Improving Structural Plausibility in 3D Molecule Generation via Property-Conditioned Training with Distorted Molecules

Lucy Vost
Department of Statistics
University of Oxford
Oxford, UK

Vijil Chenthamarakshan
IBM Research
Yorktown Heights
New York, USA

Payel Das
IBM Research
Yorktown Heights
New York, USA

Charlotte M. Deane*
Department of Statistics
University of Oxford
Oxford, UK

Abstract

Traditional drug design methods are costly and time-consuming due to their reliance on trial-and-error processes. As a result, computational methods, including diffusion models, designed for molecule generation tasks have gained significant traction. Despite their potential, they have faced criticism for producing physically implausible outputs. We alleviate this problem by conditionally training a diffusion model capable of generating molecules of varying and controllable levels of structural plausibility. This is achieved by adding distorted molecules to training datasets, and then annotating each molecule with a label representing the extent of its distortion, and hence its quality. By training the model to distinguish between favourable and unfavourable molecular conformations alongside the standard molecule generation training process, we can selectively sample molecules from the high-quality region of learned space, resulting in improvements in the validity of generated molecules. In addition to the standard two datasets used by molecule generation methods (QM9 and GEOM), we also test our method on a druglike dataset derived from ZINC. We use our conditional method with EDM, the first E(3) equivariant diffusion model for molecule generation, as well as two further models—a more recent diffusion model and a flow matching model—which were built off EDM. We demonstrate improvements in validity as assessed by RDKit parsability and the PoseBusters test suite; more broadly, though, our findings highlight the effectiveness of conditioning methods on low-quality data to improve the sampling of high-quality data.

1 Introduction

Drug design involves complex optimisation steps to obtain molecules that achieve desired biological responses. Traditional methods rely on trial-and-error, leading to high costs and limited productivity [25]. Computational approaches, especially deep learning models, aim to reduce costs and expedite processes by reducing failures. One way that such models aim to do this is by generating molecules

*Corresponding author deane@stats.ox.ac.uk

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

with desirable properties, particularly in terms of binding to their target. To achieve this, a model must first master the fundamental task of generating structurally viable molecules.

While many models historically operated in 1D or 2D space [20, 5, 6], focus has recently shifted towards developing models capable of directly outputting both atom types and coordinates in 3D. Autoregressive models were once prominent in this domain, generating 3D molecules by adding atoms and bonds iteratively [14, 22, 16]. However, such models suffer from an accumulation of errors during the generation process and do not fully capture the complexities of real-world scenarios due to their sequential nature, potentially losing global context [12, 13]. To address these limitations, recent studies have turned to diffusion models, which iteratively denoise data points sampled from a prior distribution to generate samples. Unlike autoregressive models, diffusion-based methods can simultaneously model local and global interactions between atoms. Nevertheless, diffusion in molecule generation has faced criticism for yielding implausible outputs [8, 3]. There have been ongoing efforts to improve the performance of models trained on small molecules such as those found in the QM9 dataset [19, 10, 15, 24, 11], but achieving success in generating larger molecules, as encountered in datasets like GEOM [1], remains challenging without incorporating additional techniques such as energy minimisation or docking [27].

In this paper, we focus on enhancing the ability of a diffusion model to generate plausible 3D druglike molecules. To achieve this, we use the property-conditioning method developed by Hoogeboom *et al.* [10]. Instead of conditioning a model on pre-existing properties, we condition on conformer quality, training the model to not only generate molecules, but also to distinguish high- and low-quality chemical structures.

To achieve this, we generate distorted versions of each of the three datasets we evaluate the method on: QM9, GEOM, and a subset of ZINC. We sample molecules from each dataset and apply random offsets to their original coordinates, based on a maximum distortion value. Each distorted molecule is assigned a label representing the degree of warping applied and is added back to the dataset. Non-distorted molecules are also labeled, identifying them as high-quality conformers. Using these datasets of molecules with varying levels of quality, we train property-conditioned models, encouraging the model to learn to label molecule validity while simultaneously training it to generate molecules.

First, we evaluate our conditioning method with EDM, the first E(3) equivariant diffusion model for molecule generation [10]. We then test it on two additional models: a geometry-complete diffusion model [15] and a flow matching method [21], both designed to enhance the structural plausibility of generated molecules. We also employ two datasets of druglike molecules: the GEOM dataset, and another derived from the ZINC database. This ensures the method extends beyond the scope of smaller molecules found in QM9.

Our findings demonstrate that across the models tested, conditioning a model with low-quality conformers enables it to discern between favourable and unfavourable molecular conformations. This allows us to target the area of the learned space corresponding to high-quality molecules, resulting in an improvement of the validity of generated molecules. More broadly, this demonstrates the potential of supplementing molecule generation methodologies not solely with examples of desired molecules but also with instances exemplifying undesired outcomes.

2 Methods

2.1 Generation of 3D molecules

Hoogeboom *et al.* [10] introduced the first E(3)-equivariant diffusion model (EDM) for generating 3D small molecules. Since then, significant efforts have been made to modify the original EDM, whether to adapt the method for structure-based drug design [4, 7, 12] or to enhance the validity of the generated molecules [17]. Notable examples of the latter include GCDM (Geometry-Complete Diffusion Model) [15] and EquiFM [21]. GCDM addresses the limitations of diffusion models that rely on molecule-agnostic and non-geometric graph neural networks (GNNs) for 3D graph denoising by introducing a geometry-complete approach. In contrast, EquiFM focuses on the issue of unstable probability dynamics in existing diffusion models by incorporating geometric flow matching, merging the advantages of equivariant modeling with stabilised probability dynamics.

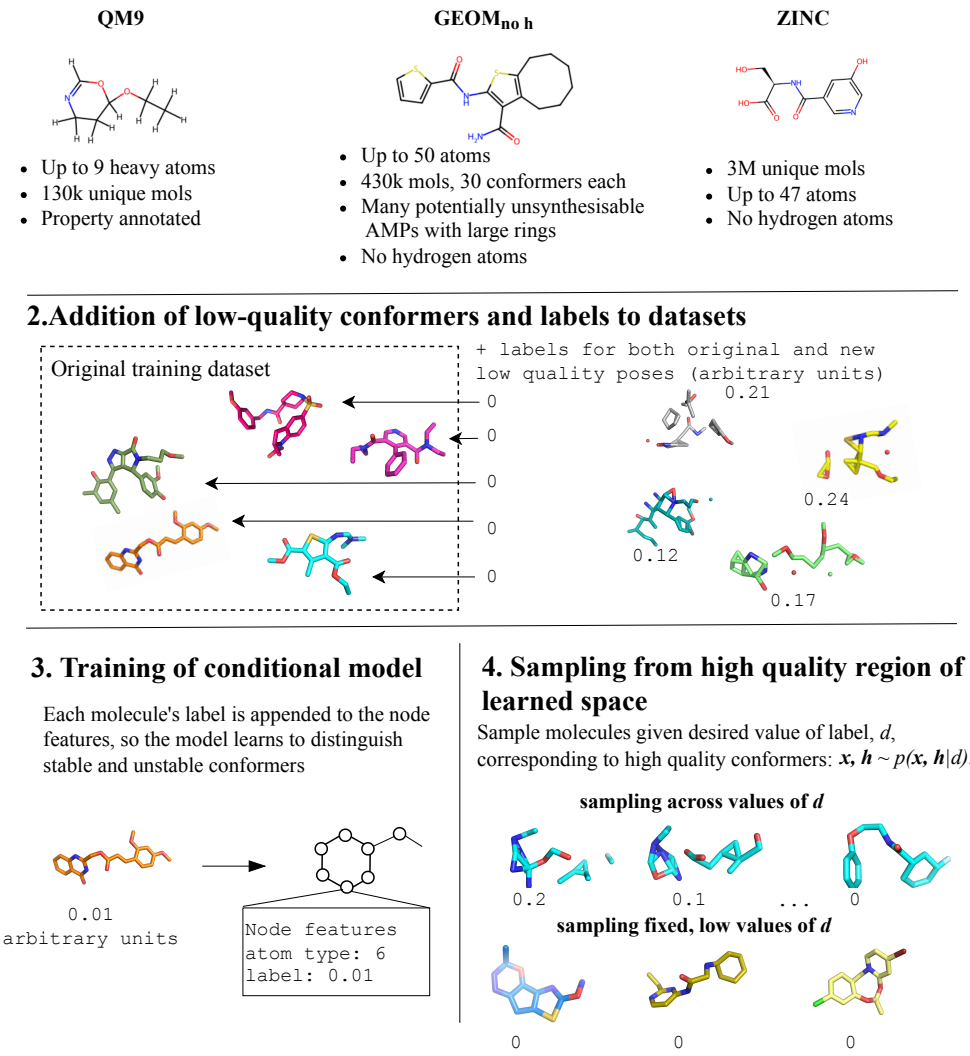

Figure 1: An overall schema of the methods used with (a) the datasets used to train both the unconditional and conditional models, (b) the generation of high energy conformers and their addition to the datasets, (c) the training of the conditional model and (d) the conditional inference

2.2 Conditioning on conformer quality

The authors of EDM developed an extension to their method to carry out conditional molecule generation. In this instance, property annotations are included alongside each of the molecules in the training dataset, and at inference, molecules can be generated with a desired value of this property. We use this property-conditioning method to train models conditioned on conformer quality. To implement this, we first generated datasets with 3D conformers of molecules of variable quality levels, and corresponding annotations. We generated distorted versions of a subset of molecules from each of the datasets we used. For each molecule, its 3D coordinates, represented as $C = \{(x_i, y_i, z_i)\}$ where i denotes the atom index, were obtained. Subsequently, a random number D within the range of 0 to D_{max} angstroms, labelled as the maximum distortion, was sampled:

$$D \sim U(0, D_{max})$$

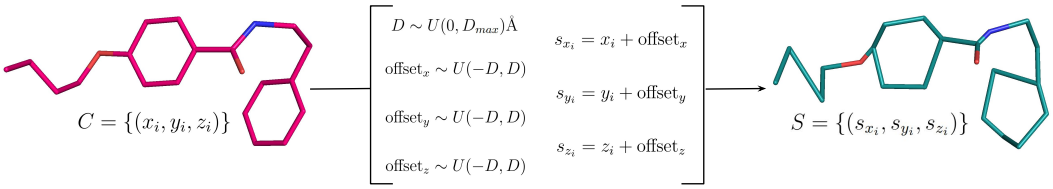

Figure 2: Diagram depicting the process of coordinate distortion for a molecule in three-dimensional space. The process involves the following steps: first, sampling the maximum distortion (*D*) from a uniform distribution between 0 and D_{max} Angstrom. Second, generating random offsets within the range of [-*D*, *D*] for each dimension of the original coordinates, C. Third, applying these offsets to the original coordinates of each atom in the molecule, resulting in a distorted conformer, S.

This value represents the maximum distance in angstrom that could be added to atoms in that molecule: in other words, the sampled distortion value determines the maximum extent of perturbation to be applied to the molecule's structure. Following this, random offsets were generated within the range of 0 to the sampled distortion, D, for each dimension of every atom's coordinates:

$$\text{offset}_x, \text{offset}_y, \text{offset}_z \sim U(-D, D)$$

These offsets were then applied to the original coordinates:

$$s_{x_i} = x_i + \text{offset}_x;\ s_{y_i} = y_i + \text{offset}_y;\ s_{z_i} = z_i + \text{offset}_z$$

Resulting in a 'distorted' version of the molecule. This distorted molecule, along with its corresponding sampled distortion value D, was subsequently added to the training set. Following the generation of the distorted datasets, we use the property-conditioning training protocol outlined by Hoogeboom *et al.* to train on them, using the distortion factor D as the property of interest, and follow the sampling protocol to generate molecules corresponding to D = 0Å.

2.3 Assessment metrics

For each trained model, we generated 100 molecules. The generated molecules were passed through RDKit, resulting in an RDKit sanitisation pass rate. All molecules were then passed through PoseBusters; however, the first step of the PoseBusters pipeline is to sanitise all molecules, so molecules that fail this automatically fail all subsequent tests. We report the number of molecules that pass RDKit sanitisation as well as all 7 non RDKit sanitisation PoseBusters tests. Finally, we also calculate an internal diversity score with MOSES[18].

2.4 Datasets

2.4.1 QM9

The QM9 dataset [19] is a widely used benchmark dataset in quantum chemistry and machine learning research. It consists of quantum-mechanical properties and 3D conformers of 130,000 small organic molecules with an average of 17.5 atoms (8.2 heavy atoms).

The QM9 dataset has been extensively used to develop and validate machine learning models for molecular property prediction. However, it has also recently become the central benchmark for *de novo* molecule generation, particularly in the development of diffusion models [10, 15].

2.4.2 GEOM

While QM9 features only smaller-than-druglike molecules, GEOM [1] is a larger-scale dataset of molecular conformers. It features 430,000 molecules, of which 317,928 are mid-sized organic molecules from AICures and MoleculeNet [26], and 133,258 molecules are from QM9, resulting in an average molecule size of 44.4 atoms (20.1 heavy atoms). For each molecule, a variable number of conformers are given along with their approximate internal energy as calculated with XTB [2]. From

Dataset	Pass rate	
	RDKit	PoseBusters
QM9	91%	81%
$\text{GEOM}_{\text{no h}}$	90%	70%
ZINC	64%	40%

Table 1: Performance comparison of the unconditional EDM across diverse molecular datasets, including QM9 (comprising small molecules), $\text{GEOM}_{\text{no h}}$ (representing larger compounds) , and a subset of the ZINC database (encompassing highly drug-like molecules).

this dataset, Hoogeboom *et al.* [10] retain the 30 lowest energy conformations for each molecule in their work.

Similar to Peng et al. [17], we use a version of GEOM from which hydrogen have been removed ($\text{GEOM}_{\text{no h}}$), as the positions of hydrogen atoms can often be inferred with a high level of confidence [9]. This not only reduces the computational demand of training, but also facilitates more effective learning of heavy atom placements. This leads to the $\text{GEOM}_{\text{no h}}$ dataset becoming the quickest to train on among the three druglike datasets. We therefore use the $\text{GEOM}_{\text{no h}}$ dataset for conducting ablation tests.

2.4.3 ZINC

ZINC [23] is a database of commercially-available compounds containing over 230 million purchasable compounds in ready-to-dock, 3D formats.

We generate a training set by selecting a subset of 660,000 molecules from the druglike catalog of the ZINC database. Unlike GEOM, this subset is curated without repeat conformers. Hydrogen atoms are not included, and the average molecule comprises 26.8 heavy atoms.

3 Results and discussion

In this section, we evaluate the performance of EDM, both conditional and non-conditional, on QM9, $\text{GEOM}_{\text{no h}}$ and ZINC. We begin by training the non-conditional model on all datasets and evaluating them as outlined above. We next performed a series of ablation tests on the $\text{GEOM}_{\text{no h}}$ dataset. These tests are used to identify a sensible ratio and distortion level of the distorted molecules. Additionally, we verify that we can sample from both the high-quality and low-quality areas of the learned space to confirm there is a discernible difference.

Subsequently, using the optimal ratio and distortion level identified from the ablation tests, we train conditional models for all three datasets: QM9, $\text{GEOM}_{\text{no h}}$, and our ZINC subset. Finally, we assess the broader applicability of the quality conditioning method by assessing it with GCDM and EquiFM.

3.1 Performance with no conditioning

First, we assess the performance of EDM when trained without conditioning on all three datasets. For QM9, we use the pretrained model provided by Hoogeboom *et al*. Table 1 shows that the QM9 dataset—comprised of smaller-than-druglike molecules—has the highest baseline model performance, with RDKit and PoseBusters pass rates of 91% and 81%, respectively (see SI for a full breakdown of the results). The baseline model also has a high performance when trained on the $\text{GEOM}_{\text{no h}}$ dataset, with pass rates of 90% and 70%.

The non-conditioned model trained on the ZINC subset exhibits a much lower RDKit sanitisation pass rate of 64%, and relatively poor PoseBusters pass rate of 40%. Unlike EDM trained on $\text{GEOM}_{\text{no h}}$, which mostly faced connectivity issues, the most common failures of the molecules from the ZINC subset trained model are in bond lengths and bond angles, passing tests for only 44% and 48% of occurrences, respectively (see SI for a full breakdown of the results).

This decrease in performance compared to $\text{GEOM}_{\text{no h}}$ could be due to the increased diversity of the dataset: our ZINC dataset does not include repeat conformers of the same molecule, whereas $\text{GEOM}_{\text{no h}}$ includes up to 30 conformers of each unique molecule. The model might therefore allocate

more attention to learning atom types—designing the actual molecule—as opposed to improving its capacity for 3D conformer generation. However, a potentially more likely explanation is the actual molecules that make up the $\text{GEOM}_{\text{no h}}$ dataset. The ZINC subset comprises entirely medium-sized compounds, whereas $\text{GEOM}_{\text{no h}}$, in addition to containing some medium-sized molecules, also incorporates the entirety of QM9, and as Table 4 shows, EDM has a high performance using the QM9 dataset. The inclusion of QM9 in $\text{GEOM}_{\text{no h}}$ also means that the average molecule size is smaller than that of ZINC subset, which may also give EDM an advantage on $\text{GEOM}_{\text{no h}}$ type molecules compared to the larger molecules in ZINC.

Overall, our findings indicate that while the baseline, non-conditional EDM model demonstrates proficiency in generating small to medium-sized compounds, its performance deteriorates on a dataset comprising of exclusively medium-sized molecules, resulting in the generation of numerous molecules with physically implausible bond lengths and angles. We next explore steering the model away from generating physically implausible molecules via our conditioning method.

3.2 Ablation tests

To identify the optimal proportion of distorted molecules and the required degree of distortion for effective conditional training, we performed ablation studies using the $\text{GEOM}_{\text{no h}}$ dataset. This dataset was selected due to its inclusion of drug-like molecule sizes, unlike QM9, whilst being a more computationally tractable set to train than the ZINC dataset.

We introduced varying numbers of distorted molecules at different distortion levels (ranging from 0Å, indicating no distortion, to the maximum distortion, D_{max}Å) into the original $\text{GEOM}_{\text{no h}}$ dataset. We defined dataset ratios based on the number of distorted and original molecules: for example, a 1:50 ratio indicates one distorted molecule was added for every fifty original molecules. We evaluated each model's performance by training conditioned models and sampling 100 molecules, ensuring that the samples were from the low-distortion-factor region of the learned space (formally, enforcing $D = 0$Å).

The model trained on a dataset with a ratio of 1:50 distorted molecules and a maximum distortion of 0.25Å exhibited the joint highest RDKit parsability rate of 97%, and the highest PoseBusters pass rate at 81%. While several models reached 97% RDKit sanitisation rates (namely 1:20, $D_{max} = 0.5$Å and 1:50, $D_{max} = 0.5$Å), these models exhibited slightly lower PoseBusters pass rates (75% and 78%, respectively). Increasing or decreasing D_{max} further resulted in PoseBusters performance decreasing across all ratios, primarily due to failures in the internal energy test.

This observation suggests that if the training includes molecules that are too distorted, the model does not effectively learn to distinguish between subtly flawed and acceptable molecular structures. Distorted molecules should therefore still bear some resemblance to realistic conformers, albeit with deliberately infeasible bond lengths and angles. On the other hand, insufficient distortion compromises the effectiveness of the conditioning classifier, and the models struggle to distinguish between high-quality and low-quality conformations, leading to poor performance in generating desirable molecules.

These results demonstrate the concept of conditioned training on negative data, and give an idea of the extent of distortion and frequency of distorted molecules to add. We used a ratio of 1:50, and $D_{max} = 0.25$Å for all subsequent tests, but note that any dataset would likely benefit from different exact values of these parameters.

We also examined the quality of molecules generated when sampling from the low-quality region of the learned space (formally, $D = D_{max}$Å). The molecules sampled using D= D_{max}Å are, as expected, worse than both the conditioned models and the baseline model in terms of PoseBusters pass rates, with the highest reaching only 53%. This poor performance is mainly attributed to failures in the internal energy test.

The RDKit parsability rates of certain models' molecules (specifically 1:20, $D = 0.1$Å and 1:50, $D = 0.1$Å) surpass the baseline model. This observation underscores the importance of incorporating comprehensive evaluations, such as those encompassed by the PoseBusters suite, in the assessment of generative models.

Maximum distortion permitted, D_{max} (Å)	Ratio of distorted:non-distorted molecules					
	1:20		1:50		1:100	
	Pass Rate		Pass Rate		Pass Rate	
	RDKit	PoseBusters	RDKit	PoseBusters	RDKit	PoseBusters
0.1	96%	73%	96%	77%	96%	77%
0.25	95%	52%	**97%**	**81%**	96%	77%
0.5	97%	75%	97%	78%	95%	68%
1	93%	57%	89%	54%	62%	8%

Table 2: Performance comparison of EDM trained conditionally on $\text{GEOM}_{\text{no h}}$ using a distortion factor, D, and sampled with D=0Å across various ratios of distorted:non-distorted molecules and maximum distortion values in angstrom.

Maximum distortion permitted, D_{max} (Å)	Ratio of distorted:non-distorted molecules					
	1:20		1:50		1:100	
	Pass Rate		Pass Rate		Pass Rate	
	RDKit	PoseBusters	RDKit	PoseBusters	RDKit	PoseBusters
0.1	91%	46%	**96%**	**53%**	81%	26%
0.25	81%	2%	81%	2%	72%	4%
0.5	49%	0%	28%	0%	38%	0%
1	41%	0%	46%	0%	29%	0%

Table 3: Performance comparison of EDM trained conditionally on $\text{GEOM}_{\text{no h}}$ using a distortion factor, D, and sampled with D=D_{max}Å across various ratios of distorted:non-distorted molecules and maximum distortion values in angstrom.

Having established the parameters to use for distorted molecules—both in terms of quality and extent of distortion—that should be included in a dataset to conditionally train EDM, and shown that we can conditionally sample from the high and low-quality areas of the learned space, we move on to applying this method to other datasets.

3.3 Conditioning on distortion factor

Dataset		Pass Rate	
		RDKit	PoseBusters
QM9	baseline	**91%**	**81%**
	conditioned	73%	53%
$\text{GEOM}_{\text{no h}}$	baseline	90%	70%
	conditioned	**97%**	**81%**
ZINC	baseline	64%	40%
	conditioned	**90%**	**63%**

Table 4: Performance comparison of EDM trained on diverse molecular datasets using the baseline model with no conditioning, and EDM conditionally trained on distortion factor using a dataset generated with D_{max} = 0.25Å and 1:50 distorted molecules, and sampled with D=0Å. The highest performance for each dataset is shown in bold.

Building upon the insights gained from the ablation tests, we generated conditional versions of each dataset by using the parameters D_{max} = 0.25Å and a ratio of distorted:non distorted molecules of 1:50. We trained conditional models on each of these modified datasets, sampled 100 molecules from each, and evaluated their quality using RDKit and PoseBusters. The results of this evaluation are presented in table 4.

When using the QM9 dataset, the highest-performing molecules on both RDKit and PoseBusters tests were generated using the non-conditioned EDM model. Conditional training resulted in a relatively uniform decrease in performance across all PoseBusters tests. This may be attributed to the fact that EDM was specifically developed to perform effectively with QM9. Further, the molecules in this dataset are small (smaller than 9 heavy atoms), which our earlier results suggest facilitates the

model's ability to distinguish between high-quality and low-quality conformers without needing examples of the latter.

The models trained conditionally on $\text{GEOM}_{\text{no h}}$ and ZINC both generated molecules with improved RDKit sanitisation rates and PoseBusters scores. For $\text{GEOM}_{\text{no h}}$, as discussed in the ablation tests (see SI), the biggest issues in the baseline were bond angles (80%), followed by bond lengths and internal energy (86%). Conditioning improved performance across all these metrics, reaching 94%, 95%, and 90%, respectively.

Training conditionally with ZINC also resulted in significant improvements across the PoseBusters tests: the lowest pass rates for both the baseline and conditioned models were in bond length/angles, steric clash, and internal energy. The most notable increase in performance for the conditioned model was the improved pass rate for atom connectivity, with this score jumping from 59% to 96%.

Having demonstrated that the conditioning method enhances the structural plausibility of generated molecules when EDM is trained on ZINC or $\text{GEOM}_{\text{no h}}$, we extend our investigation to determine whether this improvement holds for other models.

3.4 Testing the Conditioning Method on Additional Models

To evaluate the broader applicability of our method, we apply it to two other models: GCDM [15] and EquiFM [21]. The performance of these models when trained on $\text{GEOM}_{\text{no h}}$ and ZINC is presented in Tables 5a and 5b, respectively.

		RDKit	PoseBusters
$\text{GEOM}_{\text{no h}}$	baseline	**100%**	83%
	conditioned	94%	**89%**
ZINC	baseline	78%	64%
	conditioned	**100%**	**78%**

(a) Performance of GCDM

		RDKit	PoseBusters
$\text{GEOM}_{\text{no h}}$	baseline	**97%**	**57%**
	conditioned	95%	44%
ZINC	baseline	76%	39%
	conditioned	**96%**	**87%**

(b) Performance of EquiFM

Table 5: Performance comparison of GCDM and EquiFM when trained on $\text{GEOM}_{\text{no h}}$ and our ZINC subset using the default setup (baseline) or conditionally trained on distortion factor using a dataset generated with D_{max} = 0.25Å and 1:50 distorted molecules, and sampled with D=0Å. The highest performance for each dataset is shown in bold.

For both datasets, the molecules generated by GCDM trained with conditioning outperform those produced by the baseline model as assessed with PoseBusters. The improvement observed with the conditioned model is even more pronounced than with the original EDM. This aligns with findings from the GCDM paper, which found that their model generates not only more stable molecules but also more property-specific ones when property conditioning is applied. This suggests that GCDM is more effective at distinguishing between different property values, and in this case, between high- and low-quality regions of the learned space. Training EquiFM using the conditional method does not improve the plausibility of generated molecules when using $\text{GEOM}_{\text{no h}}$, in which many molecules suffer from connectivity issues (see SI for full breakdown of results). It does, however, improve the plausability of generated molecules when using the ZINC dataset, by a margin similar to that shown by EDM.

In conclusion, our conditioning method that was developed and tested with EDM is able to, without modification, enhance molecular plausibility across different models when looking at $\text{GEOM}_{\text{no h}}$ and ZINC, with GCDM showing particularly notable improvements. These results suggest that the conditioning approach is broadly applicable and beneficial.

4 Conclusions

In this work, we have demonstrated the effectiveness of including low-quality conformers in a training set and conditioning a diffusion model on a label representing conformer quality to enhance the generation of high-quality druglike molecules. By leveraging datasets derived from GEOM and ZINC, alongside a conditioning method proposed by Hoogeboom *et al.*, we have successfully improved the

validity of generated molecules. Our approach, which focuses on sampling molecules with labels corresponding to low distortion factors, leads to enhancements in RDKit parsability and validity as assessed by PoseBusters for the original EDM, as well as for a subsequent diffusion model, GCDM, and a flow-matching model, EquiFM.

Our findings underscore the importance of considering the quality of conformers in molecule generation processes. The results show that by training models to discern between favorable and unfavorable molecular conformations, we can selectively sample from the high-quality region of learned space, resulting in significant improvements in the validity of generated molecules.

Moving forward, further research could explore additional conditioning methods and datasets to continue improving the quality and diversity of generated molecules. Additionally, investigating the applicability of our approach to other areas of molecular design and exploration could yield valuable insights for drug discovery and beyond. Overall, our study provides a promising avenue for generating valid drug-sized molecules efficiently and effectively.

References

[1] Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, April 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01288-4. URL https://www.nature.com/articles/s41597-022-01288-4. Publisher: Nature Publishing Group.

[2] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science*, 11(2):e1493, 2021. ISSN 1759-0884. doi: 10.1002/wcms.1493. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1493. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1493.

[3] Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. ISSN 2041-6520, 2041-6539. doi: 10.1039/D3SC04185A. URL http://xlink.rsc.org/?DOI=D3SC04185A.

[4] Ziqi Chen, Bo Peng, Srinivasan Parthasarathy, and Xia Ning. Shape-conditioned 3D Molecule Generation via Equivariant Diffusion Models, October 2023. URL http://arxiv.org/abs/2308.11890. arXiv:2308.11890 [cs, q-bio].

[5] Vijil Chenthamarakshan, Payel Das, Samuel Hoffman, Hendrik Strobelt, Inkit Padhi, Kar Wai Lim, Benjamin Hoover, Matteo Manica, Jannis Born, Teodoro Laino, and Aleksandra Mojsilovic. CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 4320–4332. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2d16ad1968844a4300e9a490588ff9f8-Abstract.html.

[6] Orion Dollar, Nisarg Joshi, David A. C. Beck, and Jim Pfaendtner. Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24):8362–8372, 2021. doi: 10.1039/D1SC01050F. URL https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc01050f. Publisher: Royal Society of Chemistry.

[7] Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design. In *Proceedings of the 40th International Conference on Machine Learning*, pages 11827–11846. PMLR, July 2023. URL https://proceedings.mlr.press/v202/guan23a.html. ISSN: 2640-3498.

[8] Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Lio, and Tom Blundell. Benchmarking Generated Poses: How Rational is Structure-based Drug Design with Generative Models?, August 2023. URL http://arxiv.org/abs/2308.07413. arXiv:2308.07413 [q-bio].

[9] Xibing He, Elizabeth Hatcher, Lars Eriksson, Göran Widmalm, and Alexander D. MacKerell. Bifurcated Hydrogen Bonding and Asymmetric Fluctuations in a Carbohydrate Crystal Studied via X-ray Crystallography and Computational analysis. *The journal of physical chemistry. B*, 117(25):7546–7553, June 2013. ISSN 1520-6106. doi: 10.1021/jp403719g. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771504/.

[10] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D, June 2022. URL http://arxiv.org/abs/2203.17003. arXiv:2203.17003 [cs, q-bio, stat].

[11] Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. MDM: Molecular Diffusion Model for 3D Molecule Generation, September 2022. URL http://arxiv.org/abs/2209.05710. arXiv:2209.05710 [cs, q-bio].

[12] Haitao Lin, Yufei Huang, Meng Liu, Xuanjing Li, Shuiwang Ji, and Stan Z. Li. DiffBP: Generative Diffusion of 3D Molecules for Target Protein Binding, December 2022. URL http://arxiv.org/abs/2211.11214. arXiv:2211.11214 [cs, q-bio].

[13] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. Constrained Graph Variational Autoencoders for Molecule Design, March 2019. URL http://arxiv.org/abs/1805.09076. arXiv:1805.09076 [cs, stat].

[14] Youzhi Luo and Shuiwang Ji. An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch. 2022.

[15] Alex Morehead and Jianlin Cheng. Geometry-Complete Diffusion for 3D Molecule Generation and Optimization, June 2023. URL http://arxiv.org/abs/2302.04313. arXiv:2302.04313 [cs, q-bio, stat].

[16] Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets, May 2022. URL http://arxiv.org/abs/2205.07249. arXiv:2205.07249 [cs, q-bio].

[17] Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation, May 2023. URL http://arxiv.org/abs/2305.07508. arXiv:2305.07508 [cs, q-bio].

[18] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 11, December 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.565644. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.565644/full. Publisher: Frontiers.

[19] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1): 140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL https://www.nature.com/articles/sdata201422. Number: 1 Publisher: Nature Publishing Group.

[20] Jerret Ross, Brian Belgodere, Samuel C. Hoffman, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das. GP-MoLFormer: A Foundation Model For Molecular Generation, April 2024. URL http://arxiv.org/abs/2405.04912. arXiv:2405.04912 [physics, q-bio].

[21] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models, March 2023. URL http://arxiv.org/abs/2303.01469. arXiv:2303.01469 [cs, stat].

[22] Jacob O. Spiegel and Jacob D. Durrant. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of Cheminformatics*, 12(1):25, April 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00429-4. URL https://doi.org/10.1186/s13321-020-00429-4.

[23] Teague Sterling and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, November 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00559. URL https://doi.org/10.1021/acs.jcim.5b00559. Publisher: American Chemical Society.

[24] Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation. In Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 560–576, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43415-0. doi: 10.1007/978-3-031-43415-0_33.

[25] Olivier J. Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844, March 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.1166. URL https://jamanetwork.com/journals/jama/fullarticle/2762311.

[26] 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. ISSN 2041-6520, 2041-6539. doi: 10.1039/C7SC02664A. URL https://xlink.rsc.org/?DOI=C7SC02664A.

[27] Yael Ziv, Brian Marsden, and Charlotte M. Deane. MolSnapper: Conditioning Diffusion for Structure Based Drug Design, March 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.03.28.586278.

A Supplementary Information

A.1 Conditioning on internal energy

The eXtended Tight Binding (XTB) program [2] is a computational chemistry software package used for molecular modeling and simulations. It is based on the semi-empirical tight-binding approach, which approximates the electronic structure of molecules using a simplified set of parameters derived from quantum mechanics. XTB extends traditional tight-binding methods by incorporating additional empirical corrections to improve the accuracy of calculated properties. It is well-suited for studying large molecular systems where the computational cost of more accurate methods such as density functional theory (DFT) becomes prohibitive.

We use the default implementation of XTB to perform singlepoint energy calculations for each of the molecules, both distorted and original, for QM9, $\text{GEOM}_{\text{no h}}$, and our ZINC dataset. For the original GEOM dataset, the original conformers have energy annotations calculated with the same method, so we only carry out this calculation for the distorted molecules. We carry out this process using two different annotation types: a 'distortion factor', a quantity that represents the extent to which the coordinates have been altered, and an internal energy value, obtained by scoring both original and distorted conformers with the extended tight binding program (XTB) [2].

We've observed that for medium-sized, drug-like compounds, conditioning on a distance-based distortion factor results in improvements for RDKit sanitisation and PoseBusters tests. To investigate whether using a potentially more meaningful label—specifically, an internal energy value obtained using XTB—improves the conditioned models, we followed the previous distortion process. For one in every fifty molecules in each dataset, we generated a distorted version by distorting each atom's coordinates by up to 0.25 Å and added these distorted molecules back to the dataset. These distorted molecules were then passed through an XTB single-point energy calculation. The same calculation was applied to all high-quality, non-distorted molecules in all datasets, excluding the original GEOM dataset, which already has energy annotations calculated with XTB. We then trained each of the models conditionally, and once trained, we sampled 100 molecules from each, this time enforcing D = E_{min}, where E_{min} is a fixed value for each dataset corresponding to the lowest internal energy annotation of any molecule in it. Then we once again assessed all 100 molecules using RDKit and PoseBusters (Table 6).

As observed when conditioning on distortion factor, both QM9 and the original GEOM dataset generated lower quality molecules when conditioning with internal energy was carried out. QM9 saw a further decrease in performance when using internal energy, while molecules generated by the model trained on GEOM saw a slight boost in RDKit performance but still exhibited a PoseBusters pass rate of 0%, ultimately being outperformed by the baseline molecules.

Conversely, conditionally training EDM on $\text{GEOM}_{\text{no h}}$ with internal energy resulted in increased performance compared to both the baseline and the conditioned model using the distance distortion factor. The RDKit and PoseBusters pass rates reached 98% and 84%, respectively. The model trained on ZINC, however, exhibited a decrease in both RDKit and PoseBusters pass rate. These pass rates were also lower than those of the molecules generated with the ZINC model trained conditionally on distortion factor.

When using the distance based distortion factor, we enforced D=0Å when sampling, as this represented molecules that had not undergone any distortion. The distortion factor, ranging from 0Å to D_{max}Å, provides a straightforward measure of how much the structure of a molecule has been altered. In contrast, internal energy values, although physically meaningful, are challenging to compare directly between different molecules. Each molecule can have a low internal energy corresponding to a high-quality conformer and a high one corresponding to a low-quality conformer, but the lowest energy conformers of some molecules may still have higher energy values than the highest energy annotations of others. This inconsistency likely contributed to the observed performance decreases when conditioning on internal energy.

A.2 Full PoseBusters outputs

Below we provide the full outputs for each model's molecules when assessed with PoseBusters.

Dataset		RDKit	PoseBusters
QM9	baseline	**91**%	**81**%
	XTB	57%	19%
GEOM	baseline	**82**%	**24**%
	XTB	34%	0%
$\text{GEOM}_{\text{no h}}$	baseline	90%	70%
	XTB	**98**%	**84**%
ZINC	baseline	64%	**40**%
	XTB	**65%**	36%

Table 6: Performance comparison of EDM trained on diverse molecular datasets using a baseline (the unconditional EDM), and using a conditioned model, for which the model is trained on an XTB internal energy estimate. The highest performance for each dataset is shown in bold.

Dataset	Sanitisation	All Atoms Connected	Bond Lengths	Bond Angles	Internal Steric Clash	Aromatic Ring Flatness	Double Bond Flatness	Internal Energy	All Tests Passed	Diversity
Performance with no conditioning										
QM9	91%	100%	91%	91%	91%	91%	91%	81%	81%	0.88
GEOM	82%	44%	76%	74%	75%	82%	82%	66%	24%	0.85
$\text{GEOM}_{\text{no h}}$	90%	90%	86%	80%	91%	90%	89%	86%	70%	0.85
ZINC	64%	59%	44%	48%	49%	64%	64%	51%	40%	0.84
Ablation tests										
1:20, D_{max} = 0.1Å	96%	93%	94%	92%	96%	96%	96%	82%	73%	0.85
1:20, D_{max} = 0.25Å	95%	73%	88%	90%	95%	95%	95%	82%	52%	0.85
1:20, D_{max} = 0.5Å	97%	95%	94%	89%	97%	97%	97%	88%	75%	0.85
1:20, D_{max} = 1Å	93%	88%	85%	85%	92%	93%	93%	76%	57%	0.86
1:50, D_{max} = 0.1Å	96%	93%	94%	94%	96%	96%	96%	86%	77%	0.85
1:50, D_{max} = 0.25Å	97%	91%	95%	94%	96%	96%	96%	90%	81%	0.85
1:50, D_{max} = 0.5Å	97%	90%	94%	96%	97%	97%	97%	91%	78%	0.85
1:50, D_{max} = 1Å	89%	81%	84%	81%	89%	89%	89%	73%	54%	0.85
1:100, D_{max} = 0.1Å	96%	92%	95%	94%	96%	96%	96%	87%	77%	0.85
1:100, D_{max} = 0.25Å	96%	94%	95%	93%	96%	96%	96%	84%	77%	0.85
1:100, D_{max} = 0.5Å	95%	86%	94%	94%	95%	95%	95%	81%	68%	0.84
1:100, D_{max} = 1Å	62%	72%	42%	31%	56%	62%	62%	46%	8%	0.84
Performance with distortion factor conditioning										
QM9	73%	86%	71%	69%	73%	73%	73%	67%	53%	0.86
GEOM	10%	20%	40%	80%	80%	100%	100%	60%	0%	0.70
$\text{GEOM}_{\text{no h}}$	97%	91%	95%	94%	96%	96%	96%	90%	81%	0.85
ZINC	90%	96%	83%	87%	87%	90%	90%	75%	63%	0.84
Performance with energy conditioning										
QM9	57%	40%	56%	57%	56%	57%	57%	53%	19%	0.78
GEOM	34%	4%	0%	24%	25%	34%	34%	13%	0%	0.83
$\text{GEOM}_{\text{no h}}$	98%	96%	96%	96%	98%	98%	98%	88%	84%	0.85
ZINC	65%	21%	12%	41%	53%	65%	65%	39%	36%	0.86
GCDM										
$\text{GEOM}_{\text{no h}}$ baseline	100%	92%	100%	100%	95%	100%	100%	90%	83%	0.89
$\text{GEOM}_{\text{no h}}$ conditioned	94%	100%	89%	94%	94%	94%	94%	94%	89%	0.86
ZINC baseline	78%	70%	92%	88%	100%	100%	100%	87%	64%	0.75
ZINC conditioned	100%	100%	100%	100%	100%	100%	100%	78%	78%	0.76
EquiFM										
$\text{GEOM}_{\text{no h}}$ baseline	97%	93%	96%	95%	100%	100%	100%	91%	57%	0.85
$\text{GEOM}_{\text{no h}}$ conditioned	95%	72%	83%	82%	91%	95%	95%	77%	44%	0.83
ZINC baseline	76%	62%	61%	72%	71%	100%	100%	78%	39%	0.73
ZINC conditioned	96%	97%	94%	95%	96%	96%	96%	90%	87%	0.85