Lyrebird: Toward Robust and Generalizable 3D Molecular Conformer Generation via Equivariant Flows

Vedanth M. Nilabh*
Rowan Scientific Corporation
nilabh.v@northeastern.edu

Elias L. Mann
Rowan Scientific Corporation
eli@rowansci.com

Abstract

Recent generative models for 3D molecular conformer generation have made impressive progress, but data and benchmarks are limited and often fail to evaluate usefulness and trustworthiness as computational chemistry tools. We introduce Lyrebird, a general purpose model for 3D molecular conformer generation built on the ET-Flow (Equivariant Transformer Flow) architecture, and evaluate generalization by training jointly on Butina split datasets of drug like molecules from GEOM-Drugs and GEOM-QM9 and macrocyclic peptides from CREMP. Additionally, we introduce a macrocyclic conformer generation benchmark set: MPCONF196GEN, derived from the MPCONF196 energy benchmark set. We also introduce an energy based benchmark that evaluates both conformer sampling within the lowest energy basin and the degree of structural relaxation in generated conformers. Lyrebird matches state of the art ML methods and outperforms ETKDGv3 on coverage and matching metrics for drug like molecules, and improves performance on macrocycles over models only trained on GEOM-Drugs.

1 Introduction

A fundamental task in computational chemistry and drug discovery is conformational sampling: the generation and search of the low energy minima of the potential energy surface of 3D structures of a given molecule conditioned on the SMILES string. Traditional approaches involve molecular dynamics based simulations such as CREST[1], which generate accurate conformers but are compute intensive, while rule based approaches like ETKDG[2] are fast but less accurate for larger molecules. Recent generative models balance speed and accuracy: Torsional Diffusion[3] performs diffusion on torsional angles atop RDKit initialized structures; MCF[4] learns conformer fields with a PerceiverIO transformer; ET-Flow[5] uses equivariant flow matching based on TorchMD-Net[6], matching MCF at a fraction of the model size (8.3M vs 242M parameters); and RINGER[7] targets macrocyclic peptides using internal coordinate transformations. However, two issues limit the reliability of reported results. First, the prevailing random split for GEOM-Drugs and GEOM-QM9[8] can place structurally near identical molecules into train and test sets, testing interpolation rather than generalization, analogous to the issues of overestimation bias for scaffold splits identified for virtual screening[9]. Second, existing models are trained on single datasets, leaving open whether performance transfers across molecular classes where distributional shift is significant. We make two central claims: (1) how you split conformer generation data matters significantly for realistic performance estimation, and (2) training jointly on chemically diverse datasets

*Corresponding author, Work done as an intern at Rowan Scientific Corporation.

can improve coverage on underrepresented molecule classes without destabilizing learning. Lyrebird validates these claims but is not intended as a new state of the art architecture, it uses ET-Flow without modification. Our results show meaningful gains on macrocycles from multi dataset training, but also reveal that current data scales remain far from sufficient for a truly general purpose conformer search model. Our contributions:

1. Lyrebird: A single ET-Flow-based model trained jointly on GEOM-Drugs, GEOM-QM9, and CREMP[10], demonstrating that multi dataset training can improve coverage without destabilizing learning.
2. Butina clustered splits: New splits for all three datasets based on Butina clustering on Morgan fingerprints with Tanimoto similarity, providing more rigorous generalization tests than random or Murcko scaffold splits.
3. MPCONF196GEN: A macrocycle benchmark with CREST generated reference ensembles for 13 challenging molecules.
4. Energy benchmark: A physically grounded benchmark measuring whether generated conformers pre and post relaxation are near physically reasonable minima under GFN2-xTB[11] optimization. Concurrent work by Reidenbach et al.[12] independently introduced energy based evaluation for *de novo* molecule generation; our benchmark differs in targeting *conditional* conformer generation with basin sampling relative to ground truth ensembles (see Section A.4.3).

Code, trained model weights, data splits, and benchmarks are publicly available under the MIT License.[12]

2 Methods

2.1 Data Preparation

Most ML conformer generation methods use a standardized random split, which is deceptively optimistic. If two structurally similar molecules are separated into train and test, the model can perform well without truly generalizing[13]. Murcko scaffolds offer a partial remedy, but studies have found they still lead to overoptimistic estimates[9]. We instead use Butina clustering, a greedy algorithm on Morgan fingerprints with Tanimoto similarity that groups molecules by chemical similarity in fingerprint space, ensuring test molecules are structurally distinct from training data. We assign entire clusters to train/validation/test partitions (80/10/10). Computing pairwise similarity matrices at dataset scale requires substantial memory (multiple TB of RAM for GEOM-Drugs), so we plan to release precomputed splits publicly. We generate Butina clusters for GEOM-QM9 (133k small drug like molecules), GEOM-Drugs (304k medium sized drug like molecules), and CREMP[10] (36k macrocyclic peptides with 4, 5, and 6 monomers). Training sets are combined; test sets are evaluated separately.

2.2 Model

We adopt ET-Flow's architecture[5] (8.3M parameters): an O(3) equivariant flow matching model based on TorchMD-Net's Equivariant Transformer[6], with post hoc chirality correction comparing oriented volumes against RDKit stereochemistry tags. Atom features derive strictly from SMILES via MolFromSmiles to avoid test time leakage. Edges combine molecular bonds with spatial neighbors within 10 Å, encoded via radial basis functions with smooth cutoff. See Section A.2 for full details. We exclude MCF[4] and RINGER[7] from comparisons because their featurization pipelines pass ground truth RDKit Mol objects at test time, potentially leaking 3D structural information into the input representation. Featurization pipelines for conformer generation should derive all features from the SMILES string or molecular graph alone. Following ET-Flow, we use linear interpolation between a

[1]https://github.com/rowansci/lyrebird
[2]https://github.com/rowansci/MPCONF196GEN-benchmark

harmonic prior (see appendix) and target conformers, with Kabsch alignment to optimize transport cost. We trained for 750 epochs with AdamW, using cosine annealing from 5e-4 to 1e-8 over the first 500 epochs and holding constant thereafter. For all training sets, we use the top 30 Boltzmann weighted conformers per molecule, seeing each molecule once per epoch with a uniformly sampled conformer. For test sets, we keep the entire CREST generated ensemble, except for CREMP where we keep the top 100 conformers due to the large size of some ensembles (>1000). We observe that including all three datasets benefits training stability. Training on GEOM-Drugs and CREMP without GEOM-QM9 was less stable in preliminary experiments, suggesting a curriculum like effect where QM9 provides easy examples yielding stable gradients, GEOM-Drugs contributes medium difficulty examples, and CREMP introduces the most challenging structures. We emphasize that this observation is empirical rather than a formal theoretical claim, investigating the mechanisms behind multi dataset stabilization is an interesting direction for future work.

3 Experiments and Results

3.1 Experimental Setup

Following standard practice[3–5], we evaluate using RMSD based coverage (COV) and average minimum RMSD (AMR) metrics. For each test molecule with K ground truth conformers, we generate $2K$ candidates. Coverage Recall measures the fraction of reference conformers matched within threshold δ by at least one generated conformer; Coverage Precision does the reverse. AMR averages the minimum RMSD between each conformer and its nearest match. Lower AMR indicates better structural agreement. See[14] for formal definitions. We compare against ETKDG V3[2] (with macrocycle torsional preferences), Torsional Diffusion trained on Drugs, and ET-Flow trained on Drugs. For the Butina split Drugs benchmark, we can only compare with ETKDG because models trained on the old random split cannot be directly compared, and retraining is expensive. Training was done on a node of 8xA100-80GB GPU on RunPod, and benchmarking was done on A100 instances on Runpod and Modal.

3.2 Standard Drug Like Benchmarks

Table 1: GEOM-QM9 Butina-split test set ($\delta = 0.5$ Å). Best results bold.

	Recall				**Precision**			
Method	**Cov. (%) ↑**		**AMR (Å) ↓**		**Cov. (%) ↑**		**AMR (Å) ↓**	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Torsional Diffusion	86.91	**100**	0.195	0.163	82.64	**100**	0.244	0.218
ET-Flow	87.02	**100**	0.208	0.143	71.75	87.50	0.334	0.283
RDKit ETKDG	87.99	**100**	0.228	0.175	**90.82**	**100**	0.224	0.181
Lyrebird (Ours)	**92.99**	**100**	**0.101**	**0.028**	86.99	**100**	**0.161**	**0.047**

Table 2: GEOM-Drugs Butina split test set ($\delta = 0.75$ Å). Best results bold.

	Recall				**Precision**			
Method	**Cov. (%) ↑**		**AMR (Å) ↓**		**Cov. (%) ↑**		**AMR (Å) ↓**	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
RDKit ETKDG	50.4	50.0	0.901	0.802	49.7	50.0	0.871	0.753
Lyrebird (Ours)	**67.6**	**74.2**	**0.594**	**0.563**	**51.8**	50.0	**0.783**	**0.750**

On QM9 (Table 1), Lyrebird achieves the best recall coverage and AMR, with ETKDG

showing stronger mean precision coverage. Comparing Lyrebird to models trained on GEOM-Drugs is informative for deployment but is not a fair architectural comparison, since Lyrebird uses the same ET-Flow architecture and the primary difference is training data.

On GEOM-Drugs (Table 2), Lyrebird substantially outperforms ETKDG. Comparing our Butina splits to previously reported random split results (see[5]) reveals a meaningful gap, consistent with random splits overestimating generalization.

3.3 Macrocycles

Table 3: CREMP test set (macrocyclic peptides). ET-Flow precision capped at 6 Å due to numerical issues. Best results bold.

	Recall AMR (Å) ↓		**Precision AMR (Å) ↓**	
Method	Mean	Med.	Mean	Med.
RDKit ETKDG	4.69	4.68	4.73	4.71
ET-Flow	4.13	4.07	>6	>6
Lyrebird (Ours)	**2.34**	**2.33**	**2.82**	**2.81**

Table 4: MPCONF196GEN benchmark (13 challenging macrocycles with CREST reference ensembles). Best results bold.

	Recall AMR (Å) ↓		**Precision AMR (Å) ↓**	
Method	Mean	Med.	Mean	Med.
RDKit ETKDG	3.79	3.71	4.01	3.91
Torsional Diffusion	2.71	**2.58**	3.13	**2.95**
ET-Flow	2.60	3.33	2.83	3.59
Lyrebird (Ours)	**2.54**	2.96	**2.80**	3.56

On CREMP (Table 3), Lyrebird substantially outperforms ETKDG and ET-Flow. Torsional Diffusion could not produce valid structures on CREMP, likely because macrocyclic ring constraints violate its assumption of torsional angle independence[3]. On MPCONF196GEN (Table 4), Lyrebird achieves the best mean AMR while Torsional Diffusion shows better median performance. All AMR values exceed 2.5 Å, indicating that MPCONF196GEN presents a meaningful challenge for future models.

3.4 Large Drug Like Molecules and Energy

Table 5: GEOM-XL test set (molecules with >100 atoms). Best results bold.

	Recall AMR (Å) ↓		**Precision AMR (Å) ↓**	
Method	Mean	Med.	Mean	Med.
RDKit ETKDG	2.92	2.62	3.35	3.15
Torsional Diffusion	**2.05**	**1.86**	**2.94**	**2.78**
ET-Flow	2.31	1.93	3.31	2.84
Lyrebird (Ours)	2.42	2.07	3.27	2.87

Table 6: Energy benchmark on 20 GEOM-XL molecules. Left: average GFN2-xTB energy change after optimization; values are negative because optimization lowers energy, and *closer to zero* indicates conformers already near a local minimum (ETKDG is best). Right: energy difference between each method's lowest energy generated conformer and the ground truth minimum after optimization; negative means the method found a lower energy basin (Torsional Diffusion is best). See Section A.4.3 for comparison with related benchmarks.

	Energy Change (kcal/mol) ↓		**Δ to GT Min. ↓**
Method	Per-mol	Per-conf	Per-mol
RDKit ETKDG	**−0.281**	**−0.229**	0.016
Torsional Diffusion	−3.992	−2.874	**−0.423**
ET-Flow	−7.966	−5.379	−0.145
Lyrebird (Ours)	−8.134	−5.337	−0.116

On GEOM-XL (Table 5), Torsional Diffusion achieves the best results. Lyrebird is competitive but does not surpass ET-Flow, indicating that additional training data alone is insufficient for large, flexible molecules, architectural improvements may be necessary. However, ETKDG and Torsional Diffusion failed to generate valid structures for 25 of 102 test molecules, while flow based methods succeeded on all.

The energy benchmark (Table 6) reveals complementary trade offs between conformer quality and diversity. ETKDG produces conformers closest to local minima (smallest relaxation energy), but these conformers do not reach the lowest energy basins, consistent with mode collapse to a narrow region of conformational space. Flow based methods generate more structurally diverse conformers that require more optimization but reach lower energy basins overall. Torsional Diffusion achieves the best balance, finding the lowest energy conformers while requiring less relaxation than flow based methods. We note that GFN2-xTB is a semi empirical approximation, but as the same level of theory used to generate the CREST reference ensembles, it ensures internal consistency[11].

4 Discussion and Conclusion

This work argues that data splitting strategy substantially impacts reported performance and that joint training on chemically diverse datasets can improve coverage on underrepresented classes. Lyrebird uses ET-Flow without modification and claims no architectural novelty. We view the primary contributions as data centric: the Butina clustered splits, the MPCONF196GEN benchmark, and the energy based evaluation framework, which we hope will be useful for future model development regardless of architecture. On macrocycles, flow based methods prove more robust than Torsional Diffusion, which struggles when ring constraints violate torsional independence assumptions. The energy benchmark reveals that ETKDG produces locally robust but mode collapsed ensembles; flow based methods better explore conformational space at the cost of requiring optimization; Torsional Diffusion bridges the gap but inherits ETKDG's dependence on reasonable initial structures. Limitations are clear: on GEOM-XL, Lyrebird does not improve over ET-Flow despite additional data (Table 5), and Torsional Diffusion achieves the best energy results (Table 6). Multi dataset training is far from sufficient for a truly general purpose conformer search model. For future work, we see promise in larger training sets following OMOL-25[15], physical priors, and hardware efficient equivariant architectures.

Acknowledgements

The authors thank Corin Wagen, Jonathon Vandezande, and Ari Wagen for helpful discussions and early edits of the manuscript.

Bibliography

[1] P. Pracht *et al.*, "CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space," *Journal of Chemical Physics*, vol. 160, no. 11, p. 114110, 2024, doi: 10.1063/5.0197592.

[2] S. Wang, J. Witek, G. A. Landrum, and S. Riniker, "Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences," *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 2044–2058, Apr. 2020.

[3] B. Jing, G. Corso, J. Chang, R. Barzilay, and T. Jaakkola, "Torsional Diffusion for Molecular Conformer Generation." [Online]. Available: https://arxiv.org/abs/2206.01729

[4] Y. Wang, A. A. Elhag, N. Jaitly, J. M. Susskind, and M. A. Bautista, "Swallowing the Bitter Pill: Simplified Scalable Conformer Generation." [Online]. Available: https://arxiv.org/abs/2311.17932

[5] M. Hassan, N. Shenoy, J. Lee, H. Stark, S. Thaler, and D. Beaini, "ET-Flow: Equivariant Flow-Matching for Molecular Conformer Generation." [Online]. Available: https://arxiv.org/abs/2410.22388

[6] P. Thölke and G. D. Fabritiis, "TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials." [Online]. Available: https://arxiv.org/abs/2202.02541

[7] C. A. Grambow, H. Weir, N. L. Diamant, G. Scalia, T. Biancalani, and K. V. Chuang, "Accurate and Efficient Structural Ensemble Generation of Macrocyclic Peptides using Internal Coordinate Diffusion." [Online]. Available: https://arxiv.org/abs/2305.19800

[8] S. Axelrod and R. Gómez-Bombarelli, "GEOM, energy-annotated molecular conformations for property prediction and molecular generation," *Scientific Data*, vol. 9, no. 1, p. 185, 2022, doi: 10.1038/s41597-022-01288-4.

[9] Q. Guo, S. Hernandez-Hernandez, and P. J. Ballester, "Scaffold Splits Overestimate Virtual Screening Performance." [Online]. Available: https://arxiv.org/abs/2406.00873

[10] C. A. Grambow, H. Weir, C. N. Cunningham, T. Biancalani, and K. V. Chuang, "CREMP: Conformer-rotamer ensembles of macrocyclic peptides for machine learning," *Scientific Data*, vol. 11, no. 1, p. 859, 2024, doi: 10.1038/s41597-024-03698-y.

[11] C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions," *Journal of Chemical Theory and Computation*, vol. 15, no. 3, pp. 1652–1671, 2019, doi: 10.1021/acs.jctc.8b01176.

[12] D. Reidenbach, F. Nikitin, O. Isayev, and S. Paliwal, "Applications of Modular Co-Design for De Novo 3D Molecule Generation." [Online]. Available: https://arxiv.org/abs/2505.18392

[13] "Some Thoughts on Splitting Chemical Datasets — practicalcheminformatics.blogspot.com." 2024.

[14] M. Xu, S. Luo, Y. Bengio, J. Peng, and J. Tang, "Learning Neural Generative Dynamics for Molecular Conformation Generation." [Online]. Available: https://arxiv.org/abs/2102.10240

[15] D. S. Levine *et al.*, "The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models." [Online]. Available: https://arxiv.org/abs/2505.08762

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models." [Online]. Available: https://arxiv.org/abs/2006.11239

[17] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models." [Online]. Available: https://arxiv.org/abs/2010.02502

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations." [Online]. Available: https://arxiv.org/abs/2011.13456

[19] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling." [Online]. Available: https://arxiv.org/abs/2210.02747

[20] X. Liu, C. Gong, and Q. Liu, "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." [Online]. Available: https://arxiv.org/abs/2209.03003

[21] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden, "Stochastic Interpolants: A Unifying Framework for Flows and Diffusions." [Online]. Available: https://arxiv.org/abs/2303.08797

[22] A. Tong *et al.*, "Improving and generalizing flow-based generative models with minibatch optimal transport." [Online]. Available: https://arxiv.org/abs/2302.00482

[23] R. T. Q. Chen and Y. Lipman, "Flow Matching on General Geometries." [Online]. Available: https://arxiv.org/abs/2302.03660

[24] D. Haviv, A.-A. Pooladian, D. Pe'er, and B. Amos, "Wasserstein Flow Matching: Generative modeling over families of distributions." [Online]. Available: https://arxiv.org/abs/2411.00698

[25] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Q. Chen, "Multisample Flow Matching: Straightening Flows with Minibatch Couplings." [Online]. Available: https://arxiv.org/abs/2304.14772

[26] J. Řezáč, D. B\ím, O. Gutten, and L. Rul\íšek, "Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set," *Journal of Chemical Theory and Computation*, vol. 14, no. 3, pp. 1254–1266, 2018, doi: 10.1021/acs.jctc.7b01074.

[27] E. L. Mann, C. C. Wagen, J. E. Vandezande, A. M. Wagen, and S. C. Schneider, "Egret-1: Pretrained Neural Network Potentials for Efficient and Accurate Bioorganic Simulation," *arXiv preprint arXiv:2504.20955*, 2025, doi: 10.48550/arXiv.2504.20955.

A Appendix

A.1 Background: Flow Matching and Diffusion

Diffusion models enable high quality generation by approximating the stochastic differential equation (SDE) mapping a simple noise distribution (e.g., Gaussian) to data, also known as the reverse time process (Equation 3). This was originally proposed as learning to denoise data directly via a Markov chain setup that allows for computable likelihoods and a variational bound (DDPM)[16]. Later work (DDIM) formulated this as an implicit probabilistic model that does not require Markovian sampling, though at the cost of exact likelihood computation[17]. Diffusion modeling was later formalized as score matching[18], learning the score function (Equation 2) of the diffused data at each time step. During inference, the learned score integrates the reverse time stochastic differential equation to generate samples from noise. While diffusion models have shown excellent results, they come with inherent drawbacks: often requiring longer training time to converge due to SDE noise, as well as more sampling steps at inference. DDIM can be recovered as a discretization of the probability flow ODE (Equation 4) integrated using forward Euler, which is a deterministic version of Equation 3. Similarly, DDPM can be viewed as a discretization of the reverse time SDE integrated using Euler–Maruyama.

$$\mathrm{d}x_t = f(x_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t \tag{1}$$

$$s_\theta(x_t, t) = \nabla_{x_t} \log p_{t(x_t)} \tag{2}$$

$$\mathrm{d}x_t = \left(f(x_t, t) - g(t)^2 s_\theta(x_t, t)\right)\mathrm{d}t + g(t)\,\mathrm{d} \mid W_t \tag{3}$$

$$dx_t = \left(f(x_t, t) - \frac{1}{2} * g(t)^2 * s_\theta(x_t, t)\right)dt \tag{4}$$

Flow matching is a method that has recently become popular as an alternative to diffusion[19–21]. It trains a flow by directly learning the velocity field at random steps along an interpolation between noise and data, then integrating the vector field at inference time using standard ODE solvers. This has proven to be a highly stable and scalable general purpose generative framework. Variants of flow matching have been proposed with different target computations (OT maps)[22], interpolation styles (geodesic, Wasserstein)[23,24], and arbitrary source distributions[25], making it more flexible than diffusion.

A.2 Network Architecture Details

A.2.1 ET-Flow and Equivariant Machine Learning

ET-Flow builds on TorchMD-Net's equivariant transformer. The model is O(3) equivariant, with post hoc chirality correction comparing oriented volumes against RDKit tags.

A function $f : V \to W$ is ***G*-equivariant** if $f(\rho(g) \cdot x) = \tau(g) \cdot f(x)$ for representations ρ, τ. For conformer generation, SO(3) equivariance ensures rotating inputs produces equivalently rotated outputs.

A.2.2 Input Processing

The network processes molecular structures with the following components:

- **Inputs**: Atom scalars (RDKit flags, atomic number z, time t), vector coordinates $\boldsymbol{x}_i \in \mathbb{R}^3$
- **Edges**: Bonds ∪ neighbors within cutoff radius $r_c = 10$ Å
- **Distances**: Transformed via RBFs with smooth cutoff function
- **Time step**: Incorporated as input to the embedding layer

- **Features**: Strictly derived from `MolFromSmiles` (no test time leakage)

A.2.3 Embedding Layer

The embedding layer maps atomic numbers and attributes to learned representations:

$$z_i = \text{embed}_{\text{int}(z_i)} \tag{5}$$

$$h_i = \text{MLP}(h_i) \tag{6}$$

$$n_i = \sum_{j=1}^{N} \text{embed}_{\text{nbh}(z_j)} \cdot g\big(d_{ij}, l_{ij}\big) \tag{7}$$

A.2.3.1 Distance Encoding and Radial Basis Functions

Distance expansion using K exponential RBFs:

$$e_k^{\text{RBF}}(d) = \exp\Big(-\beta \cdot (d - \mu_k)^2\Big), \quad k = 1, ..., K \tag{8}$$

The interaction function combines distance and edge information:

$$g(d_{ij}, l_{ij}) = W^F \left[\varphi(d_{ij}) e_1^{\text{RBF}}(d_{ij}), ..., \varphi(d_{ij}) e_K^{\text{RBF}}(d_{ij}), l_{ij}\right] \tag{9}$$

A.2.3.2 Smooth Cutoff Function

$$\varphi(d) = \begin{cases} \frac{1}{2}\Big(\cos\Big(\pi \frac{d_{i,j}}{r_c}\Big) + 1\Big) & \text{if } 0 \le d_{i,j} \le r_c \\ 0 & \text{if } d_{i,j} > r_c \end{cases} \tag{10}$$

A.2.3.3 Final Embedding

$$\boldsymbol{x}_i = W^C \left[\text{embed}_{\text{int}(z_i)}, h_i, t, n_i\right] \tag{11}$$

A.2.4 Attention Mechanism

Multi head dot product attention with distance modulation.

$$\boldsymbol{x}_i' = \text{LayerNorm}(\text{MLP}([\boldsymbol{x}_i, h_i, t])) \tag{12}$$

$$\boldsymbol{Q} = \text{LayerNorm}\big(W^Q \boldsymbol{x}_i'\big), \quad \boldsymbol{K} = \text{LayerNorm}\big(W^K \boldsymbol{x}_j'\big), \quad \boldsymbol{V} = W^V \boldsymbol{x}_j' \tag{13}$$

Distance Filters:

$$D^K = \sigma\Big(W^{D_K} e^{\text{RBF}(d_{ij})} + b^{D_K}\Big), \quad D^V = \sigma\Big(W^{D_V} e^{\text{RBF}(d_{ij})} + b^{D_V}\Big) \tag{14}$$

Attention Scores:

$$\text{dot}(\boldsymbol{Q}, \boldsymbol{K}, D^K) = \sum_{k=1}^{F} Q_k \cdot K_k \cdot D_k^K \tag{15}$$

$$A_{ij} = \text{SiLU}(\text{dot}(\boldsymbol{Q}, \boldsymbol{K}, D^K)) \cdot \varphi\big(d_{ij}\big) \tag{16}$$

Value Processing And Aggregation:
$s_{ij}^1, s_{ij}^2, s_{ij}^3 = \text{split}\Big(\boldsymbol{V}_j \cdot D_{ij}^V\Big)$, then $y_i = W^O \sum_j A_{ij} \cdot s_{ij}^3$

A.2.5 Update Layer

Scalar Update: Split attention output:

$$q_i^1, q_i^2, q_i^3 = \text{split}(y_i) \tag{17}$$

$$\Delta \boldsymbol{x}_i = q_i^1 + q_i^2 \cdot \langle U_1 \boldsymbol{v}_i \cdot U_2 \boldsymbol{v}_i \rangle \tag{18}$$

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \Delta \boldsymbol{x}_i \tag{19}$$

Vector Update using normalized displacements $\boldsymbol{u}_{ij} = \frac{\boldsymbol{r}_j - \boldsymbol{r}_i}{\max(\|\boldsymbol{r}_j - \boldsymbol{r}_i\|, \varepsilon)}$:

$$\boldsymbol{w}_i = \sum_j \left[s_{ij}^1 \boldsymbol{v}_j + s_{ij}^2 \boldsymbol{u}_{ij} \right], \quad \Delta \boldsymbol{v}_i = \boldsymbol{w}_i + q_i^3 U_3 \boldsymbol{v}_i, \quad \boldsymbol{v}_i \leftarrow \boldsymbol{v}_i + \Delta \boldsymbol{v}_i \tag{20}$$

A.2.6 Output Layer (Velocity Field)

The output layer uses gated equivariant blocks to produce the velocity field:

$$\boldsymbol{x}_{\text{i,updated}}, \boldsymbol{w}_i = \text{split}(\text{MLP}([\boldsymbol{x}_i, U_1 \boldsymbol{v}_i])) \tag{21}$$

$$\boldsymbol{v}_{\text{i,updated}} = U_2 \boldsymbol{v}_i \cdot \boldsymbol{w}_i \tag{22}$$

Note: O(3) equivariant output; chirality correction applied post hoc by comparing oriented volumes with RDKit tags. If mismatch detected, apply reflection to correct.

A.3 Flow Matching Training Details

A.3.1 Interpolation Path

We use linear interpolation between harmonic prior and target with Gaussian noise and the noise schedule σ:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1 + \sigma\sqrt{t(1-t)}\boldsymbol{z} \tag{23}$$

A.3.2 Harmonic Prior

Samples respect molecular topology via graph Laplacian $L = D - A$. With eigendecomposition $L = P\Lambda P^T$:

$$\boldsymbol{x}_0 = P\left(\text{diag}\left(\frac{1}{\sqrt{\lambda}}\right) \cdot \boldsymbol{z}\right), \quad \boldsymbol{z} \sim \mathcal{N}(0, I) \tag{24}$$

(Zero eigenvalues set to zero; repeated independently for each coordinate.)

A.3.3 Alignment

$\boldsymbol{x}_0 \leftarrow \text{RMSDAlign}(\boldsymbol{x}_0, \boldsymbol{x}_1)$ via Kabsch algorithm.

A.3.4 Target Velocity Field

$$\boldsymbol{v}_t(\boldsymbol{x}_t) = \boldsymbol{x}_1 - \boldsymbol{x}_0 + \frac{1-2t}{2\sqrt{t(1-t)}} \cdot \boldsymbol{z} \tag{25}$$

A.3.5 Training Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1], \boldsymbol{x}\sim\rho_t}\left[\|\boldsymbol{v}_\theta(t, \boldsymbol{x}_t) - \boldsymbol{v}_t(\boldsymbol{x}_t)\|_2^2\right] \tag{26}$$

A.4 Benchmark Details

A.4.1 MPCONF196GEN

We introduce MPCONF196GEN, a challenging benchmark for macrocycle conformer generation based on the MPCONF196 energy benchmark[26], containing 13 molecules. We ran CREST-iMTD-GC[1] with GFN2-xTB[11] to compute conformer ensembles for each molecule

and computed conformer energies with the Egret-1[27] neural network potential, which achieved DFT like performance on the MPCONF196 energy benchmark.

A.4.2 Energy Benchmark Methodology

The energy benchmark addresses two questions that coverage matching metrics do not: (1) whether methods produce reasonable structures, and (2) whether these structures are at energy minima or can be optimized to them. We use GFN2-xTB, the same level of theory used by CREST to generate reference ensembles, noting that it is a semi empirical method with known limitations relative to DFT but ensuring internal consistency.

Procedure:

1. Compute initial GFN2-xTB energies for $2n$ generated conformers (where n is the number of ground truth CREST conformers)
2. Run GFN2-xTB geometry optimization
3. Compute energies of optimized structures
4. Compare to ground truth minimum energy structure

The left columns in Table 6 report average GFN2-xTB energy change after optimization. Since optimization always lowers energy, all values are negative; values closer to zero indicate conformers that were already near a local minimum before optimization. The right column compares each method's lowest energy generated conformer to the ground truth CREST minimum per molecule; negative values indicate the method found a lower energy basin than the reference.

We used a subset of 20 GEOM-XL molecules due to computational constraints (storing all intermediate structures is memory intensive). This benchmark focuses on conformer stability and relaxation to minima, rather than ensemble weighted properties post optimization.

A.4.3 Comparison with Related Energy Based Benchmarks

Concurrent work by Reidenbach et al.[12] independently introduced energy based structural evaluation for generative molecular models. Both approaches share the intuition that GFN2-xTB relaxation can probe whether generated structures are physically reasonable, but they address different problem settings:

1. **Task**: Reidenbach et al. evaluate *unconditional de novo molecule generation*, where both the molecular graph and 3D structure are generated simultaneously. Our benchmark evaluates *conditional conformer generation*, where the molecular graph is given and only 3D coordinates are predicted.
2. **Metrics**: Reidenbach et al. report relaxation energy (ΔE_{relax}) and structural errors (bond lengths, bond angles, dihedral angles) between generated and optimized structures of the same generated molecule. Our benchmark reports energy change upon optimization (assessing proximity to local minima) and additionally compares the lowest energy generated conformer to the ground truth CREST minimum, assessing basin sampling quality, a comparison specific to the conditional setting where reference ensembles exist.
3. **Scope**: Their evaluation spans thousands of unconditionally generated molecules of varying topology. Ours evaluates a fixed set of 20 molecules with known reference conformer ensembles, enabling direct comparison of generated vs. reference energy basins.

The two benchmarks thus address a similar concern, that RMSD based metrics alone do not capture physical reasonableness, but do so in distinct domains with different evaluation criteria.