Normastic

EXPLICABILITÉ DES DÉCISIONS D'UN GNN, APPLICATION À LA CHÉMOINFORMATIQUE

Mariana Brito Azevedo[1]

[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

mariana.brito-azevedo@unicaen.fr

GREYC Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs

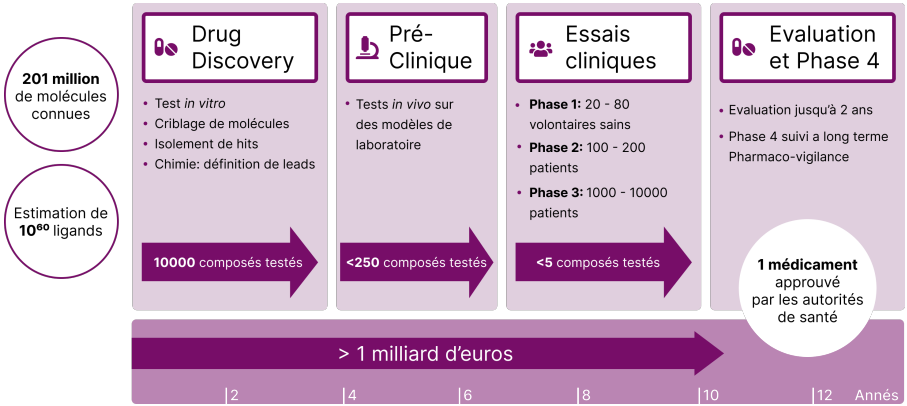
Outline

1. Contextualisation
2. Aperçu de la thèse
3. Molécules sous forme de graphes
4. GNN pour la classification de l'activité
5. Explication des résultats obtenus

Contextualisation

L'objectif et l'importance du sujet

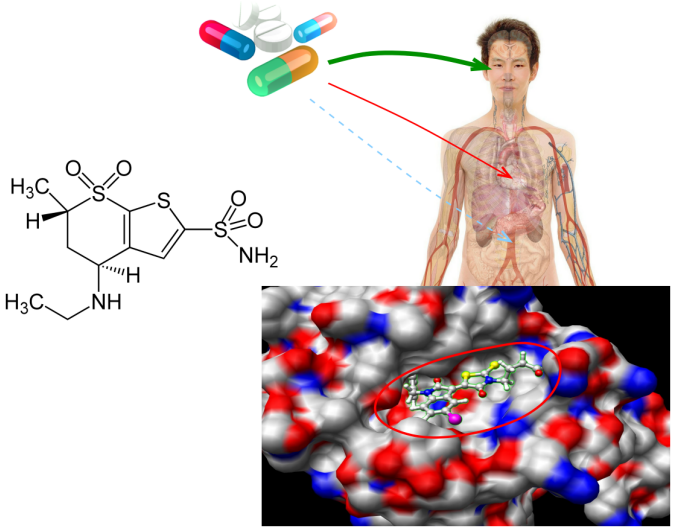
Figure: Le processus de développement d'un médicament[1]



[1]Adapté de: https://www.synergielyoncancer.fr/

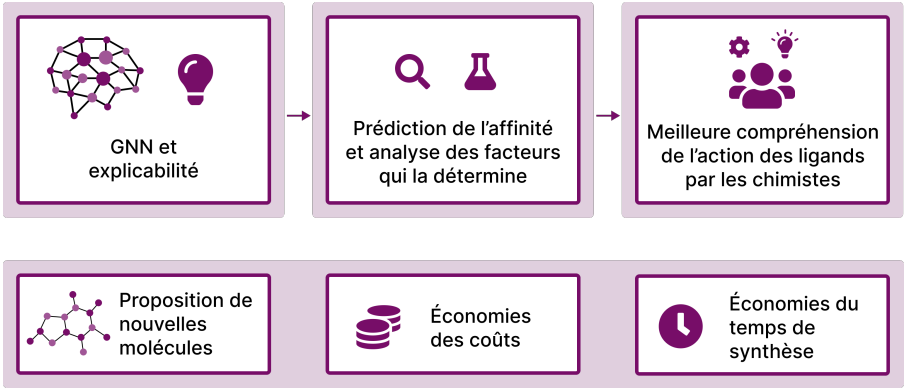
Contextualisation

L'importance et les applications



Contextualisation

L'importance et les applications



Aperçu de la thèse

Vision globale et vision détailée

VISION GLOBALE

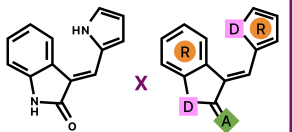
Chémoinformatique

Graph Neural Networks (GNNs)

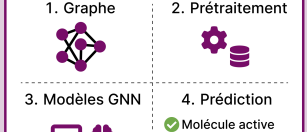
Explicabilité des décisions des GNNs

VISION DÉTAILÉE

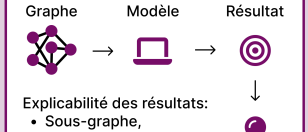
1. Molécules sous forme des graphes



2. Application des GNNs pour prédiction de l'activité

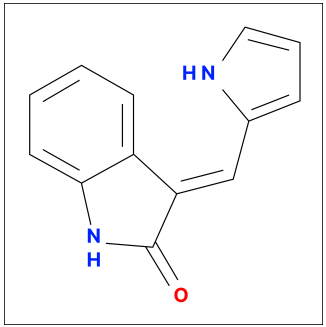


3. Explications des résultats obtenus



Molécules sous forme de graphes

Graphes moléculaires[2]

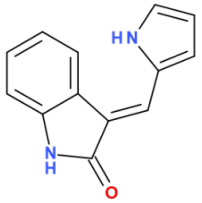


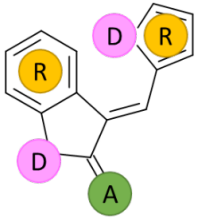
- Type d'atome;
- Degré;
- Charge;
- Hybridation;
- Hydrogènes liés;
- Aromaticité;
- Type de liaison.

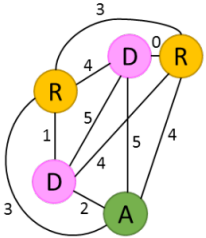
[2]Graph Neural Networks for the Prediction of Molecular Structure–Property Relationships (2023)

Molécules sous forme de graphes

CGPF: Complete Graph of Pharmacophoric Features[3]







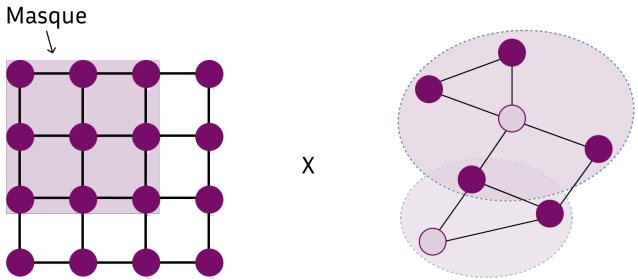
- **A**: accepteur de liaisons hydrogène;
- **D**: donneur de liaisons hydrogène;
- **N**: groupes ionisables chargés négativement;
- **P**: groupes ionisables chargés positivement;
- **H**: régions hydrophobes;
- **R**: anneaux aromatiques;

[3]The Pharmacophore Network: A Computational Method for Exploring Structure–Activity Relationships from a Large Chemical Data Set (2018)

GNN pour la classification de l'activité

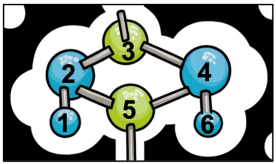
Qu'est-ce qu'un GNN ?

- Le GNN est un type de réseau de neurones utilisé pour apprendre des représentations utiles en utilisant des graphes.
- Les défis liés à l'utilisation des graphes :



GNN pour la classification de l'activité

Les matrices pour représenter un graphe[4]



Molecular graph
of six atoms

	f1	f2	f3	f4
1				
2				
3				
4				
5				
6				

Node feature
matrix

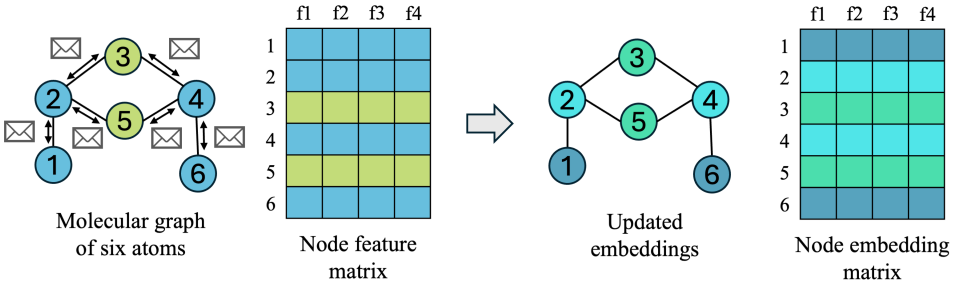
	1	2	3	4	5	6
1	1	1	0	0	0	0
2	1	1	1	0	1	0
3	0	1	1	1	0	0
4	0	0	1	1	1	1
5	0	1	0	1	1	0
6	0	0	0	1	0	1

Adjacency
matrix

[4]Source: https://syedarizvi.com/blog/2024/gnn-basics/

GNN pour la classification de l'activité

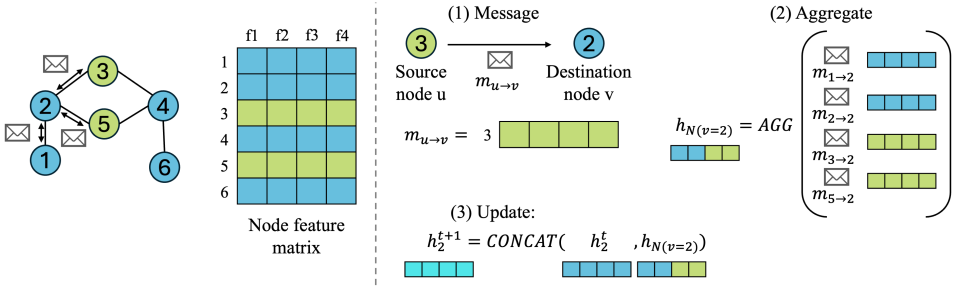
Comment fonctionne une convolution[5]



[5]Source: https://syedarizvi.com/blog/2024/gnn-basics/

GNN pour la classification de l'activité

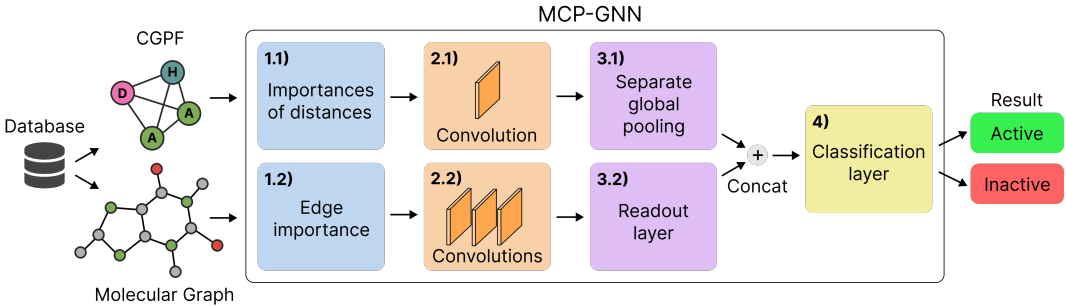
Comment fonctionne une convolution[6]



[6]Source: https://syedarizvi.com/blog/2024/gnn-basics/

GNN pour la classification de l'activité

Le modèle MCP-GNN[7]



[7] Graph Neural Network Based on Molecular and Pharmacophoric Features for Drug Design Application (2025 - Accepté)

GNN pour la classification de l'activité

L'importance des arêtes

Couche linéaire

$$y = \mathbf{x}\mathbf{W}^T$$

- L'application d'une couche linéaire pour apprendre l'importance des informations des arêtes;
- $\mathbf{x}$: vecteur d'entrée;
- $\mathbf{y}$: valeur de sortie;
- $\mathbf{W}$: matrice de poids à apprendre;

GraphConv - Pour les CGPFs

$$\mathbf{x}_i^{'} = \mathbf{W}_1\mathbf{x}_i + \mathbf{W}_2 \sum_{j \in \mathcal{N}(i)} e_{i,j} \cdot \mathbf{x}_j$$

- $\mathbf{x}_i^{'}$: nouvelle représentation du sommet i
- $\mathbf{x}_i$: vecteur de caractéristiques original du sommet i;
- $\mathbf{W}_1$: matrice de poids pour le sommet i;
- $\mathcal{N}(i)$: voisinage du sommet i;
- $\mathbf{x}_j$: vecteur de caractéristiques du voisin j;
- $\mathbf{W}_2$: matrice de poids pour les voisins du sommet i;
- $e_{i,j}$: poids de l'arête entre les sommets i et j;

GINEConv - Pour les graphes moléculaires

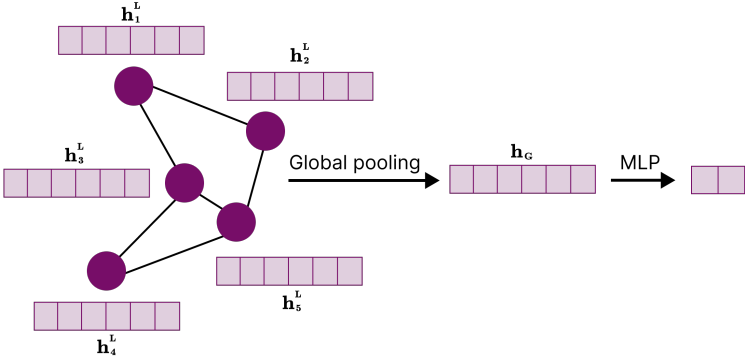
$$\mathbf{x}_i' = \text{MLP}\left((1+\epsilon)\cdot \mathbf{x}_i + \sum_{j\in\mathcal{N}(i)} \text{ReLU}(\mathbf{x}_j + \mathbf{e}_{i,j})\right)$$

- $\mathbf{x}_i'$: nouvelle représentation du sommet i
- $\mathbf{x}_i$: vecteur de caractéristiques original du sommet i;
- $\mathbf{x}_j$: vecteur de caractéristiques du voisin j;
- $\mathcal{N}(i)$: voisinage du sommet i;
- $e_{i,j}$: poids de l'arête entre les sommets i et j;
- ϵ: paramètre entraînable pour ajuster le poids de l'information propre au sommet;
- $ReLU$: fonction d'activation.

GNN pour la classification de l'activité

L'opération de global pooling

- Transformation des résultats de la dernière couche de convolution sont transformés en un vecteur de taille fixe.
- $\mathbf{h}_{\mathcal{G}} = POOL(\mathbf{h}_i^L | i \in \mathcal{V}^L)$



GNN pour la classification de l'activité

Les premiers résultats

Dataset	Model	ACC	AUC	MCC	MEAN F1
AKT1	RG-MPNN	0.875±0.016	0.935±0.012	0.748±0.032	0.873±0.016
3967 mol.	AttentiveFP	0.877±0.010	0.941±0.006	0.753±0.021	0.872±0.018
2175/1792	MCP-GNN	**0.892±0.012**	**0.949±0.009**	**0.783±0.023**	**0.890±0.012**
AURKA	RG-MPNN	0.829±0.030	0.889±0.034	0.644±0.063	0.821±0.034
3886 mol.	AttentiveFP	0.846±0.016	0.916±0.013	0.681±0.033	0.840±0.017
2290/1596	MCP-GNN	**0.860±0.016**	**0.928±0.014**	**0.710±0.034**	**0.854±0.017**
BCR-ABL	RG-MPNN	0.868±0.025	0.936±0.015	0.739±0.049	0.868±0.028
1479 mol.	AttentiveFP	0.858±0.052	0.926±0.028	0.721± 0.104	0.877±0.027
773/706	MCP-GNN	**0.883±0.024**	**0.949±0.017**	**0.761±0.050**	**0.878±0.025**
BRAF	RG-MPNN	0.895±0.020	0.934±0.033	0.709±0.060	0.877±0.023
4824 mol.	AttentiveFP	0.922±0.013	0.955±0.014	0.783±0.038	0.894±0.024
3629/1132	MCP-GNN	**0.932±0.011**	**0.960±0.013**	**0.808±0.033**	**0.901±0.016**
BTK	RG-MPNN	0.898±0.023	0.944±0.018	0.749±0.051	0.870±0.023
3640 mol.	AttentiveFP	0.902±0.018	0.947±0.015	0.759±0.043	0.876±0.024
1892/1748	MCP-GNN	**0.915±0.023**	**0.955±0.013**	**0.788±0.053**	**0.886±0.021**
CDK2	RG-MPNN	0.774±0.040	0.853±0.036	0.552±0.081	0.763±0.035
2603 mol.	AttentiveFP	0.796±0.021	0.873±0.015	0.595±0.041	0.791±0.023
1303/1300	MCP-GNN	**0.830±0.017**	**0.905±0.013**	**0.662±0.035**	**0.827±0.032**
CK1	RG-MPNN	**0.832±0.049**	0.780±0.064	0.397±0.190	0.626±0.120
807 mol.	AttentiveFP	0.827±0.039	0.833±0.055	0.439±0.111	0.699±0.055
155/632	MCP-GNN	0.825±0.036	**0.852±0.029**	**0.517±0.062**	**0.718±0.044**
MAP4K2	RG-MPNN	0.768±0.044	0.770±0.083	0.413±0.165	0.679±0.097
897 mol.	AttentiveFP	0.777±0.045	0.834±0.049	0.470±0.113	0.742±0.059
280/617	MCP-GNN	**0.791±0.063**	**0.846±0.069**	**0.511±0.148**	**0.748±0.049**
EGFR	RG-MPNN	0.854±0.014	0.923±0.013	0.709±0.028	0.854±0.015
8788 mol.	AttentiveFP	0.860±0.014	0.938±0.009	0.721±0.028	0.860±0.014
4513/4275	MCP-GNN	**0.874±0.011**	**0.946±0.006**	**0.749±0.023**	**0.874±0.011**
PIM1	RG-MPNN	0.881±0.018	0.942± 0.012	0.708±0.042	0.869±0.017
4507 mol.	AttentiveFP	0.908±0.014	0.956±0.011	0.763±0.039	0.883±0.019
3321/1186	MCP-GNN	**0.919±0.012**	**0.970±0.009**	**0.786±0.031**	**0.892±0.016**

GNN pour la classification de l'activité

Conclusion pour les GNNs

- Nous avons un classifieur avec des **bons résultats** pour faire la prédiction de l'activité d'une molécule;
- Le classifieur est **plus simples** (225.000 paramètres), mais **plus efficace** que d'autres modèles de la littérature, comme RG-MPNN (1.950.000 paramètres) et AttentiveFP (524.000 paramètres);

Explication des résultats obtenus

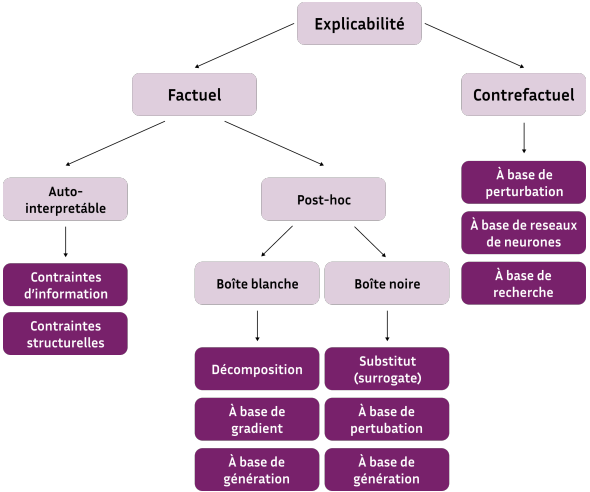
Biblio

- **L'objectif:** comprendre pourquoi une molécule est active sur une cible biologique.
- Différents algorithmes dans la littérature pour l'explicabilité des graphes:

1. **Contrefactuel**: trouver le plus petit changement dans les graphes d'entrée qui modifie la prédiction du modèle;
2. **Factuel (post-hoc)**: trouver une explication à partir de l'entrée, de la sortie et parfois des paramètres internes du modèle;
3. **Factuel (self-interpretable)**: processus d'explicabilité intrinsèque au modèle;

Explication des résultats obtenus

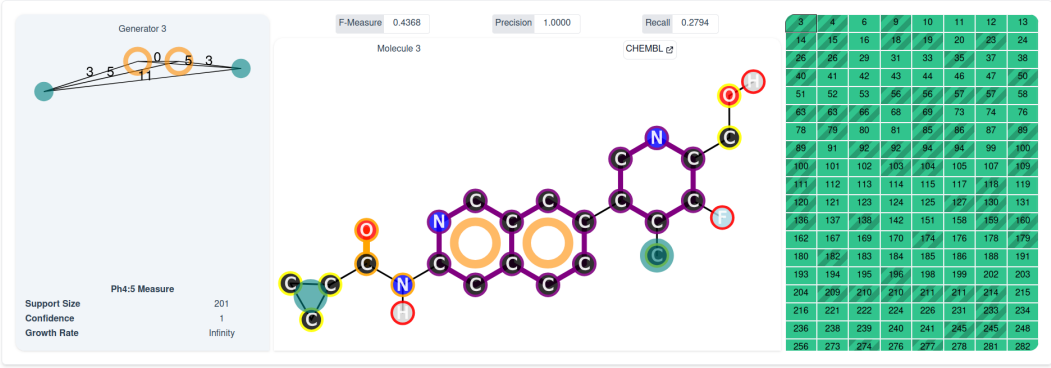
Biblio[8]



[8] A Survey on Explainability of Graph Neural Networks (2023)

Explication des résultats obtenus

Premiers résultats



The end

[thank you]

Any Questions ?

mariana.brito-azevedo@unicaen.fr