# Interpretable GNN with Stable Concepts as Neurons

Sergei O. Kuznetsov[1,*,†], Mariia M. Zueva[1,†]

[1]Higher School of Economics (HSE University), 11 Pokrovskii Bulvar, Moscow, 109025, Russian Federation

**Abstract**

A method for constructing interpretable graph neural networks based on an ordered subsets of "interesting" formal concepts is proposed. Interestingness indices for selecting concepts are compared, namely frequency (extent size) and $\Delta$-stability. The experimental results demonstrate that $\Delta$-stability provides superior performance of the generated networks at the inference stage.

**Keywords**

Concept lattice, $\Delta$-stability, Pattern Structures, Graph neural networks

## 1. Introduction

Contemporary neural networks offer high performance but suffer from low interpretability, high memory demands, and computational inefficiencies. To address this, research has focused on compact and efficient architectures. Notable advances include self-distillation [1], Logic Explained Networks (LENs) [2], Concept Distillation Modules (CDM) [3], compact complex-frequency domain models [4], and lattice-based pruning frameworks [5].

Integrating Formal Concept Analysis (FCA) with neural networks has also been explored in [6, 7, 8]. Architectures based on concept lattices [9] later incorporated interestingness measures [10], with $\Delta$-stability shown to improve learning quality [11].

Concept lattices enable network compression and inherent hierarchical structural information. In [12] authors successfully integrated lattices into GNNs to exploit these hierarchies. Their approach relied on frequency-based concept selection, which may overlook meaningful patterns. We address this by replacing frequency with $\Delta$-stability-based concept distillation and applying pattern structures for interpretable data processing.

The paper is organized as follows: Section 2 covers $\Delta$-stability theory. Section 3 details our methodology. Section 4 describes the experimental part. Section 5 concludes and outlines future work.

## 2. Theoretical Background

f First we recall some basic definitions of Formal Concept Analysis (FCA) and Pattern Structures.

*Formal Context* is a triple $(G, M, I)$ where $G$ is a set of objects, $M$ is a set of attributes, and $I \subseteq G \times M$ is a set of pairs $(g, m)$ representing that object $g$ has attribute $m$.

The space of object descriptions $\underline{\mathbb{D}} = (\mathbb{D}, \subseteq)$ is the powerset of attributes $\mathbb{D} = \wp(M)$, ordered by inclusion $\subseteq$. The description space $\underline{\mathbb{D}} = (\mathbb{D}, \subseteq)$ creates a lattice, so, for every pair of descriptions $d_1, d_2 \in \mathbb{D}$ there is exactly one meet and one join: $\exists! d_\wedge \in \mathbb{D}, d_1 \cap d_2 = D_\wedge$ and $\exists! d_\vee \in \mathbb{D}, d_1 \vee d_2 = d_\vee$.

*Description space* consists of a description set $\mathbb{D}$ with operation $\sqcap$ that defines a complete meet semilattice on $\mathbb{D}$, i.e., for any pair of descriptions $d_1, d_2 \in \mathbb{D}$ there exists meet (infimum):

$\exists! d_\wedge \in , d_1 \sqcap d_2 = d_\wedge$.

The operation $\sqcap$ defines natural order $\sqsubseteq$: $X \sqsubseteq Y \iff X \sqcap Y = X$.

*The Third International Joint Conference on Conceptual Knowledge Structures (CONCEPTS), August 31 – September 4, 2026, Montpellier, France*

*Corresponding author.

†These authors contributed equally.

skuznetsov@hse.ru (S. O. Kuznetsov); m.zueva@hse.ru (M. M. Zueva)

0000-0003-3284-9001 (S. O. Kuznetsov); 0009-0000-6332-9936 (M. M. Zueva)

The description space $\underline{\mathbb{D}}$ equipped with the set of objects $G$ and a mapping $\delta : G \to \mathbb{D}$ forms a *pattern structure* [13] $(G, \underline{\mathbb{D}}, \delta)$, which can be considered as a generalization of a formal context $(G, M, I)$.

For any pattern description $d \in \mathbb{D}$, one can define the pattern extent $d^\diamond$, and for any subset of objects $A \subseteq G$, one can define the pattern intent $A^\diamond$:

$$d^\diamond = \{g \in G \mid d \sqsubseteq \delta(g)\}, \quad A^\diamond = \sqcap\{\delta(g) \mid g \in A\}. \tag{1}$$

A pair of corresponding pattern extent $A$ and pattern intent $d$ forms a *pattern concept*: $(A, d)$, where $A^\diamond = d, d^\diamond = A$.

The *Interval Pattern Structure* uses the description space $\underline{\mathbb{D}}_{\text{int}}$ of intervals bounded by real numbers $\mathbb{D}_{\text{int}}$ ordered by interval subsumption $\sqsubseteq_{\text{int}}$:

$$\mathbb{D}_{\text{int}} = \{[l, r] \mid l, r \in \mathbb{R}, l \leq r\}, \text{ and } \forall [l_1, r_1], [l_2, r_2] \in \mathbb{D}_{\text{int}}, [l_1, r_1] \sqsubseteq_{\text{int}} [l_2, r_2]$$

$$\iff [l_1, r_1] \supseteq [l_2, r_2].$$

The *Cartesian Pattern Structure* allows one to combine various description spaces $\underline{\mathbb{D}}_1, \underline{\mathbb{D}}_2, \ldots, \underline{\mathbb{D}}_n$ in a single description space $\underline{\mathbb{D}}_\times = (\mathbb{D}_\times, \sqsubseteq_\times)$ such that:

$$\mathbb{D}_\times = \bigtimes_{1 \leq i \leq n} \mathbb{D}_i \text{ and } \forall d, E \in \mathbb{D}_\times, d \sqsubseteq_\times E \iff \bigwedge_{1 \leq i \leq n} d_i \sqsubseteq_i E_i.$$

*Support* is the number of objects described by a binary description $B \subseteq M$, or the number of objects described by a pattern description $d \in \mathbb{D}$:

$$\texttt{supp}(B) = |B'|, \forall B \subseteq M, \quad \texttt{supp}(d) = |d^\diamond|, \forall d \in \mathbb{D} \tag{2}$$

(Intentional) *stability* of a formal concept $(A, B)$ is defined in [14] as the percentage of subsets of the concept extent $A$ with maximal common description $B$:

$$\texttt{stab}_i(A, B) \coloneqq \frac{|\{A_1 \subseteq A \mid A_1' = B\}|}{2^{|A|}}. \tag{3}$$

Since computing stability is computationally hard (a #P-complete problem), several approximations were considered, including (intentional) $\Delta$-stability [15], an easily computable value monotonically related to the (intentional) concept stability:

$$\Delta\texttt{stab}_\texttt{i}(A, B) \coloneqq \min_{\substack{A_1'' = A_1 \\ A_1 \subset A}} (|A| - |A_1|). \tag{4}$$

Indeed,

$$\texttt{stab}_i(A, B) \times 2^{|A|} \leq 2^{|A|} - 2^{|A| - \Delta\texttt{stab}_i(A,B)} = 2^{|A|} \cdot (1 - 2^{-\Delta\texttt{stab}_i(A,B)}).$$

So, $\texttt{stab}_i(A, B) \leq 1 - 2^{-\Delta\texttt{stab}_i(A,B)}$ and

$$\Delta\texttt{stab}_i(A, B) \geq -log(1 - \texttt{stab}_i(A, B)$$

In Figure 1 we give a schematic representation of the powerset of extent $A$ and intentional $\Delta$-stability of the respective concept $(A, B)$.

Informally, the value of (intensional) $\Delta$-stability says "how many objects from $A$ one should delete at least to make intent $B$ a bit more precise".

The notion of *extensional stability* is dual to the intensional one:

$$\texttt{stab}_e(A, B) \coloneqq \frac{|\{B_1 \subseteq B \mid B_1' = A\}|}{2^{|B|}}. \tag{5}$$

For pattern structures stability is defined in a similar way.

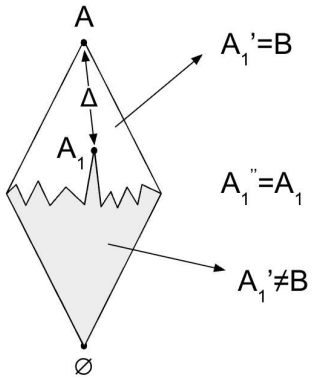


**Figure 1:** Powerset of $A$ and $\Delta$-stability.

## 3. Methodology

Our approach builds upon the architectural framework proposed by [12], introducing two key modifications: (1) we replace the generalized one-sided concept lattice with a pattern structure approach to handle numerical data directly, and (2) we employ a $\Delta$-stability-based filtering mechanism for concept selection instead of frequency-based pruning.

The model adopts a two-branch GCN architecture. The first branch constructs a graph using cosine similarity between attribute vectors of the objects and processes it with a standard GCN. The second branch constructs a concept-based graph from a set of $\Delta$-stable pattern concepts. To obtain these concepts, we use pattern structures for attributes and apply the gSofia algorithm [16], which allows us to control the number of extracted concepts via a $\Delta$-stability threshold. From the resulting set of $\Delta$-stable concepts, we build an object-object adjacency matrix based on parent-child relationships in the concept lattice. This matrix is then fed into a separate GCN branch. The learned representations from both branches are concatenated and passed through a fully connected layer with softmax activation for classification.

Unlike purely black-box models, our approach enables post-hoc interpretability: the contribution of each $\Delta$-stable concept to the classification decision can be traced through the network activations.

## 4. Experimental Part

In our experiments, the number of selected concepts $k$ ranges from 5 to 20, which allows us to study model behavior as we transition from a more detailed to a more compressed representation of the data.

For the GCN architecture, we used the PyTorch library with Kaiming Normal initialization and Hyperbolic tangent activation function. The experiments were run on MacOS 14.2.1 (23C71), Apple M2.

**Table 1**
The datasets used in experiments.

| Dataset | # objects | # attributes | # target values |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Zoo | 101 | 16 | 7 |
| Lenses | 24 | 4 | 3 |
| Soybean | 47 | 35 | 4 |
| Car | 1728 | 6 | 4 |
| Balance Scale | 625 | 4 | 3 |
| Monk | 432 | 6 | 2 |

To evaluate the performance of the concept selection method based on the $\Delta$-stability index compared

to the support-based baseline, we conducted experiments on seven datasets from the UCI repository: Iris, Zoo, Lenses, Soybean, Car, Balance Scale, and Monk (Table 1). For each dataset, we varied the number of selected concepts $k$ from 5 to 20.

The experimental results demonstrate that the $\Delta$-stability-based selection strategy consistently outperforms or matches the frequency-based baseline across all datasets. For the Iris dataset, $\Delta$-stability achieved its highest accuracy of 0.8317 at $k = 12$, compared to 0.8167 for frequency-based selection. For Zoo, the improvement was more pronounced: $\Delta$-stability reached 0.8707 at $k = 10$, while frequency-based selection peaked at only 0.8000. For Soybean, $\Delta$-stability attained 0.9474, considerably outperforming the frequency-based approach (0.8263). On the remaining four datasets — Balance Scale, Monk, Lenses, and Car — both selection methods produced identical concept lists, resulting in equivalent performance. This occurs when the $\Delta$-stability filtering stage does not eliminate any concepts from the frequency-ranked list, indicating that for these particular datasets, the most frequent concepts also possess high $\Delta$-stability. Overall, the weighted F1-scores follow the same trend as accuracy, confirming that $\Delta$-stability is a robust criterion for concept selection across diverse data characteristics.

Figure 2 illustrates the comparison of accuracy and weighted F1-score between the frequency-based and $\Delta$-stability-based selection strategies on the Zoo dataset for $k$ ranging from 5 to 20. The results demonstrate that $\Delta$-stability consistently outperforms frequency-based selection across all values of $k$, with the largest difference observed at $k = 16$, where $\Delta$-stability achieves 0.8634 accuracy compared to 0.7448 for frequency-based selection.

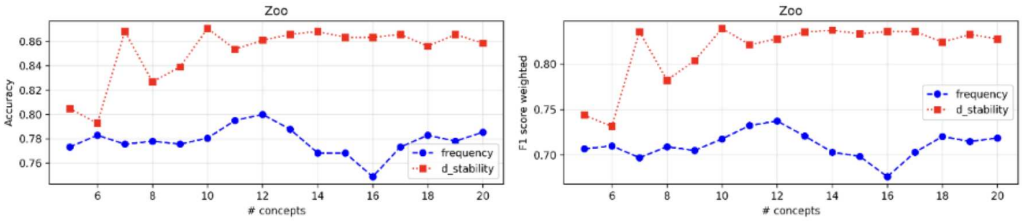


**Figure 2:** Accuracy and weighted F1-score comparison between frequency-based and $\Delta$-stability-based concept selection approaches on the Zoo dataset for $k \in [5, 20]$.

## 5. Conclusion

In this paper we have proposed an interpretable GCN architecture with neurons based on stable concepts. In our computer experiments with several datasets we have demonstrated that the initial selection of formal concepts using $\Delta$-stability followed by subsequent frequency-based selection is an effective method for selecting interesting concepts, which outperforms purely frequency-based methods or is on par with them. This result highlights the potential of the proposed approach for constructing compact concept-based neural networks, with additional benefits in interpretability of results and the ability to handle complex data represented by pattern structures.

## Data availability

The code used for the experiments is publicly available on GitHub https://github.com/messalika/Interpretable-GNN-with-Stable-Concepts-as-Neurons.

## Acknowledgments

Support from the Basic Research Program of HSE University is gratefully acknowledged.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References

[1] L. Zhang, C. Bao, K. Ma, Self-distillation: Towards efficient and compact neural networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 4388–4403.
[2] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822.
[3] L. C. Magister, P. Barbiero, D. Kazhdan, F. Siciliano, G. Ciravegna, F. Silvestri, M. Jamnik, P. Liò, Concept distillation in graph neural networks, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 233–255.
[4] F. E. Wahab, Z. Ye, N. Saleem, R. Ullah, Compact deep neural networks for real-time speech enhancement on resource-limited devices, Speech Communication 156 (2024) 103008.
[5] R. Sengupta, Lattice-based pruning in recurrent neural networks via poset modeling, arXiv preprint arXiv:2502.16525 (2025).
[6] S. Rudolph, Using fca for encoding closure operators into neural networks, in: International Conference on Conceptual Structures, Springer, 2007, pp. 321–332.
[7] G. G. Towell, J. W. Shavlik, Knowledge-based artificial neural networks, Artificial intelligence 70 (1994) 119–165.
[8] D. Endres, P. Foldiak, Interpreting the neural code with formal concept analysis, Advances in Neural Information Processing Systems 21 (2008).
[9] S. O. Kuznetsov, N. Makhazhanov, M. Ushakov, On neural network architecture based on concept lattices, in: International Symposium on Methodologies for Intelligent Systems, Springer, 2017, pp. 653–663.
[10] S. O. Kuznetsov, T. P. Makhalova, Concept interestingness measures: a comparative study., in: CLA, volume 1466, 2015, pp. 59–72.
[11] M. Zueva, S. Kuznetsov, Interestingness indices for building neural networks based on concept lattices, Automation and Remote Control 85 (2024) 272–278.
[12] M. Shao, Z. Hu, W. Wu, H. Liu, Graph neural networks induced by concept lattices for classification, International Journal of Approximate Reasoning 154 (2023) 262–276.

[13] B. Ganter, S. O. Kuznetsov, Pattern structures and their projections, in: International conference on conceptual structures, Springer, 2001, pp. 129–142.
[14] S. O. Kuznetsov, On stability of a formal concept, Annals of Mathematics and Artificial Intelligence 49 (2007) 101–115.
[15] A. Buzmakov, S. O. Kuznetsov, A. Napoli, Scalable estimates of concept stability, in: International Conference on Formal Concept Analysis, Springer, 2014, pp. 157–172.
[16] A. Buzmakov, S. O. Kuznetsov, A. Napoli, Fast generation of best interval patterns for nonmonotonic constraints, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 157–172.