ADT: ALL-ATOM DIFFUSION TOKENIZER FOR 3D MOLECULAR STRUCTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete structural tokenization provides a promising route to autoregressive 3D molecular generation. Yet, existing tokenizers face a trade-off between compact representation and all-atom reconstruction: atom-level tokens preserve atomic details but produce long sequences, while compact residue-level protein tokens typically omit side-chain geometry. To overcome this limitation, we propose **A**ll-atom **D**iffusion **T**okenizer (**ADT**), which decouples token granularity from reconstruction resolution. ADT uses one structural token per atom for small molecules and one token per protein residue that jointly represents backbone and side-chain geometry. To mitigate geometric distortion introduced by quantization, we leverage the generative capacity of diffusion models to recover fine-grained atomic geometry beyond deterministic decoding from discrete tokens. Extensive experiments across small molecules and proteins demonstrate that ADT achieves strong reconstruction from compact discrete representations while enabling high-quality autoregressive generation of both small molecules and all-atom protein structures. These results establish compact structural tokens as an effective discrete interface for all-atom molecular reconstruction and generation.

1 INTRODUCTION

Generating three-dimensional molecular structures is a central task in drug discovery and protein engineering (Su et al., 2026; Lin et al., 2025b; Luo et al., 2025; Huang et al., 2024). Molecular interactions depend critically on atomic geometry, from small-molecule conformations to protein backbones and side chains (Jin et al., 2018; Irwin et al., 2022). Many existing approaches directly model molecular structures in continuous coordinate space (Hoogeboom et al., 2022; Xu et al., 2023; Dunn & Koes, 2024; Morehead & Cheng, 2024). While effective, continuous coordinate representations do not naturally fit the autoregressive framework of large language models. Structural tokenization instead maps 3D geometry into discrete representations, enabling molecular structures to be generated autoregressively as token sequences and decoded into atomic coordinates. This formulation allows molecular structure generation to leverage scalable language-modeling paradigms demonstrated across language and biomolecular modeling (Kaplan et al., 2020; Hoffmann et al., 2022; Bhatnagar et al., 2026).

Recent molecular tokenizers encode geometry at different granularities, as shown in Fig. 1. Atom-level tokenizers (Liu et al., 2025; Geikyan & Khachatrian, 2026) assign structural codes to individual atoms, preserving all-atom coverage with sequence lengths that scale with atom count. For proteins, this yields substantially longer sequences than residue-level representations and requires more autoregressive prediction steps. Backbone-based protein tokenizers (Wang et al., 2025; Hayes et al., 2025; Gao et al., 2025) use compact residue-level codes but do not explicitly encode side-chain conformations, limiting faithful reconstruction of the original all-atom structure. Recent approaches further incorporate generative decoders to improve backbone reconstruction (Dilip et al., 2026b) and enable adaptive compression of Cα structures (Dilip et al., 2026a).

These limitations raise two challenges for compact all-atom tokenization. First, *compact all-atom encoding*: side-chain conformations are essential for describing molecular interactions, yet encoding every atom separately produces long token sequences for proteins. A useful representation must retain these atomic details while controlling sequence length.

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.

Second, *accurate reconstruction under quantization error*: mapping continuous conformations to discrete codes loses geometric information, making it difficult to recover precise local geometry and overall molecular structure.

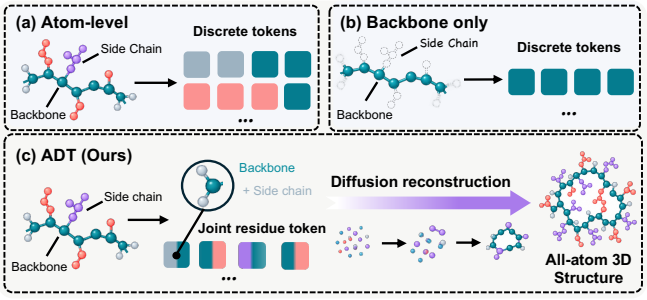

Figure 1: Illustration of tokenizer granularity. *(a)* Atom-level tokenization assigns one token to each atom. *(b)* Backbone-only tokenization omits side-chain geometry. *(c)* ADT jointly encodes backbone and side-chain geometry into one token per residue and reconstructs the all-atom structure.

We address these challenges by introducing **A**ll-atom **D**iffusion **T**okenizer (ADT), which decouples token granularity from reconstruction resolution. For proteins, we jointly encode backbone and side-chain geometry into one structural token per residue. To mitigate geometric distortion introduced by quantization, we formulate reconstruction as conditional generative modeling, exploiting the generative capacity of a diffusion decoder to recover fine-grained atomic geometry from discrete structural tokens and chemical identities.

Specifically, ADT encodes all-atom geometry into one structural token per chemical unit: an atom for small molecules and a residue for proteins. To make these tokens better suited for generative modeling, we introduce a bitwise mutual-information regularizer that promotes effective code utilization. Conditioned on these tokens and their chemical identities, a diffusion decoder reconstructs atomic geometry from the discrete bottleneck. For proteins, the decoder first models residue-level context and then lifts each residue representation to atom-level features, enabling all-atom reconstruction from a single token per residue. Beyond reconstruction, we use the learned tokens for autoregressive molecular generation, where the chemical identity sequence is generated first, followed by the corresponding structural tokens conditioned on the complete chemical sequence. Together, these designs enable compact molecular tokens to support all-atom reconstruction and generation without tying token granularity to atomic resolution.

We comprehensively evaluate our framework on reconstruction and generation across small molecules and proteins. Our experiments show that ADT achieves accurate reconstruction from compact discrete representations, including all-atom protein reconstruction from a single structural token per residue. Beyond reconstruction, our experiments demonstrate that the learned tokens enable high-quality autoregressive generation of both small molecules and all-atom protein structures. Our main contributions are summarized as follows: 1) We introduce a unified discrete representation for small molecules and proteins that decouples token granularity from reconstruction resolution, using one structural token per atom for small molecules and per residue for proteins. 2) We formulate reconstruction as conditional generation and introduce a coarse-to-fine diffusion decoder that recovers fine-grained atomic geometry from discrete structural tokens. 3) Our experiments demonstrate that compact structural tokens support reconstruction and autoregressive generation of both small molecules and all-atom protein structures.

2 Related Work

2.1 Molecular Tokenization

Molecular geometry can be serialized through coordinate strings or invariant geometric sequences (Flam-Shepherd & Aspuru-Guzik, 2023; Zholus et al., 2025; Li et al., 2025), while learned tokenizers compress continuous geometry into discrete structural codes. Such representations have been explored for small molecules (Gao et al., 2026; Geikyan & Khachatrian, 2026) and protein structures, including VQ-based backbone tokenization (Gaujac et al., 2024), ESM3 (Hayes et al., 2025), DPLM-2 (Wang et al., 2025), and ProToken (Lin et al., 2025a). Recent methods further extend structural coverage beyond protein backbones: Bio2Token uses atom-level quantized representations across biomolecular systems (Liu et al., 2025), while BiHiTo introduces hierarchical structural codes (Zheng et al., 2026). At residue granularity, SLAE learns compact local all-atom representations (Chen et al., 2026), and related backmapping methods reconstruct atomistic struc-

tures from compressed representations (Han et al., 2026). We instead formulate molecular tokenization over chemical units, pairing each atom or residue identity with a discrete structural code and decoupling token granularity from all-atom reconstruction resolution, thereby providing a common discrete interface across molecular systems.

2.2 3D Molecular Generation

3D molecular generation has been explored in both continuous coordinate space and discrete token space. Early autoregressive approaches construct molecular structures sequentially by predicting atom types and coordinates (Gebauer et al., 2019; 2022), while diffusion and flow models directly generate molecular geometry for small molecules (Hoogeboom et al., 2022; Peng et al., 2023; Dunn & Koes, 2024) and protein structures (Yim et al., 2023; Watson et al., 2023; Chu et al., 2024; Geffner et al., 2026; Li et al., 2026). More recently, discrete structural representations bridge 3D geometry and language modeling by converting molecular geometry into token sequences for autoregressive generation. This paradigm has been explored for small molecules (Gao et al., 2026) and proteins (Gaujac et al., 2024; Hayes et al., 2025; Wang et al., 2025; Dilip et al., 2026b;a). However, discrete protein generation largely operates at Cα or backbone resolution, omitting side-chain and local atomic details needed to represent interaction interfaces and functional sites (Chu et al., 2024). This limits its applicability to structure-based design tasks that require atomic-level detail. We address this limitation by combining compact autoregressive structural tokens with generative all-atom decoding.

3 Preliminaries

3D Molecular Representation. A 3D molecular structure is represented as a sequence of L structural units, $\mathcal{M} = (\mathcal{U}_1, \ldots, \mathcal{U}_L)$, where each structural unit $\mathcal{U}_i$ contains a set of atoms indexed by $\mathcal{A}_i$, with Cartesian coordinates $\mathbf{x}_{i,j} \in \mathbb{R}^3$ for $j \in \mathcal{A}_i$. For small molecules, each structural unit corresponds to a single atom, whereas for proteins, each structural unit corresponds to a residue containing multiple atoms. We denote the chemical identities of the structural units by $C = (c_1, \ldots, c_L)$, corresponding to atom types for small molecules and amino-acid types for proteins. The atomic geometry is represented as $\mathbf{X} \in \mathbb{R}^{L \times A \times 3}$, where A denotes the number of atomic identities per structural unit.

Discrete Molecular Tokenization. Given a molecular structure $\mathcal{M}$, a discrete molecular tokenizer maps it into continuous latent representations through an encoder f_θ, followed by a quantization function Q that converts the latent representations into a sequence of discrete tokens:

$$\mathbf{H} = f_\theta(\mathcal{M}), \qquad \mathbf{S} = Q(\mathbf{H}) = (s_1, \ldots, s_L), \quad s_i \in \{0, \ldots, K-1\}, \tag{1}$$

where K denotes the size of the discrete vocabulary. A decoder g_ϕ reconstructs the atomic geometry from the discrete structural tokens $\mathbf{S}$, conditioned on the structural-unit identities $\mathbf{C}$.

4 Methodology

As shown in Fig. 2, we introduce All-atom Diffusion Tokenizer (ADT), which decouples token granularity from reconstruction resolution. Each small-molecule atom and protein residue is represented by a single structural token, while a diffusion decoder mitigates quantization-induced distortion to recover fine-grained all-atom geometry. This design provides a common discrete interface across molecular modalities while retaining all-atom reconstruction.

4.1 Unified Molecular Structure Tokenization

Unified Structure Representation. Small molecules and proteins exhibit different structural granularities, with atom-level units for small molecules and residue-level units for proteins. Following the structural-unit representation defined above, we jointly embed the chemical identity and atomic geometry of each structural unit into a common sequence space:

$$\mathbf{H}^0 = \mathrm{MLP}\left([\mathbf{X}; \mathbf{M}]\right) + \mathbf{E}(\mathbf{C}), \qquad \mathbf{H} = \mathrm{TransformerEnc}(\mathbf{H}^0) \in \mathbb{R}^{L \times d}. \tag{2}$$

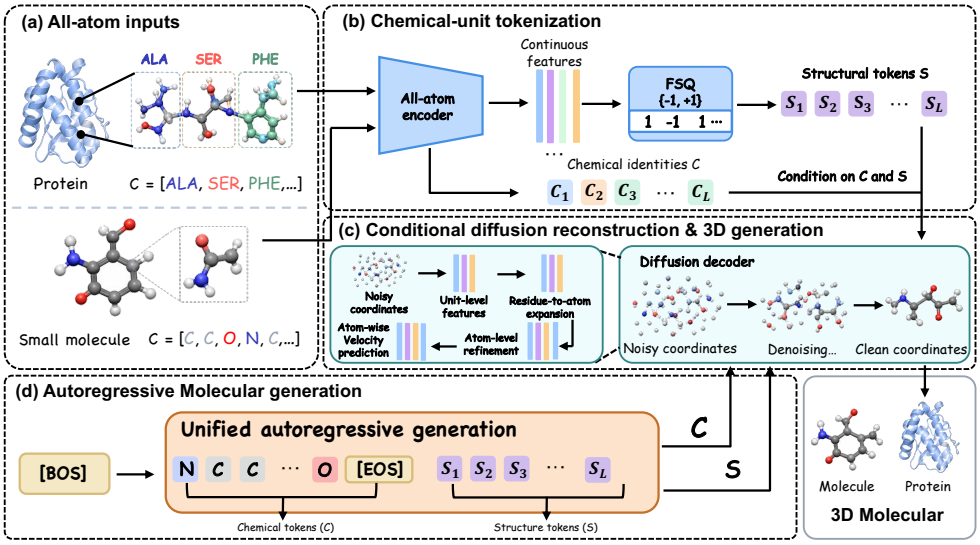

Figure 2: Overview of ADT. *(a)* All-atom structures are represented at chemical-unit granularity. *(b)* Each unit is mapped to a discrete structural token via FSQ. *(c)* Conditional diffusion reconstructs all-atom geometry, with residue-to-atom expansion for proteins. *(d)* ADT enables autoregressive generation of all-atom molecular structures.

Here, $\mathbf{M}$ is a binary mask indicating valid atoms, and $\mathbf{E}(\mathbf{C})$ denotes the structural-unit identity embeddings, corresponding to atom types for small molecules and amino-acid types for proteins. This formulation yields one contextual representation per structural unit while preserving its underlying atomic geometry.

Discrete Molecular Tokenization. We further map the unified continuous representations into a common discrete space using finite scalar quantization (FSQ) (Mentzer et al., 2024). Specifically, given the contextual representation $\mathbf{H}$, we project each structural unit into an m-dimensional quantization space, independently quantize each dimension into two symmetric levels, and map the resulting binary code to a discrete token:

$$\mathbf{Z} = \mathrm{MLP}(\mathbf{H}) \in \mathbb{R}^{L\times m}, \qquad \widehat{z}_{i,j} = Q(z_{i,j}) \in \{-1, +1\}, \qquad s_i = \sum_{j=1}^{m} \frac{\widehat{z}_{i,j}+1}{2}\, 2^{j-1}. \tag{3}$$

This yields a structural token sequence $\mathbf{S} = (s_1, \ldots, s_L)$ over the same vocabulary of size $K = 2^m$, providing a common discrete representation for structural units across molecular modalities.

4.2 Molecular Structure Reconstruction

Coarse-to-Fine Diffusion Reconstruction. We reconstruct all-atom geometry from discrete structural tokens through coarse-to-fine diffusion decoding. The decoder progressively lifts token-level representations to atom-level representations, decoupling token granularity from reconstruction resolution. Given the discrete structural tokens $\mathbf{S}$, we recover their quantized representations $\widehat{\mathbf{Z}}$ and combine them with the corresponding chemical identity embeddings to form the conditioning representations:

$$\mathbf{D} = \mathrm{MLP}(\widehat{\mathbf{Z}}) + \mathbf{E}_{\mathrm{dec}}(\mathbf{C}) \in \mathbb{R}^{L\times d}, \qquad \mathbf{P}_t = \mathcal{D}(\mathbf{X}_t; \mathbf{D}, t) \in \mathbb{R}^{L\times d}, \tag{4}$$

where $\mathbf{E}_{\mathrm{dec}}$ is a learnable decoder-side embedding of the chemical identities $\mathbf{C}$, and $\mathbf{D}$ denotes the resulting conditioning representations. $\mathbf{X}_t$ denotes the noisy molecular coordinates at diffusion time t, which are projected into structural-unit representations and processed by the diffusion Transformer $\mathcal{D}$. The resulting $\mathbf{P}_t$ contains one contextual representation for each structural unit.

Since each protein token represents an entire residue, directly decoding all atoms from a single residue-level representation creates a bottleneck for fine-grained atomic reconstruction. We there-

fore adopt a coarse-to-fine strategy that expands each intermediate residue representation into A atom-wise representations, conditioned on learnable atom embeddings and noisy coordinates:

$$\begin{aligned} \mathbf{U}^0_{t,i,a} &= \mathbf{W}\mathbf{P}_{t,i} + \mathbf{e}_a + \mathrm{MLP}_1(\mathbf{x}_{t,i,a}), \\ \mathbf{U}_{t,i} = \mathrm{Transformer}(\mathbf{U}^0_{t,i}), &\qquad \widehat{\mathbf{v}}_{t,i,a} = \mathrm{MLP}_2(\mathbf{U}_{t,i,a}) \in \mathbb{R}^3. \end{aligned} \tag{5}$$

Here, $\mathbf{e}_a$ is a learnable embedding that distinguishes individual atoms within each residue. The remaining decoder layers refine the expanded atom-wise representations, which are finally projected to atom-wise denoising velocities. This transition from $\mathbb{R}^{L\times d}$ to $\mathbb{R}^{L\times A\times d}$ decouples residue-level tokenization from all-atom reconstruction resolution.

Reconstruction Objective. We train each diffusion decoder using a velocity-prediction objective. Conditioned on the discrete structural representation, the diffusion decoder predicts the velocity of each atom from its noisy coordinates. For a molecule with N atoms, the objective is:

$$\mathcal{L}_{\mathrm{recon}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\mathbf{v}}_{t,i} - \mathbf{v}_{t,i}\|_2^2\right], \tag{6}$$

where $\widehat{\mathbf{v}}_{t,i}$ and $\mathbf{v}_{t,i}$ denote the predicted and target velocities of the i-th atom, respectively. This objective propagates atom-level reconstruction supervision through the discrete bottleneck, encouraging each token to retain the structural information required for recovering molecular geometry.

Mutual-Information Regularization. High reconstruction fidelity alone does not guarantee effective generative modeling in the discrete space. We therefore introduce a bitwise mutual-information regularizer to encourage diverse code utilization across structural units while maintaining confident individual assignments. For the j-th pre-quantization scalar $z_{i,j}$ of structural unit $\mathcal{U}_i$, we define the soft binary assignment:

$$p_{i,j} = q(B_j = 1 \mid \mathbf{h}_i) = \frac{\exp(z_{i,j}/\tau)}{\exp(z_{i,j}/\tau) + \exp(-z_{i,j}/\tau)}, \tag{7}$$

where B_j denotes the j-th binary quantization variable and τ is the temperature. Using the binary Shannon entropy $h(p) = -p\log p - (1-p)\log(1-p)$, we estimate the normalized bitwise mutual information over all valid structural units in each mini-batch as:

$$\mathcal{L}_{\mathrm{MI}} = -\frac{1}{m\log 2}\sum_{j=1}^{m}\left[h\left(\mathbb{E}_i[p_{i,j}]\right) - \mathbb{E}_i\left[h(p_{i,j})\right]\right]. \tag{8}$$

The first term encourages diverse bit utilization across structural units, while the second encourages confident assignments for individual units. The resulting $\mathcal{L}_{\mathrm{MI}}$ is used as a regularization term during tokenizer training.

Protein Geometry Regularization. For proteins, we complement the reconstruction objective with a backbone loss $\mathcal{L}_{\mathrm{bb}}$ and auxiliary geometric losses on bond lengths, bond angles, steric clashes, and C_α pairwise distances. The resulting training objective is:

$$\mathcal{L}_{\mathrm{prot}} = \mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{MI}}\mathcal{L}_{\mathrm{MI}} + \boldsymbol{\omega}^\top[\mathcal{L}_{\mathrm{bb}}, \mathcal{L}_{\mathrm{bond}}, \mathcal{L}_{\mathrm{angle}}, \mathcal{L}_{\mathrm{clash}}, \mathcal{L}_{\mathrm{dist}}]^\top. \tag{9}$$

where $\boldsymbol{\omega} \in \mathbb{R}^5$ balances the five auxiliary terms. For small-molecule tokenizer, the training objective is $\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{MI}}\mathcal{L}_{\mathrm{MI}}$. Detailed definitions are provided in Appendix C.2.

4.3 Autoregressive Molecule Generation

Discrete structural tokens allow both chemical identity and 3D geometry to be generated within the same autoregressive sequence-modeling paradigm. Building on this representation, we factorize molecular generation into chemical sequence generation followed by structural token generation conditioned on the complete chemical sequence. Let $\mathbf{C} = (c_1, \ldots, c_L, \mathrm{EOS})$ and $\mathbf{S} = (s_1, \ldots, s_L)$ denote the chemical and structural sequences, respectively. We factorize their joint distribution as:

$$p_\theta(\mathbf{C}, \mathbf{S}) = p_\theta(\mathbf{C})\, p_\theta(\mathbf{S} \mid \mathbf{C}) = \left[\prod_{i=1}^{L+1} p_\theta(c_i \mid \mathbf{C}_{<i})\right]\left[\prod_{i=1}^{L} p_\theta(s_i \mid \mathbf{C}, \mathbf{S}_{<i})\right]. \tag{10}$$

Thus, each structural prediction has access to chemical information from all molecular positions. Both autoregressive processes are trained jointly with teacher forcing:

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{i=1}^{L+1} \log p_\theta(c_i \mid \mathbf{C}_{<i}) - \beta \sum_{i=1}^{L} \log p_\theta(s_i \mid \mathbf{C}, \mathbf{S}_{<i}). \tag{11}$$

where β balances the chemical and structural autoregressive objectives.

At inference, chemical tokens are sampled until EOS, determining the sequence length L. The model then generates one structural token for each chemical position, and the molecular decoder converts $(\mathbf{S}, \mathbf{C})$ into 3D coordinates $\mathbf{X}$. For structure generation from a given chemical sequence, $\mathbf{C}$ is fixed and only the structural autoregression is performed.

5 Experiments

We evaluate ADT on both small molecules and proteins, covering 3D structure reconstruction and autoregressive generation. We report reconstruction and generation results, followed by ablation studies on the key design choices of ADT.

5.1 Experiment Setup

Datasets. We conduct experiments on benchmarks spanning small molecules and proteins. For small molecules, we use QM9 (Ramakrishnan et al., 2014), containing approximately 130K organic molecules with up to nine heavy atoms, and ∇^2DFT (Khrabrov et al., 2024), using 1.9M molecules for training and the three test splits for evaluation on unseen molecules with 8–27 heavy atoms. For proteins, we use AFDB SwissProt (Varadi et al., 2022), with CASP14 (Kryshtafovych et al., 2021)/CASP15 (Kryshtafovych et al., 2023) held out for protein reconstruction evaluation. We exclude all evaluation targets from the tokenizer training set and remove training proteins with at least 80% sequence identity to evaluation target, as measured over Foldseek (Van Kempen et al., 2024). Further details are provided in Appendix C.

Evaluation Metrics. Small-molecule reconstruction is evaluated using structure match, validity, and RMSD. For protein reconstruction, we report Cα-lDDT, backbone and all-atom RMSD, all-atom lDDT, TM-score, GDT-TS, and GDT-HA. Small-molecule generation is evaluated using validity, uniqueness, diversity, and PoseBusters validity, while protein generation metrics are specified in the corresponding experiments. Detailed definitions and evaluation protocols are provided in Appendix C.4.

Training and Evaluation. We use a codebook size of 8,192 and Transformer-based encoders with diffusion decoders for both molecular modalities. The tokenizers are optimized with AdamW with learning rate 3×10^{-4} on a single NVIDIA H200 GPU. Detailed model architectures, optimization settings, loss weights, and evaluation protocols are provided in Appendix C.1.

5.2 Reconstruction on small molecules and all-atom proteins

Reconstruction on small molecules. We evaluate molecular reconstruction on QM9 and the three ∇^2DFT test splits, covering both in-distribution molecules and generalization to unseen structures and scaffolds. Tables 1 and 2 reveal three main results: 1) Our method consistently outperforms existing discrete tokenizers. Compared with CoordToken, it reduces reconstruction RMSD by 28.2% on QM9 and 23.8–25.6% across the three ∇^2DFT test splits, while achieving a higher match rate on QM9 and maintaining 100% validity on ∇^2DFT. 2) Discrete tokenization does not preclude high reconstruction accuracy. On QM9, our method surpasses the continuous GeoLDM and ADiT baselines in both match rate and RMSD. 3) The advantage extends beyond QM9 to larger and more diverse molecules. Across the $\nabla^2 DFT$ test splits, including unseen structures and out-of-distribution

Table 1: Reconstruction results on QM9.

Type	Method	Match ↑	RMSD ↓
Continuous	GeoLDM (Xu et al., 2023)	87.2	0.1820
	ADiT (Joshi et al., 2025)	94.6	0.0890
	SinAE (Ren et al., 2026)	100.0	0.0007
Discrete	Mol-StrucTok (Gao et al., 2026)	90.7	0.1280
	CoordToken Geikyan et al. (2026)	99.4	0.0478
	ADT (Ours)	**99.7**	**0.0343**

Table 2: Reconstruction performance on the ∇^2DFT benchmark.

Method	Conformer		Structure		Scaffolds	
	RMSD ↓	Valid ↑	RMSD ↓	Valid ↑	RMSD ↓	Valid ↑
mol2token (Liu et al., 2025)	0.2010	47.1	0.2010	47.1	0.2010	47.1
CoordToken Geikyan et al. (2026)	0.0671	100.0	0.0673	100.0	0.0671	100.0
ADT (Ours)	**0.0499**	**100.0**	**0.0512**	**100.0**	**0.0511**	**100.0**

Table 3: Protein reconstruction performance on CASP14 and CASP15. DPLM-2 variants recover all-atom structures using different backbone-to-all-atom reconstruction methods.

	Method	Cα-lDDT ↑	BB RMSD ↓	AA lDDT ↑	AA RMSD ↓	TM-score ↑	GDT-TS ↑	GDT-HA ↑
CASP14	DPLM-2 (Wang et al., 2025)	0.8820	2.8576	–	–	0.8860	0.7718	0.5724
	+ Rosetta fixbb	0.8820	2.8576	0.7532	3.4918	0.8859	0.7717	0.5723
	+ PDBFixer	0.8820	2.8576	0.6632	3.9657	0.8859	0.7717	0.5723
	+ cg2all (nofix)	0.8820	2.8093	0.7710	3.4120	0.8859	0.7718	0.5722
	+ cg2all (fixed)	0.8819	2.8576	0.7611	3.4358	0.8859	0.7717	0.5723
	ADT (Ours)	**0.8970**	**0.9628**	**0.8223**	**1.3157**	**0.9695**	**0.9398**	**0.7686**
CASP15	DPLM-2 (Wang et al., 2025)	0.8933	4.0223	–	–	0.8644	0.7032	0.5104
	+ Rosetta fixbb	0.8933	4.0223	0.7671	4.5931	0.8644	0.7032	0.5104
	+ PDBFixer	0.8933	4.0223	0.6767	5.0213	0.8644	0.7032	0.5104
	+ cg2all (nofix)	0.8933	3.9796	0.7879	4.5037	0.8644	0.7029	0.5101
	+ cg2all (fixed)	0.8933	4.0224	0.7772	4.5288	0.8644	0.7032	0.5104
	ADT (Ours)	**0.8972**	**1.5191**	**0.8258**	**1.8261**	**0.9470**	**0.9001**	**0.7285**

scaffolds, ADT reduces RMSD over existing discrete tokenizers by 23.8–25.6% while maintaining complete validity.

All-atom protein reconstruction. Table 3 shows that residue-level tokenization can enable accurate all-atom reconstruction while preserving backbone fidelity. Since DPLM-2 reconstructs only protein backbones, we recover all atoms from the reconstructed backbone using PDBFixer, cg2all (Heo & Feig, 2024), and Rosetta (Leaver-Fay et al., 2011). We additionally evaluate cg2all with the input backbone fixed, denoted as cg2all (fixed). Compared with DPLM-2 followed by all-atom completion, our method substantially reduces all-atom and backbone RMSD on CASP14 and CASP15, while maintaining strong backbone lDDT. Notably, each residue is represented by only a single structural token yet decoded into its full atomic geometry. These results demonstrate that compact token granularity need not constrain reconstruction resolution, supporting our design of decoupling residue-level encoding from all-atom decoding.

5.3 Generation on small molecules and all-atom proteins

Generation on small molecules. We evaluate unconditional molecular generation on QM9. As shown in Table 4, ADT outperforms the evaluated discrete baselines across all four metrics, improving Valid and PB Valid over Mol-StrucTok from 0.97/0.88 to 1.00/0.94. ADT is also competitive with continuous generative models, matching the best PB Valid while maintaining high uniqueness and diversity. Notably, SinAE achieves substantially lower reconstruction error but comparable generation performance, indicating that reconstruction fidelity alone does not determine generative quality. Overall, these results demonstrate that ADT provides a strong discrete representation for autoregressive 3D molecular generation.

Sequence-Conditioned Protein Generation. We train Qwen3.5-0.8B (Team, 2026) on tokenized AFDB structures for sequence-conditioned protein structure generation, with training details provided in Appendix C.3. To isolate the effect of structural tokenization, we use the same Qwen3.5-0.8B backbone with either the DPLM-2 tokenizer or ADT, while keeping all other components fixed. Given an amino-acid sequence, the model autoregressively predicts one structural token per residue. ADT directly decodes these tokens into all-atom coordinates using the frozen diffusion decoder, whereas DPLM-2 outputs are further converted to all-atom structures using the side-chain reconstruction methods described in Section 5.2.

We evaluate on 183 CAMEO22 targets using sequence-length matching, Cα-lDDT, backbone lDDT, TM-score, and all-atom lDDT. As shown in Table 5, ADT achieves an all-atom lDDT of 0.543,

Table 4: Unconditional molecular generation results on QM9.

Type	Method	Valid ↑	Unique ↑	Div. ↑	PB Valid ↑
Continuous	EQGAT-diff (Le et al., 2024)	0.99	1.00	0.89	0.94
	SemlaFlow (Irwin et al., 2024)	0.99	0.95	0.89	0.94
	FlowMol (Dunn & Koes, 2024)	0.97	0.97	0.92	0.94
	ADiT (Joshi et al., 2025)	0.97	0.97	0.90	0.92
	SinAE (Ren et al., 2026)	0.97	0.97	0.91	0.94
Discrete	Mol-StrucTok (Gao et al., 2026)	0.97	0.85	0.87	0.88
	CoordToken Geikyan et al. (2026)	0.95	0.93	0.90	0.86
	ADT (Ours)	**1.00**	**0.96**	**0.91**	**0.94**

Table 5: Sequence-conditioned protein structure generation on CAMEO22.

Method	Length Match	Cα-lDDT	BB-lDDT	TM-score	AA-lDDT
DPLM-2	1.0	0.5607	0.5683	0.633	–
+ Rosetta fixbb	1.0	0.5607	0.5683	0.633	0.482
+ PDBFixer	1.0	0.5607	0.5683	0.633	0.372
+ cg2all nofix	1.0	0.5607	0.5683	0.633	0.503
+ cg2all fixed	1.0	0.5607	0.5683	0.633	0.496
ADT (Ours)	**1.0**	**0.5740**	**0.5801**	**0.665**	**0.543**

outperforming all post-processed DPLM-2 variants, including cg2all (nofix) at 0.503. ADT also improves Cα-lDDT from 0.5607 to 0.5740 and TM-score from 0.633 to 0.665, while maintaining exact sequence-length matching. These results show that a single residue-level structural token can jointly represent backbone and side-chain geometry, enabling direct all-atom generation without a separate side-chain reconstruction stage.

5.4 Effect of All-Atom Joint Encoding

We ablate how backbone and side-chain geometry should be represented by comparing three encoding strategies under the same residue-level tokenization framework: encoding backbone geometry only, encoding backbone and side chains separately (Separate BB+SC), and jointly encoding all-atom geometry into a single structural token per residue. Separate BB+SC independently encodes backbone and side-chain geometry into two residue-level structural tokens, whereas our joint variant represents both with a single token. All variants use the same model architecture and training protocol, differing only in the structural information encoded by the discrete representation. Table 6 reveals three key observations: 1) **Geometric Complementarity**: Joint all-atom encoding consistently outperforms backbone-only encoding across all metrics, while achieving comparable side-chain reconstruction to separate encoding and stronger overall backbone fidelity, revealing complementary geometric information between backbone and side chains. 2) **Improved Structural Fidelity**: Compared with separate encoding, joint encoding better preserves global structural fidelity, as reflected by consistently higher Cα-lDDT, TM-score, GDT-TS, and GDT-HA. 3) **Compact Representation**: A single residue-level discrete token enables high-fidelity all-atom reconstruction, decoupling token granularity from reconstruction resolution.

5.5 Analysis and Ablation Studies

Effect of Bitwise Regularization. We investigate the effect of bitwise regularization on reconstruction and downstream generation. As shown in Fig. 3, mild regularization improves PB Valid, suggesting that shaping the discrete representation with the bitwise mutual-information objective facilitates downstream autoregressive modeling. Both PB Valid and reconstruction RMSD remain relatively stable across a broad range of λ, indicating limited sensitivity to the exact regularization strength. At larger λ, reconstruction RMSD gradually increases without further improvements in PB Valid, motivating the use of moderate regularization.

Table 6: Ablation of protein structural encoding strategies on CASP14 and CASP15.

	Method	Cα-lDDT ↑	BB RMSD ↓	AA lDDT ↑	AA RMSD ↓	SC RMSD ↓	TM-score ↑	GDT-TS ↑	GDT-HA ↑
CASP14	Backbone only	0.8802	1.2804	0.7699	1.9269	2.3858	0.9555	0.9038	0.7241
	Separate BB+SC	0.8762	1.3105	**0.8262**	1.5075	1.6763	0.9549	0.9106	0.7238
	Joint All-atom (ADT)	**0.8970**	**0.9628**	0.8223	**1.3157**	**1.5938**	**0.9695**	**0.9398**	**0.7686**
CASP15	Backbone only	0.8864	1.6517	0.7866	2.2386	2.6672	0.9365	0.8875	0.7150
	Separate BB+SC	0.8827	**1.4884**	**0.8306**	**1.7711**	**2.0341**	0.9416	0.8896	0.7065
	Joint All-atom (ADT)	**0.8972**	1.5191	0.8258	1.8261	2.0779	**0.9470**	**0.9001**	**0.7285**

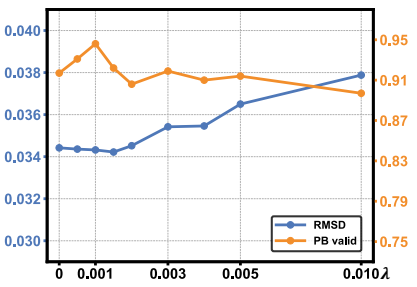

Figure 3: Effect of MI regularization on reconstruction and generation.

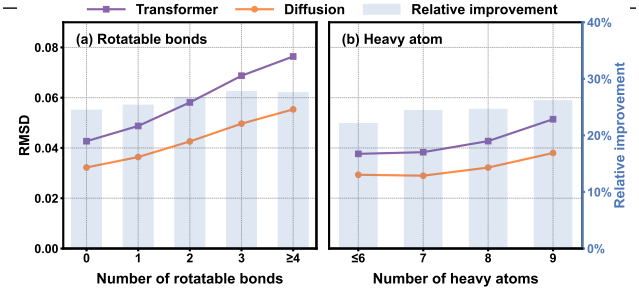

Figure 4: Reconstruction gains from generative decoding across molecular complexity.

Ablating Diffusion Reconstruction. To isolate the contribution of generative decoding, we replace the conditional diffusion decoder with a Transformer that directly predicts atomic coordinates, while keeping the remaining architectural and training settings unchanged. As shown in Table 7, diffusion decoding consistently outperforms the direct Transformer decoder across different depths, with the improvement increasing as the number of decoder layers grows. This trend further supports the effectiveness of diffusion decoding in mitigating quantization-induced geometric distortion.

We further stratify the QM9 test set by the number of heavy atoms and rotatable bonds to examine how this benefit scales with molecular complexity, as illustrated in Fig. 4. Diffusion decoding yields larger absolute reductions in reconstruction error as complexity increases, while maintaining broadly comparable relative improvements. The growing absolute advantage on more challenging molecules suggests that generative decoding becomes increasingly valuable as the geometric information discarded by discrete quantization becomes harder to recover.

Table 7: Effect of decoder depth on QM9 reconstruction.

Decoder	Layers	Match ↑	RMSD ↓	PB Valid ↑	RMSD Gain ↑
Transformer	4	97.5	0.0695	0.88	**27.19%**
Diffusion	4	**98.3**	**0.0506**	**0.89**	
Transformer	8	98.4	0.0636	0.90	**33.18%**
Diffusion	8	**99.1**	**0.0425**	**0.91**	
Transformer	12	98.9	0.0584	0.92	**41.27%**
Diffusion	12	**99.7**	**0.0343**	**0.93**	

6 Conclusion

We introduced All-atom Diffusion Tokenizer (ADT), a discrete tokenizer for 3D molecular structures that decouples token granularity from reconstruction resolution. ADT represents small molecules with one structural token per atom and proteins with one token per residue, while retaining all-atom reconstruction through conditional diffusion decoding. By leveraging generative decoding to mitigate geometric information loss introduced by quantization, ADT achieves accurate reconstruction from compact discrete representations. The resulting tokens further support autoregressive generation of both small molecules and all-atom protein structures. Together, our results demonstrate that compact structural tokens can serve as an effective discrete interface for all-atom molecular modeling, bridging efficient token-based generation with fine-grained atomic geometry.

REFERENCES

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel Curran, Alexander Hoffnagle, Kyle Ching, Michael Martyn, Stephen Nayfach, Jeffrey Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. *Advances in Neural Information Processing Systems*, 38:46109–46145, 2026.

Yilin Chen, Tianyu Lu, Cizhang Zhao, Hannah Wayment-Steele, and Po-Ssu Huang. SLAE: Strictly local all-atom environment for protein representation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4HDRkXxqNs.

Alexander E Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, 2024.

Rohit Dilip, Ayush Varshney, and David Van Valen. Adaptive protein tokenization. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=UG06mAANLH.

Rohit Dilip, Evan Zhang, Ayush Varshney, and David Van Valen. Flow autoencoders are effective protein tokenizers. In *International Conference on Learning Representations*, volume 2026, pp. 145901–145926, 2026b.

Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *ArXiv*, pp. arXiv–2404, 2024.

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.

Kaiyuan Gao, Yusong Wang, Haoxiang Guan, Zun Wang, Qizhi Pei, John Hopcroft, Kun He, and Lijun Wu. Tokenizing 3d molecule structure with quantized spherical coordinates. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 291–301, 2026.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. Foldtoken: Learning protein language via vector quantization and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 219–227, 2025.

Benoit Gaujac, Jérémie Donà, Liviu Copoiu, Timothy Atkinson, Thomas Pierrot, and Thomas D Barrett. Learning the language of protein structure. *arXiv preprint arXiv:2405.15840*, 2024.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 67350–67395, 2026.

Filya Geikyan and Hrant Khachatrian. Accurate tokenization of 3d small organic molecules. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026.

Xu Han, Yuancheng Sun, Kai Chen, Yuxuan Ren, Kang Liu, and Qiwei Ye. Constraint decoupled latent diffusion for protein backmapping. *Journal of Chemical Theory and Computation*, 22(1): 624–637, 2026.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Lim Heo and Michael Feig. One bead per residue can describe all-atom protein structures. *Structure*, 32(1):97–111, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2-d and 3-d graph diffusion models for complete molecule generation. *IEEE transactions on neural networks and learning systems*, 35(9):11857–11871, 2024.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=89QPmZjIhv.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, Sergey Nikolenko, Elena Tutubalina, and Artur Kadurin. ∇^2dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. *Advances in Neural Information Processing Systems*, 37:36869–36889, 2024.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *International Conference on Learning Representations*, volume 2024, pp. 26244–26266, 2024.

Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, volume 487, pp. 545–574. Elsevier, 2011.

Shaoning Li, Le Zhuo, Yusong Wang, Mingyu Li, Xinheng He, Fandi Wu, Hongsheng Li, and Pheng-Ann Heng. Proteinae: Protein diffusion autoencoders for structure encoding. In *International Conference on Learning Representations*, volume 2026, pp. 69939–69960, 2026.

Xiner Li, Limei Wang, Youzhi Luo, Carl Edwards, Shurui Gui, Yuchao Lin, Heng Ji, and Shuiwang Ji. Geometry informed tokenization of molecules for language model generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4umRQdvuW5.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, Zicheng Ma, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Jun Zhang, and Yi Qin Gao. Unifying sequence-structure coding for advanced protein engineering via a multimodal diffusion transformer. *Chemical Science*, 16(24):11087–11102, 2025a.

Xiaohan Lin, Yijie Xia, Yanheng Li, Yu-Peng Huang, Shuo Liu, Jun Zhang, and Yi Qin Gao. In-silico 3d molecular editing through physics-informed and preference-aligned generative foundation models. *Nature Communications*, 16(1):6043, 2025b.

Andrew Liu, Axel Elaldi, Nathan Russell, and Olivia Viessmann. Bio2token: All-atom tokenization of any biomolecular structure with mamba. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025. URL https://openreview.net/forum?id=EkV2sPrMy5.

Yanchen Luo, Zhiyuan Liu, Yi Zhao, Sihang Li, Hengxing Cai, Kenji Kawaguchi, Tat-Seng Chua, Yang Zhang, and Xiang Wang. Towards unified and lossless latent space for 3d molecular latent diffusion modeling. *Advances in Neural Information Processing Systems*, 38:30718–30748, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3d molecule generation and optimization. *Communications Chemistry*, 7(1):150, 2024.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. *ICML*, 2023.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Yuxuan Ren, Fan Yang, Jianhua Yao, and Yatao Bian. Sinae: A single-architecture flow-matching autoencoder for cross-domain atomic systems. *arXiv preprint arXiv:2607.12380*, 2026.

Xavier Robin, Juergen Haas, Rafal Gumienny, Anna Smolinski, Gerardo Tauriello, and Torsten Schwede. Continuous automated model evaluation (cameo)—perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1977–1986, 2021.

Qun Su, Qiaolin Gou, Hui Zhang, Jike Wang, Huiyong Sun, Renling Hu, Rui Qin, Huanxiang Liu, Tingjun Hou, and Yu Kang. Lamgen: Llm-based 3d molecular generation for multi-target drug design. *Nature communications*, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations*, volume 2025, pp. 35463–35490, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi S. Jaakkola. Se(3) diffusion model with application to protein backbone generation. In *ICML*, pp. 40001–40039, 2023. URL https://proceedings.mlr.press/v202/yim23a.html.

Ruochong Zheng, Yutian Liu, Yian Zhao, Zhiwei Nie, Xuehan Hou, Chang Liu, Siwei Ma, Youdong Mao, and Jie Chen. Bihito: Biomolecular hierarchy-inspired tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28848–28856, 2026.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26083–26091, 2025.

ADT: All-Atom Diffusion Tokenizer for 3D Molecular Structures

Appendix

The appendix is organized as follows:

- Appendix A provides further experimental results, including fine-grained reconstruction analyses, comparisons across token granularities and discrete protein tokenizers, and qualitative results.
- Appendix B presents additional ablation studies on structural information in discrete tokens, coarse-to-fine decoding, discrete capacity, and autoregressive factorization.
- Appendix C outlines additional implementation details, including model architectures, training settings, and evaluation metrics.
- Appendix D contains the limitations of our work.
- Appendix E describes the use of large language models during manuscript preparation.

A Further Experimental Results

A.1 Comparison Across Token Granularities

Atom-level tokenization provides direct coverage of all-atom protein geometry, but requires one structural token per atom, resulting in substantially longer sequences than residue-level tokenization. On CASP14 and CASP15, proteins contain on average 2,141.3 and 2,598.5 atoms but only 268.6 and 336.5 residues, respectively, requiring approximately **8.0**$\times$ and **7.7**$\times$ as many structural tokens as residue-level tokenization. Table 8 compares all-atom reconstruction across these token granularities. Atom-level tokenizers achieve lower reconstruction error, whereas ADT directly reconstructs complete atomic geometry from a single structural token per residue and substantially improves over residue-level backbone tokenization followed by external all-atom completion.

Table 8: All-atom protein reconstruction across token granularities on CASP14 and CASP15.

	Model	CASP14		CASP15	
		AA RMSD ↓	TM-score ↑	AA RMSD ↓	TM-score ↑
Atom-level	Bio2Token (Liu et al., 2025)	0.644	0.983	0.693	0.981
	BiHiTo (Zheng et al., 2026)	0.515	0.987	0.585	0.986
Residue-level	DPLM-2 (Wang et al., 2025)	–	0.8860	–	0.8644
	+ cg2all (no fix)	3.4120	0.8859	4.5037	0.8644
	+ PDBFixer	3.9657	0.8859	5.0213	0.8644
	+ cg2all (fixed)	3.4358	0.8859	4.5288	0.8644
	+ Rosetta fixbb	3.4918	0.8859	4.5931	0.8644
	ADT (ours)	1.3157	0.9695	1.8261	0.9470

Table 9: Protein structure reconstruction performance on CASP14 and CASP15.

	CASP14					CASP15				
Method	**Cα-lDDT ↑**	**BB RMSD ↓**	**TM-score ↑**	**GDT-TS ↑**	**GDT-HA ↑**	**Cα-lDDT ↑**	**BB RMSD ↓**	**TM-score ↑**	**GDT-TS ↑**	**GDT-HA ↑**
ESM3	0.9653	1.9934	0.9552	0.9041	0.8068	0.9672	2.0685	0.9455	0.8693	0.7470
DPLM-2	0.8820	2.8576	0.8860	0.7718	0.5724	0.8933	4.0223	0.8644	0.7032	0.5104
ADT (Ours)	0.8970	**0.9628**	**0.9695**	**0.9398**	0.7686	0.8972	**1.5191**	**0.9470**	**0.9001**	0.7285

Table 10: Unconditional protein structure generation.

Method	**Des. ↑**	**Div. ↑**	**DPT ↓**	**Nov. ↑**
ESM3	0.61	127	0.37	**0.84**
DPLM-2	0.63	130	0.37	0.72
ADT (Ours)	**0.68**	**139**	**0.35**	0.76

A.2 Comparison with Discrete Protein Tokenizers

For completeness, we additionally compare three discrete protein tokenizers, ESM3 (Hayes et al., 2025), DPLM-2 (Wang et al., 2025), and ADT, on protein structure reconstruction. ESM3 is pre-trained at substantially larger scale, covering 2.78 billion protein sequences, whereas our tokenizer is trained on only 330K structures from AFDB SwissProt. Nevertheless, ESM3 reports all-atom reconstruction RMSDs of approximately 1.30 and 1.80 Å on CASP14 and CASP15, respectively, compared with 1.32 and 1.83 Å for our method, indicating comparable all-atom reconstruction accuracy. Since the publicly released ESM3 structure decoder exposes only backbone coordinates, we restrict the directly evaluated comparison in Table 9 to backbone reconstruction.

More importantly, under direct evaluation, our method achieves lower backbone RMSD and higher TM-score and GDT-TS than ESM3 on both CASP14 and CASP15, while ESM3 achieves higher Cα-lDDT. This contrast indicates that ESM3 better preserves local backbone distances, whereas our method better preserves global structural similarity. All metrics are averaged over individual proteins.

A.3 Unconditional Protein Generation

To further evaluate the generative capability of ADT, we perform unconditional protein structure generation using a multimodal discrete diffusion model over amino-acid and ADT structural tokens, and compare with discrete protein generative models ESM3 and DPLM-2. Following prior work (Ren et al., 2026), we generate 1,000 protein structures and evaluate designability (scRMSD ≤ 2.0Å), diversity, DPT, and novelty. As shown in Table 10, ADT achieves the highest designability and diversity while maintaining competitive DPT and novelty. These results demonstrate that ADT supports diverse and designable protein generation across different discrete generative paradigms.

A.4 Qualitative Results

Figure 5 presents samples autoregressively generated by ADT across different molecular domains. On QM9, ADT generates diverse molecular topologies with multiple rings and complex bonding patterns, including samples with 3 rings and 11 bonds. On proteins, ADT scales to substantially larger structures and generates distinct global folds with nontrivial three-dimensional conformations. These results qualitatively demonstrate the capability of ADT to autoregressively generate complex structured objects across different structural scales and domains.

B Additional Ablation Studies

B.1 Structural Information in Discrete Tokens

To assess whether ADT tokens encode all-atom conformational information, we keep the amino-acid identities unchanged while randomly permuting the structural tokens within each protein, denoted

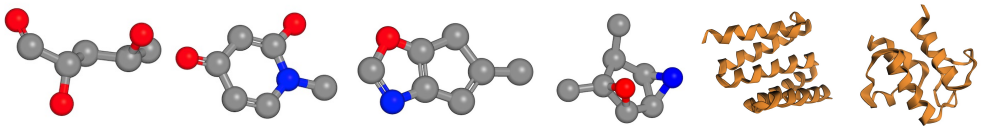

Figure 5: Qualitative results of samples generated by ADT for small molecules from QM9 and proteins.

Table 11: Ablation of structural tokens on protein structure reconstruction.

	Method	Cα lDDT ↑	BB RMSD ↓	AA lDDT ↑	AA RMSD ↓	SC RMSD ↓	TM-score ↑	GDT-TS ↑	GDT-HA ↑
CASP14	Shuffled	0.3100	21.166	0.2978	21.709	22.278	0.2253	0.0373	0.0058
	ADT	**0.8970**	**0.9628**	**0.8223**	**1.3157**	**1.5938**	**0.9695**	**0.9398**	**0.7686**
CASP15	Shuffled	0.3233	22.245	0.3103	22.733	23.274	0.2250	0.0354	0.0057
	ADT	**0.8972**	**1.5191**	**0.8258**	**1.8261**	**2.0779**	**0.9470**	**0.9001**	**0.7285**

as *Shuffled tokens*. As shown in Table 11, shuffling the structural tokens substantially degrades reconstruction across all metrics, indicating that the decoder relies on residue-specific geometric information carried by the structural tokens.

B.2 Fine-grained Geometric Reconstruction on QM9

We further evaluate fine-grained geometric fidelity in Table 12. Diffusion decoding reduces RMSD from 0.0584 to 0.0343 while also improving structure match, PoseBusters validity, and errors in bond lengths, bond angles, and dihedral angles. These results show that generative decoding more effectively reconstructs fine-grained geometry from the discrete representation than direct coordinate prediction.

Table 12: Fine-grained Geometric Reconstruction on QM9.

Decoder	Match ↑	RMSD ↓	PB Valid ↑	Bond-length ↓	Bond-angle ↓	Dihedral ↓
Transformer	98.9	0.0584	0.92	0.02219	1.4971	4.5234
Diffusion	**99.7**	**0.0343**	**0.93**	**0.01557**	**1.0929**	**3.7622**

B.3 Effect of Coarse-to-Fine Decoding

To understand how residue-level representations can be progressively lifted to all-atom geometry, we investigate how decoder layers should be allocated between residue-level modeling and atom-level refinement. Table 13 presents an ablation of our coarse-to-fine decoding design, where Coarse:Fine denotes the number of residue-level and atom-level decoder layers, respectively. The results yield two key observations: 1) Benefit of coarse-to-fine decoding. Coarse-to-fine configurations outperform both residue-only (12:0) and atom-only (0:12), showing that residue-level modeling and atom-level refinement provide complementary benefits for protein reconstruction. 2) Lightweight atomic refinement is sufficient. The 10:2 configuration consistently outperforms 8:4 and achieves the best performance across all reported metrics. This suggests that most structural modeling can be performed at the residue level, with only a lightweight atom-level stage needed for fine-grained refinement.

B.4 Effect of Discrete Capacity

To examine the effect of discrete capacity on residue-level structural tokens, we vary the codebook size while keeping one structural token per residue. As shown in Table 14, reconstruction consistently improves as the codebook grows from 2048 to 16384 across both CASP14 and CASP15. This trend suggests that increasing the representational capacity of each structural token better preserves geometric information through the discrete bottleneck. A codebook of 8192 already provides strong

Table 13: Ablation of coarse-to-fine diffusion decoding on protein structure reconstruction.

	Coarse:Fine	Cα lDDT ↑	BB RMSD ↓	AA lDDT ↑	AA RMSD ↓	TM-score ↑	GDT-TS ↑	GDT-HA ↑
CASP14	12:0	0.8568	1.1196	0.7934	1.5559	0.9528	0.9027	0.7089
	8:4	0.8892	0.9876	0.8161	1.3574	0.9668	0.9324	0.7571
	10:2	**0.8970**	**0.9628**	**0.8223**	**1.3157**	**0.9695**	**0.9398**	**0.7686**
	0:12	0.8624	1.0738	0.8067	1.5126	0.9587	0.9068	0.7115
CASP15	12:0	0.8574	1.7668	0.7967	2.1594	0.9307	0.8630	0.6688
	8:4	0.8895	1.5583	0.8196	1.8840	0.9444	0.8930	0.7170
	10:2	**0.8972**	**1.5191**	**0.8258**	**1.8261**	**0.9470**	**0.9001**	**0.7285**
	0:12	0.8629	1.6945	0.8101	2.0995	0.9365	0.8671	0.6714

Table 14: Effect of discrete codebook size on protein structure reconstruction.

	Codebook	Cα lDDT ↑	BB RMSD ↓	AA lDDT ↑	AA RMSD ↓	TM-score ↑	GDT-TS ↑	GDT-HA ↑
CASP14	2048	0.8634	1.0342	0.7864	1.4734	0.9563	0.9172	0.7243
	4096	0.8721	0.9818	0.8013	1.3869	0.9634	0.9195	0.7431
	8192	0.8970	0.9628	0.8223	1.3157	0.9695	0.9398	0.7686
	16384	**0.9056**	**0.9467**	**0.8315**	**1.2768**	**0.9711**	**0.9584**	**0.7869**
CASP15	2048	0.8628	1.6324	0.7899	2.0413	0.9338	0.8812	0.6853
	4096	0.8706	1.5587	0.8049	1.9284	0.9409	0.8876	0.7041
	8192	0.8972	1.5191	0.8258	1.8261	0.9470	0.9001	0.7285
	16384	**0.9039**	**1.4636**	**0.8388**	**1.7382**	**0.9491**	**0.9173**	**0.7467**

reconstruction performance, although increasing the vocabulary to 16384 yields further improvements. We adopt 8192 as the default to balance reconstruction fidelity with vocabulary size, given the larger prediction space that an expanded vocabulary introduces for autoregressive modeling.

B.5 Ablation of Autoregressive Factorization

Table 15: Effect of autoregressive factorization on QM9 generation.

Strategy	Valid ↑	Unique ↑	Div. ↑	PB Valid ↑
Interleaved AR	0.94	0.93	0.90	0.88
G2C-AR	0.92	0.89	0.91	0.85
C2G-AR	**1.00**	**0.96**	**0.91**	**0.94**

We examine the effect of autoregressive factorization on QM9 by comparing three generation strategies: interleaved autoregression, geometry-to-chemical autoregression (G2C-AR), and our Chemistry-to-Geometry autoregression (C2G-AR). All variants differ only in autoregressive factorization, with all other settings kept fixed. The interleaved AR follows $(s_1, c_1, \ldots, s_L, c_L)$; G2C-AR generates the complete structural sequence before the chemical sequence; and C2G-AR generates the complete chemical sequence before the structural sequence.

Table 15 highlights two findings. First, C2G-AR consistently performs best, improving all metrics over interleaved AR and directly validating our Chemistry-to-Geometry factorization. Second, the gain is not a generic benefit of two-stage generation: reversing the order with G2C-AR reduces validity to 0.92 and PB validity to 0.85, below the interleaved baseline. These results isolate the benefit of generating structural tokens after the complete chemical sequence and directly support the C2G factorization used in our model.

C More Implementation Details

C.1 Tokenizer Model Architecture

Both the small-molecule and protein tokenizers use pre-norm Transformer encoders, binary FSQ, and Diffusion Transformer (DiT) decoders, with 12 layers in both the encoder and decoder. The small-molecule tokenizer uses $d_{\text{model}} = 1536$, while the protein tokenizer uses $d_{\text{model}} = 256$. For proteins, the 12-layer DiT decoder consists of 10 residue-level layers followed by 2 intra-residue atom-level layers. Both tokenizers use 13-bit binary FSQ, corresponding to a codebook size of 8192.

C.2 Training Details

Both the small-molecule and protein tokenizers are optimized using AdamW with a learning rate of 3×10^{-4}. The mutual-information regularizer is applied to both tokenizers with $\lambda_{\text{MI}} = 10^{-3}$. The temperature of mutual information τ is set to 1.0.

Small-molecule tokenizer. The small-molecule tokenizer is trained with the diffusion reconstruction objective and the mutual-information regularizer.

Protein tokenize. For protein tokenizer training, we construct a high-confidence subset of AFDB Swiss-Prot by retaining monomeric proteins with 60–1,024 modeled residues and mean pLDDT above 85, resulting in 330,154 protein structures In addition to the diffusion reconstruction objective and mutual-information regularizer, the protein tokenizer uses auxiliary geometric losses. Specifically, we define

$$\mathcal{L}_{\text{bond}} = \frac{1}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} (\widehat{d}_{ij} - d_{ij})^2, \qquad \mathcal{L}_{\text{angle}} = \frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}} (\cos\widehat{\theta}_{ijk} - \cos\theta_{ijk})^2,$$

$$\mathcal{L}_{\text{clash}} = \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} [\max(0, \delta_{ij} - \widehat{d}_{ij})]^2, \qquad \mathcal{L}_{\text{dist}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} (\widehat{d}^{\,\text{C}_\alpha}_{ij} - d^{\text{C}_\alpha}_{ij})^2.$$

Here, $\mathcal{B}$, $\mathcal{T}$, $\mathcal{N}$, and $\mathcal{P}$ denote bonded atom pairs, bond-angle triplets, non-bonded atom pairs, and C_α pairs, respectively. The weights for the backbone, bond-length, bond-angle, clash and distance are set to 1.0, 0.1, 0.1, 0.1 and 0.01, respectively.

C.3 Autoregressive Generation Details

We train an unconditional autoregressive model on QM9 to evaluate the generative capability of the learned discrete structural representation. Each molecule is represented by a sequence of atom identities $C = (c_1, \ldots, c_L)$ and the corresponding structural tokens $S = (s_1, \ldots, s_L)$ obtained from the pretrained tokenizer. Each atom is associated with one structural token. The tokenizer uses 13-bit binary FSQ, corresponding to a structural vocabulary of $2^{13} = 8192$ discrete codes. We set $\beta = 1.0$.

Small-Molecule Autoregressive Generation. We use a 12-layer Transformer with a hidden dimension of 768, 12 attention heads, and a dropout rate of 0.1, trained from scratch. The autoregressive model is trained on the QM9 training split using AdamW with a batch size of 64 and a learning rate of 10^{-4}. Teacher-forced prediction performance is evaluated on the validation set containing 12,981 molecules.

Unconditional Protein Generation. For unconditional protein generation, we train a multimodal discrete diffusion model over amino-acid and ADT structural tokens, following the generation paradigm of prior discrete protein models to enable a controlled comparison. The model consists of 33 bidirectional Transformer layers with a hidden dimension of 1280 and 20 attention heads. During training, amino-acid and structural tokens are corrupted according to an absorbing discrete diffusion process, and the model learns to recover the clean tokens from their corrupted counterparts. At inference, generation starts from masked amino-acid and structural tokens and iteratively denoises both

modalities to jointly generate a protein sequence and its corresponding structural-token sequence. The generated structural tokens are decoded into all-atom coordinates by the frozen ADT diffusion decoder. The model is trained with a learning rate of 1×10^{-4}.

Sequence-Conditioned All-atom Protein Autoregressive Generation. The training set is constructed from the first 80 shards of AFDB v6 with pLDDT ≥ 90, comprising approximately 32.2M structures and 10.1B residues after preprocessing. We remove training structures with sequence identity greater than 40% to any CAMEO22 target, requiring at least 80% alignment coverage of the evaluation target. Each training example consists of an amino-acid sequence paired with structural tokens obtained from the pretrained ADT tokenizer. We evaluate sequence-conditioned structure generation on 183 CAMEO22 targets (Robin et al., 2021).

We use a Qwen3.5-0.8B model as the backbone with a hidden dimension of 1024. During training, the amino-acid sequence is provided as context and the cross-entropy loss is applied only to the target structural-token sequence. We optimize the model using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and weight decay 0.1. We use a global batch size of 2048, a peak learning rate of 10^{-4}, and cosine learning-rate decay to 10^{-5}, with a 5% warmup period. Gradients are clipped to a maximum norm of 1.0. At inference, given an amino-acid sequence, the model autoregressively generates structural tokens, which are then decoded into all-atom coordinates by the frozen ADT diffusion decoder. We report greedy decoding in Table 5 for direct comparison with DPLM-2.

C.4 Evaluation Metrics

C.4.1 Reconstruction Metrics

RMSD. For a structure with N atoms, ground-truth coordinates $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and reconstructed coordinates $\widehat{\mathbf{X}} \in \mathbb{R}^{N\times 3}$, we report the Kabsch-aligned RMSD:

$$\mathrm{RMSD}(\mathbf{X}, \widehat{\mathbf{X}}) = \min_{\mathbf{R}\in SO(3),\, \mathbf{t}\in\mathbb{R}^3} \frac{1}{\sqrt{N}} \left\| \mathbf{X} - \left(\widehat{\mathbf{X}}\mathbf{R}^\top + \mathbf{1}\mathbf{t}^\top \right) \right\|_F .$$

For proteins, RMSD is computed over the corresponding atom subset, including backbone (BB RMSD), all atoms (AA RMSD), or side-chain atoms (SC RMSD).

Structure Match Rate. For molecular reconstruction on QM9, a reconstruction is considered successfully matched if a valid chemical and atomic correspondence can be established with the reference molecule using a matching tolerance of 0.01 Å. The tolerance is used for structure matching rather than as a cutoff on the reconstruction RMSD. Match rate is the fraction of successfully matched molecules, while RMSD is computed after optimal alignment and averaged over the matched subset.

lDDT. The local Distance Difference Test (lDDT) measures the preservation of local interatomic distances between the reconstructed and reference structures without requiring global alignment. We report Cα-lDDT over Cα atoms, BB-lDDT over backbone atoms, and AA-lDDT over all atoms.

TM-score. TM-score measures the global structural similarity between the reconstructed and reference protein structures, with higher values indicating greater agreement in the overall fold.

GDT. The Global Distance Test measures the fraction of residues that can be superimposed within predefined distance thresholds. We report GDT-TS and GDT-HA, with GDT-HA using stricter distance thresholds to assess higher-accuracy structural agreement.

C.4.2 Generation Metrics

Valid. Fraction of generated molecules that satisfy chemical validity, including valid valence assignments and a consistent molecular graph, as determined by RDKit sanitization.

Unique. Fraction of valid generated molecules that are chemically distinct, measured by unique canonical SMILES.

Novelty. Fraction of valid generated molecules that are not present in the training set, determined by canonical SMILES.

Diversity. Mean pairwise molecular dissimilarity among valid generated molecules, measured using Tanimoto similarity of Morgan fingerprints:

$$\text{Diversity} = 1 - \frac{2}{N(N-1)} \sum_{i<j} \text{Tanimoto}(\mathbf{f}_i, \mathbf{f}_j),$$

where $\mathbf{f}_i$ denotes the Morgan fingerprint (radius 2, 2048 bits) of molecule i.

PB Valid. Fraction of generated molecules that pass the PoseBusters validity checks, providing a more stringent assessment of physicochemical plausibility through chemical consistency, intramolecular clashes, bond lengths, and bond angles.

Self-Consistency. For each generated protein structure, we design eight sequences using ProteinMPNN and predict their structures with ESMFold. We report **scRMSD** as the minimum Kabsch-aligned Cα RMSD and **scTM** as the maximum TM-score between the generated structure and the predicted structures of its designed sequences.

Designability. Fraction of generated protein structures with $\text{scRMSD} \leq 2.0\,\text{Å}$, indicating that at least one designed sequence folds consistently with the generated structure.

Protein Diversity. Number of distinct structural clusters among designable generated structures, obtained by clustering with pairwise TM-score at a cutoff of 0.5. Higher values indicate broader structural coverage.

Protein Novelty. Structural dissimilarity of designable generated structures to known protein structures, defined as

$$\text{Novelty} = \frac{1}{|\mathcal{D}|} \sum_{\widehat{\mathbf{X}} \in \mathcal{D}} \left[1 - \max_{\mathbf{P} \in \text{PDB}} \text{TM}(\widehat{\mathbf{X}}, \mathbf{P}) \right],$$

where $\mathcal{D}$ denotes the set of designable generated structures. Higher values indicate greater structural deviation from the closest known PDB structures.

Designable Pairwise TM-score (DPT). Designable Pairwise TM-score (DPT) measures the mean pairwise TM-score among designable samples (scRMSD ≤ 2.0 $\AA$), where a lower value indicates greater structural diversity among valid generations.

D LIMITATIONS

The focus of this work is to introduce and validate ADT as a compact discrete representation for all-atom molecular structures. To this end, we study two representative molecular modalities, small molecules and proteins, spanning atom-level and residue-level tokenization. A natural extension would be to expand ADT to other biomolecular modalities, such as nucleic acids and molecular complexes, toward a shared discrete representation across diverse molecular systems. Such a representation could enable joint autoregressive modeling of different molecular modalities and potentially benefit from structural knowledge transferred across them. While our experiments cover reconstruction, unconditional small-molecule generation, and sequence-conditioned protein structure generation, another promising direction is to incorporate functional or property conditioning for structure-based molecular design. We leave these directions for future investigation.

E USAGE OF LLMS

Large language models were employed as an auxiliary tool to support manuscript preparation, including grammar checking, sentence refinement, and clarification of technical descriptions. All AI-suggested text was carefully reviewed and revised by the authors to ensure accuracy, clarity, and scientific integrity.