Doctoral Dissertation

Doctoral Program in Bioengineering and Medical-Surgical Sciences (35th Cycle)

Molecular Level Insights into Taste Perception and Beyond

By

Lorenzo Pallante

Supervisor(s):

Prof. Marco A. Deriu, Supervisor

Prof. Umberto Morbiducci, Co-Supervisor

Doctoral Examination Committee:

Prof. Antonella Di Pizio, Referee, Leibniz Institute for Food Systems Biology, Technical University of Munich, Freising 85354, Germany

Dr. Francesco Gentile, Referee, Department of Chemistry and Biomolecular Sciences, University of Ottawa, Canada

Politecnico di Torino – Università degli Studi di Torino

2023

Declaration

I hereby declare that the contents and organization of this dissertation constitute my original work and do not compromise in any way the rights of third parties, including those relating to the security of personal data.

This doctoral dissertation is based on material that was previously published and for which the author holds the right for inclusion. Original publications are acknowledged in the pertinent chapters.

Lorenzo Pallante

Turin, 2023

* This dissertation is presented in partial fulfilment of the requirements for the **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo) and the University of Turin (UniTo).

alla mia famiglia, in tutte le sue forme…

Acknowledgements

I would like to express my sincerest gratitude to my supervisors at Politecnico di Torino, namely Prof. Marco A. Deriu and Prof. Umberto Morbiducci. Their unwavering support, insightful discussions, and expert guidance were paramount to my accomplishments during these years. I am also immensely grateful for their dedication to fostering a healthy and collaborative work environment, as well as for their efforts in facilitating safe international mobility periods, even amidst the challenges posed by the global pandemic throughout my PhD years.

I am immensely grateful to the individuals who granted me the opportunity to participate in the VIRTUOUS project and engage in mobility periods abroad, specifically in Switzerland and Greece. These experiences not only fostered significant professional and personal development but also facilitated the establishment of invaluable and fruitful collaborations that extend into the future. I wish to express my heartfelt appreciation to Konstantinos Theofilatos from Insybio, Emanuele Mottola from Missing Tech, Gianvito Grasso and Dario Piga from Dalle Molle Institute for Artificial Intelligence (IDSIA - USI/SUPSI), as well as all the talented young researchers I encountered throughout this journey, including Lampros, Aigli, Gabriele, Filip, Stefano, Akis, and many others.

A special thank you goes out to all the lab members and colleagues at Politecnico, including Michela, Eric, Marcello, Marco, Kostas, and Taimoor. These individuals played a significant role in mitigating the frustrations and challenges of the PhD journey, while also contributing greatly to my technical skills. I cherish the friendships and collaborations have formed and look forward to many more years of working together.

I am grateful for the talented young researchers at the Department of Mechanical and Aerospace Engineering at Polito, particularly the 5th floor gang and all its past and present members. Their expertise, kindness, and support made each day in the office a joy, and they have set an example for me as both researchers and human beings since the beginning of my career.

Finally, I am indebted to all the people who have been there for me since the beginning of my PhD, through ups and downs, and who have always been there at important times in my life. For this I thank my family, in all its forms and facets, from blood ties to lifelong relationships, which endure through the years and will forever be lifeblood in my journey.

The present PhD Thesis has been developed as part of the VIRTUOUS project, funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie-RISE Grant Agreement No. 872181 (https://www.virtuoush2020.com/).

Abstract

Taste perception is a complex and multi-layered process which involves several components from the molecular level up to subcellular, cellular, and tissue levels. At the molecular scale, taste perception is triggered by interactions between food compounds and taste receptors located on the tongue's papillae. Taste is the primary driver of food intake since the five basic taste sensations, i.e. sweet, bitter, umami, sour and salty, are related to specific nutritional needs or control strategies. Understanding the molecular features and mechanisms that determine the activation of taste receptors and the resulting gustatory sensation is crucial to comprehend why specific substances are perceived with a particular taste and how food consumption is regulated. Moreover, due to the interconnected relationship between food intake and health status, the investigation of the molecular effects of food tastants on secondary actors involved in specific diseases or interested functions appears particularly fascinating. In a broader context, this comprehensive knowledge has the potential to pave the way towards, for example, designing the taste of foods, creating ingredients that are less harmful to health, controlling food quality, improving the intake of drug treatments by enlightening their taste, or engineering personalized diets coupled with traditional pharmacological treatments to target molecular types of machinery involved in specific diseases.

In recent years, molecular modelling and machine learning have emerged as fundamental methodologies for elucidating the molecular properties that underlie specific macromolecular functions. In the context of this doctoral thesis, we have applied these methodologies to investigate taste-related molecular actors. Molecular modelling has been employed herein to establish a computational framework aimed at characterizing the interactions between a receptor and its agonist and search for similar binding pockets in protein databases of interest. We applied our methodology to a human bitter taste receptor bound with its agonist to screen the complete repertoire of solved human proteins for potential off-targets that possess similar binding sites. Starting from the methodology just mentioned, we subsequently explored the effects of natural compounds on the molecular structure and dynamics of specific proteins implicated in neurodegenerative diseases. Conversely, machine learning has been employed within a ligand-centred perspective to comprehend the physiochemical properties of food tastants that underlie their taste. To this end, we have developed specialized machine learning-based tools to predict three fundamental taste modalities, namely umami, bitter, and sweet, of a given molecule based on its molecular structure. These findings have

been or will soon be incorporated into the web platform (https://virtuous.isi.gr) which has been developed as a component of the VIRTUOUS project (https://virtuoush2020.com). The project, funded by the European Union (EU), strives to establish a comprehensive platform that amalgamates various levels and methodologies of investigation to predict the organoleptic profile of Mediterranean ingredients based on their chemical composition. This project aims to advance our comprehension of how the chemical structure of food influences our perception of taste, encompassing both the molecular realm and intricate sensory encounters that contribute to the overall taste profile.

In summary, employing a computational approach that integrates molecular modelling and machine learning, the current doctoral thesis has yielded insights into the molecular foundations of taste perception and its potential impact on secondary targets. This work serves as a foundational step towards a comprehensive, multi-level, and interdisciplinary exploration of taste, with the overarching goal of unravelling the intricate processes that link taste perception, food intake, and overall health status.

Contents

List of Figures

Figures in Appendix

List of Tables

Tables in Appendix

Chapter I

Introduction

1.1 Taste Perception, Food Intake & Uptake, and the VIRTUOUS Project

Taste is a multifaceted experience that involves the gustatory, olfactory, and trigeminal systems, and plays a critical role in regulating food intake by assessing its nutritional value and potential harm. The five primary tastes - sweet, umami, bitter, sour, and salty - serve precise purposes in this system. Sweet taste indicates the presence of sugars and carbohydrates, umami is linked to protein content, bitter typically signals potentially harmful substances, sour detects acids and prevents the ingestion of spoiled food, and salty controls mineral intake necessary for proper bodily function. However, taste perception is also an extraordinarily composite and multiscale process that involves molecular, subcellular, cellular, and tissue-level actors of the gustatory system. At the molecular level, taste perception starts through the interaction of chemical substances from the ingested foods dispersed in saliva with specific proteins, called taste receptors, located on gustatory papillae on the tongue. Specific signal transduction pathways exist for each taste type and are mediated by taste receptors, which trigger the activation of taste receptor cells.

Besides, food intake and uptake are crucial concepts in the field of nutrition, and the relationship between these two concepts is complex and multifactorial. Food intake refers to the amount and type of food that an individual consumes, while food uptake refers to the process by which nutrients are absorbed and utilized by the body. The uptake of nutrients mainly occurs in the gastrointestinal tract, which is a highly dynamic and complex system that is influenced by many factors, including genetics, gut microbiota, and lifestyle choices. The process of food uptake involves several steps, including absorption, transportation, and metabolism of

nutrients. These steps are regulated by a range of factors, including hormones, enzymes, and transporters. For example, the hormone insulin plays a critical role in the uptake of glucose by cells, while bile acids facilitate the absorption of fats and fat-soluble vitamins. Understanding the mechanisms of food uptake and the factors that influence it is essential for developing effective strategies to prevent and manage nutrition-related diseases. Advances in technology and research have led to a better understanding of the mechanisms of food uptake, and new approaches, such as personalized nutrition, are being developed to optimize food intake and uptake for optimal health outcomes. Personalized nutrition is an emerging field that aims to develop tailored dietary interventions based on an individual's unique genetic, metabolic, and lifestyle factors. This approach has the potential to improve the uptake of nutrients and reduce the risk of nutrition-related diseases, such as obesity, type 2 diabetes, and cardiovascular diseases. In summary, food intake and uptake are critical concepts in the field of nutrition, and understanding the mechanisms that regulate these processes is essential for promoting optimal health outcomes. Advances in technology and research are providing new insights into the complex interactions between diet, metabolism, and health, and personalized nutrition is emerging as a promising approach for optimizing food intake and uptake.

Investigating the molecular mechanisms underlying taste perception is critical to understanding the complex relationship between food uptake and intake. Taste perception is a multifaceted process that involves various molecular receptors, enzymes, and signalling pathways, which play a crucial role in determining an individual's food preferences and, consequently, their food intake. Understanding the molecular basis of taste perception can offer insights into the factors that influence food uptake and intake, such as the impact of genetic variation on taste perception or environmental factors on food preferences. Moreover, such research can inform the development of new strategies to enhance nutrient uptake by modifying the taste and sensory properties of foods, such as creating low-sugar alternatives to sweet foods. The emerging field of bioengineering has a vital role in nutrition research, providing innovative solutions to enhance the nutritional value and sensory properties of foods. For example, by using novel technologies to create functional foods that can deliver specific nutrients and bioactive compounds more efficiently, we can modify the taste and texture of foods to make them more appealing to consumers and potentially increase their intake of essential nutrients. Therefore, understanding the molecular mechanisms of taste perception is of significant interest as it can enable the design and development of innovative food products that promote optimal nutrition and health outcomes. This view is the foundation of an EU-funded project, named VIRTUOUS (https://virtuoush2020.com/), aimed at developing a virtual tongue as an integrated computational framework to screen among selected natural compounds and food ingredients of the Mediterranean diet (e.g., olive oil or wine) for compounds able

to target taste receptors. More in detail, the proposed intelligent computational platform, by integrating drug discovery techniques, machine learning classifiers, algorithms for big data, cloud computing, and experimental data will predict the organoleptic profile of a selected type of food based on its chemical composition. The outcomes of this project will increase the understanding of the mechanisms driving the information transfer from the chemistry level, where food molecular constituents bind taste receptors, toward the cascade of molecular, supramolecular, and cellular events emerging as an elaborated sensation which strongly contribute to the food organoleptic profile.

1.2 Thesis Aim and Outline

The present PhD Thesis is inserted in the framework of the previously mentioned VIRTUOUS project and specifically aims at investigating the molecular level of taste perception. We considered molecular modelling and machine learning as the main methodologies able to unravel the molecular features and modes of action of the actors involved in the taste perception process. On one hand, we employed molecular modelling to investigate the interactions between tastants and small natural compounds linked to the diet with different taste receptors and proteins. On the other hand, we take advantage of machine learning-based methodologies to develop specific tools to predict the taste of a query molecule from its chemical structure.

In greater detail, the thesis is divided into the following chapters:

- **Chapter I**, i.e. the present chapter, aims at introducing the main topics, the overall organisation and the scientific rationale of the thesis.
- **Chapter II** is dedicated to the molecular modelling investigation of the taste perception actors. After a comprehensive review of the major scientific advances in the field, two novel studies are presented. First, we describe a novel computational pipeline, named VirtuousPocketome, to screen the human proteome for binding sites similar to the one of a query protein-ligand complex, which in our case was one of the bitter taste receptors bound to a bitter compound. Subsequently, we evaluated the role of natural compounds on off-targets not involved in the taste prediction by investigating their impact on the structure of amyloid aggregates.
- **Chapter III** is devoted to the analysis of machine learning in the field of taste prediction. We started reviewing the main scientific works in this field to retrieve state-of-the-art databases of tastants, pinpoint the most used ML-based methods and highlight the major open issues and unmet needs. Two novel algorithms, i.e. VirtuousUmami and VirtuousSweetBitter, to predict the umami and the sweet/bitter tastes respectively are then presented.

- **Chapter IV** collects the conclusions of the present thesis, summarising the major achieved results and the future perspectives.
- **Chapter V** contains the list of the scientific references cited in the present thesis.
- **Chapter VI** is the Appendix of the Thesis, containing additional or supplementary information for the previous chapters.
- **Chapter VII** is the PhD Portfolio providing a summary of the main results achieved and activities carried out throughout the PhD period.

Chapter II

Molecular Modelling for Investigating Taste Perception and Beyond

This chapter deals with molecular modelling applied to the field of taste perception. In the first section, we summarised the recent scientific advances in the field, with specific attention to the modelling and simulation of the main taste receptor candidates and their interactions with relative tastants or other small compounds. Based on these premises, we were also interested in understanding whether other proteins or receptors not involved in taste perception shared similar binding sites with those exhibited by the main taste receptor candidates to investigate possible roles of tastants beyond the mere taste perception and to identify relevant secondary or off-targets modulated by food ingredients or natural compounds. This holds significant importance in the broader comprehension of the route of food in the body, spanning from intake to uptake and its subsequent impacts on health. While the molecular interactions between taste receptors and their corresponding tastants dictate the consumption of specific foods, we were also interested in predicting the trajectory of food ingredients within the body by evaluating their possible interactions with secondary targets. Therefore, in section 2.2, we introduced a novel computational pipeline to screen the human proteome for similar binding sites compared to a query protein-ligand complex. The proposed workflow was applied to a recently solved human bitter taste receptor, namely TAS2R46, bound to a bitter agonist to pinpoint proteins not directly involved in the gustatory system and sharing a similar binding site. In addition to the identification of possible off-targets, we were also interested in evaluating how specific protein

structures and dynamics are influenced by tastants or food-related compounds. Therefore, in section 2.3, molecular modelling and dynamics were employed to investigate the impact of small compounds on proteins not involved in taste perception. As a case study, we considered the S-shaped polymorphism of the Aβ42 amyloid fibril, and we employed molecular docking and dynamics to characterise the destabilising action of 57 natural ligands on its structure.

2.1 Molecular basis of taste perception

The present section is based on the following scientific publication:

***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. (2021). **On the human taste perception: Molecular-level understanding empowered by computational methods.** Trends in Food Science & Technology, 116, 445–459. https://doi.org/10.1016/j.tifs.2021.07.013*

***Author's contribution to the publication:** Pallante L. contributed to every stage of the study, from its conceptualization to the rationalisation of the data, up to the drafting and revision of the manuscript.*

The perception of taste is a prime example of complex signal transduction at the subcellular level, involving an intricate network of molecular machinery, which can be investigated to great extent by the tools provided by Computational Molecular Modelling. The present section summarises the current knowledge on the molecular mechanisms at the root of taste transduction, in particular involving taste receptors, highly specialised proteins driving the activation/deactivation of specific cell signalling pathways and ultimately leading to the perception of the five principal tastes: sweet, umami, bitter, salty and sour. The former three are detected by similar G protein-coupled receptors, while the latter two are transduced by ion channels. The main objective of the present section is to provide a general overview of the molecular structures investigated to date of all taste receptors and the techniques employed for their molecular modelling. In addition, we provide an analysis of the various ligands known to date for the above-listed receptors, including how they are activated in the presence of their target molecule. In the last years, numerous advances have been made in molecular research and computational investigation of ligand-receptor interaction related to taste receptors. This section aims at outlining the progress in scientific knowledge about taste perception and understanding the molecular mechanisms involved in the transfer of taste information.

2.1.1 Introduction

Taste is a complex phenomenon described as a gustatory sensation related to the perception of flavours, which are defined by the combination of sensations coming from the olfactory, gustatory, and trigeminal systems. Taste is one of the most critical control systems able to regulate substance intake, evaluating the healthiness

and nutritional content of food and preventing the ingestion of harmful or toxic elements [1]. The five basic commonly recognised tastes are sweet, umami, bitter, sour and salty, each associated with an essential bodily function. Sweet taste identifies the presence of sugars and carbohydrates, i.e. energetic food. Umami, described as savoury (the taste of cooked meat and broths), is linked to the food's protein content. Bitter taste is generally associated with unpleasant flavour and substances potentially dangerous to the body, such as spoiled food or poisons. However, bitter taste represents a very complex sensation, also associated with substances not harmful to the body, such as coffee, untreated olives, unsweetened cocoa, citrus peel, etc. Sour recognises acids and prevents ingestion of spoiled foods. Salty taste controls sodium and other minerals intake, which play a central role in maintaining the body water balance and blood circulation.

Taste perception is an extraordinarily composite and multiscale process that involves molecular, subcellular, cellular, and tissue-level actors of the gustatory system. Taste arises from chemical substances dissolved in saliva interacting with specific proteins, i.e. taste receptors, which trigger the activation of taste receptor cells (TRCs) located on gustatory papillae, modified epithelial cells distributed throughout the oral mucosa, especially on the tongue. Specific signal transduction pathways, mediated by taste receptors, exist for each taste type: sweet, umami and bitter are determined by organic molecules, and their receptors are G protein-coupled receptors (GPCRs), while sour and salty tastes arise from the presence of ions, detected by ion channels [2]. The activation of the taste receptor cells triggers a specific and taste-related cascade of events reaching the nervous system and ultimately leading to taste perception. In this context, investigating how ligand-protein interactions may drive molecular events (e.g. protein conformational changes) related to activation/deactivation of taste receptors is a crucial step towards a deeper comprehension of the biological nature of taste perception and more in general human nutrition. In this context, molecular modelling, due to a detailed atomistic resolution, represents a powerful tool to shed light on the molecular mode of action of different tastants and the structure-to-function relationships driving the signal transduction at the receptor level. Molecular modelling includes several theoretical and computational methods aimed at representing or mimicking the behaviour of biomolecules, including proteins, DNA, small ligands and polymers [3]. Molecular modelling methods are based on an atomistically-resolved description of the molecular systems, which can best be defined by direct experimental techniques. However, if the structure of interest is not already experimentally solved, it is necessary to employ some predictive method to derive a plausible molecular structure. To this end, homology modelling (HM) presents a widely-employed method to predict the 3D structures of a specific protein, called the *target,* starting from its amino acid sequence. This technique requires a solved 3D structure, the *template,* of a similar macromolecule to model the desired structure. The method accuracy depends on the sequence identity

between the target sequence and the template, as well as the sequence alignment [4]. Furthermore, Molecular Dynamics (MD) is a well-known *in silico* technique to investigate molecular systems' conformational dynamics. The time evolution of the system is obtained by the numerical solution of classical Newtonian dynamics, providing information on the thermodynamic and dynamic properties of the investigated system [5]. Due to this atomistic resolution, MD is a crucial tool for characterising the relationship between the molecular structure of an atomistic system and its function to shed light on important molecular processes and mechanisms, including protein-ligand binding, protein folding, conformational changes driving receptor activation/inhibition, etc. [5–9]. Along with MD simulations, several computational methods, including molecular docking, structure- or ligand-based virtual screening, virtual mutagenesis, machine learning-based methods, etc., have been developed and widely applied specifically to elucidate protein-ligand binding processes and characterise ligand properties and affinity for a specific receptor [10,11].

In the context of investigating taste receptors through molecular modelling, the first issue to be addressed is the receptors' atomistic structure definition, mainly due to the challenging nature of the experimental purification of GPCRs. Indeed, only 89 out of the ~800 GPCRs in the human genome have been solved [12]. This lack is usually compensated through HM, and good models can be obtained for template sequence identities higher than 30% [4]. Nevertheless, literature studies highlighted that transmembrane proteins display strong conservation of structures even at low-sequence identity (below 20%), thus suggesting that it is possible to get accurate 3D models of the TM regions by HM even in these cases [13]. In this context, several recently developed conformational and sampling prediction models have been released and customised for the GPCR structure prediction [14,15].

Apart from the wider-known databases for proteins and ligands, data concerning atomistic models related to taste are collected in many dedicated databases, such as BitterDB [16], containing both bitter receptors and relative ligands, SuperSweet [17] or SweetenersDB [18], collecting sweet compounds.

As previously mentioned, each taste is mediated by a specific receptor, expressed on specific taste cells: sweet and umami are transduced by class-C GPCRs, bitter by class-A/class-F GPCRs, whereas sour and salty are both detected by ion channels [19]. Table 2.1 summarises the primary taste receptors involved in taste transduction and example of tastants. The table also includes information regarding available 3D structures and taste cells expressing a specific receptor. The schematic representation of the main receptor candidates for each taste is shown in Figure 2.1.

Table 2.1. Summary of mammalian taste receptors

CELL TYPE	RECEPTOR(S)	AVAILABLE 3D STRUCTURES	EXAMPLES OF TASTANTS

SWEET	II	TAS1R2 + TAS1R3	No	Natural sugars (glucose, sucrose, sucralose, maltose) Artificial sweeteners (aspartame, neotame, monellin) Sweet proteins (brazzein, monellin, thaumatin, curculin) D-amino acids (D-Phenylalanine, D-alanine, D-serine)
UMAMI	II	TAS1R1 + TAS1R3, brain-mGluR1, brain-mGluR4, taste-mGluR1, taste-mGluR4, GPRC6A, CaSR and GPR92	No	Amino acids (aspartate, L-glutamate, L-AP4, glycine, L-amino acids) Dipeptide and tripeptide (short peptides) Nucleotide enhancer (IMP, GMP, AMP) Organic acids (lactic, succinic, propionic acids)
BITTER	II	25 TAS2Rs	**BitterDB** (HM)	Diphenidol, Lupolon, Quinine, Benzoin, Arborescin, Noscapine, Quassin, Artemorin, Caffeine, Arglabin, Absinthin, Cucurbitacin B, Coumarin, Chlorpheniramine, Papaverine, Adlupolone and polyphenolic compounds (Vescalagin, Castalagin, protocatechuic acid).
SALTY	I*	ENaC, CALHM1/3	**RCSB** ENaC: 6WTH	Sodium chloride (NaCl), lithium chloride (LiCl)
SOUR	III	OTOP1	**RCSB** 6NF4, 6NF4, 6O84	Acids (e.g. citric acid, tartaric acid, acetic acid, hydrochloric acid)

*Taste cells dedicated explicitly to salty taste perception are not clearly determined. In the past, several studies highlighted the absence of the ENaC expression in taste cells II and III [20], thus leading to the hypothesis that salty taste cells belong to type I [21]. However, other studies demonstrated type I cells are not-excitable and their major role is a support function [22]. Therefore, further investigations are needed to clarify the specific type of salty taste cells.

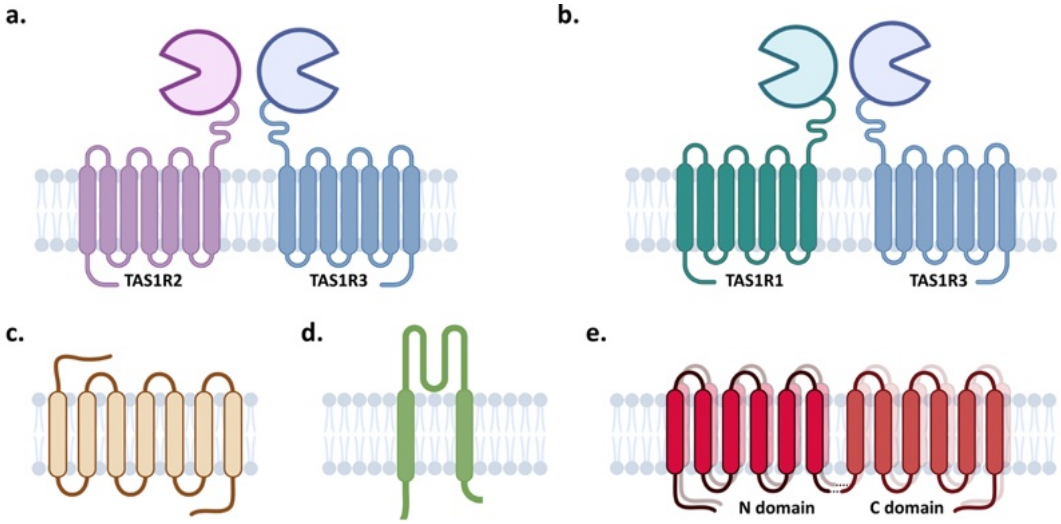


Figure 2.1. Schematic representation of the main receptor candidates for each taste, (a) sweet (TAS1R2-TAS1R3, GPCR of class C), (b) umami (TAS1R1-TAS1R3, GPCR of class C), (c) bitter (TAS2Rs, GPCR of class A/class F), (d) salty (αENaC), (e) sour (OTOP1).

The present review aims at providing a comprehensive picture of recent molecular modelling efforts related to the main taste receptor candidates. Data regarding 3D atomic models and main findings from molecular modelling investigations will be reported and rationalised for each receptor candidate. It is worth mentioning that discussed receptors cover only a limited range of possible receptors, transducers, and proteins essential to taste perception.

2.1.2 Sweet taste receptor

Sweet taste receptor is a heterodimer of TAS1R2 and TAS1R3, encoded by genes tas1r2 and tas1r3. This receptor belongs to the C family of GPCR. Its structure includes seven transmembrane helices (TMD), a large extracellular N-terminus composed of a Venus flytrap module (VFTM) and a cysteine-rich domain (CRD) connected to the transmembrane domain [19,23] (Figure 2.2a).

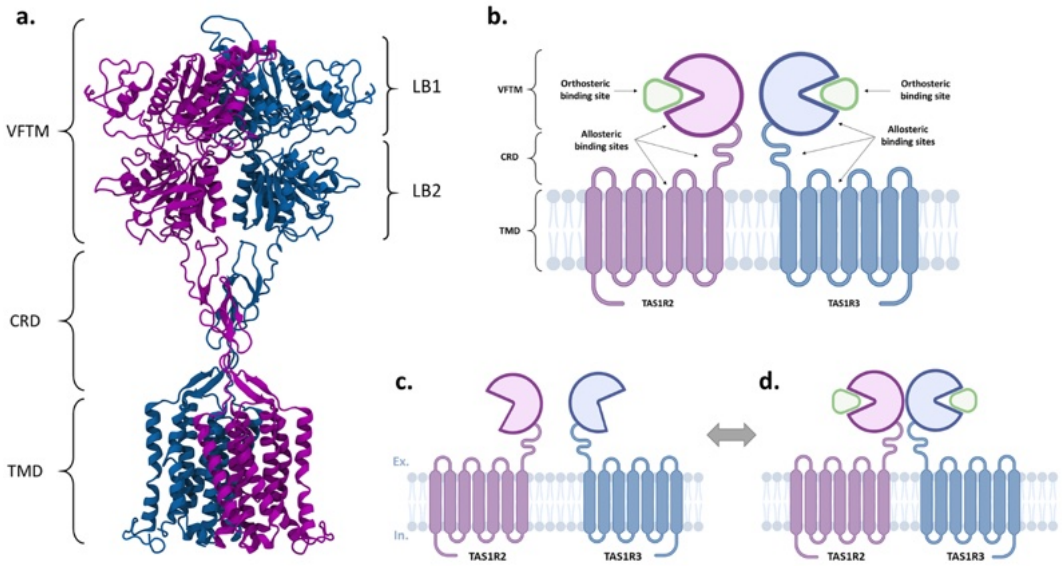


Figure 2.2. (a) 3D molecular representation of the sweet receptor, in purple the TAS1R2 and blue the TAS1R3. The structure consists of the Venus flytrap module (VFTM) with the two lobes (LB1 and LB2), the cysteine-rich domain (CRD) and the transmembrane domain (TMD). (b) Representation of the main binding sites of the sweet taste receptor. The figure at the bottom right is the representation of the activation process of the sweet taste receptor. The receptor evolves from (c) the resting state (open-open conformation) to (d) the active one (close-close conformation) after the binding of the sweet tastants (green) in the VFTM binding pocket.

This receptor responds to many compounds, including natural sugars, such as glucose, sucrose, fructose and sugar alcohols, glycosides, e.g. stevioside and glycyrrhizin, the D-amino acids, e.g. D-tryptophan and D-phenylalanine, peptides, proteins (monellin and brazzein among the most known sweet proteins) and artificial chemical compounds, such as sucralose, aspartame, neotame, saccharin and cyclamate [24,25]. The sweet receptor has an active site in the VFTM, also called the *orthosteric site*, into which small sugars and different sweeteners are suggested to bind [25,26]. Artificial sweeteners, such as stevioside and aspartame, preferentially bind to the VFTM of the TAS1R2 subunit, whereas natural sugars, such as glucose and sucrose, bind to both VFTMs of TAS1R2 and TAS1R3 [19,27,28]. There are also allosteric binding sites within the transmembrane nucleus of the TAS1R3 subunit that can enhance sweet ligands' activity in the orthosteric site [29,30]. The location of the different binding sites is schematically represented in Figure 2.2b. It is worth mentioning that sweet proteins, such as brazzein, monellin and thaumatin, exhibit a different mechanism of action if compared to small, sweet ligands. More in detail, the CRD of TAS1R3 has a crucial role in the interaction with brazzein and thaumatin, and mutations in this region affect also receptor activity toward monellin [31–34].

Receptor 3D structure and conformational dynamics

The first molecular models of sweet taste receptors came out at the beginning of the 21st century. In 2002, Temussi predicted the structure of human TAS1R2-TAS1R3 receptor starting from the free form II of a metabotropic glutamate receptor of subtype 1 (mGluR1, PDB ID: 1EWV) [35] and showed a stabilising effect of the active form of the receptor by docking three different sweet proteins, i.e. brazzein, monellin, thaumatin [36]. From these first results, several groups attempted to improve the HM process's reliability and obtain higher quality structures. In 2010, Zhang and colleagues employed the HM to predict the molecular structure of the TAS1R2-VFTM using several crystal structures of mGluR1, mGluR3, and mGluR7 (PDB IDs: 1EWK, 1EWT, 1EWV, and 3KS9 for mGluR1; 2E4U, 2E4V, 2E4W, 2E4X, and 2E4Y for mGluR3; and 2E4Z for mGluR7) [30]. Masuda and co-workers constructed the VFTM structures for the TAS1R2-TAS1R3 structure, both considering the active (glutamate-bound) and inactive (glutamate-unbound) forms of a mGluR1 (PDB: 1EWT and 1EWK, respectively) [25]. The active form of the heterodimer was constructed selecting the closed and the open forms for TAS1R2 and TAS1R3 respectively, whereas the open form of the crystal structure of mGluR1 was used to construct the inactive form of TAS1R2 and TAS1R3. Moreover, sweet small ligands were docked into the ligand-binding cleft of the TAS1R2 model in the same spot where the glutamate was in the mGluR1. Shrivastav and colleagues compared different HM and threading based methods (SWISS-MODEL, CPHmodels, Modeller, Geno3D, EsyPred 3D, HHpred, LOOPP, Phyre, I-TASSER, and Prime) [37]. The best tools were I-TASSER, CPH Model, SWISS-MODEL and Prime. In 2015, Maillet et al. built human TAS1R2-TAS1R3 VFTMs (open/closed and closed-open forms) by HM with MODELLER [38] starting from the mGluR1-VFTM crystal structure (PDB ID: 1EWK) [35 39]. They generated missing loops with MODELLER and imposed the disulfide bonds selected from the mGluR1 structure (C67-C109, C378-C394, and C432-C439). In 2017, Kim and colleagues predicted the 3D structure of the full-length TAS1R2-TAS1R3 heterodimer, including the Venus flytrap module (VFTM) in the closed–open (co) active conformation, the cysteine-rich domains (CRDs), and the transmembrane domains (TMDs) at the TM56/TM56 interface [27]. To determine the TMD structure of the sweet receptor, they predicted the ensemble of 25 stable structures for the TMD of all TAS1R1s, -2s, and -3s, and constructed the TMD heterodimer for TM45/TM45 and TM56/TM56 interfaces based on GPCR dimers from crystal structures of class A mu-opioid receptor (PDB ID: 4DKL). For the VFTM they used the structure of a mGluR1 bound to glutamate (PDB ID: 1EWK) as a template of the closed-open (co) active state and predicted the binding pose of different agonists (sucrose and stevioside). Finally, to construct the full-length heterodimer receptor, they positioned the VFDs/CRDs on top of the TMD heterodimer and coupled the bonds. In the same year, with a similar approach, Chéron et al. built a full-length sweet receptor using X-ray structures of mGluR VFD (PDB IDs: 1EWT and 1EWK), open and closed receptor states, and the mGluR1 TMD (PDB ID: 4OR2)

as templates [40]. In 2017, the Medaka fish TAS1R2-TAS1R3 sweet taste receptor (PDB ID: 5X2M) was solved by x-ray diffraction, thus providing a new, more realistic template for the VFTM [41]. In 2019, Kashani-Amin et al. introduced a new enhanced model of the full-length sweet receptor based on the most recent templates [42]. More in detail, the Medaka fish structure was chosen for the VFTM model, ensuring better models than the other tested mGluR templates, whereas PDB entry 5K5T (human calcium-sensing receptor) and 4OR2 (mGluR1) were selected for the CRD and TMD, respectively. In the same year, Perez-Aguilar and colleagues constructed and characterised a full-length structural model of the TAS1R2−TAS1R3 receptor, including both the transmembrane (TM) and extracellular (EC) domains of the heterodimer, using comparative modelling and extensive all-atom molecular dynamics simulations [43]. Models of the VFTM for the TAS1R2 receptor were generated by HM using the structures of the metabotropic glutamate receptors 1 (PDB ID: 1EWK) and 3 (PDB ID: 2E4U) as well as the $GABA_B1b$ and $GABA_B2$ receptors (PDB ID: 4MS4), whereas the metabotropic glutamate receptor 3 (PDB ID: 2E4U.pdb) was used for the cysteine-rich domain. Several crystallographic structures from different GPCRs were used (PDB IDs: 4GPO, 4OR2, 3ODU, 4DKL) for the transmembrane domain. It is worth mentioning that the full-length dimer structure of mGluR5 and CaSR for different activations states have recently been solved, paving the way towards more detailed and high-quality models for modelling full-length sweet taste receptors [44,45].

The atomistic resolution of the above-mentioned molecular modelling techniques can be straightforward to shed light on the molecular mechanisms driving the activation of the taste receptors. After ligand binding into the VFTM orthosteric binding sites, the receptor undergoes a series of conformational changes evolving from an inactive/resting state to an active one. In the resting state, the VFTM domains are both in an open configuration (no ligand docked in), resulting in the so-called *open-open conformation.* On the other hand, in the active conformation, at least one sweet compound is docked into one orthosteric binding site, resulting in its closure: if both the VFTM domains are docked to the ligands (e.g. in the case of natural sugars), the active state is characterised by the *closed-closed conformation*; otherwise, if only one VFTM is docked to the sweet tastant (e.g. in the case of artificial sweeteners), the receptor structure is called *closed-open conformation* [25]. The transition from the resting state to the active one in the VFTM leads to the approach of the VFTMs of the two monomers, especially in the ligand-binding (LB) domain 2, and then propagates through the cysteine-rich domain to the transmembrane module. This process ultimately leads to the approaching of the TMs, which trigger the activation of the coupled G protein and the subsequent intracellular pathway [27]. The activation process of the sweet taste receptor, which is fairly similar to all GPCR of class C, is schematically represented in Figure 2.2c,d. Masuda and co-workers using molecular dynamics and molecular docking to characterise the modes of binding between human sweet taste receptor and low-

molecular-weight sweet compounds suggesting a similar activation mechanism to that of mGluR1: the interaction at the core of lobes LB1 and LB2 appears to be essential for reception of all the sweeteners, and the interaction at the entry of LB1 and LB2 would reinforce the formation of the closed structure of the receptor for activation [25]. Kim and colleagues highlighted that the agonist binding into the orthosteric site of the VFTM domain of TAS1R2 leads to major conformational changes, during which the transmembrane domain (TMD) transforms from the TM56 interface to the TM6 interface, as similarly suggested for class C mGluRs [27]. After the ligand binding, the bottom part of the VFTM of the TAS1R3 is pushed toward the bottom part of the VFTM of the TAS1R2, transmitting these changes up to the TAS1R3 TMD (coupled to the G protein). Interestingly, fixing the atoms of either VFTM of TAS1R3 or CRD of TAS1R3 prevents this activation, whereas fixing CRD of TAS1R2 has no effects. Therefore, this study clarified the allosteric influence of the main structural changes of the TAS1R2 VFTM on the TAS1R3 TMD, putatively coupled to the G protein. Similarly, Perez-Aguilar and colleagues remarked that the protomers rotate respectively to each other (clockwise from the extracellular perspective), reducing the distance between the TM6 helices, especially at the extracellular helical segment [43]. However, the authors also pointed out the importance of protein-protein contacts from each protomer's TM5 helices. Interestingly, a similar transition from the inactive state mediated by TM4 and TM5 to the TM6-driven interface in the active state was highlighted in previous literature regarding similar class C GPCRs (mGluRs) [46], and in a recently characterised mGluR5 structure [44]. It is worth mentioning that Perez-Aguilar and co-workers also suggested that, contrary to the mGluRs where full activation is proposed to be reached only when both subunits in the homodimer are bound to an agonist [47], the heterodimeric receptors only require the agonist binding in one of the protomers for their full activation, according to previous literature on other class C GPCRs [43,48]

Ligand-protein interaction investigations

The VFTM contains an orthosteric site for ligand recognition, and sweet tastant can bind both TAS1R2 and TAS1R3 with distinct affinities and structural rearrangements [28]. Liu et al. identified crucial residues (S40, V66, I67, and D142 in the human model) for the species-dependent response of two artificial sweeteners, aspartame and neotame [49]. It is worth mentioning that partially overlapping results were obtained by Zhang and co-workers, which indicated seven key residues for the sucrose and sucralose binding (S40, Y103, D142, D278, E302, P277, and R383) [30]. In line with these results, Masuda et al. conducted mutagenesis studies for screening the residues responsible for sweeteners recognition, highlighting 10 remarkable residues (Y103, D142, S144, S165, P277, D278, E302, D307, E382, and R383) [25]. The proposed model uses five acidic residues (D142, D278, E302, D307, or E382) for agonists recognition: aspartame, D-Trp, and sucralose share LB1 residues (Y103 and D142) and LB2 residues (D278, E302, and D307) for binding, but specific supplementary residues are required for ligand-specific

interaction with the receptor (S144 for aspartame and P277 for sucralose). It is worth mentioning that E302 and S144 have also been previously reported as essential residues for aspartame (and neotame) recognition [50]. In 2015, Maillet and co-workers ultimately identified 11 critical residues in the TAS1R2 VFTM (S40, Y103, D142, S144, S165, S168, Y215, D278, E302, D307, and R383) in and proximal to the binding pocket that is pivotal for ligand recognition and activity of aspartame [39]. More recently, Chéron et al. investigated the orthosteric and allosteric binding sites by computing the volume of TAS1R2 and TAS1R3 binding pockets and providing a list of key residues for sweeteners interactions [40]. More in detail, they remarked that the orthosteric binding pockets in the open form are big enough to allow the binding of small as well as large sweeteners and that both the TAS1R2 and TAS1R3 cavities are hydrophilic. They also identified a secondary cavity close to the main pocket, which is similar to a pocket found on mGluR4 [51]. On the other hand, they highlighted in the TAS1R3 TMD a principal binding pocket and a smaller one in the TAS1R2 model. This finding elucidated why some sweeteners, including small ligands such as lactisole and cyclamate, can fit into the TAS1R3 binding pocket but not into the TAS1R2. In the same year, Kim et al. identified the VFTM orthosteric binding sites of sucrose and stevioside, underlining strong hydrogen bonds to nearby hydrophilic residues D142 and E302, in line with the aforementioned studies. They also remarked a much stronger binding for stevioside than for sucrose, perhaps explaining why stevioside is 210–300 times sweeter than sucrose [27].

Besides orthosteric ligands, positive allosteric modulators (PAMs), targeting different sites, influence taste receptors functions. These molecules are generally tasteless ligands, which bind to the periphery of the orthosteric binding sites with high selectivity, thereby changing the receptor's spatial conformation and enhancing receptor agonism by its activators. Hence, PAMs might be exploited to reduce dietary sugar intake or create high-intensity sweeteners [29,30]. In this context, Yamada et al., using a massive high-throughput screening campaign boosted by molecular docking, pointed out the ability of a novel class of compounds, namely unnatural tripeptide-PAMs, to enhance the sweetness of sucrose [52]. On the other hand, several studies focused their attention on the main receptor domains specifically dedicated to the recognition of possible modulators or allosteric regulators. Particular attention has been paid to the binding sites for cyclamate and lactisole, which are sweet agonist and antagonists, respectively. Jiang et al., using both experimental and computational techniques, including chimaeras, directed mutagenesis and molecular modelling, identified key residues within the transmembrane domain of TAS1R3 that determine responsiveness to lactisole and cyclamate, interestingly finding that the two revealed binding sites are substantially overlapped [53,54]. Moreover, Chéron et al. characterised the structure and dynamics of the allosteric binding pocket of the TAS1R3 sweet taste receptor both in the absence and presence of cyclamate. Molecular dynamics simulations revealed

significant variations in a network of conserved residues not directly implicated in the ligand-binding but unequivocally involved in the receptor function and the allosteric signalling mechanism [55]. These works suggested a critical role of the TAS1R3 transmembrane domain in receptor activation. Interestingly, Winning et al. also remarked the role of the heptahelical domain of human TAS1R3 for the activation of the sweet receptor by neohesperidin dihydrochalcone, which was shown to bind in the same binding sites as the sweetener cyclamate and the inhibitor lactisole. Residues involved in the ligand-binding are also implicated in the binding of allosteric modulators in other class C GPCRs, suggesting common architecture and function of the heptahelical domains of class C GPCRs [56]. Finally, Nakagita et al. characterised the molecular mechanism underlying the sweet taste inhibition of lactisole and a few of its derivatives against the TAS1R3 transmembrane domain [57]. The higher inhibitory potency of investigated inhibitors was mainly due to stabilising interactions in the ligand pocket of the TAS1R3 transmembrane domain and increasing the hydrophobic contacts. On the other hand, Zhao et al. underlined the crucial role of the heptahelical domain of TAS1R2 in mediating the species-dependent sensitivity to sweet regulators, such as the amiloride [58]. Moreover, Zhang et al. investigated the functional domains of sweet taste receptor for the interaction with enhancer molecules [30]. Their molecular modelling and mutagenesis studies revealed the ligand-binding pocket and the binding mode of two sweet taste enhancers, SE-2 and SE-3, into the TAS1R2 VFTM. They identified critical residues near the lips of the lobes involved in lobe-to-lobe interactions or lobe enhancer interactions and underlined a similar action mechanism to that of the umami taste enhancers. Interestingly, they remarked a cooperative binding between orthosteric and allosteric molecules: sweeteners bind near the LB1-LB2 interface, leading to an initial closure of the VFTM domain, whereas enhancer molecules bind near the opening of the pocket and further stabilise the closed conformation by strengthening the hydrophobic interactions between the two lobes. Furthermore, Koizumi and colleagues investigated the unique behaviour of Miraculin, a homodimeric protein isolated from the red berries of *Richadella dulcifica*, which is tasteless at neutral pH but demonstrates an acid-induced sweetness: at neutral pH, Miraculin works as an antagonist, whereas the switching towards acidic pH changes the molecule into an agonist, triggering the sweet sensation [59]. The taste-modifying activity to convert sour stimuli to sweetness was revealed by chimeric receptors and molecular modelling methods, which indicated a major role of the amino-terminal domain of the TAS1R2 for the ligand binding.

2.1.3 Umami taste receptor

The first time that the word *umami* was used was in 1908 by a Japanese chemist, DR. Kikunae Ikeda, who discovered that glutamic acid evokes a unique taste sensation. Therefore, he created the new word umami by combining two words:

umai, delicious or savoury, and *mi*, taste [60]. Only in 2002, the umami taste was recognised as the fifth basic taste.

Initially, only the class C GPCR heterodimer TAS1R1-TAS1R3 was considered as the umami taste receptor, but nowadays eight different types of receptors are accounted as umami taste receptor candidates [61]. Among these receptors, several class C GPCR homodimers have been proposed, such as metabotropic glutamate receptors, including *brain-mGluR1*, *brain-mGluR4*, *taste-mGluR1* and *taste-mGluR4*, the GPCR group 6 subtype A (*GPRC6A*) and the calcium-sensing receptor (*CaSR*). Finally, a non-dimeric structure, namely the GPR92, a class A GPCR, was also indicated. Since most of the above receptors belong to class C GPCRs, we decided to focus our discussion on class C GPCRs in the following.

The first molecule found to have an umami taste was monosodium glutamate (MSG); later, it was found that other amino acids such as aspartic acid and theanine also exhibit the same taste. At the end of the twentieth century, researchers observed that even small peptides could improve food taste. To date, there are 98 peptides identified as bearing umami taste, usually divided based on their number of amino acid residues [61]. A significant discovery was that nucleotides also represent significant mediators of typical umami taste, particularly inosine monophosphate (IMP) and guanosine monophosphate (GMP), which are mainly found in meat and vegetables, respectively. However, the latter two act synergistically with MSG [62].

The putative binding site for these ligands is located in the extracellular part of the umami receptor. In detail, two binding sites have been distinguished: an orthosteric one, located in the TAS1R1 VFTM, and multiple allosteric binding sites that are located in the VFTM and CRD of both chains. For instance, IMP and GMP simply have the role of enhancing taste perception by creating a synergistic action with MSG.

Receptor 3D structure and conformational dynamics

Just like all receptors belonging to class-C GPCRs, they feature the same 3D architecture, comprising the VFTM, the CRD and the TMD, and also the same structure-activity relationship, switching from an active state in which the receptor is in a conformation known as '*closed-open'*, to an inactive state in which the receptor is in a conformation known as '*open-open*'. The 3D molecular representation of the umami taste receptor is shown in Figure 2.3.

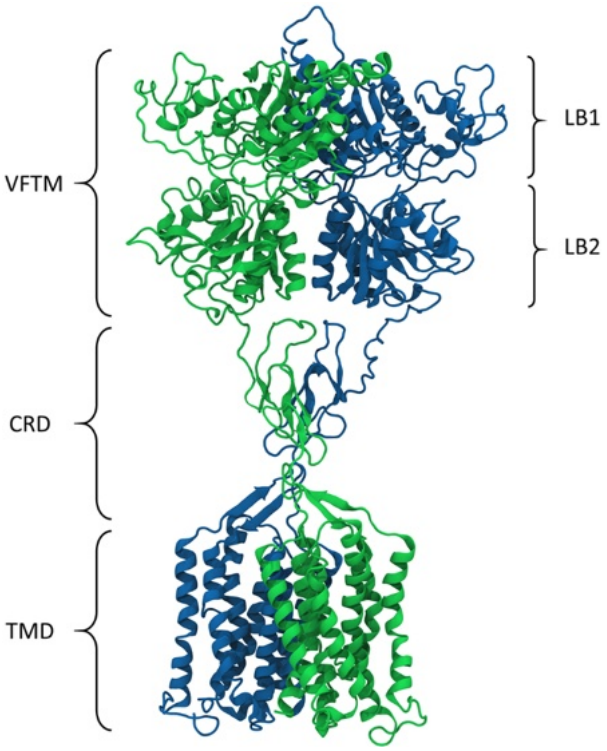


Figure 2.3. 3D molecular representation of one of the main umami receptor candidates, in green the TAS1R1 and blue the TAS1R3. The structure consists of the Venus flytrap module (VFTM) with the two lobes (LB1 and LB2), the cysteine-rich domain (CRD) and the transmembrane domain (TMD).

TAS1R1-TAS1R3 Heterodimer

Concerning the TAS1R1-TAS1R3 heterodimer, the only available structures of the human umami taste receptor stem from HM, as no crystallographic structure of this receptor exists to date. Kunishima et al. were the first to create a model of the receptor's VFTM domain from the free-form II structure of a metabotropic glutamate receptor of subtype 1 (mGluR1, PDB ID: 1EWK). The reference structure is not human and has an identity of around 17% to both TAS1R1 and TAS1R3 [35]. Many authors in the wake of these studies have continued to use mGluR1 as a reference structure; Zhang et al. have also used other metabotropic glutamate receptor subtypes such as subtype 3, mGluR3 (PDB ID: 2E4U), and subtype 7, mGluR7 (PDB ID: 2E4Z), in both open and closed forms [63]. The identity is 20% and 23% respectively. Despite low identity, the authors used the mGluR1 template since their hypothesis assumes that not only does the position of glutamate in the binding site, in the VFTM domain, between LB1 and LB2, remain the same, but also the pocket residues are conserved between the TAS1Rs and mGluRs family of proteins.

As the crystallographic structures of the extracellular part of the fish sweet receptors (PDB ID: 5X2P [41]) were already present, in 2019, Liu and co-workers used this template to create the umami model receptor. Unlike glutamate receptors, this template has a higher percentage of identity, around 33% [64].

However, all the mentioned models only include homology models of the extracellular part, i.e. the VFTM. Thus, no complete model of the best-known umami receptor, the TAS1R1-TAS1R3 heterodimer, exists to this date.

mGluR1 (brain and taste isoform) and mGluR4 (brain and taste isoform)

These two metabotropic glutamate receptors, the mGluR1 and the mGluR4, belong to two different groups of mGluRs, based on their activity and structure: group I and group III, respectively. The two types of isoforms are a little different from each other; the taste isoform does not have the same typical opening that all other models have, indeed the VFTM part is truncated and therefore has a slightly lower affinity to L-glutamate than other receptors. This receptor's crystallographic structures are plenty and have been used to create the TAS1R1-TAS1R3 heterodimer structure.

CaSR and GPRC6A

Like TAS1R1-TAS1R3, these two receptors belong to class-C GPCRs, and are identical in structure to the umami receptor; the only difference is that they are homodimers, so the two chains are identical. Bystrova and co-workers have shown that these two receptors also respond to different ligands, including L-amino acids and peptides. Geng et al., in 2016 released the crystallography structure of human calcium receptor comprising only the extracellular part in both the active (PDB ID: 5K5S) and inactive (PDB ID: 5K5T) forms [65]. More recently, the precise crystal structure of CaSR was determined for each activation states, i.e. closed-closed, open-closed, and open-open [45]. CaSR was found in different tissue, including the parathyroid gland and kidney [66].

Ligand-protein interaction investigations

Generally, in humans, the umami receptor is activated by monosodium L-glutamate (MSG). However, other amino acids can also be stimulated, such as aspartate, or by some organic acids, including lactic, succinic, and propionic acids. On the other hand, esters such as guanosine 5'-monophosphate (GMP) and inosine 5'-monophosphate (IMP) can increase the taste [67].

As in the case of the sweet receptor, the umami receptor features an orthosteric binding site located in the VFTM of both chains, TAS1R1 and TAS1R3, as well as an allosteric binding site in the TMD and CRD, following the same scheme of the sweet receptor in Figure 2.2b. When umami-enhanced peptides bind in the allosteric sites, they cause a conformational rearrangement in the receptor, which amplifies the orthosteric transduction pathway by increasing the active sites' affinity for the

umami tastants. For example, Töle and colleagues reported how allosterically bound cyclamate enhances the receptor activation by L-glutamate bound in the VFTM orthosteric site [19]. Also, IMP and GMP are capable of binding in the allosteric site and improving taste signal transduction by stabilising the closed conformation of TAS1R1 [62]. Moreover, Toda et al. showed that methional, a typical taste of cheeses, could potentially bind at two distinct sites in the transmembrane domain of TAS1R1 and served as a positive allosteric modulator (PAM) of the human umami receptor, but as a negative allosteric modulator (NAM) in mice [68].

As for the chain of conformational events beginning with ligands binding in the VFTM and ultimately leading to downstream signal transduction, different models have been proposed: Zhang et al. reported that the closure of the VFTM of TAS1R1 and TAS1R3 occurs as a two-stage process, starting with the initial positioning of glutamate in the VFTM LB1, occurring in μs timescales, followed by further positional optimisation inside the cleft, requiring ms timescales [63]. Cascales and his colleagues have shown with MD simulations that the closure mechanism, thus the activation of the umami receptor, is achieved by Form 1 in which the TAS1R1 chain has a closed conformation while TAS1R3 has an open conformation, as previously described [69].

2.1.4 Bitter taste receptor

Bitter taste receptors are members of another family of GPCRs called the taste 2 receptor family (*TAS2Rs*) [70]. Many discussions have been carried out regarding their belonging to a specific class of GPCRs: some authors place them within the class F of GPCRs, consisting of frizzled and smoothened proteins; others place them in the broader class A of GPCRs, rhodopsin-like and, recently, the online database GPCRdb (https://gpcrdb.org/) even created a new sub-family called class T for these receptors. Due to their functional principles and the position of the binding site, they resemble those of class A GPCRs to which visual and odorant sensory receptors also belong, but this is not the case for their sequence similarity [19,71]. Its structure includes short extracellular N-terminus and intracellular C-terminus, seven transmembrane helix (TMD) which are connected by three Extracellular Loops (ECLs) and three Intracellular Loops (ICLs) [72]. The most conserved component between class A GPCR and bitter receptors is the 7 TMD bundle which forms the structural core, binds ligands in the extracellular (EC) region and permit the transduction of information due to the intracellular (IC) region [73]. The comparison shows that important class A motifs and highly conserved disulfide bridge that facilitates GPCRs structure stabilisation are missing. On the other hand, the TAS2Rs specific conserved residue may have an essential role in stabilising the inactive conformation of bitter receptors.

The number of TAS2R genes varies largely across species [16,74]. Among the different species, not only does the number of genes coding for the bitter receptor change,

but there are also differences on where the genes that encode TAS2Rs are; in humans, they are coded by chromosomes 5, 7 and 12 while in mice by 2, 6 and 15. The number of bitter compounds that humans can perceive is much larger than the number of human genes; this makes us understand that every bitter receptor responds to more than one ligand [23,74]. TAS2Rs constitute an interesting subgroup of GPCR because they have many known agonists and few antagonists. Besides, this ligands' activity is usually in the micromolar range, higher than the typical nanomolar ranges of most GPCR ligands [73].

Due to the large number of TAS2Rs, the large quantity of naturally occurring bitter-tasting substances and the presence of three *generalist* receptors - TAS2R10, TAS2R14 and TAS2R46 - recognising about one-third of all bitter compounds, heterodimerization of bitter taste receptors may not be necessary to extend their already great receptive capacity [75]. However, *in vitro* experiments revealed that TAS2Rs bitter taste receptor form oligomers (approximately 325 homodimeric and heterodimeric receptors), but it is not yet known if TAS2Rs heteromeric receptors contribute to a broader detectable agonist spectrum [76].

Moreover, some authors noticed that some bitter compounds could both activate the TAS2R receptor and be able to interact with the cell membrane's ion channels, so they may also function as bitter receptors [23]. Additionally, studies have shown that TAS2Rs are not only in the taste buds but also expressed in extra-oral tissue, including heart, skeletal and smooth muscle [77]. The distribution of TAS2Rs is variable in different kinds of muscle cells, but TAS2R3, TAS2R4, TAS2R5, TAS2R10, TAS2R13, TAS2R19 and TAS2R50 are always present in a moderate way, while TAS2R14 is highly expressed in all the human body. Moreover, previous literature pointed out the expression of TAS2Rs on human airway smooth muscle [78] and smooth muscle tissue along the mouse gut and in human gastric smooth muscle cells, suggesting a possible role of TAS2Rs as targets to alter gastrointestinal motility and hence hunger sensation [79]. Moreover, TAS2Rs are also related to muscles contraction or relaxation in other organs such as the bladder [80]. Bitter molecules are usually considered poisonous substances, yet there are non-toxic ones with beneficial effects on the human body. For this reason, a better understanding of the bitter taste receptor transduction may lead to the design of specific drugs with an acceptable taste and having an essential role in muscle-related diseases.

Receptor 3D structure and conformational dynamics

At present, one of the major obstacles for the molecular modelling of bitter taste receptors is the lack of experimentally solved structures representing the 25 bitter receptors. As a matter of fact, only the molecular models by homology modelling have been developed for 23 out of 25 human bitter receptors. Those models are publicly available in the BitterDB [16] which also provides information concerning bitter receptors and related ligands. Only two receptors, the TAS2R45 and

TAS2R19, are not included in the database. The 3D molecular representation of TAS2R3 is shown in Figure 2.4a and a detailed list of all the human bitter taste receptors along with alternative names is reported in Table A - 6.2.1.

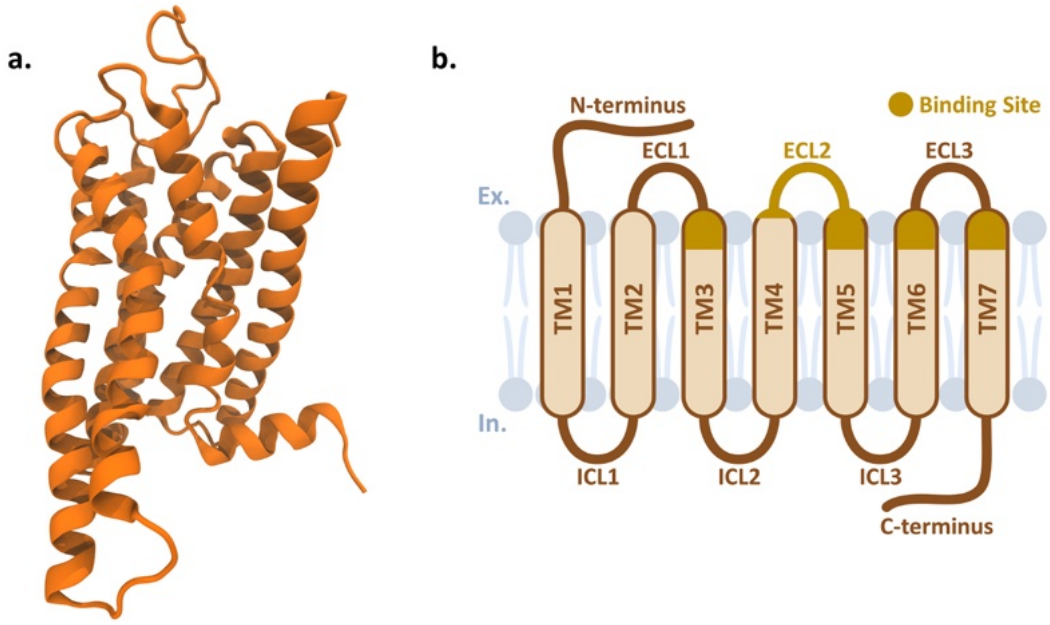


Figure 2.4. (a) 3D homology model of the TAS2R3 bitter receptor (PDB from BitterDB [16]. (b) Schematic representation of the bitter taste receptor, including the extra- and intra-cellular loops (ECLs and ICLs), the transmembrane (TM) helices, and the main structures involved in the ligand binding.

The 3D structure of the TAS2R14 bitter receptor stored in the BittterDB was modelled using the β2 adrenergic receptor (PDB ID: 3SN6), another class-A GPCR, as a template. This model was subsequently used as a template for the other receptors, which were built using MEDELLER [81] and then manually adjusted. Other groups followed similar homology modelling strategies using other experimental templates: Pydi et al. built the TAS2R4 receptor using Rhodopsin (PDB ID: 1U19) and opsin (PDB ID: 3DQB) as templates, in their active and inactive conformation respectively; Wang and co-workers modelled the TAS2R7 receptor using the serotonin receptor template (PDB ID: 6BQC) [82,83]. Similarly to previous literature [82], bovine rhodopsin and opsin were employed as templates to model TAS2R4 and TAS2R1, whereas Squid rhodopsin (PDB ID: 2Z73) was chosen as the template for TAS2R14 [84]. Moreover, in line with the homology modelling strategy used for BitterDB, other groups used the β2 adrenergic receptor (PDB ID: 3SN6) as the template to model the TAS2R16 [85] and, coherently with the GPCRdb homology modelling pipeline, TAS2R5, TAS2R7, TAS2R14, and TAS2R39 were modelled starting from the β2 adrenergic receptor (PDB ID: 3SN6), the serotonin 2B receptor (PDB ID: 5TUD), and the mu-opioid receptor (PDB ID: 5C1M) [86].

The aforementioned models were used as starting point for structure-to-function molecular studies aimed at exploring the conformational behaviour of bitter taste receptors and highlighting crucial residues/structures involved in receptor activation. For example, a *multipoint stimulation model*, similar to the one previously proposed for the sweet receptor [87], has been suggested for the activation of bitter receptors by steviol glycosides (SG) and other water-soluble molecules: at the beginning, ligands stimulate extracellular residues and, subsequently, the allosteric modulation of the transmembrane site is triggered [84]. Furthermore, crucial residues, i.e. H94 in helix 3 and E264 in helix 7, for the activation of the receptor driven by metallic ions, have been remarked [83].

Ligand-protein interaction investigations

Although all bitter taste receptors are characterized by one single binding site for bitter ligands, the high number of TAS2R receptors allow for the recognition of a huge number of bitter compounds [74]. More in detail, bitter taste receptors are activated by a wide variety of chemically different agonists. This affinity toward a huge range of chemical structures may be achieved with various interaction types between different ligands in the binding pocket [71]. Bitter taste receptors can be divided into *promiscuous*, activated by a multitude of chemically different compounds, and *selective*, activated by few chemicals [74]. Examples of promiscuous receptors are the TAS2R10, TAS2R14 and TAS2R46. Each TAS2R receptor has specific patterns for the recognition of related bitter substances, but numerous compounds can activate several TAS2Rs [74]. The selectivity and promiscuity profile of bitter taste receptors and their ligands has been recently explored by chemoinformatics approaches [88]. More in detail, results highlighted that almost all selective bitter receptors are activated only by promiscuous compounds, i.e., those ligands targeting more than one TAS2R. Instead, promiscuous receptors are activated by both promiscuous and selective binders [88]. The relevance of the ligand promiscuity investigation lies primarily in the possibility of a rational ligand design specifically aimed at modifying their chemical structure according to specific needs. On the other hand, characterisation of the molecular features defining receptor promiscuity may be pivotal for understanding the ability of bitter receptors to identify the huge variety of bitter tastants. The receptor promiscuity can be accessed with the so-called promiscuity index (PI), i.e. the number of bitter compounds that activate the receptor divided by the total number of molecules considered. On the other hand, the diversity of the ligand set can be measured with another promiscuity index, namely the PI_{NUS}, calculated as the number of unique scaffolds (NUS) for each receptor divided by the total number of NUS [88,89]. Previous literature on class A GPCR identified a correlation between the binding site characteristics and the variety of antagonists. In particular, the number of unique scaffolds, that measures the number and variability of antagonists, was demonstrated to be correlated to the exposure and hydrophobicity of the binding site and opposed to the number of hydrogen bond donors [89]. Interestingly, despite

the lack of structural data that limits a full investigation of TAS2Rs, Di Pizio et al. suggested that the aforementioned properties of the binding site correlate also with the TAS2R-promiscuity [88].

The ability of the bitter receptors to detect a huge variety of ligands is made possible by single-point mutations in the binding pocket that can improve or reduce affinity towards a specific ligand [90]. Despite the raised hypothesis that bitter receptors could have more than one binding site to accept the huge variety of bitter agonists, Slack and colleagues demonstrated the existence of a unique binding pocket [91]. Several studies were also performed to identify the binding pocket of bitter receptors through the use of point mutations on TAS2R16. These studies highlighted the binding site involves seven residues belonging to TM III, V and VI and in particular at least three of them interact directly with salicin [92]. This prediction was also confirmed by experimental studies and functional analyses on mutant receptors that led to the identification of residues responsible for the agonist selectivity and activation of TAS2R46, TAS2R43, and TAS2R31 [71]. Most structure-function studies involving bitter taste receptors have confirmed the binding pocket of TAS2Rs is located in the extracellular side of the TM bundle, between TMs III, V, VI and VII (as shown in Figure 2.4b), which is the canonical site of class A GPCRs. Indeed, several investigations on TAS2R14, TAS2R10 and TAS2R46, the most examined receptors, experimentally confirmed the involvement of residues present in the above mentioned TMs [71,90,93], but also suggested an involvement of TM II for TAS2R14 and TAS2R46 receptors, which might be explained by the more spacious pocket shape, as already reported for TAS2R14 [94]. It is worth mentioning that the residue composition of the above-mentioned binding site is highly different in every TAS2Rs, suggesting the possibility of the detection of different ligands with a variety of agonist-specific interactions patterns [90]. Several investigations highlighted that residues belonging to the ECL2, the longest loop in the extracellular side of the receptor, significantly contribute to ligand binding and activation of TAS2Rs: Liu and colleagues demonstrated that residues N167, T169 and W170 could influence ligand binding in TAS2R7 [95], and previously Karaman et al. showed that residues N163 and N172, located in ECL2, present the same function in TA2R14 [94]. Moreover, computational studies highlighted the type of interactions between the receptors and some ligands and major conformational changes related to ligand-driven activation. For example, Chen and colleagues investigated the possible activation mechanism of TAS2R16 in the presence of its agonist and antagonist, i.e. salicin and probenecid respectively, docked into its active pocket [85]. Acevedo and co-workers investigated steviol glycosides (SG), non-caloric sweeteners derived from plants, which demonstrated in in *in vitro* studies a specific affinity towards TAS2R4 and TAS2R14. This ability makes these compounds able to generate, in addition to their sweetening effect, also an unpleasant bitter taste [84]. They showed that SGs have only one site for orthosteric binding and SGs only bind to TAS2R4 and TAS2R14 and not to TAS2R1.

Moreover, they remarked a negative correlation between protein-ligands binding energies and bitterness intensity, but again not for TAS2R1. Therefore, this research pointed out that the binding site of TAS2R1, mainly inserted in the transmembrane region, is not tailored for this type of sweeteners and other water-soluble molecules, e.g. caffeine or quinine. They also observed a crucial role of the ligand size compared to the dimension of the binding site, underlining that SGs with more sugars have less affinity for bitter taste receptors. Moreover, steered molecular dynamics simulations highlighted a major difference in affinity between stevioside and rebaudioside A: the former is characterised by stronger interaction with the receptor if compared to the latter due to the formation of more hydrogen bonds at the binding site of both receptors [84]. Other bitter ligands particularly important for their nutritional properties are polyphenols, which are present for example in coffee, wine, or red fruits. Soares and his colleagues investigated the bitterness of different classes of 16 polyphenolic compounds through the activation of TAS2Rs and pointed out their stimulation on bitter taste receptors. They also noticed that the condensed tannins, a subclass of the flavonoids/flavanols, specifically activates the TAS2R5, whereas the hydrolyzable tannins, in particular the ellagitannins, triggers the TAS2R7 [86].

In literature, it is reported that bitter receptors may have only one binding site for agonists and antagonists, due to the type of interactions with a selected residue depending on the ligand nature [95]. However, some studies suggest that there may be an additional *vestibular binding site* located in the extracellular part of the receptor. Sandal and co-authors proposed that agonists can transiently occupy this site and be prefiltered before the introduction into the canonical binding site and that these two sites may have a role in discrimination of different agonists of TAS2R46 [96].

The interaction between TAS2Rs and bitter tastants also depends on several factors, e.g. type of ligands, membrane lipids and movements of TMs and ECLs. Indeed, Pydi et al. suggested cholesterol sensitivity of T2Rs and remarked a crucial role of cholesterol in the cell membrane for the interaction between amino acid [97].

2.1.5 Sour taste receptor

Sour sensing is particularly important in the taste system for monitoring the functional state of body fluids. Even if a lot of progress has been made in the studying and discovery of the molecular mechanisms behind sweet, bitter and umami tastes, sour taste is still poorly understood [98]. Sour taste is detected by type III cells and it is essential in regulating the intake of H^+ ions [1].

During the past decades, several membrane ion channels have been proposed as sour taste transducer, including epithelial sodium channel (ENaC), Acid-Sensing Ion Channel (ASIC), two-pore domain potassium (K2P) channels, H^+ gated calcium channels. In the recent past, the polycystic kidney disease 2-like1 ion channel

(PKD2L1) was identified as a putative sour taste receptor [98–100]. However, a direct role for PKD2L1 or its partner, the PKD1L3, in sour transduction was not supported by subsequent studies on knocked out mice [101]. Nevertheless, PKD2L1 is still considered a useful marker for sour taste cells (type III cells) [102].

More recently, a tremendous breakthrough was achieved from Tu and co-workers, who have discovered that transduction of sour taste in mice involves permeation of H^+ through a proton selective ion channel, a protein named Otopetrin1 (OTOP1) [103]. OTOP1 is specifically expressed in type III taste cells, it generates a proton current across the membrane in response to extracellular acidification, and it is sensitive to Zn^{2+}, which is a crucial factor for the proton current related to sour perception [103]. Using PKD2L1 as a molecular identifier for sour-responsive taste cells, different research groups [104,105] confirmed OTOP1 as the necessary transduction channel underlying sour taste. OTOP1 belongs to the Otopetrins family, which also comprises two other ortholog proteins, i.e. OTOP2 and OTOP3 [103]. Human OTOP1 (hOTOP1) forms a channel with similar properties to murine OTOP1, and murine OTOP2 and OTOP3 share 30 to 34% amino-acid identity with murine OTOP1[103].

Receptor 3D structure and conformational dynamics

Only a few 3D atomistic structures are currently available in the RCSB database for the Otopetrin family. As far as we know, the main structures are the zebrafish OTOP1 and the chicken OTOP3 (PDB entries: 6NF4 and 6NF6) [106], and the Xenopus Tropicalis OTOP3 (PDB entry: 6O84) [107].

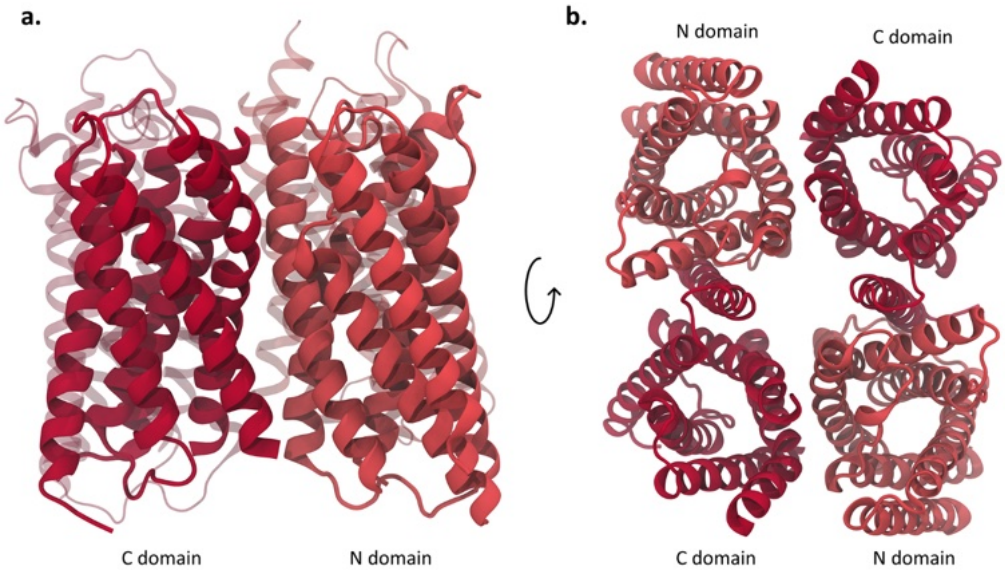


Figure 2.5. Frontal (a) and top (b) views of the 3D molecular structure of OTOP1 (PDB entry: 6NF4). Each subunit is formed by two structurally homologous domains, i.e. the N domain (light shade) and C domain (dark shade).

Recently, Chen et al. characterised the first molecular structure of the OTOP family due to the cryo-EM experimental determination of the *Xenopus Tropicalis* OTOP3 (XtOTOP3) (PDB entry: 6O84) [107]. They highlighted that XtOTOP3 adopts a unique two-pore architecture forming a homodimer: each subunit is composed of 12 transmembrane helices divided into two structurally homologues halves representing the C and the N domains, which surround a highly hydrophobic tunnel filled with lipids. It is worth mentioning that, from Wheatley, half of the plasma membrane, both subunits contain solvent-accessible cavities that are enclosed by TMs 2–6 (N-pore) and TMs 8–12 (C-pore), respectively.

Subsequent studies analysed and compared zebrafish OTOP1 (zfOTOP1) and chicken OTOP3 (chOTOP1), which are 30% identical to each other by sequence and share 44% and 59% identity with human OTOP1 and OTOP3, respectively [106]. Their results were achieved due to the direct analysis on the full-length OTOP1 (PDB entry: 6NF4), represented in Figure 2.5a and Figure 2.5b, and OTOP3 (PDB entry: 6NF6) [106]. Observed structures are very close to the ones highlighted by Chen et al, thus suggesting a common topological organisation to all Otopetrin family members.

The receptor function allowing the proton transfer across the membrane is still under debate. It is supposed that protons can flow through a 'hopping' mechanism along a hydrogen-bonded network made by water molecules and/or amino acid side-chain moieties. Two structurally analogous vestibule-shaped openings in each OTOP1 and OTOP3 subunits could represent loci for proton permeation, one housed by the N domain and the other by the C domain [106]. Interestingly, the same pattern is shared by XtOTOP3 [107]. Both domains contain numerous polar and charged residues: the region of hydrophobic residues could potentially be a hydrophobic plug that regulates water/ions accessibility. Another feature of the putative permeation pathways within the N and C domains is the *constriction triads* composed of glutamine–asparagine–tyrosine, which we abbreviate as the QNY triad. Respectively for the N- and the C-pore, they are formed by residues:

- Q174/N204/Y268 and Q433/N528/Y571 in zfOTOP1
- Q175/N205/Y266 and Q429/N503/Y546 in chOTOP3
- Q232/D262/Y322 and Q558/N623/Y666 in XtOTOP3

Their side chains are sufficiently close to interact directly or through intervening waters. The function of these triads is uncertain, but a role in proton transfer seems possible considering it is conserved in both N and C domains for both zfOTOP1 and chOTOP3 [106]. In XtOTOP3, instead, even if both pores could potentially function as proton permeation pathways, the C-pore constriction triad could probably be more crucial in determining channel activity and proton permeation in XtOTOP3 [107]. Further investigations are needed to clarify the specific role of each

triad in both pores and to possibly depict a general working mechanism for all the otopetrin family members.

Another aspect to be considered deals with water permeation from the extracellular *milieu* through the N and C domain vestibules. During MD simulations, the presence of water is continuously observed at the intrasubunit interface of the N and C domains in zfOTOP1 and chOTOP3, but not at the intersubunit interface, where water permeation through the central tunnel was completely blocked by the cholesterol molecules [106]. Similar behaviour was also reported for the XtOTOP3 [107]. The stochastic formation of a water wire during molecular dynamics simulation suggests also that proton conduction could occur through a water-hopping mechanism [106].

In conclusion, molecular dynamics simulations shed light on the molecular mechanisms for proton conduction, pointing to three main possible mechanisms: aqueous vestibules in the N and C domains, and the intra-subunit interface [106]. However, it is still unclear which of these three pathways or their combination allow the flux of proton currents.

2.1.6 Salty taste receptor

Salty taste controls sodium and other mineral intakes, which play a central role in maintaining the body water balance and blood circulation. In this context, the sodium ion (Na^+) is an essential mineral regulating the osmolality of the extracellular fluid and takes part in many physiological processes. Since Na^+ is constantly excreted from the body, it is paramount to properly integrate the ion's loss to effectively maintain the bodily balance through the diet. Na^+ specifically elicits the salty taste sensation, which guides the intake of this important mineral [108]. Salty perception may trigger both attraction and repulsion towards the source. At high concentrations, saltiness usually results in a negative reaction, whereas at low to moderate concentrations, saltiness is attractive [109]. Chemically, the salt that is usually regarded as the main trigger of a salty perception is sodium chloride (NaCl) and other salts also feature more compound gustative footprints, for example by triggering also bitter or sour sensations.

In mice, the attractiveness of salty sensation is selectively triggered by sodium and inhibited by amiloride. Since amiloride is a potent inhibitor of the ENaC, it has been proposed as a crucial component of the salty receptor machinery [109,110]. The expression of ENaCs in humans is mostly on the apical surface of epithelial tissues throughout the body. ENaC belongs to the ENaC/Degenerin (DEG) family, which include also well-known ASIC. These receptors are characterised by subunits that consist of short intracellular N- and C-termini, two membrane-spanning helices, and a large cysteine-rich extracellular domain (ECD) that can form homo- or heterotrimeric ion channels [111,112]. The ENaC receptor has three homologous

subunits α, β and γ or δ [113]. This ion channel allows the passage of Na ions, maintaining the right concentration of salt and water in the body.

In mice, the salty attraction is mediated by the α subunit of the epithelial sodium channel (α-ENaC) [20] and exhibiting sensitivity to amiloride [114]. Therefore, in rodents, attraction to low sodium is blocked by amiloride, and knockout mice lose this attraction [20]. On the other hand, appetitive salty taste is not sensitive to amiloride in humans [114], and an additional ENaC gene, the δ gene, is found in their genomes, leading to the expression of both the amiloride-sensitive α- and the less sensitive δ-ENaC subunits in human taste cells [115]. Moreover, in rodents' model, ENaC should be found at the apical membrane of taste cells [116], whereas some pieces of evidence suggest that only the δ-subunit localises to the taste pore region in human taste buds and other ENaC subunits seem to be segregated in the basolateral compartment, thus suggesting the δ-subunit as a possible salty taste receptor. In light of these considerations, it is still under debate if all the subunits are required to form a functional sodium receptor [117]. In conclusion, the ENaC is probably involved in human sodium detection, but no certain evidence has defined in which stage of the perception process. The lack of the amiloride effect [114] and the presence of α-, β-, and γ-subunit only in the basolateral portion of taste buds [115] seem to favour a role for ENaC downstream of the initial receptive events [108].

Very recently, Nomura et al. showed that sodium taste signalling in mice is independent of Ca^{2+} concentration (in contrast to the taste perception mediated by type II and type III cells) and only voltage-dependent [118]. This study demonstrates that the Na^{+} entry through ENaC leads to depolarization, driving the subsequent generation of the action potential by voltage-gated ion channels. Interestingly, the authors showed that the co-expression of the voltage-gated neurotransmitter-release channel (CALHM1/3) and ENaC, both required for amiloride-sensitive salty taste transduction, is essential to identify salty taste cells. These findings represent a big step forward in the salty taste pathway perception, notwithstanding ENaC has still to be proven as the principal sensor for salty taste in humans [114,115,119] and the apparent insensitivity to amiloride of salty taste in humans has not been explained yet [120].

Receptor 3D structure and conformational dynamics

The first crystal structure of ENaC was solved by cryo-electron microscopy (cryo-EM) at a nominal resolution of 3.9 Å (PDB entry: 6BQN). The ion channel is composed of a large extracellular domain and a narrow transmembrane domain, characterised by a 1:1:1 stoichiometry of α:β:γ subunits arranged in a counter-clockwise manner (Figure 2.6) [112].

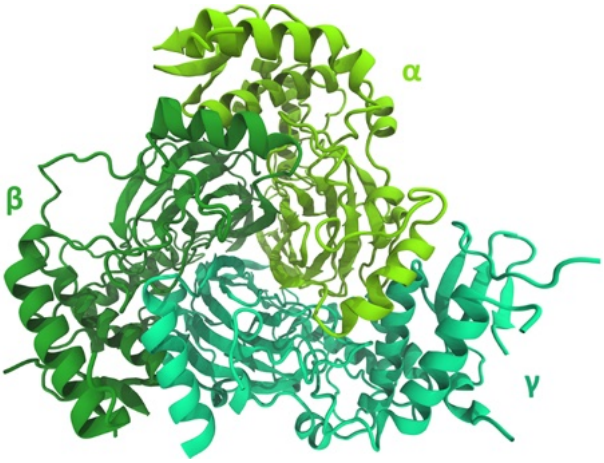


Figure 2.6. Representation of the 3D molecular structure of the trimeric ENaC (PDB entry: 6WTH), comprising the α:β:γ subunits arranged in a counter-clockwise manner.

The same group recently solved the molecular structure of ENaC by cryo-EM at 3 Å (PDB entry: 6WTH), showing that the α subunit has a primary functional module consisting of the finger and the Gating Release of Inhibition by Proteolysis (GRIP) domains, which strongly separate the behaviour of this receptor from close relative ASICs. The module is bifurcated by the α2 helix dividing two distinct regulatory sites: Na^+ and the inhibitory peptide. Removal of the inhibitory peptide perturbs the Na^+ site via the α2 helix highlighting the critical role of the α2 helix in regulating ENaC function [121]. However, the experimental resolution of the transmembrane domain (TMD) and the cytosolic domain (CD) is still missing. Future improvements on the above-mentioned structures might pave the way towards the full-length channel and gain fruitful insight to understand the mechanistic link between the removal of inhibitory peptides in the cysteine-rich extracellular domain (ECD) and channel gating.

2.1.7 Conclusions

In this work, we provided a comprehensive review of the main findings in the molecular modelling of taste receptors. The work is focused on the main candidates commonly discussed in the literature, i.e. GPCRs for sweet (TAS1R2-TAS1R3), umami (TAS1R1-TAS1R3) and bitter (TAS2Rs), OTOP1 for sour and ENaC for salty. It is worth mentioning that discussed receptors cover only a limited range of possible receptors, transducers and proteins essential to the taste perception process. Just to name a few, sugars are also transduced by sugar transporters [122], the salty taste has amiloride-sensitive and amiloride-insensitive components [118], and sour taste most certainly involves mechanisms other than OTOP1, such as intracellular acidification and blockage of KIR2.1 [102]. The presence of other key players, as well as the identification of other possible basic tastes, makes the understanding of the taste perception still incomplete and lacking, and a lot of work is still needed to get

to a more granular and comprehensive knowledge. Interestingly, the existence of a sixth taste quality linked to fat perception has been recently highlighted [123–125]. In addition, some studies have remarked that the ability to detect fatty acids is reduced in response to a high-fat diet [126]. In this context, the fat taste seems pivotal for the connection between fat intake and health status, specifically linked to overweight or obesity. Therefore, further studies related to fat taste may provide new bases for controlling the development of obesity, one of the main causes of global disease burden, including cardiovascular diseases, cancer and diabetes [127].

At present, the main findings on the receptor function come from computational and/or combined computational/experimental studies focusing on the structure-to-function relationships and ligand-protein binding investigations. We deeply discussed the need for developing high-quality molecular structures as a crucial step in molecular modelling and described the most recent experimentally-solved and *in silico*-derived structures for each taste receptor candidate. Out of the mentioned players of taste transduction, the only available experimental structure is the VFTM domain of the sweet receptor of the medaka fish, which represents a fundamental starting point for most computational investigations of sweet taste transduction mechanisms [41]. Conversely, umami and bitter receptors have been studied through HM relying on experimental templates sharing some degree of sequence similarity. Homology models for bitter receptors are publicly available from BitterDB. Of note, the comparably low reported sequence identities for these models to their respective templates, compared to other HM applications, are not detrimental to the quality of the reported studies, due to the nature of the involved receptors. Lastly, some experimentally obtained molecular structures of both human and non-human salty and sour receptors are currently available in the RCSB, and pioneered many computational studies investigating their molecular mechanisms.

A better comprehension of taste receptor molecular behaviour and ligand-driven activity modulation is a crucial scientific challenge in the wider research concerning the complex mechanisms that drive toward the cascade of supramolecular, cellular, and tissue-level events emerging as an elaborated taste sensation. The molecular-scale investigation is a first, irreplaceable step and computational molecular modelling, due to its atomistic resolution, represents a powerful tool to explore receptor structure-to-function relationships and to elucidate ligand roles in driving taste receptor activity. This type of investigation allows to quantitatively characterize the ligand-binding process, thermodynamics and kinetics of the binding mechanism, binding modes, and ligand-target interaction properties, along with quantitative measures of receptor activation/inhibition, local and global protein rearrangements, correlations between receptor domains, transition pathways between active-resting conformations, etc. Ligand-receptor binding investigations allow the evaluation of food molecular constituents in terms of specificity, selectivity and multi-target features and shed light on the natural role of taste receptors in preserving life by discriminating between healthy and dangerous

foods. Despite the enormous progress made in recent years, especially in molecular research and in the computational investigation of ligand-receptor interaction related to taste receptors, the scientific knowledge remained rather granular and unable to explain the latter in a holistic fashion. Thus, it remains of crucial interest to correctly frame the mechanisms involved in the transfer of taste information from the chemistry level, where food molecular constituents bind taste receptors, to molecular-, supramolecular- and cellular-level events, which ultimately manifest as a composite perception strongly linked to the food organoleptic profile.

2.2 VirtuousPocketome

The present section is based on the manuscript under preparation:

Pallante, L. *†, Cannariato, M.* *†, Androutsos, L., Zizzi, E.A., Hada, X., Mavroudi, S., Grasso, G., Theofilatos, K., & Deriu, M. A. (2023).* ***VirtuousPocketome: A Computational Tool for Screening Protein-ligand Complexes to Identify Similar Binding Sites****.*

† Lorenzo Pallante and Marco Cannariato contributed equally to this study.

Author's contribution to the publication: *Pallante L. supervised and wrote all the python codes underlying the pipeline, contributed to the design of the framework, conducted the molecular dynamics simulations, analyzed the results, wrote, and revised the manuscript.*

Based on the pieces of literature reported in the previous section, we were interested in defining a computational protocol to analyse the specific protein-ligand interactions between taste receptors and tastants underlying the protein activation. Starting from this information, we desired to design an automatic pipeline to retrieve other proteins outside the gustatory system sharing similar residues patterns in their binding pockets compared to taste receptors to highlight possible off-targets potentially triggered by food tastants. In this section, we present a computational pipeline, named VirtuousPocketome, able to analyse a protein-ligand complex and screen the entire currently solved human proteome for proteins sharing similar binding sites.

Protein residues in a binding pocket define the spectrum of ligands that can effectively bind the protein structure and eventually modify its function. Different proteins involved in different functions can share similar binding pockets and can be triggered by similar ligands. Therefore, recognizing structural similarities in proteins can be a valuable strategy for gaining insights into protein function and activation mechanisms. The characterization of amino acid arrangements of a binding pocket and the identification of similar patterns in different structures can increase our understanding regarding specific protein-ligand interactions, predict off-target effects, and facilitate the development of more selective and effective therapeutic agents. Several computational methods quantifying the global or local similarity of protein cavities have been developed in the past years. Nonetheless, the utilization of these methodologies is substantially hindered by their intricate nature, the inherent impracticality of automating the search for amino acid patterns, and the inability to evaluate the dynamics of the protein-ligand systems under scrutiny. Here, we present a general and automatic computational pipeline, named VirtuousPocketome, aimed at screening huge databases of proteins for similar binding pockets starting from an interested protein-ligand complex. The proposed pipeline can automatically detect the most important protein residues involved in ligand binding, also form a molecular dynamics trajectory, and screen for similar binding sites in the solved human proteome, evaluating the accessibility and reliability of the retrieved pockets. We demonstrate the pipeline's potential by exploring a recently solved human bitter taste receptor, i.e. the TAS2R46,

complexed with strychnine, screening the entire solved human proteome for similar binding sites. The application of this kind of analysis on receptors involved in taste perception appears particularly interesting, considering the close connection between diet and the maintenance of homeostasis or the onset of diseases, such as cardiovascular disorders and diabetes. Therefore, VirtuousPocketome was here employed to investigate the potential roles of food molecules within domains not directly related to the gustatory system and identify the most similar targets. We pinpointed 145 proteins sharing similar binding sites compared to the analysed bitter taste receptor and the functional enrichment analysis highlighted the main biological processes, molecular functions and cellular components related to the retrieved proteins. This work represents the foundation for future studies aimed at understanding the effective role of tastants outside the gustatory system: this will pave the way towards the rationalization of the diet as a supplement to standard pharmacological treatments and the design of novel tastants-inspired compounds to target proteins involved in specific diseases or disorders. The proposed pipeline will be released soon as a publicly accessible webserver, can be applied to any protein-ligand complex, and could be easily expanded in the future to screen any database of protein structures.

2.2.1 Introduction

In the field of structural biology, it is widely recognized that there is a strong relationship between the three-dimensional structure of a protein and its function [128]. Therefore, the recognition and analysis of structural similarities in proteins can represent a valuable strategy to gain insights into protein functions. In particular, the importance of characterizing specific amino acid arrangements of a binding pocket and identifying similar patterns in different structures lies in its potential to improve the understanding of protein-ligand interactions, predict off-target effects, and facilitate the development of more selective and effective therapeutic agents. By identifying binding pockets with similar amino acid patterns, researchers can predict potential off-target proteins for a given ligand, which can help in the design of drugs with minimal side effects. Additionally, the characterization of amino acid arrangements in binding pockets can contribute to the development of structure-based drug design methods to engineer drugs targeting specific protein families or selective ligands for individual proteins. This level of selectivity can be essential in the treatment of diseases, particularly when multiple proteins share similar functions but are involved in distinct physiological processes. As an example, we recently employed an embryonic version of the workflow presented here to understand the druggability of a query-binding site searching for similar motifs in proteins able to bind ligands of interest [129].

In the past years, several computational methods quantifying the global or local similarity of protein cavities have been developed. All these methods share three general methodological steps: (i) three-dimensional analysis of the structures of

interest; (ii) structure comparison; (iii) quantification of similarity through a metric (a scoring function). Many different representations of a given binding site are possible with varying degrees of retained information, e.g. the type of amino acid residues that interact with the ligand, or representing the binding site through a surface onto which the physical-chemical characteristics are projected, and even considering protein-ligand interactions. The first two methods can be regarded as structure-based, i.e. they stem from observing the structure of the protein. As far as the actual comparison strategies are concerned, these 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. 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 [130]. A summary of the main similarity search methods available in the literature is reported in the Supplementary Information (Table A - 6.3.1).

The possibility of screening a high number of proteins for similar binding pockets can be particularly helpful and fruitful for those complex mechanisms and processes which may involve similar receptors and ligands for very different functions. In this context, we decided to turn our attention and use the proposed pipeline to explore some of the main actors involved in taste perception, due to the strong relationship between food intake and homeostasis regulation, disease onset, immune response and metabolism. The sense of taste is a sensory modality that plays a fundamental role in discriminating ingestible substances and nutrients from potentially harmful substances that must be avoided, especially in omnivorous species given the range of their feeding strategies [131]. Humans, in particular, can perceive five primary taste qualities, i.e. sweet, umami, bitter, salty, and sour, through the interaction of molecules contained in food and specialized proteins, namely taste receptors, located on the papillae of the tongue. However, taste receptors are also expressed in other tissues besides the oral cavity, including the skin [132,133], brain [134], pancreas [135,136], heart [137,138], urethra [139], airway [78] and gastric [79] smooth muscle cells.

Moreover, taste receptors are not only involved in the gustatory function, but they participate in other regulatory activities, such as regulation of metabolic activity [140,141], innate immune response and bronchodilatation [140,142], diabetes and obesity, glucose level maintenance, appetite regulation, as well as hormone release [143], and muscle contraction/relaxation [80].

In the present work, we decided to focus our attention on the bitter taste perception and relative actors, given the high scientific output produced in recent years concerning this specific taste sensation. Bitter taste receptors are the proteins responsible for the recognition of bitter foods, normally associated with potentially harmful substances. From a structural point of view, bitter taste receptors are GPCRs belonging to the taste 2 receptor family (*TAS2Rs*) [70] and are characterized by seven transmembrane helices (TMD) connected by three Intracellular Loops (ICLs) and three Extracellular Loops (ECLs) [72]. The structural core of the 7 TMD bundle is conserved across class-A GPCR and TAS2Rs. This core plays a fundamental role in the ligand binding in the extracellular (EC) region and information transduction in the intracellular (IC) region [73]. Bitter taste receptors can be activated by a multitude of different agonists through various interaction types in their unique orthosteric binding pocket [71,144]. Based on the chemical heterogeneity of their agonists, TAS2Rs have been distinguished into *promiscuous*, such as TAS2R10, TAS2R14, and TAS2R46, which are activated by a variety of chemically diverse compounds, and *selective*, activated instead by a limited number of similar compounds [74].

Herein, we propose an automated pipeline, named VirtuousPocketome, to screen databases of protein structures to identify amino acid patterns that are similar to the ones forming the ligand binding site of a query receptor. Compared to previous literature, the novelty of this work resides in three main aspects: (i) the proposed pipeline accounts for the dynamics of the protein-ligand interaction by considering multiple binding site configurations obtained from a molecular dynamics (MD) trajectory; (ii) the identification of the crucial protein-ligand binding interactions is completely automatic; (iii) the results of the similarity search are filtered using an ad-hoc multi-step filtering process to exclude patterns unlikely to bind the ligand of interest. The algorithm builds structural motifs of the target binding site of the query receptor-ligand complex after clustering a molecular dynamics trajectory and then searches for similar patterns inside a specific protein database using the ASSAM code [145] followed by additional ad-hoc filtering steps. We applied our computational pipeline to screen the currently solved human proteome for proteins that exhibit a highly similar local amino acid pattern to the one lining the strychnine binding site in the human TAS2R46 bitter taste receptor. The rationale of the work is to explore the taste transduction pathway with a proteomic perspective to elucidate the possible role of tastants beyond the mere taste perception and to investigate whether other classes of proteins have a conserved ability to recognize such ligands, with possible implications in nutrition, homeostasis, and disease.

2.2.2 Materials and Methods

VirtuousPocketome Workflow

Overall Workflow

The proposed computational pipeline requires as mandatory inputs the coordinates (PDB file) of a protein-ligand complex and the chain label(s) uniquely identifying the receptor(s) and the ligand in the provided structure. Additionally, the user can provide the molecular dynamics trajectory (GROMACS xtc/trr/pdb file) of the complex system and additional custom parameters (details in the following).

At present, the code is limited to the screening of the entire human proteome, but it can be easily expanded to any PDB-like database.

The overall workflow of the algorithm designed in the present work is divided into four main steps, each of which will be described in detail in the following paragraphs:

1. Motifs Creation
2. Similarity Search
3. Multi-step Filtering
4. Functional Enrichment and Signalling Pathway Analyses

The main output consists of a txt file collecting the PDBs of the identified proteins sharing similar accessible binding site(s) compared to the query receptor-ligand complex. Additionally, the code provides as output the list of unique UniProt ids related to the retrieved proteins, plots summarising the main results (see the Results section), and visualisation states of the molecular system under investigation.

The overall workflow is represented in Figure 2.7 in a flow chart representation.

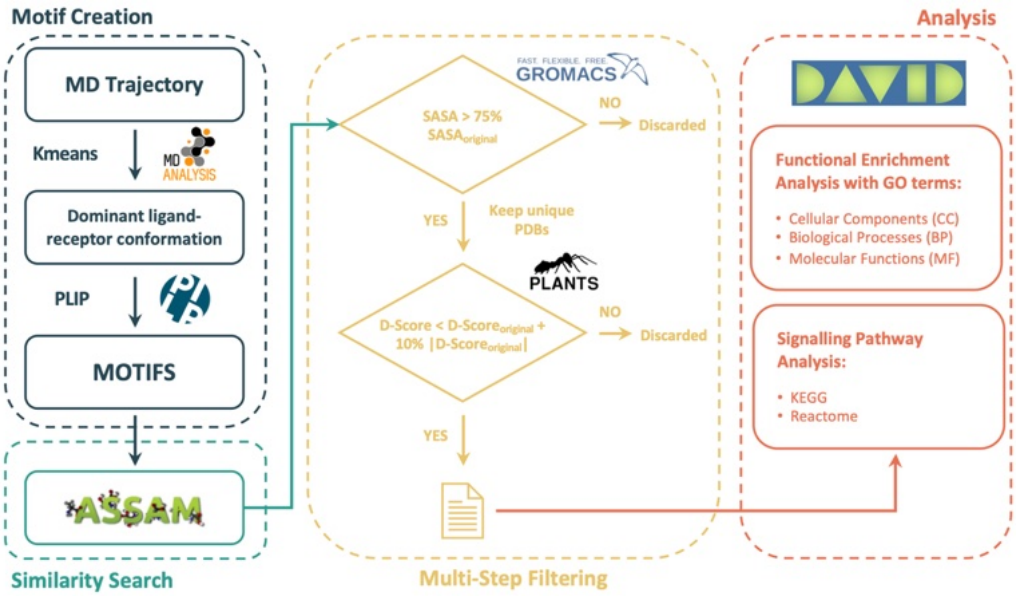


Figure 2.7. Flow chart of the overall workflow of VirtuousPocketome.

For the implementation of this workflow, we mainly used GROMACS functionalities [146], MDAnalysis modules [147], and pdb-tools [148]. The ASSAM source code was kindly provided by Prof. M. Firdaus Raih from the Molecular Function Regulation Lab (http://mfrlab.org).

Step 1 - Motifs Creation

Starting from the provided PDB input file and the MD simulation (if present) of the protein-ligand complex, a list of residues defining the binding site is retrieved based on the distance between the ligand and the receptor. The default distance threshold is set to 10 Å, but this can be overridden with a custom value provided by the user as an additional parameter. If the user only provides a single PDB file of the protein-ligand complex, a single binding site is defined according to the chosen distance threshold between the ligand and the protein. Conversely, if the user provides also the MD trajectory, the obtained subset of coordinates from each frame of the simulation identifying the ligand and the residues of the protein binding site are clusterized using the K-Means algorithms implemented in MDAnalysis [147]. The user can set the desired number of clusters (k), otherwise, the optimal number of clusters is retrieved using the silhouette method ranging from a minimum of 2 clusters up to a maximum of 12 clusters. The silhouette method, as provided by the built-in scikit-learn function (sklearn.metrics.silhouette_score), is an example of an evaluation metric to indicate if clusters are well-defined. This method computes the mean distance between a sample and all other points within the same class (a) and the mean distance between a sample and all other points in the nearest neighbouring cluster (b). The Silhouette Coefficient, s, for a single sample, is then calculated as reported in Equation (2.1:

$$s = \frac{b - a}{\max(a, b)} \tag{2.1}$$

Then, the total Silhouette Coefficient is calculated as the mean of the Silhouette Coefficient for each sample. A higher Silhouette Coefficient score is indicative of denser and well-separated clusters, aligning with the conventional understanding of a cluster. It suggests that the samples within each cluster are tightly packed and distinct from samples in other clusters. This reflects a higher degree of cohesion within clusters and a greater separation between them. Therefore, the algorithm calculates the silhouette coefficient for different numbers of clusters (from 2 to 12) and then chooses the optimal number of clusters according to the best-achieved silhouette coefficient score. The subsequent centroids of the clusters are saved.

Starting from the previously defined binding sites, the subset of residues forming the binding site is further refined, highlighting those residues involved in non-covalent interactions with the ligand. These interacting residues are retrieved using the Protein-Ligand Interaction Profiler (PLIP) software [149], which detects hydrogen bonds, hydrophobic contacts, pi-stacking, pi-cation interactions, salt bridges, water bridges, metal complexes, and halogen bonds between ligands and targets. This two steps analysis allows the definition of the most relevant protein residues for the interaction with the ligand. This subset of residues will be defined as *motifs* in the following. If only the PDB file is passed by the user, a single motif will be created; if also an MD trajectory is provided, a motif will be produced for each centroid of the binding site.

Step 2 - Similarity Search

In the second step, the similarity search for each extracted motif is carried out using the ASSAM software. The fundamental principles underlying the search methodologies of ASSAM are described in previous literature [150,151]. In brief, the protein structure is represented as a graph, where nodes represent individual amino acid side chains and the geometric relationships between nodes form the edges of the graph. Each node is composed of two pseudo-atoms, which generate vectors that correspond to the nodes in the graph. The positions of these pseudo-atoms are strategically chosen to emphasize the functional portion of the corresponding side chain. The geometric relationships between pairs of residues are defined by distances calculated between their corresponding vectors, and these relationships are represented as the edges of the graph. If we let S, M and E denote the start, middle and end of a vector, the edges of the graph encompass five components: SS, SE, ES, EE, and MM distances, although only a subset of these distances is typically used to specify a query pattern. ASSAM employs a maximal common subgraph (MCS) approach using the Bron and Kerbosch MCS algorithm to enumerate all possible correspondences with similar protein patterns[152]. After an extensive

analysis of the available methods from previous literature (see also Table A - 6.3.1), 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. a coordinate file containing the residues of interest in PDB format, (ii) the possibility to screen against any desired database in PDB format, (iii) the comparably fast time to solution, and (iv) the possibility of considering and preserving both right-handed and left-handed orientations of the alpha-helices to retrieve the superpositions. A left-handed α-helical bundle superimposed onto a right-handed one is not equivalent, particularly concerning the handedness of the two groupings of amino acids. However, when it comes to chemical activity, two groupings of amino acids may exhibit the same behaviour despite having different handedness. The crucial factor, in this case, is the distance between the individual residues [145]. Thanks to the source code kindly provided by the ASSAM developers, the proposed pipeline does not rely on the ASSAM web-based analysis tools and runs entirely offline on-premises.

The output of the ASSAM search is formatted as a list, in which each row corresponds to a *hit* protein found in the screened protein database, along with its PDB accession ID, which presents a match with the residues of the input motif, 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 between query residues and hit residues after alignment. Finally, additional pieces of information are retrieved, namely 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 publication and the EC classification.

Step 3 - Multi-step Filtering

To further refine the output from the previous steps and select only the protein hits with accessible and high-affinity binding sites, two additional filtering steps, i.e. (i) the SASA and (ii) the docking filters, have been implemented.

The first step involves the calculation of the Solvent-Accessible Surface Area (SASA) using GROMACS [146]. In detail, for each hit protein, the SASA, evaluated in nm^2, is calculated by measuring the value for all the residues matching the residues of the query protein's motifs. Values equal to zero indicate that the cleft identified by the residues is not solvent-exposed but is rather buried in the structure of the protein and therefore likely inaccessible for the solvent or 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 ligand binding. Only hits with the binding site having a SASA greater than a predefined threshold of the SASA value of the corresponding query motif are retained. The SASA criterion is summarised by equation (*2.2*):

$$SASA_{HIT} > SASA_{query} * SASA_{threshold} \quad (2.2)$$

Where $SASA_{HIT}$ corresponds to the calculated SASA value of the given hit motif, $SASA_{query}$ is the SASA value of the corresponding motif of the query protein-ligand complex, and $SASA_{threshold}$ is the threshold to select only hits with the desired SASA. $SASA_{threshold}$ is set to 75% by default, but the user can specify a custom value in the additional input parameters. The selection of this threshold was imposed as a compromise to obtain a reasonable number of protein hits that could be effectively and rationally considered, while retaining targets with motifs that are solvent-exposed and prone to ligand docking. The decision to set as default a high SASA threshold is based on the importance of solvent-accessible surface area (SASA) in assessing the capability of a protein binding site to accommodate ligands. Considering only SASA values considerably lower than the original protein binding site could compromise the algorithm's ability to retain only those binding sites most similar to the reference complex and thus more prone to ligand binding. Setting the SASA threshold to higher values, on the other hand, may lead to the exclusion of potentially promising binding sites and protein hits. As a result, users have the flexibility to adjust this threshold based on their screening objectives.

The second filtering step relies on the molecular docking of the original ligand in the query complex onto the retrieved hits from the previous steps. The docking procedure was implemented using SPORES and PLANTS [153,154]. PLANTS software was chosen since it exhibited excellent performance in terms of pose prediction and time to solution compared to other molecular docking software[155,156] and since it has been already successfully used for molecular docking and virtual screening campaigns on GPCRs[157–161]. These qualities make PLANTS particularly well-suited for the present work. Only hits with binding sites exhibiting a docking score (DSCORE) below a specified docking threshold are retained, thus only preserving binding sites with high affinity for the investigated ligand. The docking threshold is set by default as 10% of the original protein-ligand complex, if only a PDB is provided as input, or the 10% of the average PLANTS docking scores between the centroids of the original protein-ligand complex, if also the MD trajectory was specified. The docking filtering criterion is therefore defined by equation (*2.3*):

$$DSCORE_{HIT} < \overline{DSCORE}_{query} + \left|\overline{DSCORE}_{query}\right| * DSCORE_{threshold} \quad (2.3)$$

Step 4 - Functional Enrichment and Signalling Pathway Analyses

In this step, functional enrichment and signalling pathway analyses were performed to collect information regarding roles, functions, distributions, expressions, and pathways in which the identified targets from the previous steps are involved. The

pipeline automatically retrieves the unique UniProt IDs of the protein hits (since several PDB codes can correspond to the same protein) and searches for related genes using the DAVID functional annotation program [162,163]. We decided to focus our attention on the Gene Ontology (GO) terms [164] related to the Cellular Components (CC), Biological Processes (BP) and Molecular Functions (MF). The resulting GO terms were used to identify significantly enriched functional categories, using a statistical approach that considers the size of the protein list, the number of genes associated with each GO term, and the background gene set. We also performed pathway analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Reactome databases to identify the most represented pathways [165–167]. GO terms and KEGG pathways with corrected p-value < 0.1 were considered to be significantly enriched. The correction of the p-values was performed using the Benjamini-Hochberg FDR adjustment method [168]. The pipeline automatically generates separate plots with the above-mentioned analysis.

Database Curation

The developed tool can screen a database of target proteins in the PDB format for binding pocket similarities against a query protein-ligand complex. For the present work, we have collected and refined all the experimentally-solved PDB structures belonging to the human proteome. We first retrieved all the solved PDB codes belonging to *homo sapiens* from the NCBI database [169] using a dedicated python API (Bio.Entrez package), and obtained a total of 60159 entries (1st February 2023). We downloaded all the found queries from the RCSB database (http://www.rcsb.org/) [170], reaching a total of 59267 structures (892 PDBs were not available). Then, the database was cleaned by removing (i) the chains in the PDBs not belonging to the human organism (such as in the case of protein chimaeras), (ii) multiple models from each PDB, (iii) ANISOU and HETATM lines, (iv) alternative locations of the atoms in the PDB by preserving the ones with the highest occupancy. At the end of this cleaning protocol, we ended up with 58972 PDB files.

Molecular Modelling and Dynamics

The previously described workflow was applied to search for proteins sharing similar binding pockets with a human bitter taste receptor. We employed the recently-solved TAS2R46 human bitter taste receptor bound with an agonist bitter compound, named strychnine (PDB ID: 7XP6) [171]. We first removed undesired molecules and structures from the PDB, preserving only the bitter taste receptor and strychnine. Since the receptor structure presents some missing residues (157-172), we downloaded the relative model (sequence P59540 of Homo Sapiens hTAS2R46) from the AlphaFold Protein Structure Database [172]. The AlphaFold model was then aligned to the receptor in the 7XP6 PDB. Missing residues in the original experimental structure were then built by homology modelling with MOE (Molecular Operating Environment) software [173] using the relative portion of the AlphaFold model as a template, thus filling the gap in the original experimental

structure. Then, the structure of strychnine was refined using MOE, assigning the correct protonation at neutral pH and salt concentration of 0.15 M.

The obtained complex was embedded into a POPC membrane using the CHARMM-GUI web server [174]. The receptor was aligned with the membrane bilayer using the OPM web server included in the CHARMM-GUI preparation protocol. The final number of POPC lipids (165) was set by a trial-and-error approach to creating a box large enough to meet the minimum image convention for the protein during the MD simulation: the final box size was set to 8.36 x 8.36 x 11.47 nm, with the z direction perpendicular to the cell membrane plane. The system box was filled with water and neutralized using NA^+ and CL^- ions at 0.15 M physiological concentration.

We used the AMBER forcefield to describe the molecular system: in detail, we used the Lipid-21 forcefield [175] for lipids, the AMBER19SB [176] for the protein, ions and water and the General Amber Force Field (GAFF2) forcefield [177] to obtain the topology for strychnine, as implemented directly in the CHARMM-GUI suite.

The simulation workflow suggested by CHARMM-GUI was followed during minimisation, equilibration with position restraints and final simulation production. First, the system was energy-minimized using the steepest descent algorithm for 5000 steps. Then, six equilibration steps were performed gradually reducing the position restraints on the lipids and protein-heavy atoms (from 1000 to 0 $kJmol^{-1}nm^{-1}$ for lipids, from 4000 to 50 $kJmol^{-1}nm^{-1}$ for the protein backbone and from 2000 to 0 $kJmol^{-1}nm^{-1}$ for protein side-chain heavy atoms). The system was equilibrated in the NVT ensemble for 250 ps with a conservative timestep of 1 fs, using the Berendsen thermostat [178] with a coupling time constant of 1 ps and a reference temperature of 303.15 K, which is above the phase-transition temperature for POPC, and subsequently in the NPT ensemble for 125 ps with the same 1 fs timestep, followed by a further simulation of 375 ps with a 2 fs timestep, using the Berendsen thermostat with the same parameters as before and the Berendsen barostat [178] with semi-isotropic pressure coupling at 1 atm with a coupling time constant of 5 ps. Overall, the systems underwent 750 ps of equilibration.

The TAS2R46 bitter taste receptor system was then simulated for 400 ns without position restraints with a 2-fs time step in the NPT ensemble. Temperature coupling was done with the Nose-Hoover algorithm [179], whereas, pressure coupling with the Parrinello-Rahman algorithm [180]. The PME algorithm was used for electrostatic interactions with a cut-off of 0.9 nm. A reciprocal grid of 72 x 72 x 96 cells was used with 4th-order B-spline interpolation. A single cut-off of 0.9 nm was used for Van der Waals interactions. LINCS (LINear Constraint Solver) algorithm for h-bonds [181] was applied in each simulation step.

Three simulation replicas were performed to ensure the reproducibility of the simulations and to enlarge the simulation statics. All Molecular Dynamics (MD) simulations were performed using GROMACS [146] and the Visual Molecular

Dynamics (VMD) package was employed for the visual inspection of the simulated systems [182].

2.2.3 Results

Conformational Dynamics

To ensure the convergence of the MD simulations, we evaluated the RMSD and the cluster analysis as measures of simulation equilibrium for all the MD replicas. In detail, we calculated the RMSD of the protein backbone and the number of clusters during the last 50 ns using the linkage method with an RMSD cutoff of 0.15 nm. The RMSD trends reached a plateau (Figure A - 6.3.1) and only a single cluster was obtained during the last part of the simulation for all the replicas. The last 50 ns of each simulation replica were therefore considered as structural equilibrium and were concatenated to obtain a final 150 ns-long trajectory representing the ensemble of protein conformations. These concatenated trajectories were used for the subsequent analysis and to search for similar binding pockets within the human proteome through the pipeline described herein.

Motifs Creation

In the first step of the proposed pipeline, the motifs composed of the most important protein residues interacting with strychnine were identified. In particular, starting from the above-mentioned ensemble trajectory, residues within 10 Å from the position of the ligand have been extracted and their conformations clustered using the K-Means algorithm from MDAnalysis [147]. Three clusters were identified as the optimum number through the silhouette method. After extracting the motif cluster centroids, we used PLIP to narrow down the most important non-covalent interactions which define the final three motifs. In all three of them, the ligand formed a salt-bridge interaction with residue GLU265$^{7.39}$ and two hydrophobic interactions with residues TYR85$^{3.29}$ and TRP88$^{3.32}$, highlighting the stability of these contacts throughout the MD simulation. Furthermore, additional hydrophobic interactions were found with residues THR69$^{2.64}$, PHE252$^{6.58}$, and PHE261$^{7.35}$ for motif 0 and with VAL61$^{2.56}$, and VAL249$^{6.55}$ for motif 2 (Figure 2.8). The three defined motifs were used for the subsequent similarity search step with ASSAM.

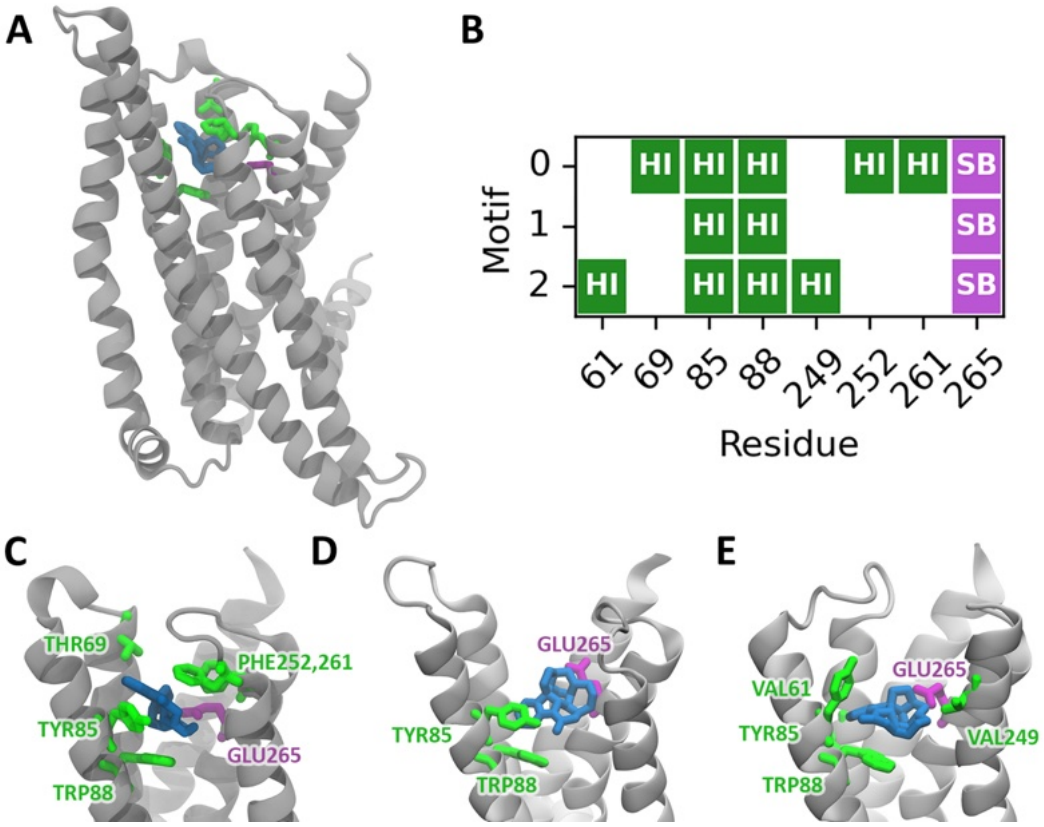


Figure 2.8. Main interactions defining the three motifs for the bitter taste receptor interacting with strychnine. (A) Representative snapshot of the bitter-ligand complex, (B) PLIP interaction analysis identifying Hydrophobic Interaction (HI) and Saltbridge Interaction (SB), (C, D, E) site views of the three motifs identified. The bitter taste receptor is represented in grey, the strychnine in blue and the interacting residues in green (hydrophobic interactions) and purple (salt bridges).

Similarity Search and Multi-step Filtering

The similarity search against the entire human proteome consisting of 58972 structures (see the Step 4 - Functional Enrichment and Signalling Pathway Analyses section) resulted in a total of 6718 hits using the ASSAM code. The subsequent steps, i.e. the SASA and docking filtering steps, yielded a total of 1852 and 257 hits respectively (Figure 2.9A). In detail, we adopted a SASA threshold of 0.75, thus preserving all protein hits having a SASA in the binding pocket of at least 75% of the SASA of the original query binding site; on the other hand, we chose a docking threshold equal to 0.1, thus keeping the hits whose docking score would not differ by more than the 10% from the docking score of the query protein/ligand complex. The best hit in terms of docking score after the multi-step filtering process, namely PDB 3A4S, is represented in Figure 2.9B, highlighting the correspondence between the original motifs in the bitter taste receptor and the relative matching residues in the identified protein.

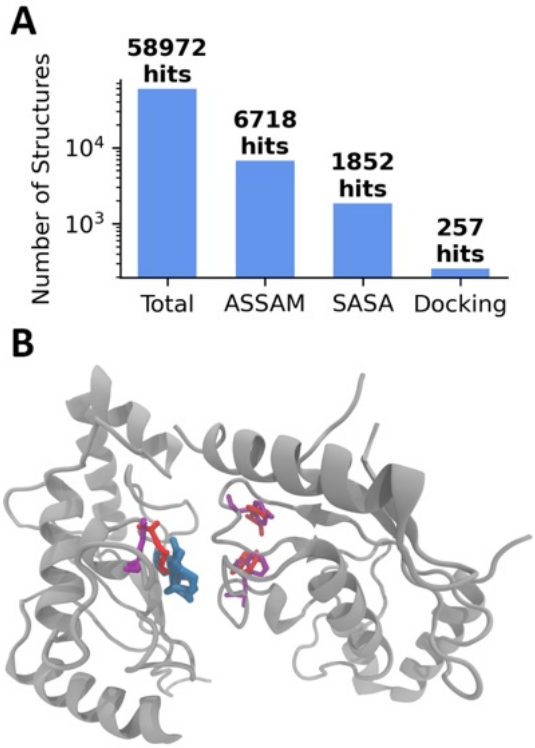


Figure 2.9. (A) Number of total structures in the original human proteome database and number of selected hits from the similarity search and the subsequent multi-step filtering. (B) Binding site view of the strychnine bound to the best hit (PDB: 3A4S) according to the docking score at the end of the multi-step filtering process. Protein is rendered in grey, strychnine in blue, residues in the original motif of the bitter taste receptor in red and matching residues in the 3A4S structure in violet.

The pipeline automatically generated a report file which includes the list of all the hits at the end of the screening process with additional information, such as PDB ids with doi of the relative publication, protein classes, shared residues between hits and query, SASA and Docking Scores. The complete list of the 257 retrieved protein hits is reported in Table A - 6.3.2.

Functional Enrichment and Signaling Pathway Analyses

VirtuousPocketome retrieved 145 unique Uniprot IDs relative to the previously identified hits, meaning that multiple PDBs at the end of the multistep filtering process corresponded to the same protein. The DAVID software was then employed to analyse the Gene Ontology terms and the signalling pathways data as described in the Material and Methods section.

The functional enrichment analysis revealed that the input genes were significantly enriched for a total of 16 GO terms in the Cellular Components category, 10 terms in the Biological Processes category and 14 terms in the Molecular Functions category, based on the corrected p-value. In addition, 0 KEGG and 0 Reactome

pathways were found to be significantly enriched at the same p-value threshold. The best 5 GO terms for each of the above-mentioned categories are represented in Figure 2.10, whereas all the significantly retrieved GO terms are represented in the Supplementary Information (Figure A - 6.3.2 and Figure A - 6.3.3). Regarding the Biological Processes (BP), most of the retrieved genes are related to metabolic processes (40.9%), including organic substance, cellular and nitrogen compound metabolic processes. Besides, the most represented Molecular Functions (MF) are related to the binding of different species, such as proteins, small compounds or ions, and to enzyme activity, such as transferase, hydrolase, and oxidoreductase. Finally, regarding the last analysed GO term, the most represented Cellular Components (CC) are cytoplasm and membrane (39.0%).

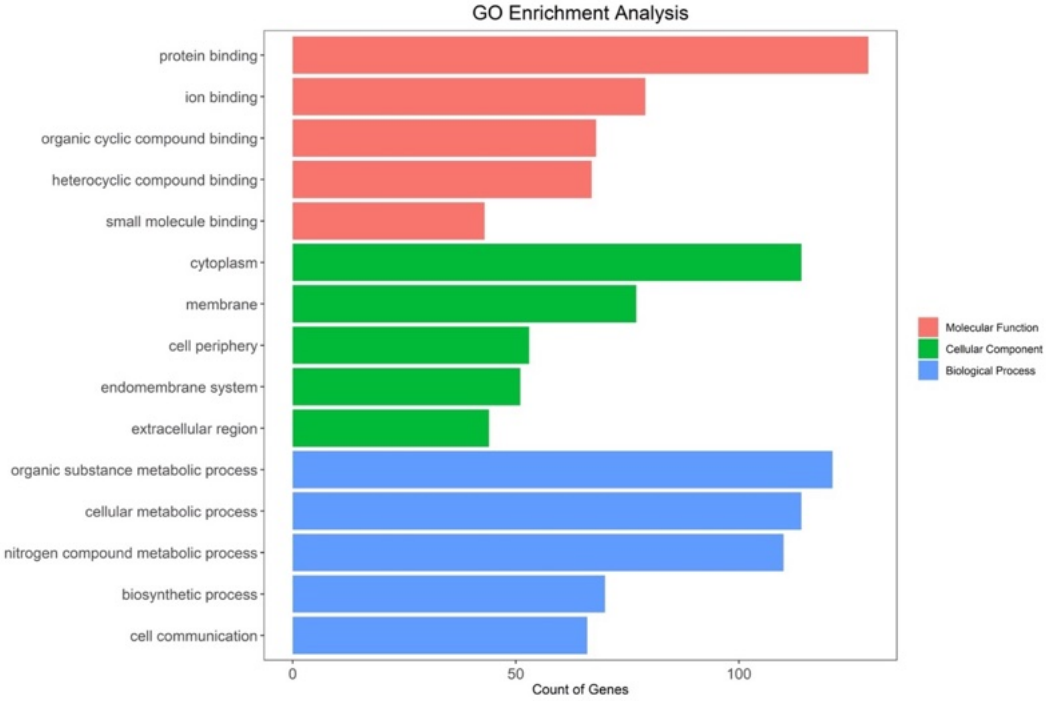


Figure 2.10. Bar plots representing the best 5 retrieved GO terms for each category in the third level of the GO hierarchy relative to Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC).

2.2.4 Discussion

Herein, we developed a novel computational pipeline, named VirtuousPocketome, to screen a desired database for proteins sharing similar binding sites with a specific protein of interest. VirtuousPocketome is based on four major steps: (i) the motifs creation step, which identified the most important residues involved in the interaction of the protein-ligand complex under investigation; (ii) the similarity search step, in which the human solved proteome is screened for similar motifs compared to the ones retrieved in the previous step; (iii) the multi-step filtering step,

which preserves only the protein hits with binding sites effectively accessible and with a certain docking affinity for the query ligand; (iv) functional enrichment and signalling pathway analyses, which identifies the most relevant cellular components, molecular functions, biological processes and signalling pathways related to the protein hits identified in the previous steps. VirtuousPocketome takes as input the molecular structure (and eventually also the molecular dynamics trajectory) of a protein-ligand complex and gives as output a list of PDB structures sharing similar binding sites. The developed protocol is automatic and can be applied to any protein-ligand complex.

To evaluate the developed code and prove its potential, we here selected the human TAS2R46 bitter taste receptor bound to a bitter ligand, i.e. strychnine and we screened the entire human proteome to search for similar structural motifs. The choice for this molecular system was driven by the recent experimental determination of the TAS2R46-strychnine complex structure [171], as well as the fact that strychnine is experimentally known to target not only TAS2R46 [71] but also other bitter taste receptors, such as TAS2R10 [183], and even other proteins [184]. Moreover, bitter taste receptors have been widely investigated also by means of computational molecular modelling in past years, ensuring enough data for comparison and evaluation of some of the results of the platorm [144,183,185–188]. Furthermore, the strong relationship between food intake and the health status, including the regulation of homeostasis and metabolism, makes the chosen molecular machinery a particularly intriguing and relevant testbed for our computational screening pipeline to pinpoint possible secondary targets outside the gustatory system for food-related tastants.

The first step of the proposed pipeline, i.e. the motifs creation step, applied to the TAS2R46 bitter taste receptor bound to strychnine pinpointed three major motifs of residues mostly involved in the ligand binding. In particular, all three motifs shared a salt-bridge interaction with residue GLU265$^{7.39}$ and two hydrophobic interactions with residues TYR85$^{3.29}$ and TRP88$^{3.32}$, whereas the first motif comprised also hydrophobic interactions with residues THR69$^{2.64}$, PHE252$^{6.58}$, and PHE261$^{7.35}$ and the third motif with VAL61$^{2.56}$, and VAL249$^{6.55}$ (Figure 2.8). Interestingly, some of these residues have already been suggested by previous literature to be important interactions for ligand binding of bitter taste receptors. In detail, GLU265$^{7.39}$ and TRP88$^{3.32}$ were demonstrated to be pivotal in TAS2R46 activation by strychnine [171]. Moreover, mutagenesis studies have reported the importance of residue GLU265$^{7.39}$ for agonist responsiveness of TAS2R46 [71] and the same position has been also linked to ligand binding for similar receptors, including hydroxytryptamine (5-HT) receptors [189,190], adrenergic receptors [191,192], purinergic receptors [193,194], and cholecystokinin-B (CCK-B)/gastrin receptor [195]. Moreover, residue TRP88$^{3.32}$ is widely conserved among TAS2Rs and was found to be crucial in the activation of TAS2R43, TAS2R30 and TAS2R46 [71,196]. The importance of residue in position 3.29 (TYR85 for TAS2R46) was confirmed also

for TAS2R10, which shows a similar binding site to TAS2R46 and is also activated by strychnine [183]. These pieces of the literature confirmed the reliability of the motifs creation step of the proposed pipeline to pinpoint the most important and relevant residues involved in the ligand binding. It is worth mentioning that the possibility of analysing a molecular dynamics trajectory allows the identification of multiple motifs, three in the present case, one for each identified cluster centroid, ensuring a more exhaustive sampling of the protein-ligand interactions.

Starting from the identified motifs, similar amino acid patterns were searched in the entire currently solved human proteome, consisting of 58972 structures. The similarity search using the ASSAM code resulted in 6718 hits. Then, the multi-step filtering reduced the number of detected sites to a total of 257 PDB structures, which represent 0.44% of the original database and 3.83% of the structures in ASSAM code output (Figure 2.9A). Therefore, the presented approach was revealed to be effective in discarding a high number of spurious proteins whose matching motifs were unlikely to bind strychnine, due to insufficient solvent exposure or low predicted affinity obtained from molecular docking.

The functional enrichment analysis allowed us to pinpoint the main biological processes, molecular functions, and cellular components related to the gene expressing the retrieved hit proteins at the end of the VirtuousPocketome pipeline (Figure 2.10). Most of the biological processes highlighted are connected to metabolic processes, which seems an intriguing result considering the strong relationship between taste perception, food intake and metabolism. Indeed, these results might indicate that strychnine has not only the ability to activate the TAS2R46 bitter taste receptor and elicit the bitter taste sensation, but it should be also the potential trigger or modulator of proteins directly involved in the metabolic processes. Moreover, the protein hits are mainly involved in molecular functions related to the binding of proteins or other small compounds and are localized mainly in the cytoplasm and membrane. These results seem reasonable considering that TAS2R46 is a transmembrane protein and a promiscuous bitter taste receptor, able to bind a wide spectrum of chemical compounds. In light of these results, it is worth mentioning that strychnine is known to bind also glycine and acetylcholine receptors [184], which are membrane proteins involved in transport and signalling functions. Therefore, the presented pipeline was able to detect similar binding sites in receptors with the same localization and biological function of known strychnine targets. In addition, strychnine is a rather promiscuous ligand with also anti-plasmodial and anti-cancer activity and other yet-unresolved molecular targets [197].

2.2.5 Conclusions

In the present work, we developed a novel, general and automatic pipeline for identifying proteins sharing similar binding pockets with a query receptor-ligand complex by screening the entire solved human proteome. The developed pipeline will be soon released as a webserver service and will be easily expanded in the

future to other protein databases, such as the AlphaFold Protein Structure Database [172]. We used the TAS2R46-strychnine complex as a testbed for the proposed method to investigate if other proteins except the taste receptors might share similar binding pockets for the recognition of tastants. The proposed methodology allows for a deep investigation of the TAS2R46-specific residues needed for the strychnine binding to pinpoint other human proteins sharing similar binding pockets and investigate the potential roles that this tastant could play in contexts beyond the gustatory system. The retrieved proteins could be further analysed to predict whether the interaction with the compound of interest results in their activation or modulation. This will increase our understanding of possible secondary effects of tastants beyond the mere taste perception and their impact on the biological processes and molecular functions in which the retrieved hit proteins are involved. This approach can also assist the design of specific foods and ingredients to develop personalised treatments able to target desired proteins or receptors involved in specific processes or diseases with the ultimate goal of accessing the potential of the diet as a supplement to traditional pharmacological treatments.

2.3 The Impact of Natural Compounds on S-Shaped Aβ42 Fibril

The present section is based on the following scientific publication:

*Muscat, S.[†], **Pallante, L.[†]**, Stojceski, F., Danani, A., Grasso, G., & Deriu, M. A. (2020). **The Impact of Natural Compounds on S-Shaped Aβ42 Fibril: From Molecular Docking to Biophysical Characterization.** International Journal of Molecular Sciences, 21(6), 2017. https://doi.org/10.3390/ijms21062017*

[†] Stefano Muscat and Lorenzo Pallante contributed equally to this study.

***Author's contribution to the publication:** Pallante L. performed all the molecular dynamics simulations, analyzed, and rationalized the data, wrote, commented, and revised the manuscript.*

In the previous section, we presented a novel computational pipeline to identify possible off-targets sharing similar binding sites compared to a query protein-ligand complex. The proposed approach is particularly interesting since it allows us to understand if an investigated ligand can potentially bind also other receptors besides its primary one. However, the proposed methodology does not evaluate the molecular influence of the compound on the target protein. Hence, in this section, our focus lay in comprehending the molecular foundation and evaluating the influence of small ligands on targets associated with specific pathologies. In this context, we used molecular modelling to assess the impact of selected natural compounds on a specific polymorphism of amyloid fibrils. The focus on natural compounds has been driven by their growing popularity for treating diseases or for being coupled to conventional pharmacological treatments, as well as their association with the diet, particularly the Mediterranean one, which fits perfectly with the objectives of this doctoral thesis work as well as those of the VIRTUOUS project. Given that natural compounds have inherent efficacy in influencing some pathological conditions, a thorough understanding of their molecular features that underlie their actions and their trajectory in the human body can ultimately aid in the rational design of diets and strategies to selectively target disease-specific proteins.

The pursuit of effective strategies inhibiting the amyloidogenic process in neurodegenerative disorders, such as Alzheimer's disease (AD), remains one of the main unsolved issues, and only a few drugs have been demonstrated to delay the degeneration of the cognitive system. Moreover, most therapies induce severe side effects and are not effective at all stages of the illness. The need to find novel and reliable drugs appears therefore of primary importance. In this context, natural compounds have shown interesting beneficial effects on the onset and progression of neurodegenerative diseases, exhibiting a great inhibitory activity on the formation of amyloid aggregates, and proving to be effective in many preclinical

and clinical studies. However, their inhibitory mechanism is still unclear. In this work, ensemble docking and molecular dynamics simulations on S-shaped $A\beta_{42}$ fibrils have been carried out to evaluate the influence of several natural compounds on amyloid conformational behaviour. A deep understanding of the interaction mechanisms between natural compounds and Aβ aggregates may play a key role to pave the way for the design, discovery and optimization strategies toward an efficient destabilization of toxic amyloid assemblies.

2.3.1 Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia. The mechanism of Alzheimer's onset and progression is still unclear, and several hypotheses have been proposed. One of the most accredited theories is the amyloid cascade hypothesis [198], which identifies as the main cause of AD progression the misfolding and the extracellular aggregation of Amyloid-β (Aβ) peptides from the cleavage of amyloid precursor protein (APP), as well as the intracellular deposition of the misfolded tau protein in neurofibrillary tangles. The Aβ aggregation leads to the formation of oligomeric toxic species, which can further aggregate in more ordered structures, called fibrils or fibres [199], up to the formation of extracellular senile plaques [200,201]. Among different lengths of Aβ peptides, senile aggregates are mostly made by the $A\beta_{40}$ fibrils, but the most toxic species are the $A\beta_{42}$ ones, due to their intrinsic tendency to self-assembly [202]. The stability of these structures is strongly linked with the progression and severity of the disease, and in the last years, many efforts have been made to characterize the molecular stability of amyloid aggregates [203–211].

In the past, several strategies have been developed to reduce or prevent Aβ production and to destabilize Aβ aggregates, including immunotherapeutic vaccines [212,213], antibodies [214,215], peptides [216,217], nanoparticles [218–220] and compounds targeting Aβ secretases [221,222] and Aβ aggregation [223–228]. However, some of these approaches have shown serious side effects [229,230] and poor permeability through the blood-brain barrier (BBB) [231]. In this context, small molecules based on natural compounds are promising inhibitors with minimal side effects and increased BBB permeability [232]. Several in vitro and in vivo studies have highlighted the potential therapeutic effects of natural compounds against neurodegenerative diseases, including AD [233–239]. However, their effects affect several aspects associated with AD, and their molecular mechanism of action is still not clear, consequently reducing the percentage of compounds at the clinical trial stage [200]. Hence, deep characterization of the molecular structure of amyloid aggregates and their interactions with promising compounds, such as natural ones, is of primary importance for the design of new efficient strategies against neurodegenerative diseases [240].

In this regard, computational methods, such as molecular dynamics (MD) simulations, thanks to a detailed molecular resolution, could represent a powerful tool to shed light on the molecular mechanisms characterizing physiological and pathological phenomena [241]. Thanks to these methods, several small molecules have been proposed as amyloid anti-aggregating agents [242]. A promising inhibitor, referred to as wgx-50, has shown destabilizing effects against Aβ fibrils and inhibition of neural apoptosis and apoptotic gene expression [243–245]. Moreover, polyproline chains have demonstrated a conversion mechanism of the Aβ secondary structure from beta-sheet to random coil, highlighting the stabilizing role of amyloid C-terminal residues [246]. Sharma et al. have evaluated the stoichiometric ratio of caffeine to the Aβ-derived switch-peptide by a combination of experimental and computational approaches, observing the peptide disaggregation when the caffeine stoichiometric is ten times higher than the peptide one [247]. Furthermore, curcumin-like compounds have been synthetized and tested on $A\beta_{40}$ showing two binding sites, one in the 17–21 region and one near the Met35 [248], which have been previously observed by experimental and computational works [249,250]. Finally, the interactions between homotaurine, scyllo-inositol and the $A\beta_{42}$ peptide at the monomer level have been extensively investigated by very long replica exchange MD with solute tempering simulations of 160 μs for each system, showing conformational changes of the $A\beta_{42}$ monomer through a nonspecific binding mechanism [228].

In this context, it is worth mentioning that molecular modelling investigations have focused mostly on a specific $A\beta_{42}$ polymorphic structure, called U-shaped fibril [251]. However, the $A\beta_{42}$ may arrange also in other polymorphic structures, such as the S-shaped structural rearrangement [252]. Interestingly, recent works have indicated that the S-shaped structure is characterized by superior conformational and mechanical stability with respect to the U-shaped one, suggesting a correlation between structural stability and toxicity [208,209,253].

Based on the above-mentioned premises, this research work investigates the binding and action mechanisms of 57 natural compounds targeting the S-shape $A\beta_{42}$ fibril by ensemble docking and molecular dynamics (MD) simulations. Ligands were selected starting from previous literature, and in particular in vivo or in vitro data showing their effect on the onset and progression of several neurological diseases, including AD [200,237,254–256]. Our results revealed the ligand's specific mechanisms of action on amyloid aggregates. More in detail, ligands can be distinguished based on their ability to disrupt or preserve the ordered conformational structure of the amyloid fibril.

2.3.2 Materials and Methods

Molecular Dynamics Setup

The atomistic structure of the S-shape $A\beta_{42}$ fibril was obtained from solid-state NMR (PDB ID: 2MXU [251]). Only five of the twelve chains of $A\beta_{42}$ were extracted as previously done in the literature [208]. Water molecules were added to a periodic cubic box with sides of 8 nm. The total system charge was neutralized adding Na^+ and Cl^- ions at a concentration of 150 mM. The AMBER99-ILDN force-field [257] and TIP3P model [258] were employed to define protein and water molecule topologies, respectively. A time step of 2 fs was used together with the LINCS constraints algorithm [259]. All subsequent operations were performed three times obtaining three different replicas in order to increase the statistics of the MD data. The systems were minimized using the steepest descent method and then simulated with position restraints on protein heavy atoms for 200 ps in NVT ensemble using the V-rescale coupling method [260] to maintain a temperature of 300 K. We further simulated the system with the above-described position restraints in NPT ensemble for 400 ps using the V-rescale thermostat [260] and isotropic Berendsen barostat [261] to maintain temperature (300 K) and pressure (1 bar), respectively. Finally, MD simulations were performed without any restraints for 200 ns under NPT ensample, using the V-rescale [260] and Parrinello-Rahman [262] coupling methods. The short-ranged Van der Waals (VDW) interactions were cut off after 1 nm and long-ranged electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method [263]. All simulations were carried out by GROMACS 2018 software package [264], while the Visual Molecular Dynamics (VMD) package was employed for the visual inspection of the simulated systems [265].

Molecular Docking Protocol

For each of the above-mentioned replicas, a cluster analysis was performed during the last 50 ns using linkage method [264] and a RMSD cut-off of 0.1 nm, as done previously in the literature [266]. The centroid of the most populated cluster for each replica was assumed as the starting receptor configuration (see Figure A - 6.4.1).

Structures of the 57 investigated ligands were downloaded from the PubChem database [267] and their protonation state was computed using Molecular Operating Environment software (MOE). The complete list of chosen natural compounds with their physiological charge was reported in the Supporting Information (Table A - 6.4.1). Then, ligand topologies were obtained with Antechamber [268,269] using the General Amber Force Field (GAFF) [268] and AM1-BCC charge method [270], as applied in previous studies [245,271–274].

Due to the lack of experimentally known binding sites for the $A\beta_{42}$ fibril, we opted for a blind docking approach. In this approach, the docking search is not confined to predefined binding sites but is extended to the entire protein and uses a higher exhaustiveness compared to standard docking protocols to ensure a more exhaustive exploration of the conformational space and a more accurate representation of the ligand's binding modes. More in detail, molecular docking was carried out by AutoDock Vina [275] using 64 as exhaustiveness and a search box of 9

nm x 9 nm x 12 nm, located at the centre of the protein and able to cover the entire protein surface. AutoDock Vina, widely recognized as one of the most popular molecular docking software among the scientific community, was selected for its simplicity and flexibility in defining the search space without the need for a predefined binding site. Additionally, AutoDock Vina offers the advantage of allowing for increased exhaustiveness in the search, which is especially advantageous for blind docking applications like the one conducted in this study. Each ligand was docked to the three different centroid configurations and only the best mode in terms of Vina binding affinity was selected, obtaining 57 receptor–ligand complexes.

Binding Energy Estimation and Protein-Compounds Conformational Dynamics

Each receptor–ligand complex system was followed by solvation, neutralization, energy minimization, position-restrained MD and a short MD production of 1 ns, with the same setup described in the molecular dynamics setup section. Then, the receptor–ligand binding energy was estimated by the MM–GBSA method [276] using parameters from previous literature [277–279].

The ten best ligands in terms of binding energy were further characterized by a MD simulation of 150 ns, in order to highlight the conformational changes of the amyloid fibril induced by the presence of natural compounds. Finally, the three starting receptor configurations were simulated using the above-mentioned protocol for 150 ns in order to compare the structural effects in the absence of ligands. Three replicas for each system were performed to check the reliability of the results. A simulations summary is reported in the Supporting Information (Table A - 6.4.2).

Order Parameter

The $A\beta_{42}$ fibrils were characterized by a regular shape, repeated in each chain. In order to estimate the structural order of the pentamer, an order parameter (ordP) was calculated similarly to previous works [208,210,211,280]:

$$ordP = \left\langle \frac{\sum_{c=1}^{Nc}(\sum_{n=1}^{Nr}(\overrightarrow{CoM_0 - C\alpha_0}) \cdot (\overrightarrow{CoM_t - C\alpha_t}))}{Nr\ Nc} \right\rangle_t \quad (2.4)$$

Here, the arrow represents the connecting vector of the centre of mass (CoM) position and alpha carbon (Cα) position of the n-th residue and of the c-th chain. The ordP is the dot product averaged along the observation time interval, the number of residues (NR) and the number of the chains (NC). Values of ordP close to 1 indicated an alignment close to the initial structure, i.e., aligned fibres along the fibril axis z. Values of ordP lower than 1 indicated a gradual structure distortion.

2.3.3 Results

We first performed long MD simulations of the compound-free $A\beta_{42}$ pentamer in order to obtain a representative structure for the docking protocol. The conformational stability of the three independent replicas of $A\beta_{42}$ was demonstrated by the RMSD in the Supporting Information (Figure A - 6.4.1). The MM–GBSA binding energy estimation of the 57 ligand–receptor complexes was computed during 1 ns of MD simulation, as previously done in the literature [279], and the results are reported in the Supporting Information (Table A - 6.4.1). The best ten compounds, characterized by the lowest values of binding energy, were selected (Figure 2.11) and further characterized by long MD simulations of 150 ns. Docking poses of the best compounds and their interaction maps are reported in the Appendix (Figure A - 6.4.3 and Figure A - 6.4.4). See the Materials and Methods section for further details.

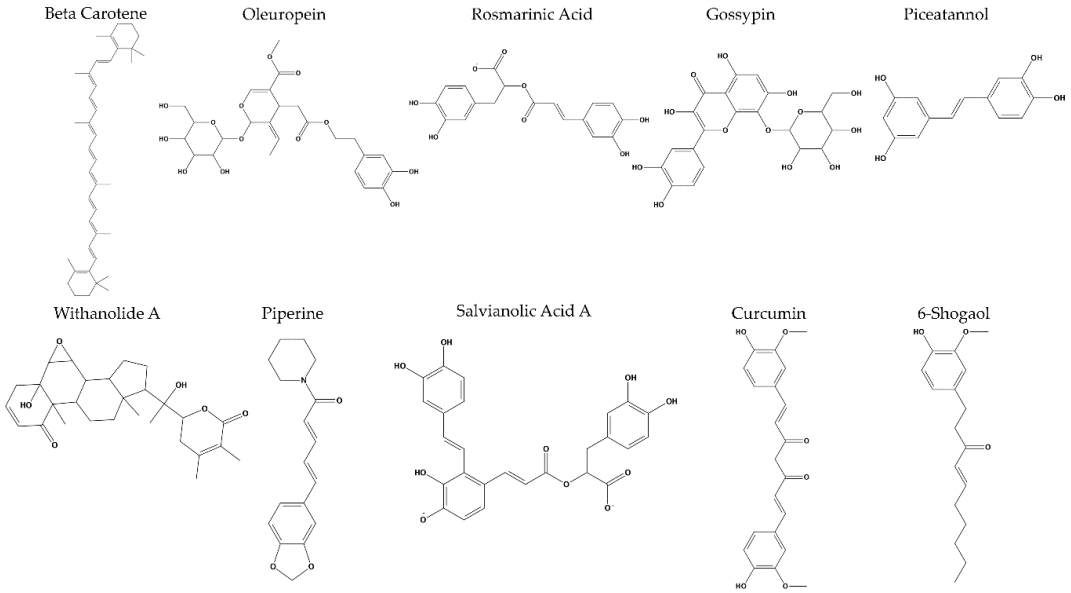


Figure 2.11. The ten best natural compounds that exhibited the lowest MM–GBSA binding energies for the selected S-shape amyloid fibril.

The MD simulations have pointed out three different mechanisms of action on the structural stability of the amyloid fibril. More in detail, (I) 6-shogaol and oleuropein were able to dock between adjacent receptor chains, inducing a considerable destabilizing effect on the whole protein; (II) curcumin, gossypin and piceatannol disrupted the ordered structure of the amyloid fibril after binding into a pocket formed by the protein S-shape; and (III) the remaining compounds, i.e., salvianolic acid A, beta-carotene, piperine, rosmarinic acid and withanolide A, did not result in remarkable protein conformational changes, thus suggesting a binding pocket stabilization. Representative snapshots of the three different mechanisms of action are reported in Figure 2.12.

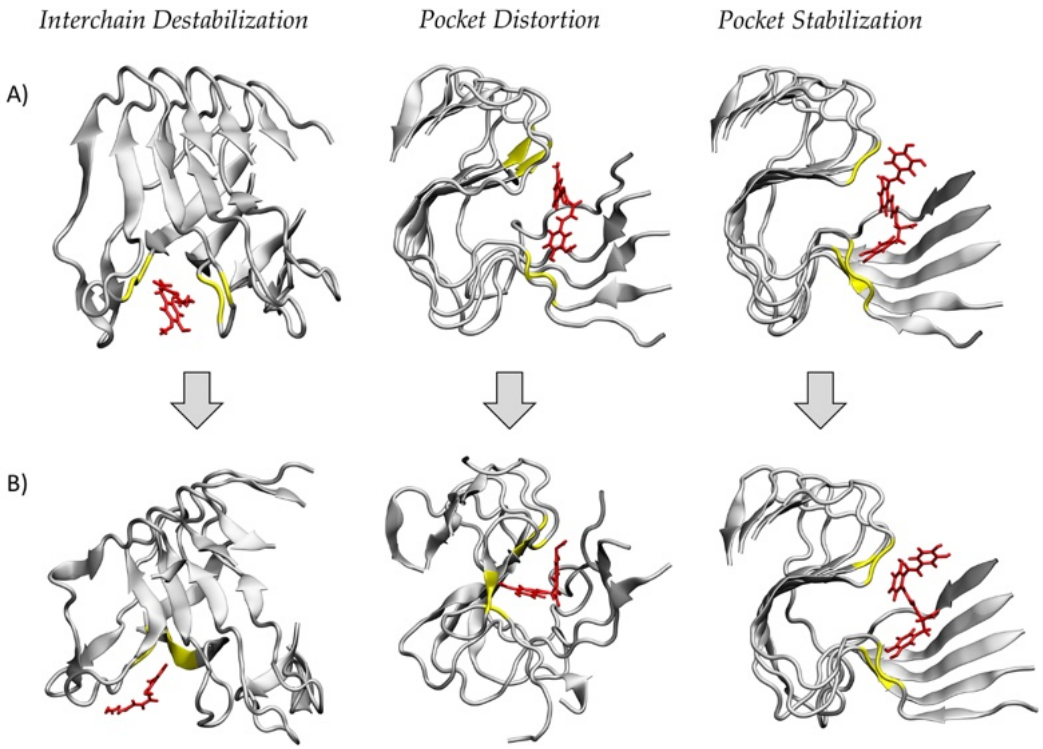


Figure 2.12. Representative snapshots of the three different mechanisms of action of selected natural compounds: interchain destabilization, pocket distortion and pocket stabilization. For each mechanism, the (A) starting configurations after the docking protocol and the (B) final structures after 150 ns of molecular dynamics (MD) simulation are shown. The ligands are represented in red, while the amyloid fibrils and their residues within 0.35 nm from the ligand are represented in grey and yellow, respectively.

In order to quantify the effects of the investigated compounds on the structural order of amyloid aggregates, three parameters were calculated on the last 25 ns of the MD simulations: (1) the beta-sheet structure probability; (2) the order parameter, calculated as described in Materials and Methods; and (3) the inter-chain interaction area, which measured the average contact surface between adjacent protein chains with a distance cut-off of 0.35 nm. All the above-mentioned analyses have been computed for the wild type protein, i.e., the ligand-free structure, and for all ligand–receptor complexes (Figure 2.13).

In Figure 2.13A, the beta-sheet probability is shown. The $A\beta_{42}$ wild type was characterized by a beta structure percentage of 37.9% ± 3.6% and each compound exhibited different effects on the protein conformational stability. Among the investigated compounds, only oleuropein, gossypin, piceatannol, curcumin and 6-shogaol proved to remarkably reduce the amyloid beta structure content. Similar trends were obtained for both the geometric order parameter and the inter-chain interaction area (Figure 2.13B,C). Therefore, oleuropein, gossypin, piceatannol, curcumin and 6-shogaol showed destabilizing effects by inducing a reduction in

terms of (i) the protein beta sheet structure content, (ii) the fibril order (quantified by the estimated order parameter) and (iii) the inter-chain interaction surface. On the above-mentioned three indicators, all the other investigated compounds demonstrated a negligible impact on the amyloid structure.

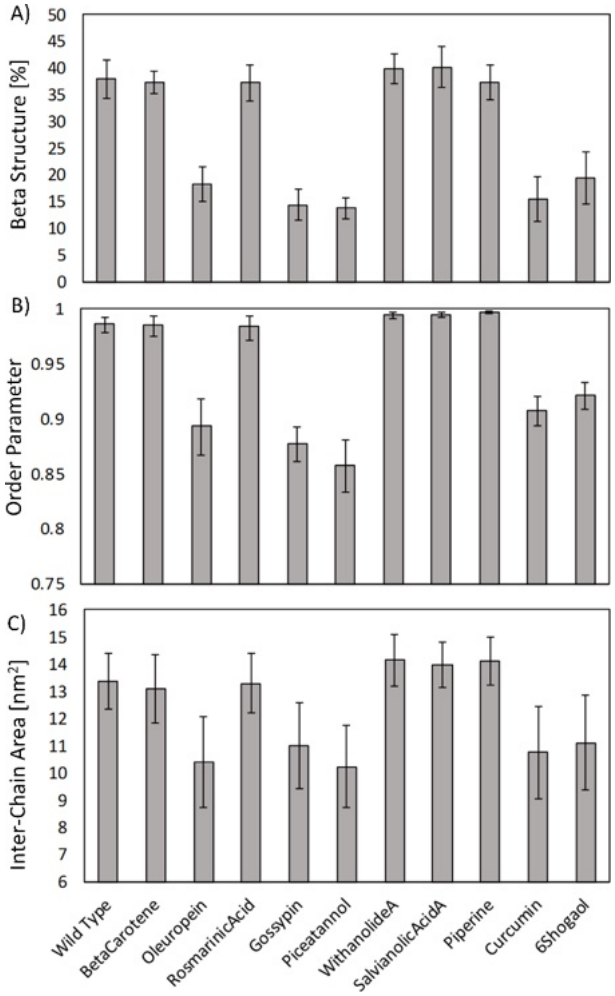


Figure 2.13. (A) Beta-sheet structure probability, (B) order parameter and (C) inter-chain interaction area for the wild-type amyloid fibrils and all the receptor-ligand complexes.

From the visual inspection of the ligand-receptor binding mechanisms, it was evident that different compounds interacted with different areas of the protein, making contact with distinct residues. Hence, to better quantify these differences, the ligand-receptor contact probability was evaluated. For each MD trajectory frame, the distance between all $A\beta_{42}$ residues and the considered compound was calculated, and the contact was counted if this distance was below a cut-off of 0.35 nm. The contact probability was then defined as the total number of contacts divided by the total number of simulation frames. The contact probability between each ligand and protein residue is reported in Figure 2.14. It should be noted that we have

not distinguished between the same residues of different protein chains, and therefore the obtained heatmap underlines the probability of interaction of a specific residue with the considered compound. This method was chosen since the original structure of the fibrillar aggregate is formed by the repetition of identical laterally bound chains. Therefore, the only noteworthy information is the residue type involved in the interaction and not the chain that it belongs to. Furthermore, it is worth noticing that 6-shogaol detached from its binding site during the simulation. Therefore, its contact probability was estimated only for the frames in which the compound was effectively docked into the binding site. In detail, the contact probability map identified two main binding areas: residues E11–F19 and residues I32–L34. Residues mainly involved in the binding process are mostly non-polar (50%) and basic (25%), suggesting that these properties are of primary importance for an effective ligand binding. Moreover, it is important to mention that 6-shogaol and oleuropein interacted less with the above-mentioned residues, since they were docked between adjacent protein chains. In particular, 6-shogaol was shown to mostly interact with the H14–G25 region, whereas oleuropein was found mainly bound to the V18–V24 and N27–I31 ones. The other compounds, instead, were buried into the binding pocket identified by the amyloid fibril S-shape and interacted with similar residues. However, rosmarinic acid expressed a slightly different behaviour, mostly interacting with the chain edge and showing a tendency to not penetrate into the binding pocket.

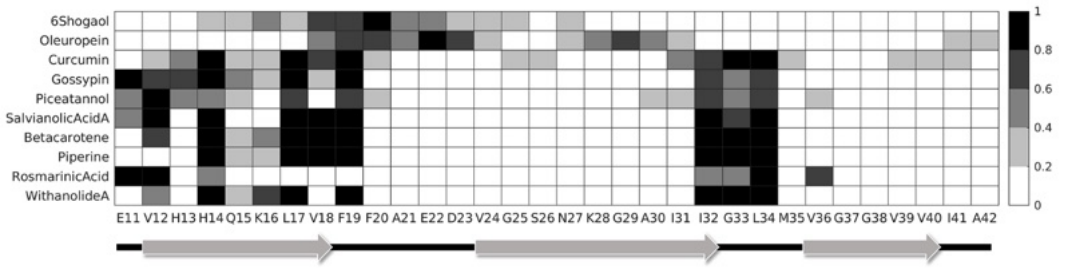


Figure 2.14. Contact probability between the selected natural compounds and the amyloid residues during the MD simulations.

In order to better characterize ligand properties, which are at the basis of their mechanisms, pharmacophore modelling was performed using LigandScout software [281]. This method allowed the definition of shared features among the properties of a series of compounds. More in detail, Figure 2.15A represents the common features between 6-shogaol and oleuropein, which were able to dock between adjacent chains (mechanism I). Otherwise, Figure 2.15B shows the common features between the ligands that led to the pocket distortion (mechanism II). All destabilizing compounds share six common features, i.e., three H bond acceptors (HBA), one H bond donor (HBD), one aromatic ring (AR) and one hydrophobic interaction (H). It is worth mentioning that the first shared features model was characterized by one more hydrophobic feature than the second one. Hence, this property can probably be related to the ability of 6-shogaol and

oleuropein to interpose between protein chains. Moreover, adding any one of the compounds that exhibited lower or no destabilizing effects (mechanism III), the shared pharmacophore model lost at least one "H bond acceptor" feature. For this reason, this characteristic seems crucial in the definition of the ligand destabilizing activity on the protein conformation.

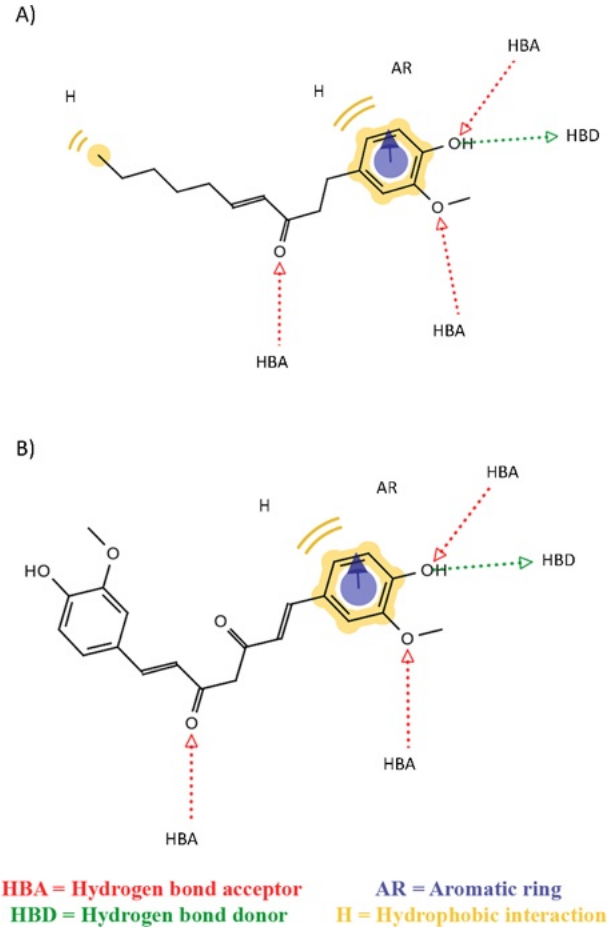


Figure 2.15. Pharmacophore model based on shared features between (A) 6-shogaol and oleuropein and (B) curcumin, gossypin and piceatannol. HBA identifies a hydrogen bond acceptor, HBD a hydrogen bond donor, AR an aromatic ring and H a hydrophobic interaction.

2.3.4 Discussion

Among several neurological disorders, AD is one of the most common forms of dementia. Even though the causes of the Alzheimer's onset and progression are still under debate, according to the amyloid cascade hypothesis [198], the amyloidogenic process that leads to the formation of extracellular aggregates of Aβ peptides is considered one of the main markers of Alzheimer's occurrence and severity. Until now, only two strategies are used to provide symptomatic relief to AD patients: acetylcholinesterase inhibitors, to maintain the level of acetylcholine in the brain,

and N-methyl-D-aspartate receptor antagonists, to prevent excitotoxicity [282]. Unfortunately, serious side effects and poor effectiveness in some phases of the disease have been detected [237,283,284]. Between different lengths of ordered and disordered amyloid peptides, the $A\beta_{42}$ fibril is known to be the most toxic due to its tendency to self-assembly into ordered structures [202]. Moreover, this structure presents two different structural rearrangements, S-shaped and the U-shaped [251,252]. In the last years, several studies have investigated the different behaviour of these polymorphisms and the S-shaped form has demonstrated greater conformational and mechanical stability than the U-shaped form [208,209,253]. Therefore, the S-shaped structure represents the primary target for pharmacological treatments, aimed to reduce the amyloidogenic process and interfere with the amyloid aggregates' stability. In this context, the search for destabilizers of Aβ fibrils may provide fruitful insights in the open research for treatments targeting AD. Natural compounds have shown promising effects, proving to be effective in many in vitro and in vivo studies with minimal side effects and increased blood brain barrier permeability [232]. However, the molecular mechanism of action of these compounds is still unclear and several computational studies have tried to characterize their effects on different amyloid aggregates [228,240,242–244,246–250]. In this work, a combination of ensemble docking and MD simulations has been applied to evaluate the influence of 57 promising compounds on preformed S-shape $A\beta_{42}$ fibrils [200,237,254–256]. We identify three different mechanisms of action for the best ten natural compounds: (I) inter-chain destabilization, (II) pocket distortion and (III) pocket stabilization. In particular, 6-shogaol and oleuropein (mechanism I) are able to disrupt the protein ordered structure docking between adjacent fibril chains; curcumin, gossypin and piceatannol (mechanism II) dock into a binding pocket identified by the amyloid S-shape, affecting the whole protein conformation; the other ligands (mechanism III), instead, preserve or slightly influence the conformational state of the amyloid fibril. In this way, we find out that only 6-shogaol, oleuropein, curcumin, gossypin and piceatannol appreciably affect the protein stability, reducing the percentual content of beta sheets, the order parameter value and the inter-chain interaction area if compared to the wild type structure. It is worth remarking that the ligands belonging to mechanisms I and II have demonstrated similar conformational effects on the amyloid fibril, inducing similar reductions of beta-sheet structure content, order parameter and inter-chain interaction area. Therefore, all the identified destabilizing compounds, i.e., 6-shogaol, curcumin, gossypin, oleuropein and piceatannol, are underlined for further investigations. Moreover, compound shared features may be used for determining a pharmacophore model to rationally design novel compounds, hopefully characterized by a more effective destabilizing strength on Aβ toxic assemblies. To remark on the importance of the selected compounds' chemical features, it is worth mentioning that brazilin, a modulator of the amyloid fibril conformation [285], shares five common features with the here characterized destabilizing ligands belonging to classes I and II (see also Figure A - 6.4.5).

The remaining compounds seem to stabilize the binding pocket, maintaining an ordered structure of the amyloid aggregate. Concerning previous literature [228,245,248,286], the present research expresses some aspects of novelty. In particular, this work considers the S-shaped polymorphism as the ligand target. Previous computational works have mostly studied a different amyloid polymorphism, namely the U-shaped one [244,245,248], which might be less stable than the S-shaped one [208,209]. Moreover, this work also provides a comprehensive comparative investigation on a considerable number of natural inhibitors, investigating their binding and action mechanisms.

Most of the selected compounds have shown antioxidant and anti-inflammatory properties in vivo [200,287–290]. Recent studies have remarked the destabilization action of curcumin on $A\beta_{40}$ and $A\beta_{42}$ [291], also with advanced amyloid accumulation [286,292]. Furthermore, it has been observed that oleuropein acts against the formation of toxic oligomer and amyloid fibrils, favouring the formation of non-toxic aggregates and improving cognitive functions [200,288,293–295]. It is worth mentioning that mechanism-III compounds have proven beneficial effects on AD onset and progression. Therefore, their mechanism of action probably alters the fibril structure in a different stage of the pathology or mostly affects other important factors of the disease, including the oxidative stress, the tau hyperphosphorylation, the α-secretase expression and the β-secretase activity.

Most ligands, except for 6-shogaol and oleuropein, interact with common residues in two main binding areas, identified by residues E11–F19 and I32–L34. Therefore, these residues seem crucial for the definition of the binding pocket and for the effective binding of the investigated compounds. Similar regions have been identified by previous studies about curcumin-like compounds in a complex with the U-shape polymorphism [248,296]. Moreover, in vitro studies have shown the key role of F19 and F20 for efficient $A\beta_{42}$ polymerization [297]. Finally, the pharmacophore modelling highlights common features between the destabilizing compounds, outlining the presence of an additional hydrophobic characteristic for the mechanism-I compounds. This feature could probably be related to the ability of 6-shogaol and oleuropein to fit between protein chains. All ligands share aromatic characteristics that have been seen to be important for interacting with amyloidogenic aggregates [298]. Moreover, the presence of three H-bond acceptor features is common to all the destabilizing compounds, but not for the other ones, suggesting that this could be an important characteristic for an effective binding to the amyloid fibril and for the activity of the investigated chemicals.

As a limit of the present research, it is worth mentioning that molecular simulations employed in this work cannot represent the entire amyloidogenic process and the present study is focused on the action of natural compounds on preformed amyloid fibrils. Only experimental studies may access the proper time and length scales to correctly describe the whole fibrillogenic process. However, the present research

represents a meaningful comparative investigation with atomic resolution, which helps the ligand screening workflow by elucidating binding and action mechanisms of a considerable number of existing natural compounds already considered in the AD research field. Future developments might consider combinations of different compounds, as well as the effect of different compound concentrations to better clarify potential cooperative or competitive mechanisms.

2.3.5 Conclusions

In this work, molecular modelling techniques were employed to screen 57 natural compounds. Five ligands, i.e., 6-shogaol, oleuropein, curcumin, gossypin and piceatannol, showed a remarkable destabilizing activity on the $A\beta_{42}$ S-shape polymorphism. Two different destabilizing modes of action of the inspected ligands were revealed. Finally, throughout pharmacophore modelling, the main common features of the highlighted ligands have been identified. These chemical features can be considered for further rational search/design of amyloid destabilizing agents. For greater detail, future studies may expand the database of the investigated compounds, including possible other interesting natural ligands characterized by shared chemical features with respect to those identified in this work. Moreover, further works may consider the effects of ligand concentration, combinations of destabilizing compounds or the presence of different species of metal ions involved in the ligand-target binding mechanism.

Chapter III

Taste Prediction Empowered by Machine Learning

Taste perception is determined at the molecular level by the interaction of specific tastants with their relative taste receptors. The chemical structure of food ingredients is the primary driver for their recognition and the subsequent activation of the molecular and supra-molecular processes that ultimately lead to taste perception. For this reason, the present chapter aims at understanding the molecular features underlying specific taste sensations: for this purpose, machine learning methods were considered and applied in the field of taste prediction. First, we provided a comprehensive overview of the main taste-related databases and ML-based models to predict the taste of molecules available from previous literature. Building upon these scientific foundations, our investigation centred on forecasting three fundamental taste sensations: umami, bitter, and sweet. This selection was predicated on the abundance of comprehensive data pertaining to these three basic tastes and their perception, which is driven by the interactions between tastants and G protein-coupled receptors (GPCRs). In contrast, the mechanisms underlying the perception of the remaining two tastes, namely sour and salty, are still subject to debate within the scientific community, as they appear to involve a broader range of factors, which hinder the simple development of ligand-based taste predictors. In section 3.2, we developed a novel ML-based tool to predict the umami taste, whereas, in section 3.3, a sweet/bitter taste predictor is presented.

To facilitate the reading of this chapter for those who may not be highly experienced in the field of machine learning, a brief introduction to the topic, specifically

focusing on the methods used in the novel works presented in this chapter, is provided in the Appendix (Section 6.1).

3.1 Machine learning for Taste Prediction

The present section is based on the following scientific publication:

*Malavolta, M., **Pallante, L.**, Mavkov, B., Stojceski, F., Grasso, G., Korfiati, A., Mavroudi, S., Kalogeras, A., Alexakos, C., Martos, V., Amoroso, D., Di Benedetto, G., Piga, D., Theofilatos, K., & Deriu, M. A. (2022). **A survey on computational taste predictors**. European Food Research and Technology, 248(9), 2215–2235. https://doi.org/10.1007/s00217-022-04044-5.*

***Author's contribution to the publication:** Pallante L. contributed to every stage of the study, from its conceptualization to the rationalisation of the data, up to the drafting and revision of the manuscript.*

Taste is a sensory modality crucial for nutrition and survival since it allows the discrimination between healthy foods and toxic substances thanks to five tastes, i.e. sweet, bitter, umami, salty and sour, associated with distinct nutritional or physiological needs. Today, taste prediction plays a key role in several fields, e.g. medical, industrial or pharmaceutical, but the complexity of the taste perception process, its multidisciplinary nature and the high number of potentially relevant players and features at the basis of the taste sensation make taste prediction a very complex task. In this context, the emerging capabilities of machine learning have provided fruitful insights in this field of research, allowing to consider and integrate a very large number of variables and identifying hidden correlations underlying the perception of a particular taste. This section aims at summarizing the latest advances in taste prediction, analysing available food-related databases and taste prediction tools developed in recent years.

3.1.1 Introduction

Taste is a crucial sense involved in the perception of food and is a sensory modality that participates in the regulation of the intake of substances, avoiding indigestible or harmful ingredients and identifying safe and healthy nutrients. Taste is determined by the gustatory system and participates in the overall perception of the flavour together with smell (olfactory system) and touch (trigeminal system) [299]. Chemicals derived from food ingestion trigger the taste perception process, starting in the oral cavity, where they bind specific proteins placed on the taste buds of the tongue [19]. The five principal tastes are bitter, sweet, sour, umami and salty, with each one being detected by specific receptors. Other tastes, such as fat taste, might be considered basic ones since they arise from the combination of somatosensory and gustation perceptions [124,125]. Each taste is linked to a vital somatic function. In general, the sweet taste is associated with the presence of energy-rich food; the bitter taste is usually linked to potentially dangerous compounds and unpleasant

flavour; umami is connected with the protein content in food; sour helps in the detection of spoiled food and acid tastants in general; lastly, salty taste monitors the intake of sodium and other minerals [1]. Moreover, taste is also supported by the sense of smell in the evaluation of foods or substances, and chemosignal detection is used by animals and humans to identify threats [300,301]. As an example, repulsive odours to humans, such as the ones generated from cadaverine, putrescine and other biogenic diamines, indicate the presence of bacterial contamination [302]. Taste sensation relies on the affinity of taste compounds for taste receptors depending on their structure. Since small variations in tastant chemistry result in drastic modifications of perceived taste, ligand-based methods, merging molecular descriptors and taste information, represent powerful data-driven tools to effectively implement machine learning (ML) algorithms with the capacity to predict taste. Such methods can be applied, for example, to screen huge databases of small compounds (e.g. ZINC15, DrugBank, ChEMBL) to select promising tastants or to rationally drive the design of novel compounds with specific functional properties and the desired taste.

Nutritious foods usually have an appetitive taste, e.g. sweet, umami, and lower concentrations of sodium and acids, whereas toxic substances generally present an unpleasant flavour such as bitter tastants, high concentrations of sodium and sour taste stimuli. Moreover, a healthy diet, such as the Mediterranean one, has been associated with beneficial impacts on human health status [303,304]. Taste prediction is therefore of paramount importance not only for the food industry but also for the medicine, pharmaceutical and biotechnology sectors. Regarding the industrial food sector, sensory evaluation is commonly applied to access the flavour of foods. Usually, it involves the measurement and evaluation of the sensory properties of foods and other materials [305,306]. The type of analysis role is crucial to address specific consumers' needs or market demands, evaluate food products, ensure high-quality products and establish the minimum shelf life of a product, food obsolescence or spoilage [307]. However, traditional methods cannot evaluate investigated food in a precise quantitative way, but only in a qualitative manner [308]. Moreover, the sensory evaluation typically requires many sensory professionals to reach a more objectiveness, with consequent problems generated by intra- and inter-operator variability, long lead times and high costs [305,308]. Thus, it is crucial to develop rational, fast and cost-effective methods to assess the food quality and its related properties, including taste. Moreover, concerning the nutritional and health field, the sweetness prediction might point out novel promising sweeteners with low caloric value to reduce the caloric intake derived from the ingestion of naturally occurring or added sugars, in line with the recommendations of the World Health Organization [309]. Indeed, the excessive consumption of added sugars is normally linked to an increase in body weight [310], obesity [311,312], and severe pathologies, such as diabetes or cardiovascular disease [313,314]. Other examples linked to the importance of taste prediction include bitter masking molecules. Indeed, the bitter

taste is one of the main problems for pharmaceutical industries due to its unpleasant taste, which represents one of the main barriers to taking medications, especially for children and the elderly population [315]. Furthermore, a change in taste perception might be caused by the onset of other pathologies, such as in the case of the loss and/or impairment of taste function after COVID-19 infection [316].

This work aims to summarise the main recent efforts in the in-silico taste prediction, starting from an overview of the major taste or food-related molecules databases and the implemented ML-based prediction tools.

3.1.2 Taste and Food-Related Databases

The first essential step for the implementation of ML-based tools is the definition of reliable and as comprehensible as possible databases (DBs) with information concerning the taste of each entry. In the past years, several databases of small compounds related to their specific taste sensations in foods have been developed. In this section, the authors pinpoint the major databases and their characteristics, which are summarised in Table 3.1.

Table 3.1. Summary of the main taste databases with weblinks, present tastes, the relative number of molecules and the possibility to download data.

Reference	Link	Taste	No.	Download
SuperSweet [17]	/	Sweet	8000	No
SweetenersDB [18]	https://bit.ly/32fG9af	Sweet	316	No
BitterDB [16]	https://bit.ly/3FinsB6	Bitter	1041	Yes
BTP640 [317]	https://bit.ly/3pogTrj	Bitter Non-Bitter	320 320	Yes
Rodgers Database [318]	/	Bitter	682	/
Umami Database	https://bit.ly/3FhePa1	Umami	800	No
UMP442 [319]	https://bit.ly/3yK6EAk	Umami Non-Umami	104 304	Yes
TastesDB [320]	/	Sweet Bitter Tasteless	435 81 133	/
Fenaroli's Handbook of Flavor Ingredients [321]	/	Sweet Bitter Tasteless	426 33 3	/

In Table 3.2, other databases, that do not contain precise information regarding the taste associated with each element but are related to food ingredients and widely used in taste prediction, are reported.

Table 3.2. Summary of the main databases related to food or commonly used by taste prediction tools, with weblinks, the number of compounds collected in each DB and the possibility to download data.

Reference	Link	No.	Download
FooDB	https://foodb.ca/	28 k	Yes
Super Natural II [322]	https://bioinf-applied.charite.de/supernatural_new/	326 k	No
FlavorDB [323]	https://cosylab.iiitd.edu.in/flavordb/	~26 k	No
PhytoHub	http://phytohub.eu/	1863	Yes

Phenol-Explorer [324]	http://phenol-explorer.eu/	501	Yes
BIOPEP-UWM [325]	https://biochemia.uwm.edu.pl/	4321	No
ChEMBL [326]	https://www.ebi.ac.uk/chembl/	~17 M	Yes
DrugBank [327]	https://go.drugbank.com/	~500 k	Yes
ZINC15 [328]	https://zinc.docking.org/	230 M	Yes
PhytoLab	https://www.phytolab.com/en/	~1300	Yes
Natural product atlas [329]	https://www.npatlas.org/	~24 k	Yes

Most used and tested databases (DBs) are described in detail in the following paragraphs.

a. SuperSweet

SuperSweet (https://bioinformatics.charite.de/sweet/) contains more than 8000 artificial and natural sweet compounds [17]. The dataset includes the number of calories, the physicochemical properties, the glycemic index, the origin, the 3D design, and other information regarding molecular receptors and targets. Sweet-tasting chemicals were taken from the literature and freely accessible data sets. The web server interface offers a very user-friendly search and a sw*eet tree* which groups the sweet substances into three main families, (carbohydrates, peptides and small molecules).

b. SweetenersDB

SweetenersDB (http://sebfiorucci.free.fr/SweetenersDB/) is a database of 316 sugars and sweeteners from 17 chemical families. Compounds were aggregated with their 2D structure or with a 3D structure using Marvin Sketch (ChemAxon) and the protonation state was defined at a pH of 6.5, according to the common pH value found in the saliva. Two natural compounds of the SweetenersDB are also present in Super Natural II (entries: 105.620, 325.102) [18].

Each sweetener has also an assigned sweetness value, indicated as logS. This value is the logarithm of the ratio between the concentrations of the considered compound and sucrose, used as a reference. In this way, this value reflects the relative sweetness of a specific compound if compared to sucrose. From a physicochemical analysis of the database, an intense sweetener has low molecular weight and a hydrophobic core. Natural sweeteners are the molecules with the highest molecular weight, and they are capable of forming more hydrogen bonds than all the sweeteners in the SweetenersDB.

c. BitterDB

BitterDB (http://bitterdb.agri.huji.ac.il/dbbitter.php) is a free source containing information about bitter taste molecules and their receptors [330]. In 2019, an upgrade of the database was made with an increase in the number of compounds (from the initial 550 to 1041) and the insertion of new features, including for example data belonging to different species rather than humans (mouse, cat, chicken).

BitterDB contains now about 1041 molecules collected from over 100 publications. For each compound, the DB provides different information, such as molecular properties, identifiers (SMILES, IUPAC name, InChIKey, CAS number and the primary sequence of proteins), cross-links, qualitative bitterness category (i.e. bittersweet, extremely bitter, slightly bitter etc.), origin (from a natural source or synthetic), and different file formats for download (SDF, image, smiles, etc.). Furthermore, toxicity data were added from the Acute Oral Toxicity Database when available, reporting experimental rat LD_{50} values as described in previous literature [331].

Most of the SMILES were taken from PubChem and the remaining ones were generated through the CycloPs server, after drawing the molecules on ChemSketch or ChemAxon. Regarding the other identifiers, the ones not available in PubChem were processed using RDKit (http://www.rdkit.org).

d. BTP640

BTP640 collects 320 experimentally confirmed bitter peptides and 320 non-bitter peptides. Bitter peptides were retrieved from various literature and peptides including ambiguous residues (e.g. B, X, Z and U) or duplicated peptide sequences were discarded. Since few experimental data concerning non-bitter peptides are available, the negative dataset was built starting from BIOPEP dataset, which contains biologically active peptide sequences (4304) widely used in the food and nutrition field (Minkiewicz et al., 2019). From this dataset, 320 peptides were randomly extracted to build the negative dataset.

e. Rodgers Database

This database was collected from previous literature and patents, including researches in BIOSIS, Food Science and Technology Abstracts, databases of internal reports at Unilever and Derwent World Patents Index (WPIDS) [318]. Structures were obtained from SciFinder, where possible, or constructed with ChemDraw. After the removal of synthetic analogues, the final database contains 649 bitter molecules. It is worth mentioning that additional 33 molecules were then considered by the authors and made public in the original paper, whereas the other 649 remain non-public. Unfortunately, no webserver or online data repository is available for this database.

f. Umami Database

The Umami Database (https://www.umamiinfo.com/umamidb/) is developed by the Umami Information Center, founded with the support of the Umami Manufacturers Association of Japan in 1982. The Umami Database was created with the idea of providing information about the umami taste in foods and, currently, about 800 items are listed in the database. Amino acids in foods are mainly of two types, i.e. ones joined together to form proteins and free amino acids, that have a more pronounced flavour. In this context, free glutamate has a remarkable umami taste. Umami Database reports also the score of free glutamate and other free amino acids which affect food taste. In addition, there are inosinate and guanylate scores, which synergistically increase umami perception. Sources for the Umami Database include public academic papers and scores analysed by a research laboratory upon request of the Umami Information Center.

g. UMP442

This dataset was constructed for the development of the iUmami prediction tool [319]. The umami set merges several experimentally validated umami peptides from the literature [61,332–336] and the BIOPEP-UWM database [325]. On the other hand, the non-umami peptides dataset is made by the bitter peptides from the positive set of BTP640 [337]. After removing peptides with non-standard letters and redundant sequences, the final UMP442 database collects 140 umami and 304 non-umami peptides. The dataset was made publicly available on GitHub (https://github.com/Shoombuatong/Dataset-Code/tree/master/iUmami).

h. TastesDB

TastesDB is an experimental database comprising 727 chemicals, with their respective experimental taste class, retrieved from several scientific publications [320]. Since all incorrect molecules or those with problematic molecular structures were removed, the final TastesDB contains 649 molecules, specifically 435 sweet, 81 bitter and 133 tasteless. For each entry of the database, the DB provides the commercial name, the SMILE, the tasting class (sweet, tasteless, bitter) and the literature reference.

3.1.3 Existent Machine Learning-based Predictors

In the past years, several studies have developed ML-based algorithms to predict the taste of specific molecules starting from their chemical structure. In this section, we will review in detail the main recent literature in the field of taste prediction. Where no precise name has been defined for the tools discussed, we have decided to use the first author's name of the reference publication for simplicity (Table 3.3).

Table 3.3. Summary of the main recent taste prediction tools, including methods, datasets and molecular descriptors employed (see also Table A - 6.5.1 for further information).

Reference	Method	Taste	No.	Descriptors

Chéron Sweet Regressor [18]	Sweet Regressor (RF, SVR)	**Sweet**	316	Dragon
Rojas Sweet Predictor [320]	Sweet Classifier (QSTR)	**Sweet** **Non-Sweet**	435 214	ECFP, Dragon
Goel Sweet Regressor [338]	Sweet Regressor (GFA, ANN)	**Sweet**	487	Material Studio
e-Sweet [339] [https://bit.ly/3wFy4ER]	Sweet Classifier (KNN, SVM, GBM, RF, DNN)	**Sweet** **Non-Sweet**	530 850	ECFP
Predisweet [340] [https://bit.ly/3reop7a]	Sweet Regressor (AB)	**Sweet**	316	Dragon, RDKit, Mordred, ChemoPy
BitterX [341] [https://bit.ly/3wJYa9O]	Bitter Classifier (SVM)	**Bitter** **Non-Bitter**	539 539	[342]
BitterPredict [343] [https://bit.ly/3igrzmQ]	Bitter Classifier (AB)	**Bitter** **Non-Bitter**	691 1952	Canvas (Schrödinger)
e-Bitter [344] [https://bit.ly/3epWzQq]	Bitter Classifier (KNN, SVM, RF, GBM, DNN)	**Bitter** **Non-Bitter**	707 592	ECFP
iBitter-SCM [317] [https://bit.ly/2VGyXAg]	Bitter Peptides Classifier (SCM)	**Bitter** **Non-Bitter**	320 320	Dipeptide composition (DPC)
BERT4Bitter [345] [https://bit.ly/2WecTxf]	Bitter Peptides Classifier (BERT)	**Bitter** **Non-Bitter**	320 320	Dipeptide composition (DPC)
iBitter-Fuse [346] [https://bit.ly/3BmC547]	Bitter Peptides Classifier (SVM)	**Bitter** **Non-Bitter**	320 320	DPC, AAC, PAAC, APAAC, AAI
BitterIntense [347]	Bitter Intensity Classifier (XGBoost)	**VB** **NVB**	246 404	Canvas (Schrödinger)
iUmami-SCM [319] [https://bit.ly/3hJs9uf]	Umami Classifier (SCM)	**Umami** **Non-Umami**	140 304	Dipeptide composition (DPC)
BitterSweetForest [348]	Bitter/Sweet Classifier (RF)	**Sweet** **Bitter**	517 685	RDKit (Binary fingerprints)

BitterSweet [349] [https://bit.ly/3rd7Att]	Bitter/Sweet Classifier (AB, RF)	**Bitter** **Non-Bitter** **Sweet** **Non-Sweet** **Non-Umami**	918 1510 1205 1171 304	Canvas, Dragon, ECFP, ChemoPy
VirtualTaste [350] [https://bit.ly/2UfVFPi]	Multi-taste classifier (RF)	**Sweet** **Bitter** **Sour**	2011 1612 1347	*not reported*

In the following, each of the aforementioned tools is examined in detail, dividing the discussion into a brief introduction, the "*Data preparation and model construction*" section and the "*Model performance*" section.

Sweet Prediction

a. Chéron Sweet Regressor

In this work, a Sweet Predictor was created using a new QSAR model [18]. This model, also applied to external datasets (SuperSweet and SuperNatural II), allowed to point out the main physio-chemical features of sweeteners related to their potency.

Data preparation and model construction

The curated dataset of sweet compounds resulted in the creation of the SweetenersDB, which is constituted of 316 compounds with known sweetness values relative to sucrose (see *Taste and Food-Related Databases* chapter for further details). The compounds' SMILES were firstly collected in a 2D database, and subsequently, 3D representations were created using Marvin, ChemAxon (https://www.chemaxon.com), choosing the three with the lowest energy. The protonation state was set at the physiological salivary pH value (6.5).

Dragon descriptors (http://www.talete.mi.it/products/dragon_description.htm) were calculated for both the 2D and 3D databases. All features with a correlation greater than 0.9 were removed, obtaining 244 descriptors for the 2D molecules and 265 descriptors for the 3D structures. Finally, all descriptors were normalised.

The dataset was randomly divided with a 70:30 ratio and the leave-one-out method was used for the cross-validation. Support Vector Regression (SVR) and Random Forest (RF) were optimised on the training set and the test set was used for the model performance evaluation.

Model performance

Performance evaluation was obtained using the squared of the correlation coefficient (R^2). Notably, the SVR reached a slightly better performance than RF

on the test set. It is worth mentioning that the models on 2D and 3D datasets reached similar performances, suggesting the 2D approach as the best option for fast screening since it is much less time-consuming. More in detail, the RF 2D, SVR 2D, RF 3D and SVR 3D models obtained correlation coefficients on test sets of 0.74, 0.83, 0.76, 0.85, respectively.

To evaluate the model applicability domain, SweetenersDB was compared with SuperSweet and SuperNatural II. Interestingly, 99.5% of the molecules from SuperSweet are similar to structures in SweetenersDB, whereas only about 34% of the SuperNatural II database belong to the chemical space defined by SweetenersDB. This analysis confirmed the importance of associating an applicability domain to a prediction model to measure the reliability of the prediction.

b. Rojas Sweet Predictor

The present Quantitative Structure-Taste Relationship (QSTR) model is a specialist framework created to foresee the pleasantness of synthetic compounds [320]. It can likewise be utilized to gain a comprehensive understanding between atomic design and pleasantness and defining novel sugars. This sweetness prediction model is the first QSTR model that considers both molecular descriptors and extended connectivity footprints, performing a structure similarity analysis in combination with the model prediction.

Data preparation and model construction

The starting dataset is TastesDB (see also *Taste and Food-Related Databases* chapter for further details). The dataset includes 649 molecules: 435 sweet, 81 bitter and 133 tasteless; the latter two classes were combined into a non-sweet class. Extended-connectivity fingerprints (ECFPs) [351] and classical molecular descriptors, i.e. Dragon 7 (3763 total descriptors) (https://chm.kode-solutions.net/pf/dragon-7-0/), were used to describe the molecules of the dataset. In all cases, the 2D representation was preferred to the 3D one, to get a conformation-independent molecular representation.

Exploration of the data and similarity analysis was performed using the Multidimensional Scaling (MDS), whereas the Partial Least Squares Discriminant Analysis (PLSDA) and N-Nearest Neighbors (N3) were employed as classifiers. Finally, the V-WSP unsupervised variable reduction method and the Genetic Algorithms-Variable Subset Selection (GA-VSS) technique were used as dimensionality reduction techniques to retrieve the most informative molecular descriptors.

The dataset was divided into three parts, maintaining the proportion of the classes: the training set consisting of 488 compounds (161 non-sweet and 327 sweet molecules), the test set consisting of 161 molecules (53 non-sweet and 108 sweet

molecules), and finally, the last part was used as an external dataset. Moreover, a 5-fold CV was employed for the GA-VSS and the Monte Carlo (leave-many-out) random sub-sampling validation of the system. These methods iteratively and randomly divided the molecules into training (80%) and evaluation (20%) sets.

Model performance

Specificity (SP), sensitivity (SN) and non-error rate (NER), which is more efficient in the case of unbalanced datasets, were used as performance metrics. The two final models, made with six molecular descriptors, were chosen based on the NER classification parameter. Since PLSDA and N3 are based on distinct methods and descriptors, a consensus analysis was employed to improve prediction [352]. Therefore, a molecule was classified if both models showed the same result and not classified otherwise.

Performance in calibration (SE = 79.2%, NER = 85.2%, SP = 91.3%, not assigned = 33%), in cross-validation (SE = 77.2%, NER = 83.1%, SP = 89.0%, not assigned = 32%), in the Monte Carlo validation (NER = 88.7%, SE = 92.7, SP = 84.8%, non-assigned = 20.5%) and in the 161 test molecules (NER = 84.8%, SE = 88.0%, SP = 81.6%, non-assigned = 19.3%) confirm the model stability. The consensus analysis improved the overall performance of the model. Notably, the model calibration was performed only on a cluster of the complete dataset, that was derived from the MDS analysis. The remaining part of the original dataset was classified using similarities scores combined with the aforementioned models.

c. Goel Sweet Regressor

The present Sweet Regressor tool is a QSAR model able to estimate the relative sweetness level of a test compound with respect to the sweetness of sucrose [338]. This tool can act as a pre-processing step in the design of new sweeteners by pointing out their crucial structural requirements.

Data preparation and model construction

The dataset was collected from several publications [353–357], resulting in 487 unique molecules with relative sweetness compared to sucrose (ranging from −0.699 to 5.334) calculated as described for the SweetenersDB (see also the *Taste and Food-Related Databases* chapter). Compounds SMILE were converted to 3D structures and Material Studio v6.0 was used to calculate 564 molecular descriptors. After performing a correlation analysis, only 61 descriptors were maintained and, after removing outliers, the remaining 455 molecules were randomly divided between the training and test sets (70:30 ratio).

Two QSAR models were developed using Artificial Neural Network (ANN) and Genetic Function Approximation (GFA) algorithms.

Model performance

Performance was assessed using the correlation coefficient for the training set, leave-one-out method and test set ($R_{training}^2$, R_{cv}^2 and R_{test}^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Square Error (MSE). Statistical parameters of ANN ($R_{training}^2 = 0.889$, $R_{test}^2 = 0.831$) and GFA with linear spline ($R_{training}^2 = 0.864$, $R_{test}^2 = 0.832$) were comparable and both QSAR models demonstrated a reasonable prediction accuracy. GFA allows the development of numerous models, autonomously selecting features and broadening the number of terms used in model construction and easily interpreting the data. On the other hand, ANN can further explain hidden relationships between complicated data and depict patterns and trends, but it lacks interpretability.

d. e-Sweet

e-Sweet is a free tool to predict the sweet taste of analysed chemicals and their relative pleasantness (RS) [339].

Data preparation and model construction

The entire dataset includes 1380 compounds, divided into sweet and non-sweet. Sweet compounds are 530 sugars curated from SuperSweet, SweetenersDB and previous literature [320,348], while 850 non-sweet comprise 718 entries from BitterDB and 132 recovered from the literature [320].

The 80:20 data splitting scheme was adopted, resulting in 883 compounds for training and internal 5-fold CV and 221 compounds for the test. Features were built using Extended-connectivity Fingerprints (ECFP) [351] and subsequently selected by their importance in a trained RF model. Implemented algorithms comprised KNN, SVM, GBM, RF and DNN with different splitting procedures (19 different splits for the former four models and 3 different splits for DNN) to reduce the bias yielded by specific splits. A total of 1312 models were first assessed individually and subsequently combined to form a pool of 4 Consensus Models (CM), that leverage the combination of individual models to improve overall classification performance.

Model performance

Model performance was assessed based on widely used metrics (F1-score, specificity, sensitivity, accuracy, precision, Matthews Correlation coefficient (MCC), Non-Error Rate (NER)). F1-score metric was chosen as the final algorithm selection criterion and, subsequently, a Y-randomization test was performed for a direct assessment of model robustness.

The split-averaged performance of the best CM on the test set reached 91% of accuracy, 90% precision, 94% specificity, 86% sensitivity, F1-score of 88%, MCC of 81%, and 90% NER, all with 95% confidence intervals of ± 1%.

e. Predisweet

Predisweet is a free-available web server (http://chemosimserver.unice.fr/predisweet/) capable of predicting sweet taste and the relative sweetness (in logarithmic scale) of compounds [340]. The applicability, reliability, and decidability domains have been used to estimate the quality of each prediction.

Data preparation and model construction

The tool is based on the SweetenersDB, collecting 316 compounds with their relative sweetness (see also *Taste and Food-Related Databases* chapter) [18]. Two other databases, namely Super-Natural II database [322] and the *phyproof* catalogue from PhytoLab (https://www.phytolab.com/en/), were considered as an external dataset (4796 natural compounds).

Each compound was collected as SMILE and sanitized using RDKit (https://www.rdkit.org/). The protonation state was predicted using ChemAxon (http://www.chemaxon.com/) at the physiological pH of saliva (pH=6.5). Molecules were standardized using the *flatkinson* standardiser (https://github.com/flatkinson/standardiser) and further processed with a Python package, removing salts and applying specific rules to normalize the structures. Molecular descriptors were calculated using Dragon v6.0.38, RDKit, Mordred [358] and ChemoPy [359] packages. Descriptors from the latter three methods (506) were defined as "open source" descriptors, as opposed to Dragon ones (635).

The Sphere Exclusion clustering algorithm divided the SweetenersDB into the training set (252 compounds) and the test set (64 compounds).

Several regression algorithms were tested, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF) and Adaptative Boosting with a Decision Tree base estimator (AB), and the 5-fold CV was employed to avoid asymmetric sampling and overfitting.

Model Performance

Predictive performance is assessed through Golbraikh and Tropsha criteria [360] and the AdaBoost Tree was considered the best method for both models. The obtained models reach R^2 higher than 0.6 (0.74 for the Open-source and 0.75 for the Dragon model) and Q^2 higher than 0.5 (0.84 for the Open-source and 0.79 for the Dragon model) for both models. Notably, since less information was available regarding potent sweeteners, developed models perform worst for high sweetness values. The open-source and the Dragon models reached similar performances showing good prediction on the test set. Therefore, the open-source version was used for the webserver (http://chemosimserver.unice.fr/predisweet/) implementation of the algorithm.

The quality of the prediction for each query is evaluated based on three metrics: (i) the *applicability domain*, which measures if the investigated compound is in the

range of descriptors of the training set, (ii) the *reliability domain*, which considers the density of information around the compound, and (iii) the *decidability domain*, which is the confidence of the prediction. The resulting quality of the prediction is also reported for the user when using the webserver platform.

Bitter Prediction

a. BitterX

BitterX is a web-based platform (http://mdl.shsmu.edu.cn/BitterX/) available for free [341]. This tool implements two different models, i.e. the *bitterant verification model*, which allows the identification of a bitter compound, and the *TAS2R recognition model*, which predicts the possible human bitter taste receptors, among the 25 known TAS2Rs. Such predictions were validated experimentally.

Data preparation and model construction

The interaction between the TAS2Rs and bitter compounds were curated from PubMed and BitterDB. A total of 539 bitter compounds were obtained to constitute the positive set, and their molecular structures were achieved from Pubchem. The negative set included 20 true non-bitterants (in-house experimental validation) and 519 molecules from the Available Chemicals Directory (ACD, http://www.accelrys.com). The final dataset contained 1078 compounds, equally divided into positive and negative bitter molecules. Molecular structures were obtained from PubChem and processed with in-hose program Checker and ChemAxon's Standardizer (http://www.chemaxon.com). On the other hand, the initial dataset for the TAS2R prediction model was taken from the literature and includes 2379 negative and 260 positive bitterant-TAS2R interactions. Due to the huge difference between the two dataset sizes, all the 260 non-redundant experimentally verified bitterant-TAS2R interactions were considered, while just 260 bitterant-TAS2R couples were selected as negatives to balance the dataset.

Physiochemical descriptors for compounds were chosen based on the Handbook of Molecular Descriptors [342], resulting in 46 and 20 descriptors for the bitterant verification and TAS2R recognition models, respectively. Moreover, 15 descriptors were used for the TAS2R representation in the TAS2R recognition model. The descriptors were chosen using a Feature Selection (FS) method based on a Genetic Algorithm (GA).

BitterX employs the support vector machine (SVM) classifier: the training was divided into two categories (+1 and -1) that represent the classification between bitter and non-bitter (and the bitterant-TAS2R interaction or not). The SVM + sigmoid method was implemented to value the probability that a molecule is bitter in the bitterant authentication and the probability of a bitterant binding to TAS2Rs in the TAS2R recognition model.

Data for algorithm training and test were used by adopting an 80:20 splitting strategy. For the bitterant verification model, this results in 862 and 216 compounds for training and test, respectively. To avoid any error derived from a particular data splitting, the other two partitions were made randomly from the general database, always following the equipartition between bitter and non-bitter molecules. Similarly, the TAS2R recognition dataset was also divided with a 4:1 ratio. Lastly, a 5-fold CV ensured the robustness of the classification algorithm.

Model performance

To evaluate and compare the model, four indices were calculated: precision, accuracy, sensitivity and specificity. Furthermore, the trade-offs between SE and SP were assessed by performing the ROC curve and calculating the AUC value.

Considering both the training (5-CV) and test sets, the *bitterant verification model* reached specificity, precision and sensitivity values above 90%, AUC above 94% and accuracy above 87%, whereas the *TAS2R recognition model* reached above 76% for the accuracy, above 75% for precision and specificity, above 78% for sensitivity and above 81% for AUC.

b. BitterPredict

BitterPredict is a bitter prediction tool published in 2017 [343]. This work is developed in the commercial MATLAB environment and the code is available on GitHub (https://github.com/Niv-Lab/BitterPredict1). The users should provide an Excel or CSV file with calculated properties by the commercial Schrödinger software and QikProp package.

Among random molecules screened by the tool, a high percentage is represented by bitter compounds. These include many synthetic molecules (66% of drugs are bitter) and natural compounds (up to 77%). It is worth mentioning the relatively high percentage of bitter food (38%), considering the natural aversion of humans towards the bitter taste.

Despite its great functionality, high accuracy (~80%) and the ability to screen a general chemical space, this tool presents some limitations, including (i) the prediction of only molecules in defined chemical space, named Bitter Domain, (ii) the inability to discriminate between weakly and strong bitter compounds, (iii) an unbalanced dataset (positive set three times smaller than the negative set).

Data preparation and model construction

The dataset was processed using Maestro, Epik and LigPrep (Schrödinger), removing uninterested structures and assigning the correct protonation state according to pH 7.0 ± 0.5. Then, non-neutralised molecules and molecules with identical descriptors were removed from the dataset to allow the calculation of the QikProp descriptors and obtain a non-redundant dataset. The whole prediction was made within a restricted chemical space called *Bitter Domain* to identify a region

in which 97% of the bitter molecules is included. This domain is defined by a hydrophobicity (AlogP) range of $-3 \leq \text{AlogP} \leq 7$ and a molecular weight $\text{MW} \leq 700$. This preparation procedure was applied to each database considered to build the final dataset.

Bitter Set (*positive set*) includes all molecules considered as bitter (691): 632 structures from BitterDB and 59 molecules from literature [361]. On the other hand, the Non-Bitter Flavors set (*negative set*) consisting of 1917 non-bitter molecules: "probably not bitter" compounds gathered from Fenaroli's handbook of Flavor ingredients (1451), sweet (336) and tasteless (130) molecules from Rojas et al. The *validation set* consisted of Bitter New, i.e. 23 molecules stored recently from several publications to increase BitterDB, UNIMI set, i.e. 56 synthesized molecules, and the Phytochemical Dictionary, consisting of 26 non-bitter and 49 bitter compounds inside the Bitter Domain. Moreover, a set of data was collected for the *sensory evaluation*. 1047 molecules were retrieved from the Sigma-Aldrich flavours and fragrances catalogue (https://www.sigmaaldrich.com/IT/it/applications/food-and-beverage-testing-and-manufacturing/flavor-and-fragrance-formulation). After data curation, 264 entries were selected as the Bitter Domain. Finally, a data set for *prospective prediction* was collected merging DrugBank approved drugs, FooDB, Natural Products Dataset from ZINC15 and ChEBI. Compounds in the Bitter Domain from these widely used DBs were 1375, 13588, 27474 and 27015, respectively.

Molecular descriptors (59) were calculated with Canvas (Schrödinger) and QikProp (Schrödinger).

The final input dataset was divided into 30% (test set) and 70% (training set) randomly, following the hold-out method and preserving the original proportions. To avoid overfitting, models were optimised by evaluating their performance only in the hold-out test.

The algorithm implemented by BitterPredict is AdaBoost. The ensemble method models used are Fitensemble and TreeBagger, which combines the outcomes of several decision trees, decreasing the impacts of overfitting and enhancing the generalization ability of the model.

Model performance

The two parameters applied to evaluate the classification models were sensitivity (SE) and specificity (SP). For the training set, the SE was 91% and SP 94%, for the test set SE was 77% and SP 86%. Furthermore, a model evaluation based only on the non-bitter datasets was made among sweet, tasteless and non-bitter flavours. In this context, the dataset that shows a better specificity was the non-bitter flavours, with a value of 86%.

The BitterPredict study also analyzed the impact of diverse descriptors estimating their contribution in reducing the error. Notably, the most important descriptor was the total charge and most of the bitter molecules were positively charged presenting an ammonium ion at physiological pH. Moreover, QikProp descriptors linked to the compound toxicity seems to have a greater impact on the model if compared to general properties descriptors.

The validation of BitterPredict was performed in three phases, as follows.

(i) *Validation using external sets* (see *Data preparation and model construction* for further details) tested the algorithm on datasets never seen before by the algorithm to avoid overfitting. Excellent performance was achieved, with a specificity of 69%-85% and sensitivity of 74%-98%.

(ii) *Validation by literature mining* consisted of a selection of 60 compounds from a DrugBank set of FDA approved drugs, half of which with the best and half of which with the worst score of bitter prediction according to BitterPredict. The results from literature research indicated that almost 60% of the top 30 bitter molecules were declared to have a bitter taste, while only 20% of the 30 non-bitter molecules had a probable indication of a bitter taste.

(iii) For the *validation by taste tests* (*sensory evaluation*), 12 participants were selected to evaluate the taste of 6 compounds predicted as non-bitter by BitterPredict among the 264 compounds taken from fragrances catalogue and Sigma-Aldrich flavours (see also *Data preparation and model construction* section). None of the six compounds differed in bitterness from the control (water) with the Dunnett test (alpha = 0.05), whereas the Quinine (established bitter molecule) demonstrate a considerably higher bitterness compared to water.

The three validation protocols indicated that BitterPredict allows obtaining reliable and satisfactory performance both for the bitter and non-bitter prediction.

Finally, the BitterPredict classifier was applied to DrugBank approved, FooDB, Natural Products Dataset from ZINC15 and ChEBI datasets (see also see *Data preparation and model construction*). The results highlight that the percentages of bitter molecules are respectively 65.94%, 38.36%, 77.21% and 43.71%.

c. e-Bitter

Developed by the same research group that created e-Sweet [339], e-Bitter is a free graphic program published in 2018 for bitter prediction, which natively implements the ECFP fingerprint and the analysis of the structural features [344]. Differently from other works [341,343], e-Bitter considers only experimentally confirmed non-bitterants, i.e. 592 compounds comprising tasteless, sweet and non-bitter chemicals. e-Bitter code is publicly available

(https://www.dropbox.com/sh/3sebvza3qzmazda/AADgpCRXJtHAJzS8DK_P-q0ka?dl=0).

Data preparation and model construction

The dataset contains experimentally confirmed 707 bitter compounds, derived mostly from BitterDB [16] and literature research [318,320], and 592 non-bitter compounds (132 tasteless, 17 non-bitter and 443 sweet). Sweet compounds were obtained from SweetnersDB, SuperSweet and previous literature [361,362]. The same compounds but with a different taste or from different datasets, or compounds such as salts or ions, were excluded, while all structures containing common elements were retained. e-Bitter uses Extended-Connectivity Fingerprints (ECFPs) as molecular descriptors [351]. Similarly to e-Sweet [339], the implemented algorithms were KNN, SVM, GBM, RF and DNN. Models were tested both with and without feature selection [363].

The splitting of the dataset follows the same criterion as previously described in e-Sweet: 1030 compounds (556 bitter and 474 non-bitter), i.e. 80% of the initial dataset, were employed as training data and for internal validation while the remaining ones, consisting of 259 compounds (141 bitter and 118 non-bitter) were used for performance testing.

Model performance

The metrics employed to assess the model performance include precision, Matthews correlation coefficient (MCC), sensitivity, accuracy, specificity, F1-score and ΔF1 – score (difference between the F1-score in cross-validation and the test set). Moreover, the reliability of the developed models was accessed using the Y-randomisation test, as for the e-Sweet model. Finally, an applicability domain based on the Tanimoto similarity was implemented to avoid non-reliable predictions on compounds highly diverse from compounds in the dataset.

Starting from an initial set of 1312 models, 96 averaged models (over adopted data partitioning schemes), and 9 consensus models were tested. Best performance was obtained by top average models trained with DNN3 (ACC = 92.0%, SP = 80.8%, SE = 98%, MCC = 82.3%, F1-score = 94.1%).

d. iBitter-SCM

The iBitter-SCM tool is the first computational model that provides a prediction of the peptides' bitter taste starting with their AA sequence independently from structural and functional information [317]. iBitter-SCM is freely available as a web server (http://camt.pythonanywhere.com/iBitter-SCM) and all codes and datasets are also on GitHub (https://github.com/Shoombuatong2527/Benchmark-datasets).

Data preparation and model construction

The dataset BTP640 (see also *Taste and Food-Related Databases* chapter for further information) includes 640 molecules equally divided between bitter and non-bitter (training set 80% and test set 20%).

Features representation was realized using the dipeptide composition (DPC). iBitter-SCM is based on the scoring card method (SCM), which enabled robust protein and peptide function prediction and analysis without any information regarding their structure and relying instead on so-called propensity scores of individual peptides and amino acids. More in detail, after preparing a training dataset and an independent dataset, the workflow started by determining the initial propensity scores (init-DPS) of dipeptides using statistics and subsequently applying Genetic Algorithms (GAs) to refine and optimize the score to the so-called optimized dipeptide propensity score (opti-DPS). Finally, the individual amino acid propensity score was again extracted by statistical methods enabling the final discrimination between bitter and non-bitter peptides employing a weighted sum with opti-DPS. In a nutshell, these scores represent the link between peptide composition and function, by directly quantifying the contribution of individual amino acids on the physical-chemical characteristics. Furthermore, informative physicochemical properties (PCPs) of individual amino acids, i.e. their direct involvement in fundamental biological reactions and pathways, were taken from the amino acid index database (AAindex) [364].

Model performance

The model was assessed with several performance metrics: accuracy (ACC), specificity (SP), sensitivity (SE) and Matthew coefficient correlation (MCC) and AUC.

The performance of the opti-DPS and the init-DPS were compared using the 10-fold cross-validation and the independent test. The best model (opti-DPS) was chosen based on the best performance on the 10-fold CV and independent test sets (ACC = 84.38%, SE = 84.38, SP = 84.38, MCC = 68.8%, AUC = 90.4%). Notably, the best opti-DPS outperforms init-DPS with enhancements on ACC, SN, SP and MCC and iBitter-SCM, compared with other traditional ML models (KNN, NB, DT, SVM, RF), demonstrated better performance and greater robustness.

e. BERT4Bitter

After creating iBitter-SCM, the same research group developed BERT4Bitter, a similar tool for the classification of bitter peptides [345]. BERT4Bitter dataset is publicly available and the developed model is freely accessible through a user-friendly web server interfacehttp://pmlab.pythonanywhere.com/BERT4Bitter.

Data preparation and model construction

The dataset used to develop the BERT4Bitter model is the same used for the iBitter-SCM method, i.e. the BTP640 [317] (see also *Taste and food-related databases*

chapter for further details). Using the same 80:20 splitting ratio, the BTP640 dataset was randomly divided for training and testing.

The peptide sequence featurization was achieved through the natural language processing (NLP) techniques, specifically using Pep2Vec [365] and FastText [366]. Each of the 20 amino acids is considered as a word and each peptide sequence was translated into a sentence (an n-dimensional word vector). In the same framework, the importance of each amino acid in the analyzed sequences was evaluated with the TFIDF method [367].

Three different deep-learning-based models, i.e. convolutional neural network (CNN), long short-term memory (LSTM) neural network and BERT-based model, were implemented using different numbers of layers (6, 5, 12, respectively) and rationally compared.

Model performance

Model evaluation was accessed both in the 10-fold CV and independent test-set. According to the cross-validation performance, the BERT model outperformed the CNN and LSTM ones in all the evaluation metrics (ACC = 0.86, AUC = 0.92, SP = 0.85, SE = 0.868, MCC = 0.72).

Notably, considering the independent test-set performance, BERT4Bitter outperformed the iBitter-SCM tool with ACC and MCC of 0.92 and 0.84, respectively, demonstrating a stronger predictive ability in discriminating bitter and non-bitter peptides.

f. iBitter-Fuse

After the development of iBitter-SCM and BERT4Bitter, the same research group implemented an improved bitter/non-bitter peptides predictor, called iBitter-Fuse [346]. This model overcomes some of the main drawbacks of the previous ones, including the generalization capability linked to the feature representation, over-fitting, redundancy and the overall performance. Exploiting several feature encoding schemes and customized algorithms to identify the most informative features, iBitter-Fuse outperformed both iBitter-SCM and BERT4Bitter, establishing itself, at the moment, as the best tool for the prediction of bitter peptides.

Data preparation and model construction

BTP640 was again used as starting dataset as done for iBitter-SCM and BERT4Bitter and the same 80:20 partition scheme was applied to effectively compare their performance.

Five feature encoding methods, including dipeptide composition (DPC), pseudo amino acid composition (PAAC), amino acid composition (AAC), physicochemical properties from AAindex (AAI) and amphiphilic pseudo amino acid composition

(APAAC), and a merged feature (DPC + PAAC + AAC + AAI + APAAC), were calculated to consider both composition and physicochemical properties. The model is based on an SVM algorithm and the feature selection was performed using a customised GA algorithm using self-assessment-report (GA-SAR) [368].

Model performance

The fused feature allows obtaining the best performance (ACC, MCC, AUC) on the cross-validation, outperforming the other five feature encoding methods. To reduce the number of fused features (994), the GA-SAR was applied and 36 features were consequently maintained.

Performance evaluation demonstrated that iBitter-Fuse outperformed the previous tools for predicting the bitterness of peptides, i.e. iBitter-SCM and BERT4Bitter, suggesting that it is a more reliable and accurate tool. More in detail, the present SVM-based model reached ACC, SE, SP, MCC and AUC of 93.0%, 93.8%, 92.2%, 85.9% and 93.3%, respectively.

g. BitterIntense

BitterIntense is a unique tool able to quantify the bitter intensity of a query molecule, discriminating between two classes, i.e. "not very bitter" (NVB) and "very bitter" (VB) [347]. This tool is paramount not only for food research but also for pharma and biotechnology industries: the ability to predict the level of bitterness during the drug discovery process represents a promising opportunity for reducing delays, animal use and financial costs. In fact, the intensely bitter taste is often associated with difficulties in taking medication, especially for children and elderly people. BitterIntense, published at the end of 2020, was also applied to widely known databases, such as DrugBank, and specific COVID-19 drug candidate datasets, highlighting interesting considerations regarding bitter intensity and toxicity.

Data preparation and model construction

The screening of bitter compounds was performed employing the rat brief-access taste aversion (BATA) model, obtaining 34 compounds. The dataset collects BitterDB and AnalytiCon's repository of natural compounds and counts about 180 molecules with a specified bitter intensity. The bitter recognition threshold is 0.1 mM: below this concentration, the molecules were considered "very bitter" (VB), whereas above this value "not very bitter" (NVB). Molecules without quantitative information were assigned to VB/NVB classes according to the taste descriptions. A non-bitter database of 152 randomly selected compounds from the negative set of BitterPredict was added to NVB class.

Moreover, external datasets have been screened using the optimized model. Toxicity data includes the FocTox dataset, i.e. extremely hazardous compounds and FAO/WHO food contaminants, and the CombiTox dataset, i.e. a combination of

DSSTox-the Distributed Structure-Searchable Toxicity Database and Toxin and Toxin-Target Database version 2.0 (T3DB) [331]. Hepatotoxicity data was retrieved from FDA's DILIrank dataset [369]. Finally, external datasets include DrugBank [327], consisting of approved and experimental drugs, COVID 19 drugs and their targets retrieved from ''Coronavirus Information. IUPHAR/BPS Guide to Pharmacology" and Natural products atlas (NPatlas, version 2019_08) [329].

SMILES were processed using Maestro (Schrödinger) (3D reconstruction, protonation at pH 7.0 ± 0.5, removal of additional molecules and generation of conformers). Molecular descriptors were calculated using Canvas (Schrödinger) and were divided into three groups, i.e. Physicochemical, Ligfilter and QikProp. Compounds not having one of these were excluded. Feature selection was performed using the feature importance gain score, obtaining a total of 55 features (from the starting 235).

The dataset was randomly divided into a training set (169 VB and 324 NVB), a test dataset (43 VB and 80 NVB) and the hold-out set for an external evaluation (31 VB and 74 NVB).

The algorithm used was the Extreme Gradient Boosting (XGBoost). Logarithmic loss and binary classification error rate were selected to monitor step by step the algorithm performance and stop it when the improvement subsides. Parameters of the models were tuned through a 10-fold CV.

Model performance

Performance evaluation was made for the three different datasets, i.e. training set (with 10 fold CV), test set and hold-out set, and for each of them accuracy (ACC), precision, sensitivity (SENS) and F1 score were calculated (ACC over 80% in all sets, PRC: 80%, 71%, 63%; SE: 85%, 86%, 77%; F1-score: 82 ± 5%, 78%, 70%, respectively). From these results, there are more false positives than false negatives, indicating the maximization of the identification of very bitter compounds.

From the analysis of the feature importance, the algorithm pointed out the role of the molecule's size and molar refractivity (a measure of polarizability) in determining the bitterness level, suggesting also a correlation between molecule size and bitter intensity.

BitterIntense applied to toxic databases, i.e. FocTox and CombiTox datasets, revealed that only a small portion (about 10%) of toxic substances are intensively bitter. The use of the BitterIntense model to the DILIrank dataset allows the evaluation of the correlation between bitterness level and hepatotoxicity, showing that most of the drugs (729) were classified as NVB. Then, approved and experimental compounds from Drugbank database (10170 compounds) and natural compounds from NPatlas (24805 compounds) were screened, showing that almost half of microbial natural products, but only 23.7% of drug candidates are predicted

as VB. Finally, 34 potential drug candidates against COVID-19 retrieved from "Coronavirus Information – IUPHAR/BPS Guide to Pharmacology" were classified, showing that 41.2% of them are likely VB, thus significantly higher than the percentage of VB drug compounds from Drugbank, suggesting a possible involvement of the bitter taste and bitter receptors in this disease.

Umami Prediction

a. iUmami-SCM

iUmami-SCM is the first umami taste predictor based on umami peptide primary sequence information [319]. iUmami-SCM is a webserver (http://camt.pythonanywhere.com/iUmami-SCM) and related datasets are available on GitHub (https://github.com/Shoombuatong/Dataset-Code/tree/master/iUmami).

Data preparation and model construction

The dataset, known as UMP442, contains 140 proved umami peptides and 304 bitter structures taken from iBitter-SCM as negative samples (see also *Taste and food-related databases* for further details). Interestingly, the peptide length of both positive and negative samples is less than 10 amino acid residues. UMP442 was divided randomly into two parts keeping the unbalancing between the positive and negative data: the training set, made up of 80% of the dataset, was employed for the generation of an initial scoring card with a statistical approach and its optimization through a GA algorithm and the independent set (UMP-IND), composed of 20% of the dataset, was employed for performance evaluation. Dipeptides propensity scores and informative physicochemical properties were employed as features in this model.

Model performance

Prediction performance depends on the optimal dipeptide propensity score (opti-DPS), therefore 10 opti-DPS were evaluated with a 10-fold CV and compared with the initial dipeptide propensity score (init-DPS). Notably, compared to other traditional ML methods (DT, KNN, MLP, NB, SVM, and RF), iUmami-SCM demonstrated better performance.

iUmami-SCM reached on the test set accuracy of 86%, MCC of 68%, AUC of 89.8%, sensitivity of 71.4% and a specificity of 93.4%. All these reported performances were calculated on the opti-DPS. However, due to the reduced numbers of peptides used for the model construction, iUmami-SCM presents as a major shortcoming a limited ability to correctly generalise the prediction.

Multi-Taste Prediction

a. BitterSweet Forest

BitterSweetForest is an open-access model based on KNIME created in 2018. This machine learning classifier predicts the sweetness and the bitterness of chemical compounds using binary fingerprints [348].

Data preparation and model construction

The dataset contains 517 artificial and natural sweet compounds, derived from SuperSweet, and 685 bitter molecules, taken from BitterDB. Instant Jchem software was employed for molecules standardization. All duplicated molecules were removed. Four different binary fingerprints were calculated with RDKit node in KNIME: Morgan fingerprint, Atom pair fingerprints, Torsion fingerprint and Morgan Feat fingerprints. Training and test sets were obtained with an 80:20 partitioning scheme, keeping the balance between the two classes. To avoid overfitting, a leave-one-out cross-validation (LOO) was performed. A Random Forest with Tree Ensemble Learner and Predictor nodes in KNIME [370] was implemented and a Bayesian-based features detection was applied to analyse the important and frequent features.

Model performance

The model was evaluated with several performances: accuracy, sensitivity, specificity, precision, F-score, ROC-AUC and Cohen's kappa. The BitterSweet model reached accuracy of 96.7%, AUC of 98% and sensitivity of 91% and 97% for sweet and bitter prediction respectively. Bayesian-based feature detection emphasised the independence between the top 10 features of sweet and bitter molecules, despite the two molecule sets appearing to show similar characteristics.

The performance was also calculated in an external validation set, which includes bitter, sweet and tasteless molecules. Despite tasteless molecules are not included in the training dataset, the model provided good results, and this suggests the features employed are specific for bitter and sweet prediction. Interestingly, the screening of SuperNatural II, DrugBank approved drug molecules and ProTox, including oral toxicity compounds, showed that most molecules exhibited bitter features and toxic substances are normally bitter.

b. BitterSweet

BitterSweet is a freely accessible tool created in 2019 to classify bitter-sweet molecules [349]. To boost the progress in the knowledge of bitter-sweet taste molecular basis, the creators of this tool make all datasets, models and even end-to-end software publicly available (https://cosylab.iiitd.edu.in/bittersweet/; https://github.com/cosylabiiit/bittersweet).

Data preparation and model construction

The dataset was created avoiding two problems observed in previous studies: the use of *unverified flavour molecules* in the training dataset, as happened in BitterPredict and BitterX, and the use of *only experimentally verified data*, leading

to a drastic reduction in the dataset size. The dataset contains around thousands of chemicals among bitter, non-bitter, sweet and non-sweet compounds retrieved from literature [318,320,343], pre-existing databases, i.e. SuperSweet, The Good Scents Company Database, BitterDB and books, i.e. Fenaroli's Handbook of Flavor Ingredient And Biochemical Targets of Plant Bioactive Compounds. Moreover, as the control for bitter and sweet prediction, tasteless and contrasting taste compounds, derived from ToxNet, TastesDB and Fenaroli's Handbook of Flavor Ingredient, were introduced in the dataset. The canonical SMILES were extracted through OpenBabel [371]. Duplicate structures, peptides, molecules with only three atoms and salt ions were removed, while only the lowest energy conformer for each molecule was retained. The chirality of the molecule was preserved. The 3D conformation and protonation state at physiological pH (7 ± 0.5) were carried out using Epik [372] and LigPrep (Schrödinger).

The training dataset for bitter/non-bitter prediction included 813 bitter molecules as positive data and 1444 sweet and tasteless molecules as the negative set, while for sweet/non-sweet prediction it consisted of 1139 sweet molecules as the positive set and 1066 bitter and tasteless compounds as the negative set. The test dataset was formed by 105 bitter and 66 non-bitter structures in the bitter prediction and 108 sweet and 53 bitter/tasteless molecules in sweet prediction. Moreover, a 5-fold stratified CV was performed to assess the model parameters.

A five-set of molecular descriptors, both commercial and open-source, was employed to create an exhaustive set of features: Physicochemical and ADMET descriptors from Canvas, Extended Connectivity Fingerprints (ECFP), 2D Molecular Descriptors and 2D/3D Molecular Descriptors from Dragon 2D and Dragon 2D/3D and 2D Topological and Structural Features from ChemoPy. Due to the high number of molecular descriptors, the Boruta algorithm [373] was employed to remove irrelevant features and principal component analysis (PCA) to get the maximum variance. Three different ML-based models were employed, i.e. Random Forest (RF), Ridge Logistic Regression (RLR) and Adaboost (AB). For each algorithm and each prediction, the five-set of molecular descriptors were evaluated separately.

Model performance

BitterSweet model performance was evaluated employing several metrics, including the Area Under the Precision-Recall Curve (PR-AUC), ROC-AUC, F1-score, sensitivity and specificity. The models that best discriminate the sweet non-sweet dichotomy were AB and RF trained after the Boruta algorithm, while PCA performed better than the Boruta algorithm only when coupled with RLR. In contrast, the algorithm that best predicts bitter taste was RLR, while the RF performed well across all molecular descriptor sets. Furthermore, PCA would seem to perform better than the Boruta algorithm. The best descriptors for the sweet

prediction were Dragon 2D features, whereas the open-source ChemoPy performed better for bitter prediction.

In conclusion, the best BitterSweet model (with AB algorithm after the Boruta feature selection and Dragon 2D/3D features) achieved these performances: ROC-AUC of 88.3%, PR-AUC of 95%, the sensitivity of 79%, the specificity of 88% and the F1-score of 86%. However, in their online tool, they employed ChemoPy descriptors with the RF-PCA algorithm as the performance of open-source descriptors were comparable to those obtained through proprietary software. The results achieved with the BitterSweet model: ROC-AUC of 84% and 88%, PR-AUC of 93% and 93%, the sensitivity of 59% and 79%, the specificity of 94% and 85% and the F1-score of 73% and 84%, all the results were reported for sweet and bitter prediction, respectively.

The BitterSweet model was also applied to several specialized chemical databases, i.e. SuperSweet, FlavorDB, FooDB, DSSTox, SuperNatural II and DrugBank, revealing that the majority of natural, toxic, and drug-like molecules are bitter, whereas for food molecules there was the same amount of bitter and sweet molecules.

In conclusion, despite the high accuracy of the BitterSweet open-source predictors, its utility is limited to individual compounds, and not for different compounds when present in a mixture.

c. Virtual Taste

VirtualTaste platform is the first freely available web server able to predict three taste qualities (sweet, bitter and sour), thanks to three dedicated tools, i.e. VirtualSweet, VirtualBitter and VirtualSour, respectively (http://virtualtaste.charite.de/VirtualTaste/)[350]. The input of the web-based platform is the two-dimensional structure of the chemical compound and the output is the prediction of the chemical's taste profile and the targeted TAS2R receptors in case of a bitter prediction.

Data preparation and model construction

The dataset contains 2011 sweet compounds, collected from the SuperSweet database and BitterSweetForest tool, 1612 bitter molecules derived from the BitterDB database and BitterSweetForest tool, and 1347 sour compounds obtained from ChEMBL [326] and manually edited from literature sources [374]. Furthermore, the bitter receptor data contains 356 ligands that interact with TAS2Rs receptor extracted from BitterDB, ChEMBL and literature. Different structures were removed from the database, such as ambiguous compounds, salt and inconclusive entries, and then standardised through RDkit in KNIME [370]. Each dataset was split into two parts, preserving the positive/negative ratio: the training set made up of 80% of each set of molecules, i.e. 1068 molecules for sweet, 1289 compounds for

bitter and 1214 structures for sour, and the remaining chemicals were employed for the external validation set. The inactive dataset used in each model were different: bitter and tasteless compounds were used as inactive compounds for the sweet prediction and sweet and tasteless compounds were employed as inactive compounds for the bitter prediction, while the sour prediction used a ligand-based approach due to the pH and acid influence present in foods. Moreover, a 10-fold CV was applied for model optimisation, keeping the ratio of active and inactive structures constant.

Each VirtualTaste model was based on an RF algorithm, following BitterSweetForest, the previous tool developed by the same research group [348]. To deal with the negative effect of the unbalanced dataset, different data sampling methods were applied: the Synthetic Minority Over-Sampling Technique-using Tanimoto Coefficient (SMOTETC) technique for VirtualSweet, the Synthetic Minority Over-Sampling Technique-using Value Difference Metric (SMOTEVDM) method for VirtualBitter and the Augmented Random Over Sampling (AugRandOS) method for VirtualSour [375]. A similarity-based method was employed for the prediction of bitter receptors [376]: the similarity between the query molecule and known bitter compounds is evaluated using the Tanimoto Coefficient and the relative target bitter receptor is then consequently predicted.

VirtualSweet and VirtualBitter models were also used for taste prediction of approved drugs and natural compounds - 1969 chemical compounds from DrugBank database and 326000 from SuperNatural II.

Model performance

Five performance metrics both in the 10-fold CV and in the external evaluation set were utilized. VirtualSweet reached on the external validation 95% for ROC-AUC, 89% for the accuracy, 92% for specificity, 86% for sensitivity and 88% for F1-score; VirtualBitter 96% for ROC-AUC, 90% for the accuracy, 97% for specificity, 88% for sensitivity and 88% for F1-score; VirtualSour 99% for ROC-AUC, 97% for the accuracy, 99% for specificity, 80% for sensitivity and 84% for F1-score. In conclusion, VirtualTaste is the first tool able to predict with reliable results three different taste qualities and achieve comparable or better performance compared to similar tools.

3.1.4 Discussion

In this section, a detailed comparison between all the above-described taste prediction tools is provided. The performance of the classification and regressor models is summarized in Table 3.4 and Table 3.5, respectively. It is noteworthy that these comparative results were not obtained on the same datasets and different

evaluation metrics were used in each analysed work. Data in the tables refer to performance on the test set.

Table 3.4. Performance on the test set of the taste prediction classification tools.

Taste	Tool	Performance (%)							
		AUC	**SE**	**SP**	**ACC**	**PRC**	**NER**	**F1**	**MCC**
Sweet	Rojas Sweet Predictor	/	88	82	/	/	85	/	/
	eSweet	/	86	94	91	90	90	88	81
Bitter	BitterX	95	92	91	92	91	/	/	/
	BitterPredict	/	77	86	/	/	/	/	/
	e-Bitter	/	98	81	92	/	/	94	82
	iBitter-SCM	90	84	84	84	/	/	/	69
	BERT4Bitter	96	94	91	92	/	/	/	84
	iBitter-Fuse	93	94	92	93	/	/	/	86
	BitterIntense	/	86	81	83	71	/	78	/
Bitter-Sweet	BitterSweet Forest	98	91	97	97	/	94	92	/
	BitterSweet (Sweet)	84	59	94	/	/	77	73	/
	BitterSweet (Bitter)	88	79	85	/	/	82	84	/
Umami	iUmami-SCM	90	71	93	87	/	/	/	68
Sweet-Bitter-Sour	VirtualSweet	95	86	92	89	/	/	88	/
	VirtualBitter	96	88	97	90	/	/	88	/
	VirtualSour	99	80	99	97	/	/	84	/

Table 3.5. Performance on the test sets of the sweet prediction regression tools.

Tool	Performance		
	R^2	**MAE**	**MSE**
Chéron Sweet Regressor	0.85	/	/

Goel Sweet Regressor	0.83	0.39	0.23
Predisweet	0.74	0.50	0.44

At present, it is evident that there is a net prevalence of tools for predicting sweet and bitter tastes. It is worth noting that only one example to predict the umami taste (iUmami-SCM) and one for the sour taste (VirtualSour in VirtualTaste) exist, and no tools for predicting the saltiness have been released, as far as the authors know. Moreover, the definition of a regression algorithm was possible only for the sweet taste (Chéron Sweet Regressor, Goel Sweet Regressor and PrediSweet) since, to date, no database for other taste sensations provides quantitative data concerning the level of the perceived taste. However, BitterIntense, despite being a classification algorithm, discriminates between "*very bitter*" and "*non very bitter*" compounds, thus accessing the level of bitterness of query molecules.

Despite all prediction tools employ a different methodology, a common structural features can be noticed among all of them, which is typical of most ML workflows: (i) the definition of a compound database also including the respective taste, preferably experimentally validated; (ii) the compound featurization, i.e. the derivation of effective molecular descriptors; (iii) the dataset splitting into training and test sets (and in some cases also a validation set); (iv) the choice of the ML method for the classification/regression; (v) performance evaluation and validation. It is worth mentioning that most of the discussed algorithms followed the guidelines defined by the Organization for Economic Co-operation and Development (OECD), which indicates the strategies for correct development and validation of robust QSAR models: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation, if possible [377].

Several tools and methods, both proprietary and open-source, were used to derive molecular descriptors, including Dragon, Canvas (Schrödinger), Extended-connectivity Fingerprint (ECFP), RDKit, Mordred and ChemoPy. It is important to note that open-source descriptors (RDKit, Mordred, ChemoPy) have been shown not to remarkably affect the performance of the PrediSweet model and to reach similar results if compared to results obtained with Dragon descriptors [340]. Similarly, in BitterSweet the best descriptors for the sweet prediction were Dragon 2D features, but the open-source ChemoPy performed better for bitter prediction [349]. This represents a very important achievement in making these tools available to a wide audience and in broadening the horizons of research in this field. Furthermore, 2D molecular descriptors are less time-consuming to be computed and less subject to variations caused by slightly different molecules 3D conformations. On the other hand, 3D descriptors can also account for specific molecule conformations, such as different conformers/isomers, and spatial

properties. Therefore, the possibility to obtain very good results also using 2D descriptors allows designing faster tools suitable for screening very large databases.

Several algorithms have been applied for taste prediction, including RF, SVM, SVR, QSTR, GFA, ANN, KNN, GBM, DNN, AB, SCM, XGBoost. Multiple Linear Regression (MLR) and Support Vector Machine (SVM) are among the first models for binary classification. These models were exceeded by tree-based models, i.e. Random Forest (RF) or AdaBoost (AB), and Neural Network (NN), which support multiclass classification and work very well in the non-linear range if they have a sufficiently large number of database elements. Generally, NN and SVM perform better with continuous and multidimensional features but they need a large sample size to increase their prediction accuracy [378]. Even though NNs, and in particular ANNs and DNNs, are being widely employed in taste prediction, they are characterised by difficulties in optimising parameters, a high computational cost and are less explainable. Moreover, probabilistic methods, i.e. Naive Bayes, are not widely used in taste prediction. These methods work well with less training data but would be better employed when the features are mutually independent.

To enhance model performance and to increase the understanding of the model, a feature selection was normally applied, such as the V-WSP unsupervised variable reduction method and genetic algorithm-based technique [320,341,346], feature importance obtained from the RF [339,344], the Boruta algorithm and the PCA [349]. However, none of these approaches consider the multi-objective nature of the dimensionality reduction techniques and thus fail to balance between the objectives of optimizing prediction performance measured in multiple classifications and regression metrics, minimizing the number of selected features and maximizing the overall interpretability/explainability of the derived prediction models. Moreover, not only feature selection but also normalisation/standardisation can further improve model performance, such as in the case of the *flatkinson* standardisation method [340], as well as modern techniques to handle the class imbalances in the data such as SMOTE [379]. Finally, all existing methods lacked a strict definition of negative datasets with most of them using random compounds as negative datasets and thus jeopardizing the prediction performance and generalization properties of the models.

As also defined by the OECD guidelines, another relevant aspect in the development of the prediction tool is the definition of the *applicability domain* (AD), which indicates the reliability of the prediction evaluating if investigated compounds are within the chemical space of the training data. In this context, PrediSweet, along with an applicability domain, developed a *reliability domain*, which considers the density of information around the compound, and a *decidability domain*, which evaluates the confidence of the prediction. The PrediSweet chemical space of the dataset used (SweetenersDB) was compared with the most comprehensive sweet database (SuperSweet): more than 99.5% of the compounds

in SuperSweet are structurally similar to a representative structure in the SweetenersDB, suggesting the large sweeteners spectrum covered by the SweetenersDB. e-Bitter and e-Sweet used the ECFP based Tanimoto-similarity between the query and the five closest neighbouring molecules in the training set. Similarly, the Rojas Sweet Predictor developed an AD using a threshold on the Jaccard-Tanimoto average distance between the query molecule and the compounds in the dataset. BitterPredict AD, known as *Bitter Domain*, includes molecules with molecular weight MW $\leq$ 700 and hydrophobicity -3 $\leq$ AlogP $\leq$ 7: all used datasets were previously filtered using this domain to ensure the reliability of the prediction. In BitterSweet, a query molecule is considered inside the applicability domain, if its median Euclidean distance from similar compounds in the training set is below a selected threshold. Interestingly, BitterSweet covered a remarkably wider applicability domain than the Rojas Sweet Predictor while achieving similar performance.

One of the main advantages of the reported prediction tools is their ability to fast screening huge databases of compounds and to estimate the number of compounds associated with a specific taste. A granular and detailed screening was performed by BitterPredict on DrugBank approved (1375 compounds), FooDB (13588 compounds), Natural Products Dataset from ZINC15 (27474 compounds) and ChEBI (27015 compounds) datasets, showing that the percentages of bitter molecules within the Bitter Domain found in these databases are 65.94%, 38.36%, 77.21% and 43.71%, respectively. Moreover, since bitter taste is associated with toxic compounds or compliance problems, the same authors used their subsequent tool, namely BitterIntense, to screen toxic databases, i.e. FocTox, CombiTox datasets and DILIrank, experimental compounds from DrugBank database (10170 compounds), natural compounds from NPatlas (24805 compounds) and 34 potential drug candidates against COVID-19 retrieved from "Coronavirus Information – IUPHAR/BPS Guide to Pharmacology". Interestingly, only a small portion of toxic compounds are intensively bitter, but 41.2% of COVID-19 candidate drugs were predicted as very bitter (VB). Moreover, BitterSweet was applied to several specialized chemical databases (SuperSweet, FlavorDB, FooDB, DSSTox, SuperNatural II and DrugBank), revealing that most natural, toxic, and drug-like chemicals are bitter, whereas the same amount of bitter and sweet molecules are present in foods. BitterSweet Forest was applied on SuperNatural II, DrugBank approved drug molecules and ProTox, and showed that toxic substances are typically bitter. In line with the previous results, VirtualSweet and VirtualBitter models were applied on approved drugs from DrugBank and natural compounds from SuperNatural II: notably, most of the approved drugs and most of the natural compounds were predicted as bitter and only a small portion of these databases was classified as sweet.

3.1.5 Conclusions

The present review aims at summarising the main scientific advances in the field of taste prediction supported by ML-based algorithms. We discussed the main available database containing food-related compounds and molecules with known taste, the main tools employed to predict the taste.

From the analysis of the databases, we pointed out two specific databases for the sweet taste, i.e. SuperSweet, which is the most comprehensive DB for sweeteners, and SweetenersDB, which collects 316 sweeteners with a relative value of sweetness. For the bitter taste, BitterDB represents the most granular and complete database, with a very intuitive and user-friendly web server that allows the download of more than a thousand bitter compounds. BitterDB, as well as BitterPredict and BitterIntense, was developed by the Niv Lab (https://biochem-food-nutrition.agri.huji.ac.il/mashaniv), which provided incredible progress in the comprehension of the bitter taste in recent years. A lot of effort has been made to develop methods specifically for the prediction of bitter peptides and another research group has continuously improved its tools publishing three consecutive works, namely iBitter-SCM, BERT4Bitter and iBitter-Fuse, in the last few years. These tools are paramount for the fast and reliable classification of huge databases of bitter peptides and for their rational de novo design, especially considering the emerging role of this class of compounds in the drug and nutritional research field. Furthermore, the UMP442 database developed during the implementation of iUmami-SCM is probably the most complete and ready-to-use database of umami and non-umami molecules since it is available from GitHub. In this context, the Umami Database seem a very promising source of information, but the availability of data is limited, it is impossible to obtain data from the webserver and no umami prediction tool which uses this source has been found in previous literature. It would be incredibly valuable to have access to the resources of such a database in the future. Finally, as far as the authors know, no publicly available databases for sour and salty tastes are available and the only attempt to generate a sour dataset was made in the development of VirtualTaste. However, the used sour dataset has not been made public. To date, the multiplicity and diversity of sources make it very complex to obtain a unified DB collecting a huge amount of compounds for each taste sensation. The authors insist also on the need for developing complete databases that include all the relevant information for each entry (SMILES, InChI, IUPAC nomenclature, etc.) to avoid any possible error in compound processing. Moreover, the definition of exhaustive databases would be essential for a correct definition of the molecular descriptors to be employed, due to the great number and variety of both open-source and proprietary descriptors.

Similarly to the taste databases, prediction tools for sweet and bitter have been more developed during the last years. Among several examples of sweet and bitter classification tools, some proposed methods can even predict the level of sweetness

(Chéron Sweet Regressor, Goel Sweet Regressor, PrediSweet) and BitterIntense can discriminate between “very bitter” and “non very bitter compounds”. Only a few reported tools were able to discriminate more than one taste sensation. BitterSweet and BitterSweet Forest are interesting examples of tools able to consider the dichotomy of sweet and bitter tastes [349]. These tools can be pivotal for the detection of natural and synthetic compounds with a pleasant taste and without adverse effects. Furthermore, VirtualTaste is the only available tool able to predict three taste sensations (sweet, bitter and sour) and the only one able to predict the sour taste (VirtualSour).

Among the 16 reported taste prediction tools, BitterX, BitterSweet, PrediSweet, iBitter-SCM, BERT4Bitter, iBitter-Fuse, iUmami-SCM and VirtualTaste provide web server applications, which allow the taste prediction using the SMILE/Fasta format, directly drawing the molecule or by uploading a file. On the other hand, eSweet, eBitter and BitterPredict only provide freely accessible code available from Dropbox or GitHub. From the authors' point of view, the development of a web interface represents a considerable strength, since it allows the tool to be used even by people who are not experts in the use of these applications. Finally, BitterSweetForest, Rojas Sweet Predictor, Goel Sweet Regressor, Chéron Sweet Regressor and BitterIntense do not have any web server or code publicly available.

It is worth mentioning that in addition to the five basic tastes, other taste qualities may be important and related to specific food ingredients. In this context, some recent publications suggested the fat taste as another basic taste quality [123–125]. Interestingly, fatty acid detection seems to decrease as a consequence of a fat-rich diet and with a great impact on obesity disease [126]. Therefore, the prediction of fat taste would represent a groundbreaking objective for future tools, considering the impact of fat intake on human health status.

Furthermore, the use of methods capable of predicting the molecular interactions of tastants and relative taste receptors could lead to significant improvements in the predictive capabilities of these tools and to great strides in understanding the physicochemical characteristics and mechanisms underlying taste perception.

3.2 VirtuousUmami

The present section is based on the following scientific publication:

***Pallante, L.**, Korfiati, A., Androutsos, L., Stojceski, F., Bompotas, A., Giannikos, I., Raftopoulos, C., Malavolta, M., Grasso, G., Mavroudi, S., Kalogeras, A., Martos, V., Amoroso, D., Piga, D., Theofilatos, K., & Deriu, M. A. (2022). **Toward a general and interpretable umami taste predictor using a multi-objective machine learning approach**. Scientific Reports, 12(1), 21735. https://doi.org/10.1038/s41598-022-25935-3.*

***Author's contribution to the publication:** Pallante L. worked on data preprocessing, developing a python library for preprocessing molecules and calculating molecular descriptors, and implementing the applicability domain. He also took part in the conception of the work, the analysis of the results, their critical discussion, the writing of the manuscript and its revisions.*

Starting from the recent advances in the field of taste prediction reported in the previous section, we present herein a novel ML-based tool, named VirtuousUmami, to predict the umami taste of a query compound starting from its molecular structure. Umami taste is one of the five basic taste modalities normally linked to the protein content in food. The implementation of fast and cost-effective tools for the prediction of the umami taste of a molecule remains extremely interesting to understand the molecular basis of this taste and to effectively rationalise the production and consumption of specific foods and ingredients. However, the only examples of umami predictors available in the literature rely on the amino acid sequence of the analysed peptides, limiting the applicability of the models. In the present study, we developed a novel ML-based algorithm, named VirtuousUmami, able to predict the umami taste of a query compound starting from its SMILES representation, thus opening the possibility of potentially using such a model on any database through a standard and more general molecular description. Herein, we have tested our model on five databases related to foods or natural compounds. The proposed tool will pave the way toward the rationalisation of the molecular features underlying the umami taste and toward the design of specific peptide-inspired compounds with specific taste properties.

3.2.1 Introduction

Umami taste is one of the five basic taste modalities and it is typically associated with the protein contents of foods. The term "umami" originates from a Japanese word that means "pleasant savoury taste", "mouthfulness" or "delicious" [380]. Umami has been linked for several years to the taste of Asiatic traditional foods or cheese and it was recognized as the fifth basic taste modality - along with sweet, bitter, salty and sour - only in 2002 to describe a pleasant or glutamate-like taste [381]. Since the umami taste is commonly linked to the food protein content, it represents an interesting taste modality, especially for, but not limited to, food industries: considering the laboriousness of traditional experimental techniques, it is pivotal to

develop fast, reliable and cost-effective methodologies able to predict the taste of food ingredients or general compounds with the ultimate goal of identifying and characterizing their chemical profile. Several experimental methods, including MALDI-TOF-MS and reversed-phase high-performance liquid chromatography (RP-HPLC) analysis, are widely used to identify and characterize peptides with umami sensory properties [382,383]. However, traditional experimental methods for characterizing and profiling from a chemical point of view the umami peptides are expensive, time-consuming, and arduous. In this context, the *in-silico* techniques have been pointed out as elicit methods to screen massive databases of compounds and retrieve specific information regarding their activity or properties through the employment of machine learning algorithms. Quantitative structure-activity relationships/quantitative structure-property relationships (QSAR/QSPR) methods aim at determining a relationship between the biological activity or the physicochemical property, respectively, and a set of descriptive features (descriptors) linked to the molecular structure of the investigated molecules [384]. In this regard, the guidelines defined by the Organization for Economic Co-operation and Development (OECD) indicate the strategies for the correct development and validation of robust QSAR models: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation, if possible [377].

Regarding the *in-silico* prediction of taste based on the molecular structure of compounds, a lot of advancements have been accomplished [385]. For example, several publications deal with the prediction of the sweet taste [386–392], the bitter taste [393–400], and the bitter/sweet dichotomy [401,402]. However, as far as the authors know, there are few attempts made by the scientific community to predict the umami taste, which are represented by the iUmami-SCM [403] and the UMPred-FRL [404] predictors. The iUmami-SCM tool predicts the umami/non-umami taste of peptides based on their primary amino acid sequence employing a scoring card method (SCM) in conjunction with the propensity scores of amino acids and dipeptides. For its design, this tool is limited to the prediction of only peptides, which however represent the candidate par excellence of umami taste. Another effort again focused on umami peptide identification is the UMPred-FRL tool, which demonstrates a higher feature discriminative capability to capture the key information about umami peptides and superior performance compared to the iUmami-SCM. However, a method for screening databases of general molecules or predicting the taste of peptides with small chemical deviation from their original structures is needed to pinpoint the major physio-chemical properties related to the occurrence of the umami taste and allow the identification of umami-related compounds from bigger pools of potential compounds. The present work is therefore based on these premises and is devoted to developing an efficient tool to predict the umami/non-umami taste of query molecules based on their chemical structure described using

the standard SMILES representation and commonly employed molecular descriptors. An ensemble dimensionality reduction and classification techniques were used to train and test the umami taste prediction model, minimizing the number of physicochemical features used as inputs and allowing the identification of the most important features related to the umami taste. The minimization of the inputs makes the prediction models simpler, reducing thus the risk of overfitting, and enables the incorporation of the prediction models in a web interface enlarging the ensemble of possible end-users. The developed tool, named VirtuousUmami, paves the way toward the possibility of analyzing different types of compounds and rationalising the chemical-physical characteristics at the basis of umami taste perception to design new ingredients and molecules with specific taste properties.

3.2.2 Materials and Methods

Data curation

For an effective comparison with previous literature dealing with umami taste predictors, the UMP442 database, also used for iUmami-SCM[319] and UMPred-FRL[404] predictors, was employed. The UMP442 dataset is freely accessible from GitHub https://github.com/Shoombuatong/Dataset-Code/tree/master/iUmami) and collects 442 peptides (140 umami and 302 non-umami): umami molecules are gathered from previous literature[61,332–336] and the BIOPEP-UWM database[325], whereas non-umami peptides are the bitter peptides from the positive set of the BTP640 database[337] (see also Table A - 6.6.1). The peptides were gathered using their amino acid sequences and then converted into their SMILES representation using the RDKit package (http://www.rdkit.org). Then, they were processed with the ChEMBL Structure Pipeline [405] (https://github.com/chembl/ChEMBL_Structure_Pipeline) to highlight possible issues in the retrieved molecular structure and to standardise the SMILES representation for the entire dataset. The latter protocol runs a molecule checker on the compound structure, standardizes chemical structures and generates the parent molecule representation based on a set of predefined rules.

Among 442 umami (140) and non-umami (302) peptides available in the UMP442 dataset, 352 ligands were used for training. The remaining 90 peptides were used for external testing to examine the generalization properties of the trained models. Of the 352 training samples, 240 were non-umami samples, and 112 were umami samples. Because there is an imbalance in the total number of samples of the two classes, we oversampled the umami class, creating synthetic data to boost the umami class. These synthetic data were created by selecting random samples from the umami class and duplicating them, a method of random oversampling for the minority class. The resulting training dataset had 240 non-umami samples and 240 umami samples. Of the 90 testing samples, 62 were non-umami samples, and 28

were umami samples. The summary of the final dataset is also reported in Table A - 6.6.2.

Molecular descriptors and dimensionality reduction

The calculation of the features for each one of the molecules was achieved using 1613 2D Mordred descriptors. The dataset was pre-processed to be used as input to the machine learning model. In particular, features with a high percentage of missing values (>30%) were filtered, while the remaining missing values were imputed using the kNN-impute method with k=20 [406]. Then, data were arithmetically normalized to the interval of [0-1]. Given the huge number of total features, i.e. 1613, compared to the size of the training dataset, an initial univariate filtering approach was deployed. The statistical analysis was performed on the umami vs non-umami peptides of the training set with the limma eBayes method [407], and correction of p-values for multiple testing was performed using the Benjamini-Hochberg FDR adjustment method [408] to calculate q-values. For both p- and q- values a threshold of 0.05 was applied. We used also four different feature selection methods, i.e. the Wilcoxon Rank Sum Test[409], kBest, JMIM[410] and MRMR[411], to further reduce the dimensionality of the training dataset. These methods were iteratively tested using an in-house evolutionary optimization algorithm (50 individuals and 100 generations) which identified the best combination of feature selection techniques among the above-mentioned alternatives. The results of these methods are used at every generation of the evolutionary algorithm for every individual to reduce the features in the training process. In this way, we are confident that at each run we select the most important features for our problem.

Data preprocessing, statistical analysis and the generation of additional plots, such as ROC curves and bean plots, were performed using the InSyBio Biomarkers Tool (see also the reference Manual for further details at https://www.insybio.com/biomarkers.html).

Model construction and performance evaluation

The classification models were generated with the hybrid combination of heuristic optimization and nonlinear machine learning classification methods incorporated in the InSyBio Biomarkers tool (https://www.insybio.com/biomarkers.html). Specifically, we used an ensemble dimensionality reduction technique employing a heuristic multi-objective Pareto-based evolutionary optimization algorithm [412] to (a) identify the optimal feature subset to be used as input to the classifiers, (b) select the most appropriate classifier among Support Vector Machines (SVM) and Random Forests and (c) select the optimal parameters for the classifier, namely C and gamma of SVM and number of trees for Random Forests. This approach allows both an unbiased and an optimized selection of the classification method and its parameters. The multi-objective Pareto-based approach was deployed to handle the multiple objectives of maximization of predictive performance, minimization of

selected features and simplicity of the classification model, revealing all the non-dominated solutions of the above-stated optimization goals. The weights used for the goals were Selected Features Number Minimization 5, Accuracy (ACC) 10, F1 score 5, F2 score 1, Precision (PRC) 1, Recall (REC) 10, ROC-AUC (AUC) 1, Number of SVs or Trees Minimization 1, which enable better handling of the imbalanced nature of our classification problem. The outcomes are multiple models performing equally well (namely, the Pareto set of optimal solutions) on the user-defined goals. After having defined the best models in terms of performance metrics, we developed ensemble models (EMs) to further improve the prediction performance. In greater detail, an ensemble model is built by combining two different single models: the final prediction probabilities of the ensemble model for the positive and negative classes is the average of the prediction probabilities coming from the two combined models. The final predicted class is therefore the one with the highest probability score.

A population of 50 individuals was used for the evolutionary algorithm and a maximum number of 100 generations was used as the termination criterion. To deal with the stochastic nature of the proposed algorithm, five different runs were conducted and the results presented are the average performance of these runs. Convergence of the algorithm (average performance less than 5% different to best performing individual) was noted after 30 generations for each independent run demonstrating that the maximum number of generations used was adequate for this problem. Additional parameters of the evolutionary algorithm were set to their default values as suggested by the InSyBio Biomarkers tool user manual (arithmetic crossover probability: 0, mutation probability: 0.01, two-point crossover probability: 0.9). Stratified 10-fold cross-validation was used to train and test the prediction models. To deal with the class-imbalanced nature of our data, in each cross-validation iteration, we applied random oversampling of the minority class in the 9 folds which were used to train the models. Further details on the implementation of the trained models and a summary of the performance metrics used are available in the Supplementary Information.

Applicability domain

In the present work, following the guidelines defined by the Organization for Economic Co-operation and Development (OECD) [377], we developed an applicability domain (AD) to provide information regarding the reliability of the prediction. We used an average-similarity approach already employed in previous recent literature in the taste prediction field [389,395]. More in detail, the AD was built as follows: (i) the Morgan Fingerprints (1024 bits, radius 2) were calculated using RDKit for all the compounds in the training set; (ii) a similarity score was then evaluated between each molecule in the training and test sets and the previously-defined fingerprints using the Tanimoto similarity index from RDKit; (iii) then the average similarity score was computed by averaging the similarity scores of the 5

most similar couple of compounds. The distribution of the average similarity scores for the training and test sets was used to identify a similarity threshold to discriminate between query compounds inside or outside the domain of applicability of the developed model. The AD check is performed every time before running the model to assess the reliability of the prediction and the output of the AD control is given to the user.

External Datasets

Several external datasets have been considered for testing the usability of the developed umami predictor. In particular, we chose some databases related to foods or natural products:

1. **FooDB** (https://foodb.ca/) is the world's largest and most comprehensive resource on food constituents, chemistry and biology (more than 70k compounds).

2. **FlavorDB** (https://cosylab.iiitd.edu.in/flavordb/) comprises 25595 flavour molecules. For the present work, we considered only 2939 molecules related to natural ingredients.

3. **PhenolExplorer** (http://phenol-explorer.eu) collects a comprehensive database of polyphenols contained in foods. We considered only compounds having composition data (SMILES), i.e. 489 compounds.

4. **Natural Product Atlas** (https://www.npatlas.org/) includes microbially-derived natural products published in peer-reviewed primary scientific literature. We downloaded 32552 natural compounds.

5. **PhytoHub** (https://phytohub.eu/) is a freely available electronic database containing detailed information about dietary phytochemicals and their human and animal metabolites. We downloaded 2110 compounds.

Each database was first checked for missing SMILES or data, standardised with the ChEMBL Structure Pipeline and, finally, the Mordred descriptors were calculated as done for the starting umami/non-umami dataset. Before running the model prediction, each dataset was screened to access the portion inside the model applicability domain and the prediction was then performed only in the above-mentioned portion.

3.2.3 Results

Dimensionality reduction

As described in the Methods section, the statistical analysis to reduce the number of employed molecular descriptors was performed on the training set with the

limma eBayes method [407]. Moreover, the correction of p-values for multiple testing to get q-values was applied using the Benjamini-Hochberg FDR adjustment method [408]. Setting the q-value threshold to 0.05, we identified 324 statistically significant differentiated features. This analysis is shown in Figure 3.1 in a volcano plot representation with the log2 of the Fold Change (log2FC) on the x-axis and the negative value of the logarithm of the p- or q-values on the y-axis. The log2FC was calculated for each feature by applying the log base 2 to the ratio between the average value of the feature for the umami class and the average of the non-umami class. P-values (Figure 3.1a) and q-values (Figure 3.1b) less than or equal to 0.05 denoted statistically significant differences between umami and non-umami samples, whereas positive log2FC values denote upregulated features, i..e features with higher values in umami than non-umami compounds, and negative log2FC values indicate downregulated features. In this view, the most informative features in the volcano plots are located at the top and far from the zero value of the x-axis. The detailed list of the prioritized molecular descriptors is available in the GitHub repository (https://github.com/lorenzopallante/VirtuousUmami) within the "data" folder ("umami_prioritized_list_of_descriptors.csv").

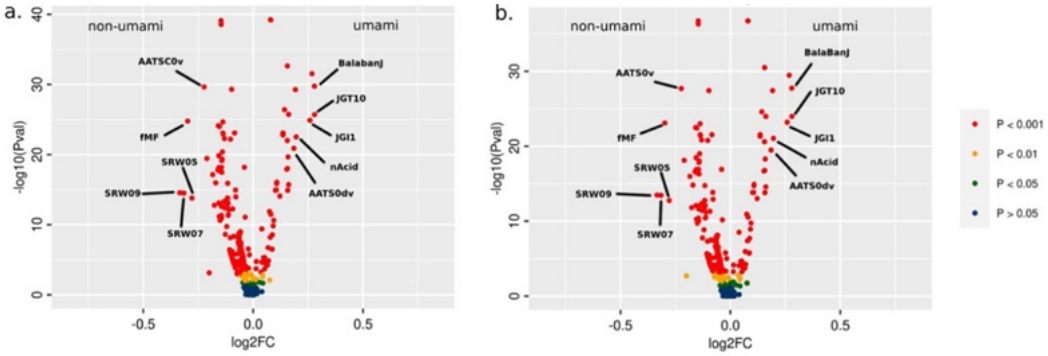


Figure 3.1. Volcano plots of the statistical analysis of the descriptors on the umami versus non-umami samples for the training set (a) with the standard limma eBayes method using p-values and (b) with correction of p-values using the Benjamini-Hochberg FDR adjustment method to calculate q-values. Only the 5 most upregulated and 5 most downregulated features are labelled for the sake of clarity.

Model performance

We developed 5 different SVM models with a specific number of selected features and support vectors (see also Table A - 6.6.3). After accessing the performance of the single SVM models (Table A - 6.6.4), we developed 10 ensemble models (EMs) by taking all the possible combinations between the SVM models (1 and 2; 1 and 3; 2 and 4; etc..) and evaluated the relative performance (Table A - 6.6.5). The EM_{3-5}, i.e. the ensemble model created combining SVM models 3 and 5, achieved the best performance and was selected as the final model. A summary of the model performance for the EM_{3-5} is reported in Table 3.6 and the relative ROC curves are represented in Figure 3.2.

Table 3.6. Summary of model performance using the ensemble model $EM_{3\text{-}5}$ obtained from the combination of SVM models 3 and 5. For the training set and the 10-fold cross-validation mean values and standard deviations are presented. The test set comprises the 90 left-out samples not used for training.

	ACC	***Spec***	***Sens***	***F1***	***F2***	***AUC***
10-fold CV	95.86%±1.89	96.70%±2.91	95.07%±1.06	95.73%±1.81	95.28%±0.88	0.96±0.02
Test	**87.64%**	**91.80%**	**78.57%**	**79.31%**	**80.99%**	**0.85**

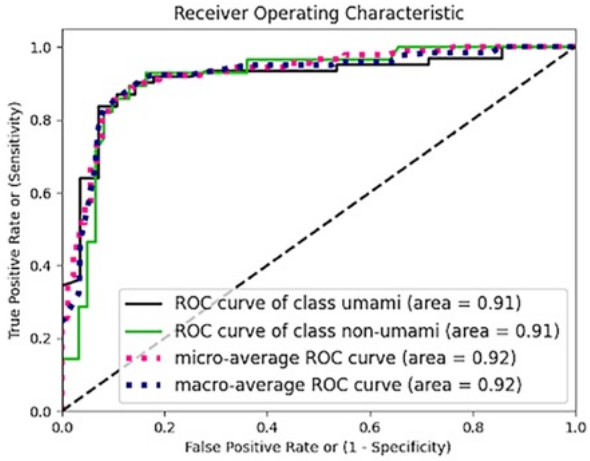


Figure 3.2. Receiver Operating Characteristic Curve of the umami versus non-umami classification.

Feature Importance

The selected features on which the predictions rely are 12 and include ATSC1m, Xch_6d, Mi, SaaCH, SMR_VSA1, JGI1, FilterItLogS, JGT10, AATSC0m, AATSC0v, Mp, fragCpx. The selected features are summarized in Table 3.7 also reporting the level of importance evaluated with the calculation of the SHAP values [413]. The distributions of the 12 features for the umami and non-umami samples are represented in Figure A - 6.6.1 and Figure A - 6.6.2.

Table 3.7. Features selected according to the best model. SHAP values represent the contribution of each feature to the prediction. The greater the value, the higher the contribution.

ID	**Name**	**Module Class**	**Description**	**SHAP importance**
1	ATSC1m	Autocorrelation	Centered moreau-broto autocorrelation of lag 1 weighted by mass	0.1090

2	AATSC0m	Autocorrelation	Averaged and centered moreau-broto autocorrelation of lag 0 weighted by mass	0.0821
3	AATSC0v	Autocorrelation	Averaged and centered moreau-broto autocorrelation of lag 0 weighted by vdw volume	0.0416
4	JGI1	TopologicalCharge	1-ordered mean topological charge	0.0331
5	JGT10	TopologicalCharge	10-ordered global topological charge	0.0323
6	SMR_VSA1	MoeType	MOE MR VSA Descriptor 1 (-inf < x < 1.29)	0.0296
7	Mi	Constitutional	Mean of constitutional weighted by ionization potential	0.0264
8	FilterItLogS	LogS	Filter-it™ LogS	0.0176
9	Mp	Constitutional	Mean of constitutional weighted by polarizability	0.0174
10	SaaCH	Estate	Sum of aaCH	0.0170
11	Xch-6d	Chi	6-ordered Chi chain weighted by sigma electrons	0.0122
12	fragCpx	FragmentComplexity	Fragment complexity	0.0083

Among the 12 selected features, the most frequent descriptor class represents internal autocorrelation properties (ATSC1m, AATSC0m, AATSC0v), calculated by the so-called Autocorrelation of a Topological Structure (ATS), which describes how a property is distributed along with the topological structure. In particular, the autocorrelation descriptors were computed using the Moreau-Broto autocorrelation weighted by mass (ATSC1m and AATSC0m) or Van der Waals volume (AATSC0v). Interestingly, the three autocorrelation properties were also retrieved among the first eight prioritized features from the initial univariate filtering. The Xch-6d descriptor belongs to the Chi descriptors family, which are topological indexes based on the molecular connectivity approach [414]. Molecular connectivity methods quantify molecular structures based on the topological and electronic characters of the atoms in the molecule. The molecule is represented by the hydrogen-suppressed graph (molecular skeleton) and the key feature in the quantitation of the graph is the characterization of the atom in the molecular

skeleton. The molecular graph may be decomposed into fragments called subgraphs, such as a skeletal bond, a pair of adjacent bonds, etc., that determine the possibility of defining different orders of the indexes: thus, the order of the Chi index is the number of edges in the corresponding subgraph. Mi and Mp are instead the mean of constitutional properties, i.e. the ionization potential and the polarizability, respectively. SaaCH descriptor is an Electropological State (Estate) index [415], which is a combination of electronic, topological and valance state information. In particular, this descriptor is calculated for specific atoms types: in this case, SaaCH stands for the sum of E-state indices for the CH in an aromatic ring. The SMR_VSA1 descriptor is a MOE type descriptor that uses a combination of the Wildman-Crippen Molar Refractivity (MR) [416], which is a measure of the total polarizability of a mole of a substance, and the Van der Waals surface area contribution. Two other descriptors, namely JGI1 and JGT10, deal instead with the compounds' topological charge considered at the first and 10th orders, respectively. FilterItLogS descriptor is derived from a program designed for filtering out molecules with unwanted properties. The program is packaged with several pre-programmed molecular properties that can be used for filtering, including (i) physicochemical parameters, such as logP, topological polar surface area criteria, number of hydrogen bond acceptors and donors, and Lipinski's rule-of-five; (ii) graph-based properties, including ring-based parameters and rotatable bond criteria; (iii) selection criteria through smarts patterns; (iv) Similarity criteria; (v) three-dimensional distances between user-definable fragments (https://github.com/silicos-it/filter-it). Finally, the fragCpx descriptor is a fragment complexity descriptor which is calculated as:

$$\text{fragCpx} = |B^2 - A^2 + A| + \frac{H}{100} \tag{3.1}$$

where A is the number of atoms, B is the number of bonds, and H is the number of heteroatoms [417].

Hierarchical clustering of the selected features allows grouping of the 12 features in three subgroups, i.e. (i) AATSC0v, ATSC1m, Mp, (ii) fragCpx, SMR_VSA1, AATSC0m, SaaCH, Xch-6d, (iii) JGI1, JGT10, Mi, FilterItLogS (see also Figure A - 6.6.3).

To represent the dataset's chemical space and underline the role of the feature importance analysis in simplifying the discrimination between the umami and non-umami, we used the tSNE dimensionality reduction technique [418] on the starting dataset taking into account all descriptors and only the best 12 above-mentioned features (Figure 3.3).

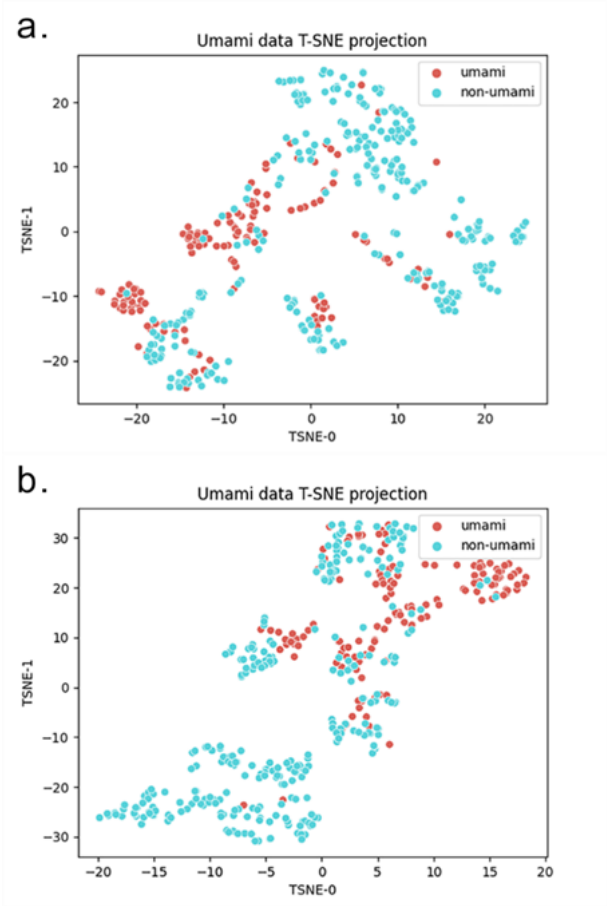


Figure 3.3. tSNE applied to the umami and non-umami samples for the whole dataset taking into account (a) all molecular descriptors (1613 features) and (b) the best 12 selected descriptors derived from the feature selection process. The selected feature subset (b) results in a remarkably better ability in discriminating between umami and non-umami compounds.

Applicability Domain (AD)

To effectively define the applicability domain (AD) of the model, we evaluated the average similarity scores of both training and test sets compared to the training sets fingerprints, as described in the Material and Methods section. The analysis reported in Figure 3.4 allowed us to establish a correct average similarity threshold (i.e. 0.4) to effectively determine if a query compound falls inside or outside the AD based on the average similarities of the employed dataset. In particular, if the average similarity score of a query compound is below the imposed threshold, then the query compound is considered outside the AD; otherwise, the compound is considered within the AD.

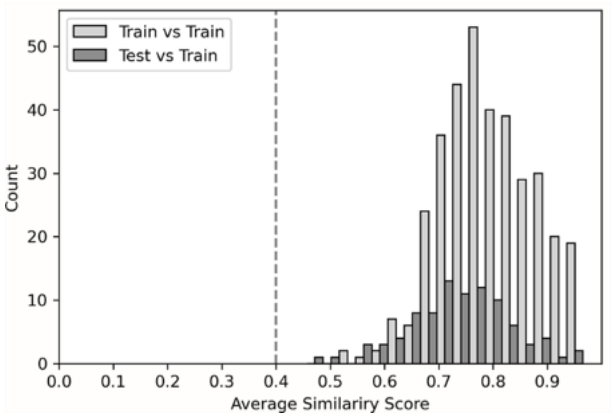


Figure 3.4. Histograms of average similarity scores of training and test sets. The average similarity score is derived by averaging the Tanimoto similarity score between the five most similar compounds in the training set. The light grey histogram represents the distribution of the average similarity scores for all the compounds composing the training set, whereas the dark grey histogram the distribution for the test set. The lower limit of the above-mentioned distributions allows for determining the similarity threshold of the applicability domain.

External datasets

The external datasets, i.e. FlavorDB, FooDB, NPAtlas, PhenolExplorer and PhytoHub, were processed as reported in the Materials and Methods section. Results are summarised in the following.

1. **FlavorDB**. After removing 380 compounds with issues from the ChEMBL structure pipeline, we got 2599 compounds. Checking the AD of the umami model, we pointed out that only 0.92% (24/2599) of the FlavorDB molecules are inside the umami AD. Our model predicted 9 of the 24 compounds (36%) as umami.

2. **FooDB**. Among the 70k chemicals included in the dataset, we preserved 69309 molecules after removing missing SMILES, duplicate compounds, and molecules with structure errors according to RDKit import functionality and high issues based on the ChEMBL Structure Pipeline. Only 1.09% (757/69309) of these molecules fall inside the AD of the model. 48% of these molecules (366/757) were then predicted as umami.

3. **Natural Product Atlas**. After running the ChEMBL structure pipeline, we preserved 32491 compounds. 1.52% (495/32491) of the molecules are inside the AD of the model and 17.3% of these molecules (86/495) were then predicted as umami.

4. **PhenolExplorer.** We first removed 3 compounds with issues according to the ChEMBL structure pipeline, obtaining 489 compounds. According to the AD, only 0.61% (3/489) of the PhenolExplorer molecules are inside the AD of the model. None of these molecules was predicted as umami.

5. **PhytoHub**. From the original dataset of 2110 compounds, we removed compounds with missing SMILES (294) or high issue scores from the ChEMBL structure pipeline (70), resulting in a database of 1746 molecules. Only a small percentage, i.e. 1.03% (18/1746), of the PhytoHub molecules are inside the applicability domain of the umami model. Just one molecule among the 18 compounds (5.5%) was predicted as umami.

Predicted umami compounds for each of the external DBs are available in the GitHub repository (https://github.com/lorenzopallante/VirtuousUmami) within the "data" folder.

Virtuous Umami Platform

The developed umami predictor was embedded into a web-based interface, namely the Virtuous Umami platform (https://virtuous.isi.gr/#/umami). This is a graphical, user-friendly interface for running analyses for chemical compounds expressed in various notations, including SMILES, FASTA format, InChI, SMARTS or PubChem compound name. If the PubChem name is provided by the user, the algorithm queries the database for the requested compound retrieving the relative canonical SMILES to run the umami prediction model. The platform is built using open-source programming solutions and is divided into two main components, i.e. the front-end and the back-end. The front-end is the part of the application that is visible to the users and runs on their devices. It provides them with the option to type compounds directly to an input field or to upload a text file containing each compound in a different line. After the analysis takes place, the results are presented in a tabular form that reports the query compound SMILES, its 2D molecular representation, the verification of the domain of applicability (True/False), the result of the umami prediction (Yes/No) and two buttons allowing the user to download the databases collecting all the calculated Mordred molecular descriptors or the best 12 on which the prediction relies. For developing the front-end, the Ionic framework was selected because it offers a wide variety of UI components that can be used to create user-friendly applications suitable both for browsers and mobile devices. The second main component, the backend, consists of a web service that runs on the cloud and is implemented using the lightweight yet powerful Flask micro-framework. It is responsible for receiving the input sent by the front-end, running the Virtuous Umami Analyser and returning the results to the front-end. To enable the aforementioned exchange of information, it provides a RESTful API that accepts and transmits data in the form of JavaScript Object Notation (JSON).

3.2.4 Discussion

Machine Learning methods have proven to play a key role in the development of prediction tools and digital support systems in a variety of application areas, including nutrition and agri-food research[419–426]. In this context, here, we developed

a novel machine-learning-driven umami taste predictor, named VirtuousUmami, to identify umami/non-umami compounds based on the SMILES representation. The classification model was generated with the hybrid combination of heuristic optimization and nonlinear machine learning classification methods, allowing both an unbiased and an optimized selection of the classification method and its parameters.

Starting from the UMP442 database [319], which collects 442 peptides, we used the Mordred molecular descriptors to obtain the features: the Mordred library is open source and demonstrated high computational efficiency and stability [358]. Moreover, we decided to only compute 2D molecular descriptors to avoid the impact of compound optimization and parameters related to the three-dimensional properties of molecules. The exhaustive list of the employed Mordred descriptors is available at https://mordred-descriptor.github.io/documentation/master/descriptors.html. The 2D Mordred descriptors provide information on compounds, such as basic information about molecules (molecular weight, number of individual types of atoms, types of bonds, degree of hybridization, spectral diameter, detour index, number of hydrogen donors and acceptors, molecular distance edge between different types of atoms, polarizability of atoms and bonds, and topological polar surface) and other features derived from symbolic representations (Zagreb index, adjacency matrix descriptors, Moreau–Mroto descriptors, Moran coefficients, Geary coefficients, and descriptors describing the Burden matrix and Barysz matrix) [427]. It is worth mentioning that other previous works successfully obtained good results in the field of taste prediction using only 2D molecular descriptors [387,402]: this represents a great step forward since 2D molecular descriptors are less expensive from a computational point of view and not affected by variations in the three-dimensional molecular structures.

Since the number of molecular descriptors (1613) was extremely higher than the number of compounds in the dataset (442), the limma eBayes statistical analysis was employed to reduce the total number of descriptors to 324, boosting the performance of the subsequent refined model. The best performance obtained from an ensemble of models in terms of accuracy (ACC), specificity (Spec), and sensitivity (Sens) scores are in good agreement with the state of the art [403,404]. In this context, to provide a comparison with previously developed umami prediction tools, iUmami-SCM[403] and UMPred-FRL[404] were assessed with the VirtuousUmami test set (Table A - 6.6.6). Comparing the evaluated metrics, the three algorithms showed overall similar performance, in terms of accuracy (ACC), specificity (Spec), sensitivity (Sens), F1 and F2 scores with all values roughly in the range of 80%-90%. Moreover, one of the major novelties of VirtuousUmami relies on its generalizability and applicability. In greater detail, its ability to process several types of molecular structure notations, including SMILES, FASTA, InChI, SMARTS or PubChem name allows screening for any type of compound, thus opening up the possibility to screen a wide range of molecular databases for

detecting umami compounds. In this context, we employed the VirtuousUmami predictor on five different external databases related to food or natural compounds, i.e. FlavorDB, FoodDB, Natural Product Atlas, PhenolExplorer and PhytoHub, highlighting compounds with umami character. Another important advantage of the proposed model relies on its explainability.

The usage of general molecular descriptors from the Mordred library and the employment of dimensionality reduction algorithms, such as statistical significance analysis and the SHAP feature importance, allowed the definition of a reduced number of interpretable features on which the model relies: in this case, the best model was able to achieve the above classification scores with only 12 features. Figure 3.3 graphically remarks on the importance of the feature selection procedure: the selected feature subset (Figure 3.3b) can discriminate remarkably better between umami and non-umami taste if compared to the tSNE analysis taking into account all the descriptors (Figure 3.3a). Despite the remarkable reduction in the number of features, it still remains complex to intuitively highlight the chemical and physical properties of umami/non-umami compounds related to the 12 most important features. In this sense, it will be very important in future studies to be able to use simpler descriptors in order to improve the explainability of the model. The definition of a small subset of important molecular features profoundly differentiates the approach proposed by previously developed methods, such as iUmami-SCM[403] and the UMPred-FRL[404], which based their predictive models only on the peptide sequences. While the possibility of optimising a predictive model on the peptide sequence alone is a great advantage in terms of model simplification, it also makes it very complicated to pinpoint the chemical-physical characteristics underlying molecules' properties and thus explain the model prediction coming from the machine learning black box.

Moreover, following the guidelines defined by the Organization for Economic Co-operation and Development (OECD) [377], we also developed an applicability domain (AD) to provide information regarding the reliability of the prediction. From this analysis, we pointed out that the distribution of the average similarities of training and test sets are similar in shape, denoting that the dataset is homogeneous and correctly repartitioned between training and test sets (Figure 3.4). The distribution of the average similarity scores towards elevated values suggests a high similarity among the compounds composing the dataset and, therefore, a quite narrow chemical space of the umami database. In this context, the development of an applicability score ensures reliable predictions for compounds within the above-mentioned domain. The above-mentioned limited spectrum is a direct consequence of the limited number of umami/non-umami compounds available from previous literature and composing our training dataset. In particular, the limited number of positive samples in the dataset (only 28 umami compounds in the test set and 112 in the training set) limits the accessible chemical space of the umami samples in the

training phase and the subsequent prediction ability of the model on the positive class, causing differences in the sensitivity scores in the test (78.6%) and training (roughly 95.1%) sets. In this case, the model sensitivity was particularly affected by the considerably few positive samples in the test set. The reduced number of compounds in the employed dataset, i.e. UMP442, is an important limitation of the present as well as previously developed umami predictors: likely, a larger size of the umami dataset will result in higher performance. Nevertheless, it is worth mentioning that the VirtuousUmami sensitivity (78.6%) is in the agreement or higher than the ones of UMPred-FRL[2] (78.6%) and iUmami-SCM[1] (71.4%) respectively, when tested against the VirutousUmami test set (see also Table A - 6.6.6). In conclusion, future extensions in available experimental data concerning umami/non-umami compounds will be pivotal to enlarging the investigated chemical space and the applicability of ML-driven methodologies, such as VirtuousUmami. Furthermore, the absence of non-peptide compounds within the validation dataset used to assess the model's performance represents a limitation in understanding the generalization capability of VirtuousUmami beyond peptides. In conclusion, future extensions in available experimental data concerning umami/non-umami compounds will be essential for enlarging the investigated chemical space to increase the predicting performance, extend the applicability and test the generalization ability of the VirtuousUmami tool.

Finally, the development of a user-friendly web interface (https://virtuous.isi.gr/#/umami) stems from the idea of making the umami prediction model usable even for users not experienced or familiar with the use of technical python codes (also available in the GitHub repository at https://github.com/lorenzopallante/VirtuousUmami).

In summary, VirtuousUmami will be a powerful tool to fast screen any compound database for the discovery of a wide range of candidate compounds with potential umami sensory properties. In a broader view, it is worth mentioning that the method developed within this work is fully generalizable to the prediction of other taste sensations since it is based on the SMILES format, a standard description and widely used by the scientific community: the present tool, therefore, lays the foundations for the creation of a general tool for the prediction of the five basic tastes.

3.3 VirtuousSweetBitter

The present section is based on the following scientific publication:

*Maroni, G., **Pallante, L.**, Di Benedetto, G., Deriu, M. A., Piga, D., & Grasso, G. (2022). **Informed classification of sweeteners/bitterants compounds via explainable machine learning**. Current Research in Food Science, 5, 2270–2280. https://doi.org/10.1016/j.crfs.2022.11.014.*

***Author's contribution to the publication:** Pallante L. worked on data preprocessing, developing a python library for preprocessing molecules and calculating molecular descriptors, and implementing the applicability domain. He also took part in the conception of the work, the analysis of the results, their critical discussion, the writing of the manuscript and its revisions.*

Similar to the umami taste in the previous section, we present here the development of a new ML-based predictor for the prediction of sweet and bitter tastes. Among all the taste perceptions, the dichotomy of sweet and bitter tastes has been the subject of several machine learning studies for classification purposes. While previous studies have provided accurate sweeteners/bitterants classifiers, there is ample scope to enhance these models by enriching the understanding of the molecular basis of bitter-sweet tastes. Towards these goals, our study focuses on the development and testing of several machine learning strategies coupled with the novel SHapley Additive exPlanations (SHAP) for a rational sweetness/bitterness classification. This allows the identification of the chemical descriptors of interest by allowing a more informed approach toward the rational design and screening of sweeteners/bitterants. To support future research in this field, we make all datasets and machine learning models publicly available and present an easy-to-use code for bitter-sweet taste prediction.

3.3.1 Introduction

Bitter and sweet tastes along with umami, saltiness, and acidity represent the fundamental taste senses[428], which are linked to specific biological and survival needs. For example, the bitter taste has evolved to protect organisms from the consumption of potentially poisonous substances, whereas the sweet taste is normally associated with the energetic and caloric content of foods. Both sweet and bitter molecules are recognized by G-protein coupled receptors (GPCR), but while taste receptors type 2 (TAS2Rs) are primarily responsible for detecting bitter tastants, the TAS1R2/TAS1R3 heterodimer belonging to class-C GPCR is known to be involved in the sensation of sweetness[429]. These receptors are located on apical membranes of taste receptor cells located in the taste buds. Human gustatory systems are characterized by the dichotomy between sweet and bitter tastes with an innate preference for sweet tastes and an aversion to bitter tastes. The sensation of bitter-sweet taste is an emerging property arising from complex molecular interactions of a compound with these receptors. Besides the oral cavity, taste

receptors are also present in other body parts such as the urethra [139], skin [132,133], brain [134], heart [137,138], and pancreas [135,136]. As well as their primary role in taste perception (in the oral cavity), such receptors are also implicated in diabetes and obesity by virtue of their roles in nutrient perception, glucose level maintenance, appetite regulation, as well as hormone release [143].

As a food additive for a long time, sweeteners have been widely used in the food industry [430]. There are lots of controversies and challenges relating to the sweetener industry in recent years, though improvements in technologies have greatly accelerated its development. When developing sweeteners, not only do they need to taste sweet, but they also need to have no harmful side effects, which increased the demand for the development of new sweeteners in the food industry. Within this framework, finding compounds with a pleasant gradient of bitter-sweet flavor may lead to the development of low-calorie sweeteners and bitter masking molecules.

The design and development pipeline of sweeteners usually follows the following pathway: extraction, separation, and identification of potential molecules from natural plants and synthesis. The previously-mentioned procedures are highly expensive and require complex chemical or biological characterization of the samples. Within this view, it is clear that computational prediction and simulation of potential compounds in the early stage could accelerate the design and development process of sweetener molecules[431].

In silico methodological approaches for the bitterant prediction include structure-based, ligand-based and machine-learning methods[432,433]; of particular interest for the present study are these latter approaches. Naive Bayes approach and circular fingerprint have been carried out in literature to classify bitterness by using a dataset of about 600 bitterants taken from a proprietary database and more than 10000 non-bitterants randomly selected from the MDL Drug Data Repository (MDDR)[434]. The model was characterized by accuracy, precision, specificity, and sensitivity of 88, 24, 89, and 72% respectively in the five-fold cross-validation. Although the previously mentioned study reports the first bitterant prediction algorithm based on a quite large dataset, the work didn't provide a prediction tool that can be used by users to test their molecules. Huang et al. addressed this issue by developing the first online toolkit of bitterness prediction called “BitterX”. The web application uses a Support Vector Machine (SVM) approach [435] on physicochemical descriptors [436]. In their study, the dataset is composed of 539 publicly available bitterants and 539 non-bitterants taken from the Available Chemicals Directory (ACD) database. The computational model offers remarkable accuracy and precision of more than 91% and sensitivity within the range of 91-94% on the test set. However, several small molecules considered as the non-bitterants are still not experimentally tested. The adaptive ensemble machine-learning method “Adaptive Boosting” (AdaBoost) was applied in another study to build a bitterness classifier called “BitterPredict”[437]. The model was trained on 12 basic physicochemical descriptors and 47 Schrödinger QikProp descriptors [437]. The BitterDB[438,439], in combination with the data from

Rojas et al.[440] was used to identify bitterants, while most of the non-bitterants (1,360 non-bitter flavours) were still hypothetical [441]. The prediction model gives the accuracy (83%), precision (66%), specificity (86%), and sensitivity (77%) on the test set. Recently, the consensus voting strategy based on multiple ML models has been used in literature to perform the bitterant classification task considering a dataset of experimentally confirmed bitterants and non-bitterants[441].

Regarding the bitter/sweet dichotomy and the in-silico taste prediction, three major examples have been recently published, i.e., BitterSweetForest[442], BitterSweet[443] and VirtualTaste[444]. BitterSweetForest and VirtualTaste are based on the random forest classification algorithm and Morgan molecular fingerprints, whereas BitterSweet is based on Dragon molecular descriptors and relies on the Adaboost method. BitterSweetForest was able to reach incredibly high predictive performance (e.g., AUROC of 0.98 both in cross-validation and external validation), but with a relatively low number of compounds in the dataset (517 artificial and natural sweet compounds and 685 bitter molecules). On the other hand, BitterSweet remarkably enlarged the bitter/sweet dataset collecting positive sets of 813 bitter and 1139 sweet molecules but achieving lower performance compared to BitterSweetForest. Virtual Taste extends the previous work of the same authors, BitterSweet Forest, with a richer dataset and develops three models based on the random forest algorithm, Morgan molecular fingerprints and different data sampling methods for bitter/non-bitter prediction (VirtualBitter model achieving AUROC values of cross-validation and external validation of 0.97 and 0.96, respectively), sweet/non-sweet prediction (VirtualSweet model achieving AUROC values of cross-validation and external validation of 0.97 and 0.96, respectively) and sour/non-sour prediction (VirtualSour model achieving AUROC values of cross-validation and external validation of 0.97 and 0.99, respectively). In VirtualTaste and BitterSweetForest the authors train different models on different families of descriptors so that the final subset of features is the one that provides the best performing model. Finally, to understand which features contribute the most to the change in the expected class a Bayesian-based feature analysis was employed in which the relative frequency of important features for each class was calculated taking the feature position and occurrence within the class and the relative feature frequency of that particular feature with respect to the other classes. In BitterSweet, on the other hand, a feature selection method and a feature compression method were compared, in the first relevant features for bitter-sweet prediction were identified using the Boruta algorithm[445], in the second Principal Component Analysis (PCA) was used to reduce the dimensionality of the feature space. Finally, to explain the most impacting features of the model, a global feature ranking based on random forest relative feature importance with a mean decrease in Gini impurity was used.

The present study focuses on the development and testing of several machine learning strategies for sweetness/bitterness classification starting from the

collection of compounds from several datasets available in the literature. Compound features were computed by using molecular descriptors from open-source libraries starting from the SMILES representations. The main contributions lie in the methods used for the feature selection and the interpretation of the resulting models. Previously discussed model explanation approaches based on random forest impurity-based feature importance provide only global interpretations in the form of feature relevance ranking for the model, furthermore they suffer from known disadvantages such as underestimation of the relative importance of features due to multicollinearity, and bias towards high cardinality features[446] . In this work, in order to improve the interpretability of the final model, we propose a method of sequential selection of relevant and uncorrelated or weakly correlated features based on hierarchical clustering on the feature's Spearman rank-order and two-sample Kolmogorov - Smirnov test. As a method of explanation, we propose to use the novel SHapley Additive exPlanations (SHAP) [447] approach which, in addition to having a solid mathematical background, provides a wider range of both global interpretation tools, such as feature importance graphs, summary and partial dependence plots, and local interpretation tools such as visualizations of the contribution of each single features in the bitter/sweet prediction of a single molecule in the dataset. This allows the identification of the chemical descriptors of interest by allowing a more informed approach to the design and screening of sweeteners/bitterants.

3.3.2 Materials and Methods

Database and data curation

The employed dataset collects compounds from several previous pieces of literature. In particular, we gathered compounds from (i) Biochemical Targets of Plant Bioactive Compounds by Gideon Polya [448], (ii) BitterDB [439], (iii) Fenaroli Handbook of Flavor Ingredient [449], (iv) DB by Rodgers et al. [434], (v) DB by Rojas et al. [450], (vi) SuperSweet [451], (vii) The Good Scents Company Database (http://www.thegoodscentscompany.com/), (viii) DB by Wiener et al. [437], (ix) SweetenersDB [452]. The resulted starting database collected a total of 3130 compounds (1764 sweet and 1366 bitter) with their SMILES description. We then checked all the SMILES using the RDKit library (http://www.rdkit.org), removing compounds with incorrect SMILES, searching for the relative correct SMILES in the PubChem database and removing duplicates. Then, the SMILES were processed with the ChEMBL Structure Pipeline [453] (https://github.com/chembl/ChEMBL_Structure_Pipeline) to highlight possible issues in the retrieved molecular structure and to standardize the SMILES representation for the entire dataset. The latter protocol runs a molecule checker on the compound structure, standardizes chemical structures and generates the parent molecule representation based on a set of predefined rules. At the end of this preprocessing pipeline, we obtain a final dataset of 2686 compounds (1415 sweet

and 1271 bitter). A summary of the final collected compounds from each of the above-mentioned databases is reported in Table A - 6.7.1. It is worth mentioning that a similar approach to dataset creation was adopted in previous literature [443]. We added some new compounds from new sources or updated version of the selected DBs. compared to the previous work, we increased the total number of compounds by 500, adding 168 sweet compounds and 355 bitter compounds.

Molecular descriptors

Starting from the SMILES representations, compound features were computed by using molecular descriptors from open-source libraries, i.e. RDkit (http://www.rdkit.org), pybel [454] and Mordred [455]. In detail, we decided to focus on 2D molecular descriptors, using 208 descriptors from RDKit, 25 from pybel and 1826 from Mordred, obtaining a total of 2059 molecular features per molecule. We focused our attention only on the 2D molecular descriptors to avoid the impact of compound optimization and parameters related to the three-dimensional properties of molecules. The 2D descriptors provide fundamental chemical information in terms of molecular weight, number of individual types of atoms, types of bonds, degree of hybridization, spectral diameter, detour index, number of hydrogen donors and acceptors, molecular distance edge between different types of atoms, the polarizability of atoms and bonds, and topological polar surface. Moreover, other features derived from a symbolic representation were also considered such as the Zagreb index, adjacency matrix descriptors, Moreau–Mroto descriptors, Moran coefficients, Geary coefficients, and descriptors describing the Burden matrix and Barysz matrix [427]. It is worth mentioning that other previous works successfully obtained good results in the field of taste prediction using only 2D molecular descriptors [431,443]: this represents a great step forward since 2D molecular descriptors are less expensive from a computational point of view and not affected by variations in the three-dimensional molecular structures. However, 2D descriptors are not able to catch variations in the molecular three-dimensional arrangements of investigated molecules. This could be potentially important in the bitter/sweet taste prediction field, since some compounds can elicit both taste sensations depending on modifications in their 3D structural properties, including isomerism [456 457,458 459–462]. Nevertheless, to avoid any possible misclassification for the above-mentioned type of compounds and employ only the 2D molecular descriptors, as also mentioned in the *Data Cleaning* section, we have not considered 70 compounds with identical 2D descriptors but different tastes. The inclusion of also 3D descriptors might be considered in the future to include compounds able to trigger sweet or bitter taste depending on their three-dimensional rearrangements.

Data cleaning

The resulting raw dataset, consisting of 2686 samples and 2060 columns (2059 features + 1 target column) was cleaned with the following procedures. First, 713 duplicate rows (or groups of more than 2 identical rows) were identified. 643 of

them had the target variable duplicated, while the remaining 70 had a different target variable. Of the former, only one sample per group of duplicate rows was kept in the dataset, while the latter were entirely removed from the dataset to avoid ambiguity. Afterwards, all columns with a percentage of missing values greater than or equal to 95% have been removed from the dataset along with all columns with zero or almost zero variance, i.e., constant or near-constant columns such that for 99% or more of the samples the same numerical value is present in the dataset. Finally, all the columns with duplicate values (or groups with more than 2 identical columns) have been collapsed into a single column to avoid redundancy. Bitter and sweet classes have been replaced with the numeric values of 0 and 1, respectively. The cleaned dataset was thus reduced to 2195 samples and 1403 columns (1402 numerical features + 1 binary target column).

Validation strategies and evaluation criteria

Stratified 5-fold cross-validation was used for training and hyper-parameter tuning. Stratification allows for the preservation of the classes' proportion in the created folds. Repeated 10-times 10-fold stratified cross-validation using different randomization of the data at each repetition was used for statistical comparison of modelling results and model selection. Models were evaluated using as primary evaluation criteria threshold-independent metrics such as Area Under Receiver Operating Characteristic Curve (AUROC) and Area Under Precision-Recall Curve (AUPRC), along with F1-score, Precision, and Recall.

Modelling

We tested:

- two conventional statistical approaches, namely, a parametric logistic regression model and a non-parametric k-nearest neighbours algorithm;
- two tree-based machine learning models, namely a random forest and a gradient boosting machine (LightGBM implementation);
- a deep learning model, i.e., multilayer perceptron (MLP).

A brief description of each model is provided in the following.

Logistic regression provides the probability of a certain class, where the log-odd is a linear combination of the input features. As a consequence, the class decision boundary is a linear function of the inputs. Linearity makes the estimation procedure simple and the results easy to understand and interpret. However, the correctness of the model depends on strong assumptions about the data including normality, independence, linearity and homoscedasticity. In a k-nearest neighbours classification algorithm, a new sample is assigned to the most common class among its k nearest neighbours. In random forest and gradient boosting machines, the prediction of the target variable is given as the result of an ensemble of weak models which are typically decision trees. A random forest fits several decision tree

classifiers on various sub-samples of the dataset in parallel and then combines the trained classifiers to improve predictive accuracy and control over-fitting. In a Gradient Boosting model, decision trees are trained consecutively in a forward stage-wise fashion, where each new tree is fitted to the predecessor's (pseudo) residual error, allowing sequential optimization of an arbitrary differentiable cost function through gradient descent. Artificial neural networks (ANNs) are computing systems characterized by elementary units (called neurons) interconnected through edges with adjustable weights. Such neurons are organized in layers that perform different types of mathematical transformations at their inputs. Typically, the weights of a neural network are adjusted through variants of the gradient descent algorithm, with gradients computed using the backpropagation algorithm. A multi-layer perceptron (MLP) is an ANN with multiple layers between the input and output layers. The MLP used in the study has a basic architecture of 2 fully connected layers with 100 neurons and ReLu activation functions. Adam optimizer[463] was used to optimize the weight parameters.

Training of all the models mentioned above was implemented in Python 3.9.7, with scikit-learn 1.0.1 and LightGBM 3.3.1 libraries.

Statistical analysis

To evaluate statistically significant differences between the performance of the models, we compared their AUROC scores by running a statistical test. To statistically compare the performance of a pair of models, we used the Nadeau and Bengio's corrected t-test[464]. This test takes into account the non-independence of the 100 AUROC scores of the individual models, obtained by evaluating the models on the same folds with repeated 10-times 10-fold stratified cross-validation. Finally, for pairwise comparison of all models, we ran the same statistical test multiple times by applying a Bonferroni correction to the computation of the p-values. The significance level was set to $p<0.05$.

Feature selection

Initially, the models are trained using all the 1402 input features, and the best learning algorithm is selected for further analysis. Indeed, the overall objective of this work is to build a model as accurate and interpretable as possible. Thus, it was necessary to select a small subset of features sufficiently informative to have an accuracy comparable to the one achieved by using all the 1402 input features. Furthermore, in order to increase interpretability and prevent underestimation of the relative importance of features due to multicollinearity[465] the selected features should be ideally uncorrelated or weakly correlated to each other. To achieve this, we used sequential feature selection combined with hierarchical clustering [466] on some features' correlation index. In this work, we used the Spearman rank-order index to take into account non-linear relationships between pairs of features. This allowed us to construct a small subset of uncorrelated or weakly correlated features

by choosing a given number of clusters and keeping a single feature from each cluster.

Different strategies can be used to select one representative feature from a particular cluster, either through automated methods or domain expert knowledge. In this work, we used an automated strategy. For each cluster, features have been ranked according to their univariate predictivity of the target variable and the most predictive one was picked. The predictivity of a feature was estimated with a two-sample Kolmogorov – Smirnov test[467] which empirically measures the distance between the two distribution functions of the considered feature, one referring to sweet and the other referring to bitter instances. The greater the distance between these two empirical distributions, the greater the probability that the sweet and bitter samples are drawn from different distributions, and the greater the univariate capability of the considered feature in predicting the target variable.

Feature importance analysis

To measure and rank the importance of each variable and explain their contribution to the individual predictions of the best performing model, we used SHAP (SHapley Additive exPlanations) values[447], a recent model-agnostic explanation methodology with a solid theoretical foundation and desirable properties. The SHAP explanation method computes Shapley values from the coalitional game theory conceptualized by the economist Lloyd Shapley, hence the name. The feature values of a sample act as players in a coalition and Shapley values tell us how to fairly distribute the resulted prediction among the features. An important feature is that the Shapley values are calculated as an addictive feature attribution method. For machine learning models, this means that SHAP values of all the input features will always sum up to the difference between baseline (expected) model output and the current model output for the prediction being explained. Furthermore, SHAP values are consistent, which means that features that are unambiguously more important are guaranteed to have a higher SHAP value. Operationally, for a single instance x, given a model f that outputs a prediction value $\hat{y}$, SHAP decomposes this prediction into the sum of a baseline value with the contributions that each feature has to the prediction, that is:

$$\hat{y} = y_{base} + \phi(x_1) + \phi(x_2) + \phi(x_3) + \cdots \quad (1)$$

where $y_{base} = E[f(X)]$ is the expected value of the predictions of all the training data X and $\phi(x_j)$ is the SHAP value corresponding to the j-th feature. In our study, positive SHAP values $\phi(x_j) > 0$ implies a positive contribution to the sweetness of the molecule, while negative SHAP values imply a positive contribution to the bitterness of the molecule. $|\phi(x_j)|$ gives the magnitude of the contribution. The specific formula for the calculation of $\phi(x_j)$ is given by the following expression:

$$\phi(x_j) = \sum_{S \subset N \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{j\}) - f_x(S)] \quad (2)$$

where N is the set of all input features with M its dimension, S is a subset of N of dimension $|S|$, $f_x(S) = E[\,f(X) \mid X^S = x^S\,]$ is the expected value of the predictions conditioned on the subset S of input features with known values x^S and $f_x(S \cup \{j\})$ is the same but with feature j added to subset S. Finally, the SHAP value for feature j is computed as a weighted average over all possible feature subsets S that don't include feature j already.

A comparison between the different models investigating the bitter/sweet dichotomy is reported in Table A - 6.7.2, highlighting the sources used for the construction of the dataset, the employed molecular descriptors for features computing, and the methods/approaches used for features selection, model building and model interpretation.

Applicability domain

In the current work, we developed an applicability domain (AD) to provide additional information about prediction reliability. An average-similarity approach already employed in previous recent literature in the taste prediction field [441,468] was considered. The AD was created considering a random 90:10 dataset partitioning into training and validation sets according to the 10-fold cross-validation employed in the model development. (i) the Morgan Fingerprints (1024 bits, radius 2) were calculated using RDKit for all the compounds in the dataset set; (ii) a similarity score was then evaluated between each molecule in the training and validation sets and the previously-defined fingerprints using the Tanimoto similarity index from RDKit; (iii) then the average similarity score was computed by averaging the similarity scores of the 5 most similar couple of compounds. The distribution of the average similarity scores for the training and validation sets was used to identify a similarity threshold to discriminate between query compounds inside or outside the domain of applicability.

3.3.3 Results and Discussion

Data pre-processing and missing values handling

Different strategies for data pre-processing and imputation of missing values were used according to the different learning models employed. For logistic regression, k-nearest neighbours, and multi-layer perceptron the outliers were treated with 90% winsorization, i.e., each variable was clipped at its 5th and 95th percentile, then each feature was scaled with min-max normalization. For gradient boosting and random forest, only 90% of winsorization was applied. The missing values were particularly severe for the molecular distance edge descriptors and the atom type e-

state descriptors, respectively 60.8% and 40.4% of missing values on average between the descriptors. These were generated due to chemical or structural characteristics of the molecule that makes the computation of a particular descriptor not possible, thus resulting in *missing not at random* (MNAR) values. For this reason, missing values have been imputed with a constant out-of-distribution value, namely: -1 for logistic regression, k-nearest neighbours and multi-layer perceptron; and -99999 for the random forest. For gradient boosting the missing values were automatically handled by the LightGBM implementation.

Model performances

The complete performances of the tested models, computed with repeated 10-times 10-fold stratified cross-validation and averaged on the folds, are summarized in Figure 3.5A-B, and the *receiver operating characteristic* (ROC) curve and *precision-recall* (PR) curve are shown in Figure 3.5C-D. Gradient boosting achieved an AUROC of 0.950 (95% CI [0.930, 0.970]); random forest achieved an AUROC of 0.942 (95% CI [0.916, 0.968]); MPL achieved an AUROC of 0.934 (95% CI [0.906, 0.962]); logistic regression and k-nearest neighbours classifier achieved an AUROC of 0.924 (95% CI [0.894, 0.954]) and (95% CI [0.880, 0.944]), respectively.

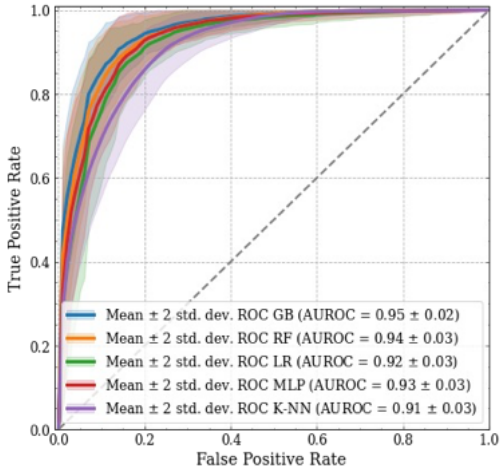
(A)

Model	AUROC	AUPRC	Precision	Recall	F1 score
Gradient boosting	**0.950**	**0.943**	**0.866**	**0.893**	**0.883**
Random forest	0.942	0.930	0.863	0.869	0.871
Logistic regression	0.924	0.905	0.838	0.877	0.860
Multi-layer perceptron	0.934	0.917	0.848	0.868	0.862
K-nearest neighbors	0.912	0.879	0.791	0.875	0.830

(B)

Model 1	Model 2	t-statistic	p-value
Gradient boosting	Random forest	3.095	0.013
Gradient boosting	Logistic regression	6.828	<0.001
Gradient boosting	Multi-layer perceptron	4.789	<0.001
Gradient boosting	K-nearest neighbors	8.602	<0.001
Random forest	Logistic regression	4.747	<0.001
Random forest	Multi-layer perceptron	2.394	0.093
Random forest	K-nearest neighbors	6.867	<0.001
Logistic regression	Multi-layer perceptron	-2.556	0.061
Logistic regression	K-nearest neighbors	2.527	0.065
Multi-layer perceptron	K-nearest neighbors	4.668	<0.001

(C)



(D)

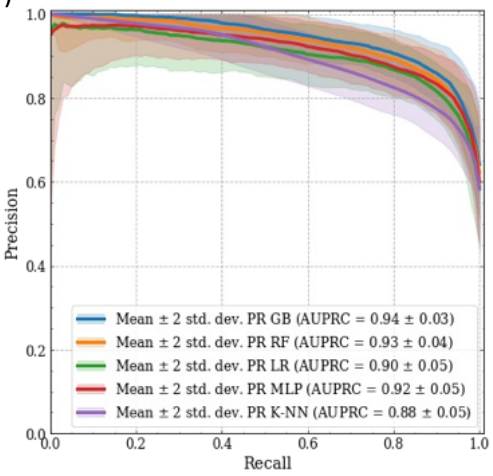


Figure 3.5. (A) Average model performance. (B) Pairwise comparison of all model performance with Nadeau and Bengio's corrected t-test and Bonferroni correction. (C) Solid lines and shaded areas represent the average receiver operating characteristics curves and their 95% confidence intervals. (D) Solid lines and shaded areas represent the average precision-recall curves and their 95% confidence intervals. Abbreviations: GB, gradient boosting, RF, random forest, LR, logistic regression, MLP, multi-layer perceptron, K-NN, k-nearest neighbours.

A direct comparison between our approach and the literature in this field is not completely fair, as performance evaluation is not performed on the same testing data. However, we provide in the following useful overview to contextualize the results achieved by our approach compared to the ones available in the literature

and to give an indication of the performance achievable for this type of classification problem. BitterSweetForest[442] achieved higher metrics (AUROC = 0.98, F1 = 0.92-0.95, ACC = 0.97), but with a remarkably lower number of samples in the database (517 artificial and natural sweet compounds and 685 bitter molecules), limiting the exploration of the bitter/sweet chemical space. BitterSweet[443] obtained different performances for the sweet/non-sweet and the bitter/non-bitter predictions. In particular, for the sweet/non-sweet prediction, BitterSweet achieved AUPRC of 0.93, AUROC of 0.85, F1 score of 0.77 and regarding the bitter/non-bitter prediction AUPRC of 0.93, AUROC of 0.88, F1 score of 0.86.

Feature selection

To select a small subset of uncorrelated or weakly correlated informative features, we used sequential feature selection combined with hierarchical clustering on the feature's Spearman rank-order correlations, as described in the following steps.

i. First, the feature correlation matrix was constructed using Spearman rank-order correlations and, for each feature, the predictive capacity of the target variable was estimated through a two-sample Kolmogorov – Smirnov test. In Figure 2A (first line), the variables piPC4 (conventional bond order ID number of order 4), GATS1d (Geary autocorrelation coefficient of lag 1 weighted by sigma electrons) and MPC5 (molecular path count of order 5) are shown, characterized by high values of the Kolmogorov – Smirnov statistic and high separation between the empirical distributions of samples with sweet target and samples with the bitter target. The second line of the same figure shows the variables CIC1 (1-ordered complementary information content), MATS7are (Moran autocorrelation coefficient of lag 7 weighted), and AATSC7s (Broto autocorrelation of lag 7 weighted by intrinsic state) characterized by low Kolmogorov – Smirnov statistic values and high overlap between the empirical distributions of sweet and bitter samples. The variable with the highest estimated predictive capacity (piPC4) was selected and used to train a LightGBM model. The resulting performances were computed with 5-fold cross-validation and stored.

ii. After converting the correlation matrix to a distance matrix, hierarchical clustering using Ward's linkage was performed and two clusters were selected. From these, the 2 most representative features were picked based on their estimated predictive capacity and used to train a LightGBM model and compute cross-validation performances, together with the intra-cluster mean absolute correlation of the features.

iii. The process described in step (ii) is repeated for 3, 4, … clusters, until each cluster is atomic i.e., it contained a single feature.

The results of this procedure are shown in Figure 3.6B, where we can observe that with a limited number of features it is still possible to approach the performance of the reference model trained with the entire set of features, which reinforces the fact that groups of features are redundant.

Finally, we have arbitrarily chosen 29 features as a good compromise between model performance (AUROC = 0.944), simplicity and interpretability. Figure A - 6.7.1 shows the correlation matrix of the selected features and the absolute values of the feature's Spearman rank-order correlations with an average absolute correlation between the variables of 0.19. This shows that the implemented procedure allowed us to develop a model with weakly correlated features as inputs.

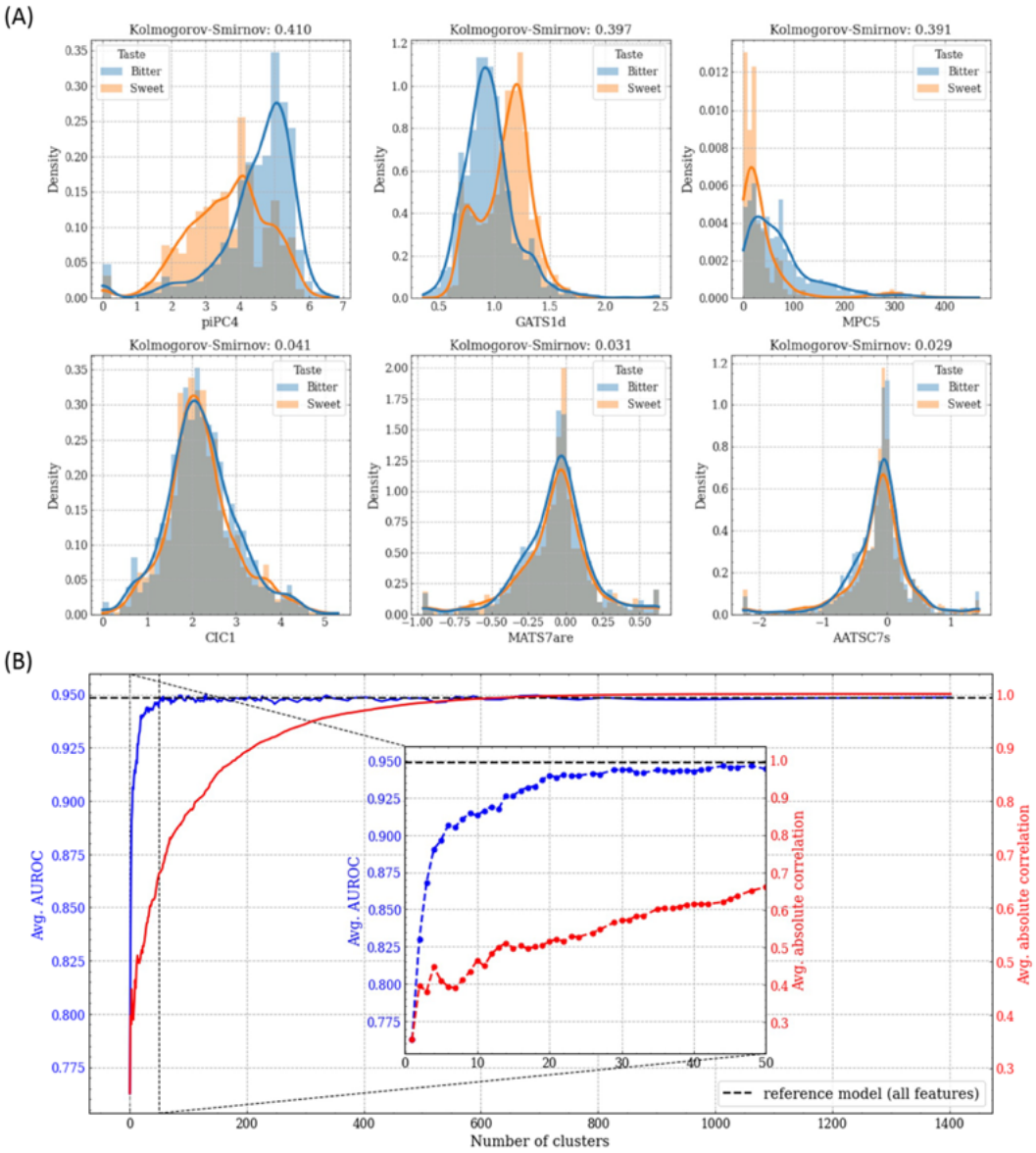


Figure 3.6. (A) Kernel density estimation of the sweet vs bitter molecules empirical distributions for features with high Kolmogorov – Smirnov statistic (first row) and low Kolmogorov – Smirnov statistic (last row). (B) Feature selection algorithm results. Average AUROC values (blue left y-axis) and average absolute intra-cluster correlation (red right y-axis) as the number of clusters increases. The zoom represents the progress of the algorithm until the first 50 clusters are reached.

Global interpretation

The SHAP explanation method aims to explain the prediction of a single instance by estimating, for each feature, its contribution to the prediction, called SHAP value (in this study the prediction associated with a molecule corresponds to the probability predicted by the model that that molecule is sweet). By combining SHAP values computed for each sample of the dataset, we obtain a matrix with one

row per sample and one column per feature. From the analysis of this matrix, it is possible to obtain global explanations of the entire model. The SHAP feature importance bar plot shown in Figure 3.7A reports the features in descending order of importance computed as the average across the data of the absolute SHAP values. BCUTi-1h (first highest eigenvalue of Burden matrix weighted by ionization potential) and MINdO (the minimum value of the atom type E-state descriptor[469] linked to the presence of the atom group double bonded with Oxygen) have been identified as the most impacting features, followed by ATSC5c (centred Moreau-Broto autocorrelation of lag 5 weighted by gasteiger charge), MATS2s (Moran autocorrelation coefficient of lag 2 weighted by intrinsic state), MINssO (the minimum value of the atom type E-state descriptor[469] linked to the presence of the -O- atom group, MDEC- 13 (molecular distance edge between all primary and tertiary carbons), MPC5 (molecular path count of order 5), GATS2v (Geary autocorrelation of lag 2 weighted by van der Waals volumes) and GATS1d (Geary autocorrelation coefficient of lag 1 weighted by sigma electrons). All other features were considered less impacting on predictions. In the SHAP summary plot of Figure 3.7B, in which each sample is depicted as a point where the position on the x-axis represents the impact on the prediction in the form of SHAP value and the colour represents the intensity (blue for low values to red for high values) of the value assumed by a feature, feature importance is combined with the directional relationship between values assumed by a feature and impact on predictions. Among the most impacting variables, ATSC5c, MATS2s and GATS2v are positively correlated with the sweetness of a molecule, while BCUTi-1h and MINdO are positively correlated with the bitterness of a molecule.

The empirical form of the relationship between feature values and impact on model predictions can be studied for each feature with the SHAP dependence plots, where each data instance is represented by a point with a position on the x-axis the value assumed by the feature and the position on the y-axis the corresponding Shapley value. The SHAP dependence plots for the 4 most representative features are presented in Figure 3.8.

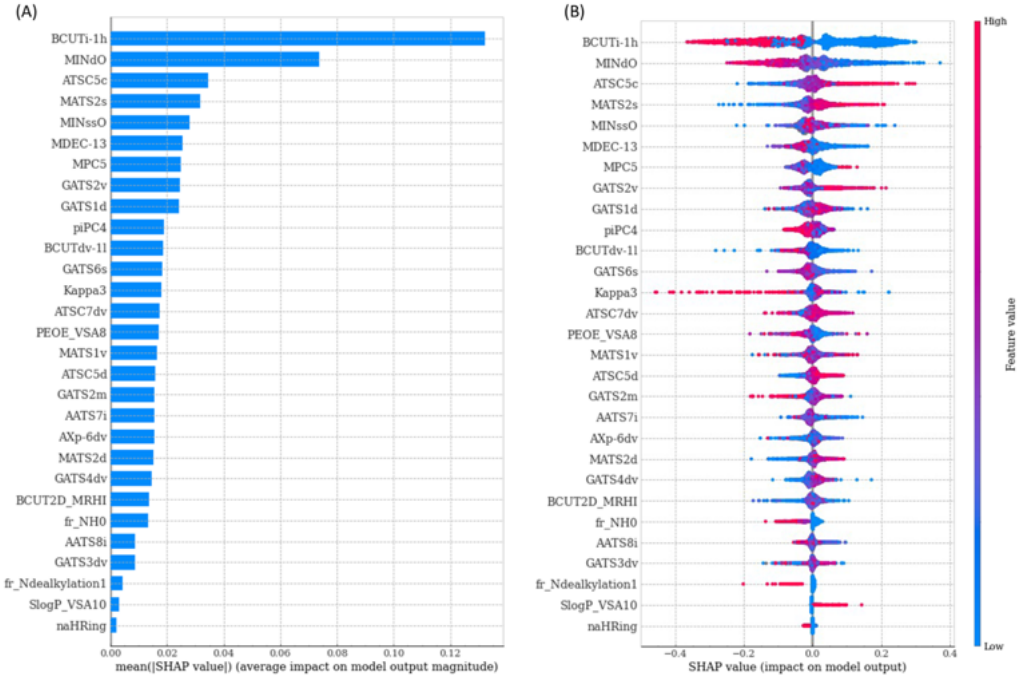


Figure 3.7. SHAP feature importance plots. (A) The left bar plot represents a ranking of the importance of the variables with their average impact on model prediction. (B) The right dot plot represents each data point with the signed contribution of each variable to the model prediction: blue colour indicates low values for a variable whereas red colour indicates high values.

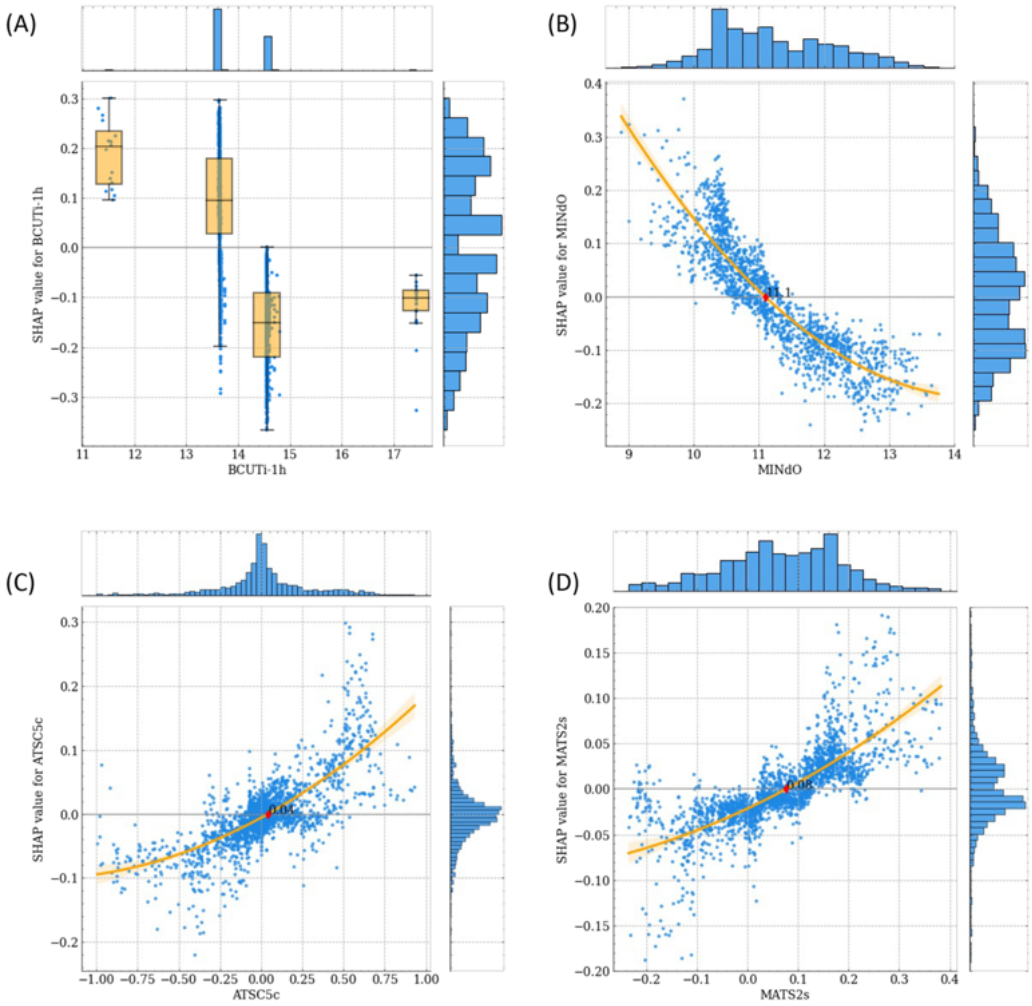


Figure 3.8. SHAP dependence plots of the 4 most representative features. (A) BCUTi-1h, (B) MINdO, (C) ATSC5c, (D) MATS2s. For discrete and mixed variables, values are plotted with a scatter plot and box plots with whiskers enclosing points belonging to different levels (A). For continuous variables, values are plotted with a scatter plot and an orange regression line with shaded 95% confidence intervals (B, C, D). A red diamond marks a cut-off point of the feature. Empirical distributions of feature and SHAP values are represented with histograms on the top and right of each plot.

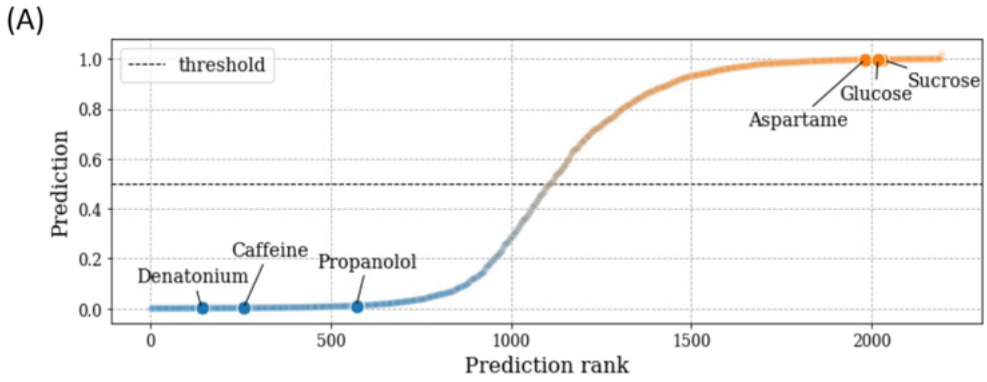
Local Interpretation

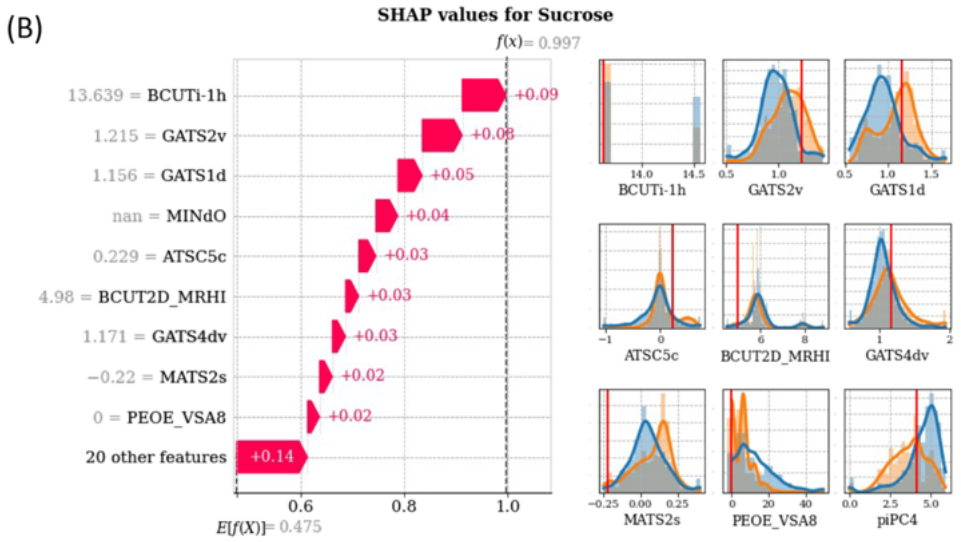
Finally, in this paragraph, we report a local interpretative analysis of the final model using as case studies six representative molecules (Figure 3.9A and Figure A - 6.7.3):

- thre sweet molecules, i.e., Sucrose, Glucose, and Aspartame;
- three bitter molecules, i.e., Propanolol, Caffein, and Denatonium.

Figure 3.9A shows the out-of-sample predictions of the entire dataset obtained in cross-validation and ordered according to the prediction ranking. The considered four reference molecules are highlighted in the plot with their sweet/bitter target correctly predicted by the model.

The SHAP profiles of the representative molecules are shown in the left panel of Figure 3.9B-C and Figure A - 6.7.3A-D. The most impacting features for the prediction are shown on the y-axis and the corresponding SHAP values are displayed through coloured arrows with their cumulative value reported on the x-axis. Positive SHAP values, represented with red arrows, indicate a positive contribution to the predicted sweetness of the molecule, while negative SHAP values, represented with blue arrows, indicate a positive contribution to the predicted bitterness of the molecule. Empirical distributions of the most impacting features are reported in the right panels of Figure 3.9B-C and Figure A - 6.7.3A-D. The orange colour distributions correspond to the sweet molecules in the dataset, while the blue ones correspond to the bitter molecules. The vertical solid red lines highlight the value assumed by the feature in the corresponding molecule. If the value is missing, the feature is skipped. For these molecules, the most impacting features contribute with the same sign to the prediction. Moreover, the most impacting feature is unanimously BCUTi-1h (first highest eigenvalue of Burden matrix weighted by ionization potential). For bitter molecules, other common impacting features are MINdO (the minimum value of the atom type E-state descriptor[469] linked to the presence of the atom group double bonded with Oxygen) and MPC5 (molecular path count of order 5), while for sweet molecules they are GATS2v (Geary autocorrelation of lag 2 weighted by van der Waals volumes) and GATS1d (Geary autocorrelation coefficient of lag 1 weighted by sigma electrons).


(A)
threshold
Prediction
Prediction rank
Sucrose
Glucose
Aspartame
Caffeine
Propanolol
Denatonium


(B)
SHAP values for Sucrose
f(x)= 0.997
13.639 = BCUTi-1h
1.215 = GATS2v
1.156 = GATS1d
nan = MINdO
0.229 = ATSC5c
4.98 = BCUT2D_MRHI
1.171 = GATS4dv
−0.22 = MATS2s
0 = PEOE_VSA8
20 other features
E[f(X)]= 0.475
BCUTi-1h
GATS2v
GATS1d
ATSC5c
BCUT2D_MRHI
GATS4dv
MATS2s
PEOE_VSA8
piPC4

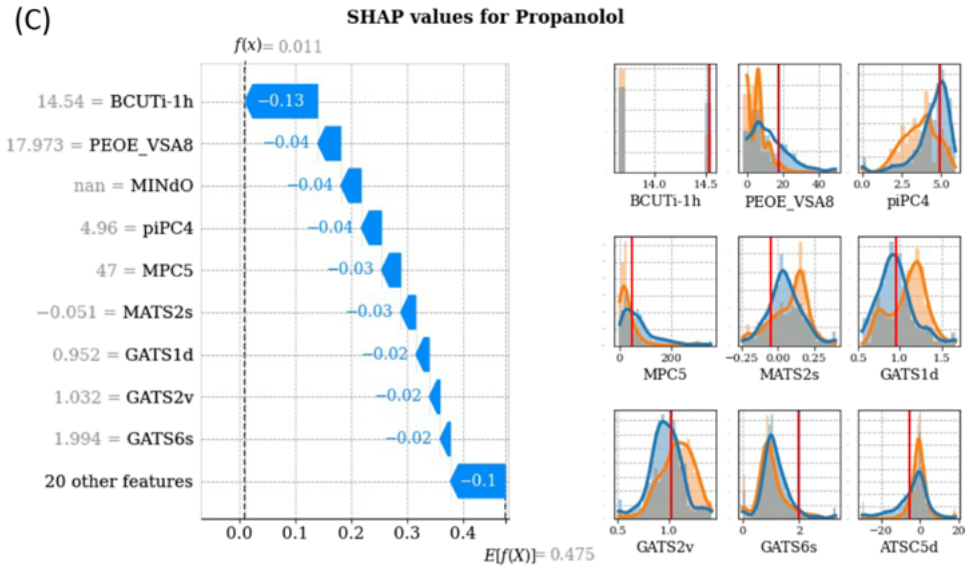

(C)
SHAP values for Propanolol
f(x)= 0.011
14.54 = BCUTi-1h
17.973 = PEOE_VSA8
nan = MINdO
4.96 = piPC4
47 = MPC5
−0.051 = MATS2s
0.952 = GATS1d
1.032 = GATS2v
1.994 = GATS6s
20 other features
E[f(X)]= 0.475
BCUTi-1h
PEOE_VSA8
piPC4
MPC5
MATS2s
GATS1d
GATS2v
GATS6s
ATSC5d

Figure 3.9 Prediction rank for the molecules of the entire dataset (x-axis) vs out-of-sample predicted sweetness probability (y-axis). Reference molecule prediction are highlighted. SHAP profiles of two representative molecules: Sucrose (B) and Propanolol (C). For each figure, SHAP values are shown in the left panel and impacting feature distributions in the right panel, with values assumed by the features highlighted with solid red lines.

3.3.4 Conclusion

The sweet-bitter dichotomy is an extremely fascinating aspect of taste perception: while the sweet taste is commonly associated with a pleasant sensation linked to the energetic content of foods, bitter is a complex control system normally related to the ability to avoid toxic or possibly harmful substances. In this work, we have further investigated this attractive mechanism to shed light on the molecular features determining the taste of a specific molecule. Therefore, we developed a machine-learning-based classifier able to discriminate between the bitter and sweet tastes of a query compound based on its molecular structure. The implemented tool is based on the widely used SMILES representation and employs open-source molecular descriptors to calculate the features on which the model relies. Thanks to statistical analysis methods, feature selection and analysis techniques, we were able to pinpoint a reduced number of molecular features determining the bitter or sweet taste and, together with the SHAP explainability method, we underlined the impact of the selected features, providing an informed and interpretable classification. In the process of designing new molecules, it is difficult to make use of the selected features as they are not intuitive. This issue, however, is not related to the selection of features, but rather to the use of molecular features in 2D. This point could be adequately addressed not only by simplifying the input molecular features, which will inevitably reduce algorithm performance, but also by taking advantage of a number of scientific studies focused on machine learning decoders able to reconstruct the chemical information starting from the 2D features of the molecule. Additionally, a generative model could be added to the computational pipeline to suggest appropriate chemical changes to achieve the desired taste. Addressing the previously-mentioned challenges represent the future development of this work, and we hope that our study will provide a starting point for potential studies in this field. The developed model will therefore pave the way toward the rational design and screening of sweet/bitter molecules through the molecular understanding of the physical and chemical characteristics underlying the perception of these tastes. To ensure the reproducibility of the results and to allow the usage of the developed model, we publicly release the Python scripts, along with the employed datasets and supplementary material on GitHub (https://github.com/gabribg88/VirtuousSweetBitter). The sweet/bitter classifier will be also implemented into a user-friendly webserver to allow its usage even to non-expert or technical users. In a broader view, this tool will be integrated into the framework of an EU-funded project, named VIRTUOUS ([64]), which aims at creating an intelligent computational platform by integrating molecular modelling

methods, drug discovery techniques, machine learning classifiers, algorithms for big data, cloud computing, and experimental data to predict the organoleptic profile of selected types of food based on their chemical composition. In conclusion, the present work represents a crucial starting point in the definition of a *virtual tongue* able to predict the taste of specific ingredients and general compounds with the ultimate goal of shedding light on the mechanisms and hidden relationships at the basis of the taste perception process.

Chapter IV

Conclusions

Taste perception is a complex experience that involves the gustatory, olfactory, and trigeminal systems, and serves to regulate food intake by assessing its nutritional value and potential harm. The gustatory system recognizes the five primary tastes - sweet, umami, bitter, sour, and salty - which each have specific functions. Taste perception is a multi-level process that involves molecular, subcellular, cellular, and tissue-level actors within the gustatory system. At the molecular level, taste perception is initiated by the interaction between chemical substances from ingested foods and specific proteins known as taste receptors, which are located on the gustatory papillae of the tongue. Each taste type has its signal transduction pathway that is mediated by taste receptors, leading to the activation of taste receptor cells. Investigating the molecular mechanisms of taste perception is crucial in illuminating the complex interplay between food uptake and intake and developing strategies to optimize nutrition and health outcomes.

This vision is embraced by the EU-funded project, VIRTUOUS (https://virtuoush2020.com/), within which this doctoral thesis is embedded. The primary goal of this project is the development of a virtual tongue as a comprehensive computational framework to screen selected natural compounds and food ingredients from the Mediterranean diet, such as olive oil or wine, for their ability to target taste receptors. The proposed platform should be able to integrate drug discovery techniques, machine learning classifiers, algorithms for big data, cloud computing, and experimental data to predict the organoleptic profile of a given food type based on its chemical composition. Through this project, a greater understanding of the mechanisms driving the transfer of information from the molecular level, where food constituents bind to taste receptors, to the cascade of

molecular, supramolecular, and cellular events that lead to an elaborated sensation contributing to the food's organoleptic profile will be achieved.

The present PhD Thesis is focused on the investigation of taste perception from its molecular-level perspective. The primary methodologies considered and employed were molecular modelling and machine learning. Molecular modelling was utilized to investigate the interactions between tastants or small natural compounds found in the diet with various taste receptors and off-target proteins. On the other hand, machine learning-based methodologies were considered to predict the taste of a molecule only based on its chemical structure. The scientific core of the present thesis consists of Chapter II and Chapter III which are respectively dedicated to the discussion and application of molecular modelling and machine learning-based methodologies to the above-mentioned scopes. In the following, a summary of the main results and conclusions from each chapters' sections, together with a critical discussion of the overall achievements, is provided.

In *2.1 - Molecular basis of taste perception*, we reviewed the major scientific advances in the molecular modelling of taste receptors and their interactions with specific tastants. This section focuses on the primary candidates discussed in the literature for taste receptors, including GPCRs for sweet, umami, and bitter, OTOP1 for sour, and ENaC for salty. However, these receptors cover a limited range of possible receptors, transducers, and proteins essential to the taste perception process. The presence of other key players and the identification of other possible basic tastes suggest that the understanding of taste perception is still incomplete and lacking, and further research is needed to achieve granular and comprehensive knowledge. Currently, computational and/or combined computational/experimental studies focusing on the structure-to-function relationships and ligand-protein binding investigations provide the main findings on taste receptor function. The need for developing high-quality molecular structures is a crucial step in molecular modelling, and the section described the most recent experimentally solved and in silico-derived structures for each taste receptor candidate. Among the players of taste transduction, only a few structures have been experimentally solved and most of the computational works rely on models coming from computationally-predicted structures. This section remarked that understanding the molecular behaviour and activity modulation of taste receptors is a crucial scientific challenge in research on the complex mechanisms that lead to the emergence of a taste sensation at the supramolecular, cellular, and tissue levels. In this context, this section provided pieces of evidence that computational molecular modelling is a powerful tool due to its atomistic resolution and enables the exploration of receptor structure-to-function relationships and ligand roles in taste receptor activity. This type of investigation allows for the quantitative characterization of the ligand-binding process, thermodynamics and kinetics of the binding mechanism, binding modes, and ligand-target interaction

properties, among others. Ligand-receptor binding investigations can evaluate food molecular constituents in terms of specificity, selectivity, multi-target features, and the natural role of taste receptors in discriminating between healthy and dangerous foods. Despite significant progress in molecular research and computational investigation of ligand-receptor interactions related to taste receptors, the scientific knowledge remains rather incomplete and unable to explain the mechanisms holistically. Therefore, it remains crucial to accurately frame the transfer of taste information from the chemistry level, where food molecular constituents bind to taste receptors, to molecular-, supramolecular-, and cellular-level events that ultimately contribute to the composite perception strongly linked to the food's organoleptic profile.

In *2.2 - VirtuousPocketome,* we were interested in understanding which are the taste receptors' residues needed for the effective tastant recognition and to understand if other proteins outside the gustatory systems share some similarities and might be considered as possible secondary target for food tastants. In this section, we therefore presented a novel computational pipeline, named VirtuousPocketome, to screen the human-solved proteome for similar binding sites to a query protein-ligand complex of interest. We applied the proposed framework to a recently solved human bitter taste receptor, namely the TAS2R46, complexed with a bitter taste compound, i.e. the strychnine. Starting from different molecular dynamics simulations, VirtuousPocketome was able to identify crucial important residues in the binding pocket of the bitter taste receptor needed for strychnine binding. The retrieved amino acidic patterns were used for the subsequent screening for similar proteins of the entire solved human proteome, consisting in 58972 structures at the time of work development, finally selecting 145 protein hits. The functional enrichment analysis allowed us to pinpoint the main biological processes, molecular functions, and cellular components related to the gene expressing the retrieved hit proteins. Most of the highlighted biological processes are connected to metabolic processes, which is an intriguing finding given the well-established relationship between taste perception, food intake, and metabolism. These results suggest that strychnine may not only activate the TAS2R46 bitter taste receptor and induce a bitter taste sensation, but it may also have the potential to trigger or modulate proteins that are directly involved in metabolic processes. Furthermore, the identified protein hits are mainly associated with molecular functions related to protein or small compound binding and are predominantly located in the cytoplasm and membrane. These findings are reasonable considering that TAS2R46 is a transmembrane protein and a broad bitter taste receptor that can bind a wide range of chemical compounds. Notably, it is experimentally known that strychnine can bind other membrane proteins besides the bitter one and exhibit also other types of activity besides the gustatory sensation, suggesting a reasonable functioning of the proposed pipeline. Further analysis of the retrieved proteins could involve predicting whether interaction with the compound of interest results in their

activation or modulation. This would enhance our understanding of potential secondary effects of tastants beyond their primary taste perception, and how they can impact the biological processes and molecular functions that the retrieved hit proteins are involved in. This approach can also aid in the design of tailored foods and ingredients to create personalized treatments that can target specific proteins or receptors involved in particular processes or diseases.

In *2.3 - The Impact of Natural Compounds on S-Shaped Aβ42 Fibril*, we used molecular modelling to evaluate the impact of a set of natural compounds on some off-targets not directly involved in the taste perception process. In detail, we assessed the role and the molecular mode of action of 57 natural compounds on the structural stability of S-shaped Aβ42 amyloid fibrils. Five ligands, namely 6-shogaol, oleuropein, curcumin, gossypin, and piceatannol, demonstrated significant destabilizing activity on the Aβ42 S-shape polymorphism, with two distinct destabilizing modes of action. This type of investigation highlighted the main common chemical features to be considered in the rational design of specific, naturally inspired compounds for targeting and destabilizing amyloid aggregates.

Section *3.1 - Machine learning for Taste Prediction* summarised the main scientific works in taste prediction using machine learning algorithms. The available databases containing food-related compounds and molecules with known taste, as well as the main tools employed to predict the taste, are discussed, highlighting specific sources for the different basic tastes. We insisted on the necessity for developing complete databases comprising all the relevant information for each entry (SMILES, InChI, IUPAC nomenclature, etc.) to avoid any possible error in compound processing. A correct specification of the molecular descriptors to be used would also require the development of extensive databases due to the large number and variety of both open-source and proprietary descriptors. We found out that most of the taste predictors developed in recent years were designed specifically for bitter and sweet tastes and interestingly some pieces of literature worked on the bitter/sweet dichotomy by developing prediction tools able to classify both sensations. On the other hand, only a few and limited attempts were made to predict the umami taste and no specific tools for the prediction of sour and salty tastes have been retrieved. Within this section, we have highlighted how the development of taste prediction tools is still incomplete and yet unstructured: at present, there is no tool capable of predicting all five main taste sensations, and few attempts are publicly accessible and usable by a broad spectrum of users, and the possibility of developing solutions capable of indicating the physio-chemical characteristics underlying a specific taste sensation remains an unsolved issue.

In section *3.2 - VirtuousUmami*, a novel machine learning-based umami taste predictor is presented. The developed tool, named VirtuousUmami, can predict the umami taste of a query compound based on its chemical structure. The molecular

structures are featurized using 2D molecular descriptors using an open-source program. The final classification model was created using a hybrid of nonlinear machine learning and heuristic optimisation techniques, enabling both an unbiased and optimised choice of the classification technique and its parameters. VirtuousUmami generalizability and applicability are one of its main novelties. More specifically, the developed tool can screen for any sort of chemical and can analyse a variety of molecular structure notations, including SMILES, FASTA, InChI, SMARTS, or PubChem name. This makes it possible to check a variety of molecular databases for umami compounds, such as the ones screened in this section (FlavorDB, FoodDB, Natural Product Atlas, PhenolExplorer, and PhytoHub), that are connected to food or natural chemicals. The explainability of the suggested paradigm is another key benefit: the application of dimensionality reduction strategies, such as statistical significance analysis, and the use of SHAP feature importance highlighted the most important molecular features on which the model relies, opening the way towards an interpretable model. However, the complexity of some of the molecular descriptors used still makes the model not easy to understand and future work will be needed to optimise models based on easily interpretable features to facilitate the definition of the chemical characteristics underlying umami perception. In conclusion, VirtuousUmami will be an effective tool for quickly screening any database of chemical compounds for the identification of a variety of candidate compounds with possible umami sensory qualities. In a broader sense, it is important to note that the method created in this work is completely generalizable to the prediction of other taste sensations because it is based on the SMILES format, a widely accepted molecular description in the scientific community. The current tool, therefore, lays the groundwork for the creation of a general tool for the prediction of the five basic tastes.

Section *3.3 - VirtuousSweetBitter* described the development of an ML-based tool to predict the sweet or bitter tastes, which are commonly associated with opposite sensations, i.e. a pleasant sensation linked to the energetic content of foods and an unpleasant taste related to the toxic or harmful substances, respectively. Similar to the previous section, we used open-source molecular descriptors to translate the molecular structure into a machine learning-readable form and statistical analysis coupled with the SHAP feature importance method to pinpoint the most important and informative features. The final model, optimised accessing the performance of several machine learning-based algorithms (logistic regression, non-parametric k-nearest neighbours' algorithm, random forest, gradient boosting machine, multilayer perceptron), relies on only 29 molecular descriptors and achieved performance in line with similar tools in the literature.

The two above-mentioned tools, i.e. VirtuousUmami and VirtuousSweetBitter, have been conceived with a very similar structure to be easily integrated in the future into a unified tool able to predict the three taste sensations (sweet, bitter, and umami) at the same time. In this context, the recent release of a new database of

compounds collecting sweet, bitter, umami, sour, salty and tasteless compounds opens the way to the possibility of developing a novel ML-based tool able to predict the taste of a query compound or tastant among the five basic taste sensations [470]. Moreover, VirtuousUmami and VirtuousSweetBitter have been already embedded into a user-friendly webserver interface (https://virtuous.isi.gr/) to allow their use by the public. The platform was designed according to the VIRTUOUS project goals, with particular attention to its usability and accessibility to a wide spectrum of users. The platform is currently composed of three main tools: (i) VirtuousFoods (https://virtuous.isi.gr/#/foods), which allows for a dynamic exploration of the FooDB (https://foodb.ca/) giving the possibility of predicting the taste of all its compounds using the Virtuous taste prediction tools; (ii) VirtuousUmami (https://virtuous.isi.gr/#/umami) and (iii) VirtuousSweetBitter (https://virtuous.isi.gr/#/sweetbitter), which can be executed analysing single entries typed directly on the web interface or a text file submitted by the user. It is worth mentioning that VirtuousPocketome described in 2.2 will be inserted as a fourth tool inside the platform, allowing its use by a wide range of users even non-experts in the field.

In summary, the present PhD thesis deals with the molecular-level perspective of taste perception. Molecular modelling has been considered and exploited to explore the structure and dynamics of taste receptors (section 2.1), characterise their specific ligand-receptor interactions and retrieve off-targets sharing similar patterns in the binding sites (section 2.2), and evaluate the impact of specific small molecules on the structure of proteins not directly involved in taste perception (section 2.3). Conversely, machine learning methods have been applied in this field to develop novel classifiers able to predict the taste of a query molecule from its molecular structure (sections 3.2 and 3.3). This level of investigation allowed us to reveal some of the molecular features and modes of actions enabling food tastants to target taste receptors and exhibit a specific taste sensation. Despite the remarkable scientific advances made in recent years in the field of taste perception, a lot of work is still needed to fill the gaps in our knowledge related to the molecular mechanisms underlying taste perception.

Therefore, the present PhD thesis represents only the first step toward a comprehensive and granular understanding of taste perception from its molecular perspective. In particular, we are currently working on the molecular modelling and dynamics of all the main taste receptor candidates and on the definition of a more general and interpretable taste predictor able to discriminate among the five basic taste sensations and give information relating to the level of perception for a specific taste. These future steps will be included in the holistic view of the VIRTUOUS project with the ultimate goal of making its platform increasingly precise and granular. Increasing the understanding of the molecular basis of taste perception will help in determining the individual's food preferences and improving the intake of specific nutrients by altering the taste of specific foods consequently. The

possibility to design ingredients with specific tastes might also impact the health outcomes, for example, designing novel sweeteners with less caloric content, developing enriched foods with specific functional or nutritional content, or coupling pharmacological treatments with a personalised diet according to user demand. It is therefore totally reasonable that the molecular investigation of taste perception should play a fundamental role in future years in the fields of nutrition, precision medicine, the food market and beyond.

Chapter V

References

1. Roper, S. D. Taste: Mammalian Taste Bud Physiology. in *Reference Module in Neuroscience and Biobehavioral Psychology* 887–893 (Elsevier, 2017). doi:10.1016/B978-0-12-809324-5.02908-4.

2. Munger, S. D. & Zufall, F. *Chemosensory Transduction - The Detection of Odors, Tastes, and Other Chemostimuli*. (2016).

3. Barbosa, N. S. V., Lima, E. R. de A. & Tavares, F. W. Molecular Modeling in Chemical Engineering. in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Elsevier, 2017). doi:10.1016/B978-0-12-409547-2.13915-0.

4. Forrest, L. R., Tang, C. L. & Honig, B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.* **91**, 508–517 (2006).

5. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).

6. Chow, E., Klepeis, J. L., Rendleman, C. A., Dror, R. O. & Shaw, D. E. 9.6 New Technologies for Molecular Dynamics Simulations. in *Comprehensive Biophysics* 86–104 (Elsevier, 2012). doi:10.1016/B978-0-12-374920-8.00908-5.

7. van Gunsteren, W. F. *et al.* Biomolecular Modeling: Goals, Problems, Perspectives. *Angew. Chem. Int. Ed.* **45**, 4064–4092 (2006).

8. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **41**, 429–452 (2012).

9. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).

10. Ben Shoshan-Galeczki, Y. & Niv, M. Y. Structure-based screening for discovery of sweet compounds. *Food Chem.* **315**, 126286 (2020).

11. Di Pizio, A. *et al.* Ligand binding modes from low resolution GPCR models and mutagenesis: Chicken bitter taste receptor as a test-case. *Sci. Rep.* **7**, 1–11 (2017).

12. Kooistra, A. J. *et al.* GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.* 1–9 (2020) doi:10.1093/nar/gkaa1080.

13. Olivella, M., Gonzalez, A., Pardo, L. & Deupi, X. Relation between sequence and structure in membrane proteins. *Bioinformatics* **29**, 1589–1592 (2013).

14. Chen, K.-Y. M., Sun, J., Salvo, J. S., Baker, D. & Barth, P. High-Resolution Modeling of Transmembrane Helical Protein Structures from Distant Homologues. *PLoS Comput. Biol.* **10**, e1003636 (2014).

15. Esguerra, M., Siretskiy, A., Bello, X., Sallander, J. & Gutiérrez-de-Terán, H. GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors. *Nucleic Acids Res.* **44**, W455–W462 (2016).

16. Dagan-Wiener, A. *et al.* BitterDB: taste ligands and receptors database in 2019. *Nucleic Acids Res.* **47**, D1179–D1185 (2019).

17. Ahmed, J. *et al.* SuperSweet-A resource on natural and artificial sweetening agents. *Nucleic Acids Res.* **39**, 377–382 (2011).

18. Chéron, J.-B., Casciuc, I., Golebiowski, J., Antonczak, S. & Fiorucci, S. Sweetness prediction of natural compounds. *Food Chem.* **221**, 1421–1425 (2017).

19. Töle, J. C., Behrens, M. & Meyerhof, W. Taste receptor function. in 173–185 (2019). doi:10.1016/B978-0-444-63855-7.00011-3.

20. Chandrashekar, J. *et al.* The cells and peripheral representation of sodium taste in mice. *Nature* **464**, 297–301 (2010).

21. Roper, S. D. & Chaudhari, N. Taste buds: cells, signals and synapses. *Nat. Rev. Neurosci.* **18**, 485–497 (2017).

22. Vandenbeuch, A. *et al.* Role of the ectonucleotidase NTPDase2 in taste bud function. *Proc. Natl. Acad. Sci.* **110**, 14789–14794 (2013).

23. Bachmanov, A. *et al.* Genetics of Taste Receptors. *Curr. Pharm. Des.* **20**, 2669–2683 (2014).

24. Di Pizio, A., Ben Shoshan-Galeczki, Y., Hayes, J. E. & Niv, M. Y. Bitter and sweet tasting molecules: It's complicated. *Neurosci. Lett.* **700**, 56–63 (2019).

25. Masuda, K. *et al.* Characterization of the Modes of Binding between Human Sweet Taste Receptor and Low-Molecular-Weight Sweet Compounds. *PLoS ONE* **7**, e35380 (2012).

26. Gravina, S. A., Yep, G. L. & Khan, M. Human biology of taste. *Ann. Saudi Med.* (2013) doi:10.5144/0256-4947.2013.217.

27. 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. *Proc. Natl. Acad. Sci.* **114**, 2568–2573 (2017).

28. Nie, Y., Vigues, S., Hobbs, J. R., Conn, G. L. & Munger, S. D. Distinct Contributions of T1R2 and T1R3 Taste Receptor Subunits to the Detection of Sweet Stimuli. *Curr. Biol.* **15**, 1948–1952 (2005).

29. Servant, G. *et al.* Positive allosteric modulators of the human sweet taste receptor enhance sweet taste. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4746–4751 (2010).

30. Zhang, F. *et al.* Molecular mechanism of the sweet taste enhancers. *Proc. Natl. Acad. Sci.* **107**, 4752–4757 (2010).

31. Jiang, P. *et al.* The Cysteine-rich Region of T1R3 Determines Responses to Intensely Sweet Proteins. *J. Biol. Chem.* **279**, 45068–45075 (2004).

32. Ohta, K., Masuda, T., Tani, F. & Kitabatake, N. The cysteine-rich domain of human T1R3 is necessary for the interaction between human T1R2–T1R3 sweet receptors and a sweet-tasting protein, thaumatin. *Biochem. Biophys. Res. Commun.* **406**, 435–438 (2011).

33. Masuda, T. *et al.* Five amino acid residues in cysteine-rich domain of human T1R3 were involved in the response for sweet-tasting protein, thaumatin. *Biochimie* **95**, 1502–1505 (2013).

34. Assadi-Porter, F. M. *et al.* Key Amino Acid Residues Involved in Multi-Point Binding Interactions between Brazzein, a Sweet Protein, and the T1R2–T1R3 Human Sweet Receptor. *J. Mol. Biol.* **398**, 584–599 (2010).

35. Kunishima, N. *et al.* Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* **407**, 971–977 (2000).

36. Temussi, P. A. Why are sweet proteins sweet? Interaction of brazzein, monellin and thaumatin with the T1R2-T1R3 receptor. *FEBS Lett.* **526**, 1–4 (2002).

37. Shrivastav, A. & Srivastava, S. Human sweet taste receptor: Complete structure prediction and evaluation. *Int. J. Chem. Anal. Sci.* **4**, 24–32 (2013).

38. Šali, A. *et al.* Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

39. Maillet, E. L. *et al.* Characterization of the binding site of aspartame in the human sweet taste receptor. *Chem. Senses* **40**, 577–586 (2015).

40. Chéron, J.-B., Golebiowski, J., Antonczak, S. & Fiorucci, S. The anatomy of mammalian sweet taste receptors. *Proteins Struct. Funct. Bioinforma.* **85**, 332–341 (2017).

41. Nuemket, N. *et al.* Structural basis for perception of diverse chemical substances by T1r taste receptors. *Nat. Commun.* **8**, 15530 (2017).

42. Kashani-Amin, E., Sakhteman, A., Larijani, B. & Ebrahim-Habibi, A. Introducing a New Model of Sweet Taste Receptor, a Class C G-protein Coupled Receptor (C GPCR). *Cell Biochem. Biophys.* 227–243 (2019) doi:10.1007/s12013-019-00872-7.

43. Perez-Aguilar, J. M., Kang, S., Zhang, L. & Zhou, R. Modeling and Structural Characterization of the Sweet Taste Receptor Heterodimer. *ACS Chem. Neurosci.* **10**, 4579–4592 (2019).

44. Koehl, A. *et al.* Structural insights into the activation of metabotropic glutamate receptors. *Nature* **566**, 79–84 (2019).

45. Ling, S. *et al.* Structural mechanism of cooperative activation of the human calcium-sensing receptor by Ca2+ ions and L-tryptophan. *Cell Res.* **31**, 383–394 (2021).

46. Xue, L. *et al.* Major ligand-induced rearrangement of the heptahelical domain interface in a GPCR dimer. *Nat. Chem. Biol.* **11**, 134–140 (2015).

47. Vafabakhsh, R., Levitz, J. & Isacoff, E. Y. Conformational dynamics of a class C G-protein-coupled receptor. *Nature* **524**, 497–501 (2015).

48. Pin, J. P. & Bettler, B. Organization and functions of mGlu and GABA B receptor complexes. *Nature* **540**, 60–68 (2016).

49. Liu, B. *et al.* Molecular Mechanism of Species-Dependent Sweet Taste toward Artificial Sweeteners. *J. Neurosci.* **31**, 11070–11076 (2011).

50. Xu, H. *et al.* Different functional roles of T1R subunits in the heteromeric taste receptors. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14258–14263 (2004).

51. Acher, F. C., Selvam, C., Pin, J.-P., Goudet, C. & Bertrand, H.-O. A critical pocket close to the glutamate binding site of mGlu receptors opens new possibilities for agonist design. *Neuropharmacology* **60**, 102–107 (2011).

52. Yamada, K. *et al.* Unnatural Tripeptides as Potent Positive Allosteric Modulators of T1R2/T1R3. *ACS Med. Chem. Lett.* **10**, 800–805 (2019).

53. Jiang, P. *et al.* Lactisole Interacts with the Transmembrane Domains of Human T1R3 to Inhibit Sweet Taste. *J. Biol. Chem.* **280**, 15238–15246 (2005).

54. Jiang, P. *et al.* Identification of the Cyclamate Interaction Site within the Transmembrane Domain of the Human Sweet Taste Receptor Subunit T1R3. *J. Biol. Chem.* **280**, 34296–34305 (2005).

55. Chéron, J.-B. *et al.* Conserved Residues Control the T1R3-Specific Allosteric Signaling Pathway of the Mammalian Sweet-Taste Receptor. *Chem. Senses* **44**, 303–310 (2019).

56. Winnig, M., Bufe, B., Kratochwil, N. A., Slack, J. P. & Meyerhof, W. The binding site for neohesperidin dihydrochalcone at the human sweet taste receptor. *BMC Struct. Biol.* **7**, 1–12 (2007).

57. Nakagita, T. *et al.* Structural insights into the differences among lactisole derivatives in inhibitory mechanisms against the human sweet taste receptor. *PLOS ONE* **14**, e0213552 (2019).

58. Zhao, M., Xu, X.-Q., Meng, X.-Y. & Liu, B. The Heptahelical Domain of the Sweet Taste Receptor T1R2 Is a New Allosteric Binding Site for the Sweet Taste Modulator Amiloride That Modulates Sweet Taste in a Species-Dependent Manner. *J. Mol. Neurosci.* **66**, 207–213 (2018).

59. Koizumi, A. *et al.* Human sweet taste receptor mediates acid-induced sweetness of miraculin. *Proc. Natl. Acad. Sci.* **108**, 16819–16824 (2011).

60. Ikeda, K. New Seasonings. *Chem. Senses* **27**, 847–849 (2002).

61. Zhang, J., Sun-Waterhouse, D., Su, G. & Zhao, M. New insight into umami receptor, umami/umami-enhancing peptides and their derivatives: A review. *Trends Food Sci. Technol.* **88**, 429–438 (2019).

62. Spaggiari, G., Di Pizio, A. & Cozzini, P. Sweet, umami and bitter taste receptors: State of the art of in silico molecular modeling approaches. *Trends Food Sci. Technol.* **96**, 21–29 (2020).

63. Zhang, F. *et al.* Molecular mechanism for the umami taste synergism. *Proc. Natl. Acad. Sci.* **105**, 20930–20934 (2008).

64. Liu, H., Da, L.-T. & Liu, Y. Understanding the molecular mechanism of umami recognition by T1R1-T1R3 using molecular dynamics simulations. *Biochem. Biophys. Res. Commun.* **514**, 967–973 (2019).

65. Geng, Y. *et al.* Structural mechanism of ligand activation in human calcium-sensing receptor. *eLife* **5**, 1–25 (2016).

66. Bystrova, M. F., Romanov, R. A., Rogachevskaja, O. A., Churbanov, G. D. & Kolesnikov, S. S. Functional expression of the extracellular-Ca 2+ -sensing receptor in mouse taste cells. *J. Cell Sci.* **123**, 972–982 (2010).

67. Fábián, T. K., Beck, A., Fejérdy, P., Hermann, P. & Fábián, G. Molecular mechanisms of taste recognition: Considerations about the role of saliva. *Int. J. Mol. Sci.* **16**, 5945–5974 (2015).

68. Toda, Y. *et al.* Positive/Negative Allosteric Modulation Switching in an Umami Taste Receptor (T1R1/T1R3) by a Natural Flavor Compound, Methional. *Sci. Rep.* **8**, 11796 (2018).

69. López Cascales, J. J., Oliveira Costa, S. D., de Groot, B. L. & Walters, D. E. Binding of glutamate to the umami receptor. *Biophys. Chem.* **152**, 139–144 (2010).

70. Chandrashekar, J. *et al.* T2Rs Function as Bitter Taste Receptors. *Cell* **100**, 703–711 (2000).

71. Brockhoff, A., Behrens, M., Niv, M. Y. & Meyerhof, W. Structural requirements of bitter taste receptor activation. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11110–11115 (2010).

72. Zhang, H. *et al.* Structure of the full-length glucagon class B G-protein-coupled receptor. *Nature* **546**, 259–264 (2017).

73. Di Pizio, A. *et al.* Comparing Class A GPCRs to bitter taste receptors. in *Biophysical Methods in Cell Biology* vol. 132 401–427 (Elsevier Ltd, 2016).

74. Meyerhof, W. *et al.* The molecular receptive ranges of human TAS2R bitter taste receptors. *Chem. Senses* **35**, 157–170 (2009).

75. Behrens, M. & Meyerhof, W. Vertebrate Bitter Taste Receptors: Keys for Survival in Changing Environments. *J. Agric. Food Chem.* **66**, 2204–2213 (2018).

76. Kuhn, C., Bufe, B., Batram, C. & Meyerhof, W. Oligomerization of TAS2R Bitter Taste Receptors. *Chem. Senses* **35**, 395–406 (2010).

77. Behrens, M. & Meyerhof, W. Bitter taste receptor research comes of age: From characterization to modulation of TAS2Rs. *Semin. Cell Dev. Biol.* **24**, 215–221 (2013).

78. Deshpande, D. A. *et al.* Bitter taste receptors on airway smooth muscle bronchodilate by localized calcium signaling and reverse obstruction. *Nat. Med.* **16**, 1299–1304 (2010).

79. Avau, B. *et al.* Targeting extra-oral bitter taste receptors modulates gastrointestinal motility with effects on satiation. *Sci. Rep.* **5**, 1–12 (2015).

80. Zhai, K. *et al.* Activation of bitter taste receptors (tas2rs) relaxes detrusor smooth muscle and suppresses overactive bladder symptoms. *Oncotarget* **7**, 21156–21167 (2016).

81. Kelm, S., Shi, J. & Deane, C. M. MEDELLER: Homology-based coordinate generation for membrane proteins. *Bioinformatics* **26**, 2833–2840 (2010).

82. Pydi, S. P. *et al.* Amino Acid Derivatives as Bitter Taste Receptor (T2R) Blockers. *J. Biol. Chem.* **289**, 25054–25066 (2014).

83. Wang, Y. *et al.* Metal Ions Activate the Human Taste Receptor TAS2R7. *Chem. Senses* **44**, 339–347 (2019).

84. Acevedo, W., González-Nilo, F. & Agosin, E. Docking and Molecular Dynamics of Steviol Glycoside–Human Bitter Receptor Interactions. *J. Agric. Food Chem.* **64**, 7585–7596 (2016).

85. Chen, Z. *et al.* Insights into the binding of agonist and antagonist to TAS2R16 receptor: a molecular simulation study. *Mol. Simul.* **44**, 322–329 (2018).

86. Soares, S. *et al.* Human Bitter Taste Receptors Are Activated by Different Classes of Polyphenols. *J. Agric. Food Chem.* **66**, 8814–8823 (2018).

87. Mayank & Jaitak, V. Interaction model of steviol glycosides from Stevia rebaudiana (Bertoni) with sweet taste receptors: A computational approach. *Phytochemistry* **116**, 12–20 (2015).

88. Di Pizio, A. & Niv, M. Y. Promiscuity and selectivity of bitter molecules and their receptors. *Bioorg. Med. Chem.* **23**, 4082–4091 (2015).

89. Levit, A., Beuming, T., Krilov, G., Sherman, W. & Niv, M. Y. Predicting GPCR Promiscuity Using Binding Site Features. *J. Chem. Inf. Model.* **54**, 184–194 (2014).

90. Born, S., Levit, A., Niv, M. Y., Meyerhof, W. & Behrens, M. The human bitter taste receptor TAS2R10 is tailored to accommodate numerous diverse ligands. *J. Neurosci.* **33**, 201–213 (2013).

91. Slack, J. P. *et al.* Modulation of Bitter Taste Perception by a Small Molecule hTAS2R Antagonist. *Curr. Biol.* **20**, 1104–1109 (2010).

92. Sakurai, T. *et al.* Characterization of the β-D-Glucopyranoside binding site of the human bitter taste receptor hTAS2R16. *J. Biol. Chem.* **285**, 28373–28378 (2010).

93. Nowak, S. *et al.* Reengineering the ligand sensitivity of the broadly tuned human bitter taste receptor TAS2R14. *Biochim. Biophys. Acta - Gen. Subj.* **1862**, 2162–2173 (2018).

94. Karaman, R. *et al.* Probing the Binding Pocket of the Broadly Tuned Human Bitter Taste Receptor TAS2R14 by Chemical Modification of Cognate Agonists. *Chem. Biol. Drug Des.* 66–75 (2016) doi:10.1111/cbdd.12734.

95. Liu, K., Jaggupilli, A., Premnath, D. & Chelikani, P. Plasticity of the ligand binding pocket in the bitter taste receptor T2R7. *Biochim. Biophys. Acta BBA - Biomembr.* **1860**, 991–999 (2018).

96. Sandal, M. *et al.* Evidence for a Transient Additional Ligand Binding Site in the TAS2R46 Bitter Taste Receptor. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00472.

97. Pydi, S. P. *et al.* Cholesterol modulates bitter taste receptor function. *Biochim. Biophys. Acta - Biomembr.* **1858**, 2081–2087 (2016).

98. Ishimaru, Y. *et al.* Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. *Proc. Natl. Acad. Sci.* **103**, 12569–12574 (2006).

99. Huang, A. L. *et al.* The cells and logic for mammalian sour taste detection. *Nature* **442**, 934–938 (2006).

100. LopezJimenez, N. D. *et al.* Two members of the TRPP family of ion channels, Pkd1l3 and Pkd2l1, are co-expressed in a subset of taste receptor cells. *J. Neurochem.* **98**, 68–77 (2006).

101. Horio, N. *et al.* Sour taste responses in mice lacking pkd channels. *PLoS ONE* **6**, 1–10 (2011).

102. Ye, W. *et al.* The K + channel K IR 2.1 functions in tandem with proton influx to mediate sour taste transduction. *Proc. Natl. Acad. Sci.* **113**, E229–E238 (2016).

103. Tu, Y.-H. *et al.* An evolutionarily conserved gene family encodes proton-selective ion channels. *Science* **359**, 1047–1050 (2018).

104. Teng, B. *et al.* Cellular and Neural Responses to Sour Stimuli Require the Proton Channel Otop1. *Curr. Biol.* **29**, 3647-3656.e5 (2019).

105. Zhang, J. *et al.* Sour Sensing from the Tongue to the Brain. *Cell* **179**, 392-402.e15 (2019).

106. Saotome, K. *et al.* Structures of the otopetrin proton channels Otop1 and Otop3. *Nat. Struct. Mol. Biol.* **26**, 518–525 (2019).

107. Chen, Q., Zeng, W., She, J., Bai, X. & Jiang, Y. Structural and functional characterization of an otopetrin family proton channel. *eLife* **8**, (2019).

108. Bigiani, A. Does ENaC Work as Sodium Taste Receptor in Humans? *Nutrients* **12**, 1195 (2020).

109. Lindemann, B. Receptors and transduction in taste. *Nature* vol. 413 219–225 Preprint at https://doi.org/10.1038/35093032 (2001).

110. Yoshida, R. *et al.* NaCl responsive taste cells in the mouse fungiform taste buds. *Neuroscience* **159**, 795–803 (2009).

111. Jasti, J., Furukawa, H., Gonzales, E. B. & Gouaux, E. Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. *Nature* **449**, 316–323 (2007).

112. Noreng, S., Bharadwaj, A., Posert, R., Yoshioka, C. & Baconguis, I. Structure of the human epithelial sodium channel by cryo-electron microscopy. *eLife* **7**, 1–23 (2018).

113. Hanukoglu, I. & Hanukoglu, A. Epithelial sodium channel (ENaC) family: Phylogeny, structurefunction, tissue distribution, and associated inherited diseases. *April* (2016) doi:10.1016/j.gene.2015.12.061.Epithelial.

114. Halpern, B. P. Amiloride and vertebrate gustatory responses to NaCl. *Neurosci. Biobehav. Rev.* **23**, 5–47 (1998).

115. Stähler, F. *et al.* A Role of the Epithelial Sodium Channel in Human Salt Taste Transduction? *Chemosens. Percept.* **1**, 78–90 (2008).

116. Witt, M. Anatomy and development of the human taste system. in *Handbook of Clinical Neurology* vol. 164 147–171 (Elsevier B.V., 2019).

117. Lossow, K., Hermans-Borgmeyer, I., Meyerhof, W. & Behrens, M. Segregated Expression of ENaC Subunits in Taste Cells. *Chem. Senses* **45**, 235–248 (2020).

118. Nomura, K., Nakanishi, M., Ishidate, F., Iwata, K. & Taruno, A. All-Electrical Ca2+-Independent Signal Transduction Mediates Attractive Sodium Taste in Taste Buds. *Neuron* **106**, 816-829.e6 (2020).

119. Schiffman, S. S., Lockhead, E. & Maes, F. W. Amiloride reduces the taste intensity of Na+ and Li+ salts and sweeteners. *Proc. Natl. Acad. Sci.* **80**, 6136–6140 (1983).

120. Liman, E. R. Salty Taste: From Transduction to Transmitter Release, Hold the Calcium. *Neuron* **106**, 709–711 (2020).

121. Noreng, S., Posert, R., Bharadwaj, A., Houser, A. & Baconguis, I. Molecular principles of assembly, activation, and inhibition in epithelial sodium channel. *eLife* **9**, 1–23 (2020).

122. Yee, K. K., Sukumaran, S. K., Kotha, R., Gilbertson, T. A. & Margolskee, R. F. Glucose transporters and ATP-gated K+ (KATP) metabolic sensors are present in type 1 taste receptor 3 (T1r3)-expressing taste cells. *Proc. Natl. Acad. Sci.* **108**, 5431–5436 (2011).

123. Liu, D., Archer, N., Duesing, K., Hannan, G. & Keast, R. Mechanism of fat taste perception: Association with diet and obesity. *Prog. Lipid Res.* **63**, 41–49 (2016).

124. Besnard, P., Passilly-Degrace, P. & Khan, N. A. Taste of Fat: A Sixth Taste Modality? *Physiol. Rev.* **96**, 151–176 (2016).

125. Khan, A. S., Keast, R. & Khan, N. A. Preference for dietary fat: From detection to disease. *Prog. Lipid Res.* **78**, 101032 (2020).

126. Newman, L. P., Bolhuis, D. P., Torres, S. J. & Keast, R. S. J. Dietary fat restriction increases fat taste sensitivity in people with obesity. *Obesity* **24**, 328–334 (2016).

127. Ng, M. *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* **384**, 766–781 (2014).

128. Artymiuk, P. J., Poirrette, A. R., Rice, D. W. & Willett, P. A polymerase I palm in adenylyl cyclase? *Nature* **388**, 33–34 (1997).

129. Cannariato, M., Miceli, M. & Deriu, M. A. In silico investigation of Alsin RLD conformational dynamics and phosphoinositides binding mechanism. *PLOS ONE* **17**, e0270955 (2022).

130. Ehrt, C., Brinkjost, T. & Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLOS Comput. Biol.* **14**, e1006483 (2018).

131. Breslin, P. A. S. An Evolutionary Perspective on Food and Human Taste. *Curr. Biol.* **23**, R409–R418 (2013).

132. Ho, H. K. *et al.* Functionally expressed bitter taste receptor TAS2R14 in human epidermal keratinocytes serves as a chemosensory receptor. *Exp. Dermatol.* **30**, 216–225 (2021).

133. Shaw, L. *et al.* Personalized expression of bitter 'taste' receptors in human skin. *PLOS ONE* **13**, e0205322 (2018).

134. Singh, N., Vrontakis, M., Parkinson, F. & Chelikani, P. Functional bitter taste receptors are expressed in brain cells. *Biochem. Biophys. Res. Commun.* **406**, 146–151 (2011).

135. Kyriazis, G. A., Soundarapandian, M. M. & Tyrberg, B. Sweet taste receptor signaling in beta cells mediates fructose-induced potentiation of glucose-stimulated insulin secretion. *Proc. Natl. Acad. Sci.* **109**, (2012).

136. Kyriazis, G. A., Smith, K. R., Tyrberg, B., Hussain, T. & Pratley, R. E. Sweet Taste Receptors Regulate Basal Insulin Secretion and Contribute to Compensatory Insulin Hypersecretion During the Development of Diabetes in Male Mice. *Endocrinology* **155**, 2112–2121 (2014).

137. Foster, S. R. *et al.* Expression, Regulation and Putative Nutrient-Sensing Function of Taste GPCRs in the Heart. *PLoS ONE* **8**, e64579 (2013).

138. Foster, S. R. *et al.* Bitter taste receptor agonists elicit G-protein-dependent negative inotropy in the murine heart. *FASEB J.* **28**, 4497–4508 (2014).

139. Deckmann, K. *et al.* Bitter triggers acetylcholine release from polymodal urethral chemosensory cells and bladder reflexes. *Proc. Natl. Acad. Sci.* **111**, 8287–8292 (2014).

140. Iwatsuki, K. & Uneyama, H. Sense of Taste in the Gastrointestinal Tract. *J. Pharmacol. Sci.* **118**, 123–128 (2012).

141. Raka, F., Farr, S., Kelly, J., Stoianov, A. & Adeli, K. Metabolic control via nutrient-sensing mechanisms: role of taste receptors and the gut-brain neuroendocrine axis. *Am. J. Physiol.-Endocrinol. Metab.* **317**, E559–E572 (2019).

142. Carey, R. M. & Lee, R. J. Taste Receptors in Upper Airway Innate Immunity. *Nutrients* **11**, 2017 (2019).

143. Behrens, M. & Lang, T. Extra-Oral Taste Receptors—Function, Disease, and Perspectives. *Front. Nutr.* **9**, 881177 (2022).

144. Slack, J. P. *et al.* Modulation of Bitter Taste Perception by a Small Molecule hTAS2R Antagonist. *Curr. Biol.* **20**, 1104–1109 (2010).

145. 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 Res.* **40**, W380–W386 (2012).

146. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

147. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. in 98–105 (2016). doi:10.25080/Majora-629e541a-00e.

148. Rodrigues, J. P. G. L. M., Teixeira, J. M. C., Trellet, M. & Bonvin, A. M. J. J. pdb-tools: a swiss army knife for molecular structures. *F1000Research* **7**, 1961 (2018).

149. Adasme, M. F. *et al.* PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **49**, W530–W534 (2021).

150. 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 Res.* **40**, W380–W386 (2012).

151. Spriggs, R. V., Artymiuk, P. J. & Willett, P. Searching for Patterns of Amino Acids in 3D Protein Structures. *J. Chem. Inf. Comput. Sci.* **43**, 412–421 (2003).

152. Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **16**, 575–577 (1973).

153. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. in *Ant Colony Optimization and Swarm Intelligence* (eds. Dorigo, M. et al.) vol. 4150 247–258 (Springer Berlin Heidelberg, 2006).

154. Korb, O., Stützle, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.* **1**, 115–134 (2007).

155. Çınaroğlu, S. S. & Timuçin, E. Comparative Assessment of Seven Docking Programs on a Nonredundant Metalloprotein Subset of the PDBbind Refined. *J. Chem. Inf. Model.* **59**, 3846–3859 (2019).

156. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11**, 425–45 (1997).

157. Sirci, F. *et al.* Virtual Fragment Screening: Discovery of Histamine H_3 Receptor Ligands Using Ligand-Based and Protein-Based Molecular Fingerprints. *J. Chem. Inf. Model.* **52**, 3308–3324 (2012).

158. Kooistra, A. J., Leurs, R., De Esch, I. J. P. & De Graaf, C. Structure-Based Prediction of G-Protein-Coupled Receptor Ligand Function: A β-Adrenoceptor Case Study. *J. Chem. Inf. Model.* **55**, 1045–1061 (2015).

159. Bassani, D., Pavan, M., Sturlese, M. & Moro, S. Sodium or Not Sodium: Should Its Presence Affect the Accuracy of Pose Prediction in Docking GPCR Antagonists? *Pharmaceuticals* **15**, 346 (2022).

160. Kooistra, A. J. *et al.* Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* **6**, 28288 (2016).

161. De Graaf, C. *et al.* Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H_1 Receptor. *J. Med. Chem.* **54**, 8195–8206 (2011).

162. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

163. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).

164. The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Res.* **36**, D440–D444 (2008).

165. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

166. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

167. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

168. Ferreira, J. A. & Zwinderman, A. H. On the Benjamini–Hochberg method. *Ann. Stat.* **34**, 1827–1849 (2006).

169. NCBI Resource Coordinators *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

170. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

171. Xu, W. *et al.* Structural basis for strychnine activation of human bitter taste receptor TAS2R46. *Science* **377**, 1298–1304 (2022).

172. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

173. Molecular Operating Environment (MOE), 2022.02 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2023. (2022).

174. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

175. Dickson, C. J., Walker, R. C. & Gould, I. R. Lipid21: Complex Lipid Membrane Simulations with AMBER. *J. Chem. Theory Comput.* **18**, 1726–1736 (2022).

176. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).

177. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

178. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

179. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).

180. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

181. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

182. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

183. Born, S., Levit, A., Niv, M. Y., Meyerhof, W. & Behrens, M. The human bitter taste receptor TAS2R10 is tailored to accommodate numerous diverse ligands. *J. Neurosci.* **33**, 201–213 (2013).

184. Jensen, A. A., Gharagozloo, P., Birdsall, N. J. M. & Zlotos, D. P. Pharmacological characterisation of strychnine and brucine analogues at glycine and α7 nicotinic acetylcholine receptors. *Eur. J. Pharmacol.* **539**, 27–33 (2006).

185. DiPizio, A. *et al.* Rational design of agonists for bitter taste receptor TAS2R14: from modeling to bench and back. *Cell. Mol. Life Sci.* **77**, (2020).

186. Fierro, F., Giorgetti, A., Carloni, P., Meyerhof, W. & Alfonso-Prieto, M. Dual binding mode of "bitter sugars" to their human bitter taste receptor target. *Sci. Rep.* **9**, 1–16 (2019).

187. Nicoli, A., Dunkel, A., Giorgino, T., de Graaf, C. & Di Pizio, A. Classification Model for the Second Extracellular Loop of Class A GPCRs. *J. Chem. Inf. Model.* **62**, 511–522 (2022).

188. Topin, J. *et al.* Functional molecular switches of mammalian G protein-coupled bitter-taste receptors. *Cell. Mol. Life Sci.* **6**, 2020.10.23.348706 (2021).

189. Kuipers, W. *et al.* Study of the interaction between aryloxypropanolamines and Asn386 in helix VII of the human 5-hydroxytryptamine1A receptor. *Mol. Pharmacol.* **51**, 889–896 (1997).

190. Oksenberg, D. *et al.* A single amino-acid difference confers major pharmacological variation between human and rodent 5-HT1B receptors. *Nature* **360**, 161–163 (1992).

191. Suryanarayana, S., Daunt, D. A., Von Zastrow, M. & Kobilka, B. K. A point mutation in the seventh hydrophobic domain of the alpha 2 adrenergic receptor

increases its affinity for a family of beta receptor antagonists. *J. Biol. Chem.* **266**, 15488–15492 (1991).

192. Suryanarayana, S. & Kobilka, B. K. Amino acid substitutions at position 312 in the seventh hydrophobic segment of the beta 2-adrenergic receptor modify ligand-binding specificity. *Mol. Pharmacol.* **44**, 111–114 (1993).

193. Erb, L. *et al.* Site-directed mutagenesis of P2U purinoceptors. Positively charged amino acids in transmembrane helices 6 and 7 affect agonist potency and specificity. *J. Biol. Chem.* **270**, 4185–4188 (1995).

194. Jiang, Q. *et al.* A Mutational Analysis of Residues Essential for Ligand Recognition at the Human P2Y $_1$ Receptor. *Mol. Pharmacol.* **52**, 499–507 (1997).

195. Kopin, A. S., McBride, E. W., Quinn, S. M., Kolakowski, L. F. & Beinborn, M. The role of the cholecystokinin-B/gastrin receptor transmembrane domains in determining affinity for subtype-selective ligands. *J. Biol. Chem.* **270**, 5019–5023 (1995).

196. Pronin, A. N. Identification of Ligands for Two Human Bitter T2R Receptors. *Chem. Senses* **29**, 583–593 (2004).

197. Zlotos, D. P., Mandour, Y. M. & Jensen, A. A. Strychnine and its mono- and dimeric analogues: a pharmaco-chemical perspective. *Nat. Prod. Rep.* **39**, 1910–1937 (2022).

198. Selkoe, D. J. The molecular pathology of Alzheimer's disease. *Neuron* **6**, 487–498 (1991).

199. Cummings, J. L. Alzheimer's Disease. *N. Engl. J. Med.* **351**, 56–67 (2004).

200. Andrade, S., Ramalho, M. J., Loureiro, J. A. & Do Carmo Pereira, M. Natural compounds for alzheimer's disease therapy: A systematic review of preclinical and clinical studies. *Int. J. Mol. Sci.* **20**, (2019).

201. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 595–608 (2016).

202. Cohen, S. I. A. *et al.* Proliferation of amyloid- 42 aggregates occurs through a secondary nucleation mechanism. *Proc. Natl. Acad. Sci.* **110**, 9758–9763 (2013).

203. Schütz, A. K. *et al.* Atomic-Resolution Three-Dimensional Structure of Amyloid $β$ Fibrils Bearing the Osaka Mutation. *Angew. Chem. Int. Ed.* **54**, 331–335 (2015).

204. Guo, S. & Akhremitchev, B. B. Packing density and structural heterogeneity of insulin amyloid fibrils measured by AFM nanoindentation. *Biomacromolecules* **7**, 1630–1636 (2006).

205. VandenAkker, C. C., Engel, M. F. M., Velikov, K. P., Bonn, M. & Koenderink, G. H. Morphology and Persistence Length of Amyloid Fibrils

Are Correlated to Peptide Molecular Structure. *J. Am. Chem. Soc.* **133**, 18030–18033 (2011).

206. Palhano, F. L., Lee, J., Grimster, N. P. & Kelly, J. W. Toward the Molecular Mechanism(s) by Which EGCG Treatment Remodels Mature Amyloid Fibrils. *J. Am. Chem. Soc.* **135**, 7503–7510 (2013).

207. Wang, T., Jo, H., DeGrado, W. F. & Hong, M. Water Distribution, Dynamics, and Interactions with Alzheimer's β-Amyloid Fibrils Investigated by Solid-State NMR. *J. Am. Chem. Soc.* **139**, 6242–6252 (2017).

208. Grasso, G. *et al.* Conformational Dynamics and Stability of U-Shaped and S-Shaped Amyloid β Assemblies. *Int. J. Mol. Sci.* **19**, 571 (2018).

209. Grasso, G. *et al.* The Role of Structural Polymorphism in Driving the Mechanical Performance of the Alzheimer's Beta Amyloid Fibrils. *Front. Bioeng. Biotechnol.* **7**, 83 (2019).

210. Miceli, M., Muscat, S., Morbiducci, U., Cavaglià, M. & Deriu, M. A. Ultrasonic waves effect on S-shaped β-amyloids conformational dynamics by non-equilibrium molecular dynamics. *J. Mol. Graph. Model.* **96**, 107518 (2020).

211. Muscat, S., Stojceski, F. & Danani, A. Elucidating the Effect of Static Electric Field on Amyloid Beta 1–42 Supramolecular Assembly. *J. Mol. Graph. Model.* **96**, 107535 (2020).

212. Lambracht-Washington, D. & Rosenberg, R. N. Advances in the development of vaccines for Alzheimer's disease. *Discov. Med.* **15**, 319–26 (2013).

213. Wang, C. Y. *et al.* UB-311, a novel UBITh ® amyloid β peptide vaccine for mild Alzheimer's disease. *Alzheimers Dement. Transl. Res. Clin. Interv.* **3**, 262–272 (2017).

214. Gold, M. Phase II clinical trials of anti–amyloid β antibodies: When is enough, enough? *Alzheimers Dement. Transl. Res. Clin. Interv.* **3**, 402–409 (2017).

215. van Dyck, C. H. Anti-Amyloid-β Monoclonal Antibodies for Alzheimer's Disease: Pitfalls and Promise. *Biol. Psychiatry* **83**, 311–319 (2018).

216. Rajasekhar, K., Madhu, C. & Govindaraju, T. Natural Tripeptide-Based Inhibitor of Multifaceted Amyloid β Toxicity. *ACS Chem. Neurosci.* **7**, 1300–1310 (2016).

217. Viet, M. H., Ngo, S. T., Lam, N. S. & Li, M. S. Inhibition of Aggregation of Amyloid Peptides by Beta-Sheet Breaker Peptides and Their Binding Affinity. *J. Phys. Chem. B* **115**, 7433–7446 (2011).

218. Lin, D. *et al.* Interaction Dynamics in Inhibiting the Aggregation of Aβ Peptides by SWCNTs: A Combined Experimental and Coarse-Grained Molecular Dynamic Simulation Study. *ACS Chem. Neurosci.* **7**, 1232–1240 (2016).

219. Liu, F., Wang, W., Sang, J., Jia, L. & Lu, F. Hydroxylated Single-Walled Carbon Nanotubes Inhibit Aβ 42 Fibrillogenesis, Disaggregate Mature Fibrils, and Protect against Aβ 42 -Induced Cytotoxicity. *ACS Chem. Neurosci.* **10**, 588–598 (2019).

220. Xiong, N., Dong, X. Y., Zheng, J., Liu, F. F. & Sun, Y. Design of LVFFARK and LVFFARK-Functionalized Nanoparticles for Inhibiting Amyloid β-Protein Fibrillation and Cytotoxicity. *ACS Appl. Mater. Interfaces* **7**, 5650–5662 (2015).

221. MacLeod, R., Hillert, E.-K., Cameron, R. T. & Baillie, G. S. The role and therapeutic targeting of α-, β- and γ-secretase in Alzheimer's disease. *Future Sci. OA* **1**, fso.15.9 (2015).

222. Cui, J. *et al.* Targeting the γ-/β-secretase interaction reduces β-amyloid generation and ameliorates Alzheimer's disease-related pathogenesis. *Cell Discov.* **1**, 15021 (2015).

223. Nie, Q., Du, X. G. & Geng, M. Y. Small molecule inhibitors of amyloid β peptide aggregation as a potential therapeutic strategy for Alzheimer's disease. *Acta Pharmacol. Sin.* **32**, 545–551 (2011).

224. Zhu, M. *et al.* Identification of small-molecule binding pockets in the soluble monomeric form of the Aβ42 peptide. *J. Chem. Phys.* **139**, (2013).

225. Doig, A. J. & Derreumaux, P. Inhibition of protein aggregation and amyloid formation by small molecules. *Curr. Opin. Struct. Biol.* **30**, 50–56 (2015).

226. Habchi, J. *et al.* Systematic development of small molecules to inhibit specific microscopic steps of Aβ42 aggregation in Alzheimer's disease. *Proc. Natl. Acad. Sci.* **114**, E200–E208 (2017).

227. Liu, F., Ma, Z., Sang, J. & Lu, F. Edaravone inhibits the conformational transition of amyloid-β42: insights from molecular dynamics simulations. *J. Biomol. Struct. Dyn.* **0**, 1–12 (2019).

228. Liang, C., Savinov, S. N., Fejzo, J., Eyles, S. J. & Chen, J. Modulation of Amyloid-β42 Conformation by Small Molecules Through Nonspecific Binding. *J. Chem. Theory Comput.* **15**, 5169–5174 (2019).

229. Orgogozo, J. M. *et al.* Subacute meningoencephalitis in a subset of patients with AD after Aβ42 immunization. *Neurology* **61**, 46–54 (2003).

230. Frid, P., Anisimov, S. V. & Popovic, N. Congo red and protein aggregation in neurodegenerative diseases. *Brain Res. Rev.* **53**, 135–160 (2007).

231. Zenaro, E., Piacentino, G. & Constantin, G. The blood-brain barrier in Alzheimer's disease. *Neurobiol. Dis.* **107**, 41–56 (2017).

232. Bui, T. T. & Nguyen, T. H. Natural product for the treatment of Alzheimer's disease. *J. Basic Clin. Physiol. Pharmacol.* **28**, 413–423 (2017).

233. Butler, M. S., Robertson, A. A. B. & Cooper, M. A. Natural product and natural product derived drugs in clinical trials. *Nat Prod Rep* **31**, 1612–1661 (2014).

234. Andrade, S., Ramalho, M. J., Loureiro, J. A. & Pereira, M. C. Interaction of natural compounds with biomembrane models: A biophysical approach for the Alzheimer's disease therapy. *Colloids Surf. B Biointerfaces* **180**, 83–92 (2019).

235. Rasool, M. *et al.* Recent Updates in the Treatment of Neurodegenerative Disorders Using Natural Compounds. *Evid. Based Complement. Alternat. Med.* **2014**, 1–7 (2014).

236. Hiremathad, A. A Review: Natural Compounds as Anti-Alzheimer´s Disease Agents. *Curr. Nutr. Food Sci.* **13**, (2017).

237. Deb, S. *et al.* Therapeutic implications of anti-inflammatory natural products in Alzheimer's disease. in *Discovery and Development of Anti-Inflammatory Agents from Natural Products* 241–258 (Elsevier, 2019). doi:10.1016/B978-0-12-816992-6.00008-5.

238. Mourtas, S. *et al.* Curcumin-decorated nanoliposomes with very high affinity for amyloid-β1-42 peptide. *Biomaterials* **32**, 1635–1645 (2011).

239. Ringman, J., Frautschy, S., Cole, G., Masterman, D. & Cummings, J. A Potential Role of the Curry Spice Curcumin in Alzheimers Disease. *Curr. Alzheimer Res.* **2**, 131–136 (2005).

240. Knowles, T. P. *et al.* Role of intermolecular forces in defining material properties of protein nanofibrils. *Science* **318**, 1900–1903 (2007).

241. Bidone, T. C., Kim, T., Deriu, M. A., Morbiducci, U. & Kamm, R. D. Multiscale impact of nucleotides and cations on the conformational equilibrium, elasticity and rheology of actin filaments and crosslinked networks. *Biomech. Model. Mechanobiol.* **14**, 1143–1155 (2015).

242. Fan, H.-M., Xu, Q. & Wei, D.-Q. Recent Studies on Mechanisms of New Drug Candidates for Alzheimer's Disease Interacting with Amyloid-β Protofibrils Using Molecular Dynamics Simulations. in 135–151 (2017). doi:10.1007/978-94-024-1045-7_6.

243. Tang, M. *et al.* A Novel Drug Candidate for Alzheimer's Disease Treatment: gx-50 Derived from Zanthoxylum Bungeanum. *J. Alzheimers Dis.* **34**, 203–213 (2013).

244. Hou, S., Gu, R.-X. & Wei, D.-Q. Inhibition of β-Amyloid Channels with a Drug Candidate wgx-50 Revealed by Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **57**, 2811–2821 (2017).

245. Fan, H.-M. *et al.* Destabilization of Alzheimer's Aβ42 Protofibrils with a Novel Drug Candidate wgx-50 by Molecular Dynamics Simulations. *J. Phys. Chem. B* **119**, 11196–11202 (2015).

246. Kanchi, P. K. & Dasmahapatra, A. K. Polyproline chains destabilize the Alzheimer's amyloid-β protofibrils: A molecular dynamics simulation study. *J. Mol. Graph. Model.* **93**, 107456 (2019).

247. Sharma, B., Kalita, S., Paul, A., Mandal, B. & Paul, S. The role of caffeine as an inhibitor in the aggregation of amyloid forming peptides: a unified molecular dynamics simulation and experimental study. *RSC Adv.* **6**, 78548–78558 (2016).

248. Battisti, A. *et al.* Curcumin-like compounds designed to modify amyloid beta peptide aggregation patterns. *RSC Adv.* **7**, 31714–31724 (2017).

249. Masuda, Y. *et al.* Solid-state NMR analysis of interaction sites of curcumin and 42-residue amyloid β-protein fibrils. *Bioorg. Med. Chem.* **19**, 5967–5974 (2011).

250. Rao, P. P. N., Mohamed, T., Teckwani, K. & Tin, G. Curcumin Binding to Beta Amyloid: A Computational Study. *Chem. Biol. Drug Des.* **86**, 813–820 (2015).

251. Xiao, Y. *et al.* Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* **22**, 499–505 (2015).

252. Fändrich, M. *et al.* Amyloid fibril polymorphism: a challenge for molecular imaging and therapy. *J. Intern. Med.* **283**, 218–237 (2018).

253. Acosta, D. M. Á. V., Vega, B. C., Basurto, J. C., Morales, L. G. F. & Rosales Hernández, M. C. Recent Advances by In Silico and In Vitro Studies of Amyloid-β 1-42 Fibril Depicted a S-Shape Conformation. *Int. J. Mol. Sci.* **19**, 2415 (2018).

254. Hiremathad, A. A Review: Natural Compounds as Anti-Alzheimer´s Disease Agents. *Curr. Nutr. Food Sci.* **13**, (2017).

255. Rasool, M. *et al.* Recent Updates in the Treatment of Neurodegenerative Disorders Using Natural Compounds. *Evid. Based Complement. Alternat. Med.* **2014**, 1–7 (2014).

256. Andrade, S., Ramalho, M. J., Loureiro, J. A. & Pereira, M. C. Interaction of natural compounds with biomembrane models: A biophysical approach for the Alzheimer's disease therapy. *Colloids Surf. B Biointerfaces* **180**, 83–92 (2019).

257. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

258. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).

259. Hess, B., Hess, B., Bekker, H., Berendsen, H. J. C. C. & Fraaije, J. G. E. M. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

260. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 14101 (2007).

261. Berendsen, H. J. C. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

262. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

263. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089 (1993).

264. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

265. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 27-28,33-38 (1996).

266. Grasso, G. *et al.* Cell penetrating peptide modulation of membrane biomechanics by Molecular dynamics. *J. Biomech.* **73**, 137–144 (2018).

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

268. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

269. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).

270. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **23**, 1623–1641 (2002).

271. Klejborowska, G. *et al.* Synthesis, biological evaluation and molecular docking studies of new amides of 4-bromothiocolchicine as anticancer agents. *Bioorg. Med. Chem.* **76**, 115144 (2019).

272. Sahakyan, H., Abelyan, N., Arakelov, V., Arakelov, G. & Nazaryan, K. In silico study of colchicine resistance molecular mechanisms caused by tubulin structural polymorphism. *PLOS ONE* **14**, e0221532 (2019).

273. Kumbhar, B. V., Borogaon, A., Panda, D. & Kunwar, A. Exploring the Origin of Differential Binding Affinities of Human Tubulin Isotypes αβII, αβIII and αβIV for DAMA-Colchicine Using Homology Modelling, Molecular Docking and Molecular Dynamics Simulations. *PLOS ONE* **11**, e0156048 (2016).

274. Gajewski, M. M., Tuszynski, J. A., Barakat, K., Huzil, J. T. & Klobukowski, M. Interactions of laulimalide, peloruside, and their derivatives with the isoforms of β-tubulin. *Can. J. Chem.* **91**, 511–517 (2013).

275. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

276. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).

277. Sun, H. *et al.* Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.* **20**, 14450–14460 (2018).

278. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem.* **32**, 866–877 (2011).

279. Su, P. C., Tsai, C. C., Mehboob, S., Hevener, K. E. & Johnson, M. E. Comparison of radii sets, entropy, QM methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of F. tularensis enoyl-ACP reductase (FabI). *J. Comput. Chem.* **36**, 1859–1873 (2015).

280. Grasso, G., Leanza, L., Morbiducci, U., Danani, A. & Deriu, M. A. Aminoacid Substitutions in the Glycine Zipper Affect the Conformational Stability of Amyloid Beta Fibrils. *J. Biomol. Struct. Dyn.* 1–13 (2019) doi:10.1080/07391102.2019.1671224.

281. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **45**, 160–169 (2005).

282. Kumar, A., Singh, A., & Ekavali. A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacol. Rep.* **67**, 195–203 (2015).

283. Auld, D. S., Kornecook, T. J., Bastianetto, S. & Quirion, R. Alzheimer's disease and the basal forebrain cholinergic system: relations to β-amyloid peptides, cognition, and treatment strategies. *Prog. Neurobiol.* **68**, 209–245 (2002).

284. Farlow, M. R., Miller, M. L. & Pejovic, V. Treatment Options in Alzheimer’s Disease: Maximizing Benefit, Managing Expectations. *Dement. Geriatr. Cogn. Disord.* **25**, 408–422 (2008).

285. Du, W. J. *et al.* Brazilin inhibits amyloid β-protein fibrillogenesis, remodels amyloid fibrils and reduces amyloid cytotoxicity. *Sci. Rep.* **5**, 1–10 (2015).

286. Tavanti, F., Pedone, A. & Menziani, M. Computational Insight into the Effect of Natural Compounds on the Destabilization of Preformed Amyloid-β(1–40) Fibrils. *Molecules* **23**, 1320 (2018).

287. Pourkhodadad, S. *et al.* Neuroprotective effects of oleuropein against cognitive dysfunction induced by colchicine in hippocampal CA1 area in rats. *J. Physiol. Sci.* **66**, 397–405 (2016).

288. Luccarini, I. *et al.* Oleuropein aglycone counteracts Aβ42 toxicity in the rat brain. *Neurosci. Lett.* **558**, 67–72 (2014).

289. Ferrándiz, M. L. & Alcaraz, M. J. Anti-inflammatory activity and inhibition of arachidonic acid metabolism by flavonoids. *Agents Actions* **32**, 283–288 (1991).

290. Thamizhiniyan, V., Vijayaraghavan, K. & Subramanian, S. P. Gossypin, a flavonol glucoside protects pancreatic beta-cells from glucotoxicity in streptozotocin-induced experimental diabetes in rats. *Biomed. Prev. Nutr.* **2**, 239–245 (2012).

291. Ono, K., Hasegawa, K., Naiki, H. & Yamada, M. Curcumin Has Potent Anti-Amyloidogenic Effects for Alzheimer's β-Amyloid Fibrils In Vitro. *J. Neurosci. Res.* (2004) doi:10.1002/jnr.20025.

292. Yang, F. *et al.* Curcumin inhibits formation of amyloid β oligomers and fibrils, binds plaques, and reduces amyloid in vivo. *J. Biol. Chem.* (2005) doi:10.1074/jbc.M404751200.

293. Diomede, L., Rigacci, S., Romeo, M., Stefani, M. & Salmona, M. Oleuropein Aglycone Protects Transgenic C. elegans Strains Expressing Aβ42 by Reducing Plaque Load and Motor Deficit. *PLoS ONE* **8**, (2013).

294. Rigacci, S. *et al.* Aβ(1-42) aggregates into non-toxic amyloid assemblies in the presence of the natural polyphenol oleuropein aglycon. *Curr. Alzheimer Res.* **8**, 841–52 (2011).

295. Pantano, D. *et al.* Oleuropein aglycone and polyphenols from olive mill waste water ameliorate cognitive deficits and neuropathology. *Br. J. Clin. Pharmacol.* **83**, 54–62 (2017).

296. Masuda, Y. *et al.* Solid-state NMR analysis of interaction sites of curcumin and 42-residue amyloid β-protein fibrils. *Bioorg. Med. Chem.* **19**, 5967–5974 (2011).

297. Kumar, J., Namsechi, R. & Sim, V. L. Structure-based peptide design to modulate amyloid beta aggregation and reduce cytotoxicity. *PLoS ONE* **10**, 1–18 (2015).

298. Leri, M., Natalello, A., Bruzzone, E., Stefani, M. & Bucciantini, M. Oleuropein aglycone and hydroxytyrosol interfere differently with toxic Aβ 1-42 aggregation. *Food Chem. Toxicol.* **129**, 1–12 (2019).

299. Doty, R. L. & Bromley, S. M. Taste. in *Encyclopedia of the Neurological Sciences* 394–396 (Elsevier, 2014). doi:10.1016/B978-0-12-385157-4.00073-7.

300. Wisman, A. & Shrira, I. The smell of death: evidence that putrescine elicits threat management mechanisms. *Front. Psychol.* **6**, 1–11 (2015).

301. Stevenson, R. J. An Initial Evaluation of the Functions of Human Olfaction. *Chem. Senses* **35**, 3–20 (2010).

302. Hussain, A. *et al.* High-affinity olfactory receptor for the death-associated odor cadaverine. *Proc. Natl. Acad. Sci.* **110**, 19579–19584 (2013).

303. Tosti, V., Bertozzi, B. & Fontana, L. Health Benefits of the Mediterranean Diet: Metabolic and Molecular Mechanisms. *J. Gerontol. Ser. A* **73**, 318–326 (2018).

304. Mentella, Scaldaferri, Ricci, Gasbarrini & Miggiano. Cancer and Mediterranean Diet: A Review. *Nutrients* **11**, 2059 (2019).

305. Yang, X. & Boyle, R. A. Sensory Evaluation of Oils/Fats and Oil/Fat–Based Foods. in *Oxidative Stability and Shelf Life of Foods Containing Oils and Fats* 157–185 (Elsevier, 2016). doi:10.1016/B978-1-63067-056-6.00003-3.

306. Stone, H. & Sidel, J. L. Introduction to Sensory Evaluation. in *Sensory Evaluation Practices* 1–19 (Elsevier, 2004). doi:10.1016/B978-012672690-9/50005-6.

307. Feiner, G. Sensory evaluation of meat products. in *Meat Products Handbook* 565–568 (Elsevier, 2006). doi:10.1533/9781845691721.3.565.

308. Vivek, K., Subbarao, K. V., Routray, W., Kamini, N. R. & Dash, K. K. Application of Fuzzy Logic in Sensory Evaluation of Food Products: a Comprehensive Study. *Food Bioprocess Technol.* **13**, 1–29 (2020).

309. Mahato, D. K. *et al.* Sugar Reduction in Dairy Food: An Overview with Flavoured Milk as an Example. *Foods* **9**, 1400 (2020).

310. Dello Russo, M. *et al.* The Impact of Adding Sugars to Milk and Fruit on Adiposity and Diet Quality in Children: A Cross-Sectional and Longitudinal Analysis of the Identification and Prevention of Dietary- and Lifestyle-Induced Health Effects in Children and Infants (IDEFICS) St. *Nutrients* **10**, 1350 (2018).

311. Malik, V. S., Schulze, M. B. & Hu, F. B. Intake of sugar-sweetened beverages and weight gain: A systematic review. *Am. J. Clin. Nutr.* **84**, 274–288 (2006).

312. Te Morenga, L., Mallard, S. & Mann, J. Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *BMJ* **346**, e7492–e7492 (2012).

313. Hu, F. B. & Malik, V. S. Sugar-sweetened beverages and risk of obesity and type 2 diabetes: Epidemiologic evidence. *Physiol. Behav.* **100**, 47–54 (2010).

314. Malik, V. S., Popkin, B. M., Bray, G. A., Després, J.-P. & Hu, F. B. Sugar-Sweetened Beverages, Obesity, Type 2 Diabetes Mellitus, and Cardiovascular Disease Risk. *Circulation* **121**, 1356–1364 (2010).

315. Mennella, J. A., Spector, A. C., Reed, D. R. & Coldwell, S. E. The Bad Taste of Medicines: Overview of Basic Research on Bitter Taste. *Clin. Ther.* **35**, 1225–1246 (2013).

316. Xydakis, M. S. *et al.* Smell and taste dysfunction in patients with COVID-19. *Lancet Infect. Dis.* **20**, 1015–1016 (2020).

317. Charoenkwan, P. *et al.* iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **112**, 2813–2822 (2020).

318. Rodgers, S., Glen, R. C. & Bender, A. Characterizing bitterness: Identification of key structural features and development of a classification model. *J. Chem. Inf. Model.* **46**, 569–576 (2006).

319. Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method

with Propensity Scores of Dipeptides. *J. Chem. Inf. Model.* **60**, 6666–6678 (2020).

320. Rojas, C. *et al.* A QSTR-based expert system to predict sweetness of molecules. *Front. Chem.* **5**, 1–12 (2017).

321. Burdock, G. A. *Fenaroli's Handbook of Flavor Ingredients*. (CRC Press, 2016). doi:10.1201/9781439847503.

322. Banerjee, P. *et al.* Super Natural II-a database of natural products. *Nucleic Acids Res.* **43**, D935–D939 (2015).

323. Garg, N. *et al.* FlavorDB: a database of flavor molecules. *Nucleic Acids Res.* **46**, D1210–D1216 (2018).

324. Rothwell, J. A. *et al.* Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database* **2013**, bat070–bat070 (2013).

325. Minkiewicz, P., Iwaniak, A. & Darewicz, M. BIOPEP-UWM database of bioactive peptides: Current opportunities. *Int. J. Mol. Sci.* **20**, (2019).

326. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).

327. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

328. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

329. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).

330. Wiener, A., Shudler, M., Levit, A. & Niv, M. Y. BitterDB: A database of bitter compounds. *Nucleic Acids Res.* **40**, 413–419 (2012).

331. Nissim, I., Dagan-Wiener, A. & Niv, M. Y. The taste of toxicity: A quantitative analysis of bitter and toxic molecules. *IUBMB Life* **69**, 938–946 (2017).

332. Nakata, T. *et al.* Role of Basic and Acidic Fragments in Delicious Peptides (Lys-Gly-Asp Glu-Glu-Ser-Leu-Ala) and the Taste Behavior of Sodium and Potassium Salts in Acidic Oligopeptides. *Biosci. Biotechnol. Biochem.* **59**, 689–693 (1995).

333. Yu, Z. *et al.* Taste, umami-enhance effect and amino acid sequence of peptides separated from silkworm pupa hydrolysate. *Food Res. Int.* **108**, 144–150 (2018).

334. Yu, X., Zhang, L., Miao, X., Li, Y. & Liu, Y. The structure features of umami hexapeptides for the T1R1/T1R3 receptor. *Food Chem.* **221**, 599–605 (2017).

335. Zhang, J., Zhao, M., Su, G. & Lin, L. Identification and taste characteristics of novel umami and umami-enhancing peptides separated from peanut protein isolate hydrolysate by consecutive chromatography and UPLC–ESI–QTOF–MS/MS. *Food Chem.* **278**, 674–682 (2019).

336. Dang, Y. *et al.* Establishment of new assessment method for the synergistic effect between umami peptides and monosodium glutamate using electronic tongue. *Food Res. Int.* **121**, 20–27 (2019).

337. Charoenkwan, P., Kanthawong, S., Schaduangrat, N., Yana, J. & Shoombuatong, W. PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. *Cells* **9**, 353 (2020).

338. Goel, A., Gajula, K., Gupta, R. & Rai, B. In-silico prediction of sweetness using structure-activity relationship models. *Food Chem.* **253**, 127–131 (2018).

339. Zheng, S., Chang, W., Xu, W., Xu, Y. & Lin, F. e-Sweet: A machine-learning based platform for the prediction of sweetener and its relative sweetness. *Front. Chem.* **7**, 1–14 (2019).

340. Bouysset, C., Belloir, C., Antonczak, S., Briand, L. & Fiorucci, S. Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food Chem.* **324**, 126864 (2020).

341. Huang, W. *et al.* BitterX: A tool for understanding bitter taste in humans. *Sci. Rep.* **6**, 1–8 (2016).

342. Todeschini, R. & Consonni, V. Methods and Principles in Medicinal Chemistry. in 438–438 (2007). doi:10.1002/9783527610907.scard.

343. Dagan-Wiener, A. *et al.* Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.* **7**, 1–13 (2017).

344. Zheng, S. *et al.* e-Bitter: Bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* **6**, 1–18 (2018).

345. Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B. & Shoombuatong, W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab133.

346. Charoenkwan, P. *et al.* iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *Int. J. Mol. Sci.* **22**, 8958 (2021).

347. Margulis, E. *et al.* Intense bitterness of molecules: Machine learning for expediting drug discovery. *Comput. Struct. Biotechnol. J.* **19**, 568–576 (2021).

348. Banerjee, P. & Preissner, R. Bitter sweet forest: A Random Forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Front. Chem.* **6**, 1–10 (2018).

349. Tuwani, R., Wadhwa, S. & Bagler, G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Sci. Rep.* **9**, 7155 (2019).

350. Fritz, F., Preissner, R. & Banerjee, P. VirtualTaste: a web server for the prediction of organoleptic properties of chemical compounds. *Nucleic Acids Res.* 1–6 (2021) doi:10.1093/nar/gkab292.

351. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

352. Baurin, N. *et al.* 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **44**, 276–285 (2004).

353. Drew, M. G. B. *et al.* Quantitative Structure−Activity Relationship Studies of Sulfamates RNHSO 3 Na: Distinction between Sweet, Sweet-Bitter, and Bitter Molecules. *J. Agric. Food Chem.* **46**, 3016–3026 (1998).

354. Yang, X., Chong, Y., Yan, A. & Chen, J. In-silico prediction of sweetness of sugars and sweeteners. *Food Chem.* **128**, 653–658 (2011).

355. Iwamura, H. Structure-sweetness relationship of L-aspartyl dipeptide analogs. A receptor site topology. *J. Med. Chem.* **24**, 572–583 (1981).

356. Vepuri, S. B., Tawari, N. R. & Degani, M. S. Quantitative Structure–Activity Relationship Study of Some Aspartic Acid Analogues to Correlate and Predict their Sweetness Potency. *QSAR Comb. Sci.* **26**, 204–214 (2007).

357. Kinghorn, A. D. & Soejarto, D. D. Discovery of terpenoid and phenolic sweeteners from plants. *Pure Appl. Chem.* **74**, 1169–1179 (2002).

358. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **10**, 4 (2018).

359. Cao, D.-S., Xu, Q.-S., Hu, Q.-N. & Liang, Y.-Z. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* **29**, 1092–1094 (2013).

360. Golbraikh, A. & Tropsha, A. Beware of q2! in *Journal of Molecular Graphics and Modelling* vol. 20 269–276 (2002).

361. Rojas, C., Tripaldi, P. & Duchowicz, P. R. A New QSPR Study on Relative Sweetness. *Int. J. Quant. Struct.-Prop. Relatsh.* **1**, 78–93 (2016).

362. Zhong, M., Chong, Y., Nie, X., Yan, A. & Yuan, Q. Prediction of sweetness by multilinear regression analysis and support vector machine. *J. Food Sci.* **78**, (2013).

363. Teixeira, A. L., Leal, J. P. & Falcao, A. O. Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminformatics* **5**, 9 (2013).

364. Kawashima, S. AAindex: Amino Acid index database. *Nucleic Acids Res.* **28**, 374–374 (2000).

365. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *1st Int. Conf. Learn. Represent. ICLR 2013 - Workshop Track Proc.* (2013).

366. Asgari, E. & Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* **10**, e0141287 (2015).

367. Aizawa, A. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manag.* **39**, 45–65 (2003).

368. Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T. & Shoombuatong, W. IQSP: A sequence-based tool for the prediction and analysis of quorum sensing peptides via chou's 5-steps rule and informative physicochemical properties. *Int. J. Mol. Sci.* **21**, (2020).

369. Chen, M. *et al.* DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **21**, 648–653 (2016).

370. Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. in *Studies in Classification, Data Analysis, and Knowledge Organization* 319–326 (2008). doi:10.1007/978-3-540-78246-9_38.

371. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).

372. Shelley, J. C. *et al.* Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **21**, 681–691 (2007).

373. Kursa, M. B., Jankowski, A. & Rudnicki, W. R. Boruta - A system for feature selection. *Fundam. Informaticae* **101**, 271–285 (2010).

374. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).

375. Banerjee, P., Dehnbostel, F. O. & Preissner, R. Prediction is a balancing act: Importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. *Front. Chem.* **6**, 1–11 (2018).

376. Mathai, N. & Kirchmair, J. Similarity-Based Methods and Machine Learning Approaches for Target Prediction in Early Drug Discovery: Performance and Scope. *Int. J. Mol. Sci.* **21**, 3585 (2020).

377. European Commission Environment Directorate General. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. (OECD, 2014). doi:10.1787/9789264085442-en.

378. F.Y, O. *et al.* Supervised Machine Learning Algorithms: Classification and Comparison. *Int. J. Comput. Trends Technol.* **48**, 128–138 (2017).

379. Wang, Q., Luo, Z., Huang, J., Feng, Y. & Liu, Z. A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. *Comput. Intell. Neurosci.* **2017**, 1–11 (2017).

380. Zhang, J., Sun-Waterhouse, D., Su, G. & Zhao, M. New insight into umami receptor, umami/umami-enhancing peptides and their derivatives: A review. *Trends Food Sci. Technol.* **88**, 429–438 (2019).

381. Temussi, P. A. The good taste of peptides. *J. Pept. Sci.* **18**, 73–82 (2012).

382. Wang, W., Zhou, X. & Liu, Y. Characterization and evaluation of umami taste: A review. *TrAC - Trends Anal. Chem.* **127**, 115876 (2020).

383. Dang, Y., Gao, X., Ma, F. & Wu, X. Comparison of umami taste peptides in water-soluble extractions of Jinhua and Parma hams. *Lwt* **60**, 1179–1186 (2015).

384. Quintero, F. A., Patel, S. J., Muñoz, F. & Sam Mannan, M. Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System. *Ind. Eng. Chem. Res.* **51**, 16101–16115 (2012).

385. Malavolta, M. *et al.* A survey on computational taste predictors. *Eur. Food Res. Technol.* (2022) doi:10.1007/s00217-022-04044-5.

386. Rojas, C., Tripaldi, P. & Duchowicz, P. R. A New QSPR Study on Relative Sweetness. *Int. J. Quant. Struct.-Prop. Relatsh.* **1**, 78–93 (2016).

387. Bouysset, C., Belloir, C., Antonczak, S., Briand, L. & Fiorucci, S. Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food Chem.* **324**, 126864 (2020).

388. Zhong, M., Chong, Y., Nie, X., Yan, A. & Yuan, Q. Prediction of sweetness by multilinear regression analysis and support vector machine. *J. Food Sci.* **78**, (2013).

389. Zheng, S., Chang, W., Xu, W., Xu, Y. & Lin, F. e-Sweet: A machine-learning based platform for the prediction of sweetener and its relative sweetness. *Front. Chem.* **7**, 1–14 (2019).

390. Rojas, C. *et al.* A QSTR-based expert system to predict sweetness of molecules. *Front. Chem.* **5**, 1–12 (2017).

391. Chéron, J.-B., Casciuc, I., Golebiowski, J., Antonczak, S. & Fiorucci, S. Sweetness prediction of natural compounds. *Food Chem.* **221**, 1421–1425 (2017).

392. Goel, A., Gajula, K., Gupta, R. & Rai, B. In-silico prediction of sweetness using structure-activity relationship models. *Food Chem.* **253**, 127–131 (2018).

393. Huang, W. *et al.* BitterX: A tool for understanding bitter taste in humans. *Sci. Rep.* **6**, 1–8 (2016).

394. Margulis, E. *et al.* Intense bitterness of molecules: Machine learning for expediting drug discovery. *Comput. Struct. Biotechnol. J.* **19**, 568–576 (2021).

395. Zheng, S. *et al.* e-Bitter: Bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* **6**, 1–18 (2018).

396. Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B. & Shoombuatong, W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab133.

397. Charoenkwan, P. *et al.* iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **112**, 2813–2822 (2020).

398. Charoenkwan, P. *et al.* iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *Int. J. Mol. Sci.* **22**, 8958 (2021).

399. Rodgers, S., Glen, R. C. & Bender, A. Characterizing bitterness: Identification of key structural features and development of a classification model. *J. Chem. Inf. Model.* **46**, 569–576 (2006).

400. Dagan-Wiener, A. *et al.* Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.* **7**, 1–13 (2017).

401. Banerjee, P. & Preissner, R. Bitter sweet forest: A Random Forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Front. Chem.* **6**, 1–10 (2018).

402. Tuwani, R., Wadhwa, S. & Bagler, G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Sci. Rep.* **9**, 7155 (2019).

403. Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. *J. Chem. Inf. Model.* **60**, 6666–6678 (2020).

404. Charoenkwan, P. *et al.* UMpred-FRL: A new approach for accurate prediction of umami peptides using feature representation learning. *Int. J. Mol. Sci.* **22**, 13124 (2021).

405. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).

406. Zhang, S. Nearest neighbor selection for iteratively kNN imputation. *J. Syst. Softw.* **85**, 2541–2552 (2012).

407. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (Springer-Verlag, 2005). doi:10.1007/0-387-29362-0_23.

408. Ferreira, J. A. & Zwinderman, A. H. On the Benjamini–Hochberg method. *Ann. Stat.* **34**, 1827–1849 (2006).

409. Haynes, W. Wilcoxon Rank Sum Test. in *Encyclopedia of Systems Biology* 2354–2355 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_1185.

410. Bennasar, M., Hicks, Y. & Setchi, R. Feature selection using Joint Mutual Information Maximisation. *Expert Syst. Appl.* **42**, 8520–8532 (2015).

411. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003* vol. 3 523–528 (IEEE Comput. Soc, 2003).

412. Corthésy, J. *et al.* An Adaptive Pipeline To Maximize Isobaric Tagging Data in Large-Scale MS-Based Proteomics. *J. Proteome Res.* **17**, 2165–2173 (2018).

413. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

414. Hall, L. H. & Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. in vol. 2 367–422 (2007).

415. Hall, L. H. & Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **35**, 1039–1045 (1995).

416. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).

417. Nilakantan, R. *et al.* A family of ring system-based structural fragments for use in structure - Activity studies: Database mining and recursive partitioning. *J. Chem. Inf. Model.* **46**, 1069–1077 (2006).

418. der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).

419. Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S. & Kurata, H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.* **103**, 225–234 (2020).

420. Hasan, M. M. *et al.* Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* **22**, 1–16 (2021).

421. Mota-Merlo, M. & Martos, V. Use of machine learning to establish limits in the classification of hyperaccumulator plants growing on serpentine, gypsum and dolomite soils. *Mediterr. Bot.* **42**, e67609 (2021).

422. Michelucci, U., Sperti, M., Piga, D., Venturini, F. & Deriu, M. A. A Model-Agnostic Algorithm for Bayes Error Determination in Binary Classification. *Algorithms* **14**, 301 (2021).

423. Venturini, F. *et al.* Exploration of Spanish Olive Oil Quality with a Miniaturized Low-Cost Fluorescence Sensor and Machine Learning Techniques. *Foods* **10**, 1010 (2021).

424. Ahmad, A., Ordoñez, J., Cartujo, P. & Martos, V. Remotely Piloted Aircraft (RPA) in Agriculture: A Pursuit of Sustainability. *Agronomy* **11**, 7 (2020).

425. Martos, V., Ahmad, A., Cartujo, P. & Ordoñez, J. Ensuring Agricultural Sustainability through Remote Sensing in the Era of Agriculture 5.0. *Appl. Sci.* **11**, 5911 (2021).

426. Pallante, L. *et al.* On the human taste perception: Molecular-level understanding empowered by computational methods. *Trends Food Sci. Technol.* **116**, 445–459 (2021).

427. Czub, N., Pacławski, A., Szlęk, J. & Mendyk, A. Curated Database and Preliminary AutoML QSAR Model for 5-HT1A Receptor. *Pharmaceutics* **13**, 1711 (2021).

428. Besnard, P., Passilly-Degrace, P. & Khan, N. A. Taste of Fat: A Sixth Taste Modality? *Physiol. Rev.* **96**, 151–176 (2016).

429. Li, X. *et al.* Human receptors for sweet and umami taste. *Proc. Natl. Acad. Sci.* **99**, 4692–4696 (2002).

430. Carocho, M., Morales, P. & Ferreira, I. C. F. R. Sweeteners as food additives in the XXI century: A review of what is known, and what is to come. *Food Chem. Toxicol.* **107**, 302–317 (2017).

431. Bouysset, C., Belloir, C., Antonczak, S., Briand, L. & Fiorucci, S. Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food Chem.* **324**, 126864 (2020).

432. Bahia, M. S., Nissim, I. & Niv, M. Y. Bitterness prediction in-silico : A step towards better drugs. *Int. J. Pharm.* **536**, 526–529 (2018).

433. Malavolta, M. *et al.* A survey on computational taste predictors. *Eur. Food Res. Technol.* (2022) doi:10.1007/s00217-022-04044-5.

434. Rodgers, S., Glen, R. C. & Bender, A. Characterizing Bitterness: Identification of Key Structural Features and Development of a Classification Model. *J. Chem. Inf. Model.* **46**, 569–576 (2006).

435. Vapnik, V. N. *The Nature of Statistical Learning Theory*. (Springer New York, 1995). doi:10.1007/978-1-4757-2440-0.

436. Huang, W. *et al.* BitterX: a tool for understanding bitter taste in humans. *Sci. Rep.* **6**, 23450 (2016).

437. Dagan-Wiener, A. *et al.* Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.* **7**, 12074 (2017).

438. Dagan-Wiener, A. *et al.* BitterDB: taste ligands and receptors database in 2019. *Nucleic Acids Res.* **47**, D1179–D1185 (2019).

439. Wiener, A., Shudler, M., Levit, A. & Niv, M. Y. BitterDB: a database of bitter compounds. *Nucleic Acids Res.* **40**, D413–D419 (2012).

440. Rojas, C. *et al.* Quantitative structure–activity relationships to predict sweet and non-sweet tastes. *Theor. Chem. Acc.* **135**, 66 (2016).

441. Zheng, S. *et al.* e-Bitter: Bitterant Prediction by the Consensus Voting From the Machine-Learning Methods. *Front. Chem.* **6**, 82 (2018).

442. Banerjee, P. & Preissner, R. BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds. *Front. Chem.* **6**, 93 (2018).

443. Tuwani, R., Wadhwa, S. & Bagler, G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Sci. Rep.* **9**, 7155 (2019).

444. Fritz, F., Preissner, R. & Banerjee, P. VirtualTaste: a web server for the prediction of organoleptic properties of chemical compounds. *Nucleic Acids Res.* **49**, W679–W684 (2021).

445. Kursa, M. B., Jankowski, A. & Rudnicki, W. R. Boruta – A System for Feature Selection. *Fundam. Informaticae* **101**, 271–285 (2010).

446. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).

447. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017) doi:10.48550/ARXIV.1705.07874.

448. Polya, G. *Biochemical Targets of Plant Bioactive Compounds: A Pharmacological Reference Guide to Sites of Action and Biological Effects*. (CRC Press, 2003). doi:10.1201/9780203013717.

449. Burdock, G. A. *Fenaroli's Handbook of Flavor Ingredients*. (CRC Press, 2016). doi:10.1201/9781439847503.

450. Rojas, C. *et al.* A QSTR-Based Expert System to Predict Sweetness of Molecules. *Front. Chem.* **5**, 53 (2017).

451. Ahmed, J. *et al.* SuperSweet--a resource on natural and artificial sweetening agents. *Nucleic Acids Res.* **39**, D377–D382 (2011).

452. Chéron, J.-B., Casciuc, I., Golebiowski, J., Antonczak, S. & Fiorucci, S. Sweetness prediction of natural compounds. *Food Chem.* **221**, 1421–1425 (2017).

453. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).

454. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2**, 5 (2008).

455. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **10**, 4 (2018).

456. Temussi, P. A. The good taste of peptides: PEPTIDES TASTE. *J. Pept. Sci.* **18**, 73–82 (2012).

457. Naim, M., Rogatka, H., Yamamoto, T. & Zehavi, U. Taste responses to neohesperidin dihydrochalcone in rats and baboon monkeys. *Physiol. Behav.* **28**, 979–986 (1982).

458. Shin, W., Kim, S. J., Shin, J. M. & Kim, S.-H. Structure-Taste Correlations in Sweet Dihydrochalcone, Sweet Dihydroisocoumarin, and Bitter Flavone Compounds. *J. Med. Chem.* **38**, 4325–4331 (1995).

459. Bachmanov, A. A. *et al.* Genetics of Amino Acid Taste and Appetite. *Adv. Nutr. Int. Rev. J.* **7**, 806S-822S (2016).

460. Kawai, M., Sekine-Hayakawa, Y., Okiyama, A. & Ninomiya, Y. Gustatory sensation of l- and d-amino acids in humans. *Amino Acids* **43**, 2349–2358 (2012).

461. Schiffman, S. S., Clark, T. B. & Gagnon, J. Influence of chirality of amino acids on the growth of perceived taste intensity with concentration. *Physiol. Behav.* **28**, 457–465 (1982).

462. DuBois, G. E. Molecular mechanism of sweetness sensation. *Physiol. Behav.* **164**, 453–463 (2016).

463. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014) doi:10.48550/ARXIV.1412.6980.

464. Nadeau, C. Inference for the Generalization Error. *Mach. Learn.* **52**, 239–281 (2003).

465. Gregorutti, B., Michel, B. & Saint-Pierre, P. Correlation and variable importance in random forests. (2013) doi:10.48550/ARXIV.1310.5726.

466. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data Min. Knowl. Discov.* **2**, 86–97 (2012).

467. Justel, A., Peña, D. & Zamar, R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Stat. Probab. Lett.* **35**, 251–259 (1997).

468. Zheng, S., Chang, W., Xu, W., Xu, Y. & Lin, F. e-Sweet: A Machine-Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness. *Front. Chem.* **7**, 35 (2019).

469. Hall, L. H. & Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **35**, 1039–1045 (1995).

470. Rojas, C. *et al.* ChemTastesDB: A curated database of molecular tastants. *Food Chem. Mol. Sci.* **4**, 100090 (2022).

471. Freidman, J. H., Bentley, J. L. & Finkel, R. A. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Softw.* **3**, 209–226 (1977).

472. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **52**, 99–115 (1990).

473. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).

474. Jambon, M., Imberty, A., Deléage, G. & Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures: Detection of 3D Sites in Protein Structures. *Proteins Struct. Funct. Bioinforma.* **52**, 137–145 (2003).

475. Stark, A., Sunyaev, S. & Russell, R. B. A Model for Statistical Significance of Local Similarities in Structure. *J. Mol. Biol.* **326**, 1307–1316 (2003).

476. Kinoshita, K. Identification of Protein Biochemical Function by Searching the Similar Shape and Electrostatic Potential on the Molecular Surface of Proteins. *Seibutsu Butsuri* **44**, 150–154 (2004).

477. Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

478. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **33**, W337–W341 (2005).

479. Lisewski, A. M. & Lichtarge, O. Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Res.* **34**, e152–e152 (2006).

480. Nagarajan, D. & Chandra, N. PocketMatch (version 2.0): A parallel algorithm for the detection of structural similarities between protein ligand binding-sites. in *2013 National Conference on Parallel Computing Technologies (PARCOMPTECH)* 1–6 (IEEE, 2013). doi:10.1109/ParCompTech.2013.6621397.

481. Yeturu, K. & Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics* **9**, 543 (2008).

482. 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. *Nucleic Acids Res.* **36**, W260–W264 (2008).

483. Tseng, Y. Y., Dundas, J. & Liang, J. Predicting Protein Function and Binding Profile via Matching of Local Evolutionary and Geometric Surface Patterns. *J. Mol. Biol.* **387**, 451–464 (2009).

484. Tseng, Y. Y., Chen, Z. J. & Li, W.-H. f POP: footprinting functional pockets of proteins by comparative spatial patterns. *Nucleic Acids Res.* **38**, D288–D295 (2010).

485. Das, S., Krein, M. P. & Breneman, C. M. PESDserv: a server for high-throughput comparison of protein binding site surfaces. *Bioinformatics* **26**, 1913–1914 (2010).

486. Standley, D. M., Yamashita, R., Kinjo, A. R., Toh, H. & Nakamura, H. SeSAW: balancing sequence and structural information in protein functional mapping. *Bioinformatics* **26**, 1258–1259 (2010).

487. Moll, M., Bryant, D. H. & Kavraki, L. E. The LabelHash algorithm for substructure matching. *BMC Bioinformatics* **11**, 555 (2010).

488. Weill, N. & Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **50**, 123–135 (2010).

489. Konc, J. & Janezic, D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.* **40**, W214–W221 (2012).

490. Ito, J.-I., Tabei, Y., Shimizu, K., Tsuda, K. & Tomii, K. PoSSuM: a database of similar protein-ligand binding and putative pockets. *Nucleic Acids Res.* **40**, D541–D548 (2012).

491. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40**, W471–W477 (2012).

492. Lin, Y., Yoo, S. & Sanchez, R. SiteComp: a server for ligand binding site analysis in protein structures. *Bioinformatics* **28**, 1172–1173 (2012).

493. 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. *F1000Research* **2**, 117 (2013).

494. 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 Res.* **41**, W256–W265 (2013).

495. 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 Res.* **41**, W432–W440 (2013).

496. Gao, M. & Skolnick, J. APoc: large-scale identification of similar protein pockets. *Bioinformatics* **29**, 597–604 (2013).

497. Caprari, S., Toti, D., Viet Hung, L., Di Stefano, M. & Polticelli, F. ASSIST: a fast versatile local structural comparison tool. *Bioinformatics* **30**, 1022–1024 (2014).

498. Chartier, M., Adriansen, E. & Najmanovich, R. IsoMIF Finder: online detection of binding site molecular interaction field similarities. *Bioinformatics* **32**, 621–623 (2016).

499. Chartier, M. & Najmanovich, R. Detection of Binding Site Molecular Interaction Field Similarities. *J. Chem. Inf. Model.* **55**, 1600–1615 (2015).

500. Lee, H. S. & Im, W. G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design: G-LoSA. *Protein Sci.* **25**, 865–876 (2016).

501. 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. *J. Cheminformatics* **8**, 19 (2016).

502. Rey, J., Rasolohery, I., Tufféry, P., Guyon, F. & Moroy, G. PatchSearch: a web server for off-target protein identification. *Nucleic Acids Res.* **47**, W365–W372 (2019).

503. 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 Res.* **47**, W350–W356 (2019).

504. Simonovsky, M. & Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* **60**, 2356–2366 (2020).

505. Mullard, A. The drug-maker's guide to the galaxy. *Nature* **549**, 445–447 (2017).

Chapter VI

Appendix

6.1 Introduction to Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) that empowers computers to learn from data and algorithms, mimicking the way humans learn. It is closely related to computational statistics, which focuses on using computers to make predictions, and mathematical optimization, which connects models to the field of statistics. ML is particularly suited for real-world problems that involve high complexity and large amounts of data.

Learning, whether by humans or machines, involves acquiring new knowledge or modifying existing behaviours. While humans learn from experience, machines rely on data. ML enables computers to automatically improve their performance by learning from data and adjusting their actions accordingly. Through this process, ML models continuously enhance their capabilities and solve complex problems.

There are three main types of ML techniques: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, a model is trained using labelled data, where the desired outputs are already known. This trained model can then make predictions on new, unseen data. Unsupervised learning, on the other hand, involves training models without labelled data, aiming to discover patterns or relationships within the data. Lastly, reinforcement learning focuses on training models through a reward-based system, where the model learns optimal actions by interacting with an environment and receiving feedback on its performance.

By leveraging these ML techniques, computers can autonomously learn from data, make predictions, and continually improve their performance, offering great potential for solving a wide range of real-world problems.

Supervised learning can be further divided into classification and regression tasks. Classification aims to predict the labels or categories of new instances based on previous observations. The predicted labels are categorical values that determine the membership of an instance in a specific group. Regression, on the other hand, deals with predicting continuous outcomes. In binary classification, the algorithm learns to distinguish between two possible classes. This decision function is learned by the supervised learning algorithm. In linear regression, the goal is to find a relationship between predictive variables and a continuous target variable to make predictions. Unsupervised learning, on the other hand, involves exploring the structure of data to extract meaningful information without the use of labels or a reward function. Clustering is a technique used in unsupervised learning to organize a large amount of information into meaningful subgroups called clusters, without any prior knowledge of their group memberships. Objects within the same cluster exhibit a certain level of similarity, while being distinct from objects in other clusters.

Machine learning algorithms generally follow a common workflow structured in multiple phases. The *pre-processing phase* is an essential step because raw data is often not in the desired format or uniformity required for optimal operation of learning algorithms. This phase involves transforming features into a standardized range, such as [0,1], or into a standard normal distribution with a mean of zero and unit variance. Additionally, feature reduction techniques may be applied to eliminate redundant or highly correlated features, aiming to improve the performance of the model. During this phase, the data is typically divided into a training set and a test set. The training set is used to optimize the algorithm, while the test set is used to evaluate the generalization capability of the trained model. In the *learning phase*, different algorithms are compared based on their performance using a chosen metric. Cross-validation is often employed to further divide the dataset into training and validation sets, enabling the estimation of the model's generalization capabilities. Model tuning, involving the adjustment of hyperparameters, is also conducted during this phase to enhance the model's performance. Finally, the *evaluation phase* utilizes the test set to estimate the model's performance on unseen data. If the model demonstrates satisfactory performance, it can be deployed to predict new data.

6.1.1 Pre-processing Phase

When working with machine learning in real-life applications, it is essential to address the issue of *missing values*. Failure to handle missing data appropriately can lead to unpredictable outcomes. The simplest approach is to remove features or training examples with missing values from the dataset. However, this can result in the loss of valuable information and potentially diminish the effectiveness of the model. To mitigate this, imputation techniques are employed to estimate or fill in the missing values. For continuous features, a common method is to substitute the

missing values with the mean or median of the available data within the same feature. In the case of binary features, the missing values can be replaced with the most frequently observed value among the non-missing instances. By using these imputation techniques, missing data can be managed effectively, allowing for robust analysis and modelling. Nevertheless, it is important to acknowledge that imputation introduces some level of uncertainty and potential bias, as the missing values are estimated based on the available data. Therefore, careful consideration and evaluation of the chosen imputation approach are paramount.

In the pre-processing stage, special consideration needs to be given to nominal characteristics, which are characterized by the absence of a specific order. Integer encoding alone is inadequate in handling such data, as it assumes a natural hierarchy among categories, leading to suboptimal performance. To avoid imposing a hierarchical order, the *One-Hot Encoding* technique can be employed. This encoding creates a binary feature for each possible category, assigning a value of 1 to the corresponding feature for each sample in its original category. However, it is important to note that this encoding can introduce multicollinearity, which can cause issues for certain methods that rely on matrix inversion.

Another crucial step in pre-processing is the *partitioning of the dataset*. The available dataset is divided into a training set and a test set. The model is trained using the training set, and then predictions are made on the test set to evaluate the model's performance on unseen data. Determining the appropriate size of the test set is a challenging decision, as a smaller test set may result in less accurate estimation of the model's generalization error.

Feature scaling is also an essential step in the pre-processing phase. Scaling ensures that all features have a comparable range and distribution, which is particularly important for algorithms that are sensitive to differences in feature scales. Common scaling techniques include normalization or standardization, which transform the features to a common scale, such as [0,1] or a standard normal distribution with mean zero and unit variance. To ensure optimal performance of machine learning algorithms, except for decision trees and random forests, it is beneficial to have features that are on the same scale. This can be achieved through two common approaches: *normalization* and *standardization*. In the case of normalization, the aim is to scale the features within a specific range, typically [0,1]. The scaling process involves transforming the values of each feature column according to the following procedure:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

where $x^{(i)}$ is a given example's feature, x_{min} is the lowest value of the column of features and x_{max} the highest one.

On the other hand, the standardization approach is particularly beneficial for linear models that initialize weights at zero. It involves centering the columns of features around zero and scaling them to have a standard deviation of one. This process aids in the updating of weights during the learning phase. The standardization procedure for each example is as follows:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

where μ_x is the sample mean considering a certain column of features and σ_x is the related standard deviation.

In the pre-processing phase, *dimensionality reduction* plays a crucial role in mitigating model complexity and preventing overfitting. Dimensionality reduction can be achieved through feature selection and feature extraction techniques. Feature selection involves selecting a subset of the original features. Sequential Backward Selection (SBS) is an approach where features are sequentially removed based on the least degradation in performance after their removal. This process continues until the desired number of features is reached. Another approach for selecting relevant features is feature importance, which can be evaluated using techniques like random forest. Feature extraction, on the other hand, involves creating a new subset of features by extracting information from the original dataset. This is done by projecting the data into a new feature space. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are commonly used linear techniques for feature extraction. LDA is a supervised technique, while PCA is unsupervised. In the case of non-linear problems, Kernel Principal Component Analysis (KPCA) is employed. KPCA performs a non-linear mapping of the data into a higher dimensional space, and then PCA is applied in this new space to project the data into a lower-dimensional space, where a linear classifier can be applied.

Class Imbalance Problem

Real-world datasets often exhibit class imbalance, particularly in medical fields where the minority class represents positive events that are of significant interest. Handling class imbalance can be challenging when building an effective model. Although accuracy can appear high when the majority class is well classified, it may not provide a comprehensive performance evaluation.

One approach to address class imbalance is through resampling techniques. Two common strategies are oversampling the minority class or undersampling the majority class. In the case of oversampling, the samples from the minority class are duplicated until the two classes have an equal number of samples. However, this method can be crude as it does not introduce new information to the model,

potentially leading to overfitting. Additionally, oversampling may identify similar and potentially overrepresented regions in the feature space.

Another approach is data augmentation, which involves generating synthetic samples from existing ones. Synthetic Minority Oversampling Technique (SMOTE) is a widely used method in data augmentation. SMOTE operates in the feature space rather than the data space, unlike traditional resampling techniques. It selects an example from the minority class and randomly chooses one of its neighboring examples from the same class. The distance between the two examples is calculated and multiplied by a random number between 0 and 1. The result is added to the initial distance, generating a synthetic example. This process continues until the two classes have an equal number of examples, providing a more balanced dataset for training the model.

6.1.2 Learning Phase

Here we present the best practices for model optimization. The model is initially trained on a dedicated training dataset, and its performance is then evaluated on unseen data. One of the key challenges in this process is *overfitting*, where the model becomes overly complex (with a high number of parameters) and fails to generalize well to test data, resulting in high variance. On the other hand, underfitting is also a significant concern, where the model lacks the necessary complexity to effectively capture patterns in the training data, leading to high bias. Variance measures the variability of the model's predictions when trained on different datasets, while bias represents the systematic error unrelated to randomness. To strike a suitable balance between bias and variance, it is essential to assess the model's generalization capabilities using techniques such as the holdout method and k-fold cross-validation. With the holdout method, the dataset is divided into a training set and a test set. Ideally, it is recommended to further partition the data into three parts: a training set, a validation set, and a test set. The training set is used to configure various models, while the validation set is used to evaluate the model's performance repeatedly after training, considering different hyperparameter values. Once satisfactory hyperparameters are determined, the model's generalization performance is estimated on the test set, providing a more reliable estimate of its effectiveness. In *k-fold cross-validation*, the training dataset is randomly divided into k non-overlapping folds, without replacement. In each iteration, k-1 folds are used for training the model, while the remaining fold is used for performance evaluation. This process is repeated k times, resulting in k models and corresponding performance estimates. The average performance across the k models is then calculated, using independent test folds, to obtain a performance estimate that is less sensitive to the specific partitioning of the data. K-fold cross-validation is a valuable technique for model optimization. Once the optimal hyperparameters are determined, the model is retrained on the entire training dataset, providing a final performance estimate using an independent test dataset.

This approach offers the advantage of ensuring that each example in the training dataset is used once for both training and validation, resulting in a low variance estimate of the model's performance.

Optimal parameters for the model can be find also using the *grid search technique,* which is employed to find the optimal combination of hyperparameters from a predefined list, with the goal of maximizing the model's performance. However, a drawback of this technique is its computational cost, as it involves evaluating the performance of all possible combinations of hyperparameters. This process can be time-consuming and resource-intensive, especially when dealing with large hyperparameter spaces.

Nested cross-validation is a technique that combines k-fold cross-validation with grid search to select the best machine learning algorithm among several options. It is particularly useful when choosing between different algorithms. The process involves an external k-fold cross-validation loop that splits the data into training and test folds. Within this loop, there is an internal loop that performs k-fold cross-validation on the training fold to select the optimal model. The performance of each algorithm is evaluated and compared using the inner cross-validation loop, and the best-performing algorithm is selected based on the results. This nested approach helps prevent overfitting and provides a more reliable estimate of the model's performance on unseen data.

6.1.3 Evaluation and Prediction Phases

The evaluation phase aims to assess the effectiveness of the obtained model. It is important to select appropriate metrics based on the specific properties being analyzed. Merely calculating the metrics is insufficient; it is essential to ensure that the results align with the domain of interest. The interpretation of the results is a crucial aspect of this phase. Metrics are often grouped together to facilitate a comprehensive evaluation.

In machine learning, one commonly consulted metric is the confusion matrix, which is a square matrix that presents the counts of *true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP)*. True negatives represent correctly identified negative cases, while true positives represent correctly identified positive cases. False negatives are positive cases identified as negative, and false positives are negative cases identified as positive. In medical contexts, a value of 1 is typically associated with positivity to an event.

Accuracy is a metric that quantifies the overall mistakes made by the classifier and is defined as follows:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

The complementary metric to accuracy is the error metric. The accuracy metric, however, fails to differentiate between false positives (FP) and false negatives (FN). While this lack of distinction is acceptable in datasets with a balanced distribution of both classes, it becomes problematic in unbalanced datasets. This is particularly relevant in medical classification, where the minority class represents positivity to an event. In the case of unbalanced datasets, the true positive rate (TPR) and false positive rate (FPR) metrics are more informative. These metrics are defined as follows:

$$FPR = \frac{FP}{FP + TN}$$
$$TPR = \frac{TP}{FN + TP}$$

Precision and recall are metrics that provide insights into the number of true positives (TP) and true negatives (TN). Precision evaluates how accurately the model classifies examples as positives, while recall assesses how effectively the model identifies the positive examples that should have been classified as such:

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

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

The receiver operating characteristic (ROC) curves are valuable tools for comparing and selecting models based on their performance. Specificity is one of the measures used in ROC analysis:

$$Specificity = \frac{TN}{FP + TN}$$

The ROC curve is obtained for binary classification by plotting the false positive rate (1-specificity) on the x-axis and the true positive rate (sensitivity) on the y-axis. These metrics are calculated for different threshold values in the range [0, 1].

For example, if a threshold of 0.5 is used, samples with a predicted output greater than or equal to 0.5 are classified as positive, while samples with a predicted output less than 0.5 are classified as negative. This threshold value represents one point on the ROC curve.

By varying the threshold, different points on the ROC curve are obtained, reflecting the trade-off between sensitivity and specificity. The ROC curve provides a visual representation of the model's performance across different threshold values, allowing for the comparison of different models and the selection of an optimal operating point.

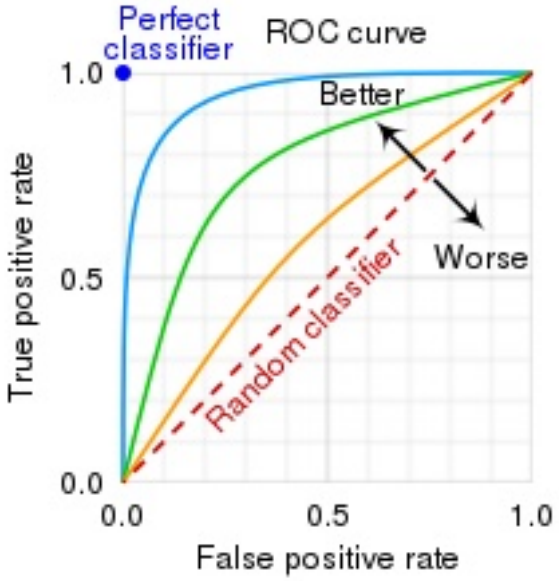


Figure A - 6.1.1. ROC Curve, Ideal Classifier, Real Classifier, Random Guess

Figure A - 6.1.1 illustrates the performance of different classifiers using the ROC curve. The random guess classifier (orange) represents the worst-case scenario, where the area under the curve (AUC) is 0.5. This indicates that the classifier performs no better than random chance. On the other hand, the ideal classifier (red) represents the optimal case, where there are no classification errors, and the AUC is equal to 1. This signifies perfect discrimination between positive and negative samples. Real classifiers (blue) should aim to achieve results that are closer to the ideal case or at least fall between the best and worst cases. The AUC values for these classifiers will lie between 0.5 and 1, indicating their performance relative to random guessing and the ideal classifier.

6.1.4 Examples of Classifiers

Several methods and classifiers have been proposed in the past years for machine learning applications. In this section, we will provide a brief overview of some of the most used machine learning classifiers. Table A - 6.1.1 provides a quick overview of most used ML algorithms, whereas following sections will describe more in detail some of the models used in the present PhD thesis.

Table A - 6.1.1. Different ML algorithms

Logistic regression	Logistic regression is a commonly employed linear method, frequently utilized in medical applications. It predicts the probability of an object belonging to a specific class, making it suitable for binary classification tasks. Logistic regression is known for its ability to avoid overfitting, distinguishing it from other linear models, which predicts continuous values. The output of logistic regression is a probability score that

	quantifies the likelihood of an object belonging to a particular class. This probability is obtained by applying a logistic function (sigmoid function) to a linear combination of the input features. The logistic function ensures that the predicted probabilities fall within the range of [0, 1], making it suitable for class probability estimation.
K-nearest neighbours (KNN)	K-Nearest Neighbors (KNN) is a supervised learning algorithm used for classification and regression tasks. It operates based on the principle that objects with similar characteristics tend to belong to the same class or have similar outcomes. When applying KNN, the algorithm considers the k nearest training examples in the feature space to the new test data. The distance between the query data and the training samples is calculated using a distance metric, commonly the Euclidean distance. KNN is known for its simplicity and ease of implementation, making it a popular choice in various domains. It can be applied to both classification and regression problems, where it predicts the class membership or the numerical value of a target variable, respectively.
Support Vector Machine (SVM)	Support Vector Machine (SVM) is a versatile algorithm for classification and regression tasks. It constructs hyperplanes to separate classes, maximizing the margin between them. SVM handles linear and non-linear problems using different kernels. It's robust against noise and outliers, focusing on support vectors near the decision boundary. SVMs handle high-dimensional data effectively, even with more features than samples. They offer interpretability by examining support vectors. However, SVMs can be computationally expensive, requiring careful parameter tuning. Overall, SVMs provide accurate predictions, generalization, and insights into various data types.
Random forest	Random Forest is an algorithm that combines multiple decision trees by utilizing the bootstrap method. It constructs an ensemble of decision trees, where each tree is trained on a randomly selected subset of the original data. This process involves creating multiple examples of the same size as the original dataset through random sampling with replacement. Each decision tree in the Random Forest is built independently, making predictions based on a subset of features at each node. The final prediction is determined by aggregating the predictions of all the individual trees, either through majority voting (for classification tasks) or averaging (for regression tasks). Random Forest is known for its high accuracy and robustness. By combining the predictions of multiple trees, it reduces the risk of overfitting that can occur with a single decision tree. The ensemble nature of Random Forest allows it to capture complex relationships in the data and make accurate predictions. However, it is important to note that Random Forest can still be susceptible to overfitting, especially if the number of trees in the ensemble is too high or if the individual trees are allowed to grow too deep. Careful parameter tuning, such as limiting the depth of the trees or controlling the number of features used at each split, can help mitigate this issue. Overall, Random Forest is a powerful algorithm that provides

	a balance between accuracy and generalization. Its ability to handle high-dimensional data and capture nonlinear relationships makes it a popular choice in various domains.
Adaptive boosting (AdaBoost)	AdaBoost, also known as Adaptive Boosting, is a statistical classification meta-algorithm that can be utilized in conjunction with other learning algorithms. Its main objective is to improve the predictive power of the model by focusing on training data samples that are most valuable for classification. Unlike traditional training methods that consider the entire dataset, AdaBoost selectively incorporates samples that contribute to enhancing the model's performance. By doing so, it optimizes the execution time by avoiding the calculation of irrelevant subsets of the original dataset. This selective approach allows AdaBoost to allocate more attention to the samples that are crucial for improving the overall accuracy of the model. By iteratively training weak learners and adjusting the weights assigned to each sample, AdaBoost effectively combines the outputs of multiple weak classifiers to create a stronger ensemble classifier. This iterative process ensures that the subsequent classifiers focus on the misclassified samples from previous iterations, resulting in a refined and more accurate final model. Overall, AdaBoost is a powerful technique that can significantly enhance the performance of learning algorithms by selectively utilizing informative samples, thereby improving execution time and predictive accuracy.
Extreme Gradient Boosting (XGBoost)	XGBoost (Extreme Gradient Boosting) is a powerful gradient boosting algorithm widely used for supervised learning tasks. It builds an ensemble of weak decision tree models in a sequential manner, optimizing a loss function through gradient descent. XGBoost excels in handling both classification and regression problems, demonstrating excellent predictive performance. It effectively handles missing data and supports various objective functions and evaluation metrics. XGBoost incorporates regularization techniques to prevent overfitting, such as shrinkage and column subsampling. It allows for parallel processing and offers flexibility in customizing the boosting process. Despite its computational complexity, XGBoost is known for its scalability and efficiency. It is widely regarded as a top-performing algorithm in various machine learning competitions and real-world applications.
Probabilistic models (Naïve Bayes)	Naïve Bayes is a probabilistic model commonly used for classification tasks. It operates based on the principles of Bayes' theorem and conditional probability. Naïve Bayes assumes that the features are conditionally independent given the class, which simplifies the modelling process. In the context of Naïve Bayes, the model demonstrates the existence of a specific type of mathematical object by showing that if objects are randomly selected from a certain class, the probability of obtaining an object of the desired type is significantly greater than zero. The model uses probability calculations to make this determination. It is important to note that even though probability is used

	in the model, the conclusion reached by Naïve Bayes is considered certain and without error. This makes Naïve Bayes a powerful and reliable method for classification tasks.
Deep learning (neural networks)	Deep learning and neural networks are versatile statistical models inspired by the structure of biological neural networks. Composed of interconnected artificial neurons, it learns to extract complex patterns from data. With hierarchical representations, it captures simple to abstract concepts. It excels in image classification, language processing, and more. Deep learning models are trained using optimization algorithms to minimize prediction errors. They require labelled data and computational resources. Once trained, they make accurate predictions on new data. Deep learning has transformed fields like healthcare, finance, and robotics, achieving state-of-the-art performance. Advancements in hardware and algorithms continue to expand its capabilities in understanding and processing complex information.

K-Nearest Neighbour

The K-nearest neighbour (KNN) classifier is a straightforward machine learning algorithm known for its simplicity. It falls under the category of lazy learning systems, as it does not learn a specific decision function from training data. Instead, it relies on the proximity of labelled examples to make predictions. KNN is considered a non-parametric model since it doesn't have a fixed number of parameters and instead adapts based on the number of training data points. Specifically, KNN is classified as an instance-based learning model, a subset of non-parametric models. The main steps of KNN involve selecting a value for K, which represents the number of nearest neighbours to consider, and a distance metric, such as the Euclidean distance. Next, KNN identifies the K nearest neighbours to the subject being classified. The label of the subject is then determined through majority voting among its neighbours. Figure A - 6.1.2 represents the classification method employed by KNN.

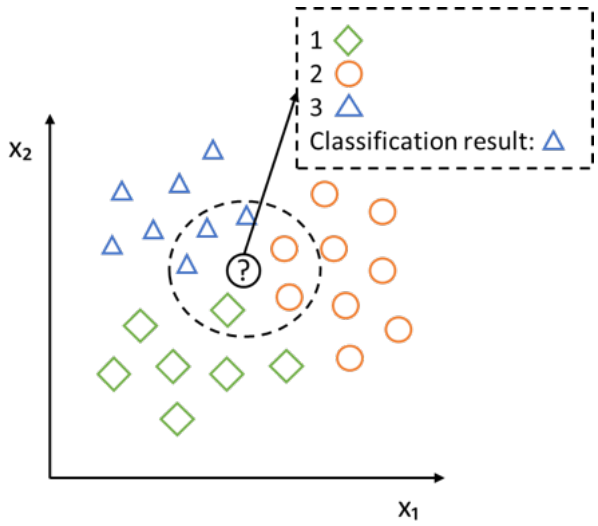


Figure A - 6.1.2. A practical example of classification by means of k-nearest neighbor classifier.

The memorization-based approach of the K-nearest neighbor (KNN) classifier offers the advantage of immediate adaptation to new data. As new instances are collected, the classifier can readily incorporate them into its decision-making process. However, this approach also has its drawbacks. One such limitation is the computational complexity involved in classifying new instances. As the number of training samples increases, the time required for classification grows linearly, which can become a challenge for large datasets. Additionally, the space required to store the memorized training samples can become an issue, particularly in datasets with high dimensions. To mitigate these challenges, various implementations of KNN have been developed, some of which leverage efficient data structures like KD-trees[471].

Choosing the right value of K in K-nearest neighbour (KNN) classification is a crucial step in balancing the trade-off between underfitting and overfitting. A small value of K may result in overfitting, as the decision boundary becomes too sensitive to local variations in the training data. On the other hand, a large value of K may lead to underfitting, as it considers a broader neighbourhood and may overlook important local patterns.

In addition to selecting the optimal K, choosing an appropriate distance metric is essential for accurate classification. The Euclidean distance is a commonly used measure in KNN, particularly for continuous feature spaces. It calculates the straight-line distance between two data points in a multidimensional space. However, it's important to note that the Euclidean distance assumes that all features are equally important, and that the data is linearly separable.

For datasets with different characteristics, alternative distance metrics such as Manhattan distance (city block distance) or Minkowski distance can be considered. These metrics allow for different degrees of emphasis on specific features or can handle specific data structures more effectively. It is crucial to evaluate the dataset's nature and domain knowledge to determine the most suitable distance metric for achieving optimal performance in the KNN classifier.

K-nearest neighbor (KNN) algorithm can indeed be susceptible to overfitting, especially when the feature space becomes sparse and the nearest neighbors are located far apart, leading to instability in the estimation. One effective approach to mitigate this issue is to reduce the dimensionality of the feature space.

Support Vector Machine

Support Vector Machine (SVM) classifier can indeed be viewed as an extension of the perceptron algorithm, one of the earliest machine learning algorithms used for classification tasks. However, SVM introduces several important enhancements and concepts that differentiate it from the perceptron[472,473]. In contrast to the perceptron, which focuses solely on minimizing misclassification errors, SVM pursues a distinct optimization objective of maximizing the margin. The margin, denoting the separation between the decision boundary (hyperplane) and the nearest data points, referred to as support vectors, is a critical component in SVM. It is defined as the distance between the decision function and these support vectors. This emphasis on margin optimization enables SVM to effectively discriminate between classes and enhance generalization performance, making it particularly advantageous when seeking clear class separation. Figure A - 6.1.3 graphically shows the above-mentioned concept.

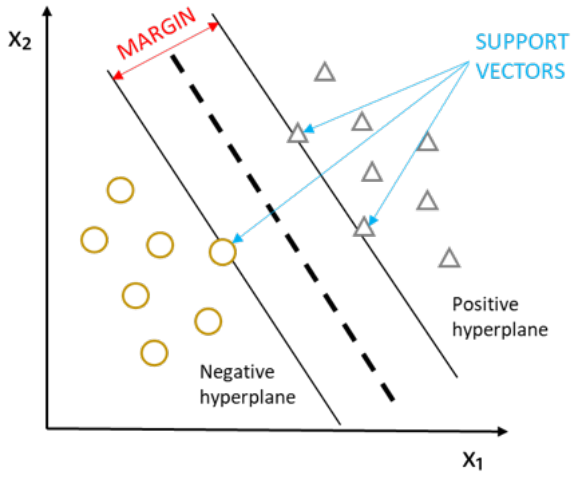


Figure A - 6.1.3 Margin maximization of the SVM model

Decision functions characterized by large margins exhibit a reduced generalization error, whereas models featuring small margins are prone to overfitting. The positive and negative hyperplanes, aligned parallel to the decision function, can be elucidated as follows:

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{pos}} = 1$$

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{neg}} = -1$$

If these functions are subtracted:

$$\boldsymbol{w^T}\left(\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}}\right) = 2$$

Normalizing the equation for **w** length, which is defined as follows:

$$\left|\left|\boldsymbol{w}\right|\right| = \sqrt{\sum_{j=1}^{m} {w_j}^2}$$

we obtain:

$$\frac{\boldsymbol{w^T}\left(\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}}\right)}{||\boldsymbol{w}||} = \frac{2}{||\boldsymbol{w}||}$$

The left-hand side of the equation can be construed as the margin, representing the distance between the positive and negative hyperplanes, which is sought to be maximized. Hence, the objective function of the Support Vector Machine (SVM) transforms into the maximization of 2 divided by the magnitude of vector w, subject to the constraint that the samples are classified correctly. This constraint can be expressed as follows:

$$w_0 + \boldsymbol{w}^T \boldsymbol{x^{(i)}} \geq 1 \; if \; y^{(i)} = 1$$

$$w_0 + \boldsymbol{w}^T \boldsymbol{x^{(i)}} < -1 \; if \; y^{(i)} = -1$$

The two equations express the requirement that all negative samples should lie on one side of the negative hyperplane, while all positive samples should be positioned on the other side of the positive hyperplane. This condition ensures a clear separation between the two classes in the feature space. By defining such strict boundaries, the SVM classifier aims to maximize the margin between the hyperplanes, leading to improved classification performance and reduced misclassification errors. This formulation emphasizes the importance of achieving a distinct separation between the classes to enhance the discriminative power of the

SVM model. Practically, minimizing $\frac{1}{2}||\boldsymbol{w}||^2$ is simpler, by means of quadratic programming.

In cases where the dataset proves to be particularly challenging, a nonlinear variant of the Support Vector Machine (SVM), known as kernel SVM, can be employed to address it. The underlying concept is that the original input space can always be transformed or mapped to a higher-dimensional feature space where the training set becomes separable. This mapping allows for the creation of new nonlinear combinations of the original features, enabling improved classification performance (Figure A - 6.1.4).

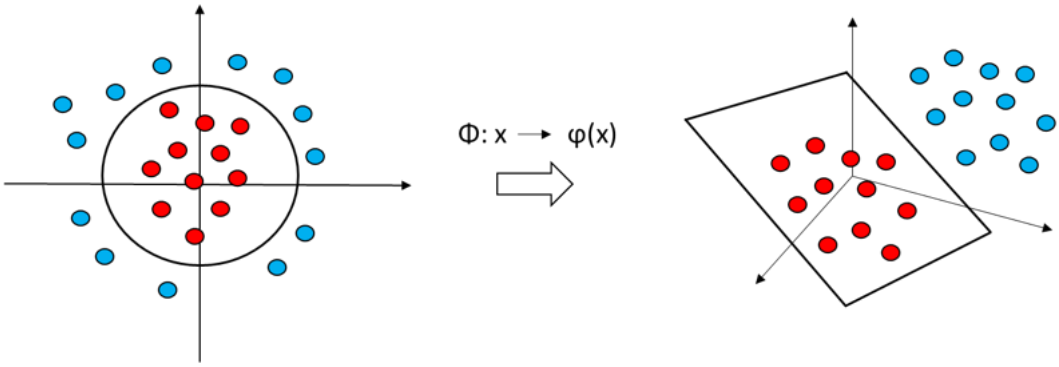


Figure A - 6.1.4. Schematic representation of the kernel function, Φ, operation used for SVM classification. On the left, the original features space is represented, whereas, on the right, the new features space after kernel transformation is shown.

In practice, the training dataset is effectively mapped to a higher-dimensional space using a kernel function Φ. This mapping is performed implicitly, meaning that the actual transformation is not explicitly computed, thereby minimizing the computational burden. The kernel functions enable the SVM algorithm to capture complex relationships and non-linear decision boundaries by introducing new features and representations. Once the dataset has been implicitly transformed, a linear SVM is trained on this new feature space. This training process aims to find an optimal linear decision boundary that effectively separates the classes in the transformed space. This linear SVM model can then be used to classify new, unseen data points by applying the same Φ transformation to these data points and using the trained SVM to make predictions. The choice of the kernel function is crucial and depends on the specific characteristics of the dataset and the problem at hand. There are various options available, such as the polynomial kernel, radial basis function (RBF) kernel, sigmoid kernel, among others. The RBF kernel, also known as the Gaussian kernel, is particularly popular due to its ability to capture complex and non-linear relationships effectively.

SVM offers notable advantages for classification tasks. One key benefit is its robust kernel framework, which enables capturing complex relationships and non-linear

patterns in data. This flexibility allows SVMs to effectively model various data distributions and decision boundaries. Moreover, SVMs demonstrate strong performance even with small training sets, as they heavily rely on support vectors, which contribute to their reliability and generalization. Nevertheless, there are limitations to consider. The computational cost of SVMs can be high for large-scale problems, necessitating substantial time and resources for training and classification. Additionally, parameter tuning for the chosen kernel function, such as determining the width of the Gaussian kernel in the case of RBF, can pose challenges. Finding the optimal parameter values often requires careful experimentation.

In summary, SVMs provide powerful classification capabilities, particularly in capturing non-linear relationships. However, one must be mindful of the computational complexity and the careful selection of kernel parameters, which can impact performance.

Decision Tree

Decision tree (DT) classifiers are effective models for achieving interpretability and making decisions. They classify data by sequentially posing a series of questions and using the answers to guide the classification process. An illustrative example of a simple decision tree is depicted in Figure A - 6.1.5.

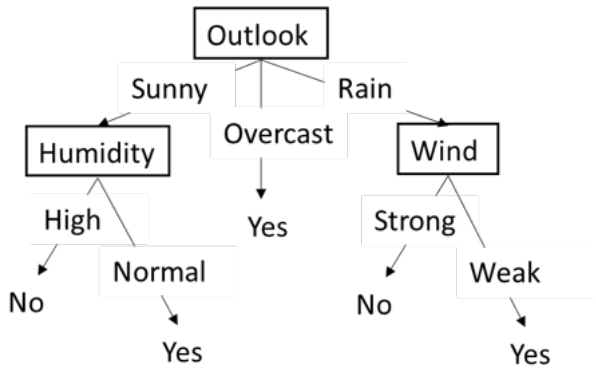


Figure A - 6.1.5 Example of a simple decision tree

Decision trees (DTs) exhibit a structure resembling a flowchart, consisting of nodes, branches, and leaves. Each node represents a test involving an attribute (i.e., a feature), while each branch emanating from a node signifies a possible outcome corresponding to a specific value of the attribute. The leaves of the tree denote nodes containing the final class label. Classification entails a sequential execution of tests, starting from the root node and terminating at a leaf node. To automatically construct DTs from data, induction methods are employed. One such approach is

the Top-Down Induction of Decision Trees, which encompasses a range of techniques for inducing DTs from a given dataset. Among these methods, ID3, developed by Quinlan, is an iterative algorithm that operates within a top-down framework. It facilitates the construction of DTs by recursively partitioning the dataset based on the most informative attributes. During the construction process, ID3 selects attributes that result in the greatest information gain, aiming to maximize the discriminatory power of the resulting tree. This iterative procedure generates a DT that effectively captures the patterns and relationships present in the training data, facilitating accurate classification of unseen instances.

The pseudo-code to construct a DT (T) from a learning set (S) is:

- If all examples in S belong to the same class C, then make a leaf labelled C;
- Otherwise
    - Select the "most informative" attribute (A)
    - Partition S according to A's values
    - Recursively construct subtrees T1, T2, … for the subsets of S (one for each value of A)

Choosing the most informative attribute involves selecting the attribute that partitions the dataset into subsets that exhibit the greatest homogeneity in terms of class labels. The process of classification requires a certain amount of information (I), and after applying attribute A, only a residual amount of information (I_{res}) is necessary to classify the object. The information gain is defined as the difference between the initial amount of information and the residual amount:

$$Gain(A) = I - I_{res}(A)$$

The most informative attribute is the one that maximizes the gain.

Entropy is defined as the averaged amount of information needed to classify an object:

$$I = -\sum_{p} p(c) log_2 p(c)$$

where p(c) is the proportion of samples belonging to class c. When all samples in a dataset belong to the same category, the entropy is minimized and equals zero. On the other hand, if the samples are evenly distributed across multiple categories, with each category containing 1/c samples (where c is the number of classes), the entropy is maximized and equals one. After applying attribute A, the dataset S is partitioned

into subsets based on the different values (v) of attribute A. Residual information is equal to the weighted sum of the amounts of information for the subsets:

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) log_2 p(c|v)$$

One way to measure the homogeneity of attributes is through information gain, where the attribute with the highest gain is considered the most informative. However, a limitation of this approach is that it tends to favor attributes with a larger number of values, leading to a bias towards such attributes. To address this limitation, a corrected measure called information gain ratio is used. This measure is obtained by dividing the gain of attribute A by its intrinsic information. By using information gain ratio, a more balanced evaluation of attributes can be achieved, considering both their gain and their inherent information.

$$GainRatio(A) = \frac{Gain(A)}{I(A)} = \frac{I - I_{res}(A)}{I(A)}$$

Another sensible measure of impurity is *Gini Index*:

$$Gini = \sum_{i \neq j} p(i)p(j),$$

where p(i) is the proportion of examples of a class and i and j are the classes. Gini is the measure of the initial information. After applying the attribute A, the resulting Gini Index is:

$$GiniGain(A) = \sum_{v} p(v) \sum_{i \neq j} p(i|v)p(j|v),$$

where p(i|v) is the information of a subset provided that a specific path has been followed. In summary, the measure of residual information in terms of GiniGain (A) is used to evaluate the attribute's contribution. Gain(A) is derived from entropy, while GainRatio (A) considers the heterogeneity between features. GiniGain (A) considers the probability of misclassifying an object.

Decision trees (DTs) have several strengths, including their ease of implementation and understanding. However, they are prone to overfitting and require data discretization due to the generation of complex decision regions. These weaknesses should be taken into consideration when utilizing DTs in practice.

Random Forest

The Random Forest (RF) classifier is an ensemble method specifically designed for decision tree (DT) classifiers. The concept behind using an ensemble of learning methods is to combine multiple weak learners to create a stronger learner with

improved generalization and reduced susceptibility to overfitting. The RF algorithm follows these main steps:

1. Randomly select n training set samples with replacement (bootstrap sample).
2. Construct a decision tree from the bootstrap sample. At each node:
    a) Randomly select d features without replacement.
    b) Divide the node based on the feature that yields the best division, determined by the objective function.
3. Repeat steps 1 and 2 k times.
4. Combine the predictions from each decision tree and assign the final label using majority voting.

In the second step, instead of evaluating functions on all features to determine the best split at each node, only a random subset of features is considered. While RFs may not offer the same level of interoperability as DTs, they provide significant advantages. Hyperparameter optimization is less demanding, and the model demonstrates resilience to noise.

AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble method that falls within the category of classifiers. Ensemble methods are designed to combine multiple individual classifiers into a meta-classifier that provides superior generalization performance compared to each individual classifier on its own.

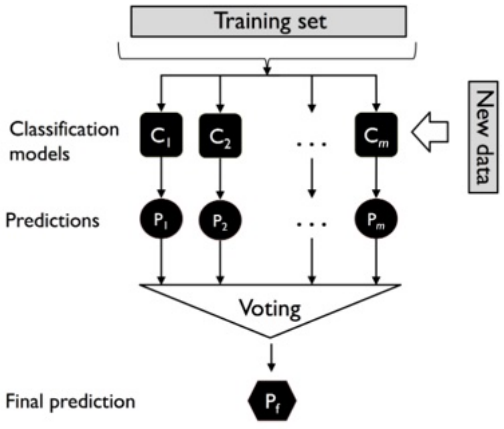


Figure A - 6.1.6. Majority Voting in Ensemble Learning.

Figure A - 6.1.5 illustrates the concept of majority voting, where the label of the class predicted by most of the classifiers is chosen. The ensemble can be constructed

using various classification algorithms or by configuring the same base algorithm with different subsets of the training data. AdaBoost, specifically, utilizes the boosting technique, which involves incorporating weak learning models into the ensemble. These models typically exhibit only a slight performance advantage over random selection. Boosting focuses on challenging training examples by allowing the weak learners to learn from misclassified examples in a staged manner, aiming to enhance the overall performance of the ensemble. The boosting procedure can be summarized as follows:

1. Extraction from the training dataset D of a random subset from the training examples d_1, without reinsertion, to train a weak learning system C_1.
2. Extraction from the training dataset of a second random training subset d_2 without reinsertion and addition of the 50% of the examples that were previously misclassified to train a weak learning system C_2.
3. Identification of the training examples, d_3, in the training dataset D on which C_1 and C_2 did not agree to train a third weak learning system C_3.
4. Combination of the weak learning models C_1, C_2, C_3 by majority vote.

AdaBoost is known for having low bias but a tendency to overfit the training data. In contrast to traditional boosting techniques, AdaBoost uses the entire training dataset to train weak learning models. At each iteration, the training examples are reweighted based on their classification performance, placing more emphasis on the misclassified examples. This iterative process allows AdaBoost to construct a strong classifier that learns from the errors made by the previous weak ensemble learning models. By continuously adjusting the weights of the training examples, AdaBoost focuses on improving its ability to correctly classify difficult instances, ultimately enhancing the overall performance of the classifier. In Figure A - 6.1.7, the process of training an AdaBoost classifier for binary classification is illustrated. In box 1, we start with a training dataset where all the examples have equal weights. The initial model is trained using this dataset. Moving to box 2, we assign higher weights to the examples that were misclassified by the previous model (depicted as circles), and lower weights to the correctly classified examples. This weighting scheme ensures that the subsequent training round focuses more on the difficult-to-classify examples. In the next training round, box 3, we observe that the weak learning model misclassifies three different examples from the circle class. As a result, these misclassified examples are assigned higher weights. Considering an AdaBoost classifier with only three boosting rounds, we proceed to combine the three weak learning models trained on different training subsets. The combination is done using a weighted majority vote, as shown in box 4. The weights assigned to each weak model reflect their performance in classifying the training examples. By iteratively updating the weights and training new weak models, AdaBoost aims to

construct a strong ensemble model that can effectively handle difficult classification instances.

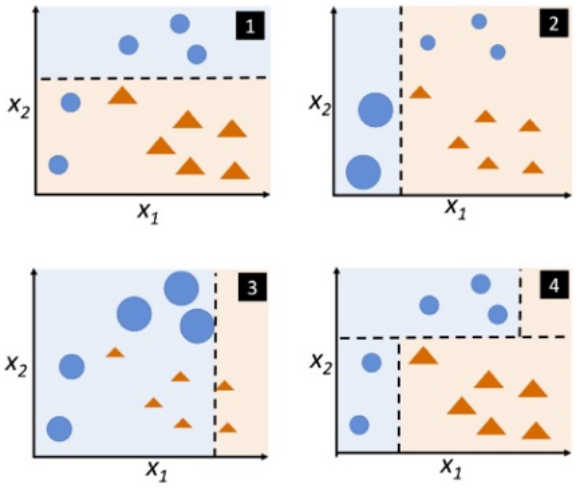


Figure A - 6.1.7. AdaBoost weights update.

XGBoost

XGBoost, short for Extreme Gradient Boosting, is an implementation of the gradient boosted trees algorithm, which combines the concepts of decision trees and gradient boosting. This classifier is commonly employed for solving regression or classification problems using a supervised learning approach. One of the notable advantages of XGBoost is its interpretability, attributed to the structure of decision trees. As depicted in Figure A - 6.1.8, decision trees consist of nodes, branches, and leaves. They enable a step-by-step decision-making process based on a series of questions. Starting from the root of the tree, the algorithm identifies the most informative feature that results in the highest information gain for data partitioning. This process of subdividing the data continues iteratively at each child node until the leaves of the tree contain only samples of a single class, ensuring purity. By iteratively training decision trees and optimizing the objective function through gradient boosting, XGBoost aims to create a powerful ensemble model that can effectively capture complex patterns and make accurate predictions.

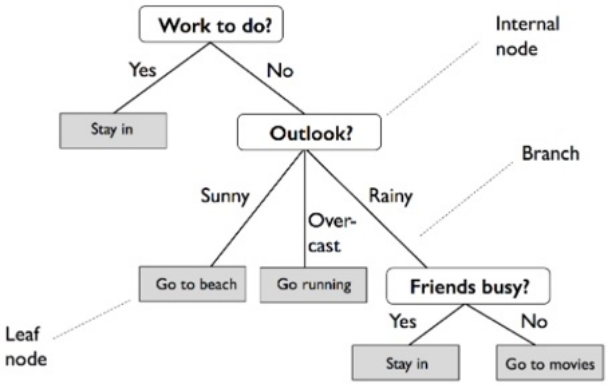


Figure A - 6.1.8. Decision Tree Classifier Workflow.

To facilitate the subdivision of nodes based on the most informative features, the objective function of the XGBoost algorithm is designed to maximize the information gain at each split. The information gain is a measure of how much the uncertainty or impurity of the data is reduced after a particular feature is used for partitioning. By selecting the feature that maximizes the information gain, XGBoost aims to create splits that result in the most significant improvement in the predictive power of the model. Maximizing the information gain allows the algorithm to make effective decisions at each node, ensuring that the subsequent splits result in more homogeneous subsets of data. This process helps in capturing the underlying patterns and relationships in the dataset, leading to improved predictive performance of the XGBoost classifier.

The information gain is defined as:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

where f is the feature used for the subdivision, D_p and D_j are the datasets of the parent node and the j-th child node respectively, I is the measure of the impurity, N_p is the total number of the parent node training examples and N_j is the number of the j-th child node training examples. However, it is frequently used the subdivision of each parent node in two child nodes identified by an information gain defined as:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

In the final analysis, the information gain is the difference between the parent node impurity and the sum of the impurities of each child node. One way to measure the impurity is the **entropy** definition:

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2\, p(i|t)$$

in which $p(i|t)$ is the proportion of the examples belonging to the i class for the t When all the examples in a subset belong to the same class, the entropy is minimized and becomes zero. On the other hand, when the classes are uniformly distributed and there is an equal proportion of examples from each class, the entropy is maximized and reaches its highest value.

XGBoost (Extreme Gradient Boosting) is a versatile machine learning algorithm widely acclaimed for its exceptional predictive accuracy. Its ability to handle complex relationships and capture nonlinear patterns in the data makes it a preferred choice for many applications. XGBoost offers the advantage of providing insights into feature importance, allowing users to identify influential variables. It incorporates regularization techniques, preventing overfitting and enhancing model generalization. Furthermore, XGBoost has built-in capabilities to handle missing data, reducing the need for extensive preprocessing. Despite its strengths, the complexity and computational requirements of XGBoost may pose challenges in terms of interpretability and training time for large datasets.

Gaussian Naïve Bayes

The Gaussian Naïve Bayes Classifier is a variant of the Naïve Bayes classifier that assumes a Gaussian (normal) distribution for continuous data. It is based on the Bayes Theorem, which is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

where:

- $P(A)$ is the probability of A occurring
- $P(B)$ is the probability of B occurring
- $P(A|B)$ is the probability of A given B
- $P(B|A)$ is the probability of B given A
- $P(A \cap B)$ is the probability of both A and B occurring

Bayes' Theorem is built upon the assumption of strong independence among features. In the context of the Gaussian Naïve Bayes Classifier, the classifier

assumes that the continuous values associated with each class are distributed according to a normal distribution. This assumption simplifies the modeling process by allowing the estimation of parameters, such as mean and variance, based on the training data. By assuming the normal distribution, the classifier can calculate the probability of a given sample belonging to a particular class using the probability density function (PDF) of the normal distribution for each feature value. The class with the highest probability density, determined using Bayes' Theorem, is assigned to the sample. In Figure A - 6.1.9, calculation of distances between each point and the class mean is depicted. Specifically, the z-score is computed, which represents the distance of a point from the mean of its respective class divided by the standard deviation of that class. This calculation allows for a standardized measure of how far each point is from the center of its class distribution.

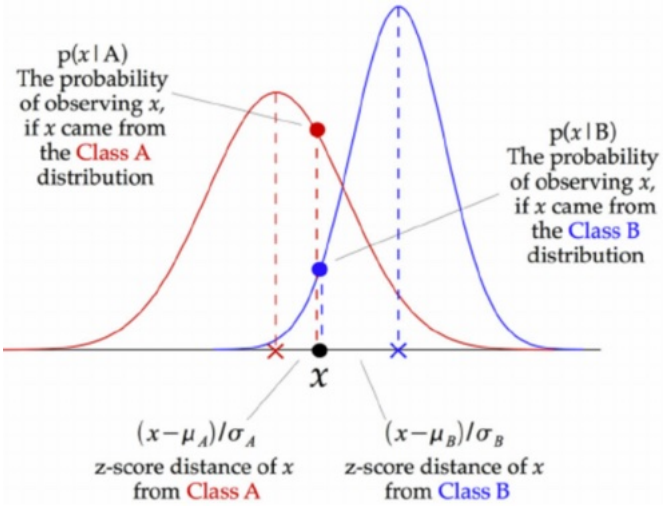


Figure A - 6.1.9. Gaussian Naïve Bayes.

6.2 Molecular basis of taste perception

List of abbreviations

- ASIC: Acid-Sensing Ion Channel
- CALHM1/3: Voltage-Gated Neurotransmitter-Release Channel
- CaSR: Calcium-Sensing Receptor
- chOTOP1: Chicken OTOP3
- CRD: Cystein Rich Domain
- EC: Extracellular
- ECL: Extracellular Loop
- ENaC: Epithelial Sodium Channel
- GMP: Guanosine 5'-Monophosphate
- GPCR: G Protein-Coupled Receptor

- GRIP: Gating Release of Inhibition by Proteolysis
- HM: Homology Modelling
- ICL: Intracellular Loop
- IMP: Inosine 5'-Monophosphate
- LB: Ligand-Binding Domain
- MD: Molecular Dynamics
- mGluR1: Metabotropic Glutamate Receptor of subtype 1
- MSG: Monosodium Glutamate
- NAM: Negative Allosteric Modulator
- NUS: Number of Unique Scaffolds
- OTOP: Otopetrin
- PAM: Positive Allosteric Modulator
- PI: Promiscuity Index
- PKD1L3: Polycystic Kidney Disease 1-Like3 Ion Channel
- PKD2L1: Polycystic Kidney Disease 2-Like1 Ion Channel
- PMD: Polycystic Mucolipin Domain
- SG: Steviol Glycosides
- TAS2R1: Taste receptor type 2 member 1
- TLC: Three Leaf Cover
- TM: Transmembrane
- TMD: Transmembrane Domain
- TPRML3: Transient Receptor Potential Cation Channel Mucolipin Subfamily Member 3
- TRC: Taste Receptor Cell
- TRPPs: Transient Receptor Potential Polycystin
- TRPV1: Transient Receptor Potential Cation Channel, Subfamily V, Member 1
- VFTM: Venus Flytrap Module
- VSLD: Voltage Sensing-Like Domain
- XtOTOP3: *Xenopus Tropicalis* OTOP3
- zfOTOP1: zebrafish OTOP1

Bitter receptors summary

Table A - 6.2.1. Summary table of the 25 human bitter taste receptors, including possible alternative nomenclature. The provided information is taken from BitterDB.

#	NAME	ALTERNATIVE NOMENCLATURE
1	TAS2R1	TRB7
2	TAS2R3	/
3	TAS2R4	/
4	TAS2R5	/
5	TAS2R7	TRB4
6	TAS2R8	TRB5
7	TAS2R9	TRB6
8	TAS2R10	TRB2
9	TAS2R13	TRB3
10	TAS2R14	TRB1
11	TAS2R16	/
12	TAS2R38	PTC, TAS2R61

13	TAS2R39	TAS2R57
14	TAS2R40	TAS2R58, GPR60
15	TAS2R41	TAS2R59
16	TAS2R42	TAS2R55
17	TAS2R43	TAS2R52
18	TAS2R44	TAS2R31, TAS2R53
19	TAS2R45	GPR59
20	TAS2R46	TAS2R54
21	TAS2R47	TAS2R30
22	TAS2R48	TAS2R19, TAS2R23
23	TAS2R49	TAS2R20, TAS2R56
24	TAS2R50	TAS2R51
25	TAS2R60	TAS2R56

6.3 VirtuousPocketome

6.3.1 Methods – Similarity Search methods in literature

Table A - 6.3.1. Summary of the main methods for the similarity search of a protein binding site in previous literature with their relative type of representation of the binding site and strategy of comparison.

Method Name	Site Representation	Strategy of Comparison
SuMo (2003)[474]	Residue-based	3D Points
PINTS (2003)[475]	Residue-based	Other
eF-seek (2004) [476]	Surface-based	Graphs
TM-Align (2005) [477]	Residue-based	Other
SiteEngine (2005) [478]	Surface-based	Graphs
ContactMetricServer [479] (2006)	Residue-based	Other
PocketMatch (2008)[480,481]	Residue-based	Other
MultiBind MAPPIS (2008)[482]	Surface-based	3D Points
PevoSOAR (2009) [483]	Surface-based	Other
fPOP (2009) [484]	Surface-based	Fingerprint
PESD-serv (2010) [485]	Interaction-based	Other
SeSAW (2010) [486]	Residue-based	Other
LabelHash (2010) [487]	Residue-based	Other
FuzCav (2010) [488]	Residue-based	Fingerprint
Pro-BIS ligand (2012) [489]	Surface-based	Graphs
PoSSuM (2012) [490]	Residue-based	Other
COFACTOR (2012) [491]	Residue-based	3D Points
SPRITE-ASSAM (2012) [145]	Residue-based	Graphs
SiteComp lin (2012) [492]	Interaction-based	Other
Iso-Cleft Finder (2013) [493]	Residue-based	Graphs
CatSId (2013) [494]	Residue-based	Graphs
IMAAAGINE (2013) [495]	Residue-based	Graphs
Apoc (2013) [496]	Residue-based	Other
ASSIST (2014) [497]	Residue-based	3D Points

IsoMIF Finder (2015) [498,499]	Interaction-based	Graphs
G-LoSA (2016) [500]	Residue-based	Graphs
Geomfinder (2016) [501]	Residue-based	Other
PatchSearch (2019) [502]	Residue-based	Graphs
Drugreposer ER (2019) [503]	Residue-based	Graphs
DeeplyTough (2020) [504]	Interaction-based	Other

6.3.2 Results - Conformational Dynamics

Structural equilibrium of the MD simulations of the bitter taste receptor bound with strychnine.

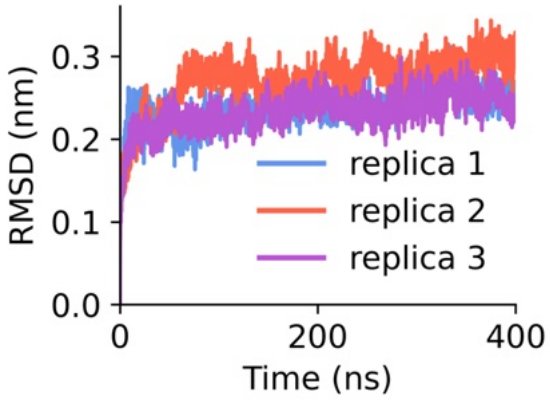


Figure A - 6.3.1. RMSD of the bitter taste receptor for the three simulation replicas performed (colored in blue, red and violet, respectively).

6.3.3 Results - Similarity Search and Multi-step Filtering

The retrieved protein hits according to the docking score (DScore) after the multi-step filtering process are reported in Table A - 6.3.2.

Table A - 6.3.2. Protein hits according to the docking score (DScore) after the multi-step filtering process.

PDB	Description	RMSD	SASA	DScore
3a4s	SUMO-CONJUGATING ENZYME UBC9	1.07	2.215	-87.7711
3b7r	LEUKOTRIENE A-4 HYDROLASE	1.25	2.155	-87.065
7bqz	SPINDLIN-1	0.75	2.278	-86.6839
2yg2	APOLIPOPROTEIN M	1.26	2.451	-86.1774
3fhe	LEUKOTRIENE A-4 HYDROLASE	1.27	1.589	-85.2467
2zfh	CUTA	1.36	1.936	-85.0248
4dpr	LEUKOTRIENE A-4 HYDROLASE	1.31	2.073	-85.0148
2r3r	CELL DIVISION PROTEIN KINASE 2	1.13	2.306	-84.9421
6o5h	LEUKOTRIENE A-4 HYDROLASE	1.3	2.348	-84.9287
3fuk	LEUKOTRIENE A-4 HYDROLASE	1.26	1.86	-84.7628
3fu5	LEUKOTRIENE A-4 HYDROLASE	1.29	2.97	-84.591
7av2	LEUKOTRIENE A-4 HYDROLASE	1.26	1.684	-84.3107
3fh7	LEUKOTRIENE A-4 HYDROLASE	1.22	2.443	-84.2115
6r2u	ZINC-ALPHA-2-GLYCOPROTEIN	1.11	1.794	-84.1411
6p5s	HOMEODOMAIN-INTERACTING PROTEIN KINASE 2	1.16	2.097	-84.1008
4r7l	LEUKOTRIENE A-4 HYDROLASE	1.25	1.298	-84.0848
4ms6	LEUKOTRIENE A-4 HYDROLASE	1.34	1.489	-83.9426
2h2u	SOLUBLE CALCIUM-ACTIVATED NUCLEOTIDASE 1	1.02	2.114	-83.8648
3chp	LEUKOTRIENE A-4 HYDROLASE	1.18	2.395	-83.3851
7av1	LEUKOTRIENE A-4 HYDROLASE	1.29	1.685	-83.3388
7kze	LEUKOTRIENE A-4 HYDROLASE	1.33	1.91	-83.2739
5bpp	LEUKOTRIENE A-4 HYDROLASE	1.26	1.274	-82.9787
3cho	LEUKOTRIENE A-4 HYDROLASE	1.24	1.414	-82.8956

5ni4	LEUKOTRIENE A-4 HYDROLASE	1.26	2.004	-82.8657
1hs6	LEUKOTRIENE A-4 HYDROLASE	1.26	2.677	-82.8356
3ftv	LEUKOTRIENE A-4 HYDROLASE	1.24	1.533	-82.6407
5ni6	LEUKOTRIENE A-4 HYDROLASE	1.23	2.295	-82.4868
2zi2	THROMBIN HEAVY CHAIN	1.16	2.086	-82.4733
7auz	LEUKOTRIENE A-4 HYDROLASE	1.28	1.45	-82.4511
6end	LEUKOTRIENE A-4 HYDROLASE	1.25	2.812	-82.2884
3ftu	LEUKOTRIENE A-4 HYDROLASE	1.25	2.925	-82.141
3ful	LEUKOTRIENE A-4 HYDROLASE	1.31	1.81	-82.1373
7qg0	NAD(+) HYDROLASE SARM1	1.39	3.449	-81.9657
4rsy	LEUKOTRIENE A-4 HYDROLASE	1.26	1.774	-81.922
7n3n	SOLUTE CARRIER FAMILY 12 MEMBER 2	1.26	2.091	-81.8164
2r24	ALDOSE REDUCTASE	1.37	1.567	-81.7647
1gw6	LEUKOTRIENE A-4 HYDROLASE	1.24	2.231	-81.7538
6enb	LEUKOTRIENE A-4 HYDROLASE	1.28	1.961	-81.7255
5kir	PROSTAGLANDIN G/H SYNTHASE 2	1.43	2.149	-81.1718
4l2l	LEUKOTRIENE A-4 HYDROLASE	1.3	1.311	-81.1157
3eq6	ACYL-COENZYME A SYNTHETASE ACSM2A	1.27	2.577	-80.8085
3fu3	LEUKOTRIENE A-4 HYDROLASE	1.29	3.235	-80.7407
2vqo	HISTONE DEACETYLASE 4	1.06	2.148	-80.5489
6p8z	GTPASE KRAS	1.34	2.009	-80.5234
3fty	LEUKOTRIENE A-4 HYDROLASE	1.25	2.088	-80.4234
3chs	LEUKOTRIENE A-4 HYDROLASE	1.21	1.943	-80.4112
6w9v	TCR-BETA CHAIN, MAJOR HISTOCOMPATIBILITY COMPLEX CLASS I-RELATED GENE	1.37	2.865	-80.3695
2r3o	CELL DIVISION PROTEIN KINASE 2	1.28	1.456	-79.8837
6h0g	PROTEIN CEREBLON, DNA DAMAGE-BINDING PROTEIN 1,DNA DAMAGE-BINDING PROTEIN 1,	1.45	2.141	-79.6843
7ure	ISOFORM 2 OF PROTEIN-SERINE O-PALMITOLEOYLTRANSFERASE	1.22	1.97	-79.589
4kfz	ANTI-LMO2 VH	1.49	2.086	-79.5841
3mph	AMILORIDE-SENSITIVE AMINE OXIDASE	1.48	1.988	-79.4227
6u3p	ACETYLCHOLINESTERASE	1.5	1.988	-79.3888
5fe7	HISTONE ACETYLTRANSFERASE KAT2B	1.42	1.618	-78.6233
6nr8	T-COMPLEX PROTEIN 1 SUBUNIT GAMMA, T-COMPLEX PROTEIN 1 SUBUNIT THETA	1.47	1.547	-78.5151
6m5o	SERINE HYDROXYMETHYLTRANSFERASE, MITOCHONDRIAL	1.37	1.652	-78.2072
5mw3	HISTONE-LYSINE N-METHYLTRANSFERASE, H3 LYSINE-79 SPECIFIC	1.36	1.407	-78.0836
3d49	THROMBIN HEAVY CHAIN	1.16	1.609	-78.0821
1msv	S-ADENOSYLMETHIONINE DECARBOXYLASE PROENZYME	1.44	1.902	-77.9728
5fe1	HISTONE ACETYLTRANSFERASE KAT2B	1.4	1.399	-77.9383
3qrt	CYCLIN-DEPENDENT KINASE 2	1.45	2.434	-77.9
3qvv	SULFOTRANSFERASE 1A1	1.45	1.562	-77.8891
7zjp	TRANSCRIPTIONAL ENHANCER FACTOR TEF-1	1.46	2.805	-77.8122
4umo	CALMODULIN, POTASSIUM VOLTAGE-GATED CHANNEL SUBFAMILY KQT MEMBER 1	1.23	1.668	-77.6608
7e9n	HEAVY CHAIN OF 35B5 FAB	1.15	2.144	-77.4416
6f5t	LYSINE-SPECIFIC DEMETHYLASE 4D	1.45	1.785	-77.2974
5m7t	PROTEIN O-GLCNACASE	1.37	2.185	-77.279
1sqm	LEUKOTRIENE A-4 HYDROLASE	1.29	1.249	-77.1773
8hkw	PEPTIDE FROM TP53-BINDING PROTEIN 1, IMPORTIN SUBUNIT ALPHA-3	1.47	1.519	-76.9298
3lnz	E3 UBIQUITIN-PROTEIN LIGASE MDM2	1.45	2.599	-76.7359
4ek3	CYCLIN-DEPENDENT KINASE 2	1.22	1.817	-76.706
3ibd	CYTOCHROME P450 2B6	1.43	1.746	-76.7021
2l12	CHROMOBOX HOMOLOG 7	1.25	3.816	-76.2799
3rvh	LYSINE-SPECIFIC DEMETHYLASE 4A	1.29	1.531	-76.2637
6qnx	COHESIN SUBUNIT SA-2, TRANSCRIPTIONAL REPRESSOR CTCF	1.18	2.616	-76.2505
6hg4	INTERLEUKIN-17 RECEPTOR C	1.48	1.43	-75.8588
4nst	CYCLIN-DEPENDENT KINASE 12, CYCLIN-K	1.46	2.276	-75.8423
4x0u	ALPHA-AMINOADIPIC SEMIALDEHYDE DEHYDROGENASE	1.42	2.124	-75.8376

7bmk	SERINE/THREONINE-PROTEIN KINASE/ENDORIBONUCLEASE IRE1	1.38	3.453	-75.8372
4bbm	CYCLIN-DEPENDENT KINASE-LIKE 2	1.31	1.714	-75.8371
2e8d	BETA-2-MICROGLOBULIN	1.49	2.138	-75.811
4xx1	FAB1 HEAVY CHAIN	1.32	1.994	-75.8052
6yaf	AP-2 COMPLEX SUBUNIT BETA	1.32	1.496	-75.7476
6qpl	SPINDLIN-1	1.45	2.241	-75.724
4mzg	SPINDLIN-1	1.41	2.286	-75.6688
1gz8	CELL DIVISION PROTEIN KINASE 2	1.19	3.287	-75.6682
6j0l	BUTYROPHILIN SUBFAMILY 3 MEMBER A3	1.19	1.654	-75.6497
6yah	AP-2 COMPLEX SUBUNIT BETA	1.32	1.466	-75.5767
6mid	MONOCLONAL ANTIBODY ZIKV-195 HEAVY CHAIN	1.48	1.145	-75.4824
6tel	HISTONE-LYSINE N-METHYLTRANSFERASE, H3 LYSINE-79 SPECIFIC	1.33	2.339	-75.2809
7rhl	CGMP-GATED CATION CHANNEL ALPHA-1	1.15	2.253	-75.1955
7zyf	LEUCYL-CYSTINYL AMINOPEPTIDASE, PREGNANCY SERUM FORM	0.94	2.367	-75.148
3u84	MENIN	1.22	2.05	-75
7qne	GABA(A) RECEPTOR SUBUNIT GAMMA-2, GAMMA-AMINOBUTYRIC ACID RECEPTOR SUBUNIT BETA-3	1.41	1.498	-74.9855
6xk9	PROTEIN CEREBLON, DNA DAMAGE-BINDING PROTEIN 1	1.38	1.429	-74.8752
6vum	STEROL O-ACYLTRANSFERASE 1	1.48	2.684	-74.8556
7p9t	5'-NUCLEOTIDASE	1.45	2.011	-74.8448
8csq	28S RIBOSOMAL PROTEIN S29, MITOCHONDRIAL	1.29	1.622	-74.8231
1cly	IGG FAB (HUMAN IGG1, KAPPA)	1.23	1.784	-74.6885
4u9v	N-ALPHA-ACETYLTRANSFERASE 40	1.31	2.191	-74.6182
4ljp	E3 UBIQUITIN-PROTEIN LIGASE RNF31	1.46	2.269	-74.4844
2r3h	CELL DIVISION PROTEIN KINASE 2	1	2.291	-74.3763
7byi	SERINE HYDROXYMETHYLTRANSFERASE, MITOCHONDRIAL	1.4	1.907	-74.2523
7qoo	CENTROMERE PROTEIN H, CENTROMERE PROTEIN I	0.72	2.216	-74.234
7ocb	SPINDLIN-1	1.44	3.322	-74.153
6v8c	ORNITHINE AMINOTRANSFERASE, MITOCHONDRIAL	1.34	1.647	-73.9414
6in3	HISTONE-LYSINE N-METHYLTRANSFERASE, H3 LYSINE-79 SPECIFIC	1.33	3.301	-73.8186
6wqz	AUTOPHAGY-RELATED PROTEIN 9A	1.22	2.699	-73.707
7r5s	CENTROMERE PROTEIN H, CENTROMERE PROTEIN I	0.96	3.945	-73.6423
7ttn	TUBULIN BETA CHAIN, T-COMPLEX PROTEIN 1 SUBUNIT GAMMA	1.34	1.756	-73.5768
1grh	GLUTATHIONE REDUCTASE	1.46	2.023	-73.4928
1grg	GLUTATHIONE REDUCTASE	1.46	1.613	-73.4914
3vv0	HISTONE-LYSINE N-METHYLTRANSFERASE SETD7	1.19	1.978	-73.342
5wbs	FRIZZLED-7,INHIBITOR PEPTIDE FZ7-21	1.09	2.043	-73.3394
3grs	GLUTATHIONE REDUCTASE	1.46	1.35	-73.3385
7m63	INDOLEAMINE 2,3-DIOXYGENASE 1	1.37	1.501	-73.2505
4pvf	SERINE HYDROXYMETHYLTRANSFERASE, MITOCHONDRIAL	1.37	2.059	-73.2492
8dwt	SPECKLE-TYPE POZ PROTEIN	1.33	2.068	-73.2317
7omn	JD1-1 VH DOMAIN	1.35	1.915	-73.2243
7q29	ANGIOTENSIN-CONVERTING ENZYME	1.5	2.346	-73.2221
6bly	CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR SUBUNIT 1	1.38	1.682	-73.1401
6ba4	HISTONE ACETYLTRANSFERASE KAT8	1.4	2.132	-73.089
5fpb	LYSINE-SPECIFIC DEMETHYLASE 4D	1.32	1.492	-72.8148
1grf	GLUTATHIONE REDUCTASE	1.45	2.048	-72.7912
2c8y	THROMBIN HEAVY CHAIN	1.17	1.538	-72.7212
6mbl	HISTONE-LYSINE N-METHYLTRANSFERASE SETD3	1.04	2.258	-72.679
2vtm	CELL DIVISION PROTEIN KINASE 2	1.44	1.645	-72.5833
7o7l	ALPHA-2-MACROGLOBULIN	1.33	1.796	-72.5426
1eak	72 KDA TYPE IV COLLAGENASE	1.43	1.489	-72.4855
5ja7	CATHEPSIN K	1.09	1.968	-72.4568
5z9w	EBOLAVIRUS NUCLEOPROTEIN (RESIDUES 19-406)	0.83	1.86	-72.4313
7lk0	ORNITHINE AMINOTRANSFERASE, MITOCHONDRIAL	1.35	2.123	-72.3516
7fcp	P5-22 ANTIBODY FAB FRAGMENT HEAVY CHAIN	1.5	1.196	-72.3062

6w3c	SERINE/THREONINE-PROTEIN KINASE/ENDORIBONUCLEASE IRE1	1.34	1.72	-72.2763
5bnj	CYCLIN-DEPENDENT KINASE 8	1.32	1.454	-72.2569
6xdb	SERINE/THREONINE-PROTEIN KINASE/ENDORIBONUCLEASE IRE1	1.42	3.229	-72.2502
7uvr	ATP-DEPENDENT CLP PROTEASE PROTEOLYTIC SUBUNIT,	1.24	1.621	-72.2437
6i8b	SPINDLIN-1	1.44	2.006	-72.077
7qdr	WD REPEAT-CONTAINING PROTEIN 61	1.47	2.207	-71.9206
7nvl	T-COMPLEX PROTEIN 1 SUBUNIT ETA	1.25	1.451	-71.8865
5jq8	CYCLIN-DEPENDENT KINASE 2	1.24	1.522	-71.813
7a3g	DIPEPTIDYL PEPTIDASE 8	1.47	2.615	-71.8083
5fe6	HISTONE ACETYLTRANSFERASE KAT2B	1.35	2.409	-71.7955
5fbh	EXTRACELLULAR CALCIUM-SENSING RECEPTOR	1.34	1.521	-71.7642
5fe2	HISTONE ACETYLTRANSFERASE KAT2B	1.34	1.605	-71.7247
5mv7	UNCONVENTIONAL MYOSIN-VIIB	1.29	2.397	-71.6435
5fy8	LYSINE-SPECIFIC DEMETHYLASE 4A	1.11	2.043	-71.6353
6h4q	LYSINE-SPECIFIC DEMETHYLASE 4A	1.25	3.312	-71.611
4udw	THROMBIN HEAVY CHAIN	1.23	2.865	-71.5643
5y3r	DNA-DEPENDENT PROTEIN KINASE CATALYTIC SUBUNIT	1.24	2.279	-71.5376
2ypt	CAAX PRENYL PROTEASE 1 HOMOLOG	1.31	2.139	-71.5061
5ka8	TYROSINE-PROTEIN PHOSPHATASE NON-RECEPTOR TYPE 1	0.99	1.969	-71.4398
2rhy	LETHAL(3)MALIGNANT BRAIN TUMOR-LIKE PROTEIN	1.32	1.6	-71.4055
8hik	ANTI-BRIL FAB LIGHT CHAIN, ANTI-BRIL FAB HEAVY CHAIN	1.25	2.178	-71.4044
5uwf	CAMP AND CAMP-INHIBITED CGMP 3',5'-CYCLIC PHOSPHODIESTERASE	1.38	2.096	-71.303
8e3i	CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR SUBUNIT 1	1.38	1.756	-71.2577
7sck	EXOSTOSIN-2	1.11	2.409	-71.1007
2wjy	REGULATOR OF NONSENSE TRANSCRIPTS 1	1.37	1.534	-71.0623
4p4h	MITOCHONDRIAL ANTIVIRAL-SIGNALING PROTEIN, PROBABLE ATP-DEPENDENT RNA HELICASE DDX58	1.24	1.584	-71.0279
7ni5	SERINE-PROTEIN KINASE ATM	1.25	2.787	-71.0278
5hnb	CYCLIN-DEPENDENT KINASE 8	1.29	1.458	-71.0242
5xez	ANTIBODY, MAB1, HEAVY CHAIN	1.43	1.533	-71.014
7nd4	COVOX-88 FAB HEAVY CHAIN	1.47	2.694	-70.9952
4ifb	BILE SALT SULFOTRANSFERASE	1.42	1.855	-70.8993
7d0p	HISTONE ACETYLTRANSFERASE KAT7	1.48	1.494	-70.899
7nvn	T-COMPLEX PROTEIN 1 SUBUNIT ETA	1.22	1.425	-70.8572
6r7n	COP9 SIGNALOSOME COMPLEX SUBUNIT 2, CULLIN-2	1.21	2.267	-70.8549
4tw0	SCAVENGER RECEPTOR CLASS B MEMBER 2	1.42	2.458	-70.8325
5q8h	DCLRE1A	1.46	1.94	-70.8299
3epa	S-ADENOSYLMETHIONINE DECARBOXYLASE BETA CHAIN, S-ADENOSYLMETHIONINE DECARBOXYLASE ALPHA CHAIN	1.44	2.078	-70.8049
1y1j	C-ALPHA-FORMYGLYCINE-GENERATING ENZYME	1.28	1.745	-70.7845
6u8w	DNA (CYTOSINE-5)-METHYLTRANSFERASE 3B, DNA (CYTOSINE-5)-METHYLTRANSFERASE 3-LIKE	1.46	1.576	-70.7426
7o7o	ALPHA-2-MACROGLOBULIN	1.23	1.581	-70.7403
6rnq	GEM-ASSOCIATED PROTEIN 5	0.84	1.908	-70.736
6h4u	LYSINE-SPECIFIC DEMETHYLASE 4A	1.29	1.781	-70.6779
4ayv	THROMBIN HEAVY CHAIN	1.29	1.922	-70.6612
7ted	ORNITHINE AMINOTRANSFERASE, MITOCHONDRIAL	1.32	3.313	-70.6456
3zmz	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1A, REST COREPRESSOR 1	1.42	1.997	-70.5755
4aw6	CAAX PRENYL PROTEASE 1 HOMOLOG	1.32	2.358	-70.5584
8b6l	TRANSMEMBRANE PROTEIN 258, DOLICHYL-DIPHOSPHOOLIGOSACCHARIDE--PROTEIN	1.32	1.701	-70.5459
5lsp	107_A07 FAB HEAVY CHAIN, 107_A07 FAB LIGHT CHAIN	1.26	1.753	-70.5312
5q8q	DCLRE1A	1.49	3.285	-70.493
2xiq	METHYLMALONYL-COA MUTASE, MITOCHONDRIAL	1.45	1.836	-70.4676
4msn	CAMP AND CAMP-INHIBITED CGMP 3',5'-CYCLIC PHOSPHODIESTERASE	1.44	2.813	-70.4675
6s8l	TUBULIN BETA-3 CHAIN	1.12	1.729	-70.4406

4ek9	HISTONE-LYSINE N-METHYLTRANSFERASE, H3 LYSINE-79 SPECIFIC	1.25	1.738	-70.4255
7sid	SERINE-PROTEIN KINASE ATM	1.37	2.352	-70.4108
7k8s	C002 FAB HEAVY CHAIN, C002 FAB LIGHT CHAIN	1.29	1.609	-70.3647
5luq	DNA-DEPENDENT PROTEIN KINASE CATALYTIC SUBUNIT,DNA-	1.05	1.73	-70.3546
7xur	SNRNA-ACTIVATING PROTEIN COMPLEX SUBUNIT 3	1.4	1.978	-70.3087
7ni6	SERINE-PROTEIN KINASE ATM	1.3	2.067	-70.3043
7m30	1-32 FAB HEAVY CHAIN	1.44	1.662	-70.2799
5ffg	INTEGRIN BETA-6, INTEGRIN ALPHA-V	1.31	2.065	-70.2695
6dv5	HEAT SHOCK PROTEIN BETA-1	1.31	1.541	-70.2648
7ye9	LIGHT CHAIN OF R1-32 FAB	1.18	2.219	-70.2502
6d01	1210 ANTIBODY, LIGHT CHAIN, 1210 ANTIBODY, HEAVY CHAIN	1.24	1.737	-70.2299
7q12	GLYCOGEN [STARCH] SYNTHASE, MUSCLE	1.44	3.126	-70.1953
6ugq	CARBONIC ANHYDRASE IX-MIMIC	1.24	1.4	-70.1773
8aql	SERINE HYDROXYMETHYLTRANSFERASE, MITOCHONDRIAL	1.32	1.257	-70.1732
7m7d	INDOLEAMINE 2,3-DIOXYGENASE 1	1.34	1.995	-70.1541
3apw	ALPHA-1-ACID GLYCOPROTEIN 2	1.43	2.657	-70.1114
4ekz	PROTEIN DISULFIDE-ISOMERASE	1.18	2.471	-70.0843
1t2a	GDP-MANNOSE 4,6 DEHYDRATASE	1.47	1.527	-70.0784
3b9f	PROTHROMBIN	1.1	2.144	-70.0644
5q4w	DCLRE1A	1.47	1.934	-70.0288
2hi8	SULFATASE-MODIFYING FACTOR 1	1.29	1.953	-70.0168
6vkg	HUMAN CARBONIC ANHYDRASE IX MIMIC	1.34	1.914	-70.0096
3vxp	HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, A-24 ALPHA CHAIN	1.49	1.638	-69.9877
6v63	ACTIN-HISTIDINE N-METHYLTRANSFERASE	0.96	3.151	-69.9484
7nvm	T-COMPLEX PROTEIN 1 SUBUNIT ETA	1.25	1.497	-69.9173
2xij	METHYLMALONYL-COA MUTASE, MITOCHONDRIAL	1.46	2.278	-69.8883
7zsc	PROLYL 4-HYDROXYLASE SUBUNIT ALPHA-2, PROTEIN DISULFIDE-ISOMERASE	1.21	1.972	-69.7845
8e3q	CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR SUBUNIT 1	1.45	1.685	-69.7186
1h4r	MERLIN	1.17	1.743	-69.6992
7czx	IG C168_LIGHT_IGKV4-1_IGKJ4,UNCHARACTERIZED PROTEIN	1.49	1.843	-69.6461
6bnb	PROTEIN CEREBLON, DNA DAMAGE-BINDING PROTEIN 1	1.45	1.865	-69.6368
6oht	3-BETA-HYDROXYSTEROID-DELTA(8),DELTA(7)-ISOMERASE	1	1.92	-69.5843
2ckj	XANTHINE OXIDOREDUCTASE	1	2.287	-69.5344
6mie	POTASSIUM VOLTAGE-GATED CHANNEL SUBFAMILY KQT MEMBER 1	1.34	2.218	-69.5254
7fem	ANGIOTENSIN-CONVERTING ENZYME 2	1.27	1.828	-69.5141
6cqd	MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE KINASE 1	1.08	2.839	-69.4974
2aii	SULFATASE MODIFYING FACTOR 1	1.29	2.389	-69.4813
2qrv	DNA (CYTOSINE-5)-METHYLTRANSFERASE 3A, DNA (CYTOSINE-5)-METHYLTRANSFERASE 3-LIKE	1.43	2.655	-69.3916
7v88	ANGIOTENSIN-CONVERTING ENZYME 2,ANGIOTENSIN-CONVERTING	1.09	2.18	-69.3707
4liq	MACROPHAGE COLONY-STIMULATING FACTOR 1 RECEPTOR	1.46	1.47	-69.3014
6cxv	INDOLEAMINE 2,3-DIOXYGENASE 1	1.44	2.39	-69.2881
7rew	ANTI-CYNO INTERLEUKIN 13 FAB HEAVY CHAIN	1.44	1.912	-69.2592
7q3n	UROMODULIN	1.32	3.582	-69.2552
5vk0	E3 UBIQUITIN-PROTEIN LIGASE MDM2	1.29	2.35	-69.2364
2rhu	LETHAL(3)MALIGNANT BRAIN TUMOR-LIKE PROTEIN	1.31	3.016	-69.2313
4x7t	OMALIZUMAB-FAB HEAVY CHAIN	1.48	1.304	-69.226
6bb2	L-LACTATE DEHYDROGENASE A CHAIN	1.04	2.177	-69.2087
5j13	ANTI-TSLP FAB-FRAGMENT, HEAVY CHAIN	0.93	1.614	-69.1306
5a7p	LYSINE-SPECIFIC DEMETHYLASE 4A	1.11	2.043	-69.0758
5l3c	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1A, REST COREPRESSOR 1	1.38	1.651	-69.0717
8d7w	PROTEIN CEREBLON, DNA DAMAGE-BINDING PROTEIN 1	1.47	1.444	-69.0315
1igr	INSULIN-LIKE GROWTH FACTOR RECEPTOR 1	1.5	1.697	-68.9379
3tiy	CYCLIN-DEPENDENT KINASE 2	1.15	1.536	-68.9332

6v01	POTASSIUM VOLTAGE-GATED CHANNEL SUBFAMILY KQT MEMBER 1	1.46	2.354	-68.9038
3fz1	CELL DIVISION PROTEIN KINASE 2	1.12	2.286	-68.892
6ohs	PHOSPHOLIPASE D2	1.44	2.056	-68.8836
3m17	IGG RECEPTOR FCRN LARGE SUBUNIT P51	1.44	2.255	-68.8318
5fwe	LYSINE-SPECIFIC DEMETHYLASE 4A	1.34	1.823	-68.8306
4u7p	DNA (CYTOSINE-5)-METHYLTRANSFERASE 3A, DNA (CYTOSINE-5)-METHYLTRANSFERASE 3-LIKE	1.45	2.049	-68.8292
7wi0	XMA01 HEAVY CHAIN VARIABLE DOMAIN	1.44	2.458	-68.8238
6b8z	TYROSINE-PROTEIN PHOSPHATASE NON-RECEPTOR TYPE 1	0.89	2.064	-68.8222
2f27	SIALIDASE 2	1.46	2.002	-68.7455
3lfs	CELL DIVISION PROTEIN KINASE 2	1.15	2.086	-68.699
4btj	TAU-TUBULIN KINASE 1	1.15	1.437	-68.6537
7ul3	HISTAMINE H2 RECEPTOR	1.43	1.995	-68.6327
4j9e	TYROSINE-PROTEIN KINASE ABL1, P17	1.44	2.116	-68.6259
5lvr	HISTONE ACETYLTRANSFERASE KAT2B	1.42	1.542	-68.625
3k3o	PHD FINGER PROTEIN 8	1.08	2.144	-68.6182
7frf	TYROSINE-PROTEIN PHOSPHATASE NON-RECEPTOR TYPE 1	1.19	2.226	-68.6141
4az2	THROMBIN HEAVY CHAIN	1.18	2.583	-68.5636
3smt	HISTONE-LYSINE N-METHYLTRANSFERASE SETD3	0.92	2.244	-68.5505

6.3.4 Results - Functional Enrich and Signal Pathway Analyses

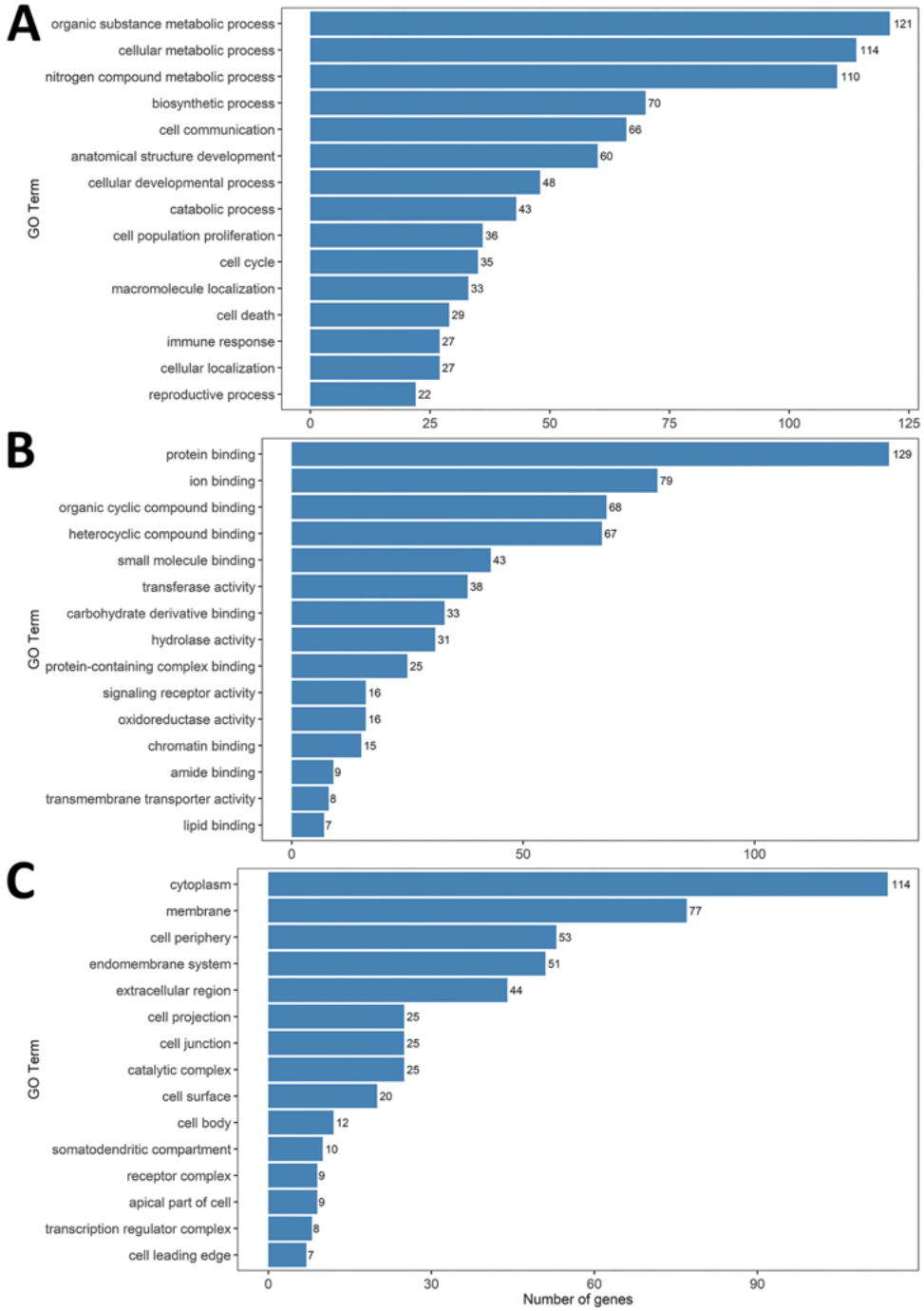


Figure A - 6.3.2.Bar plots representing the retrieved GO terms at the third level of the GO hierarchy relative to (A) Biological Processes (BP), (B) Molecular Functions (MF) and (C) Cellular Components (CC).

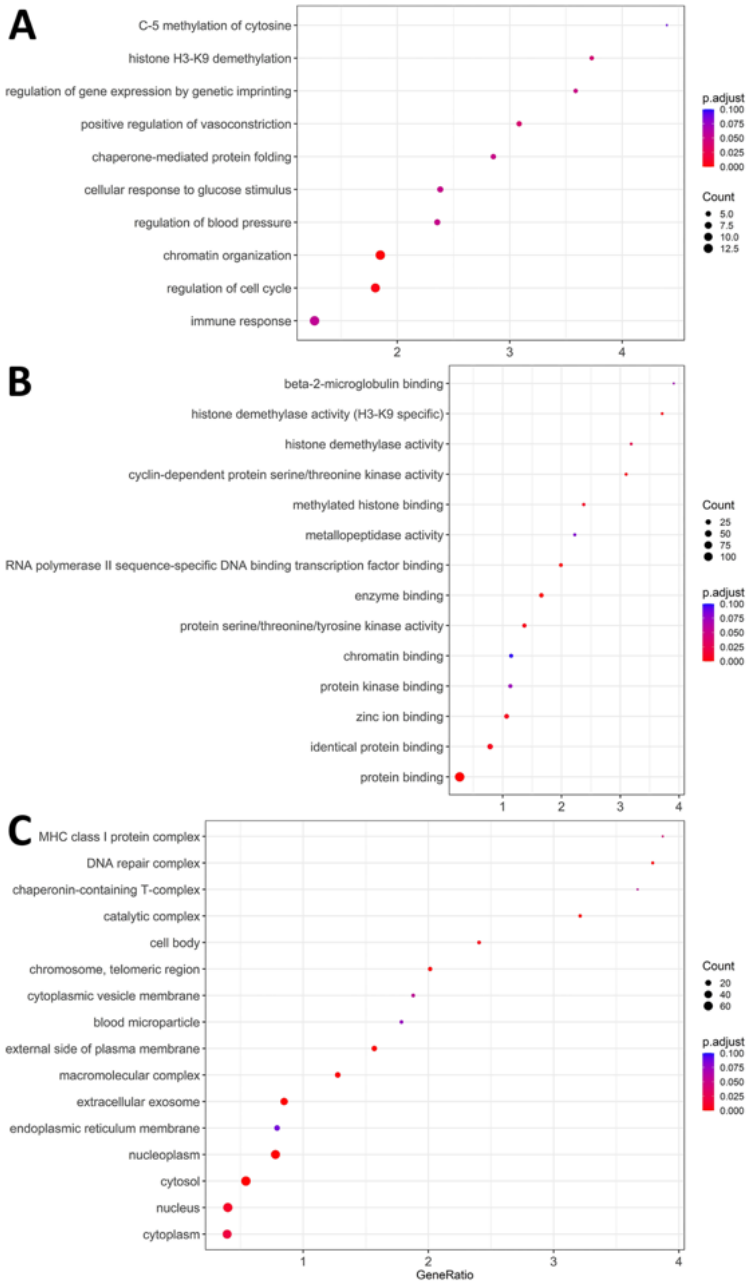


Figure A - 6.3.3. Dot plots representing the retrieved GO terms at the third level of the GO hierarchy relative to (A) Biological Processes (BP), (B) Molecular Functions (MF) and (C) Cellular Components (CC). The x-axis represents the GeneRatio, i.e. the proportion of genes in each GO term that are present in the retrieved gene list compared to the total number of genes in that GO term. The y-axis represents the statistically significant GO terms with an adjusted p-value < 0.1. The color of the dots represents the adjusted p-value (BH), red represents the smaller values, indicating higher statistical significance of the term, while blue represents larger values, indicating lower statistical significance. The size of the dots represents the number of enriched genes in the gene list associated with each GO term.

6.4 The Impact of Natural Compounds on S-Shaped Aβ42 Fibril

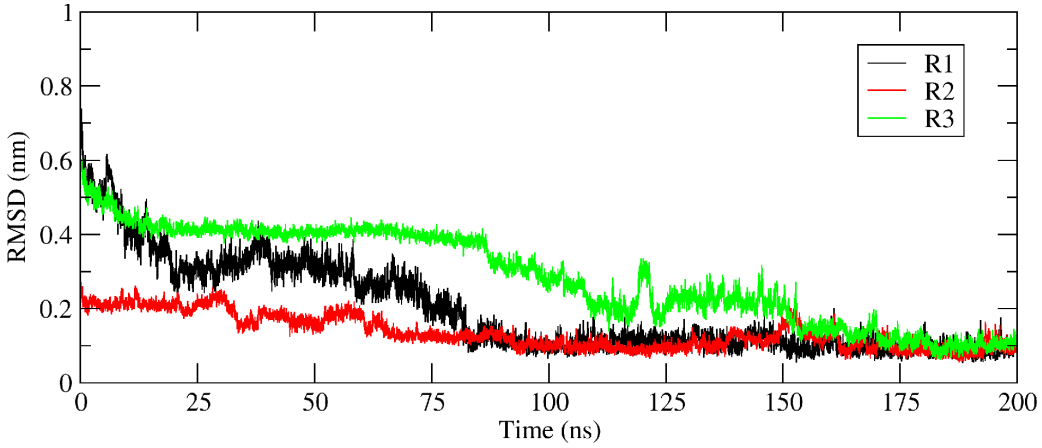


Figure A - 6.4.1. RMSD of the five chains of the 2MXU about the average structure during the last 25 ns. The first replica is represented in black, the second one in red and the third one in green. All systems reach the structural equilibrium.

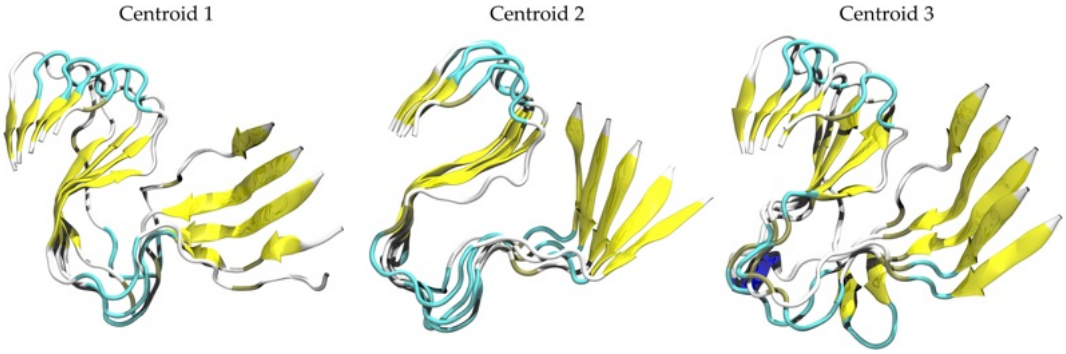


Figure A - 6.4.2. Centroids of the most populated cluster for each independent replica. Each configuration is obtained by a cluster analysis on the last 50 ns of MD simulations, using linkage method and a RMSD cut-off of 0.1nm.

Table A - 6.4.1. List of compounds with their relative binding energy and charge.

Compound	Binding Energy (kcal/mol)	Charge
Beta Carotene	-55.81	0
Oleuropein	-42.63	-1
Rosmarinic Acid	-42.61	-1
Gossypin	-40.97	-1
Piceatannol	-39.75	1
Withanolide A	-39.02	0
Salvianolic Acid A	-37.78	-1

Piperine	-35.75	0
Curcumin	-35.67	0
6-Shogaol	-34.57	-1
EGCG	-34.31	0
Myricetin	-34.28	-1
Viniferin	-34.27	-2
Epicatechin	-33.58	-1
Fisetin	-33.57	0
Diosgenin	-33.50	-1
Rutin	-33.42	-1
Asiatic Acid	-32.97	-1
Puerarin	-32.76	0
Berberine	-32.44	0
Resveratrol	-31.87	0
α-Linolenic Acid	-30.94	0
Retinal	-28.90	1
ScylloInositol	-28.54	0
Vitamine D	-28.08	-1
Rhodosin	-27.85	-1
Retinol	-27.19	0
DHA	-27.18	0
Retinoic Acid	-26.90	-2
Naringin	-26.89	-1
LTheanine	-26.61	0
Caffeic Acid	-26.41	0
Honokiol	-26.17	0
Ellagic	-26.11	0
Hydroxytyrosol	-24.82	0
Vitamine E	-24.16	0
Apigenin	-23.36	0
Quercitin	-22.33	0
NDGA	-22.01	0
Osthole	-21.97	-1
Tetracycline	-20.96	0
Baicalein	-20.91	0
Melatonin	-20.71	-1
Kaempferol	-19.98	0
Oleocanthal	-17.74	0
Naringenin	-16.60	-1
Ferulic Acid	-16.19	0
Homotaurine	-15.04	-2
Caffeine	-12.89	0
Morin	-12.50	-1
Vanillic Acid	-11.01	-1
Lipoic Acid	-9.98	0
Huperizne A	-9.53	-1
EPA	-9.19	0
Vitamine C	-7.52	0
Glycine Betaine	-0.29	0
Gallic Acid	-0.19	-1

Table A - 6.4.2. Summary of the simulated systems.

System	N. Replicas	Simulation Time [ns]
Configuration 1 – No Ligand	3	150
Configuration 2 – No Ligand	3	150
Configuration 3 – No Ligand	3	150
Beta Carotene	3	150
Oleuropein	3	150
Rosmarinic Acid	3	150
Gossypin	3	150
Piceatannol	3	150
Withanolide A	3	150
Salvianolic Acid A	3	150
Piperine	3	150
Curcumin	3	150
6-Shogaol	3	150

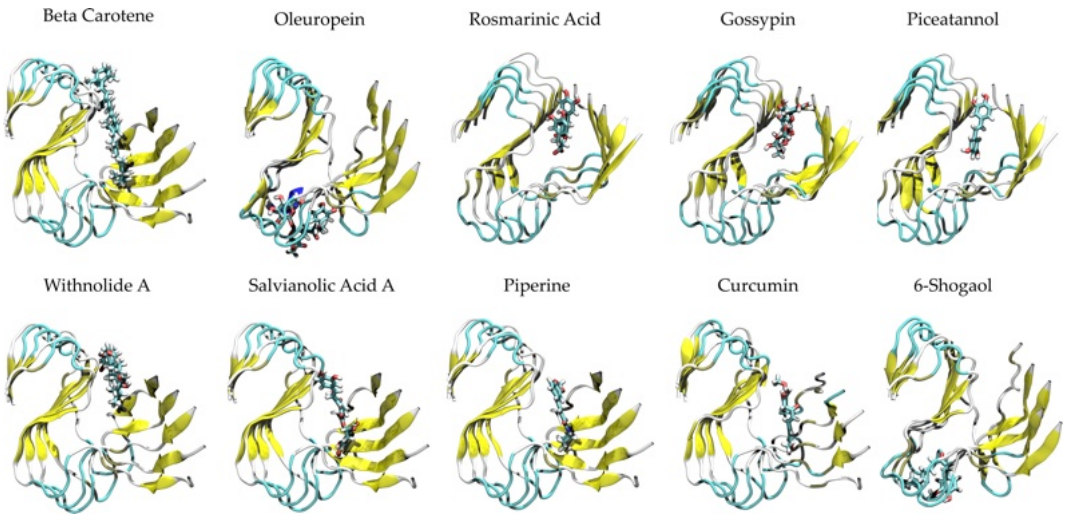


Figure A - 6.4.3. Docking poses of the best ten compounds on the amyloid fibril.

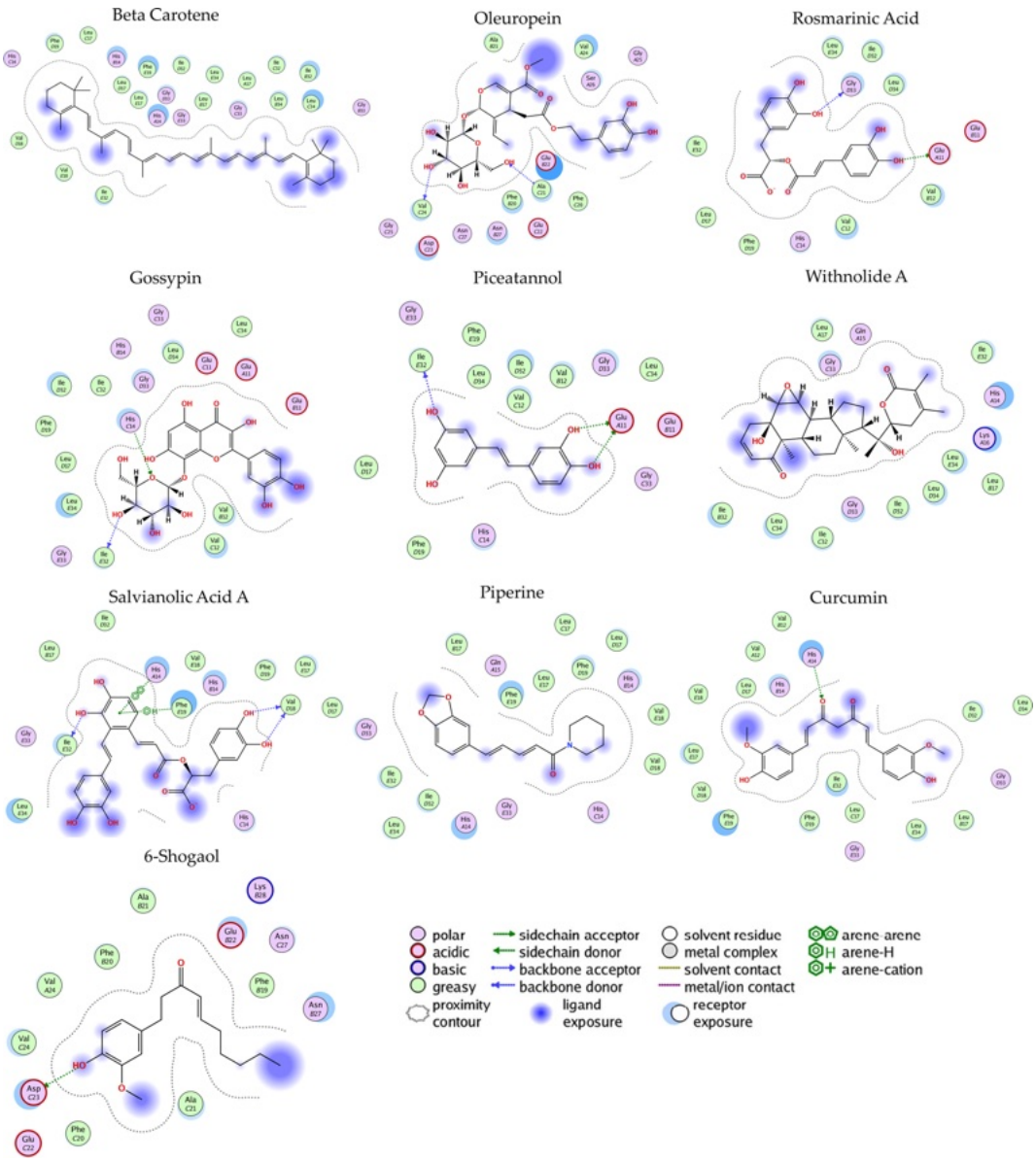


Figure A - 6.4.4. Ligand interactions maps for the best ten investigated natural compounds.

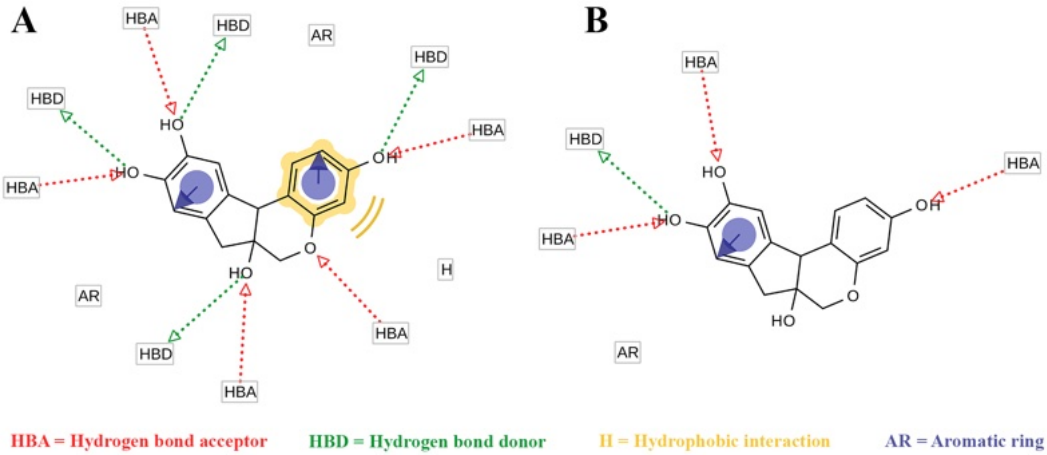


Figure A - 6.4.5. (A) Brazilin ligand-based pharmacophore and (B) shared features pharmacophore between brazilin and mechanism I and II destabilizing compounds, i.e. 6-shogaol, oleuropein, curcumin, gossypin and piceatannol.

6.5 Machine Learning for Taste Prediction

Area under the curve (AUC) measures the two-dimensional area under the ROC curve. It provides an aggregate measure of performance by examining all possible classification thresholds. The AUC is scale-invariant and is an indicator of how well predictions are classified, rather than their absolute values. The value is included between 0 and 1, and it is expressed in percentage. The more the value is far from 1, the more the model predicts wrong. An area of 0.5 corresponds to a random classifier.

Sensitivity (SE) represents the true positive rate, the number of positive data correctly classified.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Specificity (SP) represents the true negative rate, the number of negative data correctly classified.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Accuracy (ACC) is the ratio of the number of correct predictions to the total number of input samples, it represents how well the ML algorithms correctly classify the samples.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Precision (PRC), or ***Positive Predictive Value (PPV),*** represents the is the number of the real positive samples divided by the number of positive results predicted by the classifier.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Non-Error Rate (NER) represents the arithmetic mean of Sensitivity and Specificity in binary classification.

$$Non - Error\ Rate = \frac{Sensitivity + Specificity}{2}$$

F_β-score allows weighting precision and recall, especially in an unbalanced dataset.

$$F_\beta - score\ = (1 + \beta^2)\ \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

For β = 1, F_1-score (F1) is the Harmonic Mean between precision and recall. It tells how precise and robust your classifier is. High precision but lower recall means an extremely accurate classifier, but it misses a large number of instances difficult to classify. The range of the F_1-score is [0,1] and the greater it is, the better is the model performance.

Matthew's correlation coefficient (MCC) is a single-value metric that summarizes the confusion matrix. This coefficient has a high value only if it classifies correctly both positive and negative elements. When the classification is perfect, MCC value is 1.

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

Table A - 6.5.1. Summary of the main recent taste prediction tools, including the methods, the datasets and the molecular descriptors employed.

Reference	Method	Dataset			Molecular Descriptors
		Taste	Source	#	
Chéron Sweet Regressor [18]	Sweet Regressor (RF, SVR)	**Sweet**	SweetenersDB	316	Dragon
Rojas Sweet Predictor [320]	Sweet Classifier (QSTR)	**Sweet**	Sweet	435	2D molecular descriptors (ECFP), Dragon)
		Non-Sweet	Bitter	81	
			Tasteless	133	
Goel Sweet Regressor [338]	Sweet Regressor (GFA, ANN)	**Sweet**	Literature [353–357]	487	Material Studio
e-Sweet [339] [https://bit.ly/3wFy4ER]	Sweet Classifier (KNN, SVM, GBM, RF, DNN)	**Sweet**	SuperSweet	530	Extended-connectivity Fingerprint (ECFP)
			SweetenersDB		
			TasteDB		
			BitterSweet Forest		
		Non-Sweet	BitterDB	718	
			Tasteless (TastesDB)	132	
Predisweet [340] [https://bit.ly/3reop7a]	Sweet Regressor (AB)	**Sweet**	SweetenersDB	316	Dragon and open source (RDKit, Mordred, ChemoPy)
BitterX [341] [https://bit.ly/3wJYa9O]	Bitter Classifier (SVM)	**Bitter**	BitterDB	539	Descriptors from the Handbook of Molecular Descriptors (Todeschini and Consonni, 2007)
		Non-Bitter	In-house experiments	20	
			Available Chemicals Directory (ACD)	519	
BitterPredict [343] [https://bit.ly/3igrzmQ]	Bitter Classifier (AB)	**Bitter**	BitterDB	632	Canvas (Schrödinger)
			TastesDB	59	

		Non-Bitter	Fenaroli's Handbook of Flavor ingredients	1451	
			Literature	35	
			Sweet [320]	336	
			Tasteless [320]	130	
e-Bitter [344] [https://bit.ly/3epWzQq]	Bitter Classifier (KNN, SVM, RF, GBM, DNN)	**Bitter**	BitterDB	707	Extended-connectivity Fingerprint (ECFP)
			TastesDB		
			Rodgers et al., 2006		
		Non-Bitter	Tasteless (TasteDB)	132	
			Non-bitter (BitterX)	17	
			Sweet (SweetenersDB; SuperSweet; [320,362])	443	
iBitter-SCM [317] [https://bit.ly/2VGyXAg]	Bitter Peptides Classifier (SCM)	**Bitter**	Literature (peptides w/ experimental bitterness)	320	Dipeptide composition (DPC)
		Non-Bitter	Randomly from BIOPEP	320	
BERT4Bitter [345] [https://bit.ly/2WecTxf]	Bitter Peptides Classifier (BERT)	**Bitter**	Literature (peptides w/ experimental bitterness)	320	Dipeptide composition (DPC)
		Non-Bitter	Randomly from BIOPEP	320	
iBitter-Fuse [346] [https://bit.ly/3BmC547]	Bitter Peptides Classifier (SVM)	**Bitter**	Literature (peptides w/ experimental bitterness)	320	DPC, AAC, PAAC, APAAC, AAI
		Non-Bitter	Randomly from BIOPEP	320	
BitterIntense [347]	Bitter Intensity Classifier (XGBoost)	**VB**	BATA model	246	Canvas (Schrödinger)
			BitterDB		
			AnalytiCon's repository		
		NVB	Random from non-bitter of BitterPredict	404	
			BitterDB		
			AnalytiCon's repository		
BitterSweetForest [348]	Bitter/Sweet Classifier (RF)	**Sweet**	SuperSweet	517	RDKit (Binary fingerprints)
		Bitter	BitterDB	685	
BitterSweet [349] [https://bit.ly/3rd7Att]	Bitter/Sweet Classifier (AB, RF)	**Bitter**	TasteDB	918	Canvas (Physicochemical and ADMET) Dragon 7 (Extended Connectivity Fingerprints, 2D Molecular Descriptors and 3D Molecular Descriptors)
			Rodgers et al., 2006		
			Fenaroli's Handbook of Flavor Ingredient		
			Biochemical Targets of Plant Bioactive Compounds		
			BitterDB		

			The Good Scents Company Database		ChemoPy (2D Topological and Structural Features)
			BitterPredict (Phyto-Dict., Bitter-New, UNIMI)		
		Non-Bitter	Bitter Predict (Phyto-Dict., UNIMI)	1510	
		Sweet	TastesDB	1205	
			Fenaroli's Handbook of Flavor Ingredient		
			Biochemical Targets of Plant Bioactive Compounds		
			SuperSweet		
			The Good Scents Company Database		
		Non-Sweet	Tasteless (TasteDB, Fenaroli' s Handbook, ToxNet)	1171	
			Bitter molecules		
iUmami-SCM [319] [https://bit.ly/3hJs9uf]	Umami Classifier (SCM)	**Umami**	BIOPEP-UWM	140	Dipeptide composition (DPC)
		Non-Umami	Bitter: iBitter-SCM	304	
VirtualTaste [350] [https://bit.ly/2UfVFPj]	Multi-taste classifier (RF)	**Sweet**	SuperSweet	2011	*not reported*
		Bitter	BitterDB	1612	
			BitterSweet Forest		
		Sour	Manually edited from ChEMBL	1347	

6.6 VirtuousUmami

6.6.1 Materials and Methods – Data curation

Table A - 6.6.1. Summary of the starting dataset, i.e. the UMP442 database.

Class	Number	References
umami	140	Previous literature[61,332–336] and the BIOPEP-UWM database[325]
non-umami	302	Bitter peptides from BTP640 database[337]

Table A - 6.6.2. Summary of the final dataset used in the present work.

Set	Class	Number
Training	umami	240*
	non-umami	240
Test	umami	28
	non-umami	62

*Since the non-umami class is oversampled in the training set, we created synthetic data by randomly duplicating some umami compounds to balance the training dataset

6.6.2 Results - Model Construction and Performance

The evaluation functions of maximization of predictive performance, minimization of selected features and simplicity of the classification model, which were used for guiding the optimization process are the following:

- Selected Features Number Minimization (SFNM):

$$SFNM = \frac{1}{1 + Number\ of\ selected\ features}$$

- Accuracy (ACC):

$$ACC = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}$$

where Tp represents the true positives, Tn the true negatives, Fp the false positives and Fn the false negatives.

- Precision (PRC):

$$PRC = \frac{Tp}{Tp + Fp}$$

- Recall (REC):

$$REC = \frac{Tp}{Tp + Fn}$$

- F1 Score (F1):

$$F1 = \frac{2 * \mathrm{PRC} * \mathrm{REC}}{\mathrm{PRC} + \mathrm{REC}}$$

- F2 Score (F2):

$$F2 = \frac{5 * \mathrm{PRC} * \mathrm{REC}}{4 * \mathrm{PRC} + \mathrm{REC}}$$

- ROC-AUC: Area Under the Receiver Operating Characteristic curve of Sensitivity/Specificity

- Number of SVs or Trees Minimization: Number of Samples in Training Set/Number of Support Vectors of the trained Support Vector Regression Problem

We developed 5 different models summarized in Table A - 6.6.3.

Table A - 6.6.3. Summary of the 5 developed models, including the number of support vectors in the SVM implementation and the number and type of features selected by each model.

Model	***#Support Vectors***	***Selected Features***
1	340	7: AATSC0m, Mp, Mi, FilterItLogS, SMR_VSA1, JGI1, JGT10
2	482	7: AATSC0m, Mi, SaaCH, fragCpx, FilterItLogS, VSA_EState7, JGI1
3	148	8: ATSC1m, Xch_6d, Mi, SaaCH, SMR_VSA1, JGI1, FilterItLogS, JGT10
4	340	10: ATSC1Z, AATSC0m, Mp, Mi, SaaCH, fragCpx, FilterItLogS, SMR_VSA1, JGI1, JGT10
5	148	8: AATSC0m, AATSC0v, Mp, Mi, SaaCH, fragCpx, FilterItLogS, JGI1

The model performance was evaluated on the test set for all models (Table A - 6.6.4).

Table A - 6.6.4. Performance of the 5 SVM developed models.

Model	***ACC***	***Spec***	***Sens***	***F1***	***F2***	***AUC***
1	73%	93.44%	28.57%	40%	32.26%	0.61
2	77.53%	95.08%	39.29%	52.38%	43.65%	0.67
3	85.39%	90.16%	75%	76.36%	75.54%	0.83
4	73%	93.44%	28.57%	40%	32.26%	0.61
5	86.52%	90.16%	78.57%	78.57%	78.57%	0.84

To improve the predictor's performance, ten ensemble models (EMs) were built by combining two different SVM (1 and 2; 1 and 3; 2 and 4; etc..) out of the five ones developed in this work. A comparative performance analysis highlighted $EM_{3\text{-}5}$ (combination of SVM models 3 and 5) as the best ensemble model (Table A - 6.6.5)

Table A - 6.6.5. Performance of the ensemble models (EMs) optimised by combining the 5 SVM models. The ensemble model EM3-5 (combination of SVM models 3 and 5) achieved the best performance.

EM	*ACC*	*Spec*	*Sens*	*F1*	*F2*	*AUC*
$EM_{1\text{-}2}$	77.5%	93.44%	42.86%	54.55%	46.88%	0.68
$EM_{1\text{-}3}$	84.27%	88.52%	75%	75%	75%	0.82
$EM_{1\text{-}4}$	73.03%	93.44%	28.57%	40%	32.26%	0.61
$EM_{1\text{-}5}$	85.39%	88.52%	78.57%	77.19%	78.01%	0.84
$EM_{2\text{-}3}$	86.52%	90.16%	78.57%	78.57%	78.57%	0/84
$EM_{2\text{-}4}$	77.57%	93.44%	42.86%	54.55%	46.88%	0.68
$EM_{2\text{-}5}$	86.52%	90.16%	78.57%	78.57%	78.57%	0.84
$EM_{3\text{-}4}$	84.27%	88.52%	75%	75%	75%	0.82
$EM_{3\text{-}5}$	**87.64%**	**91.80%**	**78.57%**	**79.31%**	**80.99%**	**0.85**
$EM_{4\text{-}5}$	85.39%	88.52%	78.57%	77.19%	78.01%	0.84

6.6.3 Results - Feature Importance

The distributions of the 12 most significant features on which the prediction relies are represented in Figure A - 6.6.1 and Figure A - 6.6.2.

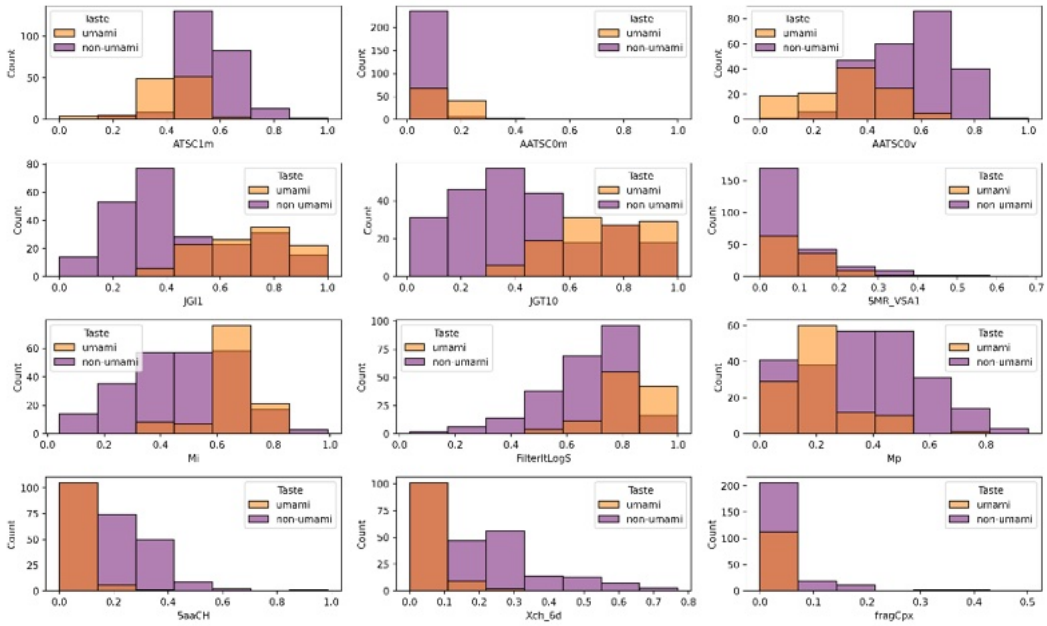


Figure A - 6.6.1. Distribution of the umami and non-umami data for the 12 most significant features

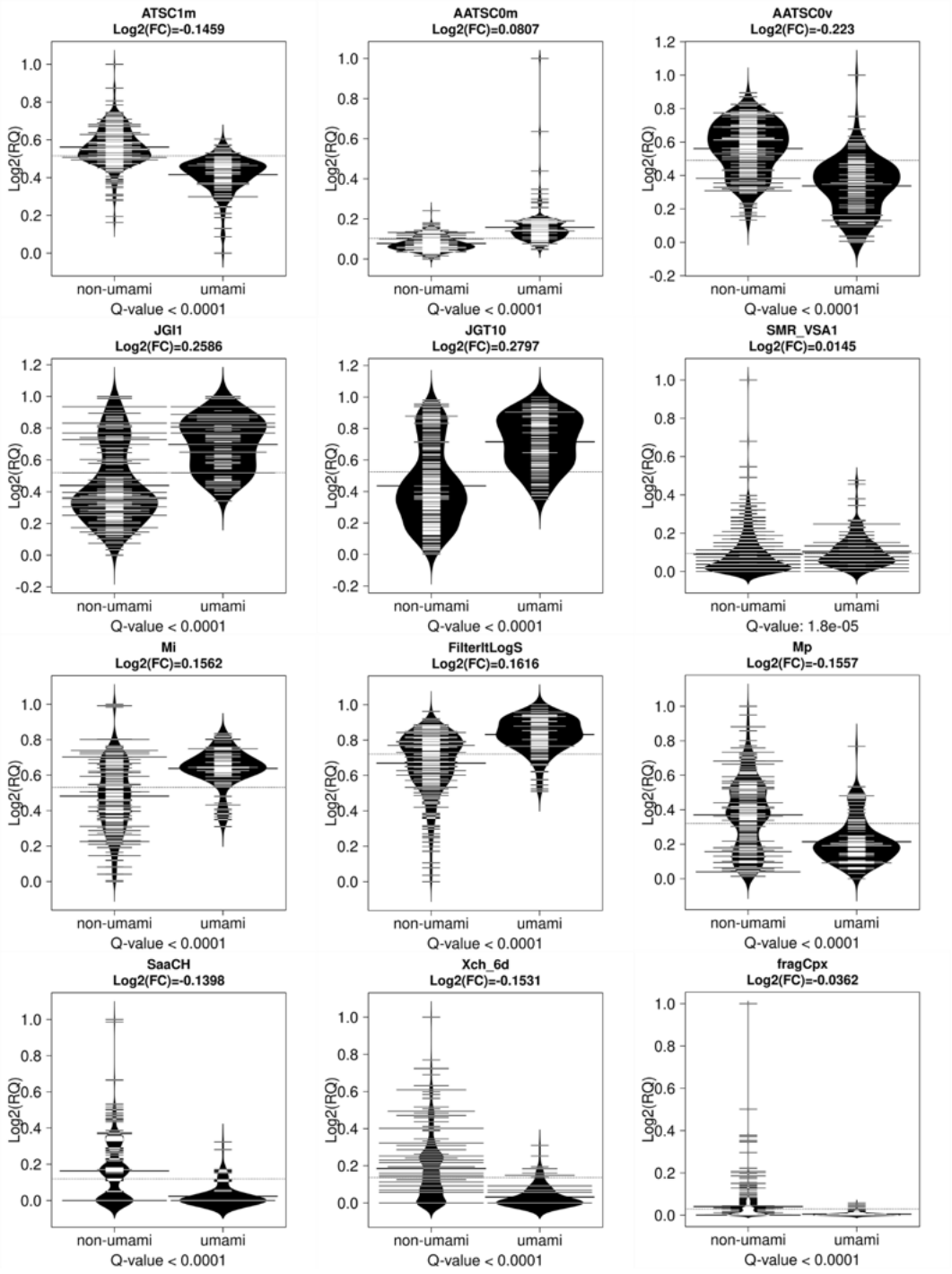


Figure A - 6.6.2. Violin plots showing the distribution of the 12 features in the umami and the non-umami compounds.

Figure A - 6.6.3 represents the hierarchical clustering of the best 12 features, highlighting three major clusters.

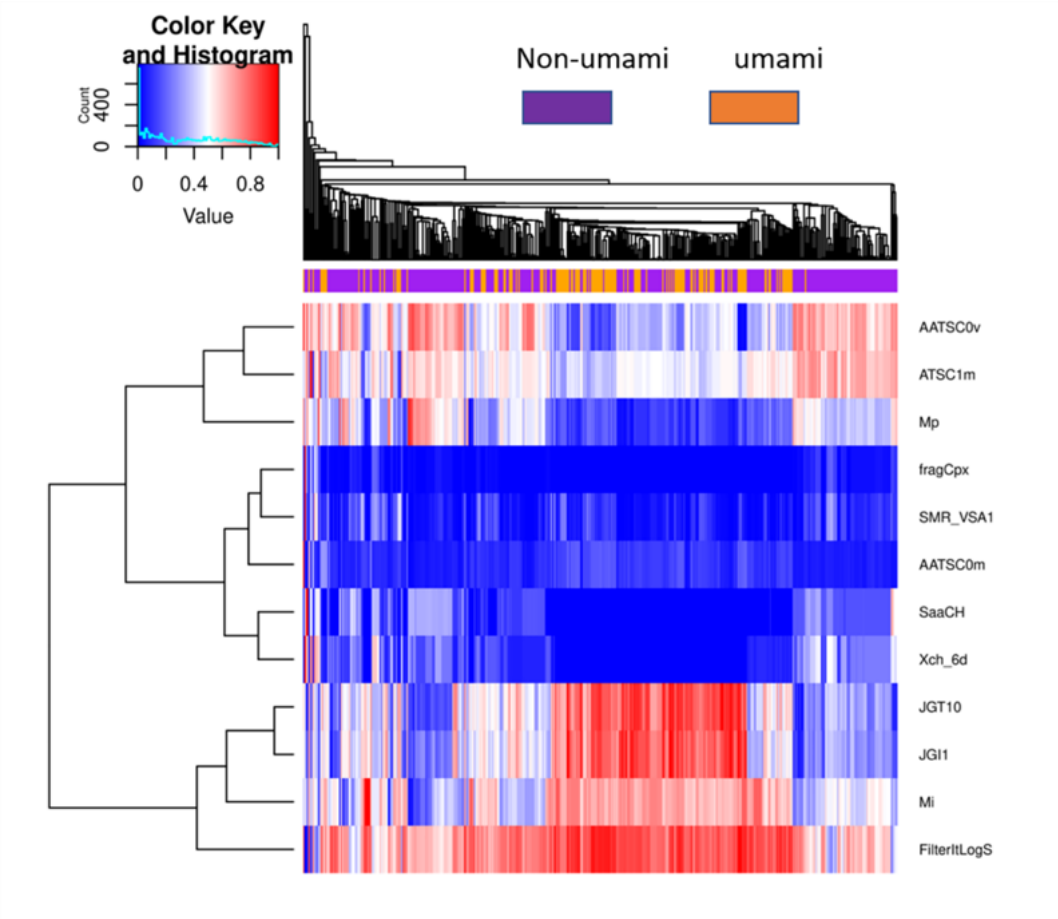


Figure A - 6.6.3. Hierarchical clustering of the selected features reveals 3 groups of features

6.6.4 Discussion

Table A - 6.6.6. Comparison between VirtuousUmami and state-of-the-art umami prediction tools on the VirtuousUmami test set.

	ACC	*Spec*	*Sens*	*F1*	*F2*
iUmami-SCM	86.7	93.5	71.4	76.9	73.5
UMPred-FRL	88.9	93.5	78.6	81.5	79.7
VirtuousUmami	**87.6**	**91.8**	**78.6**	**79.3**	**81.0**

Results indicate that considered predictors have comparable performance when tested on the VirtuousUmami independent test set. In this context, it is worth mentioning that some of the compounds present in the VirtuousUmami independent test set may come from training sets used to develop iUmami-SCM and UMPred-FRL. This could result in an unfavourable condition for the VirtuousUmami algorithm for which the test set is completely unknown. Despite this potential adverse condition, our algorithm demonstrates a predictive power at least comparable with its predecessors.

6.7 VirtuousSweetBitter

Table A - 6.7.1. Summary of the collected compounds from the selected taste databases.

Reference	Taste	#
Biochemical Targets of Plant Bioactive Compounds by Gideon Polya[448]	Bitter	39
	Sweet	32
BitterDB[439]	Bitter	1018
Fenaroli Handbook of Flavor Ingredient[449]	Bitter	16
	Sweet	419
Rodgers et al. (2006) [434]	Bitter	17
Rojas et al. (2017)[450]	Bitter	69
	Sweet	427
SuperSweet[451]	Sweet	265
The Good Scents Company Database	Bitter	37
	Sweet	153
Wiener et al. (2017)[437]	Bitter	75
SweetenersDB[452]	Sweet	119

Table A - 6.7.2. Comparison of the main bitter/sweet prediction models

Reference	Source	Molecular descriptors	Feature selection	(Best) Modelling approach	Interpretation
BitterSweetForest[442]	BitterDB and SuperSweet	Morgan, Atom-Pair, Torsion and Morgan Feat fingerprints from RDkit	Based on performance	Random Forest with Morgan fingerprint	Bayesian-based feature analysis

BitterSweet[443]	Biochemical Targets of Plant Bioactive Compounds by Gideon Polya, BitterDB, SuperSweet, Fenaroli's Handbook of Flavor Ingredients (5th Edition), Rodgers et al., Rojas et al., TOXNET, The Good Scents Company Database, Wiener et al.	ChemoPy, Dragon 2D, Dragon 2D/3D, Canvas and ECFPs	Boruta feature selection algorithm and PCA	Dragon2D/3D molecular descriptor and Boruta feature selection with Adaboost (sweet/non-sweet), Dragon2D/3D molecular descriptor and PCA with Adaboost (bitter/non-bitter)	Random forest relative feature importance with mean decrease in Gini impurity
VirtualTaste[444]	BitterDB, SuperSweet and BitterSweetForest tool	MACCS and Morgan fingerprints from RDkit	\	Random Forest	Bayesian-based feature analysis
Ours	Biochemical Targets of Plant Bioactive Compounds by Gideon Polya, BitterDB, SuperSweet, Fenaroli's Handbook of Flavor Ingredients (5th Edition), Rodgers et al., Rojas et al., The Good Scents Company Database, Wiener et al., SweetenersDB	2059 molecular descriptors from RDkit, pybel and Mordred open-source libraries	Sequential feature selection based on hierarchical clustering on the feature's Spearman rank-order and two-sample Kolmogorov - Smirnov test	Gradient Boosting (LightGBM)	Global (feature importance, dependence plots) and local interpretation (features impact on individual predictions) based on SHAP values

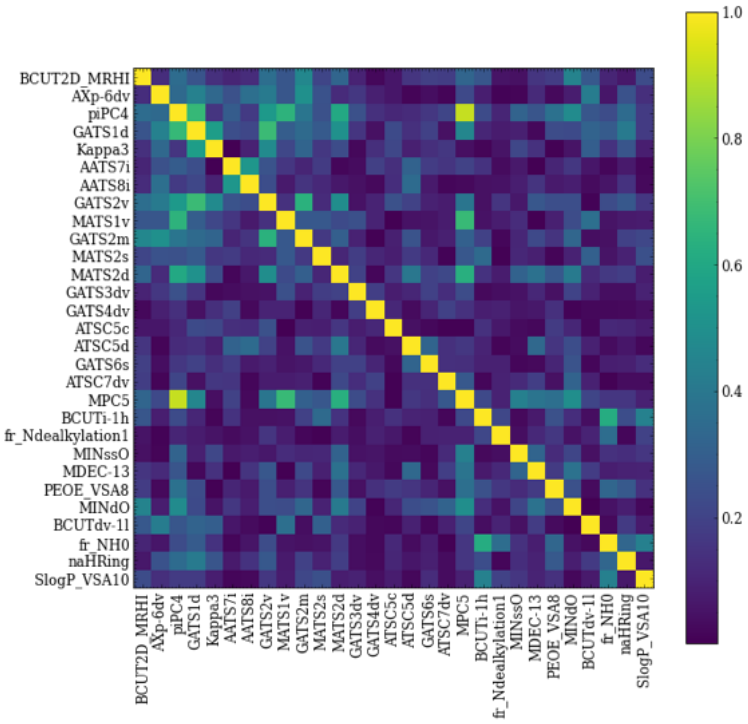


Figure A - 6.7.1. Heatmap of the selected features correlation matrix computed with Spearman's rank correlation in absolute value.

6.7.1 Validation on non-bitter/non-sweet molecules

This study is focused on the sweet/bitter dichotomy in order to isolate the most suitable variables capable of highlighting the differences between sweet and bitter compounds. However, it is also interesting to analyze the behavior of the model, fed only by the variables selected in this work, in the prediction of neither sweet nor bitter molecules. For this purpose, the original training dataset consisting of 2686 compounds (1415 sweet and 1271 bitter) was augmented by 198 additional compounds classified in the literature as neither bitter nor sweet[449,450,505] by converting the original binary classification problem into a multiclass problem with 3 labels. The final LightGBM model was retrained on this augmented dataset and performance was assessed according to a stratified 5-fold cross-validation strategy. For each class, the ROC curves are computed through a one-vs-rest method (namely, performance of the considered class against the remaining two ones) and shown in Figure A - 6.7.2, along with the macro-average ROC curve, which equally weights each point of the single ROC curve. Also, for this 3-class problem, the predictive performance is satisfactory, with an average AUROC equal to 0.92. Note that the AUROC for the Non-bitter/sweet class is slightly worse than the average

(0.89). This was expected since the features used by the predictive model was chosen only considering bitter and sweet compounds.

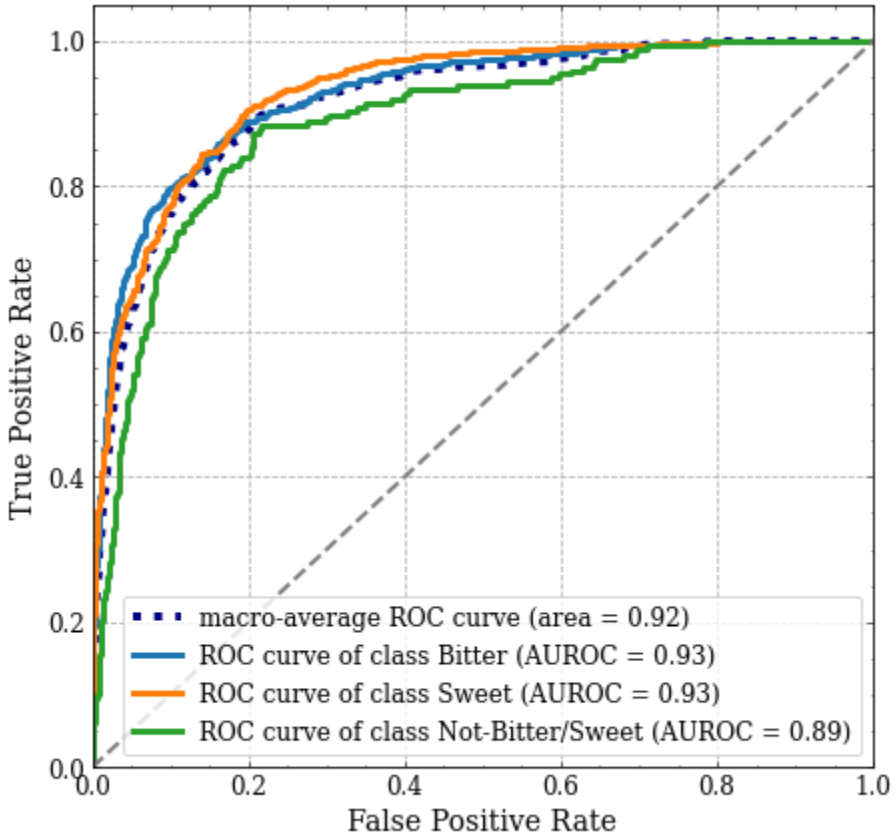


Figure A - 6.7.2. One-vs-rest ROC curves: bitter vs others (blue); sweet vs others (orange); not bitter/not-sweet vs others (green); macro-average ROC curves (dotted dark blue).

6.7.2 Local interpretation

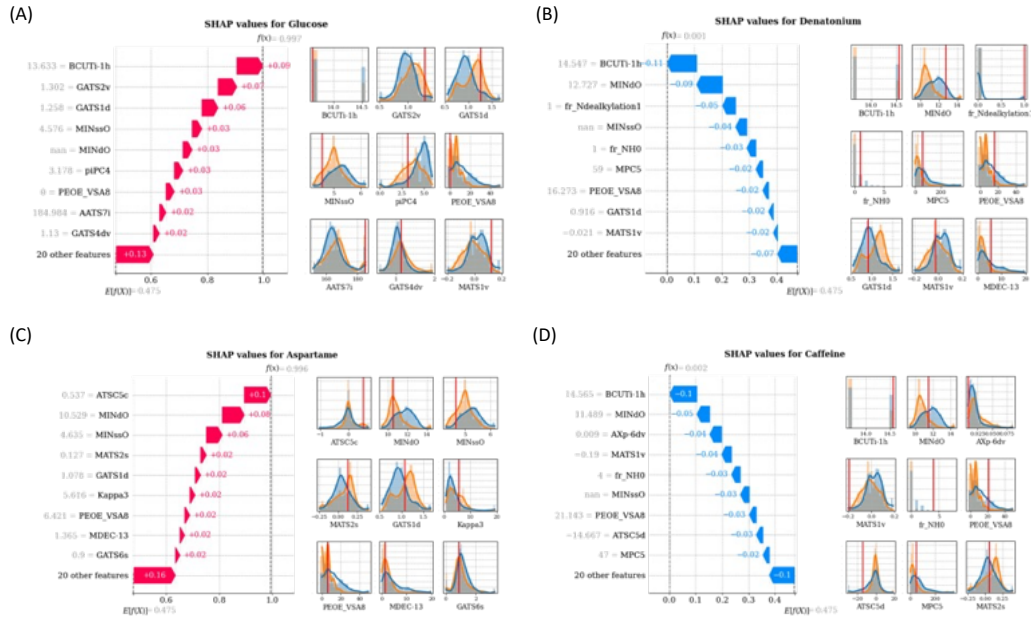


Figure A - 6.7.3. SHAP profiles of four representative molecules: Glucose (A), Denatonium (B), Aspartame (C) , and Caffeine (D). For each figure, SHAP values are shown in the left panel and impacting feature distributions in the right panel, with values assumed by the features highlighted with solid red lines.

Chapter VII

PhD Portfolio

Name	Lorenzo
Surname	Pallante
Affiliation	Politecnico di Torino, PolitoBIOMedLab, Department of Mechanical and Aerospace Engineering, Torino, 10129, Italy
Supervisors	Marco A. Deriu, Umberto Morbiducci
PhD Period	2019-2023

7.1 Peer-reviewed Scientific Publications

2022 **Pallante L**, Korfiati A, Androutsos L, et al. Toward a general and interpretable umami taste predictor using a multi-objective machine learning approach. Sci Rep. 2022;12(1):21735. doi:10.1038/s41598-022-25935-3

2022 Maroni G, **Pallante L**, Di Benedetto G, Deriu MA, Piga D, Grasso G. Informed classification of sweeteners/bitterants compounds via explainable machine learning. Curr Res Food Sci. 2022;5:2270-2280. doi:10.1016/j.crfs.2022.11.014

2022 Malavolta M, **Pallante L**, Mavkov B, et al. A survey on computational taste predictors. Eur Food Res Technol. 2022;248(9):2215-2235. doi:10.1007/s00217-022-04044-5

2021 Sztandera K, Gorzkiewicz M, Dias Martins AS, **Pallante L**, et al. Noncovalent Interactions with PAMAM and PPI Dendrimers Promote the Cellular Uptake and Photodynamic Activity of Rose Bengal: The Role of the Dendrimer Structure. J Med Chem. 2021:acs.jmedchem.1c01080. doi:10.1021/acs.jmedchem.1c01080

2021 **Pallante L**, Malavolta M, Grasso G, et al. On the human taste perception: Molecular-level understanding empowered by computational methods. Trends Food Sci Technol. 2021;116:445-459. doi:10.1016/j.tifs.2021.07.013

2020 Stojceski F, Grasso G, **Pallante L**, Danani A. Molecular and Coarse-Grained Modeling to Characterize and Optimize Dendrimer-Based Nanocarriers for Short Interfering RNA Delivery. ACS Omega. 2020;5(6):2978-2986. doi:10.1021/acsomega.9b03908

2020 **Pallante L**, Rocca A, Klejborowska G, et al. In silico Investigations of the Mode of Action of Novel Colchicine Derivatives Targeting β-Tubulin Isotypes: A Search for a Selective and Specific β-III Tubulin Ligand. Front Chem. 2020;8(February). doi:10.3389/fchem.2020.00108

2020 Muscat S, **Pallante L**, Stojceski F, Danani A, Grasso G, Deriu MA. The Impact of Natural Compounds on S-Shaped Aβ42 Fibril: From Molecular Docking to Biophysical Characterization. Int J Mol Sci. 2020; 21(6). doi:10.3390/ijms21062017

2020 Grasso G, **Pallante L**, Tuszynski JA, Morbiducci U, Deriu MA. Computational molecular modelling as a platform for a deeper understanding of protein dynamics and rational drug design. Biomed Sci Eng. Published online February 11, 2020. doi:10.4081/bse.2019.87

7.2 Manuscripts under-review or in preparation

- **Pallante L**[†], Cannariato M[†], Androutsos L, Zizzi EA, Hada X, Mavroudi S, Grasso G, Theofilatos K and Deriu MA. VirtuousPocketome: A Computational Tool for Screening Protein-ligand Complexes to Identify Similar Binding Sites

  [†] Lorenzo Pallante and Marco Cannariato contributed equally to this study.

- Androutsos L, **Pallante L**, Stojceski F, Mavroudi S, Grasso G, Deriu MA and Theofilatos K. VirtuousMultiTaste: Development of a ML-based algorithm to identify multiple taste sensations.

7.3 Scientific Awards

2020 **Cover Art**: ACS Omega, Vol. 5 N. 6, related to the publication "Molecular and Coarse-Grained Modelling to Characterize and Optimize Dendrimer-Based Nanocarriers for siRNA Delivery"

2021 **Cover Art:** Journal of Medicinal Chemistry, Vol. 64 N. 21, related to the publication "Noncovalent Interactions with PAMAM and PPI Dendrimers Promote the Cellular Uptake and Photodynamic Activity of Rose Bengal: The Role of the Dendrimer Structure"

7.4 Teaching Activities

The teaching activities listed below were carried out within the master's degree course in Biomedical Engineering at Politecnico di Torino:

- Biomechanical Design (ENG) – 28 hrs.
- Rational Drug Design (ENG) – 8 hrs.
- Biomeccanica Multiscala (ITA) – 20 hrs.

Additionally, I supervised the following Master and Bachelor Thesis Students:

- 5 Master Thesis Students (https://bit.ly/3oP63MW)
- 14 Bachelor Thesis Students

7.5 PhD Courses

7.5.1 Hard Skills

Course Name	A.Y.	Hours
Computing@Polito Workshop – HPC/Big Data/Cloud for Research	2019-20	4
Principles, materials and applications of robotics in biomedicine	2019-20	20
Multiscale modelling and coarse-graining for flow and transport PDEs (didattica di eccellenza vp)	2019-20	12
ICTP-SISSA-CECAM Workshop on Molecular Dynamics and its Applications to Biological Systems | (smr 3483)	2020-21	18
VirtuousToK – Omics, Machine Learning and Molecular Modelling Targeting Taste and Nutrition	2020-21	11
Planning, management and analysis of clinical and laboratory research	2020-21	15
Non-Extensive Statistical Mechanics	2020-21	10
High Performance Molecular Dynamics	2020-21	18
Plumed Masterclass	2020-21	16
Martini Workshop 2021	2020-21	18

7.5.2 Soft Skills

Course Name	A.Y.	Hours
Project Management	2019-20	5
Entrepreneurial finance	2020-21	5
The new Internet Society: Entering the Black-box of Digital Innovation	2020-21	6
Public Speaking	2020-21	5
Tme management	2020-21	2
Thinking out of box	2020-21	1
Research integrity	2020-22	5
Navigating the hiring process	2020-22	2
Personal branding	2020-22	1
Responsible research and innovation, the impact on social challenges	2020-22	5
Communication	2020-22	5

7.6 International Conferences and Workshops

International Conferences and Workshops	Contribution	Year
Virtuous Transfer of Knowledge (ToK) - First Workshop (oral)	Oral Pres.	2020
ICTP-SISSA-CECAM Workshop on Molecular Dynamics and its Applications to Biological Systems | (smr 3483)	Attendee	2020
CancerTO Nanoscience in Cancer Immunotherapy (oral + poster)	Oral Pres.	2021
26th Congress of the European Society of Biomechanics (oral)	Oral Pres.	2021
High Performance Molecular Dynamics	Attendee	2021
Plumed Masterclass	Attendee	2021
Martini Workshop 2021	Attendee	2021
Virtuous Transfer of Knowledge (ToK) - Second Workshop (oral)	Oral Pres.	2022
8th Annual CCPBioSim Conference Frontiers in Biomolecular Simulation 2022 (poster and flash talk)	Poster	2022
AIDD 2022 Spring School	Attendee	2022
2022 Workshop on MDAnalysis/Machine Learning	Attendee	2022

7.7 International Exchange Periods

Period	Host Institution	Country
12/2019 – 02/2020	Missing Tech Sagl.	Switzerland
10/2020 – 12/2020	Insybio PC	Greece
05/2022 – 06/2022	Missing Tech Sagl.	Switzerland
11/2022 – 12/2022	Missing Tech Sagl.	Switzerland

[*] The periods abroad were inserted in the framework of the European H2020 project MSCA-RISE VIRTUOUS (GA. 872121) for scientific activities related to specific Work Packages (WP) of the project.

About the Author

Lorenzo Pallante (Italy, born in 1994) is currently a Research Fellow at the Department of Mechanical and Aerospace Engineering of Politecnico di Torino under the supervision of Prof. Marco Agostino Deriu within the Mechanistic and Machine Learning-driven Modelling in Bioengineering (M3B) group (https://m3b.it/). Mr Lorenzo Pallante received his B.Sc. in Biomedical Engineering from Politecnico di Torino (Turin, Italy) in 2016. In 2019, he received his M.Sc. cum laude in Biomedical Engineering from Politecnico di Torino (Turin, Italy). Since November 2019, he is enrolled in a joint PhD programme of the University of Turin and the Politecnico di Torino in Bioengineering and Medical-Surgical Sciences focused on the molecular basis of subcellular mechanics, investigated using computational modelling techniques, e.g. molecular dynamics, ensemble molecular docking, metadynamics, free energy calculations, rational drug design, and artificial intelligence tools, such as machine learning-based algorithms and pathway network modelling. His research mainly focuses on the understanding structure-to-function relationships in protein physiological and pathological behaviour. He is also the author of 9 papers in peer-reviewed journals and has participated in 5 international conferences also as a speaker.