ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Graph Diffusion Counterfactual Explanation

David Bechtoldt and Sidney Bender *

Machine Learning Group, TU Berlin, 10587 Berlin, Germany

Abstract. Machine learning models that operate on graph-structured data, such as molecular graphs or social networks, often make accurate predictions but offer little insight into *why* certain predictions are made. Counterfactual explanations address this challenge by seeking the closest alternative scenario where the model's prediction would change. Although counterfactual explanations are extensively studied in tabular data and computer vision, the graph domain remains comparatively underexplored. Constructing graph counterfactuals is intrinsically difficult because graphs are discrete and non-euclidean objects. We introduce Graph Diffusion Counterfactual Explanation, a novel framework for generating counterfactual explanations on graph data, combining discrete diffusion models and classifier-free guidance. We empirically demonstrate that our method reliably generates in-distribution as well as minimally structurally different counterfactuals for both discrete classification targets and continuous properties.

1 Introduction

Explainable AI (XAI) aims to make the behavior of complex machine-learning models transparent and interpretable for humans [1]. Among existing approaches, counterfactual explanations have emerged as a particularly intuitive and actionable concept. They ask what minimal change to the input would lead to a different prediction, thereby providing insight into model behavior, robustness, and bias [2]. Although early counterfactual explanation methods largely focused on tabular data, the computer vision community has pioneered generative approaches that better align with our goals of realism and interpretability. These vision methods generate counterfactuals in the latent space of a model instead of directly perturbing input features, yielding semantic modifications that remain in the data manifold [3, 4]. For example, Jeanneret et al. (2024) [3] propose *TIME*: a method using a diffusion model with learned textual embeddings and inversion to generate counterfactuals that remain in-distribution and without access to model gradients. Conceptually, this echoes the idea by Dombrowski et al. (2024) [4], which advocates optimization in a generator-induced latent coordinate system so that counterfactual trajectories remain on the data manifold. Transferring these ideas to graphs is non-trivial because graphs are discrete, combinatorial, and non-Euclidean. Gradients in input space are ill-defined, and naive edit search becomes intractable for larger structures such as molecules [5]. Discrete diffusion models such as *DiGress* [6] partly address this by defining Markov transitions over node and edge categories and learning to denoise them

*This work was supported by BASLEARN-TU Berlin/BASF Joint Laboratory, co-financed by TU Berlin and BASF SE.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with a graph transformer. Beyond counterfactual reasoning, there also exist successful XAI methods for graphs, e.g., attribution-based approaches such as *GNN-LRP* [7], but here we focus exclusively on counterfactual explanations.

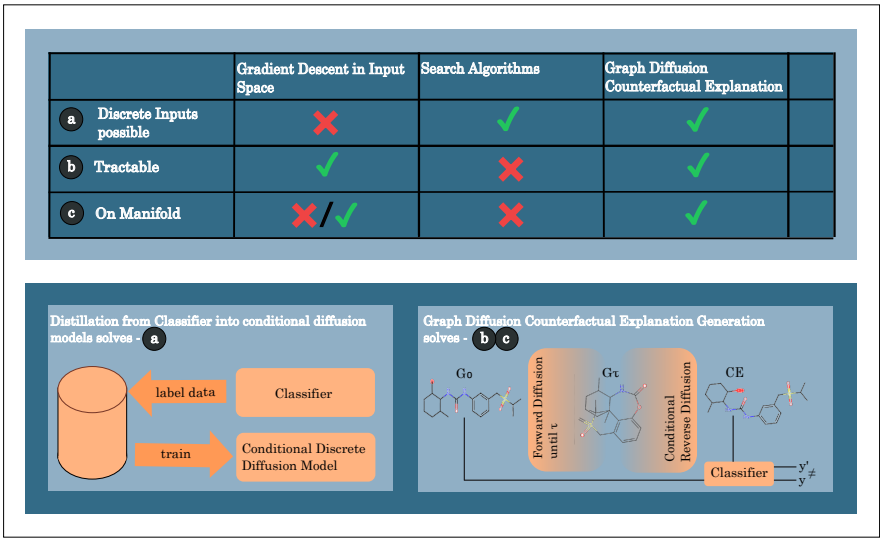

Fig. 1: Top: Comparison across (a) support for discrete inputs, (b) tractable generation, and (c) on-manifold solution of early gradient descent methods in input space, search algorithms, and our method. Bottom-left: distillation of label information into a conditional discrete diffusion model. Bottom-right: *GDCE* generation pipeline.

We present a diffusion-based framework for graph counterfactuals that unifies on-manifold generation with tractable control (Fig. 1). Concretely, we train a discrete graph diffusion model conditioned on the desired feature and thereby circumvent the lack of gradients in discrete graph space. The feature can be a classifier label or any other feature of interest. Further we replace auxiliary-classifier guidance with classifier-free guidance for graphs introduced by Ninniri et al. (2024) [5] by injecting the target condition directly into the generative model. During inference, we perturb the input graph partway along a diffusion trajectory and then guide reverse denoising toward the desired property, yielding counterfactual graphs that flip the prediction, remain close to the original graph, and respect the data manifold. In particular, there are existing counterfactual explanations for graph-structured data, such as *D4Explainer* by Chen et al. [8], but given the absence of evaluation in large-scale datasets such as *QM9* and *ZINC-250k*, we focus on ablations within the method and diagnostic checks.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Method: Graph Diffusion Counterfactual Explanation

Preliminaries: Let a graph $G = (X, E)$ be represented by one-hot encoded node and edge feature matrices $\mathbf{X} \in \mathbb{R}^{n \times a}$ and $\mathbf{E} \in \mathbb{R}^{n \times n \times b}$. Discrete diffusion defines Markov transitions over nodes and edges via transition matrices $[Q_X^t]_{ij} = q(x^t = j | x^{t-1} = i)$ and $[Q_E^t]_{ij} = q(e^t = j | e^{t-1} = i)$ [6]. The forward process applies cumulative transitions $\bar{Q}^t = Q^1 Q^2 ... Q^t$, producing noisy $G_t = (X\bar{Q}^t{}_X, E\bar{Q}^t{}_E)$. A neural denoiser parameterizes reverse transitions to sample G_{t-1} from G_t. While G_0 is a graph from the data distribution, and G_T is a complete, noisy random graph. **Classifier-free guidance for graphs**: We condition the denoiser on a target y (discrete class or continuous property) with conditioning dropout training so the same network learns both conditional and unconditional denoising. At inference, we form a guided score by linearly combining conditional and unconditional predictions with scale s. **Graph Diffusion Counterfactual Explanation**: Given an observed graph G with label y and a target y', we forward perturb a sample G_τ from $q(G_\tau | G)$ for an intermediate step $\tau \in (0, T]$. This erases fine details while preserving global structure. Secondly, we guide the reverse diffusion from τ to 0 under the condition y' using classifier-free guidance. The resulting Graph G_{CF} is our counterfactual. This design yields manifold-aware edits, meaning that large-scale structure is retained starting from G_τ, while guided denoising injects just enough flexibility to reach the target with minimal edits. The algorithm can be contemplated in Algorithm 1, where f_θ is the model's prediction of the true node and edge types, and p_θ is the model's predicted posterior.

Algorithm 1: Graph Diffusion Counterfactual Generation

Input: A graph $G = (X, E, y)$, with original feature y, target y', diffusion steps τ

Output: A graph $G_{CE} = (X, E, y')$

Compute noisy graph $G_\tau \sim (X\bar{Q}_X^\tau, E\bar{Q}_E^\tau)$ **for** $t = \tau$ **to** 1 **do**

  $\hat{p}_X, \hat{p}_E \leftarrow \phi_\theta(G_t)$; `// Reverse pass`

  $p_\theta(x_i^{t-1} \mid G_t, y_1) \leftarrow \sum_x q(x_i^{t-1} \mid x_i^t, x_i^0 = x) \cdot f_\theta(x_i^0 = x \mid G_t, y')$;

  $p_\theta(e_{ij}^{t-1} \mid G_t, y_1) \leftarrow \sum_e q(e_{ij}^{t-1} \mid e_{ij}^t, e_{ij}^0 = e) \cdot f_\theta(e_{ij}^0 = e \mid G_t, y')$;

  $G_{t-1} \sim \prod_i p_\theta(x_i^{t-1} \mid G_t, y') \cdot \prod_{i,j} p_\theta(e_{ij}^{t-1} \mid G_t, y')$;

end

return $G = (X, E, y')$

3 Proof of Concept: Planar Graphs

Data: We evaluate on all non-isomorphic planar graphs with 8 nodes. A graph is planar if it can be drawn without edge crossings on a plane. This domain lets us verify that outputs remain planar and quantify minimal structural change precisely.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Experimental setup: We train a classifier-free guided discrete diffusion model conditioned on the edge count y. For each test graph G, we request a counterfactual with $y' = |E| - 1$, generating graphs with exactly 8 nodes and the prescribed edge count. Per input, we sample 100 counterfactuals across $\tau \in \{1, 5, 10, 25, 50, 100, 200\}$. We report (i) structural validity as connected and planar, (ii) target accuracy among valid graphs, and (iii) mean Graph Edit Distance to the original. As a baseline, we simply generate 100 samples with *FreeGress* under the same y'.

Table 1: Graph Diffusion Counterfactual Generation for Planar Graphs

$y_{\text{target}} = \lvert E\rvert - 1$	**FreeGress**	τ=**1**	τ=**5**	τ=**10**	τ=**50**	τ=**100**	τ=**200**
Validity	0.78	0.14	0.55	0.87	**0.91**	0.89	0.79
Accuracy	**1.00**	0.05	0.67	0.96	**1.00**	**1.00**	**1.00**
Mean-GED	4.13	1.00	2.54	2.43	**1.57**	2.13	3.95

Results & discusion: Table 1 shows that our method yields consistently high structural validity, especially for $\tau \geq 50$ and generally outperforms *FreeGress*. Target accuracy is near-perfect at $\tau \geq 50$, while mean GED remains small (1.57). Performance decreases beyond certain noise levels, reflecting the trade-off between edit capacity and closeness. Overall, the approach produces structurally valid counterfactuals with high accuracy and minimal deviations from the original.

4 Steering Molecules into Drug-like logP Ranges

Data: We use *ZINC-250k*, a dataset of 250,00 drug-like molecules with up to 38 heavy atoms and a broad set of physicochemical properties. We target logP, the octanol-water partition coefficient, which governs the hydrophobicity and is central to the drug-likeness. We encode membership in a desirable interval $[1.7, 3.8]$, following the QED desirability range [9].

Experimental setup: We train a classifier-free guided diffusion model conditioned on a binary indicator of whether a molecule's logP lies in $[1.7, 3.8]$. For evaluation, we sample 200 test molecules outside the target range covering both overly hydrophilic and hydrophobic cases and generate 10 counterfactuals per molecule at perturbation levels $\tau = 10, 50, 100$ with $y' \in [1.7, 3.8]$.

We evaluate the outcomes using (i) structural validity, the percentage of generated molecules that are chemically valid as determined by RDKit santization checks ensuring no valence or atom-type violations, (ii) target accuracy, the percentage of valid outputs that successfully have their logP within the desired [1.7, 3.8] range and (iii) structural similarity, a measure of how close the counterfactual molecules are to the original structure. For the smaller planar graphs, we used mean GED to quantify minimal changes. However, the exact GED is computationally intractable for larger graphs since it is NP-hard. Instead, for

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

ZINC-250k we report the mean Tanimoto similarity between the original and generated molecules based on their Morgan fingerprint representations. Again, we report all metrics for *FreeGress* generated samples as a Baseline

Table 2: Graph Diffusion Counterfactual Generation for *ZINC-250k* Molecules targeting desirable logP ranges

$logP \notin [1.7, 3.8] \rightarrow logP \in [1.7, 3.8]$	**FreeGress**	τ=**10**	τ=**50**	τ=**100**
Validity	**0.71**	0.59	0.40	0.16
Accuracy	**0.55**	0.32	0.41	0.50
Similarity	0.04	**0.50**	0.44	0.28

Results & discussion: Table 2 summarizes the performance of our guided diffusion approach in steering *ZINC-250k* molecules into the desired logP range, and Figure 2 showcases *ZINC-250k* counterfactuals. A clear trade-off emerges between achieving the target property and preserving the original structure. At low noise perturbation $\tau = 10$, our method produces counterfactuals with high structural similarity to the original molecule, with a mean Tanimoto similarity of 0.5, demonstrating that only minimal edits were made, but in comparison to the baseline with hampered validity of 59% vs. 71% and accuracy of 32% vs. 55%. As we allow larger perturbations, our approach becomes more effective at altering logP with 50% at $\tau = 100$, but this comes at the cost of drastically reduced validity of only 16% and reduced Tanimoto similarity of 0.28. These trends are intuitively consistent with the task difficulty that larger structural changes are more likely to achieve a challenging property target, yet they risk breaking chemical rules or straying too far from the original molecule. In contrast, the *FreeGress* baseline, which freely generates new structures to satisfy the property, achieves higher validity and accuracy overall but produces molecules essentially unrelated to the input, as expected.

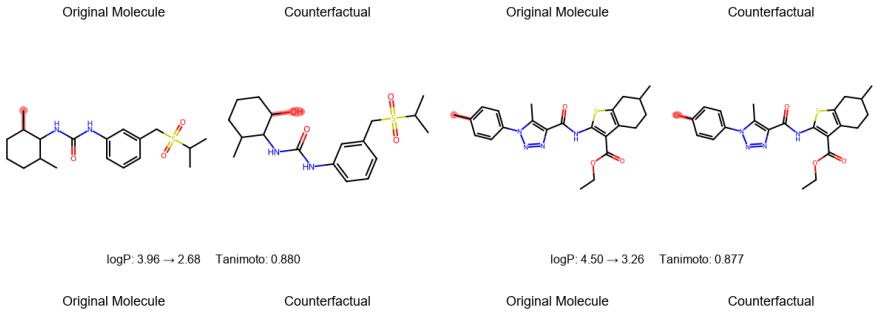

Fig. 2: Two *ZINC-250k* counterfactual examples that steer a molecule's logP into the prespecified target range.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

5 Conclusion

We presented Graph Diffusion Counterfactual Explanation, a classifier-free guided discrete diffusion framework for generating counterfactuals on graphs. The method integrates partial-noising edits with guided denoising, yielding on-manifold modifications that respect domain constraints while steering predictions toward desired targets. Across synthetic planar graphs and large drug like molecules from *ZINC-250k*. Our method produced valid, accurate, and structurally close alternatives. As expected, results reveal a fundamental trade-off: tightening similarity constraints limits achievable property shifts, whereas allowing larger perturbations improves target success at the cost of similarity and, if pushed too far, validity. Crucially, the diffusion prior helps navigate this tension by biasing edits toward realistic, high-probability regions of the graph manifold. Methodologically, our approach unifies discrete changes and continuous property conditioning within a single, streamlined generator. Overall, these findings establish diffusion-based counterfactuals as a practical and scalable tool for interrogating and shaping model behavior in the graph domain, enabling users to ask *what needs to change* and obtain actionable, domain-valid edits.

References

[1] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[2] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[3] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4757–4767, 2024.

[4] Ann-Kathrin Dombrowski, Jan E. Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic counterfactuals with generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):3257–3274, May 2024.

[5] Matteo Ninniri, Marco Podda, and Davide Bacciu. Classifier-free graph diffusion for molecular property targeting. In Albert Bifet, Jesse Davis, Tomas Krilavičius, Meelis Kull, Eirini Ntoutsi, and Indrė Žliobaitė, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 318–335, Cham, 2024. Springer Nature Switzerland.

[6] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schutt, Klaus-Robert Müller, and Grégoire Montavon. Higher-Order Explanations of Graph Neural Networks via Relevant Walks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(11):7581–7596, November 2022.

[8] Jialin Chen, Shirley Wu, Abhijit Gupta, and Zhitao Ying. D4explainer: In-distribution explanations of graph neural network via discrete denoising diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.