Frame-based Equivariant Diffusion Models for 3D Molecular Generation

Mohan Guo[*]
University of Amsterdam
mohan11.guo@gmail.com

Cong Liu[*]
University of Amsterdam
c.liu4@uva.nl

Patrick Forré
University of Amsterdam
p.d.forre@uva.nl

Abstract

Recent methods for molecular generation face a trade-off: they either enforce strict equivariance using costly architectures or relax it to gain scalability and flexibility. We propose a frame-based diffusion paradigm that enforces deterministic $\mathbb{E}(3)$-equivariance while decoupling symmetry handling from the backbone. We study Local Frame-based Diffusion (LFD), which constructs node-specific frames, and Global Frame-based Diffusion (GFD), which assigns a shared molecular frame. To enhance expressivity, we tested EdgeDiT, a Diffusion Transformer with edge-aware attention. On QM9, GFD with EdgeDiT achieves a test NLL of -137.97 ± 0.00, the lowest among evaluated baselines, and obtains decent atom stability (98.89%), molecular stability (89.39%), and validity (96.04%) within models of comparable scale. GFD converges faster than the baselines evaluated in our study and generates 10,000 molecules in 0.33 s, indicating improved sampling efficiency. These results establish frame-based diffusion as a scalable, flexible, and physically grounded paradigm for decent molecular generation.

1 Introduction

Molecular design is central to drug discovery and materials science, but traditional trial-and-error approaches are costly and slow [DiMasi et al., 2016], motivating the need for computational methods. Recent advances in deep learning have enabled accurate molecular property prediction [Luo et al., 2021, Zhou et al., 2023, Fang et al., 2022], but such models remain limited to evaluating existing compounds. Generative models aim to overcome this limitation by designing novel, chemically valid molecules. Among these, diffusion models have shown particular effectiveness in modeling complex distributions. By learning to reverse a gradual noise corruption process, they enable high-quality sample generation [Song and Ermon, 2019, Ho et al., 2020, Hoogeboom et al., 2022]. Diffusion models have achieved remarkable success in domains such as images [Ho et al., 2020, Rombach et al., 2022] and videos [Ho et al., 2022, Blattmann et al., 2023], demonstrating their ability to model high-dimensional, continuous, and structured data. These characteristics align closely with molecular generation, where data consists of 3D atomic coordinates and intricate structural patterns.

A central challenge is to incorporate symmetries into generative models. Molecules are structured as atoms in continuous 3D space connected by bonds, requiring joint modeling of geometry and relational structure. Graph neural networks (GNNs) are well suited for molecular graphs, achieving permutation equivariance [Scarselli et al., 2008]. Extensions such as Equivariant Graph Neural Networks (EGNNs) use pairwise distances as invariant features and update coordinates equivariantly, making them efficient and widely adopted in diffusion-based molecular models [Satorras et al., 2021]. Multi-Channel EGNN (MC EGNN) extends EGNNs by maintaining multiple coordinate channels throughout message passing, parameterizing the number of input, hidden, and output channels to

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Frontiers in Probabilistic Inference: Sampling Meets Learning.

enrich representation power [Levy et al., 2023]. Equivariant Diffusion Model (EDM) employs EGNNs as the backbone models in the diffusion model to respect 3D symmetries [Hoogeboom et al., 2022]. GeoLDM further extends this by operating in latent space [Xu et al., 2023]. Despite their effectiveness, GNN-based frameworks couple equivariance tightly with message passing, limiting architectural flexibility. They inherit known issues, including restricted expressivity due to the 1-Weisfeiler–Lehman bound [Balcilar et al., 2021], and over-squashing of long-range dependencies [Alon and Yahav, 2021]. These factors hinder scalability and efficiency in large systems.

Canonicalization methods achieve equivariance by mapping inputs into canonical representations, often through learned local frames [Luo et al., 2022, Kaba et al., 2023]. They offer universal approximation guarantees and can be integrated with diverse backbones via lightweight modules [Ma et al., 2024], with recent refinements in local frame construction [Lippmann et al., 2025]. However, their application to generative models such as diffusion remains limited.

SymDiff offers a more flexible alternative via stochastic symmetrization [Zhang et al., 2025]. It enforces $\mathbb{E}(3)$-equivariance through random symmetry transformations during the reverse diffusion process, allowing scalable backbones like Diffusion Transformers. Its reliance on stochastic sampling, however, risks biased symmetry coverage and reduced robustness to unseen orientations.

In this work, we introduce a frame-based diffusion paradigm that achieves deterministic $\mathbb{E}(3)$-equivariance while allowing flexible backbone architectures. We explore both global and local frame variants and show that enforcing frame alignment is key to preserving structural consistency. On the QM9 dataset [Wu et al., 2018], our Global Frame Diffusion with EdgeDiT achieves a test NLL of -137.97 ± 0.00, surpassing prior equivariant baselines while halving sampling time. These results establish frame-based designs as a scalable and expressive alternative to EGNN-based molecular diffusion models.

2 Methodology

We achieve deterministic $\mathbb{E}(3)$-equivariance in diffusion-based molecular generation by projecting atomic environments into frames, ensuring that the backbone processes only invariant representations.

Each molecule consists of atomic coordinates $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ and features $\mathbf{h} = (\mathbf{h}_1, ..., \mathbf{h}_N)$ of its N atoms, with $\mathbf{x}_i \in \mathbb{R}^3$ and $\mathbf{h}_i \in \mathbb{R}^{d_h}$. We represent the combination of coordinates and features of every node as $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_N)$, where $\mathbf{m}_i = [\mathbf{x}_i; \mathbf{h}_i]$ for atom i. Since $\mathbf{h}$ is invariant to translation and rotation, only coordinate equivariance needs to be addressed.

A transformation $g \in \mathbb{E}(3)$ is defined by $(\mathbf{R}, \mathbf{t})$ with $\mathbf{R} \in O(3)$ and $\mathbf{t} \in \mathbb{R}^3$. Under g, atomic coordinates transform as $(\mathbf{x}, \mathbf{h}) \to (\mathbf{R}\mathbf{x} + \mathbf{t}, \mathbf{h})$ In group theory, a group $\mathbb{G}$ acts on a space $\mathcal{X}$ through group actions $\mathcal{T}_g : \mathcal{X} \to \mathcal{X}$ for each $g \in \mathbb{G}$. A function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathbb{G}$-equivariant if there exist corresponding group actions $\mathcal{S}_g : \mathcal{Y} \to \mathcal{Y}$ satisfying:

$$f(\mathcal{T}_g(x)) = \mathcal{S}_g(f(x)) \quad \forall g \in \mathbb{G}, \forall x \in \mathcal{X} \tag{1}$$

In this work, we considered equivariance over $\mathbb{E}(3)$ and permutation group. For translation equivariance, we follow EDM by centering molecules at the zero-mass position, preserved under zero-centered noise, ensuring translation invariance [Hoogeboom et al., 2022, Xu et al., 2022]. For permutation equivariance, we remove positional encodings in DiTs, and since MC-EGNN is already permutation-equivariant, this property is naturally satisfied.

Our primary architecture is the Global Frame-based Diffusion Model (GFD), which constructs a single molecular frame to project coordinates into invariant representations before diffusion modeling and restores them via inverse projection. This guarantees exact $\mathbb{E}(3)$-equivariance while preserving global geometric relationships, enabling the backbone to learn consistent geometric–chemical patterns.

In addition to GFD, we study Local Frame Diffusion (LFD), where each atom has its own local frame. Without constraints LFD underperforms due to disrupted global consistency, but adding a frame alignment loss markedly improves results, supporting our hypothesis that preserving Euclidean structure is key (see Appendix A.1).

For the backbone, we employ Diffusion Transformers (DiT) due to their scalability and expressive capacity. To adapt them to molecular geometry, we apply EdgeDiT, a lightweight modification that incorporates interatomic distance and directional cues into the attention mechanism. This

Algorithm 1 Frame-based Equivariant Projection (single diffusion step)

Require: Noisy atomic coordinates $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^N$, features $\mathbf{h} = \{\mathbf{h}_i\}_{i=1}^N$, equivariant frame constructor ϕ_e, backbone ϕ_θ

Ensure: Equivariant outputs $\mathbf{y}^{\text{out}} = \{\mathbf{y}_i^{\text{out}}\}_{i=1}^N$

Construct orthogonal frame $\phi_e(\{\mathbf{x}_i, \mathbf{h}_i\}_{i=1}^N) \to \{\mathbf{O}_i\}_{i=1}^N = \{[\mathbf{u}_{i1}, \mathbf{u}_{i2}, \mathbf{u}_{i3}]\}_{i=1}^N$
for each atom $i = 1, \ldots, N$ **do**
    Equivariance: under rotation $\mathbf{R}$*, we have* $\mathbf{O}_i' = \mathbf{R}\mathbf{O}_i$
    Project coordinates into frame: $\mathbf{x}_i^{\text{in}} = \mathbf{O}_i^{-1}\mathbf{x}_i$
end for
Backbone prediction on invariant inputs: $\mathbf{y} = \phi_\theta(\{\mathbf{x}_i^{\text{in}}, \mathbf{h}_i\}_{i=1}^N)$
for each atom $i = 1, \ldots, N$ **do**
    Map back to original coordinates: $\mathbf{y}_i^{\text{out}} = \mathbf{O}_i\,\mathbf{y}_i$
end for
return $\mathbf{y}^{\text{out}}$

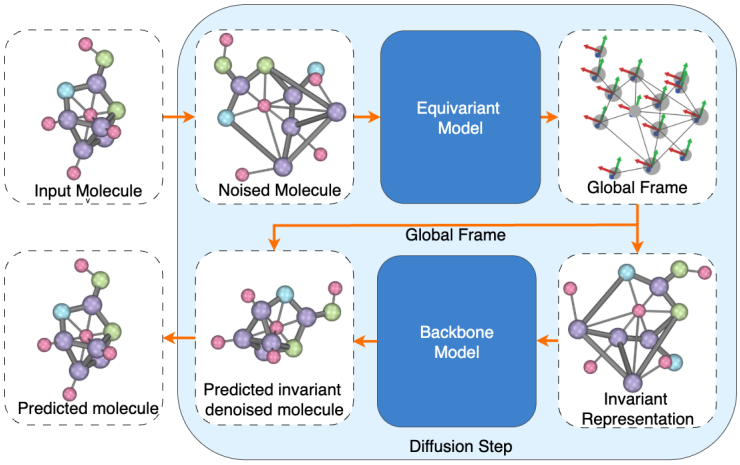

Figure 1: Global Frame-based Diffusion (GFD): an equivariant module constructs a global molecular frame from noised inputs, invariant features are derived and denoised by the backbone, and the final molecule is reconstructed via inverse frame transformation with loss to the original.

simple enhancement allows the model to capture chemical bonding patterns more effectively while maintaining the architectural flexibility of Transformer backbones.

3 Experiments

We evaluate our proposed frame-based diffusion models on the QM9 dataset, which contains 134k stable small molecules with annotated geometric and chemical properties. Performance is assessed using negative log-likelihood (NLL), atom and molecular stability, validity, and uniqueness.

As shown in Table 1, our GFD model with EdgeDiT achieves decent performance on the QM9 dataset, with test negative log-likelihood of -137.97, substantially outperforming all existing methods. Critically, these improvements are achieved using the same computational scale as the standard SymDiff baseline, indicating that the performance gains stem from methodological innovations rather than increased model capacity. This superior test NLL performance indicates that our architecture enables more effective learning, as NLL directly reflects the model's ability to capture the underlying data distribution.

Regarding molecular quality metrics, our model demonstrates superior performance compared to SymDiff at equivalent model scales: atom stability of 98.89% versus 98.74%, molecular stability of 89.39% versus 87.49%, and validity of 96.04% versus 95.75%, while maintaining comparable performance on uniqueness with 97.62% versus 97.89%. Compared to EDM, the improvements are even more substantial, with molecular stability increasing from 82.00% to 89.39% and validity

from 91.90% to 96.04%. Meanwhile, the consistently low variance across all evaluation metrics with standard deviation below 0.03 further demonstrates the reliability and stability of our deterministic equivariance approach.

Table 1: Test NLL, atom stability, molecular stability, validity and uniqueness on QM9 for 10,000 samples and 3 evaluation runs. * indicates models with double backbone scale. We omit the results for NLL where not available. Bold values indicate the best performance within each model scale/type.

Method	NLL ↓	Atom Stab. (%) ↑	Mol. Stab. (%) ↑	Val. (%) ↑	Uniq. (%) ↑
Previous Methods					
GeoLDM	–	98.90 ± 0.10	89.40 ± 0.50	93.80 ± 0.40	92.70 ± 0.50
MUDiff	$\mathbf{-135.50 \pm 2.10}$	98.80 ± 0.20	$\mathbf{89.90 \pm 1.10}$	$\mathbf{95.30 \pm 1.50}$	$\mathbf{99.10 \pm 0.50}$
END	–	$\mathbf{98.90 \pm 0.00}$	89.10 ± 0.10	94.80 ± 0.10	92.60 ± 0.20
EDM	-110.70 ± 1.50	98.70 ± 0.10	82.00 ± 0.40	91.90 ± 0.50	90.70 ± 0.60
Standard Scale Models					
SymDiff	-129.35 ± 1.07	98.74 ± 0.03	87.49 ± 0.23	95.75 ± 0.10	97.89 ± 0.26
SymDiff-H	-126.53 ± 0.90	98.57 ± 0.07	85.51 ± 0.18	95.22 ± 0.18	97.98 ± 0.09
DiT-Aug	-126.81 ± 1.69	98.64 ± 0.03	85.85 ± 0.24	95.10 ± 0.17	$\mathbf{97.98 \pm 0.08}$
DiT	-127.78 ± 2.49	98.23 ± 0.04	81.03 ± 0.25	94.71 ± 0.31	97.98 ± 0.12
GFD (Ours)	$\mathbf{-137.97 \pm 0.00}$	$\mathbf{98.89 \pm 0.01}$	$\mathbf{89.39 \pm 0.02}$	$\mathbf{96.04 \pm 0.03}$	97.62 ± 0.01
Large Scale Models					
SymDiff* (2×)	-133.79 ± 1.33	98.92 ± 0.03	89.65 ± 0.10	$\mathbf{96.36 \pm 0.27}$	$\mathbf{97.66 \pm 0.22}$
GFD* (Ours) (2×)	$\mathbf{-141.85 \pm 0.00}$	$\mathbf{98.98 \pm 0.00}$	$\mathbf{90.51 \pm 0.00}$	96.34 ± 0.01	97.61 ± 0.00
Data		99.00	95.20	97.8	100

We also compare the sampling time of our model with EDM and SymDiff baselines, as shown in Table 2. While GFD and SymDiff share the same model scale, GFD uses a deterministic equivariant module instead of SymDiff's stochastic symmetry approach. Although SymDiff achieves faster sampling speed, GFD still significantly reduces sampling time by about 50% compared to EDM, while delivering superior generation quality.

Table 2: Sampling time comparison for generating 10,000 molecules.

Method / Seconds per Sample	Mean	Std
EDM	0.59694	0.00022
SymDiff	0.20731	0.00063
GFD (Ours)	0.32585	0.00119

4 Discussion and Conclusion

Our frame-based diffusion approach consistently improves performance across evaluation criteria. On QM9, GFD with EdgeDiT achieves a test NLL of -137.97, surpassing all baselines, while also attaining higher validity, uniqueness, novelty, and stability than same-scale models. It converges faster than EDM and SymDiff, with nearly $2\times$ faster sampling. Additional experiments (Appendix A.1) confirm our hypothesis: global frames preserve molecular structure, while local frames hinder optimization unless aligned, in which case performance approaches GFD. These results demonstrate that frame-based designs combine deterministic equivariance with tangible gains in fidelity, stability, and efficiency.

While GFD achieves decent performance and efficiency, it introduces canonicalization overhead and relies on MC EGNN for frame construction, which may constrain scalability on very large datasets or extremely deep architectures. Moreover, our evaluation is currently limited to QM9. Future work includes extending experiments to larger and more diverse benchmarks and exploring more scalable frame construction modules or hybrid strategies to reduce canonicalization costs. In summary, frame-based diffusion decouples deterministic equivariance from the backbone design, enabling the use of modern high-capacity architectures. Combined with EdgeDiT, our GFD achieves decent results on QM9 and highlights the importance of preserving global structure in molecular modelling.

References

References

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Muhammet Balcilar, Pierre Héroux, Benoit Gauzere, Pascal Vasseur, Sébastien Adam, and Paul Honeine. Breaking the limits of message passing graph neural networks. In *International Conference on Machine Learning*, pages 599–608. PMLR, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, 2023.

Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pages 15546–15566. PMLR, 2023.

Daniel Levy, Sékou-Oumar Kaba, Carmelo Gonzales, Santiago Miret, and Siamak Ravanbakhsh. Using multiple vector channels improves e (n)-equivariant graph neural networks. In *ICML Workshop on Machine Learning for Astrophysics*, 2023. URL https://ml4astro.github.io/icml2023/assets/68.pdf.

Peter Lippmann, Gerrit Gerhartz, Roman Remme, and Fred A Hamprecht. Beyond canonicalization: How tensorial messages improve equivariant message passing. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. Predicting molecular conformation via dynamic graph score matching. *Advances in Neural Information Processing Systems*, 34:19784–19795, 2021.

Shitong Luo, Jiahan Li, Jiaqi Guan, Yufeng Su, Chaoran Cheng, Jian Peng, and Jianzhu Ma. Equivariant point cloud analysis via learning orientations for message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18932–18941, June 2022.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=jjcY92FX4R.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332. PMLR, 2021.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=PzcvxEMzvQC`.

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

Leo Zhang, Kianoosh Ashouritaklimi, Yee Whye Teh, and Rob Cornish. Symdiff: Equivariant diffusion via stochastic symmetrisation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=i1NNCrRxdM`.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6K2RM6wVqKu`.

A Technical Appendices and Supplementary Material

A.1 Local Frame-based Diffusion Model

A.2 Methodology

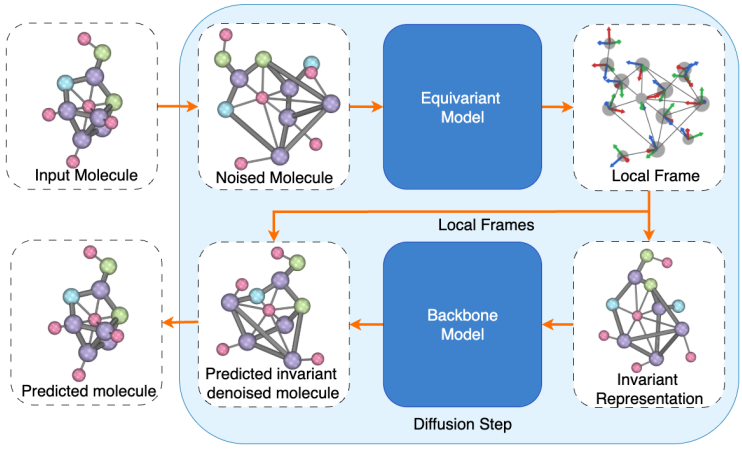

Figure 2: Local Frame-based Diffusion Model (LFD). The framework operates through the following process: (1) The input molecule is noised and then is processed by an equivariant model to construct local frames for each atom. (2) These local frames are used to derive invariant representations that capture local molecular geometry while being invariant to $\mathbb{E}(3)$ group. (3) The backbone diffusion model takes the invariant representations as inputs to predict the invariant denoised molecule. (4) Predicted molecule is obtained by applying the local frames inversely. Loss is computed between input molecule and predicted molecule.

As illustrated in Figure 2, LFD relies on an equivariant module to construct local frames for each atom. By projecting noisy coordinates into local coordinate systems, the model obtains invariant representations with respect to $\mathbb{E}(3)$ and permutation transformations. These invariant features are processed by the backbone diffusion model, and predictions are mapped back into global space via the local frames, enabling equivariance.

The training procedure is summarized in Algorithm 2, where the backbone learns to denoise projected coordinates. Sampling follows the reverse process (Algorithm 3).

Algorithm 2 Local Frame-based Diffusion (LFD) - Training

Require: Molecule $\mathbf{m}$

1: Sample $t \sim \text{Uniform}(1, T)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: $\mathbf{z}_t = \alpha_t \mathbf{m} + \sigma_t \boldsymbol{\epsilon}$
3: $\{\mathbf{O}_i\}_{i=1}^N = \phi_e(\mathbf{z}_t)$ ▷ Equivariant frames
4: Project: $\mathbf{z}_t^{\text{local}} = \{\mathbf{O}_i^{-1} \mathbf{z}_{t,i}\}$
5: Predict: $\hat{\boldsymbol{\epsilon}}^{\text{local}} = \phi_\theta(\mathbf{z}_t^{\text{local}}, t)$
6: Invert: $\hat{\boldsymbol{\epsilon}}_i = \mathbf{O}_i \hat{\boldsymbol{\epsilon}}_i^{\text{local}}$
7: Minimize $\mathcal{L} = \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|^2$

Algorithm 3 Local Frame-based Diffusion (LFD) - Sampling

1: Sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, T-1, ..., 1$ **do**
3: $\quad \{\mathbf{O}_i\} = \phi_e(\mathbf{z}_t)$
4: $\quad$ Project $\mathbf{z}_t$ into local frames, predict noise, invert
5: $\quad$ Update $\mathbf{z}_{t-1}$ via reverse diffusion step
6: **end for**
7: **return** $\mathbf{z}_0$

One central hypothesis of this work is that the superior performance of GFD compared to LFD arises from its preservation of inter-atomic relations in the Euclidean space, which maintains the natural mapping between molecular geometry and physico-chemical properties. To examine this hypothesis, we extend LFD with an additional frame alignment constraint that encourages local frames to remain consistent with a global molecular frame. Concretely, given rotation matrices $\mathbf{O}_i, \mathbf{O}_g \in SO(3)$, their relative rotation is

$$\mathbf{R}_{i,g} = \mathbf{O}_i^\top \mathbf{O}_g.$$

The angle $\theta_{i,g}$ corresponding to $\mathbf{R}_{i,g}$ can be obtained from

$$\cos\theta_{i,g} = \tfrac{1}{2}\big(\text{tr}(\mathbf{R}_{i,g}) - 1\big), \quad \sin\theta_{i,g} = \tfrac{1}{2}\|\mathbf{R}_{i,g} - \mathbf{R}_{i,g}^\top\|_F.$$

We then compute

$$\theta_{i,g} = \arctan 2(\sin\theta_{i,g}, \cos\theta_{i,g}),$$

which gives the geodesic distance on $SO(3)$ between local and global frames. The alignment loss is defined as

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\theta_{i,g}}{\pi}.$$

This term is added to the diffusion training loss to encourage consistency between local and global orientations with a weight:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{align}},$$

where $\mathcal{L}_{\text{diff}}$ is the standard diffusion loss and λ is a balancing weight.

A.3 Experimental Results

On the QM9 dataset, we evaluate our proposed frame-based diffusion models against EDM and SymDiff. As shown in Figure 3, GFD combined with EdgeDiT achieves consistently superior performance in terms of molecular stability, atom stability, and convergence speed. Vanilla LFD, in

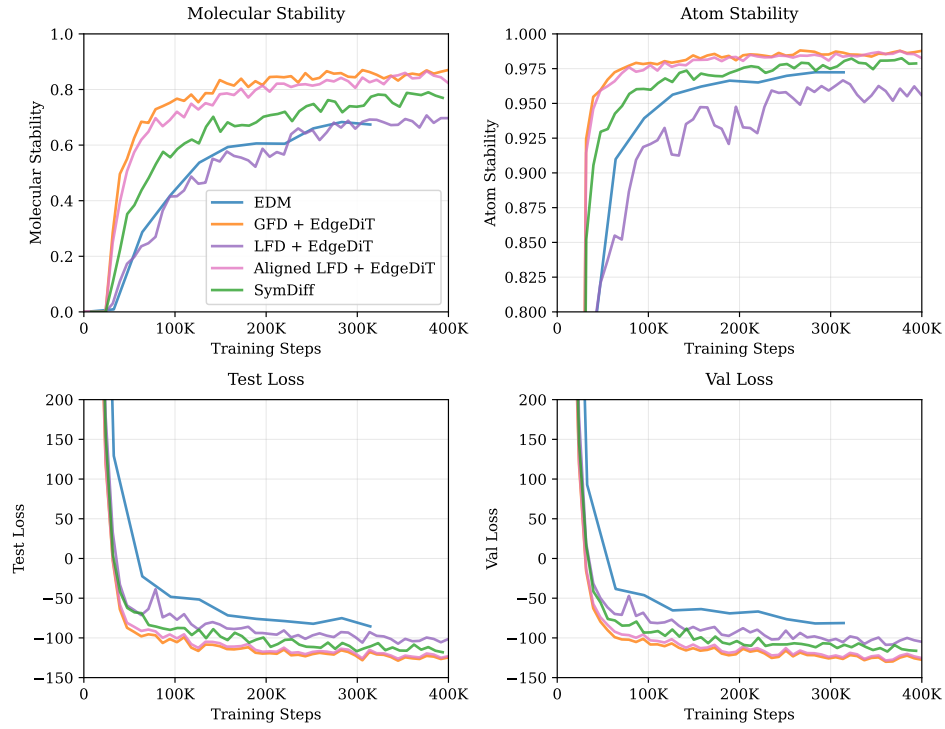

Figure 3: Training curves on QM9 comparing EDM, GFD + EdgeDiT, LFD + EdgeDiT, Aligned LFD + EdgeDiT, and SymDiff. Aligned LFD refers to LFD augmented with the proposed frame alignment loss. GFD and Aligned LFD achieve the best overall results, converging faster and attaining higher stability and lower validation loss compared to baselines. These results demonstrate that vanilla LFD substantially underperforms GFD, but incorporating frame alignment recovers performance to the level of GFD, thereby validating our hypothesis that preserving global Euclidean structure is essential for effective molecular generation.

contrast, significantly underperforms both in stability and likelihood, supporting our hypothesis that preserving the global Euclidean structure is critical for capturing meaningful molecular geometry. To further test this hypothesis, we introduced a frame alignment loss that constrains local frames to remain consistent with the global orientation. The resulting Aligned LFD model substantially improves over vanilla LFD, matching the performance of GFD in all metrics and convergence behavior. These results confirm that retaining global structural consistency is key to effective molecular generation, and that frame alignment serves as a principled mechanism to bridge the gap between local and global representations.

B Implementation details

Our models are configured to align closely with baseline settings for fair comparison. Specifically, MC EGNN consists of 3 layers, 7 internal channels, and a hidden size of 256. MPNN uses 9 layers, and 256-dimensional hidden features. Transformer and all its variants have 9 layers, 8 heads, and hidden size of 256. DiT and all its variants have 12 layers, 6 heads, and hidden size of 384. Experiments with MPNN and Transformer backbones are trained with learning rate 0.0001 and batch size 64. Experiments with DiT backbones are trained with learning rate 0.0002 and batch size 256. All experiments are conducted with AdamW optimizer on NVIDIA A100 GPUs with 40GB memory.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions: introducing a frame-based diffusion paradigm with two variants (GFD and LFD), demonstrating that GFD with EdgeDiT achieves decent likelihood and stability on QM9, and showing that frame alignment significantly improves LFD performance. These claims are supported by the experimental results reported in Section 3 and Appendix A.1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in the Discussion section: GFD introduces canonicalization overhead, relies on MC-EGNN for frame construction, which may limit scalability, and current evaluation is restricted to QM9. We also note that experiments are being extended to larger datasets.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work does not introduce new theoretical results or formal proofs. The focus is on designing a frame-based equivariant diffusion paradigm and empirically validating its effectiveness.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details of datasets (QM9), model architectures, training hyperparameters, evaluation protocols, and canonicalization procedure in the main text and Appendix. We also include ablation studies (Appendix A.1) and report metrics following standard practice, enabling faithful reproduction of our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While the QM9 dataset is publicly available, we do not release code at this stage because the project is still under active extension to additional datasets. However, we provide detailed experimental settings, model descriptions, and hyperparameters to ensure reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify training and evaluation details, including dataset splits (following EDM), model architectures, hyperparameters, and optimization settings, in Appendix to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We follow the standard protocol used by prior baselines: each model is evaluated with 10,000 generated samples across 3 independent runs, and we report the mean and standard deviation. This captures variability due to sampling and random initialization, and ensures fair comparison with existing work

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the type of compute resources used (NVIDIA A100 GPUs with 40GB memory) in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work focuses on developing molecular generative models using publicly available datasets (QM9). It does not involve sensitive data, human subjects, or dual-use applications that would raise ethical concerns. We adhered to the NeurIPS Code of Ethics throughout the research process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper proposes a methodological contribution in molecular generative modeling and is evaluated solely on the objective and widely used QM9 dataset. As such, the work itself does not have direct societal impact. Any potential positive or negative impacts would depend on downstream applications in drug or materials discovery, which are beyond the scope of this study.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work uses only the publicly available QM9 dataset and does not release any new models or datasets with foreseeable risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the original QM9 dataset and all baseline methods used in our experiments. While licenses are not explicitly listed in the main text, all assets are used under their respective academic terms of use and properly credited in the references.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This work does not release new assets (datasets, models, or benchmarks). The contribution is methodological, and all experiments are conducted on existing datasets (QM9).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This work does not involve crowdsourcing or human subjects. All experiments are conducted on publicly available molecular datasets (QM9).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve human subjects or crowdsourcing, and therefore IRB approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Large language models were not used as part of the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.