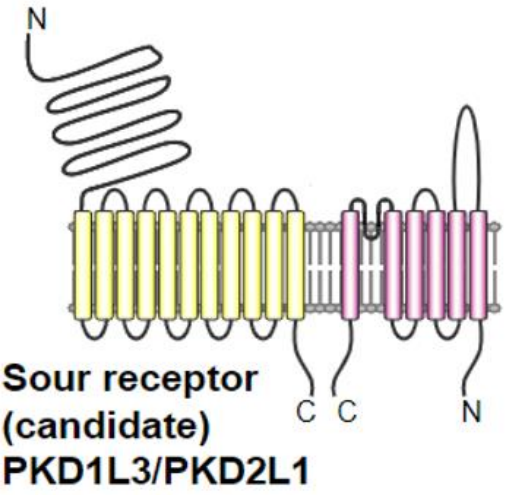


Figure 12:Sour taste receptor schematic structure[28]

The mechanism has not been fully elucidated yet, but it is has been linked to the presence of H^+ ions entering the cell[29]. At the same time, the potassium K^+ channels, that are at rest, are blocked by the proton influx. The joint action of the two channels leads to an accumulation of H^+ ions in the cytoplasm, which causes cellular acidification and the consequent depolarization of the membrane. Depolarization leads to the release of neurotransmitters that trigger synaptic transmission, mediated by the opening of voltage-gated calcium channels on the cell membrane.

2.5 The other roles of taste receptors

The hitherto described receptor mechanisms, which are at the roots of the perception of taste, certainly play a primary role in the gustatory system, but there is increasing evidence that similar mechanisms are present in different systems in which taste transduction has been adopted as a chemodetection system, such as in the cells of the airway or the cells of the gastrointestinal (GI) tract[2,25,30,31].

Multiple records have indeed shown that taste-signalling components are expressed in the GI tract. The presence of alfa-gustudicn and T2R receptors expressed in the stomach and enteroendocrine cell lines suggests that they are similarly coupled as in taste cells and therefore that there is a system similar to that of taste. Furthermore, in addition to T2R receptors, sweet and umami receptors (T1R) are also present, and in more detail the sweet receptor is involved in maintaining the level of glucose in the blood through the secretion of glucagon-like peptide-1 (GLP-1), a hormonal response[3,30].

Furthermore, other types of cells linked to taste transduction have been shown to reside in the GI tract. One example includes brush cells, which have a different morphology compared to pyramidal and elongated enteroendocrine cells, and are thought to have a chemosensory function similar to taste cells, particularly type II cells, since they express the alpha-gustducin and TRPM5 ion channels necessary for the detection of bitter and sweet flavours. Little is known about their function, but they are thought to be involved in the chemosensation of flavour molecules/nutrients. This type of cell is also present in the lungs, pancreas, and respiratory tract. The brush cells present in the latter specifically express bitter taste receptors whose activation through bitter stimuli releases acetylcholine which in turn influences the respiratory rate by lowering it[30,32].

Given the presence of taste receptors in areas of the body other than the oral cavity, in which their stimulation by food molecules does not lead to an activation of taste perception but to different signals that vary depending on the location and that in general are important in metabolism and homeostasis, it is interesting to study these systems at the molecular level to shed light on the different role that tastants can play, as well as to carry out research at the proteomic level regarding the recognition mechanisms of such tastants.

3 Materials and Methods

Over the years several theoretical and computational methods for the investigation of molecular systems at the smaller time and length scales have been introduced, which are broadly grouped under the rather general term of Molecular Modeling. Their main aim is to represent and simulate the behaviour of molecular structures, such as proteins, lipids and other relevant macromolecules, under specific conditions to analyze their physical, chemical, and mechanical properties. Many examples exist in literature where computational predictions have effectively shed light on specific molecular mechanisms, elucidated unknown structure-to-function relationships, as well as guided the design of new in-vitro experiments. Several examples FDA and EMA-approved drugs that have been discovered or designed computationally also exist. In addition, most of the Molecular Modelling tools share the great advantage of being interactive and allowing the visualization of the system of interest with single-atom resolution. Overall, also in the light of the increasing computational power and resources available to research facilities as well as the refinement of the underlying theoretical foundations, Molecular Modelling nowadays features a broad field of application ranging from chemistry to bioinformatics.

Molecular Modeling includes (i) analytical methods, such as quantum mechanics which is linked to theoretical chemistry and allows explicit modelling of subatomic components, i.e. electrons, (ii) numerical methods, such as molecular mechanics which allows more simplified modelling of the system, and (iii) other computer-based approaches. The choice of the specific method depends on the complexity of the system and the time and length scales to be analyzed. For example, even simple biological systems, such as the conformational dynamics of a protein, would be too complex to be represented at a quantum mechanics level with current hardware resources, due to the large number of particles involved (tens to hundreds of thousands of atoms). Instead, using the Molecular Mechanics approach, the computational effort is greatly reduced and the system can readily be investigated, since the macroscopic properties of the system are predicted by ignoring the behaviour of the electrons and other sub-atomic particles[31].

1.1 Molecular Mechanics

The term Molecular Mechanics (MM) describes the application of classical mechanics in determining the equilibrium of molecular structures. This method uses the position information of atoms to describe a molecular system, thus facilitating the simulation of systems with a significant number of particles: atoms are approximated as particles with a specific mass following Newton's classical laws of motion, with the bonds holding them together described as harmonic potentials.

For this reason, the MM approach is unable to describe phenomena that occur at the electronic scale, such as the formation or breaking of covalent bonds.

MM extracts the properties, such as the energy, of the investigated system by describing it in the form of a potential energy function (V) dependent on atomic positions, which includes a series of system-dependent parameters that together form what is known as a force field. A key concept is the potential energy surface, which represents the energetic state of the molecule or system as a function of its geometry. In MM the potential energy surface is the sum of two terms, i.e. of bond and non-bond interactions, describing the contributions of intra- and inter-molecular forces. Both are functions of the position *r* of the *N* particles in the system. Therefore, the potential energy $V(r^N)$ can be written as:

$$V(r^N) = V_{bond}(r^N) + V_{non-bond}(r^N)$$

The bond interactions can be of three types: they can derive from (i) bond stretching between two particles, (ii) angle bending among three particles, and (iii) bond torsion between four particles. Different models have been proposed to describe each of these terms.

$$\mathcal{V}_{bond}(r^N) = \sum_{bonds} \mathcal{V}_{bonds}(r^N) + \sum_{angles} \mathcal{V}_{angles}(r^N) + \sum_{torsion} \mathcal{V}_{torsion}(r^N)$$

Regarding the bond stretching term $\mathcal{V}_{bonds}$ that represents the interaction between two atoms connected by a covalent bond, the easiest and most common model used to describe it is Hook's law, where the interaction is modelled as a harmonic potential with the following parameters:

- k_i is the spring stiffness and indicates the resistance of the bond to being stretched,
- l_0 is the reference bond length, which is assumed when all the other force field terms are equal to zero
- l_i is the actual bond length

$$\mathcal{V}_{bonds}(l) = \frac{k_i}{2}\left(l_i - l_{i,0}\right)^2$$

The harmonic model has also been used for the angle bending term $\mathcal{V}_{angles}$. It corresponds to the triatomic unit, in particular to the angle formed by the three atoms, when two of them are both linked to the third atom. The parameters of the formulation are:

- h_i, the angle stiffness,
- $\theta_{i,0}$, the reference bond angle,

- θ_i, the actual bond angle

$$\mathcal{V}_{angles}(\theta) = \frac{h_i}{2}\left(\theta_i - \theta_{i,0}\right)^2$$

The torsional term, also called the dihedral term, refers to the rotation of a bond between four atoms(A, B, C, D) that are sequentially bonded. The torsional angle is the angle between the A-B bond and the plane identified by B-C-D atoms (*Figure 13*). The formulation is based on a sinusoidal law and the parameters are:

- V_n is the torsional stiffness,
- γ, is torsional equilibrium angle, defines the position of the minimum of the function
- ϕ, is the dihedral angle when all terms are considered.

$$\mathcal{V}_{torsion}(\phi) = \frac{V_n}{2}(1 + \cos(n\phi - \gamma))$$

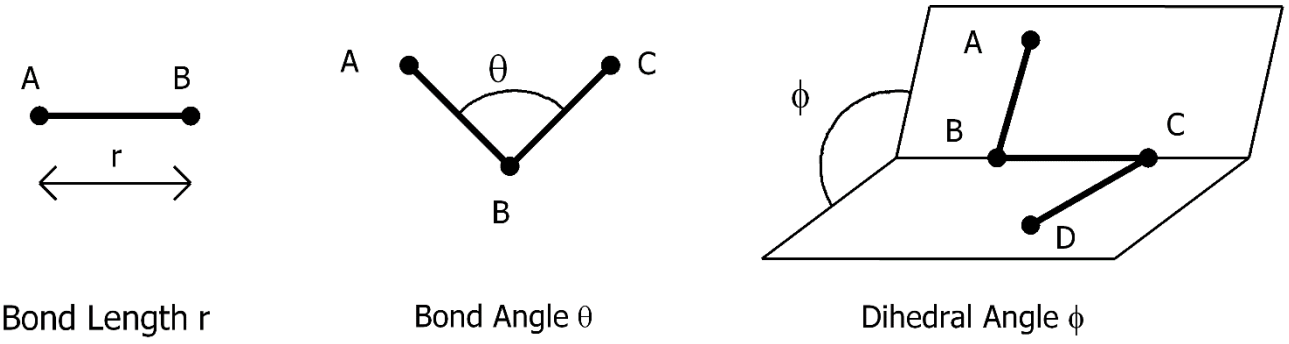


Figure 13: Representation of bond interactions. From left to right: bond length between atom A and atom B; bond angle between atoms A, B, C; and dihedral angle between four atoms A, B, C, D. Source: http://www.chem.hope.edu/~polik/doc/webmohelp16/img/blbada.png.

The non-bond interactions represent the behaviour of non-bonded atoms when they are near enough to influence each other without forming covalent bonds. The interactions are usually modelled as functions inversely proportional to the power of the distance between the atoms. Non-bond interactions consist of two terms: short-range interactions, such as Van der Waals, and long-range interactions or electrostatic interactions.

$$\mathcal{V}_{non-bond}(r^N) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathcal{V}_{electrostatic}(r_{ij}) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathcal{V}_{vdW}(r_{ij})$$

The Van der Waals interactions $\mathcal{V}_{vdW}(r_{ij})$ are attractive or repulsive interactions at variable distances. For instance, at short distances, they are repulsive and the shorter the distance between the atoms, the more the force grows exponentially; at longer distances the interactions become

attractive. The repulsive interaction is based on the "Pauli Exclusion Principle". The principle states that two electrons in an atom cannot possess the same quantum numbers. This happens when the electronic clouds of two atoms overlap and therefore generating a repulsion effect between the two nuclei. The attractive interaction is based on a weak type of intermolecular force, the London dispersion force that arises from the formation of temporary dipoles. The most common function to model the short-range interactions is the Lennard-Jones 12-6 potential:

$$\mathcal{V}_{vdW}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

The parameters are:

- r_{ij} is the interatomic distance,
- σ_{ij}, is the collision diameter,
- ε_{ij}, is the Van der Waals potential minimum.

The first term in the square bracket models the repulsive interactions and the second models the attractive interactions.

The electrostatic interactions $\mathcal{V}_{electrostatic}(r_{ij})$ are usually defined as a Coulomb potential, which describes the interaction between two charged particles and is expressed as:

$$\mathcal{V}_{electrostatic}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

The parameters are:

- r_{ij} is the distance between atoms,
- q_i q_j, are the charges of atoms i and j,
- ε_0, is the dielectric constant.

Regarding the non-bonded interactions, their calculation can involve large computational costs since their number corresponds to the second power of the number of particles N. For this reason, different methods have been proposed to decrease the computational cost of non-bonded interaction calculations, such as the introduction of a cut-off beyond which not to consider these type of interactions or the introduction of potential switches.

For the calculation of potential energy, boundary conditions are set to avoid artifacts derived from long-range interactions. In fact, during a simulation the number of particles that can be considered

is finite, so the system is placed inside a box, called a unit cell, that can be of different geometries (e.g. cubic or parallelepiped, hexagonal prism, truncated octahedron or rhombic dodecahedron). The boundary conditions are crucial in the simulation since they affect the properties of the whole system: depending on the system and the simulated conditions, these must hence be carefully defined. To get around edge effects, periodic boundary conditions (PBCs) are usually applied, which means that the simulation box is virtually surrounded by copies of itself in all directions of space (*Figure 14*). A crucial consequence of this periodicity is that each particle shall not interact with its own replica in neighboring unit cells to avoid physically inconsistent artifacts. Thus, the size of the unit cell must be tailored with great care so that the distance of any particle to its neighboring copy is above the threshold set for non-bonded interaction. This concept is referred to as the *minimum image convention* and defines the unit cell size. Furthermore, the repetition of the box in all directions in space allows the number of atoms inside the simulation box to be kept constant because if an atom leaves the unit cell a replica of it replaces it from a neighboring cell.

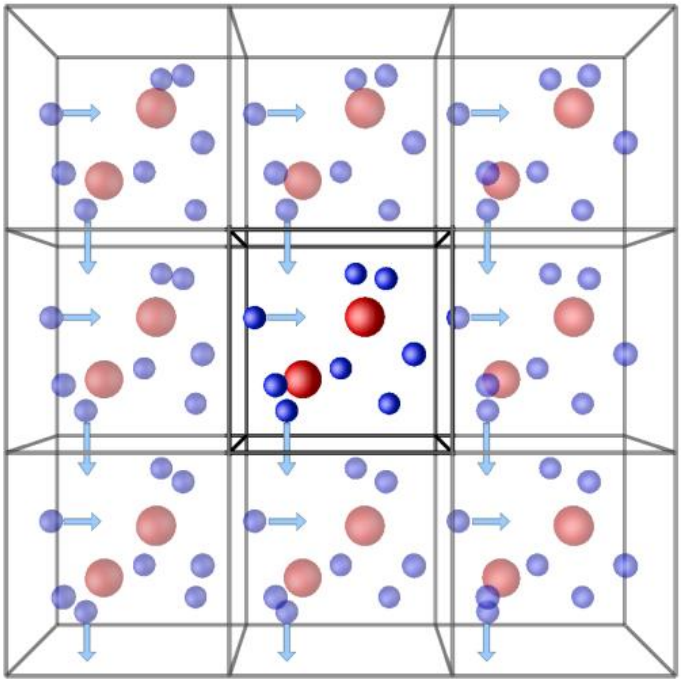


Figure 14: Schematic representation of the idea of periodic boundary conditions for a cubic box. Source: http://isaacs.sourceforge.net/phys/pbc.html

The potential energy surface in other words is the representation of the potential energy of a system on a multidimensional surface of 3N cartesian coordinates (N is the number of particles of the system). MM aims to find, through computational methods, the minima on the potential energy surface since they represent the lowest energy state accessible to the system, corresponding to a stable minimum-energy structure of the system. Therefore, the minimization of potential energy is fundamental both when it is part of a set of actions to be performed subsequently and to obtain a

system with a stable initial configuration without the presence of collision between the atoms. There are different methods to find a local minimum and they can be classified into (i) derivative and (ii) non-derivative methods.

The Simplex is an example of a non-derivative methods. It is based on the construction of a geometric figure with N+1 interconnected vertices, with N being the dimensionality of the energy function. There is a link between the figure and the movements allowed on the surface to find the minimum. However, it is not an efficient method.

As for the derivative methods, they can be distinguished into (a) first-order methods and (b) second-order methods. First-order derivative methods are based on the gradient, i.e. the first derivative of the potential energy function, since its direction indicates where the minimum lies. The absolute value of the gradient, in other words, the intensity, is an indicator of how steep the local slope is. Examples of first-order derivative methods are Steepest Descent[33] and Conjugate Gradient[34], where the basic idea in both methods is to change the coordinates of the atoms to move towards lower energy states. Second-order derivative methods are based on the information of the second derivative that expresses the curvature of the potential energy function. In comparison to the first-order derivative methods, second-order derivative methods are more accurate but require more computational effort. An example of second-order methods is the Newton-Raphson method[35]. Since all methods have their advantages and disadvantages, there is no method to be preferred *a priori*. Rather, the choice depends on the specific goal, the investigated system, the desired accuracy, the computational performance available at runtime, etc. Often, multiple methods are combined to obtain the best tradeoff between computational efficiency and accuracy.

1.2 ASA: Accessible Surface Area

ASA stands for Accessible Surface Area and defines the surface area of a biomolecule that is accessible to a solvent. It is often also referred to as solvent-accessible surface area (SASA). The concept was introduced by B.K. Lee and Fred Richards, to quantify hydrophobic burial, which is the number of hydrophobic side chains that are buried internally during the folding of a protein in the presence of an external hydrophilic solvent, and more generally to understand the folding mechanism of proteins starting from the sequence of amino acids only[36].

The accessible surface area is commonly calculated using the 'rolling ball' algorithm developed by Shrake and Rupley[37], which, as the name suggests, consists in 'scanning' the surface of the biomolecule under examination with a sphere of solvent, called a probe, having a certain radius. The radius of the sphere affects the detail of the surface found. Indeed, with a smaller sphere it is

possible to detect more surface details and this results in a larger observed surface. The most commonly used value for the radius is the one corresponding to the radius of a water molecule, i.e. 1.4 angstroms. Another characteristic that influences the measurement of the observed surface is the definition of the Van der Waals radius of the atoms constituting the biomolecule, which is used to represent the atomic surface of the biomolecule. The ASA is drawn with the centre of the solvent sphere as it moves across the surface of the biomolecule (*Figure 15*).

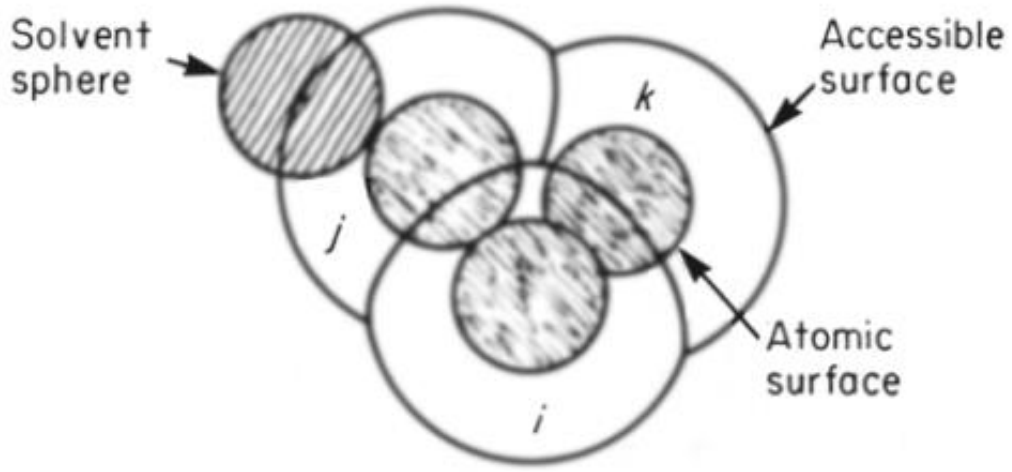


Figure 15: Representation of the accessible surface of three atoms represented as spheres, i,j,k. The volume enclosed by the accessible surface is the excluded volume.[38]

ASA is also related to the concept of the solvent-excluded surface, which defines the molecular volume from which the sphere of the solvent is excluded. This volume is obtained by putting together the Van der Waals volume, therefore the atoms, and the interstitial volume, i.e. the space between the atoms (*Figure 16*). The solvent-excluded surface is also called the Connolly surface as it was implemented three-dimensionally by Michael Connolly and Tim Richmond [38,39] .

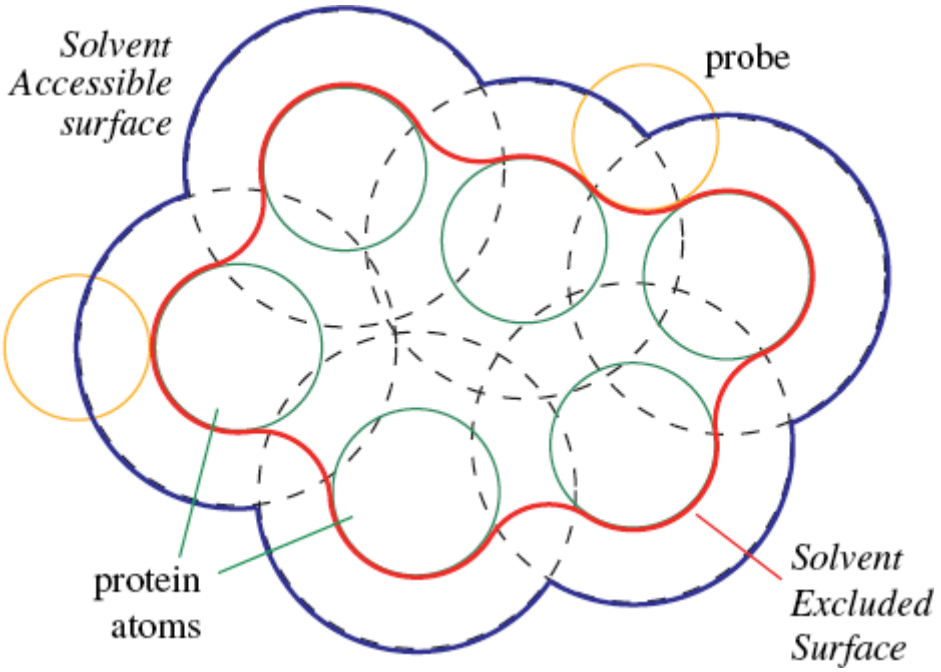


Figure 16:Representation of the Connolly surface or solvent-excluded Surface. Source: https://discourse.mcneel.com/t/2d-metaball-like-connection-between-distort-forms/58631

Several software packages, both open-source and commercial, have implemented or otherwise include the calculation of SASA, such as VMD[40], PyMOL[41], GROMACS[42], Chimera[43], MOE[44] and ChemAxon[45], just to name a few.

1.3 Molecular Docking

Molecular docking can be defined as a method that aims to predict the most energetically favourable conformation of a small molecule, i.e. ligand, in specific regions present on a macromolecule, e.g. a target protein or receptor. In particular, in most docking approaches, after the initialization of the small molecule inside the region of interest on the macromolecule, a sort of conformational adjustment takes place between ligand and receptor through the sampling of a certain extension of the conformational space available to the ligand, to (parts of) the macromolecule, or both, to obtain an optimized orientation between the two such that the free energy of the system is minimal. This means that the correct parametrization of both molecular players is crucial to obtain physically sensible results. In the past, the interaction between ligands and their receptors have been often referred to as "lock-and-key": in this analogy, one can think of the ligand as a key that must be suitable for the the receptor (the "lock"). However, since both the ligand and the receptor usually show a certain degree of flexibility in real-case scenarios, a more suitable analogy would be the *hand-in-glove*.

In addition to the free energy of the system and the stability of the bound complex, another output data of this process is some form of prediction of the binding affinity or binding energy, which is

more correctly referred to as the *score* of the docking, since *affinity* usually implies both enthalpic and entropic terms which are not always fully included in docking algorithms.

Before proceeding with the actual docking, the preparation phase of both the ligand and the receptor is fundamental, as the molecular process of ligand recognition by a molecule depends on the reciprocal three-dimensional orientations and on the nature of the non-covalent interactions that develop. Furthermore, if there are tensions, instabilities or clashes in the starting structure of the ligand and the protein, these should be solved beforehand, e.g. by simulating both under physiological conditions, as such instabilities might then give rise to numerical problems during the docking process itself.

To evaluate the potentially interesting ligand poses all docking methods use a so-called scoring function, which returns a numerical estimate of how much the given ligand conformation can constitute a favourable binding interaction. The different poses of the ligand are then ranked based on this number. Current docking protocols and softwares can be categorized based on different characteristics and features:

- scoring functions can be physics-based molecular mechanics force fields, knowledge-based on key protein-ligand interactions, or empirical;
- the ligand conformational space sampling strategy can be stochastic, i.e make random changes up to a user-defined termination policy, or systematic, i.e sample the search space in predetermined discrete incremental steps;
- size of the sampled search space, Two approaches can be distinguished, the global one which seeks the minimum energy in the space of all possible conformations, and the local one which seeks the minimum energy in a neighbourhood of the starting conformation.

Figure 17 summarizes the key steps common to all docking protocols.

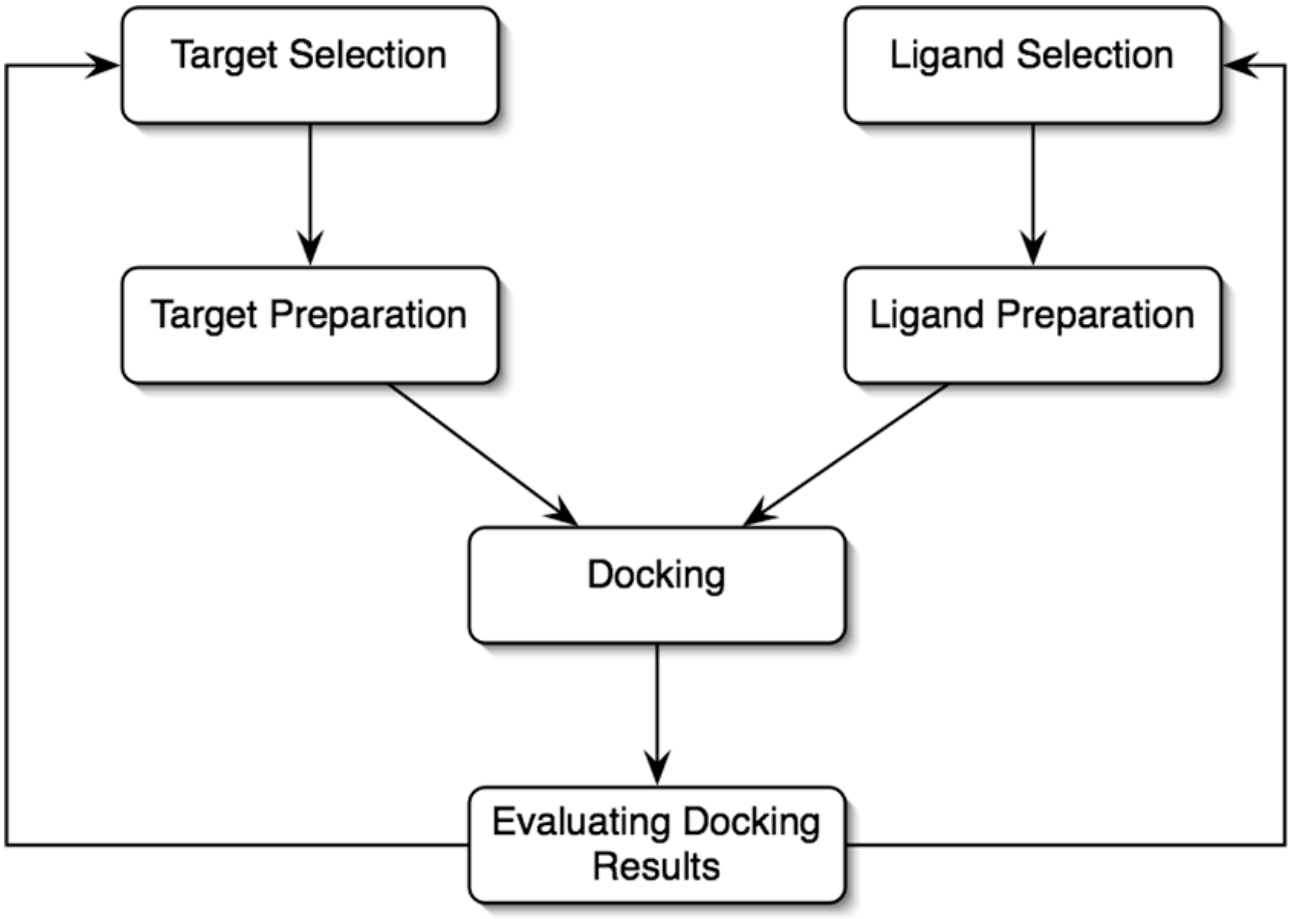


Figure 17: Representation of a docking workflow, with key steps common to all docking protocols. [46]

Given all these methodological differences, depending on the docking software used, implementations may vary. For example, the DOCK[47] software uses a systemic sampling strategy; AutoDock[48] uses two local and two global sampled space search methods and a hybrid of the two. There are many other implementations each with different characteristics.

In the present work AutoDock Vina[49,50] was used, which is an evolution of the original AutoDock software, characterized by an increase in speed of almost two orders of magnitude and an improvement in the accuracy of the binding mode predictions. The difference in performance is due to improvements in the scoring algorithm and in the search algorithm.

The scoring algorithm consists of a hybrid scoring function, i.e. empirical and knowledge-based function, inspired in the X-Score function[51], mainly differing in some terms and in the parametrization method, goimg beyond linear regression. This hybrid scoring function makes it possible to include the advantages of the two scoring functions on which it is based, that is, the empirical and the knowledge-based one. This means that poses are evaluated based on the probability, derived from experimental observations, that the pose actually occurs and based on measurements of experimental affinity. The search algorithm is a Monte-Carlo (MC) iterated search

combined with the Broyden-Fletcher-Goldfarb-Shannon (BFGS) gradient-based local optimizer. The algorithm consists of a succession of steps based on a mutation of the ligand conformation obtained in a stochastic way and a local optimization, each of which is accepted according to the *Metropolis criterio* [52], i.e. the step is assigned a probability based on the Boltzmann probability function. The general form of the *Metropolis criterion* is:

$$P \propto e^{\frac{-\Delta E}{k_B T}}$$

where ΔE is the energy difference between two random variations of the ligand conformation, k_B is the Boltzmann constant and T the absolute temperature. Local optimization is performed via the BFGS method, which uses the value of the scoring function and its gradient, i.e the derivatives of the scoring function with respect to the position and orientation of the ligand. Evaluating the gradient may take longer, but the gain in using the gradient is a significant acceleration of the optimization.

An aspect to underline is that the origin of the structure of both the ligand and the protein greatly influences the results obtained. To obtain better results, the structure of the protein should have been determined experimentally, therefore by NMR or X-ray crystallography, and obtained from the Protein Data Bank. For ligands, there are numerous databases of chemical compounds from which the structures of interest can be obtained, two of which, for example, are PubChem or DrugBank. Another important aspect affecting the docking results is the correct protonation of the ligand and receptor and the correct partial charge state of the ligand, which can be calculated by various methods (Gasteiger, AM1-BCC, etc.). It is therefore crucial to opt for the right ligand preparation method to obtain sensible results.

1.4 ASSAM - Amino acid pattern Search for Substructures And Motifs

ASSAM is a graph theoretical program that was first presented in 1993 and later implemented as a free web server available at http://mfrlab.org/grafss/assam/[53,54]. It was developed to find structural similarities of amino acid side-chains patterns between a protein or macromolecule of interest and proteins in the Protein Data Bank archive containing the currently solved proteome, enabling a proteome-wide scan of residue patterns of interest.

It is well-known that the function of biological macromolecules is directly determined by their three-dimensional structure (tertiary and quaternary structure). Hence, the presence of local structural similarities between different proteins might also suggest functional similarity, even without the identity of the amino acid sequences. More in detail, the functionality of a protein is

often associated not with its entire three-dimensional structure but with one or more restricted areas called active sites. The active sites correspond to a set of amino acids arranged in space in a specific way or pattern. Another known aspect among biological molecules is the presence of such patterns in totally unrelated proteins, which have neither sequence similarity nor global 3D structure similarity nor common evolutionary precursors, but are involved in similar chemical activities: this represents a case of convergent evolution. This is why three-dimensional amino acid patterns are the key concept of ASSAM.

The basic concept behind the ASSAM workflow is described as follows:

A. The structure of the pattern is represented through a graph. The **nodes or vertices** of the graph identify the side chains of the individual amino acids. Each node corresponds to a vector obtained from two pseudo atoms whose position identifies the functional part of the side chain corresponding to the node. The **edges or arcs** of the graph identify the geometric relationships between pairs of nodes, calculated in terms of the distances between the corresponding vectors. *Figure 18* shows an example of a graph with nodes and edges.

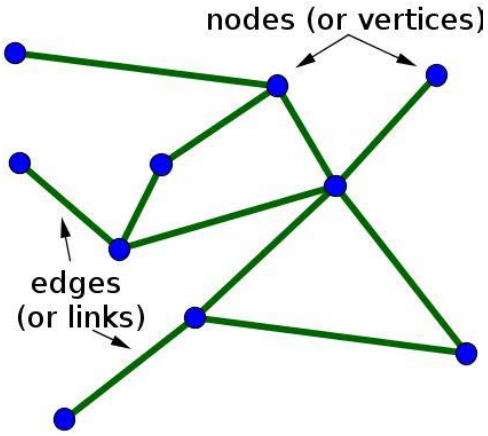


Figure 18: Example of a graph with 10 nodes and 11 edges. Source: https://mathinsight.org/definition/graph

B. The query pattern is then searched against the structures present in the Protein Data Bank for any occurrences through the maximal common subgraph approach. In graph theory, maximal common subgraph can mean (i) a graph that is an induced subgraph of two initial graphs, that is, a graph formed by the largest subset of common nodes or it can mean (ii) a graph that is a subgraph of two initial graphs that has as many edges as possible.

The side chain representation used in ASSAM is shown in *Figure 19*.

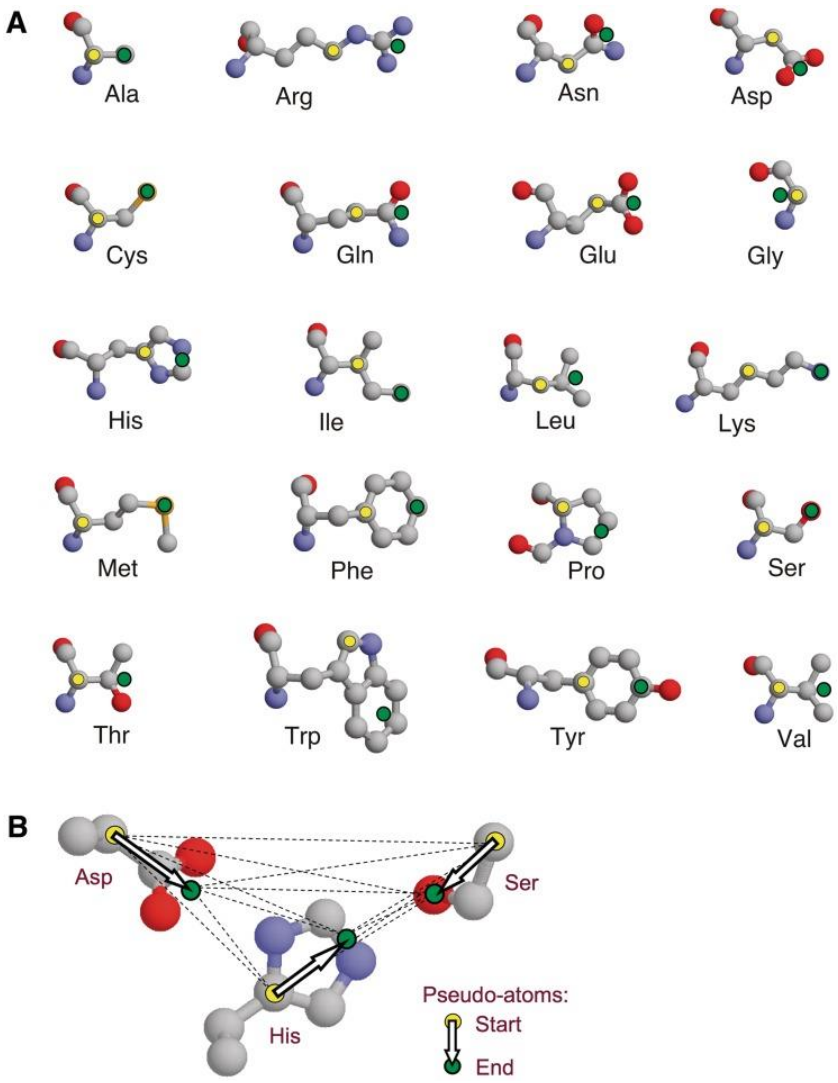


Figure 19: The side chain representation used in ASSAM. (A) Represents the 20 amino acid types with the positions of pseudo-atoms in yellow and green. (B) Diagram of an aspartate–histidine–serine pattern. The dashed lines represent the distances between pseudo-atoms.[53]

The information of the three-dimensional amino acid pattern of interest, also called the motif, can be collected in a formatted PDB file that represents the input data into the program. The motif can be formed by 2 to 12 amino acids. The search output consists of a list of protein structures, for which there is a match of residues with the residues of the query motif, ranked by the RMSD, i.e. root mean square deviation, of the matches[53] (*Figure 20*).

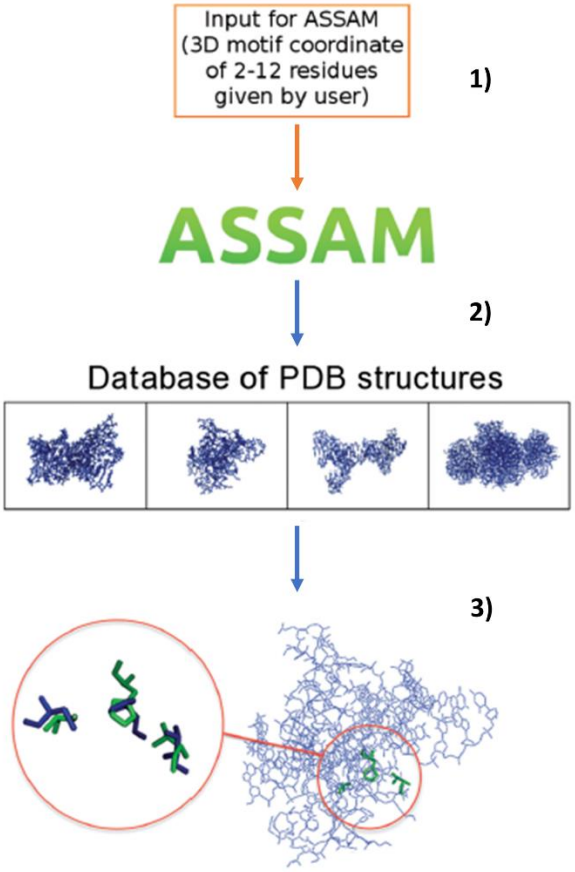


Figure 20: Diagram showing the input and output structures of ASSAM [53].

4 Sweet Taste Receptor agonist binding site similarity search to elucidate the roles of tastants in homeostasis and disease

4.1 Introduction

The sense of taste is a sensory modality that plays a fundamental role in discriminating ingestible substances, nutrients, from potentially harmful substances that must be avoided, especially in omnivorous species given the range of their feeding strategy[5]. It appears thus entirely reasonable that the sense of taste played an important role throughout the evolution of organisms, and it is well known that the basic functional mechanisms of the gustatory system have been conserved in both vertebrates and invertebrates. Humans, in particular, can perceive five primary taste qualities, i.e. sweet, umami, bitter, salty and acid, through the interaction of the molecules contained in food and the specialized sensory organs, i.e. the taste receptor cells, present on the papillae of the tongue that recognize their chemical properties.

Interestingly, sweet, umami and bitter taste receptors are also expressed in other areas of the body such as in the gastrointestinal tract, respiratory tract and endocrine system, where they are thought to be involved in activities such as nutrient detection and regulation of metabolic activity[30,55], innate immune response and bronchodilatation[25,30], and modulation of hormone secretion[2,55], respectively. This work investigates the sweet taste receptor complexed with one of its taste agonists, sucrose. The receptor is a class-C dimeric GPCR composed of the T1R2 and T1R3 monomers, consisting each of a large extracellular domain connected via a cysteine-rich domain to a transmembrane domain formed by 7 helices. The study focused on the two binding sites present in the T1R2 and T1R3 subunits in the extracellular domain.

A methodology was implemented, using algorithms and software packages belonging to the field of molecular modelling, to screen the currently solved proteome for proteins which exhibit a highly similar local amino acid pattern to the one lining the sucrose binding site in the T1R2/T1R3 dimer, in terms of both type and spatial distribution of residues. The rationale of the work is to explore the taste transduction pathway from a proteomic point of view, to elucidate the possible functions of tastants beyond the gustatory system, where they evoke the sensations of taste, and to investigate whether other classes of proteins have a conserved ability to recognize such ligands, with possible implications in nutrition, homeostasis and disease.

4.2 Materials and Methods

4.2.1 Similarity search software: ASSAM

As a first step in the present work, the state of the art of existing methodologies related to the study of protein binding sites was examined, with a special focus on strategies aimed at the comparison of different protein binding sites. Over the years, several computational methods quantifying the global or local similarity of protein cavities have been developed, and their number is growing to this day. Such methods, each with their own peculiarities and distinctive aspects, all share three general methodological steps:

1. Three-dimensional analysis of the structures of interest;
2. Comparison of such structures;
3. Quantification of similarity through a metric (e.g. a scoring function).

Among those, the most crucial step is the comparison, for two main reasons: the first is that an incorrect comparison method leads to an underestimation of the similarity score, and the second is the high computational cost in the case of multiple comparisons with a large number of structures, as can happen for example in a proteome-level search against whole databases such as the Protein Data Bank. The strategy for representation also influences the result of the comparison. Indeed, the binding site can be represented in various ways, including (a) taking the type of amino acid residues that interact with the ligand into consideration; (b) representing the binding site through a surface onto which the physical-chemical characteristics are projected; (c) by considering protein-ligand interactions. The first two methods listed are structure-based, i.e. they stem from observing the structure of the protein.

Different methods that follow the same representation scheme can follow however different comparison strategies. Some of the comparison strategies can be (a) graphical-theoretical approaches, where the maximum common subgraph is searched; (b) fingerprint approaches, where the shapes involved in the binding site are considered; (c) approaches based on labelled 3D points and geometric hashing, i.e. 3D transformations that align pairs of structures, or (d) alternative approaches. Furthermore, comparison algorithms may or may not depend on the alignment of the structures of interest. Comparison methods that rely on residues can use graphs, fingerprints, or alternative approaches. In particular, the comparison reveals the similarity between the residues, the type of residues and the atomic composition; also, such methods perform well where the sequence and atomic position of the structure of interest are well preserved. Those that rely on surfaces can instead use graphs or labelled 3D points for comparison. These methods are particularly used when dealing with binding sites in proteins that do not show significant conservation in residues, atomic

composition, orientation or folding, but show considerable selectivity towards common ligands. Indeed, in these cases, the distribution of the properties on the surface of the binding site and the shape of the binding site are determining factors for the selectivity of the ligands. And finally, methods that rely on interactions can use graphs or fingerprints for comparison[56].

In the scoring function, the final score applied to a binding site pair match might also include other properties beyond just the site comparison metric, e.g. surface similarity.

Table 1: Summary of the methods analyzed in the present work, and their respective Type of representation of the binding site and strategy of comparison between binding-sites

Method	***Type of representation***	***Strategy of comparison***
SuMo (2003) [57]	Residue-based	3D Points
PINTS (2003) [58]	Residue-based	Other
eF-seek (2004) [59]	Surface-based	Graphs
TM-Align (2005) [60]	Residue-based	Other
SiteEngine (2005) [61]	Surface-based	Graphs
ContactMetricServer (2006) [62]	Residue-based	Other
PocketMatch (2008) [63,64]	Residue-based	Other
MultiBind MAPPIS (2008) [65]	Surface-based	3D Points
PevoSOAR (2009) [66]	Surface-based	Other
fPOP (2009) [67]	Surface-based	Fingerprint
PESD-serv (2010) [68]	Interaction-based	Other
SeSAW (2010) [69]	Residue-based	Other
LabelHash (2010) [70]	Residue-based	Other
FuzCav (2010) [71]	Residue-based	Fingerprint
Pro-BIS ligand (2012) [72]	Surface-based	Graphs
PoSSuM (2012) [73]	Residue-based	Other
COFACTOR (2012) [74]	Residue-based	3D Points
SPRITE-ASAAM (2012) [53,54]	Residue-based	Graphs
SiteComp lin (2012) [75]	Interaction-based	Other
Iso-Cleft Finder (2013) [76]	Residue-based	Graphs
CatSId (2013) [77]	Residue-based	Graphs
IMAAAGINE (2013) [78]	Residue-based	Graphs
Apoc (2013) [79]	Residue-based	Other
ASSIST (2014) [80,81]	Residue-based	3D Points
IsoMIF Finder (2015) [82,83]	Interaction-based	Graphs
G-LoSA (2016) [84]	Residue-based	Graphs
Geomfinder (2016) [85]	Residue-based	Other

PatchSearch (2019) [86]	Residue-based	Graphs
Drugreposer ER (2019)[87]	Residue-based	Graphs
DeeplyTough (2020) [88]	Interaction-based	Other

After an extensive analysis of the methods presented in *Table 1*, ASSAM was chosen as the most appropriate for the present work for the following reasons: (i) the simplicity in defining the binding site to be searched, i.e. an atomic coordinate file of the residues of interest as a PDB-formatted file, (ii) the possibility to screen against the whole Protein Data Bank database, (iii) the comparably fast time to solution, and (iv) the fact that the program considers both right-handed and left-handed superpositions equally valid and both are reported in two separate result lists. Left or right overlap refers to the orientation of the alpha-helices. In particular, the superimposition of a left-handed a-helical bundle to a right-handed one are not equivalent, but, at the level of side chains, two groupings of amino acids can have the same chemical activity even without necessarily being of the same handedness, instead, the important terms are the distances between the residues[53].

4.2.2 Workflow

4.2.2.1 Gathering information from an initial model

The analyzed structure is the sweet receptor dimer composed of the T1R2 and T1R3 monomers complexed with the tastant agonist sucrose in both binding sites in the two protein subunits. The T1R2 / T1R3 model was obtained in previous work from the 6N51[89] template, which is *Metabotropic Glutamate Receptor 5 bound to L-quisqualate and Nb43*, present in the RCSB Protein Data Bank. For all the subsequent analyses, both binding sites were considered to account for their difference, despite their ability to bind the same ligand.

The first analysis that was performed is a visual analysis, using the VMD graphics software[40], which is a molecular visualization program, to frame the two ligands and consequently, the two binding sites, identified by means of a distance threshold on residues with respect to the ligands.

To produce the ASSAM input files, a list of residues of interest defining the binding site to be screened for must be obtained first, and different strategies can be exploited to do so. A first strategy consisted in obtaining a list of coordinates of the residues of the receptor located at a certain distance from the ligand. From this set of residues different smaller subsets were the extracted, composed of an increasing number of residues (from 3 to 12), as indicated in the guidelines of ASSAM. In addition, a further subset of residues was extracted, consisting of those residues involved in non-covalent interactions with the ligand and also respecting the same threshold of maximum distance from the ligand as defined above. These residues were identified

using the Protein-Ligand Interaction Profiler or PLIP software[90], which takes as an input a structure in PDB format, defined by the user or taken from the Protein Data Bank, and subsequently detects hydrogen bonds, hydrophobic contacts, pi-stacking, interactions pi-cation, salt bridges, water bridges, metal complexes and halogen bonds between ligands and targets. All the subsets of residues of interest will be referred to with the term *motif* in the subsequent descritions.

The goal of the above-mentioned strategies was to extract different residue motifs to be used as input data in the ASSAM similarity search. The motif extraction procedure was automated using the Python programming language and the MDAnalysis Python package. However, the described strategy showed a major drawback since only one conformation of the receptor-ligand complex was considered and the method of selection of sub-motifs with increasing number of residues was not optimal. Therefore, the latter step was discarded and only the motif composed of residues involved in non-covalent interactions with the ligand was considered, obtained for different conformations of the receptor-ligand complex. Such different starting conformations were obtained through a clustering procedure performed on the last 100 ns of a Molecular Dynamics trajectory of the sweet taste receptor, obtained in previous work. Clustering was performed on the residues lining the binding site. Several tests were carried out to optimize the size of this reference group, by tuning both the distance from the ligand defining the residues to be considered, and the RMSD cut-off value for clustering, gradually decreasing from 0.15 nm (*Figure 21*).

C Binding site		**D Binding site**	
Reference Group 10A			
cutoff [nm]	clusters	cutoff [nm]	clusters
0,15	1	0,15	1
0,1	2	0,1	1
0,095	22	0,095	1
0,09	177	0,09	1
0,085	624	0,085	3
---	---	0,08	12
---	---	0,075	172
---	---	0,079	26
Reference Group 5A			
cutoff [nm]	clusters	cutoff [nm]	clusters
0,15	1	0,15	1
0,1	1	0,1	1
Reference Group 3.5A			
cutoff [nm]	clusters	cutoff [nm]	clusters
0,15	1	0,15	1
0,1	1	0,1	1

Figure 21: Tests to optimize the size of the reference group, The values highlighted in blue are the final values chosen.

The final tuned values were a distance to the ligand of maximum 10 Å and a cut-off of 0.95 Å, for both binding sites. After identifying the clusters for each binding site representing the dominant conformation states, the corresponding centroids were extracted. These were subsequently used as the starting conformations from which the residue motifs representing the binding site were created, by using PLIP and MDAnalysis as previously described. In detail, for the identification of the interacting residues, the input files used in PLIP consisted of residues located within 10 Å from the ligand and the ligand itself. MDAnalysis was initially used to create a PDB formatted file containing just the binding site and its bound ligand, to be used with PLIP, and again to create a PDB formatted file containing only the interacting residues, i.e. to create the motif. *Figure 22* provides a visual representation of this process.

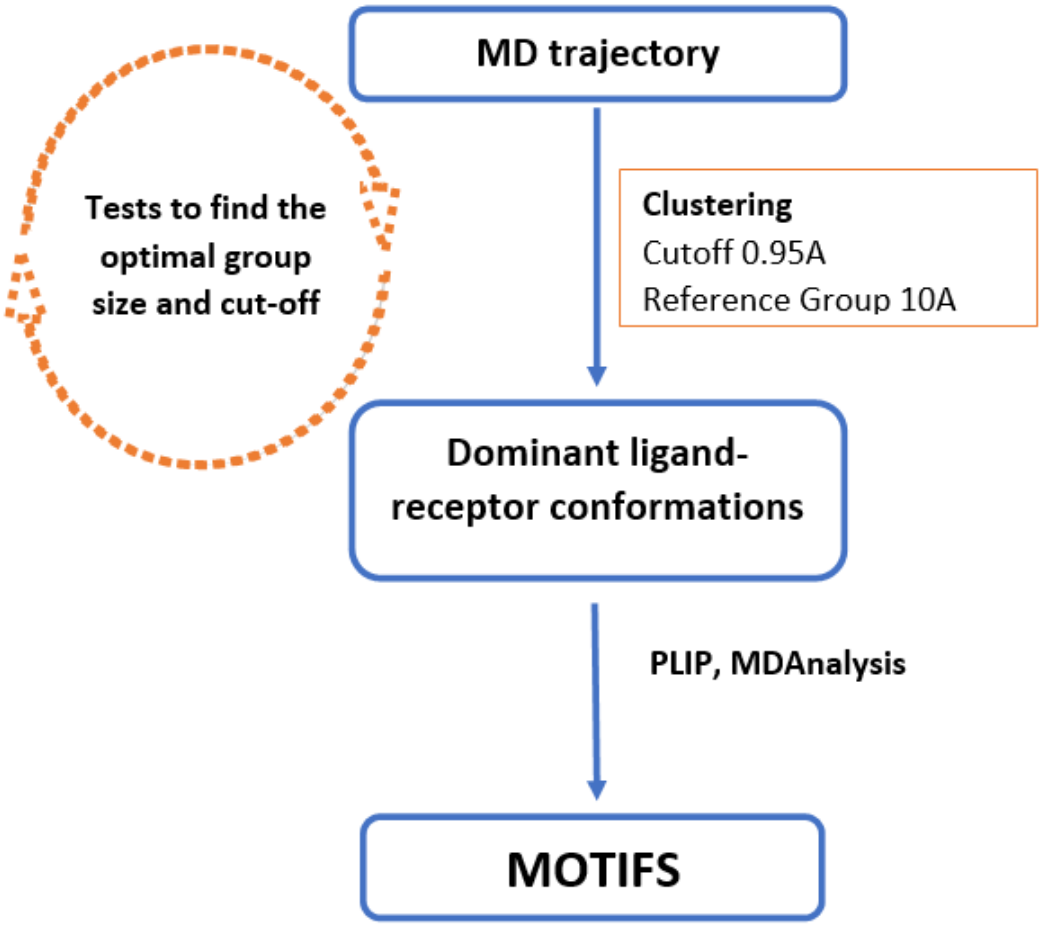
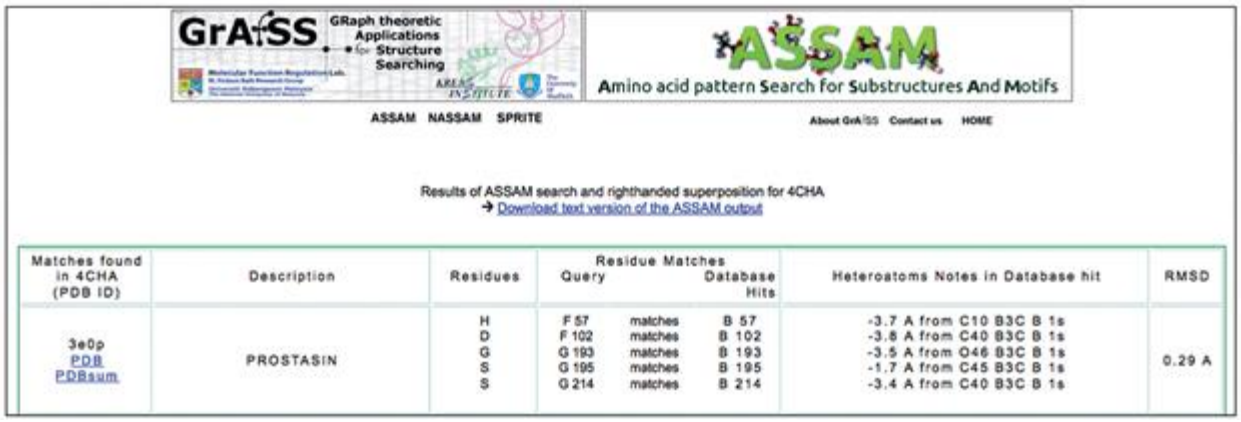


Figure 22: Flow chart of the creation of the motifs

4.2.2.2 Similarity search

Subsequently, the similarity search for each extracted motif, that is a local 3D representation of the binding site of interest, was carried out using the ASSAM webtool. For successful searches, the output results were saved in a separate file, in table format, to allow easier retrieval and

management of the data for the following analyses. The output of the ASSAM search is formatted as a list, in which each row corresponds to a *hit* protein found in the Protein Data Bank, along with its PDB accession ID, which presents a match with the residues of the input motif, which are also referred to as *query*. The matching residues between the query and the hit are also indicated, as well as the root-mean-square deviation (RMSD) value of between query residues and hit residues after alignment. For each input, ASSAM retrieves a maximum of 100 hit proteins, sorted by the number of matching residues and RMSD value. An example of the ASSAM search output is provided in *Figure 23*.


GrASS GRaph theoretic Applications for Structure Searching

ASSAM Amino acid pattern Search for Substructures And Motifs

ASSAM NASSAM SPRITE About GrASS Contact us HOME

Results of ASSAM search and righthanded superposition for 4CHA
→ Download text version of the ASSAM output

Matches found in 4CHA (PDB ID)	Description	Residues	Residue Matches Query		Database Hits	Heteroatoms Notes in Database hit	RMSD
3e0p PDB PDBsum	PROSTASIN	H	F 57	matches	B 57	-3.7 A from C10 B3C B 1s	0.29 A
		D	F 102	matches	B 102	-3.8 A from C40 B3C B 1s	
		G	G 193	matches	B 193	-3.5 A from O46 B3C B 1s	
		S	G 195	matches	B 195	-1.7 A from C45 B3C B 1s	
		S	G 214	matches	B 214	-3.4 A from C40 B3C B 1s	

Figure 23:An example of an ASSAM search output[53].

In the table-formatted file where the ASSAM output results were saved, containing the hit protein PDB ID, the matching residues of the query and the hit and the RMSD value, the following pieces of information have been subsequently added: the number of the initial conformation from which the query motif was created, and further information on the hit protein obtained from the RCSB Protein Data Bank site, such as the DOI of the corresponding pubblication, the EC classification, and the organism in which the protein is expressed. The process of creating this file has been automated as well.

4.2.2.3 Multi-step filtering

Given the large number of hit proteins found through the binding site similarity search, further filtering steps were implemented to reduce and refine this set, with the goal of extracting those hits more relevant with respect to the binding site of the receptor under analysis.

Since the hit proteins have residues matching those present in various input motifs, it is necessary to understand whether these matching residues might indeed constitute a binding site. To achieve this, the Solvent-Accesible Surface Area (SASA) was calculated, as implemented in the GROMACS software. In detail, for each hit protein, a custom index file was created containing only the

matching residues, which was then used for the calculation of the SASA, expressed in nm^2, of the group of residues. Values equal to zero indicate that the cleft identified by the residues is not exposed, but is rather buried in the structure of the protein and therefore cannot be accessible to the solvent nor the ligand. Values that are greater than zero, on the other hand, indicate that the cleft formed by the hit residues is on the surface of the corresponding protein and therefore might allow for binding the ligand. A further filtering step was carried out based on the value of the SASA calculation. To do so, the reference SASA values of the two original binding sites present in the sweet receptor were calculated in the last 100ns of the MD trajectory, to obtain two distributions of these values. From these distributions, two average SASA values were obtained which were used to establish two corresponding threshold values. Starting from these threshold values, the hits with a lower SASA value were excluded, because the binding site identified by the match residues was considered different from the corresponding binding site in the receptor. The threshold was set as 20% of the mean reference sweet receptor binding site SASA values.

The last filtering step that was carried out was based on Molecular Docking calculations. The original ligand of the receptor-ligand complex, i.e. sucrose, was docked to the site defined by the hit residues of each hit protein passing the previously mentioned SASA-based filtering step, using the AutoDock Vina docking software[49]. Among the proteins that resulted in docked sucrose with negative binding affinity, indicative of a possibly stable interaction of the latter with the protein, only those that had binding affinity values above a certain threshold were considered.

A further filtering step was lastly employed: as the ensemble of obtained proteins is expected to be quite heterogeneous, containing proteins expressed by homo sapiens as well as other organisms (as found in the PDB), the aim was to screen for human homologues for the non-human protein hits, using the BLAST (Basic Local Alignment Search Tool) program[91,92]. BLAST screens sequence databases for given sequence queries: in this case, the human genome was screened against, calculating the statistical significance of matches. In this work, only proteins with an identity greater than 80% were considered. Therefore, triple filtering was carried out in this step.

Finally, for the proteins belonging to the human proteome, excluding those obtained in the BLAST search, in particular for the ten hit proteins with the best predicted binding energy, an in-depth analysis was carried out in the UniProt bioinformatics database[93] by investigating their functional characteristics.

Figure 24 summarizes the above-described workflow using a flow chart.

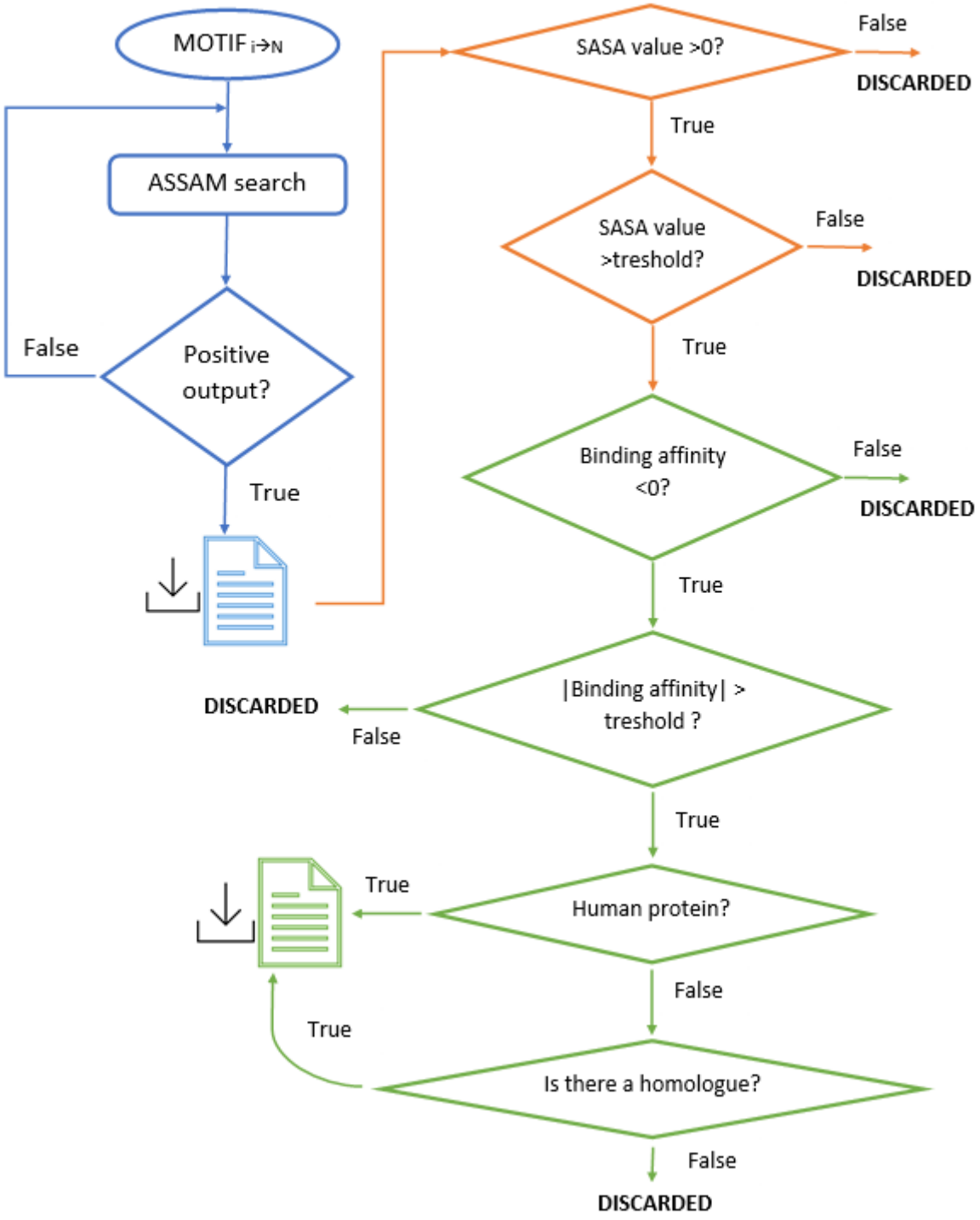


Figure 24: Flowchart of the proposed workflow. The file images indicates the archive files. The boxes in blue refer to the ASSAM search, those in orange to the filtering step of SASA, and those in green refer to the filtering step of Molecular Docking.

4.3 Results

4.3.1 Starting conformations and motifs

Figure 25 shows the renderings of the receptor-ligand complex model created in VMD and visually inspected.

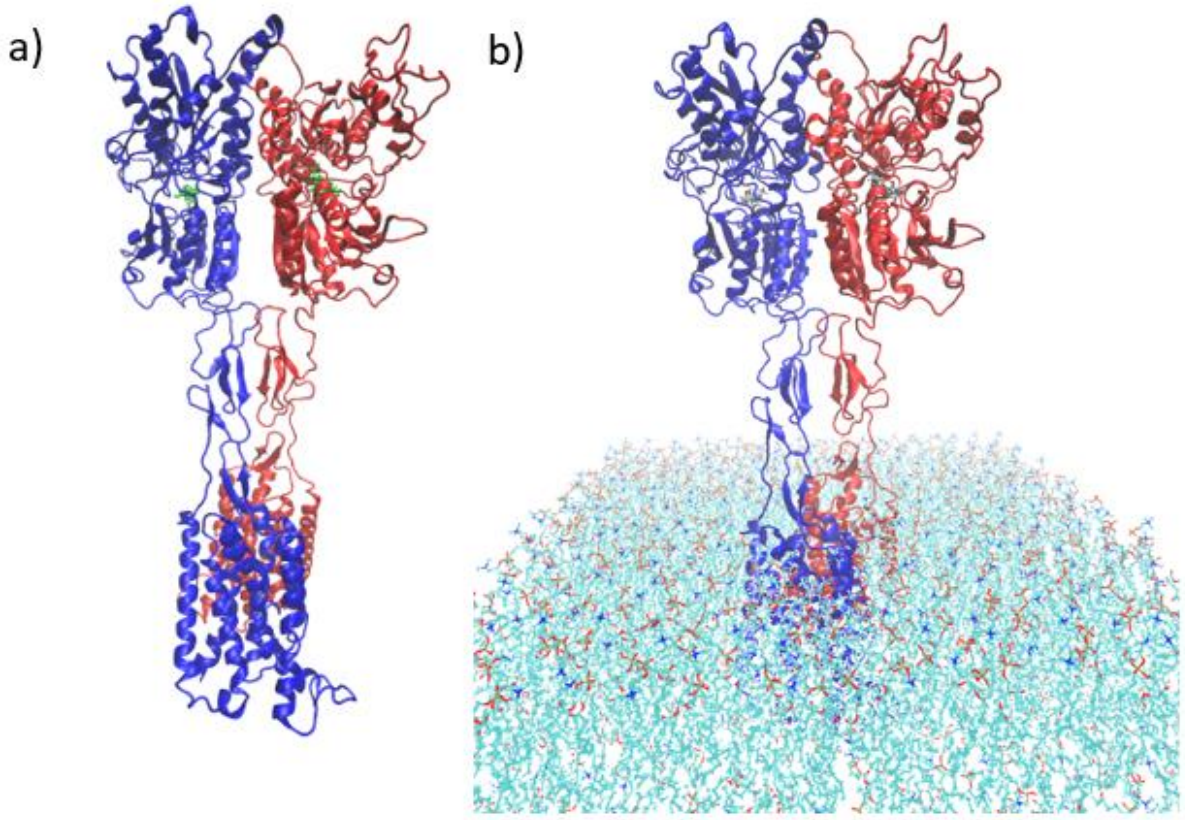


Figure 25: The ligand is shown in green. Figure a) shows the T1R2/T1R3 model, where the blue chain, chain A, corresponds to the T1R2 subunit and the red chain, chain B, corresponds to the T1R3 subunit. Figure b) shows the complex protein-membrane.

The binding sites defined by imposing a distance cutoff of 10 Å from the ligand are shown in *Figure 26*.

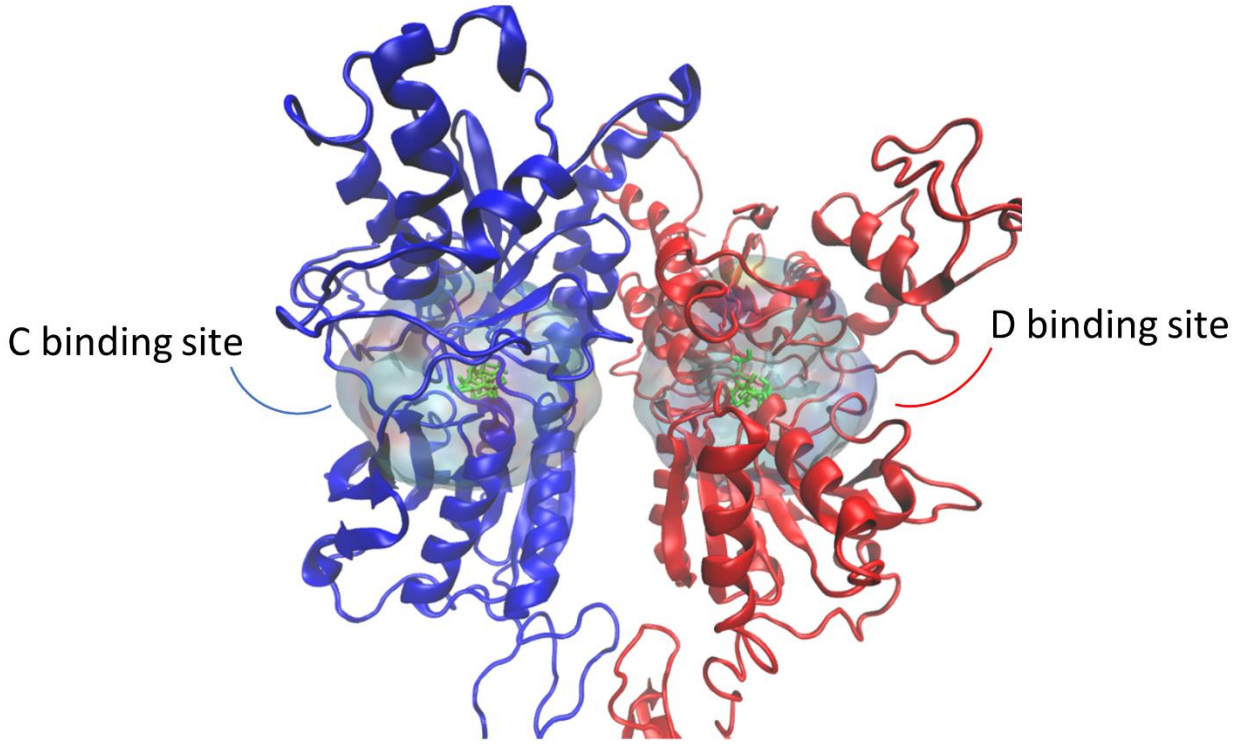


Figure 26: Focus on the ligands, sucrose, shown in green, and the two binding sites, the transparent surfaces, defined within a distance of 10 Å from the respective ligand.

The clustering procedure of the residues forming both binding sites resulted in 22 clusters for the binding site in the T1R2 subunit, the blue chain, and 1 cluster for the binding site in the T1R3 subunit, the red chain. In the description of the results reported herein, we will refer to the binding site of the T1R2 subunit as the C binding site and the binding site of the T1R3 subunit as the D binding site. Given the numeric difference of clusters of sites C and site D, a verification was carried out for site D, i.e. the clustering procedure was checked step by step manually, which confirmed the single cluster obtained.

The centroids were extracted from the clusters and then the motifs were subsequently created, 22 for the binding site C and 1 for the binding site D. Regarding binding site C, the number of interacting residues identified by PLIP and used for the creation of the motifs is on average 5 (*Figure 27*), while for binding site D, 6 interacting residues were obtained.

Binding site C	43 A ALA	44 A ASN	63 A GLU	65 A LYS	66 A VAL	67 A ILE	103 A TYR	211 A SER	212 A SER	213 A ASP	214 A THR	242 A THR
Centroid 1												
Centroid 2												
Centroid 3												
Centroid 4												
Centroid 5												
Centroid 6												
Centroid 7												
Centroid 8												
Centroid 9												
Centroid 10												
Centroid 11												
Centroid 12												
Centroid 13												
Centroid 14												
Centroid 15												
Centroid 16												
Centroid 17												
Centroid 18												
Centroid 19												
Centroid 20												
Centroid 21												
Centroid 22												

Figure 27: The figure shows in pink the interacting residues found with PLIP for the binding site C in the 22 conformations corresponding to the 22 centroids.

4.3.2 Similarity search and multi-step filtering

The initial proteome-level screening for binding sites similar to the T1R2/T1R3 dimer, performed using the ASSAM tool, resulted in a total of 2500 hits, 2400 of which for binding site C and 100 for binding site D. Out of the total 2500, those belonging to human proteome were 236 considering unique PDB IDs, 293 when including redundant PDB entries.

The SASA filtering step reduced the total number of hit proteins to 2477 by excluding all hits with a SASA equal to 0, and subsequently further down to 949 hit proteins when imposing the SASA criterion defined by equation (1) below:

$$SASA_{HIT} > \overline{SASA}_{REF} * 0.2 \quad (1)$$

Where $SASA_{HIT}$ corresponds to the calculated SASA value of the given hit motif and $\overline{SASA}_{REF}$ is the average SASA value of the sucrose binding sites in the T1R2/T1R3 dimer, i.e. the motif shall have a SASA value of at least 20% of the original binding site.

Among the 949 discovered proteins, 120 belong to the human proteome, 98 excluding redundant PDB entries.

In the subsequent docking step, the 949 input proteins were prepocessed by re-building hydrogen atom coordinates, assigning Gasteiger partial charges to both the sucrose ligand and the proteins, solving PDB formatting errors, pruning alternate atom locations and comments and fixing any missing residues in the protein structures. These preprocessing steps were performed using the OpenBabel program[94] for alternative location pruning and protonation, and using SWISS-MODEL online tool[95] to rebuild the 3D coordinates of missing residues in the proteins. Molecular Docking filtering resulted in 831 hit proteins with an overall negative predicted binding energy to sucrose. By enforcing a binding energy threshold of at least -5 kcal/mol, the number of proteins of interest was reduced to 386, of which 338 refer to binding site C and 48 refer to binding site D. Out of these 386, only 43 (corresponding to 34 non-redundant PDB IDs) belong to the human proteome. When also taking human homologues of the (animal) hit proteins into consideration, the final number of non-redundant proteins of potential interest is 44.

A graphical summary, obtained in Excel, of the decrease in the size of the initial set of hit proteins, is presented in *Figure 28*.

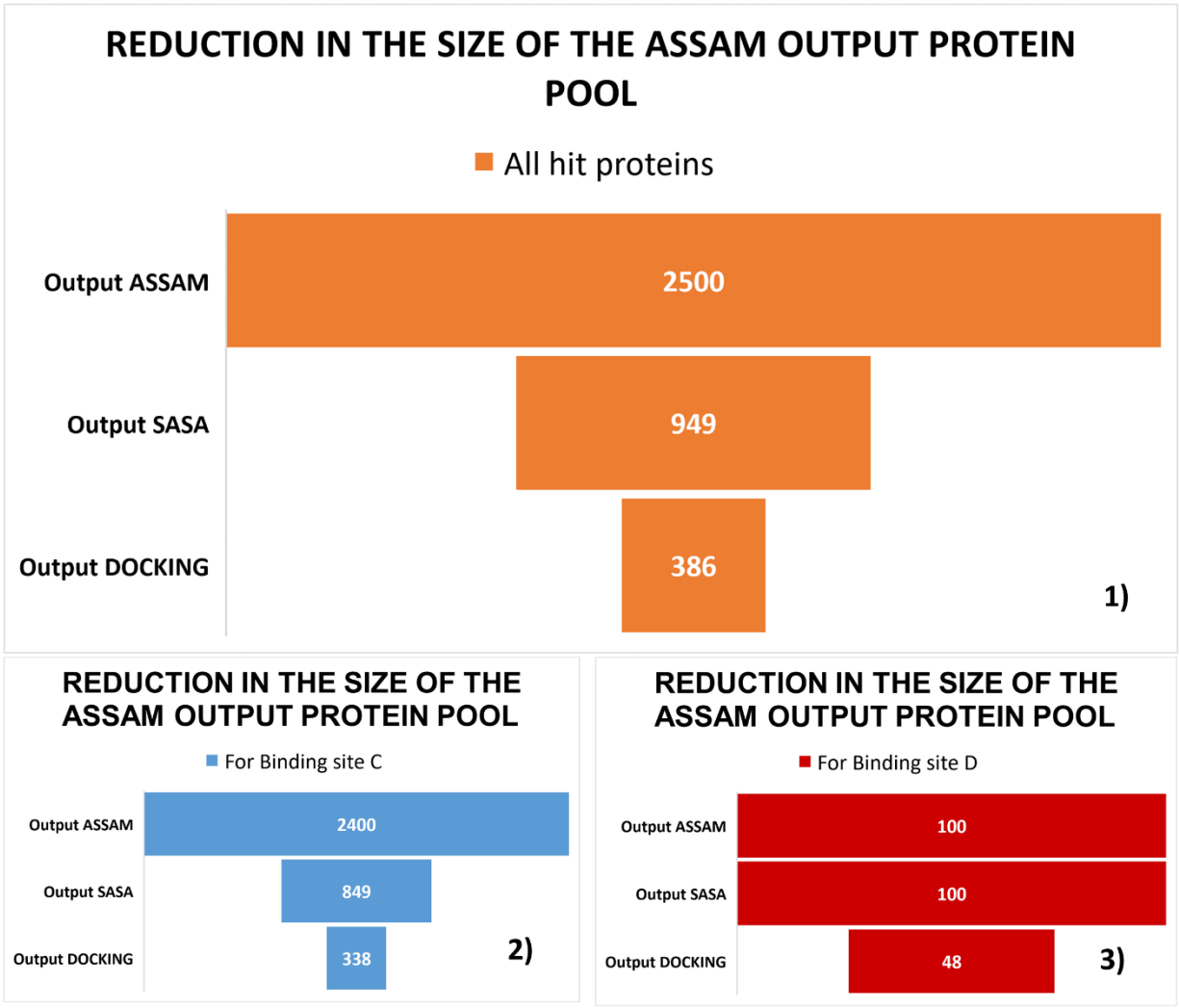


Figure 28: Reduction of hit protein pool obtained from ASSAM research results. 1) The orange cone graph represents the total hit proteins, considering the results for both binding sites; 2) the blue cone graph represent the hit proteins for the C binding site; 3) the red cone graph represent the hit proteins for the D binding site.

To further validate the binding site of the remaining hits, as identified by the residues matching the sucrose binding site present on the T1R2/T1R3, the hits were RMSD-aligned to the T1R2/T1R3 dimer by superposing the matching residues of the binding sites. This step was carried out before the search for homologous proteins. The hit proteins with the best predicted binding affinity, i.e. more negative, were considered, both among the group of proteins not belonging to the human proteome and among those belonging to the human proteome, thus selecting a total of 20 top protein hits. A more detailed protein family and classification analysis was carried out for those top hits, by inspecting the information extracted from the RCSB PDB and embedded in the ASSAM result file for every hit in earlier stages. The distribution of the main protein families of the hits is shown in *Figure 29,* and it can be noted that the majority of hits correspond to proteins associated with enzymatic activities.

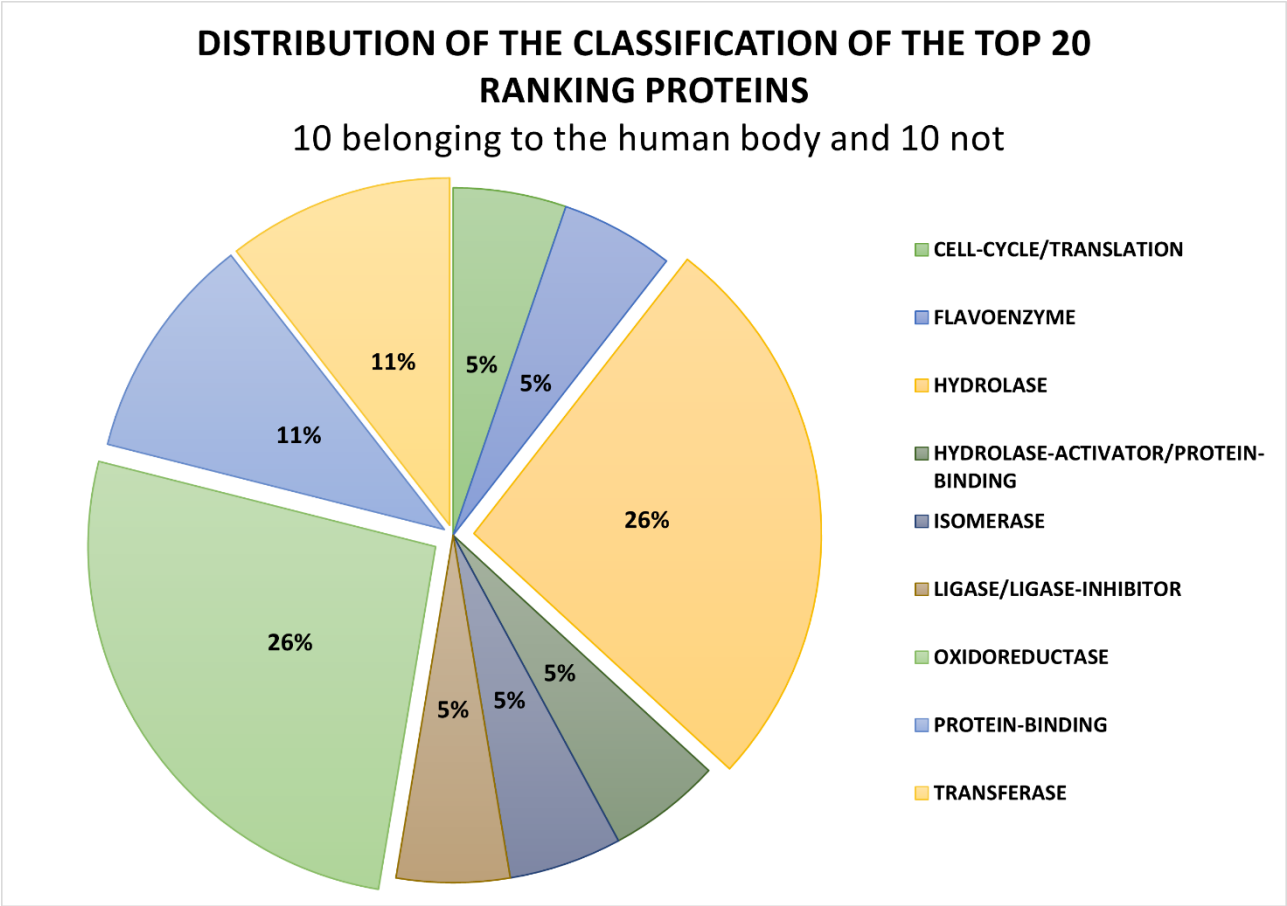


Figure 29: Distribution of the classification of proteins with better binding affinity following molecular docking analyses. The proteins in question are divided into 10 belonging to the human proteome and 10 belonging to different organisms.

The RMSD alignment of the putative binding sites on hit proteins with the corresponding reference sucrose binding site on the T1R2/T1R3 dimer was carried out by selecting (a) the binding site of the hit protein complexed with the docked ligand in the best pose and (b) the binding site of the sweet receptor corresponding to the motif for which the protein under analysis represents a hit.

The alignments were visually inspected with VMD, confirming the similarity of the two binding sites, with a good superposition of matching binding site-lining residues. *Figure 30* reports examples of this visual superposition for three hits, all belonging to the human proteome.

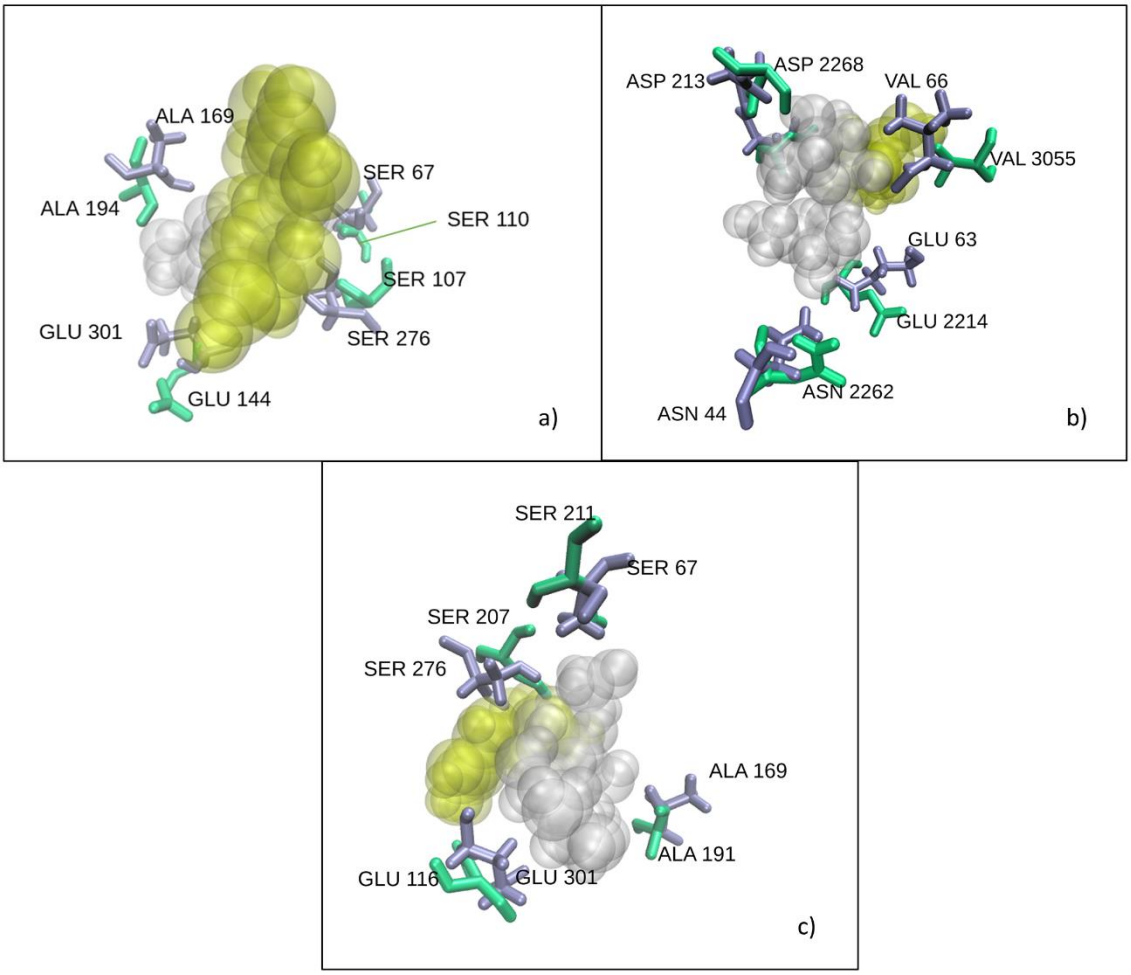


Figure 30:Overlap of the binding site of hit proteins and the corresponding binding site of the sweet receptor, based on the matching residues. The hit protein's residues are presented in green, while the sweet receptor's residues are in violet. The spheres represent the ligands, both sucrose, those in yellow are the docked ligand in the best position at the hit protein site while those in white are the sweet receptor ligand. a) Residues of hit protein 1ibr and binding site D; docking affinity = -7,0 kcal/mol; b) Residues of hit protein 2ast and binding site C; docking affinity = -6,5 kcal/mol; c) Residues of hit protein 3p8c and binding site C; docking affinity = -6,1 kcal/mol.

4.4 Discussion

Given the presence of taste receptors in regions of the human organism external to the gustatory system, where their function is not directly associated with the transduction of a taste sensation, the overarching goal of the present work is to implement an integrated and automatized methodology to perform a structure-based screening in the currently solved proteome for a specific binding site, and to deploy said methodology in a pilot investigation on the sweet taste receptor binding site, to shed light on the possible role that tastants such as sucrose may play in contexts beyond the gustatory system alone, and thus investigate the proteins not belonging to the category of taste receptors, that are able to bind such molecules.

The screening for proteins that have a binding site featuring a strong conservation of the arrangement and type of amino acid residues when compared with the sucrose binding sites present

in the sweet receptor, consisted of two main steps: first of all, after a comparative analysis of the characteristics of the currently existing methods that can perform binding site comparison, the ASSAM[53] program was selected. The main reasons are the ease with which the user can define the characteristics of the site to be searched in terms of the relative spatial position information, in PDB format, of the most significant residues, and the possibility of searching among the entire RCSB Protein Data Bank database. The term *more significant residues* herein denotes those residues, within the binding site, that are in contact with the ligand, i.e. that are involved in non-covalent interactions with the latter, such as hydrogen bonds. The set of these residues is defined as a *motif* and was obtained by analysing the protein-ligand complex with PLIP[90]. Several motifs were used as input files in the ASSAM similarity search, as obtained from different conformations extracted by means of clustering on the last 100ns of an MD trajectory of the receptor-ligand complex. Each motif is composed of 4 to 5, residues. The results of the ASSAM search are a pool of hit proteins, featuring the same residues with a highly homologous spatial arrangement with respect to those of corresponding motif. They are ranked by a similarity score given by the RMSD value of the superimposition of the matching residues. The presence of a similar spatial arrangement of residues of the same type can be indicative of a potentially similar binding site, both in terms of geometry and in terms of function.

The second phase, which consisted in extracting a subset of truly significant proteins from the ASSAM output pool composed of 2500 hit proteins, relied on a set of techniques to perform consecutive filtering phases. In detail, the SASA calculation is used as a first filter, to extract the proteins with the binding sites exposed to the solvent and, among those, to restrict the search field to those that have the binding-site shape similar to that of the respective sites present on the receptor. This operation allowed for the reduction of the size of the pool of hits by more than 60%. Molecular Docking was subsequently employed as a second filtering step, by docking the sucrose molecule, representing one natural agonist of the original sweet taste receptor binding site, to the putative binding sites found on filtered pool of hits. This analysis further narrowed the batch of proteins, as only those with negative binding energy, with a magnitude greater than 5 kcal/mol, were considered. Out of the the initial 2500 hits, the set was narrowed down to 386 proteins, approximately 15% of the size of the starting data.

Since the interest is particularly on proteins that are expressed in humans, the final set of proteins shall still be pruned by considering only (a) those directly expressed by humans in the PDB and (b) human homologues of hit proteins belonging to other animals, as found in the PDB. With these criteria, the final set of hits contained 44 non-redundant proteins, corresponding to a reduction of

the starting pool of hits by more than 98%. In the filterd set of hits (386 proteins) the analysis of the alignment of the binding sites between the hit proteins, with better binding affinity, and the sweet receptor, and the subsequent visual inspection, confirmed the actual local similarity of the binding site. Furthermore, for the hit proteins with the best predicted binding energy, a more detailed analysis of their classification and function was carried out, e.g. in terms of EC classification, revealing how most of the hits represent proteins involved in enzymatic activities, such as hydrolases, oxidoreductases or transferases. A more in-depth study of their functional characteristics, particularly for those hits belonging to the human proteome, reveals the involvement of the hit proteins in several biological processes such as cell cycle, cell division, transcription regulation, post-translational protein modification, and many others. It can be easily understood that the biological processes in which these proteins are involved are fundamental for in the lifecycle of a cell. From a more holistic perspective, it should be noted that the hit proteins are involved in functions at different hierarchical levels: (i) they play a regulatory role in the progression of the cell cycle, in particular in the G1-S and G1-M transition (hit with PDBid 2ast); (ii) they are involved in the *de-novo* biosynthesis of pyrimidine (hit with PDBid 5h2z); (iii) they are involved in the intracellular transport of proteins, in particular in the nucleocytoplasmic transport (hit with PDBid 1ibr), and in the post-translational modification of proteins such as de-ADP-ribosylation (hit with PDBid 3hfw); (iv) they are involved in the regulation of actin filament reorganization (hits with PDBid 3p8c and 4byf) which is important for processes such as cell migration, phagocytosis and intracellular motility of lipid vesicles; (v) they are involved in antiviral defence processes (hits with PDBid 2ast and 1ibr). By also considering the biological processes of homologous proteins (H.P), the field of action of the proteins involved is expanded. In particular: (vi) they are involved in the regulation of neurotransmitter release and in chemical synaptic transmission (H.P of hit with PDBid 1pk8); (vii) in the catabolic and metabolic process of glycogen (H.P of hit with PDBid 2gj4) and in the regulation of cellular metabolic processes of amino acids (H.P of hit with PDBid 4e01); (viii) in the active transport across the plasma membrane of various nutrients, in particular sodium and potassium ions and therefore maintain the cellular homeostasis of these ions and not only as these ions are involved in numerous other processes (HP of the hit with PDBid 2zxe); (ix) they are involved in the regulation of inflammatory response, in both the innate and adaptive immunity (H.P of hits with PDBid 3wpc and 4eyu); (x) are involved in purine degradation (H.P of the hit with PDBid 3nvw).

Based on these results, it seems of particular interest to further investigate the ability of these proteins to bind to sucrose in a more refined way. In particular, the goal should be to understand if the binding of sucrose to those putative hits influences to a certain extent the biological pathways in

which these proteins are involved. This could further be an innovative starting point in the development of dietary-pharmacological therapies, especially in case of involvement of these proteins in diseases or, more in general, disruption of homeostasis.

4.5 Conclusions

The present work showed an example application of structure-based ligand binding site similarity search, to elucidate the role of taste molecules in pathways beyond the perception of taste, as several experimental studies have shown the presence of taste receptors in areas of the body not belonging to the gustatory system. Therefore, a rational methodology has been developed to search for similar binding sites at the proteomic level, relying on the freely available ASSAM binding site similarity search software, followed by multiple filtering steps based on Molecular Mechanics, in particular the calculation of the SASA and Molecular Docking, to extract the most relevant proteins with respect to the binding site of the protein under investigation. The developed method was applied on a pilot investigation to the sweet receptor binding sites after extracting the key sucrose-binding residues in the receptor dimer. The final results of this study showed that most of the hits belong to classes of proteins involved in enzymatic activity (hydrolases and oxidoreductases are the most numerous): after analysing the functional characteristics of the proteins belonging to the human proteome, it is to be underlined how those are involved in several biological processes (cell cycle, cell division, post-translational protein modification, etc.). This information may shed light on the role of some tastants in the development of food-related diseases. On a higher level, the developed methodology represents an example of targeted screening technique of binding sites at the level of the entire Protein Data Bank database, with significant improvements with respect to screening procedures based solely on geometric criteria, since the screening is guided by design by the presence of specific motifs that are already be known to have a specific function, as in the case of the present work, where the function is to bind sweet molecules. Furthermore, this work presents an example of application focused on the sweet receptor and on the sucrose molecule, but the developed approach can be easily extended to other tastants, and more generally to other ligand transduction and recognition mechanisms.

5 Acknowledgements

I would like to express my gratitude to my supervisor prof. Marco Agostino Deriu for inspiring me and giving me the opportunity to conclude my university studies by working on this thesis topic. His enthusiasm motivated me during this work and increased my interest in molecular modeling.

I would like to thank my co-supervisor Eric Adriano Zizzi for his constant support and patience during the work, his help, advice in times of difficulty and his critical evaluation have been indispensable.

I would also like to thank my family, first of all, who have always supported me in the choices I have made over the years and have always been close to me in times of difficulty, helping me to face them head on and giving me the courage to do so, and my closest friends, Simona, Gabriele, Giulia, Donatella and Andrea who, despite the physical distance due to the pandemic in progress, were able to make their presence and support felt.

And finally, I want to dedicate a special thanks to my sister and my brother, without them my world would be totally different, certainly less full of the affection and closeness that binds us.

Xhesika Hada

The research has been developed as part of the VIRTUOUS project, funded by the European Union's Horizon 2020 research and innovation program under the Maria Sklodowska-Curie—RISE Grant Agreement No 872181

6 Supplementary information

Table 2: Informations about the final pool of proteins obtained after the filtering steps. In particular the corresponding motif information, matching residues between motif and hit, RMSD of the overlap and EC classification of the hit proteins belonging to the human proteome are shown.

Motif origin	Hit PDBid	Motif Matching Residues	Hit Matching Residues	RMSD [A]	EC Classification
C binding site, Centroid 1	4xoi	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	B:174:VAL, B:121:TYR, B:143:SER, B:140:THR	1,50	CELL-CYCLE
C binding site, Centroid 1	4nz6	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	A:338:VAL, A:201:TYR, A:150:SER, A:145:THR	1,38	ISOMERASE
C binding site, Centroid 1	4j37	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	A:338:VAL, A:201:TYR, A:150:SER, A:145:THR	1,51	RNA-BINDING-PROTEIN
C binding site, Centroid 1	4uwh	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	A:549:VAL, A:494:TYR, A:540:SER, A:447:THR	1,53	TRANSFERASE
C binding site, Centroid 1	1n6c	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	A:248:VAL, A:245:TYR, A:268:SER, A:266:THR	1,59	TRANSFERASE
C binding site, Centroid 1	2a74	A:66:VAL, A:103:TYR, A:212:SER, A:214:THR	D:576:VAL, D:117:TYR, D:170:SER, D:112:THR	1,07	IMMUNE-SYSTEM
C binding site, Centroid 13	3f1s	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	B:220:ASN, B:345:VAL, B:349:SER, B:328:THR	1,17	HYDROLASE-INHIBITOR/HYDROLASE
C binding site,	4byf	A:44:ASN,	A:196:ASN,	1,32	HYDROLASE

Centroid 13		A:63:GLU, A:65:LYS, A:66:VAL	A:399:GLU, A:189:LYS, A:192:VAL		
C binding site, Centroid 13	2fd6	A:44:ASN, A:65:LYS, A:66:VAL, A:214:THR	U:259:ASN, U:198:LYS, U:238:VAL, U:164:THR	1,27	IMMUNE-SYSTEM/-HYDROLASE
C binding site, Centroid 14	3f1s	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	B:220:ASN, B:345:VAL, B:349:SER, B:328:THR	1,16	HYDROLASE-INHIBITOR/HYDROLASE
C binding site, Centroid 14	3lpp	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	A:82:ASN, A:188:VAL, A:102:SER, A:104:THR	1,13	HYDROLASE
C binding site, Centroid 14	5v44	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	A:195:ASN, A:97:VAL, A:227:SER, A:220:THR	1,31	CHAPERONE
C binding site, Centroid 14	5x5o	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	A:21:ASN, A:30:VAL, A:26:SER, A:161:THR	1,32	TRANSFERASE
C binding site, Centroid 14	4pxz	A:44:ASN, A:65:LYS, A:66:VAL, A:214:THR	A:201:ASN, A:237:LYS, A:238:VAL, A:126:THR	1,35	MEMBRANE-PROTEIN
C binding site, Centroid 14	2fd6	A:44:ASN, A:65:LYS, A:66:VAL, A:214:THR	U:259:ASN, U:198:LYS, U:238:VAL, U:164:THR	1,30	IMMUNE-SYSTEM/-HYDROLASE
C binding site, Centroid 15	3f1s	A:44:ASN, A:66:VAL, A:212:SER,	B:220:ASN, B:345:VAL, B:349:SER,	1,22	HYDROLASE-INHIBITOR/HYDROLASE

		A:214:THR	B:328:THR		
C binding site, Centroid 15	4may	A:44:ASN, A:63:GLU, A:66:VAL, A:212:SER	D:60:ASN, D:81:GLU, D:17:VAL, D:85:SER	1,29	IMMUNE-SYSTEM
C binding site, Centroid 15	4may	A:44:ASN, A:63:GLU, A:66:VAL, A:214:THR	D:60:ASN, D:81:GLU, D:17:VAL, D:105:THR	1,30	IMMUNE-SYSTEM
C binding site, Centroid 15	3i2n	A:44:ASN, A:63:GLU, A:66:VAL, A:214:THR	A:225:ASN, A:273:GLU, A:311:VAL, A:277:THR	1,31	TRANSCRIPTION
C binding site, Centroid 15	2fd6	A:44:ASN, A:65:LYS, A:66:VAL, A:214:THR	U:259:ASN, U:198:LYS, U:238:VAL, U:164:THR	1,28	IMMUNE-SYSTEM/-HYDROLASE
C binding site, Centroid 16	3f1s	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	B:220:ASN, B:345:VAL, B:349:SER, B:328:THR	1,27	HYDROLASE-INHIBITOR/HYDROLASE
C binding site, Centroid 16	2fd6	A:44:ASN, A:65:LYS, A:66:VAL, A:214:THR	U:259:ASN, U:198:LYS, U:238:VAL, U:164:THR	1,28	IMMUNE-SYSTEM/-HYDROLASE
C binding site, Centroid 18	3e0c	A:63:GLU, A:65:LYS, A:212:SER, A:213:ASP, A:214:THR	A:312:GLU, A:35:LYS, A:331:SER, A:330:ASP, A:354:THR	1,74	DNA-BINDING-PROTEIN
C binding site, Centroid 20	2ast	A:44:ASN, A:63:GLU, A:66:VAL, A:212:SER,	B:2262:ASN, B:2214:GLU, C:3055:VAL, B:2241:SER,	1,60	LIGASE/LIGASE-INHIBITOR

		A:213:ASP	B:2268:ASP		
C binding site, Centroid 20	3f1s	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	B:220:ASN, B:345:VAL, B:349:SER, B:328:THR	1,13	HYDROLASE-INHIBITOR/HYDROLASE
C binding site, Centroid 20	1ivh	A:44:ASN, A:63:GLU, A:66:VAL, A:214:THR	A:11:ASN, A:84:GLU, A:82:VAL, A:23:THR	1,24	OXIDOREDUCTASE
C binding site, Centroid 21	1z70	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	X:1352:ASN, X:1157:VAL, X:1333:SER, X:1278:THR	1,38	OXIDOREDUCTASE
C binding site, Centroid 21	2aii	A:44:ASN, A:66:VAL, A:212:SER, A:214:THR	X:352:ASN, X:157:VAL, X:333:SER, X:278:THR	1,38	HYDROLASE-ACTIVATOR/PROTEIN-BINDING
C binding site, Centroid 22	2yxt	A:44:ASN, A:66:VAL, A:103:TYR, A:214:THR	A:150:ASN, A:128:VAL, A:136:TYR, A:85:THR	1,23	TRANSFERASE
C binding site, Centroid 3	4xos	A:63:GLU, A:211:SER, A:214:THR, A:242:THR	A:81:GLU, A:77:SER, A:55:THR, A:105:THR	1,50	ANTITUMOR-PROTEIN
C binding site, Centroid 3	5mj6	A:63:GLU, A:211:SER, A:214:THR, A:242:THR	A:285:GLU, A:278:SER, A:186:THR, A:280:THR	1,53	HYDROLASE
C binding site, Centroid 4	4xoi	A:63:GLU, A:103:TYR, A:212:SER, A:214:THR	B:47:GLU, B:121:TYR, B:143:SER, B:140:THR	1,21	CELL-CYCLE
C binding site,	4rfq	A:63:GLU,	A:170:GLU,	1,28	TRANSFERASE

Centroid 4		A:103:TYR, A:211:SER, A:212:SER	A:297:TYR, A:84:SER, A:83:SER		
C binding site, Centroid 4	3fed	A:63:GLU, A:103:TYR, A:211:SER, A:212:SER	A:489:GLU, A:527:TYR, A:482:SER, A:473:SER	1,34	HYDROLASE
C binding site, Centroid 4	5h2z	A:63:GLU, A:103:TYR, A:211:SER, A:212:SER	A:117:GLU, A:195:TYR, A:215:SER, A:214:SER	1,34	OXIDOREDUCTASE
D binding site, Centroid 1	3mdj	B:67:SER, B:169:ALA, B:216:ASP, B:276:SER, B:301:GLU	A:343:SER, A:279:ALA, A:335:ASP, A:339:SER, A:329:GLU	1,72	HYDROLASE/HYDROLASE-INHIBITOR
D binding site, Centroid 1	3hfw	B:67:SER, B:169:ALA, B:170:SER, B:301:GLU	A:264:SER, A:274:ALA, A:305:SER, A:25:GLU	1,00	HYDROLASE
D binding site, Centroid 1	1ibr	B:67:SER, B:169:ALA, B:276:SER, B:301:GLU	B:110:SER, B:194:ALA, B:107:SER, B:144:GLU	1,09	CELL-CYCLE/TRANSLATION
D binding site, Centroid 1	2vqm	B:67:SER, B:169:ALA, B:170:SER, B:216:ASP	A:346:SER, A:306:ALA, A:305:SER, A:354:ASP	1,11	HYDROLASE
D binding site, Centroid 1	3b6h	B:67:SER, B:169:ALA, B:276:SER, B:301:GLU	A:118:SER, A:98:ALA, A:116:SER, A:105:GLU	1,12	ISOMERASE
D binding site, Centroid 1	3p8c	B:67:SER, B:169:ALA,	A:211:SER, A:191:ALA,	1,13	PROTEIN-BINDING

		B:276:SER, B:301:GLU	A:207:SER, A:116:GLU		
D binding site, Centroid 1	2j6l	B:67:SER, B:169:ALA, B:170:SER, B:301:GLU	A:429:SER, A:468:ALA, A:481:SER, A:122:GLU	1,16	OXIDOREDUCTASE
D binding site, Centroid 1	1aye	B:169:ALA, B:170:SER, B:216:ASP, B:276:SER	A:250:ALA, A:251:SER, A:142:ASP, A:162:SER	1,14	SERINE-PROTEASE

Table 3: Information on the final protein pool obtained after the filtering steps.In particularthe corresponding motif information, residues of matching between motif and hit, RMSD of the overlap and EC classification of the hit proteins of which the human homolog has been found are shown.

Motif origin	**Hit PDBid**	**Motif Matching Residues**	**Hit Matching Residues**	**RMSD [A]**	**EC Classification**
C binding site, Centroid 4	4c2m	A:63:GLU, A:211:SER, A:212:SER, A:214:THR	A:1195:GLU, A:1578:SER, A:662:SER, A:1058:THR	1,34	TRANSCRIPTION
C binding site, Centroid 4	3wpc	A:63:GLU, A:103:TYR, A:211:SER, A:212:SER	A:464:GLU, A:554:TYR, A:636:SER, A:606:SER	1,21	DNA-BINDING-PROTEIN/DNA
C binding site, Centroid 4	4ii2	A:63:GLU, A:103:TYR, A:211:SER, A:212:SER	A:30:GLU, A:20:TYR, A:38:SER, A:36:SER	1,34	LIGASE
C binding site, Centroid 13	4c2m	A:44:ASN, A:63:GLU, A:66:VAL, A:214:THR	A:493:ASN, A:491:GLU, A:809:VAL, A:740:THR	1,2	TRANSCRIPTION
C binding site, Centroid 14	4eyu	A:44:ASN, A:66:VAL, A:212:SER,	A:1182:ASN, A:1455:VAL, A:1253:SER,	1,4	OXIDOREDUCTASE

		A:214:THR	A:1475:THR		
C binding site, Centroid 15	4e01	A:44:ASN, A:63:GLU, A:66:VAL, A:214:THR	A:89:ASN, A:141:GLU, A:136:VAL, A:342:THR	1,22	TRANSFERASE/TRANSFERASE-INHIBITOR
C binding site, Centroid 18	2aka	A:63:GLU, A:65:LYS, A:66:VAL, A:213:ASP	A:467:GLU, A:265:LYS, A:268:VAL, A:419:ASP	1,24	CONTRACTILE-PROTEIN
C binding site, Centroid 18	3nvw	A:63:GLU, A:66:VAL, A:212:SER, A:213:ASP, A:214:THR	C:1261:GLU, C:1259:VAL, C:1082:SER, C:1084:ASP, C:1083:THR	1,54	OXIDOREDUCTASE
C binding site, Centroid 18	2gj4	A:63:GLU, A:66:VAL, A:213:ASP, A:214:THR	A:716:GLU, A:718:VAL, A:693:ASP, A:671:THR	1,16	TRANSFERASE
C binding site, Centroid 22	1pk8	A:44:ASN, A:66:VAL, A:212:SER, A:213:ASP, A:214:THR	A:322:ASN, A:372:VAL, A:361:SER, A:358:ASP, A:359:THR	1,73	MEMBRANE-PROTEIN
C binding site, Centroid 22	2zxe	A:44:ASN, A:66:VAL, A:213:ASP, A:214:THR	A:384:ASN, A:719:VAL, A:214:ASP, A:247:THR	1,22	HYDROLASE/TRANSPORT-PROTEIN

7 References

(1) Calvo, S. S. C.; Egan, J. M. The Endocrinology of Taste Receptors. *Nature Reviews Endocrinology*. Nature Publishing Group April 30, 2015, pp 213–227. https://doi.org/10.1038/nrendo.2015.7.

(2) Loper, H. B.; la Sala, M.; Dotson, C.; Steinle, N. Taste Perception, Associated Hormonal Modulation, and Nutrient Intake. *Nutrition Reviews* **2015**, *73* (2), 83–91. https://doi.org/10.1093/nutrit/nuu009.

(3) Kochem, M. Type 1 Taste Receptors in Taste and Metabolism. *Annals of Nutrition and Metabolism* **2017**, *70* (3), 27–36. https://doi.org/10.1159/000478760.

(4) Reimann, F.; Tolhurst, G.; Gribble, F. M. G-Protein-Coupled Receptors in Intestinal Chemosensation. *Cell Metabolism*. April 4, 2012, pp 421–431. https://doi.org/10.1016/j.cmet.2011.12.019.

(5) Breslin, P. A. S. An Evolutionary Perspective on Food and Human Taste. *Current Biology*. May 6, 2013. https://doi.org/10.1016/j.cub.2013.04.010.

(6) Betts J. Gordon; Young Kelly A.; Wise James A.; Johnson Eddie; Poe Brandon; Kruse Dean H.; Korol Oksana; Johnson Jody E.; Womble Mark; DeSaix Peter. *Anatomy and Physiology*; OpenStax: Houston, Texas, 2013.

(7) Segovia, C.; Hutchinson, I.; Laing, D. G.; Jinks, A. L. *A Quantitative Study of Fungiform Papillae and Taste Pore Density in Adults and Children*; 2002; Vol. 138.

(8) Witt, M. Anatomy and Development of the Human Taste System. In *Handbook of Clinical Neurology*; Elsevier B.V., 2019; Vol. 164, pp 147–171. https://doi.org/10.1016/B978-0-444-63855-7.00010-1.

(9) Roper, S. D. Taste Buds as Peripheral Chemosensory Processors. *Seminars in Cell and Developmental Biology*. Elsevier Ltd 2013, pp 71–79. https://doi.org/10.1016/j.semcdb.2012.12.002.

(10) Besnard, P.; Passilly-Degrace, P.; Khan, N. A. Taste of Fat: A Sixth Taste Modality? *Physiol Rev* **2016**, *96*, 151–176. https://doi.org/10.1152/physrev.00002.2015.-An.

(11) Chaudhari, N.; Roper, S. D. The Cell Biology of Taste. *Journal of Cell Biology*. August 9, 2010, pp 285–296. https://doi.org/10.1083/jcb.201003144.

(12) Witt, M.; Reutter, K. *Anatomy of the Tongue and Taste Buds*; 2015.

(13) Luddi, A.; Governini, L.; Wilmsköttter, D.; Gudermann, T.; Boekhoff, I.; Piomboni, P. Taste Receptors: New Players in Sperm Biology. *International Journal of Molecular Sciences*. MDPI AG February 2, 2019. https://doi.org/10.3390/ijms20040967.

(14) Niki M; Yoshida R; Takai S; Ninomiya Y. Gustatory Signaling in the Periphery: Detection, Transmission, and Modulation of Taste Information. *Biol Pharm Bull.* **2010**, *33* (11), 1772–1777. https://doi.org/10.1248/bpb.33.1772.

(15) Hu, G. M.; Mai, T. L.; Chen, C. M. Visualizing the GPCR Network: Classification and Evolution. *Scientific Reports* **2017**, *7* (1). https://doi.org/10.1038/s41598-017-15707-9.

(16) Kniazeff, J.; Prézeau, L.; Rondard, P.; Pin, J. P.; Goudet, C. Dimers and beyond: The Functional Puzzles of Class C GPCRs. *Pharmacology and Therapeutics*. Elsevier Inc. 2011, pp 9–25. https://doi.org/10.1016/j.pharmthera.2011.01.006.

(17) Ellaithy, A.; Gonzalez-Maeso, J.; Logothetis, D. A.; Levitz, J. Structural and Biophysical Mechanisms of Class C G Protein-Coupled Receptor Function. *Trends in Biochemical Sciences*. Elsevier Ltd December 1, 2020, pp 1049–1064. https://doi.org/10.1016/j.tibs.2020.07.008.

(18) Patel, B. S.; Ravix, J.; Pabelick, C.; Prakash, Y. S. Class C GPCRs in the Airway. *Current Opinion in Pharmacology*. Elsevier Ltd April 1, 2020, pp 19–28. https://doi.org/10.1016/j.coph.2020.04.002.

(19) Kobayashi, Y.; Habara, M.; Ikezazki, H.; Chen, R.; Naito, Y.; Toko, K. Advanced Taste Sensors Based on Artificial Lipids with Global Selectivity to Basic Taste Qualities and High Correlation to Sensory Scores. *Sensors*. April 2010, pp 3411–3443. https://doi.org/10.3390/s100403411.

(20) Pallante, L.; Malavolta, M.; Grasso, G.; Korfiati, A.; Mavroudi, S.; Mavkov, B.; Kalogeras, A.; Alexakos, C.; Martos, V.; Amoroso, D.; di Benedetto, G.; Piga, D.; Theofilatos, K.; Deriu, M. A. On the Human Taste Perception: Molecular-Level Understanding Empowered by Computational Methods. *Trends in Food Science and Technology*. Elsevier Ltd October 1, 2021, pp 445–459. https://doi.org/10.1016/j.tifs.2021.07.013.

(21) Kim, S. K.; Chen, Y.; Abrol, R.; Goddard, W. A.; Guthrie, B. Activation Mechanism of the G Protein-Coupled Sweet Receptor Heterodimer with Sweeteners and Allosteric Agonists.

Proceedings of the National Academy of Sciences of the United States of America **2017**, *114* (10), 2568–2573. https://doi.org/10.1073/pnas.1700001114.

(22) Ikeda, K. *TRANSLATION New Seasonings*; 2002.

(23) Gravina, S. A.; Yep, G. L.; Khan, M. Human Biology of Taste. *Annals of Saudi Medicine*. May 2013, pp 217–222. https://doi.org/10.5144/0256-4947.2013.217.

(24) Töle, J. C.; Behrens, M.; Meyerhof, W. Taste Receptor Function. In *Handbook of Clinical Neurology*; Elsevier B.V., 2019; Vol. 164, pp 173–185. https://doi.org/10.1016/B978-0-444-63855-7.00011-3.

(25) Carey, R. M.; Lee, R. J. Taste Receptors in Upper Airway Innate Immunity. *Nutrients*. MDPI AG September 1, 2019. https://doi.org/10.3390/nu11092017.

(26) Schild, L. The Epithelial Sodium Channel and the Control of Sodium Balance. *Biochimica et Biophysica Acta - Molecular Basis of Disease*. December 2010, pp 1159–1165. https://doi.org/10.1016/j.bbadis.2010.06.014.

(27) Ishimaru, Y.; Inada, H.; Kubota, M.; Zhuang, H.; Tominaga, M.; Matsunami, H. *Transient Receptor Potential Family Members PKD1L3 and PKD2L1 Form a Candidate Sour Taste Receptor*; 2006.

(28) Kobayashi, Y.; Habara, M.; Ikezazki, H.; Chen, R.; Naito, Y.; Toko, K. Advanced Taste Sensors Based on Artificial Lipids with Global Selectivity to Basic Taste Qualities and High Correlation to Sensory Scores. *Sensors*. April 2010, pp 3411–3443. https://doi.org/10.3390/s100403411.

(29) Chang, R. B.; Waters, H.; Liman, E. R. A Proton Current Drives Action Potentials in Genetically Identified Sour Taste Cells. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107* (51), 22320–22325. https://doi.org/10.1073/pnas.1013664107.

(30) Iwatsuki, K.; Uneyama, H. Sense of Taste in the Gastrointestinal Tract. *Journal of Pharmacological Sciences*. Japanese Pharmacological Society 2012, pp 123–128. https://doi.org/10.1254/jphs.11R08CP.

(31) Sternini, C.; Anselmi, L.; Rozengurt, E. Enteroendocrine Cells: A Site of “taste” in Gastrointestinal Chemosensing. *Current Opinion in Endocrinology, Diabetes and Obesity*. February 2008, pp 73–78. https://doi.org/10.1097/MED.0b013e3282f43a73.

(32) Finger, T. E.; Kinnamon, S. C. Taste Isn't Just for Taste Buds Anymore. *F1000 Biology Reports*. September 1, 2011. https://doi.org/10.3410/B3-20.

(33) Fletcher, R.; Powell, M. J. D. *A Rapidly Convergent Descent Method for Minimization*.

(34) Mccarthy, J. F. *Rapid Communications Block-Conjugate-Gradient Method*; 1989; Vol. 40.

(35) Verbeke, J.; Cools, R. The Newton-Raphson Method. *International Journal of Mathematical Education in Science and Technology* **1995**, *26* (2), 177–193. https://doi.org/10.1080/0020739950260202.

(36) Con, M. L. *Solvent-Accessible Surfaces Proteins and Nucleic Ac*; 1983; Vol. 221.

(37) Shrake, A.; Rupley~, J. A. *Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin*; 1973; Vol. 79.

(38) Richmond, T. T. *Solvent Accessible Surface Area and Excluded Volume in Proteins Analytical Equations for Overlapping Spheres and Implications for the Hydrophobic Effect*; 1984; Vol. 178.

(39) Connoh, H. L. *Analytical Molecular Surface Calculation*; 1983; Vol. 548.

(40) Humphrey, W.; Dalke, A.; Schulten, K. *VMD: Visual Molecular Dynamics*; 1996.

(41) Schrödinger, L.; DeLano, W. The PyMOL Molecular Graphics System. Schrödinger 2021.

(42) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(43) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *Journal of Computational Chemistry* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(44) Chemical Computing Group ULC. Molecular Operating Environment (MOE). Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7 2021.

(45) ChemAxon Ltd. ChemAxon https://chemaxon.com/company (accessed 2021 -10 -10).

(46) Morris, G. M.; Lim-Wilby, M. *Molecular Docking*.

(47) A Ewing, T. J.; Makino, S.; Geoffrey Skillman, A.; Kuntz, I. D. *DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases*; 2001; Vol. 15.

(48) Morris, G. M.; Ruth, H.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Software News and Updates AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(49) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2009**, NA-NA. https://doi.org/10.1002/jcc.21334.

(50) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. *AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings*.

(51) Wang, R.; Lai, L.; Wang, S. *Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction*; 2002; Vol. 16.

(52) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21* (6), 1087–1092. https://doi.org/10.1063/1.1699114.

(53) Nadzirin, N.; Gardiner, E. J.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. SPRITE and ASSAM: Web Servers for Side Chain 3D-Motif Searching in Protein Structures. *Nucleic Acids Research* **2012**, *40* (W1), 380–386. https://doi.org/10.1093/nar/gks401.

(54) Spriggs, R. v.; Artymiuk, P. J.; Willett, P. Searching for Patterns of Amino Acids in 3D Protein Structures. *Journal of Chemical Information and Computer Sciences* **2003**, *43* (2), 412–421. https://doi.org/10.1021/ci0255984.

(55) Raka, F.; Farr, S.; Kelly, J.; Stoianov, A.; Adeli, K. REVIEW Role of Gut Microbiota and Gut-Brain and Gut-Liver Axes in Physiological Regulation of Inflammation, Energy Balance, and Metabolism Metabolic Control via Nutrient-Sensing Mechanisms: Role of Taste Receptors and the Gut-Brain Neuroendocrine Axis. *Am J Physiol Endocrinol Metab* **2019**, *317*, 559–572. https://doi.org/10.1152/ajpendo.00036.2019.-Nutrient.

(56) Ehrt, C.; Brinkjost, T.; Koch, O. *A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs)*; 2018; Vol. 14. https://doi.org/10.1371/journal.pcbi.1006483.

(57) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins: Structure, Function and Genetics* **2003**, *52* (2), 137–145. https://doi.org/10.1002/prot.10339.

(58) Stark, A.; Sunyaev, S.; Russell, R. B. A Model for Statistical Significance of Local Similarities in Structure. *Journal of Molecular Biology* **2003**, *326* (5), 1307–1316. https://doi.org/10.1016/S0022-2836(03)00045-7.

(59) KINOSHITA, K. Identification of Protein Biochemical Function by Searching the Similar Shape and Electrostatic Potential on the Molecular Surface of Proteins. *Seibutsu Butsuri* **2004**, *44* (4), 150–154. https://doi.org/10.2142/biophys.44.150.

(60) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Research* **2005**, *33* (7), 2302–2309. https://doi.org/10.1093/nar/gki524.

(61) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: Recognition and Comparison of Binding Sites and Protein-Protein Interfaces. *Nucleic Acids Research* **2005**, *33* (SUPPL. 2), 337–341. https://doi.org/10.1093/nar/gki482.

(62) Lisewski, A. M.; Lichtarge, O. Rapid Detection of Similarity in Protein Structure and Function through Contact Metric Distances. *Nucleic Acids Research* **2006**, *34* (22), 1–10. https://doi.org/10.1093/nar/gkl788.

(63) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinformatics* **2008**, *9*, 1–17. https://doi.org/10.1186/1471-2105-9-543.

(64) Nagarajan, D.; Chandra, N. PocketMatch (Version 2.0): A Parallel Algorithm for the Detection of Structural Similarities between Protein Ligand Binding-Sites. *2013 National Conference on Parallel Computing Technologies, PARCOMPTECH 2013* **2013**, 0–5. https://doi.org/10.1109/ParCompTech.2013.6621397.

(65) Shulman-peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. MultiBind and MAPPIS : Webservers for Multiple Alignment of Protein 3D-Binding Sites and Their Interactions. **2008**, *36* (May), 260–264. https://doi.org/10.1093/nar/gkn185.

(66) Tseng, Y. Y.; Dundas, J.; Liang, J. Predicting Protein Function and Binding Profile via Matching of Local Evolutionary and Geometric Surface Patterns. *Journal of Molecular Biology* **2009**, *387* (2), 451–464. https://doi.org/10.1016/j.jmb.2008.12.072.

(67) Tseng, Y. Y.; Chen, Z. J.; Li, W. H. FPOP: Footprinting Functional Pockets of Proteins by Comparative Spatial Patterns. *Nucleic Acids Research* **2009**, *38* (SUPPL.1), 288–295. https://doi.org/10.1093/nar/gkp900.

(68) Das, S.; Krein, M. P.; Breneman, C. M. PESDserv: A Server for High-Throughput Comparison of Protein Binding Site Surfaces. *Bioinformatics* **2010**, *26* (15), 1913–1914. https://doi.org/10.1093/bioinformatics/btq288.

(69) Standley, D. M.; Yamashita, R.; Kinjo, A. R.; Toh, H.; Nakamura, H. SeSAW: Balancing Sequence and Structural Information in Protein Functional Mapping. *Bioinformatics* **2010**, *26* (9), 1258–1259. https://doi.org/10.1093/bioinformatics/btq116.

(70) Moll, M.; Bryant, D. H.; Kavraki, L. E. The LabelHash Algorithm for Substructure Matching. *BMC Bioinformatics* **2010**, *11*. https://doi.org/10.1186/1471-2105-11-555.

(71) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *Journal of Chemical Information and Modeling* **2010**, *50* (1), 123–135. https://doi.org/10.1021/ci900349y.

(72) Konc, J.; Janežič, D. ProBiS-2012: Web Server and Web Services for Detection of Structurally Similar Binding Sites in Proteins. *Nucleic Acids Research* **2012**, *40* (W1), 214–221. https://doi.org/10.1093/nar/gks435.

(73) Ito, J. I.; Tabei, Y.; Shimizu, K.; Tsuda, K.; Tomii, K. PoSSuM: A Database of Similar Protein-Ligand Binding and Putative Pockets. *Nucleic Acids Research* **2012**, *40* (D1), 541–548. https://doi.org/10.1093/nar/gkr1130.

(74) Roy, A.; Yang, J.; Zhang, Y. COFACTOR: An Accurate Comparative Algorithm for Structure-Based Protein Function Annotation. *Nucleic Acids Research* **2012**, *40* (W1), 471–477. https://doi.org/10.1093/nar/gks372.

(75) Lin, Y.; Yoo, S.; Sanchez, R. SiteComp: A Server for Ligand Binding Site Analysis in Protein Structures. *Bioinformatics* **2012**, *28* (8), 1172–1173. https://doi.org/10.1093/bioinformatics/bts095.

(76) Kurbatova, N.; Chartier, M.; Zylber, M. I.; Najmanovich, R. IsoCleft Finder – a Web-Based Tool for the Detection and Analysis of Protein Binding-Site Geometric and Chemical Similarities [v2 ; Ref Status : Indexed , Http://F1000r.Es/13y]. **2014**, *2013* (May 2013), 1–13. https://doi.org/10.12688/f1000research.2-117.v1.

(77) Kirshner, D. A.; Nilmeier, J. P.; Lightstone, F. C. Catalytic Site Identification--a Web Server to Identify Catalytic Site Structural Matches throughout PDB. *Nucleic acids research* **2013**, *41* (Web Server issue), 256–265. https://doi.org/10.1093/nar/gkt403.

(78) Nadzirin, N.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. IMAAAGINE: A Webserver for Searching Hypothetical 3D Amino Acid Side Chain Arrangements in the Protein Data Bank. *Nucleic acids research* **2013**, *41* (Web Server issue), 432–440. https://doi.org/10.1093/nar/gkt431.

(79) Gao, M.; Skolnick, J. Structural Bioinformatics APoc : Large-Scale Identification of Similar Protein Pockets. **2013**, *29* (5), 597–604. https://doi.org/10.1093/bioinformatics/btt024.

(80) Caprari, S.; Toti, D.; Viet Hung, L.; di Stefano, M.; Polticelli, F. ASSIST: A Fast Versatile Local Structural Comparison Tool. *Bioinformatics* **2014**, *30* (7), 1022–1024. https://doi.org/10.1093/bioinformatics/btt664.

(81) Caprari, S.; Toti, D.; Viet Hung, L.; di Stefano, M.; Polticelli, F. *Supplementary Materials ASSIST: A Fast Versatile Local Structural Comparison Tool*; 2005.

(82) Chartier, M.; Najmanovich, R. Detection of Binding Site Molecular Interaction Field Similarities. *Journal of Chemical Information and Modeling* **2015**, *55* (8), 1600–1615. https://doi.org/10.1021/acs.jcim.5b00333.

(83) Chartier, M.; Adriansen, E.; Najmanovich, R. IsoMIF Finder: Online Detection of Binding Site Molecular Interaction Field Similarities. *Bioinformatics* **2016**, *32* (4), 621–623. https://doi.org/10.1093/bioinformatics/btv616.

(84) Lee, H. S.; Im, W. G-LoSA: An Efficient Computational Tool for Local Structure-Centric Biological Studies and Drug Design. *Protein Science* **2016**, *25* (4), 865–876. https://doi.org/10.1002/pro.2890.

(85) Núñez-Vivanco, G.; Valdés-Jiménez, A.; Besoaín, F.; Reyes-Parada, M. Geomfinder: A Multi-Feature Identifier of Similar Three-Dimensional Protein Patterns: A Ligand-Independent Approach. *Journal of Cheminformatics* **2016**, *8* (1), 1–15. https://doi.org/10.1186/s13321-016-0131-9.

(86) Rey, J.; Rasolohery, I.; Tufféry, P.; Guyon, F.; Moroy, G. PatchSearch: A Web Server for off-Target Protein Identification. *Nucleic Acids Research* **2019**, *47* (W1), W365–W372. https://doi.org/10.1093/nar/gkz478.

(87) Ab Ghani, N. S.; Ramlan, E. I.; Firdaus-Raih, M. Drug ReposER: A Web Server for Predicting Similar Amino Acid Arrangements to Known Drug Binding Interfaces for Potential Drug Repositioning. *Nucleic Acids Research* **2019**, *47* (W1), W350–W356. https://doi.org/10.1093/nar/gkz391.

(88) Simonovsky, M.; Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *Journal of chemical information and modeling* **2020**, *60* (4), 2356–2366. https://doi.org/10.1021/acs.jcim.9b00554.

(89) Koehl, A.; Hu, H.; Feng, D.; Sun, B.; Zhang, Y.; Robertson, M. J.; Chu, M.; Kobilka, T. S.; Laermans, T.; Steyaert, J.; Tarrasch, J.; Dutta, S.; Fonseca, R.; Weis, W. I.; Mathiesen, J. M.; Skiniotis, G.; Kobilka, B. K. Structural Insights into the Activation of Metabotropic Glutamate Receptors. *Nature* **2019**, *566* (7742), 79–84. https://doi.org/10.1038/s41586-019-0881-4.

(90) Adasme, M. F.; Linnemann, K. L.; Bolz, S. N.; Kaiser, F.; Salentin, S.; Haupt, V. J.; Schroeder, M. PLIP 2021: Expanding the Scope of the Protein-Ligand Interaction Profiler to DNA and RNA. *Nucleic Acids Research* **2021**, *49* (W1), W530–W534. https://doi.org/10.1093/nar/gkab294.

(91) Altschup, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *Basic Local Alignment Search Tool*; 1990; Vol. 215.

(92) Agarwala, R.; Barrett, T.; Beck, J.; Benson, D. A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J. R.; Bryant, S. H.; Canese, K.; Cavanaugh, M.; Charowhas, C.; Clark, K.; Dondoshansky, I.; Feolo, M.; Fitzpatrick, L.; Funk, K.; Geer, L. Y.; Gorelenkov, V.; Graeff, A.; Hlavina, W.; Holmes, B.; Johnson, M.; Kattman, B.; Khotomlianski, V.; Kimchi, A.; Kimelman, M.; Kimura, M.; Kitts, P.; Klimke, W.; Kotliarov, A.; Krasnov, S.; Kuznetsov, A.; Landrum, M. J.; Landsman, D.; Lathrop, S.; Lee, J. M.; Leubsdorf, C.; Lu, Z.; Madden, T. L.; Marchler-Bauer, A.; Malheiro, A.; Meric, P.; Karsch-Mizrachi, I.; Mnev, A.; Murphy, T.; Orris, R.; Ostell, J.; O'Sullivan, C.; Palanigobu, V.; Panchenko, A. R.; Phan, L.; Pierov, B.; Pruitt, K. D.; Rodarmer, K.; Sayers, E. W.; Schneider, V.; Schoch, C. L.; Schuler, G. D.; Sherry, S. T.; Siyan, K.; Soboleva, A.; Soussov, V.; Starchenko, G.; Tatusova, T. A.; Thibaud-Nissen, F.; Todorov, K.; Trawick, B. W.; Vakatov, D.; Ward, M.; Yaschenko, E.; Zasypkin, A.; Zbicz, K. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2018**, *46* (D1), D8–D13. https://doi.org/10.1093/nar/gkx1095.

(93) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Silva, A. da; Denny, P.; Dogan, T.; Ebenezer, T. G.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L. S.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research* **2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(94) O'boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Open Babel: An Open Chemical Toolbox*; 2011.

(95) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Research* **2018**, *46* (W1), W296–W303. https://doi.org/10.1093/nar/gky427.