AMMAS: MULTIMODAL MODELING FOR ATOMISTIC SCIENCE WITH 3D STRUCTURES AND LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

In atomistic science, the spatial arrangement of atoms determines the properties and interactions of microscopic 3D structures. Natural language provides a complementary interface for expressing structural concepts, questions, and desired geometric changes. Coupling these modalities would enable geometry-grounded 3D structure understanding and language-guided 3D structure generation. However, a modality gap remains: language modeling operates over discrete token sequences, whereas geometric modeling operates over continuous atomic coordinates. To bridge this gap, we propose **A** **M**ultimodal **M**odel for **A**tomistic **S**cience (AMMAS), which supports autoregressive generation of both discrete text and continuous 3D structures. AMMAS interleaves stacks of Shared Generalist Transformer (SGT) layers with Modality Expert Transformer (MET) layers. SGT layers jointly update the mixed sequence of atomistic and textual representations, enabling cross-modal interaction. MET layers then use modality-specific positional encoding and attention patterns to update the text-contextualized atomistic representations of each 3D structure. The updated atomistic representations feed into subsequent SGT layers, establishing a multimodal latent co-evolution flow across depth. We instantiate AMMAS for small molecules and introduce three benchmarks to evaluate geometry-grounded understanding and language-guided generation: conformational question answering, stereochemical comparison, and text-guided conformation editing. Across comparisons with eight learned baselines, AMMAS achieves the best overall performance on all benchmarks and ranks first in 14 of 15 subcategories. Against the strongest baselines, it improves the two understanding accuracies by 8.95 and 10.46 percentage points, and editing performance by 24.96 points in V-Success@3 Å and a relative 75.4% in V-NETP.

1 INTRODUCTION

In atomistic science, microscopic 3D structures are described through the identities and spatial arrangements of their constituent atoms (Barton, 1950; Larsen et al., 2017; Berman et al., 2000; Groom et al., 2016). Atomic geometry determines properties and interactions (Rahman, 1964; Car & Parrinello, 1985; Perutz, 1970; Lehn, 1988; Billesbølle et al., 2023), making the interpretation and design of these microscopic 3D structures important scientific problems (Li et al., 1999; Kuhlman et al., 2003; Schütt et al., 2018; Gasteiger et al., 2020). Natural language provides a complementary interface for expressing structural concepts (Moss, 1996), questions (Edwards et al., 2021; 2022), and desired geometric changes. Connecting these modalities can support geometry-grounded 3D structure understanding, where responses reflect a supplied atomic arrangement (Li et al., 2024), and language-guided 3D structure generation, where instructions guide the production of new 3D structures.

However, a modality gap remains: language modeling operates over discrete token sequences (Vaswani et al., 2017), whereas geometric modeling operates over continuous atomic coordinates and their spatial relationships (Zhou et al., 2023). Existing 3D structure–language approaches encode geometry as serialized coordinate text (Flam-Shepherd & Aspuru-Guzik, 2023; Zholus et al., 2024), discrete structural tokens (Pei et al., 2025; 2026), or projected geometric encoder features (Li et al., 2024; Kim et al., 2025). Across these approaches, geometry is encoded at the input stage, while subsequent language layers jointly and uniformly process structural and textual representa-

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

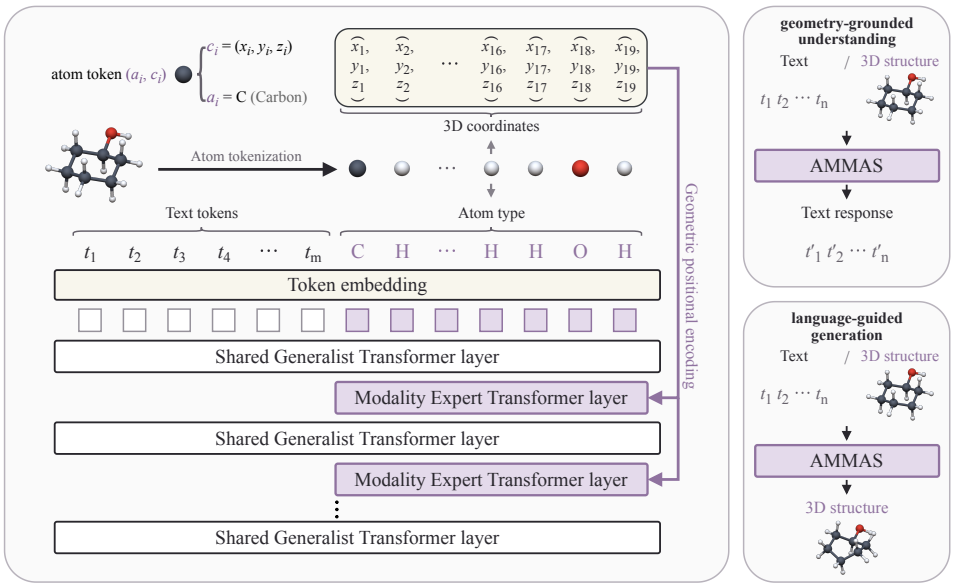

Figure 1: **AMMAS architecture and capabilities.** SGT layers jointly update atomistic and textual representations and are interleaved with MET layers, which apply geometry-specific updates within each 3D structure. This multimodal latent co-evolution flow supports geometry-grounded understanding and language-guided generation of continuous 3D structures.

tions. This motivates us to ask: why not distribute geometric modeling throughout the language backbone to better integrate 3D structures and language?

Our Contributions. We propose **A** **M**ultimodal **M**odel for **A**tomistic **S**cience (AMMAS), which supports autoregressive generation of both discrete text and continuous 3D structures. AMMAS interleaves stacks of Shared Generalist Transformer (SGT) layers for cross-modal interaction with Modality Expert Transformer (MET) layers that provide modality-specific attention patterns and positional encoding (Figure 1). SGT layers can be instantiated from a pretrained large language model (LLM) to build on its language capabilities and general knowledge. SGT layers jointly update the mixed sequence of atomistic and textual representations, enabling 3D structure-language interaction. MET layers then update these text-contextualized atomistic representations using structure-local attention and geometric positional encoding. The updated atomistic representations enter subsequent SGT layers, establishing a multimodal latent co-evolution flow across depth.

Word order carries linguistic meaning, whereas every atom in a fully observed structure should be able to attend to any other regardless of its serial position. We therefore use bidirectional attention within fully observed structures for complete geometric context and causal attention elsewhere for autoregressive prediction (Figure 2a).

We instantiate and evaluate AMMAS for small molecules, whose manageable size and conformational and stereochemical diversity (Axelrod & Gómez-Bombarelli, 2022; Moss, 1996) provide a focused setting for 3D structure–language modeling of atomistic science. We introduce three benchmarks for 3D structure-language geometry-grounded understanding and language-guided generation, spanning single-structure interpretation, cross-structure comparison, and text-guided conformation editing (Figure 3). Conformational question answering covers six families of local-to-global spatial semantics from a supplied conformation, without precomputed indices of the relevant geometric elements. Stereochemical comparison requires inferring and comparing configurations across two same-constitution structures, covering seven relation classes without preassigned stereochemical annotations. Text-guided conformation editing generates conformations from a source and instruction in two families: compactness and donor–acceptor distance.

Across eight baselines, AMMAS achieves the best overall performance on all three benchmarks and ranks first in 14 of 15 subcategories. Against the strongest task-specific baselines, it improves overall conformational question-answering accuracy and macro relation accuracy on stereochemical comparison by 8.95 and 10.46 percentage points, respectively. For editing, it improves validity-

screened success (V-Success@3 Å) by 24.96 percentage points and validity-screened target-attribute agreement (V-NETP) by 75.4% relative.

Ablations examine MET placement and observed-structure attention (Section 5.5). With geometric expert capacity fixed, AMMAS outperforms both front- and back-stacked variants in the controlled diagnostic setting, demonstrating the effectiveness of the interleaved SGT–MET architecture that establishes our multimodal latent co-evolution flow. Bidirectional observed-structure attention outperforms causal attention in overall and worst-family accuracy, confirming the necessity of complete structural context for geometry-grounded reasoning.

2 Related Work

2.1 Atomistic Science: A Shared Microscopic 3D Structural Perspective

We use atomistic science to refer to the study and modeling of physical systems through the lens of their microscopic 3D structures (Rahman, 1964; Car & Parrinello, 1985; Thompson et al., 2022). At this level, atoms serve as basic structural units, with their identities and spatial arrangements providing a common basis for representation (Larsen et al., 2017; Berman et al., 2000; Groom et al., 2016). OVITO visualizes and analyzes atomistic configurations (Stukowski, 2010), while polyhedral template matching identifies local crystal structures (Larsen et al., 2016). SchNet learns from atom types and positions to predict properties of molecules and materials (Schütt et al., 2018). Uni-3DAR develops an autoregressive framework for 3D generation and understanding across small molecules, protein-related structures, polymers, and crystals (Lu et al., 2025). This shared structural view spans scientific domains and captures geometric variation within each system. GEOM provides ensembles of accessible molecular conformations for property prediction and generation, distinguishing molecular identity from an individual 3D structure (Axelrod & Gómez-Bombarelli, 2022). Conformations can differ in torsions, local contacts, and global shape while connectivity and stereochemical configuration remain unchanged. Molecules with the same constitution can also differ in stereochemical configuration (Moss, 1996). For structure–language modeling, descriptions and comparisons must be grounded in the supplied structure, and conformational instructions must be realized in the generated geometry.

2.2 Structure–Language Modeling for Atomistic Science

Studies across scientific domains connect language with molecular strings, 2D graphs, and 3D structures. MoleculeSTM links text with SMILES or 2D molecular graphs for retrieval and property-oriented editing, without conditioning on a supplied 3D conformation (Liu et al., 2023). Prot2Text combines protein sequence and structural information to generate functional descriptions (Abdine et al., 2024), while MatterChat uses material structures and textual queries to produce property answers and structural descriptions (Tang et al., 2026). These works use language to convey scientific information and express queries or design objectives.

3D-MoLM and Mol-LLaMA condition language models on geometric encoder features, while 3D-MolT5 uses discrete structural identifiers for text or SELFIES generation. Their published models do not generate coordinates (Li et al., 2024; Kim et al., 2025; Pei et al., 2025). BioMatrix jointly models molecular and protein sequences, structures, and language in a shared discrete token space, decoding generated structural codes into 3D coordinates (Pei et al., 2026).

Multimodal architectures also differ in their organization of modality-specific computation. CogVLM adds visual experts to the attention and feed-forward layers of a pretrained language model (Wang et al., 2024). MoT combines modality-specific parameters with global self-attention (Liang et al., 2024), while Transfusion combines language-modeling and diffusion objectives within a shared Transformer (Zhou et al., 2024). AMMAS interleaves stacks of SGT layers with MET layers. SGT layers, instantiated from a pretrained LLM, jointly update atomistic and textual representations, while MET layers apply structure-local attention and geometric positional encoding to atomistic representations. Their repeated interaction establishes a multimodal latent co-evolution flow across depth. The resulting context supports geometry-grounded responses and autoregressive generation of atom types and continuous coordinates, without intermediate discrete structural codes. We also introduce model-independent small-molecule benchmarks for geometry-grounded interpretation, cross-structure comparison, and language-guided conformation editing.

3 MODEL

3.1 MULTIMODAL REPRESENTATION

Natural language uses the pretrained language model's tokenizer for input questions, instructions, and output responses. We represent an N_b-atom 3D structure b as

$$M_b = (A_b, \mathbf{C}_b), \qquad \mathbf{C}_b \in \mathbb{R}^{N_b \times 3},$$

where $A_b = (a_{b,1}, \ldots, a_{b,N_b})$ is the atom-type sequence and the i-th row of $\mathbf{C}_b$ is $\mathbf{c}_{b,i}^\top$. Each atom token $(a_{b,i}, \mathbf{c}_{b,i})$ pairs a discrete atom type with continuous coordinates $\mathbf{c}_{b,i} = (x_{b,i}, y_{b,i}, z_{b,i})^\top \in \mathbb{R}^3$. Coordinates $\mathbf{C}_b$ follow the canonicalization and task-specific alignment in Section 4.

Text tokens and atom types are embedded into hidden states $H^{(0)} \in \mathbb{R}^{L \times d}$, with sequence length L and hidden dimension d. Each structure b occupies a contiguous block, with position index set I_b. Hidden states within structure blocks and at text-token positions are termed atomistic and textual latent representations, respectively. Coordinates $\mathbf{C}_b$ provide geometric positional information to MET layers and are not serialized as numerical text. Boundary-token and task-specific sequence layouts, including coordinate-to-row alignment, are given in Appendix A.1.

3.2 SHARED GENERALIST TRANSFORMER (SGT)

Each SGT layer jointly updates atomistic and textual representations using self-attention and feed-forward parameters shared across both kinds of positions, preserving sequence length and hidden dimension. Mixed-sequence attention exchanges information across modalities and structures, subject to the mask in Section 3.5. AMMAS instantiates its SGT layers from a pretrained LLM to build on its language capabilities and general knowledge, retaining its sequence positional encoding. SGT and MET parameters are optimized jointly during multimodal training (Section 3.6).

3.3 MODALITY EXPERT TRANSFORMER (MET)

MET is a Transformer layer with modality-specific attention patterns and positional encoding. Its self-attention and feed-forward updates within a modality block preserve sequence length, token order, and hidden dimension. In AMMAS, structure-local attention and geometric positional encoding update atomistic latent representations, with coordinates fixed during each forward pass.

Structure-local attention. Within each structure block I_b, MET self-attention uses only that structure's atomistic representations (Figure 2a). SGT layers handle cross-modal and cross-structure interactions, supplying text-contextualized atomistic representations. The mask $\mathcal{A}_b$ is defined in Section 3.5.

Geometric positional encoding. MET layers use geometric positional encoding in attention. Given normalized hidden state $\widetilde{h}_{b,i}$ and coordinates $\mathbf{c}_{b,i}$, queries and keys are

$$q_{b,i} = R(\mathbf{c}_{b,i}) W_Q \widetilde{h}_{b,i}, \qquad k_{b,i} = R(\mathbf{c}_{b,i}) W_K \widetilde{h}_{b,i},$$

where W_Q and W_K are learned projections and $R(\mathbf{c})$ is a coordinate-dependent rotation. The positional contribution thus depends on spatial displacement between atoms. SGT layers retain their pretrained sequence positional encoding. Appendix A.3 gives encoding formulas and the relative-displacement identity; Appendix A.2 gives expert configurations.

3.4 MULTIMODAL LATENT CO-EVOLUTION FLOW

AMMAS interleaves stacks of SGT layers with MET layers (Figure 1). Let F_ℓ denote an SGT layer and E_ℓ the MET layer inserted after it. At a selected depth ℓ, SGT first updates the full sequence; MET then updates only the atomistic representations within each structure block:

$$\overline{H}^{(\ell)} = F_\ell\Big(H^{(\ell-1)}; \mathcal{A}_{\mathrm{SGT}}\Big), \qquad H^{(\ell)}[I_b] = E_\ell\Big(\overline{H}^{(\ell)}[I_b], \mathbf{C}_b; \mathcal{A}_b\Big).$$

Positions outside structure blocks retain their SGT outputs. Each MET layer is applied independently to all structure blocks, with separate parameters across depths. At layers without a MET

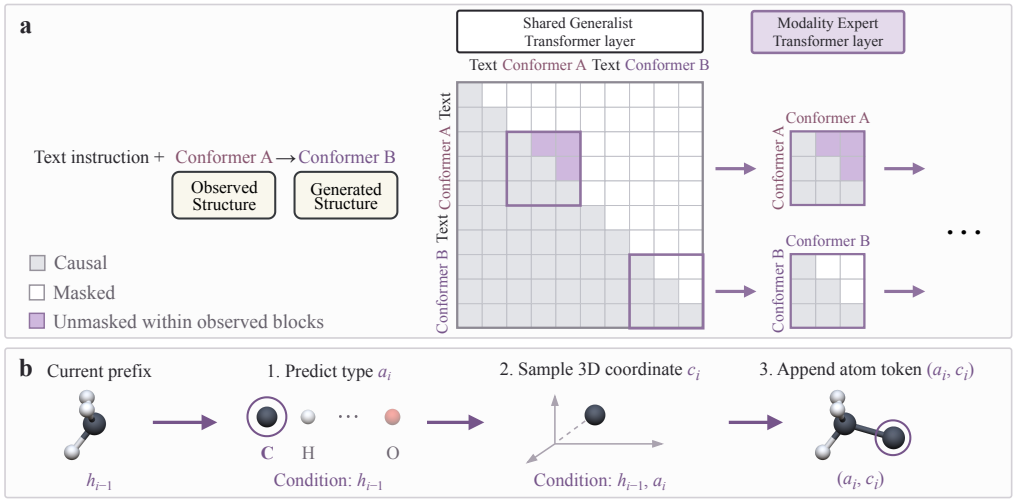

Figure 2: **AMMAS attention masking and autoregressive structure generation. (a)** The same role-aware masking rule is used in all tasks; conformation editing is shown as an example. SGT attention is bidirectional within each observed structure and causal elsewhere. MET layers use the corresponding masks restricted to individual structures. **(b)** Generation predicts an atom type, samples coordinates conditioned on the prefix and type, and appends the atom token (a_i, c_i).

update, $H^{(\ell)} = \overline{H}^{(\ell)}$. Both masks are defined in Section 3.5. MET layers apply geometric updates to atomistic representations contextualized by the text visible through preceding SGT layers. These representations then enter subsequent SGT layers for joint updates with textual representations, establishing a multimodal latent co-evolution flow across depth.

3.5 Role-Aware Attention Masking

Observed structures are fully available; generated structures grow atom by atom. A causal mask would hide later atoms within an observed structure, so we enable bidirectional attention within each observed block and causal attention elsewhere, preserving autoregressive prediction (Figure 2a). This rule applies to SGT and MET layers in all tasks. Understanding inputs are observed blocks; in editing, the source is observed and the target is generated. For valid query and key positions t, u, the SGT mask is

$$\mathcal{A}_{\mathrm{SGT}}(t,u) = \mathbf{1}[u \le t] \;\vee\; \mathbf{1}[\exists b:\; o_b = 1 \wedge t, u \in I_b]\,,$$

where $o_b = 1$ marks an observed block and a value of one permits attention. MET layers use the restriction $\mathcal{A}_b = \mathcal{A}_{\mathrm{SGT}}[I_b, I_b]$, keeping their attention within the corresponding structure block. Text, generated structures, and cross-block interactions retain causal attention in SGT layers.

3.6 Autoregressive Generation and Training

Geometry-grounded textual responses. Let $\mathcal{C}$ denote the supplied text and any observed 3D structures. A language-modeling head autoregressively predicts textual responses from the contextualized representations produced by the interleaved backbone.

Continuous-coordinate generation. For structure generation, let $\mathcal{C}_{i-1} = (\mathcal{C}, a_{<i}, \mathbf{c}_{<i})$ include the given context and previously generated atom tokens. We use a type-then-coordinate factorization:

$$p(a_i, \mathbf{c}_i \mid \mathcal{C}_{i-1}) = p(a_i \mid \mathcal{C}_{i-1})\, p(\mathbf{c}_i \mid \mathcal{C}_{i-1}, a_i).$$

A causal structure-generation tail produces a prefix state h_{i-1} from the backbone's target-prefix representations. The type head predicts an atom type or the structure-end token. For each predicted atom type, a lightweight EDM coordinate head (Karras et al., 2022; Li et al., 2025) samples its continuous 3D coordinates, conditioned on h_{i-1} and the atom-type embedding. Denoising reuses these conditions without re-evaluating the Transformer backbone. The completed atom token is appended to the prefix, and generation repeats until the structure-end token (Figure 2b). For conformation editing, $\mathcal{C}$ contains the source structure and language instruction; the reference atom-type sequence is not enforced during decoding. The detailed sampling procedure and prediction heads in A.4–A.5.

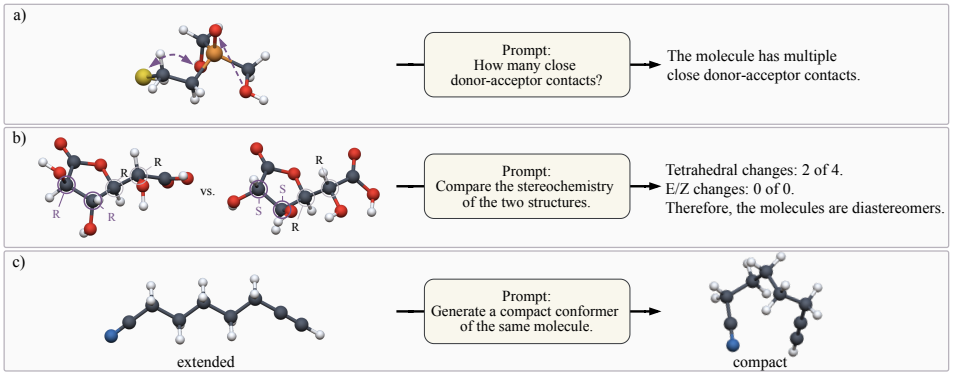

Figure 3: **Small-molecule benchmarks for 3D structure understanding and generation. (a)** Conformational question answering about donor–acceptor contacts. **(b)** Stereochemical comparison identifying a diastereomeric relationship. **(c)** Text-guided editing from an extended to a compact conformation. Dashed contact guides and R/S labels are illustrative annotations, not model inputs.

Training objectives. Text and atom-type prediction use teacher-forced cross-entropy losses, $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{type}}$. Coordinates use a noise-weighted EDM denoising loss $\mathcal{L}_{\text{coord}}$, conditioned on the ground-truth atom-token prefix and current atom type. Understanding optimizes $\mathcal{L}_{\text{text}}$, while structure generation uses the weighted sum $\mathcal{L}_{\text{gen}} = \lambda_a \mathcal{L}_{\text{type}} + \lambda_c \mathcal{L}_{\text{coord}}$. SGT and MET layers are optimized jointly with the relevant prediction modules. Loss and supervision details are in Appendices A.4–A.5; pretraining and fine-tuning are in Section 5.1.

4 BENCHMARKS

We introduce three small-molecule benchmarks to evaluate geometry-grounded understanding and language-guided 3D structure generation. Task definitions are independent of model architecture. Understanding uses PubChemQC-B3LYP structures (Nakata & Maeda, 2023); editing uses GEOM-Drug conformers (Axelrod & Gómez-Bombarelli, 2022).

Conformational question answering. Given atom identities, coordinates, and a natural-language question, the model must identify geometric elements without precomputed atom, bond, ring, or torsion indices and generate a semantic answer (Figure 3a). Six question families cover donor–acceptor contacts, hydrogen-bond candidates, torsions, ring orientation, scaffold planarity, and global shape. These conformation-dependent features cannot be determined from connectivity alone.

Stereochemical comparison. This task evaluates relational reasoning over two 3D structures with the same molecular constitution (Figure 3b). Without preassigned R/S or E/Z annotations or stereochemical-element indices, the model must infer configurations from coordinates, compare corresponding elements, and report their overall relation in a brief textual response. Seven classes cover tetrahedral, double-bond, and mixed configuration differences.

Text-guided conformation editing. Given a source conformation and instruction, the model generates a new conformation of the same molecule with the requested spatial change (Figure 3c). Two edit families target global compactness and the distance of a specified donor–acceptor pair. Source and reference conformations belong to the same molecule; reference coordinates are withheld.

Data splits and geometric preprocessing. Training, validation, and test molecules are disjoint within each benchmark. All systems use the same geometric canonicalization (Li et al., 2025) and task-specific alignment through their respective input representations. These rigid transformations preserve internal geometry and chirality; the editing source alone determines the output frame at inference. Appendix B details tasks and splits; Appendices B.1–B.3 describe preprocessing.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training setup. We initialize AMMAS's SGT layers from Qwen3-4B-Base (Yang et al., 2025) and continue pretraining on 3D molecular data, general text, and chemistry-related scientific text. Molecular and textual examples provide unconditional structure-generation and language-modeling supervision, respectively, without paired structure–instruction supervision. From this common checkpoint, we fine-tune on each benchmark separately using the objectives in Section 3.6; Appendix A.6 details the data mixture and text preprocessing.

Baselines. We compare AMMAS with eight baseline systems.

Understanding. For conformational question answering and stereochemical comparison, we evaluate ChemDFM-v1.5 (Zhao et al., 2025), ChemLLM (Zhang et al., 2024), LOGOS (Li et al., 2026), BioMatrix, 3D-MoLM, Mol-LLaMA (Kim et al., 2025), and 3D-MolT5, covering scientific language models and 3D molecule–text models.

Generation. For text-guided conformation editing, we evaluate ChemDFM-v1.5, ChemLLM, LOGOS, and BioMatrix. Three of the understanding baselines, 3D-MoLM, Mol-LLaMA, and 3D-MolT5, are excluded because they lack a dedicated 3D coordinate-generation pathway. We therefore add Qwen3 to broaden the baseline coverage as a general-purpose language-model baseline. We also include a copy-source control that returns the input conformation unchanged.

Baselines start from their authors' released checkpoints and follow their original fine-tuning strategies. All systems receive 3D-only structural inputs—atom identities and coordinates—together with the task question or instruction, through model-specific interfaces. They share task-specific splits, reference targets, parsers, and scoring rules, and are evaluated at their final task-specific checkpoints. Table 7 gives initialization and fine-tuning settings; Appendix C.2 details input adaptations. Attribution is addressed by matched controls discussed with the training protocols (Appendices C.1 and E).

Evaluation metrics. Understanding uses exact matching of parsed semantic labels, reporting overall question-answering accuracy or macro relation accuracy, together with worst-family or worst-relation accuracy and across-group population standard deviation. For editing, Success@3 Å measures reference-coordinate agreement and Normalized Edit Target Proximity (NETP) measures instruction-relevant attribute agreement. V-Success@3 Å and V-NETP additionally require atom identity, coordinate-inferred connectivity, local geometry, and source stereochemistry to pass a joint structural screen. Each model generates one candidate per example. All four metrics use the full test denominator, with missing or ineligible outputs contributing zero. Definitions, screening rules, and additional metrics are in Appendix C.3.

5.2 CONFORMATIONAL QUESTION ANSWERING

AMMAS achieves 95.81% overall accuracy, improving over ChemDFM-v1.5 (86.86%), the strongest baseline on this metric, by 8.95 percentage points (Table 1). It ranks first in all six question families, with a worst-family accuracy of 91.38%, compared with 81.04% for Mol-LLaMA. Its across-family standard deviation is the lowest (2.40 percentage points versus 3.64 for Mol-LLaMA), indicating high, balanced accuracy across families. The largest family-level gains occur on donor–acceptor contacts, hydrogen-bond candidates, and torsional profiles, where AMMAS exceeds the strongest baseline in each family by 14.75, 12.48, and 10.34 percentage points, respectively (Appendix D.1). These gains highlight more accurate identification of close-contact patterns and torsional organization in the supplied conformation.

5.3 STEREOCHEMICAL COMPARISON

AMMAS achieves 94.24% macro relation accuracy, improving over ChemDFM-v1.5 (83.78%), the strongest aggregate baseline, by 10.46 percentage points (Table 2). It ranks first in six of the seven relation classes, with a worst-relation accuracy of 89.06%, compared with 67.19% for ChemLLM. Its across-class standard deviation is the lowest (3.39 percentage points versus 10.99 for ChemLLM), indicating higher, more balanced accuracy across classes. The largest gains occur on epimers and multi-center enantiomers, where AMMAS exceeds the strongest baseline in each class by 19.38 and

Table 1: **Conformational question answering.** Accuracy in %; dispersion in percentage points.

Model	Accuracy (%)		Dispersion (pp)
	Overall ↑	Worst family ↑	Std. dev. ↓
ChemDFM-v1.5	<u>86.86</u>	80.07	5.58
ChemLLM	84.40	77.47	6.79
LOGOS	82.75	74.45	7.69
BioMatrix	68.92	42.31	13.74
3D-MoLM	66.34	58.99	4.37
Mol-LLaMA	84.04	<u>81.04</u>	<u>3.64</u>
3D-MolT5	64.90	41.85	11.61
AMMAS (Ours)	**95.81**	**91.38**	**2.40**

Table 2: **Stereochemical comparison.** Accuracy in %; dispersion in percentage points.

Model	Accuracy (%)		Dispersion (pp)
	Macro average ↑	Worst relation ↑	Std. dev. ↓
ChemDFM-v1.5	<u>83.78</u>	56.79	13.64
ChemLLM	82.04	<u>67.19</u>	<u>10.99</u>
LOGOS	80.26	42.01	19.09
BioMatrix	77.72	56.03	13.04
3D-MoLM	43.08	1.13	34.33
Mol-LLaMA	75.26	19.61	26.95
3D-MolT5	74.37	48.49	16.35
AMMAS (Ours)	**94.24**	**89.06**	**3.39**

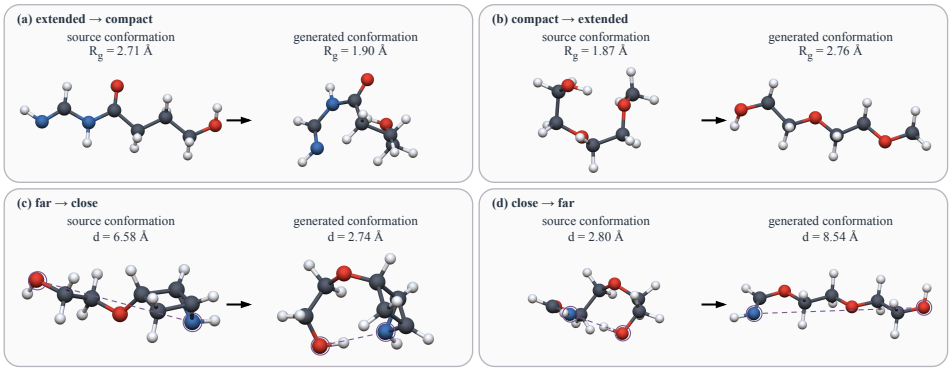

Figure 4: **Editing examples from AMMAS.** Source and generated conformations for **(a,b)** compactness editing and **(c,d)** donor–acceptor distance editing, each shown in both directions.

17.27 percentage points, respectively (Appendix D.2). Across multiple stereocenters, epimers differ at exactly one corresponding center, whereas multi-center enantiomers invert all centers. These gains highlight more accurate comparisons of multi-center configurations. A stricter evidence-level rescoring and a pairwise consistency diagnostic confirm that these gains extend to the supporting structural evidence (Appendix D.3).

5.4 Text-Guided Conformation Editing

AMMAS leads all four editing metrics in Table 3. Its V-Success@3 Å reaches 61.97%, compared with 37.02% for ChemDFM-v1.5, and its V-NETP is 0.6252 versus 0.3565, gains of 24.96 percentage points and 75.4% relative, respectively. The advantage holds in both edit families (Table 14), showing that improved target agreement is accompanied by greater structural consistency. Copy-source preserves the input but scores zero on NETP and V-NETP, distinguishing structural preservation from effective editing. A matched control with identical source conformations and different edit specifications likewise confirms that outputs track the supplied instruction (Appendix D.4). Figure 4 illustrates AMMAS's compactness and donor–acceptor distance edits in both directions; family-level and additional coordinate-error results appear in Appendix D.4.

Table 3: **Text-guided conformation editing.** V- denotes validity-screened metrics.

Model	Success@3 Å (%) ↑	V-Success@3 Å (%) ↑	NETP ↑	V-NETP ↑
Copy-source	23.80	21.89	0.0000	0.0000
Qwen3	69.79	30.12	0.6998	0.2773
ChemDFM-v1.5	73.29	<u>37.02</u>	0.7552	<u>0.3565</u>
ChemLLM	67.43	26.25	0.6896	0.2405
LOGOS	65.41	24.87	0.6388	0.2204
BioMatrix	<u>78.65</u>	32.70	<u>0.8135</u>	0.3341
AMMAS (Ours)	**83.75**	**61.97**	**0.8655**	**0.6252**

Table 4: **Architecture ablations of AMMAS.** Accuracies are percentages. Δ denotes the change in overall accuracy, in percentage points, relative to the full model and is computed before rounding.

Setting	Overall (%) ↑	Δ overall (pp)	Worst family (%) ↑
MET placement (bidirectional observed-block attention fixed)			
Interleaved (full model)	**95.81**	–	**91.38**
Front-stacked	91.37	−4.44	82.94
Back-stacked	85.12	−10.69	70.65
Observed-block attention (interleaved MET placement fixed)			
Bidirectional (full model)	**95.81**	–	**91.38**
Causal	92.01	−3.81	86.59

5.5 Ablation Studies

MET layer placement. On our diagnostic conformational question-answering test set (Table 4), we compare three layouts with the same SGT backbone and eight MET layers, fixing expert capacity and internal design. The interleaved configuration achieves 95.81% accuracy, exceeding front-stacked (91.37%) and back-stacked configurations (85.12%) by 4.44 and 10.69 percentage points, respectively. The advantage holds across all six question families, with worst-family accuracies of 91.38%, 82.94%, and 70.65% for the three configurations. Front stacking supplies geometric information to subsequent SGT layers but completes geometric updates before atomistic representations acquire textual context. Back stacking allows geometric processing to use contextualized atomistic representations, but leaves only the final SGT layer to propagate the resulting updates to textual representations. Interleaving combines these capabilities across depth: geometric updates act on text-contextualized atomistic representations and inform subsequent cross-modal processing. Together, these results demonstrate the effectiveness of the interleaved SGT–MET architecture that establishes our multimodal latent co-evolution flow.

Observed-block attention. On the same test set, replacing bidirectional attention within observed structure blocks with causal attention reduces overall accuracy by 3.81 percentage points and worst-family accuracy by 4.79 points. Causal attention prevents earlier atom positions from using all available atoms, whereas bidirectional attention gives both MET and SGT layers the complete observed structure. These gains support using complete structural context for geometry-grounded question answering. Training protocols, configurations, and per-family results are in Appendix E.

6 Conclusion

We presented AMMAS for geometry-grounded understanding and language-guided generation of continuous 3D structures in atomistic science. AMMAS interleaves stacks of SGT layers with MET layers. SGT layers jointly update atomistic and textual representations and can be instantiated from a pretrained LLM, while MET layers provide modality-specific attention patterns and positional encoding. Their repeated interaction establishes a multimodal latent co-evolution flow across depth. We instantiated AMMAS for small molecules and introduced three benchmarks for single-structure interpretation, cross-structure stereochemical comparison, and language-guided conformation editing. Its best aggregate performance among evaluated baselines on all three tasks supports its understanding and generation capabilities in this setting. Future directions include extending to proteins and materials through domain-specific geometric inductive biases in MET layers.

AI USE STATEMENT

Generative AI tools were used to support the preparation of this manuscript. In particular, they were used to polish language, improve sentence-level clarity and readability, and discuss alternative formulations during writing and revision. They were also used to assist with information retrieval and exploratory searches for relevant concepts, background knowledge, and related literature. The authors reviewed all AI-assisted revisions and independently verified factual statements, references, and scientific content against the original sources. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work is entirely computational and does not involve human participants, animal subjects, personal data, or wet-lab experiments. The proposed benchmarks evaluate 3D structure–language understanding and generation using existing public datasets. This work focuses on the representation, understanding, and generation of molecular 3D geometry using existing datasets. The reported results concern computational geometric modeling and benchmark performance rather than experimentally validated claims about molecular synthesis, biological activity, or clinical utility.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. Upon acceptance, we will release comprehensive resources on GitHub, including dataset access, experimental setup, model configurations, evaluation metrics, and checkpoints. Clear documentation and scripts will accompany these materials to enable accurate replication of all main results, thereby promoting transparency and scientific rigor.

REFERENCES

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2Text: Multimodal Protein's Function Generation with GNNs and Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10757–10765, 2024. doi: 10.1609/aaai.v38i10.28948.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022. doi: 10.1038/s41597-022-01288-4.

D. H. R. Barton. The conformation of the steroid nucleus. *Experientia*, 6(8):316–320, 1950. doi: 10.1007/BF02170915.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Christian B. Billesbølle, Claire A. de March, Wijnand J. C. van der Velden, Ning Ma, Jeevan Tewari, Claudia Llinas del Torrent, Linus Li, Bryan Faust, Nagarajan Vaidehi, Hiroaki Matsunami, and Aashish Manglik. Structural basis of odorant recognition by a human odorant receptor. *Nature*, 615(7953):742–749, 2023. doi: 10.1038/s41586-023-05798-y.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

R. Car and M. Parrinello. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Physical Review Letters*, 55(22):2471–2474, 1985. doi: 10.1103/PhysRevLett.55.2471.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 595–607, 2021. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47/.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between Molecules and Natural Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 375–413, 2022. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26/.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language Models Can Generate Molecules, Materials, and Protein Binding Sites Directly in Three Dimensions as XYZ, CIF, and PDB Files. *arXiv preprint arXiv:2305.05708*, 2023.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional Message Passing for Molecular Graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016. doi: 10.1107/S2052520616003954.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. Mol-LLaMA: Towards General Understanding of Molecules in Large Molecular Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302 (5649):1364–1368, 2003. doi: 10.1126/science.1089427.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano Eligio Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, et al. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648X/aa680e.

Peter Mahler Larsen, Søren Schmidt, and Jakob Schiøtz. Robust structural identification via polyhedral template matching. *Modelling and Simulation in Materials Science and Engineering*, 24(5): 055007, 2016. doi: 10.1088/0965-0393/24/5/055007.

Jean-Marie Lehn. Supramolecular Chemistry—Scope and Perspectives: Molecules, Supermolecules, and Molecular Devices (Nobel Lecture). *Angewandte Chemie International Edition in English*, 27(1):89–112, 1988. doi: 10.1002/anie.198800891.

Hailian Li, Mohamed Eddaoudi, M. O'Keeffe, and O. M. Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402:276–279, 1999. doi: 10.1038/46248.

Haorui Li, Weitao Du, Yuqiang Li, Hongyu Guo, and Shengchao Liu. InertialAR: Autoregressive 3D Molecule Generation with Inertial Frames. *arXiv preprint arXiv:2510.27497*, 2025.

Mingyang Li, Yurou Liu, Jieping Ye, Bing Su, Ji-Rong Wen, and Zheng Wang. Speaking the Language of Science: Toward a General-Purpose Generative Foundation Model for the Natural Sciences. *arXiv preprint arXiv:2606.16905*, 2026.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3D Molecule-Text Interpretation in Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. *arXiv preprint arXiv:2411.04996*, 2024.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal Molecule Structure-text Model for Text-based Retrieval and Editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023. doi: 10.1038/s42256-023-00759-6.

Shuqi Lu, Haowei Lin, Lin Yao, Zhifeng Gao, Xiaohong Ji, Yitao Liang, Weinan E, Linfeng Zhang, and Guolin Ke. Unified Cross-Scale 3D Generation and Understanding via Autoregressive Modeling. *arXiv preprint arXiv:2503.16278*, 2025.

M-A-P. FineFineWeb: A Comprehensive Study on Fine-grained Domain Web Corpus. Hugging Face Datasets, 2024. URL https://huggingface.co/datasets/m-a-p/FineFineWeb.

Adrian Mirza, Nawaf Alampara, Martiño Ríos-García, Mohamed Abdelalim, Jack Butler, Bethany Connolly, Tunca Dogan, Marianna Nezhurina, Bünyamin Şen, Santosh Tirunagari, Mark Worrall, Adamo Young, Philippe Schwaller, Michael Pieler, and Kevin Maik Jablonka. ChemPile: A 250GB Diverse and Curated Dataset for Chemical Foundation Models. *arXiv preprint arXiv:2505.12534*, 2025. URL https://arxiv.org/abs/2505.12534.

G. P. Moss. Basic Terminology of Stereochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*, 68(12):2193–2222, 1996. doi: 10.1351/pac199668122193.

Maho Nakata and Toshiyuki Maeda. PubChemQC B3LYP/6-31G*//PM6 Data Set: The Electronic Structures of 86 Million Molecules Using B3LYP/6-31G* Calculations. *Journal of Chemical Information and Modeling*, 63(18):5734–5754, 2023. doi: 10.1021/acs.jcim.3c00899.

Qizhi Pei, Rui Yan, Kaiyuan Gao, Jinhua Zhu, and Lijun Wu. 3D-MolT5: Leveraging Discrete Structural Information for Molecule-Text Modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Qizhi Pei, Zhimeng Zhou, Yi Duan, Yiyang Zhao, Wei Li, Han Guo, et al. BioMatrix: Towards a Comprehensive Biological Foundation Model Spanning the Modality Matrix of Sequences, Structures, and Language. *arXiv preprint arXiv:2606.22138*, 2026.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.17557.

M. F. Perutz. Stereochemistry of Cooperative Effects in Haemoglobin: Haem–Haem Interaction and the Problem of Allostery. *Nature*, 228:726–734, 1970. doi: 10.1038/228726a0.

A. Rahman. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136(2A): A405–A411, 1964. doi: 10.1103/PhysRev.136.A405.

Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet—A Deep Learning Architecture for Molecules and Materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Alexander Stukowski. Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering*, 18(1): 015012, 2010. doi: 10.1088/0965-0393/18/1/015012.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steven Farrell, Benjamin Erichson, Michael W. Mahoney, Andy Nonaka, et al. A Multimodal Large Language Model for Materials Science. *Nature Machine Intelligence*, 8(4):588–601, 2026. doi: 10.1038/s42256-026-01214-y.

The Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research*, 43(D1):D1042–D1048, 2015. doi: 10.1093/nar/gku1061.

A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022. doi: 10.1016/j.cpc.2021.108171.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, et al. CogVLM: Visual Expert for Pretrained Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-3860.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, et al. ChemLLM: A Chemical Large Language Model. *arXiv preprint arXiv:2402.06852*, 2024.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, et al. Developing ChemDFM as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6:102523, 2025. doi: 10.1016/j.xcrp.2025.102523.

Artem Zholus, Maksim Kuznetsov, Roman Schutski, Rim Shayakhmetov, Daniil Polykovskiy, Sarath Chandar, and Alex Zhavoronkov. BindGPT: A Scalable Framework for 3D Molecular Design via Language Modeling and Reinforcement Learning. *arXiv preprint arXiv:2406.03686*, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. *arXiv preprint arXiv:2408.11039*, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2023.

A Model and Training Details

A.1 Task Sequence Layouts and Visibility

The following templates specify the sequence order used by AMMAS. `[System]` and `[Assistant]` denote chat-format segments; the displayed text labels are explanatory and need not correspond to individual vocabulary tokens. The annotations `observed` and `causal` indicate structure-block roles and are not additional sequence tokens. `MOL_BOS` and `MOL_EOS` are the implementation names of the structure-begin and structure-end tokens, respectively. Coordinates accompany structure-block positions as aligned tensors rather than numerical text.

Structure block representation. For an observed structure block b occupying the sequence interval I_b, the discrete component is

$$[\texttt{MOL_BOS}, a_{b,1}, \ldots, a_{b,N_b}, \texttt{MOL_EOS}].$$

The coordinate matrix $\mathbf{C}_b \in \mathbb{R}^{N_b \times 3}$ is expanded into a block-aligned tensor $\widetilde{\mathbf{C}}_b$. Atom rows carry their corresponding coordinates; boundary and control rows carry zero placeholders, and padding rows are excluded by a validity mask m_b. Within a Transformer forward pass, coordinates are fixed inputs to the expert computation, which updates hidden states. During generation, new atom tokens extend the target prefix between prediction steps.

Mask application in task layouts. All three tasks below use the same role-aware masking rule defined in Section 3.5. Their layouts specify which structure blocks are observed and which are generated. Observed blocks allow bidirectional attention internally, while cross-block interactions

remain causal in SGT layers. MET layers attend only within their corresponding structure blocks. The placement of textual segments and structure blocks determines the context available at each position, as detailed below.

Conformational question answering.

```
[System]
[Question prefix]
[MOL_BOS, atom_1, ..., atom_N, MOL_EOS] observed
[Question suffix / answer constraint]
[Assistant]
[text answer]
```

Under SGT attention, positions in the observed structure block can attend to the system segment, the question prefix, and the entire structure block, but not to the subsequent question suffix. Answer positions can use the complete prompt, including the structure and question suffix.

Stereochemical comparison.

```
[System]
[Comparison instruction]
[MOL_BOS, atom_A1, ..., atom_AN, MOL_EOS] observed
[MOL_BOS, atom_B1, ..., atom_BN, MOL_EOS] observed
[Assistant]
[evidence description + relation answer]
```

A/B identifiers are learned embeddings added to atomistic hidden states, not additional text tokens. MET layers process the two blocks independently. SGT attention permits positions in structure B to use preceding information from structure A, but not the reverse. Both structures are available to subsequent answer positions. Bidirectionality applies within each observed block, not jointly across the two blocks. Their coordinate inputs use the common-frame view described in Appendix B.2.

Text-guided conformation editing.

```
[System]
[Editing instruction prefix]
[MOL_BOS, source_atom_1, ..., source_atom_N, MOL_EOS] observed
[Editing instruction suffix]
[Assistant]
[MOL_BOS, target_atom_1, ..., target_atom_N, MOL_EOS] causal
```

There is one observed source block and one causal target block. Under SGT attention, source positions can use the preceding instruction prefix, while target generation can use the source structure and the full editing instruction. Each MET layer processes source and target as separate blocks. At inference, the target begins at `MOL_BOS` and grows through generated atom tokens. Its atom-type sequence is predicted, not fixed to the reference. The complete target shown above represents the finished sequence, not information given in advance. During training, its atom inputs and supervision are shifted as specified in Appendix A.4.

A.2 MET Configuration

The evaluated AMMAS layout contains eight MET layers inserted after SGT layers 4, 8, 12, 16, 20, 24, 28, and 32. At each site, the same MET layer is applied independently to all structure blocks; layers at different sites have separate parameters. Each MET layer is a pre-norm layer with RMSNorm, bias-free 32-head attention, a second RMSNorm, a SwiGLU feed-forward module of width 6912, and an unconstrained learned scalar gate. The geometric positional encoding uses a rotary base of 10. Expert states match the shared hidden width of 2560.

Gated expert update. For hidden states Z_b of structure block b, the expert update used in Section 3.4 is

$$E_\ell(Z_b, \mathbf{C}_b; \mathcal{A}_b) = Z_b + g_\ell \left[T_\ell(Z_b, \widetilde{\mathbf{C}}_b; \mathcal{A}_b, m_b) - Z_b \right],$$

where T_ℓ is the pre-norm Transformer layer, $\mathcal{A}_b$ is the structure-local attention mask, and m_b marks valid block rows. The block-aligned coordinate tensor $\widetilde{\mathbf{C}}_b$ is defined in Appendix A.1. The scalar g_ℓ scales the residual update and is learned without a range constraint. Coordinate alignment and the validity mask are implicit in the main-text notation $E_\ell(\cdot, \mathbf{C}_b; \mathcal{A}_b)$; the structure-local attention mask is shown explicitly.

Structure-generation tail. The structure-generation path adds two causal structure-local Transformer layers after the interleaved SGT–MET backbone. These layers operate on target-prefix representations and feed the atom-type and coordinate heads without writing back to the shared sequence. The understanding tasks do not use this tail or the coordinate head. All tasks use the same eight-site interleaving layout.

A.3 Geometric Positional Encoding

We give the explicit rotation $R(\mathbf{c})$ used in Section 3.3. The three-axis rotary construction extends rotary query–key encoding (Su et al., 2021; Li et al., 2025) to continuous Cartesian coordinates. We describe one attention head within a fixed structure block and suppress layer, head, and block indices. Here $\mathbf{c}_i$ is the coordinate attached to token position i under the alignment in Appendix A.1. The rotation acts on feature channels; coordinates remain fixed during a Transformer forward pass.

Coordinate-dependent rotations. Let $d_h \geq 6$ be the even head width. Each Cartesian axis receives m pairs of feature channels, with r remaining channels:

$$m = \left\lfloor \frac{d_h}{6} \right\rfloor, \qquad r = d_h - 6m, \qquad \omega_s = \theta^{-s/m}, \quad s = 0, \ldots, m-1. \tag{1}$$

The rotary base is $\theta = 10$. The same fixed frequency schedule is used for all three axes and all heads. For an angle ϕ and a scalar coordinate t, define

$$\begin{aligned} R_2(\phi) &= \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \\ R_{\text{axis}}(t) &= \text{blockdiag}(R_2(\omega_0 t), \ldots, R_2(\omega_{m-1} t)), \\ R(\mathbf{c}) &= \text{blockdiag}(R_{\text{axis}}(x), R_{\text{axis}}(y), R_{\text{axis}}(z), I_r), \qquad \mathbf{c} = (x, y, z)^\top. \end{aligned} \tag{2}$$

Thus $R_{\text{axis}}(t) \in \mathbb{R}^{2m \times 2m}$ and $R(\mathbf{c}) \in \mathbb{R}^{d_h \times d_h}$. The final r channels are unchanged; the identity block is omitted when $r = 0$. In the reported configuration, $d_h = 2560/32 = 80$, so $m = 13$ and $r = 2$: the x, y, and z groups each occupy 26 consecutive channels, followed by two unrotated channels.

Queries, keys, and relative displacement. For a normalized hidden state $\widetilde{h}_i \in \mathbb{R}^d$, let $W_Q, W_K, W_V \in \mathbb{R}^{d_h \times d}$ denote the projections for this head. We compute

$$\begin{aligned} \bar{q}_i &= W_Q \widetilde{h}_i, \qquad \bar{k}_i = W_K \widetilde{h}_i, \qquad v_i = W_V \widetilde{h}_i, \\ q_i &= R(\mathbf{c}_i)\bar{q}_i, \quad k_i = R(\mathbf{c}_i)\bar{k}_i. \end{aligned} \tag{3}$$

Only queries and keys are rotated. Since $R_2(\phi)^\top R_2(\psi) = R_2(\psi - \phi)$, the block construction gives

$$\begin{aligned} R(\mathbf{c}_i)^\top R(\mathbf{c}_j) &= R(\mathbf{c}_j - \mathbf{c}_i), \\ q_i^\top k_j &= \bar{q}_i^\top R(\mathbf{c}_i)^\top R(\mathbf{c}_j)\bar{k}_j \\ &= \bar{q}_i^\top R(\mathbf{c}_j - \mathbf{c}_i)\bar{k}_j. \end{aligned} \tag{4}$$

At fixed projected features $\bar{q}_i, \bar{k}_j$, the explicit positional factor therefore depends on the relative displacement between positions i and j.

Attention and coordinate conventions. Let $\mathcal{J}_b(i)$ contain the valid positions in structure block b that query i can attend to under the mask $\mathcal{A}_b$ in Section 3.5. The attention weights and head output are

$$\begin{aligned} s_{ij} &= \frac{q_i^\top k_j}{\sqrt{d_h}}, \qquad j \in \mathcal{J}_b(i), \\ \alpha_{ij} &= \frac{\exp(s_{ij})}{\sum_{u \in \mathcal{J}_b(i)} \exp(s_{iu})}, \qquad o_i = \sum_{j \in \mathcal{J}_b(i)} \alpha_{ij} v_j. \end{aligned} \tag{5}$$

The interleaved experts use this encoding without an additional pair-distance or Nyström term. Boundary and control positions use the zero coordinate placeholders in Appendix A.1, for which $R(\mathbf{0}) = I_{d_h}$; padding positions are excluded from attention.

Coordinates are canonicalized and task-aligned as defined in Appendices B.1–B.3, with rigid-body symmetry handled by those preprocessing conventions. SGT layers retain their sequence positional encoding.

A.4 Right-Shifted Training and Autoregressive Inference

Right-shifted training. For an N-atom reference target, the atom-type and coordinate inputs are shifted by one position relative to their supervision:

```
input atom types   = [BOS, a_1*, ..., a_N*]
input coordinates  = [  0, c_1*, ..., c_N*]
type targets       = [a_1*, a_2*, ..., a_N*, EOS]
coordinate targets = [c_1*, c_2*, ..., c_N*,   0]
```

Stars denote ground-truth values; each `c_i` denotes the coordinate vector $\mathbf{c}_i$. `BOS` and `EOS` abbreviate `MOL_BOS` and `MOL_EOS`, respectively. Each `0` in a coordinate array denotes a three-dimensional zero vector. The placeholder coordinate associated with an `EOS` target does not receive coordinate loss.

Target slot	Input type	Input coordinate	Type supervision	Coordinate supervision
0	`BOS`	$\mathbf{0}$	$a_1^\star$	$\mathbf{c}_1^\star$
1	$a_1^\star$	$\mathbf{c}_1^\star$	$a_2^\star$	$\mathbf{c}_2^\star$
$i-1$	$a_{i-1}^\star$	$\mathbf{c}_{i-1}^\star$	$a_i^\star$	$\mathbf{c}_i^\star$
N	$a_N^\star$	$\mathbf{c}_N^\star$	`EOS`	None

The state at slot $i-1$ is used to predict the next atom token. Teacher forcing applies both to the preceding atom-token prefix and to the current atom type supplied separately to the coordinate denoiser. Although the complete target arrays are present in a training batch, causal masks prevent the Transformer state used for step i from accessing the current or future ground-truth atom tokens.

Prediction from the target prefix. The representation h_{i-1} in Section 3.6 is the state at the last available target-prefix position after the shared backbone, interleaved experts, and causal structure-generation tail. The first prediction uses the state at `MOL_BOS`; subsequent predictions use the state associated with the preceding atom token. The type distribution is

$$p(a_i \mid \mathcal{C}_{i-1}) = \left[\mathrm{softmax}(W_{\mathrm{type}} h_{i-1})\right]_{a_i},$$

where W_{type} is the learned type projection. The coordinate head uses the same h_{i-1} together with the selected atom type's embedding. In the editing setting, the given context $\mathcal{C}$ contains the source structure and language instruction.

Autoregressive inference. The following pseudocode describes the semantic sampling loop. It omits padding, batching, invalid-output handling, and implementation-specific stopping limits.

```
target_types = [MOL_BOS]
target_coords = [zero_vector]

repeat:
    h = last_target_state_after_backbone_MET_and_causal_tail(
        prompt, source, target_types, target_coords
    )
    next_type = sample(type_head(h))
    if next_type == MOL_EOS:
        stop
    next_coord = edm_sample(
        prefix_state = h,
        atom_type_condition = atom_type_emb(next_type)
```

```
)
append next_type and next_coord at the same new target position
```

Within `edm_sample`, the prefix state and atom-type embedding remain fixed while the noisy coordinate and noise level vary. No additional Transformer forward pass is performed between sampling the atom type and sampling its coordinate. The completed atom token enters the next Transformer prefix.

For an N-atom output, the autoregressive factorization includes a final termination probability:

$$p(A, \mathbf{C}, \mathrm{EOS} \mid \mathcal{C}) = \left[\prod_{i=1}^{N} p(a_i \mid \mathcal{C}_{i-1})\, p(\mathbf{c}_i \mid \mathcal{C}_{i-1}, a_i)\right] p(\mathrm{EOS} \mid \mathcal{C}_N).$$

A.5 Prediction Heads and Training Objectives

Text and atom-type prediction. For ground-truth response tokens $y_t^\star$ and atom types $a_i^\star$, teacher-forced cross-entropy gives

$$\mathcal{L}_{\mathrm{text}} = -\mathbb{E}_t \log p(y_t^\star \mid \mathcal{C}, y_{<t}^\star), \qquad \mathcal{L}_{\mathrm{type}} = -\mathbb{E}_i \log p\big(a_i^\star \mid \mathcal{C}_{i-1}^\star\big).$$

Here, $\star$ denotes ground-truth targets, and $\mathcal{C}_{i-1}^\star = (\mathcal{C}, a_{<i}^\star, \mathbf{c}_{<i}^\star)$ contains the given context and ground-truth atom-token prefix. The expectations average over supervised positions; atom-type supervision includes the structure-end token.

Coordinate head. The EDM head (Karras et al., 2022) uses a conditional residual MLP with three-dimensional coordinate input and output. Projected coordinates and their Fourier features form the MLP input. The noise-level embedding, prefix state, and atom-type embedding condition its residual blocks through adaptive layer normalization. The head contains no attention layers and shares its parameters across atom positions. During sampling, the prefix state and atom-type embedding are held fixed; denoising iterations evaluate only this coordinate head.

Denoising objective. Let $\epsilon \sim \mathcal{N}(0, I_3)$ and let σ be sampled from the training noise-level distribution. The prefix state $h_{i-1}^\star$ is computed from $\mathcal{C}_{i-1}^\star$ using the generation path in Appendix A.4. The EDM coordinate head predicts

$$\widehat{\mathbf{c}}_i = D_\phi\big(\mathbf{c}_i^\star + \sigma\epsilon,\ \sigma,\ h_{i-1}^\star,\ \mathrm{Emb}(a_i^\star)\big).$$

Teacher forcing applies to both the preceding atom-token prefix and the current atom type supplied separately to the denoiser. The noise-weighted denoising objective (Karras et al., 2022; Li et al., 2025) is

$$\mathcal{L}_{\mathrm{coord}} = \mathbb{E}_{i,\sigma,\epsilon}\left[w(\sigma)\, \|\widehat{\mathbf{c}}_i - \mathbf{c}_i^\star\|_2^2\right].$$

Training evaluates denoising regression in parallel over supervised atom positions, without unrolling a reverse-diffusion trajectory. The clipped noise-dependent weight is

$$w(\sigma) = \min\left(\frac{\sigma^2 + \sigma_{\mathrm{data}}^2}{(\sigma\sigma_{\mathrm{data}})^2},\ w_{\max}\right),$$

where σ_{data} is the EDM data/noise-scale parameter and $w_{\max}$ caps the weight. For conformation editing, the recorded configuration uses $\sigma_{\mathrm{data}} = 1.4$; this parameter is not a divisor applied to input coordinates (Appendix B.3).

The squared error is summed over the three coordinate components, averaged over sampled noise levels, and averaged over valid atom targets. The structure-end token receives type supervision only. Padding and conditioning-structure positions are excluded from coordinate supervision.

For editing, coordinate targets use the source-derived frame described in Appendix B.3. No Kabsch alignment is performed inside the denoising objective.

A.6 Continued-Pretraining Data

Table 5 details the source allocations of the continued-pretraining mixture; shares are measured over loss-bearing positions. The molecular stream consists of PubChemQC structures used solely for

unconditional coordinate modeling: it contains no question, relation label, editing instruction, or any paired structure–language supervision, all of which are introduced only during downstream fine-tuning. Pretraining on unlabeled structures is therefore the geometric analogue of self-supervised language-model pretraining, while all task results are reported on molecules held out from supervised fine-tuning, with stereoisomers of one constitution kept within a single split. The text stream combines general text with chemistry-related scientific text.

Table 5: **Data mixture for continued pretraining.** Shares refer to the total effective-token budget. The mixture comprises 65% 3D molecular data, 10% general text, and 25% chemistry-related scientific text.

Data source	Share
PubChemQC: 3D molecular structures (Nakata & Maeda, 2023)	65.0%
FineWeb-Edu: general text (Penedo et al., 2024)	10.0%
FineFineWeb Chemistry (M-A-P, 2024)	8.5%
EuroPMC Chemistry: full-text papers (The Europe PMC Consortium, 2015)	6.0%
EuroPMC abstracts: medicinal-chemistry keyword subset	1.5%
ChemPile Paper: remaining EuroPMC abstracts (Mirza et al., 2025)	2.5%
FineFineWeb Materials Science	4.1%
ChemPile Paper core	1.0%
ChemPile Paper bioRxiv	1.0%
ChemPile Education	0.4%
Total	**100.0%**

Text preparation and sampling. Text from the nine sources is normalized, re-tokenized with the Qwen3-4B-Base tokenizer, and packed; the text allocation is sampled according to the source quotas in Table 5, with deduplication applied across sources.

CPT–test boundary and data-leakage definition. We use *data leakage* to mean that an evaluated answer or its supervision is present during training. Under this definition there is no leakage: the CPT molecular stream is constructed to contain no question, relation, instruction, or label, and optimizes the unconditional distribution $p(\mathbf{C})$, whereas all downstream tasks evaluate conditional predictions $p(y \mid \mathbf{C}, q)$ and $p(\mathbf{C}_t \mid \mathbf{C}_s, q)$. The evaluated structures likewise do not recur as instances: CPT geometries are from the PubChemQC-PM6, while test geometries are from the different PubChemQC-B3LYP; editing is drawn from a separate source, GEOM-Drug. The CPT molecular stream is also disjoint from the downstream validation and test sets at molecule identity level. All generalization claims in this paper refer to molecules held out from supervised fine-tuning, on which task-specific supervision is introduced for the first time.

B Detailed Benchmark Definitions

Table 6 summarizes task interfaces, coverage, data sources, and complete split sizes, counting each task example or structure pair once.

Table 6: **Benchmarks for 3D molecule–text understanding and generation.** Inputs and outputs describe task requirements rather than model-specific representations. Counts refer to task examples in the complete training, validation, and test splits. The editing benchmark comprises two edit families: compactness and donor–acceptor distance. Detailed family and class definitions appear in Appendix B.

Task	Input	Requested output	Capability coverage	Source	Train / Validation / Test
Conformational Question Answering	One conformation and a question	Semantic answer	Six families spanning local-to-global spatial semantics	PubChemQC-B3LYP	450,000 / 90,000 / 90,000
Stereochemical Comparison	Two same-constitution structures and a comparison question	Brief comparison followed by a relation answer	Seven classes of configuration-level relations	PubChemQC-B3LYP	179,963 / 10,584 / 9,282
Text-Guided Conformation Editing	Source conformation and an editing instruction	New conformation of the source molecule	Two families of global or local geometry control	GEOM-Drug	320,000 / 40,000 / 40,000

Dataset splits. Within each benchmark, training, validation, and test splits are disjoint at the molecule level: all examples derived from a given molecule belong to the same split.

Offline construction and model inputs. The understanding benchmarks use coordinates supplied by PubChemQC-B3LYP; editing uses existing GEOM-Drug conformers. Molecular graphs and SMILES are used offline to identify chemical objects, select samples, establish correspondences, and construct reference labels. Geometric quantities are computed from the supplied coordinates. The resulting benchmarks provide atom identities, 3D coordinates, and task questions or instructions as model inputs, without the graphs or SMILES used for construction. Samples are selected to satisfy the task-specific criteria below. Distances are in Å and angles in degrees.

B.1 Geometric Canonicalization

All three benchmarks use the same unweighted geometric canonicalization procedure before model-specific input processing (Li et al., 2025). Principal-axis construction, deterministic axis orientation, and geometric fallbacks together establish a right-handed coordinate convention.

Unweighted centering and principal axes. Let $\mathbf{C}^{\mathrm{raw}} \in \mathbb{R}^{N\times 3}$ contain atomic coordinates $(\mathbf{c}_i^{\mathrm{raw}})^\top$ as rows. Every retained atom contributes equally. We compute the arithmetic centroid, centered coordinates, and unweighted inertia tensor:

$$\boldsymbol{\mu}^{\mathrm{raw}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_i^{\mathrm{raw}}, \qquad \mathbf{u}_i = \mathbf{c}_i^{\mathrm{raw}} - \boldsymbol{\mu}^{\mathrm{raw}},$$

$$J = \frac{1}{N}\sum_{i=1}^{N} \left(\|\mathbf{u}_i\|^2 I_3 - \mathbf{u}_i\mathbf{u}_i^\top\right).$$

The eigenvectors of J, ordered by ascending eigenvalues $\lambda_0 \le \lambda_1 \le \lambda_2$, define candidate principal axes. Atomic masses are not used.

Axis orientation and handedness. After axis selection, we resolve the first two axis signs independently using the `axis_max_abs` rule: the atom with the largest absolute coordinate on an axis is assigned a positive coordinate when its magnitude exceeds the direction tolerance. Otherwise, a fixed default sign is used. Starting from a right-handed basis, the third sign is the product of the first two, so the final frame remains right-handed.

Degenerate-axis resolution. Non-degenerate structures retain their principal axes. Numerical eigenvalue ties activate the `projected_anchor` rule: uniquely determined axes are retained, while

candidate atom and atom-pair vectors are projected into the unresolved subspace to select additional directions. Candidates are ranked by projection norm, with atom-pair vectors preferred in a tie; atomic numbers, pair distances, and indices in the current atom order provide deterministic tie-breaking. The main cases are:

Eigenvalue pattern	Retained axis	Completion rule
$\lambda_1 \simeq \lambda_2$, λ_0 distinct	Axis 0	Select a projected anchor in the plane orthogonal to axis 0.
$\lambda_0 \simeq \lambda_1$, λ_2 distinct	Axis 2	Select a projected anchor in the plane orthogonal to axis 2.
All three approximately equal	None	Select a primary geometric anchor, followed by an orthogonal secondary anchor.

If the geometry does not determine a secondary direction, a deterministic perpendicular direction completes the basis. For a strictly linear structure, this preserves the molecular axis and fixes the otherwise undetermined transverse axes. If no primary geometric anchor is available, a fixed right-handed basis is used. Coplanarity alone does not trigger the degenerate-axis procedure: coplanar structures with distinct eigenvalues retain their principal axes.

Canonical coordinates. Let $R \in SO(3)$ be the final frame matrix after axis selection, fallback handling, and sign resolution. The independently canonicalized coordinates are

$$\mathbf{C}^{\mathrm{can}} = (\mathbf{C}^{\mathrm{raw}} - \mathbf{1}(\boldsymbol{\mu}^{\mathrm{raw}})^{\top})R.$$

This coordinate transformation does not rescale the structure or reorder atoms. Proper rotations preserve internal geometry and chirality. When the geometrically selected frame is unique and transforms consistently with the input, the canonical coordinates are invariant under global $SE(3)$ transformations; geometrically undetermined directions follow the completion rules above.

B.2 Common-Frame Alignment

For stereochemical comparison and conformation editing, both structures first undergo atom-order normalization and geometric canonicalization independently. We then use the established atom correspondence to align the second canonicalized structure to the first. This produces the common-frame data views used by all evaluated systems through their respective input representations.

Proper Kabsch alignment. Let $\mathbf{C}_A^{\mathrm{can}}$ and $\mathbf{C}_B^{\mathrm{can}}$ denote the independently canonicalized coordinates, with rows indexed according to the established atom correspondence. Their arithmetic centroids are

$$\bar{\mathbf{c}}_A = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_{A,i}^{\mathrm{can}}, \qquad \bar{\mathbf{c}}_B = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_{B,i}^{\mathrm{can}}.$$

Using equal weights and row-vector coordinates, the Kabsch alignment rotation (Kabsch, 1976) is

$$Q_{B\to A} = \arg\min_{Q\in SO(3)} \left\|(\mathbf{C}_B^{\mathrm{can}} - \mathbf{1}\bar{\mathbf{c}}_B^{\top})Q - (\mathbf{C}_A^{\mathrm{can}} - \mathbf{1}\bar{\mathbf{c}}_A^{\top})\right\|_F^2.$$

The resulting common-frame coordinates are

$$\begin{aligned}\mathbf{C}_A^{\mathrm{com}} &= \mathbf{C}_A^{\mathrm{can}},\\ \mathbf{C}_B^{\mathrm{com}} &= (\mathbf{C}_B^{\mathrm{can}} - \mathbf{1}\bar{\mathbf{c}}_B^{\top})Q_{B\to A} + \mathbf{1}\bar{\mathbf{c}}_A^{\top}.\end{aligned}$$

The constraint $\det Q_{B\to A} = +1$ excludes reflections. Alignment uses the fixed atom correspondence and changes rigid-body pose while preserving each structure's internal geometry and stereochemical configuration. The centroids in these equations belong to the independently canonicalized coordinates, not the raw coordinates.

Stereochemical comparison. Molecule A defines the common coordinate frame. Molecule B is aligned to A after independent canonicalization, and both $\mathbf{C}_A^{\mathrm{com}}$ and $\mathbf{C}_B^{\mathrm{com}}$ are given as coordinate inputs. Proper alignment provides a common spatial reference without introducing a reflection that could erase a chiral distinction. The benchmark still requires configurations to be inferred from coordinates and compared across the two structures.

B.3 Conformation Editing: Source-Anchored Coordinates

Atom ordering and retained hydrogens. The source and reference target are conformers from the same GEOM-Drug molecule record. For each conformer, RDKit generates a canonical isomeric SMILES with `canonical=True`, `isomericSmiles=True`, and `allHsExplicit=True`. The atom order is obtained from `_smilesAtomOutputOrder`, and `Chem.RenumberAtoms` reorders the atoms and their coordinates accordingly. This SMILES is an offline ordering aid, not a text input to AMMAS.

The pipeline retains the explicit hydrogens already present in the input RDKit molecule. It neither removes hydrogens nor adds new coordinate-bearing atoms at this stage. Atomic local IDs are $Z-1$, where Z is the atomic number; hydrogen therefore has local ID zero. A pair is retained only if source and reference target have identical atom counts and atom-type sequences in the resulting order. Subsequent operations use the fixed correspondence between source atom k and reference-target atom k. All retained atoms, including hydrogens, participate in alignment, frame construction, coordinate supervision, and all-atom coordinate evaluation.

Independent canonicalization. After atom-order normalization, the source and reference target each undergo the unweighted geometric canonicalization in Appendix B.1. Let $\boldsymbol{\mu}_s^{\mathrm{raw}}$ and $\boldsymbol{\mu}_t^{\mathrm{raw}}$ be their raw arithmetic centroids, and let $R_s, R_t \in SO(3)$ be their independently constructed frame matrices:

$$\mathbf{C}_s^{\mathrm{can}} = (\mathbf{C}_s^{\mathrm{raw}} - \mathbf{1}(\boldsymbol{\mu}_s^{\mathrm{raw}})^\top)R_s,$$
$$\mathbf{C}_t^{\mathrm{can}} = (\mathbf{C}_t^{\mathrm{raw}} - \mathbf{1}(\boldsymbol{\mu}_t^{\mathrm{raw}})^\top)R_t.$$

Reference-target alignment. We next apply the proper Kabsch procedure in Appendix B.2 to align $\mathbf{C}_t^{\mathrm{can}}$ to $\mathbf{C}_s^{\mathrm{can}}$. Writing their canonical-coordinate centroids as

$$\bar{\mathbf{c}}_s = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_{s,i}^{\mathrm{can}}, \qquad \bar{\mathbf{c}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_{t,i}^{\mathrm{can}},$$

the stored source input and reference target are

$$\mathbf{C}_s^{\mathrm{in}} = \mathbf{C}_s^{\mathrm{can}},$$
$$\mathbf{C}_t^{\star} = (\mathbf{C}_t^{\mathrm{can}} - \mathbf{1}\bar{\mathbf{c}}_t^\top)Q_{t\to s} + \mathbf{1}\bar{\mathbf{c}}_s^\top, \qquad Q_{t\to s} \in SO(3).$$

The source remains in its independently canonicalized frame; the subsequent pairwise alignment expresses the reference target in that frame. These operations prepare coordinate supervision without fixing the generated atom-type sequence during decoding.

Units and precision. Coordinates remain in Å without radius or scale normalization. Kabsch and reference-frame calculations use float64; the resulting coordinate arrays are stored and loaded as float32. `gen3d_sigma_data=1.4` is an EDM data/noise-scale parameter.

Inference and inverse transformation. At inference, source atom-order normalization and source geometric canonicalization determine $\boldsymbol{\mu}_s^{\mathrm{raw}}$, R_s, and $\mathbf{C}_s^{\mathrm{in}}$. Reference-target canonicalization and $Q_{t\to s}$ are used only during offline reference preparation. A generated target $\widehat{\mathbf{C}}_t$ is expressed in the source-derived frame and can be returned to the original source pose by

$$\widehat{\mathbf{C}}_t^{\text{source pose}} = \widehat{\mathbf{C}}_t R_s^\top + \mathbf{1}(\boldsymbol{\mu}_s^{\mathrm{raw}})^\top.$$

This inverse uses the raw source centroid, distinct from the canonical-coordinate centroid $\bar{\mathbf{c}}_s$ used for alignment. It requires no reference target or target-alignment rotation. Restoring an original atom order, if needed, is separate from this coordinate transformation.

B.4 Conformational Question Answering: Question Families

The single-structure task uses the PubChemQC-B3LYP Hard V2 data views. Each example pairs a conformation and a question with a semantic-answer sentence corresponding to a geometry-derived label. The question identifies the property to assess without supplying the answer or precomputed atom, bond, ring, or torsion indices identifying the relevant geometric elements. Responses are mapped to the corresponding finite label space for exact-accuracy evaluation (Appendix C.3). Each

family contributes 75,000 / 15,000 / 15,000 examples to training / validation / test, and labels are balanced within each split. Since connectivity is never supplied under the 3D-only protocol, a coordinate-free input would consist solely of an unconnected bag of element identities with no bond or ring information— insufficient in principle for any of the six geometry-derived labels. Labels are balanced within each split.

Family	Operational description	Structured label space
`ring_orientation`	Aggregate orientation over planar ring pairs with disjoint atom sets	`uniformly_parallel`, `uniformly_orthogonal`, `mixed`
`scaffold_planarity`	Three-dimensional planarity of the largest atom-sharing ring system	`planar`, `moderately_warped`, `strongly_nonplanar`
`rotatable_torsion`	Conformational profile over eligible clean rotatable torsions	`anti_only`, `gauche_only`, `anti_gauche_mixed`, `cis_containing`
`donor_acceptor_distance`	Number of distinct close donor–acceptor contacts	`none`, `one`, `multiple`
`hbond_candidate`	Number of distinct intramolecular hydrogen-bond candidate contacts	`none`, `one`, `multiple`
`global_shape`	Molecule-level three-dimensional shape	`near_spherical`, `prolate`, `oblate`, `triaxial`

Donor–acceptor contacts. For both contact families, donors are N/O/S atoms bonded to at least one explicit hydrogen; acceptors are N/O/F/S/Cl atoms. These operational element-based definitions apply no additional charge, amide, or aromaticity filters. Candidate atoms must be distinct and separated by at least three bonds in the molecular graph. Contacts are deduplicated ordered pairs (D, A), with (D, A) and (A, D) counted separately. A pair is close when $d_{DA} \leq 3.5$ and far when $d_{DA} \geq 5.0$; intermediate distances provide no label evidence. Zero, one, or at least two close pairs yield `none`, `one`, or `multiple`, respectively, with `none` additionally requiring at least one far pair.

Hydrogen-bond candidates. A pair is positive if any explicit hydrogen bonded to its donor satisfies $d_{DA} \leq 3.5$, $d_{HA} \leq 2.5$, and $\angle DHA \geq 135^\circ$ (vertex at H). If no hydrogen satisfies these conditions, the pair provides negative evidence when any such hydrogen has $d_{DA} \geq 5.0$ or $d_{HA} \geq 3.5$; other triples provide no evidence. Each ordered (D, A) is counted once, with positive evidence taking precedence over negative evidence from other hydrogens. The three labels count positive pairs as above; `none` requires at least one negative-evidence pair. These are geometric candidate labels.

Rotatable torsion. The central bond is a non-ring, non-aromatic, non-conjugated heavy-atom single bond, excluding amide-type $C(=O)-N$. Each endpoint must have exactly one other heavy-atom neighbor, uniquely defining the signed torsion $\phi \in [-180^\circ, 180^\circ]$. Only defined angles in the cis ($|\phi| \leq 30^\circ$), gauche ($45^\circ \leq |\phi| \leq 75^\circ$), or anti ($|\phi| \geq 150^\circ$) intervals are retained. Each molecule must have at least two retained torsions. Any cis torsion gives `cis_containing`; otherwise, exclusively anti, exclusively gauche, or both states give `anti_only`, `gauche_only`, or `anti_gauche_mixed`. Other torsions do not participate in this aggregation.

Scaffold planarity. RDKit-perceived rings sharing an atom are merged transitively into ring systems. The largest system is the scaffold, with ties broken by the smallest canonical atom index; rings connected only by acyclic linkers remain separate. This is a ring-system definition rather than a Bemis–Murcko scaffold. For at least four scaffold atoms, let r_{plane} be their equal-weight root-mean-square perpendicular distance to the best-fit plane. The labels `planar`, `moderately_warped`, and `strongly_nonplanar` require $r_{\text{plane}} \leq 0.10$, $0.20 \leq r_{\text{plane}} \leq 0.35$, and $r_{\text{plane}} \geq 0.50$, respectively; gap intervals are not selected.

Ring orientation. Candidate pairs are all distinct RDKit-perceived rings with disjoint atom sets, without an aromaticity restriction; different pairs may reuse a ring. Each ring must have $r_{\text{plane}} \leq 0.10$. For unit plane normals, the angle $\theta = \arccos(|\mathbf{n}_1 \cdot \mathbf{n}_2|)$ is parallel at $\theta \leq 20^\circ$ and orthogonal at $\theta \geq 70^\circ$. A molecule must have at least two candidate pairs, all satisfying the planarity requirement

and one of these angle conditions. All-parallel, all-orthogonal, or both types yield `uniformly_parallel`, `uniformly_orthogonal`, or `mixed`, respectively.

Global shape. For at least three heavy atoms, let $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ be the eigenvalues of their equal-weight coordinate covariance tensor, with $\lambda_1 > 0$. Define

$$L = \frac{\lambda_1 - \lambda_2}{\lambda_1}, \qquad P = \frac{\lambda_2 - \lambda_3}{\lambda_1}, \qquad S = \frac{\lambda_3}{\lambda_1}.$$

In order, the labels are `near_spherical` if $S > L$ and $S > P$; `prolate` if $L \geq 0.60$, $L > P$, and $L > S$; `oblate` if $P \geq 0.50$, $P > L$, and $P > S$; and `triaxial` if none of the preceding conditions holds and $\max(L, P, S) \leq 0.55$. Remaining samples are not selected; `near_spherical` has no additional absolute S threshold.

Coordinate precision and maximum context length belong to the model-specific adapter rather than the shared task view.

B.5 Stereochemical Comparison: Relation Classes

The pairwise task uses structures from PubChemQC-B3LYP. Each pair shares the same molecular constitution, identified by RDKit canonical non-isomeric SMILES, and all stereoisomers of one constitution are assigned to the same split before pairs are formed. Models receive the atom identities and 3D coordinates of both structures together with the comparison question. Pairs are balanced across relation classes, with 25,709 training, 1,512 validation, and 1,326 test pairs per class.

Reference annotation. Reference relations are assigned from the 3D coordinates using RDKit stereochemistry routines and cross-checked against the source stereochemical annotations. Atom correspondence is established by canonical atom ranks on the molecular graph.

Relation class	Configuration difference
`single_center_enantiomer`	One stereocenter is present and its configuration is inverted.
`multi_center_enantiomer`	Multiple stereocenters are present and every corresponding center is inverted.
`epimer`	Multiple stereocenters are present but exactly one differs.
`diastereomer`	Several stereocenters differ, but not all.
`single_ez_isomer`	Stereocenters agree while one E/Z double bond differs.
`multi_ez_isomer`	Stereocenters agree while multiple E/Z double bonds differ.
`mixed_stereo`	Both stereocenters and E/Z double bonds differ.

The output consists of a brief comparison of the relevant configurations followed by the relation answer, and accuracy is computed on the parsed relation.

B.6 Text-Guided Conformation Editing: Edit Families

Each example associates a source conformation and a language instruction with a reference target from the same GEOM-Drug molecule record. The source and reference target have matching atom-type sequences and different coordinates by construction. AMMAS predicts the output atom-type sequence during decoding; agreement with the reference is checked during evaluation. Instructions specify the requested geometric change and, for donor–acceptor distance editing, the selected pair through a fixed template; reference target coordinates are withheld from the input. Atom correspondence, independent canonicalization, and subsequent target-to-source alignment are defined in Appendix B.3. Scoring follows the atom-type sequence, instruction-relevant attribute, and reference-coordinate criteria in Appendix C.3.

Pair selection. Pairs combine conformers of the same molecule that exhibit a clear, instruction-relevant geometric difference; both editing directions are represented.

The text-guided conformation editing dataset, `geom_drug_conformer_editing_generation_v1`, contains 320,000 training, 40,000 validation, and 40,000 test examples. It consists of two edit families, compactness and donor–acceptor distance, each contributing 160,000

training, 20,000 validation, and 20,000 test examples. These are the complete dataset splits used for editing-task training, validation, and evaluation.

Family	Instructional change	Local referent
`compactness_edit_v1`	Make an extended conformation more compact, or a compact conformation more extended.	Global goal; no local atom anchor.
`donor_acceptor_distance_edit_v1`	Move the specified intramolecular donor–acceptor pair from close to far, or from far to close.	A selected donor–acceptor pair is explicitly identified through a fixed instruction template.

C Input Contracts and Evaluation

Shared evaluation protocol. All systems use the same geometric canonicalization and task-specific common-frame data views (Section 4), through their respective input representations. Within each benchmark, systems share test examples, reference targets, output parsers, invalid-output policies, and evaluation metrics. The tasks specify what is provided and requested without prescribing an encoder, tokenization, attention mechanism, or generation algorithm. Model-specific input representations may differ, as documented below.

C.1 Initialization and Fine-Tuning Protocol

Table 7 summarizes model versions and scales, task coverage, training-set sizes, and checkpoint selection. Baselines use the checkpoint released with their original paper. AMMAS starts downstream fine-tuning from the common continued-pretraining checkpoint described in Section 5.1. We follow each baseline's original supervised fine-tuning strategy, including its optimizer, learning rate, schedule, and training duration, preserving its model-specific adaptation design. All evaluated systems use the same task-specific training splits. Each system is trained to adequately fit the task-training data and evaluated at its final task-specific checkpoint under the shared evaluation protocol. For conformation editing, both training and evaluation cover the two edit families of compactness and donor–acceptor distance.

Table 7: **Model versions, scales, and task-specific fine-tuning protocols.** Initialization entries link to released weights and identify their language backbones. Scale follows the published model size; † denotes the language-model component, excluding external structural modules. AMMAS's 5B is its rounded total parameter count. QA, Stereo, and Editing columns report the number of training examples used for each task; a dash indicates that the system is not evaluated on that task. Final denotes the checkpoint at the end of task-specific training. AMMAS is included as the reference system.

System	Initialization	Scale	QA	Stereo	Editing	Checkpoint
Qwen3	Qwen3-4B	4B	—	—	320,000	Final
ChemDFM-v1.5	ChemDFM-v1.5-8B (Llama 3)	8B	450,000	179,963	320,000	Final
ChemLLM	ChemLLM-7B-Chat-1.5-SFT (InternLM2)	7B	450,000	179,963	320,000	Final
LOGOS	LOGOS-8B (Qwen3)	8B	450,000	179,963	320,000	Final
BioMatrix	BioMatrix-4B-SFT (Qwen3)	4B†	450,000	179,963	320,000	Final
3D-MoLM	Generalist (Llama 2)	7B†	450,000	179,963	—	Final
Mol-LLaMA	Mol-Llama-3.1-8B-Instruct-Full-Weights	8B†	450,000	179,963	—	Final
3D-MolT5	3d-molt5-base (T5-v1.1 Base)	255M	450,000	179,963	—	Final
AMMAS	Common CPT checkpoint (Qwen3-4B-Base)	5B	450,000	179,963	320,000	Final

Table 8: **Fine-tuning configuration for AMMAS.** Per-task SFT hyperparameters; all runs use AdamW with cosine scheduling under DeepSpeed ZeRO-2.

Hyperparameter	QA	Stereo	Editing
Per-device micro-batch	4	4	8
Global batch size	64	64	512
SGT learning rate	1e-5	1e-5	4e-6
MET learning rate	1e-5	1e-5	1.5e-5
Weight decay	1e-5	1e-5	0
Warmup ratio	3%	3%	2%
Maximum sequence length	1536	1536	2048

Coordinate decoding uses the EDM Heun sampler with churn 40 and $\sigma_{\mathrm{data}} = 1.4$.

Attribution controls. Every baseline starts from a checkpoint carrying its own large-scale, often domain-specific pretraining; CPT brings the general Qwen3-4B backbone to comparable readiness rather than adding supervision unavailable to the baselines, while SFT splits are identical for all systems. The layout ablation (Appendix E) compares interleaving with front- and back-stacking under identical CPT protocols from the same initialization, expert count, and capacity; no layout starts from another layout's checkpoint, so it isolates placement within the geometric pathway.

Coordinate representation and training objectives. The native Qwen backbone uses one-dimensional positional encoding for token order, which does not directly encode atomic 3D positions. Our Qwen3 baseline represents coordinates as discrete text tokens and is trained with next-token cross-entropy. Continuous-coordinate diffusion instead requires a coordinate output head and a denoising objective; these are additional components beyond the native text-generation setup. In AMMAS, MET supplies geometric conditioning through the prefix states, while the EDM head also receives the noisy target coordinates and the noise level directly. Removing MET removes the only route by which the source geometry conditions the prediction, so a "same CPT and head, no MET" variant would not isolate the architecture—it would remove the geometric conditioning itself. Preserving that conditioning requires an alternative geometry-input design rather than a matched within-pathway ablation.

The pathway-as-a-whole comparison is provided by the Qwen3 editing row: the same Qwen3-4B backbone and SFT split, coordinates as serialized text, with no geometric CPT, MET, or EDM pathway. We fine-tune the released Qwen3 checkpoint on serialized-coordinate text without an additional geometric CPT stage. It scores 69.79% Success@3 Å, 30.12% V-Success, and 0.6998 NETP versus 83.75%, 61.97%, and 0.8655 for AMMAS. The two well-defined comparisons therefore bracket the architecture at both levels—the pathway absent versus present (Qwen3), and placement within the present pathway (layout ablation, under the same CPT on QA)—rather than attributing main-table gains to CPT alone.

C.2 Model-Visible Inputs

All baselines are initialized from their authors' released checkpoints and fine-tuned on the same task-specific training splits under a unified input protocol: each model receives atom identities, 3D coordinates, and the task question or instruction, with no model accessing information unavailable to the others. This fixes information content rather than encoding: text-interface models receive coordinates serialized as `element x y z` at five-decimal precision (in Å), while geometric-encoder and structure-token models receive the same information through their native interfaces. Final evaluations use untruncated inputs; maximum sequence lengths are given in Table 8. Checkpoints are listed in Table 9.

Table 9: Released checkpoints used for all baselines.

Model	Released checkpoint
Qwen3	`Qwen3-4B`
ChemDFM	`ChemDFM-v1.5-8B`
ChemLLM	`ChemLLM-7B-Chat-1_5-SFT`
LOGOS	`LOGOS-8B`
3D-MoLM	3D-MoLM generalist checkpoint
Mol-LLaMA	`Mol-Llama-3.1-8B-Instruct-Full-Weights`
3D-MolT5	`3d-molt5-base`
BioMatrix	`BioMatrix-4B-SFT`

C.3 Evaluation Metrics

Evaluation scope. The reported scores assess task-specific generalization to molecules held out from supervised fine-tuning, within the defined question, relation, and editing instruction families. They quantify semantic-label correctness, reference-based conformation editing, and instruction-relevant attribute agreement.

Understanding. Accuracy uses the complete test denominator and exact matching of the parsed semantic label. Conformational question answering accuracy is the unweighted mean across six equally sized families; macro relation accuracy is the unweighted mean across seven equally sized classes. Worst-family or worst-relation accuracy is the minimum among these group accuracies. Across-group standard deviation is the population standard deviation, measured in percentage points. Accuracies are reported as percentages. All reported differences and relative gains are computed before rounding.

Editing output eligibility. Each model generates a single candidate per editing test example, evaluated directly without best-of-N selection. Let $\mathcal{E}$ contain the test examples whose generated outputs have finite coordinates and an atom-type sequence matching the reference in length, elements, and order. This establishes atom correspondence for coordinate evaluation; connectivity and stereochemistry are checked separately below. Let $N_{\mathrm{all}} = 40{,}000$ denote the complete editing test-set size.

Instruction-relevant attribute agreement. For an evaluable output with a matching atom-type sequence, Normalized Edit Target Proximity (NETP) is defined as

$$q = \mathrm{clip}\left(1 - \frac{\left|g(\widehat{\mathbf{C}}_t) - g(\mathbf{C}_t^{\star})\right|}{|g(\mathbf{C}_s^{\mathrm{in}}) - g(\mathbf{C}_t^{\star})|}, 0, 1\right),$$

where $\mathbf{C}_s^{\text{in}}$, $\mathbf{C}_t^{\star}$, and $\widehat{\mathbf{C}}_t$ denote the source, reference target, and generated coordinate matrices, respectively, using the source-derived frame in Appendix B.3. For compactness editing, g is the radius of gyration; for distance editing, it is the distance of the specified atom pair. NETP equals one when the target attribute is matched and zero when the source attribute is unchanged. Missing outputs and those with invalid coordinates or mismatched atom counts or atom-type sequences receive zero. Aggregate NETP is the mean over all 40,000 examples, equivalently the equally weighted mean of the two family scores.

Reference-coordinate agreement. For $j \in \mathcal{E}$, let r_j denote the all-atom Kabsch RMSD, using proper rotations without reflection. We report

$$\text{Success}\,@\tau = \frac{\#\{j \in \mathcal{E} : r_j \le \tau\}}{N_{\text{all}}}, \qquad \tau \in \{1.5, 3\}\,\text{Å},$$

and the normalized Success-CDF area

$$\text{AUC}[0,3] = \frac{1}{N_{\text{all}}} \sum_{j \in \mathcal{E}} \text{clip}\left(1 - \frac{r_j}{3\,\text{Å}},\, 0,\, 1\right).$$

Examples outside $\mathcal{E}$ contribute zero to both full-denominator metrics. Family-level scores use the same definitions with the complete denominator of 20,000 examples per family. Coordinate errors are computed per molecule and averaged with equal weight over each model's evaluable outputs in $\mathcal{E}$, without filtering by the structural screen. The per-molecule coordinate mean absolute error (cMAE) is the mean of the per-atom absolute coordinate errors; coordinate errors are reported in both raw and Kabsch-aligned forms. Aggregate errors pool evaluable examples across both families. Success rates are shown as percentages when marked %, and as fractions in the additional coordinate-error tables; AUC and NETP are in $[0, 1]$. Coordinate errors are in Å.

Joint structural validity. For example j, binary indicators A_j, C_j, G_j, and S_j denote output eligibility, coordinate-inferred connectivity, heavy-atom local geometry, and source stereochemistry retention. We require all four:

$$V_j = A_j C_j G_j S_j.$$

The connectivity term passes when the undirected edge set inferred from the predicted coordinates and all provided atoms matches the source graph adjacency under the fixed atom correspondence. The geometry term uses checks adapted from PoseBusters (Buttenschoen et al., 2024): bonded and 1–3 distances must lie within pairwise molecular distance bounds, non-bonded pairs stay above the lower bound, and applicable rings and double bonds satisfy the planarity condition. The stereochemistry term passes when configurations reassigned independently from source and generated coordinates agree; changed, missing, or extra configurations fail. The same rules are applied to all stored predictions without coordinate optimization or repair; source graphs are used only for evaluation.

We combine this screen with target agreement through

$$\text{V-Success}@\tau = \frac{1}{N_{\text{all}}} \sum_{j \in \mathcal{E}} V_j\, \mathbf{1}[r_j \le \tau], \qquad \text{V-NETP} = \frac{1}{N_{\text{all}}} \sum_{j \in \mathcal{E}} V_j q_j.$$

Both retain the complete denominator, with screen failures contributing zero. V-Success measures screened coordinate agreement; V-NETP measures screened agreement in the instructed attribute. Table 10 reports the component and joint scores; Copy-source returns the input conformation unchanged.

Table 10: **Structural checks and joint scores on all 40,000 editing examples.** Bold and underline rank the learned systems; Copy-source is a control.

Model	Connectivity (%) ↑	Stereo retained (%) ↑	Heavy geometry (%) ↑	Joint V (%) ↑	V-Success@3 Å (%) ↑	V-NETP ↑
Copy-source	98.51	98.57	91.49	91.45	21.89	0.0000
Qwen3	49.67	79.32	47.29	36.98	30.12	0.2773
ChemDFM-v1.5	<u>60.95</u>	80.60	<u>56.72</u>	<u>45.05</u>	<u>37.02</u>	<u>0.3565</u>
ChemLLM	43.55	78.83	41.93	32.24	26.25	0.2405
LOGOS	41.39	76.46	40.33	30.87	24.87	0.2204
BioMatrix	47.31	**97.41**	55.42	39.96	32.70	0.3341
AMMAS (Ours)	**83.56**	<u>94.94</u>	**75.80**	**71.06**	**61.97**	**0.6252**

D Detailed Task Results

This appendix provides the family- and relation-level results underlying Section 5, together with complete aggregate editing results. Evaluation metrics and scoring rules are defined in Appendix C.3.

D.1 Conformational Question Answering Results by Family

Table 11: **Conformational question answering accuracy by family.** Each family contains 15,000 test examples. Entries are percentage accuracies. Question-family definitions are given in Appendix B.4. Bold and underline denote the best and second-best results in each row, respectively.

Question family	ChemDFM-v1.5	ChemLLM	LOGOS	BioMatrix	3D-MoLM	Mol-LLaMA	3D-MolT5	AMMAS
Donor–acceptor contacts	82.61	77.47	74.45	61.16	62.74	<u>83.02</u>	60.65	**97.77**
Hydrogen-bond candidates	<u>82.53</u>	78.63	78.07	77.69	69.61	81.50	68.09	**95.01**
Rotatable torsion	80.07	77.98	74.98	71.37	58.99	<u>81.04</u>	68.91	**91.38**
Scaffold planarity	88.41	86.65	83.94	83.19	72.09	<u>91.25</u>	78.95	**95.95**
Ring orientation	<u>92.09</u>	91.07	90.21	77.78	68.33	86.03	70.94	**99.02**
Global shape	<u>95.45</u>	94.63	94.85	42.31	66.25	81.38	41.85	**95.75**

D.2 Stereochemical Comparison Results by Relation Class

Table 12: **Stereochemical comparison accuracy by relation class.** Each class contains 1,326 test pairs. Entries are percentage accuracies. Relation-class definitions are given in Appendix B.5. Bold and underline denote the best and second-best results in each row, respectively.

Relation class	ChemDFM-v1.5	ChemLLM	LOGOS	BioMatrix	3D-MoLM	Mol-LLaMA	3D-MolT5	AMMAS
Single-center enantiomer	97.74	96.23	<u>97.89</u>	91.03	3.77	92.68	89.82	**98.79**
Epimer	56.79	<u>69.68</u>	42.01	56.03	60.18	19.61	48.49	**89.06**
Multi-center enantiomer	<u>72.55</u>	67.19	65.69	59.20	1.13	54.37	49.55	**89.82**
Diastereomer	91.40	74.36	<u>94.80</u>	85.29	82.58	**96.53**	81.98	94.12
Single E–Z isomer	81.15	<u>82.81</u>	75.41	80.62	54.22	73.45	82.20	**94.27**
Multi E–Z isomer	94.72	94.49	94.27	86.12	87.78	<u>96.30</u>	88.24	**96.46**
Mixed stereo	92.08	89.52	91.78	85.75	11.92	<u>93.89</u>	80.32	**97.13**

D.3 Evidence-Level Rescoring and Consistency Diagnostic

The headline relation accuracy scores the final relation only. We additionally rescore the same 9,282 test answers under a stricter criterion that also requires the reported tetrahedral and E/Z change and total counts to all be correct; unparseable answers count as incorrect. The relation-only and evidence-level scores are two scoring views of the identical test set: AMMAS reaches 94.24% and 91.28%, respectively. We further conduct a small consistency diagnostic on 32 groups of three differently configured molecules sharing one constitution. Each group's three pairwise comparisons are performed independently—when comparing one pair, the model sees neither the third structure nor the other answers—and a group passes only when all three answers are correct in both evidence and relation. This diagnostic is a small internal study rather than a new benchmark. We report representative language-based and structural-token systems under the shared protocol.

Table 13: **Evidence-level rescoring (9,282 test pairs) and consistency diagnostic (32 groups).** The first column requires correct evidence counts and relation; the second requires all three pairwise comparisons in a group to be correct.

Model	Evidence + relation (%)	Consistency groups
ChemDFM	77.43	4/32
ChemLLM	75.07	8/32
LOGOS	73.63	1/32
BioMatrix	66.89	7/32
AMMAS (Ours)	**91.28**	**18/32**

D.4 TEXT-GUIDED CONFORMATION EDITING RESULTS

The editing benchmark comprises two edit families, each with 20,000 test examples. Table 14 reports the joint scores by family. Table 15 gives additional aggregate metrics over all 40,000 examples; Tables 16 and 17 give the corresponding family-level metrics.

Scoring follows Appendix C.3. Success, normalized Success-CDF AUC, and NETP use the complete test denominator at the aggregate or family level. Coordinate errors are computed per molecule and averaged over each model's evaluable outputs with finite coordinates and matching atom counts and atom-type order. Aggregate errors pool evaluable examples across both families. In the additional-metric tables, bold and underline denote the best and second-best listed scores.

Table 14: **Validity-screened editing by family.** Each family contains 20,000 examples. Bold and underline rank the learned systems.

	Compactness		Donor–acceptor distance	
Model	V-Success@3 Å (%) ↑	V-NETP ↑	V-Success@3 Å (%) ↑	V-NETP ↑
Qwen3	34.88	0.3379	25.37	0.2167
ChemDFM-v1.5	<u>43.46</u>	<u>0.4335</u>	30.57	0.2794
ChemLLM	31.20	0.2997	21.29	0.1812
LOGOS	29.21	0.2765	20.54	0.1643
BioMatrix	32.97	0.3361	<u>32.43</u>	<u>0.3321</u>
AMMAS (Ours)	**65.98**	**0.6729**	**57.97**	**0.5776**

Table 15: **Additional aggregate editing metrics.** Success, AUC, and NETP are full-test fractions; coordinate errors are in Å over evaluable outputs.

Model	AUC[0, 3] ↑	Success @1.5 Å ↑	Success @3 Å ↑	NETP ↑	Raw RMSD ↓	Kabsch RMSD ↓	Raw cMAE ↓	Kabsch cMAE ↓
Copy-source	0.0490	0.0115	0.2380	0.0000	<u>3.6430</u>	3.6430	1.6019	1.6019
Qwen3	0.1838	0.0747	0.6979	0.6998	3.7865	2.6007	1.6422	1.1245
ChemDFM-v1.5	0.2047	0.0966	0.7329	0.7552	3.6962	2.5229	<u>1.5945</u>	1.0816
ChemLLM	0.1719	0.0663	0.6743	0.6896	3.9396	2.6621	1.7136	1.1566
LOGOS	0.1659	0.0613	0.6541	0.6388	3.8751	2.6401	1.6857	1.1469
BioMatrix	<u>0.2710</u>	<u>0.1813</u>	<u>0.7865</u>	<u>0.8135</u>	7.9878	<u>2.2962</u>	3.6341	<u>0.9694</u>
AMMAS	**0.2926**	**0.1919**	**0.8375**	**0.8655**	**2.7184**	**2.1909**	**1.1633**	**0.9168**

Instruction-dependence control. The source conformation does not uniquely determine the reference target. The test set contains 1,054 matched cases with identical source coordinates and atom ordering but different edit specifications; 892 cases have distinct reference targets. We conduct a target-exchange analysis on the 890 cases with evaluable predictions for both specifications, scoring each output against its instruction-matched target and against the alternative target. AMMAS attains a mean Kabsch RMSD of 2.2268 Å to the matched targets versus 2.4703 Å after exchange, and the matched pairing yields lower error in 611 of 890 cases (68.7%). The outputs therefore track the supplied edit specification rather than the source conformation alone.

Table 16: **Compactness editing: additional metrics.** Success, AUC, and NETP are fractions over 20,000 examples; errors are in Å over evaluable outputs.

Model	AUC[0, 3] ↑	Success @1.5 Å ↑	Success @3 Å ↑	NETP ↑	Raw RMSD ↓	Kabsch RMSD ↓	Raw cMAE ↓	Kabsch cMAE ↓
Qwen3	0.1999	0.0791	0.7433	0.7922	3.8549	2.5311	1.6531	1.0891
ChemDFM-v1.5	0.2248	0.1087	0.7792	0.8373	3.7485	2.4315	1.5997	1.0352
ChemLLM	0.1836	0.0667	0.7158	0.7776	4.0260	2.6045	1.7303	1.1271
LOGOS	0.1805	0.0660	0.7020	0.7399	3.9696	2.5833	1.7069	1.1186
BioMatrix	0.2861	0.1843	0.8244	0.8506	8.1653	2.2291	3.7168	0.9358
AMMAS	**0.3098**	**0.1991**	**0.8696**	**0.9127**	**2.7197**	**2.1216**	**1.1553**	**0.8791**

Table 17: **Donor–acceptor distance editing: additional metrics.** Success, AUC, and NETP are fractions over 20,000 examples; errors are in Å over evaluable outputs.

Model	AUC[0, 3] ↑	Success @1.5 Å ↑	Success @3 Å ↑	NETP ↑	Raw RMSD ↓	Kabsch RMSD ↓	Raw cMAE ↓	Kabsch cMAE ↓
Qwen3	0.1677	0.0703	0.6525	0.6074	3.7173	2.6710	1.6312	1.1603
ChemDFM-v1.5	0.1845	0.0845	0.6866	0.6731	3.6438	2.6145	1.5892	1.1282
ChemLLM	0.1601	0.0659	0.6328	0.6016	3.8528	2.7200	1.6969	1.1862
LOGOS	0.1514	0.0567	0.6061	0.5376	3.7778	2.6987	1.6639	1.1762
BioMatrix	0.2559	0.1784	0.7486	0.7764	7.8098	2.3635	3.5512	1.0032
AMMAS	**0.2754**	**0.1848**	**0.8054**	**0.8183**	**2.7172**	**2.2603**	**1.1714**	**0.9546**

E Ablation Protocols and Detailed Results

We conduct controlled architectural ablations on the conformational question-answering task, which serves as our diagnostic probe: its six balanced question families, purely discrete labels, and single-candidate evaluation isolate architectural mechanisms from generation-head capacity and sampling variation. All variants are initialized from the same Qwen3-4B-Base and independently complete the full CPT stage under identical data, compute, and optimization, followed by SFT on the identical split with the same setup, so comparisons differ only in the ablated architectural factor; no variant starts from another variant's checkpoint. The structure understanding and coordinate-generation tasks reuse the same pretrained backbone, on which the validated mechanisms are carried over. Accuracy follows Appendix C.3 with the complete test denominator; invalid responses count as incorrect. Table 18 reports correct and invalid counts, Table 19 gives expert-placement results by family, and Table 20 reports paired comparisons. Percentage-point differences are computed from unrounded accuracies.

MET layer placement. We compare three layouts with the same SGT backbone and eight MET layers. *Interleaved (AMMAS anchor)* inserts one MET layer after every four SGT layers, as specified in Appendix A.2. *Front-stacked* places all eight MET layers before the SGT backbone. *Back-stacked* places all eight MET layers between the penultimate and final SGT layers. All three retain the same MET internal design and structure-local updates. Interleaving provides the strongest coupling between geometric and linguistic computation depth, consistent with its gains across all six diagnostic families.

Observed-block attention. The causal variant replaces bidirectional attention with causal attention inside observed structure blocks in both MET layers and SGT layers. Text and cross-block attention remain causal.

Table 18: **Ablation counts on the complete 90,000-example test set.** Invalid responses remain in the denominator and count as incorrect. Accuracies are percentages. AMMAS denotes the interleaved configuration with bidirectional observed-block attention.

Configuration	Correct	Accuracy ↑	Invalid ↓
AMMAS	86,232	95.8133	1
Front-stacked experts	82,235	91.3722	1
Back-stacked experts	76,609	85.1211	2
Causal observed-block attention	82,805	92.0056	1

Table 19: **MET placement by question family.** Each family contains 15,000 test examples; entries are percentage accuracies. Question-family definitions are given in Appendix B.4. Bold denotes the best result in each row.

Question family	AMMAS	Front-stacked	Back-stacked
Donor–acceptor contacts	**97.77**	95.02	88.13
Hydrogen-bond candidates	**95.01**	90.73	87.01
Rotatable torsion	**91.38**	82.94	70.65
Scaffold planarity	**95.95**	90.16	81.55
Ring orientation	**99.02**	94.31	88.93
Global shape	**95.75**	95.07	94.46

Table 20: **Paired comparisons between ablation configurations.** Wins and losses are defined from the variant's perspective: a win denotes a correct variant response and an incorrect reference response, and a loss denotes the reverse. Cases where both systems are correct or both are incorrect do not contribute to these counts. The net change in correct predictions equals wins minus losses.

Variant	Reference	Wins	Losses	Net correct
Front-stacked experts	AMMAS	1,466	5,463	$-3,997$
Back-stacked experts	AMMAS	1,464	11,087	$-9,623$
Back-stacked experts	Front-stacked	3,046	8,672	$-5,626$
Causal observed-block attention	AMMAS	1,566	4,993	$-3,427$