Even-if Explanations for Graph Neural Networks

Gianvincenzo Alfano[1], Sergio Greco[1], Domenico Mandaglio[1], Francesco Parisi[1], Reza Shahbazian[2] and Irina Trubitsyna[1]

[1]Department of Informatics, Modeling, Electronics and System Engineering, University of Calabria, Italy
[2]Department of Humanities, University of Palermo, Italy

Abstract

Graph Neural Networks (GNNs) are the leading models for prediction tasks over graph-structured data, from molecules to social networks, which makes explaining their decisions increasingly important. Post-hoc explanations usually build on counterfactual "if only" reasoning, i.e., the smallest change that would alter a prediction, whereas the complementary semifactual "even if" reasoning, which reports the largest change that leaves a prediction unchanged, has remained almost unexplored for graphs. In this paper, we discuss semifactual reasoning for GNNs recently proposed in [1], analyse the computational complexity of the associated problems, and describe a learning-based architecture for their computation, together with the main experimental findings.

Keywords

Explainable AI, Semifactual Explanations, Graph Neural Networks, Computational Complexity

1. Introduction

Graph-structured data, where entities (nodes) are connected by pairwise relationships (edges), arise naturally in domains such as molecular property prediction, social network analysis, and recommendation systems [2, 3, 4, 5]. Graph Neural Networks (GNNs) have become the leading models for predictive tasks on graphs [6, 7, 8], yet understanding why a GNN assigns a given prediction remains a fundamental challenge in high-stakes settings where trust and transparency are essential [9].

Post-hoc explanation methods for GNNs typically provide either *factual* explanations, which identify the subgraph most responsible for a decision [10, 11, 12, 13, 14, 15, 16], or *counterfactual* explanations, which ask how minimal graph changes could alter the decision [17, 18, 15]. However, the computational complexity of these problems has been studied only rarely [19, 20]. Much less attention has been devoted to the equally important and complementary paradigm of *semifactual* "even-if" reasoning. Whereas counterfactuals identify the minimal changes that would flip a prediction, semifactuals identify the *largest* perturbations that leave it unchanged [21, 22, 23, 24]. This is particularly informative when a prediction is a desired outcome, e.g., a loan request being accepted: a semifactual reveals the input's degrees of freedom that do not affect the result [21, 24].

Semifactual reasoning has so far been studied almost exclusively in simpler settings such as tabular or binary data [21, 22, 25, 26]: formal foundations, complexity, and preference-based frameworks have been developed for binary classifiers [24] and abstract argumentation [27, 28]. None of this prior work addresses graph-structured data or GNNs [26]. The work we discuss here, originally presented in [1], addresses this missing case by defining and studying semifactual explanations for GNN-based graph and node classification, thereby extending the binary-data framework of [24] to graph-structured inputs.

Example 1. Consider graphs that model molecules, where nodes are atoms, edges are bonds, and a GNN Φ has been trained to predict water solubility. Let $\mathcal{G}$ represent *Acetone* (C_3H_6O), classified as *soluble* (Figure 1, left). A semifactual asks for the "farthest" molecule with the same atoms that is still classified as soluble: e.g., *Propanal*, a structural isomer of Acetone obtained through six edge modifications (three deletions and three additions), yet still predicted soluble. Conversely, a counterfactual asks for the

AIxIA 2026 Discussion Papers, 24th International Conference of the Italian Association for Artificial Intelligence, Perugia, Italy, 2026

g.alfano@dimes.unical.it (G. Alfano); greco@dimes.unical.it (S. Greco); d.mandaglio@dimes.unical.it (D. Mandaglio); fparisi@dimes.unical.it (F. Parisi); reza.shahbazian@unipa.it (R. Shahbazian); i.trubitsyna@dimes.unical.it (I. Trubitsyna)

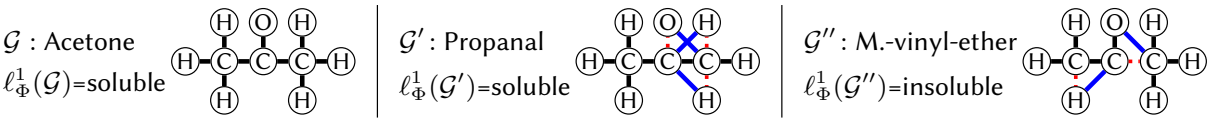

Figure 1: Acetone $\mathcal{G}$, Propanal $\mathcal{G}'$, and Methyl-vinyl-ether $\mathcal{G}''$ (Example 1): $\mathcal{G}'$ is a semifactual of $\mathcal{G}$ (distance 6), $\mathcal{G}''$ a counterfactual (distance 4). Blue/dotted-red edges denote additions/deletions [1].

"closest" isomer that is no longer soluble, e.g. *Methyl-vinyl-ether*, reachable through four modifications. The semifactual thus exposes how much structural variation the model tolerates while preserving the desirable property. □

Contributions. In this paper, we discuss our recent work on semifactual reasoning for graph-structured data [1]. Its main contributions are as follows. We formally define semifactual explanations for graph and node classification in GNNs, show that they generalize factual ones, and characterize the complexity of the associated reasoning and optimisation problems (Section 2). We then propose SEMIX, a learning-based architecture that computes semifactuals by training a neural model under a loss enforcing their defining properties (Section 3). Finally, we evaluate the approach on standard benchmarks, showing its effectiveness in both the semifactual and the factual settings (Section 4).

2. Problem Definition and Complexity Results

We represent an *attributed undirected graph* (AUG) as $\mathcal{G} = (V, E, \mathbf{X})$, where $V = \{1, \ldots, n\}$ is the set of nodes, $E \subseteq V \times V$ a symmetric set of edges, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the node feature matrix; we also denote an AUG by $\mathcal{G} = (\mathbf{A}, \mathbf{X})$, with $\mathbf{A} \in \{0,1\}^{n \times n}$ the adjacency matrix. A GNN Φ learns node representations by iteratively aggregating information from each node's neighbourhood; after κ message-passing layers, node embeddings are pooled into a graph embedding $\Phi(\mathcal{G})$ (resp. $\Phi(\mathcal{G}, u)$ for a node u) that feeds a downstream classifier. For a graph $\mathcal{G}$, the GNN induces a probability distribution over class labels C; we write $P(\hat{y}_\mathcal{G} = c \mid \Phi(\mathcal{G}))$ for the probability it assigns class $c \in C$ to $\mathcal{G}$, and $\ell^1_\Phi(\mathcal{G})$, $\ell^2_\Phi(\mathcal{G})$ for the most and second-most probable labels (and $\ell^1_\Phi(\mathcal{G}, u)$ for node classification). We assume a trained GNN Φ is given, and measure the distance between two AUGs $\mathcal{G} = (V, E, \mathbf{X})$ and $\mathcal{G}' = (V, E', \mathbf{X})$ as the cardinality of the symmetric difference of their edge sets, $|E \,\Delta\, E'| = |(E \setminus E') \cup (E' \setminus E)|$.

Intuitively, a semifactual explanation is a graph obtained from the input one by performing as many perturbations—edge additions and deletions—as possible, while keeping the same predicted class. In practice, only some changes may be admissible; this is captured by a *perturbation set* $E^p \subseteq V \times V$ of allowed edge modifications, which endows explanations with an actionability property by design [29].

Definition 1 (Semifactual for Graph Classification). Let $\mathcal{G} = (V, E, \mathbf{X})$ be an AUG, Φ a GNN, and $E^p \subseteq V \times V$ a perturbation set. A *semifactual explanation* for $\mathcal{G}$ w.r.t. Φ is a graph $\mathcal{G}' = (V, E', \mathbf{X})$ such that *(i)* $\ell^1_\Phi(\mathcal{G}') = \ell^1_\Phi(\mathcal{G})$, *(ii)* $(E \,\Delta\, E') \subseteq E^p$, and *(iii)* there is no $\mathcal{G}''$ with $|E \,\Delta\, E''| > |E \,\Delta\, E'|$ satisfying *(i)* and *(ii)*.

A *factual* explanation, which seeks the smallest subgraph independently reproducing the prediction, is recovered from Definition 1 by allowing only edge deletions ($E^p = E$); semifactuals thus *generalise* factual explanations [15, 10, 11, 16]. Semifactuals for *node* classification are defined analogously, with a target node v given as input and $\ell^1_\Phi(\cdot)$ replaced by $\ell^1_\Phi(\cdot, v)$.

We next focus on the optimisation problem that the heuristic of Section 3 ultimately addresses: computing a maximum-distance semifactual for graph classification, which we call SG.

Problem 1 (SG). Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$, a GNN Φ, and a perturbation set $E^p \subseteq V \times V$, find a graph $\mathcal{G}^* = (V, E^*, \mathbf{X})$ such that:

$$E^* = \underset{\substack{E' \subseteq V \times V \text{ subject to} \\ (E \,\Delta\, E') \subseteq E^p,\ \ell^1_\Phi(\mathcal{G}') = \ell^1_\Phi(\mathcal{G})}}{\arg\max} |E \,\Delta\, E'|,$$

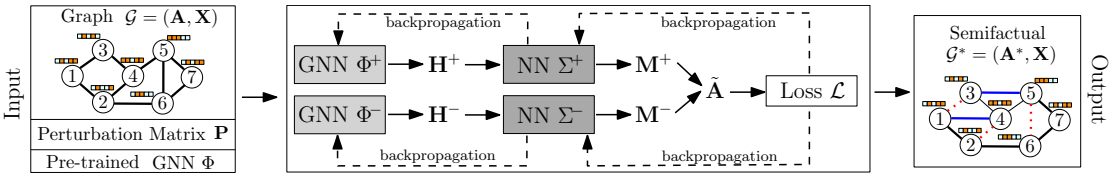

Figure 2: Overview of the proposed SEMIX. Solid/dashed arrows denote the forward/backward passes. Blue and dotted-red edges (in the output graph) denote additions and deletions w.r.t. the input graph [1].

Theorem 1. SG is $\mathbf{FP}^{\mathbf{NP}[\log n]}$-complete.

Its decision version—does a semifactual at distance at least k exist?—is **NP**-complete [27]. Similar complexity results hold for the *counterfactual* problem, where one seeks the *minimum* number of edge changes that *flip* the prediction, and for the *node-classification* setting, where a target node v is fixed and $\ell^1_\Phi(\cdot)$ is replaced by $\ell^1_\Phi(\cdot, v)$: in every case the optimisation problem is $\mathbf{FP}^{\mathbf{NP}[\log n]}$-complete and its decision version **NP**-complete [27]. In particular, this tight $\mathbf{FP}^{\mathbf{NP}[\log n]}$ characterisation strengthens the **NP**-hard bound previously known for counterfactual node classification [19], and shows that semifactual reasoning is exactly as hard as counterfactual reasoning.

3. Method

Motivated by the intractability above, the discussed work proposes a learning-based approach, termed SEMIX (Figure 2). Following the bulk of perturbation-based explanation literature [30, 17], an explanation is modelled as a mask matrix $\mathbf{M} \in [-1, 1]^{n \times n}$, where $|\mathbf{M}_{uv}|$ is the probability that edge (u, v) is added or removed, constrained by a *perturbation matrix* $\mathbf{P} \in \{-1, 0, 1\}^{n \times n}$ encoding the admissible additions $(+1)$, deletions (-1), and fixed edges (0). Writing $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{M} \in [0, 1]^{n \times n}$ for the resulting (continuous) perturbed adjacency matrix, a semifactual is read off as $\mathbf{A}^* = \mathrm{round}(\tilde{\mathbf{A}})$, where $\mathrm{round}(\cdot)$ maps each entry to the nearest value in $\{0, 1\}$.

Rather than treating $\mathbf{M}$ as free parameters, SEMIX produces it through a neural model f_θ with two parallel branches, one generating an insertion mask $\mathbf{M}^+$ and one a deletion mask $\mathbf{M}^-$, later combined into $\mathbf{M} = \mathbf{M}^+ + \mathbf{M}^-$. In each branch, a dedicated κ-layer GNN $\Phi^\pm$ processes the graph topology and node features and outputs node embeddings $\mathbf{H}^\pm = [\Phi^\pm(\mathcal{G}, u)]_{u \in V} \in \mathbb{R}^{n \times d_H}$, where each row $h^\pm_u$ is a d_H-dimensional representation of node u (with d_H possibly differing from the input dimension d) that captures the information needed to decide on perturbations of the edges incident to u. For each candidate edge (u, v), the pair $[h^\pm_u, h^\pm_v]$ is concatenated and fed to a fully connected module $\Sigma^\pm$ followed by a sigmoid σ, with the perturbation matrix $\mathbf{P}$ gating the output so that only admissible edges receive a non-zero score:

$$\mathbf{M}^+_{uv} = \mathtt{ord}(\mathbf{P}_{uv}{=}1)\, \sigma\big(\Sigma^+([h^+_u, h^+_v])\big), \quad \mathbf{M}^-_{uv} = -\,\mathtt{ord}(\mathbf{P}_{uv}{=}{-1})\, \sigma\big(\Sigma^-([h^-_u, h^-_v])\big),$$

where the operator $\mathtt{ord}(\cdot)$ returns 1 if its argument is true and 0 otherwise. The parameters θ are optimised end-to-end on a loss capturing the two requirements of a semifactual—preserving the class (by minimizing $\mathcal{L}_{\text{SF}}$) and maximising the perturbation (by minimizing $\mathcal{L}_{\text{SIM}}$):

$$\mathcal{L} = \mathcal{L}_{\text{SF}} + \lambda\, \mathcal{L}_{\text{SIM}}, \quad \mathcal{L}_{\text{SF}} = \sigma\big(\tau(\gamma + D)\big), \quad \mathcal{L}_{\text{SIM}} = 1 - \frac{\|\mathbf{M}\|_{1,1}}{\|\mathbf{P}\|_{1,1}}. \tag{1}$$

Here σ is the sigmoid, $\tau \geq 1$ a temperature, γ a margin, and

$$D = P\big(\hat{y}_{(\tilde{\mathbf{A}}, \mathbf{X})} {=} \ell^2_\Phi(\mathcal{G}) \mid \Phi(\tilde{\mathbf{A}}, \mathbf{X})\big) - P\big(\hat{y}_{(\tilde{\mathbf{A}}, \mathbf{X})} {=} \ell^1_\Phi(\mathcal{G}) \mid \Phi(\tilde{\mathbf{A}}, \mathbf{X})\big)$$

is the difference, on the perturbed graph $(\tilde{\mathbf{A}}, \mathbf{X})$, between the probabilities it assigns to the original graph's second- and top-ranked classes $\ell^2_\Phi(\mathcal{G})$ and $\ell^1_\Phi(\mathcal{G})$; thus $\mathcal{L}_{\text{SF}}$ stays small (its sigmoid argument negative) while the original label is preserved. $\mathcal{L}_{\text{SIM}}$, with $\|\cdot\|_{1,1}$ the entrywise L_1 norm, decreases as more admissible perturbations are applied, and $\lambda > 0$ balances the two.

	Semifactual setting (SCORE)				Factual setting (sufficiency)						
Dataset	SEMIX[D]	SEMIX[NN]	SEMIX[1GNN]	SEMIX	SEMIX	PGE	TAGE	CF^2	GNNE	GEM	SUBGX
Mutagenicity	.700	.720	.760	**.821**	.756	.585	.775	.726	.725	.736	**.779**
Mutag	.623	.734	.745	**.798**	**.783**	.723	.723	.697	.690	.610	.740
NCI1	.720	.757	.780	**.842**	**.823**	.663	.666	.628	.662	.617	.720
Proteins	.928	.892	.930	**.985**	.912	.931	.948	.971	**.975**	.913	.945
AIDS	.960	.973	.975	**.986**	.971	.957	.960	.981	.983	.943	**.990**
Graph-SST2	.922	**.954**	**.954**	**.954**	.965	.944	.974	.981	.982	—	**.988**
ogbg-molhiv	.882	.971	.973	**.984**	.981	**.983**	.982	.974	.975	—	—

Table 1
Explanation quality of SEMIX and competitors. Left: semifactual SCORE for SEMIX and its ablation variants. Right: sufficiency in the factual setting ($E^p = E$) for SEMIX and factual explainers PGE [11], TAGE [14], CF^2 [15], GNNE [10], GEM [12], and SUBGX [13]. Best per row *within each block* in bold; "—" denotes an out-of-memory run.

4. Experiments

The approach was evaluated on seven standard benchmark datasets for graph classification [31], using two metrics: *irrelevance* (the fraction of graphs whose explanation preserves the predicted label) and *distance* (how far the explanation is from the input), summarised by their harmonic mean SCORE; in the factual setting we report *sufficiency*, i.e. irrelevance under deletion-only perturbations. To assess the contribution of each component, an ablation compares the full two-branch model SEMIX against three progressively simplified variants: SEMIX[1GNN], which replaces the two branch-specific GNNs with a single GNN shared by both NN modules; SEMIX[NN], which removes the GNNs altogether and feeds the raw node features directly to the NN modules; and SEMIX[D], which drops both the GNNs and the NN modules, optimising the mask matrix $\mathbf{M}$ directly as free parameters via projected gradient descent.

Table 1 reports the results. In the semifactual setting (left block), SEMIX attains the best SCORE on every dataset, averaging .910 against .874, .857, and .819 for the variants—gains of 4.5%, 6.7%, and 12.9% over SEMIX[1GNN], SEMIX[NN], and SEMIX[D], with the largest gains on the structurally richer molecular datasets (Mutagenicity, Mutag, NCI1). In the factual setting (right block, $E^p = E$), SEMIX stays competitive with purpose-built factual explainers [15, 10, 11, 14], reaching the highest average sufficiency overall (0.884), ahead of the second-best method TAGE (0.861).

5. Conclusions

We have discussed a recently proposed framework [1] that brings semifactual "even-if" reasoning to Graph Neural Networks. Promising directions include alternative graph distances, node/edge-feature perturbations, preference-based explanation selection [24], decentralised settings [32] and improving the training efficiency [33].

Acknowledgments

We acknowledge financial support from PNRR MUR projects FAIR (PE0000013) and SERICS (PE00000014), project Tech4You (ECS0000009), and (MUR) PRIN 2022 project S-PIC4CHU (2022XERWK9).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude for grammar and spelling checking. The author(s) reviewed and edited the content and take(s) full responsibility for it.

References

[1] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, R. Shahbazian, I. Trubitsyna, Semifactual explanations for GNN-based classification: Formal foundations, complexity and computation, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2026.

[2] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 1082–1090.

[3] J. You, B. Liu, Z. Ying, V. Pande, J. Leskovec, Graph convolutional policy network for goal-directed molecular graph generation, Advances in Neural Information Processing Systems 31 (2018).

[4] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 34 (2018) i457–i466.

[5] F. Gullo, D. Mandaglio, A. Tagarelli, Neural discovery of balance-aware polarized communities, Machine Learning 113 (2024) 6611 – 6644. doi:10.1007/s10994-024-06581-4.

[6] L. Martirano, F. Scala, C. Comito, L. Pontieri, Who drives misinformation? key node detection with heterogeneous graph neural networks, in: International Conference on Discovery Science, Springer, 2025, pp. 47–62.

[7] L. La Cava, D. Mandaglio, L. Zangari, A. Tagarelli, Heuristic-informed mixture of experts for link prediction in multilayer networks, Information Sciences 736 (2026). doi:10.1016/j.ins.2026.123106.

[8] L. Zangari, D. Mandaglio, A. Tagarelli, Link prediction on multilayer networks through learning of within-layer and across-layer node-pair structural features and node embedding similarity, 2024, p. 924 – 935. doi:10.1145/3589334.3645646.

[9] H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in graph neural networks: A taxonomic survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 5782–5799.

[10] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, Advances in Neural Information Processing Systems 32 (2019).

[11] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, Parameterized explainer for graph neural network, Advances in Neural Information Processing Systems 33 (2020) 19620–19631.

[12] W. Lin, H. Lan, B. Li, Generative causal explanations for graph neural networks, in: International Conference on Machine Learning, PMLR, 2021, pp. 6666–6679.

[13] H. Yuan, H. Yu, J. Wang, K. Li, S. Ji, On explainability of graph neural networks via subgraph explorations, in: International Conference on Machine Learning, PMLR, 2021, pp. 12241–12252.

[14] Y. Xie, S. Katariya, X. Tang, E. Huang, N. Rao, K. Subbian, S. Ji, Task-agnostic graph explanations, Advances in Neural Information Processing Systems 35 (2022) 12027–12039.

[15] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, Y. Zhang, Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning, in: Proceedings of the ACM Web Conference, 2022, pp. 1018–1027.

[16] S. Azzolin, S. Malhotra, A. Passerini, S. Teso, Beyond topological self-explainable gnns: A formal explainability perspective, in: Proceedings of International Conference on Machine Learning (ICML), 2025, pp. 2045–2080.

[17] M. A. Prado-Romero, B. Prenkaj, G. Stilo, F. Giannotti, A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges, ACM Computing Surveys 56 (2024) 1–37.

[18] M. A. Prado-Romero, B. Prenkaj, G. Stilo, Robust stochastic graph generator for counterfactual explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 21518–21526.

[19] S. Verma, B. Armgaan, S. Medya, S. Ranu, Induce: Inductive counterfactual explanations for graph neural networks, Transactions on Machine Learning Research (2024).

[20] P. Barceló, M. Monet, J. Pérez, B. Subercaseaux, Model interpretability through the lens of computational complexity, in: Proceedings of Advances in Neural Information Processing Systems, 2020.

[21] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2023, pp. 6526–6535.
[22] E. M. Kenny, W. Huang, The utility of "even if" semi-factual explanation to optimise positive outcomes, in: Thirty-seventh Conference on Neural Information Processing Systems, 2023.
[23] R. McCloy, R. M. Byrne, Semifactual "even if" thinking, Thinking & reasoning 8 (2002) 41–67.
[24] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, R. Shahbazian, I. Trubitsyna, Even-if explanations: Formal foundations, priorities and complexity, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 15347–15355.
[25] S. Dandl, G. Casalicchio, B. Bischl, L. Bothmann, Interpretable regional descriptors: Hyperbox-based local explanations, in: Proceedings of Machine Learning and Knowledge Discovery in Databases, volume 14171, Springer, 2023, pp. 479–495.
[26] A. Artelt, M. Olsen, K. Tierney, On the hardness of computing counterfactual and semi-factual explanations in XAI, Transactions on Machine Learning Research (2025). URL: https://openreview.net/forum?id=aELzBw0q1O.
[27] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Counterfactual and semifactual explanations in abstract argumentation: Formal foundations, complexity and computation, in: Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR), 2024, pp. 14–26.
[28] G. Alfano, A. Gould, F. Leofante, A. Rago, F. Toni, Counterfactual explanations under model multiplicity and their use in computational argumentation, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), ijcai.org, 2025, pp. 4321–4329. doi:10.24963/IJCAI.2025/481.
[29] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery 38 (2024) 2770–2824.
[30] Z. Guo, Z. Wu, T. Xiao, C. Aggarwal, H. Liu, S. Wang, Counterfactual learning on graphs: A survey, Machine Intelligence Research 22 (2025) 17–59.
[31] M. Kosan, S. Verma, B. Armgaan, K. Pahwa, A. Singh, S. Medya, S. Ranu, Gnnx-bench: Unravelling the utility of perturbation-based gnn explainers through in-depth benchmarking, in: International Conference on Learning Representations, 2024.
[32] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, R. Shahbazian, I. Trubitsyna, Decentralized federated learning meets physics-informed neural networks, Knowledge-Based Systems 323 (2025) 113717.
[33] F. Scala, S. Flesca, L. Pontieri, An efficient model training framework for green ai, Machine Learning 114 (2025) 275.