GEOM-Drugs Revisited: Toward More Chemically Accurate Benchmarks for 3D Molecule Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep generative models have shown significant promise in generating valid 3D molecular structures, with the GEOM-Drugs dataset serving as a key benchmark. However, current evaluation protocols suffer from critical flaws, including incorrect valency definitions, bugs in bond order calculations, and reliance on force fields inconsistent with the reference data. In this work, we revisit GEOM-Drugs and propose a corrected evaluation framework: we identify and fix issues in data preprocessing, construct chemically accurate valency tables, and introduce a GFN2-xTB-based geometry and energy benchmark. We retrain and re-evaluate several leading models under this framework, providing updated performance metrics and practical recommendations for future benchmarking. Our results underscore the need for chemically rigorous evaluation practices in 3D molecular generation. Our recommended evaluation methods and GEOM-Drugs processing scripts will be available at the time of the workshop.

1 Introduction

Generative models for molecules are an emerging paradigm that enables the construction of novel molecules in 2D or 3D [1, 2]. These AI models learn the patterns and distribution of existing molecular data to generate previously unseen chemical structures. By encoding molecular information into mathematical representations and then sampling from a learned distribution, these models facilitate efficient exploration of vast chemical space. The field continues to evolve rapidly and is not yet mature.

The field of cheminformatics has established fundamental protocols [3, 4] and best practices [5, 6] for achieving ML models with high statistical rigor and external predictive power [4]. Here, critical steps such as data preparation, chemical structure curation, outlier detection, dataset balancing, and rigorous ML model validation must be included into the overall data workflow. Multiple studies emphasized that chemical structure curation should be treated as a separate and critical component of any cheminformatics research [6]. Seminal studies showed that accumulated errors and incorrect processing of chemical structures can significantly reduce the accuracy of ML models [7].

The GEOM data set [8] is one of the most widely used large-scale high-accuracy datasets of molecular conformations. A subset of GEOM containing drug-like molecules, known as GEOM-Drugs, has become a foundational benchmark for developing 3D molecular generative models. The frequent use of GEOM-Drugs in this field has given rise to a somewhat standardized set of metrics to evaluate the quality of generative models trained on this dataset. In this work, we identify several critical issues in how state-of-the-art 3D molecular generative models are evaluated. We believe these issues mislead the research community and limit progress in the field.

First, we highlight three major problems with the commonly used "molecular stability" metric, which measures whether atoms have valid valencies. One of the original implementations contained

a bug. This bug caused chemically implausible valencies to be counted as valid, which inflated stability scores. This flawed implementation was reused by several follow-up works [9–14], resulting in a significant body of work with misleading characterizations of model performance.

Second, many recent works lack rigorous and chemically grounded evaluation of 3D structures, which continues to hinder progress in generative modeling. Common issues include the use of oversimplified atom–atom distance lookup tables to evaluate the validity of generated 3D structures [15–20], reliance on distribution-based metrics that are difficult to interpret [10, 14], and the use of energy evaluations at inappropriate levels of theory, such as MMFF94, which is not suitable for assessing models trained on GFN2-xTB-optimized data [9, 21].

To address these issues, this paper provides:

1. A refined dataset split of GEOM-Drugs, which excludes molecules where GFN2-xTB calculations fractured the original molecule.

2. An updated molecule stability metric with a chemically accurate valency lookup table that is derived from this refined dataset.

3. An energy-based evaluation methodology for an accurate and chemically interpretable assessment of generated molecular 3D geometries.

We retrained several widely used generative models on our reprocessed dataset and updated the evaluation metrics to address previously observed issues. The relative rankings of the models remained largely consistent. However, our updates yielded practical improvements. These improvements highlight the importance of rigorous and accurate evaluation practices in the field.

2 Molecule Stability

Valency in chemistry refers to the combining capacity of an atom or element, describing how many chemical bonds it can form with other atoms. It is defined as the sum of bond orders of its covalent bonds. Due to chemical constraints (e.g., the octet rule), atoms of a given element and formal charge typically exhibit only a few plausible valencies; for instance, neutral carbon almost exclusively has a valency of 4. Molecules violating these valency constraints are chemically unstable. Thus, generative models must produce molecules adhering to these rules. A practical evaluation of generative models involves measuring the fraction of atoms with valid valencies, defined as valencies observed in the training data. A "lookup table" of valid valencies, consisting of tuples of (element, formal charge, valency), is created from the training set.

Valency can be computed as the sum of bond orders in a molecule's Kekulized form[1], where bonds are explicitly represented as single, double, or triple. This approach works reliably for molecules without aromatic bonds. When aromatic bonds are introduced, however, valency computation becomes more complex. In simple cases such as benzene, one can assume each aromatic bond contributes 1.5 to the valency, yielding the correct total (e.g., carbon atoms in benzene are correctly assigned a valency of 4). But in more complex aromatic systems, this assumption may not hold, and valency contributions can vary depending on the bonding environment and resonance structures (see Figure 1).

[1]Kekulization is the process of generating an alternative structure for a molecule where aromatic bonds are converted to alternating single and double bonds.

(a) (b)

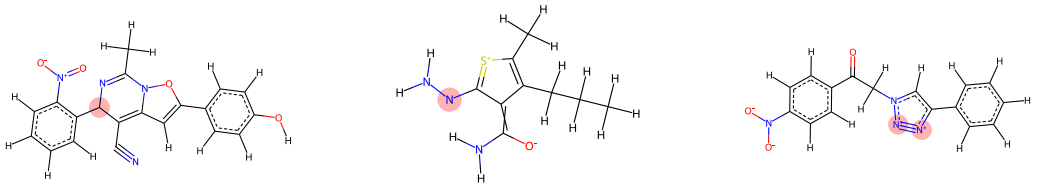
Figure 1: An example of a molecule where the assumption that aromatic bonds contribute 1.5 to atomic valency holds only partially. In the aromatic form of triphenylene (a), the green-highlighted atoms are correctly classified as stable under the 1.5 assumption, while others are misclassified. In contrast, the Kekulized representation (b) resolves the ambiguity and yields chemically accurate valency assignments across all atoms. This illustrates the limitations of the 1.5 approximation in polycyclic aromatic systems.

Hoogeboom et al. [17] first proposed a metric named "molecular stability" to evaluate the correctness of valencies in the raw output of generative models. The evaluation of "molecule stability" in terms of valencies became a norm when publishing new *de novo* molecular generative models. They noted that traditional validity metrics, defined as the fraction of molecules that can be sanitized with RDKit, can be misleading, as RDKit may implicitly adjust hydrogen counts or modify aromaticity, altering the predicted molecule. We generally support the idea of assessing raw valencies, especially for models that explicitly generate both atoms and bonds because it provides a more chemically grounded evaluation. Unlike validity, stability captures whether the generated molecules respect elemental valence constraints without relying on post-processing.

2.1 Identified Issues

We identify multiple critical issues with the valency evaluation methods used in popular molecular generative models; these issues obscure instances where generative models produce chemically implausible structures.

One of the pioneering models, MiDi, implemented a valency calculation method in which the valency contributions for all aromatic bonds were rounded to 1 instead of the intended value of 1.5. Thus, the valency computation for most atoms participating in aromatic bonds is incorrect. More importantly, it appears that the flawed valency computation was also used to construct the valency lookup table with which generated atoms are classified as "stable" or not, resulting in a lookup table with chemically implausible entries. For instance, the lookup table allows for neutral carbon with a valency of 3 and neutral nitrogen with a valency of 2. Implausible entries in the valency lookup table mask failures of the generative model and produce artificially inflated molecular stability values. Due to widespread reuse of MiDi's code, this numerical error propagated to several works including EQGAT-Diff [10], SemlaFlow [9], Megalodon [13], and FlowMol [11, 12]. Other models, such as JODO [15] and NextMol [22], computed valencies using an alternative approach based on RDKit kekulization. However, they still relied on an inappropriate lookup table for defining valid valency ranges.



(a) (b) (c)

Figure 2: Examples of molecules that pass the molecular stability test under commonly used evaluation criteria. These flawed metrics erroneously classify chemically invalid configurations as stable—including cases such as a neutral carbon with three single bonds (a), a neutral nitrogen with two single bonds (b), and a nitrogen atom with +1 charge bonded via both a triple bond and an aromatic bond (c).

2.2 Solution

Two key solutions are necessary to correct the aforementioned problems with the molecular stability metric: fixing the valency computation bug for aromatic bonds and recomputing the valency lookup table. We quantify the effects of our proposed solutions by re-evaluating models that used the faulty molecular stability metric in their original publications: EQGAT-Diff [10], Megalodon-quick [13], SemlaFlow [9], FlowMol2 [12], and Megalodon-flow [13]. The results of these reevaluations are shown in Table 1. All metrics were computed using 5,000 generated molecules per model.

Correcting the numerical bug that erroneously rounded the contribution of aromatic bonds from 1.5 to 1 (without adjusting the lookup table) causes a dramatic drop in molecular stability. This can be observed by comparing the first two columns of Table 1. Additionally, this demonstrates that neither 1 nor 1.5 provides a universally reliable estimate for the contribution of an aromatic bond to atomic valency.

We propose two strategies to address the limitations in molecular stability computation. The first strategy involves enhancing the valency lookup table by explicitly accounting for aromaticity. Instead of the conventional tuples (element, formal charge, valency), we construct a more nuanced table indexed by (element, number of aromatic bonds, formal charge, valency), with the associated values representing allowed non-aromatic bond valencies—i.e., total bond order excluding contributions from aromatic bonds (see SI Table 5). In this formulation, each atom's bonding environment is described by the tuple $(n_{\text{arom}}, v_{\text{other}})$, where n_{arom} is the number of aromatic bonds and v_{other} is the total bond order from non-aromatic bonds. For example, a carbon atom in benzene typically exhibits configurations like $(2, 1)$—two aromatic bonds and one single bond—or $(3, 0)$, as illustrated in Figure 1. Remarkably, adopting this refined lookup table results in molecular stability scores only 1–3% lower than originally reported using flawed metrics (third column in Table 1). While modest, this deviation can meaningfully influence the comparative assessment of generative models and may introduce bias into subsequent benchmark studies if left uncorrected.

An alternative approach involves retraining models on a reprocessed dataset consisting exclusively of Kekulized molecules, thereby completely removing ambiguity associated with aromaticity in valency computation. We prepared a revised version of the GEOM-Drugs dataset so that all molecules were Kekulized; there is no explicit modeling of aromatic bonds. As illustrated in Table 1, models trained on the Kekulized dataset exhibited molecular stability comparable to previously published results when valencies were computed correctly. Notably, all models except Megalodon Flow demonstrated an average 5% improvement in validity. Megalodon Flow did not show similar improvements. We hypothesize that this discrepancy arises due to smaller neural network architecture used for Megalodon Flow, a decision necessitated by limited computational resources available for this study.

We encountered another issue with GEOM-Drugs: recomputing the valency table on the raw GEOM-Drugs dataset revealed unusual valencies resulting from rare failure in the GFN2-xTB geometry optimization step used to produce the dataset. These failures produced fragmented molecules and unstable valencies such as hydrogen atom with no covalent bonds or neutral carbon with a valency of two. Examples of these instances are shown in Figure 3. We removed molecules from

GEOM-Drugs that were fragmented into multiple disconnected components due to failed GFN2-xTB geometry optimization. This led to the exclusion of 0.18% of the dataset; although this is not enough data to significantly impact model performance, the presence of these molecules alters the resulting valency lookup table.

To summarize, neither treating aromatic bonds as contributing a valence of 1 nor 1.5 yields chemically accurate results. By correcting the valency table using a refined tuple representation, which captures the number of aromatic bonds separately, the resulting molecular stability scores decrease modestly by 1 to 3%. However, since most reported stability values exceed 0.9, even such small discrepancies can have an outsized influence, potentially skewing model development and encouraging optimization against a chemically flawed metric. Notably, retraining models on a reprocessed dataset with Kekulized molecules, i.e., without explicit aromatic bonds, leads to approximately a 5% improvement in validity for 4 of 6 evaluated models. Together, these results underscore the critical importance of chemically sound preprocessing and robust evaluation protocols in the development of 3D molecular generative models.

We make available in the attached github repository the filtered GEOM-Drugs dataset with Kekulized molecules, the scripts for producing the filtered dataset from the original GEOM dataset, and an implementation of the molecular stability metric that does not permit erroneous atomic valencies.

Model	MS Original	MS 1.5 Arom	MS Arom-Dependent Valence	V&C	MS	V&C
EQGAT [10]	$0.935^{\pm 0.007}$	$0.451^{\pm 0.006}$	$0.899^{\pm 0.007}$	$0.834^{\pm 0.009}$	$0.878^{\pm 0.007}$	$0.891^{\pm 0.010}$
JODO [15]	$0.981^{\pm 0.001}$	$0.517^{\pm 0.012}$	$0.963^{\pm 0.005}$	$0.879^{\pm 0.003}$	$^{*}0.940^{\pm 0.003}$	$^{*}0.923^{\pm 0.004}$
Megalodon-quick [13]	$0.961^{\pm 0.003}$	$0.496^{\pm 0.017}$	$0.944^{\pm 0.003}$	$0.900^{\pm 0.007}$	$0.957^{\pm 0.006}$	$0.962^{\pm 0.005}$
SemlaFlow [9]	$0.980^{\pm 0.012}$	$0.608^{\pm 0.027}$	$0.969^{\pm 0.012}$	$0.920^{\pm 0.016}$	$0.974^{\pm 0.012}$	$0.975^{\pm 0.008}$
FlowMol2 [12]	$0.959^{\pm 0.007}$	$0.594^{\pm 0.009}$	$0.944^{\pm 0.007}$	$0.746^{\pm 0.010}$	$0.938^{\pm 0.005}$	$0.861^{\pm 0.012}$
Megalodon-flow [13]	$0.990^{\pm 0.003}$	$0.632^{\pm 0.011}$	$0.987^{\pm 0.004}$	$0.948^{\pm 0.003}$	$^{**}0.958^{\pm 0.004}$	$^{**}0.949^{\pm 0.002}$

Table 1: Comparison of molecular stability (MS) and connected validity (V&C) across models and processing pipelines. The left section reports results obtained using the original GEOM-Drugs dataset and evaluation code: "Original" denotes the values from metric implementations published in prior work, "1.5 Arom" reflects scores if aromatic bonds contribute 1.5 to valency, and "Arom-Dependent Valence" shows scores based on valency computed as $(n_{\text{arom}}, v_{\text{other}})$. The right section presents results obtained by retraining on fully Kekulized molecules. V&C (Valid & Connected) refers to the fraction of molecules that are both chemically valid and consist of a single connected component.

[*] JODO was trained with the EQGAT-Diff objective, using categorical diffusion instead of the original Gaussian formulation for categorical variables.

[**] Indicates results from a retrained "quick" variant, differing from the original paper which reported results for a larger model.

3 3D Molecule Evaluation

3.1 Challenges in proper and accurate 3D structure assesment

Current 3D molecular generative models face significant challenges in evaluating the geometric quality of their outputs. In particular, models trained on the GEOM-Drugs dataset often exhibit issues stemming from the evaluation protocols themselves. A widely used approach involves defining a bond length lookup table and applying fixed thresholds to assess 3D molecular stability [15–20]. However, this method is problematic for GEOM-Drugs. Only 86.5% of atoms meet the specified atom–atom distances, and just 2.8% of molecules pass the stability criterion.

Our analysis found only 272 fragmented molecules in the dataset. This indicates that GFN2-xTB geometry optimization converged successfully for most conformers. Thus, the observed bond lengths reflect the GFN2-xTB energy landscape. These values may differ from those obtained from other sources, such as the Cambridge Structural Database (CSD). This metric produces implausibly low

stability rates yet remains widely adopted in new studies. This persistence underscores the need for a more chemically faithful evaluation standard.

A more recent trend is to assess geometric quality by comparing distributions of bond lengths and angles using Wasserstein distance between generated and training data [10, 14, 23]. This approach is more principled. However, distributional metrics are often difficult to interpret, especially outside the computer science community, which makes it harder to extract chemically meaningful insights.

Other studies have proposed evaluating generated molecules by computing the relaxation energy using molecular mechanics force fields such as MMFF [9, 21, 24]. However, the choice of force field is critical. For conformers optimized with GFN2-xTB (as in GEOM-Drugs), the mean relaxation energy difference ΔE_{relax} when re-optimized with GFN2-xTB is close to zero, as expected. In contrast, MMFF evaluation yields a mean ΔE_{relax} of ≈ 16 kcal/mol. This is consistent with prior reports of MMFF errors in the 15–20 kcal/mol range relative to higher-level methods[25].

As we will demonstrate, current state-of-the-art generative models can now outperform MMFF precision on GEOM-Drugs in terms of alignment with GFN2-xTB. This renders MMFF-based comparisons unreliable and masks meaningful differences between models. However, MMFF energy can still serve as a coarse-grained filter to eliminate structurally implausible molecules, similar to its use in PoseBusters [26] for energy-based outlier detection. Given the widespread reliance on inadequate metrics, we argue that a GFN2-xTB-based evaluation pipeline is necessary for accurately assessing the practical performance of 3D molecular generative models.

3.2 GFN2-xTB energy-based geometry benchmark

GEOM-Drugs geometries are optimized using the GFN2-xTB semi-empirical quantum calculation method. Therefore, it is essential to use the same method to assess the structural integrity of generated molecules. One approach is to measure of how close a generated structure is to the closest local minima of the given energy function. To measure this we suggest to assess differences in bond lengths, bond angles, and torsion angles of generated and optimized counterparts. These quantities provide clear and interpretable measure of generated molecules for both computer scientists and computational chemists.

Bond Length Differences For each bond in the molecule, we compute the difference in bond lengths between the initial (generated) and optimized (relaxed) structures. Let r_{ij}^{init} and r_{ij}^{opt} denote the distances between atoms i and j in the initial and optimized conformations, respectively. The bond length difference Δr_{ij} is calculated as:

$$\Delta r_{ij} = \left| r_{ij}^{\text{init}} - r_{ij}^{\text{opt}} \right|$$

The average difference is reported as a result.

Bond Angle Differences For each bond angle formed by three connected atoms i, j, and k, we calculate the angle difference between the initial and optimized structures. Let $\theta_{ijk}^{\text{init}}$ and $\theta_{ijk}^{\text{opt}}$ represent the bond angles at atom j in the initial and optimized conformations, respectively. The bond angle difference $\Delta\theta_{ijk}$ is given by:

$$\Delta\theta_{ijk} = \min\left(\left| \theta_{ijk}^{\text{init}} - \theta_{ijk}^{\text{opt}} \right|, 180^\circ - \left| \theta_{ijk}^{\text{init}} - \theta_{ijk}^{\text{opt}} \right| \right)$$

As with bond lengths, the average difference is reported as a result.

Torsion Angle Differences Torsion angles involve four connected atoms i, j, k, and l. We compute the difference in torsion angles between the initial and optimized structures using:

$$\Delta\phi_{ijkl} = \min\left(\left| \phi_{ijkl}^{\text{init}} - \phi_{ijkl}^{\text{opt}} \right|, \; 360^\circ - \left| \phi_{ijkl}^{\text{init}} - \phi_{ijkl}^{\text{opt}} \right| \right)$$

where $\phi_{ijkl}^{\text{init}}$ and ϕ_{ijkl}^{opt} are the dihedral angles in the initial and optimized conformations, respectively. This formula accounts for the periodicity of dihedral angles, ensuring the smallest possible difference is used.

The average difference is reported as a result.

Results We report results for EQGAT, Megalodon-quick, SemlaFlow, FlowMol2, and Megalodon-flow, including both the median and mean relaxation energy ΔE_{relax}—the energy difference between the initial and GFN2-xTB-optimized structures—as well as structural displacement metrics discussed above (see Table 2). For each model, 5,000 molecules were evaluated, and a randomly selected subset of 5,000 molecules from GEOM-Drugs was used for baseline comparisons. To compute confidence intervals, all metrics were calculated across five equal-sized splits of 1,000 molecules each. In Table 2, the row labeled "MMFF $\rightarrow$ GFN2-xTB" quantifies geometric and energetic discrepancies between MMFF-optimized structures and their GFN2-xTB-optimized counterparts. These values highlight the structural divergence between force-field and semi-empirical optimization methods. Diffusion-based models already surpass MMFF in structural precision. Furthermore, we observe a consistent performance gap between flow-matching and diffusion-based models, even when the architecture is the same. This discrepancy has not been emphasized in prior literature. Conclusions drawn in prior works may have been influenced by the limited precision of their evaluation methodologies.

Table 2: Energy relaxation and geometric deviation metrics across generative models. Bond lengths (Å), angles (degrees), and energies (kcal/mol) are reported for valid molecules only. Diffusion-based models use 500 steps; flow-matching models use 100 steps. ΔE_{relax} denotes the energy difference between the initial and GFN2-xTB-optimized structures (i.e., the generative model's deviation from the reference energy landscape). $\Delta E_{\text{relax}}^{\text{MMFF}}$ denotes the MMFF94 energy difference between the initial structure and the structure optimized with MMFF94.

Model	Bond Length ($\times 10^{-2}$)	Bond Angles	Torsions	Median ΔE_{relax}	Mean ΔE_{relax}	Mean $\Delta E_{\text{relax}}^{\text{MMFF}}$
GEOM-Drugs	$0.00^{\pm 0.001}$	$0.001^{\pm 0.001}$	$0.01^{\pm 0.01}$	$0.000^{\pm 0.0001}$	$0.001^{\pm 0.001}$	$16.4^{\pm 0.2}$
MMFF $\rightarrow$ GFN2-xTB	$1.12^{\pm 0.01}$	$1.22^{\pm 0.004}$	$4.89^{\pm 0.10}$	$9.84^{\pm 0.06}$	$11.4^{\pm 0.2}$	$0.00^{\pm 0.05}$
EQGAT-diff	$1.00^{\pm 0.04}$	$1.15^{\pm 0.03}$	$8.58^{\pm 0.11}$	$6.40^{\pm 0.20}$	$11.1^{\pm 0.8}$	$28.4^{\pm 1.2}$
JODO	$0.77^{\pm 0.01}$	$0.83^{\pm 0.00}$	$6.01^{\pm 0.07}$	$4.74^{\pm 0.15}$	$7.04^{\pm 0.20}$	$22.1^{\pm 0.2}$
Megalodon	$0.66^{\pm 0.02}$	$0.71^{\pm 0.01}$	$5.58^{\pm 0.11}$	$3.19^{\pm 0.12}$	$5.76^{\pm 0.27}$	$21.6^{\pm 0.3}$
SemlaFlow	$3.10^{\pm 0.23}$	$2.06^{\pm 0.17}$	$6.05^{\pm 0.56}$	$32.3^{\pm 3.3}$	$91.0^{\pm 21.7}$	$69.6^{\pm 9.2}$
FlowMol2	$1.30^{\pm 0.04}$	$1.62^{\pm 0.02}$	$15.0^{\pm 0.3}$	$17.9^{\pm 0.5}$	$24.3^{\pm 0.8}$	$39.4^{\pm 1.2}$
Megalodon-flow	$2.30^{\pm 0.02}$	$1.62^{\pm 0.02}$	$5.58^{\pm 0.19}$	$20.9^{\pm 0.8}$	$46.9^{\pm 8.6}$	$45.5^{\pm 2.0}$

4 Conclusion

In this study, we revisited the GEOM-Drugs benchmark and uncovered several issues in current 3D molecular generative model evaluation pipelines. We demonstrated that widely adopted stability metrics are affected by code errors, chemically inconsistent valency tables, and reliance on postprocessed molecules, leading to inflated model performance. Furthermore, our findings suggest that MMFF-based energy benchmarks may no longer be appropriate for evaluating models trained on GFN2-xTB-optimized structures, as generative models now appear to surpass MMFF in alignment with the reference energy landscape.

To address these limitations, we proposed a refined evaluation protocol incorporating chemically sound valency definitions and GFN2-xTB-based energy and geometry assessments. Our experiments demonstrate that these corrections impact reported performance while preserving the relative rankings of models. Conversely, a high-quality dataset (error-free structures, consistent features, trustworthy labels) and relevant metrics (e.g. appropriate choice of level of theory or realistic valency lookup table) provide a solid foundation that can markedly improve model performance. We hope that this study will raise awareness about importance of chemical structure curation and pro-

cessing. We believe these improvements will foster more reliable, interpretable, and chemically meaningful progress in 3D molecular generative modeling. Our recommended evaluation methods and GEOM-Drugs processing scripts will be a vailable at the time of workshop.

References

[1] Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

[2] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[3] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57 (12):4977–5010, 2014.

[4] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, 29(6-7):476–488, 2010.

[5] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

[6] Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature chemistry*, 13(6):505–508, 2021.

[7] Douglas Young, Todd Martin, Raghuraman Venkatapathy, and Paul Harten. Are the chemical structures in your qsar correct? *QSAR & combinatorial science*, 27(11-12):1337–1345, 2008.

[8] Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

[9] Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow–efficient 3d molecular generation with latent attention and equivariant flow matching. In *The 28th International Conference on Artificial Intelligence and Statistics*.

[10] Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. *arXiv preprint arXiv:2309.17296*, 2023.

[11] Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *ArXiv*, pages arXiv–2404, 2024.

[12] Ian Dunn and David Ryan Koes. Exploring discrete flow matching for 3d de novo molecule generation. *ArXiv*, 2024. doi: 10.48550/arXiv.2411.16644.

[13] Danny Reidenbach, Filipp Nikitin, Olexandr Isayev, and Saee Gopal Paliwal. Applications of modular co-design for de novo 3d molecule generation. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*.

[14] Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer, 2023.

[15] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2-d and 3-d graph diffusion models for complete molecule generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[16] Victor Garcia Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34: 4181–4192, 2021.

[17] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

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

[19] 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. *Advances in Neural Information Processing Systems*, 36:549–568, 2023.

[20] Can Xu, Haosen Wang, Weigang Wang, Pengfei Zheng, and Hongyang Chen. Geometric-facilitated denoising diffusion model for 3d molecule generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 338–346, 2024.

[21] François Cornet, Grigory Bartosh, Mikkel Schmidt, and Christian Andersson Naesseth. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37:49429–49460, 2024.

[22] Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Next-mol: 3d diffusion meets 1d language modeling for 3d molecule generation. *arXiv preprint arXiv:2502.12638*, 2025.

[23] Julian Cremer, Tuan Le, Frank Noé, Djork-Arné Clevert, and Kristof T Schütt. Pilot: equivariant diffusion for pocket-conditioned de novo ligand generation with multi-objective guidance via importance sampling. *Chemical Science*, 15(36):14954–14967, 2024.

[24] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[25] Nicolas Foloppe and I-Jen Chen. Energy windows for computed compound conformers: covering artefacts or truly large reorganization energies? *Future Medicinal Chemistry*, 11(2): 97–118, 2019.

[26] 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.

Supplementary Information

4.1 Appendix I: Valency Lookup Tables for Stability Evaluation

To support rigorous evaluation of 3D molecular generative models, we include here a collection of empirical valency tables derived from the GEOM-Drugs dataset. These tables are used to define chemically plausible bonding patterns, detect invalid topologies, and serve as standardized references for assessing molecular stability in raw generated molecules.

Table 3: Allowed Valencies. This table summarizes the allowed valencies (i.e., number of bonds including hydrogens) observed in valid GEOM-Drugs structures. It lists configurations by element and formal charge. These values are used as a reference for atom-level and molecule-level stability metrics.

Table 4: Legacy and Invalid Valencies. This table contains valencies found in earlier versions of generative model evaluation pipelines, which include chemically implausible or legacy entries due to preprocessing bugs or failed optimization. It is frequently used to benchmark the quality of generated molecules and identify invalid valency assignments. Many recent studies reference or reuse this table directly.

Table 5: Aromatic Valency Tuples. This table enumerates all observed combinations of aromatic and non-aromatic bonds per element and charge in the dataset. Each entry is represented as a tuple , where is the count of aromatic bonds and is the total bond order from non-aromatic bonds. These tuples capture valency patterns that are otherwise ambiguous under standard counting, especially in polyaromatic and heterocyclic systems.

Together, these tables offer a robust and chemically grounded framework for interpreting stability metrics and ensuring consistency in the evaluation of 3D molecule generation pipelines. Table 4 in particular is widely used in existing benchmarking literature and reproduced here for completeness.

Table 3: Valency configurations derived from the GEOM-Drugs dataset, organized by element and formal charge. Each cell lists the allowed valencies (including implicit hydrogens) observed for a given formal charge.

Element	Charge -2	Charge -1	Charge 0	Charge $+1$	Charge $+2$	Charge $+3$
H	–	–	1	–	–	–
B	–	4	3	–	–	–
C	–	3	4	3	–	–
N	1	2	3	4	–	–
O	–	1	2	3	–	–
F	–	–	1	–	–	–
Si	–	–	4	5	–	–
P	–	–	$3, 5$	4	–	–
S	–	1	$2, 3, 6$	3	4	$2, 5$
Cl	–	–	1	2	–	–
Br	–	–	1	2	–	–
I	–	–	1	2	3	–
Bi	–	–	3	–	5	–

Table 4: Historically used but chemically implausible valency configurations by formal charge. This reference table has been widely used to assess molecular generative models. Values highlighted in red represent known incorrect or unstable configurations; values highlighted in blue were missing from historical tables but are observed in the dataset.

Element	Charge -2	Charge -1	Charge 0	Charge $+1$	Charge $+2$	Charge $+3$
H	–	0	1	0	–	–
B	–	4	3	–	–	–
C	–	3	3, 4	3	–	–
N	1	2	2, 3	2, 3, 4	–	–
O	–	1	2	3	–	–
F	–	0	1	–	–	–
Al	–	–	3	–	–	–
Si	–	–	4	5	–	–
P	–	–	3, 5	4	–	–
S	–	1, 3	2, 6	2, 3	4	5
Cl	–	–	1	2	–	–
Br	–	–	1	2	–	–
Se	–	–	2, 4, 6	–	–	–
I	–	–	1	2	3	–
Hg	–	–	1, 2	–	–	–
Bi	–	–	3	–	5	–

Table 5: Allowed valency combinations by element and number of aromatic bonds. Each cell shows normal valencies for a given atom type and number of aromatic neighbours (row) and formal charge (column). "–" indicates no observed combinations.

Element	# Aromatic	Charge -2	Charge -1	Charge 0	Charge $+1$	Charge $+2$	Charge $+3$
H	0	–	–	1	–	–	–
B	0	–	4	3	–	–	–
C	0	–	3	4	3	–	–
	2	–	1	2, 1	1	–	–
	3	–	0	0	0	–	–
N	0	1	2	3	4	–	–
	2	–	0	0, 1	0, 1, 2	–	–
	3	–	–	0	0	–	–
O	0	–	–	2	3	–	–
	2	–	–	0	–	–	–
F	0	–	–	1	–	–	–
Si	0	–	–	4	5	–	–
P	0	–	–	3, 5	4	–	–
S	0	–	1	2, 3, 6	3	4	2,5
	2	–	–	0	0, 1	–	–
	3	–	–	–	0	–	–
Cl	0	–	–	1	2	–	–
Br	0	–	–	1	2	–	–
I	0	–	–	1	2	3	–
Bi	0	–	–	3	–	5	–

4.2 Appendix II: Examples of Fractured Compounds in GEOM-Drugs

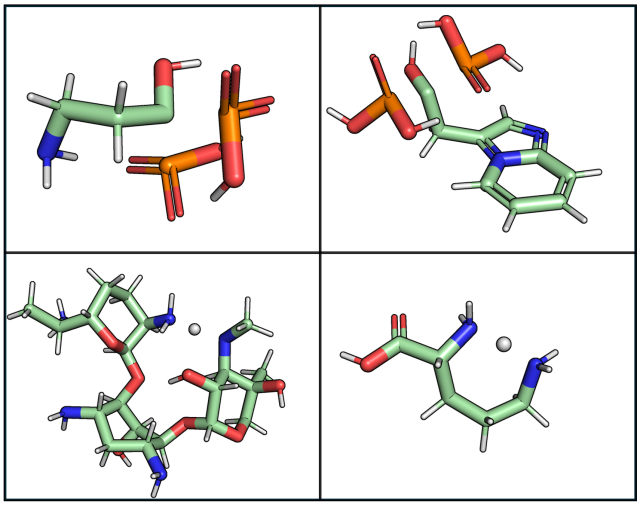


Figure 3: Examples from GEOM-Drugs where GFN2-xTB failed and resulted in fractured molecules. The first row of molecules have neutral carbon with valency 2 and those in the second row have a positively charged hydrogen with valency zero.