Prédiction de liens entre la Qualité de l'Air Intérieur et la santé à partir d'un Graphe de Connaissances et d'un Graph Neural Network

Elisa Drouot[1,2], Thierno Diallo[1], Gayo Diallo[2]

[1] Centre Scientifique et Technique du Bâtiment, [2] Bordeaux Population Health

1 INTRODUCTION

Qualité de l'Air Intérieur (QAI)
→ *Enjeux sanitaires et socio-économiques*

- Pollution de l'air intérieur → coût socio-économique de 19 milliards (pour 6 polluants retenus, OQEI 2014); 3.2 millions de décès par an en 2020 (OMS)

Objectif : comprendre et prédire les liens QAI et maladies avec des données hétérogènes ?

2 METHODE

Étapes du Workflow du modèle utilisé :

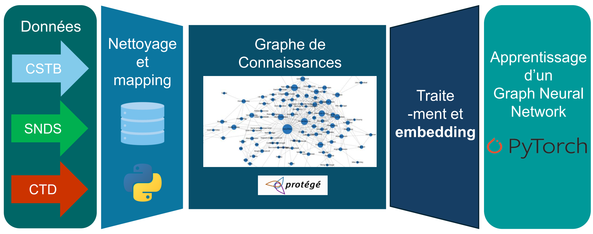


Figure 1: Schéma du workflow

3 DONNÉES ET GRAPHE DE CONNAISSANCES

Intégration de données hétérogènes issues de 3 bases de données :

- **BATENQUE :** base de données du Centre Scientifique et Technique du Bâtiment, concentration de polluants et taux de renouvellement de l'air simulé en fonction de paramètres environnementaux des résidences
- **SNDS** (Système National des Données de Santé) : données des hospitalisation agrégées → Open Data
- **CTD** (Comparative Toxicogenomics Database) : base de données toxicogénomiques qui lie les substances chimiques, gènes/phénotypes et maladies

CTD et SNDS liées par un mapping des codes maladies (CIM-10 et MeSH)

- Nombre total de nœuds : 544079
- Nombre total d'arêtes : 3542206

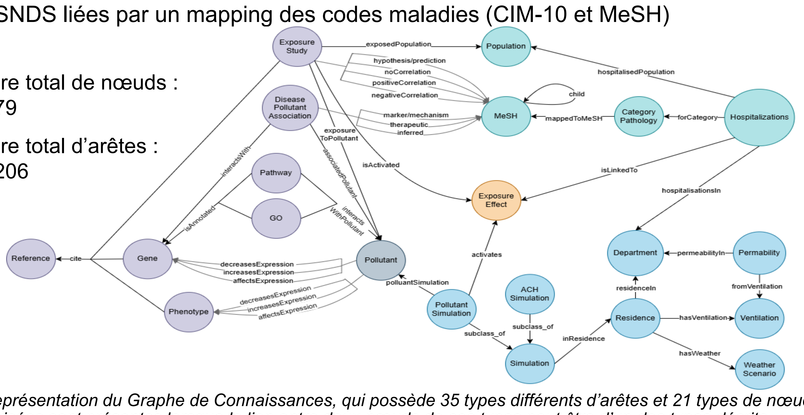
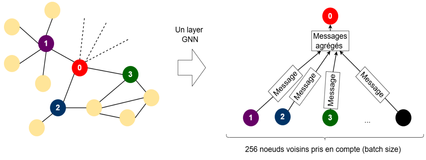


Figure 2: Représentation du Graphe de Connaissances, qui possède 35 types différents d'arêtes et 21 types de nœuds, les arêtes grisées sont présentes lorsque le lien entre deux nœuds de ces types peut être d'un des types décrits.

4 GRAPH NEURAL NETWORK

→ **Graphe hétérogène**, = plusieurs types de nœuds et plusieurs types de relations entre ces nœud. Des nœuds de différents type n'ont pas le même nombre de features et donc la même dimensionalité, donc chaque type de nœuds à ces propres vecteurs/tensors de données.



Le modèle de GNN utilisé est le GraphSAGE de la librairie **PyTorch Geometric.**
Le GNN est l'**encodeur** et le **décodeur** va prendre les embeddings de deux nœuds et calcule un score de lien (Binary Cross Entropy ici), puis ajuster les embeddings en fonction de si le score calculé est proche du celui réel (proche de 1 si l'arête existe sinon 0).

Figure 3: Représentation de l'embedding d 'un nœud par le GNN, avec une couche et une batch size de 256

5 RÉSULTATS

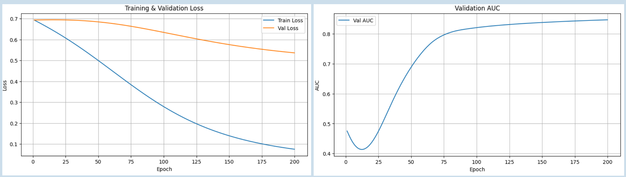


Figure 4: Courbes des loss de validation et d'entraînement ainsi que de la Area Under the Curve (AUC) sur 200 epochs.

Le modèle présenté ci-dessus a été utilisé pour étudier les arêtes de type associatedPoluttant sur 200 epochs, les valeurs obtenues pour la loss d'entraînement, de validation et l'AUC valent respectivement 0,0857, 0,5364 et 0,8473.

6 CONCLUSION ET PERSPECTIVES

L'usage d'un GNN a montré qu'il était capable de capturer la structure mais présentait une difficulté à généraliser.

Diverses perspectives sont envisagées :

RÉFÉRENCES

elisa.drouot@u-bordeaux.fr