3D Structure-based Generative Small Molecule Drug Design: Are We There Yet?

Bo Yang, Chijian Xiang, Tongtong Li, Yunong Xu, Jianing Li*

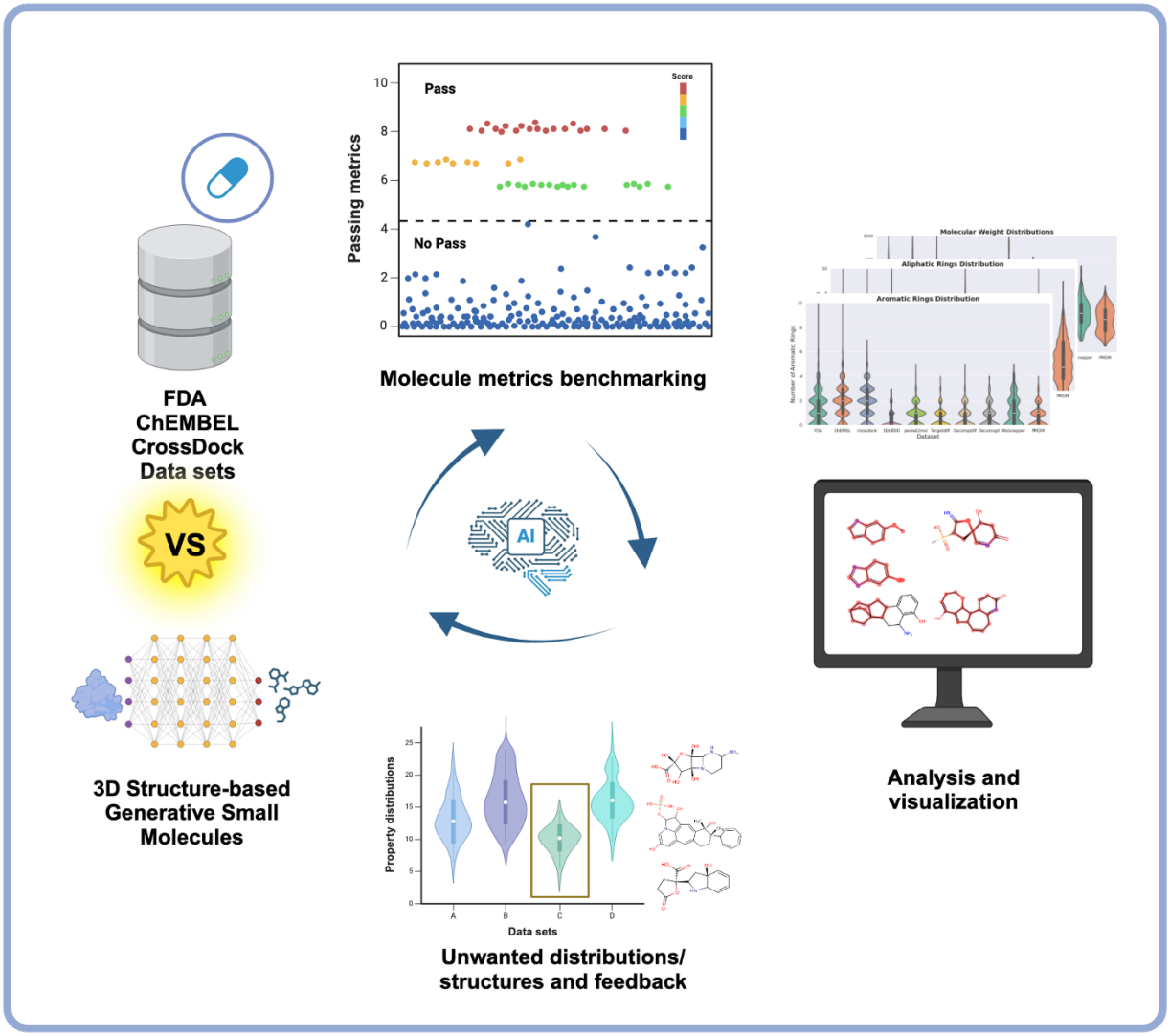
Borch Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN 47907

* Jianing Li (jianing-li@purdue.edu)

Abstract:

Structure-based drug design (SBDD) plays a crucial role in pre-clinical drug discovery. Recently developed structure-based 3D generative algorithms have the potential to streamline the SBDD process by enabling the generation of novel, druglike molecules based on the structure of a target protein's binding pocket. However, no effective metric currently exists to assess the chemical plausibility of molecules generated by these algorithms, which hampers further advancements in their development. In this study, we propose two novel metrics for evaluating the chemical plausibility of generated molecules. Combined with additional analyses, we demonstrate that existing algorithms could generate chemically implausible structures with some property distributions that deviate from those of known drug-like molecules. The metrics and analytical methods presented here offer model developers and user valuable tools to quantify the chemical plausibility of molecules they generate, ultimately facilitating the application of these algorithms in drug discovery.

Figure abstract:



Key words:

Generative algorithms, Drug discovery, Chemical plausibility, Benchmarking, Hit identification

Drug discovery is a time- and cost-intensive process, taking approximately 12-15 years and costing between $314 million and $4.46 billion to bring a novel drug to market, depending on the therapeutic area and data available, with a high failure rate[1]. With

advancements in deep learning, researchers are applying it to accelerate the drug discovery process[2]. The ultimate goal is to identify optimal candidates from an immense chemical space—estimated at over 10^{60} possible compounds[4]— that exhibit high affinity and drug-likeness for a specific target, thereby saving time and reducing costs. As a branch of deep learning in drug discovery, structure-based 3D generative algorithms focus on generating novel, druglike molecules with high affinity for target protein binding pockets, achieving significant progress in recent years (Figure 1).

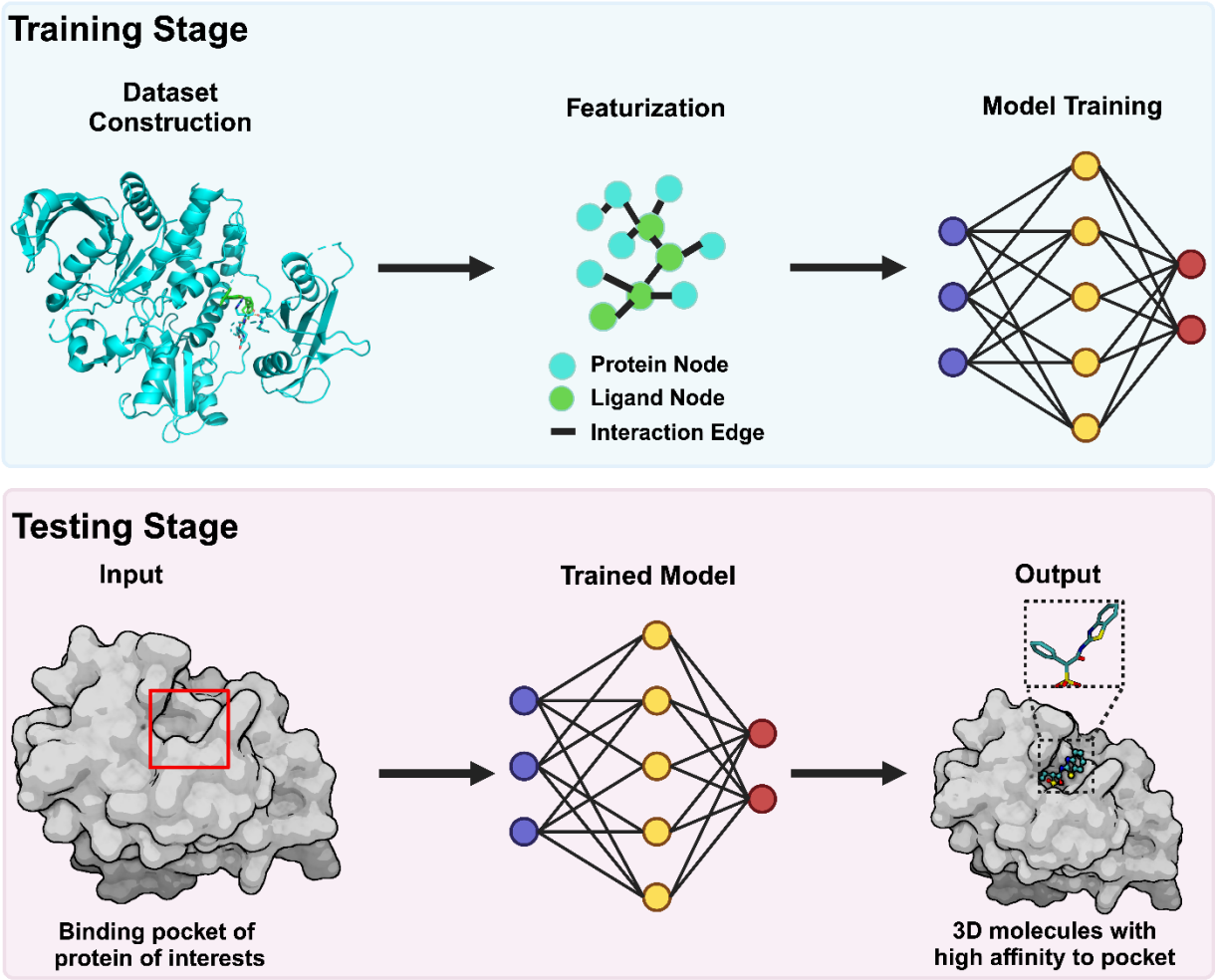


Figure 1. General workflow of structure-based 3D generative algorithms. Protein-ligand complexes are converted into machine-readable features and used as input to train the generative model. Once trained, the model takes the binding pocket of interest as input and generates 3D ligand structures that have high affinity for the given binding pocket.

Current structure-based 3D generative algorithms can be classified into several categories based on different criteria: how they represent 3D data, the strategies they employ for molecule generation, and whether they incorporate domain knowledge, such as pharmacophores or protein-ligand interactions, to guide the generation process (Figure 2). Regardless of their classification, these models are capable of generating molecules conditioned on binding pocket structures with high predicted binding affinity predicted by docking. Besides docking scores, commonly used metrics to assess the quality of generated molecules include drug-likeness[6] and synthetic accessibility[7], which are known to have their own limitations[8]. Recently, Buttenschoen et al. introduced PoseBusters, a tool designed to evaluate the physical validity of ligand poses generated by deep learning models[9]. This work has driven the development of generative models that could produce more physically plausible binding pose[10]. However, while other generative models have been shown to generate chemically implausible structures (i.e. structures that are not chemically stable or drug-like)[11], the output of structure-based 3D generative algorithms have yet to be examined rigorously from a medicinal chemistry perspective. Most studies have relied solely on the validity of SMILES strings or 3D structures as criteria for chemical plausibility, supplemented by distribution analyses of 2D structural properties. However, these approaches are insufficient to comprehensively evaluate chemical plausibility[12-14]. Additionally, the lack of effective metrics hinders the development of models that generate chemically plausible, drug-like molecules, which are crucial requirement for their application in drug discovery.

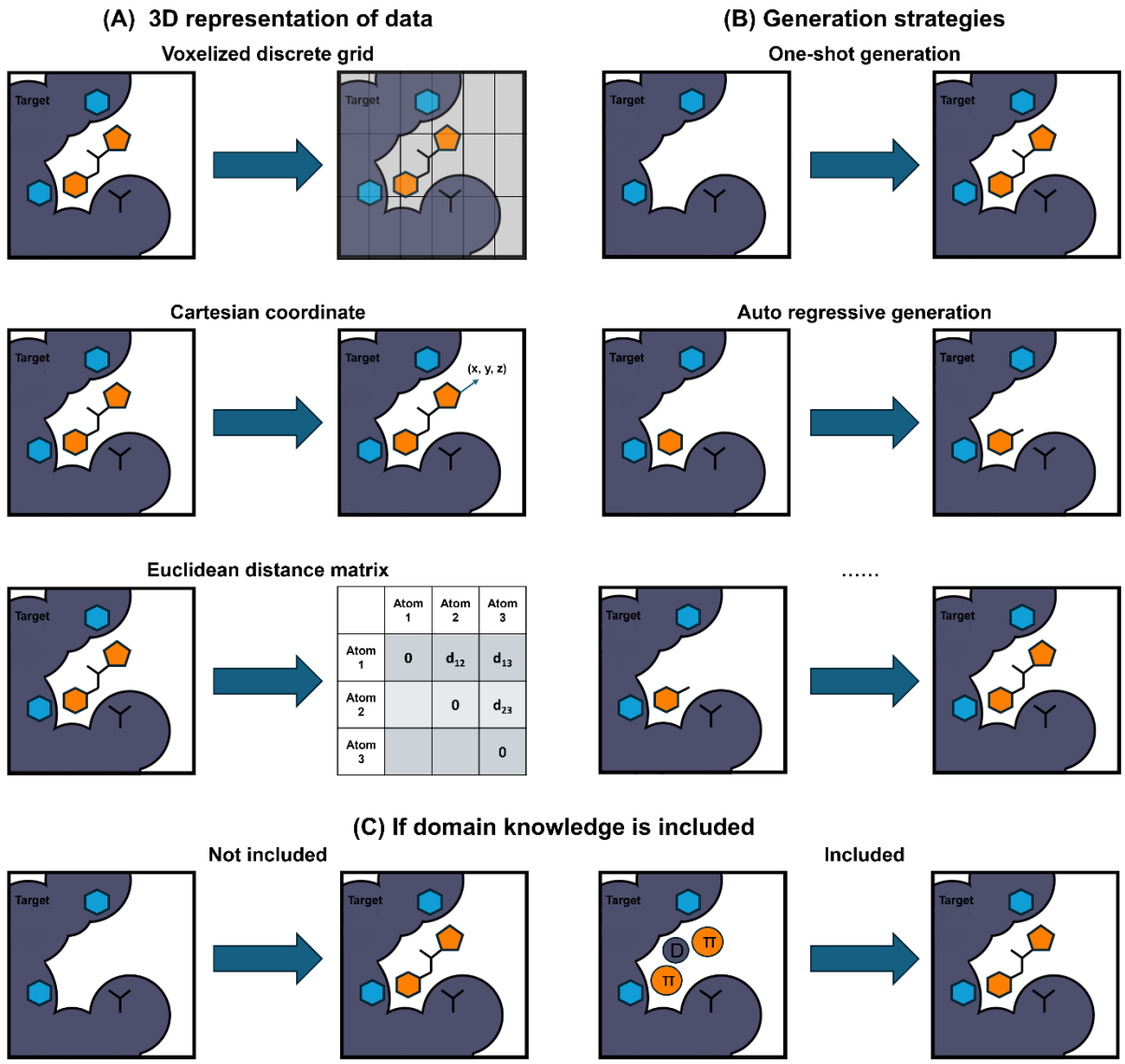


Figure 2. The classification methods of current 3D structure-based generative models. (A) Based on 3D representations of data. (B) Based on generation strategies. (C) Based on if domain knowledge is included in the generation process.

To solve the problems mentioned, we introduced two novel metrics, based on druglike molecules from ZINC20 and ZINC22 databases, for evaluating the ability of algorithms to generate chemically plausible molecules. We applied these, along with additional metrics[11, 15], to assess the chemical plausibility of molecules generated by seven structure-based 3D generative algorithms (3D-Generative-SBDD[16], Pocket2mol[17], TargetDiff[18], DecompDiff[19], PMDM[14], DecompOpt[13], MolSnapper[10]), and compared with three control groups (FDA approved small molecule drugs, ChEMBL clinical small

molecules as well as molecules from CrossDocked2020[20]) that consist of more drug-like and realistic molecules.

Novel metrics indicate that molecules generated by algorithms exhibit lower drug-likeness

To ensure a fair comparison of structure-based 3D generative models, all selected algorithms were trained and tested on the same subsets of the CrossDock2020 dataset[16, 20]. Aiming to avoid training bias, we used pre-trained models to generate 10 molecules for each of the 100 proteins in the test set, yielding a target total of 1,000 molecules for the benchmarking study. However, not all algorithms generated exactly 1,000 molecules, and some generated molecules were either invalid or duplicate. Since they are not the focus of our study, we provide these data in the supporting information (Table S1).

Three metrics are used here to assess the chemical plausibility of molecules generated by different models. The first metric involves extracting ring systems from each generated molecule and documenting their occurrence in the ChEMBL database[11, 15]. The rationale is that if a ring system does not appear frequently in ChEMBL, it is likely to be either unstable or not druglike. The second and third metrics evaluate whether the Bemis-Murcko[21] scaffolds (BM scaffolds) of the molecules appear among druglike molecules in the ZINC20 (870 million) and ZINC22 (54.6 billion) databases. If a molecule's BM scaffold is absent from these databases, it may indicate instability or a lack of druglike properties. It is also worth mentioning that the "druglike molecules" in the ZINC databases are predicted based on molecular properties and may not necessarily be truly druglike. Furthermore, some molecules in the database have not been synthesized but are predicted to be synthesizable. Additionally, it is possible for a BM scaffold to be absent from ZINC while still being chemically plausible, which is also true for ring system in ChEMBL. The results of the benchmarking study are shown in Table 1.

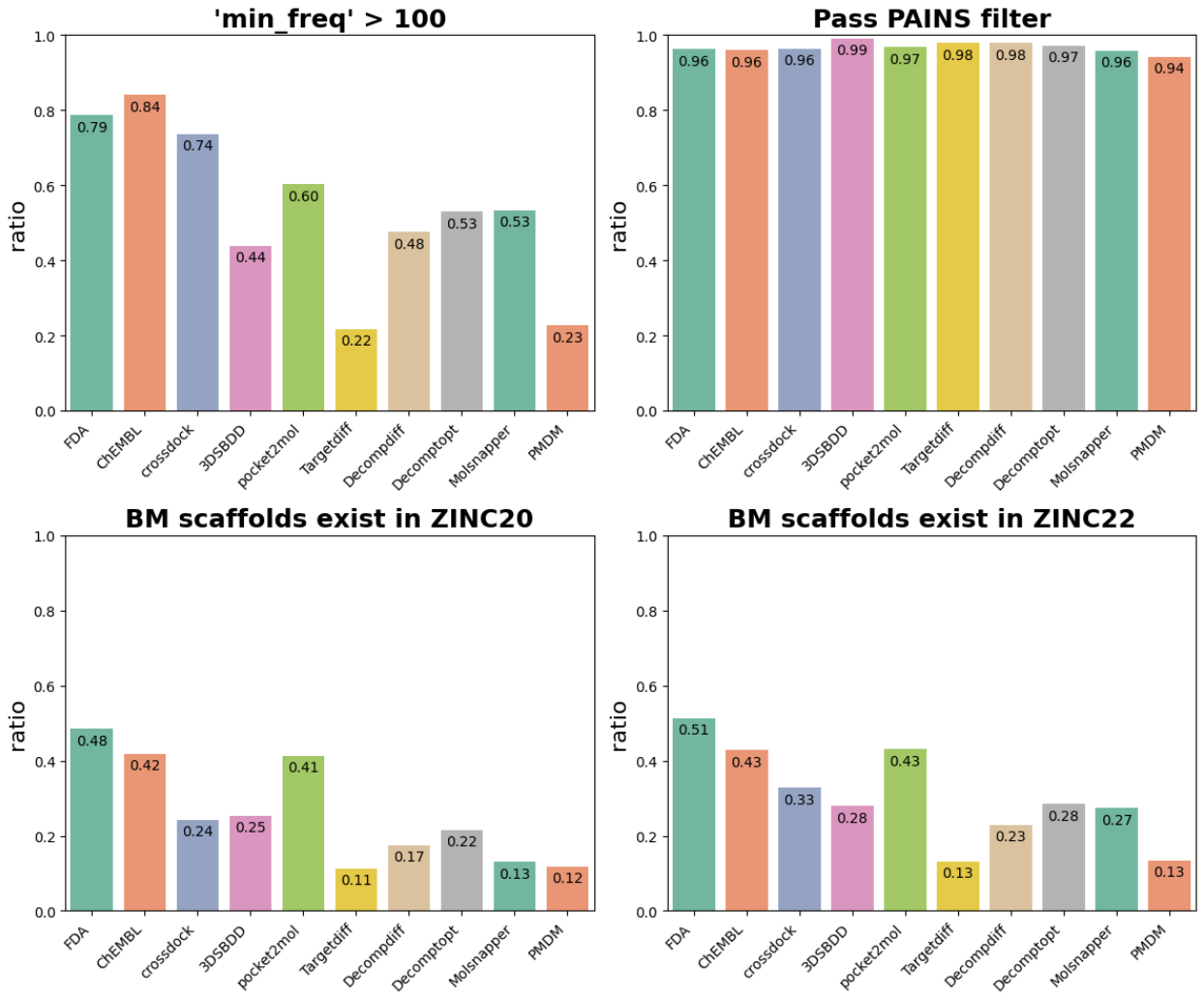


Figure 3. Benchmarking result. All values are normalized as a ratio. “min_freq” > 100: This metric is calculated by extracting the ring systems in each molecule and documenting the frequency of each ring system's occurrence in the ChEMBL database. If the least frequent ring system (some molecules may contain multiple ring systems) appears fewer than 100 times in ChEMBL, the molecule is considered potentially unstable or non-druglike. The ratio is determined by dividing the number of molecules that meet this criterion by the number of unique molecules containing ring systems. “BM scaffolds” are calculated by extracting the scaffolds of drug-like molecules from ZINC20 and ZINC22 databases. If the BM scaffolds of generated molecules are not found in these databases, it may suggest that the generated structures are either difficult to synthesize or chemically implausible. The ratio is calculated by dividing the number of BM scaffolds present in ZINC by the total number of BM scaffolds from the generated molecules.

Before examining the three metrics, one conclusion that can be drawn from the figure is that the PAINS filter does not effectively exclude problematic molecules, as indicated by the consistently high passing rates across all datasets tested (Figure 3, Figure S4).

This limitation may stem from the fact that PAINS filters, along with other widely used filters, were originally developed from high-throughput screening campaigns, where problematic structures were identified and cataloged. However, the problematic molecules generated by the model here are often chemically implausible, meaning they may not even be viable structures. As a result, these existing filters are ineffective for screening non-viable or non-existing structures.

Compared to the PAINS filter, the three metrics applied here provide more detailed insights. The three control groups consistently perform well on both metrics, except for the CrossDocked dataset in two ZINC based metrics, which may be because the CrossDocked dataset includes ligands from crystal structure but not drug molecules, thus are less druglike than molecules from other two control groups. Additionally, the control groups outperform most of the algorithms tested, with the exception of Pocket2mol and 3D_generative_SBDD. This is likely because these algorithms generate simpler structures that resemble fragments more than complete drug molecules, which we will explore in detail later. In summary, the two metrics effectively distinguish between the three control groups and the seven tested algorithms, highlighting that some generated molecules may indeed be chemically implausible.

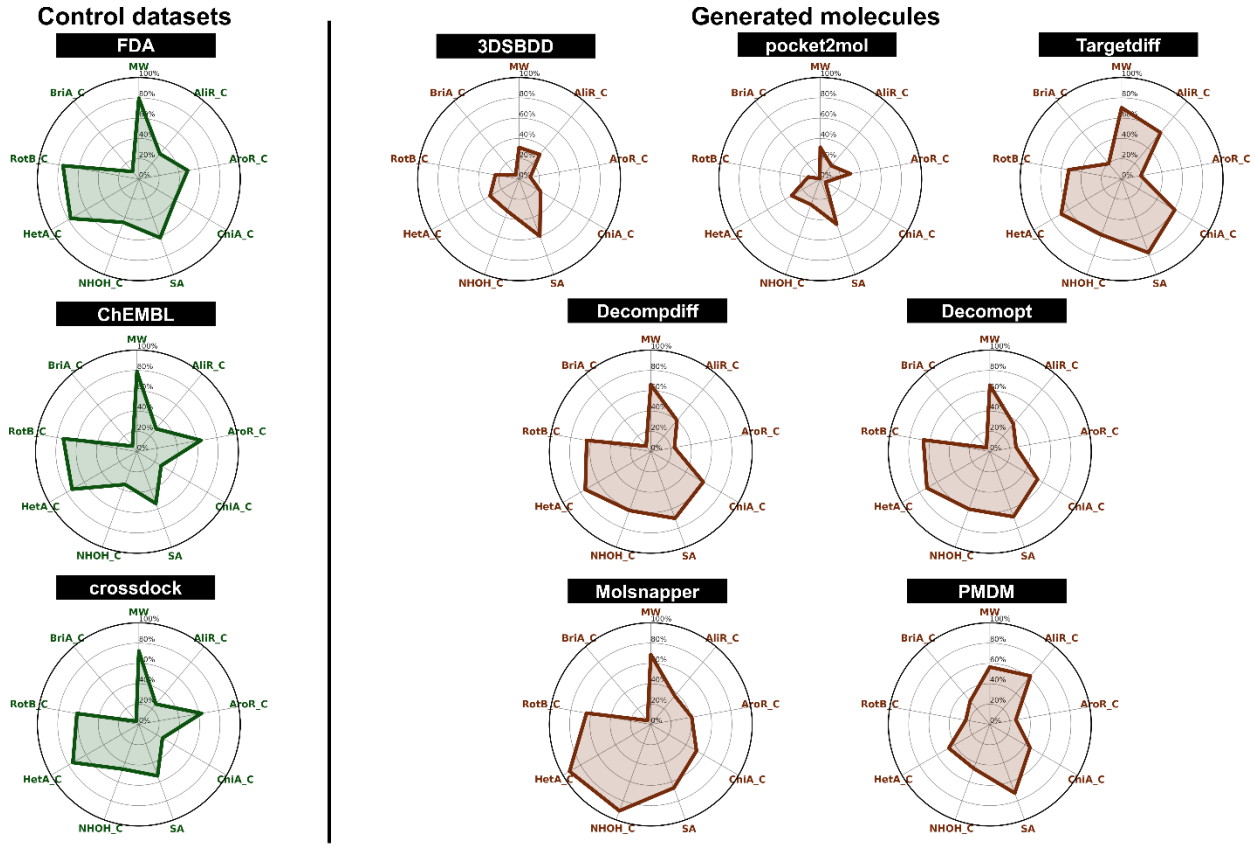


Figure 4. Overall Chemical properties distributions among datasets. MW: molecular weight; AliR_C: aliphatic ring count; AroR_C: aromatic ring count; ChiA_C: chiral atom count; SA: synthesis accessibility; NHOH_C: NH/OH group count; HetA_C: heteroatom count; RotB_C: rotatable bond count; BriA_C: bridgehead atom count.

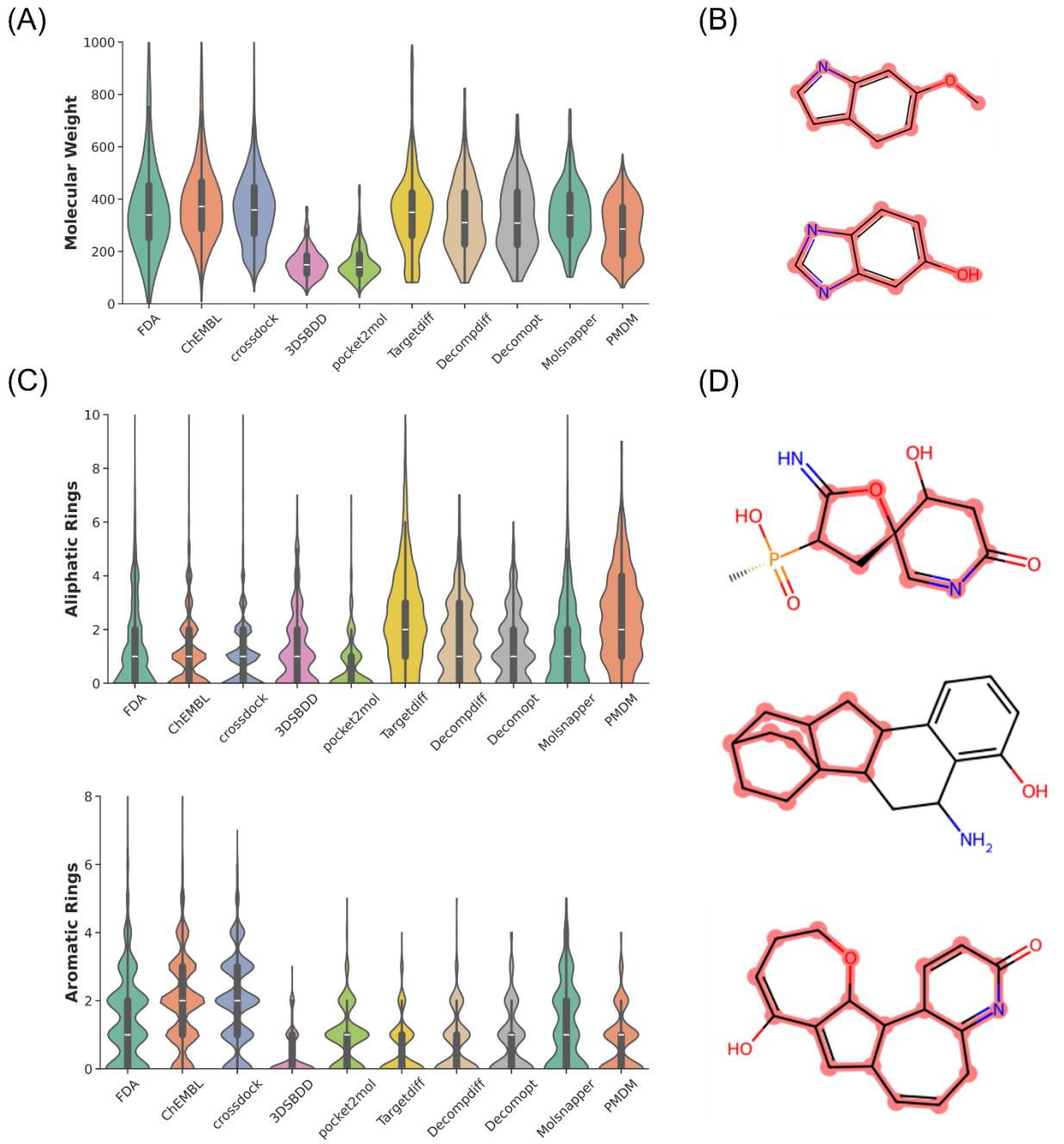


Figure 5. Chemical property distributions show difference between algorithm generated molecules and control datasets. Corresponding chemically implausible structures are selected from molecules that didn't pass all three metrics, substructures associated with the distributions are highlighted in red. (A) The molecular weight distribution across datasets. (B) Examples of molecules with low molecular weight. (C) Number of aliphatic rings and aromatic rings distributions across datasets. (D) Example of molecules with multiple aliphatic rings.

Distribution analysis reveals varying preferences among generative algorithms in molecule generation

Rather than visually inspect generated molecules one by one, to better understand why structure-based 3D generative models perform worse than control groups on the metrics, we first plotted the distributions of several structural properties and descriptors and compared them with control datasets. Overall, the control groups exhibited similar distributions, whereas the distributions of the generated datasets differed from those of the control group (Figure 4). Several example structures are selected from generated molecules that don't pass ChEMBL ring metrics we applied and showed next to the distribution. Since the aim is to highlight difference, plots with similar distributions are provided in Supporting Information (Figure S2). As noted, the 3DSBDD and Pocket2mol algorithms generated relatively small and simple molecules, as indicated by the molecular weight distributions and associated example structures (Figure 5A, Figure 5B). Additionally, most algorithms produced fewer aromatic rings, but more aliphatic rings compared to the three control groups, suggesting a preference for aliphatic ring structures (Figure 5C, Figure 5D).

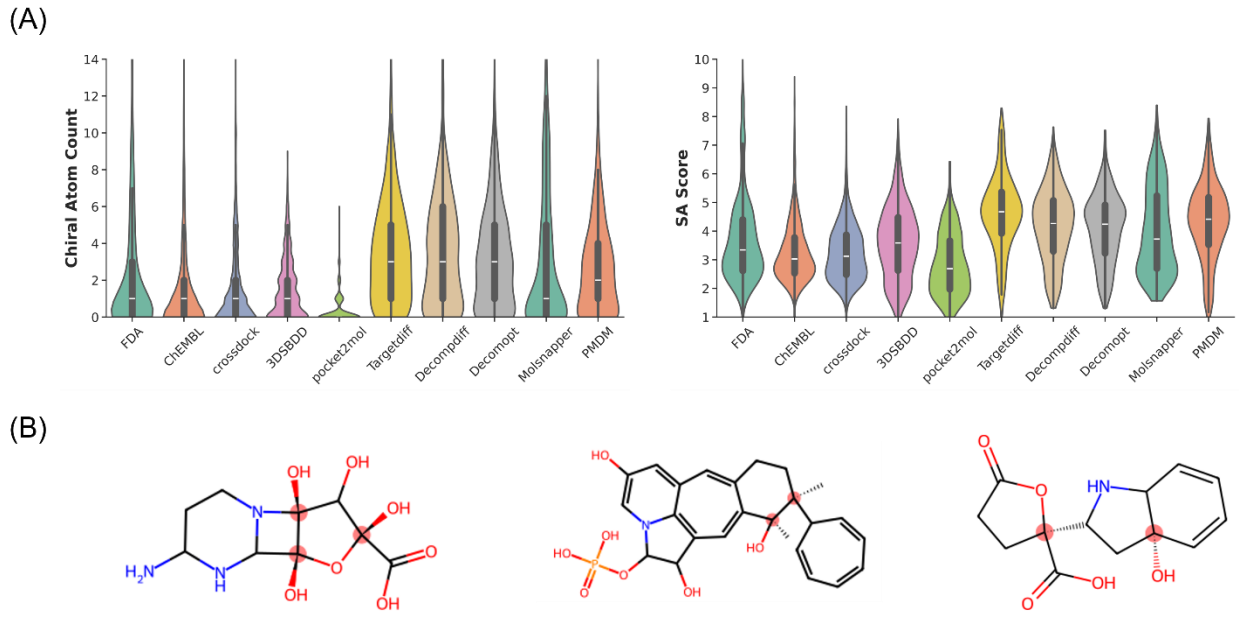


Figure 6. Distributions that are associated with structural complexity show difference between algorithm generated molecules and control datasets. Corresponding chemically implausible structures are selected from molecules that didn't pass all three metrics,

substructures associated with the distributions are highlighted in red. (A) Chiral atom count and synthetic accessibility (SA) score distributions across dataset. (B) Example of molecules with multiple chiral atoms.

In terms of structural complexity, almost all algorithms, except Pocket2mol, generated more chiral centers than the control groups (Figure 6A), indicating a tendency toward more complex molecular structures, which can be more challenging to synthesize. More chiral centers may also contribute to higher synthetic accessibility scores as Figure 6 shows, which indicate that they are harder to synthesize.

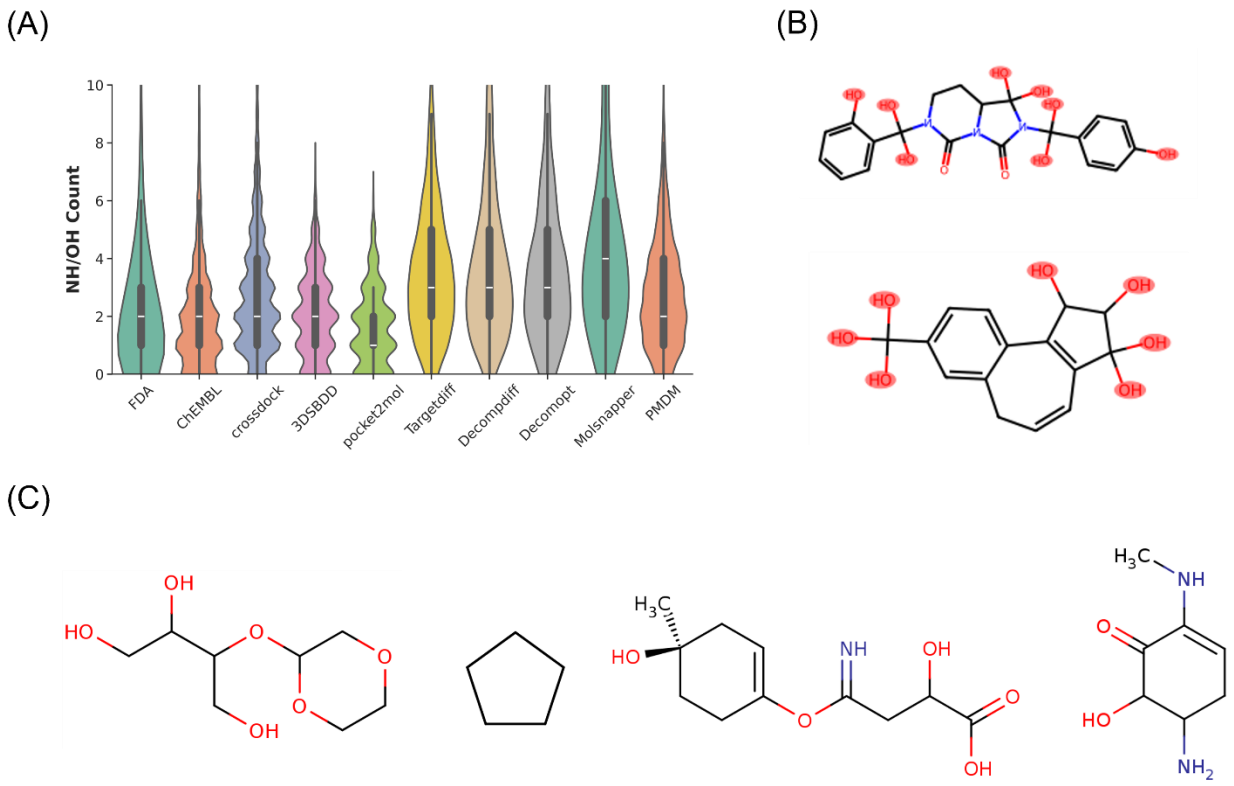


Figure 7. Specific functional groups distribution and example molecules. (A) distribution of NH and OH group count across datasets reveals the algorithm prefers generating NH and OH groups, substructures associated with the distributions are highlighted in red. (B) Example of molecules with multiple OH/NH groups. (C) Four molecules that could pass all of metrics we apply but still not druglike or hard to synthesize.

Regarding specific functional groups, almost all algorithms, aside from Pocket2mol and 3DSBDD, generated molecules with a higher frequency of NH and OH groups than the controls (Figure 7A, Figure 7B). Although these are common in drug molecules, an excess

of these groups within a single molecule may impact permeability and other pharmacokinetic (PK) profiles. Moreover, the presence of OH and NH groups may contribute to artificial high docking scores by enhancing hydrogen-bonding scores.

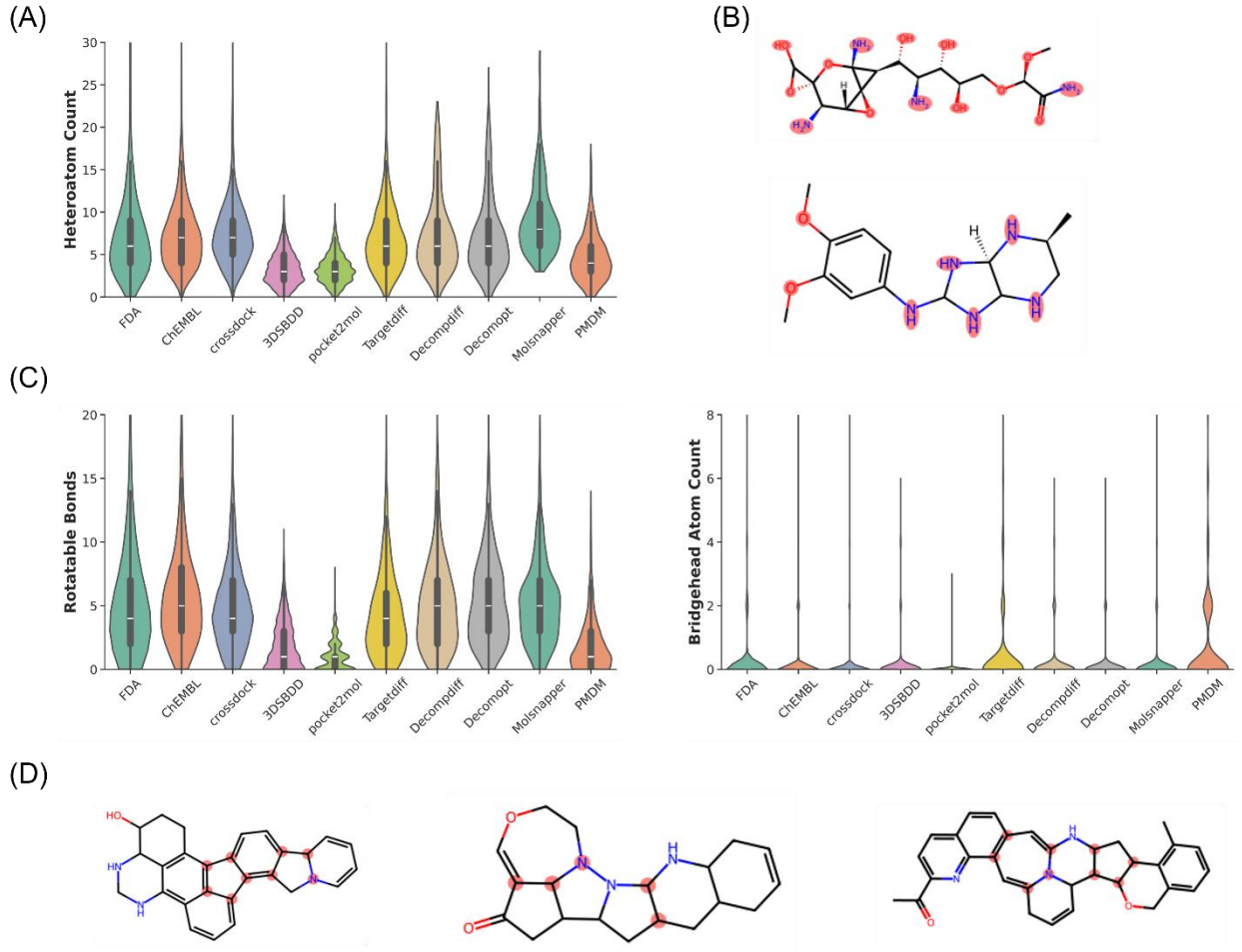


Figure 8. Distributions that provide specific information for single algorithm. Corresponding chemically implausible structures are selected from molecules that didn't pass all three metrics and, substructures associated with the distributions are highlighted in red.. (A) Molsnapper tends to generate more heteroatoms than other algorithms and control datasets. (B) Example of molecules with multiple heteroatoms. All molecules are generated by the Molsnapper algorithm. (C) PMDM generates more bridged molecules with fewer rotatable bonds. (D) Example molecules with multiple bridgehead atoms. All molecules are generated by the PMDM algorithm.

Besides overall distributions across all benchmarked algorithms, focused inspection of each algorithm also provides specific information. 3DSBDD and pocket2mol tend to

generate simple molecules, thus perform better in the three metrics we apply (Figure 4, Figure 5A). Molsnapper tends to generate more heteroatoms (Figure 8A, Figure 8B), and PMDM prefers complex ring structures with fewer rotatable bonds (Figure 8C, Figure 8D). A detailed structural and property distribution analysis could provide more information about the potential weaknesses of algorithms and determine how to improve them.

Notably, not all of the chemically unplausible molecules are flagged by three metrics we used; for instance, four structures in Figure 7C pass all three metrics but are far from druglike or would be challenging to synthesize. Therefore, it is necessary to conduct analysis of the overall distribution of structural and calculated properties, and direct inspection of individual molecules. Combined with metrics, they will provide a more comprehensive understanding of each algorithm's characteristics.

In summary, we proposed two novel metrics based on the occurrence of BM scaffolds from generated molecules within the druglike molecules from ZINC20 and ZINC22 databases. When combined with the ring occurrence metric in ChEMBL, this approach successfully distinguishes between the three control groups and the seven structure-based 3D generative models, indicating that some algorithms may produce chemically implausible molecules. Further structural and property analyses, along with visual inspection, highlight potential weaknesses across all algorithms and within specific ones. Our proposed metrics, together with the analyses presented in this paper, offer valuable insights for model developers to improve these algorithms and enhance their utility in drug discovery. The code and data used in the paper can be found at https://github.com/Kartinaa/Benchmarking_gene_model.

Method:

Data collection: FDA approved small molecule drugs control set was extracted from DrugBank 5.1.12 database. After downloading the full database, a dedicated python script was used to extract the approved small molecule drugs as InChI format and save locally. The ChEMBL clinical small molecules control set was downloaded from the ChEMBL 34 database using website filters to exclude approved small molecules, ensuring no

redundancy with the FDA-approved control set. The CrossDocked2020 control dataset was created by converting all 3D ligands from the CrossDocked2020 dataset into SMILES format and removing duplicates. All control datasets are made publicly available at https://github.com/Kartinaa/Benchmarking_gene_model.

Regarding the datasets used in the metrics: ZINC20 druglike molecules (0.87 billion) were downloaded from the ZINC20 database's Drug-like subset with the "Wait OK" label. ZINC22 druglike molecules (54.6 billion) were downloaded from the drug-like subsets in 2D trenches from the ZINC22 database. It is important to note that many of the synthesis-on-demand molecules in ZINC have not yet been synthesized. Therefore, their synthesizability is not guaranteed.

Molecule generation: As outlined in the manuscript, all algorithms compared in this study use the same training and test sets. The goal was to generate 10 molecules for each protein in the test set using pre-trained models, theoretically resulting in a total of 1,000 molecules per algorithm (10 molecules × 100 proteins in the test set). However, most algorithms did not generate exactly 1,000 molecules, and ensuring precisely 10 molecules per test protein was challenging. Since this study focuses on evaluating the chemical plausibility of the generated structures, the reasons for this variability are not discussed further.

All molecule generations were conducted using the default configurations or those recommended by the respective authors. To ensure reproducibility, we briefly outline the parameters used for several algorithms: For Molsnapper, the `num_pharma_atoms` flag was set to `20` and `clash_rate` was set to 0.1, as indicated in their GitHub repository. For Decompdiff, the `prior_mode` flag was set to `ref_prior` following configuration file. For Decompopt, the `dup_num` variable in `sample_compose.sh` was set to 0, avoiding re-docking molecules into the binding pocket to improve generation efficiency. While this setting may reduce the performance of generated molecules on docking-based metrics, we expect that it does not significantly impact the chemical plausibility of the generated structures.

Metrics incorporation and property calculations: For ring system-based metrics, ring systems for each molecule were extracted using `useful_rdkit_utils` python library,

which incorporates the ring system frequency data from the ChEMBL database. For the ZINC20 and ZINC22 druglike sets, ring system frequencies were calculated using the same Python library and served as baselines for comparison with the ring system frequencies of the generated molecules.

For BM scaffold-based metrics, the BM scaffolds were extracted from the ZINC20 and ZINC22 druglike sets using the RDKit Python library. After removing redundancies, the BM scaffolds from each dataset were saved locally and used as baselines to evaluate BM scaffold occurrences in the generated molecules. All chemical properties were calculated using the RDKit python library, and visualizations were created using Matplotlib, Seaborn, and RDKit libraries.

Author Contributions

The manuscript was prepared with contributions from all authors who have approved the final version.

Notes

The authors declare no competing financial interest.

Acknowledgments

We sincerely thank Dr. Patrick Walters for his valuable suggestions and insights, which greatly improved this manuscript. We also extend our gratitude to Dr. Brian Shoichet, Dr. Jean-Louis Reymond, Dr. Conor Colley, Yu Chen, and Ye Buehler for their helpful discussions. This work was supported by NIH R01 awards (GM129431/GM143370) and the AnalytiXIN fellowship (to J.L.).

Refence:

(1) Sertkaya, A.; Beleche, T.; Jessup, A.; Sommers, B. D. Costs of Drug Development and Research and Development Intensity in the US, 2000-2018. *JAMA Network Open* **2024**, *7* (6), e2415445-e2415445. DOI: 10.1001/jamanetworkopen.2024.15445 (acccessed 10/10/2024). Singh, N.; Vayer, P.; Tanwar, S.; Poyet, J.-L.; Tsaioun, K.; Villoutreix, B. O. Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery* **2023**, *3*, Mini Review. DOI: 10.3389/fddsv.2023.1201419.
(2) Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y. A. M. M.; Gomaa, M. M.; Hassanien, A. E. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review* **2023**, *56* (7), 5975-6037. DOI: 10.1007/s10462-022-10306-1.
(3) Khan, W.; Daud, A.; Khan, K.; Muhammad, S.; Haq, R. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal* **2023**, *4*, 100026. DOI: https://doi.org/10.1016/j.nlp.2023.100026.
(4) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* **1996**, *16* (1), 3-50. DOI: 10.1002/(sici)1098-1128(199601)16:1<3::Aid-med1>3.0.Co;2-6 From NLM.
(5) Alakhdar, A.; Poczos, B.; Washburn, N. Diffusion Models in $\textit {De Novo} $ Drug Design. *arXiv preprint arXiv:2406.08511* **2024**.
(6) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat Chem* **2012**, *4* (2), 90-98. DOI: 10.1038/nchem.1243 From NLM.
(7) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009**, *1* (1), 8. DOI: 10.1186/1758-2946-1-8.
(8) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *Journal of Chemical Information and Modeling* **2022**, *62* (12), 2973-2986. DOI: 10.1021/acs.jcim.2c00038. Zhu, W.; Wang, Y.; Niu, Y.; Zhang, L.; Liu, Z. Current Trends

and Challenges in Drug-Likeness Prediction: Are They Generalizable and Interpretable? *Health Data Science* **2023**, *3*, 0098. DOI: doi:10.34133/hds.0098.
(9) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science* **2024**, *15* (9), 3130-3139.
(10) Ziv, Y.; Marsden, B.; Deane, C. M. MolSnapper: Conditioning Diffusion for Structure Based Drug Design. *bioRxiv* **2024**, 2024.2003.2028.586278. DOI: 10.1101/2024.03.28.586278.
(11) Walters, P. Generative Molecular Design Isn't As Easy As People Make It Look. 2024.
(12) Liu, H.; Qin, Y.; Niu, Z.; Xu, M.; Wu, J.; Xiao, X.; Lei, J.; Ran, T.; Chen, H. How Good are Current Pocket-Based 3D Generative Models?: The Benchmark Set and Evaluation of Protein Pocket-Based 3D Molecular Generative Models. *Journal of Chemical Information and Modeling* **2024**. DOI: 10.1021/acs.jcim.4c01598.
(13) Zhou, X.; Cheng, X.; Yang, Y.; Bao, Y.; Wang, L.; Gu, Q. DecompOpt: Controllable and Decomposed Diffusion Models for Structure-based Molecular Optimization. *arXiv preprint arXiv:2403.13829* **2024**.
(14) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.-C.; et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nature Communications* **2024**, *15* (1), 2657. DOI: 10.1038/s41467-024-46569-1.
(15) Walters, P. Mining Ring Systems in Molecules for Fun and Profit. 2022.
(16) Luo, S.; Guan, J.; Ma, J.; Peng, J. A 3D generative model for structure-based drug design. *Advances in Neural Information Processing Systems* **2021**, *34*, 6229-6239.
(17) Peng, X.; Luo, S.; Guan, J.; Xie, Q.; Peng, J.; Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, 2022; PMLR: pp 17644-17655.
(18) Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543* **2023**.
(19) Guan, J.; Zhou, X.; Yang, Y.; Bao, Y.; Peng, J.; Ma, J.; Liu, Q.; Wang, L.; Gu, Q. DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design. In Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research; 2023.
(20) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling* **2020**, *60* (9), 4200-4215. DOI: 10.1021/acs.jcim.0c00411.
(21) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39* (15), 2887-2893. DOI: 10.1021/jm9602928.

Supporting information:

	Number of molecules	Number of valid molecules	Number of unique molecules (Inchi)	Number of molecules don't have ring system
FDA approved small molecules	2,778	2,778	2,651	345(13.0%)
ChEMBL clinical small molecules	8,236	8,236	7,700	439(5.7%)
CrossDocked dataset	11,759	11,759	11,602	720(6.2%)
3D_generative_SBDD	1,750	1,732	1,501	374(24.9%)
Pocket2mol	1,345	1,345	1,105	167(15.1%)
Targetdiff	989	989	965	97(10.0%)
Decompdiff	884	884	854	83(9.8%)
Decompopt	913	913	887	93(10.5%)
Molsnapper	849	849	791	54(6.8%)
PMDM	1,000	989	982	27(2.7%)

Table S1. Number of molecules generated by each model as well as three control groups. Number of valid molecules and number of unique molecules based on Inchi key are also calculated.

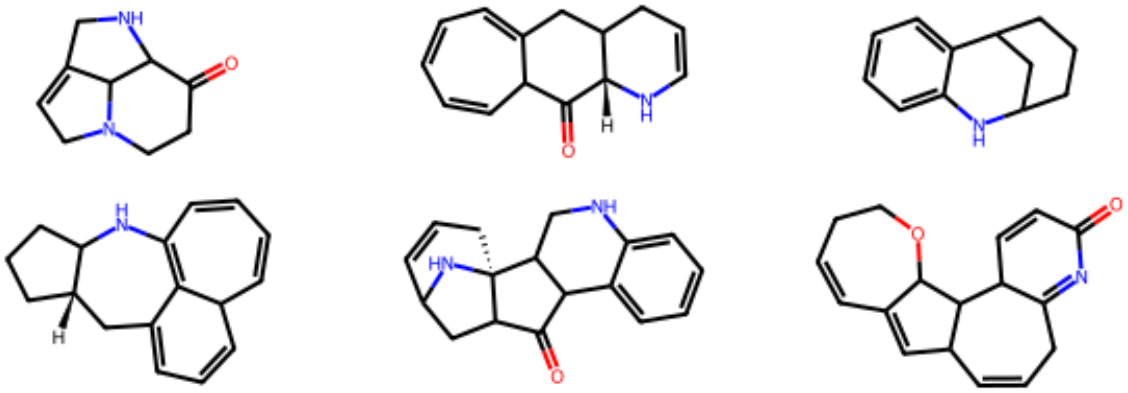


Figure S1. Examples of ring systems that appear zero time in ChEMBL database.

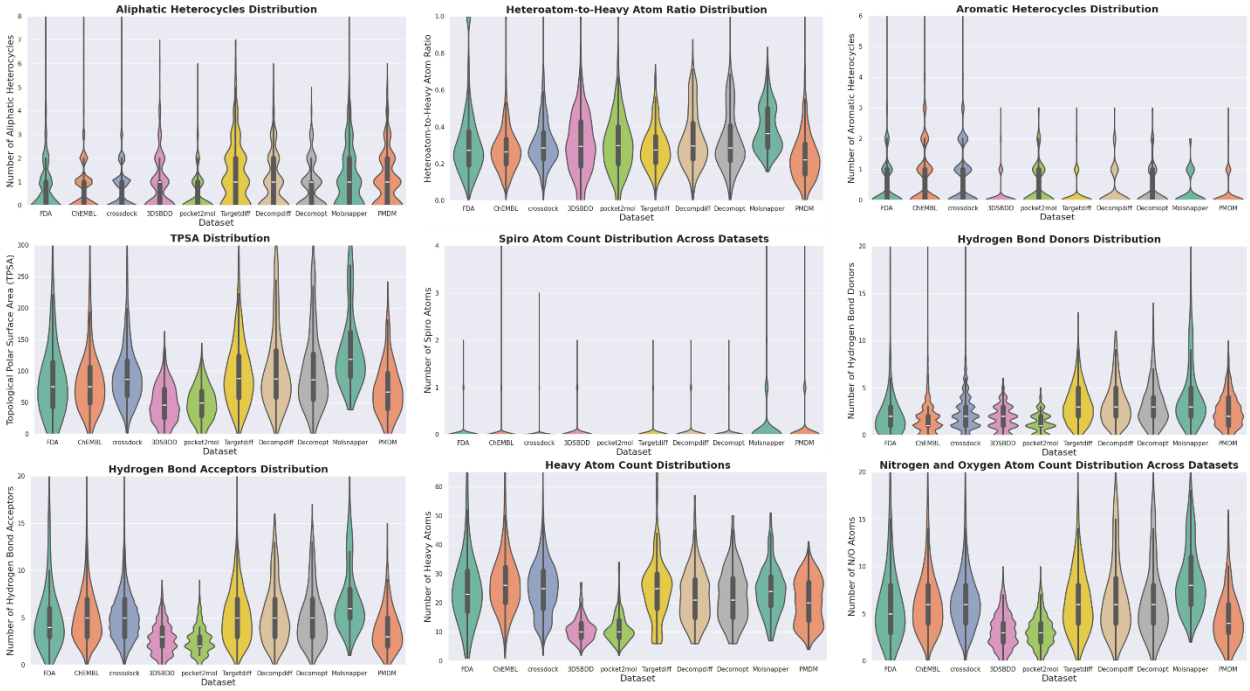


Figure S2. Other distributions calculated.

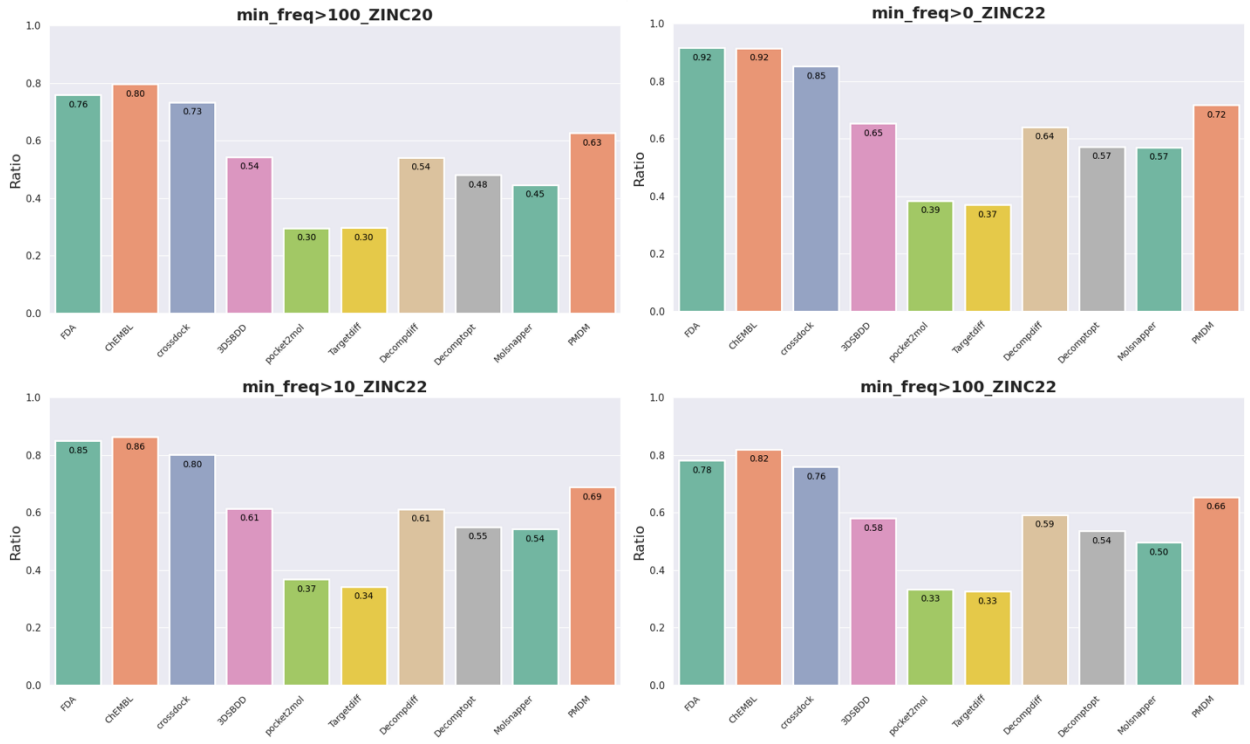


Figure S3. Benchmarking result by ring system metric using different thresholds and dataset. The conclusion remains consistent: ring system metric could differentiate between control groups and molecules generated by algorithms.

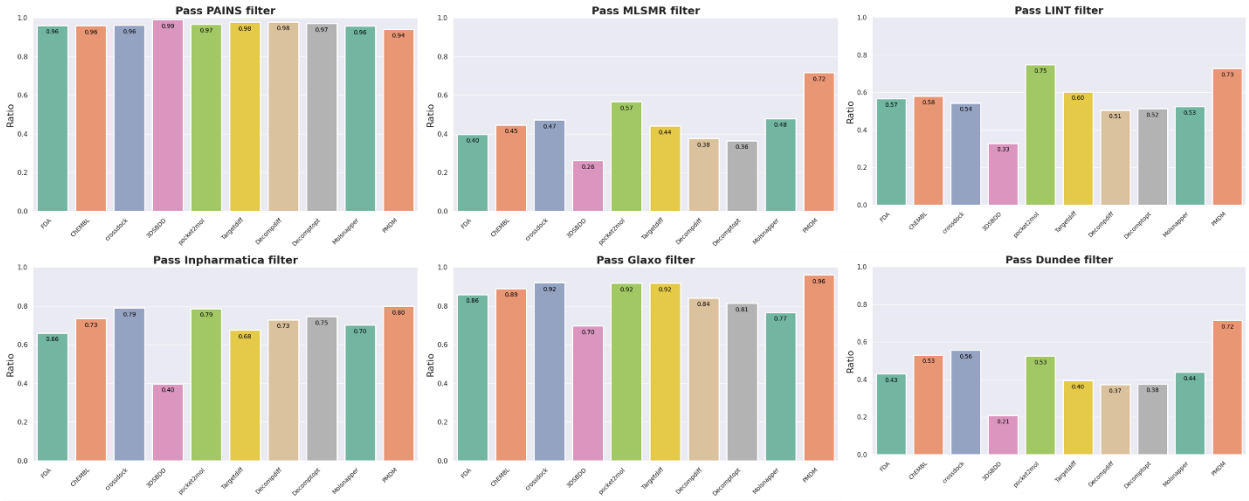


Figure S4. Commonly applied filters fail to differentiate between control groups and molecules generated by generative algorithms.

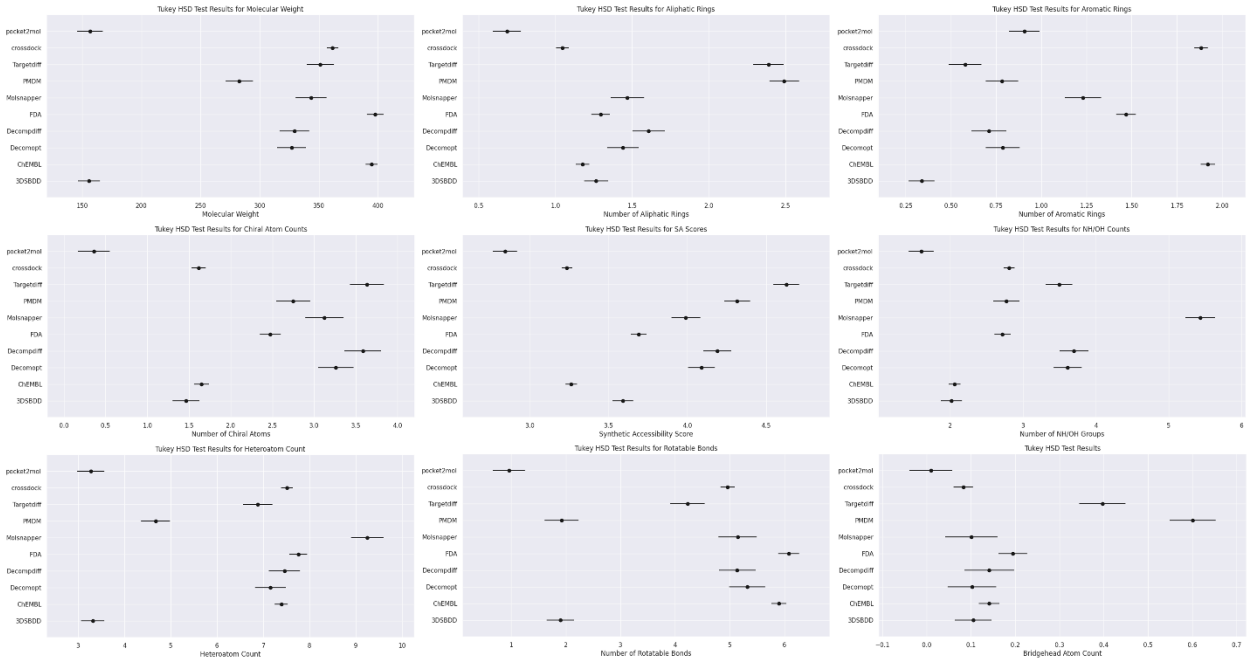


Figure S5. Tukey HSD test results for each distribution plot.