TRIDENT: Tri-Modal Molecular Representation Learning with Taxonomic Annotations and Local Correspondence

Feng Jiang [1] **Mangal Prakash** [2] **Hehuan Ma** [1] **Jianyuan Deng** [2] **Yuzhi Guo** [1] **Amina Mollaysa** [2] **Tommaso Mansi** [2] **Rui Liao** [2] **Junzhou Huang** [1]

Abstract

Molecular property prediction aims to learn representations that map chemical structures to functional properties. While multimodal learning has emerged as a powerful paradigm to learn molecular representations, prior works have largely overlooked textual and taxonomic information of molecules for representation learning. We introduce TRIDENT, a novel framework that integrates molecular SMILES, textual descriptions, and taxonomic functional annotations to learn rich molecular representations. To achieve this, we curate a comprehensive dataset of molecule-text pairs with structured, multi-level functional annotations. Instead of relying on conventional contrastive loss, TRIDENT employs a volume-based alignment objective to jointly align tri-modal features at the global level, enabling soft, geometry-aware alignment across modalities. Additionally, TRIDENT introduces a novel local alignment objective that captures detailed relationships between molecular substructures and their corresponding sub-textual descriptions. A momentum-based mechanism dynamically balances global and local alignment, enabling the model to learn both broad functional semantics and fine-grained structure-function mappings. TRIDENT achieves state-of-the-art performance on 11 downstream tasks, demonstrating the value of combining SMILES, textual, and taxonomic functional annotations for molecular property prediction.

1. Introduction

Molecular representation learning, which converts complex chemical structures into computational features, has been instrumental in advancing various aspects of drug discovery including virtual screening, and molecular design (Liu et al., 2022a; Chibani & Coudert, 2020; Shen & Nicolaou, 2019). Multi-modal molecular models further enhance representation quality by integrating structural, textual, and functional information, enabling better generalization and predictive performance (Liu et al., 2023a). These approaches hold promise for unlocking deeper insights into chemical space and accelerating the discovery of therapeutic compounds with desired properties.

However, current multimodal approaches (Luo et al., 2023; Su et al., 2022) face three key limitations: **(1) Overlooking fine-grained annotations across taxonomies**: Most existing methods simplify the representation of molecules by focusing on unified functional descriptions, neglecting the nuanced annotations provided by different taxonomic systems. The same molecule may have distinct emphases depending on the taxonomy: for example, the LOTUS Tree (Rutz et al., 2022) taxonomy highlights natural product classifications, whereas the MeSH (Medical Subject Headings) Tree (Lipscomb, 2000) taxonomy emphasizes medical functionalities of the same molecule. Ignoring these taxonomy-specific, fine-grained annotations risks reducing molecules to flat entities, thereby failing to capture the multi-faceted and structured nature of chemical functions. **(2) Alignment limitations**: Aligning modalities such as molecular structures, textual descriptions, and taxonomic functional annotations is inherently complex. Existing methods rely on pairwise alignment schemes anchored to a single modality, which struggle to model the interdependencies across all modalities (Zhu et al., 2023; Liu et al., 2022b; Xu et al., 2023; Chen et al., 2023), particularly when one modality encodes nested or multi-level information (Cicchetti et al., 2024). **(3) Neglect of local correspondences**: Many approaches focus exclusively on molecule-level alignment, disregarding the fine-grained relationships between molecular substructures (e.g., functional groups) and their corresponding sub-textual descriptions. This omission limits the expressivity of the

[1] University of Texas at Arlington, Arlington, TX, United States. [2] Johnson & Johnson Innovative Medicine, New Brunswick, NJ, United States. Correspondence to: Feng Jiang <fxj8843@mavs.uta.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

learned representations and constrains their applicability in molecular property prediction tasks.

To address these limitations, we introduce the TRIDENT (Tri-modal Representation Integrating Descriptions, Entities, and Taxonomies) framework for molecules that jointly models molecular SMILES, textual descriptions, and multi-faceted Hierarchical Taxonomic Annotation (HTA). Central to TRIDENT is the HTA modality, which organizes molecular function across hierarchical classification levels. We curate a high quality dataset of 47,269 <*SMILES, Text, HTA*> triplets from PubChem (Kim et al., 2016), annotated under 32 classification systems. To tackle the challenge of aligning these diverse modalities, TRIDENT leverages a volume-based contrastive loss, enabling soft, geometry-aware alignment of all three modalities. While recently proposed for general-purpose modality alignment (Cicchetti et al., 2024), we extend this formulation to the molecular domain for the first time, where the modalities are structurally diverse and include taxonomic semantic labels. Furthermore, TRIDENT introduces a novel local alignment module that links molecular substructures to their associated sub-textual descriptions, capturing fine-grained structure–function relationships. A momentum-based balancing mechanism dynamically integrates global and local alignments to optimize the representation learning process (see Figure 1 for an overview).

We demonstrate that TRIDENT achieves consistent and substantial improvements over existing molecular representation learning methods. Our framework sets a new benchmark, delivering state-of-the-art performance across 11 downstream molecular property prediction tasks on established benchmarks, while remaining modular and flexible, allowing integration of different modality encoders without the need for architectural modifications. We have also created a high-quality, comprehensive dataset of molecule-text-function triplets, which forms the foundation for this work and future research. To summarize, we make the following contributions:

- Introducing a Hierarchical Taxonomic Annotation (HTA) modality for molecules, supported by a newly curated high-quality multimodal dataset consisting of 47,269 <*SMILES, Text, HTA*> triplets annotated across 32 diverse taxonomic classification systems. This enables a structured, multi-level functional understanding of molecules, providing a novel resource for molecular representation learning.

- A unified global–local alignment strategy that integrates a volume-based contrastive loss for tri-modal global alignment with a novel local alignment module for substructure–subtext correspondence, dynamically balanced via a momentum-based mechanism.

- Demonstrated state-of-the-art performance across 11 molecular property prediction tasks, validating the effectiveness of hierarchical taxonomic annotations as a modality, the proposed alignment strategies, and the quality of the curated dataset.

2. Method

In this section, we provide a detailed introduction to the implementation of the TRIDENT framework, as illustrated in Figure 1 which addresses the shortcomings of existing methods in capturing a structured understanding of molecular functions across different hierarchical functional categories.

2.1. Hierarchical Taxonomic Annotation (HTA)

Traditional PubChem descriptions provide narrow functional annotations lacking broader biological and industrial context. We construct a dataset of 47,269 <*SMILES, Text, HTA*> triplets by mapping molecules across 32 hierarchical classification systems (e.g., LOTUS Tree, MeSH Tree), expanding flat descriptors into multi-domain semantic representations. HTAs are generated through GPT-4o synthesis of structured annotations, encoding cross-domain knowledge including chemical derivation, natural sources, functional applications, and regulatory associations. This process is validated by domain experts to ensure factual accuracy and captures complementary information to traditional annotations. Detailed data collection and processing procedures are provided in the Appendix.

2.2. Geometry-based Global Alignment

We aim to learn meaningful multimodal representations by jointly modeling three data modalities: molecule SMILES (M), textual descriptions (T), and HTA (H). SMILES representations utilize the encoder E_m, while both textual descriptions (T) and HTA (H) share a common text encoder E_t.

Traditional multimodal approaches typically rely on pairwise similarity metrics such as cosine similarity: $\cos(\theta_{ij}) = \frac{\langle M_i, M_j \rangle}{\|M_i\| \cdot \|M_j\|}$. However, these methods often anchor one modality and align others to it independently, failing to capture higher-order relationships across all modalities. To address this, GRAM (Cicchetti et al., 2024) introduced a geometry-based alignment approach that uses the volume of the parallelotope spanned by modality vectors as a global measure of alignment. Specifically, for three normalized embeddings m, t, and h, the volume of the parallelotope is computed as $\mathrm{Vol}(m, t, h) = \sqrt{1 - \langle m, t \rangle^2 - \langle m, h \rangle^2 - \langle t, h \rangle^2 + 2\langle m, t \rangle \langle t, h \rangle \langle h, m \rangle}$, which reflects the overall geometric alignment of the embeddings. The volume shrinks as the modalities converge and grows as they diverge. Unlike pairwise

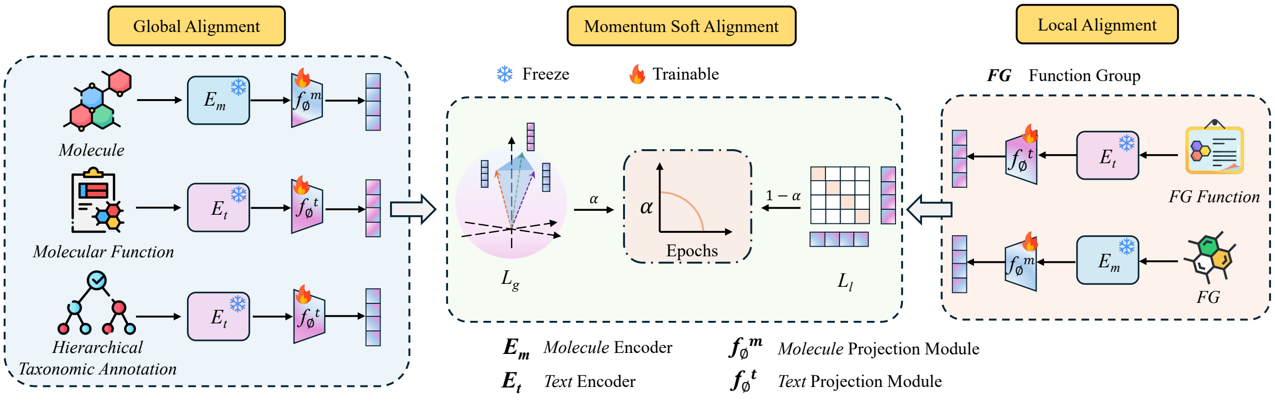


Figure 1. **Overview of TRIDENT.** TRIDENT jointly models molecular SMILES, natural language descriptions, and Hierarchical Taxonomic Annotations (HTAs) to learn rich molecular representations. The framework employs a volume-based contrastive loss for soft global tri-modal alignment and a local alignment module that links molecular substructures to sub-text spans. A momentum-based mechanism dynamically balances the contribution of global and local objectives during training. This multimodal, multi-level alignment enables precise and semantically grounded molecular understanding.

contrastive learning methods, this formulation was shown to capture the global structure of cross-modal interactions in a principled and scalable way for audio-video-text pairs (Cicchetti et al., 2024).

Global Volume-based Contrastive Loss. Following the approach introduced in GRAM (Cicchetti et al., 2024), we construct a *global contrastive objective* over the three modalities—SMILES, traditional text descriptions, and HTA annotations. Each modality is processed through a modality-specific encoder followed by a modality-specific projection head (implemented as a three-layer MLP) to map the embeddings into a shared latent space, yielding embeddings m, t, and h for SMILES, text, and HTA, respectively.

To holistically align the three modalities, we compute the volume of the parallelotope formed by the triplet of unit-normalized vectors (m, t, h). We define a *bidirectional global contrastive loss* that captures two complementary retrieval directions. In the first direction, denoted $\mathcal{L}_{\text{M2TH}}$, the model is trained to retrieve the correct semantic context—comprising both textual and taxonomic annotations—given a molecular embedding. That is, given m_i, the loss encourages the volume $\text{Vol}(m_i, t_i, h_i)$ to be smaller than volume spanned by any mismatched triplets (m_i, t_j, h_j) for $j \neq i$:

$$\mathcal{L}_{\text{M2TH}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(-\text{Vol}(m_i, t_i, h_i)/\tau)}{\sum_{j=1}^{B} \exp(-\text{Vol}(m_i, t_j, h_j)/\tau)},$$

where B is the batch size and τ is a learnable temperature parameter.

Conversely, the second direction, $\mathcal{L}_{\text{TH2M}}$, considers the retrieval of the correct molecule given the semantic context. Here, the volume of the correct triplet (m_i, t_i, h_i) is minimized relative to all volumes spanned by mismatched triples (m_j, t_i, h_i):

$$\mathcal{L}_{\text{TH2M}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(-\text{Vol}(m_i, t_i, h_i)/\tau)}{\sum_{j=1}^{B} \exp(-\text{Vol}(m_j, t_i, h_i)/\tau)}.$$

The final loss averages both directions to ensure mutual semantic alignment of all three modalities:

$$\mathcal{L}_{\text{g}} = \frac{1}{2}(\mathcal{L}_{\text{M2TH}} + \mathcal{L}_{\text{TH2M}}).$$

This bidirectional formulation encourages robust triadic alignment, capturing global structure across modalities more effectively than traditional pairwise contrastive losses.

2.3. Fine-grained Local Alignment

Although multimodal fusion effectively learns cross-modal information (Jiang et al., 2024a;b), some existing fusion methods only consider global alignment while neglecting fine-grained alignment, which may lead to suboptimal performance in capturing detailed cross-modal relationships (Dang et al., 2024; 2025; Li et al., 2023; Wang et al., 2024). While the global alignment captures the overall semantic relationships among modality embeddings, it may overlook fine-grained correspondences between molecular functional sub-groups and their sub-textual descriptions. For instance, local features such as aromatic rings, hydroxyl groups, or aliphatic chains often correspond to specific phrases in molecular descriptions or to fine-level taxonomic labels.

To address this limitation, we introduce a *local alignment contrastive loss* that complements the global volume-based

Table 1. Performance comparison on molecule property prediction. We present the ROC-AUC(%) scores of the molecular property prediction task on MoleculeNet. For baselines that report results, we directly use their reported outcomes. Note that MolCA-SMILES does not report results for the MUV and HIV datasets. The best results are marked in **bold**, and the second-best are underlined.

Method	BBBP	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	Bace	Avg
MOLFORMER	70.74±1.34	74.74±0.56	65.51±0.63	61.75±1.23	77.64±0.98	67.58±1.01	75.64±1.76	78.64±2.35	71.53
KV-PLM	70.50±0.54	72.12±1.02	55.03±1.65	59.83±0.56	89.17±2.73	54.63±4.81	65.40±1.69	75.80±2.73	67.81
MegaMolBART	68.89±0.17	73.89±0.67	63.32±0.79	59.52±1.79	78.12±4.62	61.51±2.75	71.04±1.70	82.46±0.84	69.84
MoleculeSTM-SMILES	70.75±1.90	75.71±0.89	65.17±0.37	63.70±0.81	86.60±2.28	65.69±1.46	77.02±0.44	81.99±0.41	73.33
MolFM	72.90±0.10	77.20±0.70	64.40±0.20	64.20±0.90	79.70±1.60	76.00±0.80	78.80±1.10	83.90±1.10	74.64
MoMu	70.50±2.00	75.60±0.30	63.40±0.50	60.50±0.90	79.90±4.10	70.50±1.40	75.90±0.80	76.70±2.10	71.63
Atomas	73.72±1.67	77.88±0.36	66.94±0.90	64.40±1.90	93.16±0.50	76.30±0.70	80.55±0.43	83.14±1.71	77.01
MolCA-SMILES	70.80±0.60	76.00±0.50	56.20±0.70	61.10±1.20	89.00±1.70	-	-	79.30±0.80	72.10
TRIDENT (M-S)	73.14±0.44	78.23±0.12	67.79±0.56	**64.62±0.47**	**95.75±0.71**	82.88±1.41	79.64±1.15	**84.19±0.95**	78.3
TRIDENT (M-M)	**73.95±1.01**	**79.36±0.13**	**67.80±0.37**	63.64±0.56	95.41±0.66	**83.51±0.48**	**81.63±0.52**	82.39±0.56	**78.5**

Table 2. Performance of different methods on DILI, Carcinogens, and Skin Reaction tasks, reporting AUC and Accuracy. The best results are marked in **bold**, and the second-best are underlined.

Method	**DILI (475 drugs)**		**Carcinogens (278 drugs)**		**Skin Reaction (404 drugs)**	
	AUC	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**
MOLFORMER	85.59±1.39	76.39±5.24	77.27±0.76	77.32±1.47	63.75±1.41	60.98±3.44
KV-PLM	73.46±0.61	62.50±2.08	75.18±3.71	76.01±1.75	62.88±2.30	59.76±5.17
MolT5	77.37±1.15	69.44±1.20	86.89±1.00	84.45±1.11	68.67±3.99	62.22±1.41
MoMu	80.44±2.47	75.00±4.17	80.11±1.50	78.00±2.62	61.63±1.94	56.10±3.45
MolCA-SMILES	88.34±1.28	80.56±2.40	82.00±1.80	78.76±0.52	65.13±0.88	62.20±1.72
MoleculeSTM-SMILES	91.20±2.02	84.72±2.41	83.87±1.30	81.05±0.63	67.72±0.50	61.60±0.73
Atomas	90.17±1.30	85.08±2.16	82.47±2.11	80.75±0.50	70.33±0.88	61.79±6.14
TRIDENT (M-S)	**95.08±0.70**	**86.81±2.40**	83.42±1.10	81.47±0.92	70.33±0.63	**63.42±4.22**
TRIDENT (M-M)	94.56±0.88	86.80±3.18	**87.07±0.77**	**84.62±1.07**	**72.00±1.09**	62.60±1.40

objective. Unlike GRAM (Cicchetti et al., 2024), which operates solely at the level of full modality embeddings, our method leverages the compositional nature of molecules to align substructures with their semantic counterparts in text and taxonomy.

By decomposing each molecule into interpretable substructures and anchoring them to matched textual or taxonomic segments, we encourage the model to learn fine-grained correspondences across modalities. This local supervision enforces semantic consistency not only at the global level but also within the internal structure of molecular representations.

2.3.1. Functional Group-Level Representation

We construct a dataset linking 85 functional groups with semantic descriptions through GPT-4o (Achiam et al., 2023) generation and expert review. Using RDKit, we extract functional groups from molecules and encode them through modality-specific encoders, obtaining structural embeddings $fg_1, fg_2, \ldots, fg_k$ and textual embeddings $fgt_1, fgt_2, \ldots, fgt_k$. Consolidated representations use max-pooling: $fg_{\text{pooled}} = \text{Pool}(fg_1, fg_2, \ldots, fg_k)$, $fgt_{\text{pooled}} = \text{Pool}(fgt_1, fgt_2, \ldots, fgt_k)$

2.3.2. Local Alignment Loss

Using these pooled representations, we define our bidirectional local alignment contrastive loss as follows.

$$\mathcal{L}_{FG2T} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(fg_{pooled,i} \cdot fgt_{pooled,i}/\tau)}{\sum_{j=1}^{B} \exp(fg_{pooled,i} \cdot fgt_{pooled,j}/\tau)},$$

$$\mathcal{L}_{T2FG} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(fg_{pooled,i} \cdot fgt_{pooled,i}/\tau)}{\sum_{j=1}^{B} \exp(fg_{pooled,j} \cdot fgt_{pooled,i}/\tau)},$$

$$\mathcal{L}_{\text{l}} = \frac{1}{2}(\mathcal{L}_{FG2T} + \mathcal{L}_{T2FG}),$$

where B is the batch size and τ is the temperature parameter. The bidirectional loss ensures mutual semantic grounding: $\mathcal{L}_{\text{FG2T}}$ retrieves descriptions from functional group embeddings, while $\mathcal{L}_{\text{T2FG}}$ recovers structures from text descriptions. This dual supervision encourages chemically meaningful substructure embeddings while associating them with precise textual counterparts.

2.4. Momentum-based Integration

To effectively integrate global and local alignments, we adopt a momentum-based approach that dynamically adjusts the importance of each alignment component:$L = \alpha\mathcal{L}_{\text{g}} + (1-\alpha)\mathcal{L}_{\text{l}}$, where α is a momentum coefficient that balances global and local alignments. Instead of using a fixed α, we employ an exponential moving average to update it during training: $\alpha_t = \beta\alpha_{t-1} + (1-\beta) \cdot \frac{\mathcal{L}_{\text{g}}^{(t)}}{\mathcal{L}_{\text{g}}^{(t)} + \mathcal{L}_{\text{l}}^{(t)}}$, where β is a momentum parameter (0.9), and $\mathcal{L}_{\text{g}}^{(t)}$ and $\mathcal{L}_{\text{l}}^{(t)}$ are the respective loss values at training step t. This dynamic adjustment ensures that the model focuses more on the alignment component that currently has higher loss, effectively addressing the most pressing alignment challenges at each training stage.

3. Experiments

We evaluate TRIDENT on 11 molecular property prediction tasks from MoleculeNet (Wu et al., 2018) and Therapeutics Data Commons (TDC) (Huang et al., 2021) benchmarks, comparing against recent state-of-the-art baselines including Atomas (Zhang et al., 2025). Additional experimental details, ablation studies, and comprehensive results are provided in the Appendix.

As shown in Table 1, TRIDENT achieves state-of-the-art performance across MoleculeNet tasks, substantially outperforming strong baselines such as Atomas and MolFM while achieving best-in-class performance on multiple challenging benchmarks. Table 2 demonstrates superior performance on TDC datasets, showing robustness and adaptability across different dataset scales and prediction challenges. The superior performance stems from three key innovations: (1) HTA modality provides multi-dimensional, hierarchical molecular annotations that capture richer semantic information than flat functional descriptions used by prior methods; (2) Volume-based global alignment effectively handles the complex interdependencies across three modalities, overcoming limitations of pairwise alignment schemes; (3) Local alignment captures fine-grained substructure-function relationships typically overlooked by molecule-level approaches, while momentum-based integration dynamically balances global and local objectives throughout training.

4. Conclusion

TRIDENT introduces a tri-modal molecular representation framework that unifies SMILES, textual descriptions, and hierarchical taxonomic annotations through geometry-aware volume-based global alignment and fine-grained local substructure correspondence. Trained on 47,269 triplets across 32 taxonomic systems, it achieves state-of-the-art performance on molecular property prediction benchmarks, demonstrating the value of structured, multi-level functional understanding in molecular learning. This work opens new directions for hierarchical, semantically grounded representation learning in chemical sciences.

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.

Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., and Liu, J. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

Chibani, S. and Coudert, F.-X. Machine learning approaches for the prediction of materials properties. *Apl Materials*, 8(8), 2020.

Cicchetti, G., Grassucci, E., Sigillo, L., and Comminiello, D. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*, 2024.

Dang, T. M., Guo, Y., Ma, H., Zhou, Q., Na, S., Gao, J., and Huang, J. Mfmf: Multiple foundation model fusion networks for whole slide image classification. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–8, 2024.

Dang, T. M., Zhou, Q., Guo, Y., Ma, H., Na, S., Dang, T. B., Gao, J., and Huang, J. Abnormality-aware multimodal learning for wsi classification. *Frontiers in Medicine*, 12: 1546452, 2025.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Jiang, F., Guo, Y., Ma, H., Na, S., An, W., Song, B., Han, Y., Gao, J., Wang, T., and Huang, J. Alphaepi: Enhancing b cell epitope prediction with alphafold 3. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–8, 2024a.

Jiang, F., Guo, Y., Ma, H., Na, S., Zhong, W., Han, Y., Wang, T., and Huang, J. Gte: a graph learning framework for prediction of t-cell receptors and epitopes binding specificity. *Briefings in Bioinformatics*, 25(4):bbae343, 2024b.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Li, H., Wang, C., Zhao, G., He, Z., Wang, Y., and Sun, Z. Sclera-transfuse: Fusing swin transformer and cnn for accurate sclera segmentation. In *2023 IEEE International*

Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE, 2023.

Lipscomb, C. E. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

Liu, S., Guo, H., and Tang, J. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022a.

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, 2023a.

Liu, Y., Li, S., Wu, Y., Chen, C.-W., Shan, Y., and Qie, X. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022b.

Liu, Z., Li, S., Luo, Y., Fei, H., Cao, Y., Kawaguchi, K., Wang, X., and Chua, T.-S. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023b.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

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

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Reidenbach, D., Livne, M., Ilango, R. K., Gill, M., and Israeli, J. Improving small molecule generation using mutual information machine. *arXiv preprint arXiv:2208.09016*, 2022.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Ruan, L., Hu, A., Song, Y., Zhang, L., Zheng, S., and Jin, Q. Accommodating audio modality in clip for multimodal processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9641–9649, 2023.

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J. G., Stephan, R., Page, R., Vondrášek, J., et al. The lotus initiative for open knowledge management in natural products research. *elife*, 11:e70780, 2022.

Shen, J. and Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36, 2019.

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.

Wang, C., Li, H., Zhang, Y., Zhao, G., Wang, Y., and Sun, Z. Sclera-transfuse: Fusing vision transformer and cnn for accurate sclera segmentation and recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

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.

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.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pp. 38728–38748. PMLR, 2023.

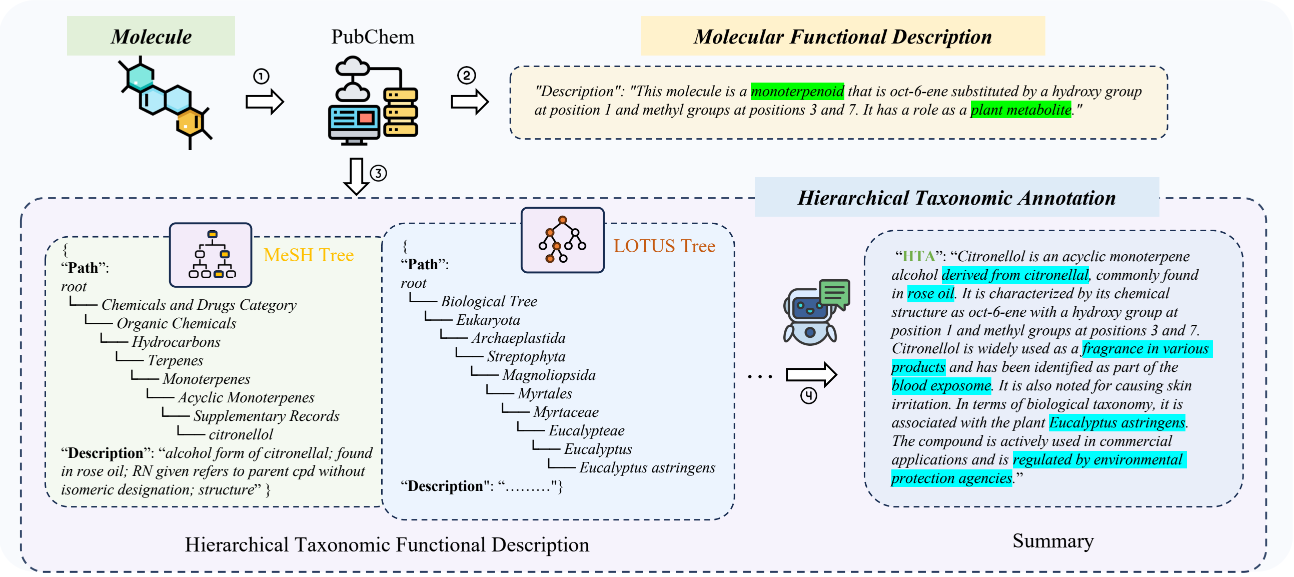
Zeng, Z., Yao, Y., Liu, Z., and Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

Zhang, Y., Ye, G., Yuan, C., Han, B., Huang, L.-K., Yao, J., Liu, W., and Rong, Y. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

A. Hierarchical Taxonomic Annotation (HTA)



"Description": "This molecule is a monoterpenoid that is oct-6-ene substituted by a hydroxy group at position 1 and methyl groups at positions 3 and 7. It has a role as a plant metabolite."

"**Description**": "*alcohol form of citronellal; found in rose oil; RN given refers to parent cpd without isomeric designation; structure*" }

"**HTA**": "*Citronellol is an acyclic monoterpene alcohol derived from citronellal, commonly found in rose oil. It is characterized by its chemical structure as oct-6-ene with a hydroxy group at position 1 and methyl groups at positions 3 and 7. Citronellol is widely used as a fragrance in various products and has been identified as part of the blood exposome. It is also noted for causing skin irritation. In terms of biological taxonomy, it is associated with the plant Eucalyptus astringens. The compound is actively used in commercial applications and is regulated by environmental protection agencies.*"

Figure 2. Traditional molecular functional descriptions are typically obtained by inputting a molecule into PubChem, where a general functional annotation is provided, as shown in Steps 1 and 2 of the figure. To achieve more comprehensive knowledge, functional annotations of the molecule are first obtained under different classification systems, as illustrated in Step 3. Then, these annotations are summarized using GPT-4o, resulting in a higher-quality textual description, as depicted in Step 4. The blue and green highlighted sections illustrate the different perspectives between traditional text and HTA text descriptions.

To enable structured, hierarchical molecular representations, we introduce the Hierarchical Taxonomic Annotation (HTA) framework, which organizes molecular functions across multiple classification levels. This setup allows the model to capture fine-grained, hierarchical semantics essential for understanding complex molecular properties and their biological roles. We curate a high-quality dataset of 47,269 <*SMILES, Text, HTA*> triplets sourced from PubChem (Kim et al., 2016). As shown in Figure 2, these triplets are annotated across 32 diverse hierarchical classification systems, providing a comprehensive, multi-level understanding of molecular behavior. Figure 2 illustrates the construction pipeline for HTA. Beginning with a molecule's SMILES representation, the molecule is queried against PubChem (Kim et al., 2016). This yields a set of traditional functional descriptions, which are typically concise, ontology-aware summaries based on cheminformatics rules. For example, citronellol is described as *a monoterpenoid... with a role as a plant metabolite.* While such descriptors are chemically accurate, they often lack broader context, such as ecological origin, industrial relevance, or toxicological implications.

To address this limitation, we augment the molecule's annotation space through structured taxonomic enrichment by mapping it into multiple biological and chemical taxonomies. For example, the LOTUS Tree (Rutz et al., 2022) highlights natural product classifications, whereas the MeSH (Medical Subject Headings) Tree (Lipscomb, 2000) emphasizes medical functionalities of the same molecule. Through this multi-perspective approach, these hierarchies expand the molecular profile beyond flat descriptors into deeply nested semantic trees spanning chemistry, biology, and pharmacology.

In the final stage, we leverage a GPT-4o (Achiam et al., 2023; Ouyang et al., 2022) to synthesize the retrieved structured annotations into a high-fidelity, human-readable HTA. Unlike traditional descriptors, HTAs encode multi-perspective knowledge: they trace the chemical derivation (e.g., from citronellal), mention natural sources (e.g., rose oil), functional applications (e.g., fragrance in various products), and regulatory or biomedical associations (e.g., environmental protection agencies, blood exposome). This generative synthesis is guided by structural prompts and validated by domain experts to ensure factual accuracy and interoperability.

Crucially, the information content in HTAs is complementary to traditional functional annotations. While the latter provides standardized yet narrow chemical definitions, HTAs integrate cross-domain knowledge that aligns better with how biological and industrial experts interpret molecular function. The results indicate that simultaneously incorporating HTAs and

Algorithm 1 Hierarchical Taxonomic Annotation: retrieval of molecule classification from PubChem.

```
Input preparation
Load CID list ........................ {from CIDs.txt file}
Configure batch size ................. {batch_size = 20}

Batch processing setup
Thread pool executor ................. {max_workers = 5}
Retry mechanism ...................... {max_retries = 3, backoff_factor = 2}
Rate limiting ........................ {0.5s between API calls, 3s batches}

API interaction
Classification headers retrieval
Endpoint ............................. {PubChem /pug_view/data/compound/{cid}/JSON/?heading=Classification}
Output ............................... {list of classification systems with HIDs}
Classification path retrieval
Endpoint ............................. {PubChem /classification_2.fcgi?hid={hid}&search_uid={cid}}
Path construction .................... {recursive traversal of parent-child nodes}
Output ............................... {hierarchical path string, description}

Result processing
Data structure ....................... {CID → {classification_system → {path, description}}}
Intermediate saving .................. {save after each batch for resumability}
Error handling ....................... {log warnings and errors, continue processing}

Output
JSON file ............................ {complete taxonomy annotation for all CIDs}
```

traditional functional annotations helps the model capture both fine-grained structural features and broader biological semantics, leading to improved performance across a range of molecular property prediction tasks.

B. Related Works

B.1. Molecule-Text Multimodal Learning

Recent advancements in molecular representation learning have demonstrated the power of multimodal approaches that integrate information from molecular graphs, SMILES strings, and textual descriptions to enhance property prediction and drug discovery. Graph Neural Networks (GNNs) have become the backbone of graph-based methods, with models like GROVER (Rong et al., 2020) and MolCLR (Wang et al., 2022) leveraging contrastive learning to produce richer molecular embeddings. Multimodal models such as KV-PLM (Zeng et al., 2022) and MolT5 (Edwards et al., 2022) treat SMILES and text as separate languages for pre-training via auto-encoding objectives, while MoMu (Su et al., 2022) and MoleculeSTM (Liu et al., 2023a) utilize independent encoders with cross-modal contrastive learning to align graphs and texts. MolFM (Luo et al., 2023) extends this paradigm by incorporating molecular structures, biomedical texts, and knowledge graphs to capture more comprehensive molecular relationships. However, despite this progress, the textual modality in existing models often derives from unstructured or single-layered descriptions, limiting the capacity to represent molecular functions across diverse biological roles and hierarchical categories. This lack of structured semantic alignment limits the ability of models to reason over complex molecular behaviors and relationships. Our work addresses this gap by introducing a high quality dataset to incorporate hierarchical taxonomic annotations for molecules, learning fine-grained hierarchical molecule-function relationships.

Algorithm 2 MultiModal Contrastive Learning: three-modal alignment with momentum integration.

```
Input: SMILES encoder, Text encoder, Category encoder
Output: Trained multimodal representations
Input modality encoders
    SMILES encoder .................... MoLFormer (768-dim)
    Text description encoder .......... SciBERT (768-dim)
    Category encoder .................. shared with text encoder (768-dim)
Projection layers
    SMILES projection ...... Linear→GELU→LayerNorm→Dropout→Linear→GELU→LayerNorm→Linear (512-dim)
    Text projection ........ Linear→GELU→LayerNorm→Dropout→Linear→GELU→LayerNorm→Linear (512-dim)
Global contrastive loss (GRAM3Modal)
    Volume computation ................ determinant of Gram matrix
    Temperature scaling ............... $\tau = 0.07$
    Volume-based alignment ............ cross entropy on negative volumes
    InfoNCE alignment ................. standard contrastive across modalities
Local functional group alignment
    Functional group detection ........ RDKit Fragments
    FG representation ................. weighted pooling of fragment embeddings
    FG contrastive loss ............... local InfoNCE between SMILES and text
Momentum-based loss integration
    Momentum coefficient .............. $\beta = 0.9$
    Initial alpha ..................... $\alpha = 0.5$
    Dynamic update .................... $\alpha = \beta \cdot \alpha_{prev} + (1 - \beta) \cdot (global\_loss/total\_loss)$
Training configuration
    Optimization ...................... Adam (lr=1e-5)
    Encoder freezing .................. both text and SMILES encoders
    Distributed training .............. DDP, NCCL backend
    Batch size ........................ 40 per GPU, multi-GPU
```

B.2. Contrastive Learning for Multimodal Alignment

Contrastive learning has emerged as a powerful strategy for aligning representations across modalities. Seminal models such as CLIP (Radford et al., 2021) demonstrated effective image-text alignment, inspiring extensions to other domains including audio (CLAP) (Elizalde et al., 2023), video (CLIP4Clip) (Luo et al., 2022), and point clouds (PointCLIP) (Zhang et al., 2022). These models typically learn by pulling semantically similar cross-modal pairs closer while pushing dissimilar ones apart. More recent approaches such as CLIP4VLA (Ruan et al., 2023), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2023) explore multimodal fusion, often anchoring learning around a central modality like images or text. GRAM (Cicchetti et al., 2024) advances this direction by introducing geometry-aware volume based contrastive objective, but it primarily focuses on audio-video-text pairs without structured semantic hierarchies. Unlike existing methods, our TRIDENT framework tackles the unique challenges of molecule-text alignment by incorporating hierarchical taxonomic relationships to capture functional semantics, and introducing global and local alignment modules with momentum-based mechanism. This enables fine-grained substructure-function correspondence and a richer multimodal embedding space tailored to molecular understanding.

C. Experimental Setup

C.1. Model Architecture

As shown in Algorithm A, our multimodal contrastive learning framework consists of the following key components:

1. **Three-modal encoding**: MoLFormer processes SMILES structures, while SciBERT encodes molecular text descriptions and category information, outputting 768-dimensional features.

2. **Feature projection**: Multi-layer MLPs project features from each modality into a 512-dimensional shared space with L2 normalization.

3. **Two-level contrastive learning**:

   - Global contrast: Applies GRAM3Modal method to calculate volume loss and InfoNCE loss across three modalities
   - Local contrast: Aligns SMILES and text representations at the functional group level

4. **Dynamic loss integration**: Employs a momentum update mechanism ($\beta = 0.9$) to adaptively adjust weights between global and local losses, with total loss $L = \alpha \cdot L_{\text{global}} + (1 - \alpha) \cdot L_{\text{local}}$.

D. Data Collection and Processing

We obtain a dataset containing 320,000 molecule-text pairs from the PubChem database and preprocess the text descriptions following the MolecularSTM method. Specifically, molecule names are replaced with “this molecule is...” or “these molecules are...” to prevent the model from recognizing molecules based solely on their names. Additionally, to create unique SMILES-text pairs, we merge molecules with the same CID (chemical identifier) and filter out text descriptions with fewer than 18 characters.

Moreover, we use PubChem’s classification system to obtain up to 32 classification descriptions for each molecule, as illustrated in Algorithm A. Ultimately, we generate 47,269 <SMILES, Text, Hierarchical Taxonomic Annotation> triplets. As shown in Figure 3, to further optimize and summarize the classification annotations, we use GPT-4 to generate summarized descriptions, resulting in high-quality HTA text descriptions.

The model is implemented in a distributed training environment, freezing pre-trained encoders and optimizing only projection layer parameters.

D.1. Training Configuration

Our multimodal contrastive learning model was trained on two NVIDIA H100 GPUs with the following configuration, as shown in Table 10.

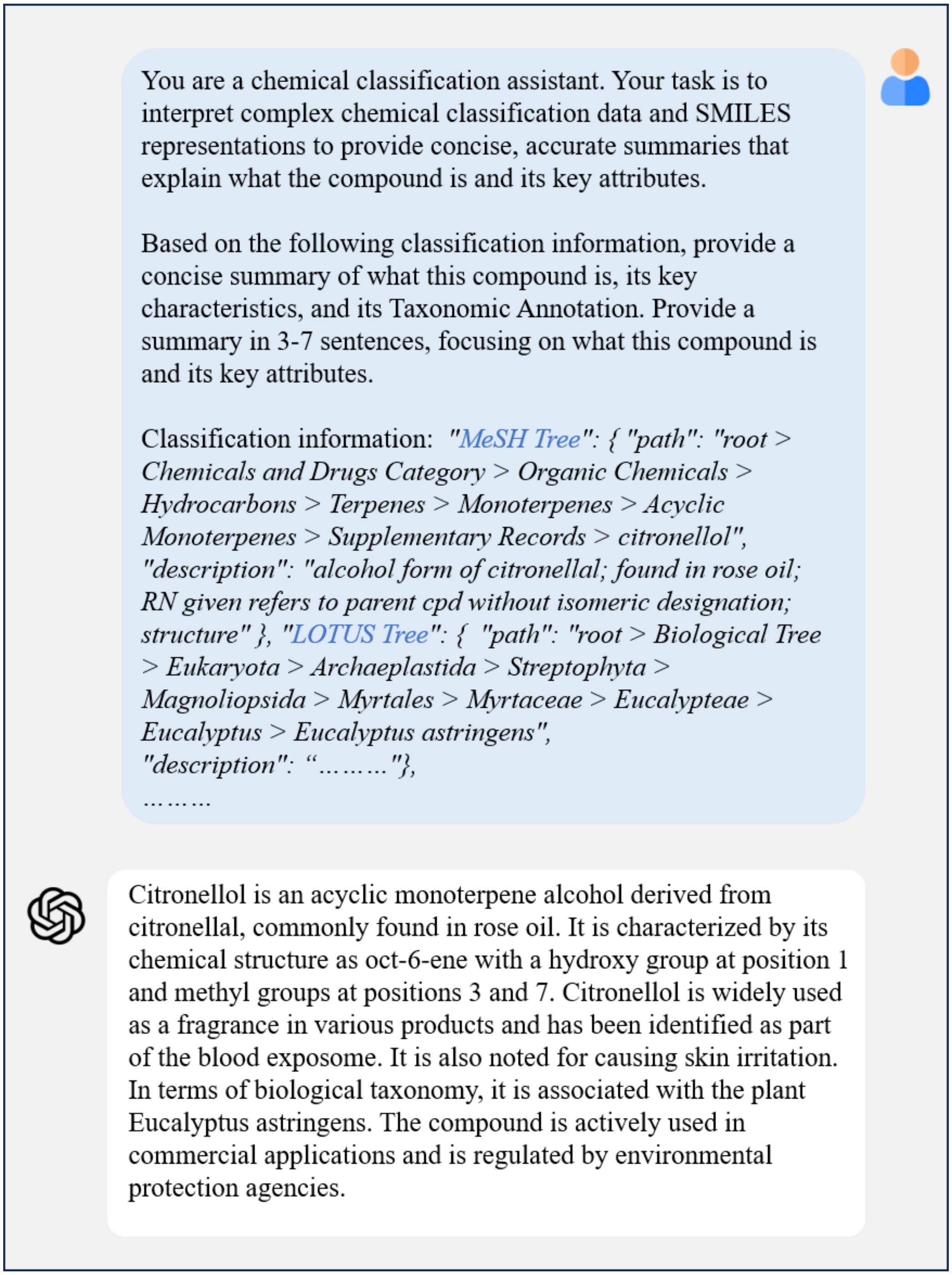
Each training epoch takes approximately 5 minutes. During training, we used DistributedSampler to ensure consistent data distribution across different GPUs and shuffled the data by setting different random seeds at the beginning of each epoch. Due to the large size of MoLFormer and SciBERT models, we adopted a strategy of freezing pre-trained encoder parameters and only training projection layer parameters, which significantly reduced computation and memory requirements while maintaining model expressiveness. We observe that the dynamic integration of global and local losses (dynamic adjustment of α value) demonstrates good adaptability during the training process, enabling reasonable balancing of the contributions from the two losses at different training stages.

D.2. Evaluation Metrics

To comprehensively evaluate the performance of our multimodal contrastive learning model on molecular property prediction tasks, we adopt appropriate evaluation metrics based on the characteristics of different datasets.

D.2.1. MoleculeNet Datasets

For binary classification tasks in MoleculeNet datasets, we employ ROC-AUC (Receiver Operating Characteristic Area Under Curve) and standard deviation as the primary evaluation metric. The ROC-AUC is calculated as follows:



You are a chemical classification assistant. Your task is to interpret complex chemical classification data and SMILES representations to provide concise, accurate summaries that explain what the compound is and its key attributes.

Based on the following classification information, provide a concise summary of what this compound is, its key characteristics, and its Taxonomic Annotation. Provide a summary in 3-7 sentences, focusing on what this compound is and its key attributes.

Classification information: *"MeSH Tree": { "path": "root > Chemicals and Drugs Category > Organic Chemicals > Hydrocarbons > Terpenes > Monoterpenes > Acyclic Monoterpenes > Supplementary Records > citronellol", "description": "alcohol form of citronellal; found in rose oil; RN given refers to parent cpd without isomeric designation; structure" }, "LOTUS Tree": { "path": "root > Biological Tree > Eukaryota > Archaeplastida > Streptophyta > Magnoliopsida > Myrtales > Myrtaceae > Eucalypteae > Eucalyptus > Eucalyptus astringens", "description": "………"},*
………

Citronellol is an acyclic monoterpene alcohol derived from citronellal, commonly found in rose oil. It is characterized by its chemical structure as oct-6-ene with a hydroxy group at position 1 and methyl groups at positions 3 and 7. Citronellol is widely used as a fragrance in various products and has been identified as part of the blood exposome. It is also noted for causing skin irritation. In terms of biological taxonomy, it is associated with the plant Eucalyptus astringens. The compound is actively used in commercial applications and is regulated by environmental protection agencies.

Figure 3. The workflow for summarizing Hierarchical Taxonomic Annotations (HTA). Using GPT-4o, detailed classification annotations are processed and summarized, resulting in high-quality HTA text descriptions for molecular data.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(t))\, dt \tag{1}$$

where the True Positive Rate (TPR) and False Positive Rate (FPR) are defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3}$$

ROC-AUC values range from 0 to 1, with values closer to 1 indicating better model performance. This metric demonstrates good robustness to class imbalance issues, making it particularly suitable for molecular property prediction tasks in the pharmaceutical domain where positive and negative samples are often unevenly distributed.

$$\mathrm{STD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

where n is the number of experiments, x_i is the result of the i-th experiment, and $\bar{x}$ is the mean of n experiments. The standard deviation reflects the stability and reliability of model performance, with smaller standard deviations indicating more stable performance across different data splits and random seeds.

D.2.2. TDC Datasets

For TDC (Therapeutics Data Commons) datasets, we employ both ROC-AUC and Accuracy as evaluation metrics:

1. **ROC-AUC**: Same definition as in MoleculeNet datasets, used to measure the model's classification performance and discriminative ability.
2. **Accuracy**: The accuracy is calculated as:

   $$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

   where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy intuitively reflects the proportion of correctly predicted samples by the model. When used in combination with ROC-AUC, it provides a more comprehensive evaluation of model performance. ROC-AUC primarily focuses on the model's ranking ability and threshold-independent performance, while accuracy directly reflects the model's classification effectiveness under specific thresholds.

E. Downstream Tasks Datasets

To comprehensively evaluate the performance of our proposed multimodal contrastive learning framework on molecular property prediction tasks, we conduct extensive experiments on two major benchmark dataset collections: MoleculeNet and TDC (Therapeutics Data Commons).

E.1. MoleculeNet Datasets

MoleculeNet is one of the most authoritative benchmark dataset collections in the field of molecular machine learning, specifically designed to evaluate the performance of molecular property prediction methods. Table 3 summarizes the detailed information of the 8 MoleculeNet datasets we used.

E.2. TDC Datasets

TDC (Therapeutics Data Commons) is a large-scale dataset collection specifically designed for therapeutics research, providing more challenging and practically valuable molecular property prediction tasks. Table 4 presents the detailed information of the 5 TDC datasets we selected.

Table 3. MoleculeNet Datasets Details

Dataset	Sample Size	Prediction Task	Task Description
BBBP	2,050	Blood-Brain Barrier Penetration	Predicts whether compounds can penetrate the blood-brain barrier
Tox21	7,831	Toxicity Assessment	Evaluates compound activity across 12 different toxicity pathways
ToxCast	8,597	Toxicity Prediction	Predicts compound toxicity across 617 biological assays
SIDER	1,427	Side Effect Prediction	Predicts adverse drug reactions covering 27 types of side effects
ClinTox	1,483	Clinical Toxicity	Evaluates clinical toxicity and FDA approval status of compounds
MUV	93,087	Biological Activity	Molecular activity prediction with 17 highly imbalanced biological targets
HIV	41,127	Antiviral Activity	Predicts compound inhibition of HIV replication
BACE	1,513	Enzyme Inhibition	Predicts β-secretase inhibitor activity for Alzheimer's disease drug discovery

Table 4. TDC Datasets Details

Dataset	Sample Size	Prediction Task	Task Description
DILI	475	Liver Injury Prediction	Predicts drug-induced liver injury, a critical safety consideration in drug development
Carcinogens	278	Carcinogenicity Prediction	Predicts compound carcinogenicity, crucial for drug and chemical safety evaluation
Skin Reaction	404	Skin Reaction Prediction	Predicts whether compounds cause skin reactions, important for topical drug development
AMES	7,255	Mutagenicity Prediction	Predicts compound mutagenicity based on Ames test, standard method for genetic toxicity
hERG	648	Cardiotoxicity Prediction	Predicts compound blocking activity against hERG potassium channels, major cause of cardiotoxicity

These datasets cover key property prediction tasks in the drug discovery process, including pharmacokinetics (ADME), toxicity, and biological activity across multiple aspects. Both dataset collections are characterized by diversity, challenging nature, standardization, and authority, and are widely recognized and used by both academia and industry.

F. Baselines

In this section, we provide descriptions of the baseline methods used for comparison in our experiments. These baselines represent current approaches in molecular representation learning and multimodal molecular modeling.

F.1. Single-Modal Baselines

MOLFORMER (Ross et al., 2022): A transformer-based model that processes SMILES string representations using masked language modeling. The model employs linear attention with rotary positional embeddings and is pre-trained on 1.1 billion molecules from PubChem and ZINC databases in an unsupervised fashion. **MegaMolBART (Reidenbach et al., 2022)**: A BART-based encoder-decoder model adapted for molecular data. It processes SMILES representations and applies bidirectional and auto-regressive transformers for molecular understanding and generation tasks.

F.2. Multimodal Baselines

MoleculeSTM (Liu et al., 2023a): A bi-modal model with separate encoders for molecular structures (SMILES/graphs) and textual descriptions. It uses contrastive learning to align structure-text pairs and is trained on over 280,000 molecule-text pairs from PubChem. **MoMu (Su et al., 2022)**: A multimodal foundation model that uses separate encoders for molecular graphs and natural language text. The model employs contrastive learning to bridge molecular structures with textual descriptions using paired molecule-text datasets. **MolFM (Luo et al., 2023)**: A tri-modal model that integrates molecular structures (2D graphs), biomedical texts, and knowledge graphs. It uses cross-modal attention mechanisms and is pre-trained with four objectives: structure-text contrastive learning, cross-modal matching, masked language modeling, and knowledge graph embedding. **KV-PLM (Zeng et al., 2022)**: A BERT-based unified framework that processes both SMILES-encoded molecular structures and natural language text through masked language modeling pre-training. The system enables cross-modal understanding between molecular structures and biomedical text. **MolCA-SMILES (Liu et al., 2023b)**: A molecular graph-language model that uses a Q-Former as a cross-modal projector to bridge graph encoders and language models. The approach employs LoRA adapters and follows a three-stage training pipeline for efficient fine-tuning. **Atomas (Zhang et al., 2025)**: A hierarchical alignment framework that introduces Adaptive Polymerization Module (APM) and Weighted Alignment Module (WAM) to learn fine-grained correspondences between SMILES and text at atom, fragment, and molecule levels. It uses a unified encoder and end-to-end training for joint alignment and generation.

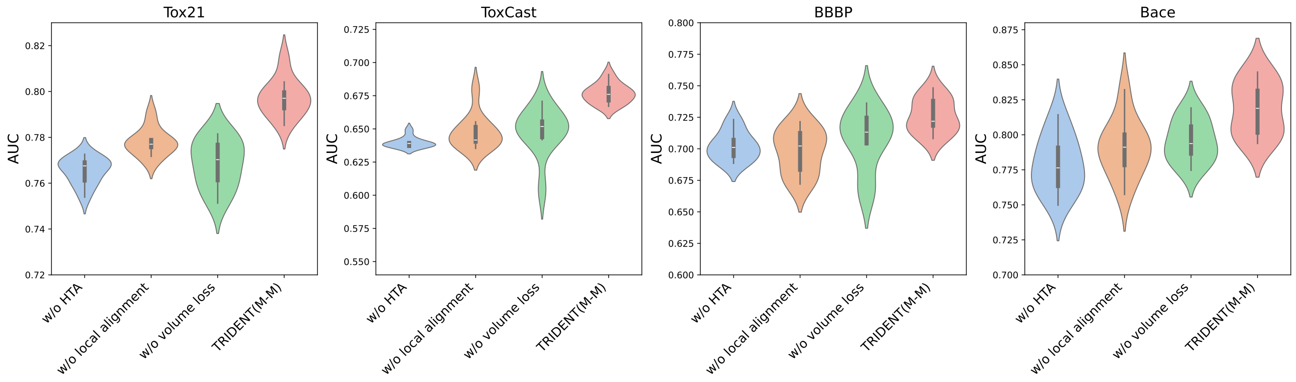


Figure 4. The ablation experiments are conducted on the Tox21, ToxCast, BBBP and Bace datasets. "`w/o HTA`" denotes that only not use hierarchical taxonomic annotation; "`w/o local alignment`" denotes that the local alignment is removed; and "`w/o volume loss`" indicates that only the volume-based loss is changed to the standard contrastive loss.

F.3. Comparison with TRIDENT

Our proposed TRIDENT framework differs from these baselines in several key aspects:

1. **Hierarchical Taxonomic Annotations**: Unlike existing methods that rely on generic textual descriptions, TRIDENT incorporates structured, multi-level functional annotations across 32 taxonomic classification systems, providing richer semantic understanding.

2. **Tri-modal Architecture**: While most baselines focus on bi-modal alignment (structure-text), TRIDENT introduces a novel tri-modal approach that jointly models SMILES, textual descriptions, and hierarchical taxonomic annotations.

3. **Volume-based Global Alignment**: Instead of traditional pairwise contrastive learning, TRIDENT employs a geometry-aware volume-based alignment objective that captures higher-order relationships across all three modalities simultaneously.

4. **Local-Global Integration**: TRIDENT uniquely combines global tri-modal alignment with fine-grained local alignment between molecular substructures and their corresponding textual descriptions, balanced through a momentum-based mechanism.

5. **Dynamic Alignment Strategy**: The momentum-based integration of global and local objectives allows TRIDENT to adaptively focus on different alignment components during training, leading to more robust representation learning.

These innovations enable TRIDENT to achieve state-of-the-art performance across 11 downstream molecular property prediction tasks, demonstrating the effectiveness of our comprehensive multimodal approach.

G. Additional Results

In this section, we present additional experimental results that complement the main findings reported in the paper. These include performance evaluations on larger-scale datasets from the TDC benchmark, more extensive ablation experiments, and additional analyses that provide deeper insights into TRIDENT's capabilities.

G.1. Ablation Study

To understand the contribution of different components in our TRIDENT framework, we conduct a detailed ablation study. We compare several model variants on representative tasks to disentangle the impact of local functional group and sub-textual description alignment loss, hierarchical taxonomic supervision as well as the volume loss for global alignment.

As shown in Figure 4, removing HTA information (w/o HTA) leads to a noticeable drop in model performance, highlighting the importance of HTA in capturing a rich, multi-level understanding of molecular behavior through hierarchical taxonomy

Table 5. Performance comparison of molecular property prediction methods based on different input modalities (SMILES, Text, and HTA) across various datasets (ROC-AUC%). Best results in **bold**.

Method	Input			Datasets				
	SMILES	Text	HTA	BBBP	Tox21	ToxCast	Sider	Bace
TRIDENT (M-M)	✓	×	✓	72.02±0.36	78.21±0.19	67.04±0.38	63.18±0.31	81.28±0.92
TRIDENT (M-M)	✓	✓	✓	**73.95±1.01**	**79.36±0.13**	**67.80±0.37**	**63.64±0.56**	**82.39±0.56**

annotations. Similarly, excluding the local-alignment component (w/o local alignment) results in a clear performance decline, showing how fine-grained alignment plays a critical role in enhancing the model's capability. Interestingly, replacing our volume loss with standard contrastive loss (w/o volume loss) causes significant instability on datasets like Tox21, ToxCast, and BBBP. This is likely because traditional alignment approaches struggle to handle multiple modalities effectively (Cicchetti et al., 2024). In addition, our momentum-based mechanism further strengthens generalization by dynamically balancing global and local objectives during training, as demonstrated in Table 6. Overall, the full TRIDENT framework consistently outperforms all ablated versions, confirming the value and necessity of each individual component.

Table 6. Strategies for combining global and local loss functions (ROC-AUC%). Sum: direct addition; Curve: sigmoid-weighted combination with increasing local loss weight; Momentum: dynamic alignment approach. Best results in **bold**.

Method	Tox21	ToxCast	BBBP	Bace
Sum	77.79±0.81	66.73±0.65	72.15±0.81	81.42±0.69
Curve	76.68±0.79	65.49±0.82	71.68±0.79	80.91±0.83
Momentum	**79.36±0.13**	**67.80±0.37**	**73.95±1.01**	**82.39±0.56**

In addition, to further explore the relationship between HTA and general molecular descriptions, we directly use HTA and molecular SMILES as inputs for the global module during pretraining, replacing the volume loss with standard contrastive learning loss while keeping other settings unchanged. The results are shown in Table 5. When using HTA as the sole text input, the model already outperforms most baselines but still falls short of the tri-modal input. This may be because HTA text and traditional molecular descriptions complement each other in terms of information representation. HTA text contains up to 32 categorical annotations, providing more diverse and multi-angled molecular information, while traditional functional descriptions are more direct and highlight the core features of molecular structures. Therefore, by simultaneously leveraging HTA text and traditional descriptions as multimodal inputs, the model captures molecular characteristics more comprehensively, thereby further improving its performance.

G.2. Performance on Larger TDC Datasets

While the main paper focused on smaller TDC datasets to demonstrate TRIDENT's data efficiency, we also evaluated our method on larger-scale molecular property prediction tasks. Table 7 presents the results on the AMES mutagenicity dataset (7,255 molecules) and the hERG cardiotoxicity dataset (648 molecules). In summary, TRIDENT's superior performance on both the large-scale AMES dataset and the moderately-sized hERG dataset demonstrates the versatility and scalability of our approach. The consistent improvements across different dataset sizes—from hundreds to thousands of molecules—validate that the tri-modal alignment strategy and hierarchical taxonomic annotations provide robust molecular representations that generalize well across various scales and prediction tasks. These results complement our findings on larger datasets and further establish TRIDENT as a powerful framework for molecular property prediction across the full spectrum of practical applications in drug discovery.

G.3. Performance without LLM Summary

To evaluate the contribution of LLM-based summarization in our HTA generation process, we conduct an ablation study comparing the performance of TRIDENT when using raw JSON taxonomic annotations versus LLM-synthesized HTA descriptions. In this experiment, we directly input the structured JSON files containing hierarchical taxonomic paths and descriptions from the 32 classification systems, bypassing the GPT-4o summarization step described in Section 3.1.

The results in Table 8 demonstrate the effectiveness of LLM-based synthesis in our HTA generation pipeline. When using raw JSON taxonomic annotations without LLM summarization (TRIDENT w/o LLM), the model achieves competitive performance but consistently underperforms compared to the full TRIDENT framework across all datasets.

Table 7. Performance of different methods on AMES and hERG tasks, reporting AUC and Accuracy. The best results are marked in **bold**, and the second-best are underlined.

Method	**AMES (7,255 drugs)**		**hERG (648 drugs)**	
	AUC	**ACC**	**AUC**	**ACC**
MOLFORMER	83.20±0.32	78.05±0.76	79.65±1.19	81.82±3.03
KV-PLM	78.23±0.90	71.70±0.94	75.87±2.76	75.30±3.08
MolT5	76.93±0.84	70.87±2.22	76.25±1.22	77.04±4.90
MoMu	77.20±0.85	70.78±0.36	75.68±1.89	73.27±3.55
MolCA-SMILES	77.62±1.49	71.74±1.07	78.40±1.84	73.94±4.38
MoleculeSTM-SMILES	83.60±1.00	77.68±0.64	79.46±4.63	79.19±4.94
Atomas	82.63±0.72	77.32±0.83	83.34±1.79	78.02±2.00
TRIDENT (M-S)	85.37±0.30	78.74±0.50	**87.60±1.20**	81.11±2.64
TRIDENT (M-M)	**86.87±0.60**	**80.20±1.44**	83.31±1.63	**83.33±2.62**

Table 8. Ablation study on the impact of LLM-based summarization in HTA generation. Comparison between using raw JSON taxonomic annotations versus LLM-synthesized HTA descriptions across molecular property prediction datasets (ROC-AUC%). Best results in **bold**.

Method	Input			Datasets				
	SMILES	Text	HTA	BBBP	Tox21	ToxCast	Sider	Bace
TRIDENT (M-M) w/o LLM	✓	✓	✓	71.89±0.56	79.01±0.33	66.86±0.75	62.78±0.45	81.12±0.69
TRIDENT (M-M)	✓	✓	✓	**73.95±1.01**	**79.36±0.13**	**67.80±0.37**	**63.64±0.56**	**82.39±0.56**

This performance gap highlights several key advantages of LLM-based summarization: (1) **Information Integration**: The LLM synthesis process effectively combines information from multiple taxonomic systems into coherent, contextually rich descriptions that capture cross-domain knowledge spanning chemistry, biology, and pharmacology. (2) **Semantic Coherence**: Raw JSON annotations often contain fragmented or inconsistent terminology across different classification systems, while LLM synthesis produces semantically coherent descriptions that are more amenable to natural language processing. (3) **Contextual Enrichment**: The synthesis process adds relevant contextual information and relationships between different taxonomic levels that may not be explicitly present in individual classification paths.

While the raw taxonomic annotations still provide valuable structural information that outperforms traditional text-only approaches, the LLM synthesis step proves crucial for maximizing the utility of hierarchical taxonomic knowledge in molecular representation learning. This finding validates our design choice to incorporate GPT-4o in the HTA generation pipeline and demonstrates that the additional computational cost of LLM synthesis is justified by the consistent performance improvements across all molecular property prediction tasks.

G.4. Impact of Tri-modal vs. Concatenated Text Architecture

To validate the necessity of our tri-modal architecture, we conduct an ablation study comparing our approach with a simpler alternative that concatenates HTA and traditional text descriptions into a single textual input. This experiment evaluates whether treating HTA and text as separate modalities provides advantages over a straightforward concatenation approach.

The results in Table 9 demonstrate the effectiveness of our tri-modal architecture over the concatenation approach. The concatenated version (TRIDENT Concatenated) combines HTA and traditional molecular descriptions into a single text input using simple string concatenation with separator tokens, then processes this unified text through the same text encoder used in our tri-modal framework. While this approach still benefits from the rich semantic information in HTA, it consistently underperforms the tri-modal architecture across all datasets. These consistent improvements highlight several key advantages of treating HTA and text as separate modalities:

Modality-Specific Representation Learning: The tri-modal architecture allows the model to learn distinct representation spaces for hierarchical taxonomic information and functional descriptions. This separation enables the capture of different

Table 9. Ablation study comparing tri-modal architecture (SMILES + Text + HTA as separate modalities) versus concatenated text approach (SMILES + concatenated HTA$\oplus$Text as single text modality). The concatenated approach combines HTA and traditional molecular descriptions using string concatenation with separator tokens, while the tri-modal approach processes each information source through separate encoders with volume-based alignment. Performance reported across molecular property prediction datasets using ROC-AUC(%). Best results in **bold**.

Method	Architecture		Datasets				
	Modalities	Text Processing	BBBP	Tox21	ToxCast	Sider	Bace
TRIDENT (Concatenated)	SMILES + Text	HTA$\oplus$Text	70.918±0.82	76.67±0.59	64.59±0.72	61.74±0.83	79.15±0.69
TRIDENT (M-M)	SMILES + Text + HTA	Separate	**73.95±1.01**	**79.36±0.13**	**67.80±0.37**	**63.64±0.56**	**82.39±0.56**

semantic aspects—taxonomic relationships in HTA versus direct functional properties in traditional text—that may require different representational strategies.

Enhanced Alignment Flexibility: The volume-based tri-modal alignment objective can capture complex geometric relationships between SMILES, text, and HTA that are not accessible when HTA and text are merged into a single modality. This geometric awareness enables more nuanced understanding of how molecular structure relates to both functional properties and taxonomic classifications.

Reduced Information Interference: Concatenation may lead to interference between the structured, multi-level taxonomic information and the more direct functional descriptions, potentially diluting the distinct contributions of each information source. Separate processing preserves the unique characteristics of each modality.

Dynamic Weighting Capabilities: The tri-modal framework allows for dynamic balancing of different information sources during training through our momentum-based mechanism, whereas concatenation fixes the relative importance of HTA and text information at the input level.

These findings validate our design choice to maintain HTA and traditional text as separate modalities, demonstrating that the additional architectural complexity of tri-modal learning is justified by consistent performance gains across all molecular property prediction tasks.

Table 10. Training Configuration Details

Parameter	**Configuration**
Hardware environment	2 × NVIDIA H100 GPU
Training framework	PyTorch DistributedDataParallel (DDP)
Communication backend	NCCL
Optimizer	Adam
Learning rate	1e-5
Batch size	40 per GPU (total batch size = 80)
Weight decay	1e-4
Training epochs	60 epochs
Training dataset size	47,269 molecule-text-HTA pairs
Gradient accumulation steps	1
Learning rate schedule	Fixed learning rate, no decay
Early stopping	Stop after 5 epochs without validation loss improvement
Loss Function Configuration	
Contrastive temperature	$\tau = 0.07$
Momentum coefficient	$\beta = 0.9$
Initial loss weight	$\alpha = 0.5$
Global loss composition	GRAM3Modal volume loss + InfoNCE loss
Local loss composition	Functional group level InfoNCE loss
Label smoothing parameter	0.1
Model Configuration	
Modality encoders	Frozen (feature extraction only)
Projection layers	Fully fine-tuned (768-dim → 512-dim)
Dropout rate	0.1
Gradient clipping	Max norm 1.0
Mixed precision training	FP16
Checkpoint saving frequency	Every 2 epochs