From Molecules to Materials and Proteins: Flow Autoencoders as Lossless and Unified Tokenizers

Yuxuan Ren [1] **Yatao Bian** [1]

[1]*Department of Computer Science, National University of Singapore.* Correspondence to: Yatao Bian ybian@nus.edu.sg.

1. Introduction

Learning compact, faithful representations of atomistic structures remains a key challenge in molecular and materials science. Existing approaches typically rely on domain-specific inductive biases [1, 2, 3]. These designs do not readily generalize across heterogeneous atomistic systems. We introduce UniAE, a flow autoencoder that unifies molecules, crystals and proteins within a single representation learning framework. By shifting reconstruction from the encoder–decoder to a flow-matching-based reconstruction process, UniAE achieves near-zero root mean square deviation (RMSD) reconstructions with accurate atom-type remapping across all domains using simple vanilla Transformers, without explicit graph modeling or domain-specific architectures. Generative models trained directly in the learned latent space achieve state-of-the-art or near state-of-the-art performance, demonstrating UniAE as a general and effective framework for atomistic representation learning and generation.

2. Methods

2.1 Unified Geometric Representation

A fundamental challenge in cross-domain atomistic modeling lies in the heterogeneous representations of three-dimensional scientific data. We represent molecules, proteins, and crystalline materials using a unified graph formulation

$$\mathcal{G} = (\mathbf{X}, \mathbf{A}),$$

where $\mathbf{X} \in \mathbb{R}^{N\times 3}$ denotes atomic coordinates and $\mathbf{A} \in \{1, \dots, K\}^N$ denotes categorical atom types.

For periodic systems, atomic positions are expressed in fractional coordinates. Lattice information is incorporated by introducing three lattice pseudo-atoms with coordinates $\mathbf{X_L} \in \mathbb{R}^{3\times 3}$ and types $\mathbf{A_L}$, yielding the augmented representation

$$\mathcal{G} = ([\mathbf{X_L}, \mathbf{X}], [\mathbf{A_L}, \mathbf{A}]).$$

For proteins, we extract C_α atoms and residue types, producing representations that conform to the same graph formulation. This unified design enables UniAE to operate across diverse atomistic domains using a shared input representation.

2.2 Flow Decoder as a Lossless Tokenizer

Conventional autoencoders directly regress coordinates $\hat{\mathbf{X}} = \text{Decoder}(\mathbf{z})$, which typically leads to "blurry" structures due to the mean-seeking nature of MSE loss. Instead, we formulate reconstruction as a conditional flow matching process.

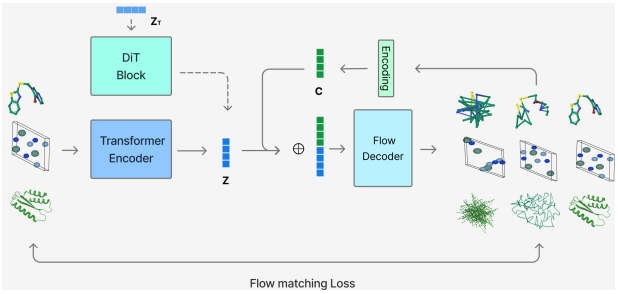

Fig. 1: The illustration of the UniAE framework

Conditional Flow Matching. The decoder ψ is trained to recover ground-truth coordinates $\mathbf{X}_1$ by learning a conditional flow that maps partially noised coordinates back to the data manifold, conditioned on the latent representation $\mathbf{z}$. We adopt Euclidean conditional flow matching (CFM) [4], where noisy coordinates are constructed as

$$\mathbf{X}_t = t \cdot \mathbf{X}_1 + (1-t) \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ t \sim \mathcal{U}(0,1).$$

Given $(\mathbf{X}_t, t, \mathbf{z})$, the decoder predicts the endpoint of the trajectory $\hat{\mathbf{X}}_1^\theta$, and is trained using the metric regression objective

$$\mathcal{L}_{\text{coord}} = \mathbb{E}_{\epsilon,t}\left[\frac{1}{N}\left\|\hat{\mathbf{X}}_1^\theta(\mathbf{X}_t, t, \mathbf{z}) - \mathbf{X}_1\right\|_2^2\right]. \tag{1}$$

At inference time, reconstruction is performed by numerically integrating the learned conditional flow, yielding high-fidelity recovery of atomic coordinates.

Atom Type Remapping via Discrete Flow Matching. To ensure accurate recovery of chemical identity, we jointly model atom types using discrete conditional flow matching based on the Discrete Flow Models (DFM) framework [5]. Ground-truth atom types $\mathbf{A}_1$ are corrupted via categorical interpolation

$$\mathbf{A}_t \sim \text{Cat}\big(t \cdot \delta(\mathbf{A}_1) + (1-t) \cdot \delta(\tfrac{1}{K})\big),$$

where $\delta(\cdot)$ denotes a one-hot encoding and K is the number of atom types. Conditioned on $(\mathbf{A}_t, t, \mathbf{z})$, the decoder predicts the endpoint distribution $\hat{\mathbf{A}}_1$, optimized using the cross-entropy loss

$$\mathcal{L}_{\text{atom}} = \mathbb{E}_t\left[-\sum_i \mathbf{A}_{1,i} \log \hat{\mathbf{A}}_{1,i}^\theta(\mathbf{A}_t, t, \mathbf{z})\right]. \tag{2}$$

Joint Training Objective. The coordinate and atom-type objectives are combined into a unified multi-objective loss

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{coord}} + \lambda_{\text{atom}} \cdot \mathcal{L}_{\text{atom}}, \tag{3}$$

Table 1: **Reconstruction fidelity across domains.** Lower RMSD is better; higher match rate is better.

Domain	Dataset	Metric	Best Baseline	UniAE
Molecules	QM9	RMSD (Å) ↓	0.0002	**0.0002**
	GEOM-Drugs	RMSD (Å) ↓	0.0008	**0.0006**
Materials	MP-20	RMSD (Å) ↓	0.4666	**0.0017**
	MP-20	Match Rate (%) ↑	84.5	**100.0**
Proteins	CASP14/15	RMSD (Å) ↓	0.23	**0.01**

where $\lambda_{\text{atom}} \in (0, 1]$ controls the relative weighting of discrete atom-type supervision. This joint conditional flow-matching formulation enables near-lossless reconstruction of both geometry and chemical identity across all considered atomistic domains.

3. Results and Conclusion

3.1 Experimental Settings

We evaluate UniAE on reconstruction and generative modeling tasks across molecules, crystalline materials, and proteins. All experiments use the same training objective and flow-matching–based reconstruction formulation across domains.

To construct training and evaluation datasets, we collect molecular structures from QM9 [6] and GEOM-Drugs [7], crystalline materials from the MP-20 database [8], and protein structures from the AlphaFold Protein Structure Database (AFDB) [9]. These datasets cover a wide range of chemical systems, from small organic molecules to periodic materials and large biomolecular structures, enabling a comprehensive evaluation of UniAE across heterogeneous atomistic domains.

3.2 Reconstruction Fidelity

We first evaluate reconstruction fidelity, which is a prerequisite for scientific usability. Reconstruction quality is measured using RMSD for continuous coordinates and structure match rate for periodic systems.

As shown in Table 1, across all domains, UniAE achieves near-lossless reconstruction. For molecules, UniAE matches or improves upon specialized molecular autoencoders. For crystalline materials, UniAE attains perfect structure match rate with RMSD below 0.002 Å. For proteins, UniAE substantially outperforms existing protein autoencoders, achieving reconstruction errors an order of magnitude lower than prior work.

3.3 Generative Modeling

We next evaluate generative performance by training latent flow matching models in the UniAE latent space.

Molecules. UniAE achieves high chemical validity and geometric consistency on molecular generation benchmarks.

Table 2: **Molecular generation performance.**

Method	Valid ↑	Novelty ↑	PB Valid ↑
ADiT [10]	0.97	0.99	0.94
FlowMol [11]	0.97	0.99	0.92
UniAE	**0.99**	**1.00**	**0.96**

Materials. On crystal generation for MP-20, UniAE achieves comparable structural stability, uniqueness, and novelty (SUN) as well as meta-stability, uniqueness, and novelty (MSUN), despite using a generic Transformer backbone.

Table 3: **Crystal generation on MP-20.**

Method	SUN ↑	MSUN ↑
FlowMM [12]	4.7	26.3
ADiT	5.3	28.2
UniAE	**5.4**	**30.2**

Proteins. UniAE achieves competitive protein backbone generation performance while discarding triangle attention module.

Table 4: **Protein generation performance.**

Method	Designability ↑	Diversity ↑
LSD [13]	0.69	203
ProteinAE [14]	0.86	228
UniAE	**0.88**	**291**

3.4 Conclusion

We introduced **UniAE**, a unified autoencoder for atomistic representation learning across molecules, crystalline materials, and proteins. By using conditional flow matching for reconstruction, UniAE separates semantic representation learning from structural fidelity, enabling a single architecture and training objective to generalize across atomistic domains and providing a unified alternative to domain-specific modeling approaches.

References

[1] Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. Geometric latent dif-

fusion models for 3d molecule generation, 2023.

[2] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal Diffusion Variational Autoencoder for Periodic Material Generation, March 2022. arXiv:2110.06197 [cs].

[3] Yuyang Zhang, Zinnia Ma, and Haipeng Gong. Topodiff: Improve protein backbone generation with topology-aware latent encoding. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[4] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023.

[5] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

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

[7] Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, April 2022.

[8] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[10] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom Diffusion Transformers:Unified generative modelling of molecules and materials.

[11] Ian Dunn and David Ryan Koes. Mixed Continuous and Categorical Flow Matching for 3D De Novo Molecule Generation, April 2024. arXiv:2404.19739 [q-bio].

[12] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating Materials with Riemannian Flow Matching, June 2024. arXiv:2406.04713 [cs].

[13] Jason Yim, Marouane Jaakik, Ge Liu, Jacob Gershon, Karsten Kreis, David Baker, Regina Barzilay, and Tommi Jaakkola. Hierarchical protein backbone generation with latent and structure diffusion, 2025.

[14] 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, 2025.