MolEmb: Multimodal Large Language Models Can Be Strong Molecular Embedding Models

Xinjian Zhao [1 2] **Xiangru Jian** [3] **Yaoyao Xu** [1 2] **Xiaozhuang Song** [1 2] **Wei Pang** [1] **Lei Bai** [1 2] **Tianshu Yu** [1 2]

Abstract

Small-molecule modeling is important across drug discovery, chemical biology, and computational life sciences, where molecular embedding models can serve as foundational infrastructure for property prediction, virtual screening, toxicity assessment, and retrieval. Most molecular encoders are specialist models built around a single molecular view, producing unconditional vectors with no language interface for varying the representation. We ask whether multimodal large language models (MLLMs), which natively process images, text, and symbolic inputs, can instead serve as *general molecular embedding models* that produce embeddings conditioned on both a molecular profile and a natural-language semantic context. We introduce **MolEmb**, a lightweight framework that adapts MLLMs by aligning molecular profiles with textual descriptions in a shared embedding space using a bidirectional contrastive objective. The resulting embedding model is competitive on molecular property prediction and supports cross-modal molecule–text retrieval in the same space. We further introduce **MolCAR**, a diagnostic benchmark for context-aware retrieval, and find that context-aware molecular embedding is primarily a data property of the supervision. These results suggest that MLLMs are not merely chemistry assistants or generators, but a viable and extensible route to general molecular embedding models for life-science workflows.

1. Introduction

Small-molecule data are a core modality in drug discovery, chemical biology, toxicology, and broader computational life-science workflows. Molecular structures shape physicochemical properties and interactions, from solubility and reactivity to binding affinity and toxicity (Wildman & Crippen, 1999; Wu et al., 2018; Sandfort et al., 2020; Bian et al., 2026b; Deng et al., 2023). As deep learning becomes increasingly central to molecular modeling, learned molecular representations have become a common interface for property prediction, virtual screening, similarity search, and retrieval-augmented scientific reasoning (Gilmer et al., 2017; Cao et al., 2024; Ying et al., 2026; Zhong et al., 2025). Molecular embedding models can serve as foundational infrastructure for these workflows. We use this term for models that provide stable, reusable embeddings, in the spirit of embedding models in natural language processing and information retrieval, rather than only an intermediate state of a task-specific prediction pipeline.

The field has produced such embeddings primarily through specialist encoders over molecular graphs, SMILES strings, or molecular geometry (Edwards et al., 2022; Zhou et al., 2023; Bian et al., 2026a). These encoders are highly effective, especially for molecular prediction, but their representation interfaces are typically fixed in advance: one molecular view is encoded into a single unconditional vector. This raises a different embedding-model question: can molecular representations be exposed through a reusable interface whose input can vary by molecular view and semantic context?

To make this question precise, we study what we call a *general molecular embedding model*. Here, *general* emphasizes two properties beyond the usual reusable-vector view of embedding models. First, multi-view input: molecules can be described through complementary forms, such as 2D depictions, SMILES strings, conformers, and textual annotations, and a flexible model should accept the views available in a given workflow. Second, semantic conditioning: scientific workflows often impose different task lenses and domain expectations on the same molecule; a toxicity-focused query should emphasize different molecular evidence than a solubility-focused one. We formalize this object in Section 3.

MLLMs (Achiam et al., 2023; Bai et al., 2025b; Comanici et al., 2025) are natural candidates for such mod-

[1]The Chinese University of Hong Kong, Shenzhen [2]Shanghai Artificial Intelligence Laboratory [3]University of Waterloo. Correspondence to: Tianshu Yu <tianshuyu@cuhk.edu.cn>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026. Copyright 2026 by the author(s).

els. They process heterogeneous inputs such as images, text, and symbolic inputs; they can receive semantic context through natural language; and their hidden states can be pooled into fixed-length embeddings. This structural fit is timely because the broader foundation-model community has begun to treat generative backbones as embedding models. LLM2Vec (BehnamGhader et al., 2024) and VLM2Vec (Jiang et al., 2024) show that generative backbones can be adapted for retrieval and representation learning. This shift is also visible in industry-scale foundation-model systems: model families such as Qwen3-VL-Embedding (Li et al., 2026) and Gemini Embedding (Lee et al., 2025) expose embedding-oriented variants alongside their generative counterparts.

In chemistry and broader scientific applications of MLLMs, however, the dominant framing remains generative: explanation, captioning, multimodal reasoning, or direct prediction (Li et al., 2025b; Tan et al., 2025; Lu et al., 2024; Li et al., 2025a; Zhou et al., 2025; Zhao et al., 2026). These capabilities are valuable, but they leave underexplored the stable vector interface that many molecular workflows require. We therefore ask whether MLLMs can serve not only as chemistry assistants or generators, but as the basis of general molecular embedding models.

We study this question through *MolEmb*, a lightweight framework for adapting MLLMs into molecular embedding models. Each molecule is represented by a multi-view profile consisting of a 2D depiction and a canonical SMILES string, while semantic context is supplied through a text instruction when needed. We extract a fixed-length representation from the MLLM and align molecular profiles with textual descriptions in a shared embedding space. For molecular learning, this route is attractive because it allows molecular embedding models to share the backbone and interface of modern foundation models, and thus to benefit in principle from progress in multimodal pretraining, scientific-domain data, inference infrastructure, and embedding-oriented training. Fig. 1 summarizes this route and the embedding-centered workflows evaluated below.

Our contributions are:

- We frame molecular representation learning through the lens of *molecular embedding models*, emphasizing embeddings as reusable deliverables rather than only intermediate states of prediction pipelines, and formalize the target object as a *general molecular embedding model*.
- We instantiate this perspective with MolEmb, a lightweight MLLM-based framework that aligns multi-view molecular profiles with textual descriptions in a shared embedding space to produce reusable molecular embeddings.
- We evaluate the MLLM-to-embedding route across molecular prediction, cross-modal retrieval, and context-aware retrieval, showing both its practical viability and that context-aware molecular embedding is primarily a data property of the supervision.

2. Related Work

We briefly position MolEmb relative to molecular representation learning, embedding-oriented adaptation of generative backbones, and multimodal foundation models for science. Molecular representation learning has largely relied on specialist graph, sequence, and geometry encoders (Hu et al., 2019; You et al., 2020; Chithrananda et al., 2020; Edwards et al., 2022; Liu et al., 2021; Zhou et al., 2023). In parallel, generative language and vision-language backbones are increasingly being adapted into embedding models for retrieval and representation learning (BehnamGhader et al., 2024; Jiang et al., 2024; Zhang et al., 2025; Li et al., 2026; Lee et al., 2025). Chemistry MLLMs, however, are still mainly studied as assistants, captioners, reasoners, or direct predictors (Su et al., 2022; Liu et al., 2023; Luo et al., 2023; Li et al., 2025b; Tan et al., 2025; Yan et al., 2025). Full related work is expanded in Appendix A.

3. General Molecular Embedding Models

The method section turns this framing into an explicit encoder design. We study *general molecular embedding models*: models that map a molecular profile and a semantic context to a reusable vector. In contrast to specialist encoders that map a fixed molecular view to one unconditional representation, this interface separates three roles: molecular identity, available observations, and the task lens under which the molecule should be represented. Let $\mathcal{P} : \mathcal{M} \to \mathcal{V}$ map each molecule to its available profile and let $\mathcal{C}$ denote admissible semantic contexts. A general molecular embedding model is

$$E_\theta : \mathcal{V} \times \mathcal{C} \to \mathbb{R}^d, \qquad \mathbf{z}_c(m) = E_\theta(\mathcal{P}(m), c),$$

and the context-aware embedding family of molecule m is $\mathcal{Z}(m) = \{\mathbf{z}_c(m) : c \in \mathcal{C}\}$. In this paper, $\mathcal{P}(m)$ contains a 2D depiction and canonical SMILES string, and c is realized as a natural-language instruction; the full profile-space construction is given in Appendix B.

This notation separates three roles that are often entangled in molecular representation learning. The molecular identity m specifies the underlying compound, the profile $\mathcal{P}(m)$ specifies which observations are available, and the context c specifies the semantic lens under which the representation is formed. Under this formulation, a fixed molecule need not correspond to a single unconditional vector; instead, useful molecular embeddings may form a context-dependent family whose members share the same downstream interface while emphasizing different scientific evidence.

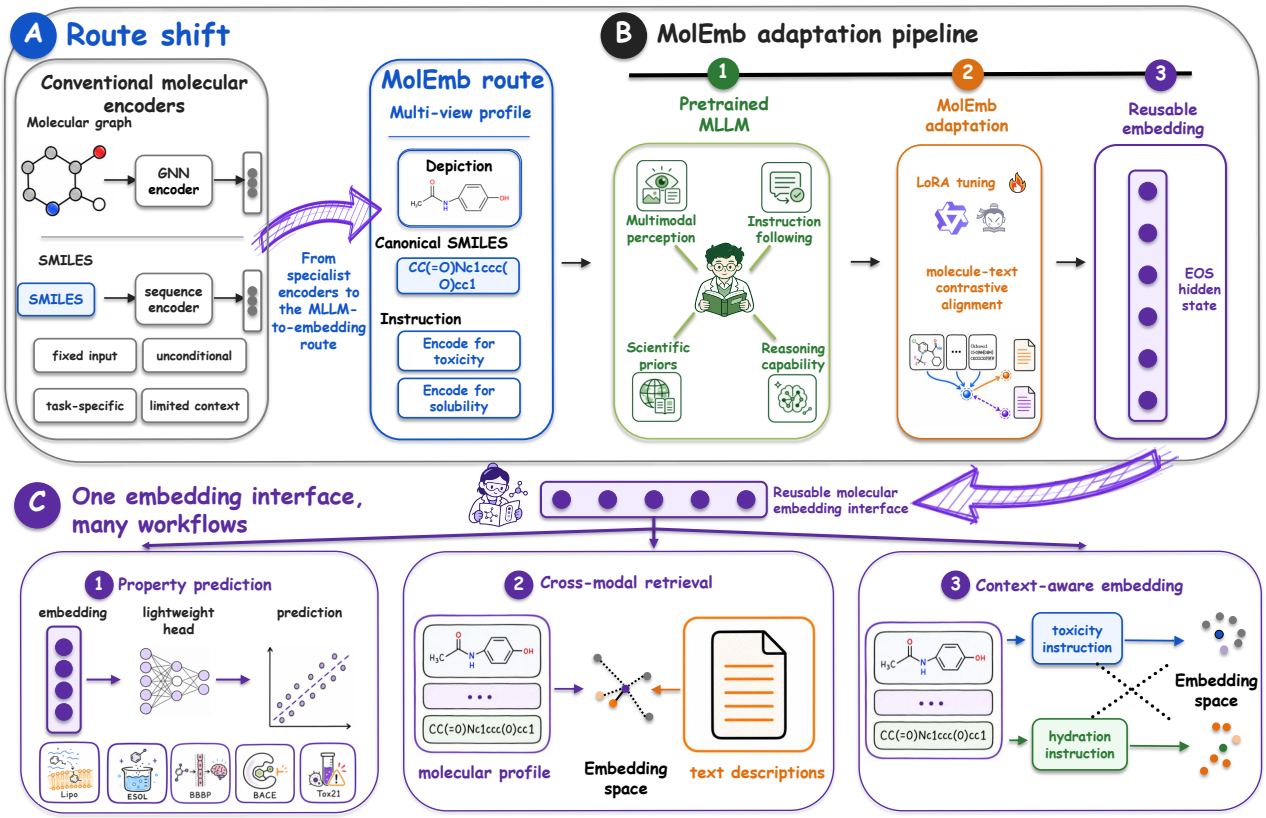

Figure 1. From specialist encoders to general molecular embedding models. (A) The MLLM-to-embedding route shifts molecular representation from fixed-view specialist encoders to an MLLM-based embedding interface that accepts multi-view molecular profiles and natural-language task instructions. (B) MolEmb instantiates this route by adapting a pretrained MLLM with lightweight LoRA modules through molecule–text contrastive alignment; the EOS hidden state is used as the molecular embedding. (C) The resulting reusable embedding interface supports property prediction, cross-modal molecule–text retrieval, and context-aware embedding by varying the task instruction.

MLLMs are natural candidates for E_θ: they accept heterogeneous token sequences, admit semantic context through natural language, and provide a reusable backbone whose multimodal pretraining and instruction-following priors can transfer to molecular adaptation. To instantiate E_θ, MolEmb formats the molecular profile and context as an MLLM input sequence

$$x(m, c) = \phi(\mathcal{P}(m), c),$$

where ϕ combines the available molecular views and the natural-language context. The MLLM produces hidden states $\mathbf{H}_\theta(x) \in \mathbb{R}^{L\times d}$, and we use end-of-sequence pooling to obtain the fixed-length embedding

$$E_\theta(\mathcal{P}(m), c) = \rho(\mathbf{H}_\theta(x(m, c))) \in \mathbb{R}^d.$$

This gives a concrete MLLM instantiation of the general embedding interface, but it does not assume that pretrained hidden states are already molecularly aligned. The central adaptation problem is to reshape the pooled MLLM representation so that molecular profiles, textual descriptions, and task-conditioned queries become comparable in a shared embedding space.

MolEmb then adapts the MLLM through molecule–text alignment. For a molecule-description pair (m_i, t_i), the molecule side is embedded as $\mathbf{z}_i^{\text{mol}} = E_\theta(\mathcal{P}(m_i), c_{\text{enc}})$ under a fixed molecule-encoding instruction, while the text side is embedded by passing the description through the same backbone, $\mathbf{z}_i^{\text{desc}} = \rho(\mathbf{H}_\theta(\phi_{\text{desc}}(t_i)))$. The trainable LoRA adapters are optimized with a bidirectional contrastive objective,

$$\mathcal{L} = \tfrac{1}{2}\big(\mathcal{L}_{\text{NCE}}(\mathbf{z}^{\text{mol}}, \mathbf{z}^{\text{desc}}) + \mathcal{L}_{\text{NCE}}(\mathbf{z}^{\text{desc}}, \mathbf{z}^{\text{mol}})\big),$$

placing molecular profiles and textual descriptions in the same embedding space.

MolEmb keeps the architecture unchanged beyond the chosen input views and alignment objective, allowing us to isolate the MLLM-to-embedding route itself. After alignment, the same encoder is reused in three ways. For molecular prediction, we attach a lightweight task head to the molecular embedding. For cross-modal retrieval, we encode molecular

profiles and text targets into the same space and rank by similarity. For context-aware embedding, we hold the molecule fixed and vary the instruction c, testing whether the embedding family $\mathcal{Z}(m)$ changes in task-meaningful directions. Full formal definitions and objective details are given in Appendix B; optimization details are in Appendix D.

4. Experiments

The three properties of a general molecular embedding model suggest three empirical questions. First, is $E_\theta(\mathcal{P}(m), c)$ useful when reused across molecular endpoints? Second, does the codomain $\mathbb{R}^d$ behave as a genuinely shared molecule–text space rather than a within-modality embedding? Third, is the context-aware family $\mathcal{Z}(m)$ non-degenerate, so that changing c moves a fixed molecule in task-meaningful directions? We answer these questions through molecular prediction, cross-modal retrieval, and MolCAR context-aware retrieval. Dataset statistics, implementation details, full tables, and examples are in Appendix C–G.

4.1. Molecular Property Prediction

We evaluate eight standard molecular property benchmarks under scaffold splits (Wu et al., 2018; Hu et al., 2020). Table 1 reports the main comparison against supervised and pretrained molecular baselines, including Direct, MolTextNet-aligned, and Mixed-aligned MolEmb variants. These rows test two requirements for the MLLM-to-embedding route: whether the multimodal molecule interface can support property learning, and whether broader molecule–text supervision improves downstream transfer. Dataset details and a modality ablation are in Appendix C and Appendix E.5.

Direct adaptation provides a useful starting point. Without molecule–text alignment, MLLM-derived embeddings already reach the range of several pretraining-based GNN methods, particularly on regression tasks. This does not by itself establish a reusable molecular embedding model, since Direct is still trained separately for each property task, but it shows that the multimodal molecule interface can be optimized for molecular property learning before introducing alignment.

Alignment benefits depend on data diversity and coverage. MolTextNet and Mixed alignment improve several endpoints, showing that molecule-description matching can reshape the MLLM representation into a more useful molecular space. The gains are not monotonic across tasks, however: Mixed alignment improves some endpoints, notably ESOL and Tox21, while MolTextNet remains stronger on others. This pattern suggests both the value and the current limitation of molecule–text data: generic description corpora transfer useful chemical cues, but still only partially cover task-conditioned molecular knowledge.

4.2. Cross-Modal Retrieval

Prediction alone does not establish a shared embedding space. We therefore evaluate molecule-to-text and text-to-molecule retrieval on the held-out MolTextNet test split (2,970 molecule–text pairs), a workflow that specialist molecular encoders cannot natively support because their unconditional vectors do not share a space with text. A general-purpose multimodal embedding model, Qwen3-VL-Embedding-8B, reaches only about 3% R@1 in both directions. After alignment on the MolTextNet corpus, all tested MolEmb backbones exceed 73% R@1, and Intern-S1-mini and Qwen3.5-0.8B exceed 83% (Appendix E.1). Among the three backbones, Intern-S1-mini, the only one with explicit scientific-domain pretraining, attains the best R@1 in both directions. The MLLM interface thus supports embedding-based molecule–text matching after molecular alignment, but cross-modal matching alone is not sufficient for a general molecular embedding model: the same molecule should also be representable differently across scientific task lenses, examined next.

The full retrieval metrics also show that R@5 and R@10 are near-saturated for aligned MolEmb models, so many remaining R@1 failures are local ranking errors rather than broad failures to retrieve a relevant molecular description. We therefore treat MolTextNet retrieval as a necessary intermediate check: it verifies that the shared codomain can support molecule–text matching, but not yet that the representation is context-aware.

4.3. Context-Aware Embedding with MolCAR

Beyond cross-modal retrieval, a reusable molecular embedding should support different scientific task lenses for the same molecule. We introduce **MolCAR**, a controlled diagnostic that holds molecule identity fixed and varies only the task instruction: a query about blood-brain barrier penetration, toxicity, or aqueous interaction may all refer to the same molecular structure but should retrieve different evidence. **MolCAR-Structured** deterministically constructs target cards from real task labels and molecular descriptors; **MolCAR-Natural** replaces those cards with Intern-S1 (235B)-generated scientific notes anchored to the same outcomes and profiles, reducing template regularity while preserving ground-truth anchors. **MolCAR-Train** is built from training splits after removing benchmark molecules by canonical SMILES, providing task-conditioned supervision without molecule leakage. The evaluation split contains 4,092 molecule–context pairs over 1,653 held-out molecules; construction details and examples are in Appendix C.4–G.

Table 1. Main downstream comparison with representative molecular representation methods. Regression columns report RMSE, where lower is better; classification columns report ROC-AUC in percent, where higher is better. For each column, the **best** and **second-best** results are highlighted.

	Regression (RMSE ↓)			Classification (ROC-AUC% ↑)				
Model	**ESOL**	**Lipo**	**FreeSolv**	**BACE**	**BBBP**	**ClinTox**	**Tox21**	**SIDER**
Supervised GNN								
S-GIN	1.173±0.057	0.757±0.018	2.755±0.349	72.97±4.00	68.17±1.48	88.14±2.51	74.91±0.51	57.60±1.40
R-GIN	1.706±0.180	1.075±0.022	7.526±2.119	75.07±2.23	64.48±2.46	72.29±4.15	71.53±0.74	62.29±1.12
Pre-training Methods								
InfoGraph	1.344±0.178	1.005±0.023	10.005±4.819	74.74±3.64	66.33±2.79	64.50±5.32	69.74±0.57	60.54±0.90
GraphCL	1.272±0.089	0.910±0.016	7.679±2.748	74.32±2.70	68.22±1.89	74.92±4.42	72.40±1.01	61.76±1.11
MVGRL	1.433±0.145	0.962±0.036	9.024±1.982	74.20±2.31	67.24±1.39	73.84±4.25	70.48±0.83	61.94±0.94
AD-GCL	1.217±0.087	0.842±0.028	5.150±0.624	76.37±2.03	68.24±1.47	80.77±3.92	71.42±0.73	63.19±0.95
JOAO	1.285±0.121	0.865±0.032	5.131±0.722	74.43±1.94	67.62±1.29	78.21±4.12	71.83±0.92	62.73±0.92
GCL-SPAN	1.218±0.052	0.802±0.019	4.531±0.463	76.74±2.02	69.59±1.34	80.28±2.42	72.83±0.62	64.87±0.88
Mole-BERT	1.015±0.030	**0.676±0.017**	–	80.80±1.40	71.90±1.60	78.90±3.00	76.80±0.50	62.80±1.10
MolTextNet	1.145±0.070	0.753±0.008	2.357±0.106	**84.70±0.10**	70.40±2.40	90.00±0.20	75.20±0.30	64.00±3.10
MolEmb (Ours)								
Intern-S1-mini (Direct)	0.827±0.003	0.700±0.008	**1.839±0.031**	75.66±1.49	73.09±0.25	**99.51±0.11**	77.69±0.59	61.28±1.31
Intern-S1-mini (MolTextNet)	0.802±0.001	0.678±0.006	1.886±0.034	81.27±0.17	**75.00±0.18**	**99.54±0.06**	77.86±0.74	**65.44±0.52**
Intern-S1-mini (Mixed)	**0.771±0.012**	**0.673±0.004**	1.965±0.080	78.65±2.15	**74.89±1.46**	99.50±0.18	**79.24±0.71**	64.78±0.08
Qwen3-VL-8B (Direct)	0.864±0.016	0.792±0.009	2.029±0.056	79.22±1.34	70.39±0.69	97.03±2.84	75.77±0.19	59.99±0.38
Qwen3-VL-8B (MolTextNet)	0.810±0.017	0.699±0.002	2.036±0.043	75.03±0.90	72.48±0.80	99.05±0.51	77.37±0.15	63.64±0.07
Qwen3-VL-8B (Mixed)	0.793±0.004	0.689±0.001	1.936±0.008	76.37±0.13	74.77±0.21	98.70±0.46	**78.08±0.07**	**64.85±0.19**
Qwen3.5-0.8B (Direct)	0.958±0.013	0.796±0.008	2.447±0.175	78.39±0.94	69.06±1.76	98.97±0.02	73.98±0.73	59.15±2.11
Qwen3.5-0.8B (MolTextNet)	0.825±0.038	0.696±0.016	**1.834±0.139**	**80.97±0.79**	71.58±2.89	99.14±0.43	76.19±0.87	62.14±0.18
Qwen3.5-0.8B (Mixed)	**0.781±0.016**	0.719±0.011	1.857±0.084	78.47±2.63	72.79±1.29	99.30±0.19	77.45±0.25	64.39±1.54

Table 2. MolCAR Context R@1 (%). Context R@1 ranks the correct task context first among documents sharing the same molecule. Base is Mixed-aligned; Continued denotes continued alignment on MolCAR-Train. Full MolCAR metrics are in Appendix E.2.

Variant	Model	Base Inst	Base NoInst	Continued Inst
Structured	Intern-S1-mini	44.4	40.4	**99.8**
Structured	Qwen3.5-0.8B	41.1	40.4	69.4
Natural	Intern-S1-mini	44.3	40.4	99.2
Natural	Qwen3.5-0.8B	40.6	40.4	72.0

Table 2 shows the key result. Context R@1 ranks the correct task context only among documents sharing the same molecule, factoring out molecule-level identification and directly testing whether the instruction changes the embedding in the intended semantic direction. Before continued alignment, both backbones remain near the 40.4% Context R@1 baseline, indicating that generic molecule–text alignment does not reliably separate different valid descriptions of the same molecule.

Continued alignment expands the context-aware embedding family. After continued alignment on MolCAR-Train, Context R@1 with task instructions reaches 99.8% (Intern-S1-mini) and 69.4% (Qwen3.5-0.8B), while NoInst stays at the 40.4% baseline. In the language of Definition B.2, the alignment-pretrained model behaves as if $\mathbf{z}_c(m) \approx \mathbf{z}_{c'}(m)$; continued alignment expands $\mathcal{Z}(m)$ along task-meaningful directions while preserving general molecule–text retrieval (Appendix E.3). The same pattern holds on MolCAR-Natural, where generated scientific notes replace fixed-schema cards; for Qwen3.5-0.8B, Natural Context R@1 is slightly higher than Structured (72.0% vs. 69.4%), consistent with language-grounding priors being better used by natural scientific text (Appendix E.2).

Context-aware embedding is a data property. This pattern is consistent with the structure of the supervision: alignment corpora pair each molecule with templated, narrow-distribution descriptions (Appendix F), so the model rarely sees the same molecule under varying task lenses. MolCAR-Train supplies the missing supervision; context-aware molecular embedding is therefore primarily a data property of the supervision, induced by task-diverse, outcome-grounded molecule–text examples.

5. Conclusion

We studied whether multimodal large language models can be adapted into reusable molecular embedding models. MolEmb takes a simple route: represent each molecule through an image depiction, SMILES, and a text context; extract an MLLM representation; and align molecule and text embeddings with a contrastive objective. Across property prediction, molecule–text retrieval, and MolCAR context-aware retrieval, the results suggest that MLLMs are not merely chemistry assistants or generators, but a viable and extensible route to general molecular embedding models. The broader implication is data-centric: task-diverse, outcome-grounded molecule–text supervision appears to be the central lever for context-aware molecular embeddings.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025a.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

BehnamGhader, P., Adlakha, V., Schmidt, F. D., Chapados, N., Mosbach, M., and Reddy, S. Llm2vec-gen: Generative embeddings from large language models. *arXiv preprint arXiv:2603.10913*, 2026.

Bian, Y., Wu, H., and Yan, J. Deep learning for affinity prediction and interface prediction in molecular interactions. *Deep Learning in Drug Design*, pp. 283–296, 2026a.

Bian, Y., Yang, N., Wu, J., and Yan, J. Deep learning for complex structure prediction in molecular interactions. In *Deep Learning in Drug Design*, pp. 297–308. Elsevier, 2026b.

Cao, Z., Sciabola, S., and Wang, Y. Large-scale pretraining improves sample efficiency of active learning-based virtual screening. *Journal of Chemical Information and Modeling*, 64(6):1882–1891, 2024.

Cheng, Z., Xiang, H., Liu, M., Zeng, L., Zeng, X., and Song, B. Remol: A chemical reaction knowledge-guided self-supervised molecular image representation learning framework. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., and Wang, F. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395, 2023.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Günther, M., Sturua, S., Akram, M. K., Mohr, I., Ungureanu, A., Wang, B., Eslami, S., Martens, S., Werk, M., Wang, N., et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 531–550, 2025.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., and Chen, W. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Lee, J., Chen, F., Dua, S., Cer, D., Shanbhogue, M., Naim, I., Ábrego, G. H., Li, Z., Chen, K., Vera, H. S., et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

Li, H., Fang, X., Li, Y., Huang, C., Wang, J., Wang, X., Bai, H., Hao, B., Lin, S., Liang, H., et al. Rxnbench: A multimodal benchmark for evaluating large language models on chemical reaction understanding from scientific literature. *arXiv preprint arXiv:2512.23565*, 2025a.

Li, J., Zhang, D., Wang, X., Hao, Z., Lei, J., Tan, Q., Zhou, C., Liu, W., Yang, Y., Xiong, X., et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 415–423, 2025b.

Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Lin, L., Chen, J., and Wang, H. Spectral augmentation for self-supervised learning on graphs. *ICLR*, 2023.

Liu, P., Tao, J., and Ren, Z. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. *Nature Machine Intelligence*, 7(2): 315–327, 2025.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Lu, X., Cao, H., Liu, Z., Bai, S., Chen, L., Yao, Y., Zheng, H.-T., and Li, Y. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3769–3789, 2024.

Luo, Y., Yang, K., Hong, M., Liu, X. Y., and Nie, Z. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C., and Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390, 2020.

Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.

Tan, Q., Zhou, D., Xia, P., Liu, W., Ouyang, W., Bai, L., Li, Y., and Fu, T. Chemmllm: Chemical multimodal large language model. *arXiv preprint arXiv:2505.16326*, 2025.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Wildman, S. A. and Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5): 868–873, 1999.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yan, Y., Wang, S., Huo, J., Ye, J., Chu, Z., Hu, X., Yu, P. S., Gomes, C., Selman, B., and Wen, Q. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025.

Ying, C., Ruan, Y., Chen, X., Wang, Y., and Yu, T. Neural graduated assignment for maximum common edge subgraphs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZVlTIyRe35.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Zhao, X., Pang, W., Xue, Z., Jian, X., Zhang, L., Xu, Y., Song, X., Wu, S., and Yu, T. The underappreciated power of vision models for graph structural understanding. *arXiv preprint arXiv:2510.24788*, 2025.

Zhao, X., Pang, W., Yu, Z., Jian, X., Song, X., Xu, Y., Xue, Z., Chen, D., Wu, S., Torr, P., et al. When vision meets graphs: A survey on graph reasoning and learning. *Authorea Preprints*, 2026.

Zhong, X., Jin, B., Ouyang, S., Shen, Y., Jin, Q., Fang, Y., Lu, Z., and Han, J. Benchmarking retrieval-augmented generation for chemistry. *arXiv preprint arXiv:2505.07671*, 2025.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023.

Zhou, Y., Wang, Y., He, X., Shen, A., Xiao, R., Li, Z., Feng, Q., Guo, Z., Yang, Y., Wu, H., et al. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning. *arXiv preprint arXiv:2506.10521*, 2025.

Zhu, Y., Chen, D., Du, Y., Wang, Y., Liu, Q., and Wu, S. Molecular contrastive pretraining with collaborative featurizations. *Journal of Chemical Information and Modeling*, 64(4):1112–1122, 2024.

Zhu, Y., Liu, G., Inae, E., and Jiang, M. Moltextnet: A two-million molecule-text dataset for multimodal molecular learning. *arXiv preprint arXiv:2506.00009*, 2025.

A. Additional Related Work

Molecular representation learning. Molecular representation learning has been widely studied through graph-, sequence-, and geometry-based encoders. Graph neural networks encode atoms and bonds directly and have been strengthened by self-supervised objectives such as attribute masking, context prediction, and graph contrastive learning (Hu et al., 2019; You et al., 2020; Wang et al., 2022; Lin et al., 2023; Zhu et al., 2024). Sequence-based molecular language models instead treat SMILES strings as chemical text, enabling scalable Transformer pretraining with masked-language or sequence-to-sequence objectives (Chithrananda et al., 2020; Ahmad et al., 2022; Edwards et al., 2022). A complementary line incorporates molecular geometry, using conformers, distances, or coordinate denoising to capture spatial molecular structure (Liu et al., 2021; Zhou et al., 2023). Beyond these molecular views, recent work also explores vision-based encoders that learn from rendered molecular diagrams or graph visualizations, showing that visual representations can capture useful structural and chemical patterns (Cheng et al., 2026; Zhao et al., 2025). These specialist encoders achieve strong property prediction performance, but they are usually tied to a fixed molecular view and produce unconditional embeddings, making it difficult to condition representations on natural-language task semantics.

From generative models to embedding models. Recent work increasingly suggests that strong generative backbones can be repurposed as high-quality embedding models (BehnamGhader et al., 2024; Jiang et al., 2024; Zhang et al., 2025; Li et al., 2026; Lee et al., 2025; Günther et al., 2025; BehnamGhader et al., 2026). In the text-only setting, methods such as LLM2Vec and NV-Embed show that decoder-style language models can be adapted into competitive text embedding models for retrieval and representation learning (BehnamGhader et al., 2024). In the multimodal setting, VLM2Vec similarly shows that vision-language models can be converted into instruction-guided embedding models for multimodal retrieval (Jiang et al., 2024). This trend is also reflected in commercial model families, where dedicated embedding variants have emerged alongside generative backbones, such as the Qwen Embedding series (Zhang et al., 2025; Li et al., 2026) and Google's Gemini embedding models (Lee et al., 2025). These results suggest that generation and embedding should not be treated as disjoint model families: a strong generative backbone can often serve as the basis of a strong embedding model after suitable adaptation.

Multimodal foundation models for science. Large multimodal models have increasingly been applied to chemistry and broader scientific domains for question answering, captioning, multimodal reasoning, and cross-modal understanding (Su et al., 2022; Liu et al., 2023; Luo et al., 2023; Li et al., 2025b; Tan et al., 2025; Yan et al., 2025; Zhao et al., 2026). Within chemistry, models such as MolFM (Luo et al., 2023) and ChemVLM (Li et al., 2025b) study multimodal learning over molecular structure, text, and related scientific signals. More broadly, scientific multimodal foundation models such as Intern-S1 extend this paradigm beyond chemistry-specific settings toward general scientific understanding (Bai et al., 2025a). Together, these works highlight an important structural advantage: multimodal large models naturally accommodate heterogeneous inputs and language-based interaction. However, much of the current literature remains generation-centric, focusing on answering, captioning, or reasoning over scientific inputs. In contrast, our work studies such backbones as embedding models: the goal is not primarily to generate chemistry-flavored text, but to produce reusable molecular representations that support prediction, retrieval, and transfer.

With that positioning in place, the next section states the formal model and alignment objective in full.

B. Formal Method Details

B.1. Definition

The main text gives a compact form of the general molecular embedding interface; this section expands the profile-space construction, MLLM instantiation, and full contrastive objective.

Standard molecular encoders typically map a molecule, under a fixed input representation, to a single unconditional vector. That abstraction is highly effective for specialist prediction tasks, but it leaves two aspects of molecular representation implicit. First, molecules are naturally multi-view objects: the same underlying molecule may be observed through a 2D depiction, a symbolic string, conformers, or other complementary annotations. Second, the most useful representation of a molecule need not be unconditional: the same molecule may require different embeddings under different semantic contexts.

Motivated by this perspective, we study what we call a *general molecular embedding model*, which maps a molecular profile together with a semantic context to a vector representation in a shared embedding space.

Definition B.1 (General Molecular Embedding Model)**.** Let $\mathcal{M}$ be a set of molecules, identified up to a chosen canonicaliza-

tion scheme. Let K be the number of possible molecular view types. For each view type k, let $\mathcal{V}_k$ denote the corresponding view space, such as 2D depictions, symbolic strings, conformers, or other molecule-specific annotations. We define the molecular profile space as

$$\mathcal{V} = \prod_{k=1}^{K} \big(\mathcal{V}_k \cup \{\varnothing_k\}\big),$$

where $\varnothing_k$ denotes that view type k is unavailable. A molecular profile map $\mathcal{P} : \mathcal{M} \to \mathcal{V}$ assigns to each molecule $m \in \mathcal{M}$ its available views $\mathcal{P}(m)$. Let $\mathcal{C}$ be a set of admissible semantic contexts, and let d be the embedding dimension. A *general molecular embedding model* is a parameterized map $E_\theta : \mathcal{V} \times \mathcal{C} \to \mathbb{R}^d$ with $\theta \in \Theta$, where Θ denotes the parameter space. For each molecule $m \in \mathcal{M}$ and context $c \in \mathcal{C}$, we write

$$\mathbf{z}_c(m) := E_\theta(\mathcal{P}(m), c) \in \mathbb{R}^d.$$

Definition B.2 (Context-Aware Embedding Family)**.** Given a general molecular embedding model E_θ, the *context-aware embedding family* of a molecule m is

$$\mathcal{Z}(m) := \{\mathbf{z}_c(m) : c \in \mathcal{C}\}.$$

Definitions B.1–B.2 separate three roles that are often entangled in standard molecular representation learning. The molecular identity m specifies the underlying molecule, the profile $\mathcal{P}(m)$ specifies the available observations of that molecule, and the context c specifies the semantic lens under which the representation is formed. Multi-view input is captured through $\mathcal{P}(m)$, semantic conditioning through c, and a common downstream interface through the shared codomain $\mathbb{R}^d$. Under this formulation, a fixed molecule need not correspond to a single unconditional vector. Instead, it may induce a family of context-dependent embeddings $\mathcal{Z}(m)$, giving rise to a context-aware embedding geometry in which changing the semantic context may move the same molecule toward different regions of the shared space while preserving a common interface for downstream use.

Whether such an object can be realized in practice is an empirical question. Section 4 tests whether multimodal large models instantiate it in a way that supports downstream prediction, retrieval, and context-aware behavior. In this paper, we instantiate $\mathcal{P}(m)$ with a 2D depiction and a canonical SMILES string, and realize c through natural-language instructions.

B.2. MLLMs as a Natural Instantiation

MLLMs are well-suited to instantiate Definition B.1 for both structural and strategic reasons. Structurally, they process heterogeneous token sequences, including image patches, text tokens, and symbolic strings. This allows the profile space $\mathcal{V}$ to include molecular depictions, SMILES, and other molecule-specific inputs, while allowing semantic contexts in $\mathcal{C}$ to be expressed through natural-language instructions. Strategically, the pretrained parameters of the MLLM provide a reusable backbone for E_θ: improvements in multimodal pretraining, instruction following, and embedding-oriented adaptation can in principle be inherited without redesigning the molecular embedding interface. More broadly, the language grounding, instruction-following behavior, and broad scientific priors of such backbones make them natural candidates for interpreting semantic contexts in $\mathcal{C}$.

To instantiate E_θ with an MLLM, we construct an input sequence

$$x(m, c) = \phi(\mathcal{P}(m), c),$$

where ϕ formats the available molecular views and semantic context into the model input. The MLLM defines a hidden-state map $\mathbf{H}_\theta : \mathcal{X} \to \mathbb{R}^{L\times d}$, where $\mathcal{X}$ is the input-token space and L is the resulting sequence length. A pooling rule $\rho : \mathbb{R}^{L\times d} \to \mathbb{R}^d$ then maps the hidden states to a fixed-length vector:

$$E_\theta(\mathcal{P}(m), c) := \rho(\mathbf{H}_\theta(x(m, c))) \in \mathbb{R}^d.$$

In our implementation, ρ selects the hidden state at the end-of-sequence position. This gives a concrete instantiation of Definition B.1 with a pretrained MLLM backbone. However, pretrained MLLM representations are not automatically aligned to molecular semantics. Section 4 empirically examines this gap through molecule–text retrieval probes.

B.3. MolEmb: Adapting MLLMs through Molecule–Text Alignment

MolEmb builds on the MLLM instantiation above and aligns molecular profiles with textual descriptions within a shared embedding space. Let $\theta_0 \in \Theta$ denote the pretrained MLLM parameters. We keep the pretrained backbone fixed and introduce a small trainable adapter for molecular adaptation. In our implementation, the adapter is parameter-efficient, with exact placement specified in the experimental setup.

We optimize the trainable adapter using paired molecule-description data. During optimization, we write the current model parameters as θ. For a molecule-description pair (m_i, t_i) from an external corpus such as MolTextNet (Zhu et al., 2025) or ChEBI-20-MM (Liu et al., 2025), the molecular embedding is obtained as

$$\mathbf{z}_i^{\mathrm{mol}} = E_\theta\big(\mathcal{P}(m_i), c_{\mathrm{enc}}\big),$$

where $c_{\mathrm{enc}} \in \mathcal{C}$ is a fixed molecule-encoding instruction shared across all molecule–text alignment pairs. Let ϕ_{desc} format a textual description into the MLLM input-token space. The description embedding is obtained by passing t_i through the same backbone:

$$\mathbf{z}_i^{\mathrm{desc}} = \rho(\mathbf{H}_\theta(\phi_{\mathrm{desc}}(t_i)))\,.$$

Both embeddings lie in the shared space $\mathbb{R}^d$. We optimize a bidirectional contrastive objective

$$\mathcal{L} = \frac{1}{2}\Big(\mathcal{L}_{\mathrm{NCE}}(\mathbf{z}^{\mathrm{mol}}, \mathbf{z}^{\mathrm{desc}}) + \mathcal{L}_{\mathrm{NCE}}(\mathbf{z}^{\mathrm{desc}}, \mathbf{z}^{\mathrm{mol}})\Big),$$

where the molecule-to-description direction is

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{z}^{\mathrm{mol}}, \mathbf{z}^{\mathrm{desc}}) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\mathrm{sim}(\mathbf{z}_i^{\mathrm{mol}}, \mathbf{z}_i^{\mathrm{desc}})/\tau\big)}{\sum_{j=1}^{B} \exp\big(\mathrm{sim}(\mathbf{z}_i^{\mathrm{mol}}, \mathbf{z}_j^{\mathrm{desc}})/\tau\big)}.$$

Here B is the batch size, τ is the temperature, and $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity. The description-to-molecule direction is defined symmetrically. After optimization, we denote the resulting parameters by θ^* and the adapted embedding model by E_{θ^*}.

MolEmb keeps the architecture unchanged beyond the chosen input views and alignment objective, allowing us to isolate the MLLM-to-embedding route itself. This design treats the pretrained MLLM as a multimodal and language-grounded backbone, whose instruction-following behavior and broad scientific priors can be exposed through lightweight molecular alignment to support molecular prediction, cross-modal retrieval, and context-aware embedding.

The aligned model supports property prediction using a lightweight task-specific head, cross-modal retrieval by ranking candidates in the shared embedding space $\mathbb{R}^d$, and context-aware embedding by varying $c \in \mathcal{C}$ to obtain different representations of the same molecule. To isolate the contribution of alignment, we also evaluate a direct adaptation baseline that bypasses molecule–text alignment and trains the MLLM directly on downstream prediction tasks.

The three properties of a general molecular embedding model in Definition B.1 suggest three corresponding empirical questions, which structure the rest of this section: (i) is $E_\theta(\mathcal{P}(m), c)$ a useful predictor when reused across endpoints (*molecular prediction*, Section 4.1)? (ii) is the codomain $\mathbb{R}^d$ a genuinely shared molecule–text space rather than a within-modality cluster (*cross-modal retrieval*, Section 4.2)? and (iii) is the context-aware embedding family $\mathcal{Z}(m) = \{\mathbf{z}_c(m) : c \in \mathcal{C}\}$ non-degenerate, in the sense that c moves the embedding in task-meaningful directions for a fixed m (*context-aware embedding*, Section 4.3)? Each subsection asks whether the MLLM-to-embedding route already supports the corresponding property, or whether targeted supervision is needed.

C. Benchmark Details

C.1. Downstream Prediction Datasets

Table 3 summarizes the eight downstream datasets used in the main prediction evaluation. All datasets are taken from the OGB / MoleculeNet benchmark suite and use the standard scaffold split.

For multi-label classification (Tox21, SIDER), we report the macro-averaged ROC-AUC across all active tasks, following the standard OGB evaluation protocol. ClinTox has two clinical trial outcome labels; we report the averaged ROC-AUC over both.

Table 3. Downstream prediction datasets. Train/valid/test sizes follow the OGB scaffold split.

Dataset	Task	Train	Valid	Test	Metric
ESOL	Regression	902	113	113	RMSE
Lipophilicity	Regression	3360	420	420	RMSE
FreeSolv	Regression	513	64	65	RMSE
BACE	Binary cls.	1210	151	152	ROC-AUC
BBBP	Binary cls.	1631	204	204	ROC-AUC
ClinTox	Binary cls.	1181	148	148	ROC-AUC
Tox21	Multi-label cls. (12)	6264	783	784	ROC-AUC
SIDER	Multi-label cls. (27)	1141	143	143	ROC-AUC

C.2. MolCAR Retrieval Metrics

The MolCAR evaluation uses five retrieval metrics, all computed over a fixed candidate pool containing all evaluation documents (molecule–context pairs). Queries are multimodal molecular profiles, optionally augmented with a task instruction (Inst) or using a generic no-instruction variant (NoInst).

- **Mol. R@1**: fraction of queries for which the ground-truth molecule appears in rank 1. Because multiple documents share the same molecule, this measures molecule-level identification rather than exact document matching.
- **Doc. R@1**: fraction of queries for which the ground-truth document (correct molecule *and* correct task context) appears in rank 1.
- **Doc. MRR**: mean reciprocal rank of the ground-truth document across all queries; tables report $100\times$MRR.
- **Doc. R@5**: fraction of queries for which the ground-truth document appears in the top-5 results.
- **Context R@1**: fraction of queries for which the ground-truth context is ranked *first among all documents sharing the same molecule*. This metric isolates instruction sensitivity from molecule identification: it asks whether the model correctly routes the query to the right task view given that the molecule is already identified. A random baseline achieves $1/K$ where K is the number of context types for that molecule. Because molecules have different numbers of available task contexts, the aggregate random baseline for Context R@1 is approximately 40.4% (1,653/4,092) rather than 12.5%.

Context R@1 is the primary metric for the context-aware embedding evaluation because it directly measures whether the model uses the task instructions as a routing signal, independently of how well it retrieves the correct molecule.

C.3. MolCAR Dataset Construction

Evaluation split (MolCAR-Structured). Molecules are held out at the canonical-SMILES level: any molecule appearing in the training splits of the eight downstream datasets is removed from the evaluation pool using exact canonical-SMILES matching. The evaluation set contains 4,092 molecule–context pairs over 1,653 unique held-out molecules drawn from the test splits of the eight task families. Each molecule is paired with all available task contexts, forming a dense multi-context evaluation structure.

Training split (MolCAR-Train). The continued-alignment corpus is built from the training splits of the same eight datasets after removing all evaluation molecules by canonical SMILES. Each training instance maps one molecular profile to one task-conditioned textual target, providing same-molecule, multi-context supervision when a molecule appears in more than one dataset.

MolCAR-Natural. Structured context documents are replaced by scientific notes generated by Intern-S1 (235B parameters). The generation is anchored to the same ground-truth outcome and molecular profile as the structured document. The generation backbone (235B) is distinct from the MolEmb backbones (Intern-S1-mini, Qwen3.5-0.8B) evaluated in the paper, which avoids confound between data generation and model evaluation.

C.4. MolCAR Construction Instructions and Templates

This subsection lists the templates and instructions used to construct MolCAR. Three ingredients are involved: the task-instructed query format, the structured target card schema, and the natural-language generation prompt for MolCAR-Natural.

Task-instructed queries. All MolCAR queries follow a single template, with only the task name varying across the eight task families:

MolCAR Query Template

`<image>`
Encode this molecule for {*task_name*}. SMILES: {*smiles*}

The eight task names are: *BACE-1 inhibition*, *blood-brain barrier penetration*, *water solubility*, *hydration free energy*, *lipophilicity*, *Tox21 toxicity outcomes*, *clinical toxicity*, and *side-effect profile*. The NoInst variant replaces `Encode this molecule for {task_name}.` with the generic instruction `Encode this molecule.` while keeping the SMILES line and image token unchanged.

Structured target schema. MolCAR-Structured target cards are generated deterministically from ground-truth task labels and molecular descriptors using the four-line schema:

MolCAR-Structured Target Template

Report family: {*family_display_name*}
Task: {*task_name*}
Observed result: {*label_sentence*}
Molecular profile: {*descriptor_sentence*}

No language model is involved at this stage: `label_sentence` is templated from the dataset's task label (e.g., a binary positive/negative phrasing for classification, or a numeric value with units for regression), and `descriptor_sentence` is a fixed-format summary of MW, cLogP, HBD, HBA, and formal charge computed from the molecule.

Natural target generation prompt. MolCAR-Natural targets are produced by prompting Intern-S1 (235B) once per (molecule, task) pair, then re-anchoring the generated text with the trusted `Observed result` and `Molecular profile` lines from the structured card. The user prompt sent to the model is:

MolCAR-Natural Generation Prompt

System. Return only the final answer text. Do not output chain-of-thought, analysis traces, or `<think>` tags.

User. Write a 2–3 sentence scientist-facing retrieval note for this molecule–task pair.
Requirements:

1. Write naturally, like a scientific database annotation or research note.
2. Preserve task/context meaning; do not invent assays, outcomes, or claims.
3. You may include light mechanistic interpretation, but only as final conclusions.
4. If you mention properties, do not enumerate full descriptor/value lists. At most 1–2 key quantitative cues are acceptable when truly necessary.
5. Describe structural characteristics only when inferable from provided context.
6. Do NOT output SMILES strings.
7. Do NOT output lines starting with "Observed result:" or "Molecular profile:".
8. Do NOT output reasoning traces, scratchpad text, or `<think>` blocks.
9. Return only the final note text.

Dataset: {*source_dataset*}
Task context: {*context_family*} – {*task_name*}
Key observation: {*observed_summary*}
Query context (SMILES hidden): {*query_with_smiles_masked*}

Reference record (factual anchors):
{*structured_target_card*}

To prevent surface-form leakage, the SMILES inside the query is replaced with `[hidden]` before being passed to the model.

Sampling parameters. Generations use temperature 0.5, top-p 1.0, top-k 50, and a 192-token cap, with thinking-mode disabled. The same sampling configuration is used for the 4,092 evaluation pairs and the 10,844 MolCAR-Train pairs, so the structured and natural variants differ only in target wording, not in molecule, task, or ground-truth outcome.

Anchor re-attachment. After generation, the trusted `Observed result` and `Molecular profile` lines from the structured card are appended to the model output. This guarantees that ground-truth task outcomes and physicochemical descriptors are preserved verbatim, while the free-form portion of the document carries the natural-language framing.

D. Implementation Details

D.1. Representation Extraction

We follow the approach of VLM2Vec (Jiang et al., 2024) and Qwen3-VL-Embedding (Li et al., 2026) and extract the hidden state at the end-of-sequence (EOS) token position as the fixed-length embedding. The EOS position is appended to the end of the input sequence. We refer to this as EOS pooling. No additional pooling (mean, CLS) is applied.

D.2. Molecule–Text Alignment Pretraining

All alignment pretraining runs use bidirectional InfoNCE contrastive loss with a fixed temperature $\tau = 0.07$. The molecular query side and the text target side are encoded by the same MLLM backbone through separate forward passes. Representations are ℓ_2-normalized before computing cosine similarity. Experiments were conducted on multi-GPU servers using the same training recipe across compared runs.

All three alignment-pretraining corpora (MolTextNet, KnowMol-100k, ChEBI-20-MM) are split with a single shared script using a 98/1/1 random partition (train/valid/test, seed 42) at the molecule–text-pair level. Following the MolTextNet experimental protocol, we use the randomly sampled MolTextNet-300K subset rather than the full MolTextNet release (Zhu et al., 2025). The Mixed condition then merges the three corpora and applies a global canonical-SMILES deduplication across all splits, so cross-corpus duplicates are removed rather than counted twice. Concrete corpus sizes after these steps are reported in Table 4. The motivation for the Mixed corpus is coverage rather than scale alone: it exposes the model to complementary forms of molecular language, including structural descriptions, functional group annotations, names, definitions, and physicochemical cues. The cross-modal retrieval probe in Section 4.2 of the main paper is evaluated on the standalone MolTextNet test split (2,970 pairs), not the Mixed test split, so the random R@1 baseline is $1/2970 \approx 0.034\%$.

Table 4. Alignment-pretraining corpus sizes (molecule–text pairs). Pre-dedupe rows are each corpus's own 98/1/1 split. Mixed reflects the post-dedupe global merge across the three sources; merging drops 5,041 duplicate pairs at the canonical-SMILES level.

Corpus	Train	Valid	Test
MolTextNet	290,962	2,969	2,970
KnowMol-100k	94,968	969	970
ChEBI-20-MM	30,630	312	314
Mixed	411,709	4,149	4,165

Optimization and adapter configuration. Across all three MLLM backbones, alignment pretraining shares the same training recipe: AdamW with weight decay 0.02, bfloat16 mixed precision, LoRA on both language and vision components, learning rate 1×10^{-4} for the LoRA parameters, query/target max length $384/512$ tokens, image size 448 px, and seed 42. LoRA rank is backbone-specific: Qwen3.5-0.8B uses $r = 4$, whereas Intern-S1-mini and Qwen3-VL-8B use $r = 8$; in all cases $\alpha = 2r$. Training runs for up to 10 epochs with early-stopping patience 3 on validation loss.

D.3. Property Prediction Adaptation

For molecular property prediction (Section 4.1), each backbone is adapted with a lightweight task head on top of LoRA adapters using the same target groups and per-backbone rank/alpha as in alignment pretraining; for the MolTextNet and Mixed conditions, LoRA weights are inherited from the alignment-pretrained checkpoint, while Direct uses a freshly initialized LoRA. Constant across all runs: AdamW with weight decay 0.02, bfloat16 mixed precision, max sequence length 512 tokens, image size 448 px, and up to 100 epochs with early stopping on validation RMSE for regression and validation ROC-AUC for classification. The task head is a two-layer MLP with hidden size 512; regression targets are standardized while classification targets are not. LoRA learning rate $\in \{5 \times 10^{-5}, 1 \times 10^{-4}\}$, head learning rate $\in \{1.5 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, and early-stopping patience $\in \{5, 10, 15\}$. Final results are averaged over 3 random seeds $\{42, 43, 44\}$.

D.4. Continued Alignment

Continued alignment uses the same contrastive objective as alignment pretraining and is initialized from an alignment-pretrained embedding model. For each MolCAR variant we run a parallel continued-alignment pass per backbone: the Structured continued-alignment checkpoint is trained on MolCAR-Train (Structured) and supplies the +Continued Alignment column of Table 6, while the Natural continued-alignment checkpoint is trained on MolCAR-Train (Natural) and supplies the +Continued Alignment column of Table 7; in each case the train and eval target distributions are matched in style. Both passes share the same recipe and differ only in the training corpus: a smaller LoRA learning rate of 5×10^{-5} to preserve the general molecular semantics acquired during alignment pretraining, and a fixed budget of 1000 optimizer steps.

E. Additional Results

E.1. Full MolTextNet Retrieval Results

Table 5. Molecule-text retrieval on the held-out MolTextNet `test` split. m2t denotes molecule-to-text retrieval, t2m text-to-molecule retrieval, and R@K denotes recall at K. All values are reported as fractions.

Model	m2t R@1	m2t R@5	m2t R@10	t2m R@1	t2m R@5	t2m R@10
Qwen3-VL-Embedding-8B	0.0347	0.0838	0.1185	0.0306	0.0828	0.1229
Qwen3.5-0.8B + MolEmb (MolTextNet)	0.8397	0.9818	0.9963	0.8458	0.9805	0.9926
Intern-S1-mini + MolEmb (MolTextNet)	0.8539	0.9811	0.9926	0.8488	0.9785	0.9923
Qwen3-VL-8B + MolEmb (MolTextNet)	0.7343	0.9549	0.9842	0.7502	0.9572	0.9828

This contrast indicates that a strong generic multimodal embedding model does not automatically provide a chemically aligned molecule–text space; molecule-specific contrastive alignment is needed to make molecular profiles and textual descriptions comparable.

E.2. Full MolCAR Results

MolCAR-Natural replaces structured target cards with scientist-facing notes synthesized by Intern-S1 (235B parameters), a large scientific-domain MLLM distinct from the MolEmb backbones evaluated in this paper. Each note is anchored to the same ground-truth outcome and molecular profile as the corresponding structured card, so the molecule, task, and observed result are held fixed while only the target text style varies. Tables 6–7 report the same Inst/NoInst/Δ evaluation protocol; both candidate pools contain 4,092 documents over 1,653 held-out molecules.

MolCAR-Natural tests whether the context-aware signal persists when the target text distribution is broadened from fixed-schema cards to generated scientific notes while preserving molecule identity, task label, and observed result. The full MolCAR metrics support the same embedding-level reading as the main text. Before continued alignment, instructed and generic queries have similar Context R@1, indicating that generic molecule–text alignment mostly learns an unconditional molecule–description geometry rather than separating the same molecule by task lens. Continued alignment changes this behavior: instructed Context R@1 increases sharply, while NoInst Context R@1 remains at the context-level baseline. In the notation of Appendix B, the observed context-aware embedding family $\mathcal{Z}(m)$ is nearly collapsed after generic alignment and expands along task-meaningful directions only after task-labeled MolCAR supervision.

Table 6. MolCAR-Structured results. **Base**: alignment-pretrained on the Mixed corpus. **+ Continued Alignment**: after continued alignment on MolCAR-Train. **Inst**: query with task-specific instruction; **NoInst**: generic query without task specification. Δ = Inst − NoInst (percentage points). **Mol. R@1**: correct molecule in top-1. **Doc. R@1**: correct document (molecule + context) in top-1. **Doc. MRR**: 100× mean reciprocal rank of the correct document. **Doc. R@5**: correct document in top-5. **Context R@1**: correct context ranked first *within* the same molecule.

Model	Metric	Base			+ Continued Alignment		
		Inst (%)	NoInst (%)	Δ (pp)	Inst (%)	NoInst (%)	Δ (pp)
Intern-S1-mini	Mol. R@1	10.5	10.6	-0.1	24.9	16.0	+8.9
	Doc. R@1	4.5	4.3	+0.2	24.8	6.3	+18.5
	Doc. MRR	11.3	10.7	+0.6	39.8	14.6	+25.2
	Doc. R@5	16.6	15.3	+1.3	57.4	21.1	+36.3
	Context R@1	44.4	40.4	+4.0	**99.8**	40.4	**+59.4**
Qwen3.5-0.8B	Mol. R@1	4.1	4.3	-0.2	12.4	10.6	+1.8
	Doc. R@1	1.8	1.7	+0.1	7.8	4.0	+3.8
	Doc. MRR	5.0	5.0	+0.0	17.4	9.2	+8.2
	Doc. R@5	6.5	6.5	+0.0	25.2	12.8	+12.4
	Context R@1	41.1	40.4	+0.7	**69.4**	40.4	**+29.0**

Table 7. MolCAR-Natural results. We report Inst / NoInst / Δ for Base and +Continued Alignment. Δ is computed from unrounded scores before display; small discrepancies are due to rounding.

Model	Metric	Base			+ Continued Alignment		
		Inst (%)	NoInst (%)	Δ (pp)	Inst (%)	NoInst (%)	Δ (pp)
Intern-S1-mini	Mol. R@1	14.4	14.4	-0.1	22.9	18.1	+4.8
	Doc. R@1	6.2	6.0	+0.2	22.8	7.5	+15.3
	Doc. MRR	14.5	13.9	+0.6	38.9	17.1	+21.8
	Doc. R@5	21.4	20.1	+1.3	57.6	25.0	+32.6
	Context R@1	44.3	40.4	+3.9	**99.2**	40.4	**+58.8**
Qwen3.5-0.8B	Mol. R@1	9.7	9.4	+0.3	21.8	16.9	+4.9
	Doc. R@1	3.9	3.7	+0.2	17.1	6.4	+10.7
	Doc. MRR	9.7	9.6	+0.1	29.6	13.6	+16.0
	Doc. R@5	14.0	14.1	-0.1	42.5	19.1	+23.4
	Context R@1	40.6	40.4	+0.2	**72.0**	40.4	**+31.6**

This pattern is consistent with the structure of the alignment supervision: as Appendix F illustrates, the descriptions in MolTextNet, KnowMol-100k, and ChEBI-20-MM are templated and narrow in linguistic distribution, so during alignment pretraining the model rarely sees the same molecule under varying task lenses. Together, these results support an embedding-level reading of context-aware molecular embedding as primarily a data property of the supervision, induced here by task-diverse, outcome-grounded molecule–text examples; the geometric view in Appendix E.4 is consistent with the same reading.

E.3. MolTextNet Retrieval Preservation after MolCAR Continued Alignment

Figure 2 compares MolTextNet m2t/t2m R@1 before and after MolCAR continued alignment for both backbones, with the MolTextNet-only checkpoint included as a reference point. Starting from the Mixed-aligned checkpoint, continued alignment on MolCAR-Train improves rather than degrades MolTextNet retrieval, indicating that the context-aware embedding capability discussed in Section 4.3 is added on top of, rather than at the expense of, the general molecule–text alignment acquired during alignment pretraining.

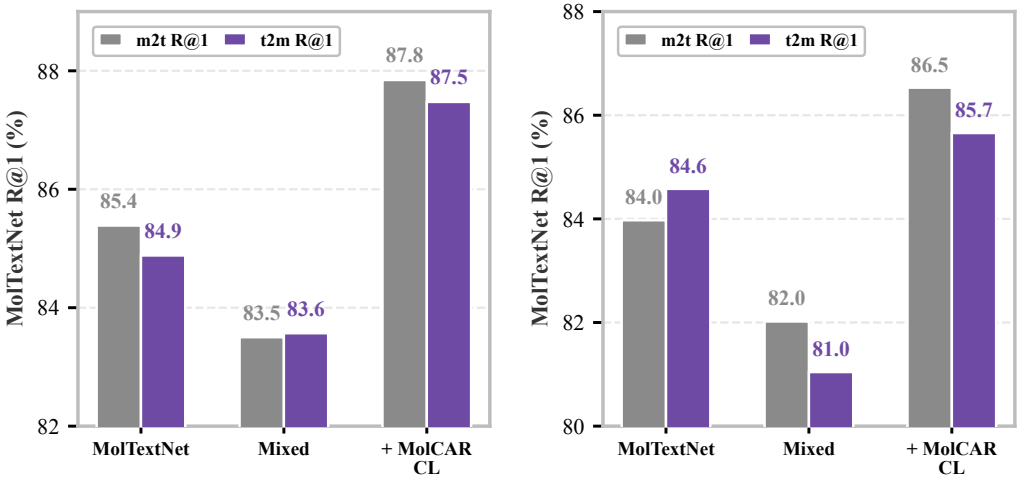

Figure 2. MolTextNet retrieval remains strong after MolCAR continued alignment, for both Intern-S1-mini (left) and Qwen3.5-0.8B (right). Bars compare m2t/t2m R@1 on the held-out MolTextNet test split for the alignment-pretrained (Mixed) checkpoint and the same checkpoint after MolCAR continued alignment; the MolTextNet-only checkpoint is shown as a reference point. Here, CL denotes continued alignment.

E.4. Controlled Instruction Probe t-SNE

To complement the retrieval-table evidence in Section 4.3, we visualize how task instructions reshape the embedding space under a controlled instruction-probe. Starting from a fixed molecule set, each unique molecule is expanded into all eight MolCAR task instructions, so changes in the resulting geometry reflect instruction conditioning rather than differences in sample composition. We then embed all (molecule, instruction) pairs through the embedding model and project them with t-SNE for visualization.

Figure 3 shows the result for Intern-S1-mini in four panels: before continued alignment with generic queries (NoInst) and with task-instructed queries (Inst), and after continued alignment under the same two query modes. Before continued alignment, the NoInst and Inst panels are visually similar, consistent with the close Inst/NoInst Context R@1 scores reported in Table 6. After continued alignment, the NoInst panel remains comparatively overlapped, while the Inst panel shows clear task-wise separation. The geometric view, therefore, tracks the quantitative result: continued alignment changes how the embedding model uses task instructions, not just how strongly it retrieves any single description. This is the same data-property reading as Appendix E.2: task-wise structure in $\mathcal{Z}(m)$ emerges only when the supervision signal varies c.

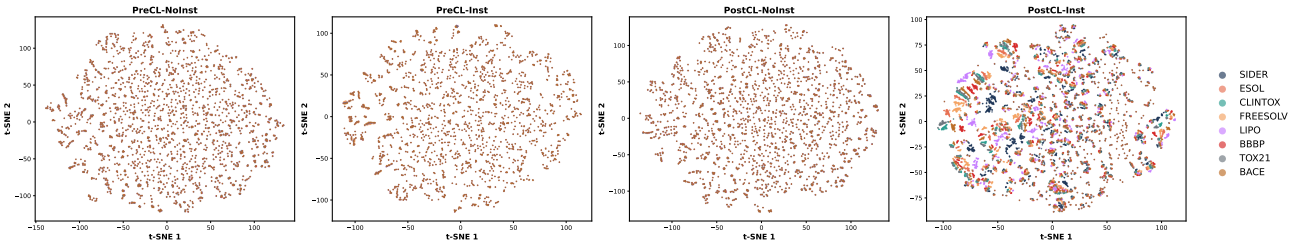

Figure 3. Controlled MolCAR instruction-probe t-SNE on a fixed molecule set for Intern-S1-mini. Each unique molecule is expanded into all eight MolCAR task instructions, so changes in geometry reflect instruction conditioning rather than sample-composition drift. The first two panels show the model before continued alignment, and the last two panels show the model after continued alignment; each setting is evaluated with generic queries (NoInst) and task-instructed queries (Inst). Clear task-wise separation appears only after continued alignment and only when task instructions are present. PreCL and PostCL denote before and after continued alignment, respectively.

E.5. Multi-view Input Ablation

Figure 4 shows an ablation comparing the full multimodal profile (2D depiction + SMILES) against a SMILES-only baseline for Intern-S1-mini under direct adaptation. The 2D depiction contributes consistent though modest gains across most regression and classification tasks, supporting the multi-view input design in Section 3.

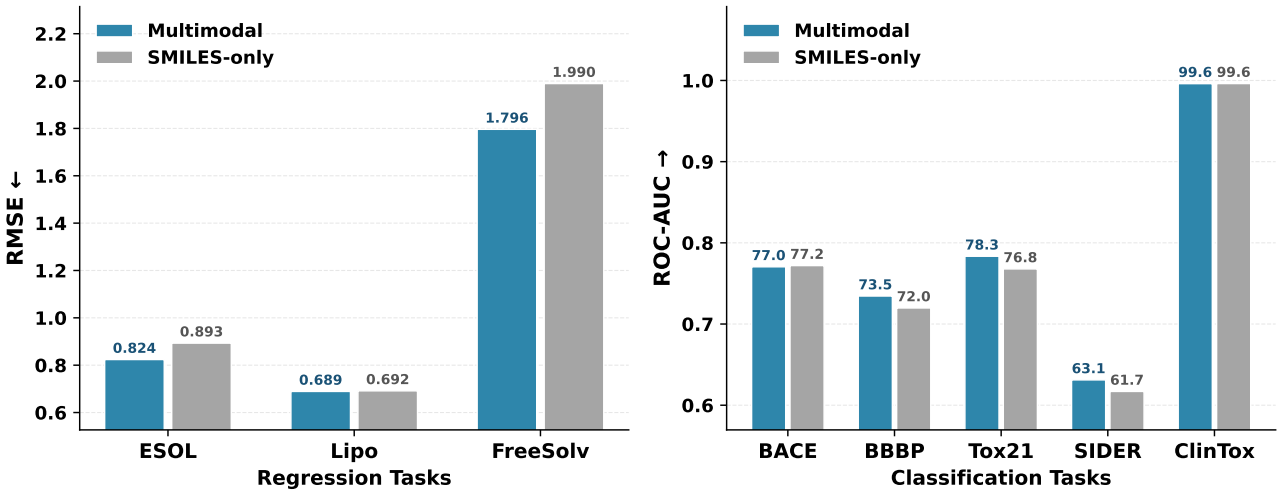

Figure 4. Multimodal profile versus SMILES-only input for Intern-S1-mini under direct adaptation. Using both the 2D depiction and canonical SMILES yields modest gains on most regression and classification tasks, supporting the value of the multi-view input interface.

F. Examples of Pretraining Corpora

We show representative full-text supervision instances from the alignment pretraining corpora used in this work. Each example preserves the multimodal molecule-side query used during training, including an explicit image token, together with the full text-side target.

MolTextNet Example

Molecule query.
<image>
Encode this molecule for reusable molecular understanding.
Focus on its structure, substructures, functional groups, and physicochemical cues.
SMILES: CCCCCCOc1cc(NC(=O)c2ccc(OC)cc2)ccc1N(C)S(C)(=O)=O

Text target.
presents a complex molecular architecture characterized by a molecular weight of 434.56 and a molecular formula of $C_{22}H_{30}N_2O_5S$. The compound exhibits a calculated logP value of 4.3, indicating moderate hydrophobicity, which may influence its membrane permeability and bioavailability. The presence of five hydrogen bond acceptors and one hydrogen bond donor suggests potential for significant molecular interactions with biological targets, while the polar surface area of 84.94 supports solubility in aqueous environments. The compound does not violate Lipinski's rule of five, suggesting favorable drug-like properties. With 11 rotatable bonds, the compound's flexibility could enhance its ability to adopt various conformations, potentially optimizing binding interactions. The structural composition includes three aromatic rings and various functional groups such as an amide, ether, sulfonamide, and a secondary amine, which may contribute to the compound's overall bioactivity. However, the antiproliferative activity against human SKBR3 cells shows an IC50 value of 19980 nM, indicating a lack of significant activity. The synthetic accessibility scores, with an SCS score of 3.8192 and SAS score of 2.0954, reflect a moderate level of synthetic complexity, suggesting that while the compound may be feasible to synthesize, it could present challenges due to its multiple functional groups and overall structural intricacies. The presence of a hexyloxy group may enhance lipophilicity, whereas the sulfonamide moiety could impart additional interactions with biological targets.

KnowMol-100k Example

Molecule query.
<image>
Encode this molecule for reusable molecular understanding.
Focus on its structure, substructures, functional groups, and physicochemical cues.
SMILES: CCC(C)CC(C)C=C(C)C1OC(c2c(O)c(C3C(O)C(O)C(O)C3O)cn(C)c2=O)CC=C1C

Text target.
IUPAC name: 3-[6-[(E)-4,6-dimethyloct-2-en-2-yl]-5-methyl-3,6-dihydro-2H-pyran-2-yl]-4-hydroxy-1-methyl-5-(2,3,4,5-tetrahydroxycyclopentyl)pyridin-2-one. Functional groups: Alkyl, Alkenyl, Phenyl, Hydroxyl, Ether, Pyridyl. The molecule consists of several distinct substructures and functional groups. The main chain is an alkenyl chain with multiple methyl groups attached, forming a branched structure. This chain is connected to a dihydropyran ring, which is a six-membered ring containing one oxygen atom. Attached to the dihydropyran ring is a pyridine ring, which is a six-membered ring containing one nitrogen atom. The pyridine ring has a hydroxyl group and a methoxy group attached to it. Additionally, there is a tetrahydroxycyclopentyl group attached to the pyridine ring, which is a five-membered ring with four hydroxyl groups. The connections between these substructures are as follows: the alkenyl chain is connected to the dihydropyran ring, which is then connected to the pyridine ring. The tetrahydroxycyclopentyl group is attached to the pyridine ring. The molecule contains alkyl, alkenyl, hydroxyl, ether, and pyridyl functional groups. The molecule exhibits moderate polarity due to the presence of multiple hydroxyl groups and a pyridine ring, which introduce polar regions, although the overall structure is somewhat balanced by nonpolar alkyl chains. The presence of hydroxyl groups and the pyridine ring suggests that the molecule has both acidic and basic sites, with the hydroxyl groups contributing to acidity and the nitrogen in the pyridine ring contributing to basicity. The molecule is likely to be soluble in polar solvents like water due to the numerous hydroxyl groups, but the nonpolar alkyl chains may also allow some solubility in nonpolar solvents. Reactivity is influenced by the hydroxyl groups, which can participate in hydrogen bonding and nucleophilic reactions, and the double bond in the alkenyl chain, which can undergo addition reactions. The molecule has stereochemical complexity due to the presence of multiple chiral centers, particularly in the tetrahydroxycyclopentyl group, leading to potential enantiomers and diastereomers. Electrophilicity is enhanced by the electron-withdrawing effects of the carbonyl and hydroxyl groups, making certain sites more susceptible to nucleophilic attack.

ChEBI-20-MM Example

Molecule query.
<image>
Encode this molecule for reusable molecular understanding.
Focus on its structure, substructures, functional groups, and physicochemical cues.

SMILES: COc1cc2c(=O)c(-c3ccc(O[C@@H]4O[C@H](C(=O)O)[C@@H](O)[C@H](O)[C@H]4O)cc3)coc2cc1O

Text target.
The molecule is a glycosyloxyisoflavone that is the glucuronide-conjugated form of the phytoestrogen glycitein. It is a 7-hydroxyisoflavone, a methoxyisoflavone, a glycosyloxyisoflavone, a beta-D-glucosiduronic acid and a monosaccharide derivative. It derives from a glycitein. IUPAC name: (2S,3S,4S,5R,6S)-3,4,5-trihydroxy-6-[4-(7-hydroxy-6-methoxy-4-oxochromen-3-yl)phenoxy]oxane-2-carboxylic acid. Physicochemical cues: XlogP 0.9 (moderate lipophilicity); polar surface area 172.0 A^2 (high polar surface area).

G. MolCAR Example

This example makes the MolCAR construction concrete on a single held-out molecule. The structured variant presents multiple candidate targets for the same molecular profile under different task lenses, while the natural variant replaces those same targets with scientist-facing notes anchored to the same observed outcomes and descriptors. The instruction query and the generic no-instruction query are both shown because MolCAR evaluates whether task language, rather than molecule identity alone, routes the query to the correct same-molecule target.

MolCAR-Structured Example

Shared molecule.
<image>
SMILES: BrC(Br)Br

Instruction query.
Encode this molecule for blood-brain barrier penetration.
SMILES: BrC(Br)Br

No-instruction query.
Encode this molecule.
SMILES: BrC(Br)Br

Candidate target cards.
Doc. A
Report family: partition and barrier transport
Task: blood-brain barrier penetration
Observed result: This molecule crosses the blood-brain barrier.
Molecular profile: MW 252.7; cLogP 2.45; HBD 0; HBA 0; formal charge 0

Doc. B
Report family: aqueous interaction
Task: water solubility
Observed result: Observed water solubility is -1.910 log mol/L.
Molecular profile: MW 252.7; cLogP 2.45; HBD 0; HBA 0; formal charge 0

Doc. C
Report family: aqueous interaction
Task: hydration free energy
Observed result: Observed hydration free energy is -2.130 kcal/mol.
Molecular profile: MW 252.7; cLogP 2.45; HBD 0; HBA 0; formal charge 0

Doc. D
Report family: safety liability
Task: Tox21 toxicity outcomes
Observed result: This molecule is positive in 0 of 12 observed Tox21 assay panel.
Molecular profile: MW 252.7; cLogP 2.45; HBD 0; HBA 0; formal charge 0

MolCAR-Natural Example

Shared molecule.
<image>
SMILES: BrC(Br)Br

Instruction query.
Encode this molecule for blood-brain barrier penetration.
SMILES: BrC(Br)Br

Candidate target notes.
Doc. A
The molecule demonstrates effective blood-brain barrier penetration, suggesting favorable physicochemical properties for central nervous system targeting. With a molecular weight of 252.7 and moderate lipophilicity (cLogP 2.45), the compound lacks hydrogen bond donors or acceptors, potentially enhancing its passive diffusion across the BBB. These characteristics align with its observed ability to cross the blood-brain barrier.

Doc. B
The molecule exhibits moderate water solubility with an observed value of -1.910 log mol/L. Its physicochemical profile suggests a neutral, lipophilic character with a molecular weight of 252.7 and a calculated logP of 2.45, lacking hydrogen bond donors or acceptors. These properties likely contribute to its solubility behavior in aqueous environments.

Doc. C
The compound exhibits a measured hydration free energy of -2.130 kcal/mol, indicating moderate hydrophilic character despite its neutral charge and lack of hydrogen bond donors or acceptors. With a molecular weight of 252.7 and a cLogP of 2.45, the molecule demonstrates balanced lipophilicity and aqueous solubility properties. These features suggest limited polar interactions with water, consistent with its structural profile.

Doc. D
The molecule exhibits no positive responses across the 12 Tox21 assay panel, suggesting a low potential for toxicity in the tested endpoints. Its molecular profile, characterized by a moderate molecular weight (252.7) and balanced lipophilicity (cLogP 2.45), indicates favorable physicochemical properties that may contribute to its non-toxic profile. The absence of hydrogen bond donors and acceptors, along with a neutral formal charge, further supports its stability and reduced reactivity in biological systems.

H. Limitations

MolEmb is evaluated as an adaptation route rather than as a fully scaled molecular foundation model. Our molecular profiles use a 2D depiction and canonical SMILES, leaving richer views such as conformers, reactions, spectra, and assay metadata for future instantiations of the profile map $\mathcal{P}(m)$. MolCAR is designed as a controlled diagnostic for context-aware embedding: it holds molecule identity fixed and varies task lenses within eight property-related task families, but it does not test transfer to unseen task families or open-ended scientific search. Finally, the results suggest that context-aware molecular embedding depends strongly on the supervision. The Mixed corpus broadens semantic coverage, but its scale and diversity remain small relative to industrial embedding pretraining, so broader context-aware embedding will require larger, more carefully curated molecule–text supervision.