How good are current pocket based 3D generative models? : The benchmark set and evaluation on protein pocket based 3D molecular generative models

Haoyang Liu[a,b#], Yifei Qin[c#], Zhangming Niu[e,f,g#], Mingyuan Xu[b], Jiaqiang Wu[c], Xianglu Xiao[f,g,h], Jinping Lei[i,*], Ting Ran[b*], Hongming Chen[b,d,j*]

[a] State Key Laboratory of Medicinal Chemical Biology and College of Life Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China
[b] Division of drug and vaccine research, Guangzhou National Laboratory, Guangzhou 315000, Guangdong, China
[c] School of pharmacy and food engineering, Wuyi University, Jiangmen 529020, Guangdong, China
[d] School of Basic Medical Sciences, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou 511436, China
[e] National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK
[f] MindRank AI, Hangzhou, Zhejiang, China
[g] AI research center, MindRank Technologies Limited, London, UK
[h] Bioengineering Department and Imperial-X, Imperial College London, London W12 7SL, UK
[i] School of Pharmaceutical Science, Sun Yat-sen University, Guangzhou 510006, China

[j] Lead contact

[#] Contributed equally

[*]Correspondence:
Hongming Chen (chen_hongming@gzlab.ac.cn); Ting Ran (ran_ting@gzlab.ac.cn); Jinping Lei (leijp@mail.sysu.edu.cn)

Abstract

The development of three-dimensional (3D) molecular generative model based on protein pockets has recently attracted a lot of attentions. This type of model aims to achieve the simultaneous generation of molecular graph and 3D binding conformation under the constraint of protein binding. Various pocket based generative models have been proposed, however, currently there is

a lack of systematic and objective evaluation metrics for these models. To address this issue, a comprehensive benchmark dataset, named as POKMOL-3D, is proposed to evaluate protein pocket based 3D molecular generative models. It includes 32 protein targets together with their known active compounds as a test set to evaluate the versatility of generation models to mimick the real-world scenario. Additionally, a series of 2D and 3D evaluation metrics was integrated to assess the quality of generated molecular structures and their binding conformations. It is expected that this work can enhance our comprehension of the effectiveness and weakness of current 3D generative models, and stimulate the discussion on challenges and useful guidance for developing next wave of molecular generative models.

Introduction

Application of deep generative model in drug design has gained widespread attention. Over the past few years, a large number of molecular generative models based on 1D/2D structures have been reported. These models mainly generate molecules by learning the structural features embodied in either 1D strings, such as the Simplified Molecular Input Line Entry System (SMILES) strings [1] and SELFIE strings[2] , or 2D molecular graphs[3, 4]. Despite substantial progress being made in improving the validity of generated molecules and the efficiency of exploring the drug-like chemical space, most of these models overlook the rich information contained in the 3D conformation of molecules. Indeed, the binding affinity of a drug molecule to the target protein is predominantly dependent on the degree of geometrical and electrostatic complementarity between their 3D conformations. Therefore, in recent two years, molecular generative models based on 3D conformations has become a hot research area.

Currently, 3D molecular generative models can be divided into three categories. One class aims to generate 3D conformations of a given 2D molecule graph[5-10]. The second class is to generate simultaneously 3D conformations and 2D graph of a molecule, which doesn't consider its binding partner, i.e. the protein pocket[11-15]. The third class strives to generate 3D conformations and 2D graph under the constrain of protein binding pocket, which has attracted most interest in the latest two years and is the ultimate goal of the so called structure based de novo molecular design. In current study, we solely focus on the third type of 3D generative model.

The LiGAN model pioneered the field by introducing a conditional variational autoencoder of 3D atomic density grids[16], in which a protein pocket is encoded by a conditional encoder network. The output atomic density grids are transformed to a 3D molecular structure using a rule-based atom fitting and bond inference algorithm. The main drawback of this model lies in its inability to maintain the equivariance on rotation and translation. In contrast, SBDD[17] and GraphBP[18] leverage a 3D equivariant graph neural network (GNN) to solve this problem. They both sequentially place atoms to a given 3D binding site, utilizing protein pocket and ligand atoms generated in previous steps as the contextual information. For GraphBP model, an anchor atom must be selected to determine position of the next generated atom, while SBDD generates the next atom on a given arbitrary position in the pocket. However, they only generate atom types and positions and utilize third-party software such as RDKit to construct bond types. Pocket2Mol[19] introduces a geometric vector perceptron (GVP)-based equivariant GNN[20, 21] to encode the 3D geometric information of protein pocket and existing fragments of the ligand in the pocket. Compared to GraphBP, it particularly involves a predictor network to infer the bond type. ResGen, which utilizes similar GVP-based architecture, further encodes protein and ligand at residue and

atom levels respectively to better capture high-level binding interactions. Different from learning the joint distribution of atom type and bond type in Pocket2Mol, ResGen decomposes the distribution as a product of multiple conditional distributions for anchor atom, atom position, atom type and bond type. PocketFlow[22] adds a layer of geometric bottleneck perceptron (GBP) to the GVP network to improve model speed and enhance information integration. It is also characterized by its AtomFlow and BondFlow modules for predicting atom type and bond type, respectively. Especially, chemical knowledge such as bond valence is explicitly integrated to guide the bond inference. SurfGen[23] represents binding pocket as protein surface and utilizes a special framework Geodesic-GNN to learn the distribution of the topological information on the surface. All these GNN-based models can be categorized as autoregressive model, varying in the way of encoding protein pocket and the decoding or sampling of atoms in the generative process.

Recently, diffusion model is an emerging deep learning technology utilizing an iterative denoising process to map noise to data and have been used for 3D molecule generation. DiffSBDD is the first 3D conditional graph diffusion model[24], in which protein pocket nodes transformed from atomic point clouds are used as conditional constraints and remain unchanged throughout the reverse diffusion process. TargetDiff is conceptually similar to DiffSBDD but employs a different diffusion formalism for the categorical atom types. Both DiffSBDD and TargetDiff map protein and ligand nodes into a joint embedding space for noise prediction, while in DiffBP[25] these two types of node are separately embedded. In addition, DiffBP introduces a new loss term to regulate the intersection between protein and ligand nodes in space. Besides, language models have also been reported for 3D structure generation. Feng *et al*. [26]developed Lingo3DMol that combines transformer-based language model architecture and deep geometric

learning technology for 3D molecular generation. A prior model was pre-trained to generate 3D molecular structures given a fragment-based SMILES string, and then fine-tuned based on protein-ligand complex data. The protein pocket and ligand embeddings are used as the input for encoder and decoder respectively.

Despite various pocket-based 3D molecular generative models reported, there is still lack of unified and comprehensive benchmark metrics to objectively evaluate the quality of generated molecules. Early-developed models, such as GraphBP and SBDD, primarily relied on common 2D/3D molecular evaluation metrics such as molecular validity, molecular docking score, druglikeness (QED)[27], synthesizability score (SAscore)[28], and structural diversity to assess the quality of generated 3D molecules. Pocket2Mol additionally performed analysis on ring size of molecules as part of quality measurement. Although TargetDiff, DiffSBDD and DiffBP were published later than Pocket2Mol, they still adopted 2D molecular evaluation metrics. ResGen and SurfGen introduced additional 2D metrics, e.g. the mean similarity between generated molecules and known active molecules, to quantify their efficiency of generating active compounds. Lingo3DMol analyzed the proportion of targets in which nearest neighbors of known active compounds can be generated. In terms of measuring the quality of 3D conformation, the Jensen-Shanon[29] divergences of bond length, bond angle, and dihedral angle of generated molecules, and docking scores of redocked compounds in binding pocket are often used as the metrics. While Lingo3DMol used "min-in-place" GlideScore[30] to evaluate the poses after minimization of the generated conformations within the pocket. In addition, ResGen and SurfGen employ extra 3D evaluation metrics such as *in situ* docking score, similarity of protein-ligand interaction fingerprints between the generated conformation and known actives, and 3D similarity

index based on the overlay between the generated conformations and ground-truth conformations.

On the other hand, most of the pocket-based 3D molecular generative models were trained and tested on the Crossdock2020 dataset[31], which is constructed by molecular docking of active ligands on PDB database to its corresponding targets. ResGen, SurfGen, and PocketFlow additionally utilized the protein pockets outside the Crossdock dataset to assess the model's potential for real-world application. In ResGen, two external pockets were selected for evaluation, while in SurfGen the evaluation was extended to 20 therapeutic targets. In PocketFlow, the authors synthesized two generated molecules for wet-lab validation, which experimentally validates the effectiveness of the 3D generative models in hit finding scenario. Particularly, the resolved crystal structures showed that the generated 3D conformations are highly similar to their active binding conformations. So far, most of the pocked-based 3D generative models employ Pocket2Mol as the baseline model for comparison, and the number of model included in their evaluation is relatively small and incomplete. For future model development, it is probably necessary to conduct a performance comparison among a larger model set under the same criteria. PoseCheck[32] is a small-scale benchmark study for this task by comparing five models including LiGAN, Pocket2Mol, 3D-SBDD, Pocket2Mol, TargetDiff and DiffSBDD. PoseCheck focused on 3D conformation evaluation using the CrossDock dataset as test set, and employed four 3D based evaluation metrics, namely steric clashes based on van der Waals distance, protein-ligand interaction fingerprints, strain energy of the generated conformations, and conformation similarity between generated and docked poses. However, PoseCheck ignored 2D metrics that can also imply the general quality of molecular structures. DrugPose is another small-scale benchmark study focusing on 3D molecular generative models, but non-specific for the protein pocket based

methods[33]. The binding similarity between pre- and post-docked poses of generated molecules, drug-likeness and synthesizability were also analyzed. Zheng *et al*. recently conducted a cross-algorithm benchmark study[32] which compared a few protein pocket based 3D molecular generative models with 1D SMILES/SELFIES and 2D molecular graph based generative methods. Despite 16 models were evaluated in their study, only several commonly used metrics, such as docking score, QED, molecular validity, were employed for the evaluation.

This study provides a comprehensive and systematic evaluation on nine 3D molecular generative models in 32 protein pockets, and the compiled benchmark dataset is called POKMOL-3D. In terms of evaluation metrics, both 2D and 3D metrics were considered and classified according to their characteristics. Given that the essence of pocket-based 3D molecular generation model is to generate molecules being able to bind specified targets, and the generated conformations should be close to their active conformations, conventional 2D and 3D evaluation metrics were expanded to include new parameters characterizing sampling speed, actives recovery and conformation quality etc. Furthermore, the widely used SMILES based generative model REINVENT [34, 35] was included as the baseline model for comparison on the 2D based metrics, providing an interesting perspective on how good current 3D based models comparing with classical SMILES based model. In summary, this work could provide a systematic and comprehensive benchmark set for evaluating 3D generative model.

Method

Model selection

Nine representative models were selected from recently published protein pocket-based 3D

169 molecular generative models spanning 2021 to 2024, which includes four distinct categories:
170 graph model, diffusion model, language model, and flow model. To demonstrate the efficiency of
171 3D generative model, the SMILES based REINVENT (version 4.0) was utilized as the base line
172 model.

Table 1. List of selected 3D molecular generative models

Model	Generative process	Model architecture	Training Set	Year
SBDD	Autoregressive	Graph Model	CrossDock2020	2021
GraphBP	Autoregressive	Graph model	CrossDock2020	2022
Pocket2Mol	Autoregressive	Graph model	CrossDock2020	2022
DiffBP	One-shot	Diffusion Model	CrossDock2020	2022
SurfGen	Autoregressive	Graph model	CrossDock2020	2023
TargeDiff	One-shot	Diffusion Model	CrossDock2020	2023
ResGen	Autoregressive	Graph model	CrossDock2020	2023
Lingo3DMol	Autoregressive	Language Model	PDBbind [36]	2024
PocketFlow	Autoregressive	Flow model	CrossDock2020	2024

173

174 **POKMOL-3D dataset**

175 In order to assess the versatility of selected models, 32 protein targets belonging to diverse
176 protein families were selected, in which five classes of target are included: kinases, non-kinase
177 enzymes, GPCRs, nuclear receptors, and protein-protein interaction targets. Given that our goal is
178 to evaluate model performance on generating molecules conditioned on the 3D information of
179 protein pocket, the targets that possess published protein-ligand complex structures were chosen.

180 For each target, one crystal structure from the RCSB PDB database[37], whose resolution is less than
181 3Å, was retrieved and only the subunit containing ligand was kept for analysis when the structure
182 comprises multiple subunits. All protein structures were optimized using the Protein Preparation
183 Wizard (PrepWizard) module in Maestro[38].

184 Moreover, active compounds of these proteins were extracted from the ChEMBL database[39]
185 and served as reference set for evaluation metrics. Active molecules were considered eligible if
186 they have a molecular weight less than 500 Da, and the target IC_{50}/EC_{50} values are less than 10
187 nM or Ki/Kd values less than or equal to 100 nM.

188 **Molecular generation**

189 In this study, the latest version of the nine models were downloaded from GitHub. During the
190 sampling process, we adhered to the default configurations for all models, with the sole exception
191 of the sampling scale which was specifically calibrated to yield 2000 molecules per run. Each
192 model was tasked with generating a minimum of 2000 molecules per target. In scenarios where
193 sampling 2000 molecules in a single run was unfeasible for certain targets, maximal three
194 sampling runs were done to expand the generation set. For molecular generation employing the
195 REINVENT model, 1000 epochs of RL, steered by the molecular docking score (GlideScore),
196 were conducted to mimic the scenario of molecular generation within protein pocket. All
197 molecules generated during the RL process were subsequently utilized for further analysis.

198 **Evaluation metrics**

199 In current study, five categories of metrics were proposed to evaluate the performance of 3D
200 molecular generative models conditioned on protein pocket. As shown in Figure 1, it encompass
201 model quality, general molecular quality, structural properties, recovery of active molecules, and

target binding related scores. Specifically, the model quality class evaluates the sampling speed and the target versatility of the model. The following two classes primarily focus on the quality of molecular structure, providing insights into the overall molecular properties, 2D topological and 3D related properties of the generated molecules. The last two classes of metrics evaluate the effectiveness of generating target binding compounds, at some extent reflecting the probability of being able to bind the target protein. Through this pool of metrics, we hope to provide a thorough benchmarking on current state-of-the-art 3D pocket-based molecular generative models, more importantly providing guidance for developing new algorithms in future.

Figure 1. Five types of evaluation metrics for compared generative models.

Model quality

To evaluate the model quality, three metrics, *i.e.* target failure rate, sampling success rate and sampling speed, were proposed. As shown in equation 1, target failure rate refers to the proportion of targets for which no molecules can be generated by the model within three sampling runs.

$$\text{target failure rate} = \frac{\text{targets wihout molecules generated}}{all\ targets} \qquad (1)$$

As shown in equation 2, sampling success rate refers to the proportion of targets for which the model can generate more than 2000 molecules within three runs.

$$\text{sampling success rate} = \frac{targets\ with\ more\ than\ 2000\ molecules\ generated}{Total\ number\ of\ targets} \qquad (2)$$

In addition, the sampling speed was defined as the average time (in seconds) required for generating one molecule. Here, the sampling time was counted for generating 100 molecules for each target under the same computational resource. For calculating the sampling speed, the employed computation resource was a linux workstation of 16-core 24GB RAM Intel Xeon Platinum 8358 2.60GHz CPU and a NVIDIA GeForce 3090 GPU.

General molecular quality

The general molecular quality set includes molecular validity, uniqueness, usability, drug-likeness, synthetic score and target based diversity. These are properties reflecting overall generation set. The calculation of properties was carried out using RDKit package[40]. Molecular validity, as defined in equation 3, refers to the proportion of valid molecules within the generated set. Molecules that can successfully go through the standardization process are considered valid.

$$\text{validity} = \frac{valid\ molecules}{generated\ molecules} \qquad (3)$$

As shown in equation 4, uniqueness refers to the proportion of unique molecules obtained after removing duplicates among the valid molecules.

$$\text{uniqueness} = \frac{unique\ molecules}{valid\ molecules} \quad (4)$$

As shown in equation 5, usability refers to the proportion of molecules containing common elements C, N, O, P, S, F, Cl, Br, I, and H. Molecules containing other elements are considered unusable, for example those containing metal elements.

$$usability = \frac{usable\ molecules}{unique\ molecules} \quad (5)$$

Furthermore, the drug-likeness score (QED) and synthetic accessibility score (SAscore) were calculated for each generated molecule using RDKit. These scores assess the drug-likeness and synthetic feasibility of the generation set, respectively.

Target based molecular diversity (TDiv) is computed utilizing equation 6 to represent the mean value of target specific diversity of generation sets:

$$\text{TDiv} = \frac{1}{\text{N}_{\text{target}}} \sum_{t=1}^{N_{target}} \left(1 - \frac{\sum_{i=1}^{N_t} \sum_{j=2}^{N_t} T_{sim}(i,j)}{{N_t}^2}\right), \text{ i} < \text{j} \quad (6)$$

where i and j refer to the indexes of two molecules in the generated molecule set for target t. The Tanimoto similarity (T_{sim}) of the Morgan fingerprints[41] is calculated based on all the pairs of molecules for the same target. This similarity is normalized on the total number of molecules in the generation set of the target. Target specific diversity score is derived from this normalized similarity. The final TDiv score is defined as the average diversity score across all targets. A higher TDiv value indicates greater diversity.

Structural properties

The structural property group includes a set of 2D topological descriptors comprising the number of heavy atom, chiral atom, ring, aromatic ring and rotatable bond, and the fraction of sp3 hybridized carbon atom (Fsp3), and their distribution was also compared. Additionally, a set of 3D

252 based geometrical properties containing the Jensen-Shanon divergency (JSD)[29] of bond length,
253 bond angle and dihedral angle was calculated. Detailed information for selected types of bond,
254 bond angle and dihedral angle can be found in Supporting Material Figure S1. The geometry
255 comparison was made between the conformations generated by the model and the low-energy
256 conformations optimized using the LigPrep module of Schrodinger package (version 2020). As
257 proposed in previous works[22], JSD measures the distance of two probability distributions and is
258 defined as in Equations 7-8.

$$M = \frac{(P + Q)}{2} \qquad (7)$$

$$JSD(P, Q) = \frac{1}{2}(D_{kl}(P||M) + D_{kl}(Q||M)) \qquad (8)$$

259 where **P** denotes the probability distribution of a 3D property of the conformations generated by
260 the model, whereas **Q** corresponds to the conformations after energy optimization. **M** represents
261 the average distribution of P and Q. The Kullback-Leibler (KL) divergence[42], denoted as D_{kl}, is
262 calculated separately to quantify the difference of either **P** or **Q** from M. The JSD value was then
263 obtained by averaging the KL divergences. A JSD value of 0 indicates that the distributions P and
264 Q are identical and a value of 1 represents completely dissimilar distributions.

265

266 ***Recovery of active molecules***

267 Recovery of actives refers the ratio of generated compounds which are similar to the actives
268 in the reference set for a specific target. Tanimoto similarity between generation set and actives in
269 the reference set of the target protein is calculated. Here, Morgan fingerprint based on two bond
270 distance was used to calculate Tanimoto similarity. For an active molecule of target t (A_t), it is
271 recovered if its similarity of any compound in generation set is larger than 0.6. Then, the recovery

rate of active molecules for target t (R_t) is calculated as Formula 9:

$$R_t = \frac{\text{number of recovered active molecules}}{total\ number\ of\ active\ molecules} \qquad (9)$$

This metric, at certain extent, can be regarded as the probability of reproducing active compound by the generative model. The ratios at molecular structure and molecular Murcko scaffold[43] level were examined respectively.

Target binding assessment

To evaluate how good the generated set can bind in its target pocket, two scoring strategies were used here. One is the *in situ* scoring strategy, which scores the generated conformations at the binding site without going through further pose optimization. The other one is the redocking strategy, which scores after redocking of generated molecules into the binding set via external docking software. The Glide docking module of Schrodinger software was used for *in situ* scoring and redocking, and the GlideScore was used as the score value[44]. In redocking, the generated conformations were first gone through the LigPrep protocol of Schrodinger software for preparation and then docked, and only one docking pose was saved for each molecule. Besides, the Root-Mean-Square-Deviation (RMSD) value between the generated conformations and the redocked conformations were also calculated without performing any conformation superposition. The RMSD value quantifies the fitness between the protein pocket and the generated ligand, as any Van der Wars clash or unmatched electrostatic interaction between ligand and protein would penalize the ligand conformation, and in this case the docking pose was used as surrogate of ground truth. In summary, the *in situ* score, redocking score and RMSD value between the generated and redocked conformations were included in the target binding assessment set of metrics.

Results and Discussion

Benchmark dataset composition

Protein pocket based 3D molecular generative model involves the encoding of 3D geometric information of protein pocket, which are extracted from experimentally determined crystal structures of various proteins. To investigate the generalizability of the investigated models to unknown targets, our benchmark dataset encompasses 32 protein pockets, among which 17 targets and 25 PDB IDs are not included in the CrossDock2020 dataset that usually used for training the pocket based 3D generative models (Figure 2, detailed structure IDs can be seen in Table S1). In addition, these targets belong to various druggable protein families, such as kinase, G protein-coupled receptor (GPCR), nuclear receptor etc., and have reported ligands as marketed or clinical drug.

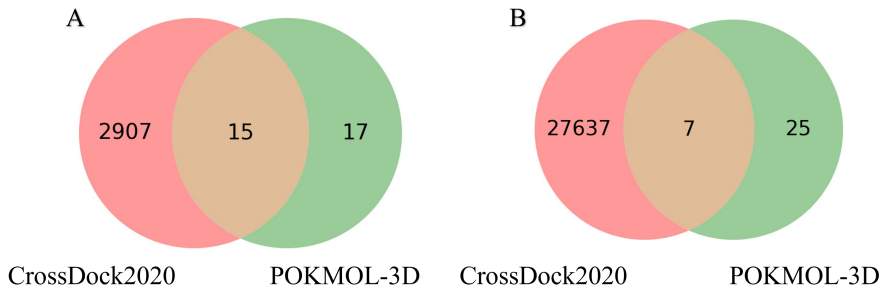


Figure 2. Target overlap between the POKMOL-3D and CrossDock2020 datasets in terms of (A) protein target name and (B) crystal structure.

Model quality

Table 2. Metrics for model quality evaluation

Model	Target failure rate	Sampling success rate	Sampling speed[a]
REINVENT	0	1	2.002

GraphBP	0	1	0.428
Pocket2Mol	0.094	0.656	7.131
PocketFlow	0	1	1.462
Lingo3DMol	0.063	0.313	2.381
DiffBP	0	1	16.338
TargetDiff	0	0.969	78.167
ResGen	0.063	0.125	29.613
SurfGen	0	0.094	28.150
SBDD	0	0.406	12.348

Note: a) estimated in second per compound

In this study, three structural unrelated metrics, i.e. target failure rate, sampling success rate and sampling speed, were proposed to evaluate model quality. As shown in Table 2, most models are able to generate compounds for all pockets so that their target failure rate is 0, while Pocket2Mol, Lingo3DMol and ResGen fail to generate molecules for a few target proteins. In detail, Pocket2Mol failed on Beta2AR, FXR and LXRB, ResGen failed on ERK2, NAMPT proteins, and Lingo3DMol failed on CDK9 and DPP4 proteins. These results suggested that these three models are not generalized good enough to deal with all targets. Additionally, the sampling success rate, defined as the fraction of targets that a model can generate over 2,000 molecules at most three runs, was employed as an additional indicator of generalizability to assess the models' capacity to generate sufficient molecules given a specified sampling size. The results indicated that models GraphBP, PocketFlow, DiffBP and the SMILES based baseline model REINVENT are able to sample over 2,000 molecules for all targets within three sampling runs, and TargetDiff also exhibits a high sampling success rate. Thus, the diffusion based and flow based models are able to generate sufficient molecules from a model-type perspective. However, the remaining models, i.e. Pocket2Mol, Lingo3DMol, ResGen, SurfGen, and SBDD, showed much lower success rate. Notably, SBDD exhibited a significantly lower sampling success rate than GraphBP that share the similar GNN architecture. This discrepancy might be attributed to the distinct approaches used by

these models to predict new atoms in the autoregressive generation process. A similar phenomenon was observed when comparing ResGen with Pocket2Mol. SurfGen, which represents protein pocket as protein surface, exhibited the lowest rate in sampling enough compounds in the pockets. Although detailed reason is not unclear, one probable reason may be that the flaws existed in generated 3D conformations make them failed in passing the internal structural validity check.

Furthermore, comparison of the sampling speed was conducted. The results showed that GraphBP exhibits fastest sampling speed, in which a molecule can be generated within one second. In contrast, SBDD was much slower than GraphBP although they share similar generative methodology. Pocket2Mol, PocketFlow, and Lingo3DMol exhibited relatively rapid sampling rate, in which a compound can be sampled in less than 10 seconds. TargetDiff showed the slowest speed, in which a compound is generated in more than one minute. Interestingly, DiffBP exhibited much faster sampling speed than diffusion based TargetDiff and graph based models ResGen and SurfGen. Furthermore, all protein pocket-based 3D molecular generative models, except GraphBP and PocketFlow, exhibited slower sampling speed than REINVENT.

General molecular quality

Table 3. Comparison of general molecular quality

Model	Validity ↑	Uniqueness ↑	Usability ↑	QED ↑	SAscore ↓	Molecule Tdiv ↑	Scaffold Tdiv ↑
Actives	1	1	0.996	0.553	3.059	0.882	0.864
REINVENT	1	0.999	1	0.603	2.763	0.944	0.933
GraphBP	0.997	0.998	0.893	0.498	5.241	0.955	0.954

Pocket2Mol	0.906	0.903	1	0.420	4.074	0.926	0.925
PocketFlow	1	0.907	1	0.471	3.084	0.938	0.917
Lingo3DMol	1	0.742	1	0.484	3.17	0.913	0.871
DiffBP	0.992	0.998	1	0.508	3.553	0.942	0.935
TargetDiff	0.998	0.58	1	0.356	5.303	0.944	0.95
ResGen	1	0.992	1	0.345	4.169	0.927	0.922
SurfGen	1	0.987	1	0.369	4.408	0.927	0.923
SBDD	0.641	0.996	1	0.357	5.937	0.915	0.914

To assess the quality of 2D structures of generated molecules, six metrics were utilized: molecular validity, uniqueness, usability, drug-likeness, synthesis accessibility, and diversity (as shown in Table 3, the distribution plots can be seen in supporting material). Molecular validity and uniqueness are fundamental metrics for evaluating generative models. The results indicated that the molecular validity of all protein pocket-based 3D models except SBDD were either equal or close to 1.0. In terms of molecular uniqueness, all models exhibited much higher performance than Lingo3DMol and TargetDiff. Especially, a new metric named molecular usability was introduced to quantify the likelihood of these models generating uncommon elements in structure. The results indicated that all models exhibit good performance on this metric, and only GraphBP generates about 10% molecules with uncommon atoms such as silicon atom.

In the assessment of drug-likeness, the QED score was averaged among the molecules generated by each model (Figure S2A). Notably, the QED scores for most 3D models were below 0.5, lower than the average value observed for known active molecules (QED score = 0.553). In

contrast, the QED score of REINVENT is superior to the average value of active molecules and higher than all 3D models. Additionally, the SAscore metric was utilized to evaluate the synthetic accessbility of generated molecules (Figure S2B). The SAscore of SBDD was significantly higher than other models, suggesting that SBDD tended to generate molecules with poorer synthesis accessbility. In contrast, PocketFlow and Lingo3DMol exhibited higher SAscore than other 3D models but still worse than the baseline REINVENT.

In evaluating molecular diversity of generation set, a pairwise similarity calculation was performed between molecules belonging to a specific target and the average molecular diversity across all targets. The results listed in Table 3 revealed that both REINVENT and 3D generative models exhibit high diversities. In summary, although 3D models exhibited similar performance on validity, uniqueness, usability and diversity with baseline model REINVENT, REINVENT model significantly showed better performance on QED and SAscore.

Structural properties

In addition to general molecular quality, it is imperative to consider fine-grained topology related structural properties for evaluation. We analyzed the distribution of several crucial 2D topological features, including the number of heavy atom, chiral centre, ring, aromatic ring, and rotatable bond, and the Fsp3 (Figure 3A-F).

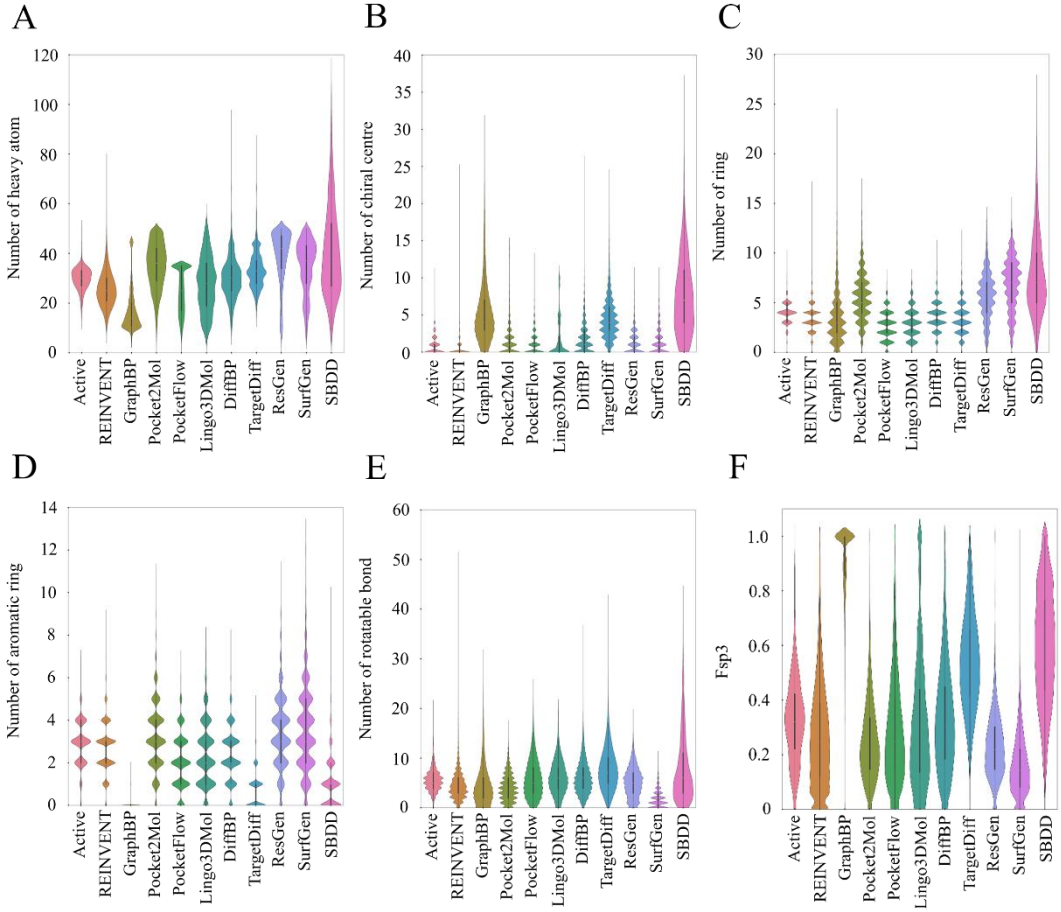


Figure 3. The violin plots for structural properties of 3D generative models including number of (A) heavy atom, (B) chiral centre, (C) ring, (D) aromatic ring, (E) rotatable bond, and (F) Fsp3.

The distribution of heavy atom number was shown in Figure 3A, REINVENT clearly exhibited most similar distribution as the active set, while most of 3D models tended to generate larger molecules than the active set, except GraphBP which generated significant portion of molecules with fewer than 20 heavy atoms. Particularly, SBDD generated molecules with a wide range from 20 to 60 heavy atoms. For Pocket2Mol, ResGen and SurfGen, the generated molecules had heavy atom between 40 and 60.

In terms of number of chiral centre, most of the 3D generative models tended to generate more chiral centre than the active set. Especially for GraphBP, SBDD and TargetDiff, the number of chiral centre in the generated compounds was obviously much larger than other models,

389 resulting in decrease of the synthetic accessbility. Whereas, REINVENT exhibited most similar
390 distribution to the active set and the generated molecules clearly had less chiral centre than the 3D
391 generative models. The ring count in a molecule serves as an indicator of its structural complexity,
392 given that most drug molecules possess at least one ring structure[45]. The ring count distribution of
393 active set fell in the range of 3-5 rings. The REINVENT model showed most similar distribution
394 to the active set, whereas the 3D models mostly exhibited much broader distribution. Especially
395 for Pocket2Mol, ResGen, SurfGen and SBDD, a substantial proportion of molecules had more
396 than five rings, resulting the increases of structural complexity and decreases of drug-likeness. In
397 contrast, PocketFlow, Lingo3DMol, DiffBP and TargetDiff exhibited distributions more close to
398 the active set than other 3D models. Interestingly, GraphBP tended to generate compounds with
399 less rings than the active set, consistent with the observation on distribution of heavy atom count.

400 Given the prevalence of aromatic rings in drug molecules[46], the number of aromatic ring is
401 also an important metric. Figure 3D showed that the distribution of aromatic ring in REINVENT
402 was quite similar to that of the active set, in which most compounds have 2-4 aromatic rings,
403 while the distribution of 3D models was deviate from the active set. Interestingly, GraphBP,
404 TargetDiff and SBDD tended to generate compounds with less number of aromatic rings
405 comparing to the active set. Pocket2Mol, ResGen and SurfGen generated a substantial fraction of
406 compounds with more than four aromatic rings, while PocketFlow, Lingo3DMol, and DiffBP
407 generated compounds with primarily one to three aromatic rings.

408 The analysis on the number of rotatable bond (Figure 3E) revealed that REINVENT exhibits
409 most similar distribution to the active set, although it still had a minor fraction of molecules
410 exceeding ten rotatable bonds. Among the 3D generative models, Pocket2Mol and GraphBP were

411 similar to the active set, while others models generated larger fraction of compounds with more
412 than ten rotatable bonds, and SurfGen generated molecules with less than two rotatable bonds.

413 The fraction of sp^3 hybridized carbon atoms is a metric that partially reflects a molecule's
414 flatness, *i.e.* the larger fraction of aromatic ring in a molecule the smaller Fsp3 value is, and it is
415 related to the success of drug in clinical trials[47]. Typically, small-molecule oral drugs harbor
416 approximately 40% of their carbon atoms in the sp^3 hybridization state[48]. Our findings (Figure 3F)
417 indicated that the proportion of sp^3-hybridized carbon atoms in active molecules primarily fell
418 within the range of 20% to 40%. The molecules generated by REINVENT model exhibited a
419 distribution closely resembling that of the active set. The 3D models Pocket2Mol, PocketFlow,
420 Lingo3DMol, DiffBP, and ResGen exhibited comparable distributions to the active set (as shown
421 in Figure 3F). SurGen had a tendency of favoring molecules containing less than 20% sp^3
422 hybridized carbon atoms, indicating larger number of aromatic ring. Whereas, GraphBP,
423 TargetDiff and SBDD showed a large fraction of compounds with high Fsp3 value, indicating
424 most of the carbon atoms in the structure are saturated carbons with few aromatic rings. The
425 analysis on the structural properties revealed that REINVENT has most close distribution to that
426 of the active set, while all the 3D generative models showed larger deviation to the active set,
427 highlighting the necessity of further improvement for current 3D generative model algorithms to
428 increase the compound quality.

429 Besides the evaluation of 2D related metrics, the quality assessing of 3D conformation is also
430 important as these 3D based models generate 3D conformation and 2D graph simultaneously. We
431 investigated the differences of distributions of bond length, bond angle, and dihedral angle
432 between the generated 3D conformations and the OPLS3 force field [49]minimized conformations,

and the JSD index was utilized to quantify the deviation of these two distributions. For most 3D models except GraphBP, the average JSD on bond length over all types of bonds is larger than 0.5 (Table 4), indicating significant difference in bond length distribution between the generated and minimized conformations. Moreover, the JSD values for bond angles and dihedral angles surpass 0.1 for all models, indicating a great deviation to the force field minimized distributions, although some 3D models are based on atomic point clouds leveraging third-party software such as OpenBabel to construct the final structures. The distributions of C≡C bond length, OC=O bond angle and cccc dihedral angle were shown in Figure 4 (more detailed analysis in Figure S3-5), the divergence could be attributed to the broader distribution on parameters in generated conformations. Only a few bond angles and dihedral angles exhibited similar distribution indicated by their small JSD metrics (Table S2-4). These results suggested that learning of bond length, bond angle and dihedral angle in the generative models may need further improvement.

Table 4. JSD divergence of bond length, bond angle and dihedral angle

Model	Bond-Length↓	Bond Angle↓	Dihedral Angle↓
GraphBP	0.431	0.348	0.133
Pocket2Mol	0.586	0.404	0.145
PocketFlow	0.672	0.303	0.266
Lingo3Dmol	0.588	0.311	0.193
DiffBP	0.586	0.378	0.159
TargetDiff	0.610	0.367	0.180
ResGen	0.647	0.341	0.143

SurfGen	0.605	0.350	0.146
SBDD	0.653	0.348	0.238

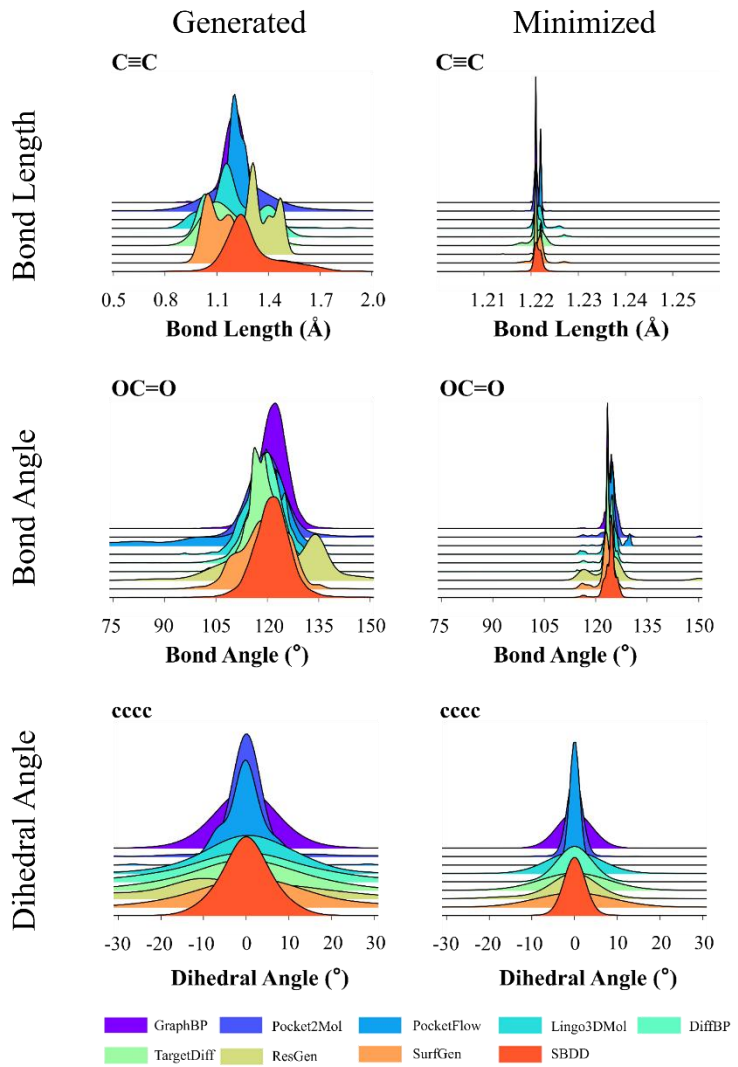


Figure 4. Distribution of the bond length of C≡C, the bond angle of OC=O and the dihedral angle of cccc. The left panel presents the distribution in generated conformations, while the right panel presents the minimized distribution.

Recovery of active molecules

One direct way of evaluating 3D generative models conditioned on protein pockets is to examine the model's capability of generating potentially bioactive molecules. For this purpose, we employed the recovery rate of active molecules, *i.e.* the percentage of successfully recovered

active molecules for a given target. One active molecule was regarded as being successfully recovered if a similar compound in the generation set could be identified given the pair similarity was larger than the user defined cut-off. This is a stricter criterion compared to the similar metric utilized in Lingo3DMol[26], which is the percentage of targets with at least one generated molecule exhibiting similarity to the actives.

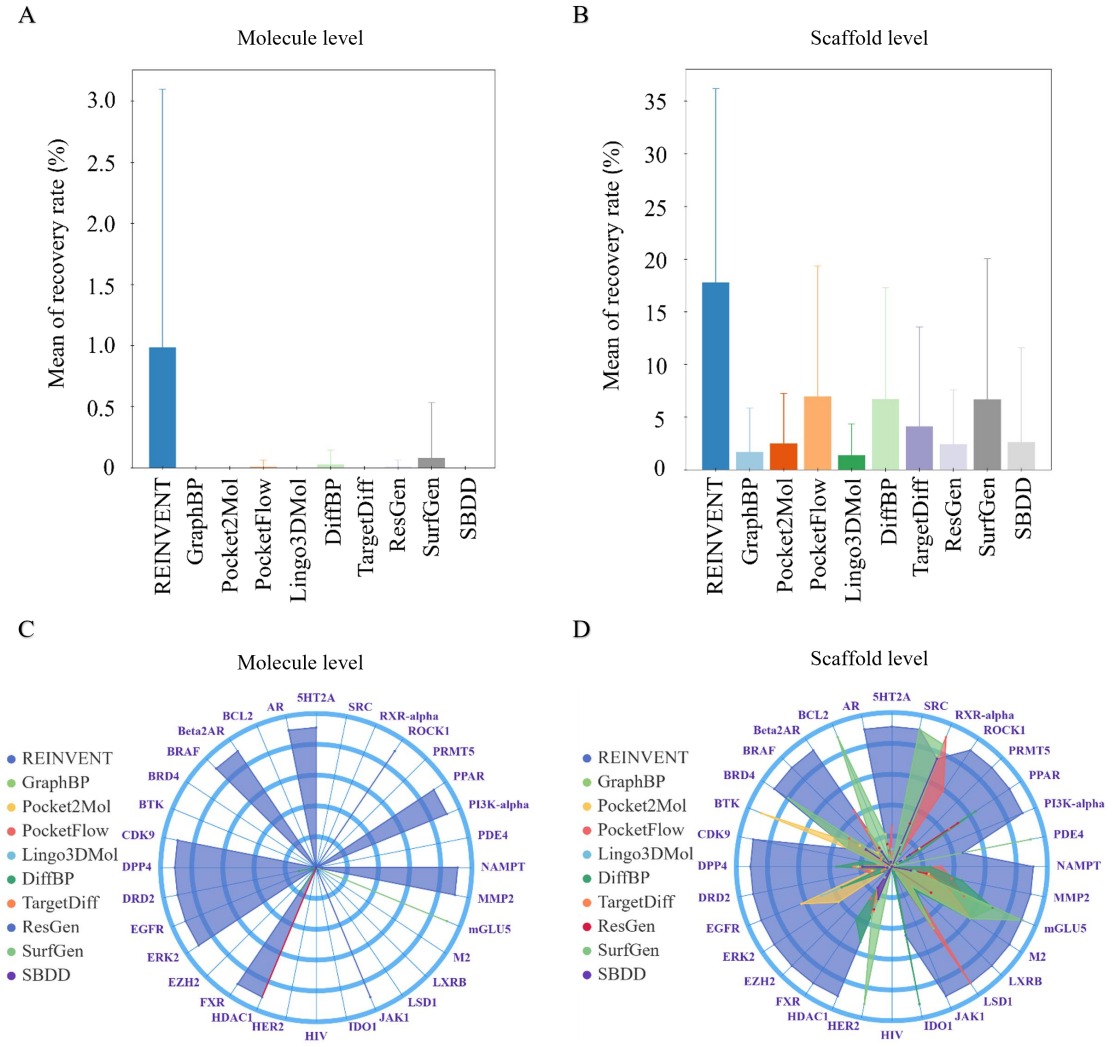


Figure 5. Statistics for recovery of active molecule. A-B) Histograms to display the average recovery rate over all targets. C-D) Radar charts to display recovery rate among the target set for all models.

In current study, the Tanimoto similarity threshold was set to 0.6. The average recovery rates at both molecule and scaffold level for these targets were shown in Figure 5A. It was obvious that

REINVENT showed the best recovery rate compared to all 3D molecular generative models at both levels, but there were still 15 targets in which REINVENT failed to recover any active molecule (Figure 5C). Specifically, GraphBP, Pocket2Mol, Lingo3DMol, TargetDiff and SBDD were unable to recover any active molecule for these 32 proteins (Table S5), while other 3D models can recovery a few actives on a few targets. The successfully recovered examples were shown in Figure 6. Compared to REINVENT, the 3D generative models tended to recover the active molecules with simple structure and relatively low similarity.

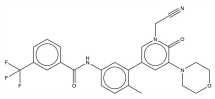
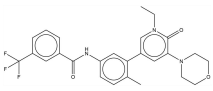
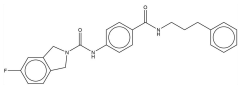
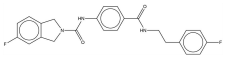
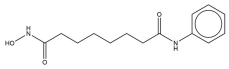
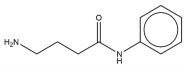
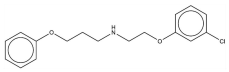
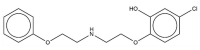
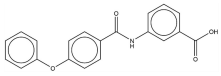
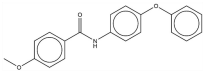
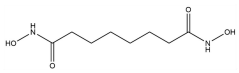
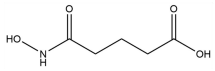
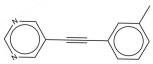
Model	Target	Active compound	Generate compound	Tanimoto similarity
REINVENT	BRAF			0.845
	NAMPT			0.782
PocketFlow	HDAC1			0.656
DiffBP	DRD2			0.610
	FXR			0.610
ResGen	HDAC1			0.650
SurfGen	mGLU5			0.621

Figure 6. Examples of successfully recovered active molecules for some models.

At Murcko scaffold level, a similar trend was observed among compared models (Figure 5B). As anticipated, REINVENT exhibited superior performance (average recovery rate is around 17 %)

than the 3D models. PocketFlow, DiffBP, and SurfGen achieved relatively better performance than other 3D models, but all 3D models exhibited avarage recovery rate lower than 10%. Furthermore, significant variability was noted among the targets for each model. In terms of target coverage, REINVENT clearly performed better on most targets (Figure 5D), while 3D models showed better performance on a limited number of targets. For example, SurfGen performed better than REINVENT on PDE4, HER2 and BCL2 proteins, Pocket2Mol on BTK and DiffBP on IDO1 etc.

Our results on active recovery rate demonstrated the limitation of current pocket based 3D models as their performance was in general inferior than baseline REINVENT. Notably, even a less stringent similarity threshold of 0.4 was applied, a similar trend was observed (Table S6), indicating the weakness of these models in learning chemical structures conditioned on the protein pocket.

Target binding assessment

The ultimate goal of generative model is to generate potential active compound to the target, therefore we introduced several protein-ligand interaction based metrics to assess the target binding capability of generated 3D conformations. Firstly, *in situ* docking score was chosen as a surrogate to quantify the binding affinity of the 3D conformations generated from those 3D models. Here, Glidescore was calculated for comparing binding affinity of conformations, and the value of 0.0 kcal/mol was set as a criterion to judge if it is favorable for the ligand to bind in a protein. A conformation is deemed to be a positive conformation (PC) if its Glidescore is less than 0.0 kcal/mol and the fraction of PC was calculated.

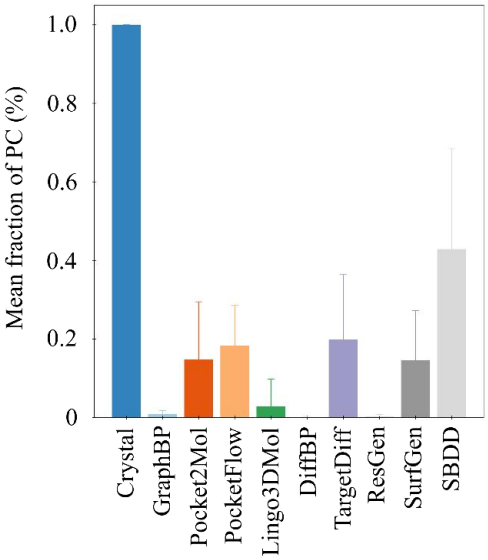


Figure 7. Histogram to display the mean fraction of PC over all targets.

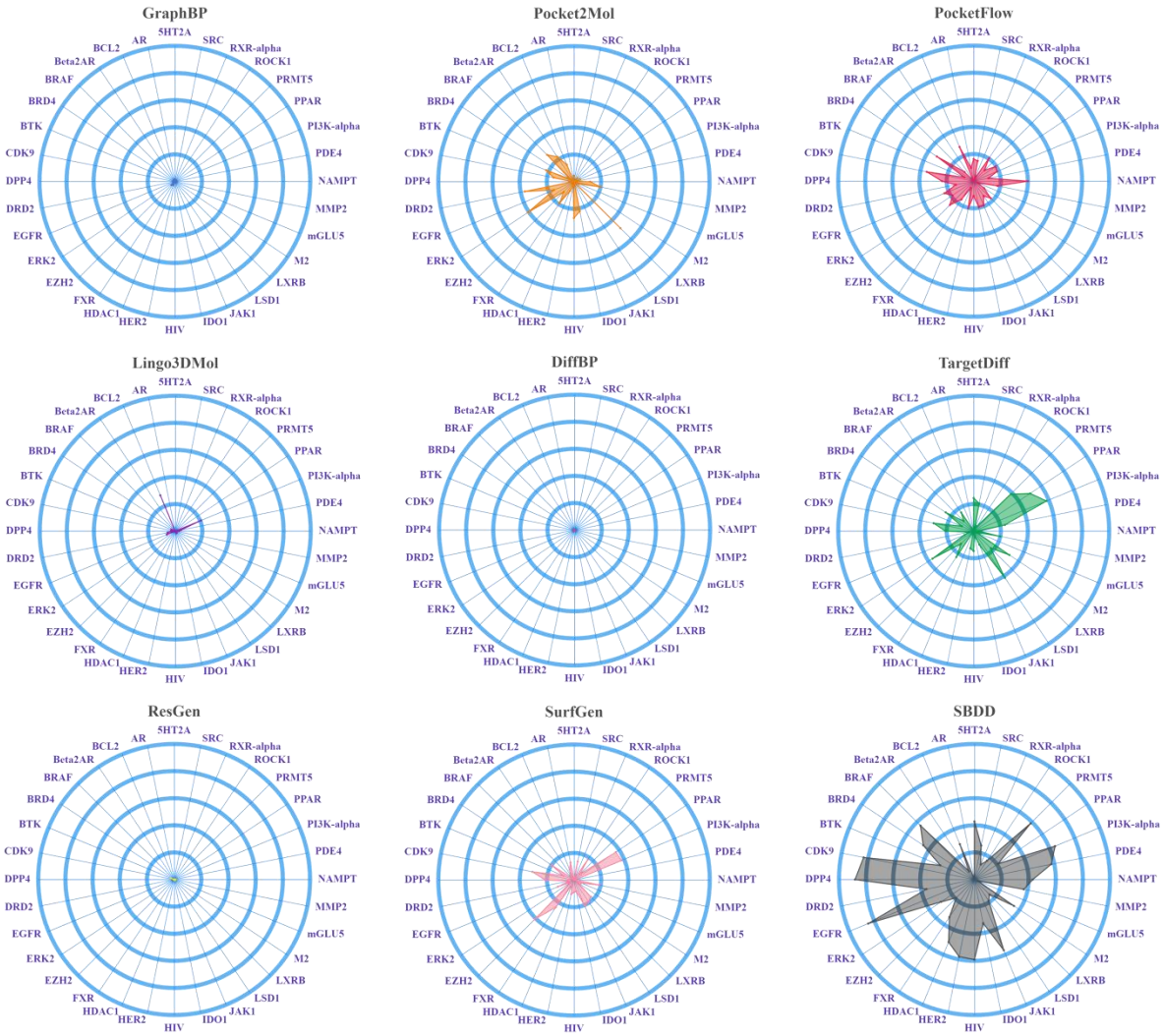


Figure8. Radar charts to display the proportion of PC of each target across the models. The circles represent five proportional levels, which are 100%, 80%, 60%, 40% and 20% from the outer layer to the centre of the chart.

501 The mean fraction of PC for each generation set could be seen in Figure 7 and the breakdown
502 of the fraction values across all 32 targets was shown in Figure 8. Among the 32 targets, GraphBP,
503 Lingo3DMol, DiffBP, and ResGen generated worst conformations and exhibited quite low
504 fraction values, while SBDD generated highest fraction of PC. A physically advantageous position
505 within the pocket was decided in SBDD model through a sampling process to estimate the
506 likelihood of atom occurrence and reduce the misplacement of ligand atoms in the pocket. SBDD
507 has on average only 40% conformations regarded as PC and only in very few targets the
508 proportion could surpass 80%. These results suggested that a lot of conformations have substantial
509 clash with protein atoms and current 3D models should strive to improve the learning of
510 protein-ligand interaction.

511 The distribution of *in situ* Glidescores for the top 10 conformations of 3D models, along with
512 that of crystal ligand conformations, was shown in Figure 9A. It can be seen from Figure 9A that
513 crystal ligands obviously exhibit best in situ docking scores, models SBDD, TargetDiff and
514 Pocket2Mol showed top three performances on *in situ* scores. Whereas, for Lingo3DMol, DiffBP
515 and ResGen with worst performances, the average in situ scores for some targets are even higher
516 than 0.0 kcal/mol. These results indicated that the learning of protein/ligand interaction is still far
517 from optimal for 3D generative models.

518 On the other hand, redocking analysis was also conducted to reproduce binding
519 conformations for the generated compounds via Glide docking in SP mode. The redocking
520 approach reflects the fitness between ligand and protein in 2D perspective as the binding
521 conformation and docking score is derived by external docking software. The distribution of
522 average redocking score for top 10 conformations was shown in Figure 9B, and scores of active

set and REINVENT were also included. The average redocking scores of all sets were between -8 and -12 kcal/mol, and differences of redocking scores among all models were small. In addition, the average score of a few 3D models such as SBDD and Pocket2Mol were lower than the active set, raising a question on whether the redocking score is really a relevant metric to evaluate generative model.

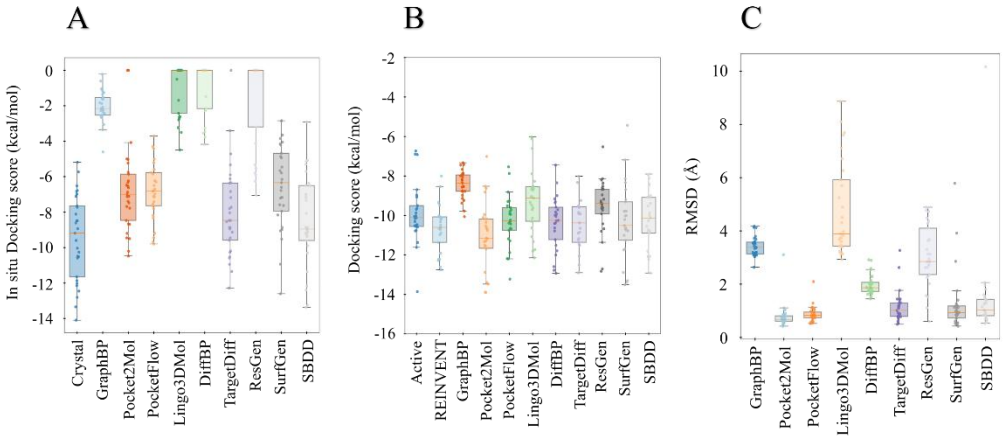


Figure 9. Distribution of protein-ligand interaction based metrics across the targets calculated based on the top 10 molecules: A: in-situ score; B: redocking score; C: the RMSD between generated and redocked conformations.

It is obvious from Figure S6A-B that the redocking score is generally lower than the corresponding in situ score, indicating that the generated conformation may be different from active conformation. A comparative analysis was then conducted to measure the root mean square deviation (RMSD) between the generated and redocked conformations (Figure S6C). As redocked conformations were generated by force field based docking software, the RMSD value can provide insights about geometrical difference between the generated 3D binding poses and poses simulated by physics principles. The average RMSD of top 10 conformations was illustrated in Figure 9C, which indicated that the RMSD values for the 32 targets fell in the range of [0, 10] Å. ResGen, Lingo3DMol and GraphBP exhibited quite large RMSDs to the docked conformations,

while other models such as Pocket2Mol and PocketFlow showed average RMSD less than 2Å, suggesting the generated conformations of these models are close to their docked conformations. In Figure 10, several examples were presented to display overlapping between generated and docked conformations for the molecule with smallest RMSD for a specific target. As shown in Figure 10, Pocket2Mol and PocketFlow exhibited much better performance on all selected targets than Lingo3DMol and ResGen. The RMSD values of Pocket2Mol and PocketFlow were smaller than 1.0 Å, suggesting high similarity between the generated and docked poses. In contrast, the poses generated by Lingo3DMol and ResGen were quite dissimilar to the redocked poses in most cases.

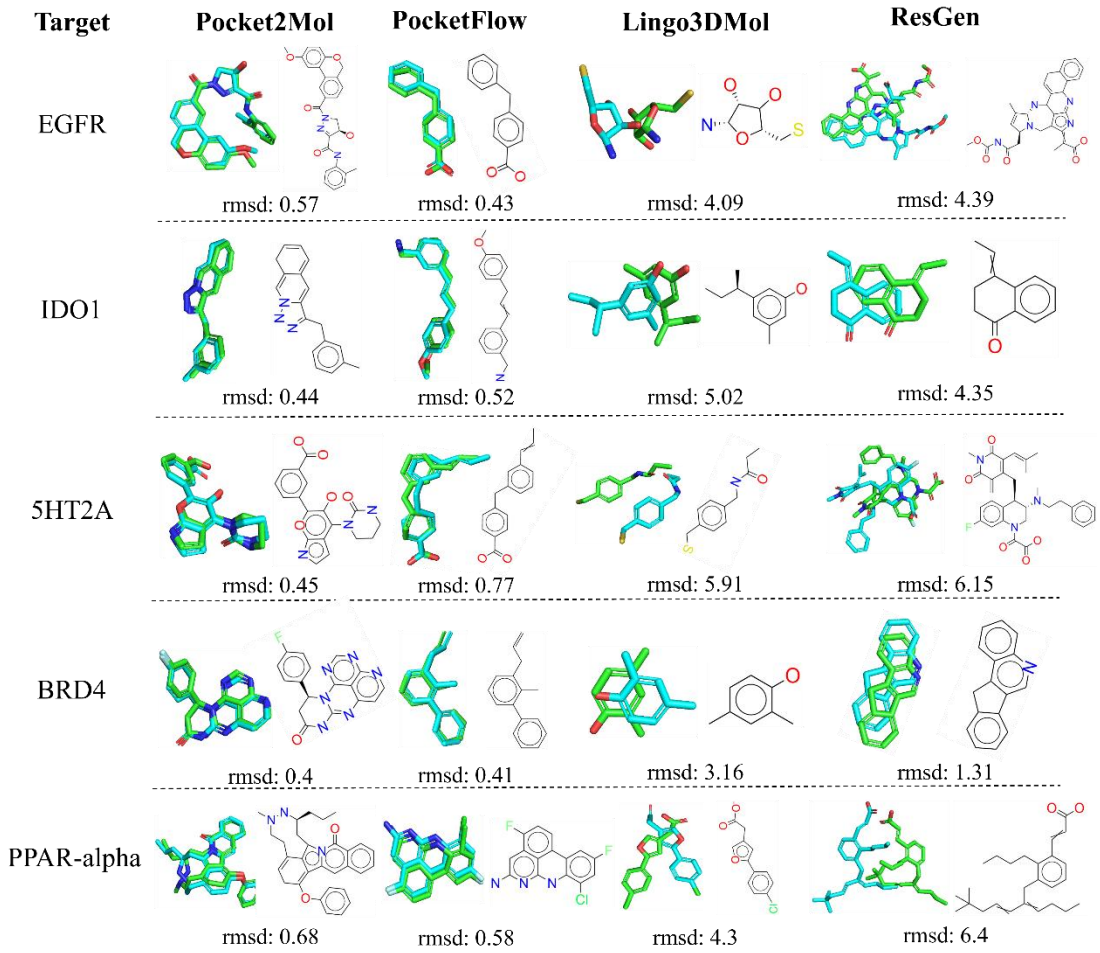


Figure 10. Examples of overlapping between generated and redocked poses for Pocket2Mol, PocketFlow,

Lingo3Dmol, and ResGen. Five targets belonging to different families were selected for the comparison.

Conclusion

In current study, a novel benchmark dataset, named POKMOL-3D, was compiled specifically for evaluating pocket based 3D molecular generative models, including a set of comprehensive metrics for measuring model quality from various perspectives. Nine recently published 3D generative models were selected for carrying out this benchmark study on 32 protein pockets, along with the SMILES based REINVENT model as the baseline. Although some promising results of pocket based 3D generative models have been reported, our benchmark study revealed some weak points existed among the selected 3D models, such as slow sampling speed, poor druggability and synthesizability of generated molecules, and failure to generate rational 3D conformations for target binding. Overall, the performance of 3D generative models on large scale of pockets is still far from satisfactory, and polishing of network architecture is needed to improve the learning of ligand-protein interaction and generalizability of current 3D generative models to enable their application on wide range of protein pockets. Through this study, we hope that the proposed evaluation framework can be useful in facilitating the future advancement of pocket based 3D molecular generative model.

Conflicts of interest

There are no conflicts to declare.

Author contributions

L. H. conceived the study, prepared the dataset and source code, and conducted the data analysis. Q. Y. and Z. N. evaluated part of models and performed data validation. M. X., X. X and J. W. also

participate in discussion. C. H., R T., and L. J. are responsible for article revision and work supervision.

Acknowledgement

We are grateful for support from the Big Data Center for Biomedical Research of Wuyi University.

Data availability

The related data and source codes of POKMOL-3D are accessible at https://github.com/haoyang9688/POKMOL-3D.

Funding

This study was supported by the National Key Research and Development Program of China (2023YFF1204902), startup and R&D Program of Guangzhou National Laboratory (YW-YWYM0205, GZNL2023A01008, GZNL2023A01005) and Overseas Experts Supporting Programs under National Research Platform (WGZJ22-001).

References

1. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
2. A. Lo, R. Pollice, A. Nigam, A. D. White, M. Krenn and A. Aspuru-Guzik, *Digital Discovery*, 2023, **2**, 897-908.
3. M. Skalic, D. Sabbadin, B. Sattarov, S. Sciabola and G. De Fabritiis, *Molecular pharmaceutics*, 2019, **16**, 4282-4291.
4. M. Xu, T. Ran and H. Chen, *Journal of Chemical Information and Modeling*, 2021, **61**, 3240-3254.
5. E. Mansimov, O. Mahmood, S. Kang and K. Cho, *Scientific reports*, 2019, **9**, 20381.

6. G. N. Simm and J. M. Hernández-Lobato, *arXiv preprint arXiv:1909.11459*, 2019.
7. T. Gogineni, Z. Xu, E. Punzalan, R. Jiang, J. Kammeraad, A. Tewari and P. Zimmerman, *Advances in Neural Information Processing Systems*, 2020, **33**, 20142-20153.
8. O. Ganea, L. Pattanaik, C. Coley, R. Barzilay, K. Jensen, W. Green and T. Jaakkola, *Advances in Neural Information Processing Systems*, 2021, **34**, 13757-13769.
9. S. Luo, C. Shi, M. Xu and J. Tang, *Advances in Neural Information Processing Systems*, 2021, **34**, 19784-19795.
10. C. Shi, S. Luo, M. Xu and J. Tang, International conference on machine learning, 2021.
11. L. Huang, H. Zhang, T. Xu and K.-C. Wong, Proceedings of the AAAI Conference on Artificial Intelligence, 2023.
12. E. Hoogeboom, V. G. Satorras, C. Vignac and M. Welling, International conference on machine learning, 2022.
13. A. Morehead and J. Cheng, *arXiv preprint arXiv:2302.04313*, 2023.
14. M. Xu, A. S. Powers, R. O. Dror, S. Ermon and J. Leskovec, International Conference on Machine Learning, 2023.
15. C. Vignac, N. Osman, L. Toni and P. Frossard, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2023.
16. M. Ragoza, T. Masuda and D. R. Koes, *Chemical science*, 2022, **13**, 2701-2713.
17. S. Luo, J. Guan, J. Ma and J. Peng, *Advances in Neural Information Processing Systems*, 2021, **34**, 6229-6239.
18. M. Liu, Y. Luo, K. Uchino, K. Maruhashi and S. Ji, *arXiv preprint arXiv:2204.09410*, 2022.
19. X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng and J. Ma, International Conference on Machine Learning, 2022.
20. B. Jing, S. Eismann, P. N. Soni and R. O. Dror, *arXiv preprint arXiv:2106.03843*, 2021.
21. C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi and L. J. Guibas, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
22. Y. Jiang, G. Zhang, J. You, H. Zhang, R. Yao, H. Xie, L. Zhang, Z. Xia, M. Dai, Y. Wu, L. Li and S. Yang, *Nature Machine Intelligence*, 2024, **6**, 326-337.
23. O. Zhang, T. Wang, G. Weng, D. Jiang, N. Wang, X. Wang, H. Zhao, J. Wu, E. Wang and G. Chen, *Nature Computational Science*, 2023, **3**, 849-859.
24. A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes and M. Welling, *arXiv preprint arXiv:2210.13695*, 2022.
25. H. Lin, Y. Huang, M. Liu, X. Li, S. Ji and S. Z. Li, *arXiv preprint arXiv:2211.11214*, 2022.
26. W. Feng, L. Wang, Z. Lin, Y. Zhu, H. Wang, J. Dong, R. Bai, H. Wang, J. Zhou and W. Peng, *Nature Machine Intelligence*, 2024, **6**, 62-73.
27. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nature chemistry*, 2012, **4**, 90-98.
28. P. Ertl and A. Schuffenhauer, *Journal of cheminformatics*, 2009, **1**, 1-11.
29. J. Lin, *IEEE Transactions on Information theory*, 1991, **37**, 145-151.
30. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley and J. K. Perry, *Journal of medicinal chemistry*, 2004, **47**, 1739-1749.
31. P. G. Francoeur, T. Masuda, J. Sunseri, A. Jia, R. B. Iovanisci, I. Snyder and D. R. Koes, *Journal of chemical information and modeling*, 2020, **60**, 4200-4215.
32. C. Harris, K. Didi, A. R. Jamasb, C. K. Joshi, S. V. Mathis, P. Lio and T. Blundell, *arXiv preprint*

arXiv:2308.07413, 2023.

33. Z. Jocys, J. Grundy and K. Farrahi, *Digital Discovery*, 2024.
34. M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, *Journal of cheminformatics*, 2017, **9**, 1-14.
35. H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *Journal of Cheminformatics*, 2024, **16**, 20.
36. R. Wang, X. Fang, Y. Lu and S. Wang, *Journal of medicinal chemistry*, 2004, **47**, 2977-2980.
37. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic acids research*, 2000, **28**, 235-242.
38. S. Maestro, *Schrödinger, LLC, New York, NY*, 2020, **2020**, 682.
39. B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez and J. F. Mosquera, *Nucleic acids research*, 2024, **52**, D1180-D1192.
40. G. Landrum, *Release*, 2013, **1**, 4.
41. D. Rogers and M. Hahn, *Journal of chemical information and modeling*, 2010, **50**, 742-754.
42. S. Kullback, 1951.
43. G. W. Bemis and M. A. Murcko, *Journal of medicinal chemistry*, 1996, **39**, 2887-2893.
44. R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, *Journal of medicinal chemistry*, 2006, **49**, 6177-6196.
45. R. D. Taylor, M. MacCoss and A. D. Lawson, *Journal of medicinal chemistry*, 2014, **57**, 5845-5859.
46. T. J. Ritchie and S. J. Macdonald, *Drug discovery today*, 2009, **14**, 1011-1020.
47. F. Lovering, J. Bikker and C. Humblet, *Journal of medicinal chemistry*, 2009, **52**, 6752-6756.
48. D. C. Kombo, K. Tallapragada, R. Jain, J. Chewning, A. A. Mazurov, J. D. Speake, T. A. Hauser and S. Toler, *Journal of chemical information and modeling*, 2013, **53**, 327-342.
49. E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren and J. L. Knight, *Journal of chemical theory and computation*, 2016, **12**, 281-296.