MOLTOKSYN: MOLECULE-TEXT RETRIEVAL WITH TEXT-CONDITIONED TOKENS AND STRUCTURE-GROUNDED SYNTHETIC CAPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural language provides a flexible way to search for molecules and describe the chemical properties that matter for a given task. However, existing molecule–text datasets rely largely on database or literature descriptions that vary widely in content and do not consistently emphasize properties that can be derived from molecular structure. We address this limitation with MOLSYNTEXT, a synthetic molecule–text dataset whose descriptions are generated from structured molecular information and screened for consistency with their source. We further introduce MOLTOKSYN, a graph–text retrieval model that represents each molecule with multiple learned molecular tokens and uses the text query to determine how these tokens are combined for scoring, allowing different aspects of the same molecule to be emphasized for different descriptions. Across bidirectional molecule–text retrieval, MOLTOKSYN consistently outperforms strong existing methods on MOLSYNTEXT and remains competitive on matched PubChem-derived text, while its learned molecular representations transfer strongly to downstream property-prediction tasks. Because retrieval alone does not reveal which chemical concepts are captured by the learned representations, we additionally introduce a controlled Semantic Benchmark covering structural, physicochemical, and relational properties. On this benchmark, MOLSYNTEXT supervision improves MOLTOKSYN on most semantic categories, with particularly strong gains for descriptor-grounded properties. Together, these results show that structure-grounded synthetic descriptions can provide informative chemical supervision, especially when combined with text-conditioned molecular representations.

1 INTRODUCTION

Molecular structure is a primary determinant of many chemical and physicochemical properties. Small structural changes can alter properties such as polarity, lipophilicity, hydrogen bonding, or molecular flexibility, and chemists naturally describe such differences using language: a desired compound may be "less lipophilic", "more polar", or contain a particular functional group. Connecting these descriptions to molecular structure therefore offers an intuitive interface for searching chemical space. A useful molecule–text model should translate between the way a chemist describes a molecule and the structural information that gives rise to those properties, enabling molecules to be retrieved from textual specifications and molecular structures to be associated with meaningful natural-language descriptions (Edwards et al., 2021; Su et al., 2022; Liu et al., 2023a).

Learning such a connection requires descriptions that consistently expose meaningful chemical information about the molecule. Existing molecule–text resources provide valuable supervision, but database- and literature-derived descriptions are often short, sparse, and only indirectly related to molecular structure. Many records focus on compound names, class labels, biological context, occurrence statements, or database-specific information, while structural and physicochemical properties are described less consistently. As a result, the supervision can be heterogeneous and may encourage models to rely on identity- or source-specific cues rather than learning the chemical properties encoded by the molecular structure. Modern language models provide an attractive way to generate descriptions at scale, but unconstrained generation introduces the opposite problem: fluent chemical text can contain claims that are not supported by the molecule. We therefore argue that

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.

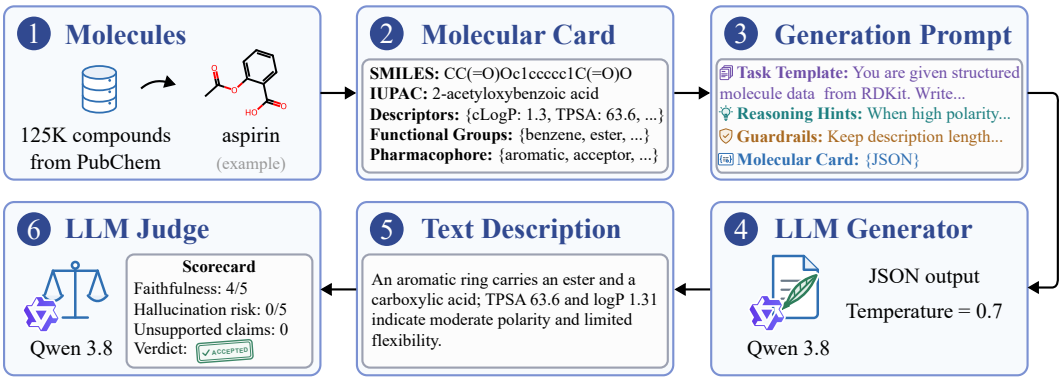


Figure 1: Overview of the MOLSYNTEXT construction pipeline. Each molecule is converted into a structured molecular card containing computed chemical information, combined with controlled generation instructions, and verbalized by Qwen3.8-27B. Candidate descriptions are checked against the source molecular information before being retained.

synthetic molecular language is most useful when generation is grounded in chemical information computed directly from the structure.

Following this idea, we introduce MOLSYNTEXT, a structure-grounded molecule–text dataset in which a language model acts as a *verbalizer* rather than a source of new chemical facts. For each molecule, deterministic descriptors, functional-group indicators, structural statistics, and other interpretable features are collected into a molecular card. The language model converts this card into a natural-language description under explicit grounding constraints, and a separate judge screens the generated text against the same molecular information. This construction allows the textual supervision to emphasize structure-derived and physicochemical properties while keeping its origin traceable to the molecular structure. Fig. 1 summarizes the generation pipeline.

Better text also exposes a limitation of dual encoders. One molecule has many valid descriptions, about hydrogen bonding, aromaticity, or flexibility, yet a dual encoder must serve all of them with one vector. MOLTOKSYN instead represents each molecule with several learned tokens and lets the text query attend over them, following the mixture-token design of Llip (Lavoie et al., 2024). The tokens come from one-way readout nodes added to a pretrained Chemprop D-MPNN (Yang et al., 2019; Heid et al., 2024; Burns et al., 2026), so message passing over the real atoms and bonds is unchanged. The query-conditioned score is the retrieval score itself, not a reranking step on top of a global embedding.

Our contributions are:

- **MOLSYNTEXT**: 216K structure-grounded descriptions of 125K molecules, generated from RDKit molecular cards, judge-filtered, and matched molecule-for-molecule with STM, a filtered subset of PubChemSTM (Liu et al., 2023a).
- **MOLTOKSYN**: a molecule–text retrieval model whose molecular representation depends on the query. On MOLSYNTEXT it improves R@1 over the strongest baseline by 9–11 points; on PubChem text it is best on 5 of 8 metrics. Its frozen features rank first or second on all seven OGB classification tasks.
- **Controlled evidence that the text matters**: with molecules, splits, and architecture fixed, replacing PubChem text with MOLSYNTEXT improves linear probing in 50 of 60 model–task pairs and raises accuracy for all six models on our new Semantic Benchmark, which tests individual chemical concepts that retrieval metrics cannot isolate.

2 RELATED WORK

Molecule–text retrieval. Early molecule–text retrieval models such as MoMu and MoleculeSTM learn one global representation for a molecule and one for its textual description, and compare them

in a shared embedding space (Su et al., 2022; Liu et al., 2023a). Later methods use richer molecular representations or stronger cross-modal interaction. MolCA connects a molecular graph encoder to text through learned query tokens (Liu et al., 2023b), MolFM introduces direct interaction between textual and molecular features (Luo et al., 2023), and 3D-MoLM incorporates three-dimensional molecular information into molecule–text learning (Li et al., 2024). More recent work such as Atomas further shows the value of fine-grained alignment between molecular and textual information (Zhang et al., 2025). Our work focuses on a complementary question: instead of assigning one fixed molecular embedding to all queries, can the molecular representation used for retrieval adapt to the description being matched?

Text-conditioned molecular representations. The idea of keeping several representations and selecting among them according to a query is well established in information retrieval (Humeau et al., 2020; Luo et al., 2022; Khattab & Zaharia, 2020). MolCA and MolFM bring related cross-modal interaction to molecular models, but use the interaction after or alongside an initial global retrieval stage (Liu et al., 2023b; Luo et al., 2023). In MOLTOKSYN, the text-conditioned interaction is the retrieval score itself. The molecular encoder produces several learned tokens that can be cached, and each text query combines them differently before the final molecule–text score is computed. Our tokens are obtained through one-way readout nodes inside the D-MPNN, so they collect complementary information from the graph without changing message passing between the real atoms and bonds.

Synthetic textual supervision. Existing molecule–text datasets often rely on descriptions collected from chemical databases, which can be short or focused on identity and database context rather than on structure-derived properties (Liu et al., 2023a). Synthetic enrichment has therefore been explored as a way to provide more informative molecular language; for example, 3D-MoLM uses GPT-generated enrichment for sparsely described PubChem molecules (Li et al., 2024). MOLSYNTEXT follows a more controlled approach. Its descriptions are generated from deterministic molecular cards and screened against the same structure-derived information. Because the same molecular identities are also paired with the STM reference text, we can study the effect of the textual supervision without changing the underlying molecule set.

3 SYNTHETIC MOLECULE–TEXT SUPERVISION

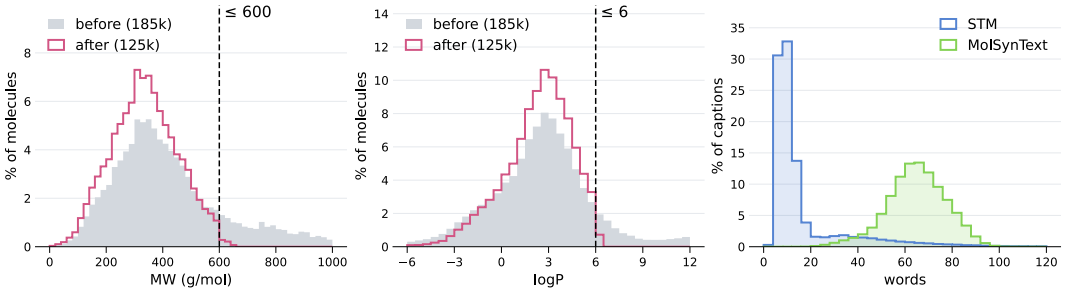
We use two text sources for the same set of molecules. STM contains PubChem-derived descriptions (Kim et al., 2025), while MOLSYNTEXT contains synthetic descriptions generated from structured molecular information. Because both text sources are paired with exactly the same molecular identities, differences between them are not caused by changes in the underlying molecule set.

3.1 MOLECULAR CORPUS AND STM REFERENCE VIEW

We start from the molecular structures released with MoleculeSTM (Liu et al., 2023a) and build a common set of 125,000 valid, unique molecules. The subset is selected mainly using a relaxed Lipinski-style filter (Lipinski et al., 2001), with thresholds of molecular weight ≤ 600, logP ≤ 6, HBD ≤ 6, and HBA ≤ 12, all computed using RDKit (RDKit Developers, 2026). This filter yields 123,328 molecules with valid text. The remaining 1,672 are sampled at random under slightly looser thresholds (molecular weight ≤ 650, logP ≤ 6.5, HBD ≤ 6, HBA ≤ 13). Figure 2(a,b) compares the molecular weight and logP distributions of the initial 185,285 molecules with those of the final set. The same molecular identifiers and structures are then used for both textual views. For the STM dataset, we reconstruct PubChemSTM-style supervision using the MoleculeSTM CID–SMILES mapping (Liu et al., 2023a). For each retained molecule, we retrieve its PubChem text, apply light normalization, and keep all available descriptions (1.23 per molecule on average).

3.2 MOLSYNTEXT: GROUNDED SYNTHETIC DESCRIPTIONS

For each molecule m_i, we construct a molecular card using deterministic RDKit-based feature extraction (RDKit Developers, 2026). The card summarizes information computable from the structure: physicochemical descriptors, ring and atom statistics, functional-group indicators, and chemically interpretable feature tags. The same source structure is used to build the card and to construct the graph consumed by the molecular encoder.



(a) Molecular weight (b) Crippen logP (c) Words per caption

Figure 2: Molecular and textual statistics of the datasets. **(a, b)** Molecular weight and Crippen logP of the initial dataset before filtering (185,285 molecules, grey) and of the final dataset after filtering (125,000 molecules, outline). Dashed lines mark the filter thresholds (MW $\leq$ 600, logP $\leq$ 6). **(c)** Words per caption over all captions of STM (154,253) and MOLSYNTEXT (216,015).

A locally hosted Qwen3.8-27B model (Yang et al., 2025; Qwen Team, 2026) then generates descriptions t_i from the card and a controlled prompt context. Each prompt combines a perspective template, optional chemistry-specific instructions, fixed hard constraints, and a serialized form of the molecular card. The templates vary the descriptive emphasis across structural motifs, functional groups, and physicochemical properties, while the hard constraints require a short English description, restrict the model to information supported by the card, and discourage unsupported biological or mechanistic claims. Generation is performed offline and the resulting text is stored with the molecular record; no language model is invoked during retrieval-model training.

This separation is important: the structured card is used only to construct the text target. The retrieval model never receives the card or the hand-computed descriptors as an additional molecular input; its molecule branch operates on the graph itself. The final construction pipeline also applies procedural checks before a candidate becomes a positive training pair. Outputs must satisfy the expected schema and contain a non-empty description. A separate Qwen3.8-27B judge compares the candidate against the molecular card and rejects descriptions containing unsupported claims or incorrectly copied information. These steps describe how MOLSYNTEXT is constructed. In total, MOLSYNTEXT contains 216,015 descriptions (1.73 per molecule on average, versus 1.23 for STM). Fig. 2(c) compares caption lengths across the two views. Full prompts, judge acceptance criteria, rejected examples, and post-processing details are given in Appendix B.

4 MOLTOKSYN: TEXT-CONDITIONED MOLECULAR TOKENS

Given paired data $\mathcal{D} = \{(m_i, t_i)\}_{i=1}^{N}$, MOLTOKSYN learns a compatibility score $s_\theta(m, t)$ that ranks matched molecule–text pairs above mismatched ones. A dual encoder maps m and t to fixed vectors and scores them by a dot product. MOLTOKSYN instead represents each molecule by K learned tokens and lets the text decide how these tokens are combined before scoring, following the mixture-token design of Llip (Lavoie et al., 2024). The same score ranks molecules for a text query and descriptions for a molecule query (Fig. 3). Throughout, i indexes molecules, j indexes descriptions and $\ell \in \{1, \ldots, K\}$ indexes molecular tokens.

4.1 MOLECULAR AND TEXT ENCODERS

Molecule branch. We use a Chemprop D-MPNN (Yang et al., 2019; Heid et al., 2024) initialized from CheMeleon (Burns et al., 2026). For a directed bond $a \to b$ with atom features x_a and bond features e_{ab}, the hidden states follow the standard update

$$h_{ab}^{(0)} = \phi\big(W_i[x_a; e_{ab}]\big), \qquad h_{ab}^{(\tau)} = \phi\Big(W_i[x_a; e_{ab}] + W_h \sum_{c \in \mathcal{N}(a)\setminus b} h_{ca}^{(\tau-1)}\Big), \tag{1}$$

for $\tau = 1, \ldots, T-1$, where ϕ is ReLU. We add K virtual readout nodes $v_1, \ldots, v_K$. Every atom a is connected to every v_ℓ by a one-way edge $a \to v_\ell$ carrying a learned feature vector ϵ_ℓ:

$$g_{a\ell}^{(0)} = \phi\big(W_i[x_a; \epsilon_\ell]\big), \qquad g_{a\ell}^{(\tau)} = \phi\Big(W_i[x_a; \epsilon_\ell] + W_h \sum_{c \in \mathcal{N}(a)} h_{ca}^{(\tau-1)}\Big). \tag{2}$$

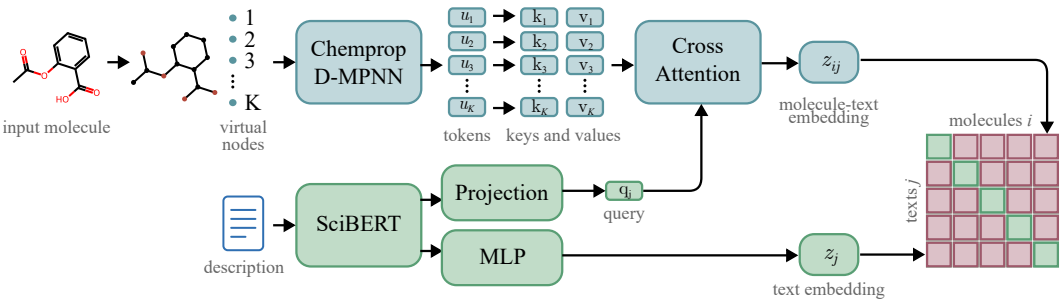


Figure 3: Architecture of MOLTOKSYN. A Chemprop graph encoder produces multiple learned molecular readouts, while the SciBERT representation conditions their cross-attention before the final molecule–text comparison.

Virtual nodes send no messages back, so the states h_{ab} of the real graph are exactly those of the unmodified D-MPNN. Each virtual node is read out with the pretrained output layer and a learned node feature ξ_ℓ, and then mapped by a slot-specific projection:

$$r_{i\ell} = \phi\Big(W_o\big[\xi_\ell; \textstyle\sum_{a\in m_i} g_{a\ell}^{(T-1)}\big]\Big), \qquad u_{i\ell} = \mathrm{LN}\big(W_\ell^u r_{i\ell} + b_\ell^u\big) \in \mathbb{R}^{d_u}, \tag{3}$$

where LN is a LayerNorm (Ba et al., 2016) shared across slots. All slots receive the same atom messages. They differ only through ϵ_ℓ, ξ_ℓ and W_ℓ^u, all learned end-to-end without predefined chemical meaning. The standard Chemprop molecule-level readout is not used.

Text branch. The text branch is SciBERT (Beltagy et al., 2019) with inputs truncated to 128 tokens. Contextual states are mean-pooled over all non-padding positions to give $e_j \in \mathbb{R}^{768}$. Two heads map e_j to a scoring vector $z_j = \mathrm{MLP}(e_j) \in \mathbb{R}^d$ (two layers, ReLU) and to an attention query $q_j = \mathrm{LN}(W_Q e_j + b_Q) \in \mathbb{R}^d$.

4.2 TEXT-CONDITIONED MOLECULAR SCORING

Each token is mapped to a key and a value by slot-specific projections, $\mathbf{k}_{i\ell} = \mathrm{LN}(W_\ell^K u_{i\ell})$ and $\mathbf{v}_{i\ell} = \mathrm{LN}(W_\ell^V u_{i\ell})$. Queries, keys and values are split into H heads of size d/H (superscript (h)). For a pair (m_i, t_j) each head attends over the K tokens:

$$\alpha_{ij\ell}^{(h)} = \frac{\exp\big(\langle q_j^{(h)}, \mathbf{k}_{i\ell}^{(h)}\rangle/\gamma\big)}{\sum_{\ell'=1}^{K} \exp\big(\langle q_j^{(h)}, \mathbf{k}_{i\ell'}^{(h)}\rangle/\gamma\big)}, \tag{4}$$

$$z_{ij} = W_O\Big[\textstyle\sum_\ell \alpha_{ij\ell}^{(1)}\mathbf{v}_{i\ell}^{(1)};\ \ldots;\ \sum_\ell \alpha_{ij\ell}^{(H)}\mathbf{v}_{i\ell}^{(H)}\Big] \in \mathbb{R}^d, \tag{5}$$

where γ is a fixed softmax temperature. The score is the cosine similarity

$$s_\theta(m_i, t_j) = s_{ij} = \big\langle \hat{z}_j, \hat{z}_{ij}\big\rangle, \qquad \hat{z} = z/\|z\|_2. \tag{6}$$

Unlike a dual encoder, z_{ij} depends on both inputs, so the same molecule can be represented differently for descriptions that emphasize different properties. The D-MPNN, the tokens $u_{i\ell}$, and the keys and values are computed once per molecule. q_j and z_j are computed once per text. Only Eqs. 4-5 are pair-specific. When a single text-independent molecule vector is needed, we use uniform attention, $\bar{z}_i = W_O\big[\frac{1}{K}\sum_\ell \mathbf{v}_{i\ell}^{(1)}; \ldots\big]$.

4.3 PAIRWISE CONTRASTIVE OBJECTIVE

For a batch of B matched pairs we score all B^2 molecule–text pairs. Following SigLIP (Zhai et al., 2023) and Llip, each score becomes a logit $o_{ij} = e^\lambda s_{ij} + \beta$ with learned log-scale λ and bias β. Let $y_{ij} = +1$ if $i = j$ and -1 otherwise; off-diagonal pairs are treated as negatives without de-duplication. The loss is

$$\mathcal{L}_{\text{sig}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} \log \sigma\big(y_{ij}\, o_{ij}\big). \tag{7}$$

Table 1: Cross-modal retrieval results (↑) on the STM full test set. Mean ± std over three scaffold splits. Best per column in bold, second best underlined.

	Molecule → Text				Text → Molecule			
Model	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
3D-MoLM	41.69 ±0.64	61.43 ±0.75	67.63 ±0.90	50.85 ±0.24	40.10 ±0.51	61.09 ±0.70	67.78 ±1.23	49.70 ±0.26
MolCA	<u>50.54 ±1.11</u>	<u>67.73 ±1.76</u>	**72.80 ±1.61**	<u>58.47 ±1.32</u>	<u>50.01 ±1.16</u>	**68.54 ±1.82**	**73.72 ±1.93**	<u>58.46 ±1.49</u>
MoleculeSTM	44.49 ±1.88	62.05 ±1.99	67.92 ±2.13	52.69 ±1.74	44.94 ±1.76	63.05 ±2.35	68.88 ±2.37	53.37 ±1.74
MoMu	50.11 ±1.31	65.78 ±1.94	70.53 ±1.87	57.40 ±1.58	49.50 ±1.09	65.98 ±2.02	70.91 ±1.93	57.15 ±1.43
MolFM	47.59 ±0.73	66.33 ±1.66	71.51 ±1.44	56.08 ±1.06	47.97 ±0.49	66.53 ±1.62	71.55 ±1.52	56.40 ±0.88
MOLTOKSYN (Ours)	**53.36 ±1.29**	**68.44 ±1.27**	<u>72.66 ±1.26</u>	**60.27 ±1.25**	**52.74 ±1.10**	<u>67.94 ±1.38</u>	<u>72.31 ±1.12</u>	**59.75 ±1.17**

Table 2: Cross-modal retrieval results (↑) on the MOLSYNTEXT full test set. Mean ± std over three scaffold splits. Best per column in bold, second best underlined.

	Molecule → Text				Text → Molecule			
Model	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
3D-MoLM	62.69 ±3.68	87.92 ±2.56	93.27 ±1.50	73.81 ±3.17	63.22 ±3.72	88.39 ±3.00	93.48 ±1.71	74.27 ±3.29
MolCA	<u>67.65 ±2.29</u>	<u>90.68 ±1.84</u>	<u>94.86 ±1.13</u>	<u>77.82 ±1.98</u>	<u>69.71 ±2.13</u>	<u>91.80 ±1.68</u>	<u>95.46 ±0.91</u>	<u>79.41 ±1.89</u>
MoleculeSTM	62.20 ±3.06	87.11 ±2.50	92.53 ±1.58	73.20 ±2.72	64.77 ±3.13	88.42 ±2.41	93.14 ±1.64	75.17 ±2.70
MoMu	66.34 ±1.97	89.26 ±1.76	93.76 ±0.88	76.42 ±1.80	67.09 ±1.92	89.65 ±1.67	93.77 ±0.92	76.99 ±1.75
MolFM	62.79 ±1.48	88.20 ±1.28	93.48 ±0.89	73.90 ±1.32	64.80 ±1.78	89.02 ±1.28	93.62 ±0.74	75.39 ±1.52
MOLTOKSYN (Ours)	**78.76 ±1.61**	**96.45 ±0.63**	**98.56 ±0.22**	**86.57 ±1.20**	**79.02 ±1.63**	**96.01 ±0.59**	**98.08 ±0.16**	**86.58 ±1.20**

The bias β offsets the imbalance between negatives and positives. At inference, s_{ij} alone ranks candidates in either direction.

5 EXPERIMENTS

We compare MOLTOKSYN on three complementary and diverse benchmarks. Cross-modal retrieval (Section 5.1) measures retrieval quality against other contrastive approaches. Linear Probing (Section 5.2) evaluates the transferability of the learned contrastive representation to the supervised property-prediction tasks. Finally, a controlled Semantic Benchmark (Section 5.3) tests whether selected chemical concepts are captured by the learned molecule–text alignment. Details about optimization, selected hyperparameters and implementation are provided in Appendix C.

5.1 CROSS-MODAL RETRIEVAL

Experimental setup. We compare MOLTOKSYN against five molecule–text retrieval baselines: MoMu (Su et al., 2022), MolFM (Luo et al., 2023), MoleculeSTM (Liu et al., 2023a), 3D-MoLM (Li et al., 2024) and MolCA (Liu et al., 2023b). All baselines are evaluated using the same train/validation/test splits. We report the mean and standard deviation over the three runs.

To separate the effect of textual supervision from that of the model architecture, we benchmark MOLTOKSYN and all baselines on two textual views defined over the same set of molecules: MOLSYNTEXT and STM. We evaluate retrieval quality using standard retrieval ranking metrics: *Recall at k* (R@k) and *Mean reciprocal rank* (MRR). We report molecule→text and text→molecule separately. R@k only checks whether the correct item is in the top k, while MRR also rewards ranking it higher. Higher values indicate better retrieval. Details about architecture, training, hyperparameters and metrics are given in Appendix C.1.

STM evaluation. Table 1 compares MOLTOKSYN against the baselines on the STM dataset. In the molecule→text direction, MOLTOKSYN achieves the best results on three of the four metrics: 53.36% R@1, 68.44% R@5, and 60.27 MRR, while its R@10 (72.66%) is marginally below MolCA (72.80%). It outperforms the strongest baseline, MolCA, by $+2.82$ pp on R@1 and $+1.80$ on MRR. In the text→molecule direction, MOLTOKSYN leads on R@1 (52.74%, $+2.73$ pp over MolCA) and MRR (59.75, $+1.29$), while MolCA is slightly ahead on R@5 and R@10. All three differences favoring MolCA lie within one standard deviation. Overall, MOLTOKSYN achieves the best result on five of the eight metrics, with the clearest gains on R@1 and MRR.

MOLSYNTEXT evaluation. Results on the MOLSYNTEXT dataset are reported in Table 2. MOLTOKSYN achieves the best results on all eight metrics, with a substantial margin over every baseline. In the molecule→text direction, MOLTOKSYN reaches 78.76% R@1 and 86.57 MRR, outperforming the strongest baseline, MolCA, by +11.11 pp on R@1 and +8.75 on MRR. In the text→molecule direction, the gap is similarly large: 79.02% R@1 (+9.31 pp over MolCA) and 86.58 MRR (+7.17). The margins on MOLSYNTEXT are considerably larger than on STM.

5.2 LINEAR PROBING

Table 3: Linear-probe (AUROC ↑) of frozen molecule representations on OGBG classification tasks Hu et al. (2020). All molecule–text models, baselines included, are pretrained on MOLSYNTEXT. Mean ± std over three checkpoints, each pretrained on a different scaffold split. Random is our architecture randomly initialized without pretraining. Best per column in bold, second best underlined.

Representation	HIV	ToxCast	Tox21	BBBP	BACE	ClinTox	SIDER
ECFP	75.10	61.64	73.77	65.93	**80.58**	81.95	60.58
MoleculeSTM	74.27±1.16	63.40±0.21	72.28±0.48	64.62±1.23	77.72±0.30	82.22±0.62	60.40±0.58
MoMu	70.47±2.21	60.29±0.16	70.42±0.57	61.45±0.79	74.75±0.70	66.34±0.77	60.45±1.29
MolCA	74.03±1.46	62.25±0.74	71.57±0.32	64.07±0.36	75.56±1.18	64.46±6.17	60.26±0.74
3D-MoLM	74.80±0.73	**64.99±0.32**	74.30±0.27	66.40±1.58	76.87±2.04	68.55±2.27	**62.10±0.33**
MolFM	71.61±0.71	62.92±0.53	71.59±0.30	59.17±2.56	76.03±0.16	60.18±1.37	59.75±0.28
MOLTOKSYN (Ours)	**76.78±0.77**	64.81±0.52	**74.51±0.06**	**66.63±1.25**	77.78±1.25	**86.59±2.22**	61.22±0.18
Random	72.37±2.19	60.21±0.36	71.07±0.92	64.47±0.42	70.85±4.97	76.18±4.57	59.42±1.06

To evaluate the transferability of the learned representations beyond retrieval, we compare the contrastive methods in terms of their ability to generalize to downstream tasks.

Experimental setup. We evaluate frozen molecular representations on standard molecular property prediction benchmarks from the OGB (Hu et al., 2020; Wu et al., 2018), comprising seven classification tasks (BBBP, BACE, ClinTox, Tox21, ToxCast, SIDER, and HIV) and three regression tasks (ESOL, FreeSolv, and Lipophilicity). We use the scaffold split provided by OGB Hu et al. (2020). To ensure a fair comparison, we evaluate every frozen representation using an identical linear-probing setup and the same hyperparameters. We compare MOLTOKSYN with the baselines introduced in Section 5.1 and additionally include ECFP fingerprints as a non-learned baseline. We also consider *Random*, which denotes representations obtained from a randomly initialized MOLTOKSYN model. For each pretrained model, we probe the three checkpoints from Section 5.1 and report the mean and standard deviation. Classification tasks are evaluated with the area under the ROC curve (*AUROC*, higher is better), and regression tasks with the root mean square error (*RMSE*, lower is better). We report results for models pretrained on both the MOLSYNTEXT and STM datasets. Further details on the baselines, datasets, evaluation protocol and extended results are provided in Appendix C.2.

Table 4: Linear-probe (RMSE ↓) of frozen molecule representations on OGBG regression tasks Hu et al. (2020). All molecule–text models, baselines included, are pretrained on MOLSYNTEXT. Mean ± std over three checkpoints, each pretrained on a different scaffold split. Random is our architecture randomly initialized without pretraining. Best per column in bold, second best underlined.

Representation	ESOL	FreeSolv	Lipophilicity
ECFP	1.582	4.013	0.875
MoleculeSTM	1.219±0.020	4.210±0.149	0.853±0.013
MoMu	1.357±0.088	3.402±0.321	0.916±0.020
MolCA	1.224±0.072	2.736±0.187	0.899±0.014
3D-MoLM	1.143±0.034	3.067±0.356	0.893±0.015
MolFM	1.419±0.021	3.232±0.235	**0.819±0.007**
MOLTOKSYN	**0.940±0.038**	**2.611±0.105**	0.879±0.013
Random	1.485±0.100	3.214±0.200	0.972±0.024

Results. MOLTOKSYN shows strong probing performance on both classification and regression tasks. When pretrained on MOLSYNTEXT, MOLTOKSYN ranks first or second on all seven classification tasks, with the best result on four of them (HIV, Tox21, BBBP, and ClinTox; Table 3). The largest gap is observed on ClinTox, where MOLTOKSYN reaches an AUROC of 86.59, compared

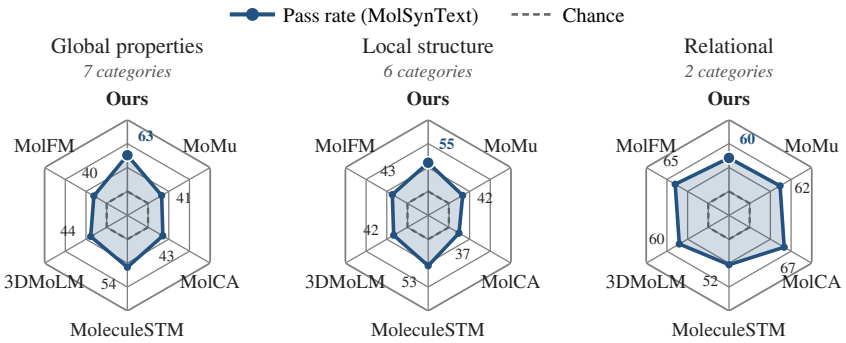


Figure 4: Controlled Semantic Benchmark aggregated by concept family (MOLSYNTEXT). Each radar summarizes one family (global properties, local structure, and relational tasks). Values denote the pass rate (%), first averaged over the categories within the family and then over three models trained on different scaffold splits. The dashed hexagon marks the chance level of 25%.

with 82.22 for the strongest competing baseline, MoleculeSTM. On the regression benchmarks, MOLTOKSYN achieves the lowest RMSE on both ESOL and FreeSolv (Table 4), improving over the next-best model by 0.203 and 0.125, respectively. On Lipophilicity, however, MolFM obtains the best result, and MOLTOKSYN is also slightly behind MoleculeSTM and ECFP. Results for STM dataset are reported in Appendix C.2.

Effect of textual supervision. Because MOLSYNTEXT and STM share the same molecules, splits, and architectures, and linear probing is evaluated on identical downstream tasks, the differences between the two pretraining views are independent of these factors and can be attributed to the textual supervision alone. Table 5 shows that pretraining on MOLSYNTEXT improves probing performance in 50 of the 60 model–task pairs (35 of 42 for classification and 15 of 18 for regression). The effect is largely consistent across architectures. These results indicate that the benefit of structure-grounded synthetic descriptions is not specific to MOLTOKSYN but carries over to a range of molecule–text architectures.

Table 5: Linear-probing gains from pretraining on MOLSYNTEXT instead of STM. Positive values indicate that MOLSYNTEXT performs better (higher AUROC, lower RMSE).

	Classification (AUROC ↑)							Regression (RMSE ↓)		
Representation	HIV	ToxCast	Tox21	BBBP	BACE	ClinTox	SIDER	ESOL	FreeSolv	Lipophilicity
MoleculeSTM	+3.81	+1.53	+0.97	+2.70	+4.63	+1.12	+1.11	+0.236	+0.380	+0.049
MoMu	+0.44	−0.29	+0.12	+2.42	+6.95	+0.29	+1.26	+0.241	+0.195	+0.042
MolCA	+3.69	+0.92	+1.96	−2.57	+2.32	+3.07	+1.39	+0.237	+0.512	+0.043
3D-MoLM	+0.84	−0.11	−0.16	+1.52	+4.77	−6.17	+1.57	−0.006	−0.061	+0.024
MolFM	+3.08	+2.69	+0.13	−1.75	+2.70	+8.06	+1.21	−0.005	+0.757	+0.048
MOLTOKSYN (Ours)	+2.75	+1.83	+1.60	+2.52	+4.67	+4.30	−0.48	+0.261	+0.507	+0.041

5.3 CONTROLLED SEMANTIC BENCHMARK

The benchmarks described above measure overall performance: cross-modal retrieval evaluates how well a model identifies the correct item in a large candidate pool, and linear probing evaluates how well the learned molecular representations transfer to property prediction. Both are useful aggregate measures, but neither shows which chemical concepts a model has learned. We therefore design a benchmark that tests whether the trained models can solve small, interpretable molecule–text tasks by selecting the correct answer among a few candidates.

Benchmark construction and Setup. The benchmark consists of 15 semantic categories with 10 test cases each, for a total of 150 test cases. The categories range from global physicochemical

properties, such as size and lipophilicity, through local structure, such as functional groups and ring topology, to relational tasks, such as analog edits and bioisosteric replacement. Further details of each category we provided in Appendix C.3 in Table 10. Each test case has a query q and four candidate answers $A = \{a_1, a_2, a_3, a_4\}$, exactly one of which is correct. The model computes a similarity score $s(q, a_j)$ for each candidate and selects the one with the highest score:

$$\hat{a} = \arg\max_{a_j \in A} s(q, a_j). \tag{8}$$

A test case is counted as correct if $\hat{a}$ is the annotated correct answer. Example queries for each category are given in Appendix C.3. We report the pass rate (fraction of cases in which $\hat{a}$ is correct; chance level 25%), and the *mean reciprocal rank* (MRR) of the correct answer. Results are averaged over the three models trained on different scaffold splits from cross-modal retrieval benchmark (Section 5.1).

Concept families. To summarize the results, we group the categories into three concept families. *Global properties* (7 categories) cover whole-molecule physicochemical descriptors: molecular size, hydrogen-bond donors and acceptors, lipophilicity, polarity (TPSA), flexibility, charge and ionization, and interaction profile. *Local structure* (6 categories) targets specific substructures and their arrangement: rings and aromaticity, functional groups, tautomers and protonation states, exact atom and group counts, ring topology, and regiochemistry. *Relational* tasks (2 categories) require comparing a molecule with a modified version of itself: analog modification and bioisosteric replacement.

Table 6: Semantic Benchmark gains of MOLSYNTEXT over STM (pass rate). Positive values indicate that MOLSYNTEXT performs better.

Model	Global	Local	Relational	Overall
MoleculeSTM	+15.2	+3.9	−20.0	+6.0
MoMu	+8.1	+0.6	−1.7	+3.8
MolCA	+4.3	−5.6	+8.3	+0.9
3D-MoLM	+8.6	+6.1	−1.7	+6.2
MolFM	+6.2	+7.8	+6.7	+6.9
MOLTOKSYN	+24.8	+5.6	−5.0	+13.1

Results. Figure 4 shows the results for models trained on MOLSYNTEXT and evaluated on Semantic Benchmark. MOLTOKSYN achieves the highest overall pass rate ($59.3 \pm 2.9\%$), ahead of MoleculeSTM ($53.3 \pm 4.7\%$), while the remaining baselines stay between 43.8% and 45.3%. The largest margin is visible on *Global properties* (62.9% vs. 54.3% for MoleculeSTM). MOLTOKSYN is also best on *Local structure* (55.0% vs. 52.8%). The *Relational* family is the only one where MOLTOKSYN (60.0%) is outperformed, by MolCA (66.7%), MolFM (65.0%) and MoMu (61.7%). Extended results in Appendix C.3 include the evaluation of models trained on STM and separate results for each of the 15 categories.

Effect of textual supervision. Following the same controlled comparison as in Section 5.2, Table 6 reports the gains from pretraining on MOLSYNTEXT instead of STM. Pretraining on MOLSYNTEXT increases the overall pass rate for all six models. The improvement is most pronounced for *Global properties*, where every model benefits, and it extends to *Local structure* for five of the six models. Overall, structure-grounded supervision improves the recognition of specific chemical concepts across architectures, and MOLTOKSYN benefits from it the most.

6 CONCLUSION

We presented a unified approach to two coupled aspects of molecule–text retrieval. MOLSYNTEXT provides controlled synthetic language derived from structured molecular information, while the matched STM view retains PubChem-derived text over the same molecular population. MOLTOKSYN complements this supervision with learned molecular tokens whose aggregation is conditioned on the text query, producing pair-specific compatibility scores rather than one fixed representation per molecule. Our experiments show that MOLTOKSYN outperforms existing contrastive molecule–text models in cross-modal retrieval on both textual views, with particularly large gains on MOLSYNTEXT. Its frozen representations transfer well to molecular property prediction, and the Controlled Semantic Benchmark shows that MOLTOKSYN captures global physicochemical and local structural concepts more reliably than the baselines.

AI USE STATEMENT

We used generative AI to generate and filter synthetic data, clean and filter data, assist with implementation source code, find relevant literature, and rewrite the manuscript to improve clarity. In particular, Qwen3.8-27B was used to generate the MOLSYNTEXT descriptions from structured molecular cards and filter them using prompts and criteria we designed (Section 3.2). We did not use generative AI to design experiments, develop theory or hypotheses, formulate mathematical claims, analyze data, or interpret results. We have reviewed AI-assisted work and take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work uses public molecular structures and database-derived or synthetically generated text and does not involve human participants or personal data. Molecule retrieval systems can nevertheless have dual-use applications. The model produces similarity scores over candidate molecules and should not be treated as experimental evidence of safety, efficacy, biological activity, or suitability for real-world use. Any released dataset and model documentation will state these limitations and the provenance of the textual supervision.

REPRODUCIBILITY STATEMENT

Appendix A reports the software and computational resources, Appendix B documents the dataset construction and generation configuration, and Appendix C lists the model details and hyperparameters used in all experiments. We also provide the code as a zipped repository in the supplementary material to support reproduction of our results.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pp. 3615–3620, 2019.

Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H. Green. Deep learning foundation models for low-data regimes from classical molecular descriptors. *Journal of Chemical Information and Modeling*, 66(17): 10692–10701, 09 2026. ISSN 1549-9596. doi: 10.1021/acs.jcim.6c01546. URL https://doi.org/10.1021/acs.jcim.6c01546.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of EMNLP*, pp. 595–607, 2021.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1): 9–17, 2024.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxgnnNFvH.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, et al. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. 235:26070–26084, 21–27 Jul 2024.

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 *International Conference on Learning Representations*, volume 2024, pp. 17352–17371, 2024.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46(1-3):3–26, 2001.

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.

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 EMNLP*, pp. 15623–15638, 2023b. doi: 10.18653/v1/2023.emnlp-main.966.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. Improving biomedical information retrieval with neural retrievers. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11038–11046, 2022.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

RDKit Developers. RDKit: Open-source cheminformatics software. `https://www.rdkit.org/`, 2026.

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.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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 *International Conference on Learning Representations*, 2025.

A Implementation and Computational Resources

Software. Source code was implemented in Python 3.12 using PyTorch 2.6.0 with CUDA 12.4. Molecules were processed with RDKit 2026.03.

Hardware and compute. All main experiments were conducted on a single gpu NVIDIA H100 80 GB using up to 32.0 GB of system RAM. Training for 100 epochs took approximately 8 hours per run.

B Dataset Construction

STM reconstruction. The STM view follows the PubChemSTM data source introduced by MoleculeSTM (Liu et al., 2023a). We use the released CID–SMILES mapping as the molecular reference, retrieve corresponding PubChem text with our parser, perform light normalization, and retain all available descriptions for each molecule in the common corpus (154,253 descriptions, 1.23 per molecule on average).

Structured molecular cards. Each molecule is parsed with RDKit and converted to a compact structured record containing deterministic molecular information. The card includes standard physicochemical descriptors, ring and atom statistics, detected functional groups, and feature-tag counts used by the generation prompt. The generator receives this card rather than an unconstrained request to infer arbitrary properties from a molecular string.

Generation configuration. Descriptions in MolSynText are generated offline with `Qwen3.8-27B` (Yang et al., 2025). The language model is used as a verbalizer of structured molecular information rather than as a source of new chemical facts. Each prompt combines a perspective template, optional chemistry-specific instructions, fixed grounding constraints, and the serialized molecular card. The templates vary the emphasis across physicochemical properties, structural motifs, functional groups, and molecular topology. Additional chemistry instructions are included with probability 0.7, with up to five sampled from a predefined pool. Generation initially produces one candidate description per molecule, temperature 0.7, and the medium reasoning setting. A separate `Qwen3.8-27B` judge compares each candidate description with its source molecular card and checks for unsupported claims or incorrectly copied information. Candidates that fail these checks are rejected.

Generator prompt. The exact wording varies with the selected perspective template, but all generation prompts follow the same structure:

```
[TEMPLATE PREFIX]
[OPTIONAL CHEMISTRY-SPECIFIC INSTRUCTIONS]
Hard constraints:
- Description must be in English.
- Keep the description to 2--4 sentences.
- Avoid unsupported claims; state uncertainty when
necessary.
- Use exact structural terminology only when supported by
the SMILES.
- Treat feature-tag and RDKit functional-group counts as
structural hints; overlapping annotations need not represent
distinct groups.
- Do not infer detailed topology or chemical states from
aggregate counts alone.
- Treat the IUPAC name only as a supporting reference;
SMILES remains the primary source of molecular connectivity.
- Return only valid JSON matching {"description": "..."}.
Molecule: [MOLECULE_CARD_JSON]
```

For example, one of the perspective templates begins:

Table 7: Representative generated descriptions rejected by the quality-control judge. Excerpts are shortened to highlight the claim responsible for rejection.

Error type	Generated description (excerpt)	Why rejected
Unsupported inference	"This stereochemically dense fused tricyclic scaffold [...] presents a modest hydrogen-bonding surface [...] that one might expect to confer appreciable water solubility."	Water solubility was inferred without direct support and was inconsistent with the high logP (≈ 4.1).
Substitution pattern	"The molecule is a symmetric, linear scaffold built around a single para-substituted benzene ring [...]"	The molecular structure contains an ortho-disubstituted benzene ring and does not support the claimed symmetry.
Conjugation	"Methyl octadeca-9,12-dienoate is a linear 18-carbon fatty acid methyl ester whose only unsaturation is a conjugated diene at positions 9 and 12."	The two double bonds are separated by a methylene group and therefore form an isolated rather than conjugated diene.
Scaffold connectivity	"The molecule is organized around a phenyl–thiazole biaryl scaffold bridged by a planar amide linker [...]"	The phenyl and thiazole rings are connected through an amide linker rather than a direct aryl–aryl bond, so describing the scaffold as biaryl is incorrect.

```
You are given structured molecule data.  Write a compact
English description in 2--4 sentences.  Do not merely
restate fields.  Select the most informative and distinctive
signals and turn them into a coherent interpretation of
the molecule.  Prioritize supported information about
physicochemical properties, hydrogen bonding, aromaticity,
flexibility, charge, lipophilicity, and structural features.
```

Judge-based quality control. Before judging, an output must be valid JSON with a non-empty description and satisfy the basic length constraints. We then use a separate `Qwen3.8-27B` instance as an accept/reject judge, with temperature 0 and reasoning disabled. The judge receives both the molecular card and the candidate description and scores faithfulness, interestingness, clarity, hallucination risk, and the number of unsupported critical claims.

A candidate is retained only when the judge returns `accepted=true`, faithfulness ≥ 4, hallucination risk ≤ 1, and zero unsupported critical claims. Otherwise, the candidate is logged as rejected and a fresh generation attempt is made.

The judge is instructed as follows:

```
You are a strict but fair medicinal-chemistry judge for
one generated molecule description.  The molecule JSON is
the source of truth.  Decide whether the candidate is good
enough to keep.
Reject serious problems such as invented atoms, rings,
functional groups, charge states, scaffolds, or descriptor
values; claims that contradict the supplied molecular
information; unsupported biological or physicochemical
claims; or invalid, unclear, or excessively long
descriptions.
Do not reject cautious, useful interpretation when it is
supported by the molecular information.
Score faithfulness, interestingness, clarity, and
hallucination risk from 0 to 5, where 0 hallucination risk
means no risk and 5 means severe hallucination.  Count
serious unsupported claims separately.
Return only valid JSON containing:  accepted,
faithfulness, interestingness, clarity, hallucination_risk,
unsupported_critical_claims, issues, and rationale_short.
MOLECULE_CARD_JSON: [MOLECULE_CARD_JSON]
CANDIDATE_JSON: [CANDIDATE_JSON]
```

Representative rejected descriptions. Table 7 shows representative errors caught by the judge. The examples are intended to illustrate different failure modes rather than estimate their frequency in the full generation process.

Description post-processing. Generated descriptions are stored as a list associated with each molecular record. To respect the 128-token input limit of `allenai/scibert_scivocab_uncased` without truncating content, we tokenize each description with the SciBERT tokenizer, including special tokens, and split descriptions that exceed the limit. Three-sentence descriptions are converted into two overlapping segments containing sentences 1–2 and 2–3, while four-sentence descriptions are split into sentences 1–2 and 3–4. Rare longer cases are divided at sentence boundaries, or at word boundaries when a single sentence exceeds the limit. This produces 216,015 descriptions for 125,000 molecules, or 1.728 descriptions per molecule on average, with every stored description containing at most 128 tokens.

Scaffold splits. We construct three scaffold splits on the final set of 125,000 molecules using Bemis–Murcko scaffolds computed from canonical RDKit SMILES without chirality. Molecules sharing the same scaffold are kept together in a single partition, ensuring zero scaffold overlap between training, validation, and test sets. Since all acyclic molecules receive an empty Murcko scaffold, the corresponding group of 16,837 molecules is always assigned to training to avoid dominating a held-out split. The remaining scaffold groups are processed from largest to smallest and assigned with seeded weighted sampling toward an 80/10/10 split, subject to a small capacity tolerance. We generate three independent splits using seeds 0, 1, and 2, with final sizes of 100,100/12,300/12,600 for seeds 0 and 1 and 100,100/12,600/12,300 for seed 2. We verify 125,000 unique canonical SMILES and zero scaffold overlap for every split.

C Experimental Details

C.1 Cross-Modal Retrieval

Dataset. The MolSynText and STM textual views cover the same $125{,}000$ molecules. We build three scaffold splits with an 80/10/10 ratio. This gives $100{,}100$ training molecules and $12{,}300$–$12{,}600$ validation and test molecules per split. Because the splits are defined over molecules, STM and MolSynText use identical partitions for each split. The same dataset splits are used to train MolTokSyn and all baselines. Since some molecules have multiple descriptions, we sample one description per molecule at each training epoch. For evaluation, we use only the first description of each molecule, so every test query has exactly one correct candidate.

Metrics. To evaluate retrieval quality, we use *Recall at* k and *Mean reciprocal rank* (MRR). Let $\mathcal{Q}$ denote the set of all queries and let r_q be the rank of the correct candidate for query $q \in \mathcal{Q}$ among all candidates in the test set. *Recall at* k (R@k, for $k \in \{1, 5, 10\}$) measures the fraction of queries for which the correct item appears among the top-k retrieved candidates:

$$\mathrm{R@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}\left[r_q \leq k\right], \tag{9}$$

where $\mathbb{1}[\cdot]$ is the indicator function. *Mean reciprocal rank* (MRR) averages the inverse rank of the correct candidate over all queries:

$$\mathrm{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{r_q}. \tag{10}$$

Training protocol. The text encoder is pretrained SciBERT (Beltagy et al., 2019). The molecule encoder is a Chemprop network with hidden size 2048 and depth 6, initialized from CheMeleon (Yang et al., 2019; Heid et al., 2024; Burns et al., 2026). Both encoders are fine-tuned. MolTokSyn uses 8 virtual molecule tokens of dimension 128, a single slot layer with slot self-attention, 8 attention heads, and a joint embedding dimension of 128. The model is trained with a SigLIP-style (Zhai et al., 2023) sigmoid pairwise loss, with a learned logit scale and bias initialized to 14.0 and -10.0, respectively.

We optimize with Adam (Kingma & Ba, 2014) (β_1=0.9, β_2=0.999) at a learning rate of 10^{-4}, using a linear warm-up over the first 6% of steps followed by linear decay to zero. Training runs for 100 epochs with a batch size of 512. Key design choices were selected based on the ablation studies in Appendix D. All baselines were retrained on our splits using public code and the training configurations reported in original publications.

Evaluation protocol. Retrieval is scored against the entire test partition (12,300 or 12,600 candidates, depending on the split). For MOLTOKSYN, we evaluate on each split the checkpoint with the highest validation R@1, averaged over both retrieval directions. Reported values are the mean and standard deviation over the three runs.

C.2 LINEAR PROBING

Datasets. We use ten MoleculeNet (Wu et al., 2018) benchmarks in the versions distributed by OGB (Hu et al., 2020): seven classification tasks (HIV, ToxCast, Tox21, BBBP, BACE, ClinTox and SIDER) and three regression tasks (ESOL, FreeSolv and Lipophilicity). We use the OGB scaffold splits.

Representations. For all contrastive models we use the retrieval checkpoints from Section 5.1. For each model, we extract the molecule vector that its retrieval code scores against the text embedding. For MOLTOKSYN, we use the 128-dimensional output of the contrastive projection head. ECFP fingerprints are computed with the RDKit Morgan fingerprint generator (radius 2, 2048 bits, binary). *Random* is the MOLTOKSYN architecture with random weights. All models are frozen, and every representation, including ECFP, is ℓ_2-normalized before probing.

Probe and metrics. The probe is a single linear layer. We use binary cross-entropy for classification and squared error for regression. We optimize parameters using L-BFGS. Classification tasks are evaluated with AUROC, regression with RMSE. For each contrastive model, we probe the three checkpoints pretrained on the three scaffold splits and report the mean and standard deviation across them. For Random, the spread is taken over the three initialization seeds. ECFP is deterministic and is reported as a single value.

Extended results. The representations pretrained on STM show a weaker pattern than observed for MOLSYNTEXT dataset (reported in Table 3). Among the learned representations, MOLTOKSYN achieves the best result on three of the seven classification tasks (HIV, ClinTox, and SIDER) and the second best on two more (ToxCast and Tox21), with 3D-MoLM as the strongest competitor (Table 8). In this setting, ECFP fingerprints remain highly competitive and outperform all learned representations on HIV and BACE. On the regression benchmarks, the pattern also differs from MOLSYNTEXT: 3D-MoLM performs best on ESOL and FreeSolv, with MOLTOKSYN second on both, while MolFM again performs best on Lipophilicity (Table 9). Overall, the MOLSYNTEXT-pretrained representation of MOLTOKSYN outperforms its STM-pretrained counterpart on nine of the ten benchmarks, with SIDER as the only exception. These results indicate that the representations learned through contrastive pretraining transfer beyond cross-modal retrieval, although the magnitude of the improvement depends on both the pretraining dataset and the downstream task.

Table 8: Linear-probe (AUROC ↑) of frozen molecule representations on OGBG classification tasks Hu et al. (2020). All molecule–text models, baselines included, are pretrained on STM. Mean ± std over three checkpoints, each pretrained on a different scaffold split. Random is our architecture randomly initialized without pretraining. Best per column in bold, second best underlined.

Representation	HIV	ToxCast	Tox21	BBBP	BACE	ClinTox	SIDER
ECFP	**75.10**	61.64	<u>73.77</u>	<u>65.93</u>	**80.58**	<u>81.95</u>	<u>60.58</u>
MoleculeSTM	70.46±0.24	61.87±0.37	71.31±0.29	61.92±2.05	73.09±1.39	81.10±1.12	59.29±0.74
MoMu	70.03±1.53	60.58±0.29	70.30±0.34	59.03±1.27	67.80±3.08	66.05±2.64	59.19±1.65
MolCA	70.34±1.37	61.33±0.47	69.61±0.97	**66.64±1.04**	73.24±1.15	61.39±1.61	58.87±2.09
3D-MoLM	73.96±1.62	**65.10±0.08**	**74.46±0.72**	64.88±0.62	72.10±1.95	74.72±2.13	60.53±0.64
MolFM	68.53±2.70	60.23±0.30	71.46±0.32	60.92±2.74	<u>73.33±2.22</u>	52.12±3.34	58.54±0.49
Ours	<u>74.03±0.10</u>	<u>62.98±0.83</u>	72.91±0.90	64.11±0.48	73.11±1.28	**82.29±2.43**	**61.70±0.54**
Random	72.37±2.19	60.21±0.36	71.07±0.92	64.47±0.42	70.85±4.97	76.18±4.57	59.42±1.06

Table 9: Linear-probe (RMSE ↓) of frozen molecule representations on OGBG regression tasks Hu et al. (2020). All molecule–text models, baselines included, are pretrained on STM. Mean ± std over three checkpoints, each pretrained on a different scaffold split. Random is our architecture randomly initialized without pretraining. Best per column in bold, second best underlined.

Representation	ESOL	FreeSolv	Lipophilicity
ECFP	1.582	4.013	<u>0.875</u>
MoleculeSTM	1.455±0.060	4.590±0.717	0.902±0.002
MoMu	1.598±0.028	3.597±0.522	0.958±0.020
MolCA	1.461±0.066	3.248±0.513	0.942±0.009
3D-MoLM	**1.137±0.028**	**3.006±0.140**	0.917±0.013
MolFM	1.414±0.028	3.989±0.276	**0.867±0.008**
MOLTOKSYN (Ours)	<u>1.201±0.083</u>	<u>3.118±0.118</u>	0.920±0.028
Random	1.485±0.100	3.214±0.200	0.972±0.024

Table 10: Semantic categories in the Controlled Semantic Benchmark. Each category contains 10 manually defined test cases.

#	**Category**	**Evaluated capability**
01	Molecule size	Comparing molecular size by atom count and molecular weight
02	H-bond donors & acceptors	Counting and locating hydrogen-bond donors and acceptors
03	Lipophilicity	Matching molecules to hydrophobic, lipophilic, or hydrophilic descriptions
04	Polarity (TPSA)	Identifying high- or low-polarity molecules, polar groups, and heteroatom-rich structures
05	Rings & aromaticity	Distinguishing aromatic, non-aromatic, fused, and polycyclic ring systems
06	Flexibility	Distinguishing acyclic torsional freedom from ring constraint
07	Charge & ionization	Recognizing formal charges, zwitterions, and ionizable groups
08	Functional groups	Recognizing groups such as carboxylic acids, amides, esters, ethers, and isothiocyanates
09	Analog modification	Identifying the analog produced by a specified edit to a fixed scaffold
10	Interaction profile	Matching molecules to interaction types: hydrophobic, electrostatic, hydrogen bonding, and π-stacking
11	Tautomerism & protonation	Distinguishing keto/enol and lactam/lactim tautomers and protonation states
12	Bioisosteric replacement	Recognizing the replacement of a local motif by a bioisostere
13	Exact counts	Determining the exact number of atoms, rings, and groups
14	Ring topology & size	Distinguishing fused, spiro, bridged, and linked ring systems; identifying ring size
15	Regiochemistry	Identifying substitution position on a ring, including relative to heteroatoms

C.3 CONTROLLED SEMANTIC BENCHMARK

Evaluation protocol. All methods are evaluated with the checkpoints trained for the retrieval experiments (Section 5.1) on the MOLSYNTEXT and STM datasets, without further fine-tuning. For each question, the query and the four candidates are embedded with the corresponding encoders, and the candidates are ranked by cosine similarity to the query. We report the *pass rate*, the percentage of questions in which the correct answer is ranked first, and the *mean reciprocal rank* (MRR) of the correct answer. Table 10 lists the 15 semantic categories used in Controlled Semantic Benchmark.

Examples and qualitative results. Figures 5 and 6 show one randomly selected test case per category, together with the answer chosen by MOLTOKSYN. Figures 7–9 report results aggregated by concept family for the models trained on MOLSYNTEXT and STM, and Figures 10–13 report results per category. In both cases, we show pass rate and MRR.

Table 11: Architecture ablation under the same InfoNCE objective. Global denotes a single query-independent molecular representation, while MOLTOKSYN uses the text-conditioned molecular-token readout with $K = 8$. Best values are shown in bold.

		Molecule → Text				Text → Molecule			
Corpus	Architecture	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
STM	Global	43.32±1.92	**67.29±2.11**	**73.16±2.02**	54.41±1.96	43.37±1.85	**67.83±2.18**	**73.47±2.08**	54.60±1.94
	MOLTOKSYN	**47.02±1.71**	66.44±1.64	72.42±1.56	**56.00±1.66**	**45.95±1.59**	66.71±1.75	72.35±1.63	**55.37±1.61**
MOLSYNTEXT	Global	66.02±2.76	93.15±0.98	96.18±0.30	78.07±2.08	68.33±2.65	94.00±1.07	96.64±0.27	79.72±2.03
	MOLTOKSYN	**71.79±2.36**	**94.12±0.74**	**96.94±0.21**	**81.52±1.72**	**73.26±2.27**	**94.23±0.82**	**96.74±0.20**	**82.47±1.65**

D ABLATION STUDIES

We perform four ablation studies to examine the main architectural and optimization choices in MOLTOKSYN. We study the effect of the text-conditioned molecular readout, the number of molecular tokens, the physical contrastive batch size, and the contrastive objective. Within each ablation, all non-investigated components are kept fixed.

D.1 TEXT-CONDITIONED MOLECULAR READOUT

We first isolate the effect of the text-conditioned molecular-token architecture. To this end, we compare MOLTOKSYN with a global-pooling variant that uses the same molecular and text encoders but compresses each molecule into a single, query-independent representation. Both variants are trained with InfoNCE, a physical batch size of 128, an accumulation factor of 2, and the same learning-rate setting. The token-based model uses $K = 8$ molecular tokens.

As shown in Table 11, the text-conditioned readout improves retrieval over the global representation, with the clearest gains on MOLSYNTEXT. There, the token-based architecture improves all eight metrics, with the largest gains in R@1 and MRR. On STM, results are more mixed: the text-conditioned readout improves R@1 and MRR in both retrieval directions, while the global representation remains slightly stronger at R@5 and R@10. We conclude that the text-conditioned readout is beneficial overall and mainly helps place the correct match at the top of the ranking. Based on these results, we use the text-conditioned readout in all main experiments.

D.2 NUMBER OF MOLECULAR TOKENS

We next vary the number of learned molecular tokens $K \in \{1, 4, 8, 16, 32\}$. All variants use the text-conditioned architecture, the SigLIP-style objective, a physical batch size of 128, an accumulation factor of 2, and the same learning-rate setting.

The effect of K differs between the two textual views. On MOLSYNTEXT, larger token sets are increasingly useful, and $K = 32$ obtains the highest central value on all eight retrieval metrics. On STM, performance is largely insensitive to the number of tokens, and the single-token variant obtains the highest central value on all metrics. Since no single value of K is best on both views, we choose $K = 8$ as a trade-off between the two, as it remains close to the best setting on both STM and MOLSYNTEXT. We therefore use $K = 8$ in MOLTOKSYN.

Table 12: Effect of the number of molecular tokens K on STM. All other architectural and optimization choices are fixed. Best values are shown in bold.

	Molecule → Text				Text → Molecule			
K	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
1	**51.54±1.72**	**68.09±1.61**	**72.99±1.54**	**59.18±1.64**	**51.19±1.58**	**68.82±1.63**	**73.40±1.51**	**59.17±1.57**
4	50.76±1.80	67.76±1.73	72.79±1.50	58.66±1.71	50.35±1.50	67.90±1.55	72.60±1.47	58.38±1.55
8	51.15±1.69	67.62±1.58	72.90±1.49	58.89±1.59	50.40±1.52	67.58±1.70	72.48±1.40	58.30±1.50
16	51.05±1.74	67.44±1.45	72.75±1.73	58.78±1.65	49.83±1.66	67.57±1.87	72.50±1.83	57.91±1.70
32	50.63±1.76	67.37±1.56	72.37±1.66	58.45±1.67	49.96±1.63	67.32±1.79	71.99±1.72	57.98±1.66

Table 13: Effect of the number of molecular tokens K on MOLSYNTEXT. All other architectural and optimization choices are fixed. Best values are shown in bold.

	Molecule → Text				Text → Molecule			
K	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
1	75.95±2.31	94.60±0.74	96.85±0.22	84.14±1.72	77.52±2.25	94.78±0.82	96.98±0.21	85.15±1.65
4	76.10±2.69	95.01±0.66	97.17±0.19	84.42±1.88	76.67±2.54	94.71±0.89	96.72±0.27	84.69±1.80
8	76.15±2.24	94.97±0.78	97.06±0.28	84.39±1.69	77.06±2.10	94.75±0.79	96.81±0.19	84.90±1.58
16	77.45±2.13	95.50±0.69	97.34±0.13	85.35±1.55	77.68±2.10	95.10±0.69	97.01±0.11	85.38±1.49
32	**77.74±2.18**	**95.83±0.72**	**97.69±0.16**	**85.78±1.59**	**78.04±2.14**	**95.40±0.74**	**97.25±0.14**	**85.77±1.54**

D.3 PHYSICAL CONTRASTIVE BATCH SIZE

For the contrastive objective, gradient accumulation and physical batch size are not equivalent. Gradient accumulation determines how many examples contribute to one optimizer update, whereas the physical batch size determines how many molecule–text pairs, and thus in-batch negatives, are present in a single evaluation of the contrastive loss. We therefore vary the physical batch size over $B \in \{32, 64, 128\}$ while keeping the effective batch size fixed at 256 through gradient accumulation.

Table 14: Effect of physical contrastive batch size at fixed effective batch size 256. All runs use $K = 8$, the SigLIP-style objective, and the same learning-rate setting. Best values are shown in bold.

		Molecule → Text				Text → Molecule			
Corpus	Physical B	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
STM	32	47.35±1.79	66.09±1.70	71.59±1.60	55.95±1.71	46.50±1.65	65.80±1.82	71.38±1.69	55.35±1.67
STM	64	49.63±1.74	67.33±1.65	72.83±1.55	57.86±1.66	48.88±1.60	67.29±1.74	72.46±1.57	57.25±1.61
STM	128	**51.15±1.69**	**67.62±1.58**	**72.90±1.49**	**58.89±1.59**	**50.40±1.52**	**67.58±1.70**	**72.48±1.40**	**58.30±1.50**
MOLSYNTEXT	32	72.91±2.57	94.45±0.85	96.88±0.25	82.30±1.84	74.11±2.45	94.01±0.88	96.47±0.24	82.87±1.76
MOLSYNTEXT	64	75.13±2.39	94.88±0.81	**97.15±0.22**	83.80±1.75	76.42±2.29	94.60±0.84	96.80±0.20	84.46±1.68
MOLSYNTEXT	128	**76.15±2.24**	**94.97±0.78**	97.06±0.28	**84.39±1.69**	**77.06±2.10**	**94.75±0.79**	**96.81±0.19**	**84.90±1.58**

As shown in Table 14, increasing the physical batch size from 32 to 128 consistently improves R@1 and MRR on both textual views and in both retrieval directions, even though the effective batch size is the same. R@5 and R@10 follow the same trend with smaller differences, as recall approaches saturation. The only exception is molecule-to-text R@10 on MOLSYNTEXT, where $B = 64$ is marginally better. We conclude that the number of in-batch negatives seen simultaneously matters for contrastive training, and that gradient accumulation alone cannot replace a larger physical batch. We therefore use the largest power-of-two batch size that fits in GPU memory, $B = 512$, in all main experiments.

D.4 CONTRASTIVE OBJECTIVE

Finally, we isolate the effect of the contrastive objective while keeping the MOLTOKSYN architecture and all other hyperparameters fixed. Both variants use $K = 8$, a physical batch size of 128, an accumulation factor of 2, an effective batch size of 256, and the same learning-rate setting. The only difference is the objective: InfoNCE or the pairwise SigLIP-style sigmoid loss.

As shown in Table 15, the SigLIP-style objective improves all metrics in both retrieval directions and on both textual views. The gains are largest for R@1 and MRR, while the differences at R@5 and R@10 are smaller, especially on MOLSYNTEXT, where recall is already close to saturation. We conclude that the pairwise sigmoid objective is the better choice for MOLTOKSYN.

E LIMITATIONS

MOLSYNTEXT is bounded by the information available in its molecular cards. The descriptions are intended to express structure- and descriptor-grounded information and should not be interpreted

Table 15: Effect of the contrastive objective with the MOLTOKSYN architecture. All runs use $K = 8$, physical batch size 128, accumulation factor 2, and the same learning-rate setting. Best values are shown in bold.

		Molecule → Text				Text → Molecule			
Corpus	Objective	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
STM	InfoNCE	47.02±1.71	66.44±1.64	72.42±1.56	56.00±1.66	45.95±1.59	66.71±1.75	72.35±1.63	55.37±1.61
	SigLIP	**51.15±1.69**	**67.62±1.58**	**72.90±1.49**	**58.89±1.59**	**50.40±1.52**	**67.58±1.70**	**72.48±1.40**	**58.30±1.50**
MOLSYNTEXT	InfoNCE	71.79±2.36	94.12±0.74	96.94±0.21	81.52±1.72	73.26±2.27	94.23±0.82	96.74±0.20	82.47±1.65
	SigLIP	**76.15±2.24**	**94.97±0.78**	**97.06±0.28**	**84.39±1.69**	**77.06±2.10**	**94.75±0.79**	**96.81±0.19**	**84.90±1.58**

as experimentally validated evidence of biological activity, toxicity, binding affinity, or clinical relevance. Constrained generation and model-based filtering can reduce obvious unsupported claims but do not guarantee that every natural-language interpretation is chemically complete.

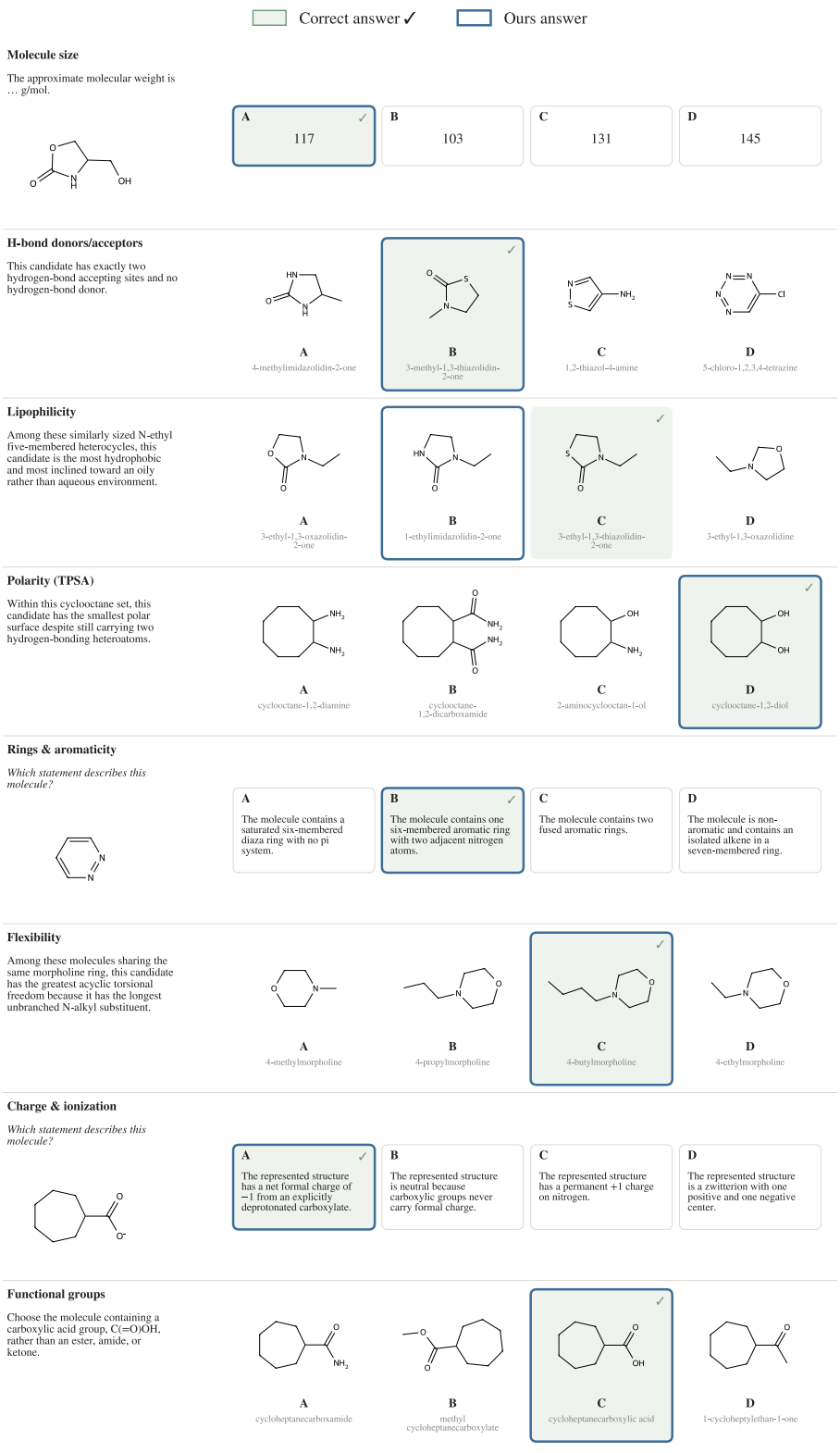


Figure 5: Qualitative examples from the Controlled Semantic Benchmark. Each question is either a molecule, answered by choosing one of four texts, or a text, answered by choosing one of four molecules. The correct answer is shaded green (✓), the answer chosen by MOLTOKSYN is framed in blue.

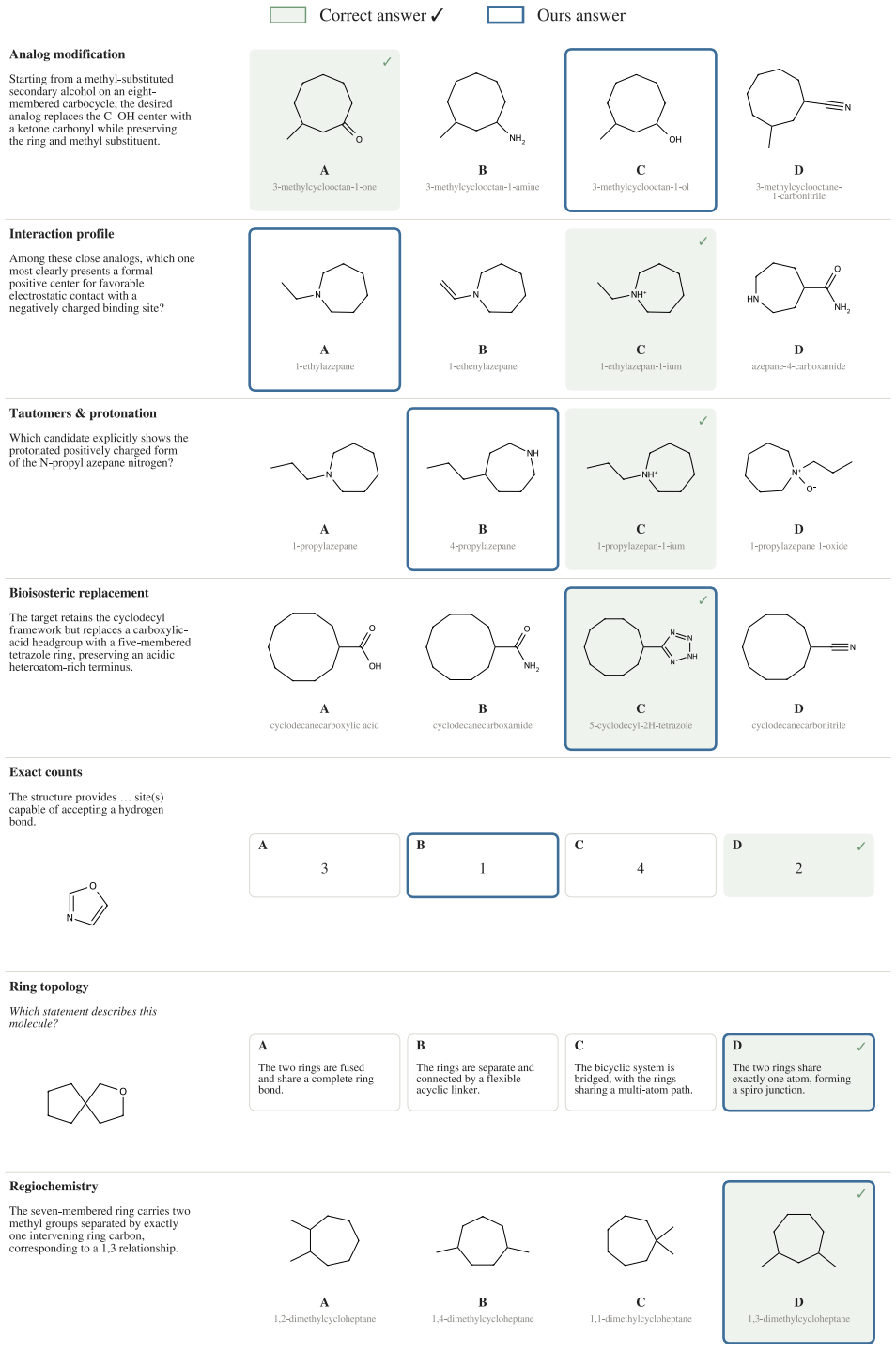


Figure 6: Qualitative examples from the Controlled Semantic Benchmark. Each question is either a molecule, answered by choosing one of four texts, or a text, answered by choosing one of four molecules. The correct answer is shaded green (✓), the answer chosen by MOLTOKSYN is framed in blue.

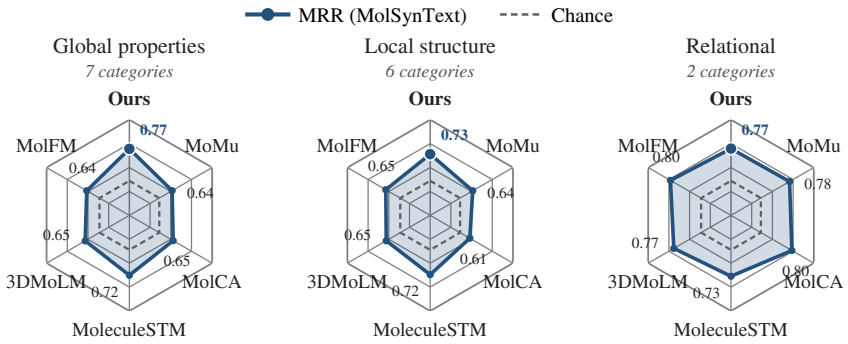


Figure 7: Controlled Semantic Benchmark aggregated by concept family (MOLSYNTEXT). Each radar summarizes one family (global properties, local structure, and relational tasks). Values denote the mean reciprocal rank (MRR) of the correct answer, first averaged over the categories within the family and then over three models trained on different scaffold splits. The dashed hexagon marks the chance level of 0.521.

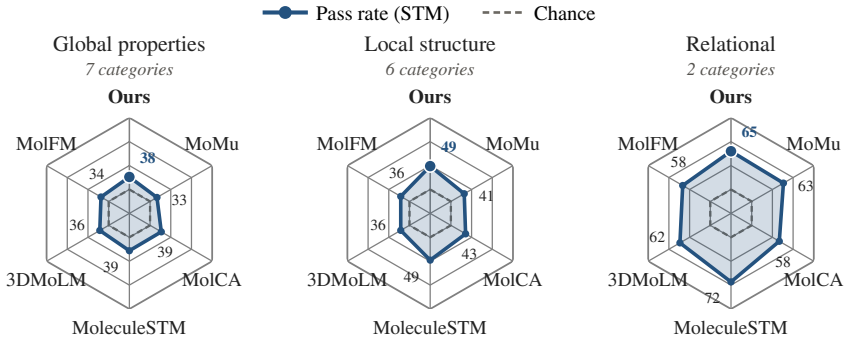


Figure 8: Controlled Semantic Benchmark aggregated by concept family (STM). Each radar summarizes one family (global properties, local structure, and relational tasks). Values denote the pass rate (%), first averaged over the categories within the family and then over three models trained on different scaffold splits. The dashed hexagon marks the chance level of 25%.

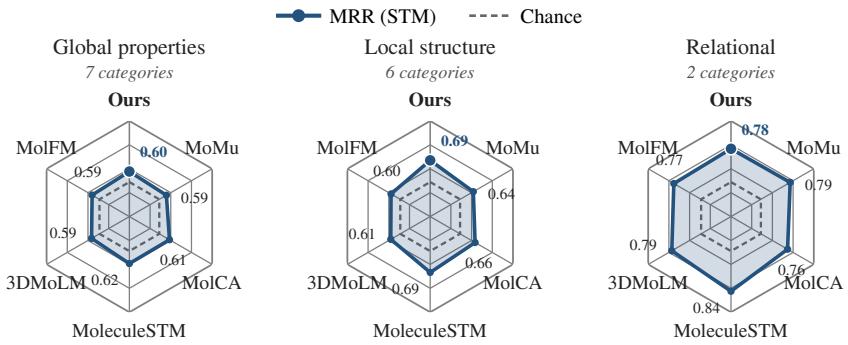


Figure 9: Controlled Semantic Benchmark aggregated by concept family (STM). Each radar summarizes one family (global properties, local structure, and relational tasks). Values denote the mean reciprocal rank (MRR) of the correct answer, first averaged over the categories within the family and then over three models trained on different scaffold splits. The dashed hexagon marks the chance level of 0.521.

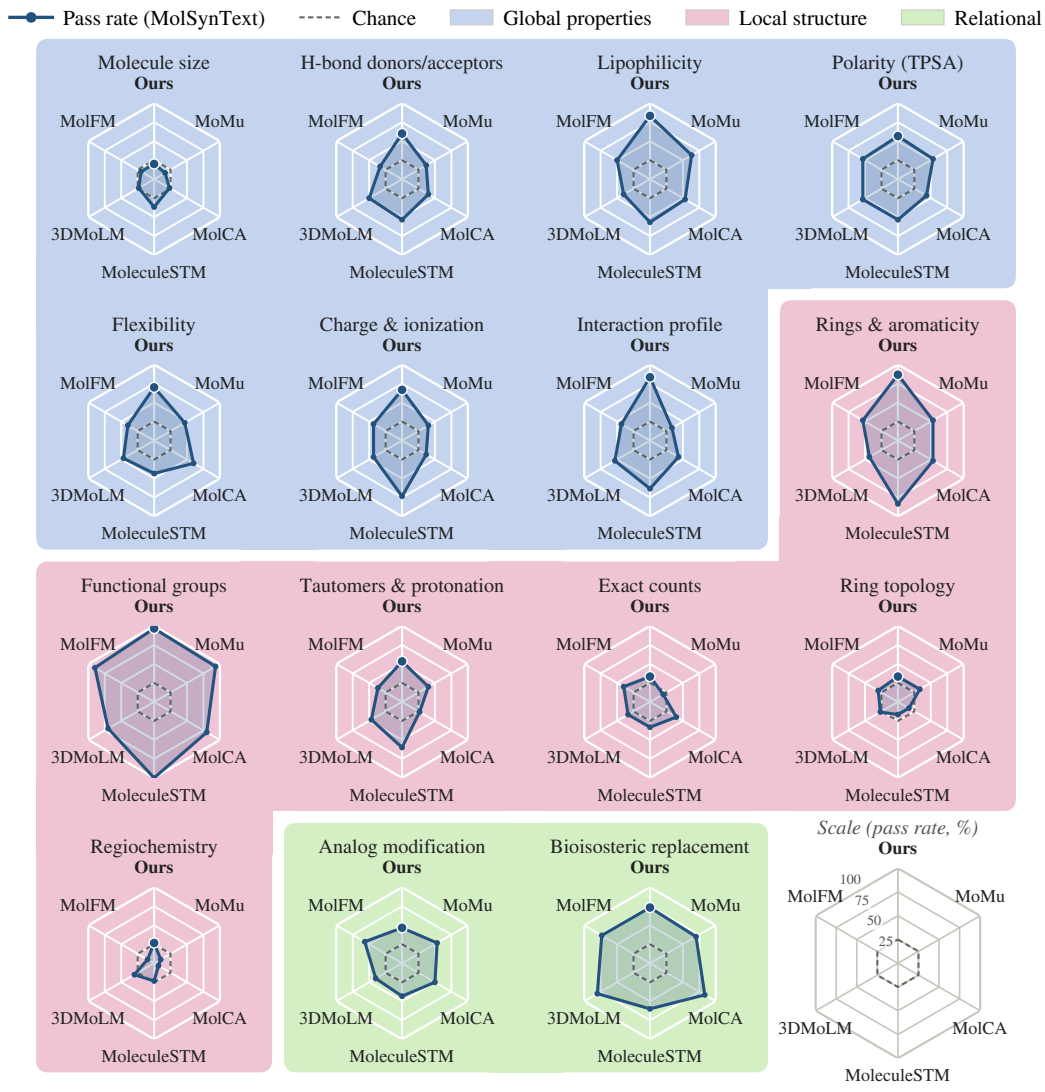


Figure 10: Controlled Semantic Benchmark results per category (MOLSYNTEXT). Each radar in the grid corresponds to one of the 15 categories. Values denote the pass rate (%), averaged over three models trained on different scaffold splits. The dashed line marks the chance level of 25%.

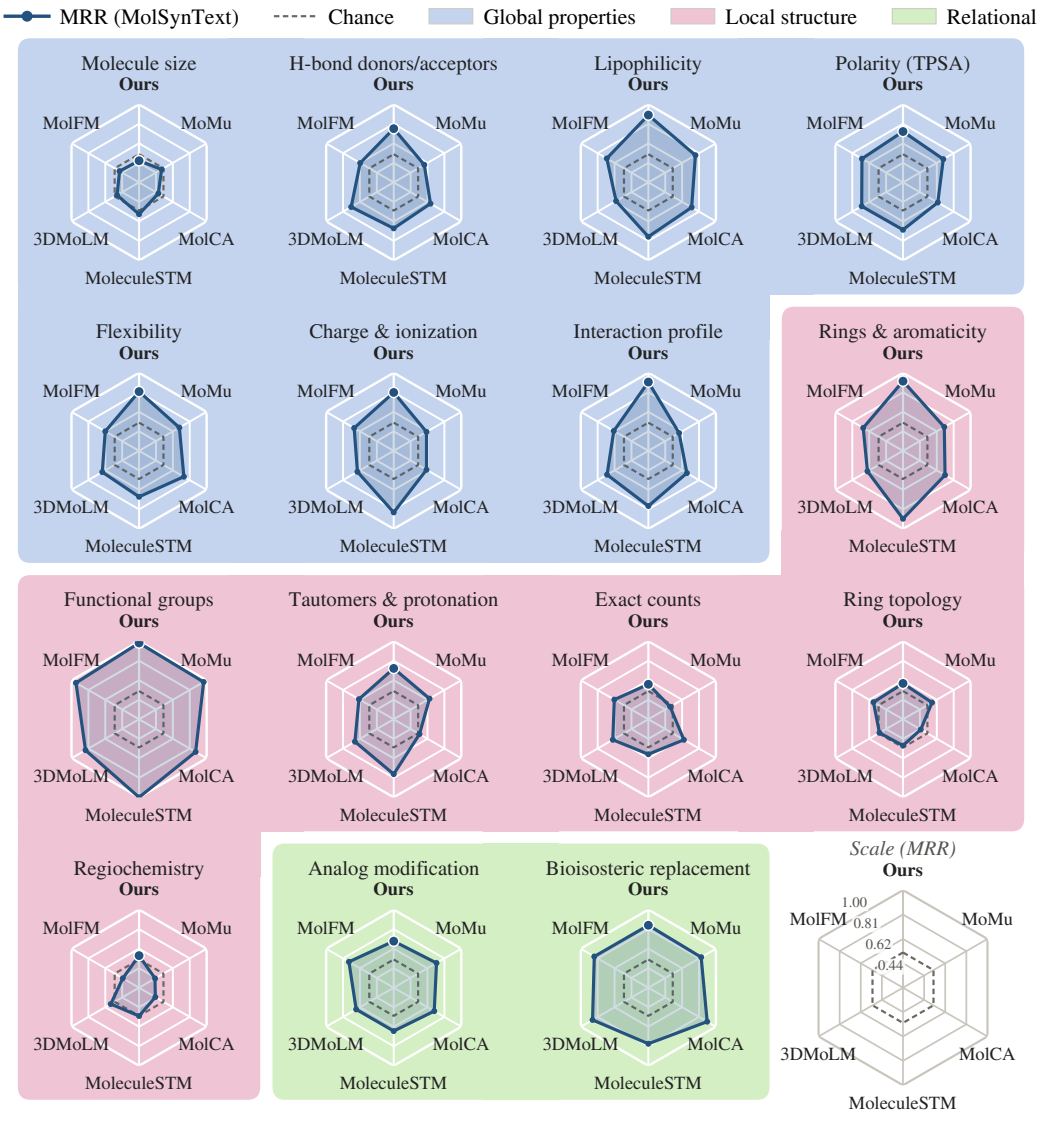


Figure 11: Controlled Semantic Benchmark results per category (MOLSYNTEXT). Each radar in the grid corresponds to one of the 15 categories. Values denote the mean reciprocal rank (MRR) of the correct answer, averaged over three models trained on different scaffold splits. The dashed line marks the chance level of 0.521.

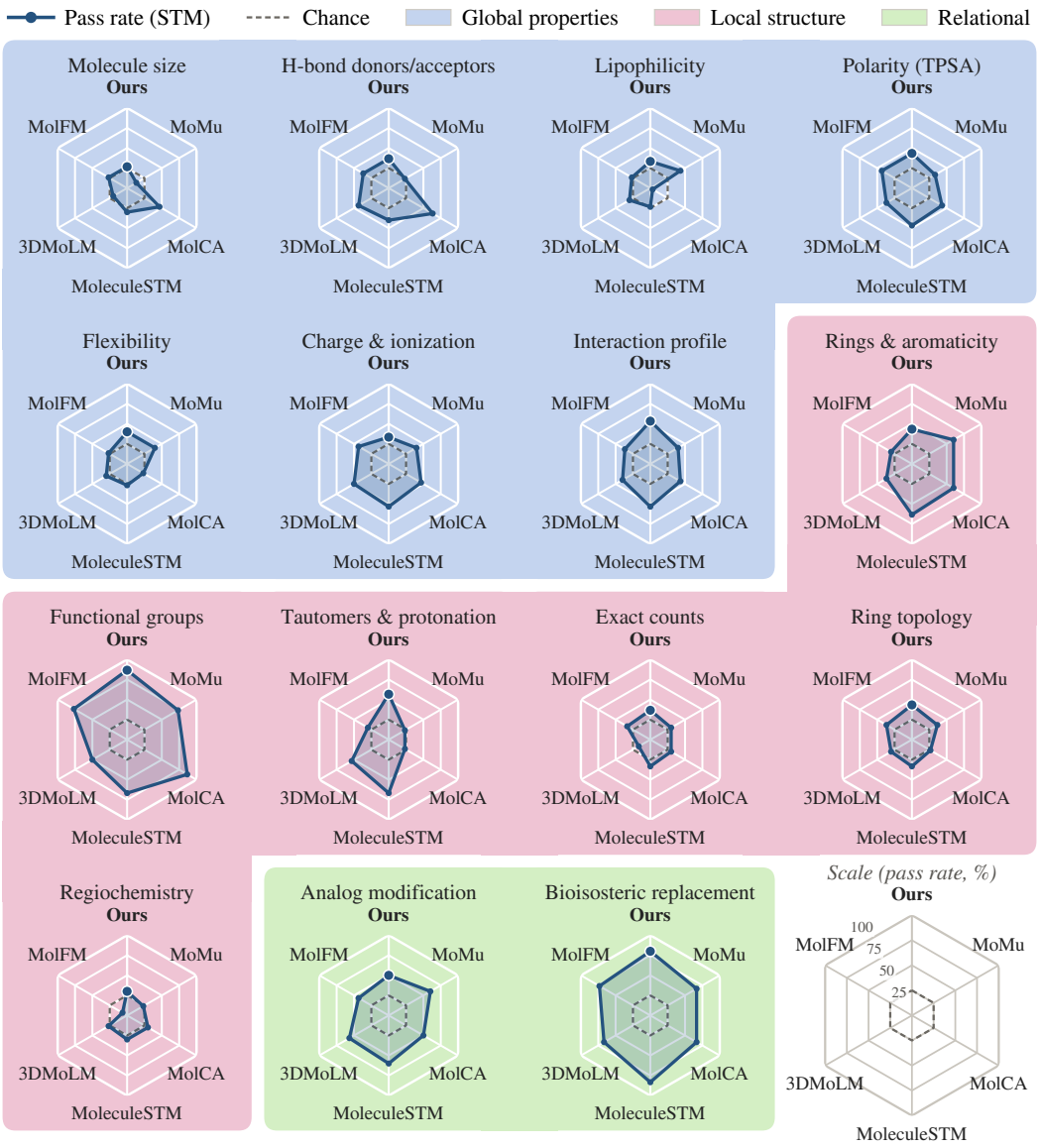


Figure 12: Controlled Semantic Benchmark results per category (STM). Each radar in the grid corresponds to one of the 15 categories. Values denote the pass rate (%), averaged over three models trained on different scaffold splits. The dashed line marks the chance level of 25%.

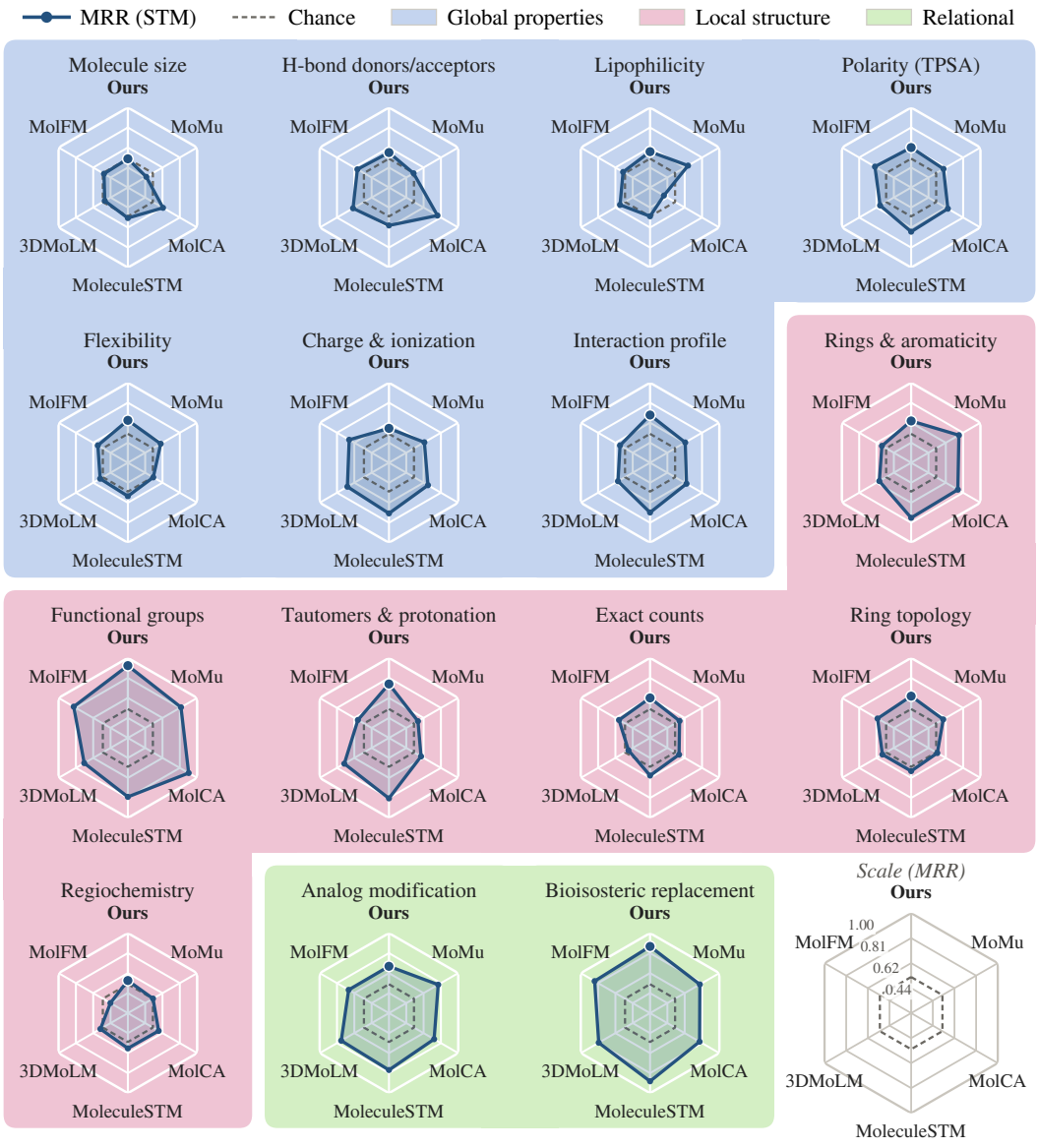


Figure 13: Controlled Semantic Benchmark results per category (STM). Each radar in the grid corresponds to one of the 15 categories. Values denote the mean reciprocal rank (MRR) of the correct answer, averaged over three models trained on different scaffold splits. The dashed line marks the chance level of 0.521.