ÉCOLE POLYTECHNIQUE DE L'UNIVERSITÉ DE NANTES
DÉPARTEMENT D'INFORMATIQUE

RAPPORT DE RECHERCHE ET DÉVELOPPEMENT

Fouille de données olfactives : clustering de molécule odorantes par GCL (Graph Constrastive Learning)

Projet de Recherche et Développement

Colin TRÈVE & Marwa TABIB

23 février 2025

encadré par Fabrice GUILLET & Angélique VILLIÈRE

— Équipe DUKE —

LABORATOIRE DES SCIENCES DU NUMÉRIQUES DE NANTES
LABORATOIRE GÉNIE DES PROCÉDÉS - ENVIRONNEMENT - AGRO-ALIMENTAIRE

coordinateur : Philippe LERAY

Fouille de données olfactives : clustering de molécule odorantes par GCL (Graph Constrastive Learning)

Projet de Recherche et Développement

Colin TRÈVE & Marwa TABIB

Résumé

Ce rapport examine l'application des réseaux de neurones graphiques (Graph Neural Networks - GNN) pour explorer la relation structure-odeur des molécules. La problématique principale réside dans l'évaluation de la capacité de modèles tels que les Graph Convolution Networks (GCN) et les Message Passing Neural Networks (MPNN) à prédire les propriétés olfactives des molécules à partir de leur structure.
Les objectifs initiaux étaient d'étudier les fondements théoriques des GNN, d'explorer les bases de données moléculaires fournies dans la littérature scientifique, et de reproduire les implémentations décrites pour valider les modèles. Les jeux de données étant déjà préparés dans les articles consultés, notre tâche se concentrait principalement sur la compréhension et la mise en œuvre des modèles.
Cependant, à ce stade du projet, l'implémentation des modèles GCN et MPNN reste en cours. Des difficultés techniques et organisationnelles ont limité notre progression dans cette tâche, notamment en raison de contraintes de coordination et d'un manque d'expérience préalable avec certains outils. Ces obstacles nous ont également empêchés de finaliser l'intégration avec OpenPOM.

Malgré ces limites, nous envisageons de poursuivre ce travail en explorant une approche contrastive pour mieux analyser la contribution des différentes parties d'une molécule à son odeur. En particulier, nous proposons d'utiliser une méthode de Graph Contrastive Learning (GCL) pour distinguer les zones de la molécule influençant l'odeur de celles ayant un impact négligeable. Cette méthode nécessitera des essais avec plusieurs algorithmes GCL pour une comparaison fiable, et les résultats obtenus seront évalués par rapport aux performances des modèles classiques GCN et MPNN.

Remerciements

Nous exprimons notre reconnaissance à M.Fabrice Guillet et Mme.Angélique Villère pour leur accompagnement tout au long de ce projet. Leurs conseils et leur disponibilité constante ont été déterminants dans notre progression. Les réunions hebdomadaires qu'ils ont organisées ont constitué des moments privilégiés de suivi et de réflexion collective, ce qui nous a aidé a mieux progresser dans ce projet.

Table des matières

1

Introduction

1.1 Présentation de la problématique

La perception olfactive représente un défi scientifique complexe où comprendre la relation entre la structure moléculaire et l'odeur nécessite une approche interdisciplinaire innovante. Comment les caractéristiques physiques d'une molécule déterminent-elles sa signature olfactive ? Les réseaux de neurones sur graphes (GNN) offrent une perspective prometteuse pour décoder ces mécanismes subtils, en cherchant à identifier les motifs structurels responsables des propriétés odorantes. Cette problématique implique de naviguer à travers la non-linéarité de la perception chimique, où l'interaction entre molécules peut générer des odeurs radicalement différentes des composés originaux, tout en développant des modèles capables de capturer cette complexité intrinsèque.

1.2 Objectifs poursuivis

Au commencement de notre projet, nous nous sommes fixé plusieurs objectifs précis :

- Comprendre et implémenter les modèles de réseaux de neurones graphiques (Graph Neural Networks - GNN), notamment les Graph Convolution Networks (GCN) et Message Passing Neural Networks (MPNN), pour analyser la relation structure-odeur.
- Explorer et manipuler les bases de données moléculaires GoodScents et Leffingwell PMP 2001, en vue de préparer un jeu de données pertinent pour notre analyse.
– Établir les points d'amélioration des modèles actuels.
- Développer un modèle capable de prédire les descripteurs olfactifs à partir de la structure moléculaire, avec un objectif de performance élevé (Proposition : GCL).

Nous évaluerons dans la suite de notre rapport dans quelle mesure ces objectifs initiaux ont été atteints, en portant un regard critique et constructif sur notre démarche et nos réalisations.

1.3 Travail réalisé

À ce stade de notre projet, nous avons accompli plusieurs étapes cruciales qui posent les fondements nécessaires à la réalisation de nos objectifs. Nous avons notamment exploré deux modèles principaux de réseaux de neurones graphiques (Graph Neural Networks - GNN) : les Graph Convolution Networks (GCN) et les Message Passing Neural Networks (MPNN). Ces deux modèles qui sert à la prédiction des descripteurs d'odeurs d'une molécule. Cette réalisation constitue une étape essentielle dans notre compréhension et notre capacité à modéliser les interactions moléculaires complexes liées à la perception olfactive.

1.4 Contribution

Ce projet vise à analyser et comparer deux modèles principaux d'apprentissage automatique (GCN et MPNN) appliqués à la prédiction des propriétés olfactives à partir de la structure moléculaire. Nos contributions principales sont les suivantes :

- Une synthèse approfondie des méthodes et modèles présentés dans la littérature, permettant une meilleure compréhension des principes sous-jacents des modèles GCN et MPNN dans le domaine de la prédiction olfactive.
- Une analyse comparative des résultats théoriques et pratiques des modèles.
- Une tentative d'implémentation des modèles sur la plateforme OpenPOM. Bien que cette phase n'ait pas abouti à un résultat fonctionnel en raison de contraintes techniques et de limitations dans les bibliothèques disponibles, cette étape a permis d'identifier les principaux obstacles liés à l'utilisation de cette plateforme pour ce type de projet.

1.5 Plan de l'étude

Le chapitre 2 constitue notre état de l'art. Nous y effectuerons un panorama exhaustif des propositions issues de la littérature scientifique concernant les modèles de réseaux de neurones graphiques. L'analyse conjointe de ces travaux nous permettra de dresser un bilan critique des approches existantes et d'identifier les pistes de recherche les plus prometteuses.
Dans le chapitre 3, nous approfondirons d'un point de vue théorique les approches les plus pertinentes identifiées précédemment.
La conclusion viendra clôturer notre travail en synthétisant nos principales contributions, en soulignant les limites rencontrées et en ouvrant des perspectives pour de futures investigations scientifiques.

2

État de l'art

La prédiction de l'odeur des molécules est un problème complexe auquel il n'est pas facile de répondre en raison de leur structure organisée sous forme de graphe. Par le passé, de nombreuses méthodes et de nombreux modèles ont été développés (forêts aléatoires, empreintes de Morgan). Actuellement, la méthode dominante pour résoudre ce problème consiste à utiliser la structure en graphe des molécules à travers des GNN (Graph Neural Networks). Parmi les articles, certains tentent même de réaliser des prédictions sur des paires de molécules[LS24].

2.1 Contexte et Problématique

2.1.1 Problème de QSOR

La prédiction des propriétés olfactives à partir de la structure moléculaire est un problème complexe et ancien, qui relève de la chimie, de la neuroscience et de l'apprentissage automatique. Contrairement à d'autres propriétés moléculaires (comme la solubilité ou la toxicité), les propriétés olfactives sont difficiles à prédire en raison de la complexité de la perception humaine et de la variabilité des récepteurs olfactifs. Les modèles traditionnels de QSOR reposent souvent sur des caractéristiques chimiques prédéfinies (comme les empreintes moléculaires ou les descripteurs Dragon/Mordred), mais ces approches ont des limites en termes de précision et de généralisation.

2.1.2 Objectif

Les trois articles visent à améliorer la prédiction des descripteurs olfactifs en utilisant des modèles de machine learning plus sophistiqués, en particulier les **réseaux de neurones sur graphes (GNN)**, qui permettent de capturer des relations structurelles complexes entre les atomes et les liaisons dans une molécule.

2.2 Approches Traditionnelles et Limites

2.2.1 Approches Classiques

— **Empreintes Moléculaires** : Les empreintes moléculaires (comme les empreintes de Morgan) sont couramment utilisées pour représenter les molécules sous forme de vecteurs fixes. Ces empreintes capturent des informations sur les environnements topologiques des atomes, mais elles sont limitées par leur nature statique et prédéfinie.

— **Random Forests (RF)** : Les forêts aléatoires sont souvent utilisées pour prédire les propriétés moléculaires à partir de ces empreintes. Bien qu'elles soient performantes, elles ne capturent pas toujours les relations complexes entre la structure et l'odeur.

— **Modèles Basés sur des Règles Empiriques** : Historiquement, des règles empiriques (comme la règle de Boelens pour les odeurs de rose) ont été utilisées pour prédire les odeurs, mais ces règles sont souvent trop simplistes et ne généralisent pas bien.

2.2.2 Limites

Les modèles traditionnels ne capturent pas bien les relations non linéaires entre la structure moléculaire et la perception olfactive. Ils sont souvent limités par la qualité et la quantité des données disponibles, ainsi que par la complexité des interactions entre les récepteurs olfactifs et les molécules.

2.3 Apport des Réseaux de Neurones sur Graphes (GNN)

2.3.1 Graph Neural Networks (GNN)

Les GNN 2.1 sont des modèles de machine learning conçus pour traiter des données structurées en graphes, comme les molécules (où les atomes sont des nœuds et les liaisons sont des arêtes). Ils permettent de capturer des informations locales et globales sur la structure moléculaire, ce qui les rend particulièrement adaptés pour la prédiction des propriétés olfactives.

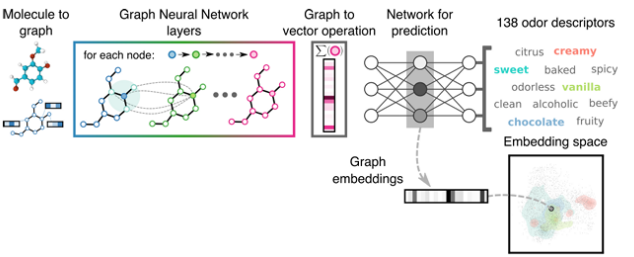

FIGURE 2.1 – Modèle du GNN

2.3.2 Avantages des GNN

— **Représentations Apprises** : Contrairement aux empreintes moléculaires prédéfinies, les GNN apprennent des représentations vectorielles (embeddings) des molécules directement à partir des données, ce qui permet de capturer des relations complexes entre la structure et l'odeur.

— **Généralisation** : Les embeddings appris par les GNN peuvent être utilisés pour des tâches de transfert learning, comme la prédiction de nouveaux descripteurs olfactifs ou l'application à d'autres ensembles de données.
— **Performance** : Les GNN surpassent les méthodes traditionnelles (comme les Random Forests) en termes de précision et de robustesse, en particulier pour les descripteurs olfactifs rares ou complexes.

2.4 Résultats et Contributions

2.4.1 Performance des Modèles

Dans les deux articles, les GNN atteignent des performances supérieures aux modèles de référence (comme les Random Forests) en termes de **AUROC** (Area Under the Receiver Operating Characteristic curve) et d'autres métriques de classification. Par exemple, dans l'article *Machine Learning for Scent*, le GNN atteint une AUROC moyenne de **0.894**, contre **0.850** pour le meilleur modèle de référence (Random Forest sur des empreintes de Morgan).

2.4.2 Carte Principale des Odeurs (Principal Odor Map, POM)

L'article [LMSL+22] introduit une **Carte Principale des Odeurs (POM)**, qui est une représentation vectorielle de l'espace olfactif apprise par le GNN. Cette carte capture les relations perceptives entre les odeurs et permet de visualiser les clusters de molécules partageant des descripteurs olfactifs similaires. La POM montre une meilleure corrélation avec les distances perceptives que les méthodes traditionnelles basées sur des empreintes chimiques.

2.4.3 Transfer Learning et Généralisation

Les embeddings appris par les GNN sont utilisés pour des tâches de transfert learning, comme la prédiction de nouveaux descripteurs olfactifs ou l'application à d'autres ensembles de données (comme le **DREAM Olfaction Challenge**). Les résultats montrent que les GNN généralisent bien à des tâches olfactives connexes, même avec des données limitées.

2.5 Une proposition

2.5.1 Présentation

Un article concernant la prédiction des odeurs utilisant uniquement la structure de graphe est apparu[BSLW19]. Cette approche néglige complètement l'aspect spatial de la molécule et se concentre sur les atomes et leurs liaisons pour tenter de déterminer l'odeur de celle-ci. Cependant, cette approche, basée sur les Graph Neural Networks (GNNs), est particulièrement adaptée au traitement des données structurées moléculaires. Leur architecture repose sur un mécanisme sophistiqué d'échange de messages entre les nœuds du graphe. Ce processus permet de capturer progressivement des informations de plus en

plus complexes sur les interactions atomiques.

L'architecture typique comprend plusieurs couches d'agrégation 2.2 :

— **Message Passing et Agrégation de message : Sortie**

$$m_v^t$$

— Chaque nœud récupère les informations de ses voisins via une fonction d'agrégation.
— Cette étape permet d'échanger des informations entre les nœuds connectés.

— **Mise à jour : Sortie**

$$h_v^T$$

— Une couche finale (souvent un réseau de neurones) est utilisé avec une fonction d'activation pour mettre à jour le nœud final.

— **Readout : Sortie**

$$y$$

— Une fonction d'agrégation combine les représentations des nœuds pour produire une représentation globale du graphe.
— Cela permet d'obtenir une vue d'ensemble des relations dans le graphe.

Le résultat du GNN sera ensuite injecté dans un perceptron multicouche pour la prédiction des odeurs. Sortie :

$$z_g$$

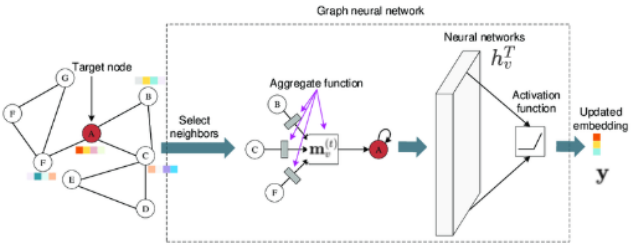

FIGURE 2.2 – Modèle du GNN

Il existe deux variantes principales du GNN, le Message Passing Neural Networks (MPNN) et le Graph Convolution Networks (GCN). Ces modèles partagent une structure commune, mais le GCN se démarque par sa simplicité architecturale.

Suite à ces travaux prometteurs sur l'utilisation de la structure en graphe des molécules et des GNN, une idée a émergé : il serait possible de prédire l'odeur d'un mélange de molécules à partir de leur structure en graphe.

Les données utilisées proviennent du site The Good Scent, via sa fonction blender. Ces données servent à entraîner un Graph Isomorphism Network (GIN), qui prend en entrée le graphe de la molécule et sa vectorisation réalisée à l'aide de la bibliothèque Python OGB (Open Graph Benchmark). La molécule traverse trois couches de message passing, avec une fonction de readout basée sur la concaténation des embeddings des molécules. Ces embeddings sont ensuite traités par un réseau de neurones multicouche (MLP) utilisant comme fonction de

coût l'entropie croisée binaire.

Pour comparer les performances, un autre modèle est utilisé comme référence : un MPNN (Message Passing Neural Network) implémenté dans l'article [BSLW19]. Ce modèle prend en entrée une paire de molécules, représentée comme un graphe avec des composantes disjointes. Après les couches de message passing, la phase de readout utilise une méthode Set2Set avec trois couches et trois étapes.

2.5.2 Analyse

Une étude comparative approfondie dans l'article [BSLW19] a évalué différentes approches d'encodage et de prédiction. Les résultats démontrent la supériorité des GNNs, avec des scores significativement meilleurs en termes d'AUROC, de F1-score et de précision 2.3.

	AUROC	Precision	F1
GNN	**0.894 [0.888, 0.902]**	**0.379 [0.351, 0.398]**	**0.360 [0.337, 0.372]**
RF-Mordred	0.850 [0.838, 0.860]	0.311 [0.288, 0.333]	0.306 [0.283, 0.319]
RF-bFP	0.832 [0.821, 0.842]	0.321 [0.293, 0.339]	0.295 [0.272, 0.308]
RF-cFP	0.845 [0.835, 0.854]	0.315 [0.280, 0.332]	0.295 [0.272, 0.311]
KNN-bFP	0.791 [0.778, 0.803]	0.328 [0.305, 0.347]	0.323 [0.299, 0.335]
KNN-cFP	0.796 [0.785, 0.809]	0.333 [0.307, 0.351]	0.316 [0.292, 0.327]

FIGURE 2.3 – Résultats des performances des GNN

Le modèle le plus performant parmi ceux testés est un GNN, qui atteint un AUROC de 0.894.

Parmi les modèles du GNN, il existe le MPNN et le GCN 2.4. D'après l'article [BSLW19], les deux modèles suivent la même structure de base : une couches de message passing -> une opération de réduction par somme (reduce-sum) qui combine les informations reçues des voisins en une seule valeur par nœud -> des couches entièrement connectées (fully connected layers) qui traitent ensuite ces informations pour produire la sortie finale du modèle. Comme c'est illustré dans la figure2.5.

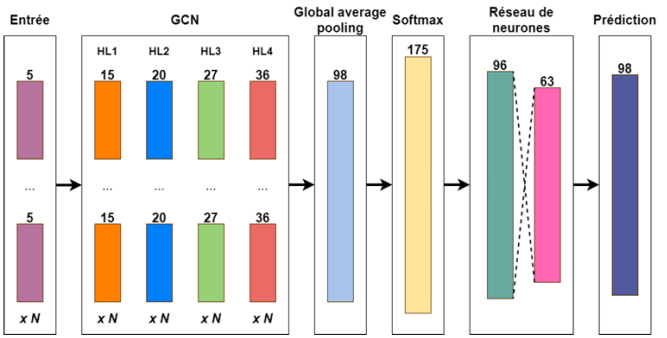

FIGURE 2.4 – Architecture du GCN - rapport Pred 2023[CJ23]

	GCN	MPNN
Message Passing Layers	concatenation message type, 4 layers of dim: [15,20,27,36], selu activation, max graph pooling	edge-conditioned matrix multiply message type, 5 layers of dim 43, GRU-update at each layer
Readout	Global sum pooling with softmax, 175 dim, one per MP layer and summed	Global sum pooling with softmax, 197 dim, one per MP layer with residual connections and summed
fully-connected neural net	2-layers of dim [96, 63] with relu, batchnorm, dropout of 0.47	3-layers of dim 392 with relu, batchnorm, dropout of 0.12 and l1/l2 regularization
Prediction	Multi-headed sigmoid, 138 tasks	
Training	Weighted-cross entropy loss, optimized with Adam, used learning rate decay with warm restarts, 300 epochs	

FIGURE 2.5 – Comparaison entre MPNN et GCN

Dans notre étude, nous avons choisi de nous concentrer sur le modèle MPNN en raison du manque d'informations précises sur le fonctionnement du GCN. De plus, le MPNN repose également sur des principes de convolution. Lors de nos recherches, nous avons trouvé qu'il existe deux versions qui ont été implémentées du MPNN et ceux sur deux années différentes. En 2022, l'article [LMSL+22] a été publié expliquant le fonctionnement de MPNN que nous avons pu traduire en 2.6.

En 2023, le git [BKSS23] a été créé pour mettre en oeuvre le modèle MPNN qui a été configuré comme dans la figure 2.7

Explication détaillée

I- Featurisation du Graphe (graph_featurizer.py) : Cette étape convertit une molécule en un graphe moléculaire avec des caractéristiques spécifiques pour les atomes et les liaisons.

Détails des caractéristiques

— Atomes → Représentés par un vecteur de **134 dimensions** (concaténation de plusieurs propriétés chimiques et structurelles).

— Liaisons (bonds) → Représentées par un vecteur de **6 dimensions**.

— Matrice d'adjacence (edge index) → Liste des connexions entre les atomes.

Dimensions des Tensors

— node_features : $(\text{num\_atoms}, 134)$

— edge_features : $(\text{num\_bonds}, 6)$

— edge_index : $(2, \text{num\_bonds} \times 2)$ (connectivité du graphe)

II- Initialisation du MPNN (mpnn_pom.py)

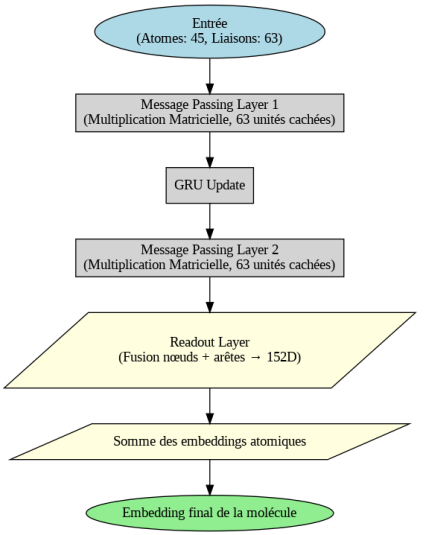

FIGURE 2.6 – étapes de messages passing dans l'article [LMSL+22]

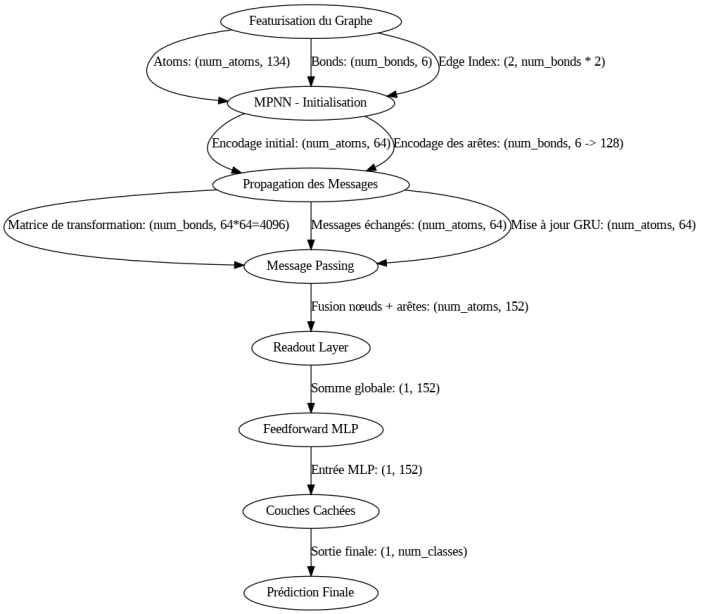

FIGURE 2.7 – étapes de messages passing dans le git [BKSS23]

Le MPNN est construit et initialisé avec des encodages pour les nœuds et les arêtes.

Composants Principaux

— Encodage initial des nœuds : node_in_feats $= 134$
— Encodage des arêtes : edge_in_feats $= 6$
— Propagation des messages avec num_step_message_passing $= 5$
— Mise à jour des nœuds avec un **GRU**
— Agrégation globale pour obtenir un embedding de la molécule

Dimensions des Tensors

— Entrée des nœuds : $(\text{num\_atoms}, 134)$
— Transformation initiale des nœuds : $(\text{num\_atoms}, 64)$
— Entrée des arêtes : $(\text{num\_bonds}, 6)$
— Messages échangés entre les nœuds : $(\text{num\_atoms}, 64)$
— Résultat final du MPNN : $(\text{num\_atoms}, 64)$

III- Propagation des Messages (pom_mpnn_gnn.py)

Chaque atome échange des informations avec ses voisins à travers les arêtes du graphe.

Processus Détaillé

— Transformation des caractéristiques des arêtes
  — edge_in_feats$(6) \rightarrow$ edge_hidden_feats(128)
  — edge_hidden_feats$(128) \rightarrow (\text{node_out_feats} \times \text{node_out_feats})(64 \times 64 = 4096)$
— Propagation des messages avec NNConv
— **Mise à jour des nœuds avec GRU** (mémoire des états précédents conservée)

Dimensions des Tensors

— Transformation des arêtes : $(\text{num\_bonds}, 128)$
— Matrice de transformation des messages : $(\text{num\_bonds}, 4096)$
— Messages échangés entre nœuds : $(\text{num\_atoms}, 64)$

IV- Readout Layer (mpnn_pom.py)

L'objectif de cette couche est de convertir le graphe en une représentation unique de la molécule.

Processus Détaillé

— Fusion des informations des nœuds et des arêtes → Embedding de 152 dimensions
— Somme de toutes les embeddings atomiques → Embedding moléculaire final

Dimensions des Tensors

— Embedding de chaque atome après fusion : $(\text{num\_atoms}, 152)$
— Embedding final de la molécule après agrégation : $(1, 152)$

V- Passage dans le Feedforward Network (model_configs.py)

L'embedding moléculaire est utilisé comme entrée pour un MLP (Multi-Layer Perceptron) permettant de réaliser la classification finale.

Processus Détaillé

— Transformation via un MLP (3 couches) avec activation ReLU et BatchNorm
— Sortie finale représentant la classification en plusieurs catégories d'odeurs

Dimensions des Tensors

— Entrée du MLP : $(1, 152)$
— Première couche cachée : $(1, 392)$
— Seconde couche cachée : $(1, 392)$
— Sortie finale : $(1, \text{num\_classes})$

Concernant les prédictions sur des paires de molécules, le GIN obtient un score d'AUROC de 0.77, légèrement supérieur à celui du MPNN, qui atteint 0.76. Cependant, lorsqu'on évalue la performance sur une seule molécule, le MPNN dépasse le GIN, et cet écart de performance devient plus marqué.

Vecteurs d'entrée

Chaque étude parmi celles qu'on a vu vectorise les atomes et les arêtes à sa manière.
Pour le GCN [gab23], les vecteurs d'entrée sont caractérisés comme suit 2.8 :

Caractéristique	Description	Dimension
Symbole	Type d'atome {'C': 0, 'O': 1, 'N': 2, 'S': 3, 'Cl': 4, 'Br': 5, 'H': 6}, encodé sous forme d'entier	1
Degré	nombre de voisins du sommet (tout atome confondu)	1
Valence Implicite	nombre de H absent du smile	1
Aromatique	Indique si l'atome appartient à un cycle aromatique (si oui: 1, 0 sinon)	1
Chiralité	Type de chiralité de l'atome: 0:CHI_UNSPECIFIED, 1:CHI_TETRAHEDRAL_CW, 2:CHI_TETRAHEDRAL_CCW	1

FIGURE 2.8 – vecteur d'entrée dans [gab23]

En ce qui concerne le MPNN, on a décidé dans [BKSS23] de diviser les vecteurs d'entrée en vecteurs de caractéristiques des atomes et vecteurs de caractéristiques des arêtes 2.9.

Caractéristiques des atomes		
Caractéristique	Description	Dimension
Valence	One-hot encoding de la valence totale de l'atome (0-6)	7
Degree	One-hot encoding du degré de l'atome (0-5)	6
Nombre d'hydrogènes	One-hot encoding du nombre d'atomes d'hydrogène voisins (0-4)	5
Charge formelle	One-hot encoding de la charge électronique (-2, -1, 0, 1, 2)	5
Numéro atomique	One-hot encoding du numéro atomique (1-100)	100
Hybridation	One-hot encoding de l'hybridation (SP, SP2, SP3, SP3D, SP3D2)	5

Caractéristiques des arêtes		
Caractéristique	Description	Dimension
Liaison simple	One-hot encoding du type de liaison single	2
Liaison double	One-hot encoding du type de liaison double	2
Liaison triple	One-hot encoding du type de liaison triple	2
Liaison aromatique	One-hot encoding du type de liaison aromatique	2

FIGURE 2.9 – vecteurs d'entrée dans [BKSS23]

Formes matricielles

Pour simplifier la compréhension des formules mathématiques nous allons passé aux formules matricielles.

1- GCN

Rappelons que la propagation des informations dans un GCN repose sur un mécanisme de passage de messages entre les nœuds. La mise à jour des représentations des nœuds suit la formule suivante :

$$h_v^{(t)} = \sigma\left(\underbrace{\sum_{u \in N(v)} \overbrace{W^{(t)} \frac{h_u^{(t-1)}}{|N(v)|}}^{\text{Message}}}_{\text{Aggregation}}\right)$$

Cette équation se divise en deux étapes essentielles :

— Message Passing : Chaque nœud reçoit les informations de ses voisins u pondérées par une matrice de transformation $W^{(t)}$.

— Aggregation : Les messages sont ensuite moyennés sur l'ensemble des voisins $N(v)$ pour produire une nouvelle représentation du nœud v, avant d'être passés par une fonction d'activation σ.

En notation matricielle, cette opération peut être exprimée comme suit :

$$\mathbf{H}^{(l)} = \hat{\mathbf{A}}\mathbf{D}^{-1}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)}$$

où $\hat{\mathbf{A}}\mathbf{D}^{-1}$ représente la normalisation de la matrice d'adjacence. Cette formulation permet d'implémenter efficacement le passage de messages dans les GCN.

Prenons l'exemple de l'Acétone qui a pour formule chimique 2.10 Voici les matrices utilisées dans l'algorithme

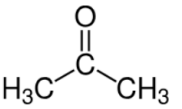

FIGURE 2.10 – Formule chimique de l'Acétone

du GCN.

1	1	0	0
1	1	1	1
0	1	1	0
0	1	0	1

FIGURE 2.11 – matrice d'adjacence normalisée Â

2	0	0	0
0	4	0	0
0	0	2	0
0	0	0	2

FIGURE 2.12 – matrice de degré normalisée D

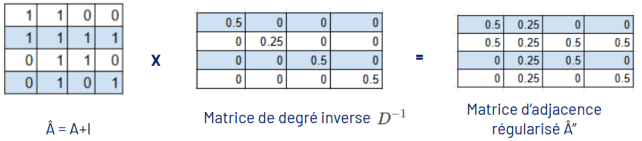

FIGURE 2.13 – calcul de la matrice d'adjacence régularisé Â"

2.6 Propagation des caractéristiques

Dans un GCN, la mise à jour des caractéristiques se fait via la propagation des messages selon l'équation :

$$\mathbf{H}^{(l)} = \hat{A}D^{-1}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)} \tag{2.1}$$

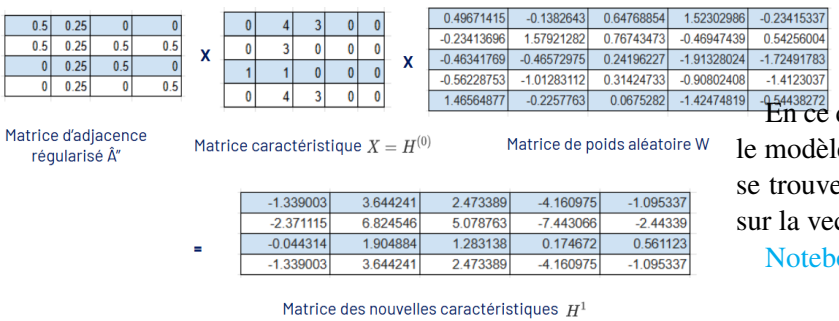

FIGURE 2.14 – calcul de la matrice caractéristique de la couche 1 $H^{(1)}$

2- MPNN Pour obtenir la matrice caractéristique de la couche (t+1) en utilisant le MPNN, il faut passer par deux étapes

— Message Passing qui a pour formule :

$$m_v^{t+1} = \sum_{u \in N(v)} W(e_{uv}) h_u^t$$

Dans sa version matricielle, cela peut être exprimé comme suit :

$$M^{(t+1)} = D^{-1/2}\hat{A}D^{-1/2}H^tW$$

— Mise à jour GRU(Gated Recurrent Unit), utilisée pour mettre à jour l'état des nœuds :

$$h_v^{t+1} = GRU(h_v^t, m_v^{t+1})$$

Et en notation matricielle :

$$H^{t+1} = GRU(H^t, M^{t+1})$$

En ce qui concerne l'exemple de l'Acétone en utilisons le modèle de MPNN, voici un lien vers Google Colab où se trouve le code qui calcule la matrice H^1 en se basant sur la vectorisation du git OpenPOM [BKSS23] :

Notebook MPNN - Colab

Intérêts des propositions

Dans le cadre de la réalisation de notre projet, l'article [BSLW19] constitue une source d'inspiration majeure. En effet, notre objectif est d'améliorer, ou du moins d'explorer, des possibilités d'amélioration de la prédiction des

odeurs à partir d'une représentation en graphe des molécules. Cette implémentation fournit un point de comparaison précieux pour évaluer les performances de notre modèle.Au départ, nous avons choisi d'explorer en profondeur le modèle GCN afin de mieux comprendre son fonctionnement et d'envisager des améliorations. Cependant, nous avons rapidement constaté un manque d'informations détaillées à son sujet, notamment dans [gab23], rédigé par des étudiants et peu approfondi. Nous avons donc décidé d'orienter nos recherches vers le modèle MPNN, d'autant plus que nous avons réalisé par la suite que ce dernier repose sur le principe de convolution utilisé dans le GCN.

L'article [LS24] présente un intérêt relatif, car il se concentre sur la prédiction de plusieurs odeurs simultanément, alors que notre objectif est d'identifier les parties spécifiques des molécules responsables de ces ressentis, un problème sensiblement différent. Néanmoins, cet article met en évidence les limites des performances du GIN sur ce type de problème, ce qui reste une observation utile pour nos travaux.

Limites des propositions

La prédiction d'odeur pour des paires de molécules repose sur des données intrigantes. En effet, la fonction blender disponible sur le site The Good Scent est mal documentée, et son fonctionnement reste non compris ou inexpliqué au sein de la communauté des parfumeurs, notamment sur les forums spécialisés. En conséquence, les résultats obtenus à partir de ces données sont probablement dénués de toute signification fiable.

L'une des limites de l'article [BSLW19] est qu'il ne prend pas en compte la représentation spatiale des molécules, un facteur qui pourrait avoir un impact sur la précision des prédictions.

2.7 Récapitulatif

Dans ce chapitre, nous avons exploré les approches contemporaines de prédiction des odeurs moléculaires à travers les Graph Neural Networks (GNNs), en analysant deux études clés portant sur la modélisation des graphes moléculaires simples et des interactions entre paires moléculaires. Ces recherches ont révélé l'efficacité remarquable des GNNs dans le traitement des données structurées, leur capacité à décoder les relations complexes entre atomes et liaisons dépassant significativement les méthodes traditionnelles. Nous avons comparé différentes variantes de GNNs, notamment le Message Passing Neural Networks (MPNN) et le Graph Convolution Networks (GCN). Chaque modèle présente des avantages et des inconvénients selon le contexte et les données utilisées. Les performances des GNNs ont été démontrées comme étant supérieures à celles des approches traditionnelles, bien que certaines limites subsistent, notamment en termes de complexité computationnelle et de prise en compte de la structure spatiale des molécules. Cette analyse approfondie des hyperparamètres et architectures des modèles constitue un point de départ essentiel pour orienter nos choix méthodologiques et explorer plus avant les potentialités des réseaux neuronaux graphiques dans la prédic-

tion des propriétés olfactives.

3

Propositions

Après avoir pris connaissance de l'état des travaux actuels, nous constatons que les modèles Graph Neural Network (GNN) sont particulièrement pertinents pour la prédiction des odeurs à partir des molécules. Ils permettent d'exploiter la structure en graphe des molécules.

3.1 Idées préliminaires

Dans l'article [OPE23], nous remarquons que certains labels, tels que "fruité", sont présents sur de nombreuses molécules et présentent une forte cooccurrence avec d'autres labels. Une question se pose alors : la présence de ces labels influence-t-elle la précision des prédictions ?

De cette observation, une idée émerge : est-il possible d'utiliser des odeurs plus spécifiques, comme "abricot", pour améliorer la prédiction des odeurs ? Pour répondre à cette question, nous envisageons d'exploiter la structure en graphe présente dans SketchOscent 3.1 [VFPG22].

Nous proposons également de déterminer quels facteurs influencent la précision des prédictions, tels que la fonction de READOUT, le nombre de couches de "message passing", ainsi que la présence ou l'absence de certaines molécules dans notre jeu de données.

3.2 Formalisation

Les odeurs peuvent être représentées sous forme de hiérarchie 3.1. Chaque odeur est une sous-classe d'une odeur plus générale. Les odeurs qui ne sont la superclasse d'aucune autre odeur sont référencées comme des odeurs individuelles, ce qui correspond à la plus petite unité de description d'une odeur. Ces odeurs sont bien plus spécifiques que celles situées plus haut dans la hiérarchie.

Si nous sommes capables de prédire correctement ces odeurs spécifiques, nous pourrions peut-être établir que toutes les superclasses de cette odeur constituent un descripteur d'odeur de cette molécule.

Cela présuppose que la prédiction de l'odeur est correcte. Cette hypothèse sera étudiée dans la section **??**.

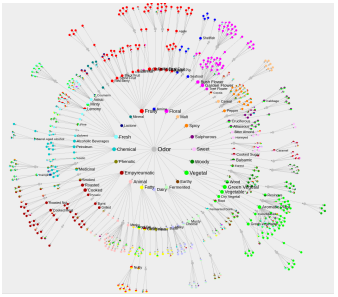

FIGURE 3.1 – Structure hiérarchique des odeurs

Pour exploiter la structure hiérarchique des odeurs, nous récupérons le fichier RDF du graphe des odeurs et sélectionnons les odeurs individuelles. Nous supprimons ensuite tous les descripteurs qui ne sont pas des odeurs individuelles, puis nous entraînons un MPNN sur ces données. Les hyperparamètres utilisés sont ceux de [BKSS23].

Grâce à cette prédiction, nous récupérons toutes les classes supérieures hiérarchiques de notre label et lui attribuons ce descripteur.

4

Expérimentation et Résultat

Ce chapitre a pour but d'expérimenter et d'identifier les facteurs influençant la précision des prédictions. Nous chercherons à mesurer l'influence de différents éléments, tels que les données utilisées, les labels à prédire, le nombre de couches de message passing, ainsi que les descripteurs d'odeurs.

4.1 Les données

Pour nos expérimentations, nous utilisons les données issues de SketchOscent. Afin de garantir la qualité du jeu de données, nous nous assurons de l'absence de doublons. Les codes CAS sont convertis en codes SMILES isomériques à l'aide de la bibliothèque Python *PubChemPy*. Les molécules pour lesquelles aucune correspondance entre CAS et SMILES n'est trouvée sont supprimées.

De plus, seuls les labels présents sur au moins 30 molécules sont conservés afin d'assurer une représentation suffisante de chaque label dans les ensembles d'entraînement et de test. Ainsi, sur les 395 descripteurs d'odeurs initiaux, il en reste 125.

Certaines molécules partagent un même code SMILES mais ont plusieurs codes CAS. Nous analysons l'impact de cette duplication sur la précision des prédictions.

Après ce filtrage, le jeu de données contient 3737 molécules, contre 3920 initialement.

4.2 Expérimentations

Afin d'identifier les facteurs influençant la prédiction des odeurs, nous faisons varier plusieurs paramètres et analysons leur impact sur la précision des résultats.

4.2.1 Modification du dataset

Cette section évalue l'impact des changements dans le jeu de données sur la précision des prédictions.

Nous comparons la performance du modèle sur trois versions du dataset :

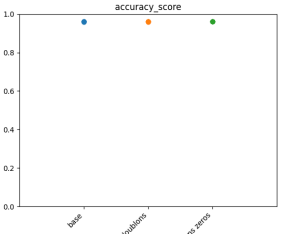

FIGURE 4.1 – Calcul de l'accuracy en fonction du dataset

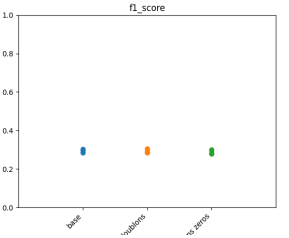

FIGURE 4.2 – Calcul du score F1 en fonction du dataset

1. Le dataset original.
2. Un dataset sans les molécules possédant un SMILES répliqué.
3. Un dataset sans les molécules ne possédant aucun descripteur d'odeur.

Les métriques utilisées pour comparer les performances sont l'AUROC, l'accuracy, le F1-score et l'indice de Jaccard.

L'objectif de cette analyse est de déterminer s'il est pertinent de conserver ou de supprimer les duplications de SMILES. En effet, si deux molécules distinctes partagent un même code SMILES mais possèdent des descripteurs odorants différents, leur fusion pourrait induire une confusion dans le modèle.

De plus, nous vérifions l'impact de la suppression des molécules sans descripteurs d'odeur. L'absence de descripteurs pourrait en soi constituer une information utile pour la prédiction odorante.

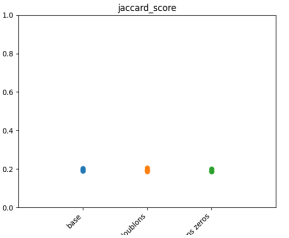

FIGURE 4.3 – Calcul de l'indice de Jaccard en fonction du dataset

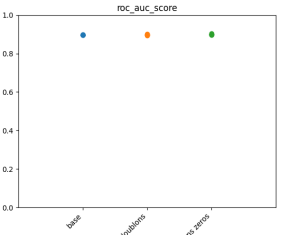

FIGURE 4.4 – Calcul du score AUROC en fonction du dataset

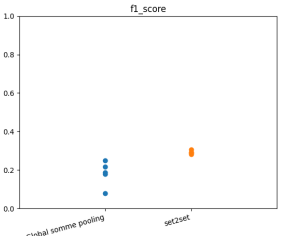

FIGURE 4.5 – F1-score des modèles en fonction de la fonction de READOUT

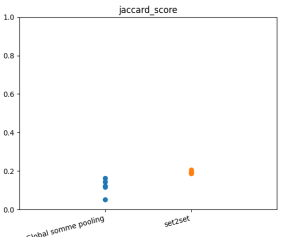

FIGURE 4.6 – Indice de Jaccard des modèles en fonction de la fonction de READOUT

4.2.2 Changement de la fonction de READOUT

Les modèles de OPENPOM [BKSS23] utilisent une fonction Set2Set composée de 2 couches et de 3 étapes. Nous nous interrogeons sur l'impact de cette fonction d'activation sur les performances des prédictions. Pour cela, nous avons remplacé la fonction Set2Set par une fonction de global sum pooling.

Les résultats obtenus sont présentés dans les figures 4.5 4.6

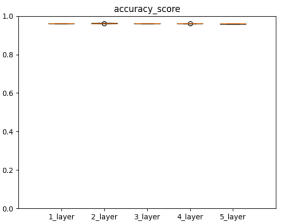

FIGURE 4.7 – Accuracy des modèles en fonction du nombre de couches

On observe que les performances sont nettement meilleures avec la fonction Set2Set, et surtout beaucoup plus stables, il y a moins de variations des performances entre les modèles.

4.2.3 Variation du nombre de couches de message passing

Nous analysons l'effet du nombre de couches de *message passing* sur la précision des prédictions. Pour cela, nous faisons varier ce paramètre de 1 à 5 couches et entraînons cinq modèles différents par configuration afin de lisser les variations de performance.

Les résultats sont illustrés dans les figures 4.7, 4.10, 4.8 et 4.9.

4.2.4 Odeurs individuelles

SketchOscent possède une hiérarchie des odeurs, avec des odeurs plus générales situées plus haut dans l'arbre hiérarchique. La question qui se pose est de savoir si

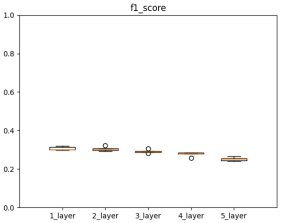

FIGURE 4.8 – F1-score des modèles en fonction du nombre de couches

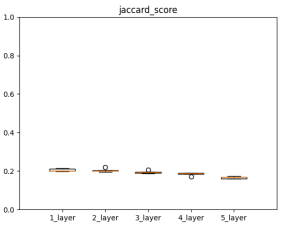

FIGURE 4.9 – Indice de Jaccard des modèles en fonction du nombre de couches

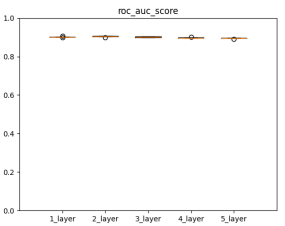

FIGURE 4.10 – AUROC des modèles en fonction du nombre de couches

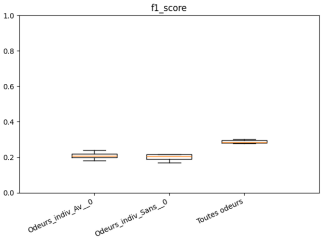

FIGURE 4.11 – Évaluation des modèles avec le score F1, le premier avec les odeurs individuelles et sans enlever de molécule, le second avec les odeurs individuelles sans les molécules qui n'ont pas de descripteur d'odeur et le dernier avec tous les descripteurs d'odeurs

les odeurs les plus générales, étant présentes sur un plus grand nombre de molécules, ont une information odorante moins précise, car elles sont plus diffusées. Nous tentons donc d'isoler les odeurs les plus spécifiques, situées en bas de cette hiérarchie. Les résultats sont présentés dans les figures 4.12 4.11 L'AUROC et l'accuracy n'apportant pas de valeur ajoutée dans ce contexte, ils ne seront pas présentés.

4.2.5 Impact des modifications du dataset

Les résultats montrent qu'il n'y a pas de différence significative dans la précision des prédictions selon la version du dataset utilisée. Cela peut s'expliquer par le fait que les modifications apportées concernent un nombre relativement faible de molécules (252 en moins sur un total de 3737), rendant l'impact négligeable. Toutefois, une lé-

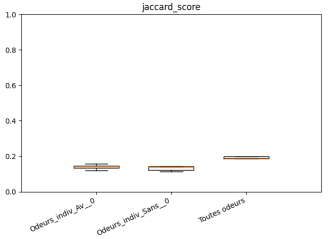

FIGURE 4.12 – Évaluation des modèles avec l'indice de Jaccard, le premier avec les odeurs individuelles et sans enlever de molécule, le second avec les odeurs individuelles sans les molécules qui n'ont pas de descripteur d'odeur et le dernier avec tous les descripteurs d'odeurs

gère augmentation du F1-score et de l'indice de Jaccard est observée.

Un point notable est que les scores d'accuracy et d'AUROC sont élevés, tandis que les scores de Jaccard et F1 sont relativement faibles. Cela s'explique probablement par le fait qu'une molécule possède en moyenne 4.6 descripteurs d'odeur, alors que le nombre total de descripteurs est de 125, ce qui entraîne un déséquilibre dans la distribution des labels prédits.

4.2.6 Impact de la variation du nombre de couches

Les résultats montrent que le nombre de couches de message passing n'a pas d'impact significatif sur la précision globale des prédictions. Toutefois, à partir de 4 couches, une légère diminution du F1-score et de l'indice de Jaccard est observée, suggérant un possible surapprentissage ou une propagation excessive de l'information.

4.2.7 Analyse de la prédiction sur des odeurs individuelles

On constate que la prédiction des odeurs individuelles présente des performances inférieures par rapport aux modèles utilisant l'ensemble des odeurs. Une hypothèse pourrait être que ces odeurs spécifiques contiennent des informations plus ciblées, rendant leur représentation moléculaire plus complexe et donc plus difficile à isoler avec nos modèles. En revanche, aucune différence significative n'a été observée dans les performances lorsqu'on conserve les molécules sans descripteurs d'odeur.

4.3 Conclusion

Suite à nos expérimentations, nous avons constaté que le prétraitement des données et la variation du nombre de couches de message passing ont un impact limité sur la précision des prédictions. En revanche, les éléments ayant un fort impact sur les performances sont les labels à prédire. Nous avons observé une baisse de la précision lors de la prédiction d'odeurs spécifiques. Pour confirmer que cela est dû à la spécificité des odeurs, il serait pertinent de mesurer les performances sur des prédictions d'odeurs plus générales. Une piste d'amélioration pourrait être l'intégration de méthodes d'apprentissage contrastif, afin de mieux distinguer les nuances entre

différentes odeurs et affiner les prédictions sur des molécules spécifiques.

Toutefois, nous avons constaté des difficultés dans la prédiction des odeurs spécifiques. Les résultats suggèrent que certaines odeurs plus précises sont plus complexes à identifier correctement. Une piste d'amélioration pourrait être d'intégrer des méthodes d'apprentissage contrastif pour mieux distinguer les nuances entre différentes odeurs et affiner les prédictions sur des molécules spécifiques.

5

Conclusion

Cette étude démontre que les Graph Neural Networks (GNN) constituent une approche prometteuse pour analyser et prédire les propriétés olfactives des molécules. En utilisant la Carte Principale des Odeurs (POM), nous avons mis en évidence une organisation cohérente de l'espace olfactif, surpassant les méthodes traditionnelles de chémoinformatique dans la représentation des similitudes entre les odeurs.

Cependant, certains défis restent à relever, notamment la prise en compte des interactions entre molécules dans les mélanges odorants. Pour améliorer encore la précision des prédictions et la généralisation des modèles, les recherches futures pourraient explorer des approches contrastives et enrichir les bases de données avec des annotations plus détaillées. Ces avancées ouvriraient la voie à des applications plus fiables dans la modélisation des odeurs.

Bibliographie

[BKSS23] Aryan Amit Barsainyan, Ritesh Kumar, Pinaki Saha, and Michael Schmuker. Openpom - open principal odor map, 2023. https://github.com/BioMachineLearning/openpom. 13, 14, 15, 17, 21, 24, 30

[BSLW19] Brian K Lee1 Richard C Gerkin Alán Aspuru-Guzik Benjamin Sanchez-Lengeling, Jennifer N Wei and Alexander B Wiltschko1r. Machine learning for scent : Learning generalizable perceptual representations of small molecules. 2019. 5, 10, 12, 17, 18, 34

[CJ23] Thomas Clouet and Gabriel Jolly. Fouille de données olfactives : Clustering de molécule odorantes par gnn (graph neural networks). Rapport de recherche et développement, École Polytechnique de l'Université de Nantes, Département d'Informatique, février 2023. 12, 30

[gab23] gabikun. Pred. https://github.com/gabikun/PRED/tree/main, 2023. 15, 18, 30

[LMSL+22] Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in human olfactory perception. *bioRxiv*, 2022. 5, 10, 13, 30, 36

[LS24] Amit Barsainyan Mrityunjay Sharma Ritesh Kumar Laura Sisson, Aryan. Olfactory label prediction on aroma-chemical pairs. 2024. 5, 8, 18, 35, 43, 44

[OPE23] principal odor map unifies diverse tasks in olfactory perception. *science*, 2023. 20

[Sis25] Laura Sisson. odor_pair, 2025. https://github.com/odor-pair/odor-pair.

[VFPG22] Angélique Villière, Catherine Fillonneau, Carole Prost, and Fabrice Guillet. Sketchoscent : towards a knowledge-based model and interactive visualisation of the odour space. In *Proceedings of the 16th Weurman Flavour Research Symposium*, 2022. 20

[YCS+21] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations, 2021.

Table des figures

Liste des tableaux

Rappels

Les réseaux neuronaux graphiques (GNN) sont une forme de réseaux neuronaux artificiels conçus pour traiter les données graphiques. Contrairement à des domaines plus couramment étudiés, tels que les images et le texte, les graphes présentent des informations hautement structurées dont il faut tenir compte lorsqu'on travaille avec eux. Les GNNs sont capables d'exploiter à la fois les informations contenues dans les composants du graphe (nœuds et arêtes) et les informations structurelles inhérentes au graphe lui-même.

B

Fiches de lecture

B.0.1 Machine Learning for Scent : Learning Generalizable Perceptual Representations of Small Molecules [BSLW19].

Introduction :

Cet article traite de la difficulté de prédire la relation entre la structure d'une molécule et son odeur, un problème connu sous le nom de Quantitative Structure-Odor Relationship (QSOR). L'objectif est de développer des modèles prédictifs robustes en utilisant des réseaux de neurones graphiques (Graph Neural Networks - GNN) pour analyser des molécules et déduire leurs propriétés odorantes. Les auteurs proposent une approche novatrice qui surpasse les méthodes existantes, avec des implications pour la neuroscience sensorielle, la chimie des fragrances et la durabilité écologique.

Résumé

La prédiction des relations entre la structure moléculaire et les odeurs constitue un défi scientifique complexe, où chaque variation moléculaire peut profondément modifier la perception sensorielle. Malgré des années de recherche en chimie, neurosciences et apprentissage automatique, ce défi reste partiellement exploré.
Dans cette perspective, les chercheurs ont développé une approche innovante s'appuyant sur deux variantes de réseaux de neurones graphiques (Graph Neural Networks - GNN). Les Graph Convolution Networks (GCN), et les Message Passing Neural Networks (MPNN). Chaque molécule est désormais considérée comme un graphe dynamique, où les atomes deviennent des nœuds interconnectés, permettant de générer des modèles capables d'anticiper 138 descripteurs olfactifs distincts.
Les résultats montrent que le modèle GCN atteint une performance prédictive remarquable, avec une aire sous la courbe ROC de 0,894, surpassant significativement les méthodes traditionnelles basées sur les empreintes moléculaires (AUROC = 0,850 pour les caractéristiques Mor-

dred).
Mais encore les embeddings des GNN montrent une organisation perceptive cohérente : les descripteurs d'odeurs proches (e.g., fruité, ananas) forment des clusters. En plus des représentations qui reflètent une hiérarchie implicite, où des descripteurs larges (e.g., floral) regroupent des sous-catégories spécifiques (e.g., jasmin, lavande).
La base de données utilisée dans cette étude provient de deux bases : GoodScents, riche de 3 786 molécules, et Leffingwell PMP 2001, avec 3 561 molécules. Après nettoyage et fusion, 5 030 molécules porteuses de 138 descripteurs ont été retenues, ouvrant la voie à une compréhension inédite de l'architecture sensorielle.

Analyse

L'article s'appuie sur une méthodologie solide et des bases de données d'experts, garantissant des résultats de haute qualité. Des comparaisons avec des approches conventionnelles (forêts aléatoires, empreintes moléculaires) valident l'efficacité des GNN grâce à des métriques de performance robustes (AUROC, précision, etc.). Les résultats incluent des intervalles de confiance, renforçant la crédibilité des analyses. Nous nous sommes appuyés sur cet article pour mieux comprendre le fonctionnement et la construction des Graph Convolutional Networks (GCN)

B.0.2 article : Olfactory Label Prediction on Aroma-Chemical Pairs[LS24]

L'article ambitionne de développer un modèle capable de prédire l'odeur résultante d'une paire de molécules. Bien qu'il existe déjà des modèles prédictifs pour une odeur à partir d'une molécule individuelle, les interactions complexes entre molécules dans un mélange rendent cette tâche beaucoup plus difficile. Ces interactions, souvent non linéaires et nombreuses, justifient l'utilisation de modèles d'intelligence artificielle. Plus précisément, les Graph Neural Networks (GNN) sont particulièrement adaptés à ce problème, car ils offrent de bons résultats et correspondent bien à la structure des molécules, qui peuvent être facilement modélisées sous forme de graphes.

Données

Les données utilisées proviennent du site The Good Scent Company, qui répertorie environ 3,5 mille molécules. Les auteurs exploitent la fonctionnalité blender de ce site pour attribuer des labels aux paires de molécules. Cette méthode leur permet de générer un ensemble de données de 165 mille paires de molécules annotées.

Une segmentation des données en ensembles d'entraînement et de test est réalisée, respectant une proportion de 50 :50. L'entraînement est effectué avec une validation croisée sur 5 partitions différentes. Bien que la segmentation soit effectuée aléatoirement, elle s'assure que chaque segment contient un échantillon représentatif des labels.

Méthodologie

Les modèles utilisés pour la prédiction sont le Graph Isomorphism Network (GIN) et le Message Passing Neural Network (MPNN). Le MPNN s'appuie sur l'implémentation décrite dans l'article "Principal Odor Map Unifies Diverse Tasks in Olfactory Perception" (Science, 381(6661) :999–1006, 2023). L'implémentation du GIN provient de la bibliothèque PyTorch Geometric.

Le GIN utilise deux couches de message passing, avec une fonction d'update basée sur un perceptron multicouche (MLP). Chaque couche a une dimension de 832. Pour chaque atome de chaque molécule, le GIN génère un embedding. Après les étapes de message passing, deux fonctions de readout sont appliquées : mean pooling et add pooling. Les résultats des deux fonctions sont concaténés.

Pour représenter une paire de molécules, les embeddings des deux molécules sont combinés aléatoirement, puis passent dans un MLP à deux couches pour obtenir un embedding final de dimension 832. La fonction de coût utilisée est l'entropie croisée binaire.

Résultats et analyse

Le MPNN surpasse systématiquement le GIN, que ce soit pour des prédictions basées sur une seule molécule ou sur une paire. L'écart de performance est encore plus marqué pour les prédictions sur une molécule unique.

La différence de performance est attribuée à la manière dont les deux modèles traitent les paires de molécules. Le MPNN représente la paire comme un graphe disjoint et utilise une phase de readout basée sur la méthode Set2Set, considérée comme plus avancée que le mean pooling ou le global sum pooling utilisés par le GIN.

Conclusion

L'article s'apparente davantage à une preuve de concept qu'à une recherche approfondie. En effet, les données utilisées pour les paires de molécules proviennent du site The Good Scent Company, mais il semble que la fonctionnalité blender n'ait pas été conçue pour prédire l'odeur de mélanges de molécules. Ces données ont été exploitées faute de disposer d'un jeu de données contenant des mélanges de molécules annotés spécifiquement pour ce type de tâche.

B.0.3 A Principal Odor Map Unifies Diverse Tasks in Human Olfactory Perception[LMSL+22]

Contexte et Problématique

L'olfaction est un sens complexe où la relation entre la structure chimique des molécules et la perception des odeurs reste mal comprise. Contrairement à la vision ou à l'audition, où les stimuli physiques (longueur d'onde, fréquence) sont bien corrélés à la perception (couleur, hauteur), la relation entre la structure chimique et la perception olfactive est souvent discontinue. Les modèles traditionnels, basés sur des propriétés physiques ou des empreintes moléculaires, ne parviennent pas à capturer ces relations complexes.

Objectif

L'objectif de cette étude est de créer une **Carte Principale des Odeurs** (*Principal Odor Map, POM*) en utilisant des réseaux de neurones graphiques (*Graph Neural Networks, GNN*) pour prédire la qualité des odeurs de nouvelles molécules. Cette carte vise à unifier les différentes tâches de perception olfactive humaine et à fournir une représentation généralisable des relations structure-odeur.

Méthodologie

Modèle de Réseau de Neurones Graphiques (GNN)

— Chaque molécule est représentée comme un graphe, où les atomes et les liaisons sont décrits par des caractéristiques spécifiques (valence, degré, nombre d'hydrogènes, etc.).
— Le modèle est entraîné sur un ensemble de données de référence de $\sim$5000 molécules, chacune décrite par plusieurs étiquettes d'odeur (ex. "crémeux", "herbacé").

Validation Prospective

— Un panel de 15 sujets entraînés a été utilisé pour évaluer la perception de 400 nouvelles molécules odorantes.
— Les performances du modèle ont été comparées à celles des panélistes humains, en utilisant des méthodes statistiques pour mesurer la concordance entre les prédictions du modèle et les moyennes du panel.

Analyse de la Carte des Odeurs (POM)

— La POM a été comparée à des modèles chémoinformatiques traditionnels (comme les empreintes de Morgan) pour évaluer sa capacité à représenter les distances perceptuelles et les hiérarchies olfactives.

Résultats

Performance du Modèle

— Le modèle GNN a atteint une performance de prédiction de la qualité des odeurs comparable à celle d'un humain moyen. Sur un ensemble de validation de 400 molécules, les prédictions du modèle correspondaient mieux à la moyenne du panel que celles du panéliste médian.
— Le modèle a surpassé les modèles chémoinformatiques traditionnels, démontrant que la POM encode efficacement une carte généralisée des relations structure-odeur.

Représentation des Relations Perceptuelles

— La POM a mieux représenté les distances perceptuelles et les hiérarchies olfactives que les modèles basés sur les empreintes de Morgan.
— Les molécules partageant une étiquette d'odeur étaient plus densément regroupées dans la POM,

indiquant une meilleure représentation des relations perceptuelles.

Généralisation à d'autres Tâches Olfactives

— La POM a été utilisée pour prédire des tâches olfactives telles que la détectabilité des odeurs, la similarité olfactive et l'applicabilité des descripteurs d'odeurs. Elle a surpassé les modèles chémoinformatiques sur plusieurs ensembles de données publiés.

Discussion

— La POM propose une carte intuitive et hiérarchique de l'espace moléculaire en termes d'odeurs, similaire à la manière dont les espaces de couleur représentent les longueurs d'onde de la lumière.
— Le modèle est robuste aux discontinuités dans les relations structure-odeur et peut être utilisé pour explorer l'espace des odeurs à grande échelle.
— Les limitations incluent l'influence de la concentration des odeurs, la présence de contaminants odorants, et la nécessité de données supplémentaires pour améliorer la précision du modèle.

Conclusion

Cette étude propose une **Carte Principale des Odeurs (POM)** basée sur des réseaux de neurones graphiques, capable de prédire avec précision la qualité des odeurs de nouvelles molécules et de généraliser à diverses tâches olfactives. Cette carte représente une avancée significative dans la compréhension de la perception olfactive et ouvre la voie à la numérisation des odeurs.

Planification

Analyse des diagramme La comparaison entre le planning prévisionnel (FigureC.1) et le planning effectif (FigureC.2) met en évidence un décalage dans les priorités et le déroulement du projet. Initialement, le plan prévoyait une implémentation directe des modèles GCN et GCL. Cependant, en cours de projet, il est apparu que la compréhension approfondie du MPNN était essentielle avant tout développement. Cela a conduit à un réajustement des tâches, avec un focus initial sur les corrections et l'analyse des représentations matricielles avant l'implémentation d'OpenPOM et du MPNN. Par conséquent, moins de temps a été alloué au développement d'une solution optimale, mais nous avons pu mieux comprendre les mécanismes internes du modèle et tester l'impact des variations des paramètres, notamment le nombre de couches de message passing.

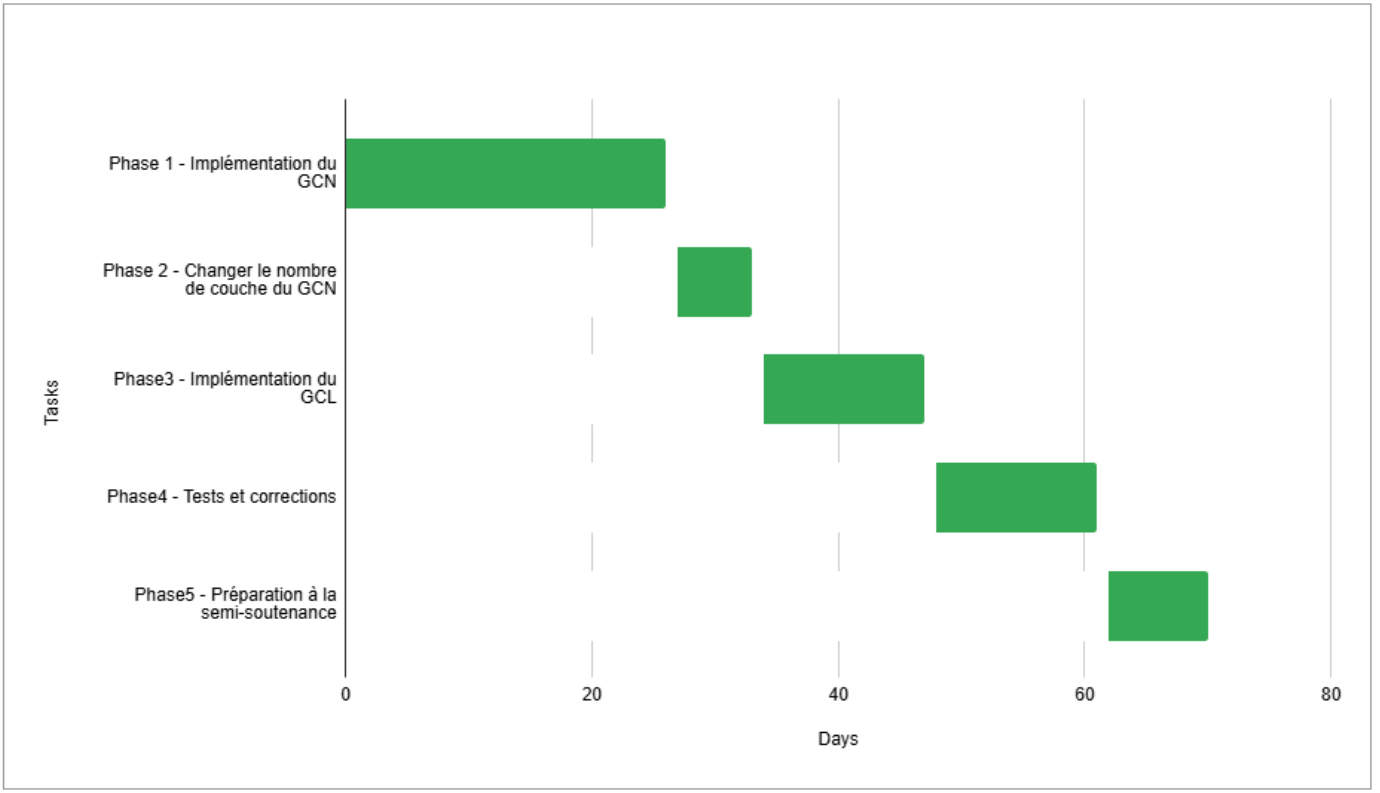

FIGURE C.1 – Planification prévisionnelle

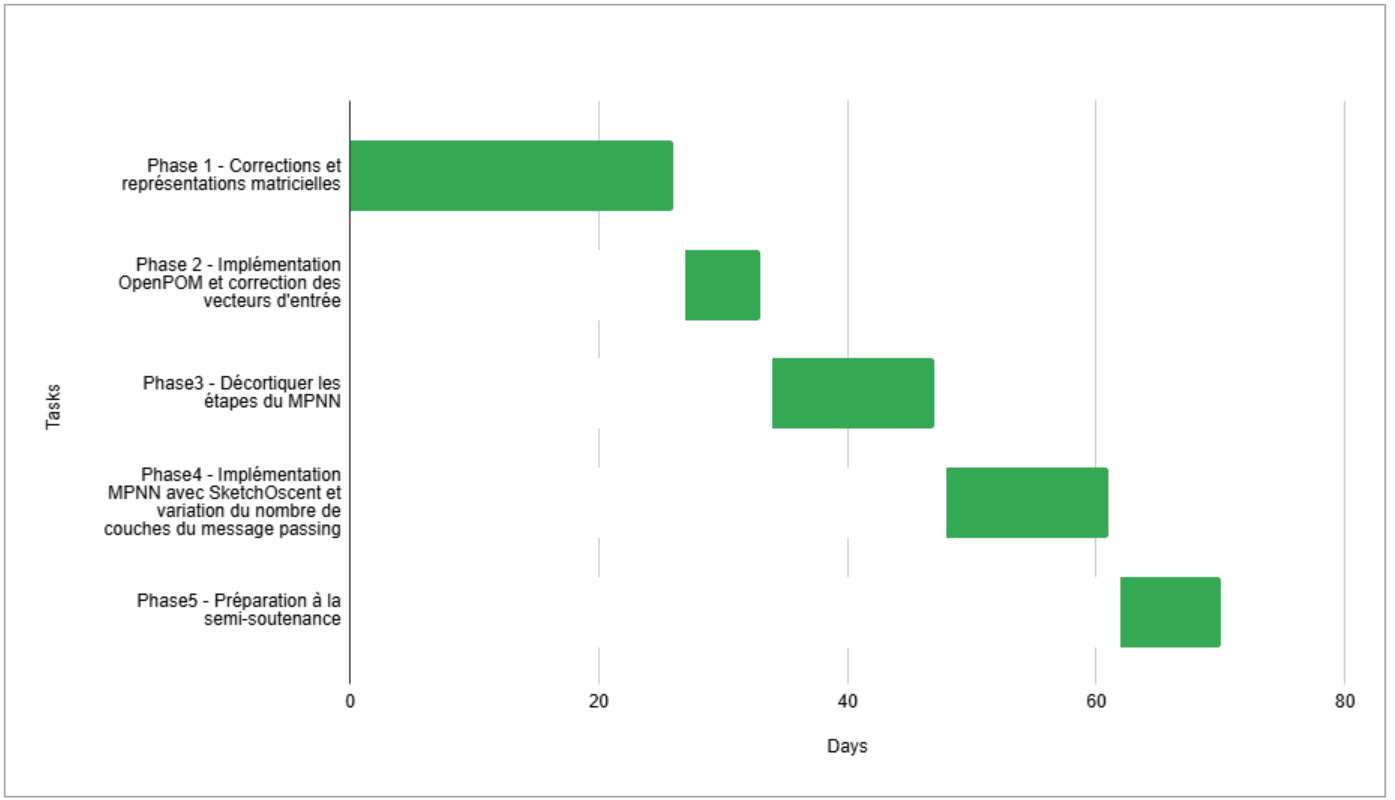

FIGURE C.2 – Planning effectif

Fiches de suivi

Fiche de suivi de la semaine 1
du 14 octobre 2024 au 20 octobre 2024

Temps de travail de Colin TRÈVE: 14 h 30 m
Temps de travail de Marwa TABIB: 13 h 00 m
Travail effectué.

- tâche 1 : Division des tâches, chacun de nous s'est concentré sur un article de son côté
- tâche 2 : Lecture des articles et des fichiers supplémentaires pour mieux comprendre le sujet
- tâche 3 : 1ère réunion de nous deux pour mettre en commun ce que que chacun a lu de son côté.
- tâche 4 : 2ème réunion avec les encadrants pour expliquer ce qu'on a compris du sujet et des articles.

Travail non effectué.

- Pour cette première semaine on a pu achevé toutes les tâches qu'on avait à faire.

Échanges avec le commanditaire.

- En discutant avec les encadrants, on a pu avoir d'autres sources pour mieux comprendre le GNN comme des cours du professeur *Leskovec* de l'université de Stanford et d'autres cours de *FIDLE*

Planification pour la semaine prochaine.

- Lire les cours de FIDLE & du professeur Leskovec ;
- Compréhension des modèles MPNN et GCN ;

Fiche de suivi de la semaine 2
du 21 octobre 2024 au 27 octobre 2024

Temps de travail de Colin TRÈVE: 14 h 00 m
Temps de travail de Marwa TABIB: 13 h 30 m
Travail effectué.

- tâche 1 :Plannification prévisionnel du projet
- tâche 2 : Chacun de son côté a lu les cours donnés par les encadrants.

— tâche 3 : Avec la base qu'on a acquis maintennat on a pus mieux comprendre à le fonctionnement des modèles cités dans les articles
— tâche 4 : Réunion de nous deux pour mettre en commun notre travail.
— tâche 5 : Réunion avec les encadrants.

Échanges avec le commanditaire.

Lors de notre réunion hebdomadaire, nos encadrants ont clarifié plusieurs points qui nous semblaient obscurs dans nos lectures d'articles et de cours, nous aidant à préciser notre compréhension et à identifier les axes essentiels de notre travail.

Planification pour la semaine prochaine.

— Implémentation du OpenPOM (le code dans un git où ils utilisent le modèle MPNN) ;
— Définir les différences entre les différents modèles du GNN (GCN, MPNN, GIN) ;

Fiche de suivi de la semaine 3
du 04 novembre 2024 au 10 novembre 2024

Temps de travail de Colin TRÈVE: 16 h 50 m
Temps de travail de Marwa TABIB: 17 h 30 m

Travail effectué.

— tâche 1 : Nous sommes passés à la comparaison des modèles séparèment : - Colin : comparaison entre MPNN et GIN en se basant sur son article[LS24] et le git OpenPOM - Marwa : comparaison entre MPNN et GCN en se basant sur son article *Machine Learning for Scent : Learning Generalizable Perceptual Representations of Small Molecules* et le git OpenPOM
— tâche 2 : Nous avons tenté d'implémenter le code du dépôt OpenPOM, mais nous avons été confrontés à des problèmes de compatibilité entre les versions des bibliothèques, ce qui a rallongé notre processus d'implémentation.
— tâche 3 : Réunion de nous deux pour mettre en commun notre travail.
— tâche 4 : Réunion avec les encadrants pour montrer l'avancement du projet.

Travail non effectué.

— Echec de l'implémentation

Échanges avec le commanditaire.

Pendant la réunion, nous nous sommes rendu compte que la comparaison que nous avons effectué était superficielle par rapport à ce qui nous a été demandé. Il ne fallait pas se contenter de la définition de chaque modèle et le nombre des couches mais plutôt connaître les vecteurs d'entrée, les vecteurs de sorties, la fonction d'activation, plus de détails sur les couches des modèles...

Planification pour la semaine prochaine.

— Faire une comparasion plus approfondie
— Chercher des informations sur le GIN
— Expliquer le "Odor Mixture Hypergraph" qui se trouve dans l'article[LS24]
— faire des recherches sur l'implémentation de l'ap-

proche contrastive

Fiche de suivi de la semaine 4
du 11 novembre 2024 au 17 novembre 2024

Temps de travail de Colin TRÈVE: 14 h 30 m
Temps de travail de Marwa TABIB: 12 h 50 m
Travail effectué.

- tâche 1 : Décortiquer le code OpenPOM pour en tirer pour l'instant : le vecteur de caractéristique des molécules car nous avons des diffucltés à comprendre le code.
- tâche 2 : Mieux comprendre les composants le graphe "Odor Mixture Hypergraph"
- tâche 3 :Réunion du binôme pour synchroniser notre travail où Colin a partagé qu'il a trouvé un autre git lié à l'article [LS24]
- tâche 4 :Réunion hebdomadaire avec les encadrants

Travail non effectué.

Nous n'avons pas pu aller plus loin dans la comparaison détaillés du à la mal compréhension du code et le manque d'information dans les articles lus.

Échanges avec le commanditaire.

Pendant notre échange avec les encadrants,il s'est avéré qu'encore une fois la comparaison n'était pas assez complète. Il manquait la construction des données(la manière dont les experts extraient les données), les différentes couche des modèles et leur taille, la quantité des odeurs encodées et les couches intermédiaires.

Planification pour la semaine prochaine.

- Comparaison des modèles
- Définir le décodeur utilisé pour chaque modèle (nombre de couche, nombre de molécule par couche)

Fiche de suivi de la semaine 5
du 18 novembre 2024 au 24 novembre 2024

Temps de travail de Colin TRÈVE: 14 h 00 m
Temps de travail de Marwa TABIB: 14 h 00 m
Travail effectué.

- Comparaison des modèles entre les articles : réalisé à 50%.
- Exploration des repository git : réalisé à 30%.
- Compréhension de la vectorisation des atomes et des liens dans une molécule : terminé à 50%.

Travail non effectué.

Renseignement sur les approches contrastives appliquées aux graphes (GCL).

Échanges avec le commanditaire.

Réunion du 21/11/24 : Échange sur la comparaison des modèles

Planification pour la semaine prochaine.

- — Apport de précision dans la comparaison des modèles de ML
- — Exploration plus en profondeur des repositories Git
- — Comprendre la fonction Blender sur le site *The Good Scent*.
- — Déterminer une valeur ajoutée.

Fiche de suivi de la semaine 6
du 25 novembre 2024 au 29 novembre 2024

Temps de travail de Colin TRÈVE: 14 h 00 m
Temps de travail de Marwa TABIB: 14 h 00 m
Travail effectué.

- — Renseignement sur les modèles Long Short Term Memory (LSTM) : difficulté moyenne, réalisé à 100%.
- — Compréhension de l'implémentation des articles : progressé à 70%.
- — Exécution des modèles de machine learning : réalisé à 30%.
- — Tableau récapitulatif des différents modèles étudiés dans les articles : réalisé à 80%.
- — Compréhension de la vectorisation des atomes et des liens dans une molécule : terminé à 100%.

Travail non effectué.

- — Renseignement sur les approches contrastives appliquées aux graphes (GCL).
- — Compréhension de la séparation entraînement/test dans un article (analyse de la librairie utilisée).
- — Apporter une piste d'amélioration

Échanges avec le commanditaire.

Réunion du 27/11/24 : Échange sur les données issues d'un article et discussion sur la pertinence des choix effectués.

Planification pour la semaine prochaine.

- — Finaliser l'exécution des implémentations.
- — Apporter des modifications aux différentes implémentations.
- — Comprendre la fonction Blender sur le site *The Good Scent*.
- — Déterminer une valeur ajoutée.

Fiche de suivi de la semaine 7
du 2 décembre 2024 au 6 décembre 2024

Temps de travail de Colin TRÈVE: 18 h 00 m
Temps de travail de Marwa TABIB: 17 h 30 m
Travail effectué.

- — Réalisation du rapport
- — Réalisation des fiches de lecture
- — Ré-implémentation d'un git

— Proposition de valeur ajoutée

Travail non effectué.

Implémentation d'une version de GCL

Échanges avec le commanditaire.

Réunion du 5/11/24 : Discussion sur la proposition de travail

Planification pour la semaine prochaine.

— Commencer à implémenter une version de GCL
— Présentation orale

Fiche de suivi de la semaine 8
du 16 décembre 2024 au 22 décembre 2024

Temps de travail de Colin TRÈVE: 13 h 30 m
Temps de travail de Marwa TABIB: 10 h 50 m

Travail effectué.

— Correction des slides (Ajout de sources, harmonisation des formules mathématiques, Ajout des informations manquantes) ;

Échanges avec le commanditaire.

Réunion le 17/12/2024 : Discussion des modifications apportées à ce jour.

Planification pour la semaine prochaine.

— Lecture de l'article "A principal odor map unifies diverse tasks in olfactory perception.Science"
— Compléter l'implémentation du GCN
— Essayer différentes nombre de couches pour le GCN
— Ajouter les formules matricielles des étapes du GNN

Fiche de suivi de la semaine 9
du 6 janvier 2025 au 12 janvier 2025

Temps de travail de Colin TRÈVE: 12 h 00 m
Temps de travail de Marwa TABIB: 11 h 00 m

Travail effectué.

— Ajout des formules matricielles et de l'exemple de l'Acétone
— Lecture de l'article "A principal odor map unifies diverse tasks in olfactory perception.Science"
— Implémentation du code utilisant le GCN

Travail non effectué.

Essayer différentes nombre de couches pour le GCN

Échanges avec le commanditaire.

Réunion du 07/01/25 : Discussion des mise à jour effectuées. L'exemple de l'Acétone n'était pas complet.

Planification pour la semaine prochaine.

— Essayer d'implémenter OpenPOM
— Essayer différentes nombre de couches pour le mesaage passing dans openPOM
— Rajouter le schéma de la molécule de l'Acétone.

— Rechercher l'encodage exact des vecteurs d'entrée

Fiche de suivi de la semaine 10
du 13 janvier 2025 au 19 janvier 2025

Temps de travail de Colin TRÈVE: 10 h 30 m
Temps de travail de Marwa TABIB: 13 h 40 m

Travail effectué.

— Implémentation de OpenPOM
— Rajout du schéma de la molécule de l'Acétone avec son calcul matriciel.
— Recherche de l'encodage exact des vecteurs d'entrée dans le git openPOM

Travail non effectué.

Essayer différents nombre de couches pour le MPNN

Échanges avec le commanditaire.

Réunion du 16/01/25 : Discussion des mise à jour effectuées. Les vecteurs d'entrée ne sont toujours pas clairs ni les étapes du MPNN. Mr. Guillet a suggéré de refaire tourner le code OpenPOM pour voir ce qui se passe à l'intérieur de la "boîte noire" (MPNN)

Planification pour la semaine prochaine.

— Essayer différentes nombre de couches pour le MPNN
— Refaire tourner le code OpenPOM pour voir ce qui se passe à l'intérieur
— Approfondir les recherches pour trouver les étapes exactes du MPNN
— Rechercher l'encodage exact des vecteurs d'entrée

Fiche de suivi de la semaine 11
du 20 janvier 2025 au 26 janvier 2025

Temps de travail de Colin TRÈVE: 13 h 30 m
Temps de travail de Marwa TABIB: 11 h 00 m

Travail effectué.

— Variation du nombre de couches de message passing
— Rajout des formules matricielles des étapes de MPNN.
— Création d'un premier schéma expliquant les étapes exactes du MPNN

Travail non effectué.

Recherche de l'encodage exact des vecteurs d'entrée

Échanges avec le commanditaire.

Réunion du 21/01/25 : Discussion des mise à jour effectuées. Le schéma des étapes openPOM n'était toujours pas aussi bien détaillé que voulu.
Discussion autour de la raison pour laquelle les scores des prédictions ne varient pas.

Planification pour la semaine prochaine.

— Essayer différentes nombre de couches pour le MPNN
— Recherche de la cause de la variation et utilisation de nouvelles métriques

Fiche de suivi de la semaine 12 du 27 janvier 2025 au 02 février 2025

Temps de travail de Colin TRÈVE: 10 h 50 m

Temps de travail de Marwa TABIB: 11 h 00 m

Travail effectué.

— Variation du nombre de couches de message passing de 1 à 5
— Réalisation de 5 modèles différents pour chaque couche.
— utilisation de seuil pour vérifier les scores de prédiction sont cohérents avec la réalité

Échanges avec le commanditaire.

Il n'y a pas eu de réunion pendant cette semaine.

Planification pour la semaine prochaine.

— Prendre en compte si la hiérarchie de l'odeur de SketchOscent
— Changer la fonction de READOUT

Fiche de suivi de la semaine 13 du 03 février 2025 au 09 février 2025

Temps de travail de Colin TRÈVE: 6 h 30 m

Temps de travail de Marwa TABIB: 11 h 00 m

Travail effectué.

— Changer la fonction de READOUT
— Réalisation du deuxième schéma des étapes du MPNN

Échanges avec le commanditaire.

Réunion du 04/02/25 : Discussion des mise à jour effectuées pendant les deux dernières semaines.

Planification pour la semaine prochaine.

— Implémenter les données SketchOscent
— Prendre en compte si la hiérarchie de l'odeur de SketchOscent
— Essayer avec les odeurs spécifiques et générales et voir la différence des résultats.

Fiche de suivi de la semaine 14 du 10 février 2025 au 16 février 2025

Temps de travail de Colin TRÈVE: 14 h 30 m

Temps de travail de Marwa TABIB: 14 h 00 m

Travail effectué.

— Prendre en compte si la hiérarchie de l'odeur de SketchOscent
— Compléter le deuxième schéma des étapes du MPNN
— rédaction du rapport

Échanges avec le commanditaire.

Réunion du 11/02/25 : Discussion des mise à jour effectuées pendant les deux dernières semaines. Pendant cette réunion, on était tous un peu perdu au niveau du fonctionnement exact du modèle du MPNN.

Planification pour la semaine prochaine.

— Revenir à la vision globale
— Reproduire les couches de openpom détaillées avec les dimensions.

Fiche de suivi de la semaine 15
du 17 février 2025 au 16 février 2025

Temps de travail de Colin TRÈVE: 21 h 50 m
Temps de travail de Marwa TABIB: 22 h 30 m

Travail effectué.

— Correction des formules matricielles
— Schématisation complète du MPNN
— Finition du code
— Rédaction du rapport
— Modification de la représentation

Le tableau D.1 récapitule le taux d'avancement du projet. Rappelons que le temps de travail théorique *minimal* correspond au temps indiqué sur la maquette pédagogique auquel on ajoute un strict minimum de 20 % correspondant au travail personnel hors emploi du temps. La partie « haute » de la fourchette correspond à 50 % de temps supplémentaire au titre du travail personnel.

	Temps prévu		Colin TRÈVE			Marwa TABIB		
	bas	haut	hebdo.	Σ	%	hebdo.	Σ	%
Semaine	h : m	h : m	h : m	h : m		h : m	h : m	
1	10 : 00	12 : 30	14 : 30	14 : 30	145 (116)	13 : 00	13 : 00	130 (104)
2	20 : 00	25 : 00	14 : 00	28 : 30	142 (114)	13 : 30	26 : 30	132 (106)
3	30 : 00	37 : 30	16 : 50	45 : 20	151 (120)	17 : 30	44 : 00	146 (117)
4	40 : 00	50 : 00	14 : 30	59 : 50	149 (119)	12 : 50	56 : 50	142 (113)
5	50 : 00	62 : 30	14 : 00	73 : 50	147 (118)	14 : 00	70 : 50	141 (113)
6	60 : 00	75 : 00	14 : 00	87 : 50	146 (117)	14 : 00	84 : 50	141 (113)
7	70 : 00	87 : 30	18 : 00	105 : 50	151 (120)	17 : 30	102 : 20	146 (116)
8	80 : 00	100 : 00	13 : 30	119 : 20	149 (119)	10 : 50	113 : 10	141 (113)
9	90 : 00	112 : 30	12 : 00	131 : 20	145 (116)	11 : 00	124 : 10	137 (110)
10	100 : 00	125 : 00	10 : 30	141 : 50	141 (113)	13 : 40	137 : 50	137 (110)
11	110 : 00	137 : 30	13 : 30	155 : 20	141 (112)	11 : 00	148 : 50	135 (108)
12	120 : 00	150 : 00	10 : 50	166 : 10	138 (110)	11 : 00	159 : 50	133 (106)
13	130 : 00	162 : 30	6 : 30	172 : 40	132 (106)	11 : 00	170 : 50	131 (105)
14	140 : 00	175 : 00	14 : 30	187 : 10	133 (106)	14 : 00	184 : 50	132 (105)
15	150 : 00	187 : 30	21 : 50	209 : 00	139 (111)	22 : 30	207 : 20	138 (110)

TABLE D.1 – Avancement du projet par rapport au temps de travail théorique minimal (respectivement haut)

Auto-contrôle et auto-évaluation

Phase I : Etude préalable, étude bibliographique et conception générale

Rapport	Organisation	Plan	Equilibre	X			
			Cohérence		X		
		Fluidité	Introductions (partielles)	X			
			Transitions		X		
			Conclusions (partielles)	X			
		Tableaux, figures	Numérotés	X			
			Légendés		X		
			Références (non "en ligne")	X			
	Rédaction	Orthographe	Coquilles		X		
			Fautes évitables		X		
			Franglais, jargon		X		
		Rédaction	Aisée		X		
			Absence de plagiat !		X		
	Bibliographie	Références	Suffisantes (nombre, intérêt)		X		
			Pérennes		X		
			Complètes (auteurs, pages..)		X		
			Conséquentes (volume)		X		
		Références dans le texte			X		
		Proposition de note haute	16,09				
		Proposition de note basse	11,09				
		Proposition de note du jury					
Projection	Organisation	Plan		X			
		Liaisons		X			
		Numérotation		X			
	Contenu	Informatif			X		
		Concis			X		
		Clair			X		
		Orthographe		X			
		Illustrations		X			
Oral	Présentation	Aisance			X		
		Tenue		X			
		Articulation, compréhension		X			
	Durée	Respect			X		
		Temps de parole équilibré			X		
	Réponses	Pertinence			X		
		Argumentation			X		
		Proposition de note haute	16,30				
		Proposition de note basse	11,48				
		Proposition de note du jury					

FIGURE E.1 – Points à contrôler à l’issue de la phase I