MIP²: INFORMATION-PRESERVING PROJECTOR FOR MOLECULAR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular large language models (LLMs) use cross-modal projectors to map pretrained molecular features into the language model embedding space. However, existing projectors do not explicitly enforce the recoverability of these features after projection. Therefore, the projectors may limit access to fine-grained features captured by molecular encoders. To address this challenge, we propose **M**olecular **I**nformation-**P**reserving **P**rojector (MIP²), a molecular-language projector that combines feature preservation with flexible nonlinear adaptation. MIP² first integrates topological and geometric features through reversible coupling, preserving both encoder representations. It then separates feature preservation and nonlinear adaptation into orthogonal subspaces, yielding an injective projection with an explicit recovery map. MIP² further retains one token per atom and enables bidirectional attention within the molecular span to support atom-level structural interactions. Experiments on molecular question answering (MoleculeQA dataset), description generation and retrosynthesis (Mol-Instructions dataset), and property prediction (Mol-Instructions and eight MoleculeNet datasets) demonstrate the effectiveness of MIP². Backbone comparisons and ablation studies further support the benefits of the proposed design.

1 INTRODUCTION

Large language models (LLMs) are increasingly used for molecular understanding, generation, and property prediction (Liu et al., 2023b; Li et al., 2024; Park et al., 2024). Recent work has extended molecular LLMs to incorporate multiple structural modalities, combining 2D graphs and 3D conformations to provide complementary topological and geometric information (Kim et al., 2026; Park et al., 2026a). These models typically use a cross-modal molecular-language projector to connect the pretrained molecular encoder to the LLM (Liu et al., 2023b; Li et al., 2024). The molecular-language projector transforms or aggregates the encoder features into molecular token embeddings in the LLM embedding space, which are then processed alongside textual inputs (Cao et al., 2025; Kim et al., 2026). Therefore, the molecular-language projector plays a crucial role in bridging molecular representations and language models.

Recent studies have explored various projector architectures. Existing molecular-language projectors commonly follow two designs: direct feature mapping and query-based aggregation (Cao et al., 2025; Liu et al., 2023b; Li et al., 2024). Direct mapping projectors use linear layers or multilayer perceptrons (MLPs) to transform molecular encoder features into embeddings compatible with the LLM (Cao et al., 2025). Query-based projectors, exemplified by Q-Former-style connectors, use learnable queries to aggregate structural representations into molecular tokens (Liu et al., 2023b; Li et al., 2024; Kim et al., 2026). Beyond feature mapping and aggregation, recent work explores dynamic token allocation to improve structural coverage (Jing et al., 2026). These approaches focus on adapting the representation and granularity of molecular features for integration with LLMs.

However, these approaches typically emphasize feature adaptation, aggregation, or token allocation, without explicitly enforcing exact recovery of the pretrained encoder features after projection (Liu et al., 2023b; Jing et al., 2026). We formulate feature preservation as exact recovery of pretrained encoder features from their projected representations, which is essential for retaining the information captured by pretrained molecular encoders. This requirement motivates a molecular-language

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

projector that guarantees feature recoverability while allowing flexible nonlinear adaptation to the LLM embedding space.

We propose **M**olecular **I**nformation-**P**reserving **P**rojector (MIP²), comprising reversible multimodal fusion, information-preserving projection, and atom-level molecular contextualization. Reversible fusion enables interactions between 2D and 3D features without discarding either encoder's information. MIP² separates feature preservation and nonlinear adaptation into orthogonal subspaces, enabling flexible alignment while ensuring exact feature recoverability. Atom-level contextualization retains one token per atom to avoid aggregation-induced information loss and enables bidirectional molecular attention to capture structural dependencies.

Our main contributions are as follows:

- We establish information preservation as an explicit design objective for molecular-language projector and show that exact feature preservation can coexist with flexible nonlinear adaptation through an injective mapping.
- We develop MIP², a molecular-language projector that combines reversible 2D-3D fusion with nonlinear adaptation in a complementary subspace. MIP² enables flexible nonlinear alignment with the LLM embedding space while preserving the information encoded in pretrained molecular features.
- We evaluate MIP² across four molecular tasks: question answering, description generation, retrosynthesis, and property prediction. Experiment across three LLM backbones demonstrate consistence generalization, while ablation studies support the benefits of the complementary subspace constraint and nonlinear adaptation.

2 METHOD

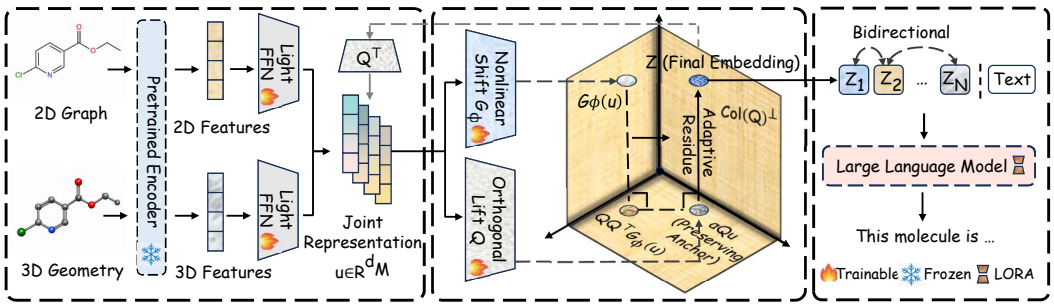

Figure 1: **Overview of proposed Molecular Information-Preserving Projector (MIP²).** Frozen 2D graph and 3D geometry encoders produce aligned atom-level representations, which interact through a reversible coupling before being lifted into the LLM embedding space. The orthogonal lift produces a preserved anchor αQu in $\mathrm{Col}(Q)$, while the nonlinear residual adapts the representation in the complementary subspace without changing the preserved coordinates. One projected token is retained per atom. Molecular tokens interact bidirectionally within the molecular span.

2.1 OVERVIEW

As illustrated in Fig. 1, frozen 2D and 3D molecular encoders produce atom-aligned features, which are passed to the MIP² projector. MIP² combines reversible fusion with constrained nonlinear projection. The fusion module enables interactions between 2D and 3D features while retaining both encoder representations. The projection module embeds the fused features into a reserved subspace of the LLM embedding space and restricts nonlinear adaptation to its orthogonal complement. This constraint is enforced in closed form, and the resulting mapping admits an explicit inverse for recovering the encoder features. The projected representations are retained as one token per atom and supplied to the LLM alongside text. Bidirectional attention within the molecular span supports interactions among atom tokens, while text generation follows the causal attention rule.

2.2 NONLINEAR INJECTIVE MOLECULAR PROJECTION

Information-preserving 2D-3D interaction. Independently pretrained 2D and 3D molecular encoders capture complementary topological and geometric information (Liu et al., 2023a; Zhou et al., 2023). Combining these representations calls for a fusion mechanism that supports nonlinear cross-modal interaction while preserving the information encoded by 2D and 3D encoders. We implement this information-preserving 2D-3D interaction through lightweight reversible additive coupling, which keeps both input representations exactly recoverable.

Let $h = [h^{2D}; h^{3D}]$ denote the pretrained topological and geometric features of the same atom, where the semicolon denotes concatenation, and h^{2D} and h^{3D} represent the 2D and 3D pretrained features, respectively. We implement the information-preserving 2D-3D interaction using two lightweight nonlinear feed-forward networks, f_θ and g_θ, with trainable parameters collectively denoted by θ. The joint representation $u = [u^{2D}; u^{3D}]$ is obtained through the coupling map

$$u = F_\theta(h) = h + \begin{bmatrix} g_\theta\left(h^{3D} + f_\theta(h^{2D})\right) \\ f_\theta(h^{2D}) \end{bmatrix}. \tag{1}$$

The nested update first incorporates topological information into the geometric features, then uses the updated geometric features to refine the topological features. This additive coupling admits an analytical inverse, ensuring exact recovery of both input representations. Architectural details are provided in Appendix C.

Information-preserving projection. The reversible interaction produces a fused representation u that retains both encoder features. To preserve this information during projection into the LLM embedding space, we separate feature preservation and nonlinear adaptation into orthogonal subspaces.

Let d_M and d_L denote the fused feature and LLM embedding dimensions, respectively, with $d_M \leq d_L$. We construct a preserved embedding αQu using a trainable matrix $Q \in \mathbb{R}^{d_L \times d_M}$ with orthonormal columns, $Q^\top Q = I_{d_M}$, and a learned scale $\alpha > 0$. Here, I_{d_M} denotes the $d_M \times d_M$ identity matrix.

To adapt αQu while preserving its coordinates in $\mathrm{Col}(Q)$, we restrict the nonlinear residual to $\mathrm{Col}(Q)^\perp$, the orthogonal complement of the column space of Q. Let $G_\phi : \mathbb{R}^{d_M} \to \mathbb{R}^{d_L}$ be a nonlinear network with trainable parameters ϕ. For each input u, we define the residual $r_\perp(u)$ as the closest vector to $G_\phi(u)$ satisfying the orthogonality constraint:

$$r_\perp(u) = \underset{v \in \mathbb{R}^{d_L}}{\arg\min} \frac{1}{2}\|v - G_\phi(u)\|_2^2 \quad \text{subject to} \quad Q^\top v = 0, \tag{2}$$

where v is a candidate residual vector and $\|\cdot\|_2$ denotes the Euclidean norm. The unique solution is the orthogonal projection of $G_\phi(u)$ onto $\mathrm{Col}(Q)^\perp$:

$$r_\perp(u) = (I_{d_L} - QQ^\top)G_\phi(u), \tag{3}$$

where I_{d_L} denotes the $d_L \times d_L$ identity matrix. This closed-form solution enforces the orthogonality constraint directly.

With a learned residual scale $\beta > 0$, the projection map Ψ produces the final atom embedding $z \in \mathbb{R}^{d_L}$. The complete molecular-language projector P composes Ψ with the reversible interaction map F_θ:

$$z = \Psi(u) = \alpha Qu + \beta r_\perp(u), \qquad P(h) = \Psi(F_\theta(h)). \tag{4}$$

Feature recoverability and separation. The projection defined above admits an explicit recovery map. Since Q has orthonormal columns, $Q^\top z$ extracts the coordinates of z along the preserved subspace. The constraint $Q^\top r_\perp(u) = 0$ eliminates the residual contribution to these coordinates. Applying $Q^\top$ to Eq. (4) and then inverting the interaction map F_θ gives

$$Q^\top z = \alpha u \quad \Longrightarrow \quad h = F_\theta^{-1}\left(\frac{1}{\alpha}Q^\top z\right). \tag{5}$$

Here, F_θ^{-1} denotes the inverse of the reversible 2D-3D interaction. This explicit left inverse establishes the injectivity of P and guarantees exact recovery of the input encoder features.

Beyond exact feature recoverability, the orthogonal decomposition also provides a distance-preservation property for the fused representations. For any two fused representations u_1, u_2, let $\Delta u = u_1 - u_2$ and $\Delta r = r_\perp(u_1) - r_\perp(u_2)$. Since $Q^\top \Delta r = 0$, the preserved and residual differences are orthogonal, yielding

$$\begin{aligned} \|\Psi(u_1) - \Psi(u_2)\|_2^2 &= \alpha^2 \|\Delta u\|_2^2 + \beta^2 \|\Delta r\|_2^2 \\ &\geq \alpha^2 \|\Delta u\|_2^2. \end{aligned} \tag{6}$$

The bound applies to u, while recovering h requires $F_\theta^- 1$. See Appendix B. for derivations and finite-precision details.

2.3 Atom-Level Molecular Contextualization

The feature recoverability established in Section 2.2 applies to individual atom representations but is not automatically guaranteed under subsequent aggregation across atoms. We therefore retain one molecular token per heavy atom at the LLM input, preserving the one-to-one correspondence between atom features and projected tokens. For a molecule containing N heavy atoms, the molecular span is given by

$$[\texttt{<mol_start>}, \texttt{<mol_atom>}_1, \ldots, \texttt{<mol_atom>}_N, \texttt{<mol_end>}]. \tag{7}$$

Each atom placeholder receives the corresponding projected representation, $E(\texttt{<mol_atom>}_i) \leftarrow P(h_i)$, while the two boundary tokens use trainable embeddings.

To give each atom token access to the full molecular context, we enable bidirectional attention within the molecular span. Let $\mathcal{S}_{\text{mol}}$ denote the set of molecular-span indices. For query position i and key position j, attention visibility is defined as follows:

$$A_{ij} = \mathbb{I}[j \leq i] \vee \mathbb{I}\left[i \in \mathcal{S}_{\text{mol}} \wedge j \in \mathcal{S}_{\text{mol}}\right]. \tag{8}$$

Molecular tokens can attend to the entire molecular span and preceding text. Text tokens retain causal attention, with those following the molecular span able to access all molecular tokens.

2.4 Training

Both molecular encoders and the original LLM backbone are frozen during training. We jointly optimize the molecular projector, molecular boundary embeddings, and lightweight LoRA adapters (Hu et al., 2022) attached to the LLM. No separate molecular-language pretraining stage is introduced.

The orthogonality of Q is maintained by retracting it onto the Stiefel manifold (Absil et al., 2008) during optimization. We initialize the nonlinear adaptation branch close to zero and the reversible interaction close to the identity, so training begins near the pretrained molecular representations rather than from strongly distorted features. The positive scales α and β are learned jointly with the projector.

Training uses the standard autoregressive language-modeling objective:

$$\mathcal{L}_{\text{LM}} = -\sum_{t \in \mathcal{Y}} \log p\left(y_t \mid y_{<t}, Z_{\text{mol}}, X\right), \tag{9}$$

where Z_{mol} denotes the projected molecular representations and X denotes the textual context.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate MIP² on four molecular tasks: question answering using MoleculeQA (Lu et al., 2024); description generation and retrosynthesis using the corresponding Mol-Instructions subsets (Fang et al., 2024); and property prediction using the HOMO, LUMO, and GAP subset of Mol-Instructions and eight MoleculeNet datasets (Wu et al., 2018). Property prediction covers classification on BACE, BBBP, ClinTox, HIV, SIDER, and Tox21, and regression on HOMO, LUMO, GAP, ESOL, and Lipophilicity. All tasks use molecular graphs and 3D conformations, while tasks from Mol-Instructions additionally use the supplied SELFIES strings. For MoleculeNet, we use the exact 8:1:1 scaffold split indices adopted by MR-MoL, enabling a split-matched comparison with its reported results. Dataset sizes, preprocessing, and split details are provided in Appendix D.

Baselines. We consider two groups of baseline models. *General-purpose and scientific language models* include the LLaMA and GPT families, Qwen3, and Galactica (Touvron et al., 2023a;b; Achiam et al., 2023; Yang et al., 2025; Taylor et al., 2022). *Molecular language models* include Mol-Instructions (Fang et al., 2024), LlaSMol (Yu et al., 2024), 3D-MoLM (Li et al., 2024), LLaMo (Park et al., 2024), Mol-LLaMA (Kim et al., 2026), MolReasoner (Zhao et al., 2025), HIGHT (Chen et al., 2025), EDT-Former (Jing et al., 2026), and MR-MoL (Park et al., 2026b). The complete baseline lists are provided in the corresponding result tables. Reference scores for question answering, description generation, retrosynthesis, and Mol-Instructions property prediction follow EDT-Former, while MoleculeNet reference scores follow MR-MoL (Park et al., 2026b).

We compare MIP² with an atom-level MLP and the Q-Former from Mol-LLaMA, using its released implementation and pretrained interface weights (Kim et al., 2026). All three projectors use the same molecular encoders and bidirectional molecular attention. MoleculeQA projector comparisons, including the unconstrained variant. Further details are provided in Appendices C and D.

Evaluation metrics. For question answering, we report overall and per-category accuracy (Lu et al., 2024). For description generation, we use BLEU, ROUGE, and METEOR (Papineni et al., 2002; Lin, 2004; Banerjee & Lavie, 2005). For retrosynthesis, we report exact match, validity, and text and fingerprint similarity metrics (Fang et al., 2024). For property prediction, we use ROC-AUC for classification, MAE for HOMO, LUMO, and GAP regression, and RMSE for ESOL and Lipophilicity regression (Fang et al., 2024; Wu et al., 2018). Metric variants, aggregation rules, and invalid-output handling are specified in Appendix E.

Implementation details. Our default backbone is Llama-3.1-8B-Instruct (Llama Team et al., 2024), with frozen MoleculeSTM GIN and Uni-Mol molecular encoders (Xu et al., 2019; Liu et al., 2023a; Zhou et al., 2023). We freeze the original LLM weights and train the projector, molecular boundary embeddings, and LoRA adapters (Hu et al., 2022). We also evaluate MLP and MIP² on Qwen3-8B and Ministral-8B. Architecture, optimization, and decoding settings are provided in Appendices C and D.

3.2 Main Results

Question answering. Our model achieves 71.28% total accuracy on MoleculeQA, exceeding the reported EDT-Former score by 2.94 percentage points, with higher accuracy in all four question categories (Table 1). Obtained using graph and 3D inputs without SMILES or SELFIES strings, these results show that the model can use structural representations to support multiple forms of molecular question answering.

Description generation. On the description subset of Mol-Instructions, our model obtains the highest reported ROUGE-1/2/L and METEOR scores, although its BLEU scores remain below the strongest reference results (Table 2). These results highlight the effectiveness of MIP^2 in generating reference-aligned molecular descriptions, demonstrating its strong capability to translate rich molecular representations into meaningful natural language.

Retrosynthesis. On the retrosynthesis subset of Mol-Instructions, our model achieves 48.7% exact match, compared with the reported 38.7% for EDT-Former, while maintaining 100% validity (Table 3). It also obtains the highest reported RDK, MACCS, and Morgan fingerprint similarities. The agreement between exact-match and fingerprint metrics indicates that predictions more closely match the reference reactants at the molecular-structure level, extending the observed benefits to structured molecular generation.

Property prediction. On the numerical property subset of Mol-Instructions, which covers HOMO, LUMO, and GAP prediction, our model achieves an MAE of 0.003529, approximately 43.1% below the reported EDT-Former value of 0.0062 (Table 2). For MoleculeNet, Table 4 evaluates six classification and two regression datasets using the same $8{:}1{:}1$ scaffold split indices as MR-MoL. MIP^2 obtains numerically better results than MR-MoL on seven datasets and matches it on Tox21. Among the generalist molecular LLMs in Table 4, MIP^2 achieves the best results on BACE, BBBP, ClinTox, HIV, SIDER, and ESOL, ties the best result on Tox21, and ranks second on Lipophilicity. The largest gains over MR-MoL are observed on ClinTox ($+12.2$ ROC-AUC points)

and HIV (+6.9 points). These results demonstrate the effectiveness of MIP^2 in preserving molecular information that is useful for diverse downstream property prediction tasks, spanning both classification and regression.

Table 1: Comparison on the MoleculeQA dataset. Prior work results are reported from EDT-Former. Accuracy (%) is reported for Structure, Source, Property, and Application questions. "T/G/3D" denotes textual molecular representation, molecular graph, and 3D geometry, respectively.

Model	Modality	Setting	Size	Structure	Source	Property	Application	Macro	Total
Random									
Random	–	–	–	24.41	22.30	23.04	24.57	23.58	24.03
Molecular LLMs									
BioMedGPT-LM	Text	SFT	7B	54.19	60.01	38.85	40.90	48.49	52.23
Mol-Inst.-Llama3.1	T/G	SFT	8B	72.79	70.82	43.08	41.22	56.98	65.31
3D-MoLM	T/G/3D	SFT	8B	73.17	70.50	44.79	44.19	58.16	65.96
LLaMo	Text	SFT	8B	70.56	66.63	44.60	45.18	56.74	63.74
Mol-LLaMA3.1	T/G/3D	SFT	8.2B	73.16	70.22	45.70	46.18	58.82	66.21
EDT-Former	T/G/3D	SFT	8.3B	74.55	72.39	50.71	48.58	61.56	68.34
General LLMs									
Galactica	Text	SFT	6.7B	32.35	41.92	31.05	28.21	33.38	33.96
BLOOM	Text	SFT	7.1B	35.01	47.51	31.46	33.56	36.89	37.31
Pythia	Text	SFT	6.9B	42.79	58.90	38.58	39.07	44.84	45.61
Llama-2-chat	Text	SFT	7B	28.75	39.84	31.33	27.71	31.91	31.54
Vicuna-v1.5	Text	SFT	13B	37.01	43.19	30.64	31.55	35.60	37.07
Large-scale General LLMs									
GPT-3.5	Text	10-Shot	–	25.60	37.60	28.04	32.22	30.87	29.29
GPT-4	Text	10-Shot	–	60.94	50.19	35.57	43.91	47.65	53.47
GPT-5	Text	10-Shot	–	62.78	53.22	36.42	46.91	49.83	56.12
Ours	**G/3D**	**SFT**	**8B**	**77.61**	**75.35**	**53.08**	**51.42**	**64.37**	**71.28**

Table 2: Comparison on the description and numerical property subsets of Mol-Instructions. Prior-work results follow EDT-Former.

	Molecular Description (↑)						Property (↓)
Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	MAE
Alpaca-7B	0.068	0.014	0.178	0.041	0.136	0.107	322.109
Baize-7B	0.064	0.015	0.189	0.053	0.148	0.106	261.343
LLaMA-2-7B	0.059	0.014	0.164	0.066	0.148	0.184	5.553
Vicuna-v1.5-13B	0.052	0.011	0.151	0.055	0.130	0.168	860.051
Galactica-6.7B	0.024	0.008	0.074	0.015	0.063	0.065	0.568
Qwen3-8B	0.098	0.029	0.207	0.050	0.157	0.173	4.737
Mol-Inst.-Llama2-7B	0.217	0.143	0.337	0.196	0.291	0.254	0.013
Mol-Inst.-Llama3.1-8B	0.419	0.361	0.719	0.646	0.709	0.637	15.059
Mol-LLaMA2-7.2B	0.433	0.385	0.711	0.649	0.601	0.601	0.0087
Mol-LLaMA3.1-8.2B	**0.445**	0.398	0.717	0.656	0.709	0.617	0.0079
MolReasoner-7B	0.438	0.322	0.553	0.366	0.482	0.475	10.323
EDT-Former-8.3B	0.424	**0.402**	0.726	0.652	0.717	0.631	0.0062
Ours	0.399	0.354	**0.744**	**0.679**	**0.735**	**0.664**	**0.003529**

3.3 Generalization Across LLM Backbones

We compare our projector with an MLP on Llama-3.1-8B-Instruct, Qwen3-8B, and Ministral-8B, using the same molecular encoders and matching the molecular attention mask within each backbone (Table 5). Our projector improves both overall and macro accuracy on all three backbones, with category-level changes varying in magnitude and direction. These aggregate gains support the effectiveness of the projector across the evaluated LLM backbones.

Table 3: Comparison on the retrosynthesis subset of Mol-Instructions. Prior-work results follow EDT-Former. Arrows indicate the preferred direction of each metric.

Model	Exact ↑	BLEU ↑	Levenshtein ↓	RDK FTS ↑	MACCS FTS ↑	Morgan FTS ↑	Validity ↑
Alpaca	0.000	0.063	46.915	0.005	0.023	0.007	0.160
Baize	0.000	0.095	44.714	0.025	0.050	0.023	0.112
ChatGLM	0.000	0.117	48.365	0.056	0.075	0.043	0.046
LLaMA	0.000	0.036	46.844	0.018	0.029	0.017	0.010
Vicuna	0.000	0.057	46.877	0.025	0.030	0.021	0.017
Galactica	0.000	0.452	34.940	0.167	0.274	0.134	0.984
Text+Chem T5	0.141	0.765	24.043	0.685	0.765	0.585	0.698
Mol-Inst.-Llama2	0.009	0.705	31.227	0.283	0.487	0.230	1.000
Mol-Inst.-Llama3.1	0.333	0.842	17.642	0.704	0.815	0.646	1.000
Mol-LLaMA-3.1	0.340	0.877	38.324	0.708	0.822	0.601	1.000
HIGHT	0.008	0.863	28.912	0.564	0.340	0.309	1.000
EDT-Former	0.387	0.930	34.202	0.721	0.836	0.670	1.000
Ours	**0.487**	**0.955**	**11.572**	**0.801**	**0.881**	**0.764**	**1.000**

Table 4: Split-matched comparison with generalist molecular LLMs on eight MoleculeNet datasets. Prior-work scores are reported from MR-MoL. Classification tasks report ROC-AUC (↑), while regression tasks report RMSE (↓).

Model	Base LLM	BACE ↑	BBBP ↑	ClinTox ↑	HIV ↑	SIDER ↑	Tox21 ↑	ESOL ↓	Lipo ↓
GIMLET	T5-small	59.0	52.6	49.9	49.5	46.8	56.3	7.557	1.545
nach0	T5-base	63.6	71.0	49.0	<u>76.1</u>	52.8	50.7	3.745	**0.874**
ChemDFM	LLaMA-13B	70.8	50.0	34.0	69.1	53.0	57.3	17.149	3.497
LlaSMol	Mistral-7B	56.2	64.9	53.0	68.2	56.4	51.2	4.306	0.879
MolecularGPT	Llama-2-7B-chat	61.6	55.7	52.9	69.8	53.6	57.0	12.812	1.149
MR-MoL	Llama-3.1-8B-Instruct	<u>82.6</u>	<u>71.8</u>	<u>73.8</u>	72.0	<u>63.3</u>	**74.2**	<u>1.210</u>	0.919
Ours	Llama-3.1-8B-Instruct	**83.0**	**72.2**	**86.0**	**78.9**	**64.7**	**74.2**	**1.164**	<u>0.918</u>

3.4 Ablation Studies

Projection design and nonlinear adaptation. Under matched bidirectional molecular attention, MIP² achieves $71.14 \pm 0.36\%$ total accuracy across three seeds, compared with $69.73 \pm 0.63\%$ for MLP and $67.66 \pm 0.91\%$ for Q-Former (Table 6). The mean gain over MLP is 1.41 percentage points with the same number of atom tokens, supporting a contribution from the projection design beyond token count. Removing the nonlinear residual reduces accuracy from 71.14% to 70.26% (Table 7). The full model improves structure and property accuracy by 1.00 and 1.87 points, respectively, while source accuracy changes little. The additional gains support learning a nonlinear adaptation for the downstream task. These results highlight the strength of MIP^2 in combining faithful molecular feature preservation with nonlinear adaptation, enabling more effective molecular understanding than conventional projection architectures.

Table 5: Projector comparison across LLM backbones on MoleculeQA. Within each backbone, both projectors use bidirectional molecular attention. Accuracy (%) is reported.

Projector	Backbone	Attention	Structure	Source	Property	Usage/App.	Macro	Total
MLP	Llama-3.1-8B-Instruct	Bidirectional	75.91	73.34	**53.21**	51.09	63.39	69.88
Ours	Llama-3.1-8B-Instruct	Bidirectional	**77.61**	**75.35**	53.08	**51.42**	**64.37**	**71.28**
MLP	Qwen3-8B	Bidirectional	71.31	69.03	50.62	48.08	59.76	65.77
Ours	Qwen3-8B	Bidirectional	**72.92**	**69.77**	**50.89**	**48.25**	**60.46**	**66.85**
MLP	Ministral-8B	Bidirectional	74.94	73.57	**48.84**	**50.08**	61.86	68.75
Ours	Ministral-8B	Bidirectional	**76.23**	**73.64**	**48.84**	49.92	**62.16**	**69.44**

Complementary subspace constraint. To test the complementary-subspace constraint, we replace the orthogonally projected residual $r_\perp(u) = (I_{d_L} - QQ^\top)G_\phi(u)$ in Eq. (3) with the unconstrained output $G_\phi(u)$, keeping all other settings fixed. The unconstrained variant achieves $70.41 \pm 0.61\%$ total and $63.41 \pm 0.36\%$ macro accuracy, 0.73 and 0.67 points below MIP², respectively (Table 7). MIP² gains most on source questions (1.69 points), followed by property and structure, but performs slightly worse on usage. The unconstrained residual exhibits a mean squared norm overlap of 0.874 with the preserved subspace and a nominal relative recovery error of 1.408 for the original encoder features, supporting the complementary-subspace constraint for recoverable nonlinear adaptation (Eq. (2)). See Appendix F.3 for definitions and full diagnostics.

Molecular contextualization. Replacing bidirectional molecular attention with a causal mask lowers total accuracy from 71.14% to 70.57% (Table 7). Bidirectional attention improves all four categories. Allowing each atom token to access the complete molecular span supports the use of context beyond preceding atoms. The projector comparisons share the same attention mask, so their gains are separate from the change in molecular attention.

Table 6: Comparison of projectors on MoleculeQA. Results are reported as mean ± standard deviation over three training seeds. Accuracy (%) is reported.

Projector	Attention	Structure	Source	Property	Usage/App.	Macro	Total
Q-Former	Bidirectional	72.58 ± 1.09	73.14 ± 0.92	51.44 ± 0.73	49.58 ± 0.98	61.68 ± 0.83	67.66 ± 0.91
MLP	Bidirectional	75.84 ± 0.73	73.77 ± 0.80	52.03 ± 1.11	$\mathbf{50.53 \pm 1.42}$	63.04 ± 0.79	69.73 ± 0.63
Ours	Bidirectional	$\mathbf{77.62 \pm 0.63}$	$\mathbf{75.18 \pm 0.30}$	$\mathbf{53.44 \pm 0.63}$	50.03 ± 1.21	$\mathbf{64.08 \pm 0.69}$	$\mathbf{71.14 \pm 0.36}$

Table 7: Component ablations of the proposed molecular projector on MoleculeQA. Results are reported as mean ± standard deviation over three training seeds. Accuracy (%) is reported.

Variant	Attention	Structure	Source	Property	Usage/App.	Macro	Total
Ours w/o nonlinear residual	Bidirectional	76.62 ± 0.18	74.83 ± 0.85	51.57 ± 2.01	49.75 ± 1.04	63.20 ± 0.54	70.26 ± 0.23
Ours (unconstrained)	Bidirectional	77.10 ± 0.82	73.49 ± 0.70	52.90 ± 0.85	$\mathbf{50.14 \pm 0.70}$	63.41 ± 0.36	70.41 ± 0.61
Ours	Causal	77.09 ± 0.35	74.71 ± 0.53	52.53 ± 0.41	49.42 ± 1.43	63.43 ± 0.29	70.57 ± 0.27
Ours	Bidirectional	$\mathbf{77.62 \pm 0.63}$	$\mathbf{75.18 \pm 0.30}$	$\mathbf{53.44 \pm 0.63}$	50.03 ± 1.21	$\mathbf{64.08 \pm 0.69}$	$\mathbf{71.14 \pm 0.36}$

Complementarity of 2D and 3D information. Combining 2D and 3D features improves MoleculeQA total accuracy over either modality alone (70.67% versus 69.39% and 69.86%); MIP² further reaches 71.28%. The full ablation and category analysis appear in Appendix F.1 (Table 11).

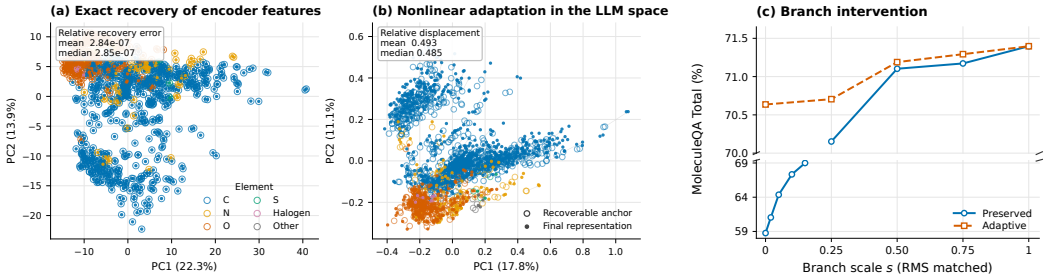

Figure 2: **Feature preservation and nonlinear adaptation.** (a) The analytical inverse recovers encoder features to numerical precision. (b) Final embeddings differ from their recoverable anchors. (c) Inference-time scaling of the preserved or adaptive branch, with molecular embedding RMS matched; $s = 1$ is the unperturbed scale.

Projection design on structured generation. On retrosynthesis, we hold the molecular encoders, attention mask, training pipeline, and decoding protocol fixed (Appendix F.2, Table 12). MIP² achieves 48.7% exact match, exceeding MLP by 4.4 percentage points and Q-Former by 6.0 points. MIP² also improves all three fingerprint similarities, including Morgan similarity from 0.7415 to

0.7635 relative to MLP. The improvements in exact match and structural similarity extend the benefit of the projection design from question answering to molecular generation.

3.5 Representation and Mechanism Analysis

Preservation without sacrificing adaptation. Figure 2(a,b) shows feature recovery and the nonlinear changes to the projected embeddings. The analytical inverse yields a mean relative recovery error of 2.84×10^{-7}, while the mean relative displacement from the recoverable anchor is 0.493. The results verify the separation between feature preservation and representation adaptation, while the interventions below further characterize the way that trained model uses these components.

Inference-time intervention on the two branches. We rescale each branch at inference time while matching molecular embedding RMS (Fig. 2(c)). Removing the preserved branch lowers total accuracy from 71.28% to 58.76% (-12.52 points), while scaling it to 0.25 yields 70.15%. This phenomenon shows that the model uses preserved features beyond their analytical recoverability. The adaptive branch has a smaller effect, Appendix F.5 further supports retaining the learned residual.

Qualitative analysis of feature retention and molecular reasoning. In Fig. 3, MIP² identifies a tetramethyl-substituted tricyclic alkene with an exocyclic methylene (left) and a pyrazole (right), while MLP and the unconstrained variant answer incorrectly. MIP² retains 75/80 and 118/130 encoder feature-neighbor relations, versus 72/80 and 101/130 for MLP and 67/80 and 105/130 for the unconstrained variant. These cases complement the aggregate results in Tables 6 and 7.

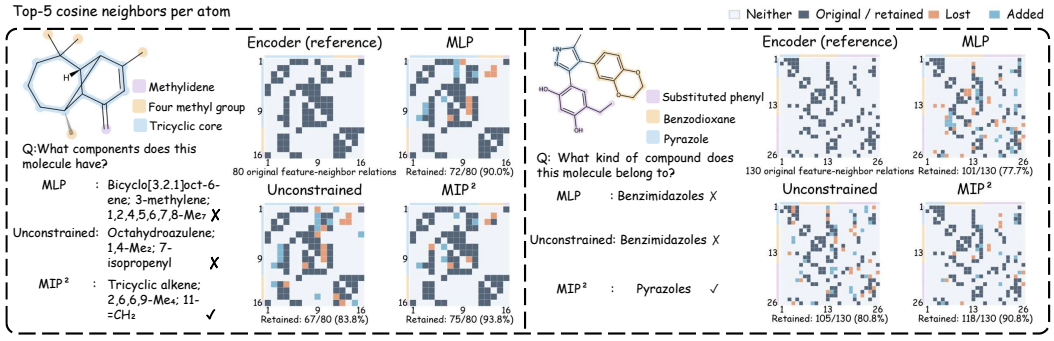

Figure 3: **Molecular question answering and encoder neighborhood retention in two illustrative examples.** Left: molecular components; right: compound class. Answers for each model appear beside each molecule. Matrices compare each atom's top-five cosine feature neighbors against the encoder reference: dark cells are retained, orange lost, and blue added relations.

Efficiency. MIP² incurs only a modest training overhead relative to the atom-level MLP, while maintaining nearly identical inference latency (77.72 vs. 77.74 ms/sample) and peak inference memory (22.95 vs. 22.94 GiB). MIP² also uses substantially fewer parameters than Q-Former (11.96M vs. 150.93M), indicating that explicit feature recoverability can be achieved without a large inference-time overhead. Further details please refer to Appendix F.6.

4 Conclusion

We present MIP², an information-preserving projector for molecular LLMs. MIP² combines reversible 2D-3D fusion with a projection that separates feature preservation and nonlinear adaptation into orthogonal subspaces. MIP² guarantees exact recovery of the input encoder features while allowing flexible adaptation to the LLM embedding space. Atom-level tokens and bidirectional molecular attention further support structural interactions. Experiments on molecular question answering, description generation, retrosynthesis, and property prediction demonstrate the effectiveness of MIP². Projector comparisons across LLM backbones and component ablations support the utility of combining information preservation with nonlinear adaptation for molecular-language alignment.

AI USE STATEMENT

Generative AI tools were used only for recommended-disclosure tasks, including language polishing, improving readability and manuscript organization, identifying relevant literature, refining scientific figures, and suggesting clearer presentation. They were not used to formulate or modify research hypotheses, design the core methodology or experiments, construct theoretical or conceptual frameworks, formulate mathematical claims or proofs, implement the proposed method, translate research content, preprocess or clean datasets, perform qualitative analysis, interpret experimental results, or generate synthetic data. All AI-assisted content was manually reviewed, verified, and revised by the authors. We take full responsibility for the final content of this work.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 354–379, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827, 2024.

Yongqiang Chen, Quanming Yao, Juzheng Zhang, James Cheng, and Yatao Bian. Hierarchical graph tokenization for molecule-language alignment. In *Forty-second International Conference on Machine Learning*, 2025.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations*, 2024.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zihao Jing, QIUHAO Zeng, Ruiyi Fang, Yan Sun, Boyu Wang, and Pingzhao Hu. Entropy-guided dynamic tokens for graph-LLM alignment in molecular understanding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. Mol-llama: Towards general understanding of molecules in large molecular language model. *Advances in Neural Information Processing Systems*, 38:26921–26960, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, Dec 2023a. ISSN 2522-5839.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Conference on Empirical Methods in Natural Language Processing*, 2023b.

Llama Team, Abhimanyu Dubey, Aaron Grattafiori, Kunal Bhalla, Soumya Chaudhuri, et al. The llama 3 herd of models, 2024.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. MoleculeQA: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3769–3789, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. Llamo: Large language model-based molecular graph assistant. *Advances in Neural Information Processing Systems*, 37: 131972–132000, 2024.

Jinyoung Park, Minseong Bae, Jeehye Na, and Hyunwoo J. Kim. Improving large molecular language model via relation-aware multimodal collaboration. *arXiv preprint arXiv:2601.12256*, 2026a.

Junwoo Park, Minyoung Shin, Cheol Soon Lee, and Sujee Lee. Multi-granular rationale-guided molecular llm for property prediction. *arXiv preprint arXiv:2608.10480*, 2026b.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

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

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*, 2024.

Yikun Zhang, Geyan Ye, Chaohao Yuan, Bo Han, Long-Kai Huang, Jianhua Yao, Wei Liu, and Yu Rong. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *International Conference on Learning Representation*, volume 2025, pp. 3866–3894, 2025.

Guojiang Zhao, Zixiang Lu, Yutang Ge, Sihang Li, Zheng Cheng, Haitao Lin, Lirong Wu, Hanchen Xia, Hengxing Cai, Wentao Guo, et al. Molreasoner: Toward effective and interpretable reasoning for molecular llms. *arXiv preprint arXiv:2508.02066*, 2025.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

A Related Work

Molecular large language models. Molecular representation learning has progressed from graph-based and geometric pretraining to structure-text alignment. Mole-BERT learns representations from molecular graphs, Uni-Mol incorporates 3D molecular geometry, and MoleculeSTM aligns molecular structures with textual descriptions Xia et al. (2023); Zhou et al. (2023); Liu et al. (2023a). Building on such representations, molecular LLMs connect structural inputs to language models in different ways: MolCA projects molecular graph features into an LLM, InstructMol uses multimodal instruction tuning, and 3D-MoLM introduces atom-level geometric information Liu et al. (2023b); Cao et al. (2025); Li et al. (2024). LLaMo further incorporates graph and motif features, while Mol-LLaMA and CoLLaMo combine complementary 2D and 3D molecular representations Park et al. (2024); Kim et al. (2026); Park et al. (2026a). As pretrained encoders capture increasingly diverse structural cues, the molecular-language projector must do more than produce LLM-compatible embeddings. The projector also determines which distinctions in the original encoder features remain accessible after projection.

Molecular projectors. Projectors bridge molecular encoder outputs and the LLM embedding space through different aggregation and adaptation strategies. Inspired by BLIP-2, MolCA and 3D-MoLM use Q-Former-style modules to summarize molecular representations with learned queries Li et al. (2023); Liu et al. (2023b); Li et al. (2024). Other work explores structural granularity: HIGHT organizes atom, motif, and molecule-level tokens, EDT-Former allocates tokens according to molecular structure, and Atomas studies fine-grained molecule-text alignment at multiple semantic levels Chen et al. (2025); Jing et al. (2026); Zhang et al. (2025). Beyond molecular applications, recent multimodal projectors have investigated the trade-off between feature aggregation and information preservation. Honeybee emphasizes local context preservation, while TokenPacker combines token compression with fine-grained feature aggregation Cha et al. (2024); Li et al. (2025). Jiang et al. further analyze how projector-based compression affects fine-grained vision-language alignment Jiang et al. (2025). These studies highlight the importance of preserving structural details during cross-modal adaptation. However, effective aggregation and alignment do not themselves guarantee that pretrained encoder features can be reconstructed from the projected representations. This distinction motivates our investigation of molecular-language projection that combines explicit feature recoverability with flexible nonlinear adaptation.

B Formal Recoverability Analysis

This section provides detailed derivations of the theoretical properties introduced in Section 2.2. We first derive the closed-form solution to the constrained nonlinear residual, then establish an explicit left inverse for the complete molecular projector. We further examine feature separation and numerical recoverability. Let $h = [h^{2D}; h^{3D}] \in \mathbb{R}^{d_M}$, let $F_\theta : \mathbb{R}^{d_M} \to \mathbb{R}^{d_M}$ be the reversible cross-modal coupling, and let $Q \in \mathbb{R}^{d_L \times d_M}$ have orthonormal columns. The projector is defined as follows:

$$P(h) = \alpha Q F_\theta(h) + \beta(I - QQ^\top)G_\phi(F_\theta(h)), \qquad Q^\top Q = I, \tag{10}$$

where $\alpha > 0$ and G_ϕ may be an arbitrary nonlinear map.

Constrained residual. For fixed u and Q, the Lagrangian of Eq. (2) is $\mathcal{L}(v, \lambda) = \frac{1}{2}\|v - G_\phi(u)\|_2^2 + \lambda^\top Q^\top v$. Stationarity and feasibility give

$$v - G_\phi(u) + Q\lambda = 0, \qquad Q^\top v = 0. \tag{11}$$

Since $Q^\top Q = I$, we obtain $\lambda = Q^\top G_\phi(u)$ and $v = (I - QQ^\top)G_\phi(u)$. Strict convexity makes this the unique minimizer.

Proposition 1 (explicit left inverse). If F_θ is invertible, $Q^\top Q = I$, and $\alpha \neq 0$, then P is injective. A left inverse on the image of P is given by

$$R(z) = F_\theta^{-1}\left(\alpha^{-1} Q^\top z\right). \tag{12}$$

Table 8: Dimensions and initialization of the default molecular interface.

Component	Input	Output	Implementation
MoleculeSTM GIN	graph	300	Frozen, atom-level 2D features
Uni-Mol	coordinates	512	Frozen, atom-level 3D features
Joint atom feature	$300 + 512$	812	Concatenation after atom alignment
f_θ	300	512	$300 \to 512 \to 512$, SiLU
g_θ	512	300	$512 \to 300 \to 300$, SiLU
Orthogonal lift Q	812	4096	Column-orthogonal, trainable
G_ϕ	812	4096	$812 \to 1624 \to 4096$, SiLU
LLM molecular tokens	N atoms	N tokens	One projected token per heavy atom

Proof. For $u = F_\theta(h)$, multiplication by $Q^\top$ removes the nonlinear residual as follows:

$$\begin{aligned} Q^\top P(h) &= \alpha Q^\top Q u + \beta Q^\top (I - QQ^\top) G_\phi(u) \\ &= \alpha u + \beta (Q^\top - Q^\top Q Q^\top) G_\phi(u) \\ &= \alpha u. \end{aligned} \tag{13}$$

It follows that $R(P(h)) = F_\theta^{-1}(F_\theta(h)) = h$. A map admitting a left inverse is injective. The result is independent of β and of the architecture or parameters of G_ϕ. □

Inverse of the reversible 2D-3D interaction. The coupling used in our implementation updates the two modalities as follows:

$$u^{3D} = h^{3D} + f_\theta(h^{2D}), \qquad u^{2D} = h^{2D} + g_\theta(u^{3D}). \tag{14}$$

Its inverse does not require either neural network to be invertible:

$$h^{2D} = u^{2D} - g_\theta(u^{3D}), \qquad h^{3D} = u^{3D} - f_\theta(h^{2D}). \tag{15}$$

Consequently, both pretrained encoder representations are recovered, rather than only their coupled coordinates. Multiple coupling blocks remain invertible by applying these inverses in reverse order.

Feature separation. For $\Delta u = u_1 - u_2$ and $\Delta r = r_\perp(u_1) - r_\perp(u_2)$, $Q^\top \Delta r = 0$ and $\|Q\Delta u\|_2 = \|\Delta u\|_2$. Expanding $\|\alpha Q \Delta u + \beta \Delta r\|_2^2$ therefore gives Eq. (6), since its cross term vanishes.

Finite-precision behavior. The guarantee above holds in exact arithmetic. We compute the complementary projection in FP32 even under BF16 training and retract Q after every optimizer update. The measured relative error $\|R(P(h)) - h\|_F / \|h\|_F$ is on the order of 10^{-7} (Section F.5), indicating that numerical error is negligible at the scale used in our experiments. Recoverability requires $d_M \le d_L$; our default setting satisfies this condition with $d_M = 812$ and $d_L = 4096$.

C Architecture and Implementation Details

C.1 Molecular encoders and atom alignment

The 2D branch is the five-layer GIN distributed with MoleculeSTM and produces a 300-dimensional feature for each heavy atom. The 3D branch uses the Uni-Mol molecular pretraining checkpoint `mol_pre_no_h_220816.pt` and produces a 512-dimensional feature for the same atom. Both encoders remain frozen in all experiments. We canonicalize each molecule once and apply the resulting atom permutation to both graph topology and coordinates. Atom symbols are checked before reordering, and the model raises an error if the two modalities do not contain the same number of atom representations. Concatenation is therefore performed only between features associated with the same heavy atom.

C.2 Initialization and scale calibration

We initialize Q with orthonormal columns. The final linear layers of f_θ, g_θ, and G_ϕ are initialized to zero, so the coupling begins as the identity and the complementary nonlinear branch initially

vanishes. The positive scales are parameterized in log space. Before optimization, we estimate a scale s_0 from the first 32 training batches:

$$s_0 = \frac{\mathrm{mean}_{x \in \mathcal{T}} \, \mathrm{RMS}(E_{\mathrm{text}}(x))}{\mathrm{mean}_{h \in \mathcal{M}} \, \mathrm{RMS}(Qh)}, \tag{16}$$

where $\mathcal{T}$ excludes padding and molecular placeholders and $\mathcal{M}$ contains molecular tokens. The data-loader RNG state is restored after calibration, ensuring that calibration does not change the subsequent training order. For the main MoleculeQA run, the text and unscaled molecular RMS values were 0.007781 and 0.390301, giving $s_0 = 0.019937$. After each optimizer step, a reduced QR decomposition retracts Q onto the Stiefel manifold. We correct column signs using the diagonal of the triangular factor, which avoids arbitrary sign flips across updates.

C.3 Molecular attention mask and cached decoding

The attention mask is constructed as an explicit Boolean tensor of shape $B \times 1 \times L \times L$. Text queries retain the decoder's causal rule. Queries inside the molecular interval additionally see every key in that interval, including later serialized atoms, but never see later question or answer tokens. Padding positions are masked as keys for all valid queries. A padded query receives a harmless self-edge to avoid an all-masked row in scaled dot-product attention; its output is discarded.

Generation first performs a complete prompt prefill using this block-causal mask. Subsequent tokens use the ordinary causal key-value cache. For a right-padded batch, the first generated token is selected from the last valid prompt position of each example, and cached RoPE positions continue from the per-example number of valid tokens. Our implementation prevents padding slots from shifting generation positions and preserves the same logits as a full causal recomputation after the molecular prefill.

C.4 Trainable parameters and optimization

The original LLM weights and both molecular encoders are frozen. We train the projector, the two molecular boundary embeddings, and LoRA adapters. Unless a task-specific setting is stated below, LoRA has rank 16, scale 32, dropout 0.05, and is attached to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `down_proj`, and `up_proj`. Optimization uses AdamW, BF16 autocasting, gradient checkpointing, a maximum gradient norm of 1.0, 100 warmup updates, and a cosine learning-rate schedule. The input limit is 1,536 tokens except for MoleculeNet, where it is 1,568 tokens. Training labels cover only assistant tokens; prompt and padding tokens receive the ignore label. The effective batch size is 32. Unless stated otherwise, evaluation uses deterministic greedy decoding; dataset-specific generation limits are given below.

The implementation environment uses Python 3.10.18, PyTorch 2.4.1 with CUDA 12.1, Transformers 4.48.3, PEFT 0.10.0, PyTorch Geometric 2.6.1, and RDKit 2025.9.1. The runtime measurements in Table 16 were obtained on a 48-GiB NVIDIA RTX 4090.

C.5 Controlled projector baselines

The MLP baseline retains one token per atom and uses $812 \to 1624 \to 4096$ with GELU activation. It has no recoverability constraint. The Q-Former baseline reproduces the released Mol-LLaMA interface: separate layer normalization for 300-dimensional MoleculeSTM and 512-dimensional Uni-Mol features, a four-layer multimodal blending module with hidden size 512 and eight heads, eight learned queries in a SciBERT-initialized Q-Former, and a $768 \to 4096$ LLM projection. We load all 521 matching interface tensors from the released Mol-LLaMA checkpoint. In controlled comparisons, the LLM LoRA adapters are initialized from scratch for every variant; only the Q-Former interface uses its released projector initialization. The linear recoverable ablation disables G_ϕ while retaining the reversible 2D-3D coupling, trainable orthogonal lift, scale calibration, LoRA configuration, and molecular attention mask.

The unconstrained MIP^2 variant retains the same architecture and training settings, including the QR update of Q, and replaces only $(I - QQ^\top)G_\phi(u)$ with $G_\phi(u)$.

Table 9: Data, prediction targets, and reported metrics. Detailed metric definitions follow in Appendix E.

Dataset or subset	Prediction target	Metrics
MoleculeQA	Multiple-choice answer	Overall, category, and macro accuracy
Mol-Instructions: description	Molecular description	BLEU-2/4, ROUGE-1/2/L, METEOR
Mol-Instructions: property	HOMO, LUMO, GAP	Macro MAE
Mol-Instructions: retrosynthesis	Reactant molecules	Exact match, validity, BLEU, Levenshtein distance, RDK/MACCS/Morgan similarity
MoleculeNet: BACE, BBBP, ClinTox, HIV, SIDER, Tox21	Classification endpoint	ROC-AUC
MoleculeNet: ESOL, Lipophilicity	Numerical property	RMSE

Table 10: Dataset sizes and task-specific training settings. Effective batch size is micro-batch size times gradient accumulation. MoleculeNet counts are query rows across 46 endpoints rather than unique molecules.

Dataset/subset	Train	Val.	Test	Epochs	Steps/epoch	Eff. batch	Input	Max atoms
MoleculeQA	49,984	0	5,786	3	1,562	32	G/3D	600
Mol-Instr. description	293,541	–	989	3	9174	32	SELFIES/G/3D	600
Mol-Instr. property	359,613	500	1,987	3	11,238	32	SELFIES/G/3D	128
Mol-Instr. retrosynthesis	128,683	–	1,000	3	4,022	32	SELFIES/G/3D	128
MoleculeNet	136,670	16,306	16,210	4†	4,271	32	G/3D	600

†MoleculeNet was configured for at most five epochs. Training stopped after the first validation-NLL degradation at epoch 4 and selected the preceding epoch-3 checkpoint.

D Datasets and Evaluation Protocols

D.1 Dataset coverage and comparison sources

Table 9 summarizes the correspondence between data, prediction targets, and evaluation metrics. We use the official 5,786-question MoleculeQA test set, the selected molecule-oriented Mol-Instructions subsets, and eight MoleculeNet datasets. The MoleculeNet splits use deterministic Bemis-Murcko scaffolds. We use the exact $8{:}1{:}1$ scaffold split indices adopted by MR-MoL.

Reference results for MoleculeQA and Mol-Instructions follow EDT-Former Jing et al. (2026); those for MoleculeNet follow MR-MoL Park et al. (2026b). These are published reference scores, not reproductions under our training setup. The main tables retain the reported model, modality, and backbone information where available. Our projector comparisons evaluate MLP, Q-Former, MIP², and its unconstrained variant with matched molecular encoders and bidirectional molecular attention. The MoleculeQA projector comparison and component ablations reports mean and standard deviation over three training seeds. Initialization differences, including the released Q-Former checkpoint, are documented in Appendix C. The cross-backbone comparison evaluates MLP and MIP² on Llama-3.1-8B-Instruct, Qwen3-8B, and Ministral-8B.

D.2 MoleculeQA

We retain all 49,984 usable official training questions and evaluate on the untouched 5,786-question official test set. Nine source records are rejected because the molecule is invalid or exceeds the 600-heavy-atom limit. The input contains the 2D graph and supplied 3D conformer but no SMILES or SELFIES. The four answer categories contain 3,113 structure, 1,343 source, 731 property, and 599 usage questions in the test set. We train for three complete epochs with micro-batch size 4, gradient accumulation 8, learning rate 10^{-4}, weight decay 0.01, and training seeds $\{1, 2, 4\}$ for the reported multi-seed result. Checkpoints are placed only at complete epoch boundaries. The main single-run result uses the final epoch-3 checkpoint rather than selecting an epoch using the test set. Evaluation uses deterministic greedy decoding with a batch size of 32 and at most eight generated tokens. We

parse the first valid A/B/C/D answer and count an example as correct only when it matches the reference option; invalid generations count as incorrect.

D.3 Mol-Instructions property prediction and retrosynthesis

Property prediction follows the official HOMO, LUMO, and GAP examples. The model receives the official SELFIES string followed by the aligned graph and conformer tokens. We generate at most 16 tokens and accept only a finite standalone number. All 1,987 test generations are valid numbers. Invalid outputs, if any, receive a property-specific penalty equal to the training range. Macro MAE averages the three per-property MAEs, while micro MAE pools all examples.

For retrosynthesis, the input consists of the official product SELFIES and its molecular tokens; SELFIES strings are not truncated. We generate at most 512 tokens using greedy decoding. Of 128,684 candidate training records, one is rejected as invalid or too large, leaving 128,683 examples. Coordinates are generated deterministically with ETKDGv3; 592 training molecules use a 2D coordinate fallback after failed 3D embedding. The official 1,000-example test set is left unchanged. The validation path is the test file only because training does not perform validation or checkpoint selection; all reported results use the complete epoch-3 checkpoint.

The molecular-description data contain 293,541 training and 989 test examples. The result in Table 2 uses the official SELFIES, graph, and 3D input described in the main text.

D.4 MoleculeNet

We use the original 8:1:1 scaffold split indices adopted by MR-MoL for BACE, BBBP, ClinTox, HIV, SIDER, Tox21, ESOL, and Lipophilicity. For the multi-label ClinTox, SIDER, and Tox21 datasets, each labeled endpoint is expanded into a separate query, following the MR-MoL protocol. The resulting joint data contain 57,234 unique molecules and 46 prediction endpoints. Molecules receive deterministic ETKDGv3 coordinates, followed by MMFF or UFF optimization when parameters are available. Molecules with more than 128 atoms and failed embeddings use deterministic 2D coordinates.

All eight datasets are trained jointly. Classification prompts request `Yes` or `No`. At evaluation, the score is the softmax probability of `Yes` computed from the two first-answer-token logits; ROC-AUC is computed per endpoint and then averaged within a multi-endpoint dataset. ESOL and Lipophilicity targets are standardized using the training mean and standard deviation. The model greedily generates a standardized number, which is transformed back to the original scale before RMSE is computed. The MoleculeNet run uses LoRA only on the four attention projections with zero LoRA dropout. Both LoRA and projector learning rates are 10^{-4}. Validation NLL decreases from 0.4140, to 0.3666, to 0.3507 over the first three epochs and rises to 0.3553 at epoch 4; we therefore report the epoch-3 test checkpoint.

E Metric Definitions

Question answering. On MoleculeQA, overall accuracy is the fraction of correct answers over all 5,786 questions. We also report accuracy for each of the four categories. Macro-averaged accuracy is their unweighted mean. The usage category is labeled Application or Usage/App. in the main tables.

Description generation. On the description subset of Mol-Instructions, we report BLEU-2 and BLEU-4 Papineni et al. (2002), ROUGE-1, ROUGE-2, and ROUGE-L Lin (2004), and METEOR Banerjee & Lavie (2005), using the standard corpus evaluators used by Mol-Instructions.

Retrosynthesis. Retrosynthesis follows the public Mol-Instructions evaluator. Generated SELFIES are trimmed to the last complete token and decoded to SMILES. Validity is the fraction that can be decoded and parsed by RDKit. Exact match compares InChI strings among valid predictions; we also record canonical exact match over all samples. BLEU is character-level corpus BLEU over valid SELFIES. Levenshtein distance is computed between canonical SMILES. RDK, MACCS, and

radius-2 Morgan scores are Tanimoto similarities between the corresponding RDKit fingerprints. Invalid predictions are excluded by the public evaluator from all metrics except validity and all-sample canonical exact match.

Property prediction. For the numerical property subset of Mol-Instructions, we first compute MAE separately for HOMO, LUMO, and GAP, then report the unweighted mean of these three MAEs in the main table. This macro average differs from pooling all test examples into one micro MAE. Per-property errors are given in Table 17; the numerical-output parser and invalid-output penalty are specified in Appendix D.

For MoleculeNet, we report ROC-AUC for BACE, BBBP, ClinTox, HIV, SIDER, and Tox21. For a dataset with multiple classification endpoints, we compute ROC-AUC separately for each endpoint and average within the dataset. We report RMSE for ESOL and Lipophilicity after converting generated predictions back to the original target scale. Scoring and target-standardization details are provided in Appendix D.

F Additional Results and Mechanism Checks

F.1 Complementarity of 2D and 3D information

Concatenating 2D and 3D features yields 70.67% total accuracy, exceeding the variants using only 2D or 3D features by 1.28 and 0.81 percentage points, respectively (Table 11). The combined representation benefits from both topology and geometry, with higher aggregate accuracy than either modality alone. MIP^2 further raises total accuracy to 71.28% and macro accuracy from 64.00% to 64.37%, supporting the value of learned interaction between the modalities before projection. The benefits of combining and fusing the modalities therefore vary across question categories.

Table 11: Modality ablation on MoleculeQA. We compare 2D, 3D, and the full model using both modalities. Accuracy (%) is reported.

Variant	2D	3D	Structure	Source	Property	Usage/App.	Macro	Total
2D only	✓		75.27	73.94	52.94	48.75	62.72	69.39
3D only		✓	75.36	74.61	52.53	**51.75**	63.56	69.86
2D & 3D Concat	✓	✓	76.65	75.06	**53.21**	51.09	64.00	70.67
Ours	✓	✓	**77.61**	**75.35**	53.08	51.42	**64.37**	**71.28**

F.2 Controlled retrosynthesis projector comparison

The projector variants share the molecular encoders, bidirectional attention, training pipeline, and decoding protocol; their Mol-Instructions retrosynthesis scores are reported in Table 12.

Table 12: Controlled projector ablation on Mol-Instructions retrosynthesis. All variants use the same bidirectional molecular contextualization. Arrows indicate the preferred direction of each metric.

Projector	Exact ↑	BLEU ↑	Levenshtein ↓	RDK FTS ↑	MACCS FTS ↑	Morgan FTS ↑	Validity ↑
Q-Former	0.427	0.9480	13.302	0.7683	0.8628	0.7288	0.999
MLP	0.443	0.9484	12.550	0.7810	0.8696	0.7415	1.000
Ours	**0.487**	**0.9548**	**11.572**	**0.8012**	**0.8814**	**0.7635**	**1.000**

F.3 Effect of the complementary subspace constraint

We construct the unconstrained variant by removing the projection of the nonlinear residual while retaining the reversible fusion, orthogonal anchor, nonlinear network, molecular attention, and training protocol. The projected embedding becomes

$$z_{\text{unc}} = \alpha Q u + \beta G_\phi(u), \qquad u = F_\theta(h). \tag{17}$$

The comparison uses three training seeds, with mean and standard deviation reported in Table 6. MIP² reaches $71.14 \pm 0.36\%$ total and $64.08 \pm 0.69\%$ macro accuracy, compared with $70.41 \pm 0.61\%$ and $63.41 \pm 0.36\%$ for the unconstrained variant. The gains of 0.73 and 0.67 percentage points support the complementary subspace constraint with the nonlinear network and remaining architecture held constant. The ablation evaluates the constraint $Q^\top r(u) = 0$ on the downstream task, with Eq. (2) providing the closest feasible residual to a given nonlinear update.

We measure residual overlap with the preserved subspace and residual magnitude relative to the anchor as

$$\rho = \frac{\|Q^\top G_\phi(u)\|_2}{\|G_\phi(u)\|_2}, \qquad \gamma = \frac{\mathrm{RMS}(\beta G_\phi(u))}{\mathrm{RMS}(\alpha Q u)}. \tag{18}$$

With $Q^\top Q = I$, the squared overlap ρ^2 equals $\|QQ^\top G_\phi(u)\|_2^2 / \|G_\phi(u)\|_2^2$. The reported mean of 0.874 therefore indicates a large proportion of residual squared norm in the preserved subspace, while $\gamma = 1.865$ indicates a residual RMS greater than the anchor RMS (Table 13). Unconstrained adaptation makes substantial use of the preserved coordinates, with only 12.6% of squared norm in the complementary subspace on average.

Applying the nominal recovery operator to Eq. (17) gives

$$\hat{u} = \alpha^{-1} Q^\top z_{\mathrm{unc}} = u + \frac{\beta}{\alpha} Q^\top G_\phi(u), \qquad \hat{h} = F_\theta^{-1}(\hat{u}). \tag{19}$$

The nominal relative recovery error for the original encoder features, $\|\hat{h} - h\|_2 / \|h\|_2$, is 1.408 ± 0.325. The error measures failure of the specified recovery map, rather than irreversible information loss or noninjectivity of the unconstrained projector. For MIP², the projected residual satisfies $Q^\top r_\perp(u) = 0$, maintaining the recovery identity independently of the nonlinear network parameters.

Table 13: Mechanism diagnostics for the unconstrained MIP² variant. Values are reported as mean $\pm$ standard deviation across three training seeds. Overlap is measured on $G_\phi(u)$, and nominal recovery is evaluated against the original encoder features h. The recovery error and squared overlap are distinct quantities.

Diagnostic	Mean $\pm$ standard deviation
Preserved subspace overlap ρ	0.935 ± 0.008
Squared norm overlap ρ^2	0.874 ± 0.016
Nominal original feature recovery error	1.408 ± 0.325
Residual to anchor RMS ratio γ	1.865 ± 0.410

F.4 MoleculeQA optimization trajectory

Table 14 reports the fixed epoch checkpoints for the main run. Accuracy improves across all three complete epochs, and the final checkpoint is used in the main comparison.

F.5 Intervention on nonlinear residual strength

For the intervention in Fig. 2(c), we multiply the trained residual by an inference-time factor s while leaving every parameter and the recoverable channel fixed. Table 15 gives the source values. Relative displacement is $\|s\beta r_\perp(u)\|_F / \|\alpha Q u\|_F$.

Using one million paired nonparametric bootstrap resamples over the 5,786 test questions, the difference between $s = 1$ and $s = 0$ is 0.50 percentage points with a 95% percentile interval of $[0.02, 0.99]$ points. The bootstrap probability that the paired difference is positive is 0.976. This comparison isolates the functional contribution of the learned nonlinear branch within a single checkpoint; it is not a substitute for independently trained ablations.

F.6 Efficiency

MIP² uses 11.96M projector parameters, compared with 7.98M for MLP and 150.93M for Q-Former (Table 16). Relative to MLP, training throughput decreases by 3.7% and training time increases by

Table 14: MoleculeQA test accuracy (%) across complete training epochs for the main seed-1 run. Checkpoints are not selected using these test results.

Checkpoint	Structure	Source	Property	Usage	Total
Epoch 1	72.86	71.78	48.97	48.75	67.09
Epoch 2	76.13	74.98	50.89	50.58	70.03
Epoch 3	77.61	75.35	53.08	51.42	71.28

Table 15: Inference-time residual intervention on the MoleculeQA epoch-3 checkpoint. Recovery remains at numerical precision while displacement and accuracy vary with residual strength.

s	Rel. disp.	Recovery error	Structure	Source	Property	Usage	Total
0.00	0.0000	4.27×10^{-7}	76.87	75.73	51.71	51.25	70.77
0.25	0.0889	4.22×10^{-7}	77.06	75.35	51.98	51.25	70.83
0.50	0.1779	4.27×10^{-7}	77.61	75.35	51.71	51.92	71.15
1.00	0.3557	4.65×10^{-7}	77.61	75.35	53.08	51.42	71.28
2.00	0.7115	6.21×10^{-7}	77.61	75.28	52.94	51.09	71.21

Table 16: Training and inference efficiency of the projector variants in our implementation. Throughput is measured in samples/s, latency in ms/sample, and peak memory in GiB.

Method	Params (M)	Train Throughput	Train Memory	Train Time	Infer Latency	Infer Throughput	Infer Memory
MLP	7.98	6.14	23.37	6:47:11	77.74	12.86	22.94
Q-Former	150.93	6.02	25.13	6:55:17	67.08	14.91	23.06
Ours	11.96	5.91	**23.32**	7:02:49	77.72	12.87	22.95

Table 17: Per-property errors on the 1,987-example Mol-Instructions property test set. Every generation is a valid finite number.

Property	Examples	MAE ↓	RMSE ↓
HOMO	684	0.003368	0.006298
LUMO	642	**0.003129**	**0.004581**
GAP	661	0.004091	0.008366
Macro	1,987	0.003529	0.006415

3.8%, while inference latency is nearly identical at 77.72 versus 77.74 ms per sample. Peak inference memory differs by 0.01 GiB, indicating that the recoverable construction adds little inference cost to the atom level MLP interface in the measured setting. Q-Former has lower inference latency at 67.08 ms per sample despite its larger parameter count, so parameter reduction alone does not imply faster decoding.