# Offre n°2025-08991

# PhD Position F/M What Do GNNs Dream of ? An Interpretability Method Based on Pattern Extraction

**Type de contrat :** Fixed-term contract

**Niveau de diplôme exigé :** Graduate degree or equivalent

**Fonction :** PhD Position

## A propos du centre ou de la direction fonctionnelle

The Inria Centre at Rennes University is one of Inria's eight centres and has more than thirty research teams. The Inria Centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

## Contexte et atouts du poste

**Within the framework of a partnership (you can choose between)**

- public with French National Research Agency (ANR)

## Mission confiée

**Assignments :**
With the help of ****, the recruited person will be taken to ****.

### Context

Graph Neural Networks (GNNs) [1] have become increasingly popular in recent years due to their ability to process graph-structured data (e.g., social networks, molecules, knowledge graphs, etc.). These models have set the new state-of-the-art in tasks such as link prediction and graph classification, achieving impressive results compared to previous approaches. However, like other neural network-based approaches, GNNs suffer from a lack of interpretability: it is nearly impossible for a human expert to understand the reasoning behind

a GNN's decision.

While numerous explainability (XAI) methods have been proposed for traditional neural networks (e.g., those processing text or images) [2,3], very few works have focused on the case of graph data. In [4], a GNN explainability method based on pattern mining [5] was proposed. This method is unique in that it directly leverages the components activated during the GNN's decision-making process, extracting patterns called ``activation rules''. These activation rules are then linked to the input graph data, enabling the generation of explanations in the form of subgraphs.

This preliminary method, however, has several limitations. First, the activation rules are only extracted for a single layer of the GNN, limiting their expressiveness. Additionally, non-activations are not taken into account, even though they can be crucial for explaining a decision. Another key limitation is the overwhelming number of activation rules generated, due to the combinatorial nature of the extraction process. A quality measure is therefore needed to select a relevant subset of rules. The current method uses a statistical-based measure on the rule set, but it fails to link the rules to the corresponding parts of the input graphs.

## Thesis Objectives

The goal of this PhD thesis is to provide human users with rich, precise, and understandable explanations of GNN decisions. The first work will be on improving the expressiveness of the activation rules extracted from the GNN. The aim is to develop a method that extracts activation rules from components across multiple GNN layers, taking into account both activations and non-activations (e.g., negative patterns [6]).

As this leads to an extremely large search space of potential rules, the proposed approach must return a small, relevant subset of rules that best explain the GNN's decision. To achieve this, techniques based on Information Theory—particularly the Minimum Description Length (MDL) principle [7]—will be explored.

A final theoretical contribution will be the development of methods to ``translate'' these expressive activation rules into the input graph space, in order to provide comprehensible explanations grounded in the original data. Targeted applications include molecular graphs and semantic web knowledge graphs. A promising direction for achieving a robust mapping between activation rules and input graphs is to leverage the knowledge stored in Large Language Models (LLMs) to capture domain-specific semantics of the graph data.

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini. The Graph Neural Network Model. In IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80 (2009).

[2] M. Tu?lio Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144

[3] Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4765-4774

[4] L. Veyrin-Forrer, A. Kamal, S. Duffner, M. Plantevit, C. Robardet. On GNN explainability with activation rules. Data Min Knowl Disc (2022).

[5] C. Aggarwal, J. Han. Frequent Pattern Mining. Springer, Cham (2014).

[6] T. Guyet, R. Quiniou. NegPSpan: efficient extraction of negative sequential patterns with embedding constraints. Data Min. Knowl. Discov. 34(2): 563-609 (2020)

[7] P. Grünwald. The Minimum Description Length Principle. The MIT Press (2007)

# Principales activités

Main activities (5 maximum)

- Develop programs
- Design experimental platforms
- Write papers
- Test, change up until validation
- Distribute the work via publications and talks
- Present the works' progress to partners

## Compétences

The candidate should have a strong interest in machine learning in general, with a particular focus on neural networks, statistics, algorithms, and programming.

## Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training

## Rémunération

Monthly gross salary amounting to 2200 euros

## Informations générales

- **Thème/Domaine :** Data and Knowledge Representation and Processing Information system (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2025-10-01
- **Durée de contrat :** 3 years
- **Date limite pour postuler :** 2025-08-31

## Contacts

- **Équipe Inria :** LACODAM
- **Directeur de thèse :**
  Cellier Peggy / Peggy.Cellier@irisa.fr

# A propos d'Inria

Inria est l’institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

# L'essentiel pour réussir

**Attention**: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

# Consignes pour postuler

Please submit online : your resume, cover letter and letters of recommendation eventually

**Sécurité défense :**
Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

**Politique de recrutement :**
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.