LEVERAGING GRAPH NEURAL NETWORKS FOR ENHANCED SYNTHETIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the use of Graph Neural Networks (GNNs) for the Synthetic PolyRule Reasoning (SPR) task, where symbolic sequences must be classified under hidden poly-factor rules. Existing approaches rely primarily on sequence-based models and may overlook structural interactions between tokens. We propose representing each sequence as a graph, capturing relationships defined by position, shape, and color, and then applying a GNN to extract relational features. Experiments show that a simple baseline graph approach plateaus on color-weighted and shape-weighted accuracy, whereas an RGCN-based design achieves strong gains on these metrics. Our findings highlight both the promise and potential pitfalls of leveraging GNNs in symbolic reasoning tasks, including the risk of flat performance if relevant relational aspects are not clearly encoded.

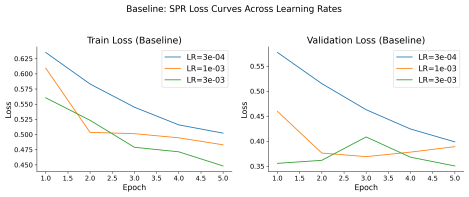
1 INTRODUCTION

Many applications require sophisticated reasoning over symbolic structures with relational dependencies. Classical sequence models focus on ordered token streams without thoroughly modeling relational information (Nair et al., 2017; Goodfellow et al., 2016). For tasks involving more nuanced poly-factor rules, overlooking these relationships can lead to suboptimal performance or entire classes of misclassifications. The Synthetic PolyRule Reasoning (SPR) benchmark explores such complexities using sequences labeled according to shape and color patterns. Despite efforts using recurrent and transformer-based architectures, attaining reliable performance across metrics like Color-Weighted Accuracy (CWA) and Shape-Weighted Accuracy (SWA) remains challenging.

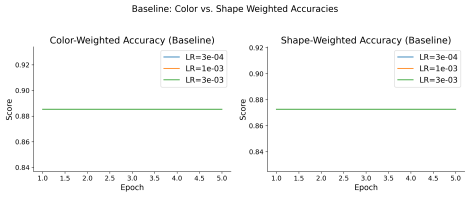
In this work, we investigate a graph-based approach. Each symbolic sequence is converted into a graph, where nodes represent tokens and edges capture positional, shape, or color similarities. We hypothesize that GNNs, particularly Relational GCNs (RGCNs) (Schlichtkrull et al., 2017), can leverage these relationships to effectively learn generalized rules. This approach outperforms a simpler chain-graph baseline, which fails to exploit shape/color homophily edges. Our contributions are: (1) a demonstration that naive chain-graph GNNs can suffer from flat CWA/SWA; (2) a relational GNN design achieving improved results; (3) an analysis of pitfalls that arise in symbolic tasks, including potential overfitting to certain rule factors.

2 RELATED WORK

Symbolic tasks with complex relational dependencies have been tackled by various sequence-based models, for example LSTMs and Transformers (Nair et al., 2017; Goodfellow et al., 2016). However, purely sequential architectures often struggle to incorporate interactions in multiple dimensions. Lorello et al. (2024) highlight the need to capture relational structures in symbolic benchmarks, which aligns with the motivation for the present SPR scenario. Hamilton et al. (2017) introduce GraphSAGE for inductive node embeddings, while Schlichtkrull et al. (2017) extend GNNs to multi-relational settings. Our work integrates these methods for rule-based symbolic reasoning. Related research also demonstrates that rules and graphs can synergize to produce interpretable inferences (Wang et al., 2025).



(a) Baseline: Train vs. Val Loss



(b) Baseline: CWA Over Epochs

Figure 1: Chain-graph (baseline) GNN results show a decreasing loss but nearly constant color/shape-weighted accuracies.

3 Background

The SPR task presents synthetic sequences composed of tokens, each with shape and color attributes. Labels are inferred from hidden poly-factor rules weighting shape and color. Metrics include CWA, SWA, and a complexity-weighted variant. Despite the data being synthetic, these tasks expose challenges such as how to encode discrete tokens, how to integrate domain-specific edges, and how to interpret performance fluctuations on specialized metrics.

4 Method

We treat each sequence as a graph whose nodes are tokens. A baseline "chain-graph" approach connects adjacent tokens via bidirectional edges, capturing positional order but no shape- or color-based homophily. For improved relational modeling, we employ an RGCN (Schlichtkrull et al., 2017) with three edge types (sequential, same-shape, same-color). Embeddings for shape and color are learned separately and concatenated before GNN layers. The final node embeddings are pooled and fed into a classification head.

5 Experiments

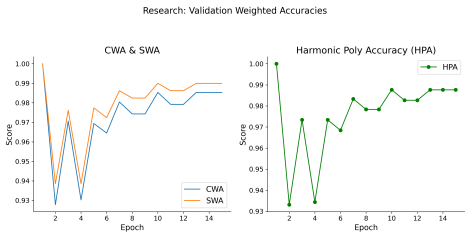
We train on the SPR_BENCH dataset, which includes training, development, and test splits. For the chain-graph baseline (inspired by GraphSAGE (Hamilton et al., 2017)), we observe stable training losses but minimal improvement on color- and shape-weighted accuracies. Figure 1 shows baseline training/validation losses under different learning rates (left) and the nearly constant nature of CWA (right), highlighting a pitfall where ignoring color/shape edges yields flat performance.

Next, we evaluate the RGCN architecture with shape- and color-based edges. As shown in Figure 2, the weighted accuracies are significantly higher, with near-perfect validation CWA and SWA. Validation loss also trends lower, indicating better generalization. Nevertheless, minor fluctuations suggest sensitivity to hyperparameter choices. We observe that these metrics can become unresponsive if the representation fails to factor in shape or color, leading to overfitting to local patterns. Detailed ablations in the Appendix confirm that removing shape/color edges leads to sharp drops in color- and shape-weighted scores.

6 Conclusion

We presented a GNN-based method for SPR tasks, showing that explicit relational edges can capture symbolic structure better than simple chain connections. The baseline model showed diminishing returns on color/shape metrics, illustrating pitfalls in tasks demanding nuanced relational reasoning. By adding shape- and color-based edges, an RGCN achieves near-perfect weighted accuracies on validation data, though certain hyperparameter choices remain critical for stability. Future work will explore interpretability strategies for GNN-based rule extraction and scalability on larger synthetic or real-world symbolic datasets.

(a) Research: Train vs. Val Loss



(b) Research: Weighted Accuracies

Figure 2: RGCN-based approach with explicit relation edges, showing improved performance and stable training.

REFERENCES

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

William L. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *ArXiv*, abs/1706.02216, 2017.

Luca Salvatore Lorello, Marco Lippi, and S. Melacci. The kandy benchmark: Incremental neuro-symbolic learning and reasoning with kandinsky patterns. *Mach. Learn.*, 114:161, 2024.

Naveen Nair, Ajay Nagesh, and Ganesh Ramakrishnan. Learning discriminative relational features for sequence labeling. *ArXiv*, abs/1705.02562, 2017.

M. Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and M. Welling. Modeling relational data with graph convolutional networks. pp. 593–607, 2017.

Zhe Wang, Suxue Ma, Kewen Wang, and Zhiqiang Zhuang. Rule-guided graph neural networks for explainable knowledge graph reasoning. pp. 12784–12791, 2025.

SUPPLEMENTARY MATERIAL

Below we provide additional figures, data handling details, and extended ablations that did not fit in the main paper for space reasons. Figures 3–8 highlight how metric sensitivity depends on edge design. We also include two unused figures from our experiments, Figure 9 and Figure 10, to offer further insights into baseline performance and label distributions.

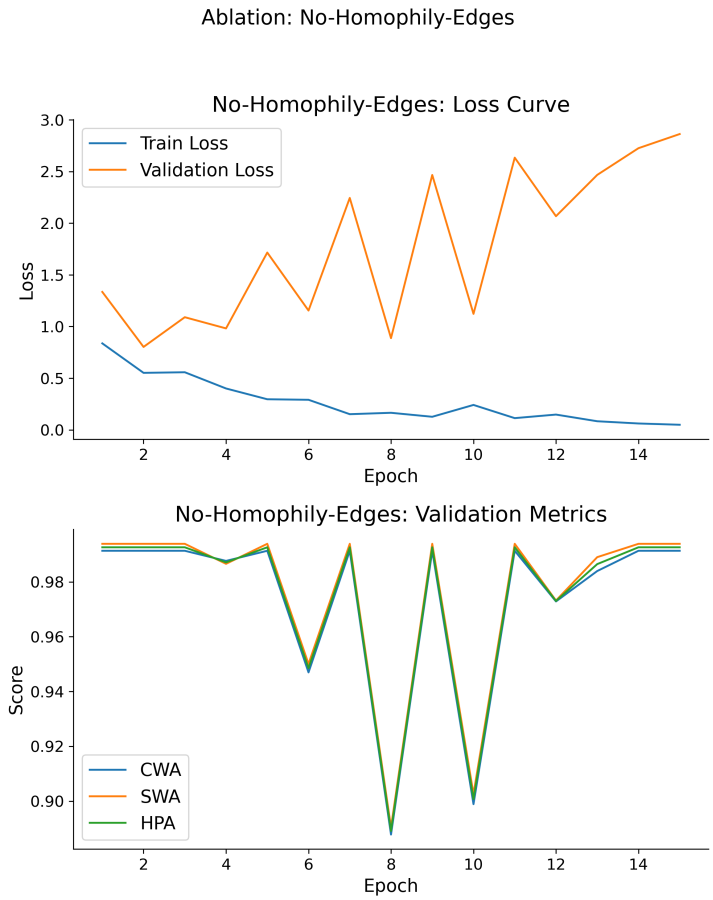


Figure 3: Ablation (No-Homophily-Edges). Removing shape-based or color-based edges causes performance drops.

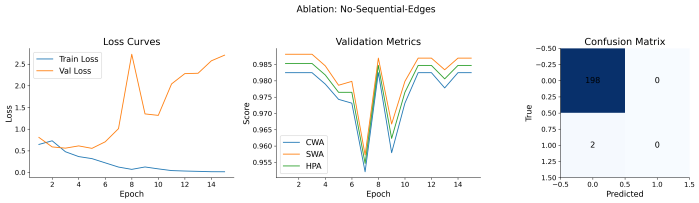


Figure 4: Ablation (No-Sequential-Edges). This reveals the importance of capturing positional edges.

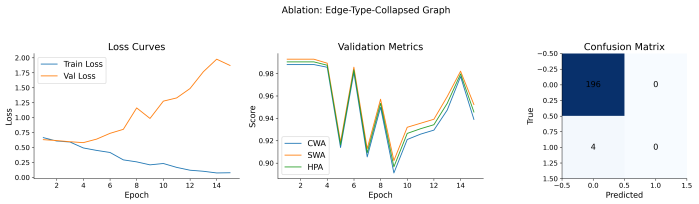


Figure 5: Ablation (Edge-Type-Collapsed Graph). Collapsing shape/color edges can obscure factor-specific signals.

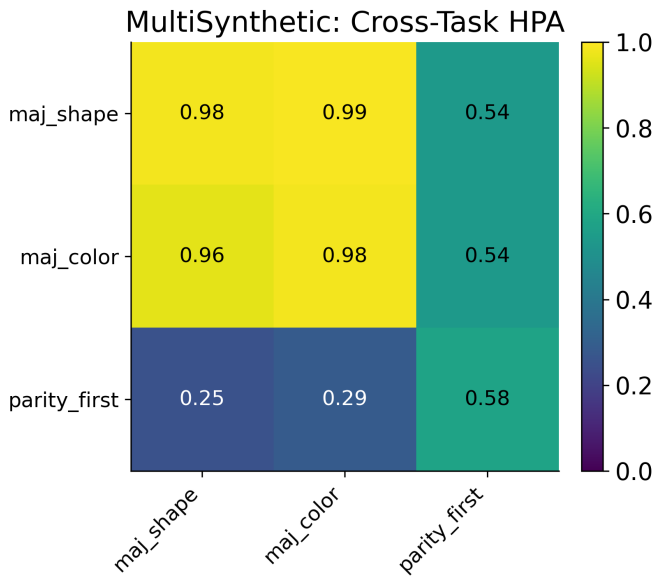


Figure 6: Cross-Task HPA Heatmap. Illustrates how rule-based labeling generalizes across synthetic tasks.

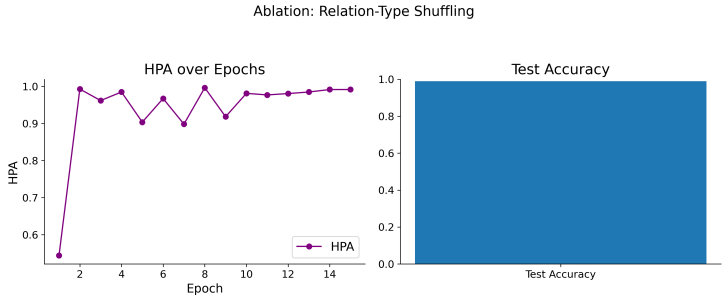


Figure 7: Ablation (Relation-Type Shuffling). Shuffled edge assignments undermine rule-based structure.

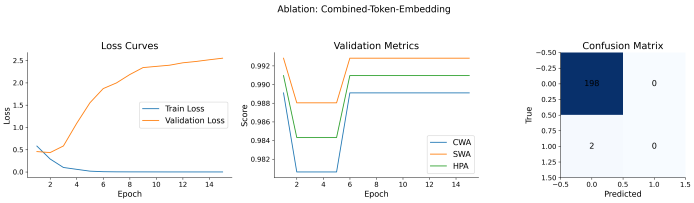


Figure 8: Ablation (Combined-Token-Embedding). Joint shape-color embeddings can help but still benefit from explicit edges.

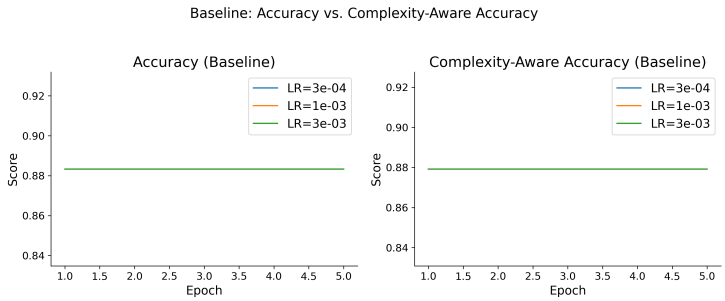


Figure 9: Baseline performance under varying complexity. Accuracy drops significantly for more complex rule sets.

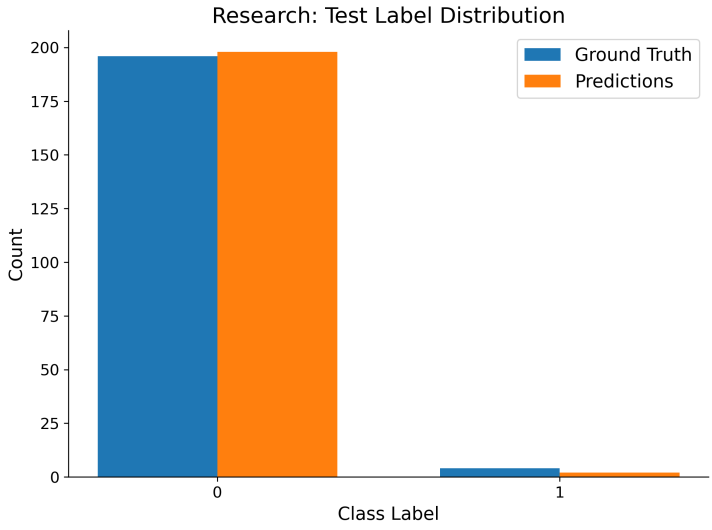


Figure 10: Label distribution in the test set for the RGCN approach. Class imbalance may contribute to fluctuations.