Compréhension des effets biologiques de molécules odorantes à l'aide d'outils computationnels et de réseaux biologiques

Marylène Rugard

Université Paris Cité

Ecole doctorale Médicament, Toxicologie, Chimie, Imageries – ED 563

Laboratoire Toxicité environnementale, cibles thérapeutiques, signalisation cellulaire et biomarqueurs, INSERM, UMR 1124

Compréhension des effets biologiques de molécules odorantes à l'aide d'outils computationnels et de réseaux biologiques

Par Marylène RUGARD

Thèse de doctorat de bio-informatique

Dirigée par Pr. Karine AUDOUZE

Et par Dr. Anne TROMELIN

Présentée et soutenue publiquement le 4 Octobre 2022

Devant un jury composé de :

Dr. Patricia DUCHAMP-VIRET, DR, Université Claude Bernard Lyon1	Rapporteur
Dr. Sébastien FIORUCCI, MCU, Université Côte d'Azur	Rapporteur
Pr. Matthieu MONTES, PU, CNAM	Examinateur
Dr. Olivier SPERANDIO, CR, Institut Pasteur	Examinateur
Dr. Gautier MOROY, MCU, Université Paris Cité	Membre invité
Dr. Boriana ATANASOVA, MCU, Université de Tours	Membre invité
Pr. Karine AUDOUZE, PU, Université Paris Cité	Directrice de thèse
Dr. Anne TROMELIN, CR, Université de Bourgogne	Co-directrice de thèse

Remerciements

Ce manuscrit est l'aboutissement de trois années d'études en tant que doctorante au sein de l'UMR 1124. Je souhaite remercier l'Université Paris Cité, l'INSERM et le projet MULTIMIX financé par l'Agence Nationale de la Recherche (ANR-18-CE21-0006-03).

Je remercie particulièrement mes directrice et co-directrice de thèse, Karine AUDOUZE et Anne TROMELIN pour avoir dirigé mes travaux de thèse durant ces trois années. Je les remercie pour leur soutien, leur confiance et pour m'avoir accordé l'opportunité de réaliser cette thèse. Un grand merci pour vos conseils, votre disponibilité et le suivi constant de mes travaux qui m'ont permis de mener à bien ces travaux.

Je souhaite également remercier tous les membres du jury, le Dr. Patricia DUCHAMP-VIRET, le Dr. Sébastien FIORUCCI, le Pr. Matthieu MONTES, le Dr. Olivier SPERANDIO, le Dr. Gautier MOROY, et le Dr. Boriana ATANASOVA, pour l'intérêt portés à mes travaux ainsi que le temps accordé à l'évaluation de ma thèse.

Un grand merci au Pr. Robert BAROUKI pour m'avoir accueillie au sein de l'unité.

Je remercie aussi le Pr. Xavier COUMOUL qui m'a acceptée au sein de son équipe. Merci pour ton soutien, ta bienveillance, ta gentillesse qui m'ont aidée à bien des égards tout au long de cette thèse.

Merci à tous les membres de l'équipe, Karine ANDREAU, Sylvie BORTOLI, Caroline CHAUVET, Martine AGGERBECK, Lawrence AGGERBECK, Etienne BLANC et tous ceux qui ne sont pas cités, pour votre soutien et votre intérêt à mes travaux durant les réunions d'équipe.

Je remercie le groupe SysTox, Patricia JEANNIN, Thomas JAYLET, Thibaut COUSTILLET et Thibault CROUZET, pour leur aide et leurs encouragements ainsi que les anciens membres Qier WU et Elias ZGHEIB. Merci pour tous ces super moments passés tous ensemble qui ont rendu cette thèse encore plus mémorable.

Je remercie Florence JORNOD qui a été présente depuis le début et qui est devenue une amie. Merci pour tout.

Et bien sûr je remercie mes amis, ma famille, mes parents, mon frère, ma belle-sœur, ma nièce et l'homme qui partage ma vie. Merci pour votre amour et votre soutien sans lesquels je n'aurais rien pu faire.

Titre

Compréhension des effets biologiques de molécules odorantes à l'aide d'outils computationnels et de réseaux biologiques

Résumé

Les molécules odorantes sont largement utilisées dans l'alimentation et la parfumerie. La compréhension de la perception des odeurs est donc un enjeu important. La perception olfactive est initiée au niveau de l'épithélium olfactif, par la liaison des molécules odorantes aux récepteurs olfactifs (RO) situés au niveau des cils des neurones olfactifs sensoriels. Les RO sont ainsi activés, déclenchant l'envoi de signaux électriques par ces neurones jusqu'au bulbe olfactif qui les transmet ensuite aux régions supérieures du cerveau. La discrimination des odeurs à partir de l'activation de quelques centaines de RO par des myriades de molécules odorantes passe par un code combinatoire selon lequel un seul odorant peut activer plusieurs RO et un seul RO reconnaît plusieurs odorants. Dans l'environnement naturel, les molécules odorantes sont généralement perçues en mélange. On distingue deux types de perception : la perception hétérogène (l'odeur spécifique de chaque constituant du mélange peut être identifiée) et la perception homogène (une seule odeur est perçue à partir du mélange). Bien que la perception homogène de mélanges d'odorants semble complexe, l'étude des caractéristiques odorantes et structurales des molécules qui les constituent participe au décryptage du code olfactif. En effet, selon le paradigme des relations structure-activité, les molécules odorantes détectées par un même RO devraient posséder des similarités structurales. Mais les RO sont encore majoritairement orphelins, et les mécanismes intervenant au niveau périphérique dans la perception olfactive restent en grande partie à expliquer. Une meilleure compréhension des interactions des RO avec leurs ligands est essentielle pour la compréhension de la perception olfactive, et de plus dépasse le champ de l'olfaction. En effet, des études révélant l'expression ectopique des RO se sont récemment multipliées. Si les RO interviennent dans d'autres processus biologiques que l'olfaction, ils pourraient également constituer des cibles thérapeutiques. En chimie médicinale, les modèles générés à partir des approches *in silico* constituent une alternative aux expérimentations longues et coûteuses. C'est pourquoi dans le cadre de cette thèse, nous avons créé différents

modèles informatiques intégratifs et prédictifs dans le but de mieux comprendre les mécanismes du processus olfactif au niveau périphérique. Afin de mieux cerner les structures moléculaires qui confèrent aux molécules odorantes leurs propriétés olfactives et biologiques, nous avons utilisé un ensemble d'approches impliquant réduction de dimension et clustering, génération de pharmacophores et construction d'un réseau biologique. Les résultats obtenus au cours de ce travail de thèse ont montré la pertinence des outils informatiques pour explorer les relations entre odorants, odeurs et RO. Ils ont permis de suggérer des hypothèses sur les modes d'interactions des composants des deux mélanges étudiés au niveau périphérique. Le réseau odorome permettra de proposer de nouvelles associations odorant-RO et des pistes utiles pour explorer le rôle biologique au sens large des molécules odorantes.

Mots clés : bioinformatique, chémoinformatique, molécules odorantes, mélanges, odeurs, perception homogène, récepteurs olfactifs, pharmacophores, classification, réseau biologique

Title

Understanding the biological effects of odorant molecules using computational tools and biological networks

Abstract

Odorant molecules are widely used in food and perfumery. The understanding of odor perception is therefore an important issue. Olfactory perception is initiated at the level of the olfactory epithelium, by the binding of odorant molecules to olfactory receptors (ORs) located at the level of the cilia of the olfactory sensory neurons. The ORs are thus activated, triggering the sending of electrical signals by these neurons to the olfactory bulb, which then transmits them to the higher regions of the brain. The discrimination of odors from the activation of a few hundred ORs by myriads of odorant molecules requires a combinatorial code according to which a single odorant can activate several ORs and a single OR recognizes several odorants. In the natural environment, odorant molecules are generally perceived as a mixture. Two types of perception are distinguished: heterogeneous perception (the specific odor of each component of the mixture can be identified) and homogeneous perception (only one odor is perceived from the mixture). Although the homogeneous perception of odorant mixtures seems complex, the study of the odorant and structural characteristics of the molecules that constitute them participates in the deciphering of the olfactory code. Indeed, according to the structure-activity relationship paradigm, odorant molecules detected by the same OR should have structural similarities. However, ORs are still mostly orphans, and the mechanisms involved at the peripheral level in olfactory perception remain largely unexplained. A better understanding of the interactions of ORs with their ligands is essential for the understanding of olfactory perception, and moreover goes beyond the field of olfaction. Indeed, studies revealing the ectopic expression of ORs have recently multiplied. If ORs are involved in biological processes other than olfaction, they could also constitute therapeutic targets. In medicinal chemistry, models generated from *in silico* approaches are an alternative to long and costly experiments. Therefore, in this thesis, we have created different integrative and predictive computational models in order to better understand the mechanisms of the olfactory process at the peripheral level. In order to better understand the molecular

structures that give odorant molecules their olfactory and biological properties, we used a set of approaches involving dimension reduction and clustering, pharmacophore generation and biological network construction. The results obtained during this thesis have shown the relevance of computational tools to explore the relationships between odorants, odors and ORs. They allowed to suggest hypotheses on the interaction modes of the components of the two mixtures studied at the peripheral level. The odorome network will allow us to propose new odorant-OR associations and useful leads to explore the biological role of odorant molecules in a broad sense.

Key words: bioinformatics, chemoinformatics, odorant molecules, mixtures, odors, homogeneous perception, olfactory receptors, pharmacophores, classification, biological network

Abréviations

Acides Aminés

A : Alanine

C : Cystéine

D : Aspartate

F : Phénylalanine

I : Isoleucine

H : Histidine

L : Leucine

M : Méthionine

N : Asparagine

P : Proline

R : Arginine

S : Sérine

T : Thréonine

V : Valine

Y : Tyrosine

Acronymes

ACP : Analyse en composantes principales

ANN : Artificial Neural Network

CAH : Classification ascendante hiérarchique

CCG : Couche des cellules granulaires

CCM : Couche des cellules mitrales

CG : Couche glomérulaire

CNO : Couche du nerf olfactif

CPE : Couche plexiforme externe

CPI : Couche plexiforme interne

FPR : Récepteurs de peptides formiques

MDS : Multidimensional Scaling

QSAR : Quantitative Structure–Activity Relationships

RCPG : Récepteurs couplés aux protéines G

RO : Récepteurs olfactifs

RV : Récepteurs voméronasaux

SAR : Structure–Activity Relationships

SNE : Stochastic Neighbor Embedding

SOM : Self-Organizing Maps

TAAR : Récepteurs associés aux traces d’amines

TF3P : Three-Dimensional Force Fields Fingerprint

t-SNE : t-distributed Stochastic Neighbor Embedding

UMAP : Uniform Manifold Approximation and Projection

VIH : Virus de l'immunodéficience humaine

Table des illustrations

SOMMAIRE

Partie 1 : Introduction

I. Le système olfactif

Le système nerveux permet à un organisme de comprendre l'environnement qui l'entoure et aide ainsi l'organisme à réagir et à interagir avec cet environnement. Le système sensoriel constitue la partie la plus fondamentale du système nerveux. Le système sensoriel est composé des cinq sens principaux (la vision, l'audition, le goût, l'olfaction et le toucher) ainsi que de sens plus complexes tels que l'équilibre, l'entéroception (informations sensorielles provenant des parois des organes internes creux) et la proprioception (position des membres) (Hellier, 2017). L'olfaction peut être considérée comme le sens primaire. Perçue par le premier nerf crânien, l'olfaction est phylogénétiquement le sens le plus ancien, et le premier des systèmes sensoriels à se développer embryologiquement chez les mammifères (Fjældstad, 2018). L'olfaction permet la perception d'odeurs issues de la détection de molécules odorantes dans l'environnement (Wooten, 2015). Cependant les organes olfactifs diffèrent chez les animaux selon leur espèce, bien que les cellules réceptrices olfactives possèdent une structure et des propriétés communes. Chez les mammifères, la détection des molécules odorantes se fait via l'épithélium olfactif qui se trouve au fond de la cavité nasale, tandis que chez la plupart des insectes, la détection est externe et au niveau des antennes (Mori, 2013). Dans ce chapitre, nous nous intéresserons au système olfactif des mammifères.

A. Physiologie du système olfactif

Le système olfactif des mammifères est complexe et constitue l'un des systèmes sensoriels les plus évolués. L'olfaction est initiée par la détection des molécules odorantes au niveau de l'épithélium olfactif (*cf*. partie Muqueuse et épithélium olfactifs) de la cavité nasale où se situent les récepteurs olfactifs et qui constitue le système olfactif périphérique (Galizia and Lledo, 2013). Les récepteurs olfactifs situés sont ainsi activés, et génèrent des signaux qui vont jusqu'au cerveau en passant par le bulbe olfactif (Lledo et al., 2005). Toutes les structures impliquées dans ce processus forment le système olfactif principal.

Un second système olfactif, appelé système voméronasal (ou système olfactif accessoire), est également présent et se situe sur la partie inférieure de la cavité nasale. Ce système est

spécialisé dans la reconnaissance des phéromones et joue donc un rôle clé dans les comportements sexuels et sociaux tels que la territorialité, l'agression et l'allaitement (Firestein, 2001; Mori, 2001). Les phéromones sont perçues par l'organe voméronasal qui transfert ensuite les signaux au bulbe olfactif accessoire qui est situé dans la partie dorsocaudale du bulbe olfactif principal. Pour autant, les deux bulbes olfactifs sont entièrement séparés sur le plan fonctionnel (Brennan, 2001). Les zones du cerveau qui reçoivent les signaux issus du bulbe olfactif accessoire diffèrent du cortex olfactif qui est la zone de projection majeure des signaux issus du bulbe olfactif principal (Mori, 2001). En effet, le bulbe olfactif accessoire envoie uniquement des projections vers l'amygdale, transmises ensuite jusqu'à l'hypothalamus (Brennan, 2001). Par conséquent, les perceptions olfactives issu de chaque système olfactif entraînent des résultats comportementaux et émotionnels différents (Mori, 2001). Cependant le rôle de l'organe voméronasal chez l'Homme est controversé (D'Aniello et al., 2017). La fonction sensorielle de l'organe voméronasal humain serait non opératoire d'après certaines études (Dulac and Torello, 2003) tandis que d'autres suggèrent une activité endocrine potentielle de l'organe voméronasal (Wessels et al., 2014).De ce fait, nous nous limiterons à la description du système olfactif principal.

1. Les molécules odorantes : la nature des stimuli

Le système olfactif est capable de détecter de nombreuses et diverses molécules odorantes, ce qui lui offre une impressionnante capacité à percevoir et à décrire notre environnement (Cowart and Rawson, 2005). La détection des molécules odorantes peut se réaliser par voie orthonasale lorsqu'elles sont inhalées directement par les narines, ou par voie rétronasale lorsqu'en mangeant ou en buvant, elles atteignent l'épithélium olfactif en passant par l'arrière-bouche (Genva et al., 2019a; Meierhenrich et al., 2005a). Plus de 7000 molécules odorantes ont été identifiées (Dinu et al., 2020) et bien que le nombre d'odeurs pouvant être perçues reste encore inconnu, il pourrait atteindre le trillion (Bushdid et al., 2014). Les molécules odorantes sont de petits composés chimiques (avec généralement un poids moléculaire inférieur à 350 Da) libérés par un substrat et qui peuvent être perçus par un organisme vivant. Il peut s'agir de petites molécules telles que des gaz (CO_2 ou H_2S), ou de molécules organiques et volatiles plus complexes (alcools, esters, éthers, cétones, aldéhydes, amine, amide, dérivés soufrés, linéaires, ramifiées, cycliques, polycycliques, hétérocycliques,

aromatiques…) (Fráter et al., 1998; Galizia and Lledo, 2013; Kermen et al., 2011). La plupart des molécules odorantes sont organiques, non ioniques, et pour les animaux terrestres hydrophobes. Cependant, les principales caractéristiques chimiques qui permettent à une molécule d'avoir une propriété odorante sont sa volatilité et sa solubilité dans le mucus recouvrant les cellules réceptrices olfactives (Cowart and Rawson, 2005; Galizia and Lledo, 2013). La volatilité d'une molécule odorante régit son aptitude à pénétrer dans la cavité nasale, et la solubilité dans le mucus définit donc sa facilité à y pénétrer afin d'accéder aux récepteurs olfactifs. Ces deux propriétés chimiques dépendent majoritairement des caractéristiques chimiques de la molécule telles que les groupements chimiques (aldéhydes, alcools…), la longueur de la chaîne carbonée ou encore la polarité des groupes latéraux (Cowart and Rawson, 2005).

Toutes ces caractéristiques influencent les propriétés odorantes d'une molécule. Cependant, il reste difficile d'établir une relation systématique entre ces caractéristiques et les qualités odorantes d'une molécule. Ainsi, des molécules présentant peu de ressemblance sur le plan structurel peuvent avoir la même odeur, et à l'inverse des molécules quasi-identiques sur le plan structurel peuvent présenter des qualités odorantes très différentes (Cowart and Rawson, 2005; Snitz et al., 2013). C'est en particulier le cas des énantiomères (Russell and Hills, 1971). Par exemple, les deux énantiomères (R)-γ-methylcyclogeranate et (S)-γ-methylcyclogeranate ont respectivement une odeur de camphre et une odeur fruitée en dépit de leurs structures très proches. A la différence de ces deux énantiomères, le 2-(1-(3,3-dimethylcyclohexyl)ethoxy)-2-methylpropyl propionate et le 3-methylcyclopentadecanone possèdent une structure très différente mais présentent tous deux une odeur musquée (Genva et al., 2019a). Il existe également des cas, où une odeur est perçue pour un seul des énantiomères, comme pour le (1R,2S)-cis methyl jasmonate qui possède une odeur de jasmin alors que le (1S,2R)-cis methyl jasmonate est inodore (Kraft and Mannschreck, 2010). La structure tridimensionnelle et l'ensemble des caractéristiques structurales d'une molécule odorante déterminent le(s) récepteur(s) olfactif(s) avec le(s)quel(s) cette molécule pourra interagir (Cowart and Rawson, 2005). Cette interaction entre une molécule odorante et un ou des récepteur(s) olfactif(s) au niveau périphérique constitue la première étape qui initie la perception d'une odeur.

2. Muqueuse et épithélium olfactifs

Située au sommet de la cavité nasale, la muqueuse olfactive est une membrane sécrétant du mucus et est également le site initial du processus olfactif (Figure 1) (Kurian et al., 2021; Salazar et al., 2019). Selon les espèces, la surface de la muqueuse olfactive varie et la taille de cette surface est directement impliquée dans la sensibilité olfactive de l'espèce. En effet, les animaux possédant une large surface épithéliale olfactive possèdent également une sensibilité olfactive élevée (espèces macrosmatiques) et à l'inverse, ceux qui possèdent une petite surface épithéliale olfactive, présentent également une faible sensibilité olfactive (espèces microsmatiques). Certains primates comme l'homme sont considérés comme des microsmates. En effet, chez l'homme, la surface de la muqueuse olfactive est d'environ 300 mm^2 soit 3 % de la surface totale de la cavité nasale. Chez les souris et les chiens, la muqueuse olfactive représente 50 à 60 % de la surface totale de la cavité nasale et peut même atteindre les 80 % chez le lapin (Chamanza and Wright, 2015; Galizia and Lledo, 2013; Leopold et al., 2000).

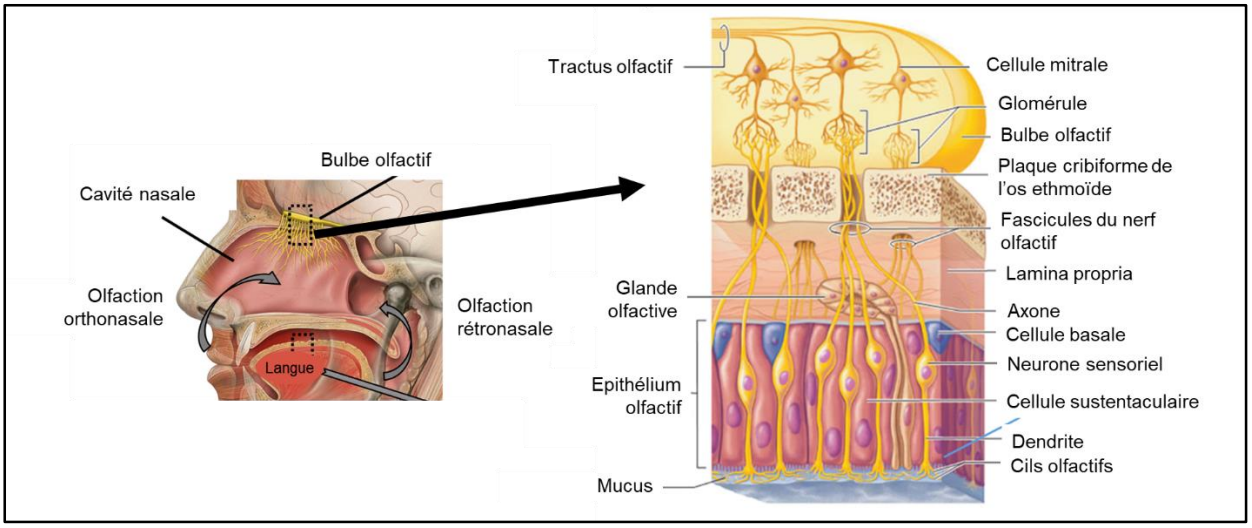


Figure 1 : Représentation du système olfactif humain (Startek et al., 2019)

La muqueuse olfactive est constituée de l'épithélium olfactif et d'une sous-muqueuse (*lamina propria*) (Kurian et al., 2021; Salazar et al., 2019). La sous-muqueuse située sous l'épithélium olfactif contient des cellules engainantes olfactives, des tissus glandulaires et caverneux, du sang et des vaisseaux lymphatiques (Kurian et al., 2021). Les cellules engainantes olfactives sont des cellules gliales qui enveloppent les axones des neurones olfactifs (Vincent et al., 2005). L'épithélium olfactif est principalement peuplé de neurones sensoriels, de cellules

sustentaculaires (cellules de soutien) et de cellules basales, dont les cellules souches olfactives (Galizia and Lledo, 2013).

Les neurones sensoriels sont des cellules éphémères d'une durée de vie allant de 30 à 60 jours. Arrivés à maturité, les neurones bipolaires sensoriels étendent une seule dendrite vers la surface épithéliale depuis leur pôle apical. A partir de cette dendrite, une multitude de cils envahissent largement la muqueuse olfactive. Les récepteurs olfactifs, qui sont situés sur les cils des neurones sensoriels interagissent avec les molécules odorantes dissoutes dans le mucus nasal. La liaison des molécules odorantes aux récepteurs olfactifs enclenche le processus d'olfaction. Ensuite, à partir de son pôle basal, le neurone sensoriel activé envoie un seul axone à travers la lame basale et la plaque cribiforme de l'os ethmoïde jusqu'au bulbe olfactif. Les axones des neurones sensoriels non myélinisés se regroupent en fascicules denses qui transmettent les signaux électriques au bulbe olfactif (Crespo et al., 2019; Feinstein and Mombaerts, 2004; Galizia and Lledo, 2013; Strotmann and Breer, 2006).

Chaque neurone sensoriel n'exprime qu'un seul type de récepteurs olfactifs (Buck and Axel, 1991), mais outre ces récepteurs olfactifs, ils expriment également tout un éventail de récepteurs différents pour les neurotransmetteurs et autres molécules de signalisation. Les récepteurs cannabinoïdes, les récepteurs cholinergiques, les récepteurs beta-adrénergiques ou encore les récepteurs purinergiques sont exprimés par ces neurones, ce qui implique une forte modulation du système olfactif dès le site d'entrée des molécules odorantes. En effet, les récepteurs cholinergiques sont liés à l'innervation autonome de l'épithélium olfactif et l'activation des récepteurs beta-adrénergiques intervient dans l'apprentissage olfactif précoce. De plus, l'activation des récepteurs purinergiques diminue la sensibilité aux odeurs, tandis que l'activation des récepteurs aux cannabinoïdes l'augmente (Czesnik et al., 2007; Galizia and Lledo, 2013; Hall, 2011; Heinbockel and Straiker, 2021; Morrison et al., 2013).

3. Le bulbe olfactif

Le bulbe olfactif est le premier relais de l'information olfactive, en recevant d'une part les signaux des axones des neurones sensoriels et d'autre part en envoyant les signaux au cortex olfactif (Heinbockel and Straiker, 2021; Mori, 1987). Il se situe dans la fosse crânienne antérieure, au-dessus de la plaque cribriforme de l'os ethmoïde et sous le lobe frontal (Huart et al., 2013). Le bulbe olfactif présente une structure laminaire et concentrique comprenant

plusieurs couches constituées de neurones, d'interneurones et de fibres afférentes (Figure 2) (Lodovichi, 2021; Mori, 1987; Scott et al., 1993; Zeppilli et al., 2021). On distingue ainsi six couches différentes nommées, de la surface vers l'intérieur :

- la couche du nerf olfactif (CNO), qui est constituée par les axones des neurones sensoriels entrants (Huart et al., 2013) ;
- la couche glomérulaire (CG), composée de glomérules et d'une région périglomérulaire. Chaque glomérule réunit les axones des neurones sensoriels exprimant le même type de récepteurs olfactifs (Huart et al., 2013). A partir des glomérules, les axones de chaque neurone sensoriel s'arborisent pour former une quinzaine de synapses avec les dendrites des cellules mitrales (mitral cells), des cellules périglomérulaires et des cellules à panache (tufted cells) (Galizia and Lledo, 2013; Huart et al., 2013) ;
- la couche plexiforme externe (CPE), qui contient le corps cellulaire des cellules à panache ainsi que les dendrites des cellules mitrales et des cellules à panache qui forment des synapses avec les interneurones (cellules juxtaglomérulaires et granulaires) ;
- la couche des cellules mitrales (CCM), constituée par le soma (corps cellulaire) des cellules mitrales (neurones olfactifs) ;
- la couche plexiforme interne (CPI), qui est composée principalement de nombreuses fibres ;
- la couche des cellules granulaires (CCG), formée par le soma des cellules granulaires qui sont les plus nombreuses dans le bulbe olfactif (Huart et al., 2013; Mori, 1987).

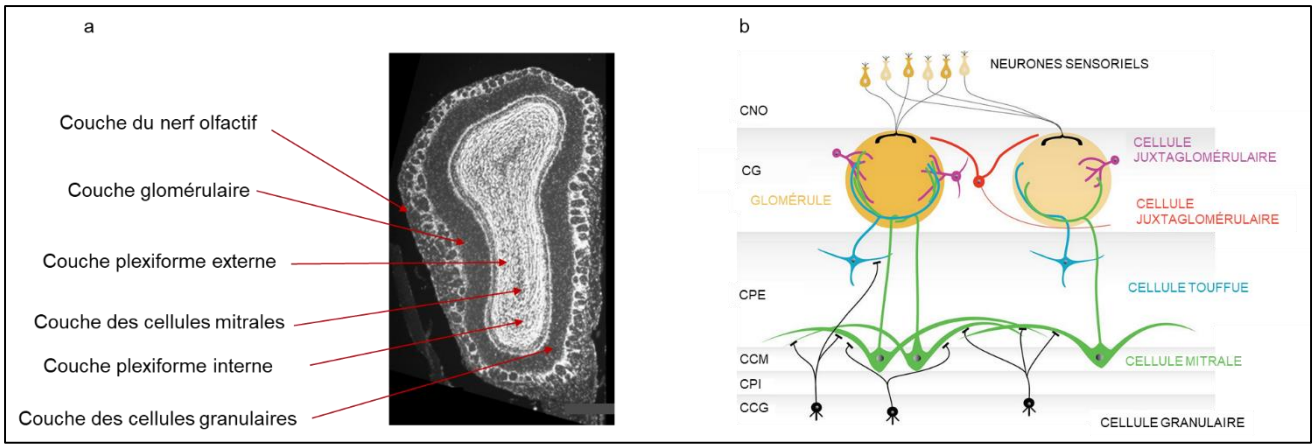


Figure 2 : Structure du bulbe olfactif. (a) Coupe coronale d'un bulbe olfactif d'une souris adulte (Hasegawa-Ishii et al., 2020). (b) Représentation de l'organisation structurelle du bulbe olfactif chez les mammifères (Nagayama et al., 2014).

Les glomérules constituent donc une structure majeure du bulbe olfactif. En effet, les projections axonales des neurones sensoriels activés vers des ensembles spécifiques de glomérules activés à leur tour, induit la formation d'un motif, créant ainsi une carte olfactive représentative de la molécule odorante (Galizia and Lledo, 2013; Johnson and Leon, 2000; Khan et al., 2010). Ces informations olfactives sont ensuite relayées au système nerveux central via le tractus olfactif formé par l'union des axones des cellules mitrales et des cellules à panache, qui les transmet au cortex olfactif (Huart et al., 2013).

Parallèlement à cette organisation structurelle laminaire, le bulbe olfactif présente également une organisation fonctionnelle compartimentée en trois domaines D_I, D_{II} (partie dorsale du bulbe olfactif) et V (partie ventrale du bulbe olfactif) (Figure 3). Ainsi, les glomérules de chaque domaine interviennent dans des réponses comportementales spécifiques de l'individu. En effet, chez la souris, l'activation des glomérules du domaine D_I entraine des comportements aversifs face à des odeurs putrides de nourriture. Les glomérules du domaine D_{II} sont impliqués dans des comportements responsifs à la peur et dans des réponses agressives à l'odeur d'autres souris mâles et ceux du domaine V sont impliqués des comportements attractifs à l'odeur de la nourriture (Mori and Sakano, 2011).

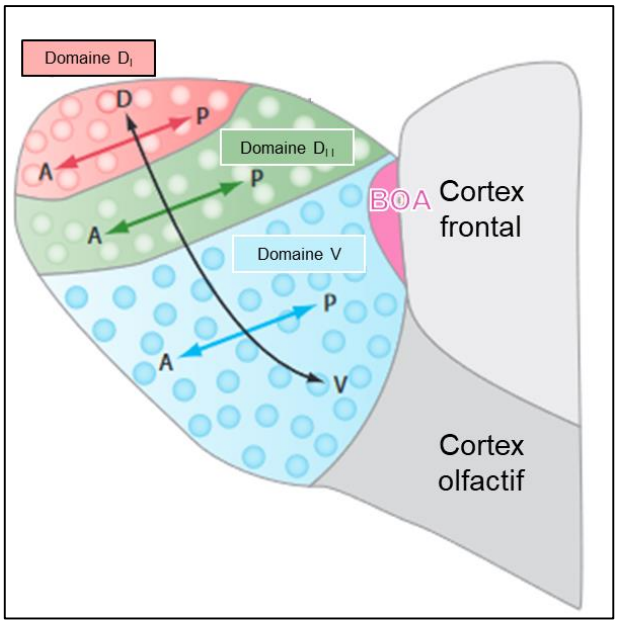


Figure 3 : Vue latérale de l'organisation des domaines du bulbe olfactif principal des rongeurs. La flèche noire indique l'axe dorso-ventral (D-V). Les flèches rouge, verte et bleue indiquent respectivement les axes antéro-postérieurs (A-P) des domaines D_I, D_{II} et V. BOA : Bulbe olfactif accessoire. (Mori and Sakano, 2011).

4. Les structures cérébrales du système olfactif

Les axones des cellules mitrales et des cellules à panache s'étendent jusqu'à un ensemble de régions corticales appelé cortex olfactif. Cette structure est constituée de sept zones distinctes : le noyau olfactif antérieur, le tubercule olfactif, les cortex piriformes antérieur et postérieur, le cortex périamygdalien, le noyau cortical antérieur de l'amygdale, et le cortex entorhinal (Cleland and Linster, 2019; Price, 2009). Les cellules à panache se projettent uniquement vers le noyau olfactif antérieur, le cortex piriforme antérieur et le tubercule olfactif alors que les cellules mitrales se projettent vers chaque zone du cortex olfactif (Cleland and Linster, 2019). Les différentes régions du cortex olfactif sont connectées entre elles via un système extensif de connections associatives permettant ensuite la transmission des informations olfactives à plusieurs autres régions du cerveau telles que le cortex orbito-frontal, l'amygdale, l'hippocampe, l'hypothalamus et le thalamus (Cleland and Linster, 2019; Price, 2009). On distingue parmi ces aires cérébrales olfactives, les structures néocorticales (cortex orbito-frontal) qui permettent une perception consciente des odeurs et les structures limbiques (amygdale, hippocampe, hypothalamus, thalamus) impliquées dans la mémoire et la composante affective d'une odeur agréable ou désagréable (Soudry et al., 2011).

B. Les récepteurs olfactifs

Cinq familles de récepteurs olfactifs ont été définies selon leur structure et leur distribution topologique : les récepteurs olfactifs (RO) dits canoniques, les récepteurs voméronasaux (RV), les récepteurs associés aux traces d'amines (TAAR), les récepteurs de peptides formiques (FPR) et les guanylate cyclases membranaires GC-D (Fleischer et al., 2009). Les quatre premiers types de récepteurs olfactifs font partie de la famille des récepteurs couplés aux protéines G (RCPG, « G protein coupled receptors, GPCR ») caractérisés par leurs sept domaines transmembranaires hydrophobes (Gaillard et al., 2004), tandis que la guanylate cyclase n'a qu'une seule hélice transmembranaire.

1. Les récepteurs olfactifs canoniques

a. Structure des récepteurs olfactifs

Les RO appartiennent à une grande famille multigénique découverte en 1991 par Linda Buck et Richard Axel (Buck and Axel, 1991). Appartenant aux RCPG de classe A, les RO représentent environ la moitié des RCPG de cette classe (Fleischer et al., 2009; Park et al., 2013). Environ un millier de gènes sont présents dans le génome des mammifères (Buck and Axel, 1991). Parmi tous ces gènes, certains sont des pseudogènes qui ne conduisent pas à l'expression de RO fonctionnels et la proportion de ces pseudogènes varie selon les espèces. En effet, chez les grands singes, environ la moitié des gènes RO sont des pseudogènes (Rouquier et al., 2000). Cette proportion augmente chez l'homme, puisque seulement près de 400 RO fonctionnels sont exprimés (Genva et al., 2019a). Chez d'autres espèces, les gènes RO sont mieux conservés comme chez la souris qui expriment près de mille RO (Glezer and Malnic, 2019).

Les protéines RO ont une longueur moyenne d'environ 320 résidus d'acides aminés qui varie principalement au niveau des parties N-terminale et C-terminale. Les RO se distinguent des autres RCPG par plusieurs motifs d'acides aminés conservés. On retrouve notamment un motif LHTPMY dans la première boucle intracellulaire, le motif MAYDRYVAIC le plus caractéristique à l'extrémité du domaine transmembranaire III, un motif SY très court à l'extrémité du domaine transmembranaire V, un segment FSTCSSH au début du domaine transmembranaire VI et PMLNPF dans le domaine transmembranaire VII. Ces séquences utilisées pour identifier

les gènes RO de nombreux génomes, bien qu'elles diffèrent légèrement entre les espèces (Fleischer et al., 2009).

Chez les mammifères, les RO peuvent être classés en deux groupes différents selon des analyses phylogénétiques : classe I et classe II (Fleischer et al., 2009). Cette classification repose sur la découverte que la grenouille (*Xenopus laevis*) possède deux groupes différents de RO : un premier similaire aux RO des poissons (classe I) et un second similaire aux RO des mammifères (classe II) (Freitag et al., 1995). Structurellement ces deux classes de récepteurs diffèrent principalement par la séquence de la deuxième boucle extracellulaire qui pourrait contribuer à la spécificité de leurs ligands. Les RO des mammifères appartiennent majoritairement à la classe II, mais les RO de classe I peuvent tout de même représenter une part importante des RO selon les espèces. Par exemple, plus de 100 RO de classe I sont présents chez l'homme et la souris, suggérant que certains RO anciens ont été maintenus et pourraient tout à fait jouer un rôle particulier chez ces espèces (Fleischer et al., 2009).

b. Liaison aux molécules odorantes et activation

Malgré la présence de motifs conservés dans les domaines transmembranaires des RO, ces domaines restent les régions les plus variables des RO (Buck and Axel, 1991). Cette variabilité des séquences des domaines transmembranaires est impliquée dans la capacité des récepteurs olfactifs à reconnaitre une large gamme de récepteurs olfactifs. En effet, plusieurs études ont permis de montrer que la poche de liaison aux molécules odorantes se situe au niveau des domaines transmembranaires III, V et VI (Abaffy et al., 2007; Katada, 2005). La fixation de la molécule odorante au récepteur olfactif induit un changement de sa conformation de la forme inactive à la forme active permettant l'interaction avec la protéine G (Bushdid et al., 2019; Fleischer et al., 2009; Ikegami et al., 2020; Le Bon et al., 2008). La protéine G (protéine intramembranaire trimérique constituée des sous-unités α, β et γ (Meierhenrich et al., 2005a) est ainsi dissociée en sous-unité α et en dimère βγ et un guanosine triphosphate est hydrolysé en guanosine diphosphate. L'hydrolyse, qui mène à la réassociation du trimère, arrête le processus d'activation (Le Bon et al., 2008). Diverses sous-unités Gα, telles que $G\alpha_{olf}$, $G\alpha_{s}$ et $G\alpha_{15/16}$ peuvent être couplés aux RO (Fleischer et al., 2009). La sous-unité $G\alpha_{olf}$ intervient fréquemment dans le processus de l'olfaction. Elle permet d'activer l'adénylate cyclase, qui convertit l'adénosine triphosphate intracellulaire en

adénosine monophosphate cyclique. A son tour, l'adénosine monophosphate cyclique induit l'ouverture d'un canal ionique situé dans la membrane plasmique du neurone sensoriel, qui permet l'entrée de cations (Na^+, Ca^{2+}...). L'entrée des cations déclenche l'ouverture des canaux chlorure, conduisant à la dépolarisation du neurone qui génère un potentiel d'action menant à un influx nerveux (Figure 4) (Hasin-Brumshtein et al., 2009; Le Bon et al., 2008).

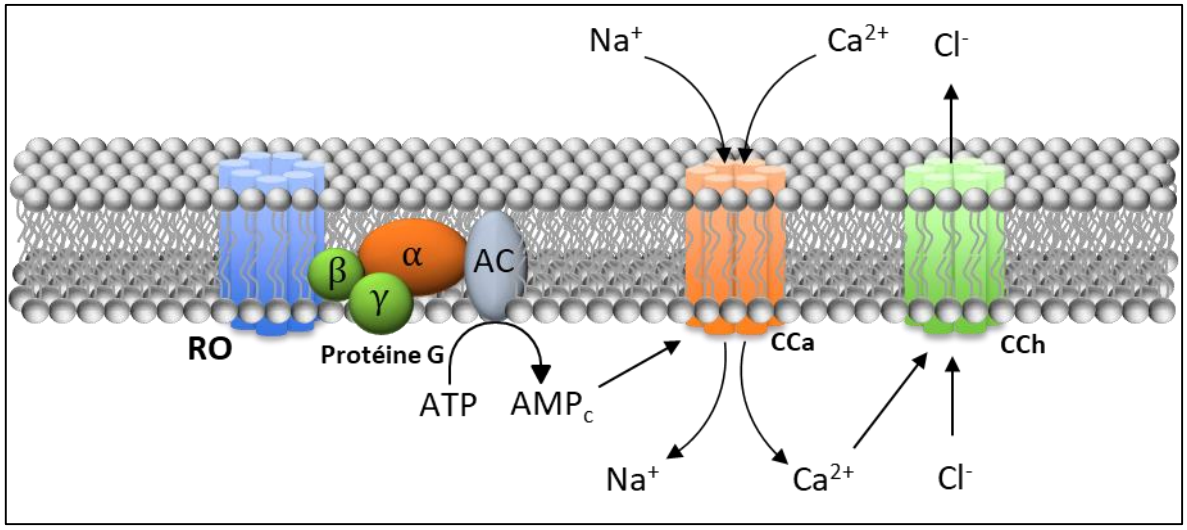


Figure 4 : Représentation schématique de la voie de transduction du signal olfactif initiée par la liaison de la molécule odorante au récepteur olfactif (RO). AC : Adénylate Cyclase. ATP : Adénosine triphosphate. AMP_c : Adénosine monophosphate cyclique. CCa : Canal cationique. CCh : Canal chlorure. (D'après Hasin-Brumshtein et al., 2009).

c. *Le code combinatoire*

Quelques centaines de récepteurs olfactifs permettent de discriminer des centaines de milliers de molécules odorantes (Bushdid et al., 2014; Mori, 2003). Cette capacité repose sur le code combinatoire (Figure 5) (Malnic et al., 1999a) selon lequel chaque récepteur olfactif peut être activé par plusieurs molécules odorantes, tandis qu'une même molécule odorante peut activer différents récepteurs olfactifs. L'odeur d'une molécule odorante est déterminée par la combinaison des récepteurs olfactifs activés par cette molécule odorante (Touhara, 2002a), et chaque molécule odorante est codée par une combinaison unique de récepteurs olfactifs. De plus, la liaison d'une molécule odorante à différents récepteurs olfactifs se fait par différentes régions et caractéristiques structurelles de la molécule odorante (Malnic et al., 1999a). Parmi les récepteurs olfactifs, on distingue deux grands types fonctionnels : les récepteurs olfactifs dits « broadly tuned » qui répondent à un large nombre de molécules odorantes structurellement différentes et les récepteurs olfactifs dits « narrowly tuned » qui

ne répondent qu'à un petit nombre de molécules structurellement proches (Saito et al., 2009). Néanmoins, moins de 20% des récepteurs olfactifs humains (soit environ 70) ont des ligands connus à ce jour (Sharma et al., 2022).

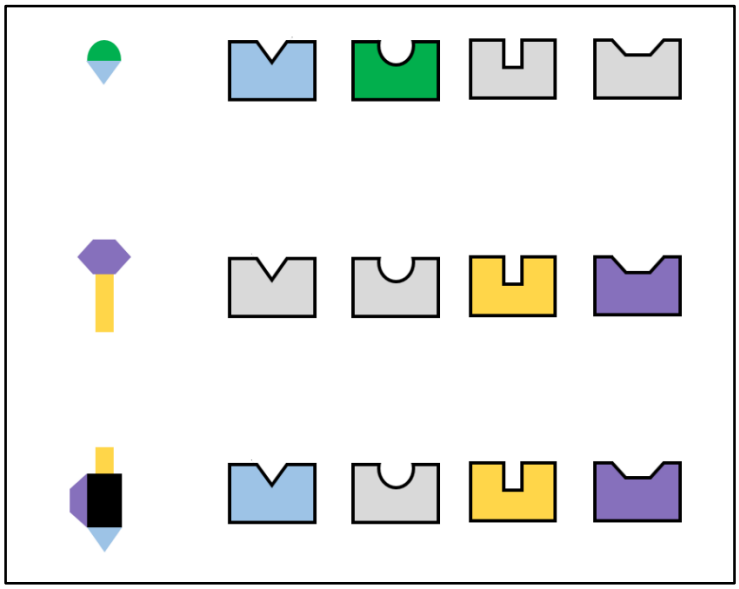


Figure 5 : Code combinatoire des molécules odorantes (D'après Malnic et al., 1999)

2. Autres récepteurs

a. Les récepteurs voméronasaux

Les récepteurs voméronasaux (RV) sont spécifiques du système voméronasal et activés par des phéromones et kairomones (substance, produite par un organisme, qui interagit avec un individu d'une autre espèce provoquant une réaction comportementale ou physiologique favorable au récepteur mais pas à l'émetteur (Ruther et al., 2002)). Pour autant, il n'est pas exclu que les RV détectent des stimuli non phéromonaux (par exemple des molécules odorantes) ou que certaines phéromones déclenchent des réponses dans d'autres systèmes, comme le système olfactif principal. Les RV sont exprimés dans l'organe voméronasal lié aux structures du système limbique. Parmi les RV, on distingue deux superfamilles : les RV de type 1 (RV1) et les RV de type 2 (RV2). Les RV1 sont exprimés dans la couche apical de l'épithélium voméronasal pour les RV1 tandis que les RV2 sont exprimés dans la couche basale de l'épithélium voméronasal. Les deux superfamilles se distinguent également par une structure

différente des gènes. Au niveau fonctionnel, les RV1 interviennent dans la détection de petites molécules volatiles impliquées dans la discrimination sexuelle et les comportements sexuels. Les RV2 sont eux associés à la détection de peptides hydrosolubles impliqués dans le contrôle de l'agression entre mâles (Luo and Katz, 2004; Silva and Antunes, 2017). Les primates, y compris les humains, semblent avoir subi la dégénérescence de la plupart du répertoire fonctionnel RV1. En effet, chez l'homme, seuls cinq gènes RV1 potentiellement fonctionnels ont été décrits (Quignon et al., 2012; Spehr and Munger, 2009). Bien qu'un récepteur humain, V1RL1, soit exprimé dans l'épithélium olfactif (Rodriguez et al., 2000), le rôle de ces récepteurs reste très incertain. Chez l'homme (et d'autres mammifères tels que le chimpanzé ou la vache), le répertoire des RV2 a complètement dégénéré (Quignon et al., 2012; Spehr and Munger, 2009).

b. Les récepteurs associés aux traces d'amines

Les récepteurs associés aux traces d'amines (Trace Amine-associated Receptors, TAAR) constituent la seconde famille de récepteurs chimio-sensoriels dans le système olfactif principal. Initialement, les TAAR sont nommés « récepteurs associés aux traces d'amines » ou « récepteurs aux traces d'amines » car plusieurs composés aminés endogènes (appelés traces d'amines) sont reconnus par certains membres du répertoire TAAR. Cependant certains récepteurs ne répondent pas à ces composés dits « traces d'amines ». Toutefois, leur localisation génomique en cluster et leur motif caractéristique dans le domaine TM7 impliquent que les TAAR, activés par des composés aminés, forment réellement une famille bien définie de récepteurs olfactifs (Dewan, 2021; Dieris et al., 2021; Fleischer et al., 2009).
Comparé aux RO canoniques, le nombre de gènes distincts de TAAR est assez faible chez les mammifères : 15 chez la souris, 17 chez le rat, 2 chez le chien et 6 chez l'homme. Ces récepteurs sont exprimés dans un sous-ensemble de neurones sensoriels dispersés dans l'épithélium olfactif et sont également exprimés dans une population de neurones du ganglion de Grueneberg (cluster de neurones situé dans le vestibule nasal). Les neurones sensoriels exprimant les TAAR utilisent une cascade de transduction similaire à celle employée par les neurones sensoriels exprimant les RO canoniques, chaque neurone n'exprimant qu'un seul type de TAAR. Mais, l'expression de ces deux types de récepteurs olfactifs semble s'exclure

mutuellement puisqu'aucun neurone sensoriel exprimant à la fois un RO et un TAAR n'a encore été identifié (Fleischer et al., 2009; Quignon et al., 2012; Spehr and Munger, 2009).

c. Les récepteurs de peptides formiques

Les récepteurs de peptides formiques (FPR) ont été décrits pour la première fois chez l'homme en 1976 (Showell et al., 1976), comme un site de liaison de haute affinité sur la surface des neutrophiles pour le peptide N-formyl formyl-méthionine-leucyl-phénylalanine (Le et al., 2002). Néanmoins, ils sont également exprimés dans l'organe voméronasal. De manière similaire aux gènes des récepteurs voméronasaux (RV), les gènes FPR y sont exprimés sélectivement dans un petit sous-ensemble de neurones qui ne co-expriment pas les récepteurs voméronasaux. Pour autant, les gènes FPR ne partagent aucune similitude de séquence avec les gènes RV. Néanmoins, les FPR semblent reconnaître certains ligands en commun avec le RV1 ou le RV2 mais en produisant des signaux spécifiquement transmis dans différentes régions du cerveau. Parallèlement, les FPR sont activés par des peptides formylés dont certains sont liés aux maladies et à l'inflammation (sécrétés par des bactéries Gram négatives et des peptides dérivés du VIH). Par conséquent, ces récepteurs pourraient permettre la détection d'aliments contaminés ou de congénères touchés par une maladie à travers leurs sécrétions corporelles (Fleischer et al., 2009; Quignon et al., 2012).

d. Les guanylate cyclase membranaires GC-D

Il existe différents types de guanylate cyclases membranaires et le sous-type GC-D est exprimé dans un sous-ensemble de neurones sensoriels de l'épithélium olfactif, désignés comme neurones GC-D. Ces neurones projettent leurs axones vers le bulbe olfactif où ils convergent vers des glomérules distincts encerclant le bulbe olfactif caudal et qui sont ainsi appelés « glomérules en collier ». Les GC-D semblent impliqués dans la détection du dioxyde de carbone (CO_2), puisque les neurones GC-D, à la différence des autres neurones sensoriels, répondent à de faibles concentrations de CO_2. Dans les neurones GC-D via l'anhydrase carbonique, le CO_2 serait converti en bicarbonate qui activerait ensuite le récepteur GC-D. Chez l'homme (comme chez plusieurs autres espèces de primates), le gène GC-D est un

pseudogène et le CO_2 est inodore pour l'homme contrairement aux rongeurs qui expriment ce gène (Fleischer et al., 2009).

3. Expression ectopique des récepteurs olfactifs

Originellement découvert pour leur rôle dans l'olfaction, les RO sont exprimés dans des tissus extérieurs à l'épithélium olfactif et peuvent donc avoir un rôle dans d'autres processus biologiques que l'olfaction (Raka et al., 2022; Tong et al., 2021). En effet, plusieurs RO (humains et murins) identifiés dans le foie et le pancréas sont impliqués dans les processus de régulation du glucose et du métabolisme des lipides (S. Zhang et al., 2021). Une étude a montré la présence d'un récepteur olfactif humain dans les artères aorte et coronaires ainsi que son rôle dans l'angiogenèse (Kim et al., 2015a). D'autres études ont montré qu'un récepteur humain présent dans les testicules intervenait dans la chimiotaxie des spermatozoïdes (Neuhaus et al., 2006; Spehr et al., 2003a).

Les RO présentent donc des fonctions multiples au sein des organismes vivants, dont un potentiel effet thérapeutique. Différents RO présents sont surexprimés dans des cellules cancéreuses (le cancer de la prostate, le cancer du sein ou le mélanome) comparées aux cellules saines et pourraient ainsi constituer des cibles thérapeutiques ou des biomarqueurs intéressants (Ranzani et al., 2017; Weber et al., 2018; Xia et al., 2001). Plusieurs études ont montré que l'activation de RO présents dans les cellules cancéreuses (cancer du poumon, cancer hépatique, cancer colorectal, leucémie myéloïde) par des ligands spécifiques diminuait leur prolifération (Kalbe et al., 2017; Manteniotis et al., 2016; Maßberg et al., 2015; Weber et al., 2017).

II. La perception odorante

La perception des odeurs résulte de la transduction, initiée par les molécules odorantes, de signaux au sein de structures hiérarchisées et interconnectées du cerveau. Une odeur est définie par trois caractéristiques principales : sa qualité, son intensité et sa valeur hédonique. La majorité des odeurs sont perçues à partir de mélanges de molécules odorantes et les caractéristiques physico-chimiques de chaque composant vont donc influer sur celles de l'odeur du mélange.

A. Les attributs de la perception odorante

1. La qualité

La qualité d'une odeur représente son identité, par exemple une odeur de rose, de fraise ou de vanille. C'est donc la qualité des odeurs qui permet de les discriminer (Holley, 2006). La qualité d'une odeur est principalement liée à la structure chimique de la molécule odorante à partir de laquelle elle est perçue (Chastrette, 1997). Les odeurs naturelles étant perçues généralement en mélange, leur perception peut découler d'un traitement analytique ou synthétique de l'information olfactive. Ainsi la perception d'un mélange peut être hétérogène (traitement analytique) ou homogène (traitement synthétique) (Berglund et al., 1976).
La perception est dite hétérogène lorsque l'odeur de chaque molécule odorante est perçue distinctement à partir du mélange. La perception homogène est définie par la perception d'une unique odeur à partir du mélange de molécules odorantes (Berglund et al., 1976). Dans le cas de la perception homogène, on distingue l'accord aromatique, lorsque l'odeur perçue est une nouvelle odeur, différente de celles des molécules odorantes du mélange (Thomas-Danguin et al., 2014), et le masquage, lorsque l'odeur d'un seul des composants du mélange est perçue (Cain and Drexler, 1974).

2. L'intensité

Une odeur peut être perçue lorsque le seuil de perception d'une molécule odorante est atteint, le seuil de perception étant la concentration minimale de la molécule odorante nécessaire pour être différenciée de manière fiable d'un échantillon vide. L'intensité olfactive caractérise la force de la perception de cette odeur (Block, 2018). L'intensité d'une odeur varie avec la concentration de la molécule odorante. Toutefois, celle-ci peut interférer avec la perception de la qualité de l'odeur. En effet, la qualité d'une odeur perçue à partir d'une molécule odorante peut être modifiée selon son intensité (Holley, 2006).
Dans le cas de mélange de molécules odorantes, l'intensité du mélange diffère généralement de la somme des intensités des composants du mélange. Comme nous l'avons vu précédemment, on distingue deux types de perception, hétérogène et homogène.

Dans le cas d'une perception hétérogène, si l'intensité de l'un des composants en mélange est plus élevée que l'intensité du composant présenté seul, on parle de synergie. Au contraire, lorsque l'intensité du composant en mélange est plus faible que l'intensité du composant présenté seul, il s'agit alors d'antagonisme ou de masquage partiel. Enfin, si l'intensité d'un composant reste identique en mélange et hors du mélange, le phénomène est appelé indépendance (Berglund et al., 1976; Thomas-Danguin et al., 2014).

En ce qui concerne les perceptions homogènes, les différentes réponses sensorielles sont nommées hyper-addition (ou synergie) si l'intensité du mélange est supérieure à la somme des intensités des odeurs de chaque constituant du mélange, addition complète si elle y est égale, et hypo-addition si elle y est inférieure (Figure 6). Il existe trois types d'hypo-addition : l'addition partielle, le compromis et la soustraction. L'addition partielle se produit lorsque l'intensité du mélange est supérieure à l'intensité du composé ayant l'intensité la plus élevée. Il s'agit de compromis, lorsque l'intensité du mélange est inférieure à celle du composé ayant l'intensité la plus élevée mais supérieure à celle du composé ayant l'intensité la plus faible. Si l'intensité du mélange est inférieure à l'intensité du composé ayant l'intensité la plus faible, il s'agit de soustraction (Berglund et al., 1976; Patte and Laffort, 1979).

$$\sigma = \frac{\text{Intensité}_{\text{mélange}}}{\text{Intensité}_A + \text{Intensité}_B}$$

$$\tau_A = \frac{\text{Intensité}_A}{\text{Intensité}_A + \text{Intensité}_B}$$

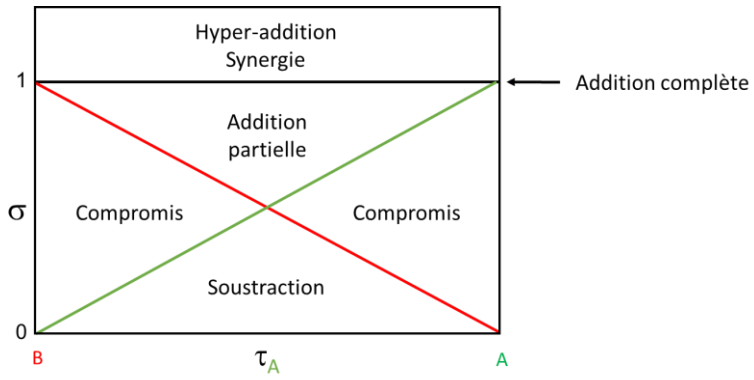


Figure 6 : Représentation de Patte et Laffort de l'intensité d'un mélange de deux composés odorants (A et B) dans le cadre d'une perception homogène. σ représente la déviation de l'addition complète. τ représente la proportion perceptuelle du composé A.

3. La valeur hédonique

Dernière caractéristique de l'odeur, la valeur hédonique définit le caractère plaisant ou déplaisant d'une odeur. La valeur hédonique d'une odeur dépend de l'apprentissage acquis lors des expériences antérieures et de la culture (Holley, 2006; Kermen et al., 2016). Cependant l'attraction ou l'évitement spontané envers une odeur est également observé pour

des substances odorantes non familières suggérant une composante innée directement attribuée à la structure de la molécule odorante et donc par extension à sa qualité (Holley, 2006; Kermen et al., 2016; Mandairon et al., 2009). De plus, l'intensité du stimulus olfactif intervient également dans l'appréciation de l'odeur. En effet, les odeurs jugées déplaisantes à faible intensité le sont davantage lorsque leur concentration et donc leur intensité augmentent. Cette relation diffère pour les odeurs agréables, puisque le caractère plaisant de ces odeurs augmente avec la concentration jusqu'à un optimum puis diminue (Holley, 2006). Au final, l'attrait ou l'aversion envers une odeur constitue un aspect important de la perception olfactive, notamment dans les comportements déclenchés par l'odeur (s'approcher ou éviter une source d'odeur) (Kermen et al., 2016).

B. Interactions dans la perception de l'odeur

La perception olfactive de mélanges de molécules odorantes est très complexe comme le montrent les différents phénomènes présentés ci-dessus. Ces effets sont dus à différentes interactions survenant à plusieurs niveaux du système olfactif. Quatre niveaux d'interactions différents ont été identifiés, de la réaction chimique jusqu'aux interactions dans le système nerveux central (Berglund et al., 1976).

1. Interactions moléculaires et réactions chimiques

Le premier niveau d'interactions implique des interactions chimiques ou physiques qui se produisent éventuellement dans le mélange (molécules instables dans leur milieu, libération à partir d'une phase condensée), et lors de la rencontre du mélange avec les voies nasales et la muqueuse olfactive (Berglund et al., 1976). Nous ne nous intéresserons ici qu'aux événements susceptibles de se produire au voisinage des récepteurs olfactifs. Dans le mucus nasal, les molécules odorantes peuvent être soumises à des conversions enzymatiques telles que la conversion des aldéhydes et des esters en acides et alcools (Robert-Hazotte et al., 2022). Par ailleurs, la solubilité des molécules odorantes est différente dans le mucus nasal et dans le mucus nasopharyngé qui possèdent une composition différente, ce qui peut entraîner une perception olfactive différente entre les olfactions rétronasale et orthonasale (Genva et al., 2019a). D'autre part, le mucus nasal ainsi que la salive sont riches en protéines de liaison

aux molécules odorantes (Odor Binding Protein ou OBP) qui se lient donc aux molécules odorantes. L'affinité des OBP pour les molécules odorantes varie selon la structure chimique de ces dernières, ajoutant à la complexité de la compréhension de la perception olfactive (Genva et al., 2019a). Il est largement accepté que les protéines de liaison aux molécules odorantes agissent comme des transporteurs et des solubilisateurs de molécules odorantes et de phéromones mais leurs rôles spécifiques et leurs modes d'action demeurent méconnus (Pelosi and Knoll, 2022).

2. Niveau périphérique du système olfactif

Le deuxième niveau d'interactions se situe à la surface des récepteurs olfactifs avec lesquels les molécules odorantes vont se lier (Berglund et al., 1976). Des interactions compétitives et non compétitives peuvent se produire au niveau des récepteurs olfactifs et ces interactions vont influer sur la réponse des neurones sensoriels au mélange.

Les interactions compétitives font intervenir deux molécules se liant au même site de liaison d'un récepteur olfactif. Il peut s'agir de deux molécules odorantes agonistes (molécules capables d'activer le récepteur) ou bien d'une molécule agoniste et d'une antagoniste (molécule capable de se lier au récepteur sans l'activer) (Thomas-Danguin et al., 2014). Par exemple, une étude a montré le récepteur humain hOR17-4 était fortement activé par le bourgeonal alors que l'undécanal est incapable de l'activer (Spehr et al., 2003a). Cependant, la stimulation du récepteur hOR17-4 avec un mélange des deux molécules n'induit aucune réponse du récepteur, ce qui suggère que l'undécanal a inhibé l'activation du récepteur par le bourgeonal.

Quant aux interactions non compétitives, elles impliquent deux molécules se fixant à deux sites de liaisons différents. Parmi ces interactions, un phénomène d'allostérie a été mis en évidence par une étude menée sur des rats (Rospars et al., 2008). Le site de liaison principal permet l'activation des récepteurs olfactifs par une molécule agoniste, alors que l'occupation du second site ne permet pas l'activation du récepteur mais peut modifier les propriétés de liaison (affinité) ou d'activation (efficacité) des agonistes au niveau du site principal (Rospars, 2013). D'autres part, des effets dose-dépendants ont également été décrits. En effet, il a été montré que la stimulation de neurones sensoriels olfactifs chez le rat par un mélange d'acétate d'isoamyle et de whisky lactone induisait des réponses différentes selon la

concentration de whisky lactone (Chaput et al., 2012). A faible concentration de whisky lactone, la réponse au mélange est plus élevée que celle à l'acétate d'isoamyle seul et à plus forte concentration de whisky lactone, la réponse au mélange est plus faible que celle à l'acétate d'isoamyle seul. Enfin, un autre exemple d'interactions non compétitives, l'antagonisme non compétitif, a également été décrit chez les insectes (Jones et al., 2012). L'effet de tous ces phénomènes montre donc leur importance dans le codage olfactif par les molécules odorantes.

3. Niveau du bulbe olfactif

Le troisième niveau d'interactions est à la périphérie du système nerveux, dans le bulbe olfactif. Les signaux issus de l'activation d'un récepteur olfactif donné et envoyés vers le bulbe olfactif peuvent interférer avec les signaux initiés par l'activation d'autres récepteurs olfactifs (Berglund et al., 1976). Ainsi, dans le bulbe olfactif, l'activité glomérulaire générée par des mélanges d'odeurs peut résulter de la simple somme de l'activité de quelques composés chimiques dominants ou résulter d'interactions inhibitrices entre différents ensembles de glomérules réactifs (Dulac, 2006). En effet, l'activité des cellules mitrales dans un glomérule donné est influencée par l'activité des cellules mitrales associées à d'autres glomérules (Davison and Katz, 2007; Economo et al., 2016), et des interactions interglomérulaires médiées par les cellules granulaires peuvent survenir. Leur activation par le glutamate libéré par les dendrites des cellules mitrales induit la libération, par les cellules granulaires, d'acide γ-aminobutyrique sur les somas et les dendrites des cellules mitrales, inhibant ainsi l'activité de ces dernières (Schoppa and Urban, 2003). Les interactions perceptives qui se produisent lors du traitement des informations issues de mélanges de molécules odorantes, résulteraient principalement de ces processus inhibiteurs. La similarité structurelle des molécules odorantes pourrait activer des modèles de chevauchement, ce qui pourrait induire une similarité perceptive tout en augmentant le potentiel d'interaction (Thomas-Danguin et al., 2014). Il a été montré qu'un mélange de molécules odorantes qui correspondent à des motifs glomérulaires similaires entraînait une inhibition latérale dans le bulbe olfactif menant à une perte d'information sur chaque molécule odorante unique (Linster and Cleland, 2004). Cette perte d'information favoriserait un motif d'activation bulbaire spécifique au mélange, ce qui induirait un code pour le mélange distinct de celui de chaque composant du mélange.

4. Niveau cérébral

Le dernier niveau d'interactions se situe dans les zones du cerveau, derniers relais de l'information olfactive. En effet, dans le cortex piriforme, les modèles d'activité neuronal lors de l'exposition à un mélange de molécules odorantes ne sont pas formés par la somme des représentations des composants du mélange et ne sont pas dominées par des réponses synergiques (Stettler and Axel, 2009). Au contraire, on observe, en réponse à un mélange de molécules odorantes, une suppression de l'activité des neurones sensibles aux odorants individuels présentés seuls (Wilson, 2003; Yoshida and Mori, 2007). Cette suppression peut en partie être attribuée aux interneurones inhibiteurs qui présentent une excitation non sélective. Elle permettrait ainsi de maintenir un nombre limité de neurones actifs indépendamment de la complexité du mélange de molécules odorantes. En effet, chaque molécule odorante étant capable d'exciter 10 % des neurones corticaux lorsqu'elle est présentée seule, une sommation sans suppression entraînerait des schémas d'activité neuronale se chevauchant massivement, qui dégraderaient ainsi le pouvoir discriminatoire (Stettler and Axel, 2009). D'autre part, l'exposition à un mélange de molécules odorantes peut également entrainer une facilitation des réponses neuronales (Yoshida and Mori, 2007).

De plus, les régions antérieures et postérieures du cortex piriforme sont très différentes dans leur organisation anatomique et possèdent des rôles distincts dans le codage des odeurs. Le codage de la structure de la molécule odorante se produit dans les régions antérieures tandis que le codage la qualité de l'odeur se produit dans les régions postérieures (Gottfried et al., 2006). Le cortex piriforme est donc une structure essentielle de la perception des mélanges de molécules odorantes puisqu'il peut contribuer à leur traitement configural c'est-à-dire le traitement des informations olfactives d'un mélange menant à de nouvelles qualités perceptuelles (Thomas-Danguin et al., 2014).

Outre le cortex piriforme, les cortex d'ordre supérieur sont également le site d'interactions d'informations olfactives. Une étude comparant le traitement cérébral des molécules de citral et de pyridine seules et de leurs mélanges a montré que les régions latérales et antérieures du cortex orbito-frontal ont répondu de manière préférentielle aux mélanges binaires. La partie antérieure du cortex orbito-frontal était activée en réponse aux mélanges et désactivée en réponse aux molécules uniques ; la partie latérale de l'OFC a répondu de façon graduelle à des différences relativement faibles dans les rapports d'intensité des deux odeurs mélangées

(Boyle et al., 2009). Ce cortex est connu pour encoder la qualité et la valeur hédonique d'une odeur (Anderson et al., 2003; Royet et al., 2003) et joue donc probablement un rôle majeur dans le traitement configural des stimuli olfactifs complexes. Une étude a également constaté que dans le cas d'un mélange de molécules odorantes contenant des composants agréables et désagréables, il y avait une représentation simultanée de la valeur hédonique positive et négative, mais dans différents cortex. En effet, l'activation du cortex orbito-frontal médian était corrélée au caractère agréable des stimuli alors que l'activation de la partie dorsale du cortex cingulaire antérieur et de la région caudale du cortex midorbito-frontal était corrélée au caractère désagréable des stimuli (Grabenhorst et al., 2007). Toutes ces interactions perceptives permettent au système olfactif de saisir efficacement les informations olfactives complexes et sont ainsi à l'origine du traitement des mélanges de molécules odorantes.

III. Modèles computationnels utilisés dans l'étude des mécanismes olfactifs

Ces dernières décennies, de nombreuses avancées ont permis une amélioration de la connaissance du système olfactif. Néanmoins, la compréhension des spécificités du processus de reconnaissance des molécules odorantes demeure encore un défi majeur. Dans ce domaine, les approches *in silico* sont essentielles afin de comprendre les mécanismes impliqués.

A. Amarrage moléculaire « Docking »

L'amarrage moléculaire, ou « docking » est une technique permettant de comprendre et de prévoir la reconnaissance moléculaire, à la fois sur le plan structurel, en trouvant les modes de liaison probables, et sur le plan thermodynamique, en prévoyant l'affinité de liaison (Figure 7). Le docking moléculaire est généralement réalisé entre une petite molécule et une macromolécule cible (Morris and Lim-Wilby, 2008). Sur la base du docking, des criblages virtuels peuvent être réalisés, permettant de sélectionner des molécules actives prédites à partir de grandes bibliothèques de structures chimiques. Les molécules sont classées en

fonction de leurs affinités pour la cible et de nouveaux ligands peuvent ainsi être identifiés (Di Pizio and Niv, 2014).

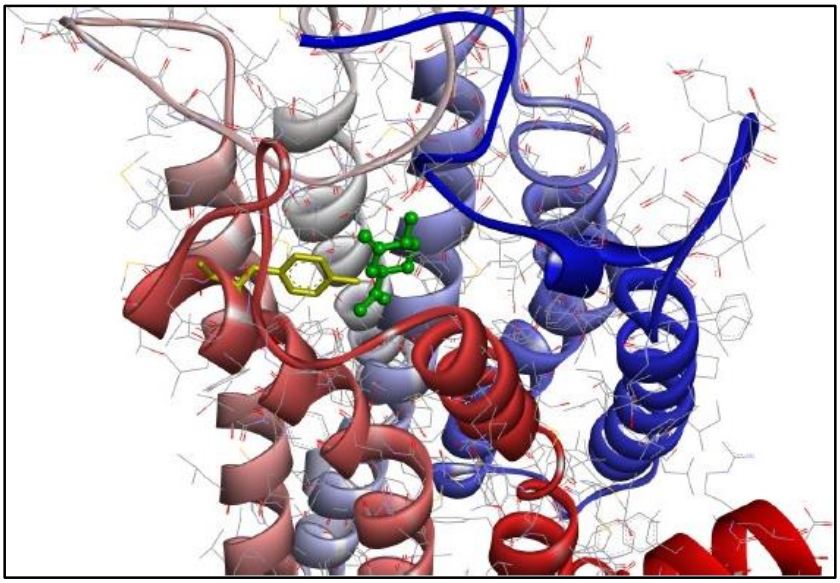


Figure 7 : Docking de la (+)-dihydrocarvone et du récepteur olfactif OR1A1. La conformation de la (+)-dihydrocarvone (en vert) située le plus près du résidu Tyr251 (en jaune) du modèle OR1A1, vue du côté N-terminal du récepteur. Les domaines transmembranaires colorés sont représentés du bleu au rouge, du N-terminal au C-terminal (Oh, 2021a).

En utilisant une technique de docking, une étude a montré une différence dans l'affinité de liaison à un récepteur olfactif entre les agonistes et les agonistes inverses, permettant ainsi de fournir une indication utile pour les tests de criblage visant à trouver de nouveaux ligands de ce récepteur (Oh, 2021b).

B. Olfactophores

Des modèles olfactophores peuvent être générés à partir des structures de molécules odorantes ayant des propriétés olfactives spécifiques. Un olfactophore est un modèle pharmacophore (*cf*. partie Pharmacophores) c'est-à-dire un ensemble de caractéristiques structurelles responsables de l'activité biologique (pharmacon) d'une molécule (Ehrlich, 1909). Dans le cas de l'olfactophore, l'activité biologique recherchée est une propriété olfactive (olfact). De la même manière que les pharmacophores, les olfactophores permettent de visualiser l'arrangement tridimensionnel des groupes moléculaires en interaction avec leurs cibles, et essentiels pour la propriété olfactive recherchée (Figure 8). En tant que tels, ils fournissent des informations sur la géométrie des poches de liaison des récepteurs et sur les types d'interaction que subissent les molécules ayant certaines propriétés olfactives, en corrélation avec l'activité d'un composé (Hauser et al., 2020; Meierhenrich et al., 2004).

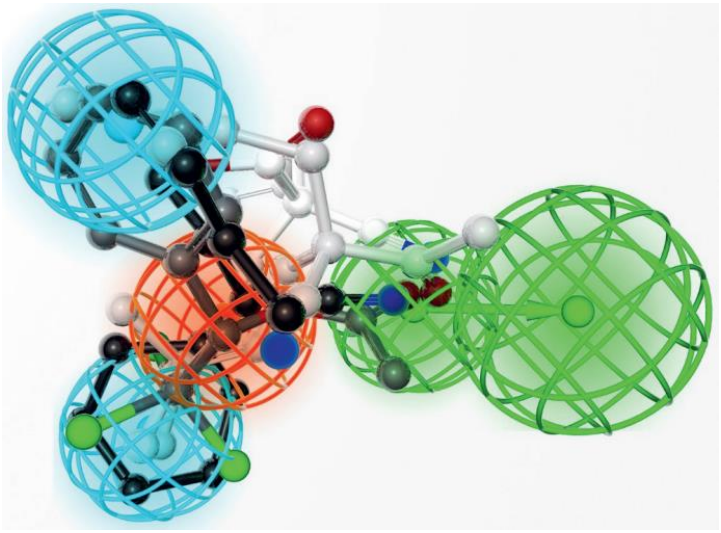


Figure 8 : Modèle d'olfactophore pour les odeurs de rose avec le Petalia (en noir), le Rosacetol (en gris foncé), le 2,2-bis(prenyl)-3-oxobutyronitrile (en gris clair), et le bis-alkynyl methyl ester (en blanc), présentant : 1 accepteur de liaison hydrogène coloré en vert, 2 groupements hydrophobes représentés en bleu, et 1 caractéristique de double liaison en orange (Hauser et al., 2020).

Par exemple, ces modèles ont permis d'identifier trois olfactophores qui ont fourni les contraintes structurelles et physicochimiques nécessaires à la conception et à la synthèse d'analogues du bois de santal (Delasalle et al., 2014).

C. Réseaux neuronaux (ANN) et réseaux neuronaux profonds

De manière analogue à l'apprentissage automatique (*cf*. partie apprentissage automatique), un réseau neuronal, profond ou non, apprend à effectuer des tâches particulières via une étape d'apprentissage, au cours de laquelle la force des connexions entre les unités est optimisée. Il est ensuite utilisé pour effectuer la même tâche sur de nouvelles entrées (Cichy and Kaiser, 2019).

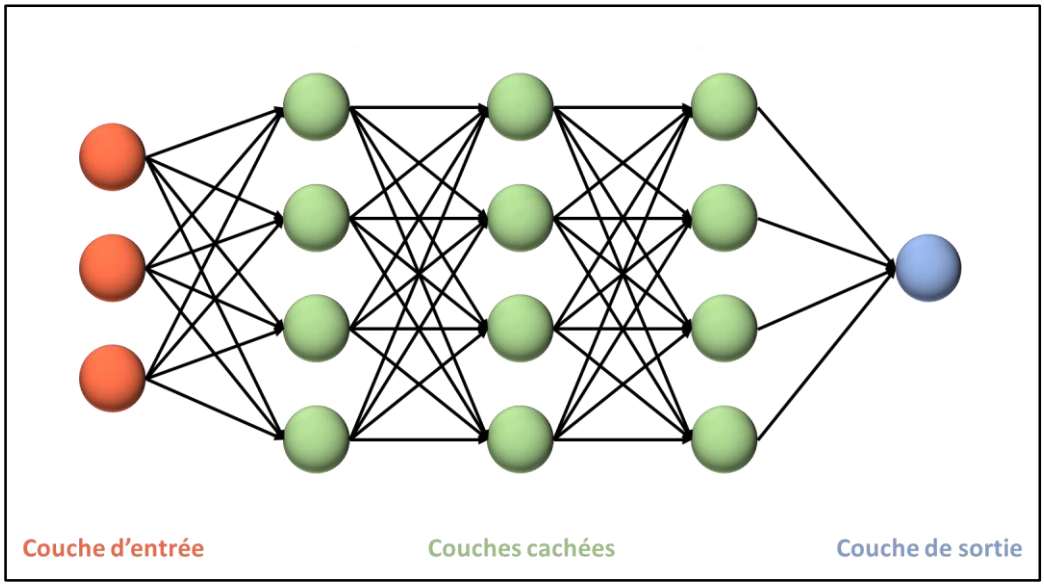


Figure 9 : Modèle général d'un réseau neuronal profond avec une couche d'entrée, plusieurs couches cachées et une couche de sortie

Les réseaux neuronaux profonds (ou Deep Learning) sont des algorithmes d'apprentissage automatique qui impliquent l'apprentissage de modèles cachés dans les données via leur exploration ainsi que l'utilisation ultérieure de ces modèles afin de classer ou prédire un événement lié à un problème que l'on se pose (Alloghani et al., 2020). L'utilisation des réseaux neuronaux profonds s'est largement développée ces dernières années (Jones, 2019). Les réseaux neuronaux profonds sont des modèles de calcul composés de nombreuses unités de traitement simples appelées neurones (inspirées du fonctionnement des neurones humains) disposées en couches interconnectées (Figure 9). Ils sont composés d'une couche d'entrée, d'une couche de sortie et de plusieurs couches cachées (Cichy and Kaiser, 2019). Les modèles composés de seulement une ou deux couches cachées, sont dits simplement réseaux neuronaux ou bien réseaux neuronaux artificiels (Artificial Neural Network, ANN) (Wang, 2003).

Plusieurs études ont utilisé les réseaux neuronaux (ANN) afin de prédire l'activité glomérulaire causée par une molécule odorante à partir de sa structure moléculaire (Soh et al., 2012) ou prédire l'odeur musquée de molécules odorantes à partir de leurs structures moléculaires (Chastrette and de Saint Laumer, 1991).

Plus récemment, les réseaux neuronaux profonds ont été utilisés dans le but de développer une approche permettant de prédire l'odeur d'une molécule à partir de son SMILES (Simplified Molecular Input Line Entry Specification : chaîne de caractères représentant la structure

moléculaire d'un composé) (Sharma et al., 2021). Une autre équipe (Wen et al., 2021) a créé un modèle basé sur les réseaux neuronaux profonds pour identifier l'odeur de molécules à partir des réponses issues d'un nez électronique, qui est un outil constitué d'un ensemble de capteurs de gaz qui détectent les gaz en mesurant la variation de la conductivité électrique.

D. Réseau biologique

Un réseau (*cf*. partie Le modèle réseau) est un modèle permettant de décrire les connexions entre les entités d'un système entier. Les entités sont représentées par des nœuds et leurs connexions par des liens (Oh and Monge, 2016). En utilisant ce modèle, une équipe a réussi à construire un réseau olfactif humain en utilisant les données produites par imagerie par résonnance magnétique fonctionnelle sur plusieurs centaines de sujets (Figure *10*). Ce réseau expose les relations entre 28 régions du cerveau humain (cortex piriforme, amygdale, thalamus) ayant un rôle dans l'olfaction, présentant ainsi les interactions les plus susceptibles d'exister entre ces régions. Trois sous-réseaux ont pu être identifiés au sein du réseau olfactif : un sous-réseau sensoriel, un sous-réseau limbique et un sous-réseau frontal. Ce réseau olfactif constitue un donc modèle représentatif de l'anatomie cérébrale fonctionnelle du système olfactif humain, et montre donc la spécialisation fonctionnelle et la ségrégation spatiale du système olfactif (Arnold et al., 2020).

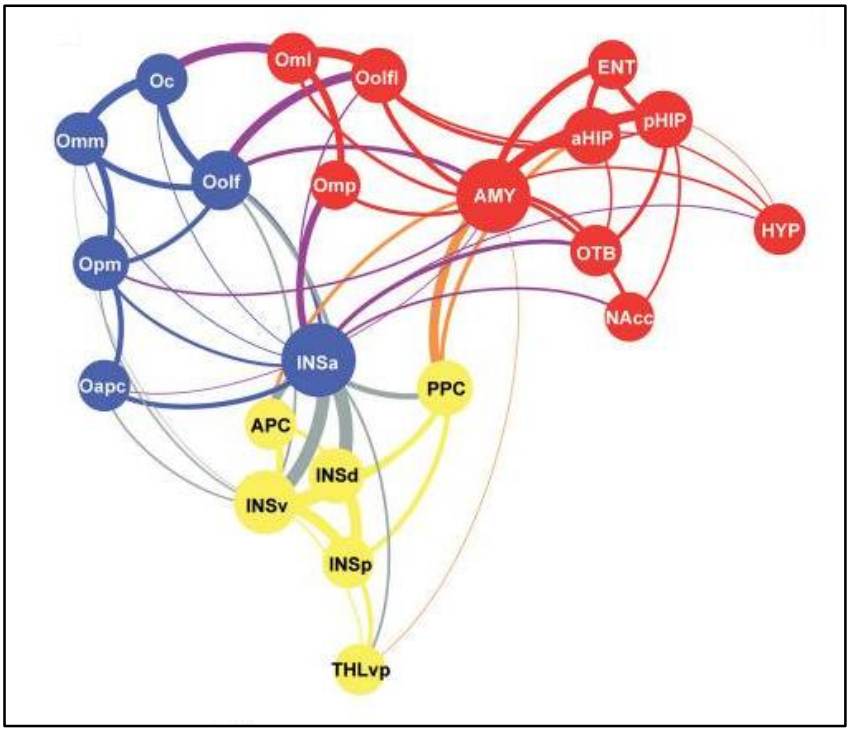


Figure 10 : Topologie du réseau olfactif. Les trois sous-réseaux sont indiqués par les trois couleurs des cercles. L'épaisseur des lignes indique la force des connexions (coefficients de corrélation moyens) et la taille des nœuds reflète la densité des connexions (nombre de connexions). AMY : Amygdale. OTB : Tubercule olfactif. HYP : Hypothalamus. ENT : Cortex entorhinal. NAcc : Noyau accumbens. pHIP : Hippocampe postérieur. aHIP : Hippocampe antérieur. PPC : Cortex piriforme postérieur. APC : Cortex piriforme antérieur. THLvp : Thalamus ventro-postérieur. INSd : Insula dorsale. INSv : Insula ventrale. INSp : Insula postérieure. INSa : Insula antérieure. Omm : Cortex orbito-frontal moyen médian. Oc : Cortex orbito-frontal central. Oolf : Cortex orbito-frontal olfactif. Opm : Cortex orbito-frontal médian postérieur. Oapc : Cortex orbito-frontal antérieur. Oml : Cortex orbito-frontal latéral médian. Oolfl : Cortex orbito-frontal olfactif latéral. Omp : Cortex orbito-frontal postérieur moyen. (Arnold et al., 2020).

IV. Objectif de la thèse

L'ensemble des informations présentées ci-dessus montre l'avancée des recherches dans le domaine de l'olfaction. Néanmoins, le rôle de la structure moléculaire d'un composé odorant sur sa qualité odorante (Genva et al., 2019a; Saini and Ramanathan, 2022; Snitz et al., 2019), les mécanismes impliqués dans la perception homogène des mélanges, accords aromatiques ou masquages (Berglund et al., 1976; Coureaud et al., 2022; Ishii et al., 2008), ou les effets biologiques des molécules odorantes au-delà du système olfactif (Di Pizio et al., 2019; Foster et al., 2014) ne sont pas parfaitement élucidés.

L'objectif de cette thèse est donc de mieux comprendre ces mécanismes à travers trois modèles différents : (1) un premier associant une technique de réduction de dimension et une méthode de classification afin d'établir un lien entre la structure d'une molécule odorante et son odeur, (2) une approche basée sur les pharmacophores pour déterminer dans quelle

mesure les mécanismes de perception homogène pourraient impliquer des cibles communes au niveau périphérique du système olfactif, et (3) un dernier modèle basée sur le réseaux dans le but d'identifier des processus biologiques dans lesquels molécules odorantes pourraient intervenir, et prédire de nouvelles cibles potentielles.

Partie 2 : Matériel et Méthodes

I. Relations structure-activité – Modélisation QSAR, 3D-QSAR et pharmacophore

Les liaisons spécifiques entre protéines et molécules (comme par exemple entre une enzyme et un substrat) sont, à divers niveaux, à l'origine de tout processus biologique (Cohen-Tannoudji, 2006). La reconnaissance spécifique nécessaire à ces liaisons serait définie par le modèle « lock and key » ou « serrure et clé », proposé en 1894 par Emil Fischer (Fischer, 1894). Ce concept repose sur le fait que les sites de reconnaissance des enzymes sont structurellement spécifiques. Ainsi un substrat ne peut se lier au site récepteur que s'il possède des fragments structurels correspondant parfaitement à ceux du site récepteur (Trinajstić et al., 1989). Ce concept est donc applicable à toute interaction ligand-récepteur. L'étude de ces interactions est donc une étape clé dans la compréhension des processus biologiques. Dans ce but, les approches « Quantitative Structure–Activity Relationships » (QSAR) sont très utilisées.

A. Approches « Quantitative Structure–Activity Relationships » QSAR

Dans les années 50, Corwin Hansch (Hansch and Muir, 1950; Muir et al., 1949) introduit le concept d'une étude qualitative, les « Structure–Activity Relationships » (SAR). Quelques années plus tard il utilise les données acquises grâce à la synthèse d'acides phénoxyacétiques diversement substitués pour effectuer des traitements statistiques afin d'établir des corrélations entre les l'activité biologique de ces acides phénoxyacétiques et une fonction intégrant leurs coefficients de partage et les constantes de substituant de Hammett. Ces travaux fondateurs ont montré qu'en utilisant deux constantes, sigma et pi qui intègrent les effets de substitutions, l'effet des substituants sur l'activité biologique d'une molécule de même type pouvait être rationalisé (Hansch et al., 1962; Hansch and Fujita, 1964). Les constantes sigma et pi constituent ainsi les premiers descripteurs moléculaires permettant de mettre en lumière des relations quantitatives entre structure et activité, et naissent alors les approches « Quantitative Structure–Activity Relationships » (QSAR) (Ekins et al., 2007).

Les approches QSAR recouvrent les approches qui impliquent la mise en œuvre de modèles permettant d'établir une corrélation entre une structure moléculaire et une activité biologique ou une propriété chimique (quantitative structure-properties relationships, QSPR). Développées ensuite largement pour la pharmacologie *in silico*, les techniques QSAR se sont répandues ensuite à tous les domaines ainsi qu'à l'olfaction (Bajgrowicz and Frater, 2000; Chastrette et al., 1990).

La construction d'un modèle QSAR repose sur l'utilisation de la structure moléculaire des composés. Pour traduire celle-ci en termes quantitatifs, il est nécessaire de générer des descripteurs moléculaires qui encodent les caractéristiques structurales. Un modèle QSAR pour une activité spécifique peut être représentée par l'équation mathématique suivante :

$Y = f(X_1, X_2, \ldots X_n)$

où Y est une variable dépendante représentant une activité ou une propriété, X_1, X_2, … X_n sont des variables indépendantes représentant les descripteurs moléculaires et f est une fonction linéaire ou non linéaire.

Dans le cas d'une relation multilinéaire, la fonction qui relie la variable dépendante aux variables indépendante prend la forme :

$Y = a0 + a1X1 + a2X2 + a3X3 + \ldots + anXn$

où a_1, a_2, … a_n sont les contributions des descripteurs et a_0 est une constante (Roy et al., 2015). Ces descripteurs constituent un ensemble de valeurs numériques qui peut représenter une large diversité d'informations. Il est donc important de sélectionner les descripteurs pertinents et non-corrélés entre eux pour construire le modèle et identifier des propriétés moléculaires essentielles à l'activité des composés (Dudek et al., 2006; Karelson et al., 1999).

Un grand nombre de descripteurs a été élaboré depuis plus de vingt ans (Consonni and Todeschini, 2000). Leurs valeurs peuvent être calculées en utilisant des logiciels tels que Dragon (Mauri et al., 2006), PaDEL-Descriptor (Yap, 2011) ou MODEL (Li et al., 2007). Au total, plus de 5000 descripteurs moléculaires ont été définis (Consonni and Todeschini, 2010) et se divisent en deux grandes familles : les descripteurs 2D et 3D.

1. Descripteurs 2D

Les descripteurs 2D sont définis à partir de la représentation topologique des composés (Mauri et al., 2006). Ils regroupent :

- les descripteurs constitutionnels, qui décrivent les éléments constituant la structure de la molécule (poids moléculaire, nombre total d'atomes présents dans la molécule, nombre de liaisons simples, doubles ou triples…) (Dudek et al., 2006)
- les descripteurs électrostatiques, qui définissent la nature électronique de la molécule (somme de la polarisabilité atomique, somme des charges partielles) (Dudek et al., 2006)
- les descripteurs topologiques, fondés sur la théorie des graphes, qui décrivent la connectivité entre les atomes de la molécule (Consonni and Todeschini, 2010; Katritzky and Gordeeva, 1993; Kier et al., 1975; Randic, 1975)
- les descripteurs quantiques, qui décrivent l'orbitale moléculaire occupée la plus élevée (HOMO : highest occupied molecular orbital) et l'orbitale moléculaire non occupée la plus basse (LUMO : lowest unoccupied molecular orbital) (Dudek et al., 2006)
- les fingerprints moléculaires, qui saisissent les propriétés structurelles, topologiques des molécules (Xue et al., 2003). Les fingerprints sont représentés sous la forme de chaîne binaire de bits où chaque bit code la présence ou l'absence d'une sous-structure dans une molécule (Bajorath, 2001; Liu and Zhou, 2008). Les fingerprints constituent une classe particulière de descripteurs moléculaires et leurs caractéristiques sont développées à la fin de cette partie.

2. Descripteurs 3D

Les descripteurs 3D sont définis à partir de la représentation géométrique des composés et des coordonnées 3D des atomes de la molécule (Mauri et al., 2006). Parmi ces descripteurs, on trouve le volume moléculaire, les indices Shadow, qui correspondent à l'aire de l'ombre projetée de la molécule sur les plans XY, XZ, YZ de l'espace 3D, ou encore des descripteurs de la surface partielle chargée (Consonni and Todeschini, 2010; Katritzky and Gordeeva, 1993; Stanton and Jurs, 1990). Ces descripteurs 3D sont indépendants de l'alignement (mais

dépendant de la conformation), et on les distingue de ceux dépendants de l'alignement des molécule qui sont présentés dans la partie 3D-QSAR (Dudek et al., 2006).

3. Fingerprints

Les représentations fingerprints qui se basent sur le graphe 2D des molécules, constituent les fingerprints 2D (Cereto-Massagué et al., 2015). Il existe plusieurs groupes de fingerprints qui se distinguent par la manière dont les sous-structures moléculaires sont identifiées. Les fingerprints topologiques capturent les fragments d'une molécule en suivant un chemin linéaire jusqu'à un nombre donné de liaisons. Chaque chemin est ensuite haché afin de déterminer les bits activés pour chaque chemin, qui formeront ensuite l'ensemble des valeurs binaires constituant le fingerprint (Cereto-Massagué et al., 2015). Les fingerprints circulaires capturent l'environnement radial de chaque atome suivant un rayon donné (Muegge and Mukherjee, 2016). Tout comme les fingerprints topologiques, les fingerprints circulaires sont des fingerprints hachés. En conséquence, un bit ne correspond pas forcément à une sous-structure spécifique mais peut être défini par plusieurs sous-structures différentes. Ce phénomène est appelé « collision de bits » (Cereto-Massagué et al., 2015). Les fingerprints basés sur les clés de sous-structure associent chaque bit à la présence ou à l'absence de sous-structures dans une molécule à partir d'une liste donnée de clés structurelles (Cereto-Massagué et al., 2015). Les fingerprints hybrides combinent des segments binaires de fingerprints obtenus à partir de différentes approches afin de créer de nouveaux fingerprints (Nisius and Bajorath, 2009).

Une seconde catégorie de fingerprints se distingue en prenant en compte les caractéristiques de la structure 3D et regroupe ainsi les Fingerprints 3D. Parmi ces fingerprints, les « Extended 3D Fingerprints » ou E3FP sont des fingerprints sphériques permettant de capturer les motifs du voisinage tridimensionnel de chaque atome de manière itérative. Les atomes voisins sont donc contenus dans un espace sphérique de rayon r appelée coquille et à chaque itération le rayon de la coquille augmente définissant ainsi un voisinage de plus en plus grand (Axen et al., 2017). Un second type de fingerprints 3D est le « Three-Dimensional Force Fields Fingerprint » (TF3P) qui décrit les informations des champs de force 3D de petites molécules. Ces fingerprints ont été développés dans le but de prédire les cibles de ligand sachant que les liaisons entre les petites molécules et leurs cibles biologiques sont régies par des forces

physiques. Les TF3P sont créés à partir d'un réseau neuronal profond dont les données d'entrée sont les grilles de champs de force moléculaire. Les grilles de champs de force moléculaires sont ensuite compressées via le modèle afin de créer une matrice à valeurs réelles qui constituera les TF3P (Wang et al., 2020). Les ECFP ont été développées dans le but de déterminer les caractéristiques moléculaires nécessaires à l'activité moléculaire (Rogers and Hahn, 2010). En particulier, les ECFP4 (ECFP utilisant un diamètre de 4) sont connus pour leur efficacité (O'Boyle and Sayle, 2016; Rogers and Hahn, 2010), notamment sur les jeux de données de référence de petites molécules (benchmarks) (Capecchi et al., 2020). Les fingerprints ECFP4 sont d'ailleurs ceux utilisés au cours de notre étude.

B. 3D QSAR, CoMFA, CoMSiA

Les approches « CoMFA », analyse comparative des champs moléculaires, et « CoMSiA », analyse comparative des indices de similarité moléculaire, ont été les premières à prendre en compte la répartition dans l'espace 3D de propriétés moléculaires (Cramer et al., 1988). Toutes deux dépendent de l'alignement des molécules et de leurs conformations. La limite des approches CoMFA et CoMSiA est que le modèle généré repose sur un seul conformère choisis arbitrairement pour chaque molécule, et ne tient pas compte de la flexibilité des molécules. Le conformère choisi est généralement celui de plus basse énergie, dont on ignore s'il est réellement le conformère actif pour la liaison au récepteur.

Contrairement aux approches CoMFA et CoMSiA, l'analyse comparative des moments moléculaires (CoMMA) (Silverman and Platt, 1996), les descripteurs moléculaires invariants holistiques pondérés (WHIM) (Todeschini et al., 1994) et les descripteurs indépendants de la grille (GRIND) (Pastor et al., 2000) sont indépendants de l'alignement des molécules.

C. Pharmacophores

Parmi les différentes approches QSAR, une approche spécifiquement tridimensionnelle fait appel à la notion de pharmacophore. Les pharmacophores sont des représentations abstraites des molécules qui associent dans un espace tridimensionnel des caractéristiques moléculaires dans une géométrie particulière. On emploie également le terme « hypothèse », c'est-à-dire qu'il s'agit d'une proposition sur ce que peuvent être l'arrangement de régions des molécules

impliquées dans l'activation du récepteur cible. Un avantage majeur est que les algorithmes prennent en compte un ensemble de conformères, et par conséquent la flexibilité des molécules.

Un pharmacophore est défini officiellement par l'Union internationale de chimie pure et appliquée (International Union of Pure and Applied Chemistry, IUPAC) comme étant « l'ensemble des caractéristiques stériques et électroniques nécessaires pour assurer les interactions supramoléculaires optimales avec une structure cible biologique spécifique et pour déclencher (ou bloquer) sa réponse biologique » (Wermuth et al., 1998).

Le concept de pharmacophore est généralement attribué à Paul Ehrlich (Ehrlich, 1909). Ce dernier aurait constaté que différents tissus étaient colorés de manière sélective par différents colorants, définissant ainsi le « chromophore » comme la partie essentielle d'une molécule permettant la coloration (Ehrlich, 1909). En transposant cette idée à la médecine, Paul Ehrlich a décrit le pharmacophore comme « une structure moléculaire qui porte (phoros) les caractéristiques essentielles responsables de l'activité biologique (pharmacon) d'un médicament » (Ehrlich, 1909). Néanmoins, c'est Peter Günd (Günd, 1977) qui donnera une définition plus générale du terme pharmacophore, le décrivant comme « un ensemble de caractéristiques structurelles d'une molécule qui est reconnu par un site récepteur et qui est responsable de l'activité biologique de cette molécule ». Les pharmacophores sont donc utilisés afin d'identifier des molécules actives remplissant certaines contraintes géométriques et chimiques leur conférant une activité biologique, qu'elle soit bénéfique ou nocive (Liu et al., 2013; Thai et al., 2013).

1. Génération des pharmacophores

Les pharmacophores sont développés à l'aide de différents logiciels, dont DISCO (Martin et al., 1993), Catalyst/HipHop et Catalyst/HypoGen (Barnum et al., 1996) à présent intégrés à Biovia Discovery Studio, PHASE (Dixon et al., 2006), LigandScout (Wolber and Langer, 2005), ou encore MOE (Molecular Operating Environment) (*MOE*, 2005). Les structures tridimensionnelles des récepteurs ne sont pas toujours connues, et le pharmacophore d'un ensemble de ligands permet d'avoir une image indirecte des caractéristiques de leur site récepteur commun.

Les différentes méthodes de modélisation qui peuvent être utilisées pour la construction des pharmacophores (Qing et al., 2014) impliquent trois étapes principales : l'extraction de « features », l'identification des motifs et la sélection des pharmacophores.

Extraction des caractéristiques moléculaires et structurales, « features »

Les « features » sont des éléments structuraux des ligands définis par leur topologie, leur fonction ou leur connectivité (Khedkar et al., 2007; van Drie, 2003). Que ce soit pour un ensemble de ligands ou pour une seule molécule, il faut tout d'abord définir la nature des caractéristiques moléculaires, « features », pertinentes pour la construction du pharmacophore (Khedkar et al., 2007). En effet six principaux types de « features » sont généralement utilisées : accepteur de liaison hydrogène, donneur de liaison hydrogène, cycle aromatique, groupement hydrophobe, ionisables positivement ou négativement. Chaque « feature » est représentée par une sphère dont le rayon définit la tolérance de déviation de la position précise de la « feature » (*Catalyst*, 2014; Qing et al., 2014).

Identification des motifs et sélection des pharmacophores

Les « features » des molécules sont associées afin de constituer des pharmacophores candidats (Khedkar et al., 2007).

Les algorithmes permettent de générer un ensemble de pharmacophores, il est donc nécessaire d'établir un classement hiérarchique en fonction de leur fiabilité. Selon les logiciels, différents critères sont utilisés pour estimer cette fiabilité. La plupart de ces algorithmes classent les pharmacophores candidats au moyen d'une fonction de score qui peut intégrer la correspondance entre les « features » et les groupements chimiques, le chevauchement des volumes des ligands, l'énergie de déformation d'une molécule pour correspondre au pharmacophore ou encore la sélectivité d'une hypothèse pharmacophore. Comme on peut le voir avec l'exemple qui va suivre, certains des termes utilisés dans les fonctions de score ont une justification physique sous-jacente mais d'autres, comme la sélectivité, sont plus subjectifs et sont inclus pour augmenter le score des pharmacophores les plus "pertinents" (Leach et al., 2010).

Par exemple, le logiciel Catalyst utilise la fonction suivante pour le calcul du score (cost) :

$Cost = eE + wW + cC$

Les coefficients *e*, *w* et *c* sont par défaut fixés à 1 mais peuvent être modifiés par l'utilisateur. La valeur *W* augmente lorsque le poids d'une « feature » s'écarte d'une valeur idéale fixée par défaut à 2,0. La composante *E* augmente avec la différence de moyenne quadratique entre les activités estimées et mesurées. Le paramètre *C* est un coût fixe qui dépend de la complexité du pharmacophore (Kurogi and Guner, 2001; Sutter et al., 2000).

A la différence, la fonction utilisée par le logiciel PHASE pour calculer le score est la suivante :

score = $F + w_vV - w_eE + w_m{}^{M-1} + w_sS$

F mesure la qualité de l'alignement des « features » dans la molécule de référence avec les « features » appariées correspondantes dans chaque molécule. *V* mesure la superposition de la molécule de référence avec chaque molécule. Ce paramètre correspond au ratio du volume commun des deux molécules sur le volume total occupé par les deux molécules. *E* est l'énergie de déformation du conformère de la molécule à partir de laquelle le pharmacophore est construit. *M* correspond au nombre de molécules qui coïncident convenablement dans l'espace avec l'hypothèse pharmacophore (qui « mappent » le pharmacophore). *S* est la sélectivité, c'est-à-dire une estimation empirique de la rareté d'une hypothèse. Cette mesure représente la fraction de molécules, dans une base de données aléatoire, susceptible de correspondre à l'hypothèse. Une faible fraction de molécules se traduit par une sélectivité élevée, indiquant que la disposition des « features » dans l'hypothèse est inhabituelle. Donc une configuration inhabituelle de l'hypothèse suggère qu'elle est nécessaire à l'activité des molécules (Dixon et al., 2006; Leach et al., 2010). Enfin, w_v, w_e, w_m et w_s sont des poids définis par l'utilisateur.

Par ailleurs, deux types de modélisation de pharmacophores se distinguent : la modélisation basée sur les ligands et la modélisation basée sur la structure.

2. Modélisation basée sur la structure

La modélisation basée sur la structure est utilisée lorsque la structure tridimensionnelle de la protéine de la protéine cible a pu être déterminée, par une méthode spectroscopique (diffractométrie des rayons X ou résonnance magnétique nucléaire) ou par cryo-microscopie (Bai et al., 2015). Dans le cas où le ligand est connu, et si la structure du complexe protéine-ligand a pu être déterminée, la modélisation est basée sur ce complexe macromolécule-ligand. Dans le cas contraire, il s'agit d'une modélisation basée sur la structure de la macromolécule

seule (Yang, 2010). Si la structure tridimensionnelle de la protéine n'a pas été déterminée expérimentalement, il est possible de générer un modèle par homologie en se basant sur la structure connue d'une protéine de la même famille (Levoin et al., 2011), en y associant une modélisation par dynamique moléculaire (Schaller et al., 2020) et de simuler la liaison avec un ligand.

a. Modélisation basée sur le complexe macromolécule-ligand

Cette modélisation implique l'utilisation du complexe de la molécule cible avec un de ses ligands, lié au site de fixation. Les pharmacophores basés sur ces complexes sont donc développés à la fois à partir des caractéristiques du site actif de la molécule cible et de celles du ligand. Cette méthode permet de définir les principales interactions entre la protéine et le ligand à partir desquels un modèle de pharmacophore commun sera dérivé (Wolber and Langer, 2005). Les pharmacophores peuvent être construits en utilisant la structure 3D d'un seul ligand mais il est préférable d'utiliser les structures de plusieurs ligands pour identifier les caractéristiques communes responsables de leur interaction avec la molécule cible.
La majorité des études utilisant la modélisation basée sur la structure se concentrent sur des récepteurs protéiques, mais d'autres structures macromoléculaires, comme les acides nucléiques, peuvent également être concernées (Schaller et al., 2020). Par exemple, des pharmacophores ont déjà été développé à partir de ligands liant le petit sillon polyamide en utilisant un complexe ligand-ADN (Spitzer et al., 2007).

b. Modélisation basée sur la macromolécule seule

Cette modélisation est utilisée lorsqu'aucun complexe macromolécule-ligand n'est disponible. Avec cette méthode, les pharmacophores sont développés à partir de la structure de la cible macromoléculaire seule (modèles atomistiques de macromolécules liées à aucun ligand, dit « apo ») (Schaller et al., 2020) et les propriétés chimiques du site actif sont analysées (Qing et al., 2014; Tintori et al., 2008).
Ensuite les structures atomiques des sites de fixation sont converties en « features » pour construire les pharmacophores.

3. Modélisation basée sur les ligands

La modélisation basée sur les ligands est utilisée lorsque la structure des protéines cibles n'est pas connue (Schaller et al., 2020). Un ensemble de plusieurs ligands, généralement appelé ensemble d'entraînement (« training set »), est alors utilisé pour développer les pharmacophores. Le choix des ligands est très important car la construction du pharmacophore est influencée par la nature des ligands, leurs fonctions chimiques et le nombre de ligands présents dans le jeu de données. Il est donc essentiel de choisir des ligands avec des structures et des activités suffisamment diverses pour constituer les données d'entrée (Khedkar et al., 2007). Au moins deux molécules actives, ou une molécule rigide très active, sont requises afin de développer des modèles basés sur les ligands. Eventuellement, des composés inactifs peuvent être ajoutés à l'ensemble de molécules (Vuorinen and Schuster, 2015).

La construction de pharmacophores à partir de plusieurs ligands nécessite l'identification des « features » communes aux ligands qui est réalisée en comparant et en combinant les « features » extraites des ligands (Khedkar et al., 2007). La première étape implique la génération des conformères de chaque ligand de l'ensemble d'entraînement afin de représenter la flexibilité conformationnelle des ligands (Yang, 2010). A partir de l'ensemble des conformères des ligands, l'algorithme de génération de pharmacophore commence par aligner les structures 3D afin de déterminer les « features » communes aux conformères situées à la même position (Schaller et al., 2020). La gestion de la flexibilité conformationnelle des ligands et l'alignement moléculaire sont les techniques clés de la modélisation basée sur les ligands (Yang, 2010). Une fois les « features » extraites les pharmacophores candidats vont pouvoir être construits.

La méthodologie de construction de pharmacophores diffère selon le logiciel utilisé. Nous avons choisi de limiter les exemples à deux logiciels commerciaux dont la comparaison a fait l'objet de plusieurs études reportées dans la littérature : Catalyst, commercialisé en 1995, et PHASE (Schrödinger) (*Phase, Schrödinger*, 2021), apparu dix ans plus tard (Dixon et al., 2006; Evans et al., 2007; Spitzer et al., 2010).

Dans les deux cas, les composés sont préalablement traités avec un module de préparation des ligands, ce qui permet d'attribuer des états de protonation appropriés pour un pH donné, déterminé par l'utilisateur. L'étape suivante consiste à générer les conformères (analyse de la

couverture conformationnelle associée à une optimisation conformationnelle par la fonction « poling » avec Catalyst (Smellie et al., 1995a, 1995b, 1995c) ; recherche de torsion de ligand par le module MacroModel pour PHASE (Evans et al., 2007; Greenidge and Weiser, 2001)).
Les deux algorithmes essentiels de Catalyst, HipHop et Hypogen sont à présent implémentés dans l'environnement Biovia Discovery Studio ; par commodité, nous emploierons l'ancienne appellation « Catalyst ». La création des hypothèses se fait en deux phases avec Catalyst. La première phase dite constructive, permet de sélectionner les hypothèses dont les « features » correspondent aux composés actifs. La seconde phase, appelée soustractive, consiste à rejeter toute hypothèse qui correspondrait au moins à la moitié des composés inactifs. Les hypothèses, qui ont été retenues suite à ces deux phases, sont ensuite scorées comme expliqué précédemment. (Kurogi and Guner, 2001).
La conception des pharmacophores avec PHASE possède des similitudes avec celle de Catalyst mais diffère sur certains points. Les composés du jeu de données doivent être défini comme étant actif ou inactif. L'utilisateur choisit le nombre de composés actifs qui doivent correspondre aux pharmacophores, le type et le nombre de chaque « feature » (0 à 3), le nombre minimum et le nombre maximum de « features » composant les hypothèses de pharmacophores (3 à 7). Une fois les hypothèses créées, un score leur est attribué, basé sur l'alignement géométrique des composés (actifs et inactifs) sur les « features » des hypothèses de pharmacophores (Evans et al., 2007).

Au cours de cette thèse, nous avons utilisé la modélisation basée sur les ligands à l'aide du logiciel PHASE version 2021-4 (Dixon et al., 2006) afin de développer des pharmacophores à partir des molécules de deux mélanges menant à une perception homogène.

4. Modèles 3D QSAR pharmacophores

Lorsque les activités des molécules sont connues et quantifiées, les pharmacophores obtenus par les algorithmes de Catalyst (HypoGen) et de PHASE peuvent être utilisés pour générer des modèles quantitatifs 3D-QSAR. Comme dans le cas des pharmacophores classiques, ces modèles proposent la distribution dans l'espace des caractéristiques communes aux ligands. De plus, ils associent aux alignements des ligands une équation reliant l'activité à des critères de superposition entre les groupements chimiques et les « features » des modèles,

permettant ainsi de prédire l'activité de nouveaux ligands potentiels par exploration de bases de données.

II. Apprentissage automatique

L'apprentissage automatique ou Machine Learning est une technologie qui consiste à développer des algorithmes informatiques capables d'émuler l'intelligence humaine. Ces algorithmes sont des processus informatiques qui utilisent des données d'entrée pour réaliser une tâche souhaitée et destinée à produire un résultat particulier. Un algorithme d'apprentissage automatique est répété plusieurs fois, ce qui lui permet d'adapter son architecture afin de devenir de plus en plus performant dans la réalisation de la tâche souhaitée (El Naqa and Murphy, 2015). On distingue parmi les algorithmes d'apprentissage automatique, les algorithmes supervisés et non supervisés.

A. Apprentissage supervisé

Pour les algorithmes supervisés, lors du processus d'adaptation, appelé apprentissage, les échantillons de données d'entrée peuvent être fournis avec les résultats souhaités. L'algorithme se configure alors de manière optimale dans le but de pouvoir retrouver le résultat souhaité lorsqu'il est présenté avec les entrées d'entraînement, mais également de pouvoir généraliser les résultats issus de l'apprentissage pour produire le résultat souhaité à partir de nouvelles données inconnues auparavant (El Naqa and Murphy, 2015).

B. Apprentissage non supervisé

Dans le cas d'un algorithme non supervisé, l'apprentissage ne repose pas sur l'association d'entrée particulière à un résultat particulier (El Naqa and Murphy, 2015). Les données ne sont donc pas catégorisées et l'apprentissage se fait via l'exploration des données d'entrée et la découverte des similarités entre elles. Donc toutes les variables utilisées dans l'analyse sont utilisées comme entrées et les similarités sont utilisées afin de définir les groupements naturels dans les données (Tarca et al., 2007). De ce fait, ces algorithmes sont utilisés pour des techniques de réduction de dimensions ou de clustering (Alloghani et al., 2020).

1. Réduction de dimensions

Les avancées scientifiques et technologiques permettent aujourd'hui de produire des données de plus en plus nombreuses et complexes. Cette complexité est caractérisée notamment par un grand nombre de variables caractérisant les objets des données (Ma and Zhu, 2013). Dans ces cas, l'analyse nécessite la réduction des dimensions des données afin d'améliorer son efficacité et sa précision. En effet, l'augmentation du nombre de variables dégrade la performance des algorithmes d'apprentissage automatique. Les techniques de réduction de dimensions permettent donc de simplifier les modèles de données afin de les analyser. Pour cela, elles identifient une représentation à faible dimension appropriée des données originales à hautes dimensions. Avec cette représentation réduite, des analyses telles que la classification, peuvent souvent donner des résultats plus précis et plus facilement interprétables (Cunningham, 2008).

Durant ce travail de thèse quatre techniques de réduction de dimensions ont été utilisées : l'analyse en composantes principales ACP, « Multidimensional Scaling » (MDS), « t-distributed Stochastic Neighbor Embedding » (t-SNE) et « Uniform Manifold Approximation and Projection » (UMAP).

L'ACP repose sur le lemme de William Johnson et Joram Lindenstrauss qui établit qu'« un ensemble de points dans un espace à hautes dimensions peut être projeté dans un sous-espace de dimensions inférieures de telle sorte que les distances relatives entre les points de données sont presque préservées » (Johnson and Lindenstrauss, 1984). L'ACP est une méthode d'analyse multivariée qui permet d'extraire les informations les plus importantes d'un jeu de données. Les variables d'origine subissent une transformation orthogonale afin de générer de nouvelles variables linéairement indépendantes appelées composantes principales (Abdi and Williams, 2010).

Contrairement à l'ACP, les méthodes MDS, t-SNE (Anowar et al., 2021) et UMAP sont des basées sur le concept de manifold. Un manifold est défini comme un espace topologique localement euclidien de n dimensions (Lee, 2010; Liao and Triantaphyllou, 2008). En d'autres termes, un manifold représente un espace où les points sont rassemblés en fonction de leurs distances euclidiennes, formant ainsi une carte continue. En fonction des proximités entre les points, différents manifolds peuvent être identifiés sur la carte (Abraham et al., 1988). Ainsi

les méthodes basées sur le manifold impliquent que les données reposent sur un manifold densément échantillonné (Sedlmair et al., 2012).

Le MDS est une technique de localisation de réseau qui cartographie le long du manifold la similarité ou la dissimilarité des paires d'objets d'un ensemble de données. La similarité (ou la dissimilarité) est ensuite convertie en distances euclidiennes entre des points dans un espace à faibles dimensions (Borg, 2018; Saeed et al., 2019).

La méthode t-SNE est une version améliorée de la méthode SNE (Stochastic Neighbor Embedding). L'algorithme SNE convertit les distances euclidiennes entre les points de données à hautes dimensions en probabilités conditionnelles représentant les similarités afin de les représenter sur le manifold. Des probabilités conditionnelles similaires sont calculées pour les points des données à basses dimensions. Une fois les probabilités calculées pour les représentations à haute et à basse dimension, l'objectif est de minimiser le décalage entre les deux vecteurs de similarités (Oliveira et al., 2018). Ce décalage se mesure avec la somme des divergences de Kullback-Leibler (mesure statistique qui quantifie la proximité entre une distribution de probabilité et une distribution modèle (Shlens, 2014)) sur tous les points en utilisant une méthode de descente de gradient. La méthode t-SNE se différencie de la méthode SNE en utilisant des gradients plus simples et en appliquant la distribution t de Student plutôt qu'une gaussienne pour calculer la similarité entre deux points dans l'espace à faibles dimensions (Oliveira et al., 2018).

La méthode UMAP est une technique de réduction de dimension relativement récente (McInnes et al., 2018) permettant de capturer précisément la structure non linéaire de grands ensembles de données. Comme déjà mentionné, UMAP est une technique basée sur le manifold et qui repose sur la géométrie riemannienne et la topologie algébrique. UMAP utilise les approximations locales des manifolds et regroupe leurs représentations locales d'ensembles simpliciaux flous pour construire une représentation topologique des données de hautes dimensions (un ensemble simplicial est la présentation combinatoire d'un espace topologique pondéré par des probabilités (Jackson, 2019). Pour une certaine représentation de basses dimensions des données, une représentation topologique équivalente peut être construite en utilisant un processus similaire. Ensuite, la disposition de la représentation des données est optimisée dans l'espace de basses dimensions permettant la minimisation de l'entropie croisée entre les deux représentations topologiques (McInnes et al., 2020). Concrètement, la réduction de dimension par UMAP commence par le calcul des distances

entre chaque point. L'algorithme considère ensuite, pour chaque point, ses plus proches voisins et attribue un poids (probabilité) de lien entre le point considéré et ses n voisins. UMAP construit ainsi un graphe pondéré et utilise ensuite sur celui-ci un algorithme de dessin fondés sur les forces pour projeter et représenter les données de manière optimale à basse dimension. Ces algorithmes sont basés sur des analogies issues de la physique. Un système de forces appliquées entre les nœuds et les liens, est utilisé. Les nœuds se repoussent comme des particules de même charge tandis que les liens tendent à rapprocher les nœuds voisins à l'instar de ressorts (Bahoken et al., 2013).

UMAP possède l'avantage d'être personnalisable et peut ainsi être adapté à chaque jeu de données. En effet, plusieurs paramètres peuvent être modifiés après l'intégration des données : la méthode de calcul de la distance, le nombre de voisins, la distance minimale pour le regroupement des points en basse dimension, ou encore le nombre de dimensions souhaité. Le choix approprié des valeurs de ces paramètres est crucial pour discerner des groupes contenant des molécules structurellement proches.

Le calcul de la distance repose sur le calcul des similarités/dissimilarités entre les objets, et déterminent la similitude entre deux ensembles à n dimensions. Pour ce faire plusieurs métriques sont disponibles. Par exemple, pour des données quantitatives on peut calculer les matrices de similarités par le Cosinus, et les matrices de dissimilarités par les distances de Canberra, de Manhattan, de Minkowski et la distance euclidienne. Les fingerprints étant des données binaires, les similarités peuvent être calculées en utilisant l'indice de Dice ou l'indice de Jaccard (Sachdeva et al., 2009; Suebsing and Hiransakolwong, 2012; Zou et al., 2004).

Nous avons retenu dans un premier temps les indices de Tanimoto/Jaccard, de Dice et la similarité Cosinus. L'indice de Tanimoto/Jaccard est l'une des mesures de similarité les plus fondamentales et les mieux adaptées (Bajusz et al., 2015) dans la comparaison de données biologiques de présence-absence (Chung et al., 2019) et a donné les meilleurs résultats lors d'essais préliminaires sur nos données. Nous l'avons par conséquent sélectionné pour notre étude.

Le paramètre « nombre de voisins » permet d'équilibrer la structure locale et la structure globale dans les données. Le nombre de voisins permet donc de déterminer la taille du voisinage local pris en compte lors de l'apprentissage de la structure multiple des données par UMAP. Ainsi, lorsque ce paramètre est faible, UMAP se concentre sur la structure très locale, même au détriment de l'image globale. Mais lorsqu'il est plus élevé, UMAP se concentre sur

des voisinages plus larges de chaque point mais en conséquence perd en précision sur la structure locale.

Le paramètre « distance minimale » contrôle la distance avec laquelle les points sont regroupés dans la représentation à faible dimension. Ce paramètre définit donc la distance minimale autorisée entre deux points de la représentation à basses dimensions. Avec des valeurs faibles, les points sont plus agglutinés les uns aux autres. Au contraire avec des valeurs plus élevées, les points sont beaucoup moins regroupés, et dans ce cas UMAP préserve mieux la structure topologique générale.

Différentes valeurs des paramètres "nombre de voisins" et "distance minimale" ont été testées, et sur la base des résultats préliminaires ainsi obtenus, le nombre de voisins et la distance minimale ont été fixés respectivement à 15 et 0.

Le paramètre « nombre de dimensions » permet de déterminer la dimensionnalité de l'espace de dimensions réduites dans lequel les données sont intégrées ; le plus souvent, deux ou trois dimensions sont utilisées.

2. Classification

La classification (ou « clustering ») est une technique dont l'objectif est d'organiser un ensemble d'éléments de données en groupes (ou « clusters »), de manière à ce que les éléments d'un même groupe soient plus similaires les uns aux autres qu'aux éléments des autres groupes (Grira et al., 2005). Cette notion de similarité entre deux éléments s'exprime par une métrique de distance. Il existe plusieurs métriques de distance telles que la distance de Manhattan ou la distance euclidienne, qui peuvent avoir un impact différent sur l'analyse des données (Gentleman and Carey, 2008).

Parmi les différentes techniques de classification existantes, la classification ascendante hiérarchique (CAH), les k-means et les cartes auto-organisatrices de Kohonen (« Self-Organizing Maps », SOM) (Kohonen, 2013, 1998) ont été utilisées au cours de notre étude.

Dans le cas de l'utilisation de la méthode CAH, une matrice de distance euclidienne 2 à 2 de chaque molécule a été calculée. Les deux classes les plus proches sont ensuite successivement regroupées jusqu'à obtenir un arbre de classification complet. Afin de regrouper ces différentes classes formées, plusieurs méthodes d'agrégation de clusters ou méthodes de

liaison peuvent être utilisée ("single linkage", "average linkage", "complete linkage", Ward…). La méthode "single linkage" calcule toutes les dissemblances par paires entre les éléments d'un premier cluster et ceux d'un second cluster et considère la plus petite des dissemblances comme distance entre les deux clusters. La méthode "centroid linkage" calcule la dissemblance entre les centroïdes de deux clusters. La méthode "complete linkage" calcule toutes les dissemblances par paires entre les éléments d'un premier cluster et ceux d'un second cluster et considère ensuite la plus grande valeur de ces dissemblances comme la distance entre les deux clusters.
Pour notre étude, la CAH a été réalisée avec le critère agrégatif "ward.D2", présenté comme étant plus performant que les autres méthodes citées (Bipul Hossen, 2015). En effet, la méthode Ward s'est avérée être celle permettant une classification la plus proche de la vraie répartition d'ensembles de données en différents groupes (Bipul Hossen, 2015). Cette méthode permet de minimiser l'inertie intra-classe et maximiser l'inertie inter-classe. En d'autres termes, la variance au sein d'un même groupe sera minimisée tandis que celle entre deux clusters sera augmentée.

L'algorithme de la méthode des k-means est un algorithme d'apprentissage non supervisé. La première étape consiste à définir un nombre de centroïdes correspondant au nombre de clusters attendus. Chaque élément du jeu de données est assigné au cluster dont le centroïde est le plus proche. A chaque itération, la moyenne de tous les points d'un même cluster est calculée afin de redéfinir de nouveaux centroïdes. Les centroïdes des clusters sont améliorés à chaque itération jusqu'à ce qu'il n'y ait plus de changement (Ordonez, 2003). Une limite de l'algorithme des k-means est qu'il nécessite de définir initialement un nombre de partitions, ce qui suppose d'en avoir une idée précise.

Les cartes auto-organisatrices SOM constituent une classe de réseau de neurones artificiels fondée sur des méthodes d'apprentissage non-supervisées qui permettent de cartographier des données à hautes dimensions de manière ordonnée sur une grille à basses dimensions. L'algorithme crée une grille bidimensionnelle avec des modèles d'observation représentés par des nœuds. Les modèles sont calculés dans le but de décrire de manière optimale le domaine des observations des données. Les modèles les plus similaires sont plus proches, et les modèles dissemblables sont plus éloignés dans la grille. Chaque élément du jeu de données

sélectionne le modèle qui lui correspond le mieux. La classification se fait ensuite suivant le partitionnement de la grille (Kohonen, 2013, 1998), et ainsi les informations topologiques et métriques les plus importantes des données sont conservées..

III. Le modèle réseau

Un réseau ou graphe est un modèle relationnel entre les composants individuels d'un système entier qui permet donc de représenter les connexions existantes entre ces différents composants. Cette représentation se traduit simplement par un ensemble de nœuds reliés entre eux par des liens (Oh and Monge, 2016). Les nœuds et les liens peuvent être pondérés ou classifiés avec des informations supplémentaires, telles que le nom, la force ou la fréquence de leur interaction. Par exemple, les connections d'un réseau social peuvent être pondérées par le nombre de messages envoyés entre deux individus (fréquence) ou la durée depuis laquelle les deux individus se connaissent (force). Mais un réseau reste une représentation simplifiée qui préserve uniquement l'essentiel des motifs de connexion donc beaucoup d'informations peuvent être perdues lors du processus de réduction d'un système à une représentation de réseau (Newman, 2010). Pour autant, ils permettent de fournir des informations cruciales dans la compréhension de la structure et des mécanismes des systèmes qu'ils représentent (Oh and Monge, 2016).

A. Origines du modèle réseau

On peut dater le début de la théorie des graphes avec l'article de Leonhard Euler (Euler, 1741) sur le problème des ponts de Königsberg. La question posée par ce problème était de déterminer s'il existait un itinéraire permettant de traverser chaque pont une seule et unique fois tout en revenant au point de départ (Figure 11). Afin de répondre à cette question, Euler a introduit de nouveaux théorèmes, constituant la théorie des graphes. Chaque secteur de la ville a été représentée par un point et le franchissement d'un pont reliant une région à une autre représenté par une ligne. Le réseau construit de cette manière comporte quatre nœuds et sept arêtes. Leonhard Euler a ainsi démontré qu'aucun itinéraire ne pouvait répondre aux critères (Théorème d'Euler) (Oh and Monge, 2016).

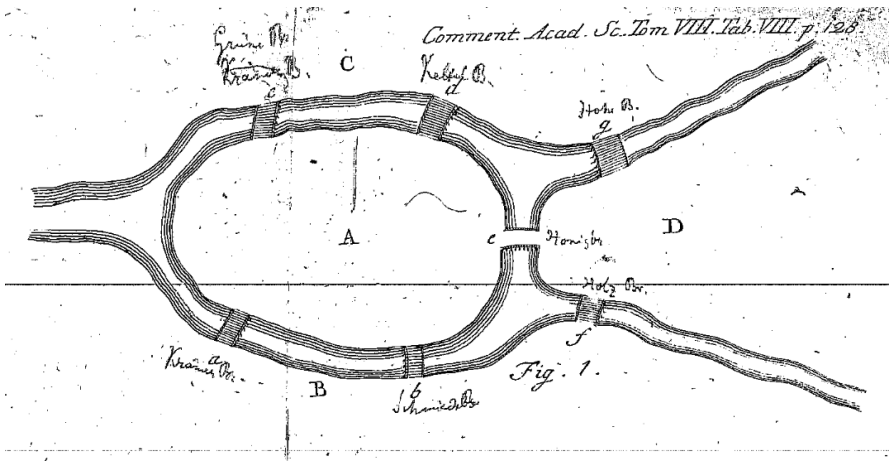


Figure 11 : Carte des sept ponts de Königsberg dessinée par Euler (tiré d'Euler,1741).

Ensuite, en 1889, le mathématicien britannique Cayley a utilisé une approche de la théorie des graphes dans l'étude d'un type particulier de graphes : les arbres (réseau sans cycles ni boucles). Pour les « Cayley-trees », comme ils sont nommés, chaque nœud a un nombre identique de liens, à l'exception des nœuds situés à la fin du réseau. Plus tard, le mathématicien français Berge (1962) a contribué de manière significative au domaine. Il a également rappelé et utilisé le nombre cyclomatique μ, qui compte essentiellement le nombre de boucles ou de cycles dans un graphique. La théorie des graphes est aujourd'hui utilisée et développée dans de nombreux domaines tels que l'étude des flux migratoires en géographie, l'étude des interactions entre médicaments et cibles protéiques en biologie ou encore l'étude des relations entre individus d'un système social en sociologie (Derrible and Kennedy, 2011).

B. Les différentes structures de réseaux

Selon les informations comprises dans le réseau, il se structure de différentes manières. Les réseaux simples (Figure 12A) sont des réseaux qui ne possèdent ni boucle (lien liant un seul même nœud) ni liens parallèles (plusieurs liens reliant deux mêmes nœuds). Il existe également d'autres représentations de réseaux plus complexes. Par exemple, si les nœuds d'un réseau sont regroupés dans deux sous-ensembles distincts, alors le réseau sera dit bipartite (Figure 12B). On parle également de réseau multipartite lorsque les nœuds sont regroupés dans plus de deux sous-ensembles et de réseau monopartite lorsque les nœuds sont compris dans un unique ensemble. Les réseaux peuvent également être orientés ou non orientés. Dans le cas d'un réseau orienté (Figure 12C), les liens possèdent une direction. Ce

type de réseau est très utile lorsque l'on veut représenter l'action d'un élément sur un autre (par exemple, pour indiquer les directions de la circulation de l'information entre différentes populations de neurones (Shih et al., 2015)). Le réseau est donc non orienté si les liens ne possèdent pas de direction. Dans un graphe pondéré (Figure 12D), des valeurs sont attribuées aux nœuds et/ou aux liens (par exemple, pour pondérer la connectivité entre différentes régions du cerveau et déterminer les régions qui interagissent le plus entre elles (Arnold et al., 2020)). Un réseau cyclique, ou tout simplement cycle (Figure 12E), est un réseau dont les nœuds forment un cycle parfait. Au contraire, un réseau acyclique est donc un réseau qui ne contient pas de cycles (Bondy and Murty, 2008).

Il existe également des réseaux basés sur la théorie des hypergraphes développée par Claude Berge (Berge, 1983) et présentée comme une généralisation de la théorie des graphes. Le concept d'hypergraphe permet de modéliser des relations entre plus de deux nœuds (Bretto et al., 2002). Par exemple, il serait possible d'utiliser un hypergraphe afin de représenter les interactions entre les molécules odorantes de mélanges homogènes avec chaque récepteur olfactif qu'elles activent.

Certains réseaux sont aussi nommés « sans échelle » (« scale-free ») car certains nœuds ont un nombre apparemment considérable de liens et qu'aucun nœud n'est typique des autres. Ces nœuds, appelés « hubs », possèdent de nombreuses connections et constituent donc les points centraux d'un réseau (Barabási and Bonabeau, 2003; Müller-Linow et al., 2008).

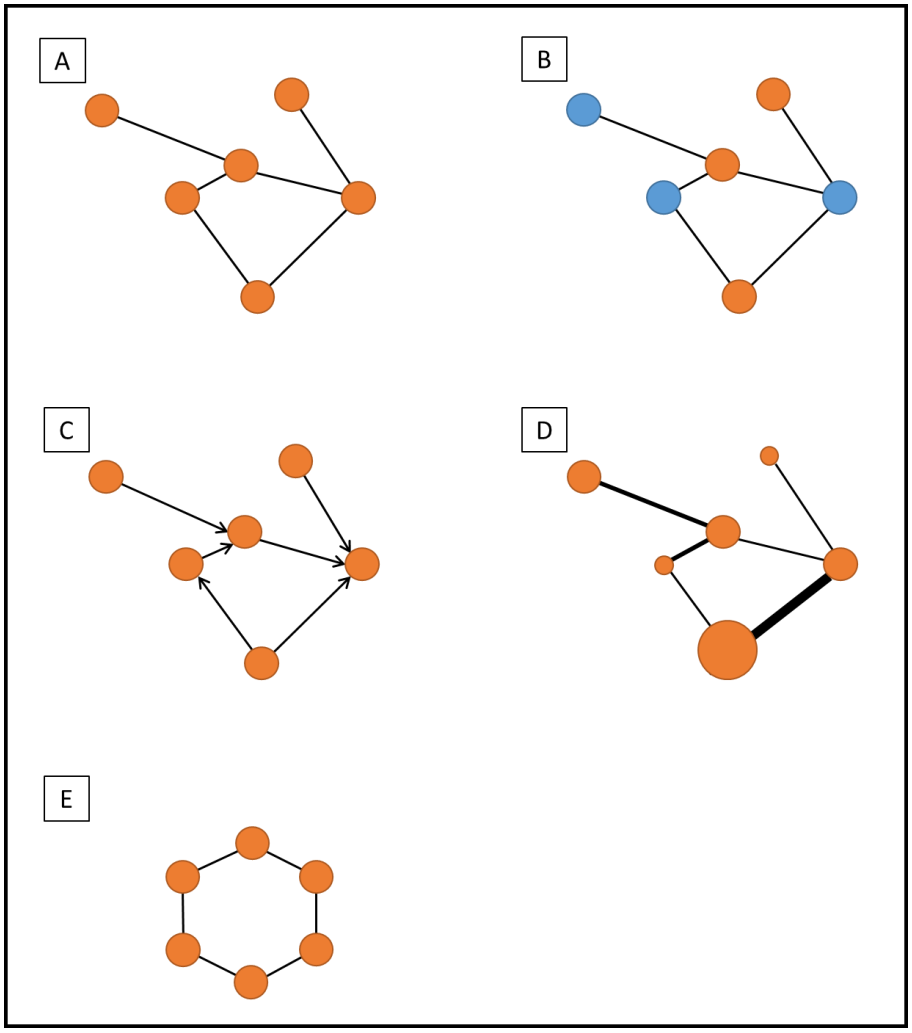


Figure 12 : Différentes structures de réseaux. A : Réseau simple. B : Réseau bipartite. C : Réseau orienté. D : Réseau pondéré. E : Réseau cyclique.

Partie 3 : Résultats

Dans cette partie sont présentées les trois études réalisées au cours de cette thèse. Chaque étude correspond à un des modèles mentionnés précédemment.

I. Classification de composés odorants à l'aide d'UMAP pour augmenter la connaissance des liens entre les odeurs et les structures moléculaires

Le décryptage du code combinatoire est un élément clé dans la compréhension de la perception olfactive, mais à l'heure actuelle peu de récepteurs olfactifs ont des ligands connus (Sharma et al., 2022). Toutefois, sachant que l'odeur d'une molécule odorante est définie par les récepteurs olfactifs qu'elle active (Malnic et al., 1999b; Touhara, 2002b), et en considérant que les molécules odorantes détectées par un même récepteur olfactif ont des structures similaires (Malnic et al., 2004), il doit être possible d'établir des relations entre la structure d'une molécule odorante et son odeur, qui permettraient de mieux comprendre la perception olfactive. Dans ce but, nous avons construit un modèle combinant une technique de réduction de dimension et une méthode de classification. Pour cela, un jeu de données répertoriant des composés odorants et leurs odeurs associées connues a été construit à partir d'informations issues de bases de données. Les structures moléculaires des composés odorants ont ensuite été encodées en fingerprints de 1024 bits à partir desquels quatre techniques de réduction dimensionnelle (PCA, MDS, t-SNE et UMAP) et deux méthodes de classification (k-means et CAH) ont été appliquées. L'association de l'UMAP avec les méthodes k-means et CAH a fourni les meilleurs résultats. En calculant la présence ou l'absence de sous-structures moléculaires chez les composés odorants, nous avons pu relier les groupements chimiques aux odeurs. Ainsi, nous avons pu observer des associations entre les notes "boisées" et "épicées" avec des structures allyliques et bicycliques, les notes "balsamiques" avec des cycles insaturés ou encore les notes "huileuses", "grasses" et "fruitées" caractérisées par des esters et de longues chaînes carbonées.

Article publié (2021)

Rugard M, Jaylet T, Taboureau O, Tromelin A, Audouze K (2021). "Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages." Plos One 16(5) e0252486 DOI 10.1371/journal.pone.0252486
Received: January 27, 2021
Accepted: May 15, 2021
Published: May 28, 2021

Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages

Marylène Rugard[1], Thomas Jaylet[1], Olivier Taboureau[2], Anne Tromelin[3], Karine Audouze [1]*
1 T3S, Inserm UMR S-1124, Universite´ de Paris, Paris, France,
2 Inserm U1133, CNRS UMR 8251,
Universite´ de Paris, Paris, France,
3 Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon,
CNRS, INRAE, Universite´ Bourgogne Franche-Comte´, Dijon, France

* karine.audouze@u-paris.fr

Abstract

This study aims to highlight the relationships between the structure of smell compounds and their odors. For this purpose, heterogeneous data sources were screened, and 6038 odorant compounds and their known associated odors (162 odor notes) were compiled, each individual molecule being represented with a set of 1024 structural fingerprint. Several dimensional reduction techniques (PCA, MDS, t-SNE and UMAP) with two clustering methods (k-means and agglomerative hierarchical clustering AHC) were assessed based on the calculated fingerprints. The combination of UMAP with k-means and AHC methods allowed to obtain a good representativeness of odors by clusters, as well as the best visualization of the proximity of odorants on the basis of their molecular structures. The presence or absence of molecular substructures has been calculated on odorant in order to link chemical groups to odors. The results of this analysis bring out some associations for both the odor notes and the chemical structures of the molecules such as "woody" and "spicy" notes with allylic and bicyclic structures, "balsamic" notes with unsaturated rings, both "sulfurous" and "citrus" with aldehydes, alcohols, carboxylic acids, amines and sulfur compounds, and "oily", "fatty" and "fruity" characterized by esters and with long carbon chains. Overall, the use of UMAP associated to clustering is a promising method to suggest hypotheses on the odorant structure-odor relationships.

Introduction

Odorant molecules are largely used in food, cosmetic and perfumes [1, 2]. Moreover, the extra-nasally expression of ORs receptors suggest their potential therapeutic interest [3].

The olfactory system can discriminate a large range of odorants of different shapes, sizes, and chemical functions [4]. The discriminatory capacity is carried out through various processes. The olfactory perception begins at the olfactory epithelium level with the activation of olfactory receptors (ORs) by the binding of odorants. The ORs are mainly expressed in olfactory cilia of the sensory olfactory neurons (OSNs); the activation of ORs triggers the transmission of signals by the

OSNs to the olfactory bulb before to be distributed to other regions of the brain such as the piriform cortex [5–8].

There are currently about 7000 odorant molecules reported [9], while number of odors able to be perceived is currently unknown, but could reach 1 trillion [10]. Besides, there are less than 2000 of functional ORs in mammals as a whole (about 400 in Human) [11]. Hence, the olfactory perception and discrimination of a such huge number of odors by a limited number of functional ORs is due to involving a combinatorial coding. The combinatorial coding is based on the fact that a single odorant is recognized by several receptors and that a single odorant receptor recognizes several odorants. So, the odor quality of different odorants are encoded by different combinations of receptors [12, 13].

Obtaining a reliable description of the odors by the overall sensory is complicated as emotional context has been reported to be very strongly associated with olfactory information [14, 15]. Indeed, studies of brain activity have shown that exposure to olfactory stimuli activates some brain structures of the limbic system linked to emotions, learning and memory. Hence, odors are difficult to describe verbally, and the words used depend on the context, the familiarities with odor, and culture-specific experiences [16, 17]. Nevertheless, the verbal description of odor remains a main way to characterize the olfactory biological activity of odorants in Human [18, 19]. According to a medicinal chemistry approach of odor perception, matching ligands to ORs are critical for understanding the olfactory system. Indeed, olfactory receptor deorphanization should aid to understand how the molecular properties of odorant molecules act on the receptor activation. However, ligands have been published for nearly 10% of the approximately 400 functional human ORs [20, 21]. Because of the difficulty to deorphanize the ORs by experiment, *in silico* approaches are a promising way, as well by ligand (odorants) approaches as by target (ORs) approaches. Assuming that odorants detected by the same OR have related structures [22], several studies have been carried out to explore relationships between the structure of odorants and their receptors by creating different models using different approaches such as Quantitative Structure–Activity Relationship (QSAR) [23], neural networks [24] and docking [25]. For example, previous studies have developed predictive models based on neural networks for camphoraceous and fruity odors [26], or using artificial intelligence [27], for example by combining fuzzy logic with Kohonen neural networks [28–30]. These hybrid methods have shown their ability to establish robust structure-odor relationships models on different series of molecules, allowing a clustering of the odors for a set of test molecules with a prediction rate of over 70%. The study of several ORs were also performed through mutagenesis, molecular modelling, and functional expression and led to identify the structure of binding site, improving the knowledge of structure-functions relationships of the ORs [25, 31–33]. Other strategies were to develop integrative systems biology based-models using existing knowledge such as ligand-protein associations and protein-protein interactions in order to

decipher the human odorome [34]. Nevertheless, despite significant advances, establishing the link between odors and molecular structures remains largely unresolved and challenging [35–37]. Our study focuses on the relationship between the structures of a large set of smell compounds and their odors. For this purpose, we built a dataset comprising more than 6000 smell compounds associated with their smell description by compiling information available in several databases [38, 39]. The structural information of the molecules was encoded into fingerprints, and a computational study aiming to analyze and visualize the smell compounds distribution in their chemical space was performed. Four-dimensional reduction techniques combined with two clustering methods were tested in order to select the most suitable approach for the present dataset. Two classical dimensional reduction methods, Principal Component Analysis (PCA) and Multidimensional Scaling (MDS), and two more recent approaches, the t-Distributed Stochastic Neighbor Embedding (t-SNE) [40], and the Uniform Manifold Approximation and Projection (UMAP) [41] were chosen. After data reduction, clustering analyses were performed individually either by k-means or by agglomerative hierarchical clustering (AHC) using the 2-dimensional space coordinates defined by each dimension reduction techniques. Then, an analysis of the distribution of odor notes and chemical functions / molecular substructures represented in the different clusters was performed. The association of the UMAP method with clustering appeared to be a relevant combination to discriminate the relationships between the structures of molecules and their odors.

Materials and methods

Data of smell compounds, odor notes and ORs

For this study, a dataset of 6038 smell compounds and 162 odor notes (of which “odorless”) having at least 5 occurrences [42] was extracted and compiled from the databases "The Good Scents Company" (access 23/01/19) [39] and "Flavor Base" (9th Edition) [38]. Data can be available upon request.

Encoding molecular structures into fingerprints

Each molecular structure was encoded into Extended-connectivity fingerprints (ECFP), i.e. in binary vector: the presence of a given function/substructure in the compound is represented by 1, while its absence is represented by 0 [43, 44]. In these fingerprints, substructures are generated by considering each atom and their neighborhood on several circular layers (up to a given diameter/radius). To calculate them, KNIME software (v 3.6.2) was used with the following parameters: radius = 2, allowing to obtain fingerprints equivalent to Extended-connectivity fingerprints 4 (ECFP4). ECFPs are fingerprints specially developed to seizure molecular features

necessary to molecular activity and particularly suited to Tanimoto similarity methods [45]. More specifically, ECFP4 are known for their efficiency [45, 46], and are among the best on small molecules benchmarks [47]. The use of bits number = 1024 associated to these fingerprints, makes it possible to obtain a fairly precise molecular structure for the study [48].

In addition, sixty-two molecular substructures of the smell compounds were computed with KNIME in the aim to identify potential relations between the odor notes and the chemical functional groups of molecules.

Dimension reduction from the 1024-bit fingerprints

To visualize the encoded smell compounds, four-dimensional reduction techniques were applied: PCA, MDS, t-SNE and UMAP. The PCA, MDS and t-SNE methods were performed using the R software (v 4.0.2) using several packages such as FactoMineR, stats and labdsv, while the UMAP method (v 0.4) [49] was applied using Python 3.7.6 with the package 'umaplearn'. The PCA is a multivariate analysis method, that allows to extract the most important information by an orthogonal transformation to generate correlated variables with new linearly independent variables called principal components [50]. MDS is a network localization technique that maps the similarity or the dissimilarity of pairs of objects from a dataset. The similarity/dissimilarity is converted into distances between points in a two-dimensional space [51, 52]. The t-SNE method is an improved version of the Stochastic Neighbor Embedding (SNE). Like the SNE, the t-SNE measures the similarity between pairs of objects of the high dimensional data and of the two-dimensional embedding. Therefore the t-SNE generates a two-dimensional embedding using gradient descent to minimize the Kullback-Liebler divergence between the vector of similarities between pairs of objects in the high dimensional data and the similarities between pairs of objects in the embedding [53]. The UMAP method is a recently developed dimension reduction technique [41] that allows to precisely capture the non-linear structure of large data sets. UMAP is a manifold technique constructed from a theoretical framework based on Riemannian geometry and algebraic topology. The manifold theory considered the following key concept: a manifold is a space where points are gathered according to their Euclidean distances, so forming a continuous map. According the proximities between the points, differentiable manifolds can be identified on the map [54]. Then, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations to construct a topological representation of the high dimensional data. Given some low dimensional representation of the data, an equivalent topological representation can be built using a similar process. Then, the layout of the data representation is optimized in the low dimensional space allowing minimization of the cross-entropy between the two topological representations [49]. Globally, UMAP starts by calculating

the distances between each point. It then considers, for each point, its n closest neighbors and assigns a weight (probability) of link between the point considered and its n neighbors. From this, UMAP builds a weighted graph and then uses a “force-based layout” algorithm on it to project and represent data optimally at low dimensions. The UMAP advantage is that it is customizable in order to be adapted to its own data. For example, several parameters can be modified after data integration: the distance calculation method, the number of the neighbors, the minimum distance for grouping points at low dimensions, or the desired number of dimensions. To calculate distances/similarities between fingerprints, three metrics seem to be the most suitable: the Tanimoto/Jaccard index, the Dice index and the Cosine coefficient [55]. The Jaccard/Tanimoto index, which represents the fraction of bits shared between 2 fingerprints, gave the best results in preliminary assays on our data, and was therefore selected for our study.

The objective of our visualization was to obtain a compromise between local and global information, in order to suitably perceive the emergence of groups containing structurally close molecules. For that, the appropriate choice of the values of the number of neighbors and the minimum distance is crucial. The “number of neighbors” parameter allows to balance local versus global structure in the data. So, at low values, UMAP concentrates on very local structure even to the detriment of the global picture. But at higher values, UMAP focuses on larger neighborhoods of each point but loses detail structure. The “minimum distance” parameter controls the distance with which the points are grouped together in the low dimensional representation. At low values, there are clumpier embeddings between the points and at higher values, the points are much less grouped and UMAP preserves more the broad topological structure. After testing of different values of these two parameters, the number of neighbors and the minimum distance were fixed to 15 and 0, respectively.

Visualization, clustering and structure-odor analysis

AHC and k-means clustering were carried out from the reduced dimensions obtained with the four previous techniques, to group structurally similar molecules. These clustering were done using R (v 4.0.2) on the two dimensional data to avoid problems associated with high-dimensional clustering [56]. For AHC, Euclidean distance matrix 2 to 2 of each molecule was calculated with the aggregative criterion "ward.D2", which seeks to minimize the intra-class inertia and maximize the inter-class inertia. The two closest classes were thus successively grouped until obtaining a complete clustering tree. Hierarchical clustering is simple and easy to use whatever the form of similarity or distance [57, 58]. This technique has great flexibility with regard to a level of granularity, and is applicable to any attribute types [58]. However, the merging of clusters is definitive. Therefore it is not possible to

correct erroneous decisions [57]. For the k-means clustering, several numbers of centroids corresponding to the numbers

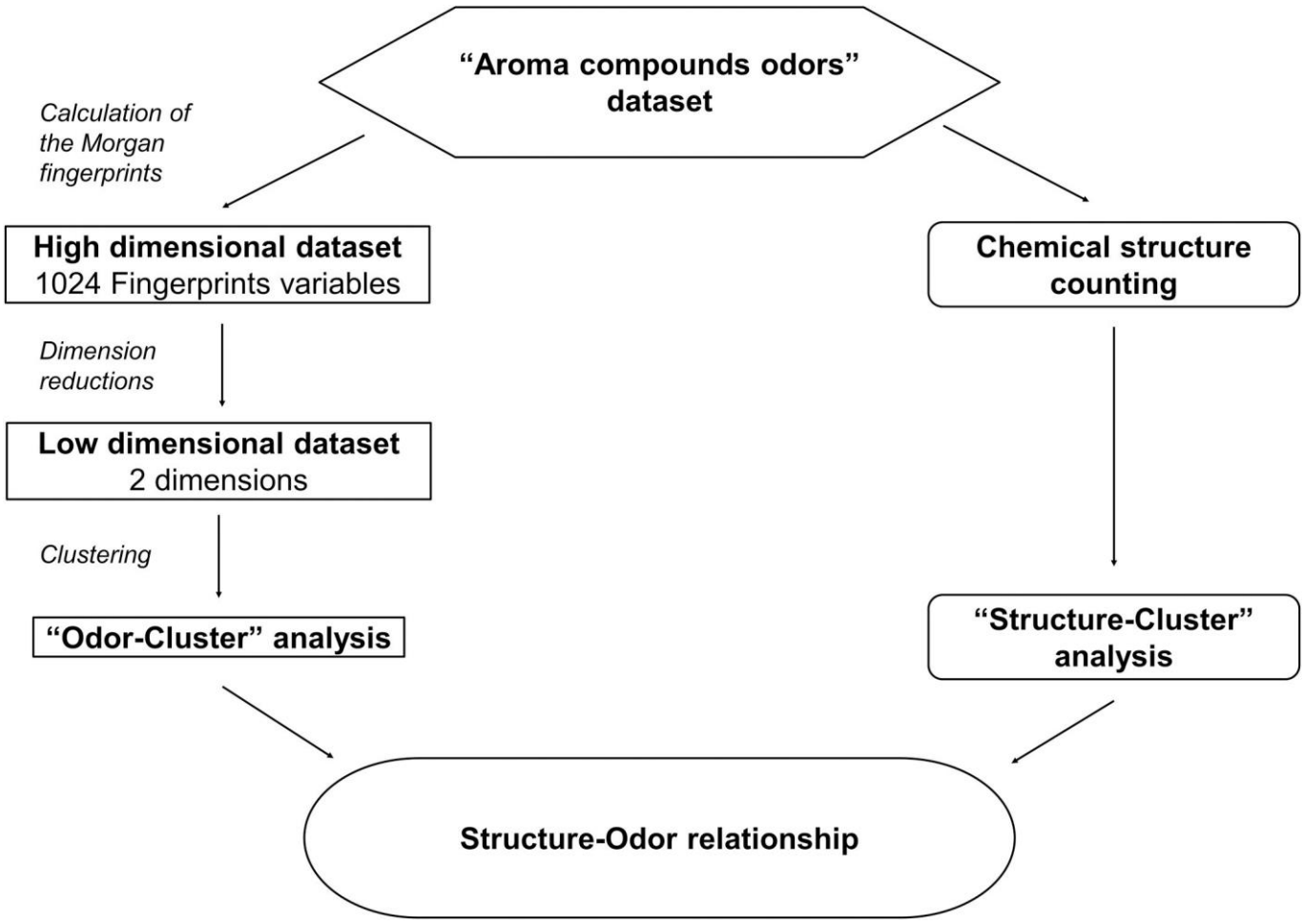


Fig 1. Representation of the workflow. On the left, reduction of the high dimensional space defined by the fingerprints and clustering; on the right, molecular substructures calculation.

https://doi.org/10.1371/journal.pone.0252486.g001

of clusters were tested and the points of the dataset were assigned to its nearest cluster at each iteration. All points of the same cluster were averaged and new centroids were recalculated. Cluster centroids were improved at each iteration until there is no more changes [59]. The kmeans algorithm is known to be sensitive to outliers, and less efficient with clusters that are not hyper-spheres [57]. To choose the optimal number of clusters, the intra-cluster variability was analyzed. The aim was to have a low intra-cluster variability to obtain homogeneous groups, but high enough so that the population within each cluster is sufficient.

Once the clusters were defined, either by AHC or k-means, we first looked into the distribution of odor notes across the clusters. Then the chemical groups/functions of the molecules belonging to the different clusters were investigated. The overall setting up protocol is described in Fig 1.

Results

Overview analysis of the dataset of odorant compounds

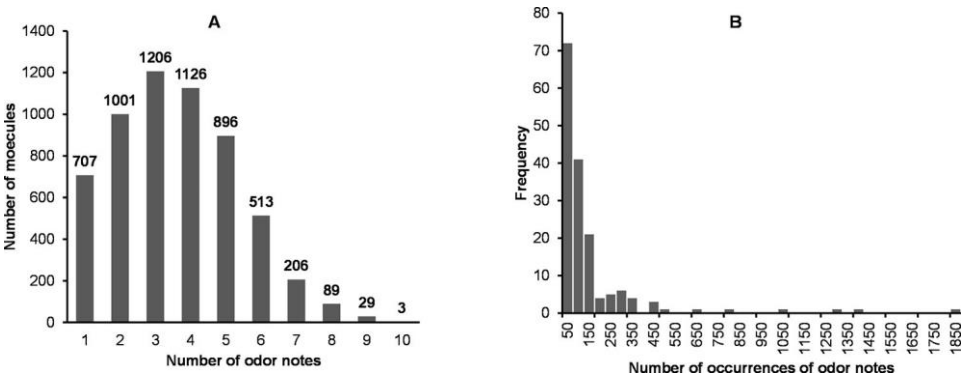


Fig 2. Distribution of the odor notes and the number of their occurrences. A: Histogram of the number of odorants according to the number of odor notes. B: Histogram of the workforce according to the number of occurrences of the odorants.

https://doi.org/10.1371/journal.pone.0252486.g002

The dataset encompasses 6038 smell compounds, of which mainly odorants, but also various smell compounds, whose sapid compounds or additives (inorganic salts, amino-acids, peptides, polymers...). Such molecules have little or no volatility, and consequently are unable to reach in vapor phase to the nasal cavity to activate the ORs. Therefore, these compounds are described "odorless". We identified 261 compounds with these characteristics. Excluding the "odorless" compounds, most odorants are described by 2 to 5 odor notes (Fig 2A). The number of occurrences of the odor notes ranges from 1828 ("fruity") to 5 ("bland" and "tallow"). Most of the odor notes have less than 150 occurrences (Fig 2B), while only 4 odor notes exceed 1000 occurrences (1828 for "fruity", 1389 for "green", 1283 for "sweet", 1010 for "floral").

Dimensions reduction, clustering and visualization of the data

The high-dimensional data provided by the 1024 calculated fingerprints, used to encode the molecular structures of the smell compounds, were reduced to two-dimensional data using four dimensional reduction techniques. Then, the two clustering methods were applied to these 2D space coordinates to group the most similar molecules according to their structure. To determine the optimal number of clusters, an "elbow" curve representing the intra-cluster variability as a function of the number of clusters was done (S1 Fig) for each dimensional reduction technique. As the elbow curve showed a variable optimal number of clusters, a Kelley penalty score was used in addition to precisely determine the optimal number of clusters (S2 Fig). The minimum score is attributed to the optimal number of clusters, which was five clusters for the t-SNE and four clusters for the three other techniques. For our study, the clustering calculations were carried out, following these numbers of

clusters, to have a good balance between variability and number of individuals per group. The assignment of smell compounds in the 2-two-dimensional space defined by the calculation of all techniques is shown in Fig 3.

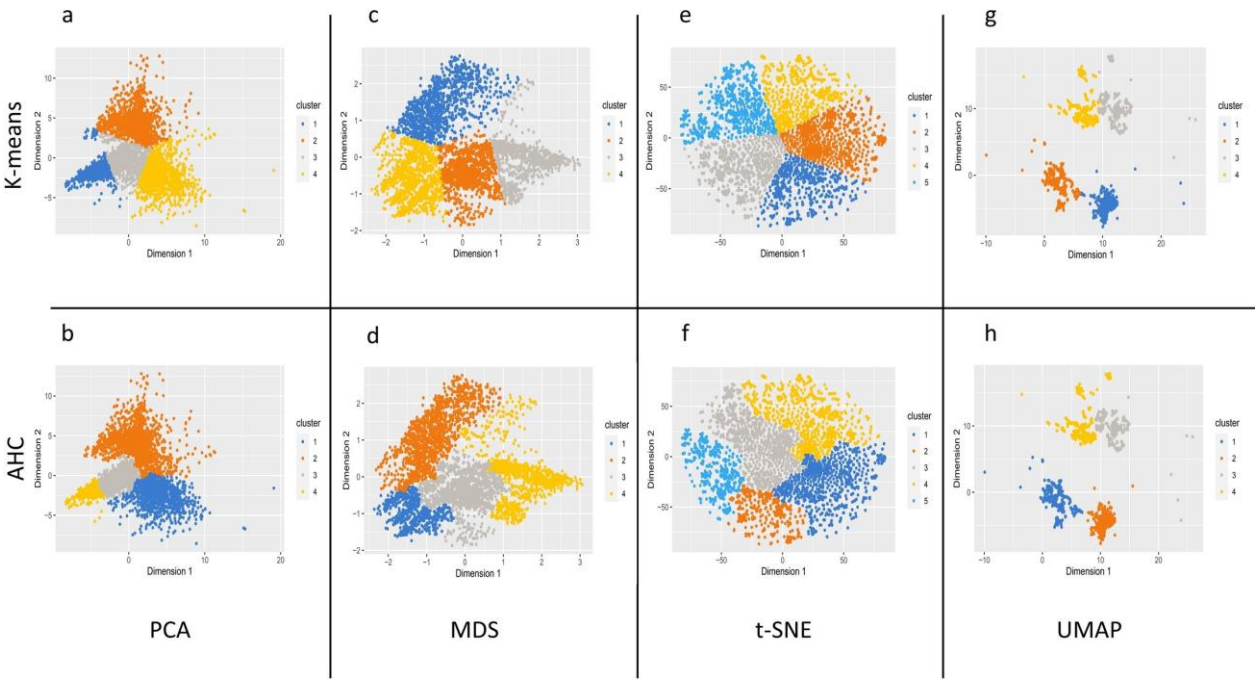


Fig 3. Visualization of the compounds-odors dataset in the 2-two dimensional spaces obtained after dimension reduction using PCA, MDS, t-SNE and UMAP. The data are colored according to the clusters produced by the k-means clustering and AHC that were carried out on the basis of the coordinate in the 2D spaces. The colors allow only to visualize the clusters easily and are specific to each method; there is no correspondence between the colors according to the several methods. The data are reported in S1 Table. (a) Clusters obtained by the PCA k-means approach: the clusters C1a, C2a, C3a and C4a encompass respectively 1523, 1466, 1622 and 1427 smell compounds; (b) Clusters obtained by PCA AHC approach: the clusters C1b, C2b, C3b and C4b encompass respectively 1461, 1756, 1997 and 824 smell compounds; (c) Clusters obtained by MDS k-means approach: the clusters C1c, C2c, C3c and C4c encompass respectively 1312, 1774, 1468 and 1484 smell compounds; (d) Clusters obtained by MDS AHC approach: the clusters C1d, C2d, C3d and C4d encompass respectively 854, 1551, 1970 and 1663 smell compounds; (e) Clusters obtained by tSNE k-means approach: the clusters C1e, C2e, C3e, C4e and C5e encompass respectively 1008, 1375, 1225, 1122 and 1308 smell compounds; (f) Clusters obtained by tSNE AHC approach: the clusters C1f, C2f, C3f, C4f and C5f encompass respectively 1480, 636, 1633, 1524 and 765 smell compounds; (g) Clusters obtained by UMAP kmeans approach: the clusters C1g, C2g, C3g and C4g encompass respectively 1597, 1344, 1454 and 1643 smell compounds; (h) Clusters obtained by UMAP AHC approach: the clusters C1h, C2h, C3h and C4h encompass respectively 1640, 1584, 1332 and 1482 smell compounds. In each chart, C1, C2, C3, C4 and C5 clusters are depicted respectively in blue, orange, grey, yellow and light blue.

https://doi.org/10.1371/journal.pone.0252486.g003

The projection of the PCA, MDS and t-SNE maps did not shown a clear separation. Instead, the UMAP technique revealed a good separation of the four groups. The color representation of the compounds by clusters displayed well defined areas using the four-dimensional reduction techniques. Nevertheless, the areas defined by each of the clustering methods were not identical when applied to a same dimensional reduction approach. Indeed, each of the two clustering methods could separate differently the 2D-spaces. Thus, to assess the homogeneity of clustering between the 2 clustering methods, intersection of two clusters were computed using the following equation:

Cx(M k-means)∩Cy(M AHC)

where x and y were cluster numbers, M referred to the dimensional reduction methods, and \ was the mathematical intersection operator. In other words, it measured the number of molecules that belonged to two clusters obtained with the two clustering methods. For example, C1a(PCA k-means)\C1b(PCA AHC) encompassed 16 common molecules. In addition, the dendrograms from each of the four AHC studies allowed to determine which clusters were aggregated, and thus which clusters were closer (S3 Fig). With PCA-AHC, clusters 1 and 2, and clusters 3 and 4 aggregated. About the MDS-AHC, clusters 4 and 2, and clusters 1 and 3 merged. The t-SNE-AHC technique showed that clusters 3 and 5 aggregated together, and then with cluster 4 in one side whereas clusters 1 and 2 joined in the other side. Finally, on the UMAP-AHC, clusters 1 and 2 were aggregated, as well as clusters 3 and 4.

Analysis of the cluster constituents: structure-odor relationships

Odor notes

We performed the analysis of the cluster composition considering two viewpoints: the frequencies of the odor notes carried by the smell compounds, and the number of molecules carrying specific odor notes. More precisely, because the number of occurrences of the odors varied in a large range from 5 to 1828, the direct comparison of the number of occurrences would not be reliable for the less frequent odor notes. Therefore, we considered two ratios (S2 Table):

$$\%\ odor\ notes = \%ON\ = \frac{\text{number of occurrences of an odor note in the cluster}}{\text{total number of occurrences of this odor}}$$

$$\%\ odorant\ molecules = \%OM = \frac{\text{number of occurrences of an odor in the cluster}}{\text{number of elements (molecules) in this cluster}}$$

For example, with the PCA-kmeans approach, there were 1523 molecules in the cluster C1. The most frequent odor note “fruity” had 1828 occurrences in the dataset and 691 in C1:

$$\%ON \text{ "fruity"} = {}^{691}/_{1828} = 37.8\%$$
$$\%OM \text{ "fruity"} = {}^{691}/_{1523} = 45.4\%$$

Besides “beefy” had 20 occurrences in the dataset, and 3 in C1:

$$\%ON \text{ "beefy"} = {}^{3}/_{20} = 15.0\%$$
$$\%OM \text{ "beefy"} = {}^{3}/_{1523} = 0.2\%$$

Thus, about 38% of “fruity” molecules were gathered in C1 and constituted 45% of this cluster. Part of the “beefy” molecules (3 odorants) were in C1 representing only 0.2% of this cluster. All the frequency values were reported in S2 Table.

To compare the effectiveness of the used techniques to discriminate the odors, radar charts were performed (Fig 4), based on the distribution of the 17 most frequent odor notes across the clusters.

These charts revealed the specificity of several odor notes according to the obtained clusters for each of the dimensional reduction methods. An overview of the specificity of odor notes was summarized by the calculation of the number of odor notes for which %ON is higher than 50. The result is displayed in Fig 5, and showed the greatest discriminant capacity of UMAP whatever the clustering method.

The analysis of the %ON values obtained for the 17 most frequent odor notes provided interesting findings. For clarity, we focused our results on UMAP, the results from the others methods being described in supplementary (S2 Table).

The clusters C1g(UMAP k-means) and C2h(UMAP AHC) were constituted of more than 60% of “balsamic” odor note, as well as “floral”, “spicy”, nutty” and “sweet” notes. Similar profiles were also observed for C2g and C1h (“woody” and “spicy” notes), C3g and C3h (“odorless”, “sulfurous”, “citrus”), and C4g and C4h (“fatty”, “waxy”, “fruity”, “green”). By combining the clusters C1g and C2h, C2g and C1h, C3g and C3h, C4g and C4h and called respectively C1g2h, C2g1h, C3hg and C4hg, the odor notes “woody” and to a lesser degree “spicy” were typical for the molecules belonging to cluster C2g1h (S4B Fig). About 66% of the occurrence of the “woody” note was gathered in cluster C2g1h while “woody” molecules represented about 26% of this cluster. About C2g1h, although it contained about 30 different odors notes, more than 90% of the molecules carried the odor notes “sandalwood” and “cedar” (S2 Table). Additionally, 54 molecules of C2g1h carried both the odor notes “woody” and “spicy” constituted near to 10% of this cluster. “Spicy” note was more frequent in

the cluster C1g2h (representing 12% of this cluster). However, “balsamic” was the odor the most represented in C1g2h (66%, S4A Fig). Besides, “nutty” (%ON 54%), “floral” (%ON 39%) and “sweet” (%ON 35%) notes were specifically more frequent in C1g2h comparing to the three other clusters.

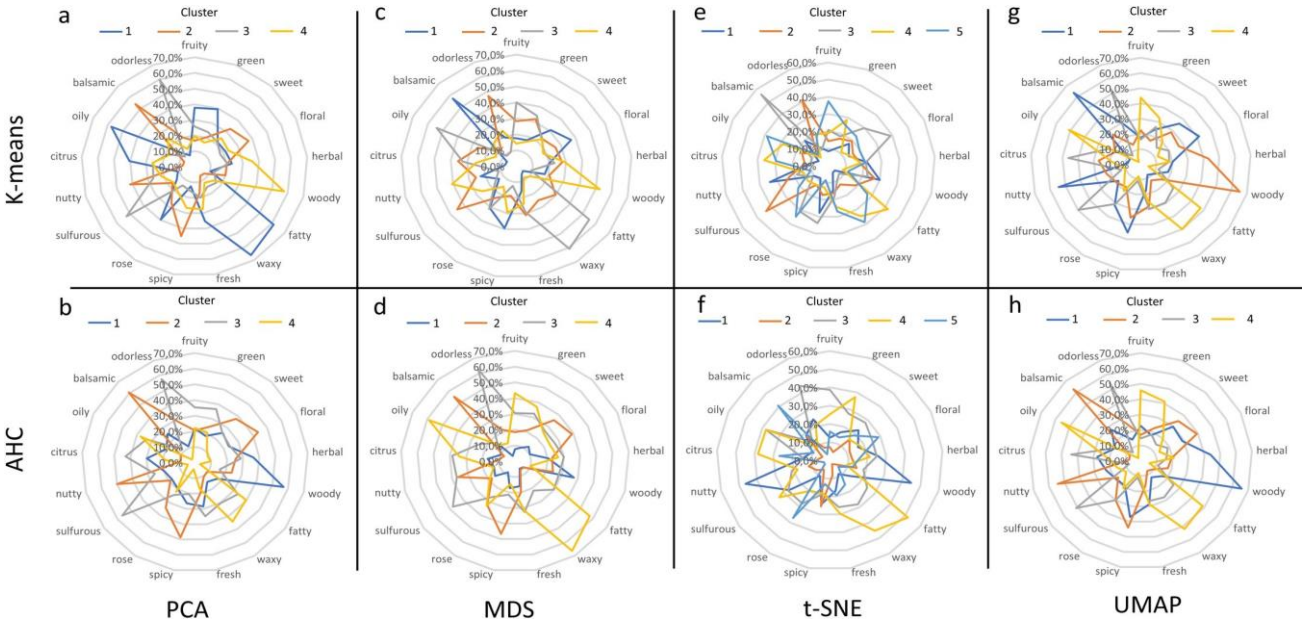


Fig 4. Radar charts of the distribution of the %ON values obtained for the 17 most frequent odor notes across the clusters. (a) Clusters obtained by PCA k-means method; (b) Clusters obtained by PCA-AHC method; (c) Clusters obtained by MDS k-means method; (d) Clusters obtained by MDS-AHC method; (e) Clusters obtained by t-SNE k-means method; (f) Clusters obtained by t-SNE-AHC method; (g) Clusters obtained by UMAP k-means method; (h) Clusters obtained by UMAP-AHC method. In each chart, C1, C2, C3, C4 and C5 clusters are depicted respectively in blue, in orange, in grey, in yellow, in light blue.

https://doi.org/10.1371/journal.pone.0252486.g004

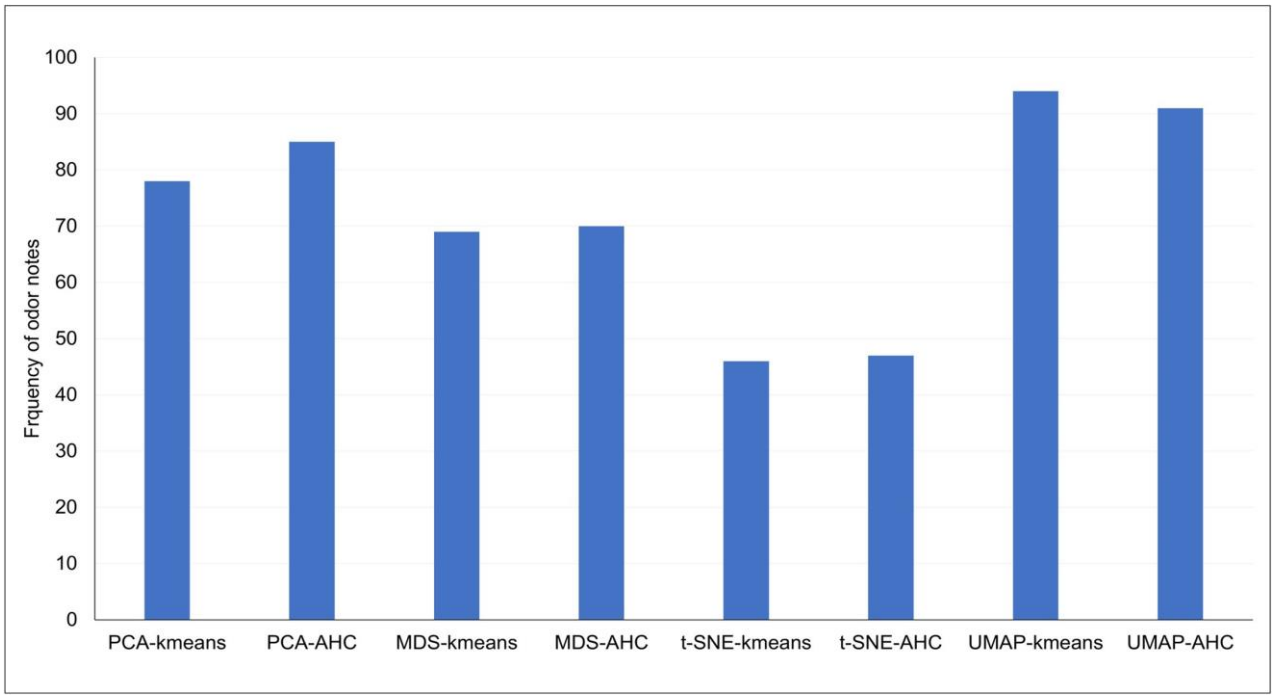


Fig 5. Histogram of the number of odor notes whose %ON is greater than 50 for each technique.

https://doi.org/10.1371/journal.pone.0252486.g005

The cluster C3hg (S4C Fig) put together "sulfurous" and "citrus" odor notes (%ON 51 and 44% respectively). In addition, more than 60% of the occurrences of "mustard", "garlic", "onion" and "alliaceous" were in this cluster, whereas "bergamot", "lemon", "orange", "mandarin" were also well represented (S2 Table). Additionally, there was about 100 odorless compounds in C3hg. Finally, the odor notes "oily", "waxy", "fatty", "fruity" and "green" bring together the main part of their occurrences in cluster C4gh (S4D Fig). The "fruity" molecules represented 57% of the cluster C4gh while "fruity" was often associated to another odor note in the odor description, especially to "green" (21%), and also to "apple" (11%). There were also some "fruity-fatty" and "fruity-waxy" associations (5 to 8%).

As presented above, several molecules could belong to intersections between two clusters, noted Cx(UMAP k-means)\Cy(UMAP AHC). Several of these overlapping clusters corresponded to similar areas of the 2D-spaces, and the belonging molecules were sharing the same odor notes. At the difference, some clusters parts were placed far from the main area of the other elements related to the same cluster. The composition of clusters calculated on the basis of UMAP coordinates were particularly well maintained across k-means and AHC clustering methods. Only 238 molecules were switched to another cluster. The areas C1g \ C1h, C1g \ C3h, C3g \ C4h, C4g \ C3h included respectively 44, 15, 158 and 21 molecules. C3g gathered more than "sulfurous" and "odorless"

molecules, while the molecules belonging to C4h were characterized by “fruity”, green”, “waxy” and “fatty” notes. The group C3g\C4h contained nor “sulfurous” nor “odorless” molecule. In opposite, “green” molecules constituted almost three quarters of this group, while “fruity” was shared by more than one third of the molecules. C1g \ C1h shared more than one third of molecules with the odor “floral” and the odor “sweet”. For the C3g \ C4h area, a large majority of molecules carried the odor “green” (123 molecules). And the area C4g \ C3h encompassed 11 molecules with the fruity odor. It was therefore the molecules carrying the "green" odor which mainly change cluster depending on the clustering method. Results from the others methods were discussed on S1 File.

Chemical structures and functions of odorants

Among the 62 chemical structures and functions of different nature shared by the smell compounds of the dataset, we selected eighteen chemical functional groups present in at least 5% of the molecules of one of the 4 clusters (S3 Table). By focusing on these eighteen chemical structures and functions, we explored their frequency depending on the clusters to which they belong. As shown in Figs 6 and 7, carbonyl compounds were predominantly present in all clusters, that was the majority of odorant molecules have carbonyl groups, and the cluster 4 owned the higher percentage (80%), mainly as ester functions. Aldehydes and alcohols were mainly in cluster 3, as well as carboxylic acids, aliphatic amines, thiols and sulfides.

Molecules having an allylic group were especially frequent in clusters 1 and 4 (45%), and to a lesser extent in cluster 3. Moreover, the cluster C1 was especially rich in bicyclic structures. The cluster C4 was characterized by molecules with long carbon chains without ramifications (60%). Conversely, the cluster C2 was lacking in allyl groups, but was remarkably rich in unsaturated rings (phenols, aryl-methyl groups, aromatic amines and alcohols, furans) while molecules belonging to other clusters were deficient in such chemical groups.

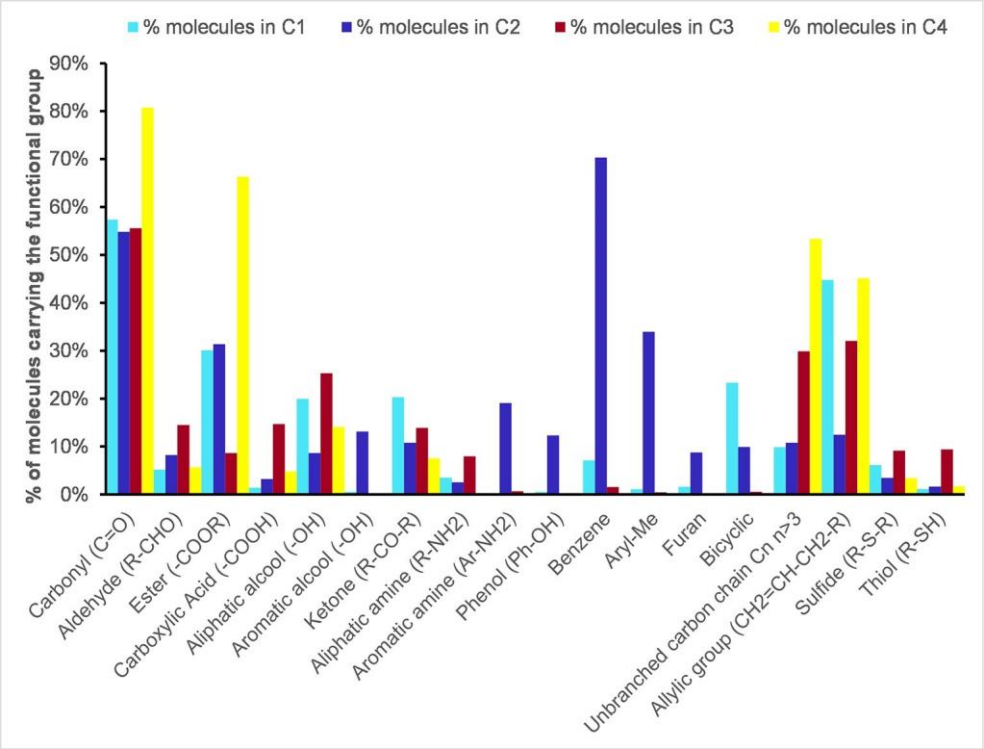


Fig 6. Histogram of the distribution of the chemical functional groups according the clusters. Only the structures present in at least 5% of the molecules of one of the 4 clusters C1, C2, C3 and C4 are represented: C1 in light blue; C2 in dark blue; C3 in dark red; C4 in yellow.

https://doi.org/10.1371/journal.pone.0252486.g006

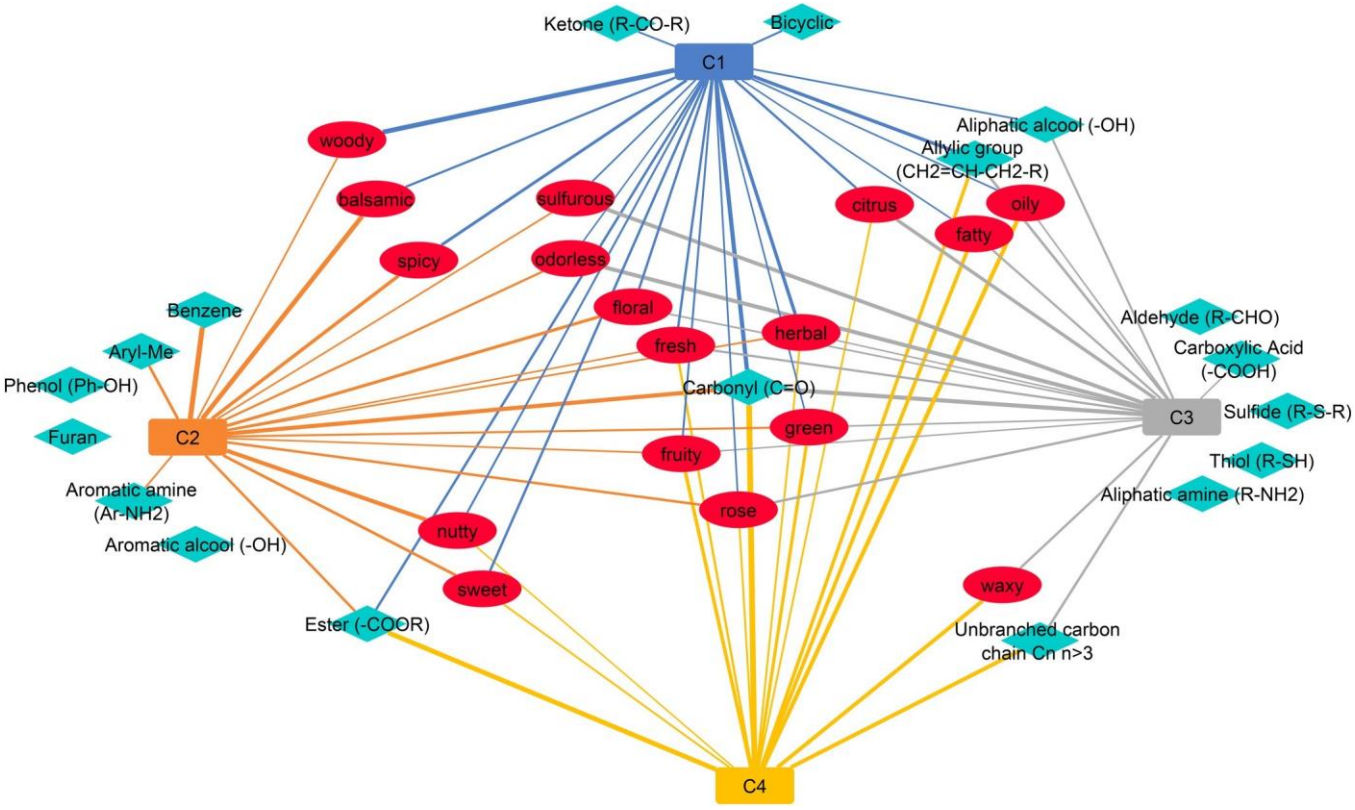


Fig 7. Network representation of the links between odor notes (red ellipse) and chemical functional groups (blue diamond). The nature of the line varies as a function of the relative frequency of occurrences. The thicker the line, the higher is the number of occurrences of an

odor note or a chemical functional group within the cluster to which it is linked. The edges are invisibly for the relative frequency of occurrences less than 0.1. The blue, orange, grey and yellow rectangles correspond respectively to clusters 1, 2, 3 and 4. The blue lines correspond to the associations between the cluster 1 and the odor notes or the cluster 1 and the chemical functional groups. The orange lines correspond to the associations between the cluster 2 and the odor notes or the cluster 2 and the chemical functional groups. The grey lines correspond to the associations between the cluster 3 and the odor notes or the cluster 3 and the chemical functional groups. The yellow lines correspond to the associations between the cluster 4 and the odor notes or the cluster 4 and the chemical functional groups.

https://doi.org/10.1371/journal.pone.0252486.g007

Discussion

Odor structure relationships in olfaction are key elements in understanding the olfactory system, an area in which there is still a great lack of knowledge [35–37].

With the aim to highlight the links between the molecular structure of smell compounds and their odor notes, we assessed four-dimensional reduction techniques applied to the molecular structures of 6038 smell compounds encoded by 1024- bit fingerprints. The spreading of smell compounds in a two-dimensional space was thus obtained for each technique. The coordinates were then used, independently, to perform a k-means and a AHC clustering, therefore providing the distribution of the smell compounds among several clusters. The visualization of the data in 2D spaces (Fig 3) showed the various areas defined by the clustering calculations, that allowed to evaluate the performance of the eight used approaches (reduction combined to clustering) to establish reliable links between molecular structures and odor notes (Figs 3–5). The less significant results were obtained using the t-SNE, as well concerning the blurred spatial arrangement of the elements in the 2D-space than the overlapping of clustering partitions obtained by k-means and AHC. The MDS and PCA calculations provided better but average results, except for PCA-AHC for which results were a slightly better. All the results and analyses put forward the precision of UMAP in aggregations of the elements according to the cluster areas that were reflected by the high degree of specificity of odor notes regarding the clusters. Indeed, as UMAP is based on the fact that manifold structure exists in the data, UMAP calculation is able to find these structures in the noise of a dataset which is suitable for data visualization. As the amount of data sampled increases, the amount of structure highlighted by noise lower [49]; therefore, the robustness of UMAP increases with the amount of data. Lastly, UMAP has the advantage of preserving the local and the global data structure, by keeping a runtime shorter than other dimension reduction techniques [60].

The characteristics of smell compounds across the UMAP clusters were examined on two points of view: the odor notes and the chemical functional groups. Analyzing the proportions of odor notes across the clusters focused on the 17 most frequent odor notes, including “odorless” quality. In parallel, 18 chemical functional groups were used to point out the main chemical features of the smell compounds. This dual approach revealed interesting specificities of the molecules according to

the cluster to which they belong. The radar charts reported in Fig 4G and 4H and in S4 Fig bring out very distinct odor profiles. Few molecules of the combined clusters C2g1h shared the odor note “woody” and are characterized by allylic chains and carbonyl and ketone chemical functions. We noted that nearly 50 molecules carried both the odor notes “woody” and “spicy”; for example, copaene (woody; spicy), thujopsene (woody; spicy; dry), isocaryophyllene (woody; spicy), which are polycyclic molecules. The odor note “spicy” was rather frequent in C2g1h, and “balsamic” was the major odor note of C1g2h while the cyclic and aromatic moieties were a distinctiveness of the molecules of C1g2h. Interestingly, the bicyclic molecules were specific to some molecules of C1g2h and C2g1h, and quite absent from the clusters C3gh and C4gh. The odor notes “nutty” and “floral”, as well as “rose”, were also characteristic of molecules of C1g2h (S2 Table). Taking together the observations related to the clusters C1g2h and C2g1h, these suggested that two types of “spicy” molecules could be discriminated both by their perception and their structures: the "spicy-woody" and the "spicybalsamic" molecules.

The cluster C3gh is peculiar in that “sulfurous” and citrus” molecules were mixed whereas “sulfurous” and “citrus” odors evoke opposing hedonic values unpleasant/pleasant [61]. C3gh is characterized by its composition on aldehydes, aliphatic alcohols and amines, carboxylic acids, and obviously organic sulfur molecules that share the sulfurous, sulfur and pungent odors. At the difference, there were very few esters. We can also note that odorless compounds that contribute to C3gh are amino acids, carboxylic acids and their salts. If excluding the effect of sulfur atom on the odor of “sulfurous” molecules, some structural features common to the carbon chains of “sulfurous” and “citrus” molecules could explain their grouping in C3gh. Further accurate examinations of the chemical structures will be needed to address this issue. The molecules that belonged to C4gh have “fruity”, “green”, “fatty” and “waxy” odor notes. As shown in a previous work [62] these odor notes were often used together in the descriptions of natural fruity odors of esters while long chains confer fatty and waxy odors. Indeed, about 50% of molecules of C4gh shared allylic or aliphatic chains, and ester function. Besides the odor “fruity” was frequently associated to “green” or “apple” in the odor descriptions, and less frequently to “fatty” or “waxy”. Obviously, no odor notes or chemical structure were specific to a cluster, which was not surprising, but it was still possible to associate certain chemical structures with certain odors (S4 Table). It could not be expected to adjust in only four groups the complexity of many thousands of odorants and several millions of perceptible odors [63]. Moreover, most molecules were described by 3 or 4 odor notes (Fig 2A), meaning that there exist “spicy-woody” and “spicy-balsamic”, “fruity-green” and “fruity-fatty” molecules, and numerous other cases [62], and that these odors can be discriminated by humans. Such associations of odors notes will be considered in a further work.

To conclude, the obtained results highlight some relationships between the structure of the molecule and odor. The UMAP dimensional reduction method associated to k-means and AHC clustering techniques allowed to obtain interesting results revealing links between molecular structures and odor qualities. Such association of k-means and AHC clustering with UMAP is the first performed on molecular fingerprints for a dataset related to odors. Therefore, the use of UMAP provides a promising way to improve the understanding of the structure-odor relationships by visualizing high quality embedding of large data sets that were previously unattainable [49]. Upcoming studies would be considered to refine the odor-structure relationships inside specific group by applying other clustering methods as Maximum Common Substructure Methods or Gaussian mixture model [64, 65]. In perspective, it would be interesting to integrate olfactory receptors on which odorant molecules interact to, in order to demonstrate structure-odor-receptor relationships. In addition, conducting this study using a 3-D dimensional reduction could provide complementary information on the structureodor relationships as an extension of the present study.

Supporting information

S1 Table. Fingerprint, coordinates in 2D spaces and clusters.
(XLSX)

S2 Table. Odor notes and occurrences.
(XLSX)

S3 Table. Distribution of the chemical groups and functions by cluster.
(DOCX)

S4 Table. Table of chemical structures associated with odors. (DOCX)

S1 Fig. "Elbow" curve. Representation of intra-cluster variability as a function of the number of clusters. The optimal number of clusters is around the bend of the curve. (DOCX)

S2 Fig. Progression of the penalty score according to the number of clusters. The minimum score is assigned to the optimal number of clusters. (DOCX)

S3 Fig. Dendrograms of the AHC of molecules for each dimension reduction technique. (DOCX)

S4 Fig. Radar charts of the distribution of the %ON values obtained for the 17 most frequent odor notes across clusters of the UMAP-kmeans and UMAP-AHC techniques. A: Comparison between C1g and C2h. B: Comparison between C2g and C1h. C: Comparison between C3g and C3h. D: Comparison between C1g and C2h. (DOCX)

S1 File.
(DOCX)

Acknowledgments

We would like to thank University of Paris, Inserm, as well as Dr. Thierry Thomas-Danguin for helpful discussions, and Dr. Elisabeth Guichard for her support.

Author Contributions

Conceptualization: Olivier Taboureau, Karine Audouze.

Data curation: Marylène Rugard.

Formal analysis: Marylène Rugard, Anne Tromelin.

Funding acquisition: Anne Tromelin.

Methodology: Thomas Jaylet.

Software: Thomas Jaylet.

Supervision: Karine Audouze.

Writing – original draft: Marylène Rugard, Thomas Jaylet, Karine Audouze.

Writing – review & editing: Olivier Taboureau, Anne Tromelin.

References

1. Braga A, Guerreiro C, Belo I. Generation of Flavors and Fragrances Through Biotransformation and De Novo Synthesis. Food Bioprocess Technol. 2018 Dec; 11(12):2217–28.
2. Armanino N, Charpentier J, Flachsmann F, Goeke A, Liniger M, Kraft P. What's Hot, What's Not: The Trends of the Past 20 Years in the Chemistry of Odorants. Angew Chem Int Ed Engl. 2020 Sep 14; 59 (38):16310–44. https://doi.org/10.1002/anie.202005719 PMID: 32453472
3. Lee S-J, Depoortere I, Hatt H. Therapeutic potential of ectopic olfactory and taste receptors. Nat Rev Drug Discov. 2019 Feb; 18(2):116–38. https://doi.org/10.1038/s41573-018-0002-3 PMID: 30504792
4. Kini A, Firestein S. The Molecular Basis of Olfaction. CHIMIA International Journal for Chemistry. 2001;453–9.
5. Buck LB. Information coding in the vertebrate olfactory system. Annu Rev Neurosci. 1996; 19:517–44. https://doi.org/10.1146/annurev.ne.19.030196.002505 PMID: 8833453
6. Firestein S. How the olfactory system makes sense of scents. Nature. 2001 Sep 13; 413(6852):211–8. https://doi.org/10.1038/35093026 PMID: 11557990
7. Lledo P-M, Gheusi G, Vincent J-D. Information processing in the mammalian olfactory system. Physiol Rev. 2005 Jan; 85(1):281–317. https://doi.org/10.1152/physrev.00008.2004 PMID: 15618482
8. Shipley MT, Ennis M, Puche AC. The Olfactory System. In: Conn PM, editor. Neuroscience in Medicine. Totowa, NJ: Humana Press; 2003. p. 579–93.
9. Dinu V, MacCalman T, Yang N, Adams GG, Yakubov GE, Harding SE, et al. Probing the effect of aroma compounds on the hydrodynamic properties of mucin glycoproteins. Eur Biophys J. 2020 Dec; 49(8):799–808. https://doi.org/10.1007/s00249-020-01475-4 PMID: 33185715
10. Bushdid C, Magnasco MO, Vosshall LB, Keller A. Humans Can Discriminate More than 1 Trillion Olfactory Stimuli. Science. 2014 Mar 21; 343(6177):1370–2. https://doi.org/10.1126/science.1249168 PMID: 24653035
11. Tromelin A. Odour perception: A review of an intricate signalling pathway: Olfactory system and odour perception. Flavour Fragr J. 2016 Mar; 31(2):107–19.
12. Malnic B, Hirono J, Sato T, Buck LB. Combinatorial receptor codes for odors. Cell. 1999 Mar 5; 96

(5):713–23. https://doi.org/10.1016/s0092-8674(00)80581-4 PMID: 10089886

13. Touhara K. Odor discrimination by G protein-coupled olfactory receptors. Microsc Res Tech. 2002 Aug 1; 58(3):135–41. https://doi.org/10.1002/jemt.10131 PMID: 12203691

14. Hamakawa M, Okamoto T. The effect of different emotional states on olfactory perception: A preliminary study. Flavour and Fragrance Journal. 2018; 33(6):420–7.

15. Ferdenzi C, Roberts SC, Schirmer A, Delplanque S, Cekic S, Porcherot C, et al. Variability of affective responses to odors: culture, gender, and olfactory knowledge. Chem Senses. 2013 Feb; 38(2):175–86. https://doi.org/10.1093/chemse/bjs083 PMID: 23196070

16. de Araujo IE, Rolls ET, Velazco MI, Margot C, Cayeux I. Cognitive modulation of olfactory processing. Neuron. 2005 May 19; 46(4):671–9. https://doi.org/10.1016/j.neuron.2005.04.021 PMID: 15944134

17. Meierhenrich UJ, Golebiowski J, Fernandez X, Cabrol-Bass D. The molecular basis of olfactory chemoreception. Angew Chem Int Ed Engl. 2004 Dec 3; 43(47):6410–2. https://doi.org/10.1002/anie. 200462322 PMID: 15578781

18. Poivet E, Tahirova N, Peterlin Z, Xu L, Zou D-J, Acree T, et al. Functional odor classification through a medicinal chemistry approach. Sci Adv. 2018; 4(2):eaao6086. https://doi.org/10.1126/sciadv.aao6086 PMID: 29487905

19. Poivet E, Peterlin Z, Tahirova N, Xu L, Altomare C, Paria A, et al. Applying medicinal chemistry strategies to understand odorant discrimination. Nat Commun. 2016 Apr 4; 7:11157. https://doi.org/10.1038/ ncomms11157 PMID: 27040654

20. Mainland JD, Li YR, Zhou T, Liu WLL, Matsunami H. Human olfactory receptor responses to odorants. Sci Data. 2015; 2:150002. https://doi.org/10.1038/sdata.2015.2 PMID: 25977809

21. Peterlin Z, Firestein S, Rogers ME. The state of the art of odorant receptor deorphanization: A report from the orphanage. Journal of General Physiology. 2014 May 1; 143(5):527–42. https://doi.org/10. 1085/jgp.201311151 PMID: 24733839

22. Malnic B, Godfrey PA, Buck LB. The human olfactory receptor gene family. Proc Natl Acad Sci U S A. 2004 Feb 24; 101(8):2584–9. https://doi.org/10.1073/pnas.0307882100 PMID: 14983052

23. Gabler S, Soelter J, Hussain T, Sachse S, Schmuker M. Physicochemical vs. Vibrational Descriptors for Prediction of Odor Receptor Responses. Mol Inform. 2013 Oct; 32(9–10):855–65. https://doi.org/10. 1002/minf.201300037 PMID: 27480237

24. Schmuker M, de Bruyne M, Ha¨hnel M, Schneider G. Predicting olfactory receptor neuron responses from odorant structure. Chem Cent J. 2007 May 4; 1:11. https://doi.org/10.1186/1752-153X-1-11 PMID: 17880742

25. Schmiedeberg K, Shirokova E, Weber H-P, Schilling B, Meyerhof W, Krautwurst D. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. J Struct Biol. 2007 Sep; 159(3):400–12. https://doi.org/10.1016/j.jsb.2007.04.013 PMID: 17601748

26. Chastrette M, Cretin D, el Aïdi C. Structure-odor relationships: using neural networks in the estimation of camphoraceous or fruity odors and olfactory thresholds of aliphatic alcohols. J Chem Inf Comput Sci. 1996 Feb; 36(1):108–13. https://doi.org/10.1021/ci950154b PMID: 8576286

27. Lo¨tsch J, Kringel D, Hummel T. Machine Learning in Human Olfactory Research. Chem Senses. 2019 Jan 1; 44(1):11–22. https://doi.org/10.1093/chemse/bjy067 PMID: 30371751

28. Audouze K, Ros F, Pintore M, Chre´tien JR. Prediction of odours of aliphatic alcohols and carbonylated compounds using fuzzy partition and self organising maps (SOM). Analusis. 2000 Sep; 28(7):625–32.

29. Pintore M, Audouze K, Ros F, Chre´tien J. Adaptive fuzzy partition in database mining: application to olfaction. Data Sci J. 2002; 1:99–110.

30. Ros F, Audouze K, Pintore M, Chre´tien JR. Hybrid systems for virtual screening: interest of fuzzy clustering applied to olfaction. SAR QSAR Environ Res. 2000; 11(3–4):281–300. https://doi.org/10.1080/ 10629360008033236 PMID: 10969876

31. Behrens M, Briand L, de March CA, Matsunami H, Yamashita A, Meyerhof W, et al. Structure–Function Relationships of Olfactory and Taste Receptors. Chemical Senses. 2018 Feb 2; 43(2):81–7. https://doi. org/10.1093/chemse/bjx083 PMID: 29342245

32. Charlier L, Topin J, Ronin C, Kim S-K, Goddard WA, Efremov R, et al. How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case. Cell Mol Life Sci. 2012 Dec; 69(24):4205–13. https://doi.org/10.1007/s00018-012-1116-0 PMID: 22926438

33. Launay G, Te´letche´a S, Wade F, Pajot-Augy E, Gibrat J-F, Sanz G. Automatic modeling of mammalian olfactory receptors and docking of odorants. Protein Eng Des Sel. 2012 Aug; 25(8):377–86. https://doi. org/10.1093/protein/gzs037 PMID: 22691703

34. Audouze K, Tromelin A, Le Bon AM, Belloir C, Petersen RK, Kristiansen K, et al. Identification of odorant-receptor interactions by global mapping of the human odorome. PLoS One. 2014; 9(4):e93037. https://doi.org/10.1371/journal.pone.0093037 PMID: 24695519

35. Licon CC, Bosc G, Sabri M, Mantel M, Fournel A, Bushdid C, et al. Chemical features mining provides new descriptive structure-odor relationships. PLoS Comput Biol. 2019; 15(4):e1006945. https://doi.org/ 10.1371/journal.pcbi.1006945 PMID: 31022180

36. Sell CS. The Relationship Between Molecular Structure and Odour. In: Chemistry and the Sense of Smell. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2014. p. 388–419.

37. Genva M, Kenne Kemene T, Deleu M, Lins L, Fauconnier M-L. Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure? Int J Mol Sci. 2019 Jun 20; 20(12). https://doi.org/10.3390/ ijms20123018 PMID: 31226833

38. Leffingwell & Associates. Flavor-Base. 9th Edition. Available online: http://www.leffingwell.com/ flavbase.htm.

39. The Good Scents Company, Available online: http://www.thegoodscentscompany.com/.

40. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(11).

41. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. JOSS. 2018 Sep 2; 3(29):861.

42. Zarzo M, Stanton DT. Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. Attention, Perception & Psychophysics. 2009 Feb 1; 71(2):225–47. https://doi.org/10.3758/APP.71.2.225 PMID: 19304614

43. Glem RC, Bender A, Arnby CH, Carlsson L, Boyer S, Smith J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs. 2006 Mar; 9(3):199–204. PMID: 16523386

44. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J Chem Doc. 1965 May; 5(2):107–13.

45. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24; 50(5):742–54. https://doi.org/10.1021/ci100050t PMID: 20426451

46. O'Boyle NM, Sayle RA. Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminform. 2016; 8:36. https://doi.org/10.1186/s13321-016-0148-0 PMID: 27382417

47. Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform. 2020 Jun 12; 12(1):43. https://doi.org/10.1186/s13321-02000445-4 PMID: 33431010

48. Knime [Internet]. Available from: http://www.knime.com

49. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat] [Internet]. 2020 Sep 17 [cited 2020 Nov 2]; Available from: http:// arxiv.org/abs/1802.03426

50. Abdi H, Williams LJ. Principal component analysis: Principal component analysis. WIREs Comp Stat. 2010 Jul; 2(4):433–59.

51. Saeed N, Nam H, Al-Naffouri TY, Alouini M-S. A State-of-the-Art Survey on Multidimensional ScalingBased Localization Techniques. IEEE Commun Surv Tutorials. 2019; 21(4):3565–83.

52. Borg I. Applied multidimensional scaling and unfolding. Springer; 2018.

53. Arora S, Hu W, Kothari P. An Analysis of the t-SNE Algorithm for Data Visualization. Proceedings of Machine Learning Research. 2018;1455–62.

54. Abraham R, Marsden JE, Ratiu T. Manifolds, Tensor Analysis, and Applications. New York, NY: Springer New York; 1988. (Marsden JE, Sirovich L, John F, editors. Applied Mathematical Sciences; vol. 75).

55. Bajusz D, Ra´cz A, He´berger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform. 2015; 7:20. https://doi.org/10.1186/s13321-015-0069-3 PMID: 26052348

56. Oskolkov N. tSNE vs. UMAP: Global Structure. Medium. 2020; Available from: https:// towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

57. Kaushik M, Mathur B. Comparative study of K-means and hierarchical clustering techniques. International journal of software and hardware research in engineering. 2014; 2(6):93–8.

58. Abbas OA. Comparisons between data clustering algorithms. International Arab Journal of Information Technology. 2008; 5(3).

59. Ordonez C. Clustering binary data streams with K-means. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery—DMKD '03 [Internet]. San Diego, California: ACM Press; 2003 [cited 2021 Mar 28]. p. 12. Available from: http://portal.acm.org/ citation.cfm?doid=882082.882087

60. Becht E, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Evaluation of UMAP as an alternative to t-SNE for single-cell data [Internet]. Bioinformatics; 2018 Apr [cited 2020 Nov 2]. Available from: http://biorxiv.org/lookup/doi/10.1101/298430

61. Khan RM, Luk C-H, Flinker A, Aggarwal A, Lapid H, Haddad R, et al. Predicting Odor Pleasantness from Odorant Structure: Pleasantness as a Reflection of the Physical World. Journal of Neuroscience. 2007 Sep 12; 27(37):10015–23. https://doi.org/10.1523/JNEUROSCI.1158-07.2007 PMID: 17855616

62. Tromelin A, Chabanet C, Audouze K, Koensgen F, Guichard E. Multivariate statistical analysis of a large odorants database aimed at revealing similarities and links between odorants and odors. Flavour Fragr J. 2018 Jan; 33(1):106–26.

63. Kermen F, Chakirian A, Sezille C, Joussain P, Le Goff G, Ziessel A, et al. Molecular complexity determines the number of olfactory notes and the pleasantness of smells. Sci Rep. 2011; 1:206. https://doi. org/10.1038/srep00206 PMID: 22355721

64. Stahl M, Mauser H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. J Chem Inf Model. 2005 May; 45(3):542–8. https://doi.org/10.1021/ci050011h PMID: 15921444

65. Li X, Luo D, Cheng Y, Wong K-Y, Hung K. Identifying the Primary Odor Perception Descriptors by MultiOutput Linear Regression Models. Applied Sciences. 2021 Apr 7; 11(8):3320.

II. Combinaison d'approches de classification et de pharmacophores pour comprendre les perceptions olfactives homogènes, et applications à un accord aromatique et à un masquage

La qualité odorante d'une molécule détermine son identité en tant que composé odorant, du moins lorsqu'elle est seule. Dans le cas d'un mélange de molécules odorantes, la perception peut être hétérogène ou homogène. La perception est qualifiée d'hétérogène lorsque les odeurs des composants d'un mélange peuvent être perçues (Berglund et al., 1976) chaque constituant du mélange conservant son identité odorante, et d'homogène lorsqu'une seule odeur est perçue à partir du mélange. Dans le cas d'une perception homogène, on distingue l'accord aromatique (lorsque l'odeur perçue est une nouvelle odeur, différente de celles des odorants du mélange) (Thomas-Danguin et al., 2014), et le masquage (lorsque l'odeur d'un seul des composants du mélange est reconnue) (Kay et al., 2005). Nous avons voulu mieux comprendre les mécanismes impliqués dans la perception homogène de mélanges de molécules odorantes. Nous nous sommes intéressés à un accord aromatique composé de six molécules (vanilline, acétate d'isoamyle, frambinone, acétate d'éthyle, beta-ionone, beta-damascenone), perçu avec une typicité de sirop de grenadine (Sinding et al., 2013), et un mélange binaire de deux molécules (whisky lactone et acétate d'isoamyle) constituant un masquage de la note fruitée de acétate d'isoamyle par la note boisée de la whisky-lactone (Atanasova et al., 2004). A partir des fingerprints calculés lors de l'étude précédente, nous avons effectué une réduction de dimension avec l'algorithme UMAP puis une classification par SOM sur les coordonnées des dimensions réduites permettant de définir des clusters spécifiques. Afin de déterminer si les molécules des mélanges partageaient des caractéristiques structurelles communes, nous avons réalisé une modélisation pharmacophore à partir d'une part, des composants des mélanges et d'autre part, à partir de sous-ensembles de molécules sélectionnées sur la base de leurs notes olfactives dans les clusters SOM contenant les composants des mélanges.

En ce qui concerne le masquage, des hypothèses pertinentes ont été obtenues suggérant que les composants du mélange pourraient avoir un ou plusieurs sites de liaison communs ; ceci suggère que la perception configurale du masquage serait au moins en partie établie au niveau périphérique.

A la différence de ce que nous avons observé dans le cas du masquage WL/IA, aucune hypothèse pertinente n'a été générée à partir des molécules constituant l'accord aromatique « grenadine ». Cela suggère que les composants de ce mélange n'ont probablement pas de sites de liaison commun au niveau périphérique, la perception configurale supposant alors une intégration du signal à des niveaux supérieurs du système olfactif (bulbe olfactif ou cerveau).

Manuscript 954590 submitted for publication in Frontiers in Ecology and Evolution, section Chemical Ecology (27/05/2022)

Combining classification and pharmacophore approaches to understand homogeneous olfactory perceptions, and applications in blending and masking mixtures

Marylène Rugard[1], Anne Tromelin*[2], Karine Audouze[1]

[1] Université de Paris Cité, T3S, Inserm UMR S-1124, F-75006 Paris, France

[2] Centre des Sciences du Goût et de l'Alimentation, Institut Agro Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France

*** Correspondence:**
Corresponding Author
anne.tromelin@inrae.fr

Keywords: UMAP, SOM, pharmacophores, olfaction, odorants, homogeneous perception

Abstract

The mechanisms involved in the homogeneous perception of odorant mixtures are largely unknown, and studying structure-odor relationships may be a key step in understanding homogeneous perception. With the aim of enhancing knowledge about blending and masking mixture perceptions, we combined classification and pharmacophore approaches. By collecting data from different sources, we built a large dataset that lists more than 5000 odorant molecules and their associated odors. Using the fingerprints that represent the structures of these molecules and uniform manifold approximation and projection (UMAP), dimension reduction was performed. The self-organizing map (SOM) classification was then carried out on the coordinates in the UMAP space, and as a result of the reduced dimensions, specific clusters were defined. A blended mixture (red cordial [RC] mixture: vanillin [V], isoamyl acetate [IA], frambinone [F], ethyl acetate [EA], beta-ionone [bI], beta-damascenone [bD]) and a masking mixture (whiskey-lactone [WL] and isoamyl acetate [IA]) were considered. Pharmacophore modeling (PHASE) was performed to examine whether common features were present for the molecules in each mixture. We used (i) subsets of the components from the aroma mixtures and (ii) subsets of a dozen molecules that were selected on the basis of their odor notes in the SOM clusters that contained the components of the mixtures. Relevant hypotheses were obtained from the subsets of the molecules

involved in masking WL-IA or those that shared the same odor notes of these components. Conversely, no relevant hypotheses were generated on the basis of the subsets constituted by the RC mixture components. Through pharmacophore comparisons, it was suggested that WL and IA could have common binding site(s); however, this is likely not the case for the RC mixture components. These results point out that a homogeneous perception can be formed by two different ways. The configural perception of WL-IA masking could take place at the peripheral level. Conversely, this is not the case for the RC mixture, in which configural perception might require signal integration at higher levels in the olfactory bulb and/or in the brain.

Introduction

The olfactory system is a complex system that allows the perception of many and varied odorant molecules (Kini and Firestein, 2001). The perception of odors involves several levels in a complex process that begins at peripheral receptors and ends in the brain (Murthy, 2011). In the first step, odor molecules bind to odor receptors, leading to the activation of the latter (Buck, 1996). This step is governed by the combinatorial code, implying that an odorant can activate several olfactory receptors and that an olfactory receptor can be activated by several different odorants (Malnic et al., 1999; Touhara, 2002). Deciphering this code remains a complicated challenge due to the presence of many orphan olfactory receptors (Peterlin et al., 2014; Mainland et al., 2015). In addition, the perception of mixed odors adds a layer of complexity to understanding the olfactory system.

The perception of odors is defined by several characteristics, including the quality, intensity and hedonic value (Berglund et al., 1976; Holley, 2006). The quality determines the identity of a scent, but in the case of an odorant mixture, the quality of the odors is expressed in a different way. Indeed, the perception of a mixture can be homogeneous or heterogeneous. A mixture is described as homogeneous when a single odor is distinguished from the mixture, and a heterogeneous mixture is when the odors of a mixture's components can be perceived (Berglund et al., 1976). In the case of a homogeneous perception, we can distinguish the blending mixture when the perceived odor is a new odor and is different from those of the odorants in the mixture (Thomas-Danguin et al., 2014), and overshadowing is when the odor of only one of the components of the mixture is recognized (Kay et al., 2005). Intensity is the strength of the perception of an odor and depends on the concentration of scent molecules.

Several studies have been carried out with the aim of better understanding the olfactory perception of odor mixtures by highlighting the molecule-receptor interactions that occur during the perception of mixtures (El Mountassir et al., 2016) or by highlighting the necessary characteristics of the odorant molecules in a mixture for perceiving the odors (Tromelin et al., 2020). However, since the ligands in the majority of receptors are still unknown, Quantitative-Structure-Activity Relationships (QSAR) methods are interesting and relevant in this type of study. Indeed, these methods make it possible to correlate biological activity with particular molecular properties (Tropsha and Wang, 2007) without needing to know the ligand–receptor complexes. Among the QSAR techniques, the use of pharmacophores has been demonstrated to be a very effective approach in the field (Gund, 1977; Yang, 2010). Indeed, a pharmacophore is defined as "*a set of structural features in a*

molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" (Wermuth et al., 1998; Leach et al., 2010).

In this study, we were interested in two particular mixtures of odorant molecules. The first mixture, constituting a blending mixture called red cordial (RC mixture), contains isoamyl acetate (IA), vanillin (V), frambinone (F), ethyl acetate (EA), beta-damascenone (bD) and beta-ionone (bI) (Le Berre et al., 2010; Sinding et al., 2013). The RC mixture is perceived as having a grenadine syrup odor. The second mixture is composed of isoamyl acetate (IA) and whiskey-lactone (WL) (Atanasova et al., 2004). This mixture, noted WL-IA, constitutes an overshadowing in which a qualitative dominance of the woody note of WL occurs when the perceived intensities of unmixed WL and IA are equal.

All of these molecules are part of a dataset that was used in a previous study (Rugard et al., 2021), and this dataset lists odorant molecules and their associated odors. With the aim of enhancing knowledge about blending and masking mixture perceptions, we combined classification and pharmacophore approaches.

By collecting data from different sources, we built a large dataset that lists 5665 odorant molecules with their associated odors. Using the fingerprints that represent the structure of these molecules, we performed a dimension reduction by uniform manifold approximation and projection (UMAP).

The combination of the UMAP technique (McInnes et al., 2018) and the k-means classification has been demonstrated to be the most relevant technique for discriminating structure-odor relationships (Rugard et al., 2021). For this reason, we again used the UMAP method in this study to obtain 2- and 3-dimensional data. In addition to the k-means classification method, we used the self-organizing map (SOM) classification (Kohonen, 1998), and the results provided by these two classifications were compared.

Pharmacophore modeling (PHASE) (Dixon et al., 2006) was performed to examine whether common features were present for the molecules in each mixture. The pharmacophore models allow us to determine the spatial disposition of the structural characteristics common to these molecules, which suggests possible binding to one or more common receptors. To this end, we used (i) subsets of the components from the aroma mixtures and (ii) subsets of a dozen molecules that were selected based on their odor notes in clusters containing the components of the mixtures. The purpose of this pharmacophore study was to determine the extent at which the key step in the configural processing of the homogeneous perception of odorant mixtures occurs at the peripheral level of the olfactory system.

Materials and Methods

Description of the dataset and molecules of interest

For this study, a dataset of 5665 smell compounds and 162 odor notes that occurred at least 5 times (Zarzo and Stanton, 2009) were extracted and compiled from the databases "The Good Scents Company" (The Good Scents Company) and "Flavor Base" (Leffingwell, 2013). Data are available upon request.

Within the database, we focused on several subsets that were defined by the classification methods and the odorant profile of the molecules. The list of the molecules of these subsets is detailed in the Results section and in Supplementary Table 1.

Table 1. List of the components of the RC mixture and the WL-IA binary mixture and their respective odor notes.

Name	CAS	Odor notes
Vanillin	8014-42-4	sweet; vanilla; creamy; chocolate
Isoamyl acetate	123-92-2	sweet; fruity; banana; solvent; pear
Frambinone	5471-51-2	sweet; berry; raspberry; ripe; floral; fruity
Ethyl acetate	141-78-6	ethereal; fruity; sweet; weedy; green; sharp; brandy; winey
beta-Damascenone	23696-85-7	fruity; floral; apple; plum; tea; rose; tobacco; natural; grape; raspberry; sweet
beta-Ionone	14901-07-6	floral; woody; sweet; fruity; berry; tropical; violet; raspberry, dry, powdery, orris
Whiskey lactone	39212-23-2	tonka; coumarinic; coconut; toasted; nutty; celery; burnt; woody; lactonic; maple*; lovage*

*Less than 5 occurrences in the database.

Fingerprint generation, dimension reduction and clustering

The structure of each molecule was converted with KNIME software (v 3.6.2) to extended-connectivity fingerprints (ECFP) of 1024 bits in the same way as performed in a previous study (Rugard et al., 2021).

The uniform manifold approximation and projection (UMAP) dimensional reduction technique was performed using the R software umapr package (version 4.1.1) (R Core Team, 2021) from the fingerprint values (McInnes et al., 2020).

The k-means and SOM (Kohonen Self-organizing map) (Kohonen, 1998) classifications were performed and based on the coordinates in the UMAP 2D and 3D spaces using R software (version 4.1.1) (R Core Team, 2021). The kohonen package was used for the SOM calculation (Wehrens and Buydens, 2007).

The visualization of the distribution of the molecules and clusters in the 2D-and 3D-UMAP spaces was performed using the package XLSTAT-3Dplot (Addinsoft, 2022).

Construction of pharmacophores

The pharmacophores were generated using PHASE Schrödinger software release 2021-4 (Dixon et al., 2006). All input molecules were defined as active using default settings. The values of the feature tolerance were 2 Å. All the default hypothesis settings were used except for the minimum number of features in the pharmacophore hypothesis, which was set at 3. The generation of conformers was performed by setting the target number of

conformers to 50 within an energy range of 21 kJ/mol. The maximum number of generated hypotheses was set to 10. The minimum number of features in the pharmacophore hypothesis was fixed to 3, and at least 50% of the active molecules had to match the hypothesis.

For our study, we used hydrogen bond acceptors (A features), hydrophobic (H features), and aromatic rings (R features). The pharmacophores were generated from all molecular conformers. The algorithm selected a conformer for each molecule to align the molecules based on the common features shared by several molecules, which thus constituted a candidate pharmacophore (van Drie, 2003). A score was then calculated and assigned to each pharmacophore to determine the best pharmacophore (Khedkar et al., 2007; Schaller et al., 2020).

Results

Overview of the dataset

The dataset includes 5665 odorants. Most odorants are described by 2 to 5 odor notes (Rugard et al., 2021). During this study, we were interested in two homogeneous mixtures, a blending mixture and a masking mixture, which were previously studied at the CSGA (Le Berre et al., 2010; Sinding et al., 2013) (Table 1).

The RC mixture is composed of vanillin (V), isoamyl acetate (IA), frambinone (F), ethyl acetate (EA), beta-damascenone (bD) and beta-ionone (bI) (Le Berre et al., 2010; Sinding et al., 2013) The experimental sensory tests have shown that each molecule has a distinct odor quality. Nevertheless, the odor of the mixture of these six components in specific proportions was judged to be similar to that of the Grenadine syrup (Le Berre et al., 2010). It has been established that the presence of isoamyl acetate, vanillin and, to a lesser degree, frambinone in the RC mixture is essential for the perception of this grenadine syrup smell (Sinding et al., 2013). In addition, the presence of the other three molecules (EA, bD, bI) typicality improves the grenadine syrup without being essential.

The masking mix is composed of whiskey-lactone (WL) and isoamyl acetate (IA) (Atanasova et al., 2004). It has been experimentally shown that when the perceived intensities of each unmixed compound were equal, the qualitative dominance was carried out by the woody note in the binary mixture.

Dimensions reduction and clustering

The molecular structures of the 5665 molecules were encoded by 1024 fingerprints, and the high-dimensional data provided were reduced to two- and three-dimensional data using the UMAP reduction method.

The clustering calculations were performed from coordinate values 2D and 3D of the molecules in the UMAP spaces using two classification algorithms, k-means and SOM.

With the k-means clustering, the partitioning of n observations into k clusters was achieved in which each observation belongs to the cluster with the nearest mean (cluster centroid). The k-means algorithm is a well-known method that is among the most popular clustering

techniques; nevertheless, the k-means algorithm is sensitive to outliers and is less efficient if clusters are not hyperspheres (Kaushik and Mathur, 2014).

SOM classification, which is less widely used, is a technique in which the visualization, exploration and clustering of large datasets is performed. The SOM maps high-dimensional data in an orderly fashion on a low-dimensional grid. The most important topology and metric information of the data is retained. The SOM creates a two-dimensional grid with observation models that are represented by nodes. The models are calculated with the aim of optimally describing the domain of observations. The most similar models are closer, and the dissimilar models are farther apart in the grid. In this sense, the SOM makes it possible to represent the similarity of data and to cluster data (Kohonen, 2013).

To refine the analysis, we carried out the classifications via k-means and SOM with 4, 9 and 16 clusters. We called the classification into 4, 9 and 16 clusters “classification at level L4”, “classification at level L9” and “classification at level L16”, respectively.

Classification at level L4

We applied the two clustering methods on both 2D and 3D-UMAP coordinates. In the two spaces, the elements are split into three main areas (Figure 1). We designated these areas Aa-2D, Ab-2D, Ac-2D, Aa-3D, Ab-3D, Ac-3D, and Ad-3D for the 2D and 3D spaces, respectively. The area Ad-3D concerns only 37 elements located at negative UMAP1-3D coordinates (X axis). The notations a, b, c, and d do not refer to a similar distribution of the elements in the related 2D and 3D areas but only to the positions in each of the two spaces (which is attributed according to increasing values first by the vertical axis and then by the X-axis). More precisely, 2561 elements of area Aa-2D are distributed mainly in Ac-3D (2530 elements) and among Ab-3D (19 elements) and Aa-3D (12 elements). Similarly, 1519 elements of the Ab-2D area (1573 elements) are located in area Aa-3D, while 37, 13 and 4 elements belong to Ad-3D, Ab-3D and Ac-3D, respectively. Finally, 1515 elements of Ac-2D are in Ab-3D, and the other elements are distributed between Ac-3D (9 elements) and Aa-3D (7 elements).

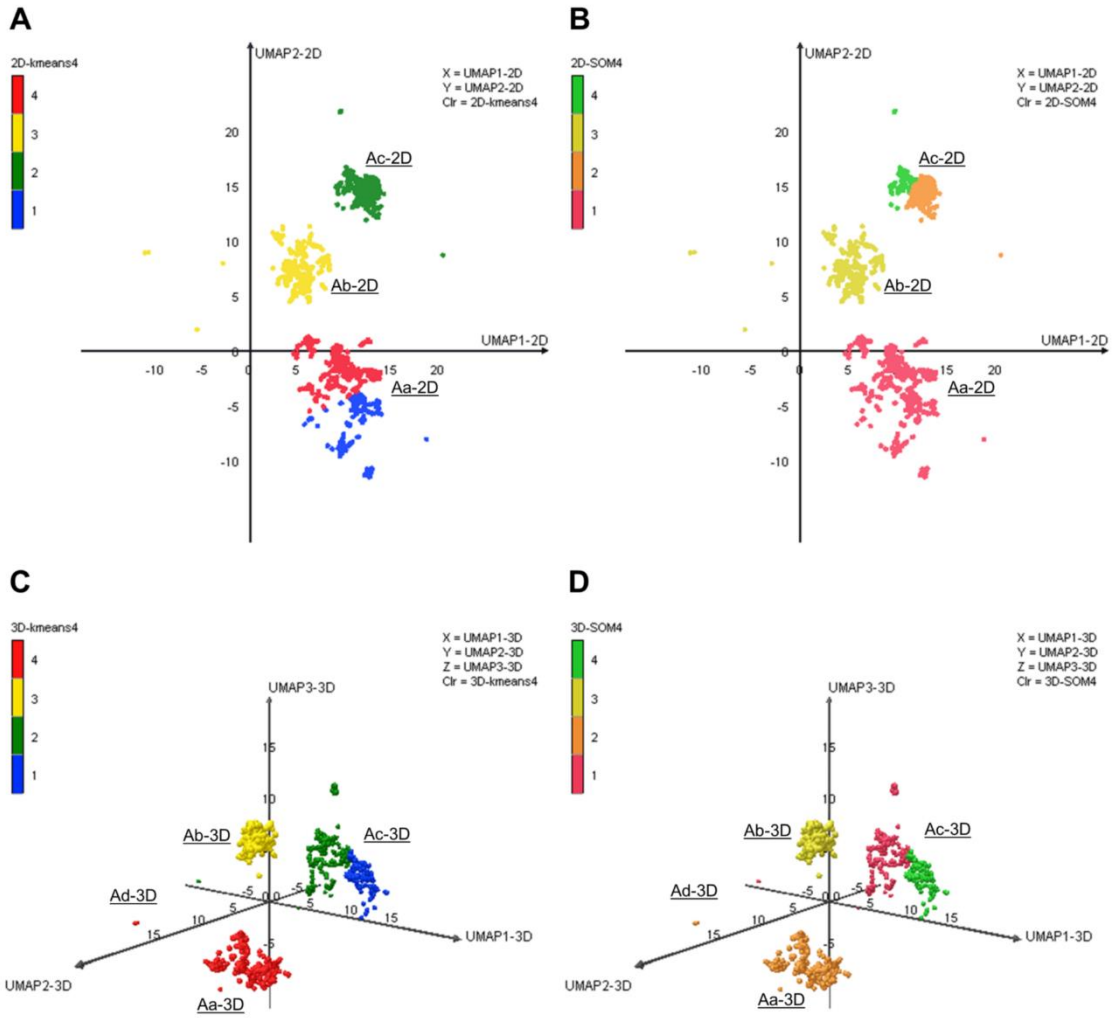


Figure 1. Visualizations of the compounds-odors dataset in the 2D and the 3D space obtained by dimension reduction using UMAP and k-means and SOM clustering. (A): k-means clustering of the 2D coordinates. (B): SOM clustering of the 2D coordinates. (C): k-means clustering of the 3D coordinates. (D): SOM clustering of the 3D coordinates.

For clarity, the clusters of each classification were named by their cluster number followed by the classification technique and then by the classification level. For example, in the UMAP 2D space, cluster 1 of the k-means classification with 4 clusters is named “2D-kmeans4-Cl-1”. A similar notation is used for clusters at levels L9 and L16.

In the UMAP-2D space, the elements of Aa-2D form the cluster 2D-SOM4-Cl-1 (2561 elements) but are split between 2D-kmeans4-Cl-1 (929 elements) and 2D-kmeans4-Cl-4 (1632 elements). Furthermore, the elements of Ac-2D gathered in 2D-kmeans4-Cl-2 (1531 elements) are split between 2D-SOM4-Cl-2 (1045 elements) and 2D-SOM4-Cl-4 (486 elements). Conversely, all elements of Ab-2D are gathered in clusters 2D-kmeans4-Cl-3 and 2D-SOM4-Cl-3, which are the only clusters identical by both k-means and SOM classifications (1573 elements).

In contrast, the distribution of the elements in the 3D-space clusters is similar for the clusters k-means and SOM. Indeed, the elements of clusters k-means4 and SOM4 located in areas Aa-3D and Ab-3D are the same. More precisely, 3D-kmeans4-Cl-3 and 3D-SOM4-Cl-3 match (1547 elements), while 3D-kmeans4-Cl-4 and 3D-SOM4-Cl-2 gather the same elements (1569). The area Ac-3D encloses (i) the clusters 3D-kmeans4-Cl-1 (932 elements) and 3D-kmeans4-Cl-2 (1617) and (ii) 3D-SOM4-Cl-1 (1612) and 3D-SOM4-Cl-4 (937 elements). Despite quite similar compositions, the clusters 3D-kmeans4-Cl-1 and 3D-SOM4-Cl-4, on the one hand, and 3D-kmeans4-Cl-2 and 3D-SOM4-Cl-1, on the other hand, are not strictly identical: five elements of 3D-kmeans4-Cl-2 belong to cluster 3D-SOM4-Cl-4.

At level L4, the elements of area Ad-3D are divided into two parts merged into 3D-Cl2-kmeans4 and 3D-Cl4-kmeans4 and 3D-SOM4-Cl-1 and 3D-SOM4-Cl-2.

The two classifications that are generated from 2D coordinates induce different distributions of the clusters, while the classifications made from 3D coordinates induce almost identical distributions. To avoid unnecessary complexity, we present the following analyses at levels L9 and L16 with 3D coordinates. The visualization of the clusters obtained at levels L9 and L16 from the 2D coordinates are displayed in Supplementary Figure 1.

Classification at level L9

The SOM classifications at levels L9 (a grid of 3 by 3) provide 7 clusters and two empty classes: 3D-SOM9-Cl-5 and 3D-SOM9-Cl-6.

The allotment of the elements across the clusters is displayed in Figure 2 A and 2 B.

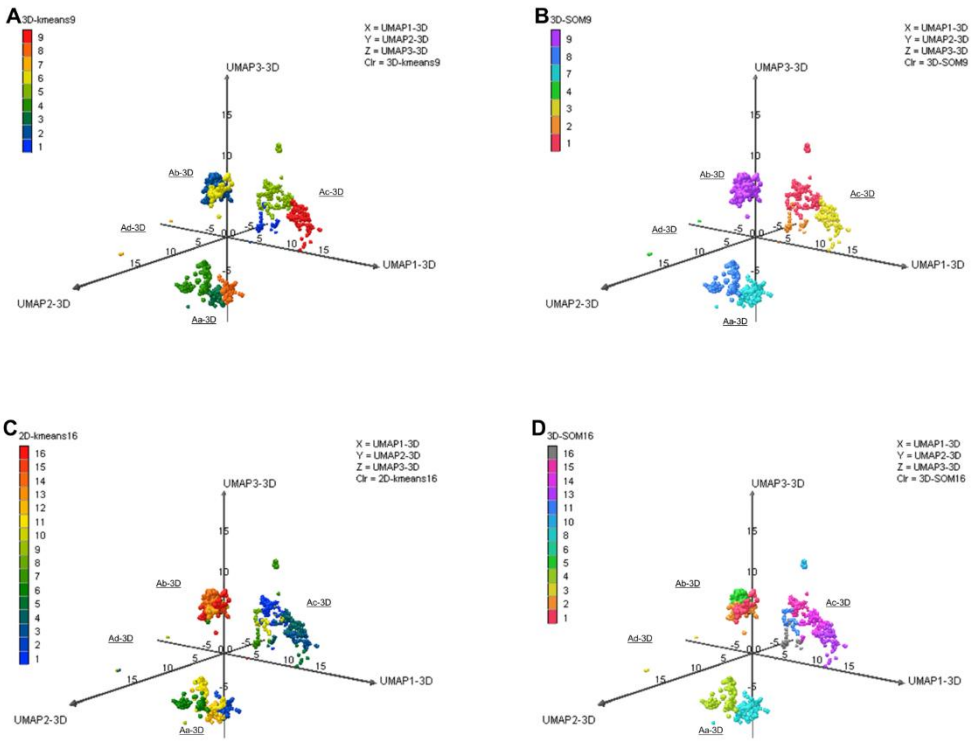


Figure 2. Visualizations of the compounds-odors dataset in the 3D space obtained using k-means and SOM clustering. At level L9 (A): 9 clusters k-means, (B): 7 clusters SOM. At level 16: (C) 16 clusters k-means, (D) 13 clusters SOM

The elements belonging to area Aa-3D are distributed between three clusters k-means and two clusters SOM: (i) 3D-kmeans9-Cl-3 (394 elements), 3D-kmeans9-Cl-4 (575 elements), 3D-kmeans9-Cl-8 (569 elements); and (ii) 3D-SOM9-Cl-7 (933 elements) and 3D-SOM9-Cl-8 (605 elements). Part of the cluster 3D-kmeans9-Cl-3 (30 elements) and the whole cluster 3D-kmeans9-Cl-4 correspond to the cluster 3D-SOM9-Cl-8. The other part of the cluster 3D-kmeans9-Cl-3 (364 elements) with the whole 3D-kmeans9-Cl-8 corresponds to the cluster 3D-SOM9-Cl-7.

The elements in the area Ab-3D are split between clusters 3D-kmeans9-Cl-2 (1029 elements) and 3D-Cl6-kmeans9 (518 elements) but constitute the cluster 3D-SOM9-Cl-9 (1547 elements).

Gathering the elements in the Ac-3D area involves three clusters by both k-means SOM classifications. The components of these clusters are quite similar but not identical. The elements belong (i) to the clusters 3D-kmeans9-Cl-1 (377 elements), 3D-kmeans9-Cl-5 (1249 elements) and 3D-kmeans9-Cl-9 (917 elements) and (ii) to the clusters 3D-SOM9-Cl-2 (377 elements), 3D-SOM9-Cl-1 (1248 elements) and 3D-SOM9-Cl-3 (918 elements). The composition of the clusters k-means differs only from the SOM clusters by one element (ethyl 2-ethyl acetoacetate), which belongs to 3D-SOM9-Cl-3, and not to 3D-SOM9-Cl-1 as the other elements of 3D-kmeans9-Cl-5.

The elements of area Ad-3D constitute clusters 3D-kmeans9-Cl-7 and 3D-SOM9-Cl-4 (both have the same 37 elements).

Classification at level L16

The SOM classifications at levels L16 (a grid of 4 by 4) provide 13 clusters and the following empty classes: 3D-SOM16-Cl-7, 3D SOM16-Cl-9- and 3D-SOM16-Cl-12.

The allotment of the elements across the clusters is displayed in Figure 2 C and 2 D.

In area Aa-3D, only clusters 3D-kmeans16-Cl-7 and 3D-SOM16-Cl-8 are identical (933 elements), and the other clusters of these areas are partly overlaid. In the following section, the intersection between two clusters is symbolized using the following notation:

3D-M'Li-Cl-x ∩ 3D-M"Lj-Cl-y,

where M refers to the classification method, L is the level of clustering, x and y are cluster numbers, and the symbol "∩" is the mathematical intersection operator. Every intersection measures the number of molecules that belong to the two clusters obtained with the two clustering methods. All overlaps between clusters are reported in Supplementary Table 2.

Table 2. Location of the molecules contained in the mixtures in the 3D-space areas and in the clusters.

Name	3D-space Area	Cluster kmeans	Cluster SOM
Vanillin (V)	Ab-3D	3D-kmeans16-Cl-8	3D-SOM16-Cl-2
Isoamyl acetate (IA)	Ac-3D	3D-kmeans16-Cl-4	3D-SOM16-Cl-13
Frambinone (F)	Ab-3D	3D-kmeans16-Cl-8	3D-SOM16-Cl-5
Ethyl acetate (EA)	Ac-3D	3D-kmeans16-Cl-4	3D-SOM16-Cl-14
beta-Damascenone (bD)	Aa-3D	3D-kmeans16-Cl-7	3D-SOM16-Cl-8
beta-Ionone (bI)	Aa-3D	3D-kmeans16-Cl-7	3D-SOM16-Cl-8
Whiskey lactone (WL)	Aa-3D	3D-kmeans16-Cl-16	3D-SOM16-Cl-4

The elements of 3D-SOM16-Cl-4 are divided between 3D-kmeans16-Cl-14 (249 elements) and 3D-kmeans16-Cl-16 (356 elements). Furthermore, 3D-SOM16-Cl-4 (605 elements) and 3D-SOM16-Cl-8 are identical to 3D-SOM9-Cl-8 and 3D-SOM9-Cl-7, respectively. This is not the case for the clusters k-means for which there is no correspondence but partial overlapping between the three clusters at level L16 (3D-kmeans16-Cl-7, 3D-kmeans16-Cl-14, 3D-kmeans16-Cl-16-) and the three clusters at level L9 (3D-kmeans9-Cl-3, 3D-kmeans9-Cl-4, 3D-kmeans9-Cl-8).

In the Ab-3D area, the elements are split into two clusters, k-means16, and four clusters, SOM16. The clusters 3D-kmeans16-Cl-8 (1029 elements) and 3D-kmeans16-Cl-11 (518 elements) are identical to the clusters 3D-kmeans9-Cl-6 and 3D-kmeans9-Cl-2, respectively.

The overlap between the four clusters SOM16 3D-SOM16-Cl-1 (430 elements), 3D SOM16-Cl-2- (587 elements), 3D-SOM16-Cl-5 (400 elements), 3D-SOM16-Cl-6 (130 elements) and the two clusters k-means16 3D-kmeans16-Cl-8 (1029 elements) and 3D-kmeans16-Cl-11- (518 elements) are reported in Supplementary Table 2.

Part of cluster 3D-SOM16-Cl-2 (499 of 587 elements) added to the entire clusters 3D SOM16-Cl-5- and 3D-SOM16-Cl-6 make the cluster 3D-kmeans16-Cl-8. The other part of the cluster 3D-SOM16-Cl-2 (88 elements) added to 3D-SOM1-Cl-16 makes the cluster 3D-kmeans16-Cl-11.

The elements of area Ac-3D are split between 10 clusters k-means and 6 clusters SOM (Supplementary Table 3). The intersection of the clusters of area Ac-3D is very complex, and there is no complete overlap between the two clusters. For example, approximately half of the cluster 3D-kmeans16-Cl-4 (379 elements) is part of the 3D-Cl13-SOM16 cluster, and the other part of the 3D-kmeans16-Cl-4 cluster (229 elements) is in the 3D-SOM16-Cl-14 cluster. The overlays are summarized in Supplementary Table 2.

The elements of area Ad-3D constitute the clusters 3D-kmeans16-Cl-15 and 3D-SOM16-Cl-3 (both 37 elements), which are identical to those of the classification at level L9.

Distribution of odor notes within clusters

The number of occurrences of the odor notes varies in a large range from 5 (tallow) to 1790 (fruity). Consequently, the direct comparison of the number of occurrences of odor notes in

each cluster is irrelevant. Hence, we considered the following relative amounts related to each cluster: (i) the relative frequency of every odor note compared to their frequency in the database and (ii) the relative frequency of odorants carrying each odor note compared to the number of molecules in the considered cluster (Rugard et al., 2021). Thus, we considered the two ratios as follows:

$$\%\ odor\ note = \%ON\ = \frac{\text{number of occurrences of an odor note in the cluster}}{\text{total number of occurrences of this odor}}$$

$$\%\ odorant\ molecule = \%OM = \frac{\text{number of occurrences of an odor in the cluster}}{\text{number of elements (molecules) in this cluster}}$$

The values of the number of occurrences of the odor notes in the database and in the clusters are reported in Supplementary Table 4.

We examined the occurrences of the 25 most frequent odor notes across the clusters k-means and SOM at the three levels of clustering. The division in increasing the number of clusters improves the odor specificities of each subset of elements. Level L16 appears to be the most appropriate to reflect the odor specificities of the subsets of odorants.

We endeavored to select a suitable classification that allowed us to obtain good contrast between the odor profiles of the clusters. In line with the aim of our study, we focused on the allotments among the clusters k-means and SOM of the molecules involved in the RC mixture and the masking mixture WL-IA.

The location of these molecules in the areas and in the clusters is reported in Table 2, and the graphs of the odor profiles that are related to the clusters containing the molecules of the mixtures are displayed in Figure 3 and in Supplementary Figure 2.

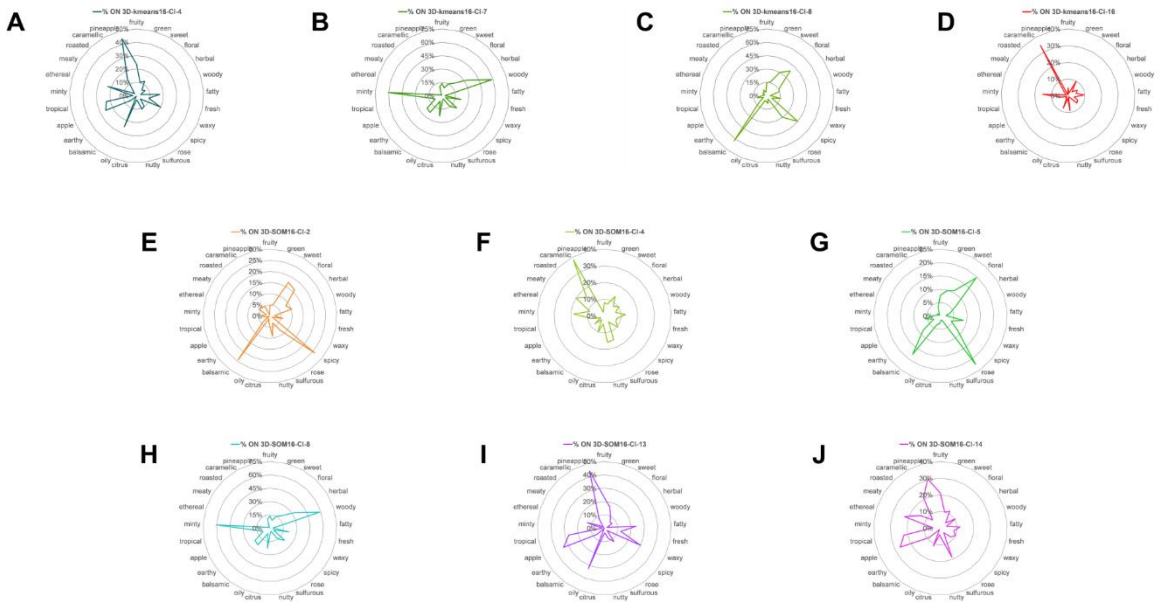


Figure 3. Radar charts of the distribution of the 25 most frequent odor notes across clusters 3D-k-means16 and 3D-SOM16 that contain the molecules of interest. (A) 3D-kmeans16-Cl-4, (B) 3D-kmeans16-Cl-7, (C) 3D-kmeans16-Cl-8, (D) 3D-kmeans16-Cl-16, (E) 3D-SOM16-Cl-2, (F 3D-SOM16-Cl-4, (G) 3D-SOM16-Cl-5, (H) 3D-SOM16-Cl-8, (I) 3D-SOM16-Cl-13, (J) 3D-SOM16-Cl-14.

The area Aa-3D encompasses two molecules from the RC mixture, bD and bI, and WL (masking mixture WL-IA); bD and bI belong to the same clusters as well by k-means (3D-kmeans16-Cl-7) as SOM classification (3D-SOM16-Cl-8). WL belongs to another cluster, 3D-kmeans16-Cl-16- and 3D-SOM16-Cl-4, for k-means and SOM classifications, respectively.

Both V and F belong to the area Ab-3D and to cluster 3D-kmeans16-Cl-8 but to two different SOM clusters (3D-SOM16-Cl-2 and 3D-SOM16-Cl-5). The cluster 3D-kmeans16-Cl-8 contains more than 60% of the molecules that share a balsamic note, as well as approximately 40% of floral and/or spicy odorants. In contrast, in the SOM16 classification, V and F are in different clusters, 3D-SOM16-Cl-2 and 3D-SOM16-Cl-5, respectively. These clusters encompass 24% and 18% of the balsamic odorants, respectively. The main difference between these two clusters is that 3D-SOM16-Cl-2 gathers a larger part of the spicy odorants than that of 3D-SOM16-Cl-2 (25% compared with 8%). Moreover, spicy odorants constitute 18% of 3D-SOM16-Cl-2. Conversely, 22% of the rose odorants are in 3D-SOM16-Cl-5 and constitute 18% of this cluster against only 5% of 3D-SOM16-Cl-2.

The two esters IA and EA are in area Ac-3D. They are gathered in the same k-means cluster (3D-kmeans16-Cl-4), but in two different SOM clusters, IA is in 3D-SOM16-Cl-13 and EA is in 3D-SOM16-Cl-14. The cluster 3D-kmeans16-Cl-4 brings together 44% of the pineapple odorants and, to a lesser extent, the odorants having the apple, oily (both 25%), ethereal, waxy (both 20%) and/or fatty (17%) notes. Using the SOM classification, clusters 3D-SOM16-Cl-13 (IA) and 3D-SOM16-Cl-14 (EA) bring together 44% and 31% of the pineapple odorants, respectively; that is, 75% of pineapple odorants are in these clusters. The odorants carrying apple, ethereal and tropical notes are also included; nevertheless, 3D-SOM16-Cl-14 contains additional caramellic, sulfurous and meaty odorants (21%, 19% and 14% versus less than 8% in 3D-SOM16-Cl-13, respectively).

Using k-means16 classification, several odorants were determined to belong to the same cluster, including the following: bD and bI (3D-kmeans16-Cl-7), V and F (3D-kmeans16-Cl-8), IA and EA (3D-kmeans16-Cl-4). Conversely, except for bD and bI, 3D-SOM16 classification places each component of the mixture in a separate cluster. This is more appropriate since the odorant profiles of the studied molecules best fit those of molecules belonging to the same cluster. We therefore selected the clusters 3D-SOM16 that contained the components of the mixture in the following.

In addition to the 25 most frequent odor notes, we examined the distribution of the odor notes involved in each of the six clusters that encompassed the molecules from the mixtures. The odor description of these seven molecules implicates that 39 odor notes have

at least 5 occurrences (Table 1). Some of these odor notes are common to the list of the 25 frequent odor notes, but several other less frequent notes are characteristic of the odorant profile for the molecules related to a specific cluster. With this in mind, we examined the distribution of the 39 odor notes according to the clusters, and the odorant profiles are displayed in Figure 3 and Figure 4.

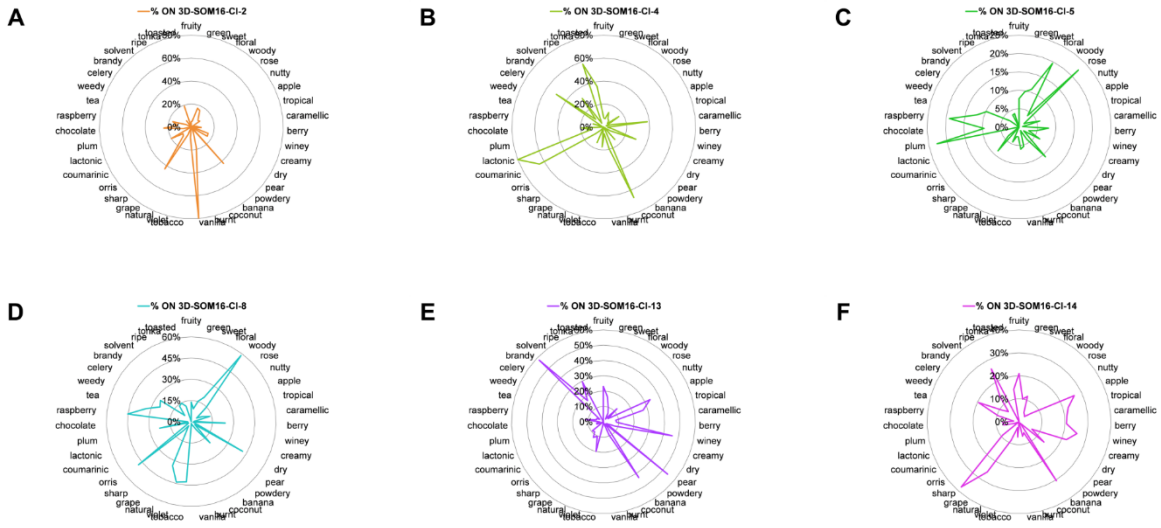


Figure 4. Radar charts of the distribution of the 39 odor notes across clusters of the 3D-SOM16 clustering that contain the molecules of interest. (A) 3D-SOM16-Cl-2, (B) 3D-SOM16-Cl-4, (C) 3D-SOM16-Cl-5, (D) 3D-SOM16-Cl-8, (E) 3D-SOM16-Cl-13, (F) 3D-SOM16-Cl-14.

Considering the odor notes of the mixture highlights the importance of the vanilla note in the cluster 3D-SOM16-Cl-2 in addition to the spicy and balsamic odor notes. Odorants carrying the spicy, balsamic and vanilla notes constitute 18%, 12% and 11% of 3D-SOM16-Cl-2, respectively. Additionally, more than 40% of the occurrences of powdery and grape notes are clustered in this cluster. The odorant description of vanillin also involves chocolate and creamy notes; the cluster 3D-SOM16-Cl-2 encompasses 16% and 24% of the occurrences of these two odor notes, respectively.

The cluster 3D-SOM16-Cl-4 (WL) encompasses nearly 40% of caramellic notes, and more than 60% of the molecules carrying the odor notes coconut (67%), coumarinic (64%), lactonic (80%), and to a lesser extent celery (50%) and tonka (60%) (Figure 3 F and Figure 4 B).

3D-SOM-Cl-5 is dominated by the notes floral, balsamic and rose, and the molecules carrying the floral odor constitute nearly 50% of the cluster (Supplementary Table 4). However, this cluster encompasses less than 25% of these three odor notes. In addition to the floral notes, the odor profile based on the 39 odor notes indicates the presence of the odor notes plum (23%) and raspberry (19%) (Figure 4 C).

More than 30% of the molecules carrying the raspberry note are gathered in the cluster 3D-SOM16-Cl-8, which also contains two molecules of the RC mixture, bI and bD. Nevertheless, the odor profile of 3D-SOM16-Cl-8 is different from that of 3D-SOM16-Cl-5 because the notes tobacco, violet and orris notes are especially frequent in 3D-SOM16-Cl8 (43, 44 and 48%, respectively) (Figure 4 D).

The esters IA and EA belong to clusters 3D-SOM16-Cl-13 and 3D-SOM16-Cl-14, respectively, which are characterized by the frequency of the fruity note (fruity odorants constitute 65% of 3D-SOM16-Cl-13 and 55% of 3D-SOM16-Cl-14, Supplementary Table 4). Both clusters group on each more than 20% apple, tropical and pineapple odorants; in other words, the whole of the two clusters contain, respectively 60, 49 and 75% of apple, tropical and pineapple molecules.

Considering the most frequent odor notes, the difference between 3D-SOM16-Cl-13 and 3D-SOM16-Cl-14 is due to various frequencies of several odor notes. First, at least 30% of the occurrences of fatty, waxy and oily molecules are gathered in 3D-SOM16-Cl-13 compared to approximately 10% in 3D-SOM16-Cl-13. Furthermore, considering the odor notes carried by the mixtures, it appears that the winey, pear and brandy odor notes are characteristic of the odorant profile of 3D-SOM16-Cl-13 (46, 54, and 58% of the odorant carrying winey, pear and brandy notes, respectively, belong to SOM16-Cl-13); a somewhat greater portion of banana molecules (% ON 43%) belong to SOM16-Cl-13 than to SOM16-Cl-14 (30%). In addition, sharp is unique to 3D-SOM16-Cl-14 (38% of the occurrences of this odor are noted in 3D-SOM16-Cl-14 compared to less than 10% in 3D-SOM16-Cl-13). Similarly, approximately 20% of odorants that have the caramellic, berry and creamy notes belong to 3D-SOM16-Cl-14, while less than 10% of such odorants are in 3D-SOM16-Cl-14.

Co-occurrences of odor notes across the clusters 3D-SOM16

In addition to the odorant profile examinations, we look at the co-occurrences of the odor notes (Supplementary Table 5). We focused on 55 odor notes (the 25 most frequent odor notes in addition to 39 odor notes of the mixtures, 10 common to both). We used the nonsymmetrical square matrix of odor notes obtained from the co-occurrences symmetric square matrix by weighting the number of associations by the frequency of occurrences to generate the heatmaps displayed in Figure 5.

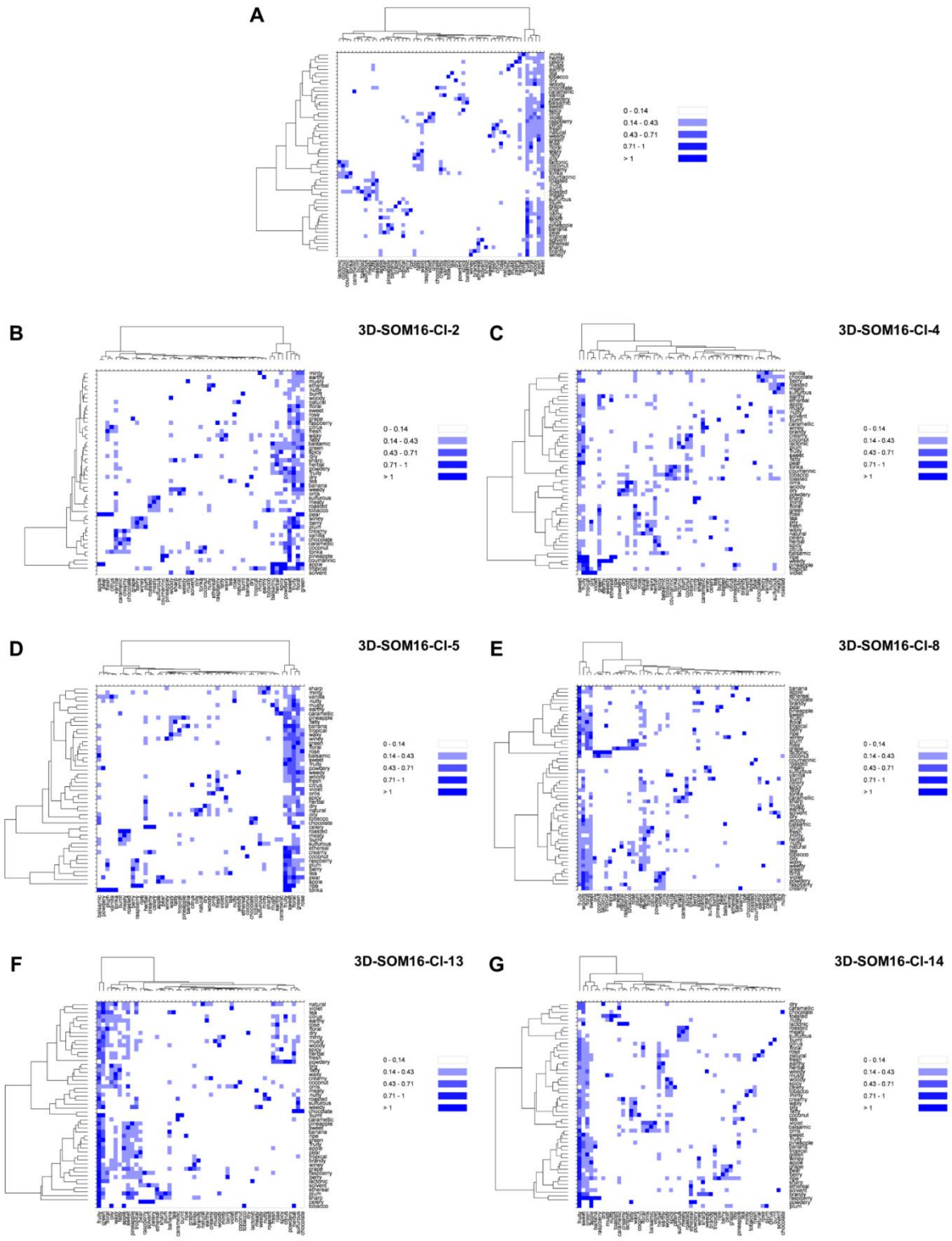


Figure 5. Heatmaps based on the nonsymmetrical square matrix of the odor notes relative co-occurrences of the 55 odor notes of the mixtures obtained from various groups of

odorants. (A) From all odorants of the database, (B) 3D-SOM16-Cl-2, (C) 3D-SOM16-Cl-4, (D) 3D-SOM16-Cl-5, (E) 3D-SOM16-Cl-8, (F) 3D-SOM16-Cl-13, (G) 3D-SOM16-Cl-14.

For the entire database, vanilla is mainly associated with a sweet note (63% of its occurrences) and with the creamy, spicy (both 25%), floral (18%), woody and balsamic (both 11%) odor notes. In the odorant descriptions of the compounds that belong to 3D-SOM16-Cl-2, the frequencies of associations are quite similar but occur slightly more frequently with creamy (27%), spicy (31%) and balsamic (19%) notes. The occurrences of vanilla to chocolate is approximately 10% as well in the entire base as in 3D-SOM-Cl2. Nevertheless, the occurrences of the chocolate notes are noticeably more frequently associated with vanilla, creamy and sweet notes in 3D-SOM-Cl2 than in the entire base. Indeed, 70% of the occurrences of chocolate notes are associated with sweet and 60% with both vanilla and creamy notes, while those are 38%, 21% and 19%, respectively, in the entire base.

The cluster 3D-SOM16-Cl-4 contains WL and is characterized by the caramellic, lactonic, coumarinic, coconut, tonka and celery odor notes. There are several co-occurrences between these odor notes, especially between coconut and coumarinic, lactonic, tonka and toasted (42%, 31%, 42% 21% of the occurrences of these four odor notes in the entire database and 59%, 33% and 60% in 3D-SOM16-Cl4, respectively). Conversely, there are few associations between these six odor notes and the celery, caramellic, woody, nutty, caramellic, burnt, celery, and toasted notes, which themselves have few co-occurrences in the database except for toasted and nutty (43% of occurrences of toasted) notes. However, only 20% of the occurrences of the toasted note are associated with the nutty note in 3D-SOM16-Cl-4. In contrast, caramellic and burnt notes are associated more frequently present in 3D-SOM16-Cl-4 than in the entire base (50% against 19% of the occurrences of the burnt note).

We considered the following odor notes in 3D-SOM16-Cl-5: fruity, sweet, floral, rose, balsamic, berry, plum, raspberry and ripe. The fruity, sweet and floral notes are strongly associated among themselves, as well in the database as in 3D-SOM16-Cl5. The rose, balsamic, berry, plum, raspberry and ripe notes are also strongly associated with fruity, sweet and floral notes. Although only 15% of occurrences of ripe notes are associated with floral notes in the base, the unique occurrence of this odor note in 3D-SOM16-Cl5 is associated with fruity, sweet and floral. Conversely, there are relatively few co-occurrences between rose, balsamic, berry, plum and raspberry notes, as well as fruity, sweet, and floral notes, among themselves. Balsamic, berry and raspberry notes are more frequently associated with the rose note in 3D-SOM16-Cl6-5 than in the entire base (25% against less than 15%); plum and rose notes have no co-occurrence in 3D-SOM16-Cl6-5. Moreover, 25% of the occurrence of the raspberry note is associated with berry and plum notes (against 14% and 7% in the entire base, respectively).

The odorant descriptions of bI and bD involve 16 odor notes, to which the minty and herbal odor notes that are specific to 3D-SOM16-Cl-8 must be added, for a total of 18 odor notes. The herbal and minty notes are well associated with each other as well in the entire base as

in 3D-SOM16-Cl-8 and slightly more in 3D-SOM16-Cl-8. These two notes are also associated with fruity, sweet and floral notes (at least 20% of their occurrences except for the minty and floral notes; 11% and 12% of the occurrences of minty occurs in the entire base and in 3D-SOM16-Cl-8, respectively). Conversely, the co-occurrences of minty and herbal notes with the odor notes of bI and bd do not exceed 5% with the exception of the association between herbal and woody, which reach 37% of the occurrences of herbal in 3D-SOM16-Cl-8 (19% in the entire base). There are few associations among the odor notes that are involved in the descriptions of both bI and bD in the entire base, but they are more frequent in 3D-SOM16-Cl-8 except for the woody note, which is well associated with most notes. The wrong association is between woody and apple notes in the entire base (4% of the occurrences of apple) but reaches 23% in 3D-SOM16-Cl-8. Finally, the woody note is the most frequent odor note associated with raspberry, especially in 3D-SOM16-Cl-8 (74% of the occurrences of raspberry against 40% in the entire base).

We focused on the most frequent 15 odor notes involved in the cluster 3D-SOM16-Cl-13 by embracing the most frequent notes in the entire base as well as those involved in the mixtures. All these odor notes are strongly associated with fruity, and in a minor part to green and sweet notes, as well in the entire base as in 3D-SOM16-Cl-13. Nevertheless, there is little more relative frequency of the association with fruity in 3D-SOM16-Cl-13 as in the entire base, but somewhat less with the green note and much less with the sweet note. This is especially the case for pear and banana notes, and these two odor notes are more associated among themselves in 3D-SOM16-Cl-13 than in the entire base. Moreover, the solvent note is approximately four times more associated with pear and banana notes, while all occurrences of solvent and ripe are associated with the fruity note in 3D-SOM16-Cl-13, against 38% and 81% in the entire base, respectively. The occurrences of fruity, green and sweet notes are more associated with winey and brandy notes in 3D-SOM16-Cl-13. Winey, pear, banana and brandy notes are slightly more associated among themselves in 3D-SOM16-Cl-13 except for winey and banana notes (for them, there is almost no difference between association) in the entire base and 3D-SOM16-Cl-13).

By combining the most frequent notes belonging to the molecules of 3D-SOM16-Cl-14, we retained 13 odor notes, of which 7 were related to EA (fruity, sweet, weedy, green, sharp, brandy and winey). As observed for the odor notes of 3D-SOM16-Cl-13, all notes are strongly connected to fruity, and the relative co-occurrences are approximately the same for the entire base and for 3D-SOM16-Cl-13 with the exception of the sulfurous note, which is two times more associated with fruity in 3D-SOM16-Cl-13 (%ON 41% in 3D-SOM16-Cl-13 against 19% in the entire base). The associations to green and sweet notes are also less frequent with the association with fruity and globally notes in 3D-SOM16-Cl-14 than in the entire base except for brandy (respectively three and two times more frequent with green and sweet in 3D-SOM16-Cl-14). Similarly, the relative co-occurrences of ethereal with pineapple, winey, sharp and weedy are two times more frequent in 3D-SOM16-Cl-14. Brandy has only two occurrences in 3D-SOM16-Cl-14, and only one with winey, sharp and weedy notes, which correspond to the odorant description of EA.

Selection of subsets of odorants based on odor profiles

Within each of these clusters, we selected approximately 10 molecules based on the odor notes of the components of the mixtures (Table 3, Supplementary Table 1). For that, within

each cluster, for each molecule, the number of common odor notes, the number of noncommon odor notes, the percentage of common odor notes and the percentage of noncommon odor notes with the molecule of the mixture were calculated.

Table 3. List of molecules with an odor profile similar to those of the molecules of interest.

Molecule of interest	Molecule's subset with similar odor profile
Whiskey lactone	7-Methyltetrahydronaphthalenone; delta-Heptalactone; Menthofurolactone; Octahydrocoumarin; Laitone; Coconut naphthalenone; (R)-tonka furanone; (+/-)-dihydromint lactone
Isoamyl acetate	2-Methylbutyl-butyrate; hexyl acetate; isobutyl propionate; methyl butyrate; isopropyl propionate; methyl 4-methyl valerate; isoamyl butyrate; propyl acetate; butyl acetate; amyl acetate
Frambinone	Anisyl isobutyrate; 4-hydroxyphenethyl alcohol; 4-(para-tolyl)-2-butanone; Tufurol acetate; 2-Methylbenzyl acetate; alpha-Methylbenzyl-propionate; Phenethyl-2-methylbutyrate; methyl 4-phenyl butyrate; benzyl acetoacetate
Vanillin	vanillyl isobutyrate; vanillin propylene glycol acetal; ethyl vanillin isobutyrate; 1-Ethoxy-2-methoxybenzene; ortho-dimethyl hydroquinone; ethyl vanillin; vanillyl acetate; vanillylidene acetone; vanillin hexylene glycol acetal; ethyl vanillin hexylene glycol acetal; ethyl vanillin propylene glycol acetal
Ethyl acetate	2-Methylbut-2-enyl-formate; Isobutyl pyruvate; methyl acetate; methyl (E)-2-butenoate; ethyl 2-methyl butyrate; 2-methyl butyl propionate; isopropyl acetate; ethyl nitrite; hexyl lactate; methyl 3-hydroxybutyrate
beta-Damascenone	plum damascone (high alpha); (Z)-alpha-damascone; Cyclohexylethyl isovalerate; Cyclohexylethyl valerate; 1-(3-(methyl thio)-butyryl)-2,6,6-trimethyl cyclohexene; 3-cyclohexene-1-carboxylic acid, 2,6,6-trimethyl-, methyl ester
beta-Ionone	beta-ionyl acetate; alpha-ionol; alpha-ionyl acetate; 3-Methylcyclohexyl acetate; beta-Irone; Campholene-acetate; Nopyl-acetate; 4-dimethyl ionone

To select subsets containing approximately ten molecules with odor notes close to those of the mixture molecules, we selected all the molecules that had at least 3 common odor notes and no more than two uncommon odor notes. We refined the selection according to the following criteria:

- If in these first criteria, more than ten molecules were selected, we again selected these molecules in which the percentage of noncommon odor notes was the smallest;
- If there was still a need to restrict the number of molecules selected, we selected the molecules that had the highest percentage of common odor notes.

These subsets of odorants, including each molecular component of the mixtures, were used to perform a pharmacophore approach.

Pharmacophore study

Pharmacophore generation

We used the PHASE module (Dixon et al., 2006) to set the chemical features of the odorants and identify the critical common features present in a set of odorants. We developed pharmacophores in the following ways: (i) from groups based on the mixture components and containing at least two molecules of the mixtures, (ii) from each molecule component of the mixtures, and (iii) from the subsets of molecules selected in the SOM16 clusters. Each molecule is in fact an ensemble of its conformers (energy range of 21 kJ/mol) and was denoted “M-c”, where M is the molecule, and c symbolizes conformers. The subsets of various molecules will be named by the initial of the referred molecule followed by the initial "s" (for subset). For example, the vanillin subset selected on the basis of odorant profile was named "V-s".

Every pharmacophore was developed following a ligand-based model with at least 3 features, and all molecules were considered active molecules. For all hypotheses, the aromatic rings were considered as hydrophobic groups. Knowing that a hydrogen bond donor is also an acceptor, we used only the hydrogen bond acceptor (A) feature. The PhaseHypoScore ranks the hypotheses from the worst to the best, allowing us to identify the most significant hypothesis. For clarity, the hypotheses relating to each molecule was named by the abbreviation "hyp" (for pharmacophore hypothesis) followed by the initial name of the molecule. For example, the hypothesis developed from the vanillin V-s subset was named "hyp-V-s".

Pharmacophores generated from subsets composed of components of the mixtures

Pharmacophores generated from the subsets based on the “Red Cordial” mixture

We considered the set of the six molecules and several subsets involving at least two molecules of the RC mixture. Only one hypothesis could be created from the V-IA-F-EA-bD-bI-s subset (Figure 6 A). The partial hit for this hypothesis concerns only IA, bD and bI, so it does not match all active molecules. The details of the pharmacophore are presented in Supplementary Table 6.

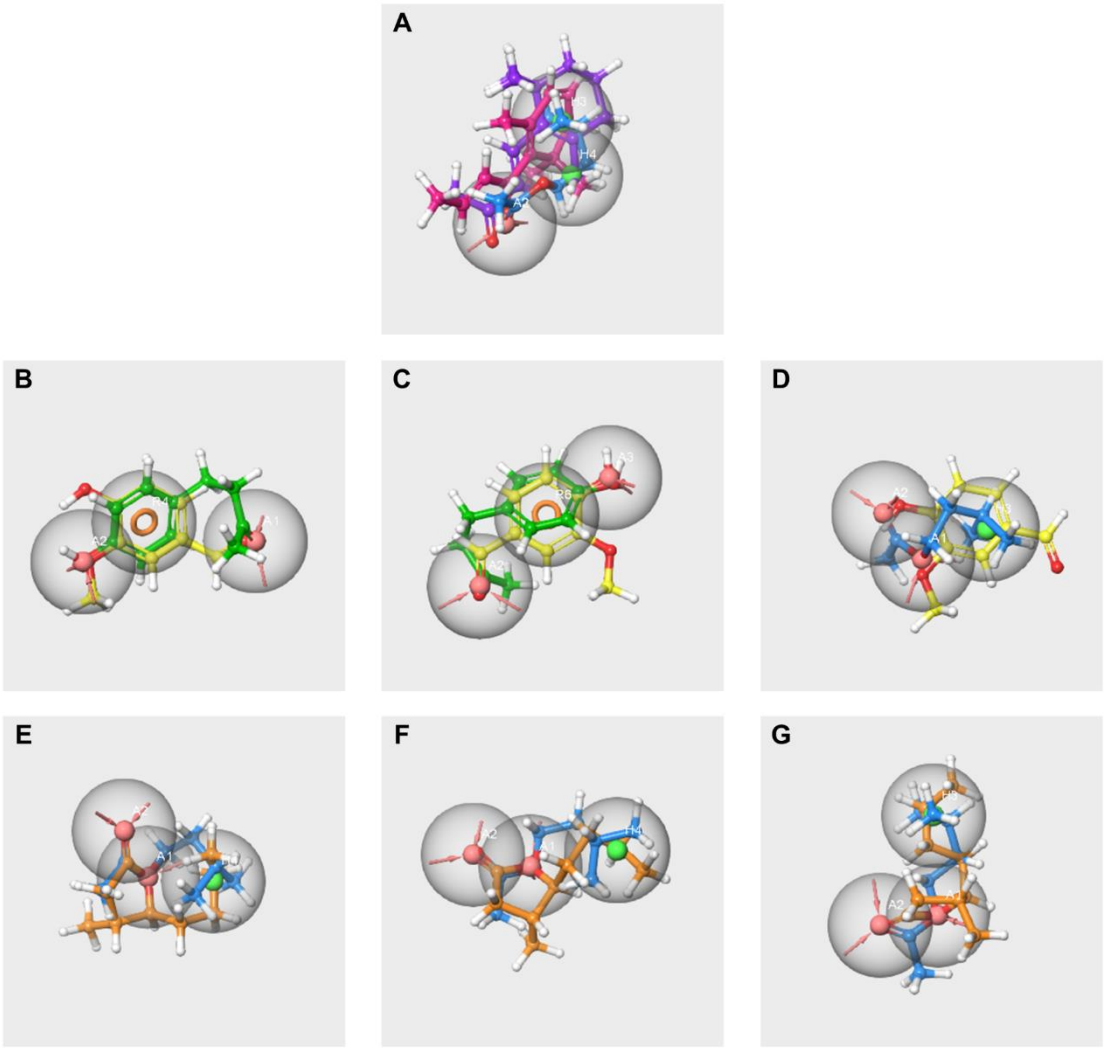


Figure 6. Pharmacophore hypotheses generated from the subsets V-IA-F-EA-bD-bI-s (A), V-IA-F-s (B, C, D) and WL-IA-s (E, F, G). V in yellow, IA in blue, F in green, bD in pink, bI in purple, WL in orange. The red spheres correspond to hydrogen bond acceptors, the green spheres correspond to hydrophobic groups, and the orange circles indicate aromatic rings R.

According to a previous study (Sinding et al., 2013), the presence of V, IA and F has a significant impact on the perception of the blending mixture. However, compared to the other two molecules, F is involved less in the perception of grenadine syrup odor. Therefore, we focused on V, IA and F to generate pharmacophore hypotheses from three subsets encompassing V and at least V and one of the two other main components of the RC mixture, the V-IA-F-s, V-IA-s and V-F-s subsets.

Three pharmacophore hypotheses were generated from the V-IA-F-s subset. The mapping of the molecules is displayed in Figure 6 (B, C, D).

The two first pharmacophores (AAR_1 and AAR_2) possess an aromatic ring and hydrogen bond acceptors and match V and F but not IA. The last hypothesis, AAH_1, which maps V and

IA, contains two hydrogen bond acceptors and a hydrophobic feature that corresponds to the aromatic cycle of V. The poor score values indicate the weak significance of the hypotheses.

The pharmacophore generations carried out from the V-IA-s and V-F-s subsets also provided poor results. The single hypothesis obtained from the V-IA-s subset is identical to hypothesis AAH_1 generated from the V-IA-F-s subset. Likewise, the two hypotheses generated from the V-F-s subset are identical to those generated from the V-IA-F-s subset (AAR_1, AAR_2).

Pharmacophores generated from the subset WL-IA-s

Nine pharmacophores were generated from the two molecules, and the first three were characterized by good PhaseHypoScore values (Supplementary Table 6). The three best hypotheses and the mapping of the molecules are displayed in Figure 6 (E, F, G).

These hypotheses consist of the same features, two A and one H, but differ in their geometries, as indicated by the distances and angle values. Indeed, the AAH_2 hypothesis has a linear geometry ($\widehat{A1A2H4}$ 149.7 deg) similar to those of hypotheses AAH_1 and AAH_3 ($\widehat{A1A2H4}$ approximately 100 deg). The distances A1-A2 vary from 2.25 Å (AAH_1 and AAH_2 hypotheses) to 2.27 Å (AAH_3 hypothesis). The distances A1-H4 are 2.58 and 3,56 Å for AAH_1 and AAH_2 hypotheses, respectively; the distance A1-H3 is 3.83 Å for AAH_3 hypothesis.

Pharmacophore generated from the conformers of each single molecule component of the RC mixture and WL-IA masking

In this case, each molecule is considered an ensemble of its conformers, and the obtained hypothesis represents the pharmacophore of its conformational space. The pharmacophore hypothesis generated from each molecule of the two mixtures is displayed in Figure 7.

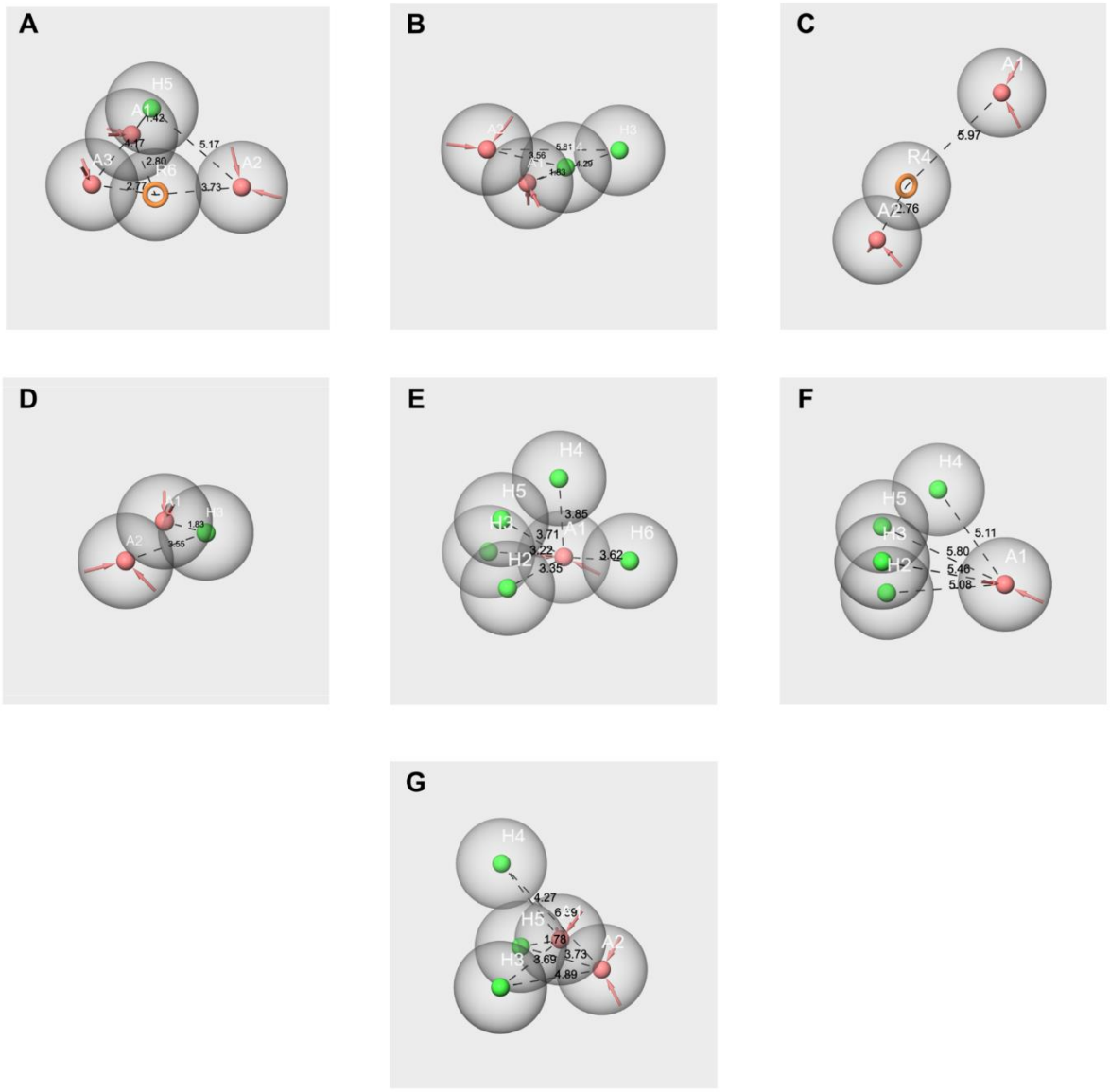


Figure 7. Pharmacophore best hypotheses generated from V-c (A), IA-c (B), F-c (C), EA-c (D), bD-c (E), bI-c (F), and WL-c (G). The red spheres correspond to the hydrogen bond acceptors. The green spheres correspond to the hydrophobic groups. Orange circles correspond to aromatic rings. The values indicate the distance in Ångström between two features. Only the distances between features A and H and A and R are represented.

All hypotheses include at least one hydrogen bond acceptor feature (A) and one hydrophobic feature (H) with the exception of hyp-F-c, which does not contain hydrophobic features but an aromatic ring (R). Conversely, the hyp-V-c hypothesis contains both R and H features. The hypotheses hyp-EA-c, hyp-bD-c, hyp-bI-c and hyp-WL-c encompass various numbers of H and A features: from two A and one H for hyp-EA-c to one A and five H for hyp-bD-c; hyp-bD-c and hyp-bI-c are richer in hydrophobic features.

All the interfeature distance values are reported in Supplementary Table 7. By comparing the distances between the A features and H features for the hypothesis of the molecular components of RC mixtures, very few close values between the A and H or R features were observed. Indeed, there are three distances A-H for hyp-V-c (1.424 Å, 4.165 Å, 5.167 Å), and only one of them matches those of hyp-IA-c (A1-H3, 4.287 Å). Conversely, the distances A3-R6 (2.767 Å) and R6-A1 (2.796) of hyp-V-c are close to the A2-R4 distance of hyp-F-c (2.758 Å), while R4-A1 of hyp-F-c (5.967 Å) is close to A2-H3 of hyp-IA-c (5.815 Å). However, the distances between the two hydrophobic features, H or R, vary greatly according to the hypotheses (R6-H5, 3.734 Å for hyp-V-c; H4-H3 2.468 Å for hyp-IA-c). Moreover, despite quite similar structures regarding the number of hydrophobic features between hyp-bD-c and hyp-bI-c, there are no common A-H distances (ranging from 3.225 to 3.853 Å and from 5.082 to 5.803 Å, respectively, for hyp-bD-c and hyp-bI-c).

In contrast, there are several close values regarding the A-H and H-H distances of the hyp-IA-c and hyp-WL-c hypotheses, including the following: 1.828 Å, 3.558 Å, and 4.287 Å for the A1-H4, A2-H4 and A1-H3 distances of hyp-IA-c, respectively, and 1.781 Å, 3.686 Å, 4.273 Å for the A1-H5, H3-A1 and A1-H4 distances of hyp-WL-c, respectively. Moreover, the H-H distances are also close (2.468 Å for hyp-IA-c and 2.361 for hyp-WL-c). A similar observation can be made for A-A distances of 2.303 Å and 2.255 Å for hyp-IA and hyp-WL, respectively.

Pharmacophores from the subsets of molecules having similar odor profiles

We developed pharmacophore hypotheses from each subset resulting from the selection of the molecules in the clusters SOM16 on the basis of their odor profiles (Table 3, Supplementary Table 1): V-s (12 molecules), IA-s (11 molecules), F-s (10 molecules), EA-s (11 molecules), bD-s (7 molecules), bI-s (10 molecules) and WL-s (9 molecules). The details of the PHASE-generated top hypotheses are reported in Supplementary Table 8 and the interfeature distances in Supplementary Table 9.

The best hypotheses obtained from each subset and the mapping of the related molecules are shown in Figure 8. The structures of the molecules align in an ordered way on each pharmacophore hypothesis as follows: the rings of each molecule are aligned on each other, and the carbon chains are in close spaces and the oxygen atoms map the A features, although in a lesser ordered arrangement in the cases of bD (Figure 9 E) and bI (Figure 9 F). All molecules of each subset map the related best significant hypotheses with the exception of frambinone, which does not map the hyp-F-s model.

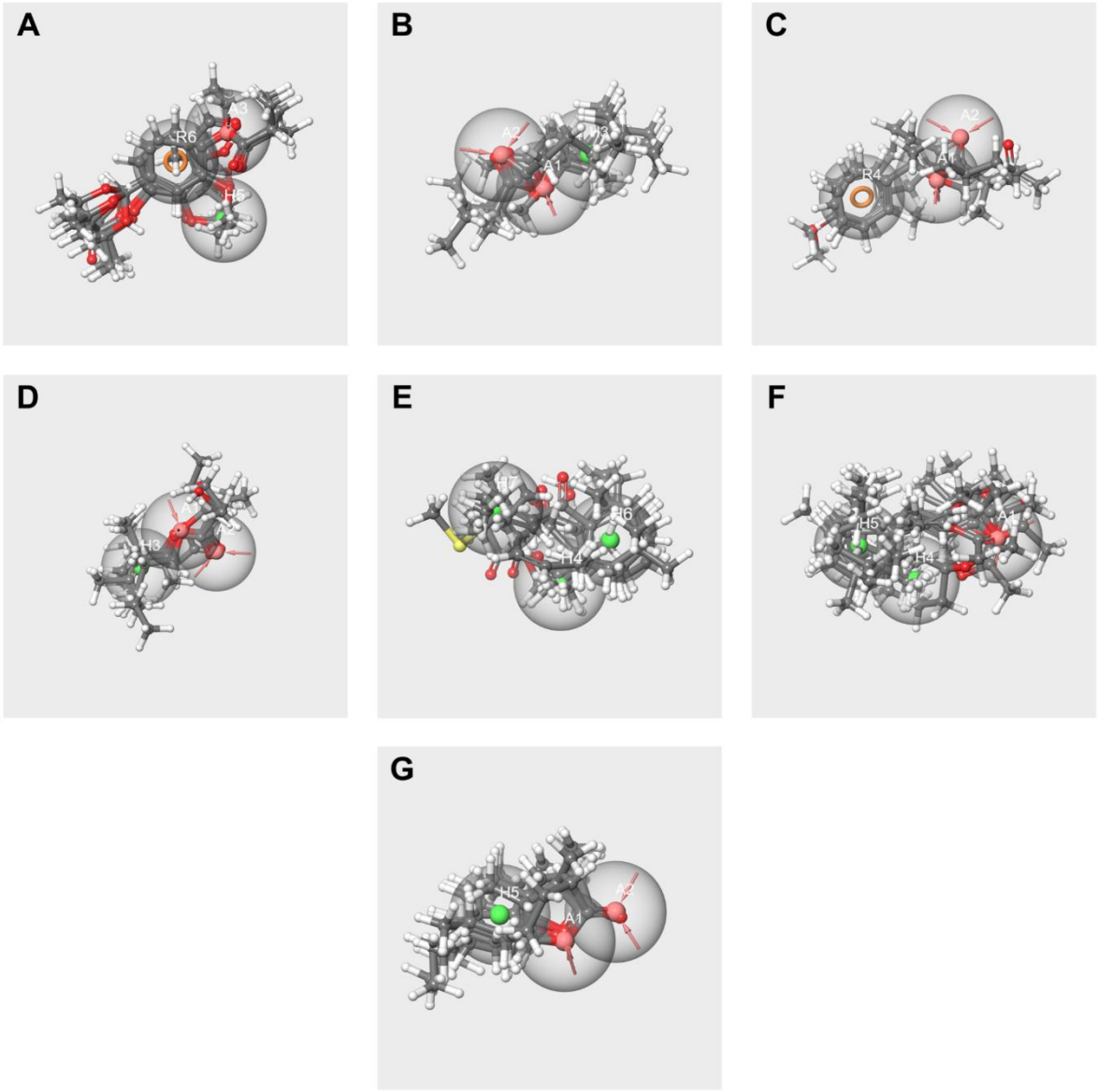


Figure 8. Best pharmacophore hypotheses obtained from the V-s (A), IA-s (B), F-s (C), EA-s (D), bD-s (E), bI-s (F) and W-s (G) subsets. The red spheres correspond to the hydrogen bond acceptors. The green spheres correspond to the hydrophobic groups. The orange circles correspond to aromatic rings.

The best hypotheses models that are related to RC mixtures were diversely made. All encompass at least one hydrogen bond acceptor A, except hyp-bD-s, which contains only H. Hyp-V-s encompasses the three features A, H and R, and hyp-bI-s possesses two H. Four hypotheses, hyp-IA-s, hyp-EA-s, hyp-WL-s and hyp-F-s, have two hydrogen bond acceptors, while hyp-F-s does not contain hydrophobic feature H but an aromatic feature R. The interfeature distances are reported in Supplementary Table 6.

Pharmacophore comparisons

To evaluate the similarities between the hypotheses, we compared the three-dimensional arrangement of the features of these hypotheses by aligning them using the "Hypothesis Alignment" task, which aligns hypotheses in pairs using one of the two hypotheses as a template. The root-mean-squared deviation (RMSD) between the features of the reference hypothesis and the second hypothesis allows us to evaluate the quality of the alignment of the structures. All RMSD values are reported in Table 4.

We focused on the three main components of the RC mixture (V, IA and F) and WL. The comparisons were performed between (i) hypotheses generated from the molecules V, IA, F, and WL; (ii) hypotheses generated from the subsets of similar odor profiles according the components of the mixtures; and (iii) hypotheses generated from the subsets of similar odor profiles with the hypothesis of the corresponding molecule.

Comparisons between the hypotheses generated from the molecules V, IA, F, and WL

We compared in pair (i) the hypotheses hyp-V-c, hyp-IA-c, hyp-F-c (RC mixture), and (ii) hyp-WL-c and hyp-IA-c (masking mixture WL-IA). The four obtained hypothesis alignments are displayed in Figure 9, and the RMSD values are reported in Table 4.

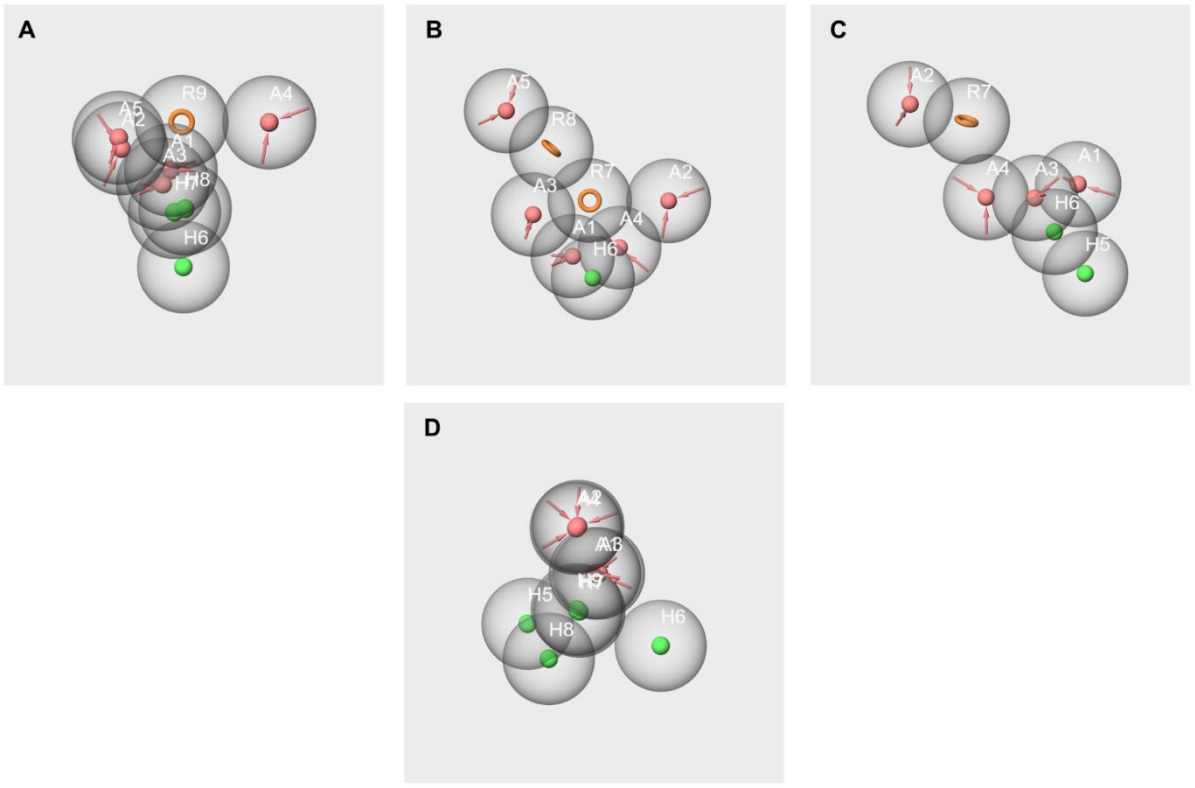


Figure 9. Pharmacophore mappings of hypotheses hyp-V-c and hyp-IA-c (A), hyp-V-c and hyp-F-c (B), hyp-IA-c and hyp-F-c (C), hyp-IA-c and hyp-WL-c (D).

A mediocre overlap was obtained for the hyp-V-c and hyp-IA-c hypotheses, corresponding to a poor RMSD value. No reliable result was obtained for hyp-V-c and hyp-F alignment or hyp-F-c and hyp-IA-c alignment. Conversely, the hyp-IA-c and hyp-WL-c hypotheses were satisfactorily aligned, as reflected by a good RMSD value.

Comparison of the hypotheses generated from the subsets of similar odor profiles

We carried out the following pairwise comparisons between the hypotheses generated from the subsets on the odor profiles of the components: (i) V-s, IA-s, F-s (RC mixture) and (ii) WL-s and IA-s (WL-IA masking). The alignments of the pharmacophores are displayed in Figure 10, and the RMSD values are displayed in Table 4.

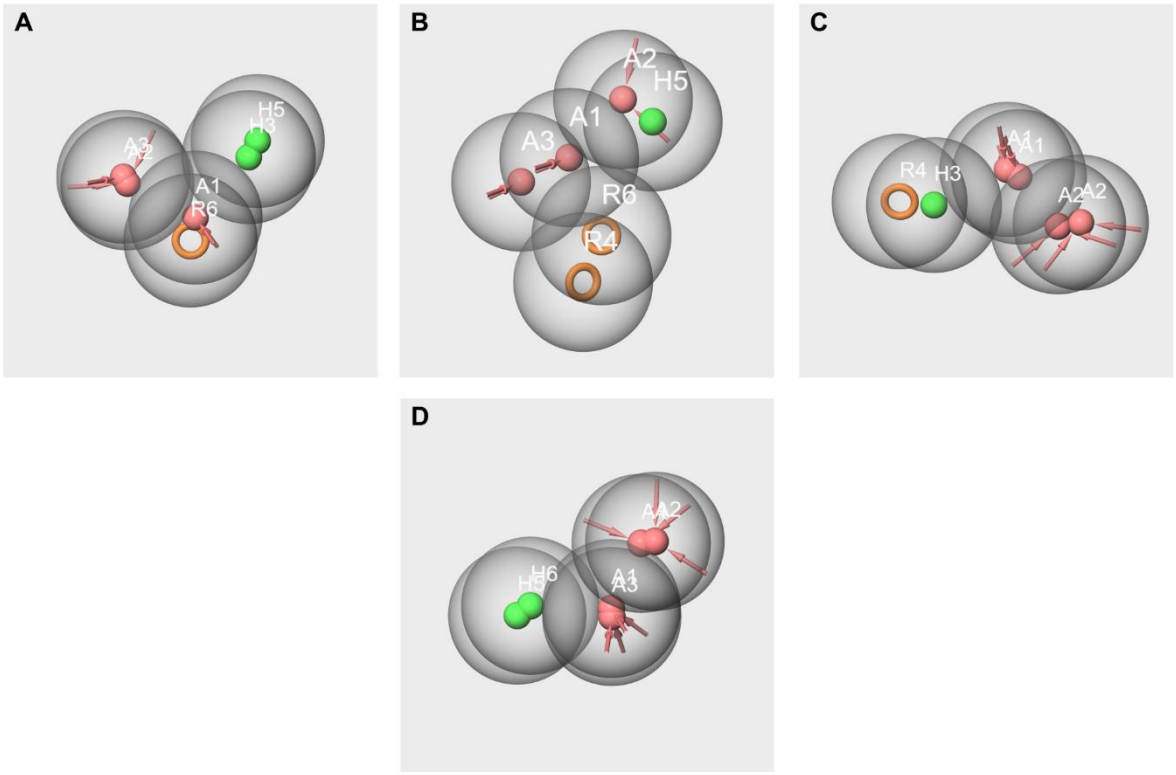


Figure 10. Alignments of the pharmacophore hypotheses hyp-V-s and hyp-IA-s (A), hyp-V-s and hyp-F-s (B), hyp-IA-s and hyp-F-s (C), hyp-WL-s and hyp-IA-s (D).

Table 4. RMSD of the alignment of the pharmacophore hypotheses.

Pair of hypotheses	RMSD
hyp-V-c and hyp-IA-c	0.5647
hyp-V-c and hyp-F-c	-
hyp-IA-c and hyp-F-c	-
WL-c and IA-c	0.1485
V-s and IA-s	0.5305
V-s and F-s	1.3647
IA-s and F-s	0.7449
WL-s and IA-s	0.3842
hyp-V-c and hyp-V-s	0.066
hyp-IA-c and hyp-IA-s	0.247
hyp-F-c and hyp-F-s	-
hyp-WL-c and hyp-WL-s	0.442

The three resulting map comparisons, which were carried out for the hypotheses that were generated from subsets V-s (Figure 10 A), IA-s (Figure 10 B), and F-s (Figure 10 C), show inadequate overlap between the hydrogen bond acceptors and the hydrophobic features.

The root-mean-squared deviation (RMSD) allows evaluation of the quality of the alignment of the features. The alignment of these hypotheses provided quite high RMSDs (Table 4), indicating an incorrect quality of the alignments. Indeed, in the cases of the three alignments, there are very close positions for (i) A and R (hyp-V-s and hyp-IA-s comparison), (ii) A and H (hyp-V-s and hyp-F-s comparison), and (iii) A and R (hyp-IA-s and hyp-F-s comparison).

Conversely, the comparison between the hyp-WL-s and hyp-IA-s hypotheses provided a satisfactory mapping (RMSD = 0.3842) due to a good overlap of their respective A and H features (RMSD = 0.440, Figure 10 C).

Comparison of the hypotheses generated from the subsets of similar odor profiles with the hypothesis of the corresponding molecule

We compared in pairs the hypotheses generated from the conformers of the components of the RC mixture (hyp-V-c, hyp-IA-c, hyp-F-c and hyp-WL-c) with the hypotheses generated from the subsets based on odor profiles (hyp-V-s, hyp-IA-s, hyp-F-s g and hyp-WL-s). The obtained hypothesis alignments are displayed in Figure 11, and the RMSD values are shown in Table 4. No reliable alignment could be achieved for F-s/F-c (no RMSD value).

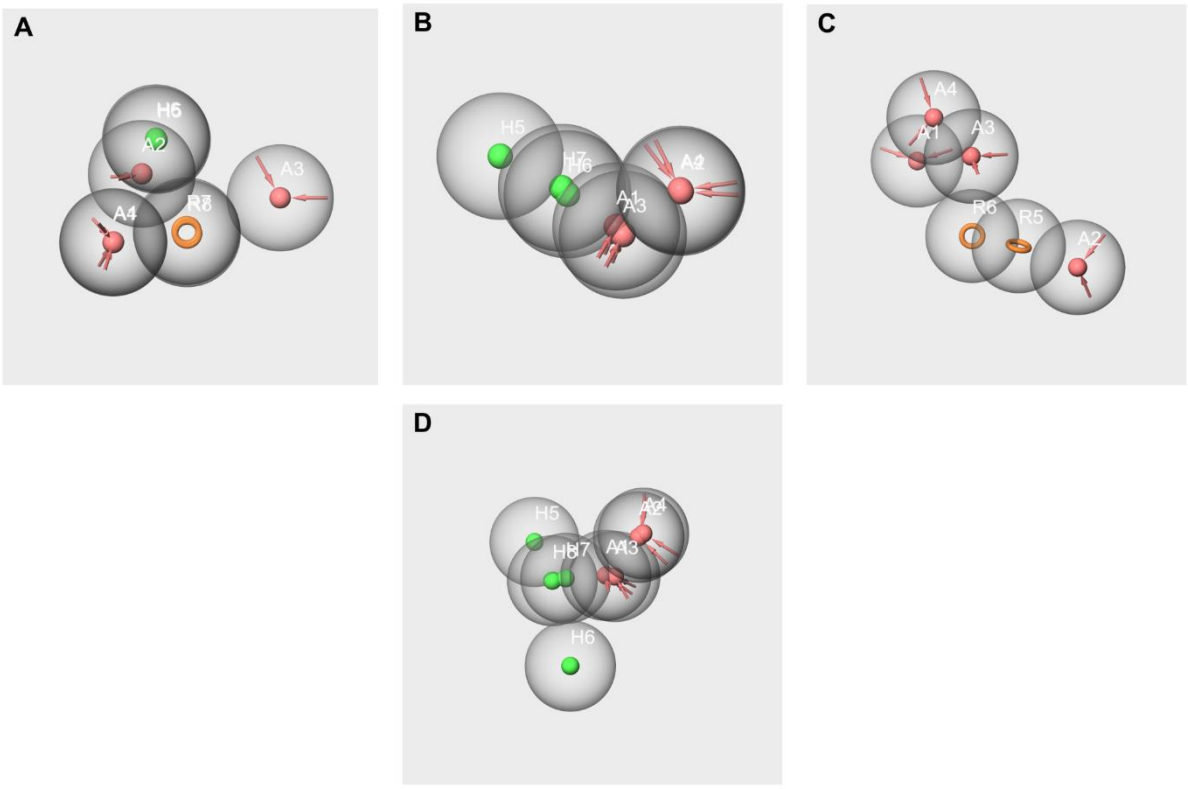


Figure 11. Pharmacophore mappings of the pharmacophore hypotheses hyp-V-c and hyp-V-s (A), hyp-IA-c and hyp-IA-s (B) and hyp-F-c and hyp-F-s (C), hyp-WL-c and hyp-WL-s (D).

Conversely, the alignment in pairs of the hypotheses hyp-V-c/hyp-V-s and hyp-IA-c/hypIA-s present very low RMSDs (Table 4), meaning that the two possess a very close structure. Indeed, there is substantial overlap for the pairs of features R7-R8, H5-H6, and A1-A4. Similarly, the alignment of hyp-IA-c and hyp-IA-s shows good mapping between H6 and H7, A1 and A3, and A2 and A4. To a lesser extent, the alignment of hyp-WL-c and hyp-WL-s is satisfactory, and the overlays between A and H features exclude only one H of each model.

Discussion

The present work qualitatively highlights the links between the structural properties of odorants and their perceived odor to improve the understanding of the homogeneous perception of aroma mixtures. To achieve this goal, we developed our work according to several approaches. First, we divided a large database containing 5665 odorants into clusters and focused on seven molecules involved in two aroma mixtures, generating a homogeneous perception. These two mixtures include (i) a blending mixture of six odorants, called the red cordial mixture (RC mixture), whose olfactory perception carries a typicality analogous to those of grenadine syrup (Le Berre et al., 2010; Sinding et al., 2013), and (ii) a binary masking mixture of whiskey lactone (WL, “woody” note) and isoamyl acetate (IA, “fruity” note) (Atanasova et al., 2004). We investigated the fit between the odor notes carried by molecules of the mixtures and the odor notes characteristic of the clusters to which they belong. Finally, we carried out a pharmacophore study to investigate and compare the structural features of the components of the two mixtures to evaluate the

possibility that molecules in the mixtures could bind to some common sites at the olfactory receptors.

In a recent study (Rugard et al., 2021), we pointed out the advantage of the use of the UMAP dimensional reduction method associated with k-means clustering as a promising way to suggest hypotheses on odorant structure-odor relationships. In that first work, we limited the dimensional reduction to 2D space and the separation of elements into four clusters. In the present work, we extended the distribution of the elements in a 3D space.

The UMAP reduction dimension also allowed us to clearly identify three areas in the 2D space as in the 3D space. It should be noted that although there is no exact correspondence, the elements of area Aa-2d are mainly in area Ac-3D; thus, almost all of the elements of Ab-2D and Ac-2D are, respectively in Aa-3D and Ab-3D.

We evaluated the reliability of two clustering methods at increasing levels of clustering, k-means and self-organizing maps (SOM). Both k-means and SOM are nonhierarchical clustering methods for which the desired number of clusters k must be predefined. Nevertheless, there are several differences and specificities of these two clustering algorithms (Xiao et al., 2005; Mingoti and Lima, 2006; Rodriguez et al., 2019). Indeed, k-means is a vector quantization method, of which a limitation is that the final classification depends on the initial selection of the number of centroids and seeds, while nodes (centroids) are independent of each other. In addition, SOM is a nonlinear mapping method and neural network model by which the clusters are formed geometrically, while the number of neurons of the output layer has a close relationship with the class number in the input stack. Moreover, the SOM can change its internal structure by the influence of elements from the same system by preserving the original topology of the dataset.

Applying the two classification methods for partitioning elements into four clusters to the 2D and 3D chemical spaces that were defined by the UMAP methods led to different cluster attributions of elements of areas Aa-2D and Ac-2D according to the use of k-means or SOM clustering. Conversely, except for a few exceptions, the elements are similarly divided into the four clusters based on 3D-UMAP space coordinates. It can be noted that the allotment of elements into nine clusters based on 2D-UMAP coordinates allows us to obtain a similar allocation among clusters for the elements belonging to areas Ac-2D (Figure 1, Supplementary Figure 1).

The UMAP method differs from most other dimensional reduction approaches in that the distribution of the elements in the UMAP space is optimized according to the dimension of this space. In other words, the coordinates of any element in the 2D-UMAP space differ from their coordinates in any plane of the 3D-UMAP space. Consequently, there should not be element masked in the background by other elements. Nevertheless, a distinctive allotment of numerous and close elements in some zones of the space will be more efficient in 3-D space than in 2D space. This could probably explain the best consistency of the division into four clusters in the 3D-UMAP space.

Another observation related to the benefit of the use of 3D-UMAP space concerns the number of effective clusters. Indeed, in the 2D-UMAP space, there are nine or sixteen clusters according to dividing at level L9 or L16, respectively. Conversely, at the same

dividing levels L9 and L16, the number of SOM clusters is decreased to 7 and 13, respectively. This could be due to the flexibility of the SOM algorithm, which, compared to k-means, can better adapt the number of clusters to the topology of the data. Moreover, the allotment of odorants in distinct clusters according their odor notes is more appropriate by SOM16 than by kmeans16 clustering, at least in the context of the present study. That does not mean that SOM decisively outperforms k-means; several groups of odorants defined by k-means, or by intersects between k-means and SOM clusters, could be useful to other analyzes in future studies.

The following SOM16 clusters encompass the components of odorants: SOM16-Cl-2 (V), SOM16-Cl-4 (WL), SOM16-Cl-5 (F), SOM16-Cl-8 (bI and bD), SOM16-Cl-13, (IA), and SOM16-Cl-14 (EA). It should be noted that only some odorants shared all major odor notes related to each cluster. For example, all molecules of cluster SOM16-Cl-2 do not have common sweet, floral, spicy and balsamic odor notes. We examined the location of the components of the mixtures with regard to the odor notes characterizing the elements of these six clusters to which they belong. As detailed below, the allotment of the components of the mixtures by cluster appeared satisfactory.

The key odor note of vanillin is vanilla, which has 83 occurrences and ranks by decreasing frequency 59th out of 162 odor notes. Approximately 80% of the occurrences of vanilla are in the cluster 3D-SOM16-Cl-2 (Figure 3 E and Figure 4 A). The belonging of vanillin to this cluster is well justified, as well on the basis of the number and relative frequency of its occurrences than in regard to the notes creamy and chocolate. These two odor notes have 15% and 24% of their occurrences in 3D-SOM16-Cl-2, respectively, where they are associated with 70% of the occurrences of chocolate, while 60% of the occurrences of each of the two notes are associated with vanilla.

The complex odorant description of WL, which belongs to 3D-SOM-Cl-4, involves 9 odor notes, of which 5 (tonka, coumarinic, coconut, celery, lactonic) are mainly characteristic of the odorant profile of the molecules belonging to this cluster (Figure 3 F and Figure 4 B). Taken as a whole, the odor notes carried by WL are more frequent in 3D-SOM-Cl-4 than in any other cluster 3D-SOM16, as illustrated in Figure 4 B. However, the woody note that was identified as the key typicity in the overshadowing of the fruity note provided by IA (Atanasova et al., 2004) is much more frequent in 3D-SOM16-Cl-8 than in 3D-SOM16-Cl-4.

Despite its name, frambinone (or raspberry ketone) does not belong to the cluster 3D-SOM16-Cl-8, which contains 45% of raspberry odorants. Conversely, F belongs to cluster 3D-SOM-Cl-5, which gathers 19% of raspberry odorants (Figure 3 G and Figure 4 C). The odor of F is described by the fruity, sweet, floral, berry, raspberry, and ripe notes. Nevertheless, the complete description reported in the Flavor-Base is "*Very sweet, fruity odor and taste reminiscent of raspberry*", highlighting the sweet note before raspberry that appears as a "reminiscent" (Leffingwell, 2013), namely, quite weak. In the same way, Arctander indicates, "*Very sweet, fruity, warm odor resembling Raspberry preserves. Sweet-fruity taste but not very powerful*" (Arctander, 1969). This suggests little raspberry typicity and the dominance of the floral note.

Both bD and bI belong to 3D-SOM16-Cl-8, which is characterized by woody and minty notes (Figure 3 H and Figure 4 D). The odorants bD and bI have one of the most complex

descriptions, involving 11 and 8 odor notes, respectively; they have four notes in common (fruity, sweet, floral, raspberry); rose, apple, tobacco, natural, grape, plum, and tea are specific to bD, while woody, tropical, berry, dry, powdery, violet and orris are specific to bI. Neither bI nor bD have a minty note, while minty odorants are obviously mainly gathered in 3D-SOM16-Cl-8. Moreover, there are few, or not, co-occurrences between minty and the odor notes involved in the odorant description of bI and bD. This would suggest that another specific group of odorants characterized by the minty note is embedded in 3D-SOM16-Cl-8 but is distinct from the odorant carrying floral notes such as violet, orris, and rose. As the raspberry note being common to F, bI and bD, and the examination of the co-occurrences of the raspberry note in the clusters 3D-SOM16-Cl-5 and 3D-SOM16-Cl-8 points out the following privileged associations: between raspberry and floral in 3D-SOM16-Cl-5 on the one hand and between raspberry and woody in 3D-SOM16-Cl-8 on the other hand. Thus, these two types of raspberry odorants would be related to unlike types of chemical structures.

IA and EA belong to two distinct 3D-SOM16 clusters (SOM16-Cl-13 and SOM16-Cl-14), but both are included in the cluster 3D-kmeans-Cl-8. The examination of the odor notes of the molecules that constitute the clusters SOM16-Cl-13 (Figure 3 I and Figure 4 E) and SOM16-Cl-14 (Figure 3 J and Figure 4 F) reveals the pertinence of the division into two groups by the SOM calculation. Indeed, the analysis of the relative frequencies (% ON) of the odorants highlighted crucial differences in the distribution of fatty, waxy, oily, winey, pear and brandy molecules, largely more frequent in SOM16-Cl-13, whereas caramellic, berry, creamy and sharp notes more specifically characterize SOM16-Cl-14.

To determine the specific structural characteristics of the molecules involved in the two mixtures, we performed a pharmacophore study using a threefold approach for purposes of comparison between the generated hypotheses. In addition to the generation of pharmacophores using single or several molecules of the mixtures, we performed pharmacophore generation using subsets of a dozen selected molecules in each of the six clusters on the basis of their odorant descriptions.

Several significant pharmacophore models were generated using the subset WL-IA. The three best significant hypotheses give various alignments of the two molecules with good overlapping of the molecules on the features. The hypotheses generated by WL and by IA share interfeature distances that allow good pharmacophore mapping of the best significant hypotheses generated from each of them. The subsets based on odorant profiles of WL and IA allowed us to obtain reliable models, as well as good alignments of the molecules, including the reference molecules WL and IA. The good results obtained by in pair comparisons of pharmacophores put forward a similar between the geometry of the hypotheses generated as well from single molecule WL and IA as from the subsets. All these statements are in favor of the assumption that WL and IA would share common binding mode(s) and/or binding site(s) at the peripheral level of olfactory systems, and this even though WL and IA belong to different structural groups.

In contrast, the case of RC mixtures appears to be far more complex and challenging. No reliable model was obtained from the six molecules constituting the RC mixture. Indeed, only one pharmacophore was obtained, which had poor significance, and only three molecules were mapped on the feature, excluding V, IA and F, while V and IA are the indispensable components required to confer a satisfactory typicality close to that of

grenadine syrup. The restriction of the training set to V, IA and F did not satisfactorily improve the result because the whole overlapping of the three molecules could not be obtained with any generated pharmacophore. Nevertheless, and despite its weak significance, a model involving two hydrogen bond acceptors and one hydrophobic group allows the mapping of both V and IA (Figure 6 D). The pharmacophore generations carried out on the single molecules and on the subsets based on odor profiles combined with the pharmacophore comparisons shed light on the structural characteristics of the odorants of the mixtures. Not surprisingly, the pharmacophores generated by molecules of the RC mixture do not share common distances between hydrogen bond acceptors and hydrophobic features. The comparisons in pairs confirm a poor overlap of the hypotheses hyp-V-c, hyp-IA-c and hyp-F-c. While reliable models, as well as good alignments of the molecules, were obtained using the V-s and IA-s subsets, F is the only molecule that is not mapped on the hyp-F-s model generated from the F-s subset. Obviously, no mapping was obtained between the hyp-F-c and hyp-F-s models. Indeed, the relative positions of the features and the interfeature distances between aromatic and hydrophobic features of the hyp-F-c model are very different from those of hyp-F-s (Supplementary Table 7 and Supplementary Table 9). This could suggest less efficient binding to some ORs and be related to the weak “raspberry” odor note described in several databases (Arctander, 1969; Leffingwell, 2013).

Such a result would suggest that all the components of the RC mixture do not share a common binding mode to the peripheral level of the olfactory system. However, how then is it possible to explain the homogeneous perception of the RC mixture? Several points can be noted.

First, the grenadine syrup, or red cordial, was originally made from pomegranate pulp, sugar and water. Recently, some red fruits, especially raspberries, have replaced pomegranate pulp in the recipe, and sometimes vanilla and/or lemon juice are added. A hypothesis could be that the mixture mimics the grenadine syrup odor by using molecules carrying sweet and raspberry odors. This would be in accordance with the use of vanilla, frambinone, beta-ionone and beta-damascenone as components of the mixture. Unfortunately, such hypothesis conflicts with the experimental study that highlighted the key role of isoamyl acetate, of which the odor is lacking the raspberry note, while frambinone intervenes to a lesser extent, and last, bI, bD and EA should be limited to simple improvement in the organoleptic quality. To summarize, the configural process of grenadine syrup odor perception of the RC mixture is consequently based on the vanilla-sweet-chocolate notes of V and fruity-pear-banana-winey notes of IA. The existence of the unique pharmacophore model mapping together V and IA suggests that these two molecules could nonetheless have a common interaction mode at the peripheral level of the olfactory system.

A second point that should be underlined is that to our knowledge, no grenadine syrup or red cordial odor notes are reported in the available databases. Moreover, there is no molecule carrying both vanilla and raspberry odor notes in the base used in the present study. Taken together, these two observations suggest that grenadine syrup odor is based on a learning process and that its image and its olfactory identity result first and foremost in the integration of the olfactory signal at the brain level. Nevertheless, because the first step of the olfactory signal takes place by activation of OR(s), a better knowledge of these targets

at the peripheral level remains necessary to progress in the understanding of such blending perception.

Conclusion

In summary, the pharmacophore approach that was applied to the two mixtures studied allowed us to propose the possibility of a common binding site for the two components of WL-IA. As a consequence, a competitive binding mechanism might occur at the OR common site(s). Conversely, such a possibility seems unclear for the components of the RC mixture. Nevertheless, the term “no common site” does not mean “no common OR target” because several binding sites can exist on the same receptors, such as orthosteric and allosteric binding sites. To test these assumptions, in vitro functional tests first performed on mouse ORs identified by RNA-seq are now in progress.

Through the present study of two specific mixtures, we demonstrated an improvement in the splitting of elements by increasing the dimensionality of the UMAP space from 2D to 3D, as well as the pertinence of the use of clustering based on 3D-UMAP coordinates. If the clustering performed using the k-means method on the basis of 3D coordinates in the 3D-UMAP space provided satisfactory results, the SOM method has been demonstrated to be the most appropriate method in the case of odorants involved in the present work. However, this statement was not seen as absolute. Indeed, k-means clustering could provide more significant results for other odorants or groups of odorants. Moreover, in a broader context, it could be interesting to focus on elements that do not belong to the intersection of clusters defined by both k-means and SOM calculations.

It remains to be investigated how refining odor-structure relationships inside the clusters, since each cluster probably encompass various subsets of odorants that would be homogenous with regard to their structures and carried odor notes. It could be efficient to split the clusters into smaller groups or to rely on the network of co-occurrences between the odor notes; the two ways might be obviously carried out as complementary.

Beyond the context of the present study of two specific mixtures, our study highlights that the approach involving the splitting of elements in 3D-UMAP space followed by gathering in clusters of increasing levels of breakdown constitutes a powerful tool to explore the odor-structure relationships.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Author Contributions

Marylène Rugard: software, data curation, formal analysis, writing-original draft preparation; Anne Tromelin: conceptualization, methodology, data curation, formal analysis, funding acquisition, supervision, writing-review & editing; Karine Audouze: conceptualization, supervision, writing-review.

Funding

The authors received funding for this work from Agence Nationale de la Recherche, ANR-18-CE21-0006, project MULTIMIX (https://anr.fr/en).

List of abbreviations

UMAP: uniform manifold approximation and projection
SOM: self-organizing map
QSAR: Quantitative-Structure-Activity Relationships
V: Vanillin
IA: isoamyl acetate
F: frambinone (other names: oxanone, raspberry ketone)
EA: ethyl acetate
bD: beta-damascenone
bI: beta-ionone
WL: whiskey lactone
RC mixture: red cordial mixture

Acknowledgments

Thanks are due to Thomas Jaylet, which has carried out the initial UMAP studies during his master internship in the team T3S Inserm UMR S-1124.

References

Addinsoft (2022). XLSTAT statistical and data analysis solution. Paris, France. Paris, France. Available: https://www.xlstat.com.

Arctander, S. (1969). Perfume and Flavor Chemicals (Aroma Chemicals) Vol 1 and 2. Carol Stream, Illinois, USA: Allured Publishing Corporation.

Atanasova, B., Thomas-Danguin, T., Langlois, D., Nicklaus, S., and Etievant, P. (2004). Perceptual interactions between fruity and woody notes of wine. Flavour Fragr. J. 19(6), 476-482. doi: 10.1002/ffj.1474.

Berglund, B., Berglund, U., and Lindvall, T. (1976). Psychological processing of odor mixtures. Psychological Review 83(6), 432-441. doi: 10.1037/0033-295x.83.6.432.

Buck, L.B. (1996). Information coding in the vertebrate olfactory system. Annu Rev Neurosci 19, 517-544. doi: 10.1146/annurev.ne.19.030196.002505.

Dixon, S.L., Smondyrev, A.M., Knoll, E.H., Rao, S.N., Shaw, D.E., and Friesner, R.A. (2006). PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20(10-11), 647-671. doi: 10.1007/s10822-006-9087-6.

El Mountassir, F., Belloir, C., Briand, L., Thomas-Danguin, T., and Le Bon, A.-M. (2016). Encoding odorant mixtures by human olfactory receptors: Encoding odorant mixtures by olfactory receptors. Flavour Fragr. J. 31(5), 400-407. doi: 10.1002/ffj.3331.

Gund, P. (1977). "Three-Dimensional Pharmacophoric Pattern Searching," in Progress in Molecular and Subcellular Biology, eds. F.E. Hahn, H. Kersten, W. Kersten & W. Szybalski. (Berlin, Heidelberg: Springer Berlin Heidelberg), 117-143.

Holley, A. (2006). Système olfactif et neurobiologie. Terrain (47), 107-122. doi: 10.4000/terrain.4271.

Kaushik, M., and Mathur, B. (2014). Comparative study of K-means and hierarchical clustering techniques. Int. J. Software Hardware Res. Eng. 2(6), 93-98.

Kay, L.M., Crk, T., and Thorngate, J. (2005). A Redefinition of Odor Mixture Quality. Behavioral Neuroscience 119(3), 726-733. doi: 10.1037/0735-7044.119.3.726.

Khedkar, S., Malde, A., Coutinho, E., and Srivastava, S. (2007). Pharmacophore Modeling in Drug Discovery and Development: An Overview. MC 3(2), 187-197. doi: 10.2174/157340607780059521.

Kini, A., and Firestein, S. (2001). The Molecular Basis of Olfaction. CHIMIA International Journal for Chemistry.

Kohonen, T. (1998). The self-organizing map. Neurocomputing 21(1-3), 1-6. doi: 10.1016/s0925-2312(98)00030-7.

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks 37, 52-65. doi: 10.1016/j.neunet.2012.09.018.

Le Berre, E., Jarmuzek, E., Béno, N., Etiévant, P., Prescott, J., and Thomas-Danguin, T. (2010). Learning Influences the Perception of Odor Mixtures. Chem. Percept. 3(3-4), 156-166. doi: 10.1007/s12078-010-9076-y.

Leach, A.R., Gillet, V.J., Lewis, R.A., and Taylor, R. (2010). Three-Dimensional Pharmacophore Methods in Drug Discovery. J. Med. Chem. 53(2), 539-558. doi: 10.1021/jm900817u.

Leffingwell (2013). Flavor-Base 9th Ed. Available: http://www.leffingwell.com/flavbase.htm.

Mainland, J.D., Li, Y.R., Zhou, T., Liu, W.L.L., and Matsunami, H. (2015). Human olfactory receptor responses to odorants. Sci Data 2, 150002. doi: 10.1038/sdata.2015.2.

Malnic, B., Hirono, J., Sato, T., and Buck, L.B. (1999). Combinatorial receptor codes for odors. Cell 96(5), 713-723. doi: 10.1016/s0092-8674(00)80581-4.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv 3, 861. doi: 10.21105/joss.00861.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. 3. doi: 10.21105/joss.00861.

Mingoti, S.A., and Lima, J.O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. Eur. J. Oper. Res. 174(3), 1742-1759. doi: 10.1016/j.ejor.2005.03.039.

Murthy, V.N. (2011). Olfactory Maps in the Brain. Annu Rev Neurosci 34(1), 233-258. doi: 10.1146/annurev-neuro-061010-113738.

Peterlin, Z., Firestein, S., and Rogers, M.E. (2014). The state of the art of odorant receptor deorphanization: A report from the orphanage. J. Gen. Physiol. 143(5), 527-542. doi: 10.1085/jgp.201311151.

R Core Team (2021). R: A Language and Environment for Statistical Computing. Available: https://www.R-project.org/.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D., et al. (2019). Clustering algorithms: A comparative approach. PLoS ONE 14(1). doi: 10.1371/journal.pone.0210236.

Rugard, M., Jaylet, T., Taboureau, O., Tromelin, A., and Audouze, K. (2021). Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. PLoS ONE 16(5), e0252486. doi: 10.1371/journal.pone.0252486.

Schaller, D., Šribar, D., Noonan, T., Deng, L., Nguyen, T.N., Pach, S., et al. (2020). Next generation 3D pharmacophore modeling. WIREs Comput Mol Sci 10(4). doi: 10.1002/wcms.1468.

Sinding, C., Thomas-Danguin, T., Chambault, A., Béno, N., Dosne, T., Chabanet, C., et al. (2013). Rabbit Neonates and Human Adults Perceive a Blending 6-Component Odor Mixture in a Comparable Manner. PLoS ONE 8(1), e53534. doi: 10.1371/journal.pone.0053534.

The Good Scents Company. Available: http://www.thegoodscentscompany.com.

Thomas-Danguin, T., Sinding, C., Romagny, S.b., El Mountassir, F., Atanasova, B., Le Berre, E., et al. (2014). The perception of odor objects in everyday life: a review on the processing of odor mixtures. Front. Psychol. 5. doi: 10.3389/fpsyg.2014.00504.

Touhara, K. (2002). Odor discrimination by G protein-coupled olfactory receptors. Microsc Res Tech 58(3), 135-141. doi: 10.1002/jemt.10131.

Tromelin, A., Koensgen, F., Audouze, K., Guichard, E., and Thomas-Danguin, T. (2020). Exploring the Characteristics of an Aroma-Blending Mixture by Investigating the Network of Shared Odors and the Molecular Features of Their Related Odorants. Molecules 25(13), 3032. doi: 10.3390/molecules25133032.

Tropsha, A., and Wang, S.X. (2007). "QSAR Modeling of GPCR Ligands: Methodologies and Examples of Applications," in GPCRs: From Deorphanization to Lead Structure Identification, eds. H. Bourne, R. Horuk, J. Kuhnke & H. Michel. (Berlin, Heidelberg: Springer Berlin Heidelberg), 49-74.

van Drie, J. (2003). Pharmacophore Discovery - Lessons Learned. CPD 9(20), 1649-1664. doi: 10.2174/1381612033454568.

Wehrens, R., and Buydens, L.M.C. (2007). Self- and super-organizing maps in R: The kohonen package. J. Stat. Softw. 21(5), 1-19.

Wermuth, G., Ganellin, C.R., Lindberg, P., and Mitscher, L.A. (1998). Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl. Chem. 70(5), 1129-1143.

Xiao, Y.D., Clauset, A., Harris, R., Bayram, E., Santago, P., and Schmitt, J.D. (2005). Supervised self-organizing maps in drug discovery. 1. Robust behavior with overdetermined data sets. J. Chem Inf. Model. 45(6)**,** 1749-1758. doi: 10.1021/ci0500839.

Yang, S.Y. (2010). Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov. Today 15(11-12)**,** 444-450. doi: 10.1016/j.drudis.2010.03.013.

Zarzo, M., and Stanton, D.T. (2009). Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. Attention, Perception & Psychophysics 71(2)**,** 225-247. doi: 10.3758/app.71.2.225.

Supplementary Material

Zip file Rugard-et-al Supplementary Material.zip (Supplementary Figure 1.tif, Supplementary Figure 2.tif, Supplementary Table 1.pdf, Supplementary Table 2.pdf, Supplementary Table 3.pdf, Supplementary Table 4.xlsx, Supplementary Table 5.xlsx, Supplementary Table 6.pdf, Supplementary Table 7.pdf, Supplementary Table 8.pdf, Supplementary Table 9.pdf).

Data Availability Statement

Data are available under all supplementary tables and files. Data compiled from the databases "The Good Scents Company" (access 23/01/19) and "Flavor Base" (9th Edition) are available upon request.

III. Prédictions de cibles biologiques des molécules odorantes : le nouvel odorome

La désorphanisation des récepteurs olfactifs constitue une étape importante de la compréhension du système olfactif et peut être facilitée par des approches *in silico*. De plus, les récepteurs olfactifs interviennent dans des processus biologiques au-delà du système olfactif (Tong et al., 2021). Plusieurs récepteurs olfactifs sont exprimés dans les testicules, le foie, le pancréas, les reins ou encore les poumons (Kim et al., 2015b; Spehr et al., 2003b; K. Zhang et al., 2021) et pourraient donc constituer des biomarqueurs ou cibles thérapeutiques pour certaines pathologies.

Nous avons développé une approche basée sur les réseaux afin de mieux comprendre les effets biologiques des molécules odorantes et de prédire de nouvelles interactions entre molécules odorantes et récepteurs olfactifs. Un réseau prédictif d'interactions entre molécules odorantes et récepteurs olfactifs a été développé à partir d'interactions connues entre des deux groupes. D'autre part, un réseau RO-RO a été construit à partir de ces interactions connues, qui a ensuite été enrichi avec des interactions protéines-protéines. Des analyses d'enrichissement biologique associées à l'utilisation de la base de données DisgeNET ont permis d'identifier les voies de signalisation et les pathologies dans lesquelles les récepteurs olfactifs et protéines étaient impliqués. Au final, des milliers d'interactions ont été prédites entre molécules odorantes et récepteurs olfactifs ont pu être prédites et plusieurs processus biologiques tels que la maladie d'Alzheimer, le lupus systémique ou encore l'athérosclérose ont également été identifiés.

Article en cours de rédaction

Prédictions de cibles biologiques des molécules odorantes : le nouvel odorome

Marylène Rugard[1], Anne Tromelin[2], Karine Audouze[1]

[1] Université de Paris Cité, T3S, Inserm UMR S-1124, F-75006 Paris, France
[2] Centre des Sciences du Goût et de l'Alimentation, Institut Agro Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France

I. Introduction

Le système olfactif des mammifères est capable de discriminer des milliers de molécules odorantes de différentes structures moléculaires (Dinu et al., 2020; Sankaran et al., 2012). Cette capacité repose sur le code combinatoire qui implique qu'un récepteur olfactif (RO) peut être activé par différentes molécules odorantes et qu'une même molécule odorante peut activer différents récepteurs olfactifs (Malnic et al., 1999). Donc l'identification des ligands des récepteurs olfactifs constitue un point clé dans la compréhension du processus d'olfaction. Cependant à ce jour moins de 20 % des près de 400 récepteurs olfactifs humains ont été désorphanisés (Sharma et al., 2022). La désorphanisation des récepteurs olfactifs constitue donc une tâche complexe et les approches *in silico* constitue une alternative encourageante dans ce domaine.

En considérant que les molécules odorantes qui activent un même récepteur olfactif ont des structures proches (Malnic et al., 2004), différentes études ont tenté d'identifier les relations entre la structure des odorants et leurs récepteurs en utilisant des modèles de relation quantitative structure-activité (QSAR) (Gabler et al., 2013), des réseaux de neurones (Schmuker et al., 2007) ou encore de l'amarrage moléculaire (docking) (Schmiedeberg et al., 2007), une technique qui permet de comprendre et prévoir la reconnaissance moléculaire d'une petite molécule par une macromolécule cible, à la fois sur le plan structurel, en prévoyant les modes de liaison et l'affinité de liaison (Morris and Lim-Wilby, 2008).

Néanmoins, au-delà de leurs rôles dans l'olfaction, les récepteurs olfactifs interviennent dans divers processus biologiques (Di Pizio et al., 2019; Drew, 2022; Foster et al., 2014; Tong et al.,

2021). En effet, l'expression de plusieurs récepteurs olfactifs a été identifié dans les testicules, le foie, le pancréas, les reins, les poumons, le cœur, et encore d'autres organes (Kim et al., 2015; Spehr et al., 2003; Zhang et al., 2021). L'identification du rôle de l'expression ectopique des récepteurs olfactifs ainsi que de leurs ligands mieux permettrait de fournir des biomarqueurs ou cibles thérapeutiques pour certaines pathologies.

Une première étude basée sur les réseaux avait déjà permis de mieux appréhender l'odorome humain (Audouze et al., 2014). L'odorome est une représentation de l'impact des molécules odorantes sur la santé humaine à travers leurs interactions avec les cibles biologiques (Audouze et al., 2014).

Dans le cadre de la présente étude, nous nous nous sommes appuyés sur cette conception de réseau pour l'élargir en intégrant davantage d'informations concernant les interactions entre les récepteurs olfactifs et les molécules odorantes. L'objectif de l'étude était de prédire de nouvelles interactions entre molécules odorantes et récepteurs olfactifs ainsi que d'identifier des processus biologiques dans lesquels les récepteurs olfactifs seraient impliqués. Pour cela, nous avons développé un réseau d'associations entre récepteurs olfactifs (réseau RO-RO) à partir d'un premier réseau d'interactions connues entre molécules odorantes et récepteurs olfactifs. Ce réseau a ensuite été enrichi avec des interactions protéines-protéines dans le but d'identifier des processus autres que ceux associés à l'olfaction. Le réseau enrichi a ensuite été utilisé afin de réaliser une classification permettant d'identifier des clusters au sein du réseau. A partir des clusters, un enrichissement biologique a permis d'identifier les voies de signalisation dans lesquelles les récepteurs olfactifs et protéines étaient impliqués. Nous avons ensuite utilisé la base de données DisGeNET afin de déterminer les pathologies humaines dans lesquelles intervenaient les protéines du réseau. Cette nouvelle version de l'odorome regroupe les interactions connues entre molécules odorantes et récepteurs olfactifs de quatre espèces de mammifères, qui sont quatre fois plus nombreuses que celles du précédent modèle (Audouze et al., 2014).

Ainsi finalement, plusieurs milliers d'interactions entre molécules odorantes et récepteurs olfactifs ont pu être prédites. De plus, nous avons pu mettre en lumière l'implication de certains récepteurs olfactifs dans des processus biologiques tels que la maladie d'Alzheimer, le lupus systémique ou encore l'athérosclérose.

II. Matériel et méthode

A. Jeu de données

Afin de créer le réseau odorome V2, nous avons développé un jeu de données répertoriant des associations entre des molécules odorantes et les récepteurs olfactifs (RO) qu'elles activent. Toutes ces associations (Tableau S1) ont été extraites à partir de la curation manuelle de la littérature, de trois bases de données « OdorDB » ("OdorDB, SenseLab. Available online: https://senselab.med.yale.edu/OdorDB/," n.d.) (02/2020), « OlfactionDB » (Modena et al., 2012) (02/2020) et « ODORactor » (Liu et al., 2011) (03/2020) ainsi qu'à partir d'un outil de text mining appelé AOP-helpFinder permettant une curation automatique de la littérature (Jornod et al., 2020). AOP-helpFinder est un programme basé sur l'intelligence artificielle et développé dans le but d'identifier, à partir des résumés des articles scientifiques de la base de données PubMed, des associations fiables entre les événements biologiques et une molécule chimique d'intérêt. Pour cela, le programme utilise la théorie des graphes afin de déterminer la distance qui sépare l'évènement biologique de la molécule d'intérêt dans le texte (Carvaillo et al., 2019). L'outil développé pour un projet de toxicologie a été adapté pour être utilisé dans le cadre de mon étude. Les évènements biologiques ne sont plus recherchés, mais à la place les récepteurs olfactifs associés aux molécules odorantes.

De ces différentes informations, nous avons choisi d'exclure les interactions relatives aux espèces non mammifères (le moustique, le nématode, la drosophile et le poisson medaka) et les interactions relatives à une espèce mammifère pour lesquelles la quantité de données était trop faible pour qu'elle ait un poids suffisant pour améliorer le modèle (1 à 3 interactions pour les espèces du chimpanzé, du macaque, du singe écureuil et de l'orang-outan). De ce fait, nous avons conservé quatre espèces de mammifères : l'homme, la souris, le rat et le chien, qui font partie des espèces mammifères les plus étudiées (Wackermannová et al., 2016). Au total, 3778 interactions ont été extraites (soit 2840 de plus que le modèle précédent) entre 793 molécules odorantes uniques et 535 récepteurs olfactifs uniques. Une large majorité de ces interactions correspondent aux espèces de l'homme et la souris puisque 1506 interactions concernent l'homme, 1852 concernent la souris et seulement 318 et 102 interactions pour le chien et le rat respectivement. Ce jeu de données peut être représenté par un modèle de réseau bipartite

reliant donc deux ensembles de nœuds différents, les molécules odorantes et les récepteurs olfactifs.

B. Mise en place du réseau RO-RO

L'analyse des réseaux bipartites est souvent réalisée à partir de leurs projections monopartites. Dans ce but, à partir du réseau bipartite, nous avons construit un réseau monopartite reliant uniquement entre eux les récepteurs olfactifs des quatre espèces (réseau RO-RO). Outre l'analyse du réseau bipartite, le réseau RO-RO sera utilisé pour l'identification d'associations prédictives entre molécules odorantes et récepteurs olfactifs. Le réseau monopartite est donc généré en conservant uniquement les nœuds représentants les RO et en établissant un lien entre deux RO s'ils sont connectés à au moins une même molécule odorante (Vogt and Mestres, 2019).

C. Intégration de données biologiques, classification et enrichissement biologique

Afin de pouvoir identifier les différents processus biologiques dans lesquels les molécules odorantes et récepteurs olfactifs pourraient être impliqués, nous avons dans un premier temps enrichi le réseau RO-RO à l'aide d'interactions protéine-protéine (PPI) relatives à chaque espèce. Ces PPI sont des liaisons physiques entre au moins deux protéines, provoquées par diverses perturbations biologiques (Hao et al., 2016). Ces interactions sont donc essentielles dans les processus biologiques au cours desquels elles participent (Bu, 2003). Les PPI ont été extraites pour chaque espèce à partir des trois bases de données STRING (Szklarczyk et al., 2021), Inweb (Lage et al., 2007) et Biogrid (Oughtred et al., 2021).

Ensuite, nous avons réalisé un clustering sur le réseau RO-RO enrichi des PPI. Le processus de clustering permet d'identifier dans le réseau des sous-graphes denses ou modules ayant une fonction biologique possible. Pour cette étape, nous avons utilisé deux algorithmes différents : le « Markov clustering » (MCL) (van Dongen, 2000) et le « Molecular Complex Detection » (MCODE) (Bader and Hogue, 2003), afin d'identifier les potentiels modules présents dans le réseau.

L'algorithme MCL identifie les clusters en simulant des flux aléatoires au sein d'un réseau. Cet algorithme repose sur l'idée que les régions denses des réseaux correspondent à des régions dans lesquelles le nombre de chemins de longueur k est relativement grand, pour un petit k. Les flux aléatoires de longueur k ont donc une probabilité plus élevée pour les chemins dont le début et la fin se trouvent dans la même région dense que pour les autres chemins (van Dongen, 2000). Autrement dit, la probabilité d'un chemin aléatoire reliant deux nœuds d'un même cluster est plus élevée que celle d'un chemin aléatoire reliant deux nœuds de deux clusters différents. Donc, en effectuant des flux aléatoires sur le réseau, il est ainsi possible de découvrir où le flux a tendance à se rassembler, et donc, où se trouvent les clusters.
L'algorithme MCODE détecte les régions densément connectées des réseaux. Pour cela, les nœuds du réseau sont pondérés par la densité du voisinage local. Ensuite à partir du réseau dont les nœuds sont pondérés, un complexe se crée avec le sommet ayant le poids le plus élevé et se déplace ensuite vers l'extérieur à partir du sommet de départ, en incluant les sommets du complexe dont le poids est supérieur à un seuil donné. Si un sommet est inclus, ses voisins sont vérifiés de la même manière afin d'identifier s'ils font partie du complexe. Le processus s'arrête lorsque plus aucun sommet ne peut être ajouté au complexe et se répète alors pour le prochain sommet pondéré le plus élevé du réseau. Et les régions les plus denses du réseau sont ainsi identifiées (Bader and Hogue, 2003).

L'identification des clusters a permis de procédé par la suite à un enrichissement biologique réalisé à l'aide du serveur web DAVID (Sherman et al., 2022). En effet, le serveur DAVID permet de réaliser une analyse d'enrichissement utilisant un test exact de Fisher modifié[1] pour identifier les termes d'annotation les plus surreprésentés associés à une liste de gènes ou de protéines. Ainsi, il est possible d'obtenir une liste des principaux thèmes biologiques associés aux gènes/protéines de la liste de départ. Il peut utiliser plusieurs classes de catégories d'annotation trouvées dans la base de connaissances DAVID comme l'ontologie des gènes, les voies de signalisation ou encore la littérature (Sherman et al., 2022). Ainsi, à partir de chaque cluster, la liste des protéines (PPI et récepteurs olfactifs) relatives à une même espèce a été

[1] Le calcul de la p-value se fait en soustrayant un gène à la liste de gènes ou de protéines

utilisée pour réaliser l'enrichissement biologique via DAVID. Donc pour chaque cluster, un enrichissement biologique est réalisé par espèce.

Afin de compléter la liste des processus biologiques probables pouvant impliquer les récepteurs olfactif et molécules odorantes, nous avons étendu le réseau en intégrant les pathologies associées aux protéines. Nous nous sommes limités aux pathologies associées aux protéines humaines, et avons extrait ces informations de la base de données DisGeNET (Pinero et al., 2015). Les informations répertoriée dans cette base de données sont issues de quatre sources différentes : des ressources curées (C) par des experts (CTD, UNIPROT…), des ressources issues d'expériences sur des modèles animaux (M), des ressources déduites d'associations variant-maladie ou de la base de données « Human Phenotype Ontology » qui fournit des informations sur les anomalies phénotypiques rencontrées dans les maladies humaines (I) et également des ressources issues de la littérature par curation automatique (L). Un score situé entre 0 et 1, est ensuite attribué à chaque association gène-pathologie. Il est calculé comme suit :

Score DisGeNET = C + M + I + L

Où C = 0,6 si $N_{sources} > 2$,

0,5 si $N_{sources} = 2$,

0,3 si $N_{sources} = 1$,

0 autrement

M = 0,2 si $N_{sources} > 0$,

0 autrement

I = 0,1 si $N_{sources} > 0$,

0 autrement

L = 0, si $N_{pubs} > 9$,

$N_{pubs} \times 0.01$ si $N_{pubs} < 9$

Avec $N_{sources}$, le nombre de sources supportant l'association gène-pathologie parmi les différentes ressources C, M et I et N_{pubs}, le nombre de publications supportant l'association gène-pathologie parmi les ressources L.

Nous avons ainsi choisi de conserver les associations ayant un score minimum de 0,5.

D. Prédiction des interactions entre récepteurs olfactifs et molécules odorantes

Les associations prédictives entre molécules odorantes et récepteurs olfactifs ont été déterminées à partir du réseau monopartite d'une part et également du réseau bipartite d'autre part. En effet, dans un premier temps un score a été attribué à chaque association du réseau RO-RO (Audouze et al., 2010), afin de quantifier ces associations. Pour cela, nous avons choisi le score pull-down qui est initialement utilisé pour évaluer la qualité des données d'interaction entre protéines issues d'expériences complexes (de Lichtenberg et al., 2005) telles que la méthode de purification par affinité en tandem couplée à la spectrométrie de masse (Gavin et al., 2002). Récemment, il a été utilisé pour quantifier des associations entre évènements médicamenteux indésirables (des événements survenant pendant le traitement médicamenteux et ayant des effets négatifs sur l'organisme) dans le but de prédire les évènements médicamenteux indésirables causés par des médicaments (Wu et al., 2020).
Le score pull-down (Spull) entre deux RO a été calculé selon l'équation suivante :

$$Spull_{RO1\text{-}RO2} = \log10\ [(N_1 \cap N_2)\ (N_1 \cup N_2) / (N_1 + 1)\ (N_2 + 1)]$$

Où N_1 représente le nombre de molécules odorantes associées à RO1, N_2 représente le nombre de molécules odorantes associées à RO2, $(N_1 \cap N_2)$ représente le nombre de molécules odorantes communes à RO1 et RO2 et $(N_1 \cup N_2)$ représente le nombre total de molécules odorantes associées à RO1 et associées à RO2. Plus le score est proche de zéro et plus l'association peut être considérée comme fiable. Pour les associations RO-RO dont le score est le plus élevé, il est possible de prédire des récepteurs olfactifs pour une molécule odorante en considérant que si une molécule odorante est un ligand d'un des deux récepteurs de l'association, alors le second récepteur est une cible potentielle de cette molécule odorante.
Le protocole global de l'étude est décrit dans la Figure 13.

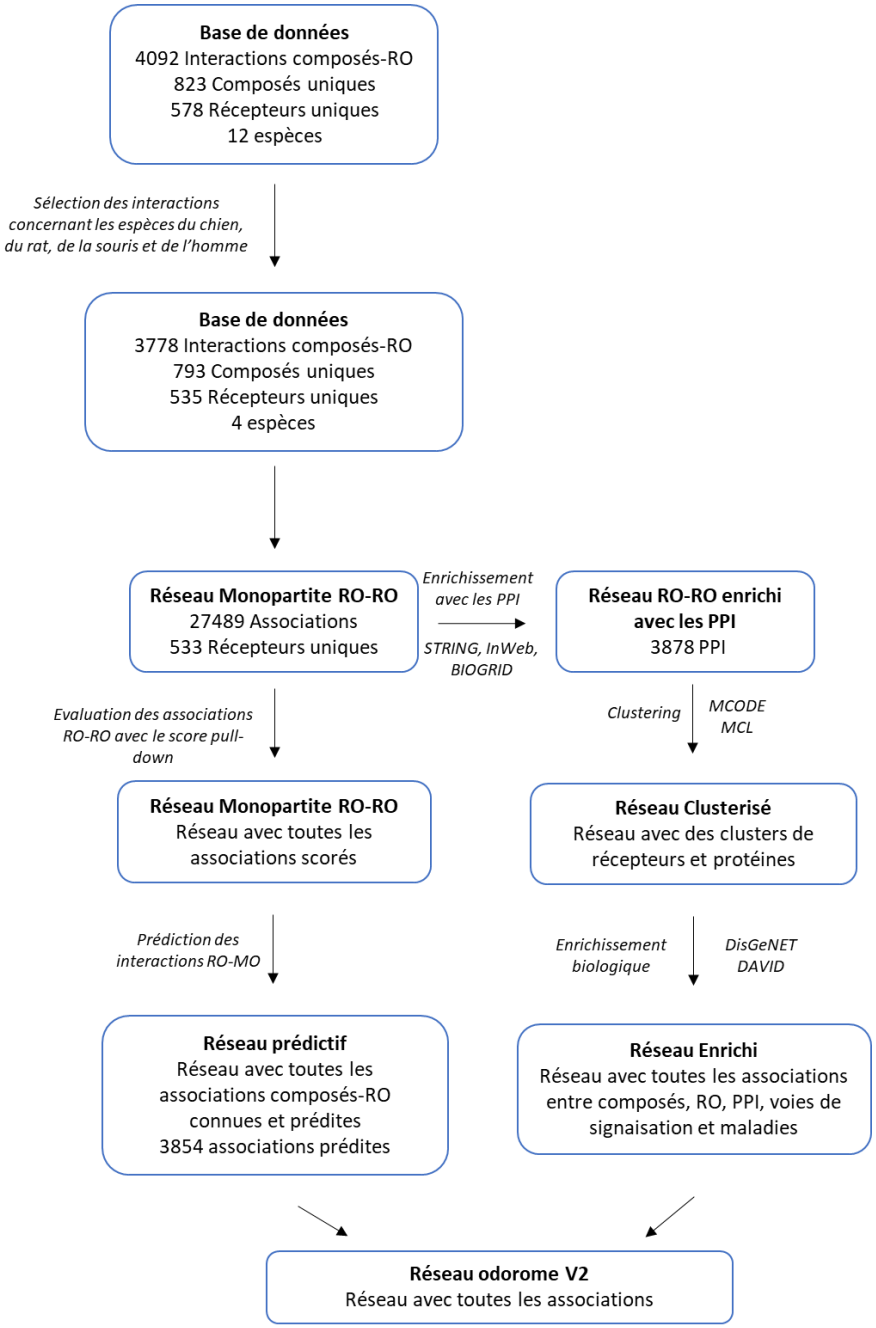


Figure 13 : Protocole global de l'étude

III. Résultats

Afin d'identifier les effets biologiques des molécules odorantes, nous avons construit un nouvel odorome V2 permettant de déterminer de potentielles cibles biologiques pour les molécules odorantes ainsi que les processus biologiques dans lesquels elles sont susceptibles d'intervenir.

A. Generation de l'odorome V2

Les premières données intégrées au modèle ont été les interactions entre molécules odorantes et récepteurs olfactifs. En considérant que deux récepteurs olfactifs qui interagissent avec une même molécule odorante, peuvent être associés l'un à l'autre, un réseau RO-RO a ainsi pu être établi. Le réseau RO-RO a ensuite été enrichi avec des interactions protéine-protéine issues des bases de données STRING, Inweb et Biogrid (Figure 14). Aucune PPI n'a pu être identifiée pour les récepteurs olfactifs canins mais au total 228 PPI ont été identifiées pour les récepteurs olfactifs des trois autres espèces. Nous avons ensuite voulu identifier les clusters présents au sein du réseau RO-RO enrichi avec les PPI.

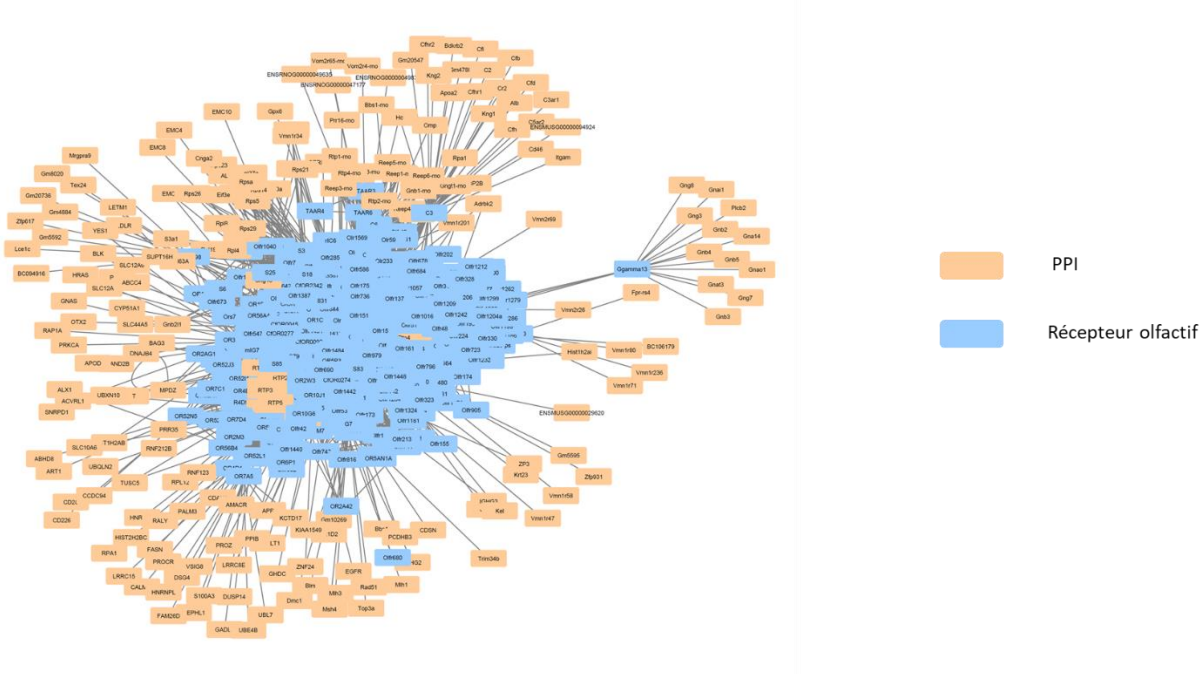


Figure 14 : Réseau RO-RO enrichi des PPI

B. Classification du réseau

Afin d'identifier ces clusters, nous avons réalisé une classification du réseau avec les deux algorithmes MCL et MCODE (Tableau 1). L'algorithme MCL a permis de clusteriser tous les nœuds du réseau au sein de 12 clusters différents. Et seuls 538 nœuds ont pu être attribués aux 13 clusters formés par l'algorithme MCODE. La répartition des éléments au sein des

clusters diffère entre les deux algorithmes puisque 90 % des éléments clusterisés avec l'algorithme MCODE sont dans les 2 plus grands clusters alors que les 2 plus grands clusters identifiés à l'aide de l'algorithme MCL représentent 58% des éléments clusterisés.

Tableau 1 : Répartition des éléments du réseau RO-RO enrichi avec les PPI selon les algorithmes de classification

Clusters	Nombre d'éléments avec la classification MCL	Nombre d'éléments avec la classification MCODE
1	60	154
2	322	331
3	121	9
4	14	8
5	95	6
6	6	5
7	5	4
8	2	3
9	4	3
10	3	3
11	110	3
12	21	6
13	/	3

C. Enrichissement biologique et intégration des pathologies

Le manque d'éléments non assignés à des clusters ainsi que la trop grosse hétérogénéité dans la répartition des éléments avec l'algorithme MCODE nous ont orienté vers le choix des clusters établis avec l'algorithme MCL pour l'enrichissement biologique.

A partir de chaque cluster, des listes des protéines appartenant à chaque espèce a été utilisée pour réaliser un enrichissement biologique par espèce. Cette étape a permis de mettre en lumière 35 voies de signalisation dans lesquelles les protéines et récepteurs olfactifs sont impliqués (Tableau S2). De plus, l'utilisation de la base de données DisGeNET a permis de déterminer l'implication des protéines humaines du réseau dans 126 pathologies (Tableau 2, Tableau S3).

Tableau 2 : Les dix associations protéine-pathologie chez l'espèce humaine les plus fiables

Protéine humaine	Pathologie	Score DisGeNET
HRAS	Syndrome de Costello	1,000
GNAS	Pseudopseudohypoparathyroïdie	1,000
PPIB	Ostéogénèse imparfaite de type IX	0,930
REEP1	Paraplégie spastique autosomique dominante de type 31	0,920
AMACR	Déficit en alpha-méthylacyl-CoA racémase	0,910
APP	Maladie d'Alzheimer	0,900
LDLR	Hypercholestérolémie familiale	0,900
LDLR	Hypercholestérolémie	0,900
BAG3	Cardiomyopathie dilatée	0,900
SLC12A6	Neuropathie avec agénésie du corps calleux	0,900

D. Prédiction de nouveaux ligands

A l'aide du score pull-down qui a été attribué aux associations du réseau RO-RO, nous avons pu établir de potentielles interactions entre les diverses molécules odorantes et récepteurs olfactifs (Figure 15). Ainsi 3854 nouvelles interactions ont pu être prédites dont 3 pour des récepteurs olfactifs du rat, 952 pour ceux de l'homme, 1404 pour ceux du chien et 1495 pour ceux de la souris. Dans le cadre du projet ANR MULTIMIX duquel dépend cette étude, les interactions prédites avec les meilleurs scores pull-down pourront être validées par mesure in vitro de l'activité fonctionnelle des récepteurs olfactifs produits par expression hétérologue dans des cellules HEK. Une des techniques couramment utilisées est dite « GloSensor™ cAMP assay » (Wang et al., 2022). Cette technique permet de détecter l'activité des RCPG via un changement de la concentration intracellulaire d'adénosine monophosphate cyclique (AMPc). Cette méthode utilise une luciférase qui contient un fragment de protéine de liaison à l'AMPc. La liaison de l'AMPc à la luciférase induit un changement de sa conformation, ce qui entraîne une augmentation de la production de lumière qui permet de mesurer l'activité des ligands au niveau du RCPG étudié (Buccioni et al., 2011).

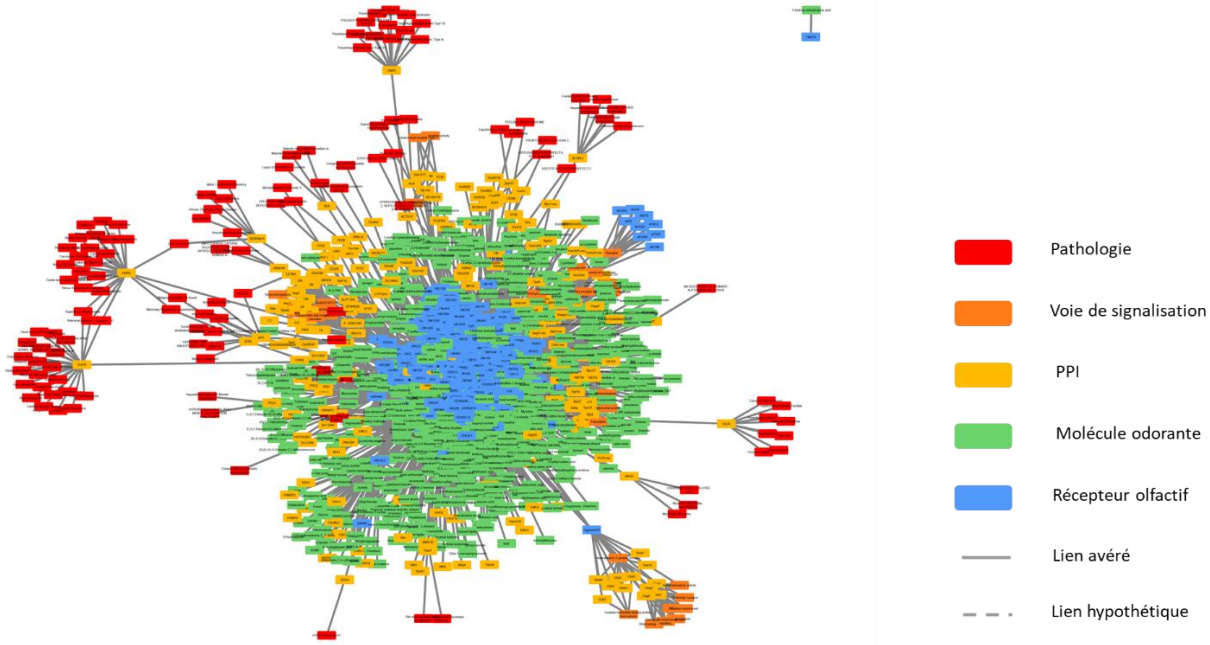


Figure 15 : Réseau odorome. Représentation des interactions connues et prédictes entre récepteurs olfactifs et molécules odorantes, des PPI et des associations avec les pathologies et voies de signalisation.

IV. Discussion

Au cours de cette étude, nous avons développé une nouvelle version de l'odorome avec plus d'interactions connues que le modèle précédent. Ce modèle odorome V2 répertorie au départ près de 4000 interactions avec 1506 interactions concernent l'homme, 1852 concernent la souris, 318 pour le chien et 102 pour le rat). Ce modèle a donc permis d'une part de mettre en évidence divers processus biologiques impliquant les récepteurs olfactifs et a fourni d'autre part de potentiels nouveaux ligands des récepteurs olfactifs.

A partir d'interactions connues entre molécules odorantes extraites de différentes sources d'informations, nous avons développé un réseau d'associations entre récepteurs olfactifs. Ce réseau RO-RO a ensuite été enrichi avec des PPI qui a ensuite été utilisé afin de réaliser une classification. L'objectif de cette classification était d'identifier de potentiels modules pouvant représenter des processus biologiques. A partir des clusters obtenus, un enrichissement biologique a permis d'identifier les voies de signalisation dans lesquelles les récepteurs olfactifs et protéines pourraient être impliqués. La base de données DisGeNet a ensuite été utilisée pour compléter l'enrichissement biologique afin de déterminer les pathologies dans

lesquelles intervenaient les protéines humaines du réseau. Ainsi plusieurs milliers d'interactions entre molécules odorantes et récepteurs olfactifs ont pu être prédites et 161 voies de signalisation et pathologies ont été identifiées.

Ce réseau odorome V2 constitue ainsi une source diverse d'informations concernant les récepteurs olfactifs et molécules odorantes. Des associations ont déjà été observées entre les acteurs du système olfactif et certaines des voies de signalisation et pathologies identifiées. Mais parmi les 10 pathologies les plus probables, nous avons pu retrouver dans la littérature ce type d'associations uniquement pour la maladie d'Alzheimer, ce qui peut suggérer que les autres pathologies pourraient constituer des pistes de recherche innovantes.

En effet, des dérégulations des récepteurs olfactifs ont déjà été remarquées dans des zones cérébrales spécifiques de sujets humains atteints de maladie d'Alzheimer. Les expressions de gènes de récepteurs olfactifs du cortex enthorinal et du cortex frontal ont été étudiées à différents stades de la progression de la maladie. Et, il a été constaté que les niveaux d'expression des ARNm du récepteur OR11H1 augmentaient et ceux des ARNm des récepteurs OR4F4 et OR10G8 diminuaient dans le cortex entorhinal aux stades avancés de la maladie par rapport aux contrôles. De plus, dans le cortex frontal, les niveaux d'expression des ARNm des récepteurs OR4F4 et OR52L1 ont également augmenté (Ansoleaga et al., 2013).

Une autre étude a montré que des troubles olfactifs apparaissaient chez des patients atteints de lupus systémique (Bombini et al., 2018). En effet, une réduction significative de la sensibilité olfactive, de la discrimination et de l'identification des odeurs chez les sujets malades comparés aux sujets sains.

D'autre part, il a également été montré que des récepteurs olfactifs intervenaient dans le processus d'athérosclérose chez la souris. L'octanal est issu de la peroxydation des acides gras qui s'accumule dans la paroi des vaisseaux provoquant l'athérosclérose. Via l'activation du récepteur olfactif Olfr2 situé dans les macrophages vasculaires, l'octanal intervient dans l'augmentation de la production d'interleukines 1β qui est un processus clé de cette maladie. Par ailleurs, chez l'homme, les concentrations circulantes d'octanal sont similaires à celles des modèles murins d'athérosclérose et chez l'homme. Donc la pathologie humaine pourrait impliquer ces mêmes mécanismes (Orecchioni et al., 2022; Rayner and Rasheed, 2022).

Toutes ces études suggèrent que le modèle odorome V2 peut apporter des pistes sérieuses dans la compréhension du rôle des molécules odorantes et des récepteurs olfactifs. En effet, les associations établies entre récepteurs olfactifs, PPI et pathologies peuvent être utilisées afin de mettre en lumière certains mécanismes pathologiques impliquant les récepteurs olfactifs et permettre ainsi d'identifier de potentielles cibles thérapeutiques. Par exemple, le syndrome de Costello est associé à la protéine HRAS (protéine impliquée dans l'activation de la transduction du signal de la protéine Ras, une enzyme catalysant l'hydrolyse de la guanosine triphosphate) qui est elle-même interagit avec le récepteur olfactif humain OR2T10. Si ce récepteur était présent dans les tissus affectés par le syndrome de Costello, il serait alors pertinent de chercher de quelle manière ce récepteur pourrait être impliqué dans cette maladie. Donc, identifier la présence des récepteurs olfactifs associés aux pathologies, dans les tissus affectés par ces mêmes pathologies serait également une information importante à fournir avec le modèle. Pour cela, il serait possible d'étudier par analyse génomique l'expression des gènes des récepteurs olfactifs dans ces tissus.

De plus, les prédictions d'interactions entre les molécules odorantes et récepteurs olfactifs doivent encore être validées expérimentalement par mesure in vitro de l'activité fonctionnelles des récepteurs (Wang et al., 2022). D'autre part, pour une molécule qui interagit avec un ou plusieurs récepteurs olfactifs, il serait également intéressant d'ajouter au modèle, les récepteurs olfactifs qui ne sont pas activés par cette même molécule odorante. Dans un premier temps, cela permettrait d'enrichir le modèle avec de nouvelles informations. Mais dans un second temps, ces données associées aux odeurs des molécules odorantes pourront également être utilisées afin de générer de nouvelles prédictions. En effet, les interactions entre molécules odorantes et récepteurs olfactifs seront converties en vecteurs binaires représentant l'activation ou la non-activation des récepteurs olfactifs. Pour une molécule odorante donnée, un vecteur binaire sera créé indiquant par un « 1 » l'activation d'un récepteur ou par un « 0 » sa non-activation. Les vecteurs binaires seront ensuite utilisés en tant que données d'entrée de réseaux neuronaux afin de prédire l'odeur de molécules odorantes. Toutes les informations fournies par le modèle pourraient alors ensuite être mises à la disposition de la communauté par le biais d'un site web.

V. Références

Ansoleaga, B., Garcia-Esparcia, P., Llorens, F., Moreno, J., Aso, E., Ferrer, I., 2013. Dysregulation of brain olfactory and taste receptors in AD, PSP and CJD, and AD-related model. Neuroscience 248, 369–382. https://doi.org/10.1016/j.neuroscience.2013.06.034

Audouze, K., Juncker, A.S., Roque, F.J.S.S.A., Krysiak-Baltyn, K., Weinhold, N., Taboureau, O., Jensen, T.S., Brunak, S., 2010. Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxicogenomics Networks. PLoS Comput Biol 6, e1000788. https://doi.org/10.1371/journal.pcbi.1000788

Audouze, K., Tromelin, A., Le Bon, A.M., Belloir, C., Petersen, R.K., Kristiansen, K., Brunak, S., Taboureau, O., 2014. Identification of odorant-receptor interactions by global mapping of the human odorome. PLoS One 9, e93037. https://doi.org/10.1371/journal.pone.0093037

Bader, G.D., Hogue, C.W., 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2. https://doi.org/10.1186/1471-2105-4-2

Bombini, M.F., Peres, F.A., Lapa, A.T., Sinicato, N.A., Quental, B.R., Pincelli, Á. de S.M., Amaral, T.N., Gomes, C.C., del Rio, A.P., Marques-Neto, J.F., Costallat, L.T.L., Fernandes, P.T., Cendes, F., Rittner, L., Appenzeller, S., 2018. Olfactory function in systemic lupus erythematosus and systemic sclerosis. A longitudinal study and review of the literature. Autoimmunity Reviews 17, 405–412. https://doi.org/10.1016/j.autrev.2018.02.002

Bu, D., 2003. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research 31, 2443–2450. https://doi.org/10.1093/nar/gkg340

Buccioni, M., Marucci, G., Dal Ben, D., Giacobbe, D., Lambertucci, C., Soverchia, L., Thomas, A., Volpini, R., Cristalli, G., 2011. Innovative functional cAMP assay for studying G protein-coupled receptors: application to the pharmacological characterization of GPR17. Purinergic Signalling 7, 463–468. https://doi.org/10.1007/s11302-011-9245-8

Carvaillo, J.-C., Barouki, R., Coumoul, X., Audouze, K., 2019. Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach. Environ Health Perspect 127, 047005. https://doi.org/10.1289/EHP4200

de Lichtenberg, U., Jensen, L.J., Brunak, S., Bork, P., 2005. Dynamic Complex Formation During the Yeast Cell Cycle. Science 307, 724–727. https://doi.org/10.1126/science.1105103

Di Pizio, A., Behrens, M., Krautwurst, D., 2019. Beyond the Flavour: The Potential Druggability of Chemosensory G Protein-Coupled Receptors. IJMS 20, 1402. https://doi.org/10.3390/ijms20061402

Dinu, V., MacCalman, T., Yang, N., Adams, G.G., Yakubov, G.E., Harding, S.E., Fisk, I.D., 2020. Probing the effect of aroma compounds on the hydrodynamic properties of mucin glycoproteins. Eur Biophys J 49, 799–808. https://doi.org/10.1007/s00249-020-01475-4

Drew, L., 2022. Olfactory receptors are not unique to the nose. Nature 606, S14–S17. https://doi.org/10.1038/d41586-022-01631-0

Foster, S.R., Roura, E., Thomas, W.G., 2014. Extrasensory perception: Odorant and taste receptors beyond the nose and mouth. Pharmacology & Therapeutics 142, 41–61. https://doi.org/10.1016/j.pharmthera.2013.11.004

Gabler, S., Soelter, J., Hussain, T., Sachse, S., Schmuker, M., 2013. Physicochemical vs. Vibrational Descriptors for Prediction of Odor Receptor Responses. Mol Inform 32, 855–865. https://doi.org/10.1002/minf.201300037

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147. https://doi.org/10.1038/415141a

Hao, T., Peng, W., Wang, Q., Wang, B., Sun, J., 2016. Reconstruction and Application of Protein–Protein Interaction Network. IJMS 17, 907. https://doi.org/10.3390/ijms17060907

Jornod, F., Rugard, M., Tamisier, L., Coumoul, X., Andersen, H.R., Barouki, R., Audouze, K., 2020. AOP4EUpest: mapping of pesticides in adverse outcome pathways using a text mining tool. Bioinformatics 36, 4379–4381. https://doi.org/10.1093/bioinformatics/btaa545

Kim, S.-H., Yoon, Y.C., Lee, A.S., Kang, N., Koo, J., Rhyu, M.-R., Park, J.-H., 2015. Expression of human olfactory receptor 10J5 in heart aorta, coronary artery, and endothelial cells and its functional role in angiogenesis. Biochem Biophys Res Commun 460, 404–408. https://doi.org/10.1016/j.bbrc.2015.03.046

Lage, K., Karlberg, E.O., Størling, Z.M., Ólason, P.Í., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., Brunak, S., 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25, 309–316. https://doi.org/10.1038/nbt1295

Liu, X., Su, X., Wang, F., Huang, Z., Wang, Q., Li, Z., Zhang, R., Wu, L., Pan, Y., Chen, Y., Zhuang, H., Chen, G., Shi, T., Zhang, J., 2011. ODORactor: a web server for deciphering olfactory coding. Bioinformatics 27, 2302–2303. https://doi.org/10.1093/bioinformatics/btr385

Malnic, B., Godfrey, P.A., Buck, L.B., 2004. The human olfactory receptor gene family. Proc Natl Acad Sci U S A 101, 2584–2589. https://doi.org/10.1073/pnas.0307882100

Malnic, B., Hirono, J., Sato, T., Buck, L.B., 1999. Combinatorial receptor codes for odors. Cell 96, 713–723. https://doi.org/10.1016/s0092-8674(00)80581-4

Modena, D., Trentini, M., Corsini, M., Bombaci, A., Giorgetti, A., 2012. OlfactionDB: A Database of Olfactory Receptors and Their Ligands. ALS 1, 1–5. https://doi.org/10.5923/j.als.20110101.01

Morris, G.M., Lim-Wilby, M., 2008. Molecular Docking, in: Kukol, A. (Ed.), Molecular Modeling of Proteins, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365–382. https://doi.org/10.1007/978-1-59745-177-2_19

OdorDB, SenseLab. Available online: https://senselab.med.yale.edu/OdorDB/, n.d.

Orecchioni, M., Kobiyama, K., Winkels, H., Ghosheh, Y., McArdle, S., Mikulski, Z., Kiosses, W.B., Fan, Z., Wen, L., Jung, Y., Roy, P., Ali, A.J., Miyamoto, Y., Mangan, M., Makings, J., Wang, Zhihao, Denn, A., Vallejo, J., Owens, M., Durant, C.P., Braumann, S., Mader, N., Li, L.,

Matsunami, H., Eckmann, L., Latz, E., Wang, Zeneng, Hazen, S.L., Ley, K., 2022. Olfactory receptor 2 in vascular macrophages drives atherosclerosis by NLRP3-dependent IL-1 production. Science 375, 214–221. https://doi.org/10.1126/science.abg3067

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., Tyers, M., 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science 30, 187–200. https://doi.org/10.1002/pro.3978

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I., 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015, bav028–bav028. https://doi.org/10.1093/database/bav028

Rayner, K.J., Rasheed, A., 2022. The scent of atherosclerosis. Science 375, 145–146. https://doi.org/10.1126/science.abn4708

Sankaran, S., Khot, L.R., Panigrahi, S., 2012. Biology and applications of olfactory sensing system: A review. Sensors and Actuators B: Chemical 171–172, 1–17. https://doi.org/10.1016/j.snb.2012.03.029

Schmiedeberg, K., Shirokova, E., Weber, H.-P., Schilling, B., Meyerhof, W., Krautwurst, D., 2007. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. J Struct Biol 159, 400–412. https://doi.org/10.1016/j.jsb.2007.04.013

Schmuker, M., de Bruyne, M., Hähnel, M., Schneider, G., 2007. Predicting olfactory receptor neuron responses from odorant structure. Chem Cent J 1, 11. https://doi.org/10.1186/1752-153X-1-11

Sharma, A., Saha, B.K., Kumar, R., Varadwaj, P.K., 2022. OlfactionBase: a repository to explore odors, odorants, olfactory receptors and odorant–receptor interactions. Nucleic Acids Research 50, D678–D686. https://doi.org/10.1093/nar/gkab763

Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., Chang, W., 2022. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research gkac194. https://doi.org/10.1093/nar/gkac194

Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J.A., Wetzel, C.H., Zimmer, R.K., Hatt, H., 2003. Identification of a Testicular Odorant Receptor Mediating Human Sperm Chemotaxis. Science 299, 2054–2058. https://doi.org/10.1126/science.1080376

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J., von Mering, C., 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research 49, D605–D612. https://doi.org/10.1093/nar/gkaa1074

Tong, T., Wang, Y., Kang, S.-G., Huang, K., 2021. Ectopic Odorant Receptor Responding to Flavor Compounds: Versatile Roles in Health and Disease. Pharmaceutics 13, 1314. https://doi.org/10.3390/pharmaceutics13081314

van Dongen, S., 2000. A Cluster Algorithm for Graphs.

Vogt, I., Mestres, J., 2019. Information Loss in Network Pharmacology. Mol. Inf. 38, 1900032. https://doi.org/10.1002/minf.201900032

Wackermannová, M., Pinc, L., Jebavý, L., 2016. Olfactory Sensitivity in Mammalian Species. Physiol Res 369–390. https://doi.org/10.33549/physiolres.932955

Wang, F.I., Ding, G., Ng, G.S., Dixon, S.J., Chidiac, P., 2022. Luciferase-based GloSensor™ cAMP assay: Temperature optimization and application to cell-based kinetic studies. Methods 203, 249–258. https://doi.org/10.1016/j.ymeth.2021.10.009

Wu, Q., Taboureau, O., Audouze, K., 2020. Development of an adverse drug event network to predict drug toxicity. Current Research in Toxicology 1, 48–55. https://doi.org/10.1016/j.crtox.2020.06.001

Zhang, K., Wang, L., Liu, Z., Geng, B., Teng, Y., Liu, X., Yi, Q., Yu, D., Chen, X., Zhao, D., Xia, Y., 2021. Mechanosensory and mechanotransductive processes mediated by ion channels in articular chondrocytes: Potential therapeutic targets for osteoarthritis. Channels 15, 339–359. https://doi.org/10.1080/19336950.2021.1903184

Partie 4 : Discussion, conclusion et perspectives

I. Discussion

Depuis l'antiquité, l'olfaction a souvent été considéré comme un sens vestigial (Majid, 2021). De ce fait, les études scientifiques sur l'olfaction ont longtemps été délaissées, comparées aux études d'autres sens tels que l'audition ou la vue. Néanmoins, depuis plusieurs décennies, notamment avec la découverte de la famille de gènes codant pour les récepteurs olfactifs par Linda Buck et Richard Axel (Buck and Axel, 1991), les études sur l'olfaction se sont multipliées, permettant ainsi de mieux comprendre ce sens (Hoskison, 2013; Meierhenrich et al., 2005b). L'olfaction est un sens particulièrement complexe, qui joue un rôle essentiel dans la consommation de nourriture, la sélection des partenaires, la reproduction et également dans la sécurité et la survie en alertant de la présence d'odeurs nocives et de prédateurs (Hoskison, 2013; Majid, 2021).

Outre son implication dans ces processus, l'olfaction peut être affectée dans certaines pathologies, et des dysfonctions olfactives peuvent ainsi être utilisées comme outil de diagnostic de maladies neurodégénératives telles que la maladie d'Alzheimer (Murphy, 2019) ou la maladie de Parkinson (Marin et al., 2018). Des pertes (anosmie) ou des modifications (dysnosmie) des perceptions olfactives font ainsi partie des symptômes constatés chez des sujets atteints du Sars-CoV-2 (COVID-19) (Gerkin et al., 2021; Parma et al., 2020). Les anosmies partielles ou complètes sont par ailleurs associées à des états dépressifs (Athanassi et al., 2021; Chen et al., 2020; Croy and Hummel, 2017; Fang et al., 2021; Schienle et al., 2018). De plus, l'obésité pourrait également être associée à des modifications de la perception olfactive comme la sensibilité olfactive, la discrimination olfactive ou l'identification des odeurs (Majid, 2021; Peng et al., 2019).

Malgré les énormes avancées établies dans le domaine de l'olfaction comme la compréhension des mécanismes de signalisation moléculaire, certains mécanismes ne sont

pas encore totalement compris, comme le rôle de la structure moléculaire d'un composé odorant sur sa qualité odorante (Genva et al., 2019b; Saini and Ramanathan, 2022; Snitz et al., 2019), les mécanismes impliqués dans la perception homogène des mélanges, accords aromatiques ou masquages (Berglund et al., 1976; Coureaud et al., 2022; Ishii et al., 2008), ou les effets biologiques des molécules odorantes au-delà du système olfactif (Di Pizio et al., 2019; Foster et al., 2014).

Dans le but de mieux appréhender ces questions, nous avons développé trois modèles permettant de : (1) mettre en lumière le lien entre la structure d'une molécule et son odeur, (2) déterminer dans quelle mesure les mécanismes de perception homogène pouvaient intervenir au niveau périphérique du système olfactif et (3) identifier des processus biologiques dans lesquels molécules odorantes pourraient intervenir, et prédire de nouvelles cibles biologiques potentielles.

L'élaboration de tels modèles est fortement complexifiée par plusieurs aspects. Tout d'abord, la difficulté à définir les qualités odorantes des molécules odorantes. En effet, les descriptions verbales des notes odorantes peuvent être diverses car elles impliquent à la fois des aspects sémantiques et affectifs, les émotions et la mémoire étant très associés à la perception odorante (Chrea, 2005; Sabiniewicz et al., 2021; Stevenson and Mahmut, 2013).

D'autre part, les interactions moléculaires et réactions chimiques subies par les molécules odorantes dans la muqueuse olfactive ou encore les interactions des signaux survenant au niveau des structures supérieures du système olfactif influencent la perception odorante (Heydel et al., 2013; Nagashima and Touhara, 2010; Robert-Hazotte et al., 2022; Stettler and Axel, 2009; Yoshida and Mori, 2007).

Enfin, de nombreuses théories ont été proposées afin de décrire les mécanismes spécifiques de la perception odorante, en impliquant des propriétés particulières des molécules odorantes (Sankaran et al., 2012; Sharma et al., 2019). Parmi elles, la théorie vibrationnelle de l'odeur repose sur le principe que l'odeur des molécules odorantes peut être différenciée en fonction de la fréquence de vibration des molécules (Dyson, 1938; Wright, 1954). Une autre théorie, la théorie basée sur la forme, implique l'odeur d'une molécule odorante est associé au nombre de récepteurs auxquels elle se lie, aux caractéristiques structurelles nécessaires pour se lier aux récepteurs et à l'intensité d'excitation du récepteur (Mori and Shepherd, 1994). Il existe également une théorie des enzymes qui admet que des enzymes actives dans

l'épithélium olfactif peuvent être inhibées par les molécules odorantes, modifiant ainsi les concentrations relatives de certaines molécules pour un récepteur olfactif (Jones and Jones, 1953; Kistiakowsky, 1950; Wendt, 1952). De plus, le complexe formé entre la molécule odorante et l'enzyme, provoque des changements de conformation de la molécule odorante (Sharma et al., 2019). Ce changement de conformation entraîne l'exposition de points enfouis de la molécule, lui permettant potentiellement de se lier à d'autres récepteurs olfactifs (Sharma et al., 2019). En effet, il a été montré que des neurones sensoriels répondant à l'octanal étaient activés de manière plus efficace par des analogues à conformation restreinte de l'octanal que par l'octanal lui-même, indiquant que les récepteurs olfactifs activés par l'octanal possèdent un filtre conformationnel (Peterlin et al., 2008).
Cette thèse a permis de présenter différents mécanismes intervenant dans la perception olfactive (comme le code combinatoire) et d'apporter des éléments de réponse aux mécanismes qui restent encore méconnus.

II. Conclusion

L'objectif de cette thèse était de mieux comprendre certains mécanismes spécifiques de l'olfaction. Pour cela, trois axes ont été étudiés à l'aide de trois approches différentes et complémentaires.

Dans un premier temps, nous avons construit un modèle, décrit dans la partie 3-I, qui associait une technique de réduction de dimension et une méthode de classification. La technique UMAP appliquées aux fingerprints encodant les structures moléculaires des composés odorants suivie par une technique de classification a permis de définir des clusters qui distinguaient également certaines notes odorantes. Nous avons pu ainsi mettre en évidence certains liens entre la structure de molécules et leurs odeurs. En effet, diverses associations ont été observées entre certains groupements chimiques et notes odorantes. Par exemple, des structures allyliques et bicycliques sont présentes dans les clusters dont les odorants sont caractérisées par les notes "boisées" et "épicées", tandis que les clusters où se trouvent des cycles insaturés sont caractérisés par notes "balsamiques", et que tandis que les esters et longues chaînes carbonées sont portés par des odorants ayant des notes "huileuses", "grasses" et "fruitées" sont associées.

Le modèle présenté dans la partie 3-II, a été développé dans le but de mieux comprendre la perception homogène de mélanges de molécules odorantes. Les résultats évoquent deux modes de fonctionnements différents entre le masquage et l'accord aromatique. En effet, les pharmacophores développés à partir des composants du masquage semblent indiquer que ces composants pourraient se lier à un même site récepteur, ce qui pourrait correspondre à des mécanismes de compétition, agonistes compétitifs ou antagonistes. Si tel est le cas, la perception configurale de ce masquage serait déterminée en partie au niveau périphérique du système olfactif. A l'inverse, les hypothèse pharmacophores générées à partir des molécules du mélange « grenadine » n'ont donné aucun résultat fiable indiquant des caractéristiques structurales et spatiales communes à ces molécules, ce qui va à l'encontre de modes de fixation commun sur des récepteurs. De ce fait, la perception configurale de cet accord « grenadine » est plus probablement due à une intégration du signal olfactif au-delà du niveau périphérique.

Le réseau odorome présenté dans la partie 3-III situe les molécules odorantes dans un espace d'effets biologiques par les liens qui ont pu être établis sur la base des études reportées dans la littérature. Par rapport à la première version de l'odorome qui avait été proposée dans une précédente étude, la construction de ce nouvel odorome a intégré un jeu de données contenant quatre fois plus d'interactions entre récepteurs olfactifs et molécules odorantes que la version précédente. Ainsi enrichi, le modèle odorome V2 va à la fois permettre de mieux comprendre les interactions entre molécules odorantes et récepteurs olfactifs, mais aussi d'en prédire des milliers et d'identifier plusieurs processus biologiques et pathologies pouvant impliquer les récepteurs olfactifs.

Les différentes études réalisées au cours de cette thèse ont permis de mieux comprendre la manière dont influe la structure des molécules odorantes dans la perception d'une odeur et d'identifier l'intervention de molécules odorantes dans des processus biologiques et de nouvelles interactions potentielles entre molécules odorantes et récepteurs olfactifs. Tous ces résultats seront ensuite utilisés afin d'initier de nouvelles études complémentaires.

III. Perspectives

En continuité avec les modèles développés au cours de cette thèse, plusieurs études pourront être mises en place afin d'approfondir les résultats obtenus.

Les résultats obtenus dans la partie 3-II, se révèlent être une piste intéressante pour les relations structure-odeur. En effet, diviser les clusters obtenus avec SOM permettraient d'affiner les liens entre la structure des molécules odorantes et leurs odeurs.

Comme mentionné dans la partie 3-III, le modèle odorome pourrait être amélioré en y ajoutant des informations concernant les récepteurs olfactifs qui ne sont pas activés par une molécule odorante. En combinant ces données avec les interactions connues entre molécules odorantes et récepteurs olfactifs, il serait possible de créer, pour chaque molécule odorante, des vecteurs binaires représentant par un « 1 » l'activation ou par un « 0 » la non-activation des récepteurs olfactifs. Et les vecteurs binaires pourront être utilisés en tant que données d'entrée de réseaux neuronaux profonds afin de prédire l'odeur de molécules odorantes. D'autre part, ces vecteurs binaires pourraient également être associés à des fingerprints représentant la structure des molécules odorantes comme calculés dans l'étude de la partie 3-I. Cela permettrait d'enrichir les données utilisées pour les réseaux neuronaux profonds et d'ainsi pouvoir potentiellement mettre en lumière un lien entre la structure des molécules, les récepteurs olfactifs qu'elles activent et leurs odeurs. Il s'agira ensuite de réaliser la validation expérimentale des prédictions d'interactions entre molécules odorantes et récepteur par la mesure *in vitro* de l'activité fonctionnelles des récepteurs olfactifs.

Partie 5 : Bibliographie

Abaffy, T., Malhotra, A., Luetje, C.W., 2007. The molecular basis for ligand specificity in a mouse olfactory receptor: a network of functionally important residues. Journal of Biological Chemistry 282, 1216–1224. https://doi.org/10.1074/jbc.M609355200

Abdi, H., Williams, L.J., 2010. Principal component analysis: Principal component analysis. WIREs Comp Stat 2, 433–459. https://doi.org/10.1002/wics.101

Abraham, R., Marsden, J.E., Ratiu, T., 1988. Manifolds, Tensor Analysis, and Applications, Applied Mathematical Sciences. Springer New York, New York, NY.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J., 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science, in: Berry, M.W., Mohamed, A., Yap, B.W. (Eds.), Supervised and Unsupervised Learning for Data Science, Unsupervised and Semi-Supervised Learning. Springer International Publishing, Cham, pp. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

Anderson, A.K., Christoff, K., Stappen, I., Panitz, D., Ghahremani, D.G., Glover, G., Gabrieli, J.D.E., Sobel, N., 2003. Dissociated neural representations of intensity and valence in human olfaction. Nat Neurosci 6, 196–202. https://doi.org/10.1038/nn1001

Anowar, F., Sadaoui, S., Selim, B., 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Computer Science Review 40, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Arnold, T.C., You, Y., Ding, M., Zuo, X.-N., de Araujo, I., Li, W., 2020. Functional Connectome Analyses Reveal the Human Olfactory Network Organization. eNeuro 7, ENEURO.0551-19.2020. https://doi.org/10.1523/ENEURO.0551-19.2020

Atanasova, B., Thomas-Danguin, T., Langlois, D., Nicklaus, S., Etievant, P., 2004. Perceptual interactions between fruity and woody notes of wine. Flavour Fragr. J. 19, 476–482. https://doi.org/10.1002/ffj.1474

Athanassi, A., Dorado Doncel, R., Bath, K.G., Mandairon, N., 2021. Relationship between depression and olfactory sensory function: a review. Chemical Senses 46, bjab044. https://doi.org/10.1093/chemse/bjab044

Axen, S.D., Huang, X.-P., Cáceres, E.L., Gendelev, L., Roth, B.L., Keiser, M.J., 2017. A Simple Representation of Three-Dimensional Molecular Structure. J. Med. Chem. 60, 7393–7409. https://doi.org/10.1021/acs.jmedchem.7b00696

Bahoken, F., Beauguitte, L., Lhomme, S., 2013. La visualisation des réseaux. Principes, enjeux et perspectives.

Bai, X., McMullan, G., Scheres, S.H.W., 2015. How cryo-EM is revolutionizing structural biology. Trends in Biochemical Sciences 40, 49–57. https://doi.org/10.1016/j.tibs.2014.10.005

Bajgrowicz, J.A., Frater, G., 2000. Chiral recognition of sandalwood odorants. Enantiomer 5, 225–234.

Bajorath, J., 2001. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. J. Chem. Inf. Comput. Sci. 41, 233–245. https://doi.org/10.1021/ci0001482

Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7, 20. https://doi.org/10.1186/s13321-015-0069-3

Barabási, A.-L., Bonabeau, E., 2003. Scale-Free Networks. Scientific American 288, 60–69.

Barnum, D., Greene, J., Smellie, A., Sprague, P., 1996. Identification of Common Functional Configurations Among Molecules. J. Chem. Inf. Comput. Sci. 36, 563–571. https://doi.org/10.1021/ci950273r

Berge, C., 1983. Graphes. Gauthier-Villars, Paris.

Berglund, B., Berglund, U., Lindvall, T., 1976. Psychological processing of odor mixtures. Psychological Review 83, 432–441. https://doi.org/10.1037/0033-295X.83.6.432

Bipul Hossen, Md., 2015. Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study. CBB 3, 88. https://doi.org/10.11648/j.cbb.20150306.12

Block, E., 2018. Molecular Basis of Mammalian Odor Discrimination: A Status Report. J. Agric. Food Chem. 66, 13346–13366. https://doi.org/10.1021/acs.jafc.8b04471

Bondy, A., Murty, U.S.R., 2008. Graph theory, Springer London. ed.

Borg, I., 2018. Applied multidimensional scaling and unfolding. Springer.

Boyle, J.A., Djordjevic, J., Olsson, M.J., Lundstrom, J.N., Jones-Gotman, M., 2009. The Human Brain Distinguishes between Single Odorants and Binary Mixtures. Cerebral Cortex 19, 66–71. https://doi.org/10.1093/cercor/bhn058

Brennan, P.A., 2001. The vomeronasal system. CMLS, Cell. Mol. Life Sci. 58, 546–555. https://doi.org/10.1007/PL00000880

Bretto, A., Cherifi, H., Aboutajdine, D., 2002. Hypergraph imaging: an overview. Pattern Recognition 35, 651–658. https://doi.org/10.1016/S0031-3203(01)00067-X

Buck, L., Axel, R., 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. Cell 65, 175–187. https://doi.org/10.1016/0092-8674(91)90418-X

Bushdid, C., de March, C.A., Topin, J., Do, M., Matsunami, H., Golebiowski, J., 2019. Mammalian class I odorant receptors exhibit a conserved vestibular-binding pocket. Cell. Mol. Life Sci. 76, 995–1004. https://doi.org/10.1007/s00018-018-2996-4

Bushdid, C., Magnasco, M.O., Vosshall, L.B., Keller, A., 2014. Humans can discriminate more than 1 trillion olfactory stimuli. Science 343, 1370–1372. https://doi.org/10.1126/science.1249168

Cain, W.S., Drexler, M., 1974. SCOPE AND EVALUATION OF ODOR COUNTERACTION AND MASKING. Ann NY Acad Sci 237, 427–439. https://doi.org/10.1111/j.1749-6632.1974.tb49876.x

Capecchi, A., Probst, D., Reymond, J.-L., 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12, 43. https://doi.org/10.1186/s13321-020-00445-4

Catalyst, 2014. . Accelrys, Inc, San Diego.

Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., Pujadas, G., 2015. Molecular fingerprint similarity search in virtual screening. Methods 71, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005

Chamanza, R., Wright, J.A., 2015. A Review of the Comparative Anatomy, Histology, Physiology and Pathology of the Nasal Cavity of Rats, Mice, Dogs and Non-human Primates.

Relevance to Inhalation Toxicology and Human Health Risk Assessment. Journal of Comparative Pathology 153, 287–314. https://doi.org/10.1016/j.jcpa.2015.08.009

Chaput, M.A., El Mountassir, F., Atanasova, B., Thomas-Danguin, T., Le Bon, A.M., Perrut, A., Ferry, B., Duchamp-Viret, P., 2012. Interactions of odorants with olfactory receptors and receptor neurons match the perceptual dynamics observed for woody and fruity odorant mixtures: Wine fruity-woody odours in the rat nose and human brain. European Journal of Neuroscience 35, 584–597. https://doi.org/10.1111/j.1460-9568.2011.07976.x

Chastrette, M., 1997. Trends in Structure-Odor Relationship. SAR and QSAR in Environmental Research 6, 215–254. https://doi.org/10.1080/10629369708033253

Chastrette, M., de Saint Laumer, J., 1991. Structure-odor relationships using neural networks. European Journal of Medicinal Chemistry 26, 829–833. https://doi.org/10.1016/0223-5234(91)90010-K

Chastrette, M., Zakarya, D., Pierre, C., 1990. Relations structure-odeur de bois de santal: recherche d'un modèle d'interaction fondé sur le concept d'hypermotif santalophore. European Journal of Medicinal Chemistry 25, 433–440. https://doi.org/10.1016/0223-5234(90)90007-P

Chen, B., Akshita, J., Han, P., Thaploo, D., Kitzler, H.H., Hummel, T., 2020. Aberrancies of Brain Network Structures in Patients with Anosmia. Brain Topogr 33, 403–411. https://doi.org/10.1007/s10548-020-00769-2

Chrea, C., 2005. Semantic, Typicality and Odor Representation: A Cross-cultural Study. Chemical Senses 30, 37–49. https://doi.org/10.1093/chemse/bjh255

Chung, N.C., Miasojedow, B., Startek, M., Gambin, A., 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC Bioinformatics 20, 644. https://doi.org/10.1186/s12859-019-3118-5

Cichy, R.M., Kaiser, D., 2019. Deep Neural Networks as Scientific Models. Trends in Cognitive Sciences 23, 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cleland, T.A., Linster, C., 2019. Central olfactory structures, in: Handbook of Clinical Neurology. Elsevier, pp. 79–96. https://doi.org/10.1016/B978-0-444-63855-7.00006-X

Cohen-Tannoudji, L., 2006. CINETIQUE DE REACTIONS LIGAND-RECEPTEUR EN SURFACE - étude fondée sur l'utilisation de colloïdes magnétiques. Université Pierre et Marie Curie - Paris VI.

Consonni, V., Todeschini, R., 2010. Molecular Descriptors, in: Puzyn, T., Leszczynski, J., Cronin, M.T. (Eds.), Recent Advances in QSAR Studies, Challenges and Advances in Computational Chemistry and Physics. Springer Netherlands, Dordrecht, pp. 29–102. https://doi.org/10.1007/978-1-4020-9783-6_3

Consonni, V., Todeschini, R. (Eds.), 2000. Handbook of molecular descriptors, Methods and principles in medicinal chemistry. Wiley-VCH, Weinheim ; New York.

Coureaud, G., Thomas-Danguin, T., Sandoz, J.-C., Wilson, D.A., 2022. Biological constraints on configural odour mixture perception. Journal of Experimental Biology 225, jeb242274. https://doi.org/10.1242/jeb.242274

Cowart, B.J., Rawson, N.E., 2005. Olfaction, in: Blackwell Handbook of Sensation and Perception. pp. 567–600.

Cramer, R.D., Patterson, D.E., Bunce, J.D., 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. 110, 5959–5967. https://doi.org/10.1021/ja00226a005

Crespo, C., Liberia, T., Blasco-Ibáñez, J.M., Nácher, J., Varea, E., 2019. Cranial Pair I: The Olfactory Nerve: THE OLFACTORY NERVE. Anat. Rec. 302, 405–427. https://doi.org/10.1002/ar.23816

Croy, I., Hummel, T., 2017. Olfaction as a marker for depression. J Neurol 264, 631–638. https://doi.org/10.1007/s00415-016-8227-8

Cunningham, P., 2008. Dimension Reduction, in: Cord, M., Cunningham, P. (Eds.), Machine Learning Techniques for Multimedia, Cognitive Technologies. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 91–112. https://doi.org/10.1007/978-3-540-75171-7_4

Czesnik, D., Schild, D., Kuduz, J., Manzini, I., 2007. Cannabinoid action in the olfactory epithelium. Proc. Natl. Acad. Sci. U.S.A. 104, 2967–2972. https://doi.org/10.1073/pnas.0609067104

D'Aniello, B., Semin, G.R., Scandurra, A., Pinelli, C., 2017. The Vomeronasal Organ: A Neglected Organ. Front. Neuroanat. 11, 70. https://doi.org/10.3389/fnana.2017.00070

Davison, I.G., Katz, L.C., 2007. Sparse and Selective Odor Coding by Mitral/Tufted Neurons in the Main Olfactory Bulb. Journal of Neuroscience 27, 2091–2101. https://doi.org/10.1523/JNEUROSCI.3779-06.2007

Delasalle, C., de March, C.A., Meierhenrich, U.J., Brevard, H., Golebiowski, J., Baldovini, N., 2014. Structure-Odor Relationships of Semisynthetic β-Santalol Analogs. Chemistry & Biodiversity 11, 1843–1860. https://doi.org/10.1002/cbdv.201400082

Derrible, S., Kennedy, C., 2011. Applications of Graph Theory and Network Science to Transit Network Design. Transport Reviews 31, 495–519. https://doi.org/10.1080/01441647.2010.543709

Dewan, A., 2021. Olfactory signaling via trace amine-associated receptors. Cell Tissue Res 383, 395–407. https://doi.org/10.1007/s00441-020-03331-5

Di Pizio, A., Behrens, M., Krautwurst, D., 2019. Beyond the Flavour: The Potential Druggability of Chemosensory G Protein-Coupled Receptors. IJMS 20, 1402. https://doi.org/10.3390/ijms20061402

Dieris, M., Kowatschew, D., Korsching, S.I., 2021. Olfactory function in the trace amine-associated receptor family (TAARs) evolved twice independently. Sci Rep 11, 7807. https://doi.org/10.1038/s41598-021-87236-5

Dinu, V., MacCalman, T., Yang, N., Adams, G.G., Yakubov, G.E., Harding, S.E., Fisk, I.D., 2020. Probing the effect of aroma compounds on the hydrodynamic properties of mucin glycoproteins. Eur Biophys J 49, 799–808. https://doi.org/10.1007/s00249-020-01475-4

Di Pizio, A., Niv, M.Y., 2014. Computational Studies of Smell and Taste Receptors. Isr. J. Chem. 54, 1205–1218. https://doi.org/10.1002/ijch.201400027

Dixon, S.L., Smondyrev, A.M., Knoll, E.H., Rao, S.N., Shaw, D.E., Friesner, R.A., 2006. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20, 647–671. https://doi.org/10.1007/s10822-006-9087-6

Dudek, A., Arodz, T., Galvez, J., 2006. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. CCHTS 9, 213–228. https://doi.org/10.2174/138620706776055539

Dulac, C., 2006. Sparse Encoding of Natural Scents. Neuron 50, 816–818. https://doi.org/10.1016/j.neuron.2006.06.002

Dulac, C., Torello, A.T., 2003. Molecular detection of pheromone signals in mammals: from genes to behaviour. Nat Rev Neurosci 4, 551–562. https://doi.org/10.1038/nrn1140

Dyson, G.M., 1938. The scientific basis of odour. J. Chem. Technol. Biotechnol. 57, 647–651. https://doi.org/10.1002/jctb.5000572802

Economo, M.N., Hansen, K.R., Wachowiak, M., 2016. Control of Mitral/Tufted Cell Output by Selective Inhibition among Olfactory Bulb Glomeruli. Neuron 91, 397–411. https://doi.org/10.1016/j.neuron.2016.06.001

Ehrlich, P., 1909. Über den jetzigen Stand der Chemotherapie. Berichte der deutschen chemischen Gesellschaft 42, 17–47.

Ekins, S., Mestres, J., Testa, B., 2007. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling: *In silico* pharmacology for drug discovery. British Journal of Pharmacology 152, 9–20. https://doi.org/10.1038/sj.bjp.0707305

El Naqa, I., Murphy, M.J., 2015. What Is Machine Learning?, in: El Naqa, I., Li, R., Murphy, M.J. (Eds.), Machine Learning in Radiation Oncology. Springer International Publishing, Cham, pp. 3–11. https://doi.org/10.1007/978-3-319-18305-3_1

Euler, L., 1741. Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae 128–140.

Evans, D.A., Doman, T.N., Thorner, D.A., Bodkin, M.J., 2007. 3D QSAR Methods: Phase and Catalyst Compared. J. Chem. Inf. Model. 47, 1248–1257. https://doi.org/10.1021/ci7000082

Fang, T.-C., Chang, M.-H., Yang, C.-P., Chen, Y.-H., Lin, C.-H., 2021. The Association of Olfactory Dysfunction With Depression, Cognition, and Disease Severity in Parkinson's Disease. Front. Neurol. 12, 779712. https://doi.org/10.3389/fneur.2021.779712

Feinstein, P., Mombaerts, P., 2004. A Contextual Model for Axonal Sorting into Glomeruli in the Mouse Olfactory System. Cell 117, 817–831. https://doi.org/10.1016/j.cell.2004.05.011

Firestein, S., 2001. How the olfactory system makes sense of scents. Nature 413, 211–218. https://doi.org/10.1038/35093026

Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dtsch. Chem. Ges. 27, 2985–2993. https://doi.org/10.1002/cber.18940270364

Fjældstad, A., 2018. Testing olfactory function and mapping the structural olfactory networks in the brain. Dan Med J 65, B5428.

Fleischer, J., Breer, H., Strotmann, J., 2009. Mammalian olfactory receptors. Front Cell Neurosci 3, 9. https://doi.org/10.3389/neuro.03.009.2009

Foster, S.R., Roura, E., Thomas, W.G., 2014. Extrasensory perception: Odorant and taste receptors beyond the nose and mouth. Pharmacology & Therapeutics 142, 41–61. https://doi.org/10.1016/j.pharmthera.2013.11.004

Fráter, G., Bajgrowicz, J.A., Kraft, P., 1998. Fragrance chemistry. Tetrahedron 54, 7633–7703. https://doi.org/10.1016/S0040-4020(98)00199-9

Freitag, J., Krieger, J., Strotmann, J., Breer, H., 1995. Two classes of olfactory receptors in xenopus laevis. Neuron 15, 1383–1392. https://doi.org/10.1016/0896-6273(95)90016-0

Gaillard, I., Rouquier, S., Giorgi, D., 2004. Olfactory receptors. Cellular and Molecular Life Sciences (CMLS) 61, 456–469. https://doi.org/10.1007/s00018-003-3273-7

Galizia, C.G., Lledo, P.-M., 2013. Olfaction, in: Galizia, C.G., Lledo, P.-M. (Eds.), Neurosciences - From Molecule to Behavior: A University Textbook. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 253–284. https://doi.org/10.1007/978-3-642-10769-6_13

Gentleman, R., Carey, V.J., 2008. Unsupervised Machine Learning, in: Bioconductor Case Studies. Springer New York, New York, NY, pp. 137–157. https://doi.org/10.1007/978-0-387-77240-0_10

Genva, M., Kenne Kemene, T., Deleu, M., Lins, L., Fauconnier, M.-L., 2019a. Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure? IJMS 20, 3018. https://doi.org/10.3390/ijms20123018

Genva, M., Kenne Kemene, T., Deleu, M., Lins, L., Fauconnier, M.-L., 2019b. Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure? IJMS 20, 3018. https://doi.org/10.3390/ijms20123018

Gerkin, R.C., Ohla, K., Veldhuizen, M.G., Joseph, P.V., Kelly, C.E., Bakke, A.J., Steele, K.E., Farruggia, M.C., Pellegrino, R., Pepino, M.Y., Bouysset, C., Soler, G.M., Pereda-Loth, V., Dibattista, M., Cooper, K.W., Croijmans, I., Di Pizio, A., Ozdener, M.H., Fjaeldstad, A.W., Lin, C., Sandell, M.A., Singh, P.B., Brindha, V.E., Olsson, S.B., Saraiva, L.R., Ahuja, G., Alwashahi, M.K., Bhutani, S., D'Errico, A., Fornazieri, M.A., Golebiowski, J., Dar Hwang, L., Öztürk, L., Roura, E., Spinelli, S., Whitcroft, K.L., Faraji, F., Fischmeister, F.P.S., Heinbockel, T., Hsieh, J.W., Huart, C., Konstantinidis, I., Menini, A., Morini, G., Olofsson, J.K., Philpott, C.M., Pierron, D., Shields, V.D.C., Voznessenskaya, V.V., Albayay, J., Altundag, A., Bensafi, M., Bock, M.A., Calcinoni, O., Fredborg, W., Laudamiel, C., Lim, J., Lundström, J.N., Macchi, A., Meyer, P., Moein, S.T., Santamaría, E., Sengupta, D., Rohlfs Dominguez, P., Yanik, H., Hummel, T., Hayes, J.E., Reed, D.R., Niv, M.Y., Munger, S.D., Parma, V., GCCR Group Author, Boesveldt, S., de Groot, J.H.B., Dinnella, C., Freiherr, J., Laktionova, T., Marino, S., Monteleone, E., Nunez-Parra, A., Abdulrahman, O., Ritchie, M., Thomas-Danguin, T., Walsh-Messinger, J., Al Abri, R., Alizadeh, R., Bignon, E., Cantone, E., Paola Cecchini, M., Chen, J., Dolors Guàrdia, M., Hoover, K.C., Karni, N., Navarro, M., Nolden, A.A., Portillo Mazal, P., Rowan, N.R., Sarabi-Jamab, A., Archer, N.S., Chen, B., Di Valerio, E.A., Feeney, E.L., Frasnelli, J., Hannum, M.E., Hopkins, C., Klein, H., Mignot, C., Mucignat, C., Ning, Y., Ozturk, E.E., Peng, M., Saatci, O., Sell, E.A., Yan, C.H., Alfaro, R., Cecchetto, C., Coureaud, G., Herriman, R.D., Justice, J.M., Kaushik, P.K., Koyama, S., Overdevest, J.B., Pirastu, N., Ramirez, V.A., Roberts, S.C., Smith, B.C., Cao, H., Wang, H., Balungwe Birindwa, P., Baguma, M., 2021. Recent Smell Loss Is the Best Predictor of COVID-19 Among Individuals With Recent Respiratory Symptoms. Chemical Senses 46, bjaa081. https://doi.org/10.1093/chemse/bjaa081

Glezer, I., Malnic, B., 2019. Olfactory receptor function, in: Handbook of Clinical Neurology. Elsevier, pp. 67–78. https://doi.org/10.1016/B978-0-444-63855-7.00005-8

Gottfried, J.A., Winston, J.S., Dolan, R.J., 2006. Dissociable Codes of Odor Quality and Odorant Structure in Human Piriform Cortex. Neuron 49, 467–479. https://doi.org/10.1016/j.neuron.2006.01.007

Grabenhorst, F., Rolls, E.T., Margot, C., da Silva, M.A.A.P., Velazco, M.I., 2007. How Pleasant and Unpleasant Stimuli Combine in Different Brain Regions: Odor Mixtures. Journal of Neuroscience 27, 13532–13540. https://doi.org/10.1523/JNEUROSCI.3337-07.2007

Greenidge, P., Weiser, J., 2001. A Comparison of Methods for Pharmacophore Generation with the Catalyst Software and their Use for 3D-QSAR Application to a Set of 4-

Aminopyridine Thrombin Inhibitors. MRMC 1, 79–87. https://doi.org/10.2174/1389557013407223

Grira, N., Crucianu, M., Boujemaa, N., 2005. Unsupervised and Semi-supervised Clustering: a brief survey. A Review of Machine Learning Techniques for Processing Multimedia Content.

Günd, P., 1977. Three-Dimensional Pharmacophoric Pattern Searching, in: Hahn, F.E., Kersten, H., Kersten, W., Szybalski, W. (Eds.), Progress in Molecular and Subcellular Biology, Progress in Molecular and Subcellular Biology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 117–143. https://doi.org/10.1007/978-3-642-66626-1_4

Hall, R.A., 2011. Autonomic Modulation of Olfactory Signaling. Sci. Signal. 4. https://doi.org/10.1126/scisignal.2001672

Hansch, C., Maloney, P.P., Fujita, T., Muir, R.M., 1962. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature 194, 178–180. https://doi.org/10.1038/194178b0

Hansch, C., Muir, R.M., 1950. THE ORTHO EFFECT IN PLANT GROWTH-REGULATORS. Plant Physiol. 25, 389–393. https://doi.org/10.1104/pp.25.3.389

Hansch, Corwin., Fujita, Toshio., 1964. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc. 86, 1616–1626. https://doi.org/10.1021/ja01062a035

Hasegawa-Ishii, S., Imamura, F., Nagayama, S., Murata, M., Shimada, A., 2020. Differential Effects of Nasal Inflammation and Odor Deprivation on Layer-Specific Degeneration of the Mouse Olfactory Bulb. eNeuro 7, ENEURO.0403-19.2020. https://doi.org/10.1523/ENEURO.0403-19.2020

Hasin-Brumshtein, Y., Lancet, D., Olender, T., 2009. Human olfaction: from genomic variation to phenotypic diversity. Trends Genet 25, 178–184. https://doi.org/10.1016/j.tig.2009.02.002

Hauser, N., Kraft, P., Carreira, E.M., 2020. The Serendipitous Discovery of a Rose Odorant. Chimia 74, 247. https://doi.org/10.2533/chimia.2020.247

Heinbockel, T., Straiker, A., 2021. Cannabinoids Regulate Sensory Processing in Early Olfactory and Visual Neural Circuits. Front. Neural Circuits 15, 662349. https://doi.org/10.3389/fncir.2021.662349

Hellier, J.L. (Ed.), 2017. The five senses and beyond: the encyclopedia of perception. Greenwood, an imprintof ABC-CLIO, LLC, Santa Barbara, California.

Heydel, J.-M., Coelho, A., Thiebaud, N., Legendre, A., Bon, A.-M.L., Faure, P., Neiers, F., Artur, Y., Golebiowski, J., Briand, L., 2013. Odorant-Binding Proteins and Xenobiotic Metabolizing Enzymes: Implications in Olfactory Perireceptor Events: Odorant-Binding Proteins and Metabolizing Enzymes. Anat. Rec. 296, 1333–1345. https://doi.org/10.1002/ar.22735

Holley, A., 2006. Système olfactif et neurobiologie. terrain 107–122. https://doi.org/10.4000/terrain.4271

Hoskison, E.E., 2013. Olfaction, pheromones and life. J. Laryngol. Otol. 127, 1156–1159. https://doi.org/10.1017/S0022215113002545

Huart, C., Rombaux, P., Hummel, T., 2013. Plasticity of the Human Olfactory System: The Olfactory Bulb. Molecules 18, 11586–11600. https://doi.org/10.3390/molecules180911586

Ikegami, K., de March, C.A., Nagai, M.H., Ghosh, S., Do, M., Sharma, R., Bruguera, E.S., Lu, Y.E., Fukutani, Y., Vaidehi, N., Yohda, M., Matsunami, H., 2020. Structural instability and divergence from conserved residues underlie intracellular retention of mammalian odorant receptors. Proc. Natl. Acad. Sci. U.S.A. 117, 2957–2967. https://doi.org/10.1073/pnas.1915520117

Ishii, A., Roudnitzky, N., Beno, N., Bensafi, M., Hummel, T., Rouby, C., Thomas-Danguin, T., 2008. Synergy and Masking in Odor Mixtures: An Electrophysiological Study of Orthonasal vs. Retronasal Perception. Chemical Senses 33, 553–561. https://doi.org/10.1093/chemse/bjn022

Jackson, A., 2019. The mathematics of UMAP.

Johnson, B.A., Leon, M., 2000. Modular representations of odorants in the glomerular layer of the rat olfactory bulb and the effects of stimulus concentration. J. Comp. Neurol. 422, 496–509. https://doi.org/10.1002/1096-9861(20000710)422:4<496::AID-CNE2>3.0.CO;2-4

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space, in: Beals, R., Beck, A., Bellow, A., Hajian, A. (Eds.), Contemporary Mathematics. American Mathematical Society, Providence, Rhode Island, pp. 189–206. https://doi.org/10.1090/conm/026/737400

Jones, D.T., 2019. Setting the standards for machine learning in biology. Nat Rev Mol Cell Biol 20, 659–660. https://doi.org/10.1038/s41580-019-0176-5

Jones, F.N., Jones, M.H., 1953. Modern Theories of Olfaction: A Critical Review. The Journal of Psychology 36, 207–241. https://doi.org/10.1080/00223980.1953.9712890

Jones, P.L., Pask, G.M., Romaine, I.M., Taylor, R.W., Reid, P.R., Waterson, A.G., Sulikowski, G.A., Zwiebel, L.J., 2012. Allosteric Antagonism of Insect Odorant Receptor Ion Channels. PLoS ONE 7, e30304. https://doi.org/10.1371/journal.pone.0030304

Kalbe, B., Schulz, V.M., Schlimm, M., Philippou, S., Jovancevic, N., Jansen, F., Scholz, P., Lübbert, H., Jarocki, M., Faissner, A., Hecker, E., Veitinger, S., Tsai, T., Osterloh, S., Hatt, H., 2017. Helional-induced activation of human olfactory receptor 2J3 promotes apoptosis and inhibits proliferation in a non-small-cell lung cancer cell line. European Journal of Cell Biology 96, 34–46. https://doi.org/10.1016/j.ejcb.2016.11.004

Karelson, M., Maran, U., Wang, Y., Katritzky, A.R., 1999. QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of CODESSA Applications. Collect. Czech. Chem. Commun. 64, 1551–1571. https://doi.org/10.1135/cccc19991551

Katada, S., 2005. Structural Basis for a Broad But Selective Ligand Spectrum of a Mouse Olfactory Receptor: Mapping the Odorant-Binding Site. Journal of Neuroscience 25, 1806–1815. https://doi.org/10.1523/JNEUROSCI.4723-04.2005

Katritzky, A.R., Gordeeva, E.V., 1993. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. Journal of chemical information and computer sciences 33, 835–857.

Kay, L.M., Crk, T., Thorngate, J., 2005. A Redefinition of Odor Mixture Quality. Behavioral Neuroscience 119, 726–733. https://doi.org/10.1037/0735-7044.119.3.726

Kermen, F., Chakirian, A., Sezille, C., Joussain, P., Le Goff, G., Ziessel, A., Chastrette, M., Mandairon, N., Didier, A., Rouby, C., Bensafi, M., 2011. Molecular complexity determines the number of olfactory notes and the pleasantness of smells. Sci Rep 1, 206. https://doi.org/10.1038/srep00206

Kermen, F., Midroit, M., Kuczewski, N., Forest, J., Thévenet, M., Sacquet, J., Benetollo, C., Richard, M., Didier, A., Mandairon, N., 2016. Topographical representation of odor hedonics in the olfactory bulb. Nat Neurosci 19, 876–878. https://doi.org/10.1038/nn.4317

Khan, A.G., Parthasarathy, K., Bhalla, U.S., 2010. Odor representations in the mammalian olfactory bulb. Wiley Interdiscip Rev Syst Biol Med 2, 603–611. https://doi.org/10.1002/wsbm.85

Khedkar, S., Malde, A., Coutinho, E., Srivastava, S., 2007. Pharmacophore Modeling in Drug Discovery and Development: An Overview. MC 3, 187–197. https://doi.org/10.2174/157340607780059521

Kier, L.B., Hall, L.H., Murray, W.J., Randi, M., 1975. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. Journal of Pharmaceutical Sciences 64, 1971–1974. https://doi.org/10.1002/jps.2600641214

Kim, S.-H., Yoon, Y.C., Lee, A.S., Kang, N., Koo, J., Rhyu, M.-R., Park, J.-H., 2015a. Expression of human olfactory receptor 10J5 in heart aorta, coronary artery, and endothelial cells and its functional role in angiogenesis. Biochemical and Biophysical Research Communications 460, 404–408. https://doi.org/10.1016/j.bbrc.2015.03.046

Kim, S.-H., Yoon, Y.C., Lee, A.S., Kang, N., Koo, J., Rhyu, M.-R., Park, J.-H., 2015b. Expression of human olfactory receptor 10J5 in heart aorta, coronary artery, and endothelial cells and its functional role in angiogenesis. Biochem Biophys Res Commun 460, 404–408. https://doi.org/10.1016/j.bbrc.2015.03.046

Kistiakowsky, G.B., 1950. On the Theory of Odors. Science 112, 154–155. https://doi.org/10.1126/science.112.2901.154-a

Kohonen, T., 2013. Essentials of the self-organizing map. Neural Networks 37, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Kohonen, T., 1998. The self-organizing map. Neurocomputing 21, 1–6. https://doi.org/10.1016/S0925-2312(98)00030-7

Kraft, P., Mannschreck, A., 2010. The Enantioselectivity of Odor Sensation: Some Examples for Undergraduate Chemistry Courses. J. Chem. Educ. 87, 598–603. https://doi.org/10.1021/ed100128v

Kurian, S.M., Naressi, R.G., Manoel, D., Barwich, A.-S., Malnic, B., Saraiva, L.R., 2021. Odor coding in the mammalian olfactory epithelium. Cell Tissue Res 383, 445–456. https://doi.org/10.1007/s00441-020-03327-1

Kurogi, Y., Guner, O., 2001. Pharmacophore Modeling and Three-dimensional Database Searching for Drug Design Using Catalyst. CMC 8, 1035–1055. https://doi.org/10.2174/0929867013372481

Le Bon, A.-M., Tromelin, A., Thomas-Danguin, T., Briand, L., 2008. Les récepteurs olfactifs et le codage des odeurs. Cahiers de Nutrition et de Diététique 43, 282–288. https://doi.org/10.1016/S0007-9960(08)75569-X

Le, Y., Murphy, P., Wang, J., 2002. Formyl-peptide receptors revisited. Trends in Immunology 23, 541–548. https://doi.org/10.1016/S1471-4906(02)02316-5

Leach, A.R., Gillet, V.J., Lewis, R.A., Taylor, R., 2010. Three-Dimensional Pharmacophore Methods in Drug Discovery. J. Med. Chem. 53, 539–558. https://doi.org/10.1021/jm900817u

Lee, J., 2010. Introduction to topological manifolds, Springer Science&Business Media. ed.

Leopold, D.A., Hummel, T., Schwob, J.E., Hong, S.C., Knecht, M., Kobal, G., 2000. Anterior Distribution of Human Olfactory Epithelium: The Laryngoscope 110, 417–421. https://doi.org/10.1097/00005537-200003000-00016

Levoin, N., Calmels, T., Krief, S., Danvy, D., Berrebi-Bertrand, I., Lecomte, J.-M., Schwartz, J.-C., Capet, M., 2011. Homology Model Versus X-ray Structure in Receptor-based Drug Design: A Retrospective Analysis with the Dopamine D3 Receptor. ACS Med. Chem. Lett. 2, 293–297. https://doi.org/10.1021/ml100288q

Li, Z.R., Han, L.Y., Xue, Y., Yap, C.W., Li, H., Jiang, L., Chen, Y.Z., 2007. MODEL—molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. Biotechnol. Bioeng. 97, 389–396. https://doi.org/10.1002/bit.21214

Liao, T.W., Triantaphyllou, E., 2008. Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications. WORLD SCIENTIFIC. https://doi.org/10.1142/6689

Linster, C., Cleland, T.A., 2004. Configurational and elemental odor mixture perception can arise from local inhibition. J Comput Neurosci 16, 39–47. https://doi.org/10.1023/b:jcns.0000004840.87570.2e

Liu, R., Zhou, D., 2008. Using Molecular Fingerprint as Descriptors in the QSPR Study of Lipophilicity. J. Chem. Inf. Model. 48, 542–549. https://doi.org/10.1021/ci700372s

Liu, X., Zhu, F., H. Ma, X., Shi, Z., Y. Yang, S., Q. Wei, Y., Z. Chen, Y., 2013. Predicting Targeted Polypharmacology for Drug Repositioning and Multi- Target Drug Discovery. CMC 20, 1646–1661. https://doi.org/10.2174/0929867311320130005

Lledo, P.-M., Gheusi, G., Vincent, J.-D., 2005. Information processing in the mammalian olfactory system. Physiol Rev 85, 281–317. https://doi.org/10.1152/physrev.00008.2004

Lodovichi, C., 2021. Topographic organization in the olfactory bulb. Cell Tissue Res 383, 457–472. https://doi.org/10.1007/s00441-020-03348-w

Luo, M., Katz, L., 2004. Encoding pheromonal signals in the mammalian vomeronasal system. Current Opinion in Neurobiology 14, 428–434. https://doi.org/10.1016/j.conb.2004.07.001

Ma, Y., Zhu, L., 2013. A Review on Dimension Reduction: *A Review on Dimension Reduction*. International Statistical Review 81, 134–150. https://doi.org/10.1111/j.1751-5823.2012.00182.x

Majid, A., 2021. Human Olfaction at the Intersection of Language, Culture, and Biology. Trends in Cognitive Sciences 25, 111–123. https://doi.org/10.1016/j.tics.2020.11.005

Malnic, B., Godfrey, P.A., Buck, L.B., 2004. The human olfactory receptor gene family. Proc Natl Acad Sci U S A 101, 2584–2589. https://doi.org/10.1073/pnas.0307882100

Malnic, B., Hirono, J., Sato, T., Buck, L.B., 1999a. Combinatorial receptor codes for odors. Cell 96, 713–723. https://doi.org/10.1016/s0092-8674(00)80581-4

Malnic, B., Hirono, J., Sato, T., Buck, L.B., 1999b. Combinatorial receptor codes for odors. Cell 96, 713–723. https://doi.org/10.1016/s0092-8674(00)80581-4

Mandairon, N., Poncelet, J., Bensafi, M., Didier, A., 2009. Humans and Mice Express Similar Olfactory Preferences. PLoS ONE 4, e4209. https://doi.org/10.1371/journal.pone.0004209

Manteniotis, S., Wojcik, S., Brauhoff, P., Möllmann, M., Petersen, L., Göthert, J., Schmiegel, W., Dührsen, U., Gisselmann, G., Hatt, H., 2016. Functional characterization of the

ectopically expressed olfactory receptor 2AT4 in human myelogenous leukemia. Cell Death Discovery 2, 15070. https://doi.org/10.1038/cddiscovery.2015.70

Marin, C., Vilas, D., Langdon, C., Alobid, I., López-Chacón, M., Haehner, A., Hummel, T., Mullol, J., 2018. Olfactory Dysfunction in Neurodegenerative Diseases. Curr Allergy Asthma Rep 18, 42. https://doi.org/10.1007/s11882-018-0796-4

Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I., Pavlik, P.A., 1993. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. J Computer-Aided Mol Des 7, 83–102. https://doi.org/10.1007/BF00141577

Maßberg, D., Simon, A., Häussinger, D., Keitel, V., Gisselmann, G., Conrad, H., Hatt, H., 2015. Monoterpene (−)-citronellal affects hepatocarcinoma cell signaling via an olfactory receptor. Archives of Biochemistry and Biophysics 566, 100–109. https://doi.org/10.1016/j.abb.2014.12.004

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: An easy approach to molecular descriptor calculations. Match 56, 237–248.

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [cs, stat].

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: Uniform Manifold Approximation and Projection. JOSS 3, 861. https://doi.org/10.21105/joss.00861

Meierhenrich, U., Golebiowski, J., Fernandez, X., Cabrol-Bass, D., 2005a. De la molécule à l'odeur : Les bases moléculaires des premières étapes de l'olfaction.

Meierhenrich, U., Golebiowski, J., Fernandez, X., Cabrol-Bass, D., 2005b. De la molécule à l'odeur : Les bases moléculaires des premières étapes de l'olfaction.

Meierhenrich, U.J., Golebiowski, J., Fernandez, X., Cabrol-Bass, D., 2004. The Molecular Basis of Olfactory Chemoreception. Angew. Chem. Int. Ed. 43, 6410–6412. https://doi.org/10.1002/anie.200462322

MOE, 2005. . Chemical Computing Group Inc, Montreal, Quebec, Canada.

Mori, I., 2013. Olfaction, in: Brenner's Encyclopedia of Genetics. Elsevier, pp. 161–163. https://doi.org/10.1016/B978-0-12-374984-0.01088-3

Mori, I., 2001. Olfaction, in: Encyclopedia of Genetics. Elsevier, pp. 1368–1370. https://doi.org/10.1006/rwgn.2001.0924

Mori, K., 2003. Grouping of odorant receptors: odour maps in the mammalian olfactory bulb. Biochemical Society Transactions 31, 134–136. https://doi.org/10.1042/bst0310134

Mori, K., 1987. Membrane and synaptic properties of identified neurons in the olfactory bulb. Progress in Neurobiology 29, 275–320. https://doi.org/10.1016/0301-0082(87)90024-4

Mori, K., Sakano, H., 2011. How is the olfactory map formed and interpreted in the mammalian brain? Annu Rev Neurosci 34, 467–499. https://doi.org/10.1146/annurev-neuro-112210-112917

Mori, K., Shepherd, G.M., 1994. Emerging principles of molecular signal processing by mitral/tufted cells in the olfactory bulb. Seminars in Cell Biology 5, 65–74. https://doi.org/10.1006/scel.1994.1009

Morris, G.M., Lim-Wilby, M., 2008. Molecular Docking, in: Kukol, A. (Ed.), Molecular Modeling of Proteins, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365–382. https://doi.org/10.1007/978-1-59745-177-2_19

Morrison, G.L., Fontaine, C.J., Harley, C.W., Yuan, Q., 2013. A role for the anterior piriform cortex in early odor preference learning: evidence for multiple olfactory learning structures in the rat pup. Journal of Neurophysiology 110, 141–152. https://doi.org/10.1152/jn.00072.2013

Muegge, I., Mukherjee, P., 2016. An overview of molecular fingerprint similarity search in virtual screening. Expert Opinion on Drug Discovery 11, 137–148. https://doi.org/10.1517/17460441.2016.1117070

Muir, R.M., Hansch, C.H., Gallup, A.H., 1949. GROWTH REGULATION BY ORGANIC COMPOUNDS. Plant Physiol. 24, 359–366. https://doi.org/10.1104/pp.24.3.359

Müller-Linow, M., Hilgetag, C.C., Hütt, M.-T., 2008. Organization of Excitable Dynamics in Hierarchical Biological Networks. PLoS Comput Biol 4, e1000190. https://doi.org/10.1371/journal.pcbi.1000190

Murphy, C., 2019. Olfactory and other sensory impairments in Alzheimer disease. Nat Rev Neurol 15, 11–24. https://doi.org/10.1038/s41582-018-0097-5

Nagashima, A., Touhara, K., 2010. Enzymatic Conversion of Odorants in Nasal Mucus Affects Olfactory Glomerular Activation Patterns and Odor Perception. Journal of Neuroscience 30, 16391–16398. https://doi.org/10.1523/JNEUROSCI.2527-10.2010

Nagayama, S., Homma, R., Imamura, F., 2014. Neuronal organization of olfactory bulb circuits. Front Neural Circuits 8, 98. https://doi.org/10.3389/fncir.2014.00098

Neuhaus, E.M., Mashukova, A., Barbour, J., Wolters, D., Hatt, H., 2006. Novel function of β-arrestin2 in the nucleus of mature spermatozoa. Journal of Cell Science 119, 3047–3056. https://doi.org/10.1242/jcs.03046

Newman, M., 2010. Networks. Oxford University Press.

Nisius, B., Bajorath, J., 2009. Molecular Fingerprint Recombination: Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types. ChemMedChem 4, 1859–1863. https://doi.org/10.1002/cmdc.200900243

O'Boyle, N.M., Sayle, R.A., 2016. Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminform 8, 36. https://doi.org/10.1186/s13321-016-0148-0

Oh, P., Monge, P., 2016. Network Theory and Models, in: Jensen, K.B., Rothenbuhler, E.W., Pooley, J.D., Craig, R.T. (Eds.), The International Encyclopedia of Communication Theory and Philosophy. Wiley, pp. 1–15. https://doi.org/10.1002/9781118766804.wbiect246

Oh, S.J., 2021a. Implications of the simple chemical structure of the odorant molecules interacting with the olfactory receptor 1A1. Genomics Inform 19, e18. https://doi.org/10.5808/gi.21033

Oh, S.J., 2021b. Computational evaluation of interactions between olfactory receptor OR2W1 and its ligands. Genomics Inform 19, e9. https://doi.org/10.5808/gi.21026

Oliveira, F.H.M., Machado, A.R.P., Andrade, A.O., 2018. On the Use of t-Distributed Stochastic Neighbor Embedding for Data Visualization and Classification of Individuals with Parkinson's Disease. Computational and Mathematical Methods in Medicine 2018, 1–17. https://doi.org/10.1155/2018/8019232

Ordonez, C., 2003. Clustering binary data streams with K-means, in: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03. Presented at the the 8th ACM SIGMOD workshop, ACM Press, San Diego, California, p. 12. https://doi.org/10.1145/882082.882087

Park, J.H., Morizumi, T., Li, Y., Hong, J.E., Pai, E.F., Hofmann, K.P., Choe, H.-W., Ernst, O.P., 2013. Opsin, a Structural Model for Olfactory Receptors? Angew. Chem. 125, 11227–11230. https://doi.org/10.1002/ange.201302374

Parma, V., Ohla, K., Veldhuizen, M.G., Niv, M.Y., Kelly, C.E., Bakke, A.J., Cooper, K.W., Bouysset, C., Pirastu, N., Dibattista, M., Kaur, R., Liuzza, M.T., Pepino, M.Y., Schöpf, V., Pereda-Loth, V., Olsson, S.B., Gerkin, R.C., Rohlfs Domínguez, P., Albayay, J., Farruggia, M.C., Bhutani, S., Fjaeldstad, A.W., Kumar, R., Menini, A., Bensafi, M., Sandell, M., Konstantinidis, I., Di Pizio, A., Genovese, F., Öztürk, L., Thomas-Danguin, T., Frasnelli, J., Boesveldt, S., Saatci, Ö., Saraiva, L.R., Lin, C., Golebiowski, J., Hwang, L.-D., Ozdener, M.H., Guàrdia, M.D., Laudamiel, C., Ritchie, M., Havlícek, J., Pierron, D., Roura, E., Navarro, M., Nolden, A.A., Lim, J., Whitcroft, K.L., Colquitt, L.R., Ferdenzi, C., Brindha, E.V., Altundag, A., Macchi, A., Nunez-Parra, A., Patel, Z.M., Fiorucci, S., Philpott, C.M., Smith, B.C., Lundström, J.N., Mucignat, C., Parker, J.K., van den Brink, M., Schmuker, M., Fischmeister, F.P.S., Heinbockel, T., Shields, V.D.C., Faraji, F., Santamaría, E., Fredborg, W.E.A., Morini, G., Olofsson, J.K., Jalessi, M., Karni, N., D'Errico, A., Alizadeh, R., Pellegrino, R., Meyer, P., Huart, C., Chen, B., Soler, G.M., Alwashahi, M.K., Welge-Lüssen, A., Freiherr, J., de Groot, J.H.B., Klein, H., Okamoto, M., Singh, P.B., Hsieh, J.W., GCCR Group Author, Abdulrahman, O., Dalton, P., Yan, C.H., Voznessenskaya, V.V., Chen, J., Sell, E.A., Walsh-Messinger, J., Archer, N.S., Koyama, S., Deary, V., Roberts, S.C., Yanık, H., Albayrak, S., Nováková, L.M., Croijmans, I., Mazal, P.P., Moein, S.T., Margulis, E., Mignot, C., Mariño, S., Georgiev, D., Kaushik, P.K., Malnic, B., Wang, H., Seyed-Allaei, S., Yoluk, N., Razzaghi-Asl, S., Justice, J.M., Restrepo, D., Reed, D.R., Hummel, T., Munger, S.D., Hayes, J.E., 2020. More Than Smell—COVID-19 Is Associated With Severe Impairment of Smell, Taste, and Chemesthesis. Chemical Senses 45, 609–622. https://doi.org/10.1093/chemse/bjaa041

Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi, S., 2000. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. J. Med. Chem. 43, 3233–3243. https://doi.org/10.1021/jm000941m

Patte, F., Laffort, P., 1979. An alternative model of olfactory quantitative interaction in binary mixtures. Chem Senses 4, 267–274. https://doi.org/10.1093/chemse/4.4.267

Pelosi, P., Knoll, W., 2022. Odorant-binding proteins of mammals. Biological Reviews 97, 20–44. https://doi.org/10.1111/brv.12787

Peng, M., Coutts, D., Wang, T., Cakmak, Y.O., 2019. Systematic review of olfactory shifts related to obesity. Obesity Reviews 20, 325–338. https://doi.org/10.1111/obr.12800

Peterlin, Z., Li, Y., Sun, G., Shah, R., Firestein, S., Ryan, K., 2008. The Importance of Odorant Conformation to the Binding and Activation of a Representative Olfactory Receptor. Chemistry & Biology 15, 1317–1327. https://doi.org/10.1016/j.chembiol.2008.10.014

Phase, Schrödinger, 2021. . LLC, New York, NY.

Price, J.L., 2009. Olfactory Higher Centers Anatomy, in: Encyclopedia of Neuroscience. Elsevier, pp. 129–136. https://doi.org/10.1016/B978-008045046-9.01692-2

Qing, X., Lee, X.Y., De Raeymaecker, J., Tame, J., Zhang, K., De Maeyer, M., Voet, A., 2014. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. JRLCR 81. https://doi.org/10.2147/JRLCR.S46843

Quignon, P., Rimbault, M., Robin, S., Galibert, F., 2012. Genetics of canine olfaction and receptor diversity. Mamm Genome 23, 132–143. https://doi.org/10.1007/s00335-011-9371-1

Raka, R.N., Wu, H., Xiao, J., Hossen, I., Cao, Y., Huang, M., Jin, J., 2022. Human ectopic olfactory receptors and their food originated ligands: a review. Critical Reviews in Food Science and Nutrition 62, 5424–5443. https://doi.org/10.1080/10408398.2021.1885007

Randic, M., 1975. Characterization of molecular branching. J. Am. Chem. Soc. 97, 6609–6615. https://doi.org/10.1021/ja00856a001

Ranzani, M., Iyer, V., Ibarra-Soria, X., Del Castillo Velasco-Herrera, M., Garnett, M., Logan, D., Adams, D.J., 2017. Revisiting olfactory receptors as putative drivers of cancer. Wellcome Open Res 2, 9. https://doi.org/10.12688/wellcomeopenres.10646.1

Robert-Hazotte, A., Faure, P., Ménétrier, F., Folia, M., Schwartz, M., Le Quéré, J.-L., Neiers, F., Thomas-Danguin, T., Heydel, J.-M., 2022. Nasal Odorant Competitive Metabolism Is Involved in the Human Olfactory Process. J. Agric. Food Chem. acs.jafc.2c02720. https://doi.org/10.1021/acs.jafc.2c02720

Rodriguez, I., Greer, C.A., Mok, M.Y., Mombaerts, P., 2000. A putative pheromone receptor gene expressed in human olfactory mucosa. Nat Genet 26, 18–19. https://doi.org/10.1038/79124

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J Chem Inf Model 50, 742–754. https://doi.org/10.1021/ci100050t

Rospars, J.-P., 2013. Interactions of Odorants with Olfactory Receptors and Other Preprocessing Mechanisms: How Complex and Difficult to Predict? Chemical Senses 38, 283–287. https://doi.org/10.1093/chemse/bjt004

Rospars, J.-P., Lansky, P., Chaput, M., Duchamp-Viret, P., 2008. Competitive and Noncompetitive Odorant Interactions in the Early Neural Coding of Odorant Mixtures. Journal of Neuroscience 28, 2659–2666. https://doi.org/10.1523/JNEUROSCI.4670-07.2008

Rouquier, S., Blancher, A., Giorgi, D., 2000. The olfactory receptor gene repertoire in primates and mouse: Evidence for reduction of the functional fraction in primates. Proc. Natl. Acad. Sci. U.S.A. 97, 2870–2874. https://doi.org/10.1073/pnas.040580197

Roy, K., Kar, S., Das, R.N., 2015. A primer on QSAR/QSPR modeling: fundamental concepts, Springer. ed.

Royet, J.-P., Plailly, J., Delon-Martin, C., Kareken, D.A., Segebarth, C., 2003. fMRI of emotional responses to odors: NeuroImage 20, 713–728. https://doi.org/10.1016/S1053-8119(03)00388-4

Russell, G.F., Hills, J.I., 1971. Odor Differences between Enantiomeric Isomers. Science 172, 1043–1044. https://doi.org/10.1126/science.172.3987.1043

Ruther, J., Meiners, T., Steidle, J.L.M., 2002. Rich in phenomena-lacking in terms. A classification of kairomones. Chemoecology 12, 161–167. https://doi.org/10.1007/PL00012664

Sabiniewicz, A., Heyne, F., Hummel, T., 2021. Odors modify emotional responses. Flavour Fragr J 36, 256–263. https://doi.org/10.1002/ffj.3640

Sachdeva, V., Freimuth, D.M., Mueller, C., 2009. Evaluating the Jaccard-Tanimoto Index on Multi-core Architectures, in: Allen, G., Nabrzyski, J., Seidel, E., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (Eds.), Computational Science – ICCS 2009, Lecture Notes in

Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 944–953. https://doi.org/10.1007/978-3-642-01970-8_95

Saeed, N., Nam, H., Al-Naffouri, T.Y., Alouini, M.-S., 2019. A State-of-the-Art Survey on Multidimensional Scaling-Based Localization Techniques. IEEE Commun. Surv. Tutorials 21, 3565–3583. https://doi.org/10.1109/COMST.2019.2921972

Saini, K., Ramanathan, V., 2022. A Review of Machine Learning Approaches to Predicting Molecular Odor in the Context of Multi-Label Classification (preprint). In Review. https://doi.org/10.21203/rs.3.rs-1492792/v1

Saito, H., Chi, Q., Zhuang, H., Matsunami, H., Mainland, J.D., 2009. Odor Coding by a Mammalian Receptor Repertoire. Sci. Signal. 2. https://doi.org/10.1126/scisignal.2000016

Salazar, I., Sanchez-Quinteiro, P., Barrios, A.W., López Amado, M., Vega, J.A., 2019. Anatomy of the olfactory mucosa, in: Handbook of Clinical Neurology. Elsevier, pp. 47–65. https://doi.org/10.1016/B978-0-444-63855-7.00004-6

Sankaran, S., Khot, L.R., Panigrahi, S., 2012. Biology and applications of olfactory sensing system: A review. Sensors and Actuators B: Chemical 171–172, 1–17. https://doi.org/10.1016/j.snb.2012.03.029

Schaller, D., Šribar, D., Noonan, T., Deng, L., Nguyen, T.N., Pach, S., Machalz, D., Bermudez, M., Wolber, G., 2020. Next generation 3D pharmacophore modeling. WIREs Comput Mol Sci 10. https://doi.org/10.1002/wcms.1468

Schienle, A., Wolf, A., Tomazic, P.V., Ille, R., 2018. Affective Personality Traits in Olfactory Dysfunction: the Role of Dysthymia and Arousal. Chem. Percept. 11, 72–76. https://doi.org/10.1007/s12078-017-9242-6

Schoppa, N.E., Urban, N.N., 2003. Dendritic processing within olfactory bulb circuits. Trends in Neurosciences 26, 501–506. https://doi.org/10.1016/S0166-2236(03)00228-5

Scott, J.W., Wellis, D.P., Riggott, M.J., Buonviso, N., 1993. Functional organization of the main olfactory bulb. Microsc. Res. Tech. 24, 142–156. https://doi.org/10.1002/jemt.1070240206

Sedlmair, M., Brehmer, M., Ingram, S., Munzner, T., 2012. Dimensionality Reduction in the Wild: Gaps and Guidance.

Sharma, A., Kumar, R., Aier, I., Semwal, R., Tyagi, P., Varadwaj, P., 2019. Sense of Smell: Structural, Functional, Mechanistic Advancements and Challenges in Human Olfactory Research. CN 17, 891–911. https://doi.org/10.2174/1570159X17666181206095626

Sharma, A., Kumar, R., Ranjta, S., Varadwaj, P.K., 2021. SMILES to Smell: Decoding the Structure–Odor Relationship of Chemical Compounds Using the Deep Neural Network Approach. J. Chem. Inf. Model. 61, 676–688. https://doi.org/10.1021/acs.jcim.0c01288

Sharma, A., Saha, B.K., Kumar, R., Varadwaj, P.K., 2022. OlfactionBase: a repository to explore odors, odorants, olfactory receptors and odorant–receptor interactions. Nucleic Acids Research 50, D678–D686. https://doi.org/10.1093/nar/gkab763

Shih, C.-T., Sporns, O., Yuan, S.-L., Su, T.-S., Lin, Y.-J., Chuang, C.-C., Wang, T.-Y., Lo, C.-C., Greenspan, R.J., Chiang, A.-S., 2015. Connectomics-Based Analysis of Information Flow in the Drosophila Brain. Current Biology 25, 1249–1258. https://doi.org/10.1016/j.cub.2015.03.021

Shlens, J., 2014. Notes on Kullback-Leibler Divergence and Likelihood.

Showell, H.J., Freer, R.J., Zigmond, S.H., Schiffmann, E., Aswanikumar, S., Corcoran, B., Becker, E.L., 1976. The structure-activity relations of synthetic peptides as chemotactic factors and inducers of lysosomal secretion for neutrophils. Journal of Experimental Medicine 143, 1154–1169. https://doi.org/10.1084/jem.143.5.1154

Silva, L., Antunes, A., 2017. Vomeronasal Receptors in Vertebrates and the Evolution of Pheromone Detection. Annu. Rev. Anim. Biosci. 5, 353–370. https://doi.org/10.1146/annurev-animal-022516-022801

Silverman, B.D., Platt, Daniel.E., 1996. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. J. Med. Chem. 39, 2129–2140. https://doi.org/10.1021/jm950589q

Sinding, C., Thomas-Danguin, T., Chambault, A., Béno, N., Dosne, T., Chabanet, C., Schaal, B., Coureaud, G., 2013. Rabbit Neonates and Human Adults Perceive a Blending 6-Component Odor Mixture in a Comparable Manner. PLoS ONE 8, e53534. https://doi.org/10.1371/journal.pone.0053534

Smellie, A., Kahn, S.D., Teig, S.L., 1995a. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. J. Chem. Inf. Comput. Sci. 35, 285–294. https://doi.org/10.1021/ci00024a018

Smellie, A., Kahn, S.D., Teig, S.L., 1995b. Analysis of Conformational Coverage. 2. Applications of Conformational Models. J. Chem. Inf. Comput. Sci. 35, 295–304. https://doi.org/10.1021/ci00024a019

Smellie, A., Teig, S.L., Towbin, P., 1995c. Poling: Promoting conformational variation. J. Comput. Chem. 16, 171–187. https://doi.org/10.1002/jcc.540160205

Snitz, K., Perl, O., Honigstein, D., Secundo, L., Ravia, A., Yablonka, A., Endevelt-Shapira, Y., Sobel, N., 2019. SmellSpace: An Odor-Based Social Network as a Platform for Collecting Olfactory Perceptual Data. Chemical Senses 44, 267–278. https://doi.org/10.1093/chemse/bjz014

Snitz, K., Yablonka, A., Weiss, T., Frumin, I., Khan, R.M., Sobel, N., 2013. Predicting odor perceptual similarity from odor structure. PLoS Comput Biol 9, e1003184. https://doi.org/10.1371/journal.pcbi.1003184

Soh, Z., Tsuji, T., Takiguchi, N., Ohtake, H., 2012. A neural network model for olfactory glomerular activity prediction. Presented at the INTERNATIONAL CONFERENCE OF COMPUTATIONAL METHODS IN SCIENCES AND ENGINEERING 2009: (ICCMSE 2009), Rhodes, Greece, pp. 1261–1264. https://doi.org/10.1063/1.4772159

Soudry, Y., Lemogne, C., Malinvaud, D., Consoli, S.-M., Bonfils, P., 2011. Olfactory system and emotion: Common substrates. European Annals of Otorhinolaryngology, Head and Neck Diseases 128, 18–23. https://doi.org/10.1016/j.anorl.2010.09.007

Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J.A., Wetzel, C.H., Zimmer, R.K., Hatt, H., 2003a. Identification of a Testicular Odorant Receptor Mediating Human Sperm Chemotaxis. Science 299, 2054–2058. https://doi.org/10.1126/science.1080376

Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J.A., Wetzel, C.H., Zimmer, R.K., Hatt, H., 2003b. Identification of a Testicular Odorant Receptor Mediating Human Sperm Chemotaxis. Science 299, 2054–2058. https://doi.org/10.1126/science.1080376

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

Spitzer, G.M., Heiss, M., Mangold, M., Markt, P., Kirchmair, J., Wolber, G., Liedl, K.R., 2010. One Concept, Three Implementations of 3D Pharmacophore-Based Virtual Screening: Distinct Coverage of Chemical Search Space. J. Chem. Inf. Model. 50, 1241–1247. https://doi.org/10.1021/ci100136b

Spitzer, G.M., Wellenzohn, B., Laggner, C., Langer, T., Liedl, K.R., 2007. DNA Minor Groove Pharmacophores Describing Sequence Specific Properties. J. Chem. Inf. Model. 47, 1580–1589. https://doi.org/10.1021/ci600500v

Stanton, D.T., Jurs, P.C., 1990. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. Anal. Chem. 62, 2323–2329. https://doi.org/10.1021/ac00220a013

Startek, J.B., Voets, T., Talavera, K., 2019. To flourish or perish: evolutionary TRiPs into the sensory biology of plant-herbivore interactions. Pflugers Arch - Eur J Physiol 471, 213–236. https://doi.org/10.1007/s00424-018-2205-1

Stettler, D.D., Axel, R., 2009. Representations of Odor in the Piriform Cortex. Neuron 63, 854–864. https://doi.org/10.1016/j.neuron.2009.09.005

Stevenson, R.J., Mahmut, M.K., 2013. The accessibility of semantic knowledge for odours that can and cannot be named. Quarterly Journal of Experimental Psychology 66, 1414–1431. https://doi.org/10.1080/17470218.2012.753097

Strotmann, J., Breer, H., 2006. Formation of glomerular maps in the olfactory system. Seminars in Cell & Developmental Biology 17, 402–410. https://doi.org/10.1016/j.semcdb.2006.04.010

Suebsing, A., Hiransakolwong, N., 2012. A novel technique for feature subset selection based on cosine similarity. Applied Mathematical Sciences 6, 6627–6655.

Sutter, J., Guner, O., Hoffman, R., Li, H., Waldman, M., 2000. Effect of Variable Weights and Tolerances on Predictive Model Generation. Pharmacophore Perception, Development, and Use in Drug Design 501–511.

Tarca, A.L., Carey, V.J., Chen, X., Romero, R., Drăghici, S., 2007. Machine Learning and Its Applications to Biology. PLoS Comput Biol 3, e116. https://doi.org/10.1371/journal.pcbi.0030116

Thai, K.-M., Ngo, T.-D., Tran, T.-D., Le, M.-T., 2013. Pharmacophore Modeling for Antitargets. CTMC 13, 1002–1014. https://doi.org/10.2174/1568026611313090004

Thomas-Danguin, T., Sinding, C., Romagny, S., El Mountassir, F., Atanasova, B., Le Berre, E., Le Bon, A.-M., Coureaud, G., 2014. The perception of odor objects in everyday life: a review on the processing of odor mixtures. Front. Psychol. 5. https://doi.org/10.3389/fpsyg.2014.00504

Tintori, C., Corradi, V., Magnani, M., Manetti, F., Botta, M., 2008. Targets Looking for Drugs: A Multistep Computational Protocol for the Development of Structure-Based Pharmacophores and Their Applications for Hit Discovery. J. Chem. Inf. Model. 48, 2166–2179. https://doi.org/10.1021/ci800105p

Todeschini, R., Lasagni, M., Marengo, E., 1994. New molecular descriptors for 2D and 3D structures. Theory. J. Chemometrics 8, 263–272. https://doi.org/10.1002/cem.1180080405

Tong, T., Wang, Y., Kang, S.-G., Huang, K., 2021. Ectopic Odorant Receptor Responding to Flavor Compounds: Versatile Roles in Health and Disease. Pharmaceutics 13, 1314. https://doi.org/10.3390/pharmaceutics13081314

Touhara, K., 2002a. Odor discrimination by G protein-coupled olfactory receptors. Microsc Res Tech 58, 135–141. https://doi.org/10.1002/jemt.10131

Touhara, K., 2002b. Odor discrimination by G protein-coupled olfactory receptors. Microsc Res Tech 58, 135–141. https://doi.org/10.1002/jemt.10131

Trinajstić, N., Nikolić, S., Carter, S., 1989. QSAR: Theory and application. Kemija u industriji/Journal of Chemists and Chemical Engineers 38, 469–484.

van Drie, J., 2003. Pharmacophore Discovery - Lessons Learned. CPD 9, 1649–1664. https://doi.org/10.2174/1381612033454568

Vincent, A.J., West, A.K., Chuah, M.I., 2005. Morphological and functional plasticity of olfactory ensheathing cells. J Neurocytol 34, 65–80. https://doi.org/10.1007/s11068-005-5048-6

Vuorinen, A., Schuster, D., 2015. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. Methods 71, 113–134. https://doi.org/10.1016/j.ymeth.2014.10.013

Wang, S.-C., 2003. Artificial Neural Network, in: Interdisciplinary Computing in Java Programming. Springer US, Boston, MA, pp. 81–100. https://doi.org/10.1007/978-1-4615-0377-4_5

Wang, Y., Hu, J., Lai, J., Li, Y., Jin, H., Zhang, Lihe, Zhang, Liangren, Liu, Z., 2020. TF3P: Three-dimensional Force Fields Fingerprint Learned by Deep Capsular Network. J. Chem. Inf. Model. 60, 2754–2765. https://doi.org/10.1021/acs.jcim.0c00005

Weber, L., Al-Refae, K., Ebbert, J., Jägers, P., Altmüller, J., Becker, C., Hahn, S., Gisselmann, G., Hatt, H., 2017. Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. PLoS ONE 12, e0172491. https://doi.org/10.1371/journal.pone.0172491

Weber, L., Maßberg, D., Becker, C., Altmüller, J., Ubrig, B., Bonatz, G., Wölk, G., Philippou, S., Tannapfel, A., Hatt, H., Gisselmann, G., 2018. Olfactory Receptors as Biomarkers in Human Breast Carcinoma Tissues. Front. Oncol. 8, 33. https://doi.org/10.3389/fonc.2018.00033

Wen, T., Mo, Z., Li, J., Liu, Q., Wu, L., Luo, D., 2021. An Odor Labeling Convolutional Encoder–Decoder for Odor Sensing in Machine Olfaction. Sensors 21, 388. https://doi.org/10.3390/s21020388

Wendt, G.R., 1952. Somesthesis and the Chemical Senses. Annu. Rev. Psychol. 3, 105–130. https://doi.org/10.1146/annurev.ps.03.020152.000541

Wermuth, C.G., Ganellin, C.R., Lindberg, P., Mitscher, L.A., 1998. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure and Applied Chemistry 70, 1129–1143. https://doi.org/10.1351/pac199870051129

Wessels, Q., Hoogland, P.V.J.M., Vorster, W., 2014. Anatomical evidence for an endocrine activity of the vomeronasal organ in humans: Anatomical Evidence for an Endocrine Activity of the VNO. Clin. Anat. 27, 856–860. https://doi.org/10.1002/ca.22382

Wilson, D.A., 2003. Rapid, Experience-Induced Enhancement in Odorant Discrimination by Anterior Piriform Cortex Neurons. Journal of Neurophysiology 90, 65–72. https://doi.org/10.1152/jn.00133.2003

Wolber, G., Langer, T., 2005. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. J. Chem. Inf. Model. 45, 160–169. https://doi.org/10.1021/ci049885e

Wooten, C., 2015. Anatomy of the Olfactory Nerves, in: Nerves and Nerve Injuries. Elsevier, pp. 273–276. https://doi.org/10.1016/B978-0-12-410390-0.00019-6

Wright, R.H., 1954. Odour and molecular vibration. I. Quantum and thermodynamic considerations. J. Appl. Chem. 4, 611–615. https://doi.org/10.1002/jctb.5010041104

Xia, C., Ma, W., Wang, F., Hua, S., Liu, M., 2001. Identification of a prostate-specific G-protein coupled receptor in prostate cancer. Oncogene 20, 5903–5907. https://doi.org/10.1038/sj.onc.1204803

Xue, L., Godden, J.W., Stahura, F.L., Bajorath, J., 2003. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. J. Chem. Inf. Comput. Sci. 43, 1151–1157. https://doi.org/10.1021/ci030285+

Yang, S.-Y., 2010. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discovery Today 15, 444–450. https://doi.org/10.1016/j.drudis.2010.03.013

Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474. https://doi.org/10.1002/jcc.21707

Yoshida, I., Mori, K., 2007. Odorant Category Profile Selectivity of Olfactory Cortex Neurons. Journal of Neuroscience 27, 9105–9114. https://doi.org/10.1523/JNEUROSCI.2720-07.2007

Zeppilli, S., Ackels, T., Attey, R., Klimpert, N., Ritola, K.D., Boeing, S., Crombach, A., Schaefer, A.T., Fleischmann, A., 2021. Molecular characterization of projection neuron subtypes in the mouse olfactory bulb. eLife 10, e65445. https://doi.org/10.7554/eLife.65445

Zhang, K., Wang, L., Liu, Z., Geng, B., Teng, Y., Liu, X., Yi, Q., Yu, D., Chen, X., Zhao, D., Xia, Y., 2021. Mechanosensory and mechanotransductive processes mediated by ion channels in articular chondrocytes: Potential therapeutic targets for osteoarthritis. Channels 15, 339–359. https://doi.org/10.1080/19336950.2021.1903184

Zhang, S., Li, L., Li, H., 2021. Role of ectopic olfactory receptors in glucose and lipid metabolism. Br J Pharmacol 178, 4792–4807. https://doi.org/10.1111/bph.15666

Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11, 178–189. https://doi.org/10.1016/s1076-6332(03)00671-8

Partie 6 : Annexes

I. Informations supplémentaires de la partie 3-I

S1 Table. Fingerprint, coordinates in 2D spaces and clusters.

(Fichier Supplementary Table1-partie3-I.xlsx)

S2 Table. Odor notes and occurrences.

(Fichier Supplementary Table2-partie3-I.xlsx)

S3 Table. Distribution of the chemical groups and functions by cluster.

S4 Table. Table of chemical structures associated with odors.

S1 Fig. “Elbow” curve.

S2 Fig. Progression of the penalty score according to the number of clusters.

S3 Fig. Dendrograms of the AHC of molecules for each dimension reduction technique.

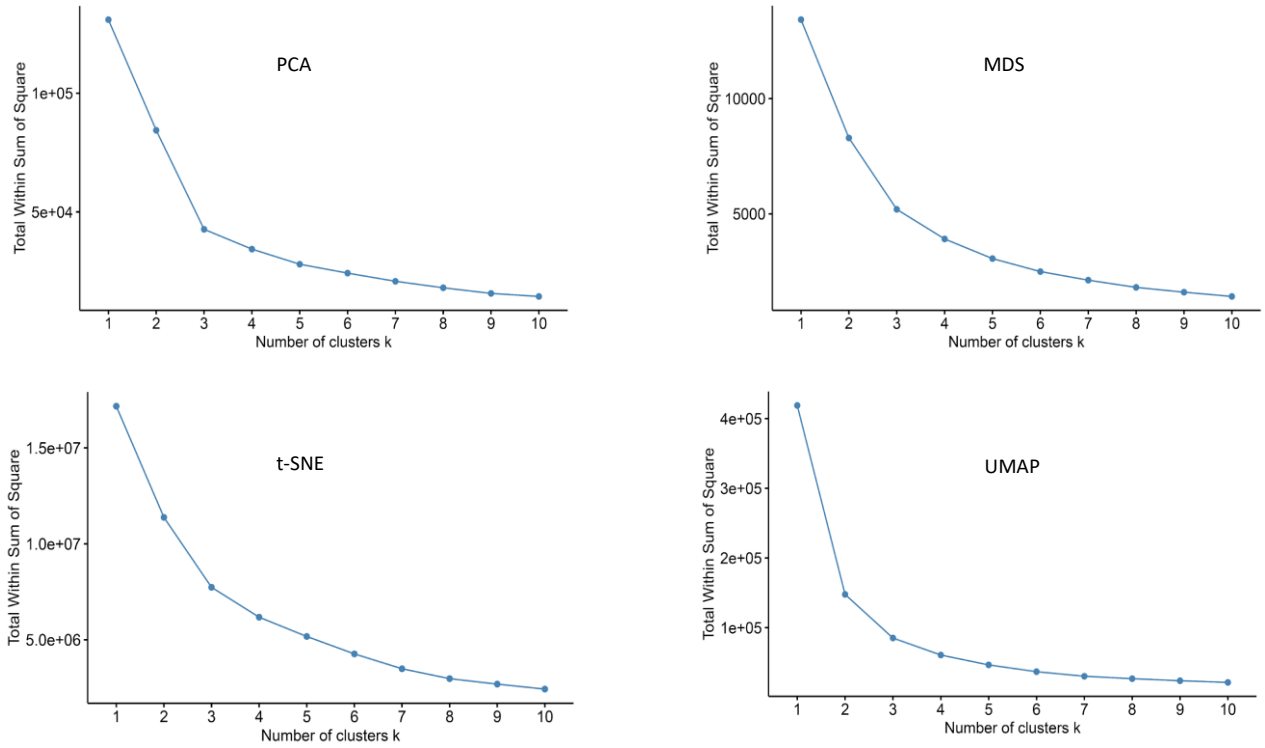
S4 Fig. Radar charts of the distribution of the %ON values obtained for the 17 most frequent odor notes across clusters of the UMAP-kmeans and UMAP-AHC techniques.

S3 Table. Distribution of the chemical groups and functions by cluster.

Chemical function	% molecules in C1	% molecules in C2	% molecules in C3	% molecules in C4
Carbonyl (C=O)	57,37%	54,80%	55,56%	80,70%
Aldehyde (R-CHO)	5,18%	8,21%	14,49%	5,67%
Ester (-COOR)	30,09%	31,38%	8,63%	66,33%
Carboxylic Acid (-COOH)	1,40%	3,22%	14,71%	4,79%
Aliphatic alcool (-OH)	19,98%	8,65%	25,30%	14,10%
Aromatic alcool (-OH)	0,55%	13,13%	0,08%	0,00%
Ketone (R-CO-R)	20,28%	10,80%	13,89%	7,56%
Aliphatic amine (R-NH2)	3,53%	2,53%	7,96%	0,27%
Aromatic amine (Ar-NH2)	0,12%	19,07%	0,68%	0,00%
Phenol (Ph-OH)	0,55%	12,31%	0,08%	0,00%
Benzene	7,13%	70,33%	1,58%	0,00%
Aryl-Me	1,10%	33,96%	0,45%	0,00%
Furan	1,58%	8,78%	0,15%	0,00%
Bicyclic	23,33%	9,91%	0,53%	0,00%
Unbranched carbon chain Cn n>3	9,87%	10,80%	29,88%	53,37%
Allylic group (CH2=CH-CH2-R)	44,76%	12,44%	32,06%	45,14%
Sulfide (R-S-R)	6,15%	3,47%	9,16%	3,44%
Thiol (R-SH)	1,16%	1,64%	9,38%	1,69%

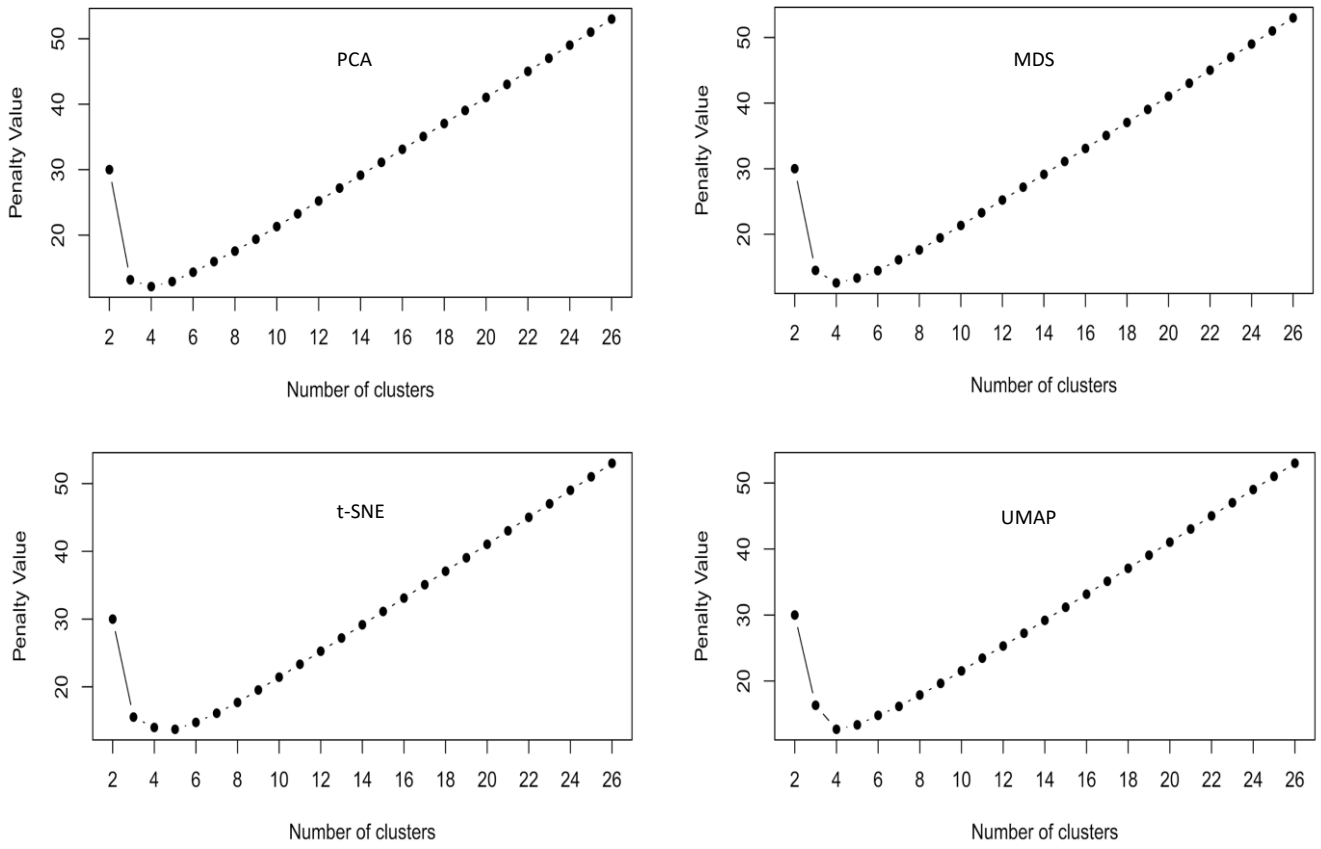
S4 Table. Table of chemical structures associated with odors.

Chemical group	Odor
Long chain	Fatty
Long chain	Waxy
Amino acids, carboxylic acids	Odorless
Ester	Fruity
Sulfur	Sulfurous
Sulfur	Pungent
Polycyclic	Woody, spicy
Allylic chain, carbonyl, ketone	Woody



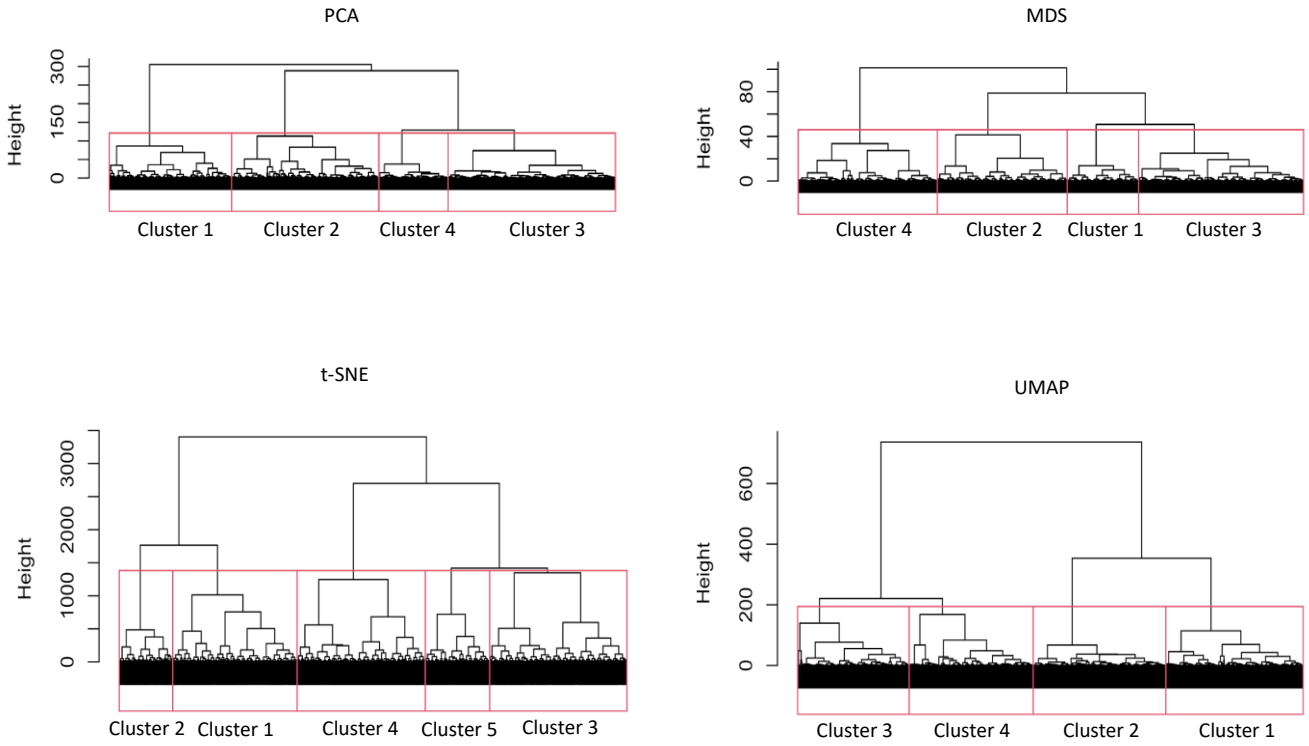
S1 Fig. "Elbow" curve.

Representation of intra-cluster variability as a function of the number of clusters. The optimal number of clusters is around the bend of the curve.

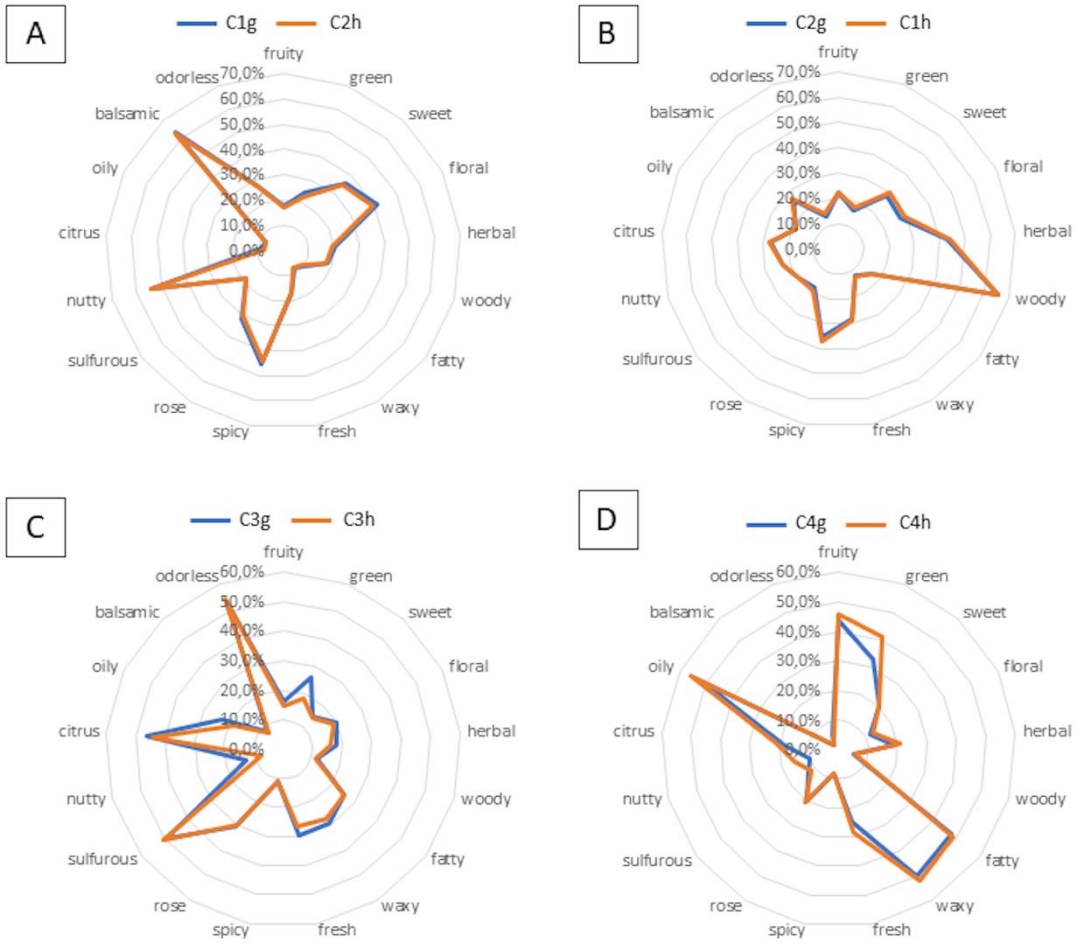


S2 Fig. Progression of the penalty score according to the number of clusters.

The minimum score is assigned to the optimal number of clusters.



S3 Fig. Dendrograms of the AHC of molecules for each dimension reduction technique.



S4 Fig. Radar charts of the distribution of the %ON values obtained for the 17 most frequent odor notes across clusters of the UMAP-kmeans and UMAP-AHC techniques.

A: Comparison between C1g and C2h. B: Comparison between C2g and C1h. C: Comparison between C3g and C3h. D: Comparison between C4g and C4h.

II. Informations supplémentaires de la partie 3-II

Table S1. List of the components of subsets of odorants based on odor profiles

Table S2. Overlap of clusters.

Table S3. List of clusters k-means and SOM of areas Ac-3D at level L16.

Table S4. Number of occurrences of the odor notes in the database and in the clusters obtained using k-means and SOM methods at the three levels of clustering.

(Fichier Supplementary Table4-partie3-II.xlsx)

Table S5. Co-occurrences symmetric square matrices of odor notes obtained for the whole database.

(Fichier Supplementary Table5-partie3-II.xlsx)

Table S6. Details of the PHASE-generated hypothesis from the subset composed of the mixture components.

Table S7. Distances between features of the hypotheses hyp-V-c, hyp-IA-c, hyp-F-c, hyp-EA-c, hyp-bD-c, hyp-bI-c, and hyp-WL-c.

Table S8. Details of the PHASE top ten hypotheses generated from V-s, IA-s, F-s, EA-s, bD-s, bI -s and W-s subsets.

Table S9. Distances between features of the hypotheses generated from V-s, IA-s, F-s, EA-s, bD-s, bI-s, and WL-s.

Fig S1. Visualizations of the compounds-odors dataset in the 2D space obtained using k-means and SOM clustering.

Fig S2. Radar charts of the distribution of the molecules carrying the 25 most frequent odor notes across clusters 3D-SOM16 that contain the molecules of interest.

Supplementary Table 1. List of the components of subsets of odorants based on odor profiles.

Subset	Name	IUPAC Name	CAS Number	Cluster 3D-SOM16	Odor description
V-s	vanillin	4-hydroxy-3-methoxybenzaldehyde	121-33-5	Cl-2	sweet, vanilla, creamy, chocolate
	vanillyl isobutyrate	(4-formyl-2-methoxyphenyl) 2-methylpropanoate	20665-85-4	Cl-2	sweet, vanilla, creamy, fruity, caramellic, chocolate
	vanillin propylene glycol acetal	2-methoxy-4-(4-methyl-1,3-dioxolan-2-yl)phenol	68527-74-2	Cl-2	sweet, vanilla, chocolate, creamy, dairy, powdery
	ethyl vanillin isobutyrate	3-ethoxy-4-hydroxybenzaldehyde; 2-methylpropanoate	188417-26-7	Cl-2	sweet, creamy, vanilla, chocolate, milky
	2-ethoxyanisole	1-ethoxy-2-methoxybenzene	17600-72-5	Cl-2	sweet, creamy, vanilla, earthy, nutty
	ortho-dimethyl hydroquinone	1,2-dimethoxybenzene	91-16-7	Cl-2	sweet, creamy, vanilla, phenolic, musty
	ethyl vanillin	3-ethoxy-4-hydroxybenzaldehyde	121-32-4	Cl-2	sweet, creamy, vanilla, caramellic
	vanillyl acetate	(4-formyl-2-methoxyphenyl) acetate	881-68-5	Cl-2	sweet, creamy, vanilla, powdery, floral
	vanillylidene acetone	(E)-4-(4-hydroxy-3-methoxyphenyl)but-3-en-2-one	1080-12-2	Cl-2	sweet, powdery, vanilla, creamy, balsamic
	vanillin hexylene glycol acetal	2-methoxy-4-(4,4,6-trimethyl-1,3-dioxan-2-yl)phenol	52514-66-6	Cl-2	sweet, creamy, vanilla, smoky
	ethyl vanillin hexylene glycol acetal	2-ethoxy-4-(4,4,6-trimethyl-1,3-dioxan-2-yl)phenol	52514-67-7	Cl-2	sweet, creamy, vanilla, phenolic
	ethyl vanillin propylene glycol acetal	2-ethoxy-4-(4-methyl-1,3-dioxolan-2-yl)phenol	68527-76-4	Cl-2	sweet, vanilla, creamy, spicy
IA-s	isoamyl acetate	3-methylbutyl acetate	123-92-2	Cl-13	sweet, fruity, banana, solvent, pear

	2-methylbutyl butyrate	2-methylbutyl butanoate	51115-64-1	Cl-13	fruity, sweet, apricot, apple, banana, pear
	hexyl acetate	hexyl acetate	142-92-7	Cl-13	fruity, green, apple, banana, sweet
	isobutyl propionate	2-methylpropyl propanoate	540-42-1	Cl-13	green, ethereal, sweet, fruity, banana
	methyl butyrate	methyl butanoate	623-42-7	Cl-13	fruity, apple, sweet, banana, pineapple
	isopropyl propionate	propan-2-yl propanoate	637-78-5	Cl-13	fruity, ethereal, sweet, pineapple, banana
	methyl 4-methyl valerate	methyl 4-methylpentanoate	2412-80-8	Cl-13	fruity, sweet, banana, pineapple, cheesy
	isoamyl butyrate	3-methylbutyl butanoate	106-27-4	Cl-13	fruity, green, apricot, pear, banana
	propyl acetate	propyl acetate	109-60-4	Cl-13	solvent, celery, fruity, raspberry, pear
	butyl acetate	butyl acetate	123-86-4	Cl-13	ethereal, solvent, fruity, banana
	amyl acetate	pentyl acetate	628-63-7	Cl-13	ethereal, fruity, pear, banana, apple
F-s	frambinone	4-(4-hydroxyphenyl)butan-2-one	5471-51-2	Cl-5	sweet, berry, raspberry, ripe, floral, fruity
	anisyl isobutyrate	Anisyl-isobutyrate	71172-26-4	Cl-5	sweet, floral, fruity
	4-hydroxyphenethyl alcohol	4-(2-hydroxyethyl)phenol	501-94-0	Cl-5	sweet, floral, fruity
	4-(para-tolyl)-2-butanone	4-(4-methylphenyl)butan-2-one	7774-79-0	Cl-5	sweet, fruity, raspberry, plum, floral
	tufurol acetate	tufurol-acetate	62346-96-7	Cl-5	sweet, floral, fresh, fruity
	2-methylbenzyl acetate	2-Methylbenzyl-acetate	17373-93-2	Cl-5	sweet, fruity, cherry, floral

	alpha-methylbenzyl propionate	1-phenylethyl propanoate	120-45-6	Cl-5	fruity, floral, sweet, green
	phenethyl 2-methylbutyrate	2-phenylethyl 2-methylbutanoate	24817-51-4	Cl-5	sweet, fruity, herbal, floral
	methyl 4-phenyl butyrate	methyl 4-phenylbutanoate	2046-17-5	Cl-5	fruity, honey, floral, sweet
	benzyl acetoacetate	benzyl 3-oxobutanoate	5396-89-4	Cl-5	sweet, banana, floral, fruity
EA-s	ethyl acetate	ethyl acetate	141-78-6	Cl-14	ethereal, fruity, sweet, weedy, green, sharp, brandy, winey
	prenyl formate	3-methylbut-2-enyl formate	68480-28-4	Cl-14	ethereal, green, fruity, winey
	Isobutyl pyruvate	isobutyl-pyruvate	13051-48-4	Cl-14	sweet, ethereal, fruity
	methyl acetate	methyl acetate	79-20-9	Cl-14	ethereal, sweet, fruity
	methyl (E)-2-butenoate	methyl (E)-but-2-enoate	623-43-8	Cl-14	sharp, green, fruity
	ethyl 2-methyl butyrate	ethyl 2-methylbutanoate	7452-79-1	Cl-14	sharp, sweet, green, apple, fruity
	2-methyl butyl propionate	2-methylbutyl propanoate	2438-20-2	Cl-14	sweet, fruity, ethereal, rummy
	isopropyl acetate	propan-2-yl acetate	108-21-4	Cl-14	ethereal, fruity, sweet, banana
	ethyl nitrite	ethyl nitrite	109-95-5	Cl-14	ethereal, sweet, rummy, fruity
	hexyl lactate	hexyl 2-hydroxypropanoate	20279-51-0	Cl-14	sweet, floral, green, fruity
	methyl 3-hydroxybutyrate	methyl 3-hydroxybutanoate	1487-49-6	Cl-14	fruity, green, apple, winey
bD-s	beta-damascenone	(E)-1-(2,6,6-trimethyl-1-cyclohexa-1,3-dienyl)but-2-en-1-one	23696-85-7	Cl-8	fruity, floral, apple, plum, tea, rose, tobacco, natural, grape, raspberry, sweet
	plum damascone (high alpha)	(E)-1-(2,4,4-trimethylcyclohex-2-en-1-yl)but-2-en-1-one	39872-57-6	Cl-8	fruity, floral, plum, rose, tobacco

	(Z)-alpha-damascone	(Z)-1-(2,6,6-trimethyl-1-cyclohex-2-enyl)but-2-en-1-one	23726-94-5	Cl-8	floral, rose, apple, fruity
	cyclohexylethyl isovalerate	1-cyclohexylethyl butanoate	Flavor-Base ; pas de CAS	Cl-8	sweet, fruity, apple
	cyclohexylethyl valerate	cyclohexylethyl-valerate	Flavor-Base ; pas de CAS	Cl-8	sweet, fruity, apple
	1-(3-(methyl thio)-butyryl)-2,6,6-trimethyl cyclohexene	3-(methylsulfanyl)-1-(2,6,6-trimethyl-1-cyclohexen-1-yl)-1-butanone	68697-67-6	Cl-8	sweet, fruity, tobacco
	3-cyclohexene-1-carboxylic acid, 2,6,6-trimethyl-, methyl ester	methyl 2,6,6-trimethylcyclohex-3-ene-1-carboxylate	815580-59-7	Cl-8	rose, fruity, floral
	beta-ionone	4-(2,6,6-trimethylcyclohexen-1-yl)but-3-en-2-one	14901-07-06	Cl-8	floral, woody, sweet, fruity, berry, tropical, violet, raspberry
	beta-ionyl acetate	4-(2,6,6-trimethyl-1-cyclohexenyl)but-3-en-2-yl acetate	22030-19-9	Cl-8	sweet, berry, raspberry, violet, woody
	alpha-ionol	4-(2,6,6-trimethyl-1-cyclohex-2-enyl)but-3-en-2-ol	25312-34-9	Cl-8	tropical, sweet, floral, violet, woody
	alpha-ionyl acetate	4-(2,6,6-trimethyl-1-cyclohex-2-enyl)but-3-en-2-yl acetate	52210-18-1	Cl-8	sweet, woody, floral, violet, berry
bI-s	3-methylcyclohexyl acetate	3-methylcyclohexyl-acetate	50539-20-3	Cl-8	fruity, sweet, berry, floral
	beta-irone	(E)-4-(2,5,6,6-tetramethyl-1-cyclohexenyl)but-3-en-2-one	79-70-9	Cl-8	woody, violet, fruity, raspberry
	campholene acetate	2-[(1S)-2,2,3-trimethylcyclopent-3-en-1-yl]ethyl acetate	36789-59-0	Cl-8	fruity, berry, woody, floral
	nopyl acetate	2-(6,6-dimethylbicyclo[3.1.1]hept-2-en-2-yl)ethyl acetate	128-51-8	Cl-8	sweet, woody, fruity, floral

	4-dimethyl ionone	(E)-4-methyl-1-(2,6,6-trimethylcyclohex-2-en-1-yl)pent-1-en-3-one	68459-99-4	Cl-8	sweet, floral, violet, woody
WL-s	whiskey lactone	5-butyl-4-methyloxolan-2-one	39212-23-2	Cl-4	tonka, coumarinic, coconut, toasted, nutty, celery, burnt, woody, lactonic
	7-mMethyltetrahydronaphthalenone	7-methyl-4,4a,5,6-tetrahydro-3H-naphthalen-2-one	34545-88-5	Cl-4	coumarinic, tonka, toasted, coconut, fruity, tobacco
	delta-heptalactone	6-ethyltetrahydropyran-2-one	3301-90-4	Cl-4	lactonic, coconut, coumarinic
	menthofurolactone	3,6-dimethyl-4,5,6,7-tetrahydro-3H-1-benzofuran-2-one	16642-41-4	Cl-4	sweet, lactonic, coumarinic, coconut
	octahydrocoumarin	3,4,4a,5,6,7,8,8a-octahydrochromen-2-one	4430-31-3	Cl-4	sweet, herbal, coumarinic, tonka, woody
	laitone	8-propan-2-yl-4-oxaspiro[4.5]decan-3-one	4625-90-5	Cl-4	woody, herbal, celery, lactonic
	coconut naphthalenone	7-methyl-4,4a,5,6-tetrahydro-3H-naphthalen-2-one	34545-88-5	Cl-4	toasted, coconut, tonka
	(R)-tonka furanone	(R)-3,6-dimethyl-5,6-dihydro-1-benzofuran-2(4H)-one	75640-26-5	Cl-4	lactonic, coumarinic, sweet, coconut
	(+/-)-dihydromint lactone	3,6-dimethyl-3a,4,5,6,7,7a-hexahydro-3H-benzofuran-2-one	92015-65-1	Cl-4	coumarinic, lactonic, tonka, phenolic, tobacco

Supplementary Table 2. Overlap of clusters.

	Overlaps between clusters	Nb of elements
clusters k-means in area Aa-3D at levels L9 and L16	3D-kmeans9-Cl-3 ∩ 3D-kmeans16-Cl-16	30
	3D-kmeans9-Cl-3 ∩ 3D-kmeans16-Cl-7	364
	3D-kmeans9-Cl-4 ∩ 3D-kmeans16-Cl-14	249
	3D-kmeans9-Cl-4 ∩ 3D-kmeans16-Cl-16	326
	3D-kmeans9-Cl-8 ∩ 3D-kmeans16-Cl-7	569
clusters kmeans16 and SOM16 in area Ab-3D at level L16	3D-kmeans16-Cl-8 ∩ 3D-SOM16-Cl-2	499
	3D-kmeans16-Cl-8 ∩ 3D-SOM16-Cl-5	400
	3D-kmeans16-Cl-8 ∩ 3D-SOM16-Cl-6	130
	3D-kmeans16-Cl-11 ∩ 3D-SOM16-Cl-1	430
clusters k-means16 and SOM16 in area Ac-3D at level L16	3D-kmeans16-Cl-1 ∩ 3D-SOM16-Cl-16	185
	3D-kmeans16-Cl-2 ∩ 3D-SOM16-Cl-11	248
	3D-kmeans16-Cl-3 ∩ 3D-SOM16-Cl-13	250
	3D-kmeans16-Cl-4 ∩ 3D-SOM16-Cl-13	379
	3D-kmeans16-Cl-4 ∩ 3D-SOM16-Cl-14	229
	3D-kmeans16-Cl-5 ∩ 3D-SOM16-Cl-11	4
	3D-kmeans16-Cl-5 ∩ 3D-SOM16-Cl-15	298
	3D-kmeans16-Cl-6 ∩ 3D-SOM16-Cl-14	146
	3D-kmeans16-Cl-9 ∩ 3D-SOM16-Cl-15	197
	3D-kmeans16-Cl-9 ∩ 3D-SOM16-Cl-16	69
	3D-kmeans16-Cl-10 ∩ 3D-SOM16-Cl-10	145
	3D-kmeans16-Cl-12 ∩ 3D-SOM16-Cl-14	269
	3D-kmeans16-Cl-12 ∩ 3D-SOM16-Cl-15	1
	3D-kmeans16-Cl-13 ∩ 3D-SOM16-Cl-14	42
	3D-kmeans16-Cl-13 ∩ 3D-SOM16-Cl-16	81

Supplementary Table 3. List of clusters k-means and SOM of areas Ac-3D at level L16.

Clusters in area Ac3D	Number of elements
3D-kmeans16-Cl-1	185
3D-kmeans16-Cl-2	248
3D-kmeans16-Cl-3	250
3D-kmeans16-Cl-4	608
3D-kmeans16-Cl-5	302
3D-kmeans16-Cl-6	146
3D-kmeans16-Cl-9	266
3D-kmeans16-Cl-10	145
3D-kmeans16-Cl-12	270
3D-kmeans16-Cl-13	123
3D-SOM16-Cl-10	145
3D-SOM16-Cl-11	252
3D-SOM16-Cl-13	629
3D-SOM16-Cl-14	686
3D-SOM16-Cl-15	496
3D-SOM16-Cl-16	335

Supplementary Table 6. Details of the PHASE-generated hypothesis from the subset composed of the mixture components.

Subset	Hypothesis	Phase Hypo Score	EF1%	BEDROC160.9	Ranked Actives
V-IA-F-EA-bI-bD-s	AHH_1	0.62	33.53	0.39	3
V-IA-F-s	AAR_1	0.78	66.87	0.50	2
	AAR_2	0.51	33.43	0.39	2
	AAH_1	0.47	33.43	0.39	2
WL-IA-s	AAH_1	1.21	100.20	1.00	2
	AAH_2	1.21	100.20	1.00	2
	AAH_3	1.20	100.20	1.00	2
	AHH_2	1.10	50.10	0.63	2
	AHH_5	1.01	50.10	0.56	2
	AAH_4	1.01	50.10	0.55	2
	AHH_1	0.90	50.10	0.54	2
	AHH_3	0.88	50.10	0.54	2
	AHH_4	0.87	50.10	0.54	2

EF1% = enrichment factor; BEDROC160.9 = Boltzmann-enhanced discrimination of receiver operating characteristic; A = hydrogen bond acceptor; H = hydrophobic; R = aromatic ring.

Supplementary Table 7. Distances between features of the hypotheses hyp-V-c, hyp-IA-c, hyp-F-c, hyp-EA-c, hyp-bD-c, hyp-bI-c, and hyp-WL-c.

Hypothesis	Feature 1	Feature 2	Distance (Å)
hyp-V-c	A1	A2	5.267
	A1	H5	1.424
	A3	A1	2.742
	A3	A2	6.452
	A3	H5	4.165
	A3	R6	2.767
	H5	A2	5.167
	R6	A1	2.796
	R6	A2	3.735
	R6	H5	3.734
hyp-IA-c	A1	H3	4.287
	A1	H4	1.828
	A2	A1	2.303
	A2	H3	5.815
	A2	H4	3.558
	H4	H3	2.468
hyp-F-c	A2	A1	8.621
	A2	R4	2.758
	R4	A1	5.967
hyp-EA-c	A2	A1	2.303
	H3	A1	1.826
	H3	A2	3.554
hyp-bD-c	A1	H2	3.353
	A1	H3	3.225
	A1	H4	3.853
	A1	H5	3.714
	H2	H3	2.584
	H2	H4	5.085
	H2	H5	2.886
	H3	H4	4.698
	H3	H5	2.501
	H4	H5	2.974
	H6	A1	3.616
	H6	H2	5.267
	H6	H3	6.397
	H6	H4	4.576
	H6	H5	5.662
hyp-bI-c	A1	H2	5.082

	A1	H3	5.464
	A1	H4	5.112
	A1	H5	5.803
	H2	H3	2.576
	H2	H4	4.980
	H2	H5	2.846
	H3	H4	4.788
	H3	H5	2.574
	H5	H4	2.964
hyp-WL-c	A1	H4	4.273
	A1	H5	1.781
	A2	A1	2.255
	A2	H4	6.392
	A2	H5	3.730
	H3	A1	3.686
	H3	A2	4.888
	H3	H4	5.864
	H3	H5	2.361
	H5	H4	3.869

Supplementary Table 8. Details of the PHASE top ten hypotheses generated from V-s, IA-s, F-s, EA-s, bD-s, bI -s and W-s subsets.

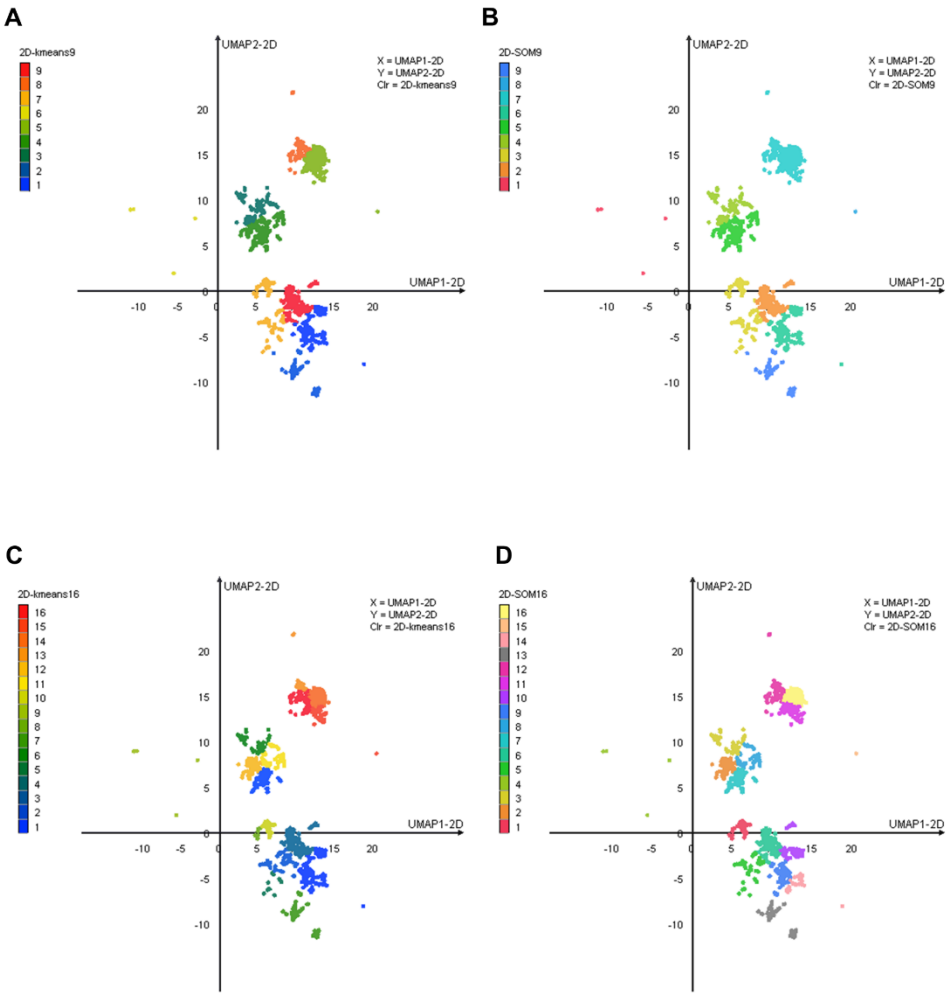
Subset	Hypothesis	Phase Hypo Score	EF1%	BEDROC160.9	Ranked Actives
V-s	AHR_3	1.14	84.33	0.97	12
	AHR_1	1.11	75.90	0.91	12
	AHR_2	1.03	75.90	0.89	11
	AAR_2	1.03	75.90	0.89	10
	AAR_1	1.03	84.33	0.97	11
	AAR_3	1.02	75.90	0.89	10
	AAHR_1	0.91	75.90	0.89	9
	AAHR_3	0.90	75.90	0.89	10
	AAAR_2	0.90	75.90	0.89	9
	AAAR_1	0.90	75.90	0.89	9
IA-s	AAH_1	1.26	91.91	0.99	11
	AAH_3	1.25	91.91	0.98	11
	AAH_2	1.17	91.91	0.96	10
F-s	AAR_1	0.93	70.70	0.84	7
	AAR_2	0.92	70.70	0.82	7
	AHR_2	0.74	50.50	0.59	7
	AAH_1	0.72	50.50	0.61	7
	AAHR_1	0.72	50.50	0.67	5
	AHR_5	0.71	50.50	0.67	6
	AHR_3	0.71	50.50	0.69	7
	AHR_1	0.69	40.40	0.56	7
	AAHR_4	0.69	40.40	0.56	6
	AAHR_2	0.69	40.40	0.61	6
EA-s	AAH_1	1.16	78.48	0.92	8
	AAH_2	1.13	78.48	0.89	8
bD-s	HHH_1	1.10	71.93	0.76	7
	AHH_2	1.07	71.93	0.79	7
	AHH_3	1.06	71.93	0.82	7
	AHHH_1	1.06	71.93	0.77	7
	AHHH_5	1.04	86.31	0.82	7
	AHH_1	0.96	71.93	0.82	7
	AHHH_2	0.93	57.54	0.74	7
	HHH_2	0.92	71.93	0.82	7
	AHHH_3	0.92	57.54	0.73	7
	AHHH_4	0.91	57.54	0.69	7
bI-s	AHH_3	1.20	89.69	0.95	9
	HHH_2	1.14	89.69	0.95	8
	HHH_1	1.14	89.69	0.94	8

Subset	Hypothesis	Phase Hypo Score	EF1%	BEDROC160.9	Ranked Actives
	AHH_2	1.14	89.69	0.94	9
	AHH_1	1.14	89.69	0.95	8
	AHHH_9	1.13	78.48	0.89	8
	HHH_3	1.08	78.48	0.88	8
	AHH_4	1.07	67.27	0.82	9
	AHH_6	0.99	67.27	0.81	9
	AHH_5	0.98	67.27	0.81	9
WL-s	AAH_1	1.08	88.20	0.85	7
	AHH_1	1.02	63.00	0.69	8
	AHH_2	0.99	75.60	0.83	7
	AHH_3	0.94	50.40	0.70	8
	AAHH_1	0.89	63.00	0.75	6
	AAH_2	0.80	37.80	0.54	6
	AHH_4	0.78	50.40	0.64	8
	AAH_3	0.62	25.20	0.40	7
	HHH_1	0.61	25.20	0.32	5
	AHH_5	0.52	25.20	0.32	8

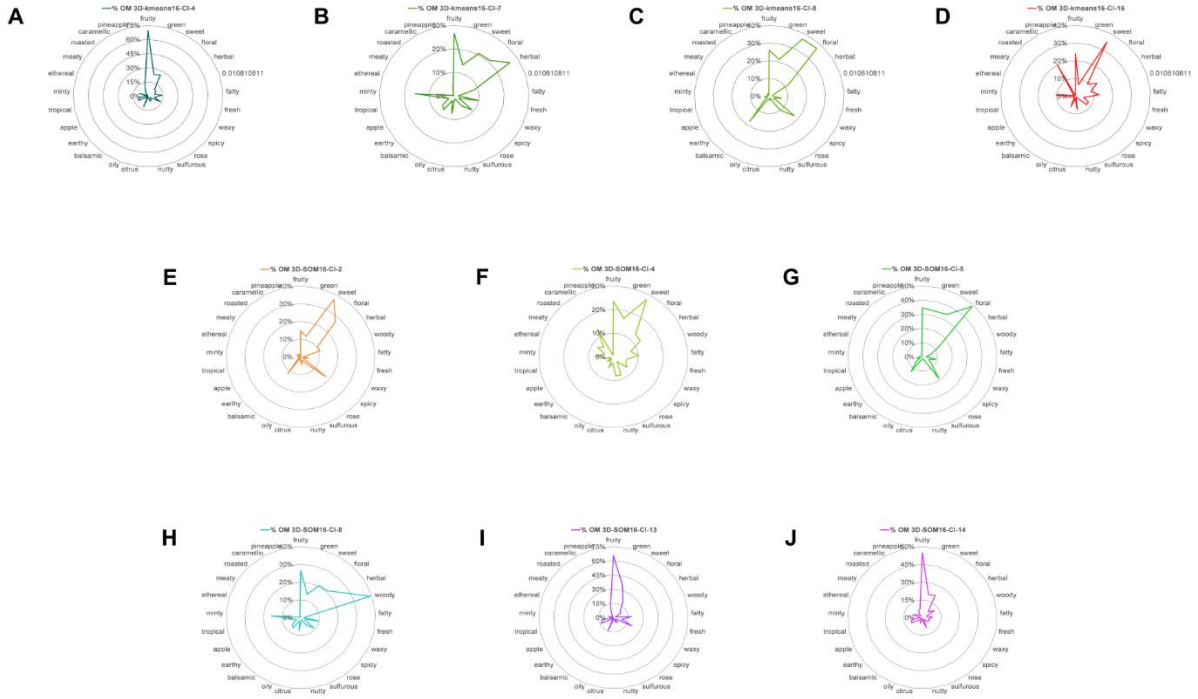
EF1% = Enrichment factor; BEDROC160.9 = Boltzmann-enhanced discrimination of receiver operating characteristic; A = hydrogen bond acceptor; H = hydrophobic; R = Aromatic ring.

Supplementary Table 9. Distances between features of the hypotheses generated from V-s, IA-s, F-s, EA-s, bD-s, bI-s, and WL-s.

Subset	Feature 1	Feature 2	Distance (Å)
Hyp-V-s	A3	R6	2.770
	H5	A3	4.146
	R6	H5	3.579
hyp-IA-s	A1	H3	2.406
	A2	A1	2.274
	H3	A2	3.707
hyp-F-s	A1	R4	3.594
	A2	A1	2.274
	R4	A2	5.367
hyp-EA-s	A1	H3	2.791
	A2	A1	2.274
	H3	A2	3.900
hyp-bD-s	H4	H6	2.835
	H6	H7	5.102
	H7	H4	4.368
hyp-bI-s	A1	H5	6.169
	H4	A1	4.125
	H4	H5	2.926
hyp-WL-s	A1	A2	2.309
	A2	H5	4.557
	H5	A1	2.767



Supplementary Figure 1. Visualizations of the compounds-odors dataset in the 2D space obtained using k-means and SOM clustering. (A) k-means clustering at level L9; (B) SOM clustering at level L9; (C) k-means clustering at level L16; (D) SOM clustering at level L16.



Supplementary Figure 2. Radar charts of the distribution of the molecules carrying the 25 most frequent odor notes across clusters 3D-SOM16 that contain the molecules of interest. (A) 3D-kmeans16-Cl-4, (B) 3D-kmeans16-Cl-7, (C) 3D-kmeans16-Cl-8, (D) 3D-kmeans16-Cl-16, (E) 3D-SOM16-Cl-2, (F 3D-SOM16-Cl-4, (G) 3D-SOM16-Cl-5, (H) 3D-SOM16-Cl-8, (I) 3D-SOM16-Cl-13, (J) 3D-SOM16-Cl-14.

III. Informations supplémentaires de la partie 3-III

Tableau S1. Liste totale des interactions extraites entre molécules odorantes et récepteurs olfactifs.

(Fichier Tableau-S1-partie-3-III.xlsx)

Tableau S2. Liste des voies de signalisation identifiées par l'enrichissement biologique.

(Fichier Tableau-S2-partie-3-III.xlsx)

Tableau S3. Liste des pathologies impliquant les protéines humaines de l'odorome.

Tableau S3. Liste des pathologies impliquant les protéines humaines de l'odorome

Protein	Disease	Score DisGeNET
HRAS	Costello syndrome (disorder)	1.000
GNAS	Pseudopseudohypoparathyroidism	1.000
PPIB	Osteogenesis imperfecta, type ix (disorder)	0.930
REEP1	Spastic paraplegia 31, autosomal dominant	0.920
AMACR	Alpha-Methylacyl-CoA Racemase Deficiency	0.910
APP	Alzheimer's Disease	0.900
LDLR	Hypercholesterolemia, Familial	0.900
LDLR	Hypercholesterolemia	0.900
BAG3	Cardiomyopathy, dilated, 1hh	0.900
SLC12A6	Corpus callosum agenesis neuronopathy	0.900
REEP6	Retinitis pigmentosa 77	0.900
LDLR	Hyperlipoproteinemia Type IIa	0.800
GNAS	Pseudohypoparathyroidism, Type Ia	0.800
GNAS	Albright's hereditary osteodystrophy	0.800
GNAS	McCune-Albright Syndrome	0.800
GNAS	Pseudohypoparathyroidism, type Ib	0.800
ALB	Hyperthyroxinemia, Familial Dysalbuminemic	0.800
HRAS	melanoma	0.800
GNAS	Osteoma cutis	0.800
BAG3	Myofibrillar Myopathy	0.800
DSG4	Hypotrichosis 6	0.800
CDAN1	Congenital dyserythropoietic anemia, type I	0.780
ACVRL1	Osler-rendu-weber syndrome 2	0.760
GNAL	Dystonia 25	0.730
GNAS	Pseudohypoparathyroidism Type 1C	0.730
HRAS	Nevus Sebaceus of Jadassohn	0.720
HRAS	Organoid Nevus Phakomatosis	0.710
EGFR	Non-Small Cell Lung Carcinoma	0.700
EGFR	Adenocarcinoma of lung (disorder)	0.700
EGFR	Glioblastoma	0.700
SERPINA1	alpha 1-Antitrypsin Deficiency	0.700
LDLR	Atherosclerosis	0.700
APP	Dementia	0.700
GNAS	Pseudohypoparathyroidism	0.700
EGFR	Squamous cell carcinoma of esophagus	0.700
BLK	Lupus Erythematosus, Systemic	0.700
HRAS	Carcinoma of bladder	0.700
BAG3	Myopathy, Myofibrillar, Bag3-Related	0.700
HRAS	Malignant neoplasm of urinary bladder	0.700
GNAS	Pseudohypoparathyroidism Type 1B	0.700

OTX2	Microphthalmia, Syndromic 5	0.700
APP	Cerebral amyloid angiopathy, app-related	0.700
GNB1	Mental retardation, autosomal dominant 42	0.700
UBQLN2	Amyotrophic lateral sclerosis 15, with or without frontotemporal dementia	0.700
CIT	Microcephaly 17, primary, autosomal recessive	0.700
MPDZ	Hydrocephalus, congenital, 2, with or without brain or eye anomalies	0.700
AMACR	Bile acid synthesis defect, congenital, 4	0.700
CDSN	Hypotrichosis Simplex of Scalp	0.700
EGFR	Inflammatory skin and bowel disease, neonatal, 2	0.700
GNAS	Acth-Independent Macronodular Adrenal Hyperplasia	0.700
REEP2	Spastic paraplegia 72, autosomal recessive	0.700
GNAS	Polyostotic fibrous dysplasia	0.650
LDLR	Fatty Liver	0.640
HRAS	Noonan Syndrome	0.640
PIGO	Hyperphosphatasia with Mental Retardation	0.640
EGFR	Head and Neck Neoplasms	0.630
CDSN	Peeling skin syndrome	0.630
BLK	Maturity onset diabetes mellitus in young	0.620
EGFR	Malignant neoplasm of lung	0.600
APP	Impaired cognition	0.600
EGFR	Lung Neoplasms	0.600
ESR2	Malignant neoplasm of breast	0.600
ALB	Diabetic Nephropathy	0.600
EGFR	Malignant Head and Neck Neoplasm	0.600
ACVRL1	Hereditary hemorrhagic telangiectasia	0.600
SERPINA1	Lung diseases	0.600
EGFR	Malignant neoplasm of esophagus	0.600
HRAS	Breast Carcinoma	0.600
HRAS	Squamous cell carcinoma	0.600
EGFR	Esophageal Neoplasms	0.600
HRAS	Malignant neoplasm of breast	0.600
HRAS	Bladder Neoplasm	0.600
LETM1	Wolf-Hirschhorn Syndrome	0.600
APP	Cerebral hemorrhage with amyloidosis, hereditary, Dutch type	0.600
HRAS	Neuroblastoma	0.600
PIGO	Hyperphosphatasia with mental retardation syndrome 2	0.600
OTX2	Pituitary hormone deficiency, combined, 6	0.600
B4GALT1	Congenital disorder of glycosylation type 2D	0.600
ALX1	Frontonasal dysplasia 3	0.600
ZP3	Oocyte maturation defect 3	0.600
KCTD17	Dystonia 26, myoclonic	0.600
BLK	Maturity-onset diabetes of the young, type 11	0.600

HRAS	Cardio-facio-cutaneous syndrome	0.570
ACVRL1	Familial primary pulmonary hypertension	0.560
HRAS	Thyroid carcinoma	0.560
HRAS	Phacomatosis pigmentokeratotica	0.540
HRAS	Carcinoma, Transitional Cell	0.530
REEP6	Retinitis Pigmentosa	0.530
ALX1	Frontonasal dysplasia	0.530
HRAS	Glioma	0.520
ACVRL1	Pulmonary Hypertension	0.510
SERPINA1	Bipolar Disorder	0.510
GNAS	Schizophrenia	0.510
REEP1	Neuronopathy, distal hereditary motor, type v	0.510
KCTD17	Myoclonic dystonia	0.510
EGFR	Carcinoma of lung	0.500
EGFR	Glioblastoma Multiforme	0.500
EGFR	Squamous cell carcinoma	0.500
EGFR	Squamous cell carcinoma of the head and neck	0.500
EGFR	Glioma	0.500
APP	Memory impairment	0.500
SERPINA1	Pulmonary Emphysema	0.500
APP	Alzheimer Disease, Early Onset	0.500
SERPINA1	Chronic Obstructive Airway Disease	0.500
LDLR	Coronary heart disease	0.500
LDLR	Coronary Artery Disease	0.500
ALB	Hypoalbuminemia	0.500
EGFR	Brain Neoplasms	0.500
EGFR	Squamous cell carcinoma of lung	0.500
UBQLN2	Amyotrophic Lateral Sclerosis	0.500
APP	Cerebral Amyloid Angiopathy	0.500
EGFR	Esophageal carcinoma	0.500
EGFR	Colorectal Neoplasms	0.500
CD28	Rheumatoid Arthritis	0.500
LDLR	Hyperlipidemia	0.500
ACVRL1	Idiopathic pulmonary arterial hypertension	0.500
SERPINA1	Liver carcinoma	0.500
EGFR	Prostatic Neoplasms	0.500
HRAS	Papilloma	0.500
HRAS	Liver carcinoma	0.500
APP	Seizures	0.500
HRAS	Adenocarcinoma of lung (disorder)	0.500
HRAS	Leukemia, Myelocytic, Acute	0.500
ACVRL1	Pulmonary arterial hypertension	0.500
LETM1	Seizures	0.500
EGFR	Giant Cell Glioblastoma	0.500

HRAS	Mammary Neoplasms, Experimental	0.500
EGFR	Cardiomyopathy, Dilated	0.500
EGFR	Acute kidney injury	0.500
ESR2	Mammary Neoplasms, Experimental	0.500
HRAS	Mouth Neoplasms	0.500
PRKCA	Cardiomegaly	0.500
HRAS	Malignant Glioma	0.500