Towards Molecular Conformer Generation with Language Models

Menua Bedrosian [1 2] **Hrant Khachatrian** [1 2]

Abstract

Generating accurate 3D conformations of small molecules from their 2D representations is a central task in computational drug discovery, impacting molecular docking, virtual screening, and property prediction. While most recent advances rely on diffusion-based generative models, these methods come with limitations such as slow sampling and architectural rigidity. In this work, we demonstrate that autoregressive language models can effectively learn to generate 3D molecular conformations from text-only data. We propose a simple yet expressive representation that combines canonical SMILES with raw 3D atomic coordinates in a unified tokenized format. Using this approach, we train language models ranging from 100 million to 1 billion parameters on a dataset curated from GEOM-Drugs. While our models currently perform slightly behind the best diffusion-based methods, they achieve competitive results and show consistent improvements with scale. We derive empirical scaling laws demonstrating that generation quality improves predictably with model size and data, suggesting a clear path toward closing the performance gap. These findings indicate that language models are a scalable and flexible alternative for 3D molecular generation, with potential for further improvement through recent advancements in large language models, such as in-context learning and post-training adaptation.

1. Introduction

The three-dimensional (3D) structure of a molecule plays a pivotal role in determining its physicochemical properties and biological activity. Accurate conformer generation is critical in many downstream applications, including molecular docking and rational drug design. Traditional approaches such as X-ray crystallography, and density functional theory (DFT) can yield highly accurate structures (Hawkins, 2017; Pracht et al., 2020), but they are expensive and impractical at scale.

To address this, deep learning-based generative methods have emerged, with diffusion models representing the current state-of-the-art (Ganea et al., 2021; Jing et al., 2022). These models have shown strong performance on benchmark datasets, but they suffer from several limitations: expensive multi-step sampling, restricted flexibility at inference, and strong architectural coupling to the training framework.

Recent works have begun exploring large language models (LLMs) as an alternative approach (Zhang et al., 2023; Zholus et al., 2024), but thus far, they have not demonstrated clear competitiveness with diffusion models on key benchmarks. In this work, we investigate the potential of LLMs for direct 3D conformer generation from molecular 2D representations. We find that while LLMs do not yet surpass the most recent diffusion-based models, their performance consistently improves with scale and training data. This trend suggests a promising path forward. Our key contributions are:

1. We introduce a token-based representation that merges 2D molecular topology and 3D conformation into a text-only format suitable for LLM training.

2. We train LangMol - foundation models from scratch (100M–1B parameters) on a curated dataset from GEOM-Drugs and show competitive performance on standard metrics.

3. We derive empirical scaling laws showing that model performance improves as the model size and the number of tokens increase, highlighting the scalability of this approach.

[1]YerevaNN Research Lab, Yerevan, Armenia [2]Yerevan State University, Yerevan, Armenia. Correspondence to: Menua Bedrosian <{menua.bedrosian}@ysu.am>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

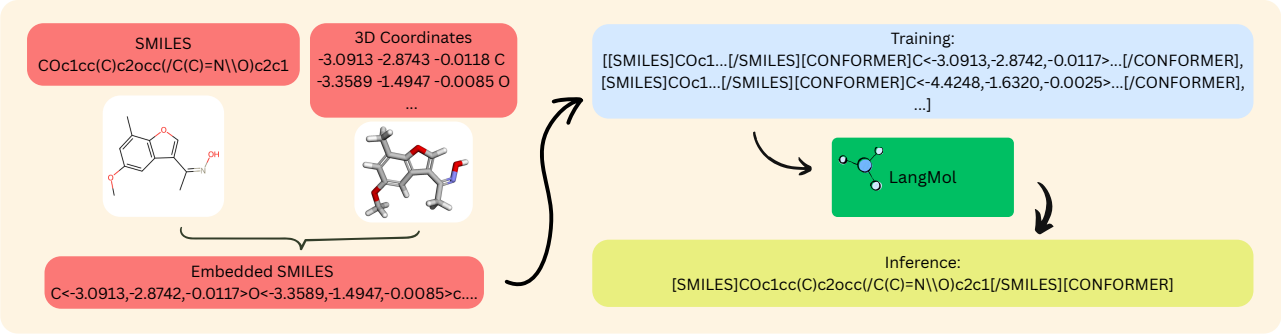

Figure 1. Overview of LangMol. The 2D SMILES and 3D conformer data get combined in the embedded SMILES format (left). The model is trained on pairs of SMILES and embedded SMILES. Inference is done by prompting the SMILES (right).

2. Related work

Molecular Conformation Generation The generation and optimization of molecular conformers have been studied in the cheminformatics field extensively by (Hawkins, 2017; Lagorce et al., 2009; Bolton et al., 2011; Cole et al., 2018; Miteva et al., 2010). For example, Axelrod & Gomez-Bombarelli (2022) report that generating conformers for a single drug-like molecule can require up to 90 core-hours, making these methods impractical for large-scale applications.

Diffusion-Based Models. Recent diffusion models like GeoMol (Ganea et al., 2021), Torsional Diffusion (Jing et al., 2022), Symphony (Daigavane et al., 2023), and ET-Flow (Hassan et al., 2024) achieve state-of-the-art accuracy by modeling torsional degrees of freedom and leveraging equivariant architectures. However, they require expensive iterative sampling and often rely on handcrafted features, which hinder scalability. Fast inference variants (Zhang & Chen, 2022) mitigate this to some extent, but do not fully eliminate the bottlenecks.

Language Models for Molecular Modeling. Large language models (LLMs) like GPT-3 (Brown et al., 2020) and LLaMA (Dubey et al., 2024) have shown strong generalization through in-context learning and prompt conditioning. In chemistry, models like ChemLactica (Guevorguian et al., 2024) and MolT5 (Edwards et al., 2022) have demonstrated the utility of language models in learning molecular syntax and semantics. Initial efforts to apply LLMs to 3D tasks, such as Tora3D (Zhang et al., 2023) and BindGPT (Zholus et al., 2024), are promising, but have yet to match the performance of diffusion-based models on standard conformer generation benchmarks.

3. Methods

Data Preparation We construct our training dataset from the GEOM-Drugs collection (Axelrod & Gomez-Bombarelli, 2022), which provides high-quality 3D conformers for drug-like molecules. Each training instance consists of two paired representations: a canonical SMILES string describing the 2D molecular graph, and an embedded SMILES string that extends this sequence by appending Cartesian 3D coordinates to each atomic symbol.

For molecules with multiple conformers, we generate one example per conformer, keeping the same canonical SMILES while varying the embedded conformer representation. We format each sequence using explicit textual delimiters—[SMILES], [/SMILES], [CONFORMER], [/CONFORMER]—as introduced in Guevorguian et al. (2024), which helps the model distinguish between structural and geometric content. A summarized example is shown in Figure 1; tokenization details are provided in Appendix A.

This format provides several advantages:

(1) It allows the model to learn directly from raw 3D coordinates bypassing the need for torsion angle preprocessing or engineered without the need for additional preprocessing or engineered features, as supported by (Wang et al., 2024; Zholus et al., 2024);

(2) It merges two distinct modalities—2D molecular topology and 3D spatial conformation—into a unified, text-only format that supports end-to-end training;

Model Training We pretrain decoder-only autoregressive language models with sizes ranging from 100 million to 1 billion parameters. These models follow the LLaMA 3.2 architecture (Dubey et al., 2024), with a reduced context length of 2048 tokens to improve training efficiency. The models are trained starting from a random initialization. Model configuration details and training hyperparameters are listed in Appendix B. We train each model for 4 epochs on the full dataset. Learning rate schedules follow warmup-stable-decay technique described by Wen et al. (2024), and checkpoint-specific decay phases are applied post hoc to

Table 1. Molecule conformer generation results on GEOM-DRUGS ($\delta = 0.75\mathring{A}$).

	Recall				Precision			
	Coverage ↑		AMR ↓		Coverage ↑		AMR ↓	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GeoDiff	42.10	37.80	0.835	0.809	24.90	14.50	1.136	1.090
GeoMol	44.60	41.40	0.875	0.834	43.00	36.40	0.928	0.841
Torsional Diff.	72.70	80.00	0.582	0.565	55.20	56.90	0.778	0.729
ET-Flow	79.53	84.57	0.452	0.419	**74.38**	**81.04**	**0.541**	**0.470**
MCF - L	**84.70**	**92.20**	**0.390**	**0.247**	66.80	71.30	0.618	0.530
LangMol-1B	59.38	58.33	0.710	0.610	57.08	57.14	0.888	0.712

stabilize final performance.

Inference At inference time, the model is prompted with a canonical SMILES string and tasked with autoregressively generating a corresponding 3D conformation. Our approach supports efficient batched prompting, enabling scalable and parallelizable generation. We use temperature sampling to balance diversity and fidelity. Appendix D explores the effects of different sampling temperatures, as well as alternative sampling strategies and their failure cases.

4. Experiments

Evaluation Setup We follow the dataset splits and preprocessing setup from Ganea et al. (2021), and process our dataset into training, validation, and test splits (243,473 / 30,433 / 1,000 molecules, respectively). Each molecule in the training set is limited to a maximum of 30 conformers to prevent class imbalance. After tokenizing the curated dataset with LlaMA 3 (including special tokens), we obtain roughly 2.5 billion tokens.

To evaluate our model, we generate 2k conformers per molecule in the test set. We then compute Coverage and Average Minimum RMSD (AMR) using RMSD-based metrics from prior literature (Axelrod & Gomez-Bombarelli, 2022; Ganea et al., 2021), applying a threshold of 0.75Å for coverage. Precision and recall are reported separately, reflecting how well the generated ensemble matches the diversity and accuracy of ground-truth conformers. Exact metric definitions are provided in Appendix E.

Performance Results Table 1 compares our largest model, LangMol (1B parameters, trained for 4 epochs), against leading diffusion-based models. Although LangMol still lags behind the best-performing methods such as MCF and ET-Flow, it achieves surprisingly competitive results—particularly in recall coverage and precision—despite requiring only a single-pass generation, without iterative sampling or equivariant feature design. Note that the trade-off between precision and recall can be controlled after pretraining, at the generation phase. See Appendix D for more details.

Does it Scale? Transformer-based language models are known to be scalable with respect to number of parameters and number of tokens with no signs of saturation (Kaplan et al., 2020; Hoffmann et al., 2022). To verify whether scaling effects hold for molecular conformation generation we train models of four sizes (100M, 170M, 380M, 1B) for 1, 2, 3 and 4 epochs of the same data. Fig. 2 shows the training losses for all 16 combinations. Furthermore, we fit $L(N, D) = A \cdot N^b \cdot D^c$ formula on these points using L-BFGS in the log-log space, and obtain $L = 11.6436 \cdot N^{-0.0641} \cdot D^{-0.0499}$. Clearly, the loss gets better on both axes.

Next, we validate that lower loss corresponds to better performance metrics. Fig. 2 (right) shows that both coverage metrics (precision and recall) increase with lower values of training loss. By combining this with the scaling law derived above we can hypothesize that larger parameter count *and* larger dataset size should continue to improve the performance of our approach. The experiments for validating this with larger N and D is left for future work.

5. Conclusion and Future Work

In this work we showed that autoregressive language models are a viable solution for molecular conformation generation. While our LangMol was not able to beat state-of-the-art metrics obtained with diffusion models, we have demonstrated a viable path forward.

We have shown that larger language models with larger data generate better conformers. One obvious direction for future work is to verify the scaling properties with larger experiments. We believe this will require significantly larger datasets, as repeating the same data for more than four epochs will bring diminishing improvements (Muennighoff et al., 2023).

In parallel, we encourage future research on improving the

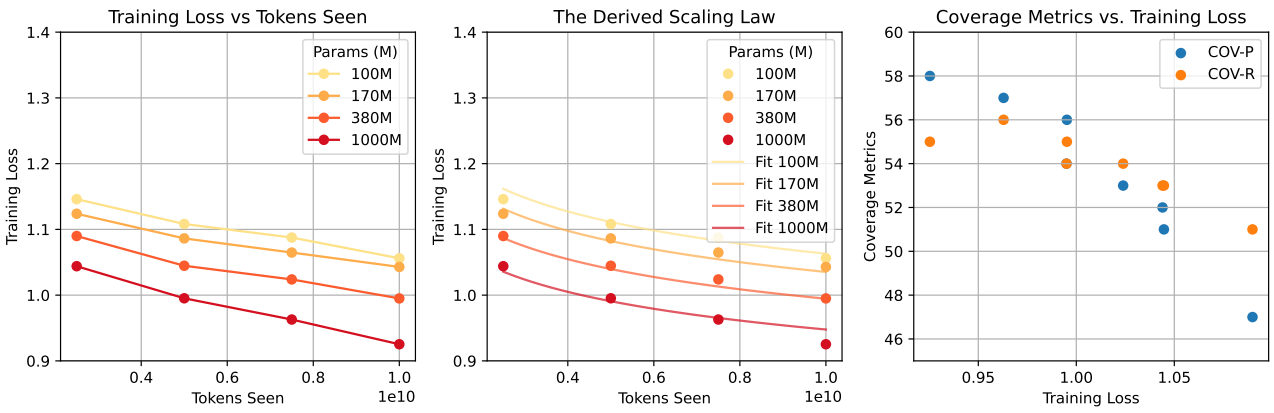

Figure 2. The dependence of training loss on model size and number of the tokens seen during training (left), scaling law derived from these values (center), and the dependence of coverage metrics on training loss (right).

scaling parameters demonstrated in this paper. Such improvements might come from better tokenizers, better filtered datasets, post-training, and in-context learning.

Next, we noticed that the conformers generated by both diffusion and autoregressive approaches suffer from physical inaccuracies. We strongly believe that tools similar to PoseBusters (Buttenschoen et al., 2024) should become an integral part of the evaluation process to identify critical errors of these generative models.

Finally, practical applications of conformer generation methods might require very high speed of execution. We believe the recipe should be the following: training the largest and most powerful models, then distilling them into smaller, high performance models. We hope this work motivates more research on scalable and accurate conformer generation.

Acknowledgements

The research was supported by the Higher Education and Science Committee of MESCS RA (Research project No 24FP-1A058). We would also like to thank Professor Ruben Abagyan for his valuable advice and insightful suggestions, which greatly contributed to the direction of this work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Axelrod, S. and Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Bolton, E. E., Kim, S., and Bryant, S. H. Pubchem3d: conformer generation. *Journal of cheminformatics*, 3: 1–16, 2011.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Buttenschoen, M., Morris, G. M., and Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Cole, J. C., Korb, O., McCabe, P., Read, M. G., and Taylor, R. Knowledge-based conformer generation using the cambridge structural database. *Journal of chemical information and modeling*, 58(3):615–629, 2018.

Daigavane, A., Kim, S., Geiger, M., and Smidt, T. Symphony: Symmetry-equivariant point-centered spherical harmonics for molecule generation. *arXiv preprint arXiv:2311.16199*, 2023.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Ganea, O., Pattanaik, L., Coley, C., Barzilay, R., Jensen, K., Green, W., and Jaakkola, T. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34: 13757–13769, 2021.

Guevorguian, P., Bedrosian, M., Fahradyan, T., Chilingaryan, G., Khachatrian, H., and Aghajanyan, A. Small molecule optimization with large language models. *arXiv preprint arXiv:2407.18897*, 2024.

Hassan, M., Shenoy, N., Lee, J., Stark, H., Thaler, S., and Beaini, D. Et-flow: Equivariant flow-matching for molecular conformer generation. *arXiv preprint arXiv:2410.22388*, 2024.

Hawkins, P. C. Conformation generation: the state of the art. *Journal of chemical information and modeling*, 57 (8):1747–1756, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35: 24240–24253, 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Lagorce, D., Pencheva, T., Villoutreix, B. O., and Miteva, M. A. Dg-ammos: A new tool to generate 3d conformation of small molecules using d istance g eometry and a utomated m olecular m echanics o ptimization for in silico s creening. *BMC Chemical Biology*, 9:1–10, 2009.

Liang, W., Liu, T., Wright, L., Constable, W., Gu, A., Huang, C.-C., Zhang, I., Feng, W., Huang, H., Wang, J., Purandare, S., Nadathur, G., and Idreos, S. Torchtitan: One-stop pytorch native solution for production ready LLM pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=SFN6Wm7YBI`.

Miteva, M. A., Guyon, F., and Tuffi¿ 1/2ry, P. Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic acids research*, 38(suppl_2):W622–W627, 2010.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Nguyen, M., Baker, A., Neo, C., Roush, A., Kirsch, A., and Shwartz-Ziv, R. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*, 2024.

Pracht, P., Bohle, F., and Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14):7169–7192, 2020.

Wang, Y., Elhag, A. A., Jaitly, N., Susskind, J. M., and Bautista, M. Á. Swallowing the bitter pill: Simplified scalable conformer generation. In *Forty-first International Conference on Machine Learning*, 2024.

Wen, K., Li, Z., Wang, J., Hall, D., Liang, P., and Ma, T. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

Zhang, Q. and Chen, Y. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Zhang, Z., Wang, G., Li, R., Ni, L., Zhang, R., Cheng, K., Ren, Q., Kong, X., Ni, S., Tong, X., et al. Tora3d: an autoregressive torsion angle prediction model for molecular 3d conformation generation. *Journal of Cheminformatics*, 15(1):57, 2023.

Zholus, A., Kuznetsov, M., Schutski, R., Shayakhmetov, R., Polykovskiy, D., Chandar, S., and Zhavoronkov, A. Bindgpt: A scalable framework for 3d molecular design via language modeling and reinforcement learning. *arXiv preprint arXiv:2406.03686*, 2024.

A. Tokenization Format

Our novel approach to molecular data embedding integrates 3D conformational information directly into the SMILES string. For each atom in a given SMILES sequence, its corresponding 3D Cartesian coordinates are appended immediately to the right of the atom symbol, enclosed within angle brackets (e.g., C<-4.4248,-1.6320,-0.0025>). This process is applied sequentially across the entire SMILES string, creating a comprehensive, conformer-aware representation. To delineate the distinct data types within this combined string, we employ specific tags: the entire SMILES string is enclosed by [SMILES] and [/SMILES] tags, while the full embedded string, including the coordinates, is wrapped by [CONFORMER] and [/CONFORMER] tags. This structured format constitutes a single data sample. To encourage the model's ability to learn conformational diversity, we generate multiple such samples for a single SMILES string by repeating the SMILES but embedding the coordinates from different conformers. For clarity in presentation, atom symbols are color-coded blue, their associated coordinates are green, and all structural tags (including ring closure numbers and the aforementioned [ ] delimiters) are rendered in red.

SMILES

`COc1cc(C)c2occ(/C(C)=N\\O)c2c1`

Coordinates

- C -4.4249 -1.632 -0.0026
- O -3.0393 -1.8671 0.0027
- C -2.1872 -0.7982 0.0015
- C -0.8374 -1.1251 0.0027
- C 0.1 -0.0993 0.0013
- C -0.355 1.2338 -0.0001
- C -1.7003 1.5809 -0.0016
- C -2.6081 0.5364 -0.0008
- C -2.1212 3.0164 0.0008
- O 0.694 2.0969 -0.0016
- C 1.8047 1.3443 -0.001
- C 1.5442 0.0001 0.0002
- C 2.5375 -1.06 -0.0008
- N 3.8049 -0.9288 -0.0001
- O 4.3565 0.3554 0.0018
- C 2.0797 -2.4855 -0.0025

Coordinates Embedded SMILES

C<-4.4248,-1.6320,-0.0025>O<-3.0392,-1.8670,0.0026>c<-2.1872,-0.7981,0.0015>1c<-2.6081,0.5364,-0.0008>c<-1.7003,1.5809,-0.0015>(C<-2.1212,3.0163,0.0007>)c<-0.3550,1.2337,-0.0001>2o<0.6940,2.0969,-0.0016>c<1.8046,1.3442,-0.0009>c<1.5441,0.0001,0.0001>(/C<2.5374,-1.0599,-0.0007>(C<2.0797,-2.4855,-0.0025>)=N<3.8049,-0.9288,-0.0001>\\O<4.3564,0.3554,0.0017>)c<0.1000,-0.0992,0.0012>2c<-0.8373,-1.1250,0.0026>1

Sample in Dataset

[SMILES]COc1cc(C)c2occ(/C(C)=N\O)c2c1[/SMILES][CONFORMER]C<-4.4248,-1.6320,-0.0025>O<-3.0392,-1.8670,0.0026>c<-2.1872,-0.7981,0.0015>1c<-2.6081,0.5364,-0.0008>c<-1.7003,1.5809,-0.0015>(C<-2.1212,3.0163,0.0007>)c<-0.3550,1.2337,-0.0001>2o<0.6940,2.0969,-0.0016>c<1.8046,1.3442,-0.0009>c<1.5441,0.0001,0.0001>(/C<2.5374,-1.0599,-0.0007>(C<2.0797,-2.4855,-0.0025>)=N<3.8049,-0.9288,-0.0001>\\O<4.3564,0.3554,0.0017>)c<0.1000,-0.0992,0.0012>2c<-0.8373,-1.1250,0.0026>1[/CONFORMER]

B. Architecture and Hyperparameters

We trained four autoregressive language models based on the LLaMA 3.2 architecture, ranging from 100 million to 1 billion parameters. The architectural configurations for each model, including the number of layers, hidden dimensions, and attention heads, are detailed in table B. All models were trained using a context length of 2048. The training hyperparameters used for all models are summarized in table B. Training was performed on 6 NVIDIA H100 80GB GPUs using the Torch-Titan infrastructure (Liang et al., 2025). The models were trained for four epochs on a dataset consisting of 2.5 billion tokens per epoch, totaling 10 billion tokens seen by each model. Inputs were packed and tokenization was done using the LLaMA 3 tokenizer.

A warmup-stable-decay (Edwards et al., 2022) learning rate schedule was used: a linear warmup over the first 200 steps, followed by a stable phase, and finally a linear decay over the last 10 percent of the total training steps. For each epoch, we saved a primary checkpoint and then applied an additional linear learning rate decay over 1000 steps to generate finer checkpoints used in scaling law analysis. This setup was chosen to isolate model behavior in the stable training regime and observe smooth transitions during the decay phase. In addition to the quantitative tables, Figure shows the training loss curves of the 1B model across the full training trajectory, including the decayed checkpoints, to illustrate training dynamics and convergence behavior.

Model	Dim	Layers	Heads	KV Heads	Params (M)
100M	512	8	8	4	103
170M	768	8	16	8	169
380M	1024	16	16	8	375
1B	2048	16	32	8	1002

Table 2. Architecture and parameter counts of models.

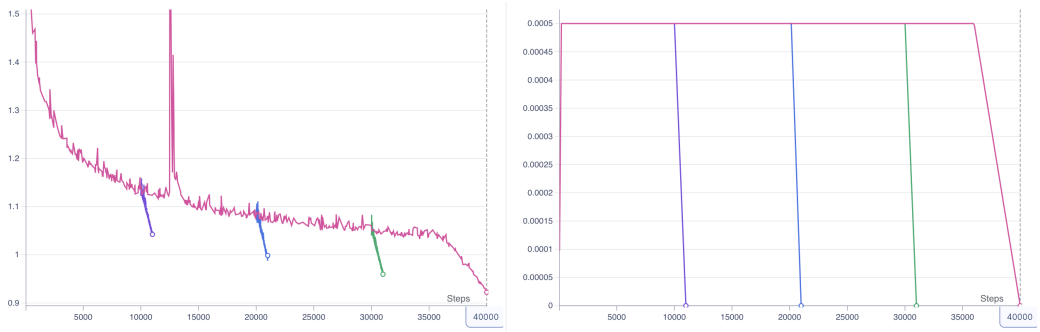

Figure 3. 1B model train loss curve (left) and learning rate schedules (right).

C. Evaluations and Failure Modes

We evaluated our models—380M and 1B parameters—across different numbers of training epochs using top-p sampling with a temperature of 1.0 and top-p value of 0.8. As shown in Table 4, we observe a clear trend: both coverage and matching metrics steadily improve with increased model size and training data. At the same time, various failure modes become less frequent.

Model	Batch Size	Ctx Len	LR	Betas	WD	Warmup
100M	96	2048	0.001	(0.9, 0.95)	0.1	200
170M	128	2048	0.001	(0.9, 0.95)	0.1	200
380M	168	2048	0.001	(0.9, 0.95)	0.1	200
1B	120	2048	0.0005	(0.9, 0.95)	0.1	200

Table 3. Training hyperparameters for each model. All models were trained for 4 epochs (10B tokens total) using AdamW and a warmup-stable-decay schedule.

We define four types of failure. A mismatch occurs when the generated SMILES string does not match the reference SMILES provided in the prompt. No EOS refers to generations that did not terminate within the maximum allowed length of 2000 tokens. Parse fail corresponds to outputs that RDKit could not parse into valid molecules. Finally, a missing mol indicates that none of the generated sequences for a given molecule were valid, leading us to omit that molecule from evaluation. This consistent improvement across metrics and failure modes motivated us to study the underlying scaling laws governing model behavior.

Model	COV-R	COV-P	MAT-R	MAT-P	Miss Mols	Mismatch	No EOS	Parse Fail
380m_1e	51	47	0.78	0.87	6	5361	6	0
380m_2e	53	51	0.77	0.82	3	5099	1	0
380m_3e	54	53	0.76	0.80	2	5921	6	1
380m_4e	54	54	0.76	0.78	2	4672	2	2
1b_1e	53	52	0.78	0.81	2	2518	8	0
1b_2e	55	56	0.77	0.76	2	3076	12	0
1b_3e	56	57	0.75	0.78	1	1781	2	1
1b_4e	55	58	0.75	0.78	2	2292	4	0

Table 4. Model evaluation metrics for various training configurations.

D. Sampling Hyperparameter Search

We experiment with different sampling strategies for our 1B parameter LangMol model to understand their effect on conformer generation quality. Specifically, we vary the top-p nucleus sampling threshold and observe a consistent inverse relationship between recall and precision: increasing the p value tends to improve recall by generating a more diverse set of conformers, but this comes at the cost of lower precision due to reduced structural accuracy. Conversely, lower p values yield more precise but less diverse outputs. We also explore minimum-p sampling (Nguyen et al., 2024), a recently introduced decoding technique designed to promote creativity in generation. However, in our setting, it did not lead to significant improvements in either recall or precision, suggesting that standard top-p sampling remains a more reliable choice for balancing diversity and accuracy in molecular conformer generation. These trends are illustrated in Figure 4, where the trade-offs between different sampling configurations are clearly visible.

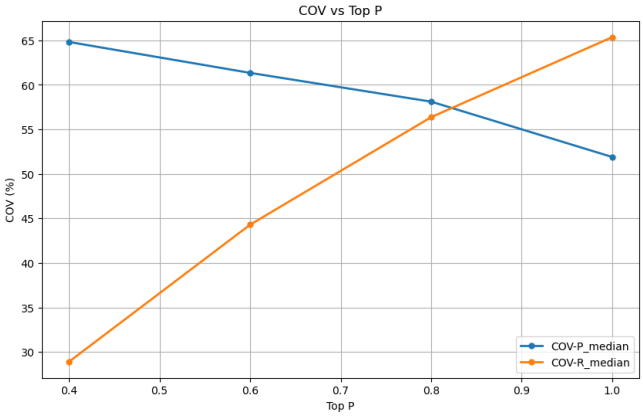

Figure 4. Coverage recall and precision vs. top p sampling values.

E. Metrics Definition

Following (Ganea et al., 2021) we utilize the following metrics to evaluate our work. Here, C_{g} represents the set of generated conformations, while C_{r} denotes the set of reference conformations. For both AMR and COV, we compute and report Recall (R) and Precision (P). Recall quantifies how well the generated conformers capture the ground-truth conformations, whereas Precision measures the proportion of generated conformers that are accurate. The precise definitions of these metrics are provided in the following equations:

$$\text{AMR-R}(C_g, C_r) = \frac{1}{|C_r|} \sum_{\mathbf{R} \in C_r} \min_{\hat{\mathbf{R}} \in C_g} \text{RMSD}(\mathbf{R}, \hat{\mathbf{R}}) \tag{1}$$

$$\text{COV-R}(C_g, C_r) = \frac{1}{|C_r|} \left| \{ \mathbf{R} \in C_r \mid \text{RMSD}(\mathbf{R}, \hat{\mathbf{R}}) < \delta, \hat{\mathbf{R}} \in C_g \} \right| \tag{2}$$

$$\text{AMR-P}(C_r, C_g) = \frac{1}{|C_g|} \sum_{\hat{\mathbf{R}} \in C_g} \min_{\mathbf{R} \in C_r} \text{RMSD}(\hat{\mathbf{R}}, \mathbf{R}) \tag{3}$$

$$\text{COV-P}(C_r, C_g) = \frac{1}{|C_g|} \left| \{ \hat{\mathbf{R}} \in C_g \mid \text{RMSD}(\hat{\mathbf{R}}, \mathbf{R}) < \delta, \mathbf{R} \in C_r \} \right| \tag{4}$$

F. Generation Visualizations

Figure 5 shows randomly selected molecules from GEOM-drugs test set. The reference molecule is depicted on the left while five of the molecules generated by LangMol-1B are randomly selected on the right.

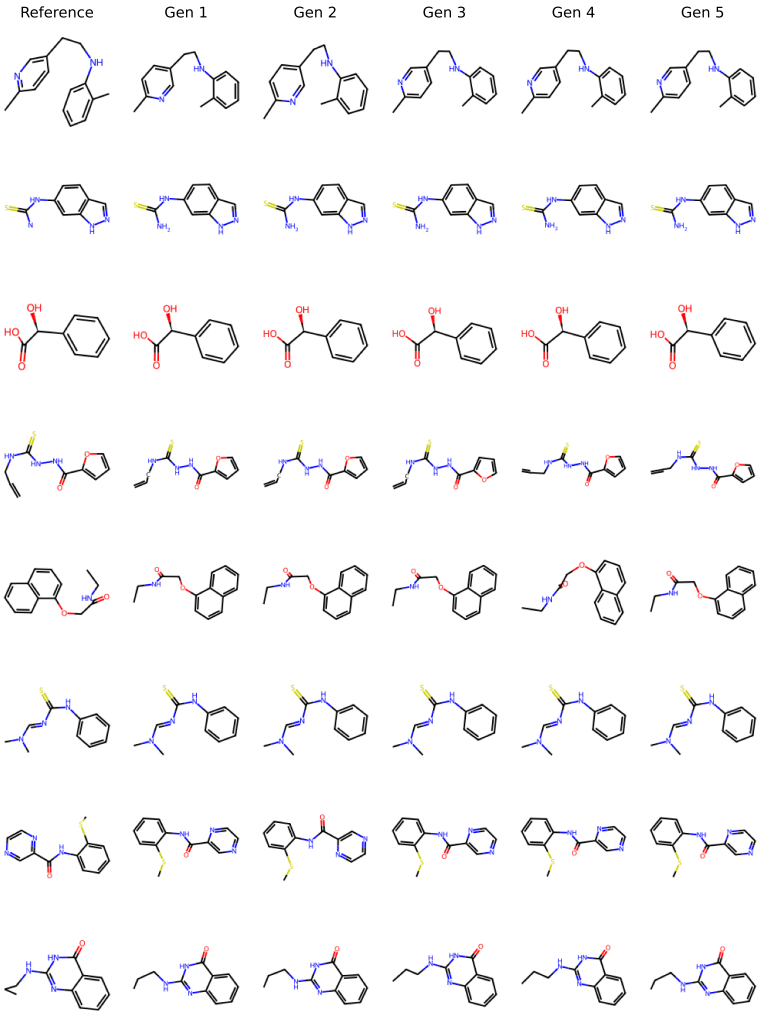

Figure 5. Reference vs. generated conformations of LangMol