AN EVALUATION OF UNCONDITIONAL 3D MOLECULAR GENERATION METHODS

Martin Buttenschoen, Yael Ziv, Garrett M. Morris, & Charlotte M. Deane
Department of Statistics
24-29 St Giles', Oxford OX1 3LB
United Kingdom

ABSTRACT

Unconditional molecular generation is a stepping stone for conditional molecular generation, which is important in *de novo* drug design. Recent unconditional 3D molecular generation methods report saturated benchmarks, suggesting it is time to re-evaluate our benchmarks and compare the latest models. We assess five recent high-performing 3D molecular generation methods (EQGAT-diff, FlowMol, GCDM, GeoLDM, and SemlaFlow), in terms of both standard benchmarks and chemical and physical validity. Overall, the best method, SemlaFlow, has a success rate of 87% in generating valid, unique, and novel molecules without post-processing and 92.4% with post-processing.

1 INTRODUCTION

Generating drug-like molecules and their conformations is a common task in rational drug design. In drug discovery, molecules need to be generated conditionally in order to satisfy desired properties, such as having a particular shape or improving upon a lead compound. Unconditional molecule generation has been a hot topic in machine learning research in recent years and many models have been proposed for this task (Anderson et al., 2019; Satorras et al., 2021; Peng et al., 2023; Song et al., 2023; Vignac et al., 2023). The focus on unconditionally generating molecules is because it serves as a stepping stone towards the conditional tasks, as adapting an unconditional model for conditional purposes is a common approach in solving machine learning problems. For example, in image generation unconditional samplers are conditioned on a user's textual input to generate images of particular types only. In drug discovery, there are already examples of this paradigm of first building an unconditional model which is then repurposed for conditional generation (also called goal-directed or controllable generation) (Hoogeboom et al., 2022; Baillif et al., 2023; Xu et al., 2023; Morehead & Cheng, 2024; Le et al., 2024; Ziv et al., 2024).

In unconditional generation, the aim is to generate a large set of valid, unique, and novel drug-like molecules. Recent molecular generation methods report saturated benchmarks (Irwin et al., 2024). However, current testing has not included the assessment of the physical and chemical validity of the output molecules. Two years ago, Baillif et al. (2023) lamented the missing 3D assessments in the two benchmarks GuacaMol (Brown et al., 2019) and MOSES (Polykovskiy et al., 2020). To improve the validation of the molecular conformations, Baillif et al. (2023) called for, but did not implement, the use of empirical chemical knowledge to assess the validity of bond lengths, bond angles, dihedral angles, and steric clashes. These geometry-based tools were, however, recently implemented in the related area of docking by the PoseBusters tool and benchmark (Buttenschoen et al., 2024) which uses the RDKit's (Landrum et al., 2024) Distance Geometry module that is also used in the ETKDG algorithm (Riniker & Landrum, 2015) to check empirically informed upper and lower bounds on various molecular geometries, including bond lengths and bond angles. Our work is not the first assessment that uses geometry-based measures, for example, Hoogeboom et al. (2022) checked generated 3D conformations against typical bond lengths, but these metrics have not been consistently applied in 3D molecular conformation papers and benchmarks.

Here we assess five recent high-performing 3D molecular generation methods (EQGAT-diff, FlowMol, GCDM, GeoLDM, and SemlaFlow) with and without post-processing, in terms of both standard benchmarks and chemical and physical validity.

2 METHODS

2.1 MODELS

Five recent deep generative models for unconditional 3D molecular generation were tested: EQGAT-diff (Le et al., 2024), FlowMol (Dunn & Koes, 2024), GCDM (Morehead & Cheng, 2024), GeoLDM (Xu et al., 2023), and SemlaFlow (Irwin et al., 2024). Although full details for each method can be found in the original publications, we provide a short description of each:

GeoLDM (Xu et al., 2023) employs a latent diffusion framework and an autoencoder with a continuous latent space. The model was designed to specifically capture roto-translational equivariance constraints with the aim of modelling molecular geometries accurately. GeoLDM does not explicitly predict bonds.

EQGAT-diff (Le et al., 2024) leverages E(3)-equivariant diffusion processes that integrate continuous atomic positions with categorical atomic elements and bond types. The authors report that their use of time-dependent loss weighting improves training convergence, sample quality, and inference time.

Geometry-Complete Diffusion Model (GCDM) (Morehead & Cheng, 2024) incorporates geometry-aware graph neural networks into a denoising process in an attempt to capture molecular geometries effectively. A reported key feature of GCDM is its ability to account for chirality.

FlowMol (Dunn & Koes, 2024) uses a flow-matching generative modelling framework. In particular, the FlowMol model combines a continuous framework for atomic positions with discrete state spaces for atom types and bonds. According to the authors, the lightweight design of the model enables high performance on large datasets such as GEOM-Drugs.

SemlaFlow (Irwin et al., 2024) trains a scalable E(3)-equivariant message-passing model using conditional flow matching. The authors claim a novel molecular size-dependent prior that enhances generative performance, and they write that the overall model has up to 2-orders-of-magnitude shorter sampling times compared to other methods.

The five models were developed for the 'Drugs' subset of the Geometric Ensemble of Molecules (GEOM) dataset curated by Axelrod & Gómez-Bombarelli (2022). EQGAT-diff, FlowMol, and SemlaFlow use the training, validation, and test splits generated by Vignac et al. (2023) and GCDM and GeoLDM use the splits generated by Anderson et al. (2019). The model weights made available by the authors of each model were used.

2.2 ASSESSMENT

The five methods were benchmarked for their ability to generate 100,000 valid, novel, unique, drug-like molecules. Only molecules that are valid, novel, and unique are counted as successes.

Validity can be divided into the validity of the molecular graph and that of the molecular conformation. The molecular graph is checked to be chemically valid—fulfils chemical valency rules, and the molecular conformation is checked to be physically valid—has valid bond geometries and low strain energies. Formally, we say that a molecule is *valid* if its molecular graph is *chemically valid* and its conformation is *physically valid*.

Chemical validity of the molecular graph is assessed using four tests. A molecular graph is *chemically valid* if 1) the generated file can be loaded with the `MolFromMolFile` function of the RDKit (Landrum et al., 2024) with the sanitization option turned off; 2) the generated RDKit molecule object can be sanitised using the RDKit's `SanitizeMol` function; 3) the molecule has all of its hydrogens added explicitly, assuming that the molecule is not a radical; and 4) the molecule is connected, that is the generated molecular graph is connected in the mathematical sense and in the chemical sense does not have more than one component (or fragment). These four tests assess whether a molecular graph is chemically valid.

Physical validity of the molecular conformation is assessed using six tests. A molecular conformation is *physically valid* if it passes the bond lengths, bond angles, planar aromatic rings, planar double bonds, internal steric clash, and internal energy tests of the PoseBusters test suite (Buttenschoen et al., 2024). Bond lengths and bond angles are compared to experimentally determined values and violations below and above 25% are flagged. Internal steric clash is measured using

typical van der Waals radii and violations below 30% are flagged. Strain energy is measured by the ratio of the observed energy and that of an ensemble of energy-minimised generated conformations. Here, the Universal Force Field (Rappe et al., 1992) was used with a threshold ratio of 100. If all of these tests of the geometry of the conformation pass, then the molecular conformation is physically valid.

Uniqueness and novelty of the molecules is assessed using canonical SMILES strings, generated by the RDKit's `MolToSmiles` function (Landrum et al., 2024). Formally, a generated molecule is *novel* if its SMILES string does not occur in the reference set, and a molecule is *unique* in a multiset of molecules if its SMILES string occurs only once. For the novelty check, the reference set is the entire GEOM Drugs set (Axelrod & Gómez-Bombarelli, 2022), which was used for the training of all of the benchmarked methods.

Two standard metrics are computed to compare the distributions of the molecules under these metrics. Drug-likeness is estimated using the quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012), and synthetic accessibility (SA) is approximated using the SAscore (Ertl & Schuffenhauer, 2009). In addition, the appendix contains the distributions of the spacial score (Krzyzanowski et al., 2023), Crippen's logP (Wildman & Crippen, 1999), the number of heavy atoms, and the conformation strain approximated using the energy ratio from PoseBusters (Buttenschoen et al., 2024). All underlying metrics were calculated using the RDKit. The distribution of the metrics are plotted using the `kdeplot` function of the Python package `seaborn` (Waskom, 2021).

Furthermore, ECFP4 count fingerprints (Morgan, 1965) with 2048 bits are calculated using the RDKit's `GetMorganGenerator` method. The most prevalent substructures in all generated molecules and reference data sets were identified using the Sort and Slice method (Dablander et al., 2024). The count vectors were projected into two dimensions using the UMAP algorithm (McInnes et al., 2020) with the settings `metric="manhattan"`, `min_dist=1.0`, and `spread=1.0` to visualize chemical space. Furthermore, the Fréchet ChemNet distance (Preuer et al., 2018) is calculated by generating canonical SMILES for all molecules (without subsampling), filtering out invalid SMILES, and using the `get_fcd` function of the `fcd` package (version 1.2.2).

As baselines, two data sets are being used: the GEOM Drugs set, which contains 301,855 molecules, including all the training data on which the five methods were trained, and the DrugBank set containing 2,066 approved drugs. Details of the two data sets can be found in Appendix C

2.3 Post-processing

The generated molecules were post-processed in three steps. First, the largest fragment was picked using the RDKit's `LargestFragmentChooser`. Second, missing explicit hydrogens were filled in with the assumption that the molecule is not a radical. Third, the molecular conformation was refined by minimising energy using the Universal Force Field (Rappe et al., 1992) implemented in the RDKit.

3 Results

All five 3D molecular generation methods generated large sets of valid, unique, and novel molecules. For example, sampling 100,000 times, SemlaFlow generated 87.0% and FlowMol 59.7% valid, unique and novel molecules without post-processing. Table 1 shows the total share of novel, unique, and valid molecules generated by each method. Note that GCDM and GeoLDM do not add all hydrogens without post-processing. For all methods, the limiting factor is validity, as uniqueness and novelty are almost perfect, with more than 99% of the valid molecules being novel and unique. SemlaFlow was also the fastest of the methods (Table 2), generating 87,523 valid molecules in 3 hours.

However, the methods are still not as good as the data sets on which they were trained or that they were aiming to capture. Both the GEOM Drugs set, which was used for training of the methods, and the DrugBank set, which is a database of known, approved drugs, have almost perfect scores. GEOM Drugs contains 99.8% chemically valid molecular graphs and 94.2% valid molecular conformations, while the DrugBank molecules have 100% valid molecular graphs and 98.8% valid molecular con-

Table 1: Validity, uniqueness, and novelty of the generated and reference molecules. The table contains the percentage of successfully generated molecules that are valid, unique, and novel, out of 100,000; for GEOM Drugs and DrugBank the percentages are out of 301,855 and 2,066 respectively. Novelty is relative to GEOM Drugs. All methods generated large numbers of valid, unique, and novel molecules, for example, SemlaFlow generated 87,523 without post-processing and GCDM generated 95,188 with post-processing.

	% Valid	% Valid & Unique	% Valid & Unique & Novel
EQGAT-diff	59.7	59.7	59.5
FlowMol	59.8	59.8	59.7
GCDM	0.2	0.2	0.2
GeoLDM	2.9	2.9	2.9
SemlaFlow	87.5	87.4	87.0
EQGAT-diff + PP	84.2	84.2	84.0
FlowMol + PP	84.2	84.2	84.1
GCDM + PP	95.2	95.2	95.2
GeoLDM + PP	69.6	69.3	69.3
SemlaFlow + PP	93.1	92.9	92.4
GEOM Drugs	94.2	93.7	0.0
DrugBank	98.8	98.2	52.6

formations (Table 3). Given that these training data scores are higher than the best model, there is still room for improvement.

The failure modes are in terms of both the generated molecular graphs and the 3D conformations are shown in Table 3. For example, EQGAT-diff produces 62.6% chemically valid molecular graphs and 82.5% physically valid conformations, leading to an overall validity of 59.7%. For the molecular graphs, the explicit hydrogens check is the largest failure mode for EQGAT-diff, GCDM, and GeoLDM (Table 4), while for the conformational checks, almost all methods show some failures in the geometry-based as well as the energy-based tests (Table 5). Some of the violations that occur are very large (pink zones in Figures 4 and 5). For example, all methods produce some bond lengths that are 20% too long or too short. Despite the fact that the internal steric clash check allows significant overlap of van der Waals radii (30%), even the best methods—SemlaFlow and GCDM—still exhibited failures. Overall, there still appears to be potential to improve validity in terms of connectedness and geometry.

Running the post-processing steps described in the Methods mitigates these results to some degree. Picking the largest fragment, adding missing hydrogens, and minimising the conformations' energy improved the results by up to 95% depending on the method. For example, SemlaFlow's performance increased by 5.4 percentage points to 92.4% and GCDM's increased to 95.2%. The improvements of GCDM and GeoLDM are mostly due to the addition of hydrogens since these two methods do not generate all hydrogens explicitly. With post-processing, the best method according to these metrics becomes GCDM. In general, post-processing improves the ability of all methods to generate a large number of valid, unique, and novel molecules.

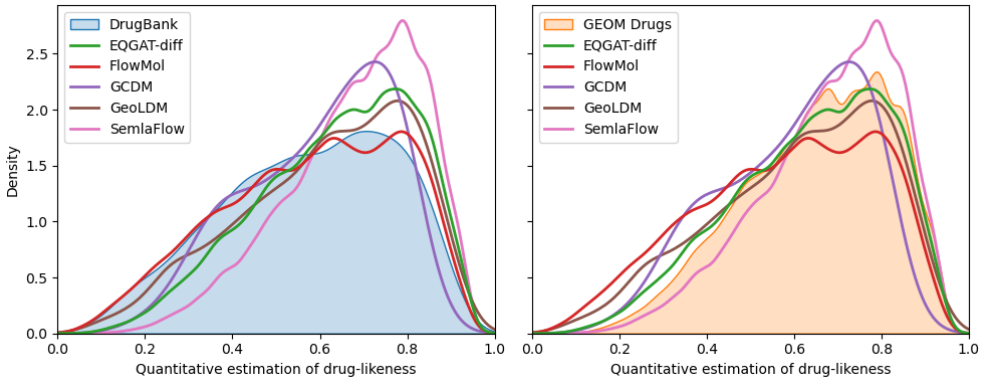
The distributions of the metrics for drug-likeness and synthetic accessibility of the generated molecules tend to those of the GEOM Drugs training data. Figure 1 shows the distributions of the QED and SAscore, which estimate drug-likeness and synthetic accessibility respectively. These plots and the additional metrics in the Appendix in Table 6 and Figures 6 and 7 show that the best methods generate molecules that adopt the same distribution under the selected measures as the training data tend to.

In terms of chemical diversity, the generated molecules of all methods tend to sit inside the molecules of the training data, but not all methods cover the space sufficiently. The UMAP projections of the molecules' fingerprints (Figure 2) show that EQGAT-diff, FlowMol, GeoLDM, and SemlaFlow cover the core of the space of GEOM Drugs but there are also outlier islands on the fringes of the DrugBank and GEOM Drugs distributions, for which the methods do not generate molecules. Also

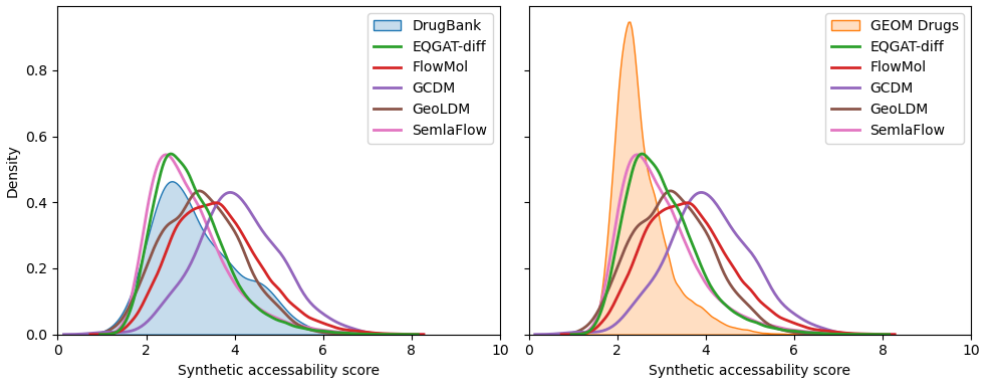
note that GCDM covers a smaller space than the other tested methods. These qualitative observations are reflected in the quantitative measurements. The calculated Fréchet ChemNet distance (Figure 3) from the GEOM Drugs training data is the smallest for the molecules generated by EQGAT-Diff (5.1) and SemlaFlow (5.1), while GCDM has the largest distance (45.2). In terms of this metric, SemlaFlow and EQGAT-diff capture the space of the training data the best.

4 Conclusion

The high observed success rates (of up to 95.2%) for generating valid, unique, and novel molecules show that these methods are already useful unconditional 3D molecular generators. The top methods are able to generate molecules with the same distribution as the training set in terms of the QED and SAscore metrics, but they appear to not fully explore the space of the training data set in terms of ECFP4-2048 count fingerprints and Fréchet ChemNet distance. Furthermore, the PoseBusters-based checks used here are still quite generous, for example allowing 30% closer overlap of the van der Waals radii of atoms. Future work should explore further improving these methods against even more stringent physical and chemical checks.



(a) Distributions of the molecules' QED. The QED is a proxy for drug-likeness. The left panel shows the DrugBank molecules in blue as a reference; the right panel shows the GEOM Drugs molecules in light orange as a reference.



(b) Distributions of the molecules' SAscore. The SAscore is a proxy for synthetic accessibility. The left panel shows the DrugBank molecules in blue as a reference; the right panel shows the GEOM Drugs molecules in light orange as a reference.

Figure 1: Distributions of the molecules in terms of drug-likeness estimated by QED and synthetic accessibility estimated by SAscore in comparison to the approved drugs in DrugBank and to the training data GEOM Drugs. All methods generate molecules that approximately sit in the same distribution under QED and SAscore as the approved drugs and the training data.

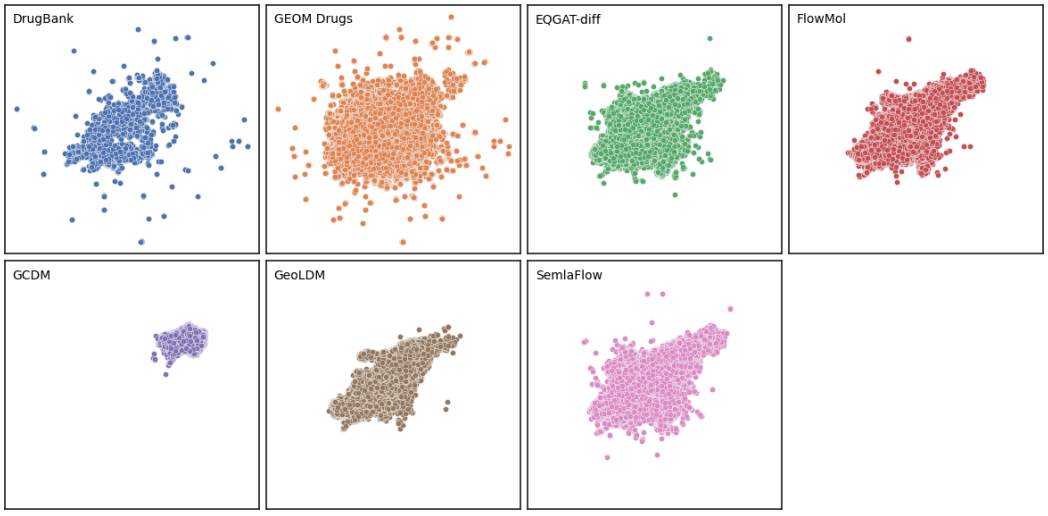


Figure 2: Visualization of the chemical space covered by the generated molecules and the reference data sets shown using the same projection for all methods. The two-dimensional map was generated from all molecules' ECFP4 2048 count fingerprints by the UMAP algorithm.

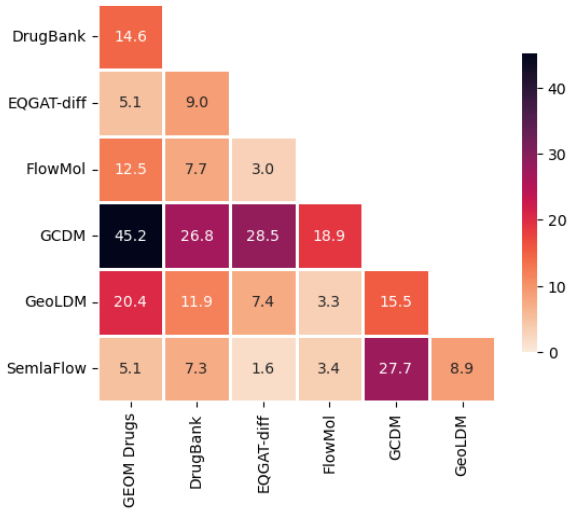


Figure 3: Fréchet ChemNet distance between the sets of generated molecules with post-processing, the training data GEOM Drugs and the approved drugs in DrugBank. SemlaFlow and EQGAT-diff capture the training data the best as they have the lowest distance (5.1) to GEOM Drugs. Note that all sets are significantly larger than the recommended data set size (5'000) to calculate this metric except DrugBank which contains 2,066 molecules.

REFERENCES

Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Deep generative models for 3D molecular structure. *Current Opinion in Structural Biology*, 80:102566, 2023.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019.

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.

Markus Dablander, Thierry Hanser, Renaud Lambiotte, and Garrett M. Morris. Sort & Slice: A simple and superior alternative to hash-based folding for extended-connectivity fingerprints. *Journal of Cheminformatics*, 16(1):135, 2024.

Ian Dunn and David R. Koes. Exploring discrete flow matching for 3D de novo molecule generation, 2024. arXiv:2411.16644.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3D molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

Adrian Krzyzanowski, Axel Pahl, Michael Grigalunas, and Herbert Waldmann. Spacial score-a comprehensive topological indicator for small-molecule complexity. *Journal of Medicinal Chemistry*, 66(18):12739–12750, 2023.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, Sereina Riniker, Peter Gedeck, Gareth Jones, Eisuke Kawashima, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Tad Hurst, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, Juuso Lehtivarjo, Guillaume Godin, Axel Pahl, François François Bérenger, and Hussein Faara. RDKit: Open-source cheminformatics. Zenodo, 2024.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3D molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020. arXiv:1802.03426.

Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3D molecule generation and optimization. *Communications Chemistry*, 7(1):150, 2024.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. MolDiff: Addressing the atom-bond inconsistency problem in 3D molecule diffusion generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27611–27629. PMLR, 2023.

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*, 11:565644, 2020.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet ChemNet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018.

A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992.

Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55 (12):2562–2574, 2015.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2021. arXiv:2102.09844.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3D molecule generation. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14170, pp. 560–576. Springer Nature Switzerland, Cham, 2023.

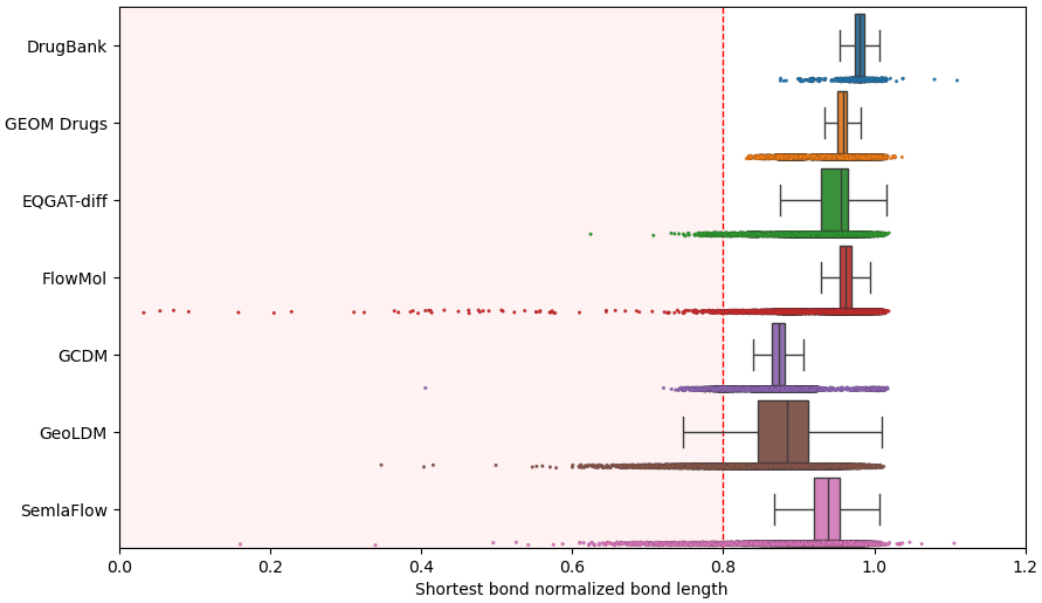
Michael L. Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6 (60):3021, 2021.

Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

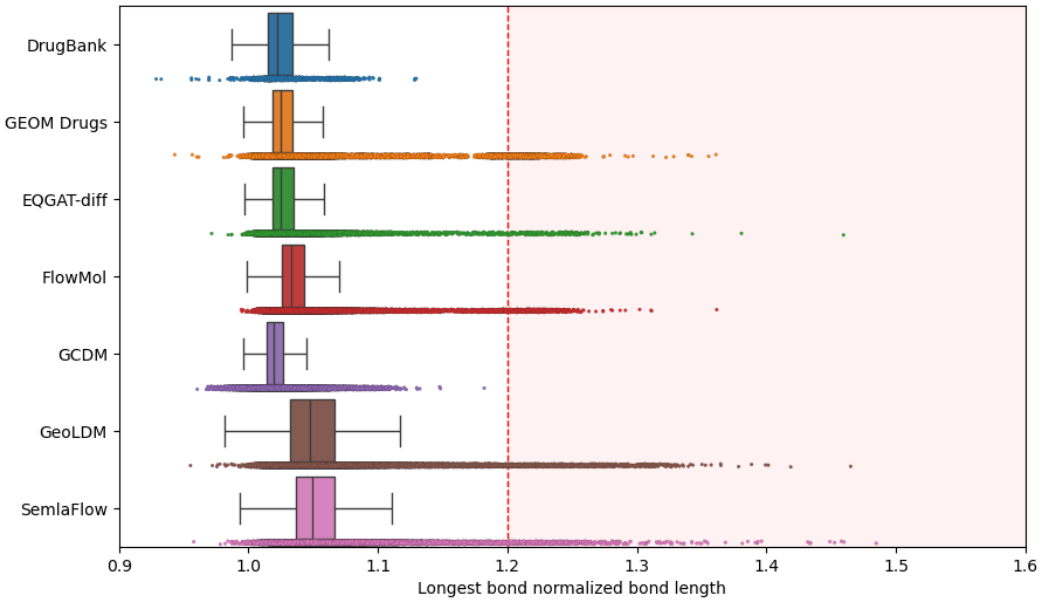
Minkai Xu, Alexander S Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38592–38610. PMLR, 2023.

Yael Ziv, Brian Marsden, and Charlotte M. Deane. MolSnapper: Conditioning diffusion for structure based drug design, 2024. bioRxiv:2024.03.28.586278.

A FURTHER RESULTS

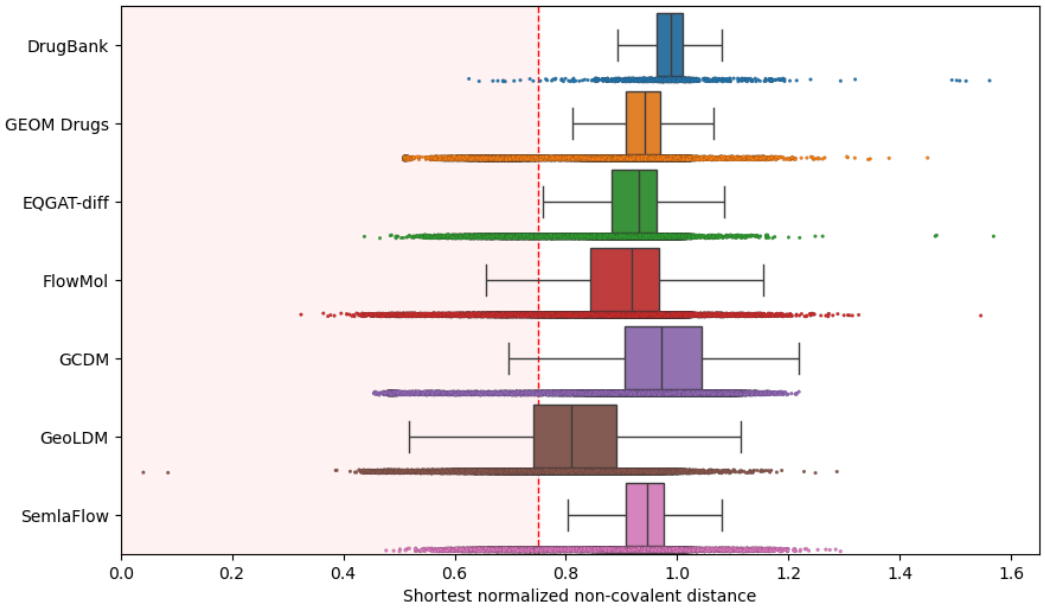


(a) Shortest normalized bond length recorded of each molecule.

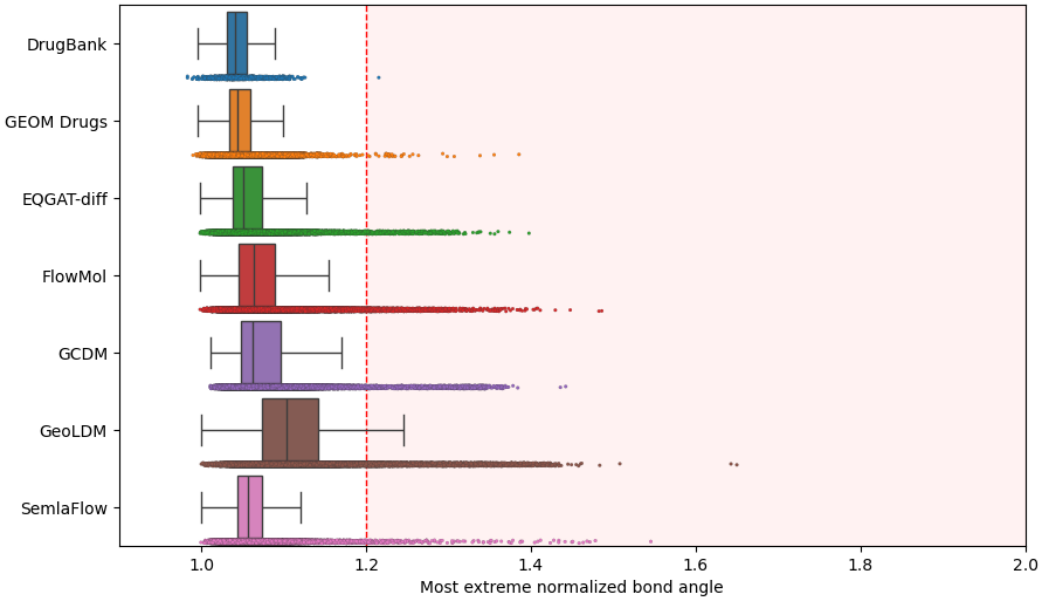


(b) Longest normalized bond length recorded of each molecule.

Figure 4: Distributions of the shortest and longest bond lengths. The values are normalized by the limits obtained from the RDKit's distance geometry module.



(a) Shortest normalized non-covalent intermolecular distances recorded in each molecule



(b) Most extreme normalized bond angle recorded of each molecule.

Figure 5: Distributions of the shortest non-covalent distances and most extreme bond angles. The values are normalized by the limits obtained from the RDKit's distance geometry module.

Table 2: Runtime (in hours) for generating 100,000 samples on a Fedora Linux server using 14 CPU cores, 100 GB RAM, and one Nvidia A100 80GB GPU. The table reports the total time required for each model to complete the generation process which is described in Appendix B.

	Runtime (h)
EQGAT-diff	123
FlowMol	6
GCDM	42
GeoLDM	176
SemlaFlow	3

Table 3: Validity of the generated and ground truth molecules. The table contains the total number of molecules that was to be generated or is contained in the data set and the proportions of molecules that pass all the chemical validity test, all physical validity tests. The last columns is the percentage of molecules that pass all the tests together.

	% Chemical	% Physical	% Valid
EQGAT-diff	62.6	82.5	59.7
FlowMol	66.4	76.2	59.8
GCDM	0.2	82.3	0.2
GeoLDM	3.2	57.0	2.9
SemlaFlow	91.3	90.8	87.5
EQGAT-diff + PP	87.1	84.2	84.2
FlowMol + PP	87.2	84.2	84.2
GCDM + PP	95.5	95.2	95.2
GeoLDM + PP	73.8	69.6	69.6
SemlaFlow + PP	94.9	93.1	93.1
GEOM Drugs	99.8	94.2	94.2
DrugBank	100.0	98.8	98.8

Table 4: Components of the chemical validity of the generated and ground truth molecules. The numbers shown are the percentages of the molecules that pass each of the tests. The last column is the percentage of molecules that pass all of the tests. The chemical tests check the molecular graph generated and they do not check the molecules' 3D conformations.

	Sanitizes	Hydrogens explicit	Connected	% Chemical
EQGAT-diff	87.1	64.4	84.4	62.6
FlowMol	86.6	81.1	68.6	66.4
GCDM	97.2	0.2	86.6	0.2
GeoLDM	72.5	5.3	37.4	3.2
SemlaFlow	94.9	93.8	92.3	91.3
EQGAT-diff + PP	87.1	87.1	87.1	87.1
FlowMol + PP	87.2	87.2	87.2	87.2
GCDM + PP	95.5	95.5	95.5	95.5
GeoLDM + PP	73.8	73.8	73.8	73.8
SemlaFlow + PP	94.9	94.9	94.9	94.9
GEOM Drugs	100.0	100.0	99.8	99.8
DrugBank	100.0	100.0	100.0	100.0

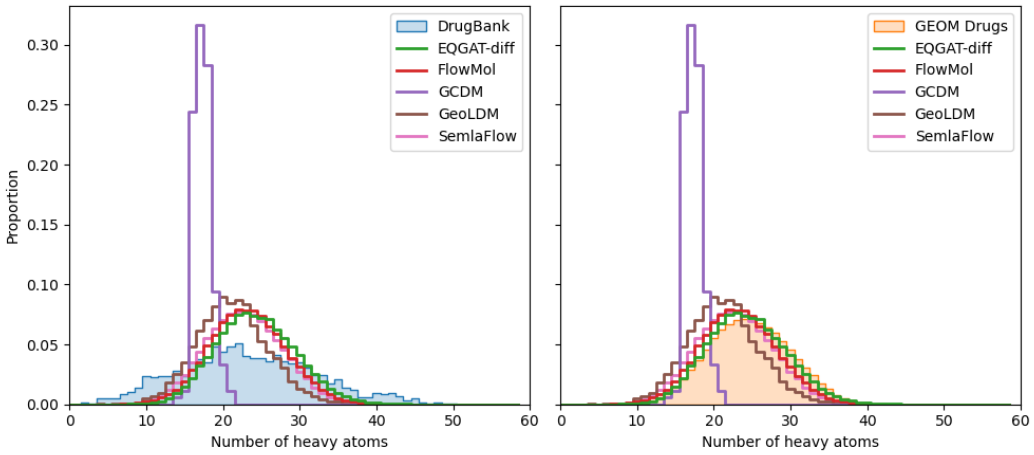
Table 5: Components of the physical validity of the generated and ground truth molecules. The numbers shown are the percentages of the molecules that pass each of the intramolecular PoseBusters tests. The last column is the percentage of molecules that pass all of the intramolecular tests. The physical tests check the 3D conformations of the generated molecules.

	Bond lengths	Bond angles	Internal steric clash	Planar aromatic rings	Planar double bonds	Internal energy	% Physical
EQGAT-diff	87.0	86.9	82.9	87.0	87.0	86.8	82.5
FlowMol	86.5	86.1	82.5	86.5	81.2	85.4	76.2
GCDM	97.2	96.4	94.8	97.2	97.2	84.2	82.3
GeoLDM	71.0	69.6	62.7	72.4	71.2	70.0	57.0
SemlaFlow	94.6	94.8	92.0	94.9	94.2	94.8	90.8
EQGAT-diff + PP	87.0	87.1	84.6	87.0	87.0	86.7	84.2
FlowMol + PP	87.2	87.2	84.9	87.2	87.1	86.8	84.2
GCDM + PP	95.5	95.5	95.5	95.5	95.5	95.2	95.2
GeoLDM + PP	73.7	73.8	70.5	73.8	73.6	73.1	69.6
SemlaFlow + PP	94.9	94.9	93.2	94.9	94.8	94.8	93.1
GEOM Drugs	100.0	100.0	94.6	100.0	99.7	100.0	94.2
DrugBank	100.0	100.0	99.7	100.0	100.0	99.1	98.8

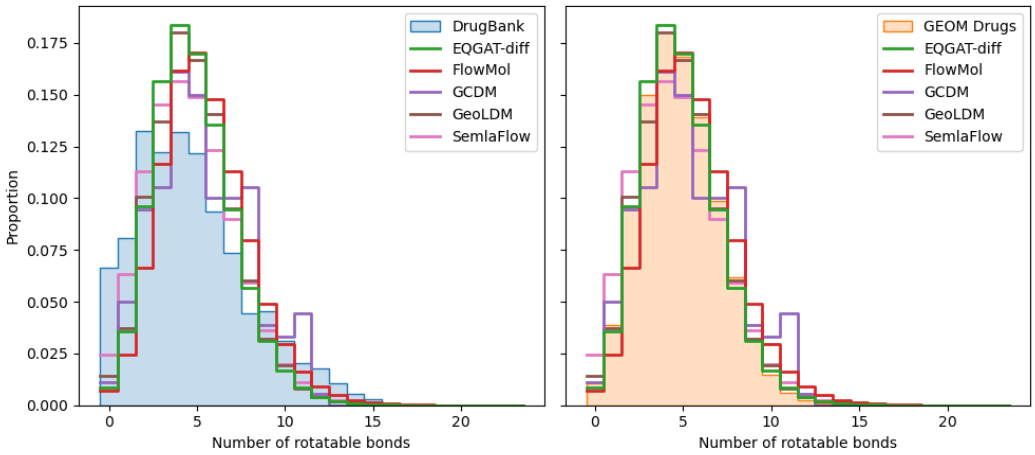
Table 6: Various molecular properties. The table contains the median and interquartile range of the SAscore, QED, Lipinski rule of five, the spacial score, molecular weight, number of heavy atoms, number of rings, and logP.

	SAscore	QED	Lipinski R5	Spacial score
EQGAT-diff	3.14 ± 1.18	0.65 ± 0.27	5.0	15.33 ± 8.59
FlowMol	3.66 ± 1.40	0.57 ± 0.34	5.0	17.65 ± 12.04
GCDM	4.99 ± 0.87	0.38 ± 0.29	5.0 ± 1.0	36.84 ± 10.55
GeoLDM	4.65 ± 1.32	0.50 ± 0.35	5.0	19.33 ± 8.82
SemlaFlow	2.82 ± 1.11	0.70 ± 0.22	5.0	15.35 ± 9.33
EQGAT-diff + PP	3.11 ± 1.22	0.65 ± 0.28	5.0	15.68 ± 9.31
FlowMol + PP	3.61 ± 1.43	0.59 ± 0.31	5.0	17.81 ± 12.81
GCDM + PP	4.96 ± 0.87	0.39 ± 0.29	5.0 ± 1.0	37.05 ± 10.61
GeoLDM + PP	4.32 ± 1.39	0.49 ± 0.33	5.0	21.88 ± 11.86
SemlaFlow + PP	2.81 ± 1.11	0.70 ± 0.22	5.0	15.38 ± 9.40
GEOM Drugs	2.39 ± 0.69	0.67 ± 0.26	5.0	13.74 ± 5.90
DrugBank	2.96 ± 1.39	0.59 ± 0.31	5.0 ± 1.0	15.69 ± 11.80

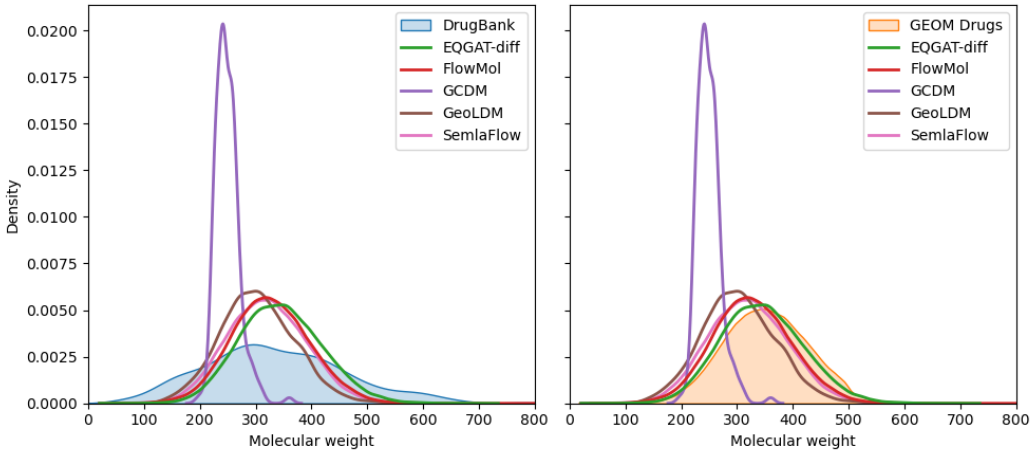
	Weight [Da]	# Heavy atoms	# Rings	LogP
EQGAT-diff	352.16 ± 104.91	25.00 ± 8.00	3.00 ± 2.00	2.55 ± 1.86
FlowMol	337.10 ± 98.07	24.00 ± 6.00	3.00 ± 1.00	1.52 ± 2.14
GCDM	354.25 ± 58.02	24.00 ± 4.00	2.00 ± 1.00	0.57 ± 2.62
GeoLDM	366.19 ± 106.01	26.00 ± 8.00	3.00 ± 2.00	1.08 ± 2.42
SemlaFlow	321.05 ± 98.01	23.00 ± 7.00	3.00 ± 1.00	2.77 ± 1.74
EQGAT-diff + PP	349.10 ± 105.00	25.00 ± 7.00	3.00 ± 2.00	2.52 ± 1.88
FlowMol + PP	316.15 ± 106.02	22.00 ± 7.00	2.00 ± 1.00	1.42 ± 2.10
GCDM + PP	348.24 ± 61.98	24.00 ± 4.00	2.00 ± 1.00	0.56 ± 2.59
GeoLDM + PP	308.03 ± 144.10	21.00 ± 10.00	2.00 ± 2.00	0.69 ± 2.34
SemlaFlow + PP	317.19 ± 98.08	23.00 ± 7.00	3.00 ± 1.00	2.74 ± 1.75
GEOM Drugs	351.23 ± 105.00	25.00 ± 7.00	3.00 ± 2.00	2.92 ± 1.64
DrugBank	324.13 ± 171.09	23.00 ± 12.00	3.00 ± 3.00	2.58 ± 2.88



(a) Number of heavy atoms.

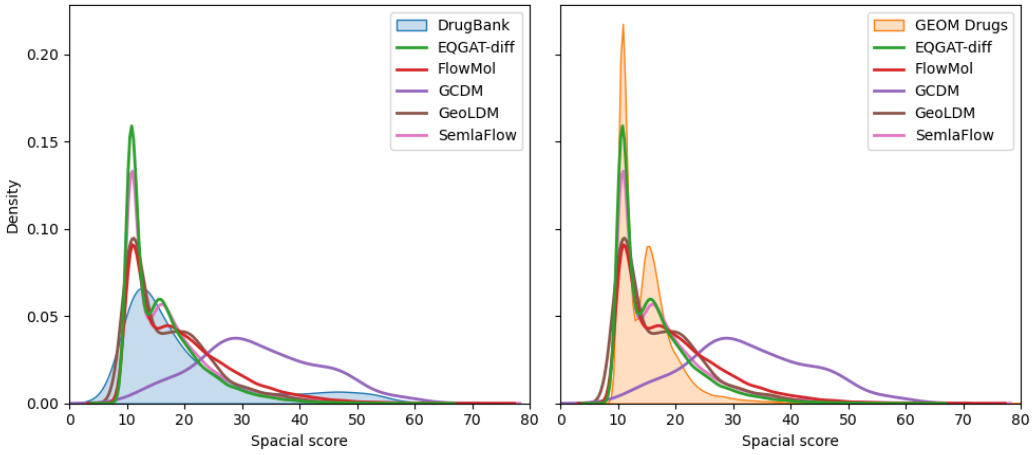


(b) Number of rotatable bonds.

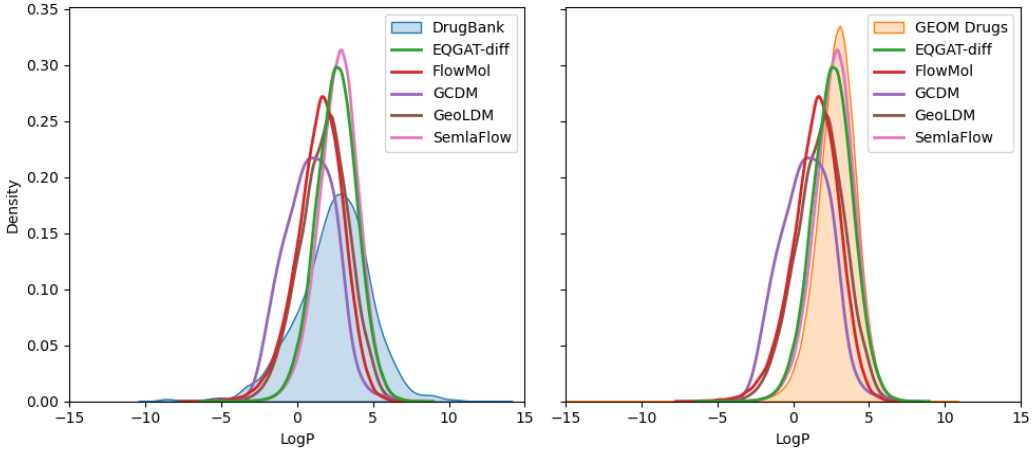


(c) Molecular weight in daltons.

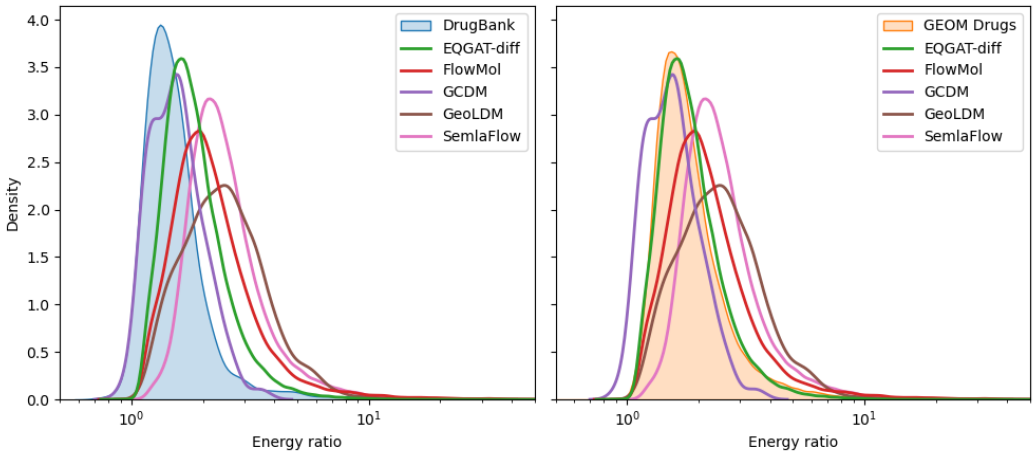
Figure 6: Distributions of the valid molecules in terms of number of heavy atoms, number of rotatable bond, and molecular weight.



(a) Molecule complexity estimated by the Spacial Score.



(b) Lipophilicity estimated by Crippen's logP.



(c) Conformational strain estimated by the energy ratio of the generated conformation relative to the average energy of an ensemble of energy minimization conformations generated with ETKDGv3. The energies were estimated using the Universal Force Field.

Figure 7: Distributions of the valid molecules in terms of synthetic accessibility, molecule complexity, and lipophilicity.

B DETAILS ON SAMPLING METHODS

This section describes how each model was used to generate 100,000 molecules.

B.1 EQGAT-DIFF

The code for EQGAT-diff with commit hash `68aea80691a8ba82e00816c82875347cbda2c2e5` was obtained from the public code repository of the authors `https://github.com/tuanle618/eqgat-diff/tree/main`. The model weights trained on the GEOM Drugs were obtained from the authors upon request.The model was run using the script `run_evaluation.py` with the setting for batch size at 100 and the dataset option to 'drugs'. All other parameters were kept at their default settings. The model was run 20 times sampling 5,000 molecules each time. The generated XYZ files were converted to SDF format and combined into a single file with Open Babel (O'Boyle et al., 2011).

B.2 FLOWMOL

The code for FlowMol with commit hash `c3503939ce409a12e558e3231b5c807f86d9fe1d` was obtained from the authors' public code repository `https://github.com/Dunni3/FlowMol`. The model weights were obtained from `https://bits.csb.pitt.edu/files/FlowMol/trained_models/` and placed in the directory as instructed by the README provided by the authors. The script `test.py` was run using `n_timesteps=250`, `nummols=100000`, and `model_dir=flowmol/trained_models/geom_ctmc` and the output was a single SDF file.

B.3 GCDM

The code for GCDM with commit hash `109d9d7625a00fb669454246fc846f348be3df0d` was obtained from the authors' public code repository `https://github.com/BioinfoMachineLearning/bio-diffusion`. The model checkpoints used were obtained from Zenodo `https://zenodo.org/record/13375913/files/GCDM_Checkpoints.tar.gz`. The model was run using script `mol_gen_sample.py`. The model was run 50 times using the seeds 123 through 173 generating 2,000 molecules each time. All other settings passed to the script were `num_nodes=44`, `all_frags=true`, `sanitize=false`, `relax=false`, `num_resamplings=1`, `jump_length=1`, and `num_timesteps=1000`. The 50 generated SDF files were concatenated using the shell command `cat`.

B.4 GEOLDM

The code for GeoLDM with commit hash `03ae2031c712a1a6c1678e747bdcdc7a7560e00b` was obtained from the authors' public code repository `https://github.com/MinkaiXu/GeoLDM`. The weights for the model trained on GEOM Drugs available in the directory were used. The model was run using the script `eval_analyze.py` with the setting for batch size at 100. The model was run 10 times sampling 10,000 molecules each time. The generated TXT files were converted to SDF format and combined into a single file with Open Babel (O'Boyle et al., 2011).

B.5 SEMLAFLOW

The code for SemlaFlow with commit hash `0c021d663f9feacbfe19e6f3527b2ad98d58ecab` was obtained from the authors' public code repository `https://github.com/rssrwn/semla-flow` The model weights for the model trained on the GEOM drugs dataset were obtained via the links in the repository. The model was run once using the 'predict' script provided by the authors and the output was a single SDF file.

C DATASETS

The Geometric Ensemble Of Molecules (GEOM) was curated by Axelrod & Gómez-Bombarelli (2022). The dataset has two subsets: the 'QM9' sub set contains 133k molecules with up to 9 heavy

(non-hydrogen) atoms and GEOM's 'Drugs' sub set contains 304k drug-like molecules with up to 91 heavy atoms and for all these molecules, at least one conformation annotated with the conformation potential energies is available. The GEOM Drugs structures used here were obtained from Irwin et al. (2024).

The DrugBank (Knox et al., 2024) contains CA approved, investigational, and withdrawn drugs and their structures. Here, DrugBank v5.1.13 released in January 2018 containing 2'185 structures of approved drugs was used. The data set used here is that of the 2'066 drugs which are provided with 3d structures and have the status 'approved'. It is released under a Creative Commons Attribution-NonCommercial 4.0 International License.