A Knowledge Graph and Graph Neural Network Framework for Air Quality-Health Relationships

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

[1] Centre Scientifique et Technique du Bâtiment, Grenoble, France
[2] BPH Inserm 1219 Research Center, Univ. Bordeaux, F-33000, Bordeaux, France

elisa.drouot@cstb.fr

Résumé

L'air que nous respirons provient majoritairement de l'environement intérieur (estimé à 90% de notre temps). Or la qualité de l'air étant un enjeu de santé publique, de nombreuses études ont mis en évidence des liens significatifs entre exposition à différents polluants et santé humaine. Dans ce contexte, nos travaux visent à intégrer des données hétérogènes environnementales temporelles, issues des bâtiments, de la toxicogénomique et de données agrégées de santé (prévalence de pathologies, causes de décès) avec une approche à base de Graphes de Connaissances et de Graph Neural Networks (GNN) afin de caractériser et prédire les associations possibles entre polluants et pathologies. Les résultats préliminaires ont montré qu'il était possible d'obtenir une perte d'entraînement (train loss) de 0.0942 et une AUC (Area Under the Curve) sur les données de validation de 0.8473.

Mots-clés

Qualité de l'Air Intérieur, Santé environnementale, Prédiction de liens, Graphes de Connaissances, GNN

Abstract

The air we breathe comes predominantly from indoor environments, where we spend an estimated 90% of our time. As indoor air quality represents a major public health concern, numerous studies have established significant links between exposure to various pollutants and human health. In this context, our research aims to integrate heterogeneous temporal environmental, from buildings, toxicogenomics, and aggregated health records (such as disease prevalence and causes of death). By employing an approach based on Knowledge Graphs and Graph Neural Networks (GNN), we seek to characterize and predict potential associations between pollutants and pathologies. Preliminary results demonstrate the ability to achieve a training loss of 0.0942 and a validation AUC (Area Under the Curve) of 0.8473.

Keywords

Indoor Air quality, Environmental Health, Link Prediction, Knowledge Graph, GNN

1 Introduction

According to the State of Global Air (SoGA) report published by the Health Effects Institute (HEI), air pollution is the second leading cause of death worldwide, responsible for 8.1 million deaths globally in 2021. More precisely, 38% of these deaths are attributed to household air pollution [6]. Indoor pollutant concentrations depend heavily on occupant behavior and building characteristics, such as construction materials and ventilation systems [9]. A major challenge in linking indoor air pollution to health outcomes lies in the vast diversity of both pollutants and their associated diseases. Consequently, most studies tend to focus either on a single disease, such as acute lower respiratory infections [11], or on a single pollutant [3].

As a result, comprehensively linking building parameters and indoor pollutant concentrations to specific pathologies remains a complex endeavor. The availability of large, heterogeneous, and evolving datasets makes computational approaches highly relevant for addressing this gap. For instance, machine learning models have been employed to predict associations between respiratory diseases, air pollution, and climatic factors [8]. In the domain of health-related link prediction, Knowledge Graphs [5] have emerged as a commonly used and powerful tool ; for example, they have been successfully utilized to map connections between pesticides and diseases [16] or outdoor pollutants and diseases [7]. Applying Graph Neural Networks (GNNs) to these knowledge structures represents a highly promising avenue for discovering complex associations. Although GNN-based methods have already been introduced in the field of air pollution, their applications have primarily been restricted to air quality forecasting rather than health outcome prediction [14].

2 Methodology

2.1 Selecting and Cross-linking Relevant Data Sources

2.1.1 Data Sources Used

To effectively capture and model the complex relationships between pollutant exposure and health outcomes, three distinct databases were integrated into our study :

— **Clinical Data from the French National Health Data System (SNDS) :** Sourced from the open-access portal, this aggregated dataset contains patient counts, reference populations, and prevalence rates for 79 distinct pathological categories. The data is rigorously stratified by biological sex, five-year age cohorts, and administrative departments.
— **The Comparative Toxicogenomics Database (CTD) :** A robust resource linking chemical exposures to biological outcomes [2]. It provides both direct and inferred chemical-disease associations, with the latter being deduced from complex interactions involving genes, phenotypes, and biological pathways.
— **Residential Exposure and Building Characteristics :** Detailed building-specific data and simulated indoor residential exposures were provided by the French Scientific and Technical Centre for Building (CSTB).

2.1.2 Mapping

A primary technical challenge involved harmonizing the health nomenclature systems between the CTD and the SNDS datasets through an integration process [12]. The CTD database relies on the Medical Subject Headings (MeSH) thesaurus—a hierarchically organized controlled vocabulary maintained by the U.S. National Library of Medicine—and occasionally employs the Online Mendelian Inheritance in Man (OMIM) coding system. In contrast, the SNDS classifies health outcomes into 79 distinct categories defined by ICD-10 (International Classification of Diseases, version 10), CCAM (Common Classification of Medical Acts), and GHM (Homogeneous Patient Groups) codes. To perform the cross-dataset mapping, the Unified Medical Language System (UMLS) Metathesaurus browser [1]facilitated the alignment of ICD-10 codes with their corresponding MeSH terms. To preserve the clinical semantic context, the hierarchical tree structure of MeSH was maintained, ensuring that each disease node remained accurately linked to its parent concept.

2.2 Knowledge Graph Construction

A Knowledge Graph (KG) [5] was designed and developed to integrate the three aforementioned datasets. Within this architecture, nodes are assigned explicit semantic classes—such as Pollutant, Gene, Disease, and Housing Unit—thereby preserving the inherent heterogeneity of the underlying data. The edges represent directed, typed relationships that encode specific biological or environmental mechanisms, rather than mere co-occurrences. A defining characteristic of this model is that relationships between identical node types can vary significantly ; for instance, a pollutant may exhibit an up-regulation (increased expression) relationship with a specific gene, while also potentially exhibiting a down-regulation (decreased expression) association under different conditions.

A central challenge in this framework is establishing a biologically and environmentally plausible link between simulated indoor pollutant concentrations and observed hospitalization trends. To bridge this gap, a mediator node designated as "Exposure Effect" was introduced. This node functions as a logical trigger : an association is only activated when the simulated residential pollutant concentration exceeds the critical threshold reported in the corresponding toxicological literature.

An additional complexity involves the temporal constraints governing pollutant exposure and subsequent hospital admission. To resolve this limitation, a directed relationship was established between the Exposure Effect node and Hospitalization events, strictly conditioned on the exposure chronologically preceding the hospital admission. To better understand temporal constraints, one perspective is to use a temporal graph neural network [10].

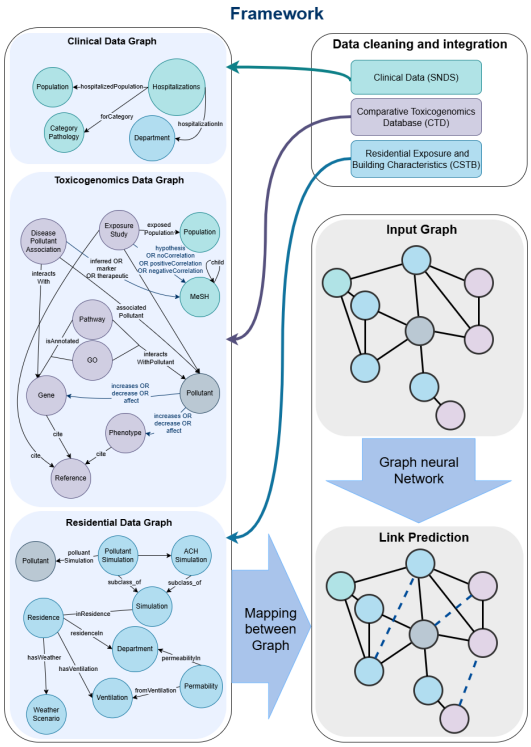

FIGURE 1 – Framework for link prediction integrating different data sources

2.3 Learning Over the Knowledge Graph

Given that the developed knowledge graph contains multiple node and edge types encoding rich semantic information, it is formally modeled as a heterogeneous graph. It is then required to translate this graph into a set of dense vectors that could then be used for the learning process. Therefore, node embeddings are first initialized as 64-dimensional vectors, without incorporating external node features. A single layer of a Heterogeneous Graph Neural Network (HeteroGNN) is then applied to learn node representations. This layer is implemented via PyTorch Geometric (PyG), an advanced graph learning library built upon the PyTorch framework. The network utilizes relation-specific message passing facilitated by a Hetero-

Conv layer. This architecture allows each distinct relation type within the heterogeneous graph to be processed independently, subsequently aggregating the extracted features to generate enriched node representations [13]. The convolutional operations are grounded in the GraphSAGE framework, which iteratively updates node embeddings by aggregating features from local neighborhoods, thereby effectively leveraging the structural topology of the graph for robust representation learning [4]. Negative sampling [15] is performed uniformly at random, which is suboptimal for Knowledge Graphs. This opens avenues for improvement, such as hard negative sampling techniques.For evaluation, edge scores are computed via a dot product between source and destination node embeddings, and optimized using binary cross-entropy with logits against the true labels.

3 Preliminary Results

The constructed knowledge graph comprises 3,462,139 edge instances and 536,924 nodes.
The GNN model was trained for 200 epochs, focusing exclusively on a specific edge type : the *associatedPollutant* relationship, which connects *DiseasePollutantAssociation* nodes to *Pollutant* nodes. This relationship was the only one using during training due to computational resources and time constraints.
The edges are split into training, validation, and test sets with proportions of 80%, 10%, and 10%, respectively. The model converged to a low training loss of 0.0857, indicating an effective extraction and learning of structural patterns within the training data. The validation loss stabilized at 0.5364, showing that there is gap between training and validation performance that may indicate a lack of ability to generalize. Furthermore, the model achieved a validation AUC (Area Under the Receiver Operating Characteristic Curve) of 0.8473 . This metric demonstrates a strong discriminative capability in binary edge classification, effectively distinguishing between positive and negative edges (i.e., the presence or absence of a valid link within the knowledge graph).

FIGURE 2 – Training and validation loss performance of the graph neural network model across epochs

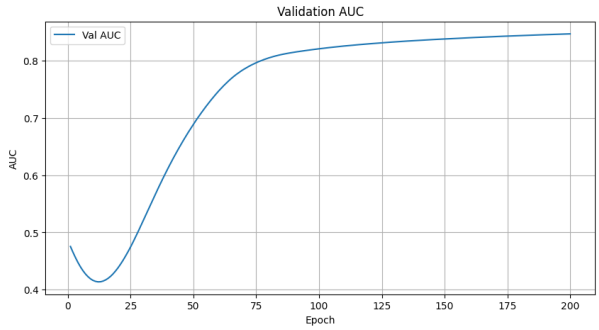

FIGURE 3 – Validation Area Under the Curve across epochs

4 Conclusion and Future Directions

Through the integration of heterogeneous data from environmental, toxicogenic, and clinical sources into a unified Knowledge Graph, the current work enables a comprehensive exploration of the relationships between these interlinked domains. The data used present certain limitations, as they do not provide a complete representation. For instance, the building data only include residential buildings. However, the proposed method, based on a GNN approach, enables effective link prediction within the knowledge graph.

Preliminary results demonstrate an ability to identify existing links, as shown by a validation AUC of 0.8473. However, the comparison between the training loss (0.0857) and validation loss (0.5364) suggests that while the GNN successfully learns structural information, its capacity for generalization remains limited. To improve the generalization capability of the GNN model, several strategies can be employed, including dropout techniques such as DropEdge, which randomly removes edges instead of nodes, as well as regularization methods such as weight decay.

Future work will focus on improving both the embeddings and the GNN architecture. Additionally, the environmental dataset will be expanded to include thermal comfort parameters. Another avenue of interest is the creation of a baseline, by comparing GNN to other models such as regression models, other Knowledge Graph embedding (i.e., TransE, RotatE), or other Neural Networks models (like R-CGN and Edge Transformers).

Références

[1] Olivier Bodenreider. The unified medical language system (umls) : integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue) :267–270, 2004.

[2] Allan Peter Davis, Caroline G. Murphy, Cynthia A. Saraceni-Richards, Michael C. Rosenstein, Thomas C. Wiegers, and Carolyn J. Mattingly. Comparative toxicogenomics database : a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research*, 37(Database issue) :D786–D792, Jan 2009.

[3] A. Garcia, E. Santa-Helena, A. De Falco, J. de Paula Ribeiro, A. Gioda, and C. R. Gioda. Toxicological effects of fine particulate matter (pm2.5) : Health risks and associated systemic injuries—systematic review. *Water, air, and soil Pollution*, 234(6), May 2023.

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc.

[5] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021.

[6] Health Effects Institute. State of global air 2024, 2024.

[7] Nareesa Karmali, Abdougafarou Mamam, and Gayo Diallo. Outdoor air quality andnbsp;health impact : Thenbsp;panorama knowledge graph based approach. In *Computational Collective Intelligence : 17th International Conference, ICCCI 2025, Ho Chi Minh City, Vietnam, November 12–15, 2025, Proceedings, Part II*, page 32–46, Berlin, Heidelberg, 2025. Springer-Verlag.

[8] Y. Ku, S. B. Kwon, J.-H. Yoon, S.-K. Mun, and M. Chang. Machine learning models for predicting the occurrence of respiratory diseases using climatic and air-pollution factors. *Clinical and Experimental Otorhinolaryngology*, 15(2) :168–176, May 2022.

[9] P. Kumar, A.B. Singh, T. Arora, S. Singh, and R. Singh. Critical review on emerging health effects associated with the indoor air quality and its sustainable management. *Science of The Total Environment*, 872 :162163, May 2023.

[10] Antonio Longa, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, Franco Scarselli, and Andrea Passerini. Graph neural networks for temporal graphs : State of the art, open challenges, and opportunities, 2023.

[11] H. Nair, D. J. Nokes, B. D. Gessner, M. Dherani, S. A. Madhi, R. J. Singleton, K. L. O'Brien, A. Roca, P. F. Wright, N. Bruce, A. Chandran, E. Theodoratou, A. Sutanto, E. R. Sedyaningsih, M. Ngama, P. K. Munywoki, C. Kartasasmita, E. A. Simões, I. Rudan, M. W. Weber, and H. Campbell. Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children : a systematic review and meta-analysis. *Lancet*, 375(9725) :1545–1555, May 2010.

[12] Inès Osman, Sadok Ben Yahia, and Gayo Diallo. Ontology integration : Approaches and challenging issues. *Information Fusion*, 71 :38–63, 2021.

[13] R. Ragesh, S. Sellamanickam, A. Iyer, R. Bairi, and V. Lingam. Hetegcn : Heterogeneous graph convolutional networks for text classification. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 860–868, New York, NY, USA, 2021. Association for Computing Machinery.

[14] J. Xu, S. Wang, N. Ying, X. Xiao, J. Zhang, J. Zhiling, Y. Cheng, and G. Zhang. Dynamic graph neural network with adaptive edge attributes for air quality prediction : A case study in china. *Heliyon*, 9 :e17746, 07 2023.

[15] Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, and Jie Tang. Understanding negative sampling in graph representation learning, 2020.

[16] D. Zhang, X. Wu, P. Chen, Q. Wang, Y. Li, C. Zhai, and G. Hao. Knowledge-driven pesticide repurposing via link prediction with pesticide graph embedding, 01 2025.