FROM RELIABLE EXPLANATIONS TO BETTER LEARNING: BRIDGING THE OOD GAP IN 3D GNN EXPLANATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural network (GNN) explainers aim to identify subgraphs that preserve model predictions and have been widely studied on 3D graphs. However, two fundamental questions remain open: (1) *Are these explanations reliable?* and (2) *What benefits can they bring to model learning?* In this paper, we show that these two questions are inherently connected through an overlooked issue in 3D GNNs: explanatory subgraphs are fundamentally out-of-distribution (OOD). Due to the cutoff-based construction of 3D proximity graphs, existing explainers break the underlying geometric constraints, producing subgraphs that violate the training distribution. As a result, the predictor behaves unreliably on such inputs, leading to uninformative or even misleading explanations. To address this, we propose **E2L**, an explanation-aware framework that bridges the OOD gap by aligning the training and explanation distributions. By integrating the explainer into the learning process, the model is directly optimized on the explanatory distribution, ensuring that explanations are in-distribution and reliable for the predictor. Building on this foundation, we further bridge the gap between explanation and representation learning, demonstrating that reliable explanations naturally induce compact and expressive subgraph structures that provide both efficiency and expressivity benefits for representation learning. Experiments on both small and large-scale 3D molecular graph benchmarks show that E2L overcomes the OOD issue, improves explanation reliability, and achieves substantial efficiency and expressivity gains. The code is released at `https://anonymous.4open.science/r/E2L-28EE`.

1 INTRODUCTION

3D graph neural networks (GNNs) have achieved remarkable success in modeling structured data in domains such as molecular modeling (Liu et al., 2022b; Schütt et al., 2017; Gasteiger et al., 2020; Brandstetter et al., 2022) and protein analysis (Zhang et al., 2023b; Jing et al., 2021). As these models grow increasingly complex, understanding their decision-making process becomes critical. To this end, GNN explainers identify compact subgraphs most responsible for a model's prediction, providing insights into model behavior and reliability (Ying et al., 2019; Luo et al., 2020; Miao et al., 2023; Qu et al., 2025). Despite significant progress, two fundamental questions remain unresolved in 3D GNN explanation: *(1) Are the explanations reliable?* and *(2) Can explanations benefit representation learning beyond interpretability?*

Regarding explanation reliability, most existing explainers operate on a post-hoc paradigm (Ying et al., 2019; Luo et al., 2020; Liu et al., 2025): a pretrained predictor is fixed, and an explainer is trained to extract sparse subgraphs that preserve its predictions as explanations. This implicitly assumes that the predictor generalizes to such subgraphs. However, in 3D settings, this assumption breaks down. Due to the cutoff-based construction of 3D proximity graphs (Schütt et al., 2017; Gasteiger et al., 2020), as defined in Sec. 3.2, extracted subgraphs typically violate the underlying graph construction rules. This directly results in an explanation-induced *out-of-distribution (OOD) shift*. Consequently, predictions on these subgraphs become unreliable, leading to unreliable explanations with large numerical errors. For example, on the QM9 dataset (Ramakrishnan et al., 2014) for dipole moment prediction, a pretrained model achieves a mean absolute error (MAE) of 0.03 on full graphs, while the same model evaluated on state-of-the-art (SOTA) post-hoc explanatory subgraphs yields an MAE as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

high as 0.42 (details in Sec. 5.1), clearly indicating a severely unreliable explanation. The explanation-induced OOD shift leads to unreliable model behavior on extracted subgraphs, undermining their trustworthiness for further application (Zheng et al., 2024; Chen et al., 2024). Meanwhile, the post-hoc paradigm fundamentally limits the practical usability of explanations for representation learning. These two issues constitute fundamental barriers between 3D GNN explanation and learning, whose connection remains largely unexplored in prior work.

To address these challenges, we revisit 3D GNN explanation from a unified perspective that jointly considers reliability and its role in learning. Our contributions are as follows: (1) we characterize the explanation-induced OOD issue and empirically demonstrate its severity in 3D GNN explanations; (2) we propose **E2L**, which fundamentally overcomes this OOD issue by unifying the training and explanation distributions to produce reliable explanations; (3) building upon E2L, we establish a principled connection between explanation and representation learning, showing for the first time that reliable explanations can serve as excellent representations, which improve both expressivity and computational efficiency. To the best of our knowledge, this is the first framework that explicitly bridges the OOD gap in 3D GNN explanation and connects explanation to representation learning. Accordingly, E2L is not designed to reproduce the decision process of a particular pretrained parameter realization; instead, it reformulates 3D GNN explanation as identifying the compact structure that the model should rely on during learning.

2 Related Work

GNN explanation in 2D and 3D graphs. GNN explanation has been extensively studied in 2D graphs. At the instance level, GNNExplainer (Ying et al., 2019) establishes the standard formulation by learning edge and node masks to extract a prediction-preserving subgraph. Follow-up works (Zhang et al., 2025; Yuan et al., 2021; Zhang et al., 2021) build on this paradigm to identify the most relevant input components for individual predictions. In parallel, model-level approaches (Yuan et al., 2020; Vasilcoiu et al., 2024; Lv et al., 2022) aim to capture global patterns of a trained model in inductive or distributional settings. Since these methods target a different scope, we focus on instance-level explanation. In contrast, **3D GNN explanation remains relatively underexplored** due to fundamental differences in graph construction and geometric constraints. In particular, 3D graphs are typically constructed using distance thresholds, which impose strict structural rules on valid graphs. Existing methods primarily adapt masking to 3D settings, including probabilistic masking tailored to spatial structures (Miao et al., 2023), node-wise masking (Liu et al., 2025), and radii optimization-based methods (Qu et al., 2025).

OOD in 2D GNN explanation. The OOD issue is increasingly recognized in 2D GNN explanation from both evaluation and generation perspectives. Prior studies show that masking or removal-based evaluation can introduce distribution shifts and distort explanation fidelity (Wu et al., 2022a; Amara et al., 2023; Zheng et al., 2024), with recent work formalizing structural distribution shifts between original and explanatory graphs (Chen et al., 2026b). Other methods alleviate this mismatch through graph augmentation or in-distribution generation, including MixupExplainer, proxy-graph methods, and diffusion-based explainers (Zhang et al., 2023a; Chen et al., 2024; 2023; 2026a). However, these approaches are primarily designed for 2D graph topology and often rely on augmented or generated graphs, which may no longer correspond to strict subgraphs of the original input. This challenge is amplified in 3D graphs, where graph augmentation or generation must also satisfy the underlying geometric principles. Constructing in-distribution proxy graphs while preserving these geometric constraints is non-trivial. Therefore, E2L avoids this difficulty by directly aligning the training and explanatory distributions without additional graph generation.

Explanation-aware graph learning. Explanation-aware learning has been extensively explored in 2D graphs, including stochastic-attention methods such as GSAT (Miao et al., 2022), information-bottleneck subgraph learning (Yu et al., 2021), prototype-based self-explaining models (Zhang et al., 2022), and graph rationalization methods (Wu et al., 2022b; Liu et al., 2022a). These methods are not specifically designed for dense 3D proximity graphs and generally require nontrivial adaptation to preserve geometry-dependent computation. More importantly, they typically introduce intrinsic explanation mechanisms rather than addressing how a pre-designed explainer can benefit predictor learning. E2L addresses this complementary problem through a distribution-alignment framework that connects an explainer with the predictor, allowing geometry-aware 3D explanations to directly participate in representation learning. Detailed conceptual and empirical comparisons appear in Appendices C and F.4.

3 PRELIMINARIES

3.1 GNN EXPLANATION

GNN explanation seeks a compact subgraph that retains information relevant to a specified target. Let a pretrained GNN Φ_γ with parameters γ map a graph G to a prediction $\hat{Y} = \Phi_\gamma(G)$, and let Y denote the ground-truth target. A graph is represented as $G = (V, A, Z)$, where V denotes the node set, $A \in \{0,1\}^{n\times n}$ is the adjacency matrix, and $Z \in \mathbb{R}^{n\times d}$ is the node feature matrix. For a chosen target T and a fixed predictor, subgraph selection can be formulated as:

$$G_S^* = \arg\min_{G_S \subseteq G} \mathcal{L}\big(T, \Phi_\gamma(G_S)\big) \quad \text{s.t.} \quad |G_S| \le B, \tag{1}$$

where $\mathcal{L}$ is a target-dependent loss and B controls the explanation size. For post-hoc prediction explanation, $T = \hat{Y}$ encourages fidelity to the pretrained model, as in GNNExplainer (Ying et al., 2019). Choosing $T = Y$ instead evaluates whether the subgraph supports accurate task prediction, reflecting the predictive compression principle underlying the graph information bottleneck (GIB) (Wu et al., 2020; Yu et al., 2021). Here, the size constraint enforces structural compactness rather than an explicit mutual-information constraint. In practice, discrete subgraph selection is commonly relaxed using a continuous edge mask $M \in [0,1]^{n\times n}$ applied to the adjacency matrix:

$$M^* = \arg\min_{M} \mathcal{L}\big(T, \Phi_\gamma(M \odot A, Z)\big) \quad \text{s.t.} \quad \|M\|_1 \le B, \tag{2}$$

where $\odot$ denotes element-wise multiplication. Additional regularization terms may encourage sparsity and discreteness. The optimized mask is then used to extract a subgraph relevant to the chosen target.

3.2 3D PROXIMITY GRAPHS

Many applications in geometric learning represent 3D structures as graphs, where nodes correspond to points (e.g., atoms) and edges encode spatial relationships. A common construction is the *3D proximity graph*, where connectivity is defined by pairwise distances in Euclidean space. Formally, given a set of points with 3D coordinates $X \in \mathbb{R}^{n\times 3}$, a proximity graph is defined as $G = (V, A, X)$, omitting node features Z for brevity. Here, V is the node set and $A \in \{0,1\}^{n\times n}$ is the adjacency matrix determined by a distance-based rule with a fixed cutoff distance r_c, i.e.,

$$A_{ij}^{r_c} = \mathbb{I}\big(\|x_i - x_j\|_2 \le r_c\big), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $x_i, x_j \in \mathbb{R}^3$ denote the coordinates of nodes i and j. This cutoff-based construction is prevalent in geometric deep learning, as it provides a simple and efficient way to encode local interactions. In practice, the cutoff radius r_c is a hyperparameter that controls the sparsity and locality of the resulting graph: larger r_c leads to denser connectivity. In 3D GNNs, r_c is commonly set to large values (e.g., 5–10Å in the QM9 dataset), resulting in **dense** proximity graphs, where each node is connected to multiple nearby nodes to capture local geometric interactions. Sparsification of 3D proximity graphs is thus particularly necessary for improving efficiency. Unless otherwise stated, this paper focuses on explanation in 3D proximity graphs, which are the most widely used representation in 3D GNNs.

Notably, this construction induces an implicit *symmetry bias*: since connectivity is determined by a single global radius, the resulting graph is necessarily undirected with $A_{ij}^{r_c} = A_{ji}^{r_c}$, and all nodes share the same receptive field radius r_c. While this design simplifies modeling, it restricts the space of admissible graph structures to those consistent with uniform, symmetric proximity.

4 E2L: FROM RELIABLE EXPLANATIONS TO BETTER LEARNING

4.1 OOD GAP IN 3D GNN EXPLANATION

Following the standard formulation of GNN explanation (Sec. 3.1), GNN explainers aim to identify a subgraph $G_S \subseteq G$ most relevant to a prediction.

Explanation-induced distribution shift. Although Eq. (1) is well-defined, it implicitly assumes that the pretrained predictor Φ_γ behaves reliably on the extracted subgraphs. We argue that this assumption is fundamentally violated in 3D GNNs due to an explanation-induced OOD shift.

Let $\mathcal{D}$ denote the training distribution over graph-label pairs (G, Y) used to train a GNN Φ_γ. Given an explanation method $\mathcal{E}$, each graph $G \sim \mathcal{D}$ induces an explanatory subgraph $G_S = \mathcal{E}(G; \Phi_\gamma)$, defining the subgraph distribution $\mathcal{D}_\mathcal{E} = \{(G_S, Y) : (G, Y) \sim \mathcal{D}\}$. In general, $\mathcal{D}_\mathcal{E} \neq \mathcal{D}$.

Definition 4.1. To assess predictive reliability on explanatory subgraphs, following the definition of domain shifting risk (Ben-David et al., 2010), we define the task-based explanation risk of a predictor Φ_γ under an explainer $\mathcal{E}$ as

$$R_{\exp}(\Phi_\gamma, \mathcal{E}) = \mathbb{E}_{(G,Y)\sim\mathcal{D}}\big[\mathcal{L}\big(Y, \Phi_\gamma(G_S)\big)\big], \qquad G_S = \mathcal{E}(G; \Phi_\gamma). \tag{4}$$

This quantity is the population counterpart of Eq. 1 with $T = Y$ and measures task prediction error on explanatory subgraphs. It therefore directly measures whether the pretrained model can produce meaningful predictions on explanatory subgraphs. However, Φ_γ is trained only to minimize the in-distribution risk on full graphs, i.e., $R_{\text{train}}(\Phi_\gamma) = \mathbb{E}_{(G,Y)\sim\mathcal{D}}[\mathcal{L}(Y, \Phi_\gamma(G))]$, but is never optimized on subgraphs sampled from $\mathcal{D}_\mathcal{E}$. **When G_S is OOD, the pretrained predictor's output $\Phi_\gamma(G_S)$ can become inaccurate and unstable, leading to a large $R_{\exp}(\Phi_\gamma, \mathcal{E})$.**

Why is the OOD gap particularly severe in 3D GNNs? This issue is significantly exacerbated in 3D graphs due to proximity-based graph construction. Standard 3D GNNs are trained on dense proximity graphs (Sec. 3.2), where edges encode local geometric interactions under a fixed radius constraint. However, explanatory subgraphs are typically obtained via masking, which can violate the underlying geometric principles and produce graphs that no longer satisfy the cutoff-based connectivity. As a result, these subgraphs fall outside the training distribution, leading to a pronounced distribution shift and unreliable model behavior.

We empirically validate this effect via a controlled simulation on molecular graphs in the QM9 dataset (Ramakrishnan et al., 2014). Specifically, we use random edge masking as a proxy for subgraph extraction in existing explainers, treating each masked graph as an explanatory candidate. To quantify the resulting distribution shift, we evaluate distributional consistency with respect to the *underlying graph construction rules*. Chemical validity is used for 2D graphs, while cutoff-based proximity consistency is used for 3D graphs. The definition of distributional consistency and experimental details are provided in Appendix A.

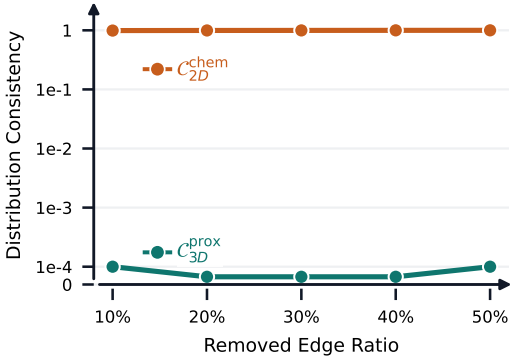

Figure 1: Comparison of simulated explanation processes in 2D chemical graphs and 3D proximity graphs on 10,000 QM9 samples. $\mathcal{C}_{2D}^{\text{chem}}$ and $\mathcal{C}_{3D}^{\text{prox}}$ denote chemical validity and proximity-based consistency, respectively. While 2D explanatory graphs remain largely valid, 3D explanatory graphs exhibit a sharp decline in consistency, indicating a severe distribution shift.

As shown in Fig. 1, from the perspective of graph construction rules, 2D subgraphs maintain high consistency across a wide range of edge removal ratios, indicating that explanations largely remain in-distribution and preserve chemical validity. In contrast, 3D graphs exhibit a rapid decline in consistency even under mild edge removal (e.g., 10%–20%), demonstrating that the resulting subgraphs quickly violate proximity constraints and fall outside the training distribution. The pretrained predictor Φ_γ is then forced to extrapolate on these OOD subgraphs, making $\Phi_\gamma(G_S)$ unreliable and leading to a large explanation risk in Eq. (4). As a result, the optimization of post-hoc explainers is no longer driven by meaningful predictive signals, but by the model's erroneous responses on OOD inputs. The corresponding explanations G_S are thus dominated by artifacts of failure under distribution shift. These *fundamentally meaningless explanations* are corroborated by empirical results, where SOTA post-hoc explainers exhibit extremely large explanatory error (e.g., SchNet's dipole moment MAE increases from 0.03 on full graphs to 0.42 on RISE explanations), despite preserving a nontrivial proportion of edges (79%).

4.2 E2L: Bridging the OOD Gap and Benefiting Representation Learning

As discussed in Sec. 4.1, the pretrained GNN Φ_γ is optimized on full proximity graphs and cannot adapt to the distribution of explanatory subgraphs while γ remains fixed. We therefore intentionally shift the setting from a fixed pretrained parameter realization to the task-relevant computation structure of a predefined GNN architecture, so that explanation reliability is measured by whether the resulting compact structure can support accurate target prediction. Then we seek a subgraph that best represents the original dense graph for the target task:

$$G_S^* = \arg\min_{G_S \subseteq G} \mathcal{L}\big(Y, \Phi(G_S)\big) \quad \text{s.t.} \quad |G_S| \le B. \tag{5}$$

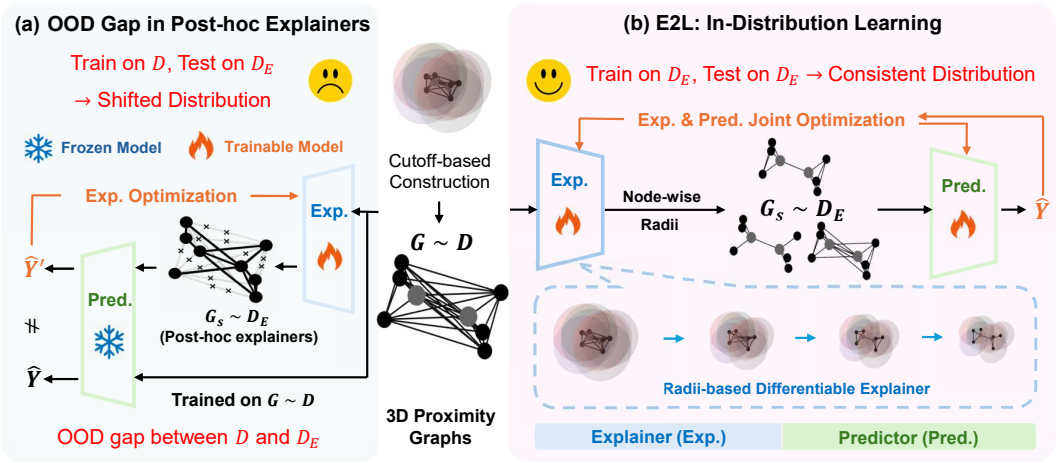

Figure 2: OOD gap and E2L framework. (a) Post-hoc explainers evaluate a pretrained model on extracted subgraphs $G_S \sim \mathcal{D}_E$, inducing a distribution shift from the training domain $\mathcal{D}$ to the explanatory domain $\mathcal{D}_E$, which leads to unreliable explanations. (b) E2L integrates a differentiable explainer into the learning process to construct G_S, and performs prediction directly on it. In particular, due to the design of the radii-based differentiable explainer, the model is trained end-to-end only on the prediction loss, ensuring consistent distribution alignment.

With this reformulation, the explanation target is no longer dependent on a fixed pretrained parameter realization Φ_γ, but the compact, task-relevant computation structure that a predefined GNN architecture can learn to rely on for accurate prediction. To realize this formulation, we propose E2L, which integrates an explainer and a predictor, as illustrated in Fig. 2:

$$\widehat{Y} = \Phi_\psi\big(\mathcal{E}_\theta(G_c)\big), \tag{6}$$

where $\mathcal{E}_\theta$ constructs an explanatory graph from G_c, and Φ_ψ learns to predict using the induced structure. Their parameters θ and ψ are jointly optimized.

Specifically, given a GNN model, E2L preserves the backbone's native aggregation operations and inserts the explainer after the first t layers. These layers process G_c to produce intermediate representations $H^{(t)}$, from which the explainer constructs G_{exp} for the remaining layers. The aggregation mechanism of 3D GNNs is provided in Appendix B. Based on the formulation in Equation 6, we provide an instantiation of E2L built with RISE (Qu et al., 2025). We introduce the explainer in Sec. 4.2.1 and the benefits of E2L in Sec. 4.2.2.

4.2.1 Explainer: Bring the Explanation in Distribution

As discussed in Sec. 4.1, existing post-hoc explainers generate subgraphs that violate the cutoff-based construction of 3D proximity graphs, resulting in out-of-distribution (OOD) inputs and unreliable predictions. To address this issue, we firstly design an explainer, which is able to work in the training process via a differentiable subgraph extraction mechanism.

Learning node-wise radii. We parameterize explanations via *node-wise radii* (Qu et al., 2025) under a fixed budget. Given the intermediate node representations after the dense prefix, $H^{(t)} = \{h_i^{(t)}\}_{i\in V}$, the explainer $\mathcal{E}_\theta$ applies a lightweight radius head R_θ to each node and directly produces node-wise radii via a graph-wise normalization:

$$r_i = r_c \cdot b \cdot |V| \cdot \frac{\exp\big(R_\theta(h_i^{(t)})\big)}{\sum_{j\in V}\exp\big(R_\theta(h_j^{(t)})\big)}. \tag{7}$$

Here, r_c is the original cutoff radius, b controls graph sparsity by controlling the normalized radius, and $|V|$ is the number of nodes. This formulation enforces a fixed total radius mass, so the explainer learns how to *reallocate* connectivity across nodes rather than scaling connectivity arbitrarily.

Explanatory subgraph construction. Given the node-wise radii $\{r_i\}_{i\in V}$, we construct the explanatory subgraph by reparameterizing the connectivity of the original cutoff graph $G_c = (V, A^{r_c}, X)$. Specifically, the explanatory subgraph G_{exp} is defined as

$$A_{ij}^{\text{exp}} = A_{ij}^{r_c} \cdot \mathbb{I}\big(\|x_i - x_j\|_2 \le r_i\big), \tag{8}$$

which yields $G_{\text{exp}} = (V, A^{\text{exp}}, X)$. Compared to the uniform cutoff construction, this formulation replaces the global radius r_c with node-wise adaptive radii, thereby inducing *asymmetric connectivity*,

i.e., in general $A_{ij}^{\text{exp}} \neq A_{ji}^{\text{exp}}$. As a result, G_{exp} is no longer a standard undirected proximity graph, but a directed graph with node-dependent receptive fields. However, the radius in Eq. (8) is not differentiable. To enable training, we adopt a differentiable relaxation:

$$A_{ij}^{\text{exp}} = A_{ij}^{r_c} \cdot sigmoid\big(k(r_i - \|x_i - x_j\|_2)\big), \tag{9}$$

which serves as a soft edge gate for gradient-based optimization. The sharpness k controls the transition between inactive and active edges, and is set to a large value (e.g., $k = 200$), making A_{ij}^{exp} closely approximate binary masks. This enables a natural transition from soft values in training to hard thresholding in inference.

Distributional extension. From the distributional perspective, the explainer $\mathcal{E}_\theta$ induces a transformation of the original data distribution:

$$\mathcal{D}_{\mathcal{E}_\theta} = \big\{(G_{\text{exp}}, Y) \,\big|\, G_{\text{exp}} = \mathcal{E}_\theta(G_c),\ (G_c, Y) \sim \mathcal{D}\big\}, \tag{10}$$

which corresponds to the family of node-wise radius graphs induced by $\{r_i\}_{i \in V}$. Compared to the cutoff-based family parameterized by a global radius r_c, this construction extends the support of admissible graphs from uniform-radius proximity graphs to a strictly richer class with node-dependent receptive fields. Consequently, the effective training distribution is no longer $\mathcal{D}$, but $\mathcal{D}_{\mathcal{E}_\theta}$. This eliminates the train-test mismatch in conventional post-hoc explanation, where a predictor trained on $\mathcal{D}$ is evaluated on subgraphs outside its training distribution.

4.2.2 Predictor: Let the Explanation Benefit Learning

Because E2L does not alter the underlying GNN architecture, the predictor works the same as the original model, while operating on the $\mathcal{D}_{\mathcal{E}_\theta}$. From the learning perspective, this distributional extension leads to Corollary 4.2, which connects the training objective and explanation risk in E2L:

Corollary 4.2. *Following Definition 4.1, the explanation risk of the joint model $(\Phi_\psi, \mathcal{E}_\theta)$ under the induced distribution $\mathcal{D}_{\mathcal{E}_\theta}$ can be represented as*

$$R_{\text{exp}}(\Phi_\psi, \mathcal{E}_\theta) = \mathbb{E}_{(G_{\text{exp}}, Y) \sim \mathcal{D}_{\mathcal{E}_\theta}} \left[\mathcal{L}\big(Y, \Phi_\psi(G_{\text{exp}})\big)\right]. \tag{11}$$

Then the training objective of E2L satisfies

$$R_{\text{train}}(\Phi_\psi, \mathcal{E}_\theta) = \mathbb{E}_{(G_c, Y) \sim \mathcal{D}} \left[\mathcal{L}\big(Y, \Phi_\psi(\mathcal{E}_\theta(G_c))\big)\right] = R_{\text{exp}}(\Phi_\psi, \mathcal{E}_\theta). \tag{12}$$

This equivalence can be directly proved by a change of variables under the explainer-induced mapping $G_{\text{exp}} = \mathcal{E}_\theta(G_c)$, which shows that the 3D GNN in E2L is optimized directly on $\mathcal{D}_{\mathcal{E}_\theta}$. Explanatory graphs are therefore no longer treated as out-of-distribution inputs, but are fully aligned with the training process. In contrast to post-hoc explainers, where explanation risk arises from evaluating a pretrained model on unseen subgraphs, *E2L converts this distribution shift into an explicit optimization objective*, ensuring that the learned predictor remains reliable on the generated explanations. Beyond overcoming the distribution mismatch, this explanation-aware formulation also fundamentally changes the input data structure of the predictor, which brings benefits to the representation learning. The detailed analysis of the connections between E2L and existing methods is provided in Sec. 2 and Appendix C.

Benefits I: computational efficiency. Since G_{exp} is constructed under a fixed budget constraint, the sparsity of the explanatory graph can be constrained via controlling b. Consequently, the aggregation in Eq. (20) is performed over fewer edges, reducing computational complexity of the predictor from $\mathcal{O}(|E_c|)$ to $\mathcal{O}(|E_{\text{exp}}|)$ with $|E_{\text{exp}}| \leq |E_c|$ in edge-based GNN aggregation. Moreover, unlike post-hoc explainers that require a multiple-stage training, E2L jointly optimizes the explainer and predictor in a single-stage training process, further reducing the overall training cost in explanation.

Benefits II: better expressivity. Another key advantage of E2L lies in the flexibility of graph construction, leading to the expressivity gain that can be formalized as:

Proposition 4.3. *Under a fixed edge budget B, the graph families generated by cutoff-based construction $\mathcal{G}_{\text{cutoff}}(B)$ and E2L-based construction $\mathcal{G}_{\text{E2L}}(B)$ satisfy*

$$\mathcal{G}_{\text{cutoff}}(B) \subsetneq \mathcal{G}_{\text{E2L}}(B). \tag{13}$$

Consequently, for any predictor class Φ_ψ operating on graph-structured inputs, the optimal value of the learning objective satisfies

$$\inf_{\Phi_\psi,\, G \in \mathcal{G}_{\text{E2L}}(B)} \mathbb{E}\big[\mathcal{L}\big(Y, \Phi_\psi(G)\big)\big] \leq \inf_{\Phi_\psi,\, G \in \mathcal{G}_{\text{cutoff}}(B)} \mathbb{E}\big[\mathcal{L}\big(Y, \Phi_\psi(G)\big)\big]. \tag{14}$$

The proof is provided in Appendix D. Here, the budget B denotes a common constraint on the complexity of the resulting graph (e.g., the number of retained edges). Different construction mechanisms may realize this constraint differently; for example, the RISE-based instantiation of E2L controls sparsity through a radius budget, whereas mask-based explainers may use top-K edge selection. As a result, cutoff-based proximity graphs are restricted to symmetric structures induced by a global threshold, while E2L admits a strictly richer family of connectivity patterns under the same resulting graph-complexity constraint. This expanded graph family enlarges the feasible space of graph structures, which leads to the objective bound in Eq. (14). In this sense, E2L provides more expressive connectivity patterns while allowing the resulting graph sparsity to be controlled.

5 EXPERIMENTS

Our experiments evaluate E2L from two complementary perspectives: explanation reliability and the benefit of explanations for learning. First, for reliability, we examine whether overcoming the OOD gap enables explanatory subgraphs to preserve, or even improve upon, the predictive performance of the original proximity graphs under substantially reduced structures. Second, we study the *benefit of explanation for learning*, investigating whether these reliable, in-distribution explanatory graphs can serve as compact yet expressive computation structures that improve both prediction efficiency and accuracy. The experimental setup is introduced in Appendix E.1.

5.1 RELIABLE EXPLANATIONS: E2L BRIDGES THE OOD GAP

Table 1: Explanatory MAE with **SchNet** on the QM9 dataset. The ratio of preserved edges is indicated in parentheses. Each property name is followed by the prediction result on cutoff-based graph. The controlled budget b is shown under the property name. The best results are highlighted in **bold**.

Explainer	μ (D): 0.03(100%)			α (bohr^3): 0.24(100%)		
	0.2	0.3	0.4	0.2	0.3	0.4
GNNExplainer	2.97(50%)	1.16(70%)	0.86(80%)	2.84(50%)	2.22(70%)	2.05(80%)
PGExplainer	1.88(46%)	1.76(65%)	1.68(83%)	10.42(46%)	9.52(64%)	9.22(82%)
LRI-Bernoulli	6.09(44%)	5.33(70%)	4.13(80%)	14.46(48%)	11.85(67%)	10.14(80%)
RISE	0.60(44%)	0.52(65%)	0.42(79%)	2.67(45%)	2.13(64%)	1.43(77%)
E2L	**0.13(16%)**	**0.06(48%)**	**0.04(72%)**	**0.30(38%)**	**0.29(47%)**	**0.15(72%)**
Explainer	ϵ_{HOMO} (eV): 0.04(100%)			ϵ_{LUMO} (eV): 0.03(100%)		
	0.2	0.3	0.4	0.2	0.3	0.4
GNNExplainer	0.33(50%)	0.24(70%)	0.15(80%)	1.18(50%)	0.77(70%)	0.21(90%)
PGExplainer	0.57(50%)	0.58(70%)	0.60(80%)	1.32(50%)	1.36(72%)	1.42(86%)
LRI-Bernoulli	0.60(48%)	0.62(67%)	0.62(80%)	1.36(50%)	1.38(80%)	1.40(90%)
RISE	0.33(44%)	0.25(64%)	0.17(77%)	0.54(45%)	0.37(66%)	0.21(80%)
E2L	**0.17(22%)**	**0.09(48%)**	**0.07(72%)**	**0.23(23%)**	**0.12(47%)**	**0.07(71%)**

Bridging the OOD gap enables E2L to produce more reliable explanations under the definition in Sec. 4. Specifically, we compare different explainers by measuring predictive performance on extracted subgraphs under controlled sparsity, where a reliable explanation should preserve the task-relevant information needed for accurate prediction with a substantially reduced graph. As shown in Tables 1, 2, and 3, E2L consistently achieves lower prediction error than existing explainers across different sparsity levels, while using comparable or fewer edges[1]. More importantly, E2L achieves predictive performance close to, and in several cases even better than, that of the original proximity graphs despite using substantially fewer edges. For example, on property α of QM9, E2L achieves an MAE of 0.15 with 72% edge preservation on SchNet and 0.05 with 53% edge preservation on DimeNet. Similar improvements are observed on GEOM-Drugs. These results show that, by overcoming the explanation-induced OOD problem, E2L extracts compact explanatory structures that reliably retain the task-relevant predictive information.

We further observe a clear distinction across backbone models that is consistent with geometric modeling principles[2]. SchNet relies solely on pairwise distance-based interactions, and therefore requires relatively dense connectivity to maintain competitive expressivity. Even under the same sparsity budget, its performance degrades more noticeably as edges are removed. In contrast, DimeNet incorporates higher-order geometric information through angular interactions, which provides stronger structural constraints and enhances expressivity. As a result, it can achieve comparable, or even

[1]Additional experiments on SEGNN and prediction uncertainty analysis are shown in Appendix F

[2]The geometric principle is formalized and proved in Appendix G

Table 2: Explanatory MAE with **DimeNet** on the QM9 dataset. The ratio of preserved edges is indicated in parentheses. Each property name is followed by the prediction result on original graphs. The controlled budget b is shown under the property name. The best results are highlighted in **bold**.

Explainer	μ (D): 0.03(100%)			α (bohr^3): 0.05(100%)		
	0.4	0.5	0.6	0.4	0.5	0.6
GNNExplainer	0.90(30%)	0.80(40%)	0.67(50%)	25.87(30%)	17.77(40%)	10.07(50%)
PGExplainer	1.20(31%)	1.41(48%)	1.42(57%)	56.00(30%)	112.81(40%)	208.19(50%)
LRI-Bernoulli	7.93(33%)	6.09(44%)	5.27(53%)	15.96(35%)	15.14(41%)	13.74(54%)
RISE	0.70(25%)	0.57(38%)	0.46(50%)	5.00(25%)	1.49(38%)	1.09(50%)
E2L	**0.04(18%)**	**0.04(43%)**	**0.04(54%)**	**0.05(22%)**	**0.05(38%)**	**0.05(53%)**
Explainer	ϵ_{HOMO} (eV): 0.03(100%)			ϵ_{LUMO} (eV): 0.02(100%)		
	0.4	0.5	0.6	0.4	0.5	0.6
GNNExplainer	1.51(40%)	1.00(50%)	0.49(70%)	1.06(40%)	0.88(50%)	0.37(70%)
PGExplainer	3.71(43%)	3.60(50%)	3.27(71%)	1.04(40%)	1.00(50%)	0.95(70%)
LRI-Bernoulli	3.65(40%)	3.55(50%)	3.13(70%)	0.94(40%)	0.93(50%)	0.91(70%)
RISE	5.09(38%)	2.73(49%)	1.49(66%)	0.75(38%)	0.73(48%)	0.20(66%)
E2L	**0.03(32%)**	**0.03(45%)**	**0.03(53%)**	**0.02(21%)**	**0.02(41%)**	**0.02(54%)**

Table 3: Explanatory MAE on the GEOM-Drugs dataset. Both backbones employ a cutoff distance of $5\mathring{A}$, resulting in an average of 981 edges per graph. The controlled budget b is shown under the property name. The preserved edge ratio is shown in parentheses.

Explainer	SchNet ($Hartree$): 0.012 (100%)			DimeNet ($Hartree$): 0.008 (100%)		
	0.4	0.5	0.6	0.4	0.6	0.7
GNNExplainer	0.11(30%)	0.11(40%)	0.10(50%)	4.22(40%)	2.29(50%)	1.87(60%)
PGExplainer	0.07(26%)	0.08(40%)	0.10(53%)	41.65(30%)	29.95(50%)	22.93(60%)
LRI-Bernoulli	17.54(20%)	17.53(40%)	17.53(50%)	24.63(30%)	12.02(50%)	9.65(60%)
RISE	0.10(20%)	0.10(31%)	0.08(43%)	0.20(27%)	0.08(41%)	0.06(56%)
E2L	**0.007(12%)**	**0.006(24%)**	**0.005(52%)**	**0.008(12%)**	**0.007(39%)**	**0.006(51%)**

superior, predictive performance with substantially fewer edges. This trend can be clearly observed on the same prediction targets across different backbones (e.g., α). For SchNet, maintaining competitive performance requires preserving a relatively high proportion of edges (e.g., above 70%). In contrast, DimeNet is able to retain competitive accuracy with substantially sparser graph structures (e.g., around 30%). These results empirically indicate that our explanatory results follow the geometric principle: higher-order features reduce the dependence on dense connectivity in geometric data.

5.2 Efficiency and Expressivity: 3D GNNs Benefit from E2L Explanations

We next investigate whether the learned explanatory graphs can benefit representation learning beyond explanation reliability. By bridging the OOD gap, E2L produces meaningful explanatory graphs that are directly optimized for prediction, allowing them to serve as compact computational structures rather than post-hoc artifacts. We evaluate this property on the EC-Protein benchmark, which contains substantially larger 3D graphs than the molecular datasets considered above and therefore provides a more challenging setting for evaluating computational efficiency. As shown in Table 4, integrating E2L with GearNet-Edge improves classification accuracy from 87.4% to 88.9%, while reducing the average number of edges from $5{,}924$ to $3{,}880$. Meanwhile, the computational cost decreases from 106.6 to 72.1 GFLOPs per sample, representing approximately 34.5% fewer edges and 32.4% fewer computations. Importantly, this efficiency gain does not come at the cost of predictive capability; instead, the sparser E2L representation achieves the improvement in accuracy. These results demonstrate that the explanatory structures learned by E2L can remove redundant geometric interactions while retaining, and even strengthening, task-relevant information. This observation is consistent with the expanded graph family characterized in Proposition 4.3: rather than relying on a uniformly dense cutoff graph, E2L reallocates connectivity according to the learned importance of individual nodes, enabling a more expressive use of a limited edge budget. Together with the results on QM9 dataset (Appendix F.2) and the comparison with proximity graphs built by smaller cutoff distances (Appendix F.3), we show that the reliable explanations from E2L can naturally serve as efficient representations for 3D GNN learning across graph scales.

Table 4: Efficiency and performance comparison on the EC dataset. We report classification accuracy (%) on the *react* task and computational efficiency. $\uparrow$ and $\downarrow$ indicate the higher and the lower are better, respectively.

Model	React $\uparrow$	Edges $\downarrow$	GFLOPs $\downarrow$
IEConv	87.2	5924	-
ProNet	86.4	5924	-
GearNet-Edge	87.4	5924	106.6
GearNet+E2L	88.9	3880	72.1

5.3 Comparisons with Alternative Strategies

Beyond comparing E2L with existing post-hoc explainers under the standard evaluation protocol, we further investigate whether its benefits can be reproduced by alternative strategies.

Post-hoc explanation followed by predictor adaptation. Beyond fairly comparing E2L with baselines under the same datasets and experimental settings, a natural question is whether the OOD issue can be addressed simply by adapting the predictor to the post-hoc explanatory distribution. Therefore, we compare this strategy with E2L. Specifically, we fine-tune the pretrained predictor on post-hoc explanatory graphs so that it is explicitly adapted to the induced subgraph distribution. As detailed in Appendix F.3, this adaptation substantially reduces the prediction error on explanatory graphs. However, E2L achieves better performance while preserving fewer edges. It indicates that predictor adaptation after explanation is insufficient to reproduce E2L. Because the post-hoc explanatory structures are optimized using the OOD response from the pretrained predictor, whereas E2L jointly optimizes the explanatory structure and predictor under the unified distribution.

Transfer of 2D explanation-aware learning to 3D graphs. Directly adapting a representative 2D explanation-aware method does not reproduce the benefits of combining geometry-aware explanation with E2L. As detailed in Appendix F.4, we adapt GSAT (Miao et al., 2022) to SchNet and evaluate it under comparable edge-retention ratios. Although GSAT also jointly optimizes explanation and prediction, E2L consistently achieves better performance across the evaluated molecular properties. These results show that directly transferring 2D explanation-aware methods to 3D proximity graphs is non-trivial and yields inferior performance. In contrast, E2L bridges the geometry-aware 3D explainers with predictor learning, enabling the learned explanatory structures to more effectively benefit 3D representation learning.

5.4 Ablation Study

E2L splits the backbone into a dense prefix and an explanation-aware suffix: the first t layers operate on the 3D proximity graph, after which the explainer predicts node-wise radii and the remaining layers compute on the induced explanation graph. We study how the explainer layer t affects predictive performance. Smaller t applies explanation earlier and is more efficient, while larger t delays sparsification and provides richer embeddings from dense-graph aggregation. As shown in Table 5, on QM9, increasing t consistently improves performance across all targets, indicating that delayed sparsification helps the explainer exploit richer intermediate representations. The gains are gradual and stable, suggesting that E2L does not rely on a precise insertion point but benefits smoothly from deeper dense-prefix processing. Meanwhile, early insertion ($t = 1$) remains competitive, with only a modest gap to later insertion. This demonstrates a robust trade-off between early sparsification for efficiency and deeper dense computation for improved predictive performance. Additionally, we provide an ablation on the sharpness k in Appendix H.

Table 5: Ablation on the explainer layer (t) in SchNet with budget $b = 0.5$. $\downarrow$ means the lower is better.

t	GFLOPs $\downarrow$	$\mu \downarrow$	$\alpha \downarrow$	$\epsilon_{\text{HOMO}} \downarrow$	$\epsilon_{\text{LUMO}} \downarrow$
1	0.091	0.0279	0.0862	0.0488	0.0458
2	0.092	0.0263	0.0845	0.0486	0.0430
3	0.094	0.0254	0.0828	0.0467	0.0423

6 Conclusion and Future Work

In this paper, we identify an explanation-induced OOD gap in post-hoc explanation of 3D GNNs: explanatory subgraphs can violate the graph structures seen during training, making the frozen predictor unreliable on them. We propose **E2L**, which jointly optimizes the explainer and predictor on the explainer-induced graph distribution, thereby removing this train–explanation mismatch. We further show that the resulting sparse, task-relevant structures can serve as efficient computation graphs for 3D representation learning.

Limitations and future work. E2L changes the explanation target from faithfully reproducing a fixed pretrained model to learning task-relevant explanatory structures jointly with the predictor. This is motivated by the fact that a pretrained predictor may provide unreliable signals on OOD explanatory subgraphs, making faithful post-hoc optimization difficult. Developing methods that can reliably explain a fixed pretrained 3D GNN while addressing this OOD issue remains an important direction. Moreover, for $SE(3)$-equivariant GNNs, our current explainer relies on scalar features for radius prediction, which limits its expressivity. Future work could leverage higher-order tensor representations (Fuchs et al., 2020) to develop more expressive equivariant explainers.

AI USE STATEMENT

In this work, we used generative AI tools to assist with paper writing and method implementation. We did not use these tools to generate research ideas, develop theoretical models, design research methodology or experiments. We have reviewed all content associated with this work, including AI-assisted text and code, for accuracy, correctness, and consistency with our intended methodology. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our results. The anonymized source code is released at `https://anonymous.4open.science/r/E2L-28EE`. Implementation details of the preliminary experiments are given in Appendix A, and details of E2L on invariant and equivariant backbones together with the full experimental settings are provided in Appendix E.

REFERENCES

Kenza Amara, Mennatallah El-Assady, and Rex Ying. Ginx-eval: Towards in-distribution evaluation of graph neural network explanations. In *XAI in Action: Past, Present, and Future Applications at NeurIPS 2023*, 2023.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=_xwr8gOBeV1`.

Jialin Chen, Shirley Wu, Abhijit Gupta, and Rex Ying. D4explainer: In-distribution explanations of graph neural network via discrete denoising diffusion. *Advances in Neural Information Processing Systems*, 36:78964–78986, 2023.

Linmao Chen, Chaobo He, Junwei Cheng, Chunying Li, and Quanlong Guan. Generating in-distribution counterfactual explanation for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20208–20216, 2026a. doi: 10.1609/aaai.v40i24.39108.

Zhuomin Chen, Jiaxing Zhang, Jingchao Ni, Xiaoting Li, Yuchen Bian, Md Mezbahul Islam, Ananda Mondal, Hua Wei, and Dongsheng Luo. Generating in-distribution proxy graphs for explaining graph neural networks. In *International Conference on Machine Learning*, pp. 7712–7730. PMLR, 2024.

Zhuomin Chen, Hojat Allah Salehi, Esteban Schafir, Xu Zheng, Jiaxing Zhang, Hua Wei, Jingchao Ni, Farhad Shirani, and Dongsheng Luo. Addressing structural distribution shift in explanations for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (9):10976–10990, 2026b. doi: 10.1109/TPAMI.2026.3690304.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33: 1970–1981, 2020.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. In *International Conference on Learning Representations*, 2021.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph rationalization with environment-based augmentations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1069–1078, 2022a. doi: 10.1145/3534678.3539347.

Xufeng Liu, Dongsheng Luo, Wenhan Gao, and Yi Liu. 3DGraphX: Explaining 3D molecular graph models via incorporating chemical priors. In *Proceedings of the 31st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2025.

Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations*, 2022b.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19620–19631. Curran Associates, Inc., 2020.

Ge Lv, Lei Chen, and Caleb Chen Cao. On glocal explainability of graph neural networks. In *International Conference on Database Systems for Advanced Applications*, pp. 648–664. Springer, 2022.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15524–15543. PMLR, 2022. URL https://proceedings.mlr.press/v162/miao22a.html.

Siqi Miao, Yunan Luo, Mia Liu, and Pan Li. Interpretable geometric deep learning via learnable randomness injection. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6u7mf9s2A9.

Jingxiang Qu, Wenhan Gao, Jiaxing Zhang, Xufeng Liu, Hua Wei, Haibin Ling, and Yi Liu. Rise: Radius of influence based subgraph extraction for 3d molecular graph explanation. In *International Conference on Machine Learning*, pp. 50744–50761. PMLR, 2025.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*, pp. 991–1001, 2017.

Ana Vasilcoiu, THF Stessen, Thies Kersten, and Batu Helvacioglu. [re] gnninterpreter: A probabilistic generative model-level explanation for graph neural networks. *Transactions on Machine Learning Research*, 2024.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in neural information processing systems*, 33:20437–20448, 2020.

Ying-Xin Wu, Xiang Wang, An Zhang, Xia Hu, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Deconfounding to explanation evaluation in graph neural networks. *arXiv preprint arXiv:2201.08802*, 2022a.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022b.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.05563.

Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery ; Data Mining*. ACM, August 2020. doi: 10.1145/3394486.3403085. URL http://dx.doi.org/10.1145/3394486.3403085.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning*, pp. 12241–12252. PMLR, 2021.

Jiaxing Zhang, Dongsheng Luo, and Hua Wei. Mixupexplainer: Generalizing explanations for graph neural networks with data augmentation. In *Proceedings of 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2023a.

Jiaxing Zhang, Xiaoou Liu, Dongsheng Luo, and Hua Wei. Is your explanation reliable: Confidence-aware explanation on graph neural networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3740–3751, 2025.

Yue Zhang, David Defazio, and Arti Ramesh. Relex: A model-agnostic relational model explainer. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 1042–1049, 2021.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. ProtGNN: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9127–9135, 2022. doi: 10.1609/aaai.v36i8.20898. URL https://ojs.aaai.org/index.php/AAAI/article/view/20898.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023b.

Xu Zheng, Farhad Shirani, Tianchun Wang, Wei Cheng, Zhuomin Chen, Haifeng Chen, Hua Wei, and Dongsheng Luo. Towards robust fidelity for evaluating explainability of graph neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

A Implementation details of Preliminary Experiments

Overview. To empirically verify the explanation-induced distribution shift discussed in Sec. 4.1, we conduct a controlled preliminary experiment on the QM9 dataset, comparing the behavior of 2D chemical graphs and 3D proximity graphs under simulated subgraph extraction. Specifically, we use a stochastic edge-removal process as the proxy for explanatory methods and evaluate whether the resulting subgraphs remain consistent with their underlying graph construction mechanisms.

Dataset and sampling. We use the QM9 dataset. From the full dataset, we randomly sample $10{,}000$ molecules. For each sampled molecule, we construct both its 2D chemical graph and its 3D proximity graph. The experiment evaluates edge removal ratios $\rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, corresponding to removing 10% to 50% of edges.

Graph construction. For each molecule, we construct two graph representations:

- **2D chemical graph.** Nodes correspond to atoms and edges correspond to chemical bonds provided by QM9. The graph is treated as undirected, with duplicate edges and self-loops removed.
- **3D proximity graph.** Nodes correspond to atoms with coordinates $\{x_i \in \mathbb{R}^3\}$. Edges are constructed using a distance-based cutoff rule:

$$A_{ij} = \mathbb{I}\big(\|x_i - x_j\|_2 \le r_c\big), \tag{15}$$

  resulting in an undirected proximity graph defined by a global cutoff radius.

This setup ensures that the 2D and 3D graphs reflect their respective generative principles: chemical bonding rules and geometric proximity.

Edge removal and graph pruning. To simulate subgraph extraction in explanation methods, we randomly remove a fraction ρ of edges from each graph without replacement. Let m denote the number of edges in the original graph. We uniformly sample and remove $\lfloor \rho m \rfloor$ edges, then prune isolated nodes (i.e., nodes with zero remaining degree). The resulting graph is reindexed before evaluation. We report the retained edge ratio after pruning.

Distributional consistency. We define distributional consistency separately for 2D and 3D graphs, based on their respective construction rules.

2D chemical consistency. For the pruned 2D graph, we reconstruct a molecule using RDKit by treating all retained edges as single bonds and disabling implicit hydrogens. We then perform a sanitization check using `Chem.SanitizeMol`. A graph is considered chemically valid if sanitization succeeds. The final consistency metric is defined as:

$$\mathcal{C}_{2D}^{\text{chem}}(\rho) = \mathbb{E}\big[\mathbf{1}(\text{RDKit-valid})\big]. \tag{16}$$

This measures whether the pruned graph remains a valid molecule under standard chemical rules.

3D proximity consistency. For the pruned 3D graph, we evaluate whether it remains consistent with a distance-threshold construction. Let

$$r^\star = \max_{(i,j)\in E_{\text{ret}}} \|x_i - x_j\|_2 \tag{17}$$

be the maximum distance among retained edges, and E_{ret} denotes the set of edges retained after random edge removal and node pruning. A valid proximity graph must include all edges between node pairs whose distance is at most $r^\star$. Therefore, we define the consistency condition as:

$$\forall (u,v), \quad \|x_u - x_v\|_2 \le r^\star \;\Rightarrow\; (u,v) \in E_{\text{ret}}. \tag{18}$$

The 3D consistency metric is then:

$$\mathcal{C}_{3D}^{\text{prox}}(\rho) = \mathbb{E}\big[\mathbf{1}(\text{threshold-consistent})\big]. \tag{19}$$

This criterion captures whether the pruned graph can still be explained by cutoff-distance rules. It is worth noting that this experiment focuses on revealing the severity of the OOD issue in 3D settings from the perspective of graph construction rules. This does not imply that 2D graphs are free from OOD effects, as distribution shift can arise from multiple factors (e.g., graph semantics and feature distributions). Instead, our results demonstrate that under data construction consistency, the OOD issue is substantially more severe in 3D proximity graphs.

B GNN AGGREGATION

We first emphasize that E2L does not modify the underlying GNN architectures. Instead, it preserves the original aggregation mechanism of the backbone model (e.g., MPNN (Gilmer et al., 2017), Graphormer (Ying et al., 2021), etc), and only changes the graph structure used by the later layers (i.e., *predictor*). Let L denote the total number of GNN layers, and let $t \in \{1, \dots, L-1\}$ be the index of intermediate layers. On the original proximity graph $G_c = (V, A^{r_c}, X)$, the first t layers follow the backbone's native aggregation rule:

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}_i^c} \phi_\psi^{(k)}\big(h_i^{(k)}, h_j^{(k)}, x_i, x_j\big)\right), \qquad k = 0, \dots, t-1, \tag{20}$$

where $\sigma(\cdot)$ denotes the activation function, $\mathcal{N}_i^c = \{j \mid A_{ij}^{r_c} = 1\}$ denotes the neighborhood induced by the original cutoff graph, and $\phi_\psi^{(k)}(\cdot)$ is the backbone-specific aggregation function. We denote the resulting intermediate graph representation after the dense prefix by $H^{(t)} = \{h_i^{(t)}\}_{i \in V}$.

C RELATIONS TO EXISTING EXPLAINERS AND EXPLANATION-AWARE LEARNING

E2L retains the conventional goal of identifying a compact and prediction-relevant subgraph, but changes the object being explained from a fixed pretrained parameter realization to the task-relevant computation structure of a predefined GNN architecture. Existing post-hoc explainers seek a subgraph that preserves the behavior of a pretrained predictor whose parameters remain fixed during explanation. However, our analysis shows that this fixed predictor is exactly where the explanation-induced OOD problem arises in 3D proximity graphs: it is trained on cutoff-based graphs but evaluated on sparse explanatory graphs outside that training distribution. Consequently, predictions on explanatory subgraphs can become unreliable, compromising the optimization signal used for subgraph selection. E2L instead allows the predictor to adapt to the explanatory distribution and evaluates whether the resulting compact structure preserves the information required for accurate target prediction. Thus, E2L studies the predictive reliability of explanatory structures rather than faithfulness to one particular pretrained parameter realization.

E2L is a framework for connecting an explainer and a predictor, rather than a new standalone explainer. The core role of E2L is to make the explanatory structure part of the GNN learning process, such that the predictor is optimized directly on the graph distribution induced by the explainer. Therefore, E2L is largely orthogonal to how the explanatory subgraph itself is parameterized. In this work, we instantiate E2L using the node-wise radius formulation of RISE (Qu et al., 2025), since it provides a natural differentiable explanation mechanism for 3D proximity graphs. However, E2L is not restricted to RISE; future differentiable 3D explainers can be incorporated into the same framework by replacing the explainer module while retaining the same distribution-alignment principle.

E2L is also different from existing explanation-aware learning methods. Methods such as GSAT (Miao et al., 2022) have already explored jointly learning interpretable structures and predictors in conventional 2D graphs, and we do not claim that joint explainer–predictor optimization itself is new. These methods typically propose a specific explanation-aware learning mechanism, such as stochastic attention and information-bottleneck regularization in GSAT. In contrast, E2L addresses a different problem: how to bridge an existing explainer with predictor learning so that the predictor is trained on the same explanatory distribution on which it is evaluated. E2L therefore provides a general distribution-alignment framework rather than prescribing a particular explanation mechanism or regularization objective.

The RISE-based instantiation realizes the E2L framework for 3D proximity GNNs, where directly transferring existing 2D explanation-aware methods is nontrivial. Unlike conventional 2D graphs with explicitly defined edges, 3D proximity graphs are substantially denser and their connectivity is induced by continuous geometry through cutoff distances. Generic edge-selection or discrete subgraph-search mechanisms therefore do not directly preserve the geometric structure underlying 3D GNN computation. The original RISE (Qu et al., 2025) addresses this geometric structure through

node-wise radii, but remains a post-hoc explainer operating on a fixed pretrained predictor. E2L instead incorporates this geometry-aware explainer into representation learning and overcomes the corresponding train–explanation distribution mismatch. Therefore, to the best of our knowledge, E2L is the first framework that explicitly identifies and overcomes the explanation-induced OOD problem in 3D GNN explanation, while its RISE-based realization provides a practical explanation-aware learning instantiation specifically designed for 3D proximity graphs. The experimental comparison between E2L and the explanation-aware method is provided in Appendix F.4.

D Proof of Proposition 4.3

Proof. We prove the claim by set inclusion. Under the edge budget B, a cutoff-based graph is constructed using a global radius r, i.e., $r_i \equiv r$ for all $i \in V$. Here, B constrains the resulting graph size, i.e., $|E(G)| \leq B$, rather than the internal sparsity-control parameter of a particular construction mechanism. For any cutoff-based graph $G \in \mathcal{G}_{\text{cutoff}}(B)$ constructed with a global radius r, choosing $r_i = r$ for every node in the E2L construction reproduces exactly the same graph. Therefore,

$$\mathcal{G}_{\text{cutoff}}(B) \subseteq \mathcal{G}_{\text{E2L}}(B). \tag{21}$$

To show strict inclusion, note that E2L allows heterogeneous node-wise radii. Consider two nodes $u, v \in V$ with $r_u \neq r_v$. Then there can exist a pair (u, v) such that

$$A^{\text{exp}}_{uv} = 1, \quad A^{\text{exp}}_{vu} = 0, \tag{22}$$

which yields an asymmetric (directed) edge. However, any cutoff-based graph satisfies symmetry due to the global radius construction:

$$A_{ij} = \mathbb{I}(\|x_i - x_j\|_2 \leq r) = \mathbb{I}(\|x_j - x_i\|_2 \leq r) = A_{ji}. \tag{23}$$

Hence, such asymmetric graphs cannot be generated by cutoff-based constructions, implying

$$\mathcal{G}_{\text{cutoff}}(B) \subsetneq \mathcal{G}_{\text{E2L}}(B). \tag{24}$$

Finally, since $\mathcal{G}_{\text{cutoff}}(B) \subseteq \mathcal{G}_{\text{E2L}}(B)$, the feasible set of the latter strictly contains that of the former. Therefore, for the minimization problem, we have

$$\inf_{\Phi_\psi,\, G \in \mathcal{G}_{\text{E2L}}(B)} \mathbb{E}\big[\mathcal{L}\big(Y, \Phi_\psi(G)\big)\big] \leq \inf_{\Phi_\psi,\, G \in \mathcal{G}_{\text{cutoff}}(B)} \mathbb{E}\big[\mathcal{L}\big(Y, \Phi_\psi(G)\big)\big], \tag{25}$$

where the inequality follows from the monotonicity of the infimum with respect to set inclusion. □

E Implementation details

We instantiate E2L on a diverse set of 3D GNN backbones, including both *invariant* models (SchNet (Schütt et al., 2017), DimeNet (Gasteiger et al., 2020), and GearNet (Zhang et al., 2023b)), and *equivariant* models (e.g., SEGNN (Brandstetter et al., 2022)). In all cases, E2L follows a unified design: it inserts a lightweight explainer into the learning process, predicts node-wise radii under a fixed budget, and constrains subsequent message passing to the induced explanation graph. Importantly, E2L does not modify the underlying aggregation functions of the backbone models, but only alters the graph structure on which they operate.

Shared mechanism. Given a graph $G = (V, A^{r_c}, X)$ with cutoff radius r_c, the explainer predicts a scalar logit ℓ_i for each node $i \in V$, which is normalized within each graph:

$$\alpha_i = \frac{\exp(\ell_i)}{\sum_{j \in V} \exp(\ell_j)}, \qquad r_i = r_c \cdot b \cdot |V| \cdot \alpha_i, \tag{26}$$

where b is the radius budget. This enforces a fixed total radius mass $\sum_i r_i = r_c \cdot b \cdot |V|$, eliminating the need for explicit sparsity regularization.

The radii induce directed edge masks:

$$m_{ij} = \sigma(k(r_i - d_{ij})), \tag{27}$$

where $d_{ij} = \|x_i - x_j\|_2$ and k controls sharpness. During training, m_{ij} is used as a soft gate; during inference, a hard explanation graph is constructed via thresholding. The remaining backbone layers operate on this explanation graph, resulting in an end-to-end explanation-aware predictor. It is noted that the explainer layer t is set to 2 in all backbones. The ablation study on t is shown in Sec. 5.4.

E.1 Experimental Setup

Datasets. We conduct experiments on widely used 3D molecular datasets, QM9 (Ramakrishnan et al., 2014), GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022), and EC-Protein (Hermosilla et al., 2021). All of them are originally represented as 3D proximity graphs constructed based on cutoff distances in backbone models. The split of QM9 and GEOM-Drugs follow Qu et al. (2025), while the split of EC-Protein follows Zhang et al. (2023b). The units for four QM9 properties (μ, α, ϵ_{HOMO}, ϵ_{LUMO}) are D, bohr^3, eV, and eV, respectively. The unit of energy prediction in GEOM-Drugs dataset is $Hartree$.

Baseline explainers. We compare E2L with standard GNN explanation methods, including GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), LRI-Bernoulli (Miao et al., 2023), and the SOTA explainer, RISE (Qu et al., 2025). The detailed introduction of backbone models is provided in Appendix E.2.

Controlled sparsity. To ensure a fair and rigorous comparison, we explicitly control sparsity via the radius budget b in Eq. 7. For every comparison, the edges preserved by E2L are at the same level, and mostly *fewer* than baseline explainers. This guarantees that E2L operates under a stronger sparsity constraint. Any improvement in predictive performance thus reflects a more effective and reliable explanation.

E.2 Backbone Models

E.2.1 Invariant Backbones

For invariant models, E2L preserves the original symmetry by ensuring that both the radius prediction and the induced graph structure depend only on invariant quantities such as distances and scalar embeddings.

SchNet (Schütt et al., 2017): In SchNet, the radius head is attached to intermediate node embeddings after a dense prefix of interaction blocks. The early layers operate on the full cutoff graph to extract geometric context, after which the predicted radii determine which edges remain active in the subsequent layers. The masking is implemented by scaling edge features during training and pruning edges during inference. Since SchNet relies solely on scalar features and distance-based filters, both the predicted radii and the resulting explanation graph are invariant to translations and rotations.

DimeNet (Gasteiger et al., 2020): For DimeNet, which maintains edge-level representations and angular interactions, the integration of E2L requires a careful transition from edge to node representations. Specifically, intermediate edge embeddings are aggregated into node-level summaries, which are then used to predict radii. The resulting edge masks are applied not only to edge features but also to triplet interactions, where each angular message is gated by the product of the corresponding edge masks. In the hard inference setting, the graph is explicitly pruned and all geometric quantities, including distances, angles, and triplet indices, are recomputed on the resulting subgraph. This design ensures that the invariant geometric structure of DimeNet is preserved throughout the explanation-aware computation.

GearNet (Zhang et al., 2023b): For GearNet, which operates on residue-level graphs and incorporates both node-level message passing and line-graph-based edge interactions, the integration of E2L is naturally aligned with its hybrid geometric design. Specifically, after an initial sequence of GearNet layers on the full residue graph, intermediate node representations are used to predict node-wise radii via a lightweight MLP head. These radii are then converted into directed edge gates based on residue distances, which are applied to both residue-level message passing and the associated line-graph edge interactions.

E.2.2 Equivariant Backbones

For equivariant models, the key requirement is to preserve the transformation behavior of features under $SE(3)$ transformations. This imposes additional constraints on the design of the explainer.

SEGNN (Brandstetter et al., 2022): In SEGNN, node and edge features are organized according to irreducible representations, and message passing is implemented via equivariant tensor operations. To integrate E2L, the radius head is restricted to consume only scalar ($0e$) channels or invariant

norms derived from higher-order features, ensuring that each predicted radius is itself invariant. The resulting edge masks are therefore scalar quantities depending only on invariant radii and distances. During message passing, these scalar masks multiply equivariant features, which preserves their transformation laws. In the hard setting, edges are pruned based solely on invariant conditions, and all geometric features can be recomputed consistently on the pruned graph. As a result, the overall model remains fully equivariant.

Across both invariant and equivariant backbones, E2L follows a unified principle: *it decouples graph construction from representation learning while preserving the original inductive biases of the backbone model.* By restricting message passing to an explanation graph learned in a fully differentiable and in-distribution manner, E2L enables a seamless integration of explanation and prediction without compromising symmetry, expressivity, or efficiency.

E.3 Experimental Settings

All experiments are conducted on a high-performance computing cluster running Ubuntu 20.04, equipped with dual Intel Xeon Silver 4316 CPUs (40 cores) and 8 NVIDIA RTX A6000 GPUs (48GB memory each). We implement all models in `PyTorch` (v2.4.1) with `PyTorch Geometric` (v2.6.1), using CUDA 12.x and cuDNN for acceleration. Experiments are conducted on the QM9, Geom-drugs and EC-Protein datasets. Models are trained using the AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-6}, and the best checkpoint is selected based on validation performance on the explanation graph. We fix random seeds for reproducibility. Importantly, no additional loss terms are introduced; the model is trained end-to-end using the original prediction objective, while sparsity is enforced implicitly through the budgeted radius parameterization.

F Additional Experimental Results

F.1 Explanatory Results on SEGNN

Table 6: Explanation reliability comparison using **SEGNN** on QM9 at budget $b = 0.6$. We report MAE (lower is better) on four targets. The preserved edge ratio is shown in parentheses. Each property name is followed by the prediction result of the original backbone model.

Explainer	μ: 0.023 (100%)	α: 0.060 (100%)	ϵ_{HOMO}: 0.024 (100%)	ϵ_{LUMO}: 0.021 (100%)
GNNExplainer	0.939(59.8%)	16.388(59.8%)	0.457(59.8%)	1.014(59.8%)
PGExplainer	0.974(60.2%)	17.673(52.2%)	0.461(60.0%)	1.129(60.5%)
LRI-Bernoulli	0.895(60.0%)	15.201(60.2%)	3.363(60.0%)	0.578(63.8%)
RISE	0.676(53.2%)	9.629(47.6%)	0.298(57.3%)	0.247(55.8%)
E2L	**0.521(49.1%)**	**5.790(44.2%)**	**0.232(54.9%)**	**0.152(51.4%)**

As shown in Table 6, E2L consistently outperforms all compared explainers on SEGNN across all targets, achieving substantially lower MAE under a stricter sparsity constraint. This trend mirrors the observations on invariant backbones, further confirming that bridging the OOD gap is the key to obtaining reliable explanations. Nevertheless, we observe that the performance of E2L on SEGNN still exhibits a noticeable gap compared to the original cutoff-based model. This limitation arises from the equivariance constraint of SEGNN: to preserve SE(3)-equivariance, the current explainer is restricted to operate only on scalar ($0e$) features when predicting node-wise radii. As a result, it cannot fully leverage higher-order geometric representations encoded in tensor features, leading to reduced expressivity of the learned explanation graph. Consequently, although E2L substantially improves over existing explainers, its performance remains inferior to the full-graph baseline. Developing more expressive explainers that can effectively utilize tensor-product representations while preserving equivariance is an important direction for future work, as discussed in Sec. 6.

F.2 Efficiency and Expressivity Experiments on QM9 Dataset

We additionally evaluate whether E2L explanations can serve as efficient computation structures on QM9. Table 7 compares the original cutoff-based SchNet and DimeNet models with their E2L counterparts in terms of predictive performance, retained graph size, and computational cost.

For SchNet, E2L with $b = 0.6$ maintains a graph structure close in size to the original cutoff graph, while substantially improving prediction on several targets. In particular, the MAE decreases from

0.033 to 0.024 on μ and from 0.235 to 0.082 on α, while remaining comparable on ϵ_{HOMO} and ϵ_{LUMO}. This result highlights the *expressivity* benefit of E2L: even without aggressive sparsification, replacing the uniform cutoff structure with learned node-wise receptive fields provides a more effective graph representation for prediction. In other words, the benefit of E2L is not restricted to removing edges; the learned allocation of connectivity itself can improve representation quality.

For DimeNet, the benefit is manifested more directly as a favorable accuracy–efficiency trade-off. With $b = 0.85$, E2L reduces the average number of edges from 279 to 234 and the computational cost from 7.229 to 6.067 GFLOPs, corresponding to reductions of approximately 16% in both quantities. Despite this sparsification, predictive performance remains essentially unchanged: the MAE changes only from 0.029 to 0.031 on μ, improves from 0.047 to 0.046 on α and from 0.028 to 0.027 on ϵ_{HOMO}, and remains 0.020 on ϵ_{LUMO}.

Taken together, these results illustrate two complementary benefits of explanation-aware graph learning. E2L can improve the expressivity of the underlying graph representation by learning task-adaptive connectivity, as observed with SchNet, while also enabling substantial sparsification with negligible predictive degradation when the backbone can effectively exploit sparse higher-order geometric information, as observed with DimeNet. This provides additional evidence that once the explanation-induced OOD gap is removed, explanatory graphs can function as compact and effective representations for 3D GNN computation rather than merely serving as post-hoc interpretations.

Table 7: Efficiency and performance comparison on QM9. We report MAE on four representative targets, together with the average number of edges and GFLOPs. ↓ means the lower is better.

Model	$\mu \downarrow$	$\alpha \downarrow$	$\epsilon_{\mathrm{HOMO}} \downarrow$	$\epsilon_{\mathrm{LUMO}} \downarrow$	Edges ↓	GFLOPs ↓
SchNet (cutoff-based)	0.033	0.235	0.041	0.034	316	0.099
SchNet-E2L ($b = 0.6$)	0.024	0.082	0.041	0.035	306	0.098
DimeNet (cutoff-based)	0.029	0.047	0.028	0.020	279	7.229
DimeNet-E2L ($b = 0.85$)	0.031	0.046	0.027	0.020	234	6.067

F.3 Fine-Tuning on Explanatory Graphs and Retraining on Uniform-Cutoff Control

We further investigate whether E2L's gains can be reproduced either by adapting the predictor to post-hoc explanatory graphs or by simply training on a uniformly sparser proximity graph. For the first comparison, we use RISE (Qu et al., 2025), a post-hoc 3D GNN explainer in our main experiments. Starting from the pretrained SchNet predictor Φ_0, we first obtain RISE explanatory graphs under budget $b = 0.4$, and then fine-tune the predictor on the induced RISE-graph distribution to obtain Φ_{FT}. We evaluate both Φ_0 and Φ_{FT} on the original proximity graphs and RISE explanatory graphs. For the second comparison, we train a separate SchNet predictor, denoted as Φ_{SC}, directly on proximity graphs constructed with a smaller uniform cutoff of $4.4\mathring{A}$, compared with the original $10\mathring{A}$ cutoff. This control maintains the standard cutoff-based graph construction while reducing the number of edges, allowing us to examine whether E2L's improvements can be explained solely by using a sparser in-distribution graph. We compare these settings with E2L under comparable or stronger sparsity constraints in Table 8.

Fine-tuning on RISE explanatory graphs mitigates the distribution mismatch, but cannot fully repair explanations that were originally extracted using unreliable OOD predictive signals. After adapting the predictor to the RISE-induced subgraph distribution, the MAE is substantially reduced compared with directly evaluating the pretrained predictor on RISE graphs, e.g., from 0.42 to 0.05 on μ and from 1.43 to 0.19 on α. However, E2L still achieves better or comparable performance while preserving fewer edges, reaching 0.04, 0.15, 0.07, and 0.07 MAE on the four properties with only approximately 71–72% of the edges. This suggests that adapting the predictor after explanation cannot fully recover the information lost during post-hoc explanation. In RISE, explanatory structures are first optimized using the responses of a pretrained predictor on OOD subgraphs, and unreliable predictive signals may therefore already influence which structures are selected before predictor adaptation. In contrast, E2L jointly aligns explanation and prediction during learning, so the explanatory graph is optimized directly under reliable target-prediction signals. Moreover, the RISE fine-tuning protocol requires a three-stage pipeline of backbone pretraining,

Table 8: Comparison of SchNet trained or adapted on different sparse graph structures on QM9. We report MAE (↓) on four molecular properties. Φ_0 denotes the original pretrained SchNet, Φ_{FT} denotes the predictor fine-tuned on RISE explanatory graphs, and Φ_{SC} denotes a predictor trained directly on proximity graphs constructed with a smaller uniform cutoff of $4.4\mathring{A}$, compared with $10\mathring{A}$ in the original setting. Preserved edge ratios are shown in parentheses.

Evaluation setting	$\mu \downarrow$	$\alpha \downarrow$	$\epsilon_{\mathrm{HOMO}} \downarrow$	$\epsilon_{\mathrm{LUMO}} \downarrow$
Φ_0 on proximity graphs	0.03 (100%)	0.24 (100%)	0.04 (100%)	0.03 (100%)
Φ_{SC} on smaller-cutoff graphs	0.03 (80%)	0.13 (80%)	0.07 (80%)	0.06 (80%)
Φ_0 on RISE graphs	0.42 (79%)	1.43 (77%)	0.17 (77%)	0.21 (80%)
Φ_{FT} on proximity graphs	0.09 (100%)	0.40 (100%)	0.12 (100%)	0.13 (100%)
Φ_{FT} on RISE graphs	0.05 (79%)	0.19 (77%)	**0.07** (77%)	**0.07** (80%)
E2L	**0.04** (72%)	**0.15** (72%)	**0.07** (72%)	**0.07** (71%)
E2L-FT	**0.03** (72%)	**0.12** (72%)	**0.05** (72%)	**0.05** (71%)

explainer optimization, and predictor adaptation, whereas E2L jointly learns the explanatory structure and predictor.

Further fine-tuning on the explanatory graphs learned by E2L shows that these structures remain useful beyond the jointly trained predictor itself. To test whether the benefit of E2L arises only from co-adaptation between its explainer and predictor, we fix the explanatory structures produced by the learned E2L explainer and further fine-tune the backbone on these graphs, denoted as E2L-FT. E2L-FT applies the backbone directly to the fixed E2L explanatory graphs, such that the original dense proximity graph is no longer required in prediction. E2L-FT further reduces the MAE from 0.04, 0.15, 0.07, and 0.07 to 0.03, 0.12, 0.05, and 0.05, respectively, without increasing the preserved edge ratio. It consistently outperforms fine-tuning on RISE explanatory graphs. These results provide additional evidence that E2L does not merely learn sparse structures specialized to a particular jointly optimized predictor; instead, its explanatory graphs retain task-relevant information that can support continued representation learning by the backbone. This supports our central hypothesis that, once the explanation-induced OOD gap is resolved, the explainer can extract compact and expressive subgraphs that serve as effective computation structures for model learning.

Simply reducing the uniform cutoff does not recover the performance of E2L-FT, showing that E2L learns more expressive and reliable sparse structures rather than merely benefiting from graph sparsification. Training SchNet directly with a reduced cutoff of $4.4\mathring{A}$ provides an important in-distribution control: unlike post-hoc masking, the resulting graphs remain valid cutoff-based proximity graphs, and the predictor is trained directly on this graph distribution. Nevertheless, despite preserving approximately 80% of the original edges, the smaller-cutoff model obtains MAEs higher than E2L-FT's results. Moreover, E2L-FT matches or improves prediction on all four properties under a stricter sparsity constraint. This result indicates that the benefit of E2L cannot be explained solely by reducing graph density or by training the predictor on a sparse in-distribution graph. A single global cutoff uniformly contracts every node's receptive field and therefore remains restricted to the original symmetric cutoff-based graph family. In contrast, E2L learns node-dependent radii and reallocates the available connectivity budget according to task-relevant information, enabling heterogeneous receptive fields and a richer set of explanatory structures.

Notably, these experiments complement, rather than replace, the standard post-hoc explanation evaluation. Replacing the original predictor with a fine-tuned predictor changes the model being explained and therefore no longer follows the conventional definition of post-hoc explanation. Similarly, the smaller-cutoff experiment is not an explanation method, but a controlled representation-learning baseline that tests whether uniform sparsification alone is sufficient. We therefore use the RISE-FT, smaller-cutoff, and E2L-FT results as supporting evidence for the reliability and representation-learning value of E2L's explanatory structures, while retaining evaluation with the original pretrained predictor when assessing conventional post-hoc explanation.

F.4 Comparison with Explanation-Aware Methods

The conceptual differences between E2L and existing explanation-aware graph learning methods, as well as the challenges of directly applying these predominantly 2D methods to 3D proximity graphs, are discussed in Appendix C. Here, we focus on an empirical comparison with GSAT (Miao et al., 2022), a representative explanation-aware method that can be adapted to our 3D setting.

To evaluate whether directly transferring an existing explanation-aware method to 3D graphs can achieve comparable performance, we adapt GSAT to SchNet on QM9. GSAT jointly learns prediction and edge probabilities through stochastic Bernoulli masks and a KL regularizer with prior parameter r. Following its interpretation procedure, we construct deterministic explanations by retaining the top-ranked 50% or 75% of edges and evaluate the predictor on the resulting hard subgraphs. We report GSAT–SchNet results with $r \in \{0.5, 0.7\}$ and compare them with E2L–SchNet under comparable edge budgets. The prior r is distinct from the edge-retention ratio.

Table 9: MAE (↓) of GSAT–SchNet and E2L–SchNet on QM9. r denotes GSAT's Bernoulli-prior parameter.

Method	r	Preserved edges	$\mu \downarrow$	$\alpha \downarrow$	$\epsilon_{\text{HOMO}} \downarrow$	$\epsilon_{\text{LUMO}} \downarrow$
GSAT–SchNet	0.5	50%	0.26	4.53	0.18	0.22
GSAT–SchNet	0.7	50%	0.15	6.23	0.19	0.22
E2L–SchNet	–	$\sim 50\%$	**0.06**	**0.29**	**0.09**	**0.12**
GSAT–SchNet	0.5	75%	0.20	3.24	0.14	0.21
GSAT–SchNet	0.7	75%	0.09	3.49	0.09	0.11
E2L–SchNet	–	$\sim 72\%$	**0.04**	**0.15**	**0.07**	**0.07**

As shown in Table 9, E2L outperforms both GSAT configurations on all four targets at both retention levels. For example, on α, E2L reduces the best reported GSAT MAE from 4.53 to 0.29 at the lower retention level and from 3.24 to 0.15 at the higher level. GSAT's stochastic-to-hard mask conversion and additional KL regularization may contribute to this gap. Moreover, because of the top-k selection mechanism, GSAT is suffering from the discrepancy of top-k selection, as introduced in Qu et al. (2025). By jointly learning geometry-aware radius selection and prediction using the prediction objective alone, E2L provides more effective sparse computation structures under this evaluation protocol.

F.5 Qualitative Analysis of Reliable Explanations

Reliable explanations learned by E2L can also align with chemically meaningful structures. As defined in our formulation, the reliability of an E2L explanation does not require the extracted graph to reproduce a predefined chemical structure or human-specified motif. Instead, a reliable explanation is a compact explanatory graph that preserves the task-relevant information required for accurate prediction, while the predictor is trained directly on the induced explanatory distribution. This definition focuses on whether the learned structure is genuinely informative to the prediction task rather than whether it necessarily matches prior human knowledge.

We further investigate whether such reliable explanations are also consistent with chemical intuition. To expose only the most essential interactions selected by E2L, we impose a substantially more stringent radius budget of $b = 0.12$ and visualize representative molecules from QM9. As shown in Fig. 3, the original 3D proximity graphs contain many additional distance-based edges beyond the actual molecular bonds. Under the strict budget, E2L removes these redundant geometric interactions and retains only the highly localized connections that are most important for prediction. Remarkably, for the representative molecules shown here, the resulting explanatory graphs recover the corresponding chemical-bond structures exactly. This observation provides complementary qualitative evidence that, although chemical correspondence is not imposed by the E2L objective, the reliable structures learned from prediction alone can naturally converge toward chemically intuitive molecular structures when the explanation is strongly constrained.

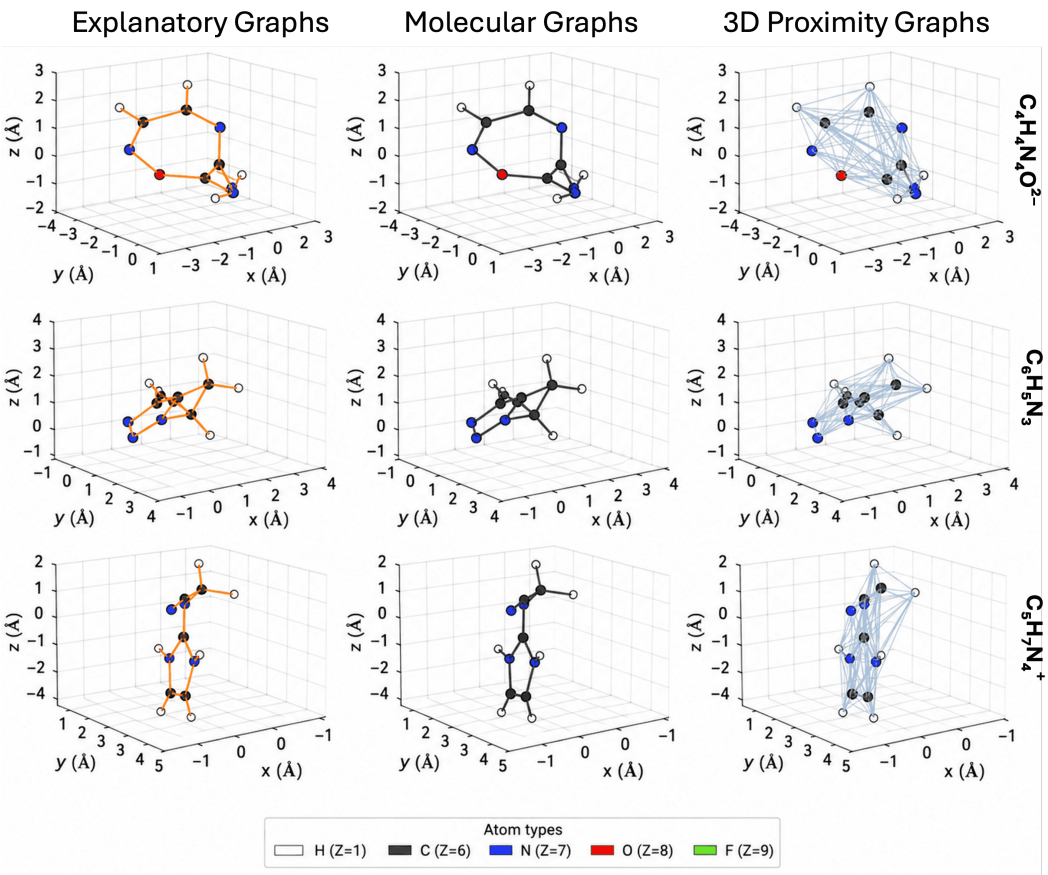

Figure 3: **Qualitative visualization of E2L explanations under a stringent budget** $b = 0.12$**.** Each row shows the E2L explanatory graph (left), the molecular graph defined by chemical bonds (middle), and the original dense 3D proximity graph (right). Although E2L is optimized only to preserve task-relevant predictive information rather than explicitly reconstruct chemical bonds, under the strict budget its explanatory graphs recover the chemical-bond structures of the representative molecules.

Table 10: Predictive performance of E2L on QM9 (SchNet backbone) under different radius budgets b. Entries report MAE $\pm\mathrm{CI}_{95}$, where CI_{95} denotes the half-width of an approximate 95% confidence interval over test molecules.

b	μ	α	ϵ_{HOMO}	ϵ_{LUMO}
0.2	0.132 ± 0.001	0.298 ± 0.006	0.171 ± 0.002	0.229 ± 0.002
0.3	0.062 ± 0.001	0.287 ± 0.006	0.090 ± 0.001	0.121 ± 0.001
0.4	0.040 ± 0.001	0.154 ± 0.005	0.067 ± 0.001	0.070 ± 0.001

F.6 Confidence Intervals for Predictive Performance

We quantify sampling uncertainty in E2L's predictive performance on explanation-induced subgraphs using approximate 95% confidence intervals. Tables 10 and 11 report results for four QM9 properties (μ, α, ϵ_{HOMO}, and ϵ_{LUMO}). For each backbone, we evaluated one saved checkpoint per property and radius budget, using budgets of $\{0.2, 0.3, 0.4\}$ for SchNet and $\{0.4, 0.5, 0.6\}$ for DimeNet. Each checkpoint was evaluated on its corresponding original test set. We report MAE $\pm\mathrm{CI}_{95}$, with confidence-interval half-widths calculated using the normal approximation $\mathrm{CI}_{95} = 1.96s/\sqrt{n}$, where s is the sample standard deviation of per-molecule absolute errors and n is the number of test molecules. SchNet's observed MAE decreases as the radius budget increases across all four properties, whereas DimeNet exhibits comparatively small changes over its evaluated budgets.

Table 11: Predictive performance of E2L on QM9 (DimeNet backbone) under different radius budgets b. Entries report MAE $\pm \mathrm{CI}_{95}$, where CI_{95} denotes the half-width of an approximate 95% confidence interval over test molecules.

b	μ	α	ϵ_{HOMO}	ϵ_{LUMO}
0.4	0.044 ± 0.001	0.052 ± 0.002	0.027 ± 0.001	0.022 ± 0.001
0.5	0.038 ± 0.001	0.052 ± 0.004	0.027 ± 0.001	0.021 ± 0.001
0.6	0.037 ± 0.001	0.054 ± 0.003	0.028 ± 0.001	0.021 ± 0.001

G Geometric Principle: Expressivity vs. Feature Order

We provide an information-theoretic explanation for the empirical observation in Sec. 5.1 that higher-order geometric backbones (e.g., DimeNet) remain accurate under E2L-induced sparsification, whereas distance-only models (e.g., SchNet) degrade more significantly.

Let $G = (V, A, X)$ denote a 3D graph with coordinates $X \in \Omega \subseteq (\mathbb{R}^3)^n$, and let $\Gamma \leq E(3)$ denote the symmetry group. We consider Γ-invariant targets $f \in \mathcal{F}_\Gamma(\Omega, \mathcal{Y})$. Under E2L, the predictor operates on an explanation graph $G_{\exp} = (V, A^{\exp}, X)$, which corresponds to a (generally directed) sparse support $D \subseteq V_{\neq}^2$ with $|D| \leq B$.

Geometric features on sparse supports. For a support D, define order-2 (distance) features:

$$\varphi_D^{(2)}(X) = (\|x_i - x_j\|_2)_{(i,j)\in D}\,. \tag{28}$$

Define order-3 (angular) features:

$$\cos \angle x_j x_i x_k = \frac{(x_j - x_i)^\top (x_k - x_i)}{\|x_j - x_i\|_2 \, \|x_k - x_i\|_2}, \tag{29}$$

and let $\varphi_D^{(3)}(X)$ denote the cumulative feature map including both distances and valid angular terms induced by D.

Budgeted identifiability. A support D is (m, f)-identifying if

$$\varphi_D^{(m)}(X) = \varphi_D^{(m)}(X') \;\Rightarrow\; f(X) = f(X'), \quad \forall X, X' \in \Omega. \tag{30}$$

Let $\mathfrak{I}_m^\Gamma(B)$ denote the class of functions identifiable under budget B using order-m features.

Proposition G.1 (Monotonicity in feature order)**.** *For any $2 \leq m_1 < m_2$,*

$$\mathfrak{I}_{m_1}^\Gamma(B) \subseteq \mathfrak{I}_{m_2}^\Gamma(B). \tag{31}$$

Proof. Let $f \in \mathfrak{I}_{m_1}^\Gamma(B)$. Then there exists a support D with $|D| \leq B$ such that D is (m_1, f)-identifying.

By construction, the order-m_2 feature map $\varphi_D^{(m_2)}$ is cumulative and contains all components of $\varphi_D^{(m_1)}$. Therefore, there exists a projection π such that

$$\varphi_D^{(m_1)}(X) = \pi\big(\varphi_D^{(m_2)}(X)\big), \quad \forall X \in \Omega. \tag{32}$$

Thus, if $\varphi_D^{(m_2)}(X) = \varphi_D^{(m_2)}(X')$, then

$$\varphi_D^{(m_1)}(X) = \varphi_D^{(m_1)}(X'). \tag{33}$$

Since D is (m_1, f)-identifying, it follows that $f(X) = f(X')$. Hence D is also (m_2, f)-identifying, implying $f \in \mathfrak{I}_{m_2}^\Gamma(B)$. □

Theorem G.2 (Strict advantage of higher-order features under sparsity)**.** *For $n \geq 3$, there exists a Γ-invariant target f and budget $B = n - 1$ such that*

$$f \in \mathfrak{I}_3^{E(3)}(B) \setminus \mathfrak{I}_2^{E(3)}(B). \tag{34}$$

Proof. This can be shown using a classical result in distance geometry: pairwise distances alone do not uniquely determine a configuration unless the measurement graph is globally rigid. In particular, a tree (e.g., a chain with $|D| = n - 1$) is not globally rigid in $\mathbb{R}^3$, so there exist non-congruent configurations that share identical edge lengths but differ in angles. Therefore, order-2 features $\phi_D^{(2)}$ cannot identify angle-dependent invariant targets. In contrast, order-3 features explicitly include angular information, which removes this ambiguity and makes such targets identifiable. This establishes the strict advantage.

Interpretation for E2L. Theorem G.2 shows that, under the same edge budget B imposed by E2L, higher-order features strictly increase the class of identifiable geometric functions. In particular, sparse supports induced by $G_{\exp}$ may be sufficient for full geometric characterization when angular information is available, but insufficient when only pairwise distances are used. This explains the empirical trend observed in Sec. 5.1: under E2L sparsification, the explanation graph $G_{\exp}$ provides a constrained support D. Backbones relying on $\varphi^{(2)}$ (e.g., SchNet) lose critical geometric information as edges are removed, whereas backbones using $\varphi^{(3)}$ (e.g., DimeNet) retain stronger identifiability from the same sparse structure. Consequently, higher-order models exhibit greater robustness to sparsification, consistent with the observed performance gap.

H Ablation Study on Sharpness

We further investigate the effect of the sharpness parameter k in the differentiable edge gate. Recall that the explanatory graph is constructed via the soft mask:

$$A_{ij}^{\exp} = A_{ij}^{r_c} \cdot sigmoid\big(k(r_i - \|x_i - x_j\|_2)\big), \tag{35}$$

The sharpness k controls the transition from soft gating to hard thresholding.

Setup. We conduct the ablation on SchNet with budget $b = 0.3$, and evaluate three values $k \in \{200, 400, 800\}$. We report the predictive performance on property μ, together with the preserved edge ratio under hard thresholding. As shown in Table 12, the predictive performance and sparsity remain largely stable across different k. Increasing k makes the soft mask modestly closer to a binary indicator, but does not significantly affect either the MAE or the preserved edge ratio. This demonstrates that E2L is robust to the choice of sharpness within a reasonable range. And $k = 200$ is sufficient for optimization.

Table 12: Ablation on sharpness k in SchNet with budget $b = 0.3$. We report MAE on μ (lower is better), preserved edge ratio, and computational cost.

k	$\mu \downarrow$	Edges $\downarrow$	GFLOPs $\downarrow$
200	0.062	47.8%	0.068
400	0.064	47.9%	0.068
800	0.063	47.6%	0.068

From an optimization perspective, although an extremely large k (e.g., $k = 10,000$) can in principle lead to gradient saturation, this issue is not severe in practice when k is chosen in a modest range. Specifically, under current settings, most effective edges satisfy $r_i \approx d_{ij}$ during training, where the optimization maintains non-vanishing gradients. As a result, the explainer can still be optimized effectively even with relatively sharp transitions.