A 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 1. Ultimately, we generate 47,269 <SMILES, Text, Hierarchical Taxonomic Annotation> triplets. As shown in Figure 4, to further optimize and summarize the classification annotations, we use GPT-4 to generate summarized descriptions, resulting in high-quality HTA text descriptions.

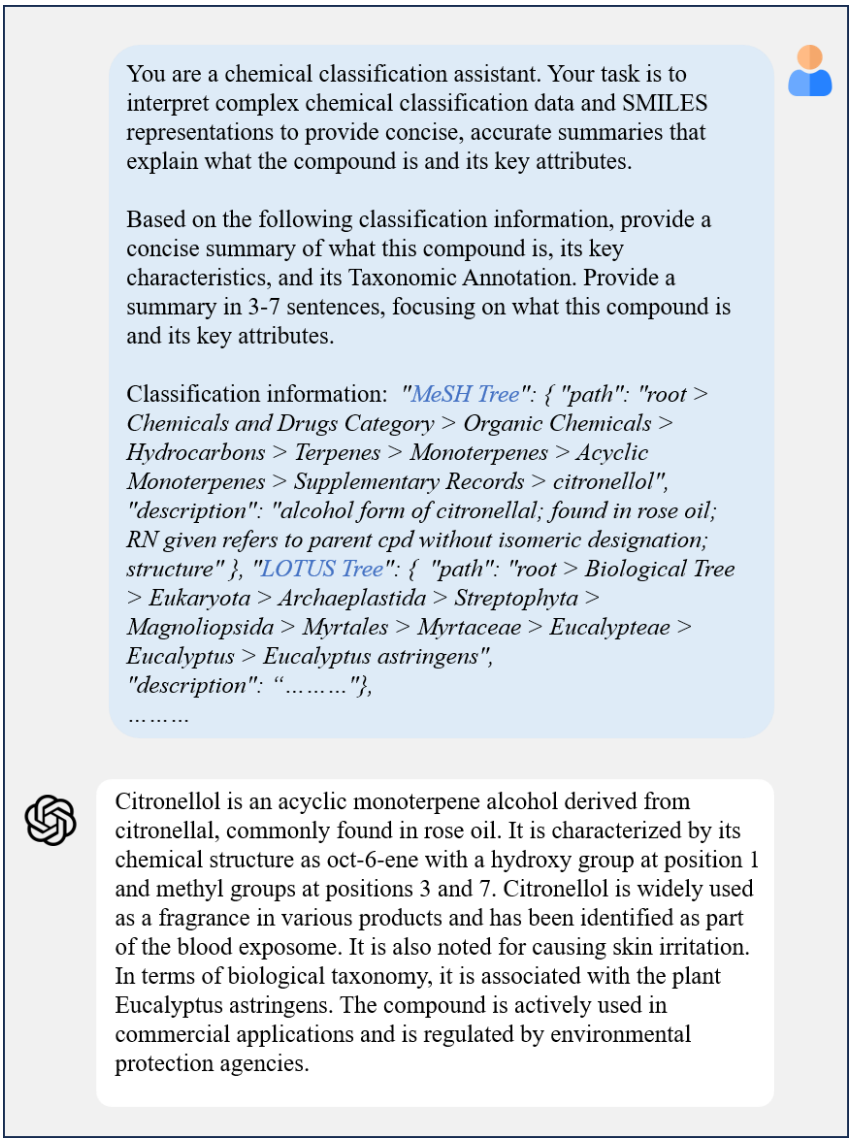

Figure 4: The workflow for summarizing Hierarchical Taxonomic Annotations (HTA). Using GPT-4o, detailed classification annotations are processed and summarized, resulting in high-quality HTA text descriptions for molecular data.

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 $\rightarrow$ {classification_system $\rightarrow$ {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}

B Experimental Setup

B.1 Model Architecture

As shown in Algorithm 2, 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 $L = \alpha \cdot L_{\text{global}} + (1 - \alpha) \cdot L_{\text{local}}$.

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

B.2 Training Configuration

Our multimodal contrastive learning model was trained on two NVIDIA H100 GPUs with the following configuration, as shown in Table 5.

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

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

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}

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.

B.3 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.

B.3.1 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:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{9}$$

Table 5: 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

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

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{11}$$

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.

$$\text{STD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

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.

Table 6: 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 7: 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

B.3.2 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:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

   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.

C 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).

C.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 6 summarizes the detailed information of the 8 MoleculeNet datasets we used.

C.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 7 presents the detailed information of the 5 TDC datasets we selected.

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.

D 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.

D.1 Single-Modal Baselines

MOLFORMER: 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**: 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.

D.2 Multimodal Baselines

MoleculeSTM: 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**: 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**: 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**: 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**: 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**: 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.

D.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.

Table 8: 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 9: 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**

E 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.

E.1 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 8 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.

E.2 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 9 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 10: Ablation study comparing tri-modal architecture (SMILES + Text + HTA as separate modalities) versus concatenated text approach (SMILES + concatenated HTA⊕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⊕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**

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.

E.3 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 10 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 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.