

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

Réalisé par: Colin Trève & Marwa TABIB  
Encadré par: Fabrice Guillet & Angélique Villière  
Coordinateur: Philippe LERAY

# Table de matière

01

**Introduction**

02

**Solutions  
proposées**

03

**Comparaison**

04

**Proposition**

05

**Diagramme de  
Gantt**

06

**Conclusion**

01

# Introduction





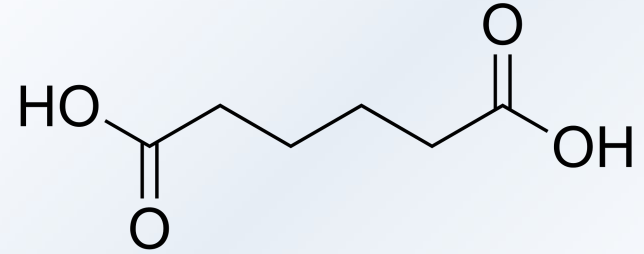
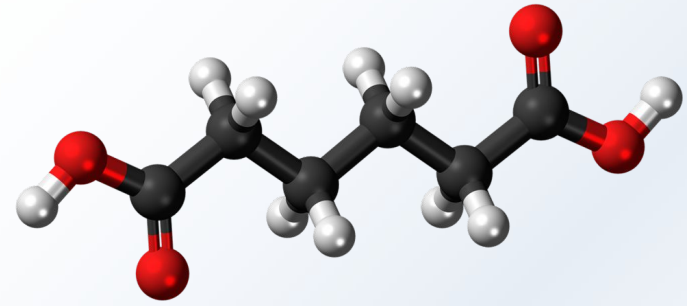
# Sujet

- Perception olfactive
- Relation entre la structure moléculaire et l'odeur

# Problématique

Reconnaître des odeurs à partir d'une molécule

Représenter une molécule par un graphe

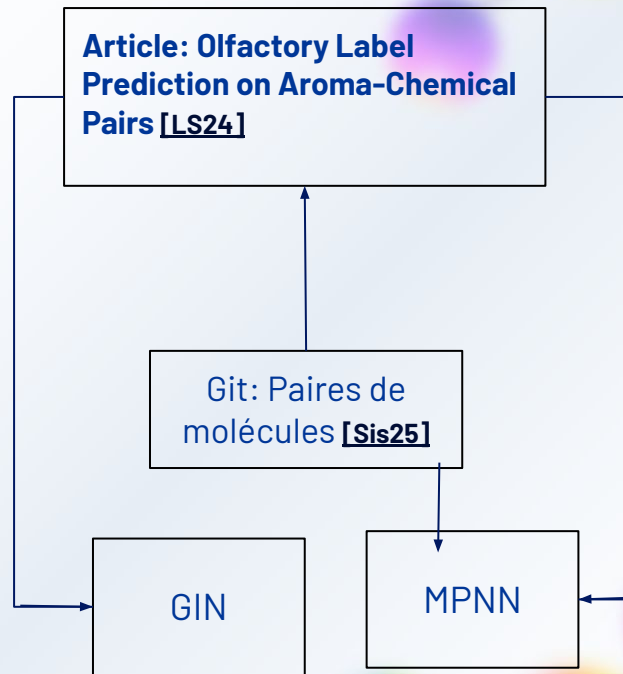
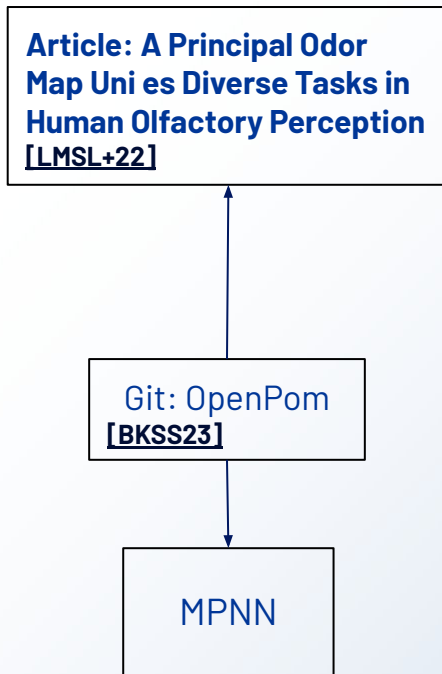
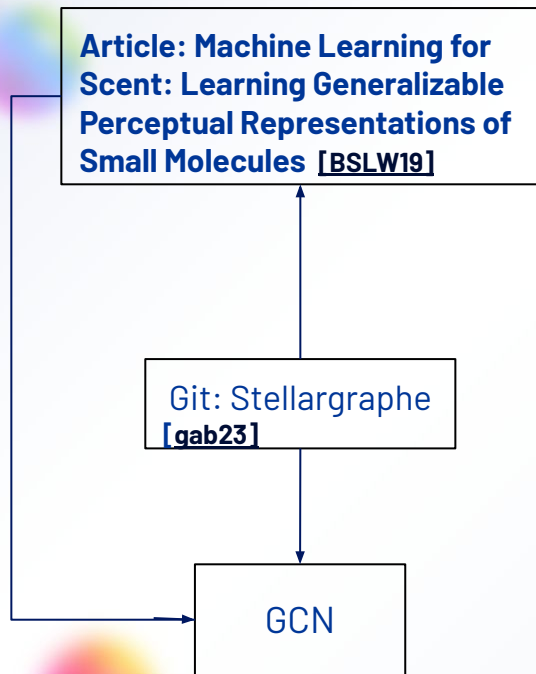


02

# Solutions proposées



# Sources



# Solutions proposées dans la littérature

- Random forest models (RF)
- k-nearest neighbor models (KNN)
- GNN

## Comparaison

	AUROC	Precision	F1
GNN	<b>0.894 [0.888, 0.902]</b>	<b>0.379 [0.351, 0.398]</b>	<b>0.360 [0.337, 0.372]</b>
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 1: Tableau de comparaison des modèles de prédiction d'odeur - article  
**[BSLW19]**



# GNN (Graph Neural Network)

- Couche de GNN
  - Message Passing
  - Agrégation des message
- Phase de READOUT (agrégation des noeuds)
- Décodeur (prédiction)

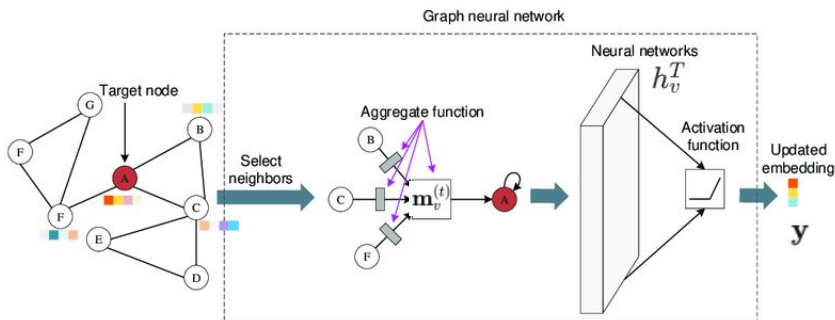


Figure 3: Processus de *Message Passing* et d'*Agrégation* dans les GNN [RSGH]

$$\mathbf{m}_v^{(t)} = \text{AGGREGATE}(\{\mathbf{h}_u^{(t-1)}, \forall u \in \mathcal{N}(v)\})$$

$$\mathbf{h}_v^{(t)} = \text{UPDATE}(\mathbf{h}_v^{(t-1)}, \mathbf{m}_v^{(t)})$$

$$\mathbf{y} = \text{READOUT}(\{\mathbf{h}_v^{(T)}, \forall v \in G\})$$

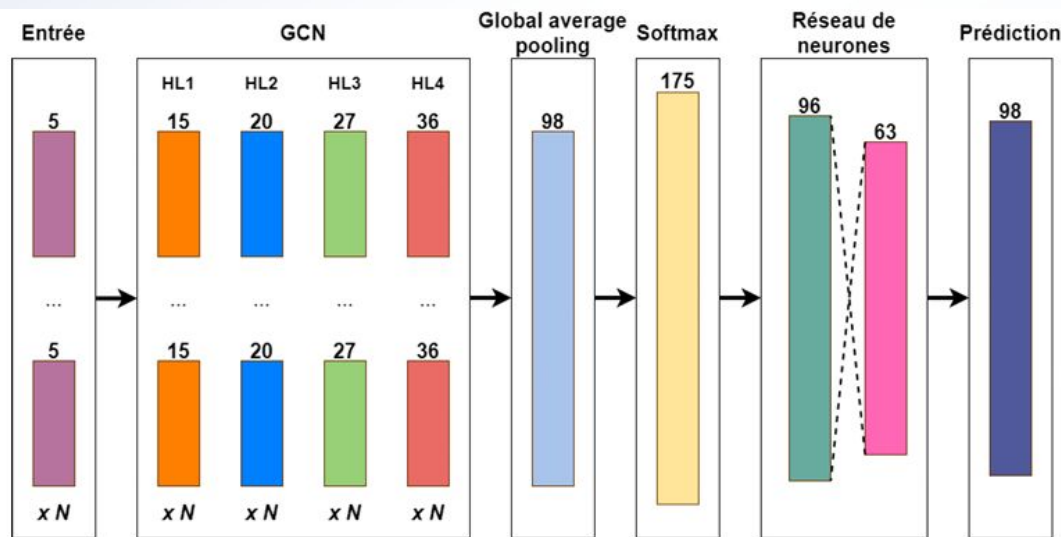
$$\mathbf{z}_g = \text{MLP}(\mathbf{y})$$

02

# Modèles de GNN



# GCN



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

$h_u^{(t)}$  est la représentation du nœud  $u$  à la couche  $t$ .  
 $N(v)$  est l'ensemble des voisins du nœud  $v$ .

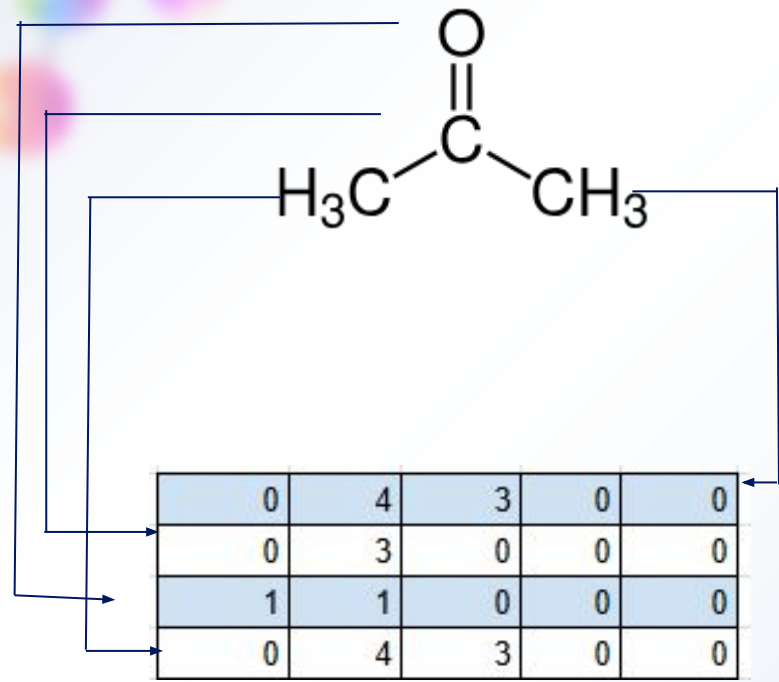
**Formule matricielle:**

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

Figure 3: Architecture du GCN- rapport PRED

2023:[CJ23]

# Exemple: Acétone(C3H6O)



0	4	3	0	0
0	3	0	0	0
1	1	0	0	0
0	4	3	0	0

Matrice caractéristique  $X = H^{(0)}$

Type (symbol)	Degré	Valence implicite	Aromatique	Chiralité
0 (C)	4	3	0	0
0 (C)	3	0	0	0
1(O)	1	0	0	0
0 (C)	4	3	0	0

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

Matrice d'adjacence A

1	0	0	0
0	3	0	0
0	0	1	0
0	0	0	1

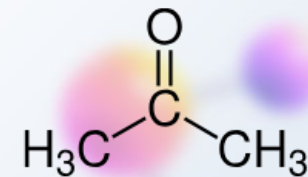
Matrice de degré D'

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

Matrice de degré normalisée D= D'+I

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

## Exemple: Acétone(C3H6O)



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

$$\hat{A} = A + I$$

x

0.5	0	0	0
0	0.25	0	0
0	0	0.5	0
0	0	0	0.5

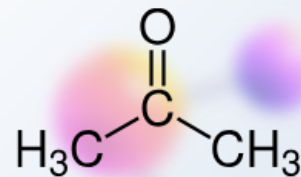
Matrice de degré inverse  $D^{-1}$

=

0.5	0.25	0	0
0.5	0.25	0.5	0.5
0	0.25	0.5	0
0	0.25	0	0.5

Matrice d'adjacence  
régularisé  $\hat{A}''$

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



0.5	0.25	0	0
0.5	0.25	0.5	0.5
0	0.25	0.5	0
0	0.25	0	0.5

$\times$

0	4	3	0	0
0	3	0	0	0
1	1	0	0	0
0	4	3	0	0

$\times$

0.49671415	-0.1382643	0.64768854	1.52302986	-0.23415337
-0.23413696	1.57921282	0.76743473	-0.46947439	0.54256004
-0.46341769	-0.46572975	0.24196227	-1.91328024	-1.72491783
-0.56228753	-1.01283112	0.31424733	-0.90802408	-1.4123037
1.46564877	-0.2257763	0.0675282	-1.42474819	-0.54438272

Matrice d'adjacence  
régularisé  $\hat{\mathbf{A}}$

Matrice caractéristique  $\mathbf{X} = \mathbf{H}^{(0)}$

Matrice de poids aléatoire  $\mathbf{W}$

$=$

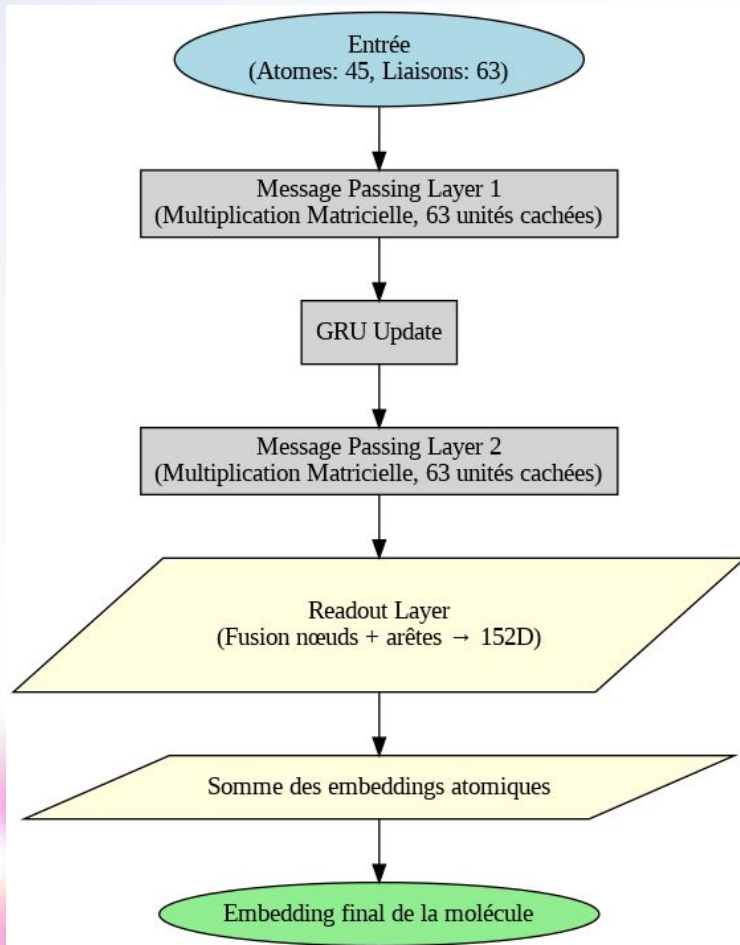
-1.339003	3.644241	2.473389	-4.160975	-1.095337
-2.371115	6.824546	5.078763	-7.443066	-2.44339
-0.044314	1.904884	1.283138	0.174672	0.561123
-1.339003	3.644241	2.473389	-4.160975	-1.095337

Matrice des nouvelles caractéristiques  $\mathbf{H}^1$

# ÉTAPES DU MPNN



## Dans l'article [LMSL+22]



## Dans le git [BKSS23]





# MPNN

## 1- Message Passing

$$m_v^{t+1} = \sum_{u \in N(v)} W(e_{uv}) h_u^t$$
$$M^{(t+1)} = D^{-1/2} \hat{A} D^{-1/2} H^t W$$

## 2- Mise à jour GRU

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

## 3- Phase du readout

$$y' = \phi \left( h_v^{(T)}, \sum_{u \in N(v)} \psi \left( h_u^{(T)}, e_{uv} \right) \right)$$
$$Y' = H^t W_z + \sum_{u \in N(v)} E W$$

## 4- Agrégation finale

$$y = \sum_{v \in V} y'$$
$$Y = \sum_{v \in V} Y'$$

# L'exemple de l'Acétone

En ce qui concerne l'exemple de l'acétone, voici un lien du google colab où se trouve le code qui calcule la matrice  $H^1$  en se basant sur la vectorisation du git OpenPOM [\[BKSS23\]](#): [Untitled9.ipynb - Colab](#)

03

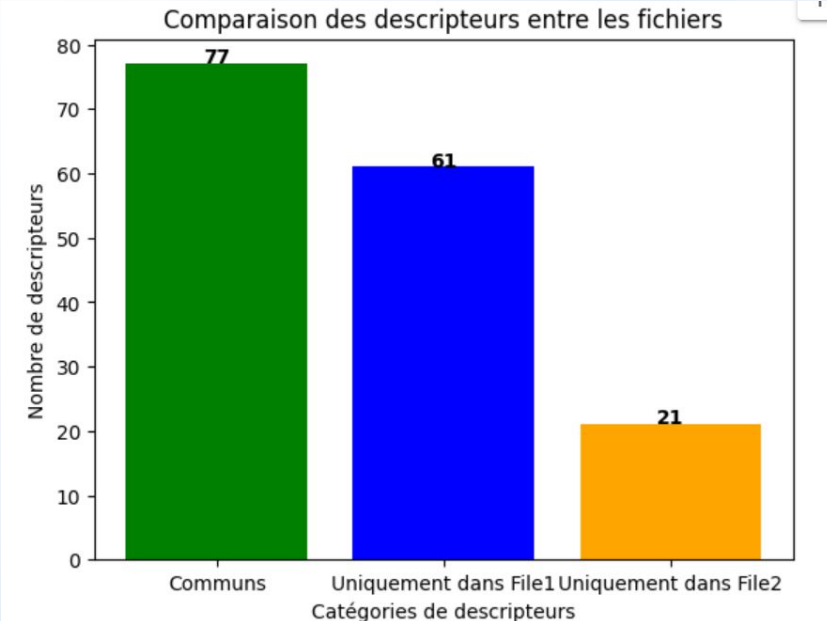
# Comparaison



openPom	stellargraphe
aldehyde	aldehyde
alliacious	alliacious
almond	almond
animal	animal
apple	apple
apricot	apricot
balsamic	balsamic
banana	banana
berry	berry
burnt	burnt
buttery	buttery
cabbage	cabbage
caramel	caramel
cedar	cedar
cheese	cheese
cherry	cherry
chocolate	chocolate
cinnamon	cinnamon
citrus	citrus
cocoa	cocoa
coconut	coconut
coffee	coffee

alcoholic	amberggris
amber	aniseed
anisc	beeswax
aromatic	bread
beefy	broth
bergamot	camphor
bitter	chemical
black currant	chicken
brandy	coriander
camphoreous	coumarin
celery	earth
chamomile	ester
clean	ether
clove	fermentative
cooked	fresh milk
cooling	jam
cortex	lactic
coumarinic	licorice
dairy	lily of the valley
dry	lime
earthyness	plastic
ethereal	
fermented	

# Descripteurs d'odeurs utilisés dans OpenPom et Stellargraphe

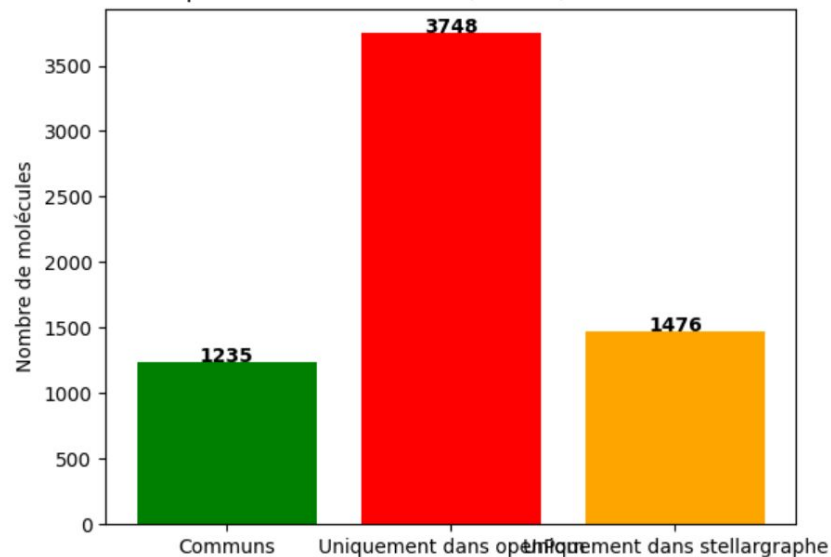


openPom	stellargraphe
C1CCCC2(CCC1)CCCO2	C1CCCC2(CCC1)CCCO2
C1CCCCC2COC2CCCC1	C1CCCCC2COC2CCCC1
C1CCNCC1	C1CCNCC1
C1CSCCS1	C1CSCCS1
C1CSCS1	C1CSCS1
C1CSCSC1	C1CSCSC1
C=CC=O	C=CC=O
C=CCC(C=O)c1ccccc1	C=CCC(C=O)c1ccccc1
C=CCCC(=O)c1ccccc1	C=CCCC(=O)c1ccccc1
C=CCCCCCCCC=O	C=CCCCCCCCC=O
C=CCCCCCCCC(=O)OCC=C	C=CCCCCCCCC(=O)OCC=C
C=CCCCCCCCC=O	C=CCCCCCCCC=O
C=CCCCCCCCCOC#N	C=CCCCCCCCCOC#N
C=CCN=C=S	C=CCN=C=S
C=CCOC(=O)CCC1CCCCC1	C=CCOC(=O)CCC1CCCCC1
C=CCOC(=O)CCCC1CCCCC1	C=CCOC(=O)CCCC1CCCCC1
C=CCOC(=O)CCCCC1CCCCC1	C=CCOC(=O)CCCCC1CCCCC1
C=CCOC(=O)COC1CCCCC1	C=CCOC(=O)COC1CCCCC1
C=CCOC(=O)COe1ccccc1	C=CCOC(=O)COe1ccccc1
C=CCOC(=O)Cc1ccccc1	C=CCOC(=O)Cc1ccccc1
C=CCOC(=O)c1ccccc1	C=CCOC(=O)c1ccccc1

# Descripteurs de molécules utilisées dans OpenPom et Stellargraphe

BrC=Cc1ccccc1	Br\C=C\c1ccccc1
C#CC(C)(O)CCC=C(C)C	BrC1ccc(NC(=O)c2ccccc2)c(c1)c3ccccc3
C#CC1(OC(C)=O)CCCCC1	C(Cc1ccccc1)Cc2ccccc2
C#CC1(OC(C)=O)CCCCC1C(C)CC	C(OCc1ccccc1)c2ccccc2
C#CCO	C(OCc1ccccc1)c2ccccc2
C#CCO.CCC(=O)CCC1C(C)=CCCC	C(SCc1ccccc1)c2ccccc2
C=CC1OCCO1c1ccccc1	C(SSCc1ccccc1)c2ccccc2
C1=CC2COC2CCCCCCCC1	C(SSCc1ccccc1)c2ccccc2
C1=NCCCC1	C(c1ccccc1)c2ccccc2
C1CC2C3CC(C2C1)C1(CCCO1)C3	C(n1ccccc1)c2ccccc2
C1CCC(SSC2CCCCC2)CC1	C/C(C=O)=C/c1ccccc1
C1CCCCC1	C/C(C=O)=C/c1ccccc1
C1CCNC1	C/C=C(/C)C(=O)OCC(C)C
C1CCOCC1	C/C=C(/C)C(=O)OCC(C)C
C1CCSC1	C/C=C(/C)C(=O)OCCc1ccccc1
C1CCSSC1	C/C=C(C)/C=C/C(C)C
C1CNCCN1	C/C=C(C)/C
C1CNCCN1.O=C(O)CCCCC(=O)O	C/C=C(C)/COC(=O)C/C=C/C
C1CS1	C/C=C/C(=O)C1=C(C)C=CCC1(C)C
C1SSSSS1	C/C=C/C(=O)C1=C(C)CCCC1(C)C
C1SCSS1	C/C=C/C(=O)C1C(=CCCC1(C)C)C
C1SSCSSS1	C/C=C/C(=O)C1C(C)=CCC1(C)C
C=C(C)C(C)C(C)C(C)C(C)C	C/C=C/C(=O)C1CCC(C)C(C)=C1C

Comparaison des molécules (SMILES) entre les fichiers





# Vecteur d'entrée de Stellargraph [gab23]

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

# Vecteurs d'entrée de OpenPOM [BKSS23]

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'hydrogène	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

# Tables de comparaison des modèles

	AUROC	Precision	Recall	F1
MPNN	0.890 [0.882, 0.898]	0.379 [0.352, 0.399]	0.387 [0.366, 0.408]	0.362 [0.335, 0.375]
GCN	0.894 [0.888, 0.902]	0.379 [0.351, 0.398]	0.390 [0.365, 0.412]	0.360 [0.337, 0.372]

Figure 5: Performances du MPNN et GCN -article [\[BSLW19\]](#)

04

# Proposition





# Modification modèle actuelle

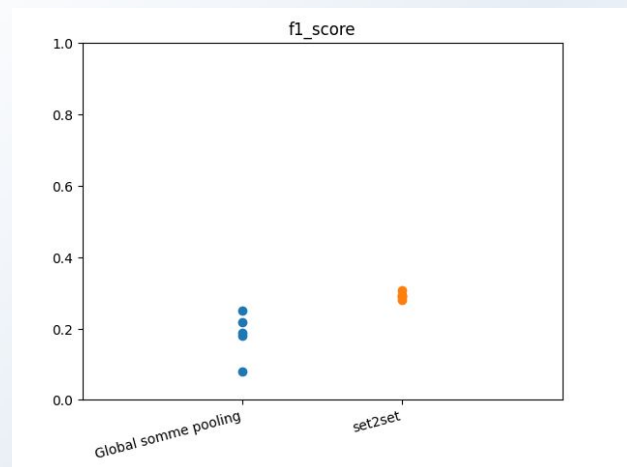
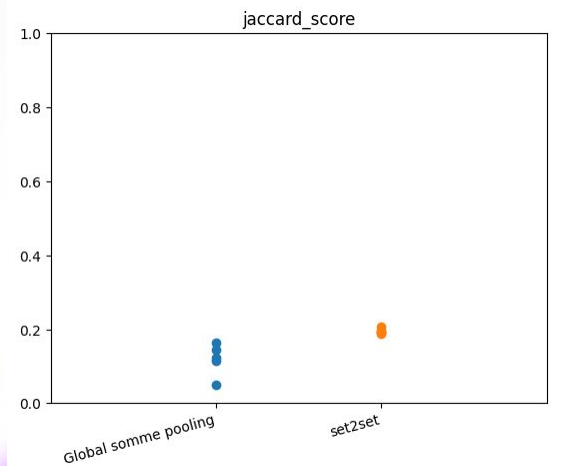
**Objectif** : Réutiliser le modèle de OPENPOM([BKSS23]) et tenter de trouver des pistes d'amélioration pour cela on va modifier :

- Les données du dataset
- La fonction de READOUT
- Le nombres de couches
- En fonction de la hiérarchisation des odeurs de SketchOscnt

# Modification de la fonction de READOUT

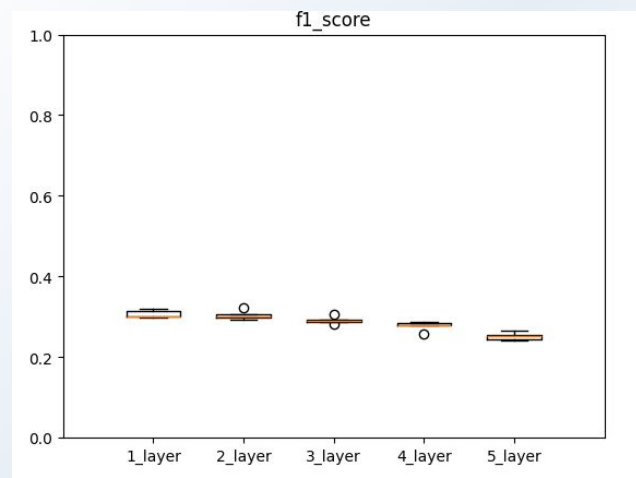
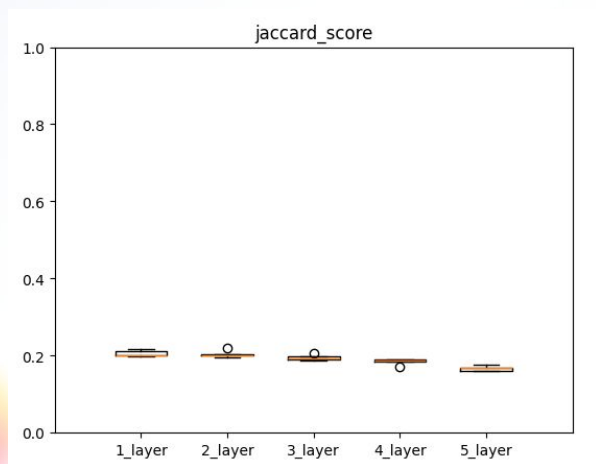
On modifie la fonction de READOUT:

- de set2set (2 couches et 3 step)
- à globale somme pooling.



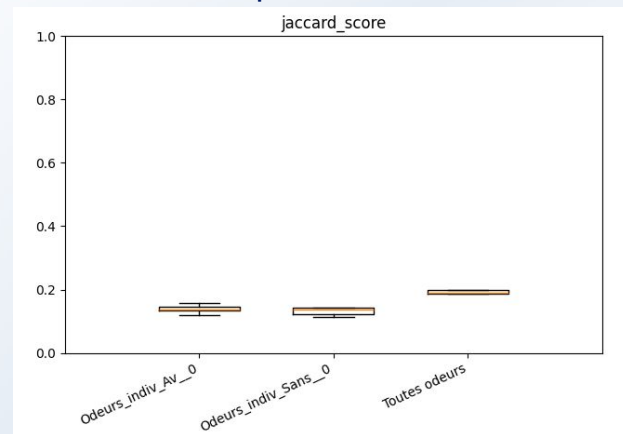
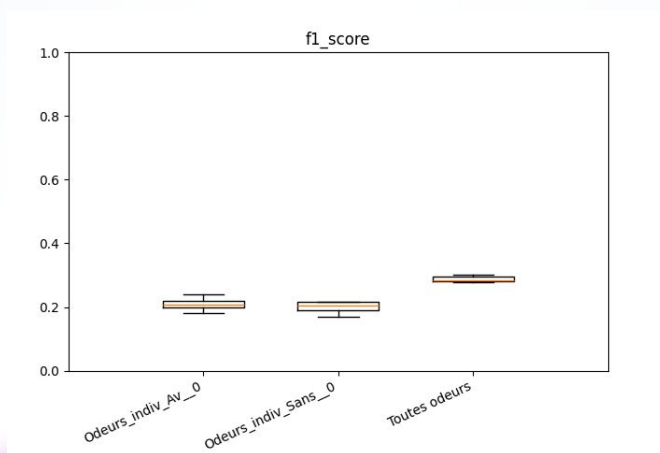
# Variation du nombre de couches de message passing

- Variation du nombre de couche de 1 à 5



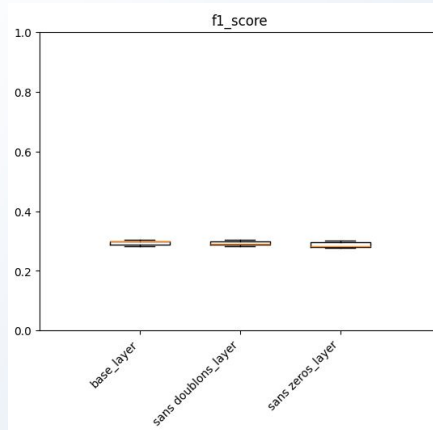
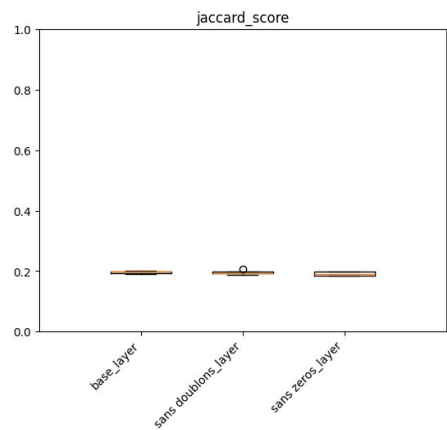
# Odeurs individuelle

- Modèles avec les odeurs individuelles et les molécules sans descripteur
- Modèles avec les odeurs individuelles et sans les molécules sans descripteur
- Modèles avec toutes les odeurs et le molécules sans descripteur



# Variation du jeu de données

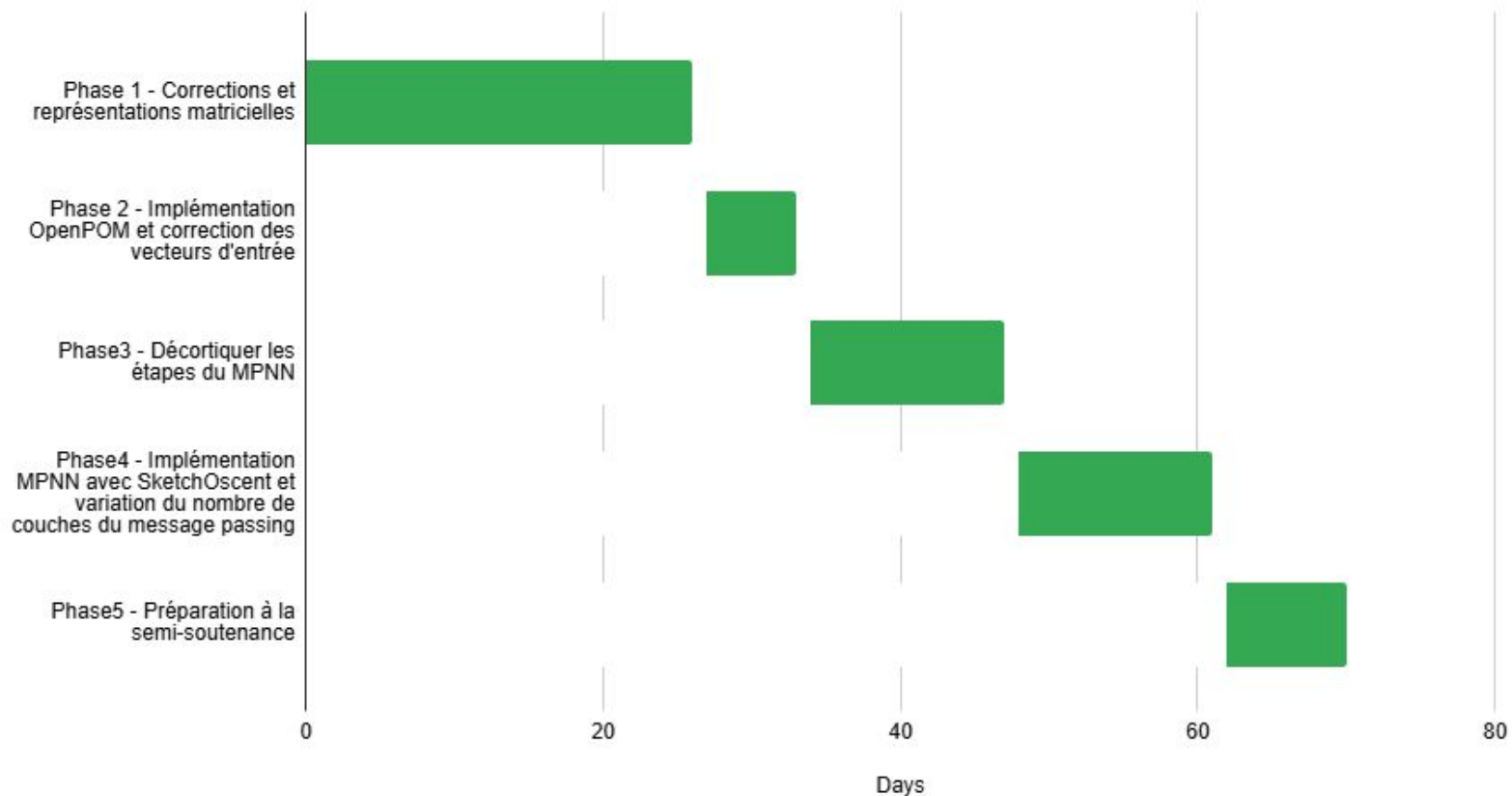
- Dataset de base
- Avec des codes SMILES doublons
- Sans les molécules sans descripteurs d'odeur



05

# Diagramme de Gantt





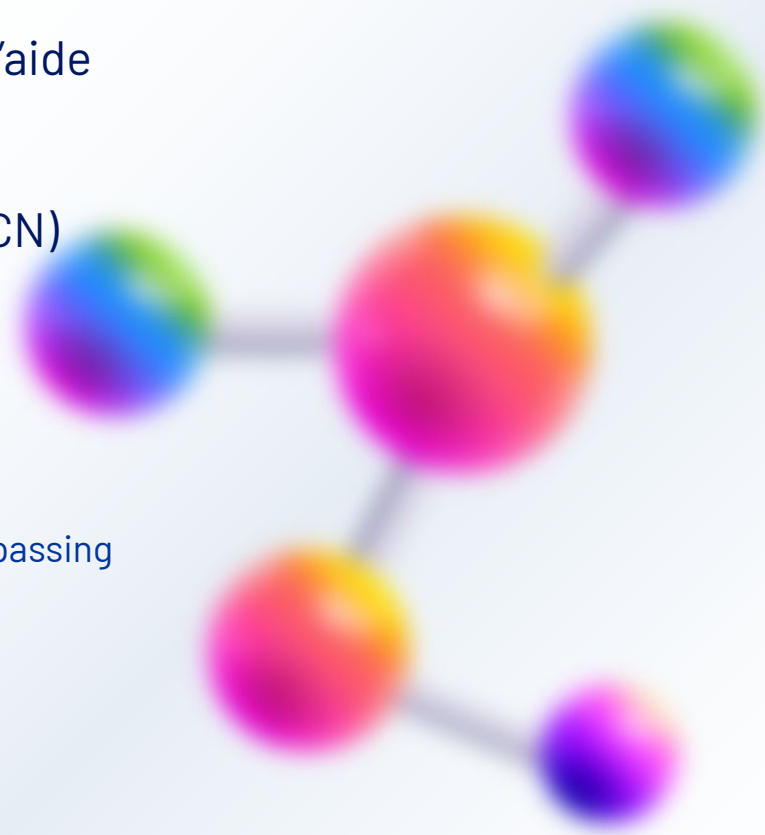
06

# Conclusion





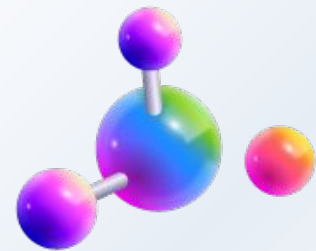
- Prédiction des descripteurs d'odeurs à l'aide du GNN
- Différents modèles du GNN ( MPNN & GCN)
- Comparaisons des modèles
- Améliorations du modèle OpenPOM
  - Modification de la fonction Readout
  - Variation du nombre de couche du message passing
  - Hiérarchisation des odeurs Sketch Oscent
  - Variation du jeu de données



07

# Bibliographie





**[BSLW19]** Brian K Lee<sup>1</sup> Richard C Gerkin Alán Aspuru-Guzik Benjamin Sanchez-Lengeling, Jennifer N Wei and Alexander B Wiltchko<sup>1r</sup>. Machine learning for scent :Learning generalizable perceptual representations of small molecules. 2019.

**[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. Wiltchko. A principal odor map unifies diverse tasks in human olfactory perception. bioRxiv, 2022.

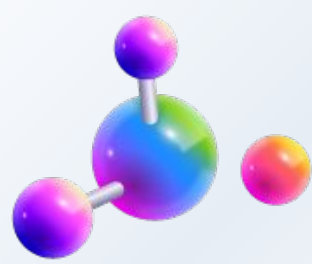
**[LS24]** Amit Barsainyan Mrityunjay Sharma Ritesh Kumar Laura Sisson, Aryan. Olfactory label prediction on aroma-chemical pairs. 2024

**[gab23]** gabikun. Pred. <https://github.com/gabikun/PRED/tree/main,2023>

**[BKSS23]** Aryan Amit Barsainyan, Ritesh Kumar, Pinaki Saha, and Michael Schmuker. Openpom - open principal odor map, 2023. <https://github.com/BioMachineLearning/openpom>.

**[Sis25]** Laura Sisson. odor\_pair, 2025. <https://github.com/odor-pair/odor-pair>.

**[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.



**[RSGH]** [https://www.researchgate.net/figure/A-typical-and-basic-architecture-and-processing-procedures-of-GNN-First-GNN-selects\\_fig2\\_352526255](https://www.researchgate.net/figure/A-typical-and-basic-architecture-and-processing-procedures-of-GNN-First-GNN-selects_fig2_352526255)

**[GNTRO]** A Gentle Introduction to Graph Neural Networks  
<https://distill.pub/2021/gnn-intro/>

07

# Annexe



# GIN(Graph Isomorphism Network)

## En annexe

la fonction de mise à jour est :

$$h_v^t = MPL^t\left((1 + \varepsilon^t) \cdot h_v^{t-1} \sum_{u \in N(u)} h_u^{t-1}\right)$$

Où epsilon est un paramètre appris dans le MLP

## En annexe

Paramètre (MPNN)	Article olfactory(MPNN)	OPENPOM(MPNN)
Entrée	?	vector(129, 6)
Nombre et taille de couche de message passing	5 couche de dimension 43	6 couche + connexion résiduelle
Phase de mise à jour	GRU-update at each layer	RELU
READOUT	Global sum pooling, soft max avec un MLP de dimension 197	Set2Set de 3 couches de 6 étapes prend en entrée: 64+64
Décodeur	Multi-headed sigmoid, 138 tasks Weighted-cross entropy loss, optimized with Adam, used learning rate decay with warm restarts, 300 epochs	FFN d'une couche de dimension 300

## En annexe

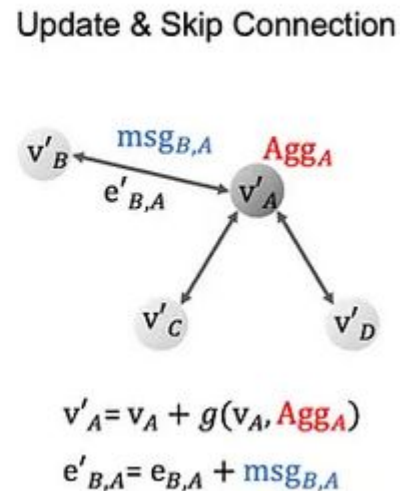
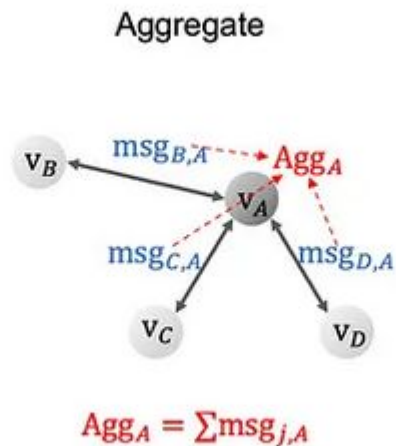
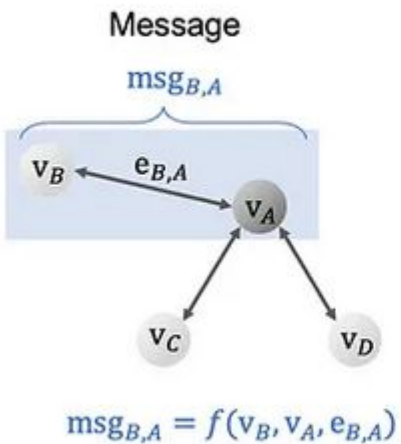
Paramètre (GCN)	Article olfactory(GCN)	Implémentation étudiant(GCN)
Entrée	?	Vecteur de dimension 15
Nombre et taille de couche de message passing	4 couches de dimension [15,20,27,36]	4 couches de dimension [15,20,27,36]
Phase de mise à jour	2 couche de dim [96, 63] avec relu, batchnorm, dropout of 0.47	SELU
READOUT	3 couche de dimension 392 avec relu, batchnorm, dropout de 0.12	GlobalAveragePooling =>vecteur(dim:98) Couche dense + softmax =>vecteur(dim: 175)
Décodeur	Multi-headed sigmoid, 138 tasks Weighted-cross entropy loss, optimized with Adam, used learning rate decay with warm restarts, 300 epochs	MLP de 3 couche [relu, relu, sigmoid] de dimension [96,63,98]



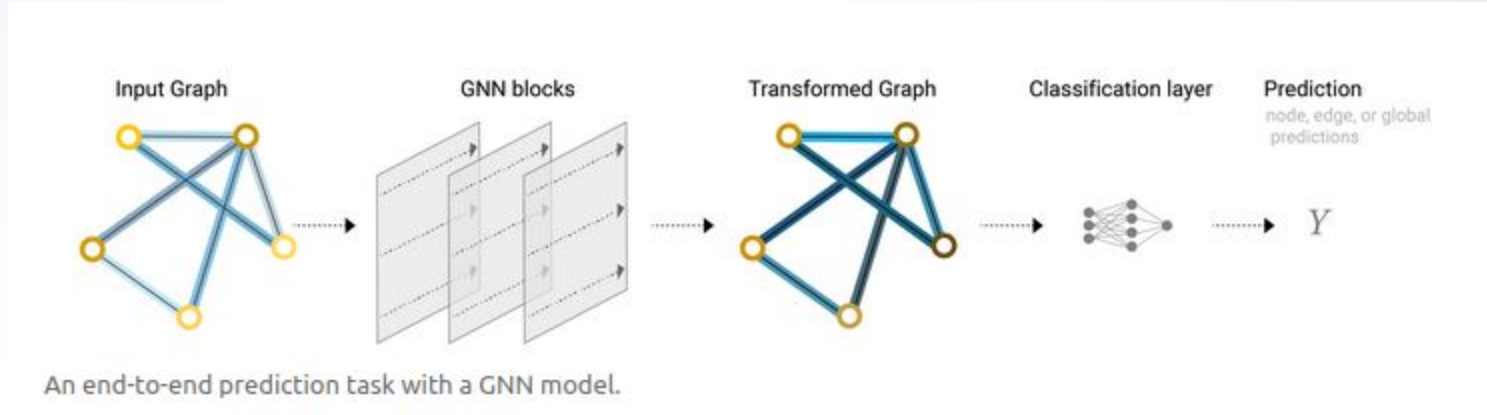
Paramètre	Article Paire de molécule	Implémentation git
Entrée	?	Vecteur atome de 9 élément
Nombre de couche de message passing	3 couche	5 couche
Phase de mise à jour	MLP de 2 couche de dimension caché 832	MLP de 2 couche de dimension 100 Fonction activation : ReLU
READOUT	Global mean + Add pool Concaténation des deux fonction et concaténation des molécule	concaténation(Global mean+ Add pool) 2 couche de MLP de taille 100. Fonction d'activation : ReLU et Rien. concaténation(2 molécules)
Décodeur	MLP de 2 couche de dimension 832 qui utilise l'entropie binaire croisée comme fonction de coût	MLP de 2 couche de dimension du nombre de classe. Et fonction d'activation ReLU et Rien

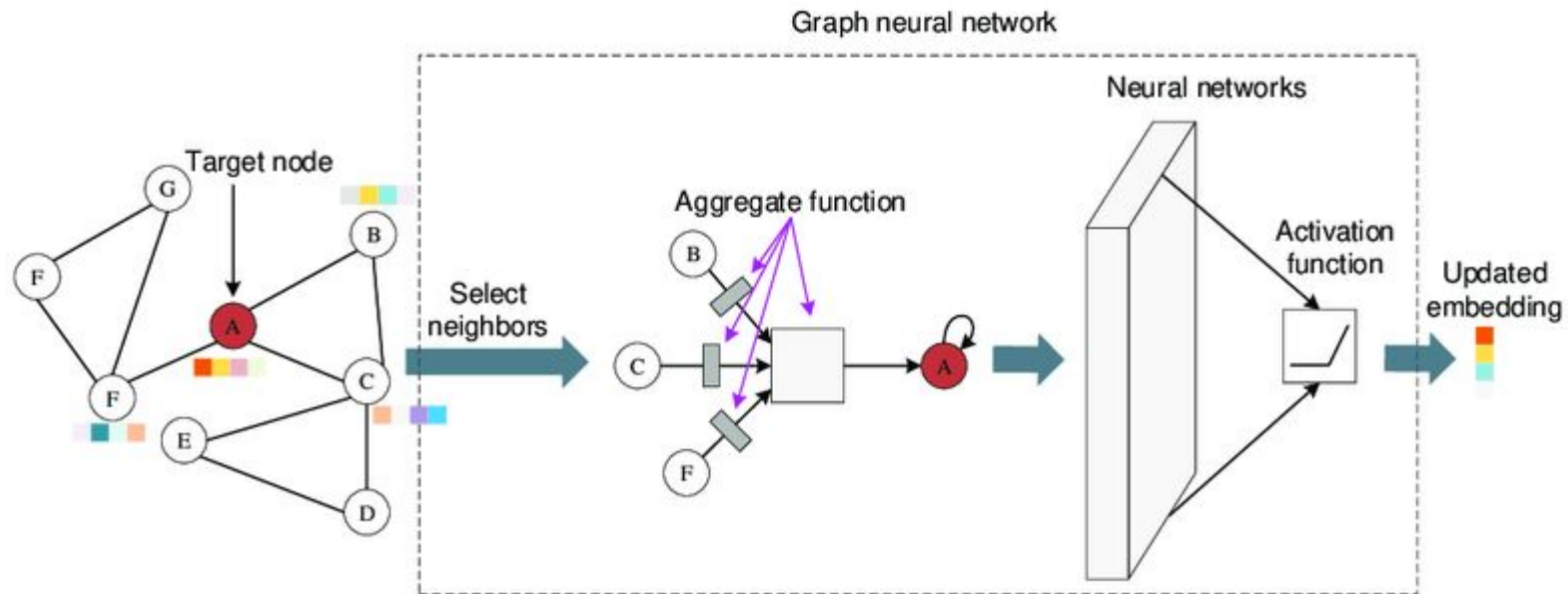
acidic	fusel	floral	vanilla
aldehydic	green	fresh	vegetable
alliaceous	herbal	fruity	waxy
amber	honey	funga	winey
animal	jammy	woody	tonka
anise	licorice	creamy	thujonic
aromatic	marine	dairy	tropical
balsamic	meaty	earthy	sulfurous
berry	medicinal	estery	spicy
bitter	melon	ethereal	sour
breathy	mentholic	fatty	solvent
brown	minty	fermented	soapy
burnt	mossy	caramellic	rummy
buttery	musk	cheesy	roasted
camphoreous	musty	chemical	powdery
cocoa	nutty	chocolate	phenolic
coconut	oily	citrus	onion
coffee	orris	coumarinic	licorice
clean	buttery	breathy	amber

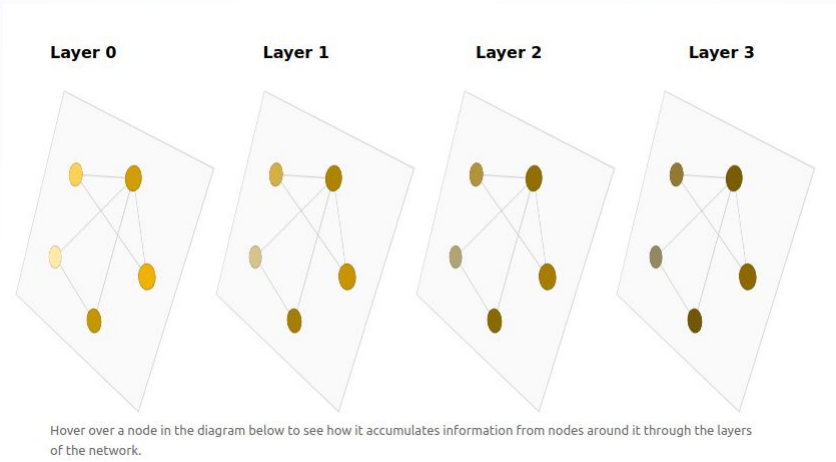
# Schema GNN



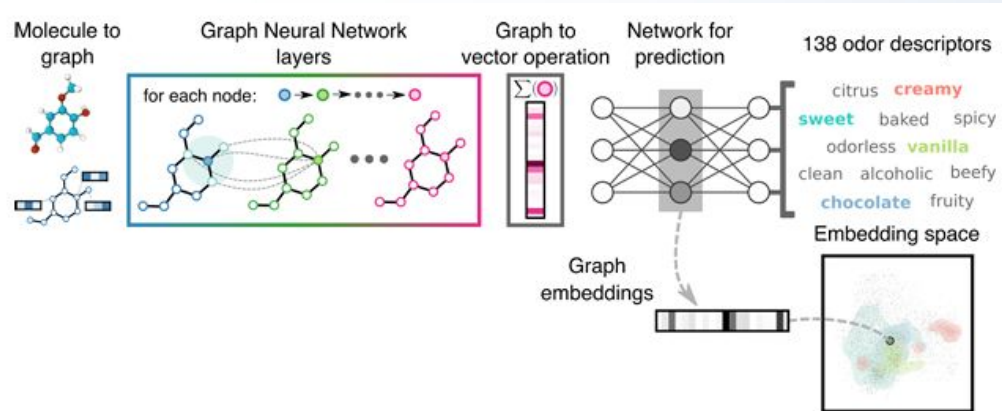
# Nouveaux schéma possible du GNN





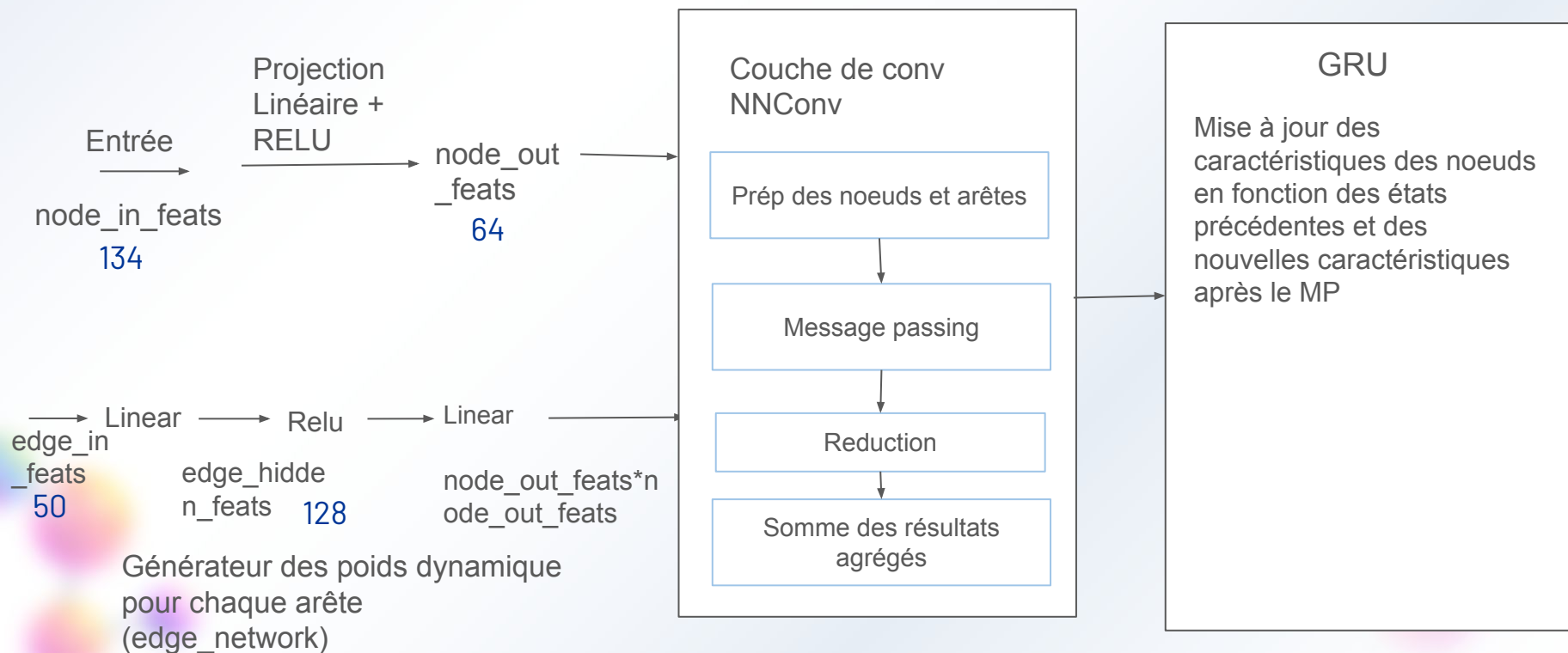


[https://www.researchgate.net/figure/A-typical-and-basic-architecture-and-processing-procedures-of-GNN-First-GNN-selects\\_fig2\\_352526255](https://www.researchgate.net/figure/A-typical-and-basic-architecture-and-processing-procedures-of-GNN-First-GNN-selects_fig2_352526255)



A supprimer

## 1 couche de message passing





# Set2Set

Fonction d'agrégation

$\mathbf{q}_t$  est initialisé à 0

$$\begin{aligned}\mathbf{q}_t &= \text{LSTM}(\mathbf{q}_{t-1}^*) \\ \alpha_{i,t} &= \text{softmax}(\mathbf{x}_i \cdot \mathbf{q}_t) \\ \mathbf{r}_t &= \sum_{i=1}^N \alpha_{i,t} \mathbf{x}_i \\ \mathbf{q}_t^* &= \mathbf{q}_t \parallel \mathbf{r}_t,\end{aligned}$$