A COUNTERFACTUAL FRAMEWORK FOR EVALUATING AND IMPROVING THE USE OF MOLECULAR STRUCTURE IN MOLECULAR GRAPH-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular graph-language models have achieved strong performance in molecular description and property prediction. These tasks depend on chemical features encoded in SMILES strings and 2D molecular graphs, including functional groups, ring systems and formal charges. However, current evaluations measure whether an answer is correct, but not whether it is derived from the corresponding chemical structures. Existing datasets also lack controlled molecular pairs that isolate a specific structural feature. In this paper, we propose a counterfactual framework to evaluate and improve how molecular graph-language models use molecular structure. We design a counterfactual benchmark from real, scaffold-matched molecular pairs to test whether models detect the absence of a functional group or another chemical feature. We further design a counterfactual learning model to distinguish molecules with and without the target chemical feature. We also develop an end-to-end training framework that transfers the learned structural information to a broader range of downstream tasks. The benchmark shows that a standard molecular graph-language model achieves only 0.485 pairwise accuracy. Our method raises this accuracy to 0.915 and reaches 0.928 on 822 external molecular pairs. It also improves molecular description, IUPAC naming, molecule–text retrieval, and property prediction. Codes are released on `https://anonymous.4open.science/r/MolCF-0D2C`.

1 INTRODUCTION

Recent molecular graph-language models connect molecular graphs with large language models and support molecular description, IUPAC name prediction, retrieval, and property prediction (Liu et al., 2023c; Park et al., 2024; Chen et al., 2025). These tasks depend on chemically defined attributes encoded by SMILES strings and 2D molecular graphs, including functional groups, ring systems, formal charges, and stereochemistry (Zhang et al., 2025; Chen et al., 2025). Changing one of these attributes can change the correct description, the systematic name, and many molecular properties.

However, current evaluations primarily ask whether final outputs match reference texts, retrieval targets, or property labels. Captioning metrics measure lexical overlap (Papineni et al., 2002; Lin, 2004; Banerjee & Lavie, 2005), retrieval metrics measure global correspondence, and property benchmarks measure final accuracy, but none intervenes on the molecular structure (Edwards et al., 2022; Liu et al., 2023c). Even models given atom-, motif-, or fragment-level inputs are validated by the same output-level scores (Chen et al., 2025; Zhang et al., 2025), so whether they learned those levels is never tested. Models can therefore assign similar descriptions to closely related molecules that differ in a chemically decisive feature, and MoleculeQA shows that generation metrics can miss such factual errors (Lu et al., 2024). Output-level metrics therefore leave open a question: *do existing molecular graph-language models use molecular structure when they produce a correct output, and if not, where between graph and text is the evidence lost?*

To examine this question, we construct a counterfactual benchmark from real, scaffold-matched molecules. Each pair contains a molecule with a target structural attribute and a closely matched molecule without it, and the same chemical statement must receive a higher score on the molecule that contains the attribute. Figure 1 previews the answer: across five published systems, ChEBI-

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.

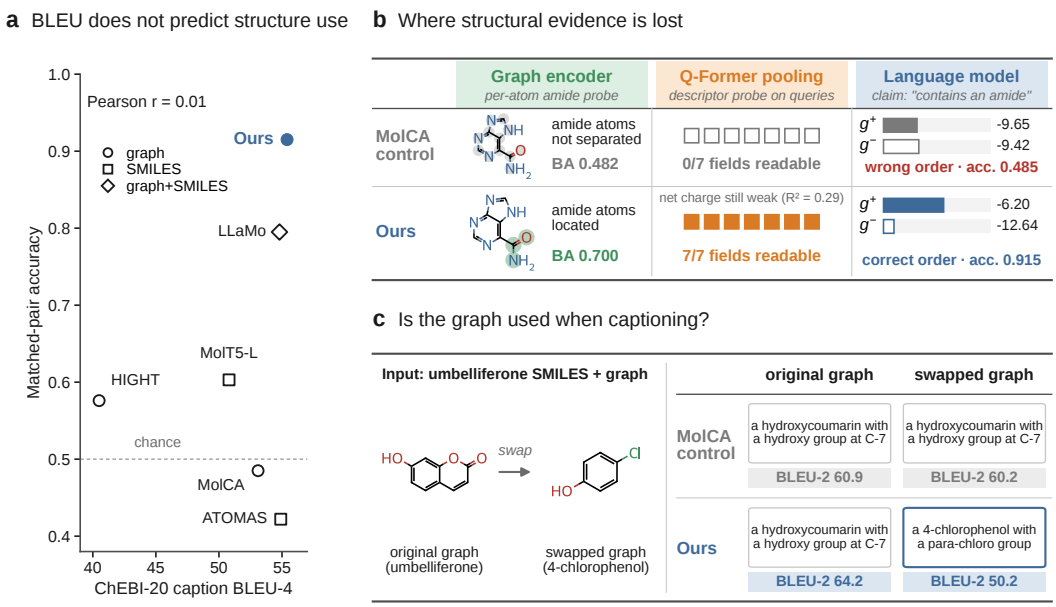

Figure 1: **Output quality does not reveal structure use.** (a) For these five systems, BLEU-4 and matched-pair accuracy have $r = 0.01$. (b) Stage-wise probes show that the control loses functional-group evidence in the graph encoder and descriptor evidence in Q-Former pooling. (c) Swapping the input graph barely changes the control caption.

20 BLEU-4 and pairwise accuracy have Pearson $r = 0.01$, a MolCA baseline reaches only 0.485, linear probes read neither functional groups nor molecular descriptors from its representations, and swapping the input graph barely changes its caption.

To address this limitation, we propose a counterfactual framework that jointly evaluates and improves the use of molecular structure. A counterfactual learning model increases the score gap of a fixed statement between molecules with and without the target feature, while structured supervision keeps functional groups readable from node tokens and molecular descriptors readable from query tokens. An end-to-end training framework transfers the resulting representations to downstream molecular tasks. Our method raises pairwise accuracy from 0.485 to 0.915. On 822 external pairs deduplicated against ChEBI-20, 619 of which flip chloro, accuracy is 0.928. Relative to the matched control, description, IUPAC naming, retrieval, and property prediction improve, with larger gains on structure-sensitive slices. Our contributions are as follows:

- We introduce a counterfactual benchmark of real, scaffold-matched molecular pairs over 21 functional-group and three molecular-descriptor axes, together with a stage-wise diagnosis that locates where structural evidence is lost (Sections 3.2 and 4.2).
- We develop a counterfactual learning model that scores one statement on both molecules of a pair under a graph-only prompt, enforces a margin between them, and supervises functional groups at the atom level and molecular descriptors at the molecule level (Sections 3.3 and 3.4).
- We show that the learned ordering generalizes to 822 external pairs and five unseen functional groups, and that end-to-end finetuning improves description, IUPAC naming, retrieval, and property prediction relative to the matched control (Sections 4.3 and 4.4).

2 Related Work

Molecular graph–language models. Early molecule–text systems learn shared spaces for retrieval or cast description and generation as translation (Edwards et al., 2021; 2022; Christofidellis et al., 2023). Later systems align graphs or strings with text through contrastive objectives and generative pretraining (Su et al., 2022; Liu et al., 2023a; Luo et al., 2023; Liu et al., 2023b; Pei et al., 2023), and

MolCA, GIT-Mol, and LLaMo connect graph encoders to language models through learned cross-modal interfaces (Liu et al., 2023c; 2024; Park et al., 2024). Instruction-tuned assistants and 3D extensions broaden the task range (Zhao et al., 2023; Cao et al., 2025; Fang et al., 2024b; Li et al., 2024). Their evaluations report output quality or global correspondence; they do not test whether a fixed statement responds to a specific structural change.

Fine-grained molecule–text learning. Recent work refines the molecule–text interface below the whole-molecule level. HIGHT tokenizes a molecular graph into atom-, motif-, and molecule-level tokens with separate adapters, augments instruction data with motif-level statements, and introduces MotifHallu, which asks yes/no questions about functional-group presence in a single molecule (Chen et al., 2025). ATOMAS clusters SMILES and text tokens into fragment- and molecule-level units and aligns the two modalities contrastively at three semantic levels (Zhang et al., 2025); MolBridge aligns extracted substructures with the text phrases that describe them, and CAMT5 tokenizes SMILES into substructure-level units for text-to-molecule generation (Park et al., 2025; Kim et al., 2025). Motif-, fragment-, and geometry-aware encoders pursue the same goal for property prediction by adding motif trees, two-level similarity, or 3D views to pretraining (Zhang et al., 2021; Ji et al., 2022; Zang et al., 2023; Liu et al., 2022; Zhou et al., 2023). These designs make functional groups and other local structures more accessible to the model, yet whether the model has actually acquired them is validated only through output-level metrics on the final task. A finer input granularity therefore does not by itself establish that the intended functional groups or hierarchy were learned.

Verifying fine-grained structure learning. Factuality benchmarks sharpen this output-level check: MoleculeQA turns reference descriptions into multiple-choice questions and finds that models with similar generation scores differ widely in factual accuracy, MolErr2Fix asks models to locate and correct annotated errors, and atomic-level NLI verifies each caption claim against its reference (Lu et al., 2024; Wu et al., 2025; Gkoumas & Liakata, 2025); MolecularIQ poses questions symbolically verifiable on the molecular graph, and FGBench asks how a property changes between two molecules that differ by a functional-group edit (Bartmann et al., 2026; Liu et al., 2025). All of them grade an answer about a given molecule, so a correct answer may rest on the SMILES string or a language prior, and none confronts the model with a statement that is false for a closely matched molecule. Counterfactual explanation methods address property predictors instead: they search for an edited molecule that flips a prediction by enumerating a local chemical space, generating candidates with reinforcement learning, or deleting graph edges (Wellawatte et al., 2022; Numeroso & Bacciu, 2021; Lucic et al., 2022), and matched molecular pairs and activity cliffs quantify whether predictors track a chemical change (Janssen et al., 2026; van Tilborg et al., 2022); both explain or stress a predictor rather than test what a graph-language model has learned. We fill this gap with counter-to-fact statements: one functional-group or descriptor statement is scored on a real molecule for which it is true and on a scaffold-matched molecule for which it is false, so the verdict reflects the learned structure rather than output quality, and the same test serves as the learning signal. Section 3 formalizes the test and Section 4.2 applies it to the models above; Appendix A gives an extended comparison.

3 Counterfactual Evaluation and Learning Framework

Figure 2 summarizes the framework: the criterion and its benchmark, learning under the same criterion, and downstream transfer.

3.1 Problem Formulation

Let $g = (V, E)$ be a 2D molecular graph, s its SMILES string, and $c = (c_1, \ldots, c_T)$ a chemical statement. A molecular graph–language model encodes g into node tokens $\mathbf{H} = f_g(g) = [\mathbf{h}_1, \ldots, \mathbf{h}_{|V|}]$, compresses them through a cross-modal connector into N_q query tokens $\mathbf{Q} = f_q(\mathbf{H}) = [\mathbf{q}_1, \ldots, \mathbf{q}_{N_q}]$, and passes $\mathbf{Q}$, optionally with s, to a language model that scores c by its mean token log-likelihood

$$S_\theta(c \mid x) = \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(c_t \mid c_{<t}, x), \qquad x \in \{g, (g, s)\}. \tag{1}$$

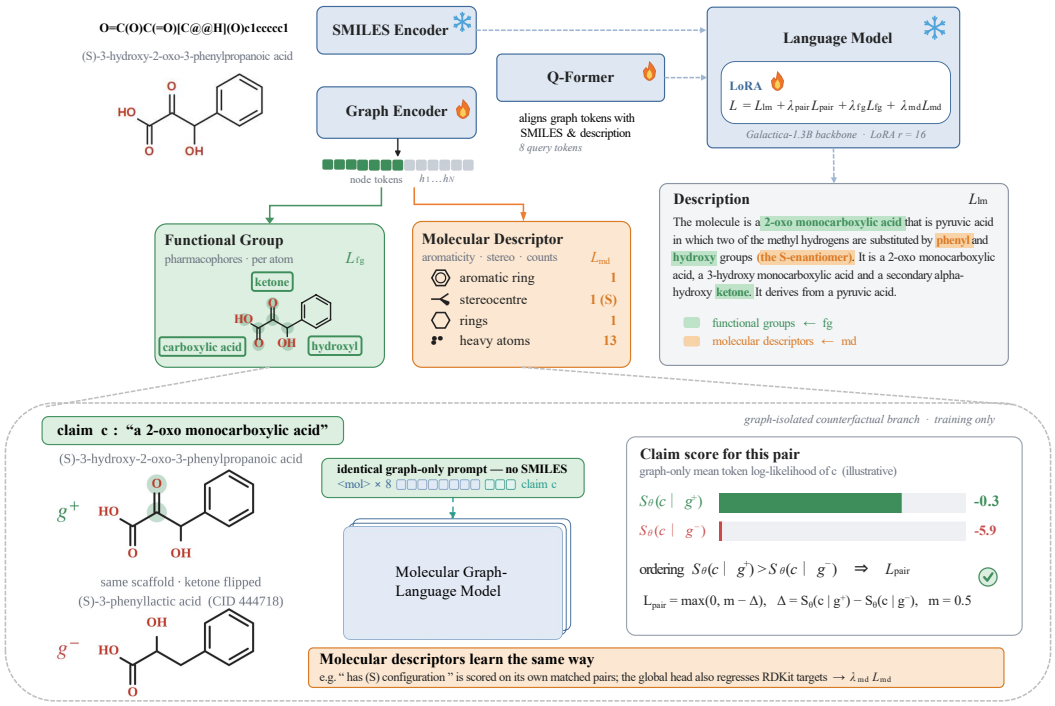

Figure 2: **Counterfactual evaluation and learning framework on a ChEBI-20 example.** Top: the graph encoder, Q-Former, and LoRA-adapted language model are trained with $\mathcal{L}_{\text{lm}}$ on the description, $\mathcal{L}_{\text{fg}}$ on per-atom functional groups, and $\mathcal{L}_{\text{md}}$ on molecular descriptors. Bottom: one claim is scored on a real scaffold-matched pair under an identical graph-only prompt, and $\mathcal{L}_{\text{pair}}$ enforces $S_\theta(c \mid g^+) > S_\theta(c \mid g^-)$, the ordering that defines the benchmark.

The graph-isolated score $S_\theta(c \mid g)$ uses the same prompt for every molecule and omits the SMILES string, so the language model cannot match a substring such as `Cl`. *Structure use* here means an ordering that survives this graph-only prompt. The MolCA GIN embeds atomic number and a chirality tag on atoms, and bond type and bond direction on edges. Formal charge is not one of those channels. The joint prompt is reported alongside. For a target attribute a with indicator $A_a(g) \in \{0, 1\}$, we pair a feature-present molecule g^+ ($A_a(g^+) = 1$) with a feature-absent molecule g^- ($A_a(g^-) = 0$). A statement c_a uses the target structure when $S_\theta(c_a \mid g^+) > S_\theta(c_a \mid g^-)$. We apply this criterion to two types of facts that are verifiable from an isolated 2D structure (Figure 2): *functional groups* (FG), local substructures anchored to specific atoms, and *molecular descriptors* (MD), global attributes such as stereochemistry, formal charge, aromaticity, and counts. Claims that need 3D conformations, assay conditions, or external knowledge fall outside this scope.

3.2 Counterfactual Benchmark

Real matched pairs. Each instance is a tuple (g^+, g^-, c_a, a) of two real, chemically valid molecules. Pair mining accepts a candidate for attribute a only if both molecules share a Bemis–Murcko scaffold key, exactly one of them carries a, and every other tracked attribute is equal: the remaining functional-group bits and the net charge, fragment, ring, aromatic-ring, stereocenter, stereobond, and carbon-chain counts. Heavy-atom count is not matched, since the target group itself adds atoms (audit in Appendix B). Among the partners of one positive molecule, the 16 most similar by Morgan-fingerprint Tanimoto similarity are kept. The statement c_a is identical for both endpoints, so the language form stays fixed while its supporting evidence changes, and unlike subgraph deletion, no invalid or distribution-shifted graph is created. FG axes cover 21 SMARTS-defined groups, such as carboxylic acids, hydroxyls, ketones, esters, amides, and halogens; MD axes cover stereochemical presence, formal-charge sign, and hydration, the descriptors that can flip within a scaffold. Each axis has a chemical definition, a statement template, and an RDKit verifier (Appendix B).

Protocol and metric. Molecules are assigned to scaffold partitions before mining, so training, validation, and test pairs never share a scaffold. For a benchmark set $\mathcal{P}$, pairwise accuracy is

$$\mathrm{Acc}_{\mathrm{pair}} = \frac{1}{|\mathcal{P}|} \sum_{(g^+, g^-, c_a, a) \in \mathcal{P}} \mathbf{1}\big[S_\theta(c_a \mid g^+) > S_\theta(c_a \mid g^-)\big], \tag{2}$$

reported overall and per attribute type, with 95% confidence intervals from resampling scaffold clusters, so pairs from one scaffold are not counted as independent evidence.

Stage-wise diagnosis. Pairwise accuracy reports whether a statement tracks its structure; three probes locate where the evidence is lost between graph and text. A linear probe on frozen node tokens $\mathbf{h}_i$ predicts the FG membership of each atom, scored by macro balanced accuracy on scaffold-held-out molecules; a linear probe on the mean query token $\bar{\mathbf{q}} = N_q^{-1} \sum_n \mathbf{q}_n$ predicts each molecular descriptor and passes a readability gate when a count reaches held-out $R^2 > 0.5$, or when a binary field reaches balanced accuracy at least 0.70 and its cluster-interval lower bound exceeds 0.5; and a graph-swap test fixes s and the prompt, replaces g by the graph of a random test molecule or shuffles its atoms, and measures the change in the decoded description. These probes separate evidence that is absent from the node tokens, lost in connector pooling, or ignored by the language model.

3.3 Counterfactual Structure Learning

Backbone and standard pathway. We instantiate the framework with MolCA (Liu et al., 2023c): f_g is a GIN graph encoder, f_q is a Q-Former with $N_q = 8$ learnable queries, and a projector maps $\mathbf{Q}$ into a LoRA-adapted Galactica-1.3B language model, which also reads s. For a reference description y of length L, the standard objective is $\mathcal{L}_{\mathrm{lm}} = -\frac{1}{L} \sum_l \log p_\theta(y_l \mid y_{<l}, \mathbf{Q}, s)$.

Graph-isolated pairwise objective. In parallel, the bottom branch of Figure 2 scores a claim with the prompt $[\mathbf{Q}; c]$ and no SMILES. Both endpoints receive this identical prompt, so the gap $\Delta_a = S_\theta(c_a \mid g^+) - S_\theta(c_a \mid g^-)$ can arise only from their graphs. We optimize

$$\mathcal{L}_{\mathrm{pair}} = \frac{1}{|\mathcal{B}_{\mathrm{pair}}|} \sum_{(g^+, g^-, c_a, a) \in \mathcal{B}_{\mathrm{pair}}} \max(0, m - \Delta_a), \tag{3}$$

with margin $m = 0.5$, which directly optimizes the ordering measured by Equation (2); the hinge is active only while a pair is misordered or separated by less than m. For each molecule in a caption batch that carries a supervised attribute, one partner is drawn from a pair bank mined inside the training scaffold partition under the constraints of Section 3.2, and only the graph-only branch scores it. Real matched molecules thus act as in-distribution hard negatives.

3.4 Functional Group and Molecular Descriptor Learning

The pairwise objective specifies which claims must change across a pair, but not where the evidence is stored. Because the control loses FG evidence in its node embeddings and MD evidence in query pooling (Section 4.2), we supervise each type at its own granularity with labels that RDKit derives from training molecules only; the labels are never used at inference.

Functional groups. Let $\mathcal{K}$ be ten SMARTS-defined groups and $M_k(g) \subseteq V$ the atoms matched by group k. Each atom receives the multi-label target $y_{i,k} = \mathbf{1}[i \in M_k(g)]$, predicted from its node token before the Q-Former, where a few atoms would otherwise be diluted by pooling:

$$\mathcal{L}_{\mathrm{fg}} = \frac{1}{|V||\mathcal{K}|} \sum_{i \in V} \sum_{k \in \mathcal{K}} \mathrm{BCE}\big(\sigma(\mathbf{w}_k^\top \mathbf{h}_i + b_k),\, y_{i,k}\big). \tag{4}$$

Molecular descriptors. RDKit computes three binary descriptors (stereochemical presence, aromatic-ring presence, hydration) and five counts (heavy atoms, rings, aromatic rings, longest carbon chain, assigned stereocenters), standardized with training statistics as $\tilde{d}_j$. Because a descriptor summarizes the whole molecule, it is predicted from the mean query token $\bar{\mathbf{q}}$, the same readout that the diagnosis probes:

$$\mathcal{L}_{\mathrm{md}} = \sum_{j \in \mathrm{bin}} \mathrm{BCE}\big(\sigma(\mathbf{u}_j^\top \bar{\mathbf{q}}), d_j\big) + \sum_{j \in \mathrm{cnt}} \mathrm{SmoothL1}\Big(\mathbf{v}_j^\top \bar{\mathbf{q}}, \tilde{d}_j\Big). \tag{5}$$

Pairs for both types. FG pairs use the claim "This molecule contains $\langle$group $k\rangle$" for the ten supervised groups and eleven further library groups. MD descriptors that can flip within a scaffold receive fixed claims such as "This molecule is chiral." Together they form $\mathcal{B}_{\text{pair}} = \mathcal{B}_{\text{FG}} \cup \mathcal{B}_{\text{MD}}$. The heads make FG and MD *readable*, $\mathcal{L}_{\text{pair}}$ makes claim scores *follow* them, and $\mathcal{L}_{\text{lm}}$ lets the description *express* them.

3.5 Overall Objective and Downstream Transfer

We optimize

$$\mathcal{L} = \mathcal{L}_{\text{lm}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{fg}}\mathcal{L}_{\text{fg}} + \lambda_{\text{md}}\mathcal{L}_{\text{md}}, \tag{6}$$

with $\lambda_{\text{pair}} = 0.05$ and $\lambda_{\text{fg}} = \lambda_{\text{md}} = 0.1$, selected on validation data. The graph encoder, Q-Former, projector, LoRA adapters ($r = 16$), and head weights are trained, while the Galactica backbone stays frozen.

End-to-end transfer. For each downstream task τ, we finetune the same parameters end to end with graph and SMILES inputs, replacing only $\mathcal{L}_{\text{lm}}$ in Equation (6) by the task loss $\mathcal{L}^{\tau}_{\text{task}}$: FG and MD labels are recomputed with RDKit for the task molecules and the pair branch keeps drawing partners from the training bank, so FG and MD information stays readable while the task loss reshapes the representation. Molecular description, IUPAC naming, and text-based generation use the sequence likelihood of the target description, name, or SMILES; molecule–text retrieval applies a symmetric InfoNCE loss between $\bar{\mathbf{q}}$ and the Q-Former text embedding; and property prediction attaches a linear classifier to the final hidden state of the language model. Matched partners, RDKit labels, and the graph-isolated branch are used only in training, so inference takes a single molecule through the task pathway with no added latency (Appendix C).

4 Experiments

Section 4.2 answers the question of Section 1; Sections 4.3 to 4.5 test whether counterfactual learning raises and generalizes structural sensitivity, whether it transfers to molecular tasks, and which components matter.

4.1 Experimental Setup

Data. The molecule–text collections are published datasets. The counterfactual pairs are mined by us and are not released with those datasets. ChEBI-20 (Edwards et al., 2022) supplies molecular description and text-based generation on its published split of 26,407/3,301/3,300 molecules. The 1,146 held-out pairs are mined from molecules in that test split: after Bemis–Murcko scaffold assignment, both endpoints must share a scaffold, exactly one carries the target attribute, and every other tracked attribute is equal, with at most 16 Morgan-Tanimoto neighbors kept per positive molecule (Section 3.2). Training pairs use the same rules inside the training split and share no scaffold with these 1,146 pairs. The 822 external pairs are not FGBench's property questions. We start from FGBench pairs (Liu et al., 2025) and remove any pair whose endpoint matches a ChEBI-20 molecule by canonical SMILES, InChIKey, or a nonempty cyclic scaffold, leaving 822 pairs in 535 scaffold clusters. PubChem324k uses the MolCA splits for captioning, IUPAC naming, and retrieval. PCdes uses its own split as an independent retrieval test (Liu et al., 2023c; Zeng et al., 2022). Eight MoleculeNet classification tasks use their published scaffold splits (Wu et al., 2018).

Baselines and metrics. The controlled baseline is MolCA with identical initialization, data, budget, and LoRA configuration but without the structural objectives. Public MolT5, ATOMAS, HIGHT, and LLaMo checkpoints are scored with the same templates, pairs, and scoring rule (Edwards et al., 2022; Zhang et al., 2025; Chen et al., 2025; Park et al., 2024). Downstream comparisons additionally report published UniMoT, MoMu, and MolFM results (Guo et al., 2025; Su et al., 2022; Luo et al., 2023); systems with different backbones or corpora provide context rather than a controlled ranking. We report pairwise accuracy with scaffold-cluster 95% intervals, BLEU, ROUGE, and METEOR (Papineni et al., 2002; Lin, 2004; Banerjee & Lavie, 2005), claim contradiction and coverage, retrieval accuracy, and ROC-AUC. Headline results average three seeds; all selection uses validation data (Appendix D).

Table 1: **Structural sensitivity of molecular graph–language models.** All models are scored with the same statement templates on the same held-out matched pairs (FG, MD, and overall) and on the 822 external pairs. BLEU-4 is ChEBI-20 description quality. **Bold** and underlined mark the best and second-best values.

Model	Input	BLEU-4	FG Acc.	MD Acc.	Overall	External
MolT5-Large	SMILES	50.8	0.628	0.541	0.603	0.611
ATOMAS-base	SMILES	54.9	0.431	0.398	0.422	0.437
HIGHT (Vicuna-7B)	Graph	40.5	0.594	0.529	0.576	0.570
HIGHT (LLaMA2-7B)	Graph	–	0.702	0.603	0.675	0.664
LLaMo	Graph + SMILES	54.8	0.818	0.731	0.795	0.781
MolCA control	Graph + SMILES	53.1	0.487	0.479	0.485	0.489
Ours	Graph + SMILES	**55.4**	**0.931**	**0.872**	**0.915**	**0.928**

4.2 Structural Sensitivity of Existing Models

Existing models are weakly sensitive to the structure behind a statement, and their output quality does not reveal it. Table 1 scores seven systems with the same templates on the same 1,146 held-out pairs and 822 external pairs. The MolCA control reaches 53.1 BLEU-4 on ChEBI-20 yet orders only 0.485 of the held-out pairs and 0.489 of the external pairs correctly, chance level. LLaMo and ATOMAS reach nearly identical BLEU-4 (54.8 and 54.9) but pairwise accuracies of 0.795 and 0.422, and across the five systems with a ChEBI-20 score BLEU-4 and pairwise accuracy have a Pearson correlation of 0.01 (Figure 1a). Hierarchical inputs help, as HIGHT and LLaMo exceed the MolCA control, yet every baseline fails on at least one fifth of the pairs and is weaker on molecular descriptors than on functional groups.

Structural evidence is lost before the language model. The probes of Section 3.2 locate the loss at two stages of the MolCA control (Figure 1b). The node-token probe reads functional-group membership with only 0.482 macro balanced accuracy on 2,798 scaffold-held-out molecules, so the graph encoder itself does not separate atoms by group. The query-token probe passes the readability gate for none of the seven molecular descriptors. Net charge and element counts remain weakly readable from mean-pooled queries after supervision ($R^2 = 0.291$ and 0.369). A raw sum of the pre-embedding atom-feature vector recovers both, but the encoder above does not receive formal charge, so this gap is not only a pooling loss.

The graph is a passenger during captioning. With the SMILES string and prompt fixed, shuffling atoms within each graph or swapping in the graph of a random test molecule changes the control BLEU-2 only from 60.9 to 59.7 and 60.2 (Figure 1c): its captions are driven by the SMILES string and the language prior. Under the same swap, our model falls from 64.2 to 50.2, so its descriptions depend on the graph pathway.

4.3 Learning and Generalization

Statements follow the structure after counterfactual learning. Training with Equation (6) raises held-out pairwise accuracy from 0.485 to 0.915 ± 0.009 over three seeds, a paired gain of 43.0 points (cluster interval [32.3, 50.2]). The gain holds for both types (FG 0.487→0.931, MD 0.479→0.872) and on the hardest pairs with Tanimoto similarity at least 0.6 (0.519→0.864). Under the joint graph and SMILES prompt, accuracy stays within 1.2 points of the graph-only value, so the gain does not exploit SMILES differences. Mismatched claims stay near chance, and pairs whose g^- is the larger molecule are ordered as well as the rest, so neither the template direction nor the larger-graph rule accounts for the gain (Appendix E). Hydroxyl, ether, ester, and ketone axes improve most, whereas amine and halogen axes remain harder (Appendix E).

Readability and ordering come from different losses. Node-level functional-group balanced accuracy rises from 0.482 to 0.700, and all seven descriptor fields pass their gates (binary fields 0.973 and 0.988; count R^2 between 0.558 and 0.773). These readouts are the supervised heads. Section 4.5 shows that removing $\mathcal{L}_{\text{pair}}$ returns pairwise accuracy to 0.503 while the heads stay readable, and removing $\mathcal{L}_{\text{fg}}$ leaves pairwise accuracy at 0.908 while functional-group balanced accuracy falls to 0.521. Chloro ordering reaches 0.851 while its node balanced accuracy stays at 0.491. Paraphrased

Table 2: **Downstream tasks.** Description (ChEBI-20), IUPAC naming and retrieval (PubChem324k), text-based molecule generation (ChEBI-20, exact match and Morgan FTS in %), and mean ROC-AUC. The MolCA row is the phase-matched control, and its property entry is the eight-task mean. Published systems use other backbones or task sets and are not ranked against this control. – denotes a result that is not reported.

	Description		IUPAC		Generation		Retrieval		Property
Model	BLEU-4	METEOR	BLEU-4	METEOR	Exact	FTS	M2T	T2M	AUC
MolT5-Large	50.8	61.4	49.7	58.5	31.8	75.0	–	–	–
MolCA	53.1	65.1	66.6	72.1	–	–	66.6	66.0	73.8
ATOMAS-base	54.9	–	–	–	34.3	77.3	–	–	77.0
LLaMo	54.8	66.6	–	73.4	–	–	–	–	–
UniMoT	58.3	70.3	–	–	–	–	69.5	69.8	75.6
Ours	55.4	67.6	70.1	75.2	38.0	79.0	72.0	72.0	77.1

templates and probes on groups that no head supervises still rise with the full model (Appendix E), so the ordering is not one training sentence and not only the trained linear head.

External and unseen-attribute generalization. On the 822 external pairs, disjoint from ChEBI-20 molecules and scaffolds, the model reaches 0.928 (95% CI [0.905, 0.949]) against 0.489 ([0.451, 0.526]) for the control, a paired gain of 43.9 points ([39.5, 48.4]). Whether individual FGBench molecules appeared in Galactica pretraining cannot be identified from the public release. The external set is dominated by chloro changes (619 pairs) and HIV-sourced molecules (711 pairs). Removing chloro pairs still yields a 19.2-point gain, removing HIV pairs a 43.2-point gain, and the macro-average over chloro, bromo, fluoro, and nitro is 29.6 points. The pooled Other slice ($n = 33$) gains 1.8 points and is not in that average. Retraining without nitro, nitrile, sulfonamide, thiol, and iodo pairs still raises accuracy on those unseen axes from 0.495 to 0.777 on 103 held-out pairs of those groups. That manifest is separate from the Other FG bucket of the 1,146-pair benchmark (Appendix E).

4.4 Downstream Molecular Tasks

We finetune end to end on five tasks. Inference takes one molecule and does not use a matched pair. Table 2 reports the standard metric for each task. Structure-sensitive slices are in Appendix F.

Molecule description. Given a 2D molecular graph, the model writes a natural-language description of the molecule (Edwards et al., 2022). We use the published ChEBI-20 split and score captions with BLEU and METEOR (Papineni et al., 2002; Banerjee & Lavie, 2005). Relative to the phase-matched MolCA control, caption quality rises, and the caption audit in the appendix moves more than BLEU.

IUPAC naming. The model maps the graph to the systematic name (Liu et al., 2023c). We use the MolCA split of PubChem324k, built from PubChem molecule–text records (Kim et al., 2023). The gain over the same control is larger than on description and falls on locants, functional-group suffixes, and stereodescriptors.

Text-based molecule generation. Given a description, the model writes a SMILES string (Edwards et al., 2022). We use ChEBI-20. Exact match and Morgan fingerprint similarity (Rogers & Hahn, 2010) are compared with published systems in Table 2. This task has no phase-matched MolCA control.

Molecule–text retrieval. The model retrieves the description of a given molecule, and the molecule of a given description (Liu et al., 2023c). Training and the main test use the MolCA split of PubChem324k. PCdes is an additional test with its own split (Zeng et al., 2022). The larger gain is on functional-group queries.

Property prediction. The model classifies bioactivity and toxicity labels from the molecular graph. We use eight binary tasks from MoleculeNet under scaffold splits, scored by mean ROC-AUC (Wu et al., 2018). Per-task scores and the BBBP label-flip pairs are in the appendix.

Table 3: **Component ablation.** FG BA is node-level functional-group balanced accuracy; MD fields counts descriptors passing the readability gate; Contr. is the direct-claim contradiction rate of decoded ChEBI-20 captions.

Variant	Pair Acc.↑	External↑	FG BA↑	MD fields↑	BLEU-2↑	Contr.↓
Full model	**0.915**	**0.928**	**0.700**	**7/7**	**64.2**	**12.8**
w/o $\mathcal{L}_{\text{pair}}$	0.503	0.508	0.694	7/7	63.1	14.6
w/o graph isolation	0.599	0.611	0.689	7/7	63.5	14.1
erasure instead of real pairs	0.322	0.351	0.672	7/7	58.4	15.9
w/o $\mathcal{L}_{\text{fg}}$	0.908	0.917	0.521	7/7	62.7	15.8
w/o $\mathcal{L}_{\text{md}}$	0.911	0.922	0.697	1/7	63.0	14.9
w/o end-to-end transfer	0.915	0.928	0.700	7/7	60.8	18.5
Caption-only control	0.485	0.489	0.482	0/7	60.9	18.5

4.5 Ablation and Analysis

Table 3 removes one component at a time under the same initialization, budget, and three seeds. Removing $\mathcal{L}_{\text{pair}}$ returns discrimination to chance while the heads stay readable. Letting the pairwise branch see SMILES caps accuracy at 0.599, and replacing real partners with erased subgraphs drops it to 0.322. Removing $\mathcal{L}_{\text{fg}}$ or $\mathcal{L}_{\text{md}}$ leaves pairwise accuracy near 0.91 but erases its own readability. Without end-to-end transfer, captions return to the control. Margins in [0.25, 1.0] are stable, half of the training pairs reach most of the gain, and training costs 1.3× the control with no inference overhead (Appendix G).

Qualitative analysis. Matched-pair captions, text-based generation, IUPAC naming, retrieval, and BBBP cases are in Appendix H. On 300 captions, inter-annotator Cohen's κ is 0.81, and verifier–human agreement is 91.7%.

5 Conclusion

We introduced a counterfactual framework that evaluates and improves the use of molecular structure in molecular graph–language models by holding one chemical statement fixed and scoring it on real, scaffold-matched molecules that differ in a target attribute. Under this test, output quality does not predict structural sensitivity, and the control's linear probes are weak on functional groups and descriptors. A graph-isolated pairwise objective raises that ordering on held-out and external pairs. Functional-group and descriptor heads make those attributes linearly readable, and end-to-end finetuning carries the matched control gains into description, naming, retrieval, and property prediction.

AI Use Statement

AI tools were used to help analyze large batches of data, to review code written by the authors for potential errors, and to correct language in the manuscript written by the authors. The authors verified all code, claims, citations, proofs, and results and remain responsible.

Ethics Statement

This work uses only public molecular datasets (ChEBI-20, PubChem324k, PCdes, MoleculeNet, and FGBench) under their original licenses, and no personal or sensitive data. The human audit was performed by chemists who volunteered and were informed of the purpose of the study. Structure-sensitive molecular models could in principle assist the design of harmful compounds; our models describe, name, retrieve, and classify existing molecules and do not generate optimized candidates for a target activity, and we release them for research use only. We report all selection rules and uncertainty estimates to support research integrity.

Reproducibility Statement

Appendix I lists the public datasets, pair manifests, training configuration, and evaluation protocol, and Appendix D gives all hyperparameters. Headline structural results and property-prediction means are reported over three training seeds, and the anonymous supplement contains the pair miner, SMARTS library, training code, and evaluators.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. MolecularIQ: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RqwEzZqMFv.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 354–379. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.25/.

Yongqiang Chen, Quanming Yao, Juzheng Zhang, James Cheng, and Yatao Bian. Hierarchical graph tokenization for molecule-language alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9664–9690. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25cf.html.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6140–6157. PMLR, 2023. URL https://proceedings.mlr.press/v202/christofidellis23a.html.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47/.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26.

Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. MolTC: Towards molecular relational modeling in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1943–1958. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.116. URL https://aclanthology.org/2024.findings-acl.116/.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=Tlsdsb6l9n.

Shikun Feng, Lixin Yang, Yanwen Huang, Yuyan Ni, Wei-Ying Ma, and Yanyan Lan. UniMAP: Universal SMILES-graph representation learning. *arXiv preprint arXiv:2310.14216*, 2023. URL https://arxiv.org/abs/2310.14216.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12994–13013. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.760. URL https://aclanthology.org/2024.findings-emnlp.760/.

Dimitris Gkoumas and Maria Liakata. Less for more: Enhanced feedback-aligned mixed LLMs for molecule caption generation and fine-grained NLI evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2882–2902. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.144. URL https://aclanthology.org/2025.acl-long.144/.

Shuhan Guo, Yatao Bian, Ruibing Wang, Nan Yin, Zhen Wang, and Quanming Yao. Unified molecule-text language model with discrete token representation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9205–9213, 2025. doi: 10.24963/ijcai.2025/1023. URL https://www.ijcai.org/proceedings/2025/1023.

Yinhan He, Zaiyi Zheng, Patrick Soga, Yaochen Zhu, Yushun Dong, and Jundong Li. Explaining graph neural networks with large language models: A counterfactual perspective on molecule graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7079–7096. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.415. URL https://aclanthology.org/2024.findings-emnlp.415/.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. Structural reasoning improves molecular understanding of LLM. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21016–21036. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1023. URL https://aclanthology.org/2025.acl-long.1023/.

Kerrin Janssen, Jan M. Wollschläger, Jonny Proppe, and Andreas H. Göller. When machine learning models learn chemistry I: Quantifying explainability with matched molecular pairs. *Digital Discovery*, 5:571–582, 2026. doi: 10.1039/D5DD00398A. URL https://pubs.rsc.org/en/content/articlehtml/2026/dd/d5dd00398a.

Zewei Ji, Runhan Shi, Jiarui Lu, Fang Li, and Yang Yang. ReLMole: Molecular representation learning based on two-level graph similarities. *Journal of Chemical Information and Modeling*, 62(22):5361–5372, 2022. doi: 10.1021/acs.jcim.2c00798. URL https://doi.org/10.1021/acs.jcim.2c00798.

Seojin Kim, Hyeontae Song, Jaehyun Nam, and Jinwoo Shin. Training text-to-molecule models with context-aware tokenization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22442–22460. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1221. URL https://aclanthology.org/2025.findings-emnlp.1221/.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956. URL https://doi.org/10.1093/nar/gkac956.

Alec Lamens and Jürgen Bajorath. Generation of molecular counterfactuals for explainable machine learning based on core-substituent recombination. *ChemMedChem*, 19(3):e202300586, 2024. doi: 10.1002/cmdc.202300586. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/cmdc.202300586.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3D molecule-text interpretation in language models. In *The Twelfth*

International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=xI4yNlkaqh.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073, 2024. doi: 10.1016/j.compbiomed.2024.108073. URL https://doi.org/10.1016/j.compbiomed.2024.108073.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023a. doi: 10.1038/s42256-023-00759-6. URL https://www.nature.com/articles/s42256-023-00759-6.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. FGBench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bcaa0c3f644f580857a56f38e305867e-Paper-Datasets_and_Benchmarks_Track.pdf.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1606–1616. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-short.138. URL https://aclanthology.org/2023.acl-short.138/.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15623–15638, 2023c. doi: 10.18653/v1/2023.emnlp-main.966. URL https://aclanthology.org/2023.emnlp-main.966.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. MoleculeQA: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3769–3789, 2024. doi: 10.18653/v1/2024.findings-emnlp.216. URL https://aclanthology.org/2024.findings-emnlp.216.

Ana Lucic, Maartje A. Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511. PMLR, 2022. URL https://proceedings.mlr.press/v151/lucic22a.html.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. MolFM: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023. URL https://arxiv.org/abs/2307.09484.

Kha-Dinh Luong and Ambuj K. Singh. Fragment-based pretraining and finetuning on molecular graphs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17584–17601, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/38ec60a949c3538e5cbb337b1b386dcf-Abstract-Conference.html.

Danilo Numeroso and Davide Bacciu. MEG: Generating molecular counterfactual explanations for deep graph networks. In *2021 International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2021. doi: 10.1109/IJCNN52387.2021.9534266. URL https://ieeexplore.ieee.org/document/9534266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Hyuntae Park, Yeachan Kim, and SangKeun Lee. Bridging the gap between molecule and textual descriptions via substructure-aware alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23459–23479. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1197. URL https://aclanthology.org/2025.emnlp-main.1197/.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J. Kim. LLaMo: Large language model-based molecular graph assistant. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WKTNdU155n.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1123. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.70. URL https://aclanthology.org/2023.emnlp-main.70/.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30458–30490. PMLR, 2023. URL https://proceedings.mlr.press/v202/seidl23a.html.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022. URL https://arxiv.org/abs/2209.05481.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, 2022. doi: 10.1021/acs.jcim.2c01073. URL https://doi.org/10.1021/acs.jcim.2c01073.

Sarveswara Rao Vangala, Sowmya Ramaswamy Krishnan, Navneet Bung, Rajgopal Srinivasan, and Arijit Roy. pBRICS: A novel fragmentation method for explainable property prediction of drug-like small molecules. *Journal of Chemical Information and Modeling*, 63(16):5066–5076, 2023. doi: 10.1021/acs.jcim.3c00689. URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c00689.

Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science*, 13:3697–3705, 2022. doi: 10.1039/D1SC05259D. URL https://pubs.rsc.org/en/content/articlehtml/2022/sc/d1sc05259d.

Yuyang Wu, Jinhui Ye, Shuhao Zhang, Lu Dai, Yonatan Bisk, and Olexandr Isayev. MolErr2Fix: Benchmarking LLM trustworthiness in chemistry via modular error detection, localization, explanation, and correction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19354–19371. Association for Computational Linguistics,

2025. doi: 10.18653/v1/2025.emnlp-main.977. URL https://aclanthology.org/2025.emnlp-main.977/.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A. URL https://doi.org/10.1039/C7SC02664A.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/391aae9ccaf95a3d80dde4c8408f2c91-Abstract-Conference.html.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6:34, 2023. doi: 10.1038/s42004-023-00825-5. URL https://www.nature.com/articles/s42004-023-00825-5.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13(1):862, 2022. URL https://www.nature.com/articles/s41467-022-28494-3.

Yikun Zhang, Geyan Ye, Chaohao Yuan, Bo Han, Long-Kai Huang, Jianhua Yao, Wei Liu, and Yu Rong. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mun3bGqdDM.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15870–15882, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/85267d349a5e647ff0a9edcb5ffd1e02-Abstract.html.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20454–20468, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/129033c7c08be683059559e8d6bfd460-Abstract-Conference.html.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.