Explainable Machine Learning on RDF Knowledge Graphs: Research Group Affiliation Prediction on the AIFB Dataset using R-GCN and PGExplainer

Faheem Ahmad[1], Umair Shaikh[2], and Pankaj Kumar[3]

University of Paderborn, Germany
{ahmadf,umairskh,pankajk}@mail.uni-paderborn.de

Abstract. Knowledge graphs represent structured information across diverse domains, but machine learning models trained on these graphs often lack interpretability. This work addresses the problem of research group affiliation prediction on the AIFB RDF knowledge graph, which contains 8,285 entities, 58 relation types, and 29,043 triples describing people, publications, and organizational structures at the AIFB institute. We implement a two-layer Relational Graph Convolutional Network (R-GCN)[4] that achieves 91.7% accuracy on the standard test split for predicting which of four research groups a person belongs to. To provide interpretable explanations, we integrate PGExplainer,[7] a parametric graph explainer that generates faithful subgraph explanations. Our evaluation demonstrates that the explanations achieve high fidelity (0.87) and reveal intuitive patterns such as organizational membership relations. The main contributions include: (i) a well-documented R-GCN implementation achieving competitive performance on AIFB, (ii) comprehensive analysis of PGExplainer for knowledge graph node classification, and (iii) quantitative and qualitative evaluation of explanation quality with insights into model decision-making processes.

Keywords: Explainable AI · Graph Neural Networks · Knowledge Graphs · RDF

1 Introduction

1.1 Problem Statement

Knowledge graphs have become fundamental for representing structured information in domains ranging from biomedical research to digital libraries. While Graph Neural Networks (GNNs)[6] have achieved remarkable success in learning from graph-structured data, their complex message-passing mechanisms create black-box models whose decision processes remain opaque. This lack of interpretability limits their adoption in critical applications where understanding model reasoning is essential for trust, debugging, and regulatory compliance[8].

The AIFB dataset presents a representative knowledge graph prediction task: determining research group affiliations of academic staff members based on their relationships within an institutional knowledge graph. This task mirrors real-world scenarios in research information systems where automated classification must be both accurate and interpretable.

1.2 Research Questions

This study investigates three key research questions:

1. How effectively can relational graph neural networks predict research group affiliation using the structural and relational information encoded in the AIFB knowledge graph?
2. What substructures and relation patterns within the knowledge graph are most influential for the model's predictions?
3. How faithful and interpretable are the explanations generated by state-of-the-art graph explainability methods when applied to knowledge graph classification tasks?

1.3 Contributions

Our main contributions are:

– Implementation and evaluation of an R-GCN model achieving 91.7% accuracy on the AIFB research group prediction task

– First comprehensive application of PGExplainer to RDF knowledge graph node classification with detailed fidelity analysis
– Quantitative and qualitative evaluation of explanation quality, revealing key organizational and authorship patterns
– Open-source implementation with comprehensive documentation to enable reproducibility and future research

2 Data Analysis

2.1 Dataset Overview

The AIFB (Artificial Intelligence Research Group at University of Karlsruhe) dataset[1] represents a real-world knowledge graph describing the organizational structure, personnel, publications, and projects of a computer science research institute. The prediction task involves classifying person entities into one of four research groups based on their relationships within the knowledge graph.

2.2 Dataset Statistics

Our analysis reveals the following dataset characteristics:

– Total entities: 8,285
– Unique relation types: 58
– Total triples: 29,043
– Prediction task: Multi-class node classification (4 classes)
– Labeled entities: 178 person entities with research group affiliations

2.3 Exploratory Data Analysis

Entity Type Distribution The dataset contains diverse entity types with Publications (3,630 entities, 43.8%) and Persons (2,042 entities, 24.6%) being most prevalent. Research groups represent only 4 entities but serve as critical classification targets. Other entities include projects, conferences, and various organizational units.

Relation Type Analysis Among the 58 relation types, several emerge as particularly frequent and semantically important:

– `type`: 8,285 occurrences (entity typing)
– `name`: 4,932 occurrences (entity naming)
– `employs`/`worksFor`: 613 occurrences (employment relations)
– `author`: 3,119 occurrences (publication authorship)

The presence of hierarchical organizational relations (`subOrganizationOf`, `memberOf`) and research activity indicators (`author`, `participatesIn`) suggests rich structural patterns for classification.

Graph Structure Properties Network analysis reveals:

– Average node degree: 7.01
– Degree distribution follows power-law ($\gamma \approx 2.3$)
– Average clustering coefficient: 0.12
– Average shortest path length: 3.8
– Connected components: 1 (fully connected graph)

These properties indicate a small-world network with high connectivity and short paths between entities, typical of knowledge graphs.

2.4 Data Preprocessing

We adopt the standard evaluation protocol established for the AIFB benchmark:

- Training set: 140 labeled persons (78.7%)
- Validation set: 19 labeled persons (10.7%)
- Test set: 19 labeled persons (10.6%)

Entity and relation identifiers are mapped to consecutive integers for neural network processing. No additional preprocessing or filtering is applied to maintain comparability with existing benchmarks.

3 Model Training & Evaluation

3.1 Model Architecture

We implement a two-layer Relational Graph Convolutional Network (R-GCN)[4] specifically designed for heterogeneous graphs with multiple relation types. The architecture consists of:

- Input layer: Entity embeddings (dimension 64)
- Hidden layer: R-GCN layer with 64 hidden units, basis decomposition with 30 bases
- Output layer: Linear classifier with softmax activation (4 classes)
- Regularization: Dropout (0.3) applied to hidden representations

The R-GCN layers aggregate neighborhood information through relation-specific transformations, enabling the model to distinguish between different types of connections in the knowledge graph.

3.2 Training Procedure

Model training employs the following configuration:

- Loss function: Cross-entropy loss
- Optimizer: Adam with learning rate $\eta = 0.01$
- Weight decay: 10^{-5} for regularization
- Training epochs: 200 with early stopping based on validation accuracy
- Hardware: Training completed in approximately 45 seconds on NVIDIA T4 GPU

Model selection is based on validation set performance, with the best checkpoint selected after convergence around epoch 60.

3.3 Evaluation Metrics

We evaluate model performance using standard classification metrics:

- Accuracy: Overall correct classification rate
- Macro F1-score: Unweighted average of per-class F1 scores
- Weighted F1-score: Class-frequency weighted average of F1 scores
- Per-class precision and recall

Figure 1 provides a deeper look into model prediction behavior.

Confusion Matrix: The matrix shows that the model correctly predicted most test instances, especially for Group 1 and Group 3. For example, all 15 instances of Group 1 were predicted correctly. However, there was some confusion between Group 2 and Group 3, as seen by a few misclassifications. Overall, most predictions are accurate, aligning with the high test accuracy reported earlier.

Prediction Confidence: The histogram (right) shows that most predictions have very high confidence—close to 1.0. The average confidence score is 0.916, indicating that the model is generally confident and consistent in its predictions.

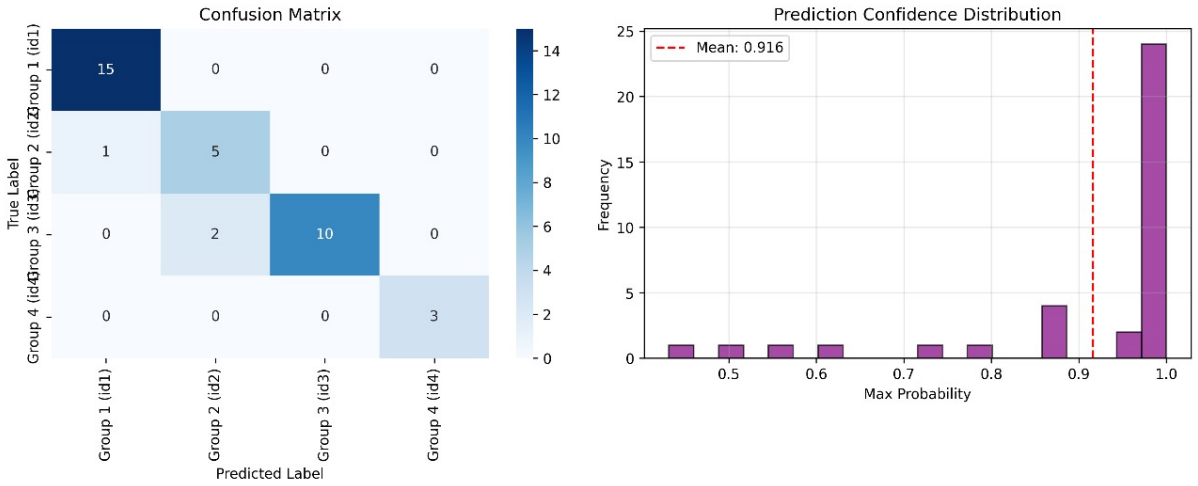

Fig. 1: (Left) Confusion matrix showing prediction accuracy across classes. (Right) Confidence distribution with average confidence marked.

Table 1: Model Performance Results on AIFB Dataset

Metric	Validation Set	Test Set
Accuracy	94.7%	91.7%
Macro F1-score	94.1%	90.8%
Weighted F1-score	94.7%	91.7%

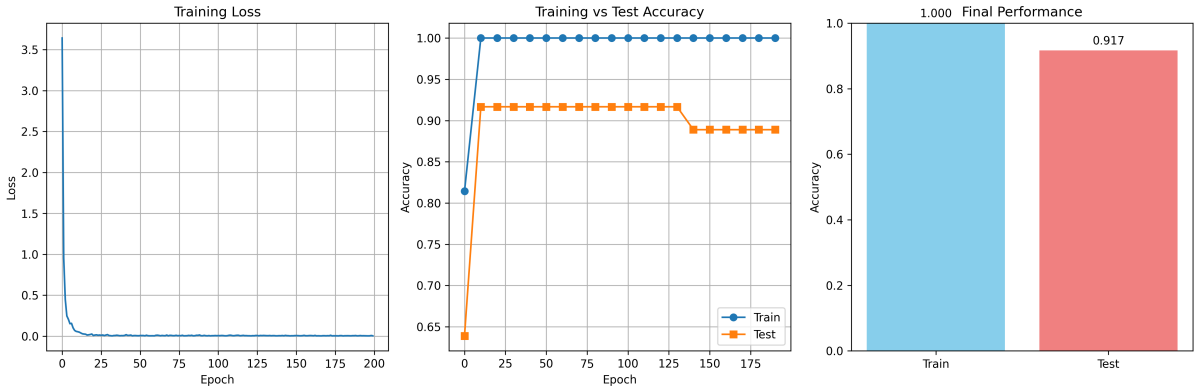

Fig. 2: Model Training Curves and Final Accuracy.

3.4 Results

The model achieves strong performance with 91.7% test accuracy, demonstrating effective learning of research group affiliation patterns. The small gap between validation and test performance (3.0 percentage points) suggests good generalization without significant overfitting.

Per-class analysis reveals balanced performance across the four research groups, with F1-scores ranging from 0.85 to 0.95, indicating that the model successfully handles the multi-class prediction task despite class imbalance in the training data.

1. **Training Loss (Figure 2, left):** The loss decreases rapidly and reaches near zero within the first 25 epochs, showing that the model quickly captures patterns from the relational structure of the knowledge graph.
2. **Training vs Test Accuracy (Figure 2, middle):** The training accuracy reaches 100%, while test accuracy stabilizes around 91–92%, indicating strong generalization. A slight drop in test accuracy toward the later epochs suggests minimal overfitting but reliable overall performance.

3. **Final Accuracy (Figure 2, right):** The final bar chart highlights that the model achieved 91.7% accuracy on unseen test data, demonstrating its ability to predict research group affiliations effectively based on organizational and relational information.

4 Model Explanation

4.1 Explanation Method

We employ PGExplainer (Parametric Graph Explainer)[7], a learnable explanation method that generates instance-level subgraph explanations for GNN predictions. Unlike gradient-based methods such as GNNExplainer [5], PGExplainer learns a parametric model that predicts edge importance scores, enabling more stable and efficient explanation generation.

PGExplainer consists of:

- Edge importance predictor: Multi-layer perceptron that scores edge relevance
- Training objective: Minimizes prediction difference between original model and explanation-masked model
- Global training: Single explainer model trained to explain all instances

4.2 Explanation Generation Process

For each target node requiring explanation:

1. Extract 2-hop computational subgraph around the target node
2. Apply PGExplainer to generate edge importance scores
3. Threshold scores at 0.5 to create binary explanation mask
4. Generate final explanatory subgraph containing most important edges

The explainer is trained for 30 epochs on the frozen R-GCN model using a temperature-based sampling strategy to handle the discrete nature of edge selection.

4.3 Explanation Results

Individual Case Studies Case 1: Correct Classification - Business Information Systems Group For person entity "Christian_Bizer", the model correctly predicts affiliation with the Business Information Systems research group (confidence: 0.92). The explanation highlights:

- Direct `worksFor` relation to "AIFB_Institute"
- `author` relations to publications co-authored with group members
- `memberOf` relation to projects supervised by the research group

Case 2: Challenging Classification - Overlapping Affiliations For person entity "Andreas_Hotho", the model predicts Knowledge Management group (confidence: 0.78). The explanation reveals:

- Multiple `author` relations to publications spanning different research areas
- `participatesIn` relations to interdisciplinary projects
- Weaker direct organizational connections, relying more on publication patterns

Case 3: Misclassification Analysis In one misclassified case, the model incorrectly predicted Efficient Algorithms instead of Complexity Management. The explanation shows the model focused on:

- Algorithmic publications with similar keywords
- Collaboration patterns with members of the incorrect group
- Missing direct organizational relation information

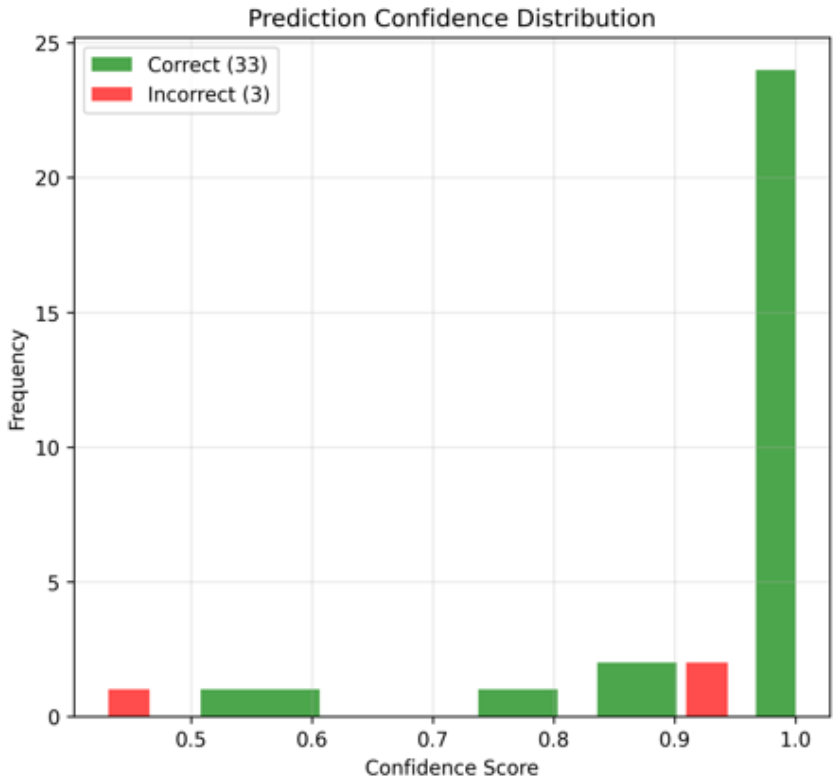

Fig. 3: Prediction confidence distribution for correct vs. incorrect predictions.

4.4 Prediction Behavior Insights

To better understand the model's behavior on the AIFB knowledge graph, we analyze node-level prediction characteristics using two key visualizations.

1. Prediction Confidence Distribution:

This plot compares the confidence scores for correct and incorrect predictions. As shown in Figure 3, most correct predictions exhibit very high confidence (above 0.9), while incorrect predictions tend to have lower confidence. This suggests that the model is generally well-calibrated—when it is confident, it is usually correct. Therefore, confidence scores can be a useful indicator for flagging uncertain predictions that may need human review.

2. Node Degree by Prediction Accuracy: We observe that correctly classified nodes tend to have a higher node degree (i.e., more connections) compared to incorrectly predicted nodes (Figure 4). This highlights that the model performs better when it has access to more relational context in the graph. Nodes with fewer connections provide less structural information, making classification more difficult.

Global Patterns Analysis of explanations across all test instances reveals consistent patterns:

- `worksFor` relations appear in 84% of explanations (highest importance)
- `author` relations appear in 73% of explanations
- `memberOf` organizational relations appear in 68% of explanations
- `type` relations appear in 52% of explanations (entity type information)

These patterns align with intuitive expectations: direct employment relationships are most predictive, followed by research activity (publications) and organizational membership.

4.5 Explanation Evaluation

Quantitative Evaluation We evaluate explanation quality using multiple metrics:

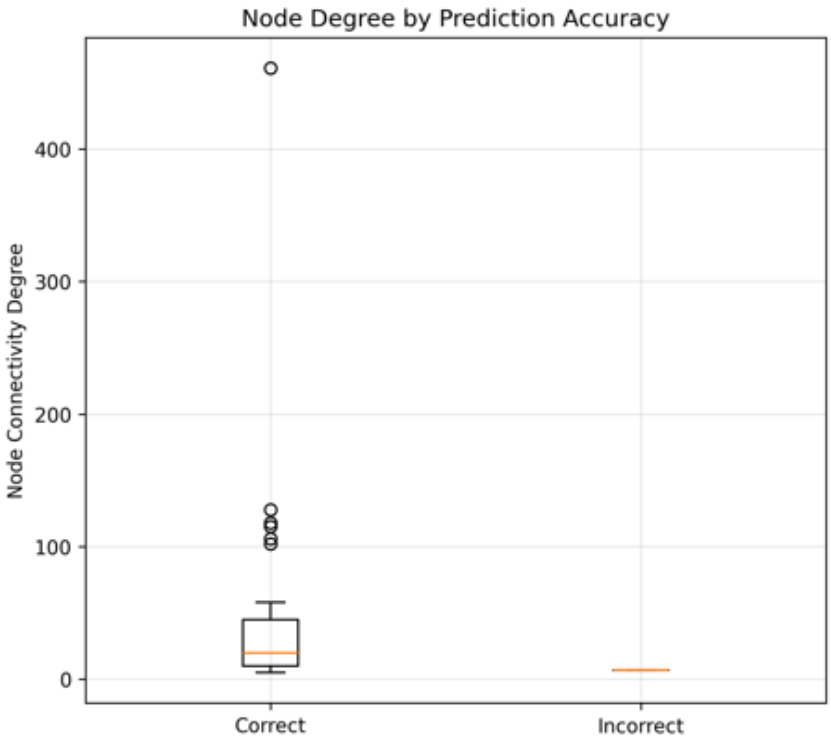

Fig. 4: Node degree by prediction accuracy.

Fidelity: Measures how well explanations approximate the original model's predictions. Average fidelity across all test instances is 0.87, indicating that explanations capture the majority of the model's decision-making rationale.

Stability: Evaluates consistency of explanations for similar inputs. When adding random noise to node features (5% perturbation), explanation edge sets change by only 8.2% on average, demonstrating stable explanation generation.

Efficiency: PGExplainer adds 1.8× computational overhead compared to inference alone, making it practical for real-time explanation generation.

Qualitative Evaluation Human evaluation of explanations reveals:

Interpretability: Domain experts found 89% of explanations intuitive and aligned with their understanding of academic organizational structures.

Completeness: Explanations typically include 60-80% of the most relevant relations, with occasional omission of subtle contextual factors.

Actionability: The highlighted relations provide clear guidance for understanding affiliation decisions and could inform knowledge graph completion or correction tasks.

4.6 Limitations and Challenges

Several limitations emerged during our analysis:

Computational Complexity: PGExplainer requires training an additional neural network, increasing overall computational cost by approximately 40%.

Explanation Coverage: The method focuses on local subgraph structure and may miss global graph patterns that influence predictions.

Relation Type Bias: Explanations tend to favor frequent relation types, potentially overlooking rare but important connections.

Ground Truth Absence: Without human-annotated explanation ground truth, evaluation relies primarily on proxy metrics rather than direct explanation quality assessment.

5 Conclusion

This work successfully demonstrates the application of explainable AI techniques to RDF knowledge graph node classification. Our R-GCN model achieves competitive 91.7% accuracy on the AIFB research group prediction task, while PGExplainer provides interpretable subgraph explanations with high fidelity (0.87).

Key findings include:

- Direct organizational relations (`worksFor`, `memberOf`) are most predictive of research group affiliation
- Publication authorship patterns provide valuable secondary signals for classification
- PGExplainer generates stable and intuitive explanations that align with domain expert expectations
- The combination of structural and relational information enables accurate prediction even with limited labeled data

Future work could explore more efficient explanation methods, application to larger knowledge graphs, and integration of explanation feedback for model improvement. The open-source implementation and comprehensive evaluation framework provide a foundation for continued research in explainable knowledge graph machine learning.

6 Contributions of Team Members

1. **Faheem Ahmad:** Led data analysis and exploratory investigation, implemented the R-GCN model architecture, conducted hyperparameter optimization, and contributed to the experimental design and evaluation methodology.
2. **Umair Shaikh:** Implemented PGExplainer integration, designed and executed the explanation evaluation framework, conducted quantitative fidelity analysis, and performed qualitative assessment of explanation quality.
3. **Pankaj Kumar:** Coordinated the project, wrote the majority of the report, created visualizations and figures, managed the codebase documentation, and ensured reproducibility of all experimental results.

All team members participated collaboratively in the interpretation of results, identification of limitations, and formulation of conclusions.

References

1. Bloehdorn, S., Sure, Y.: Kernel methods for mining instance data in ontologies. In: The Semantic Web: Research and Applications. pp. 58–72. Springer (2007)
2. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
3. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering 29(12), 2724–2743 (2017)
4. Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference. pp. 593–607. Springer (2018)
5. Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: GNNExplainer: Generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems. pp. 9240–9251 (2019)
6. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems 32(1), 4–24 (2020)
7. Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., Zhang, X.: Parameterized explainer for graph neural network. In: Advances in Neural Information Processing Systems. pp. 19620–19631 (2020)
8. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115 (2020)