3D Molecular Pocket-based Generation with Token-only Large Language Model

Jike Wang[1,#], Hao Luo[1,#], Rui Qin[1,#], Mingyang Wang[1], Xiaozhe Wan[2], Meijing Fang[1], Odin Zhang[1], Qiaolin Gou[1], Qun Su[1], Chao Shen[1], Ziyi You[1], Liwei Liu[2,*], Chang-Yu Hsieh[1,*], Tingjun Hou[1,*], Yu Kang[1,*]

[1]*College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China*

[2]*Advanced Computing and Storage Laboratory, Central Research Institute, 2012 Laboratories. Huawei Technologies Co., Ltd., Nanjing 210000, Jiangsu, China*

[#]*Equivalent authors*

Corresponding authors

Yu Kang

E-mail: yukang@zju.edu.cn

Tingjun Hou

E-mail: tingjunhou@zju.edu.cn

Chang-Yu Hsieh

E-mail: kimhsieh@zju.edu.cn

Liwei Liu

E-mail: liuliwei5@huawei.com

Abstract

The generation of three-dimensional (3D) molecules based on target structures represents a cutting-edge challenge in drug discovery. Many existing approaches often produce molecules with invalid configurations, unphysical conformations, suboptimal drug-like qualities, limited synthesizability, and require extensive generation times. To address these challenges, we present 3DSMILES-GPT, a fully language-model-driven framework for 3D molecular generation that utilizes tokens exclusively. We treat both two-dimensional (2D) and 3D molecular representations as linguistic expressions, combining them through full-dimensional representations and pre-training the model on a vast dataset encompassing tens of millions of drug-like molecules. This token-only approach enables the model to comprehensively understand the 2D and 3D characteristics of large-scale molecules. Subsequently, we fine-tune the model using pair-wise structural data of protein pockets and molecules, followed by reinforcement learning to further optimize the biophysical and chemical properties of the generated molecules. Experimental results demonstrate that 3DSMILES-GPT generates molecules that comprehensively outperform existing methods in terms of Vina docking score, drug-likeness (QED), and synthetic accessibility score (SAS). Notably, it achieves a 33% enhancement in the quantitative estimation of QED, meanwhile the Vina score maintaining its state-of-the-art performance. The generation speed is remarkably fast, with the average time approximately 0.45 seconds per generation, representing a threefold increase over the fastest existing methods. This innovative approach highlights the potential of 3DSMILES-GPT to revolutionize the generation of drug-like molecules, pushing the boundaries of 3D molecular generation in the drug discovery process.

Introduction

In recent years, deep generative models have attracted extensive attention, demonstrating remarkable advancements across diverse domains, ranging from natural language processing to video synthesis. These models exhibit remarkable proficiency in encoding and synthesizing data within continuous domains. However, as the focus shifts towards more intricate and discrete data types, notably chemical molecules, there is a growing emphasis on developing generative models capable of generating authentic and efficacious data within these realms. The progression of deep generative models has spurred the development of various methodologies aimed at tackling the challenge of molecular generation, offering a promising avenue for innovative drug molecule design.

During earlier periods, ligand-based molecular generation (LBMG) gained significant popularity. These methodologies can be categorized into two primary types based on how generated molecules are represented: graph-based molecular generation and sequence-based molecular generation. The fundamental principle involves representing molecules as graphs or sequences, thus framing the generation task as either a graph structure generation or natural language generation problem. Techniques such as Bayesian optimization (BO) and reinforcement learning are employed to guide the model in generating the desired drug molecules.

Molecules inherently possess structures resembling graphs, rendering it intuitive to express their information graphically. Consequently, methods for molecular design grounded in graph representations and traditional heuristic algorithms have long been established. For example, Brown et al. devised a molecular optimization algorithm based on molecular graphs by employing genetic algorithms in 2004[1], while in 2013, Virshup et al. introduced the ACSESS algorithm[2]. With the progression of graph neural networks (GNN) in recent years, these networks have exhibited remarkable adaptability across diverse challenges rooted in graph-structured data. DeCao et al. pioneered the integration of GNN for drug design with their 2018 proposal of MolGAN[3], thereby forging novel pathways in molecular design. As graph-based methodologies continue to advance rapidly, an escalating number of researchers

are capitalizing on molecular graph representations for drug design[4-8].

When compared to GNN-driven strategies for molecular synthesis, sequence-based methodologies offer a more succinct avenue. This stems from the fact that chemical compounds can be effectively represented through chemical languages such as the Simplified Molecular Input Line Entry System (SMILES)[9] or SELFES[10], which mirror the structure of natural language. Consequently, a plethora of scholarly works on molecular design have proposed frameworks based on recurrent neural networks (RNNs) or transformer[11]. In 2016, ChemVAE amalgamated variational autoencoders (VAEs) with BO to explore the latent state space in search of molecules with desired attributes[12]. In 2017, Olivecrona et al. harnessed reinforcement learning to fine-tune the generation process of RNN-based molecules, yielding structures similar to specified ones or possessing predetermined activities[13]. In 2021, Wang et al. shifted towards a transformer decoder instead of RNNs for generation, combining knowledge distillation and reinforcement learning to develop MCMG[14]. Over time, a variety of sequence-based molecular generation methodologies have emerged.[15-26].

However, 2D molecular generation exhibits a significant limitation since these techniques neglect the crucial 3D structural complementarity between protein pockets and molecules. Given the pivotal role of ligand-protein conformational selection in drug design, evaluating such complementary features requires an understanding grounded in the intrinsic 3D structures of protein pockets and molecules. Consequently, there has been emerging interest in 3D structure-based molecular generation.

With the advent of deep geometric learning, numerous studies on autoregressive 3D molecular generation have surfaced. For instance, Gebauer introduced G-SchNet[27], an autoregressive deep neural network that generates diverse small organic molecules by sequentially positioning atoms in Euclidean space. Subsequently, models such as LiGAN[28], GraphBP[29], SBDD[30], and Pocket2Mol[31] have been developed to directly generate molecules within pockets[32-35]. However, autoregressive methodologies are susceptible to error accumulation, which has spurred the exploration of diffusion-based 3D molecular generation approaches[36-39]. These methods facilitate the

simultaneous generation of entire molecules, rather than sequentially producing atoms. So far, these strategies have yet to effectively capture the distribution of chemical bonds, leading to the creation of impractical molecular structures.

The 3D molecular generation methods discussed above primarily rely on GNN. While language models (LMs) effectively extract abundant 2D molecular insights from extensive drug-like datasets during 2D molecular design, their ability to represent continuous 3D molecular architectures remains limited. However, with the recent proliferation of large-scale language models (LLMs), numerous studies suggest that LLMs can adeptly acquire continuous numerical representations. Born et al. presented the Regression Transformer[40], which accomplishes unified regression and prediction tasks by encoding numerical values as tokens. Furthermore, Flam-Shepherd et al. utilized Cartesian coordinates xyz token to represent the 3D structures of molecules[41]. Both methodologies have shown promising effectiveness in their respective drug design endeavors. Recently, Feng et al. proposed Lingo3DMol[42], a fragment-based LM-centric 3D molecular generation model, which exhibits promising performance surpassing that of graph network-based models during their benchmark assessments. Notably, BindGPT[43], a decoder-only language model has demonstrated remarkable success in pocket-conditioned generation of 3D molecules. This approach aligns closely with the objectives of our research and serves as concurrent work that showcases similar methodologies.

The aforementioned endeavors highlight the adeptness of LMs in discerning intricate details pertaining to the inherent 3D structural characteristics of molecules. Compared to the complex diffusion and GNN-based methods for molecular generation, autoregressive approaches grounded in LMs offer simpler and more efficient training processes. Moreover, token-only paradigms seamlessly integrate with existing universal LLMs. Consequently, we explore the feasibility of employing a simpler and more explicit method to delineate the structural features of molecules and protein pockets. Based on this foundational understanding, we affirm the capacity of LLM to apprehend pertinent positional cues associated with molecules and protein pockets. Herein, we present 3DSMILES-GPT, an innovative token-only framework designed

for explicit 3D molecular generation, firmly rooted in LLM. As shown in **Fig. 1**, the architectural blueprint of 3DSMILES-GPT centers on a transformer decoder. By framing the task of generating 2D and 3D structures as a natural language generation endeavor, 3DSMILES-GPT encodes atomic 3D coordinates as tokens, facilitating the acquisition of molecular 2D and 3D information. To maximize the intrinsic capabilities of LMs, our methodology begins with the pretraining phase of 3DSMILES-GPT using an extensive dataset with drug-like molecules. After fine-tuning on a specified protein-ligand dataset, we integrate surface atomic coordinates from pockets along with ligand molecules. Furthermore, to enhance the model's ability to extract information from protein pockets, we introduce a protein encoder as a detachable modular component. Furthermore, the application of reinforcement learning methodologies enables the refinement of generated molecules across a diverse range of properties. The experimental results demonstrate that, compared to existing state-of-the-art (SOTA) methodologies, 3DSMILES-GPT achieves optimal performance across 8 out of 10 benchmark metrics including bioactivity, drug-likeness, and synthetic accessibility. Moreover, the targeted case studies on 5 distinct protein targets further elucidate its efficacy in generating drug-like molecules with robust binding strength in practical scenarios.

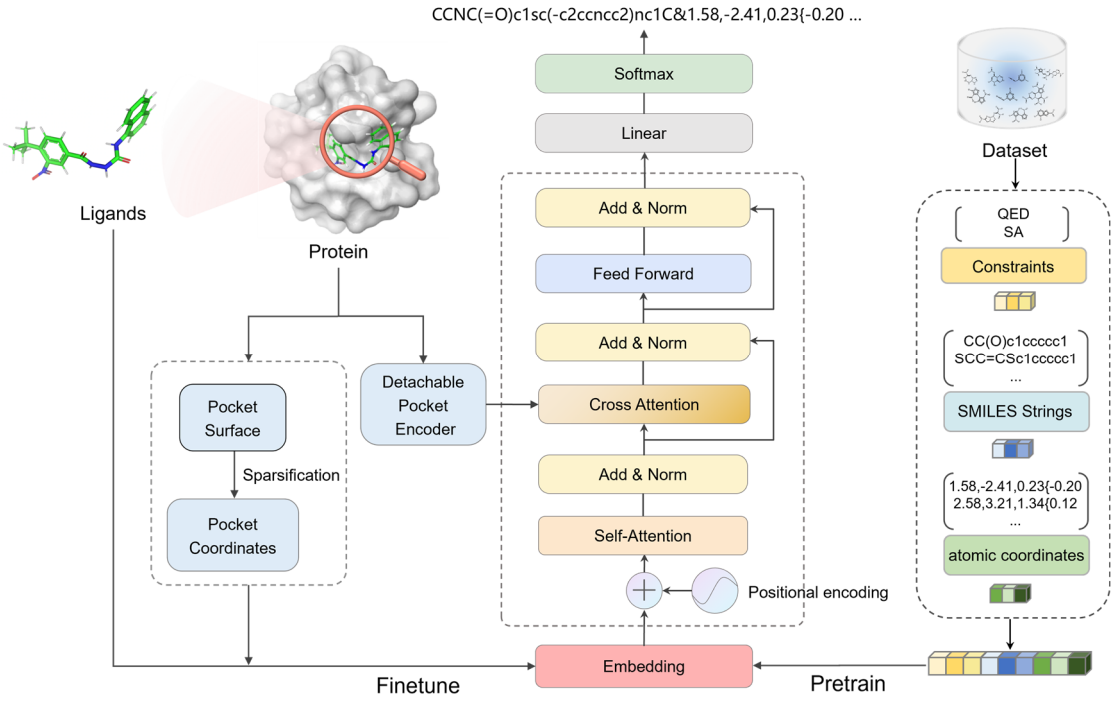


Fig. 1 | The overview of 3DSMILES-GPT.

Results

The competency of language models in generating 2D molecular configurations is undeniable, owing to their adeptness in processing discrete data. By utilizing chemical languages like SMILES as input, these models demonstrate proficiency in acquiring knowledge of the inherent 2D topological arrangements of molecules. However, the pivotal question remains whether language models can effectively capture the distribution of continuous data, including molecular conformations. Thus, in this segment, we begin by assessing the quality of conformations generated by 3DSMILES-GPT. Subsequently, we conduct an analysis of the properties and binding efficacy of the generated molecules. Finally, we evaluate the generalization capacity of 3DSMILES-GPT with respect to specific drug targets.

Quality of generated conformation

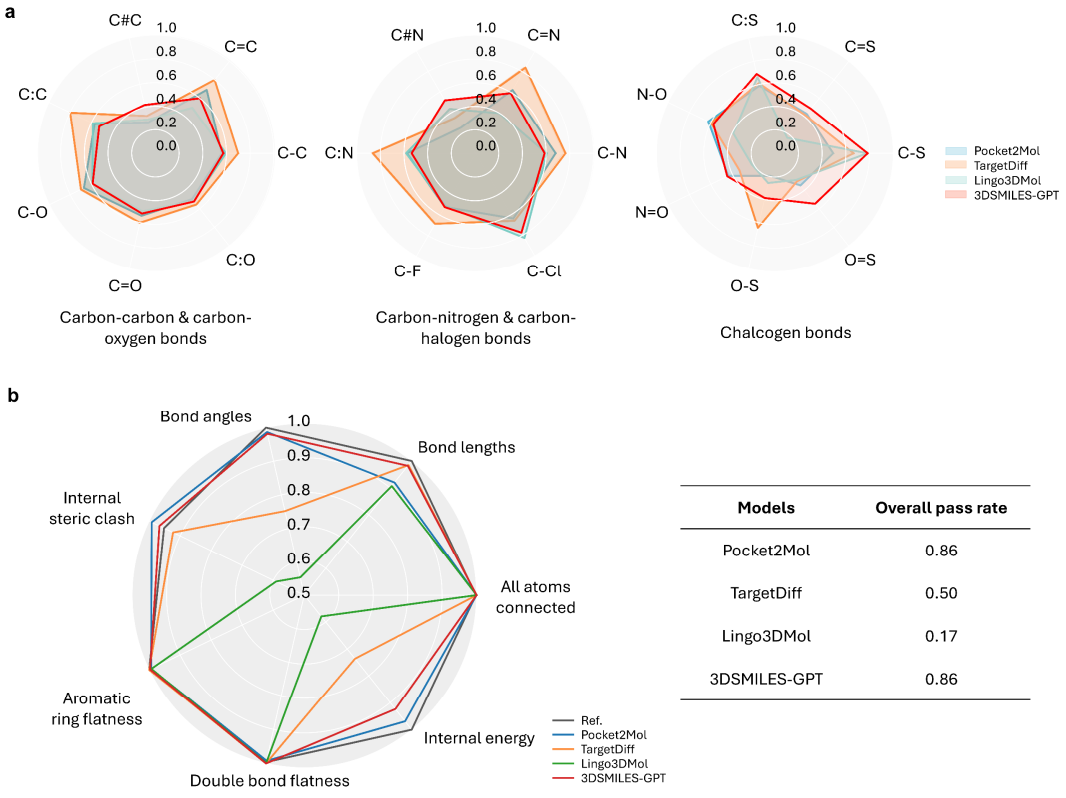
A significant challenge in DL-based molecular generation is the frequent production of non-physically plausible structures, an issue further magnified in existing token-only LLM-based methodologies. This limitation stems from the inherent difficulty that

token-only LLM-based approaches face in effectively processing continuous data, often leading to the generation of non-physical conformations. Addressing this critical concern, we aim to identify molecules that exhibit physical conformations. Building upon this premise, we delve into further exploration involving enhanced molecular affinity and desirable drug-likeness, taking into consideration the protein pocket as a constraint. In this aspect of assessment, distinct from mere generation of conformers, our focus lies predominantly on the physical plausibility of the generated molecules within the pocket.

For molecules bound within protein pockets, each type of bond lengths exhibits a robust distribution due to the constrained degrees of freedom resulting from lower flexibility. These bond lengths do not vary significantly across different constrained pockets. To evaluate the performance, we compared the distributions of some common bond lengths between the generated molecules and training molecules across various types of chemical bonds, using the Jenson-Shannon divergence (JSD)[44] as a quantification metric. We categorized the data based on the types of chemical bonds to visualize the distributions of bond lengths (**Fig. 2a**). It shows that our method maintains a most balanced overall performance with no significant weaknesses across all types of bonds, namely, no JSD values smaller than 0.4. In the first two groups shown in **Fig. 2a**, which include bonds with carbon atoms that form the basic skeleton of drug molecules, 3DSMILES-GPT generally achieved suboptimal results, slightly inferior to TargetDiff, a SOTA method based on diffusion models. Positively, with respect to chalcogen-related chemical bonds, our model achieved the best results for most of these bonds, showing its robustness. These bonds occur less frequently in drug molecules compared to other categories, but they remain crucial in key functional groups such as nitro and sulfonic groups, which are often found in antibacterial drugs. The success of 3DSMILES-GPT in these bonds is likely attributable to the GPT model's capability to retain and recall information from scarce samples, allowing it to accurately reproduce the relative positions of atoms in molecules containing these relatively rare groups.

As shown in **Table S1** in more detail, our model achieves the best performance in

over one-third of the bond length types and demonstrates comparably close to or superior performance in other bond length types compared to other models. On a global scale, our model's predictive capability for bond lengths in pocket generation tasks still lags behind TargetDiff, This discrepancy may be attributed to potential forgetting phenomena in transfer learning. The results indicate that, when compared to other methods specifically designed for pocket generation tasks utilizing GNNs or language models, the performance of our model is essentially equivalent and, in some instances, even superior. This finding underscores that using atomic coordinates as tokens for prediction can effectively reproduce the distributions of molecular bonds.



Models	Overall pass rate
Pocket2Mol	0.86
TargetDiff	0.50
Lingo3DMol	0.17
3DSMILES-GPT	0.86

Fig. 2 | **(a)** The Jensen-Shannon Divergence (JSD) of common chemical bond lengths between 3DSMILES-GPT and other models, compared to reference molecules. For ease of visual comparison, the values are presented as 1-JSD, where values closer to 1 indicate better performance. **(b)** The performance on each metrics and the overall pass rate of each model tested with PoseBusters.

For a broader assessment, we employed PoseBusters[45], a suite designed to examine the physical and chemical inconsistencies in docking and molecular generation. It offers diverse metrics for inspecting potential errors in molecular conformations. Thus, we aim for our generated molecules to achieve high pass rates across those all metrics evaluating validity, sub-structure and stereochemistry plausibility, rather than excelling solely in specific ones. As shown in **Figure 2b and Table S2**, our model consistently achieves over an 85% pass rate across multiple metrics, indicating that the majority of generated molecules adhere to the physical and chemical plausibility as observed in natural states. In contrast, other models such as Lingo3DMol and TargetDiff, while achieving optimal performance in individual metrics, exhibit subpar performance in certain specific metrics like bond angles or steric clashes, with pass rates ranging from 50% to 70%. Compared to Pocket2Mol, our model performs almost equally well across multiple metrics, achieving a pass rate of over 90% in various independent metrics.

In addition, while Pocket2Mol generated molecules with low molecular weights, molecules generated by 3DSMILES-GPT whose molecular weights were closer to the reference (reported real active molecules), thereby better aligning with the requirements of real-world drug discovery scenarios (**Table 1**). Consequently, our superior performance across various metrics and the overall higher pass rate with PoseBusters underscore the robustness and applicability of our approach. In conclusion, 3DSMILES-GPT demonstrates commendable performance in generating molecular conformations.

Molecular properties and binding mode

Initially, an assessment was conducted to evaluate the binding strength of the generated molecules by scoring them directly using AutoDock Vina with Vina score (kcal·mol^{-1}). As shown in **Table 1**, it was observed that 3DSMILES-GPT achieved notably higher average Vina scores compared to other baseline methods, even surpassing those of genuine molecules.

Molecules with large size are more likely to occupy protein pockets, leading to

higher Vina scores. This phenomenon emphasizes the importance of considering the physicochemical properties of generated molecules comprehensively. As demonstrated by Feng et al.[42], certain large, multi-ring structured molecules are unsuitable for many cases of drug development. Therefore, a thorough evaluation of the generated molecules is essential. An examination of **Table 1** indicates that the molecules generated by our model align more closely with authentic molecules in terms of molecular weight compared to other baseline methods. Regarding molecular size, TargetDiff closely resembles authentic molecules, exhibiting similar characteristics with our model. Conversely, the molecules generated by Pocket2Mol and Lingo3Dmol display undersized and oversized dimensions, respectively, in both molecular size and weight.

Table 1 | Binding energies and drug-likeness properties.

Metrics	Ref.	Pocket2Mol	TargetDiff	Lingo3DMol	3DSMILES-GPT
Mean Vina score (↓)	-7.45	-7.15	-7.11	-7.68	**-7.72**
Mean QED (↑)	0.48	0.57	0.57	0.26	**0.76**
Mean SAS (↓)	3.43	3.16	4.33	4.51	**3.07**
Drug-like molecules % (↑)	74%	94%	81%	30%	**100%**
Mol Size	22.75	17.74	22.65	40.68	23.71
Mol Weight	332.35	241.25	298.41	480.50	329.10
Validity (↑)	-	**1.00**	**0.97**	0.99	0.99
Diversity (↑)	-	**0.96**	**0.96**	0.92	0.89
BR % (↑)	-	48%	48%	**58%**	53%
BR-QED (↑)	-	0.56	0.59	0.27	**0.76**
BR-SAS (↓)	-	3.52	4.78	4.51	**3.10**

Time/s (↓)	-	13.63	12.19	1.32	**0.45**

Subsequent analyses involved a comparison of the SAS and QED of the generated molecules. As shown in **Table 1**, our model demonstrates significant advantages in both QED and SAS metrics compared to alternative pocket-aware molecular generation approaches. Notably, our model exhibits an approximate 33% improvement in QED performance over the top-performing baseline, Pocket2Mol. This highlights the superior drug-like characteristics of the molecules generated by 3DSMILES-GPT, thereby enhancing their potential pharmaceutical utility. Furthermore, when compared to other baseline models, the molecules generated by 3DSMILES-GPT also demonstrate higher SAS values, indicative of improved synthetic feasibility. Additionally, in comparison to other methodologies, our approach demonstrates a superior molecular generation speed, of only 0.45s per generation, as evaluated using an NVIDIA Tesla V100 GPU.

To further explore the interplay among the QED, SAS, and various other molecular properties, we utilized heatmap visualization (**Fig. 3**). **Fig. 3a** illustrates the relationship between the quantity of molecules and their corresponding QED and SAS values. Notably, a substantial proportion of molecules generated by 3DSMILES-GPT cluster in the bottom-left quadrant, indicating a prevalence of molecules exhibiting heightened QED and diminished SAS compared to other models. Furthermore, we investigated the influence of molecular weight on this relationship, revealing that the molecules generated by Pocket2Mol, characterized by elevated QED and reduced SAS, tend to possess smaller molecular weights (**Fig. 3b**). Subsequently, we explored the correlation between the Vina scores and QED/SAS. As depicted in **Fig. 3c**, the notably lighter coloration in the bottom-left quadrant for 3DSMILES-GPT indicates lower Vina scores and, consequently, reduced binding energies.

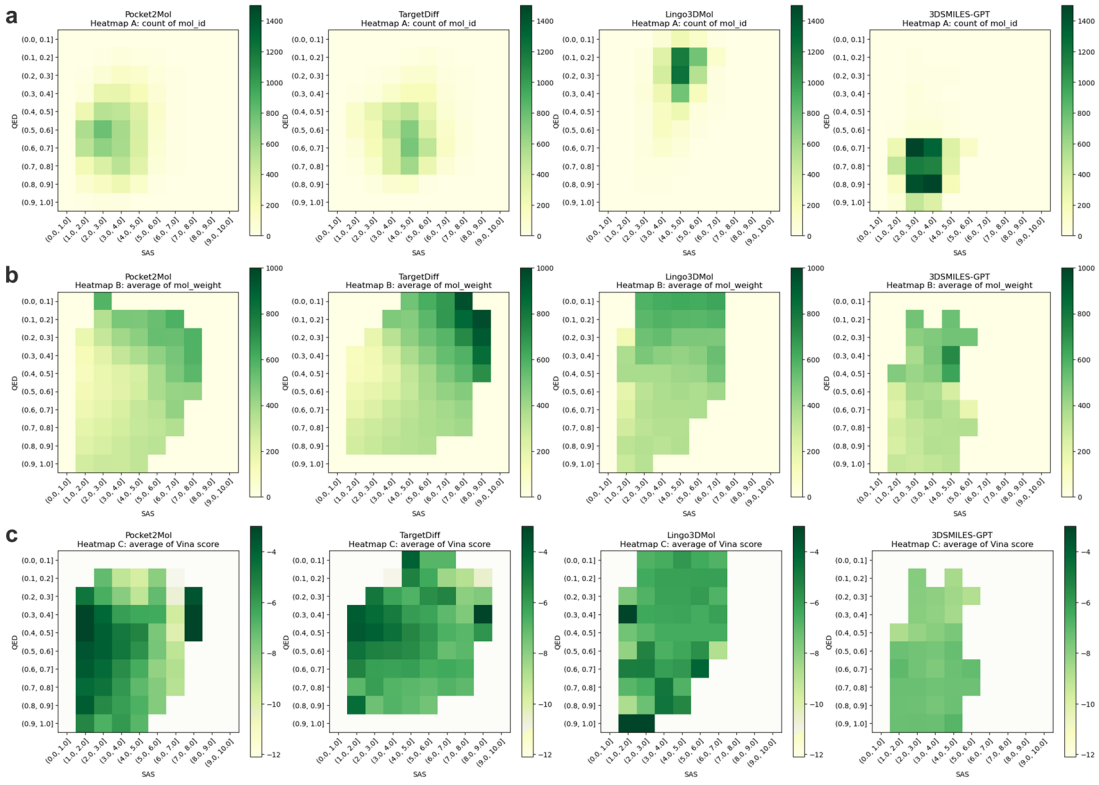


Fig. 3 | The distribution heatmaps of QED, SAS, and other properties for the molecules generated by each model. **(a)** QED, SAS and number of molecules, **(b)** QED, SAS and molecular weight, and **(c)** QED, SAS and Vina score.

In our endeavor to create molecules, our objective is to generate compounds with properties that surpass those of currently available ones. To assess the model's effectiveness in achieving this objective, we thoroughly examined the molecules generated by each model, using the ground truth molecules from the test set as a reference point. We refer to results where the affinity is superior to the reference molecules as "Better than References" (BR). The analysis revealed that our model exhibits reduced diversity compared to others, a finding that aligns with our initial expectations. This decline in diversity can be attributed to the imposition of constraints related to physicochemical properties during the training phase of 3DSMILES-GPT, coupled with additional restrictions on QED and logP during the process of molecule generation, ultimately resulting in a lower diversity of generated molecules.

We quantified the number of the molecules generated by 3DSMILES-GPT that

achieved lower Vina scores compared to the reference molecules. Notably, 53% of the molecules generated by 3DSMILES-GPT exhibited lower Vina scores compared to the reference molecules, while Pocket2Mol and TargetDiff achieved 48%. However, although the molecules generated by Lingo3DMol often exhibit higher affinity than the reference molecules, they tend to cluster within a narrow range according to the Vina scores (**Fig. 3**). This might be suitable for certain drug discovery tasks, but if the reference molecules generally have high affinity, such as with kinase targets like ATK1 and CDK2, the proportion of BR molecules may decrease. On the other hand, 3DSMILES-GPT shows the ability to explore chemical space with higher affinity for the target, which is an advantage of our model. We also computed the average QED and SAS of molecules from the BR set (**Table 1**), consistently demonstrating that 3DSMILES-GPT keeps superior performance, particularly in QED, with an improvement of approximately 33%.

In summary, 3DSMILES-GPT demonstrates the capability to generate molecules with higher binding affinity and improved QED and SAS metrics. Notably, it can produce conformations comparable to those obtained through redocking without requiring the redocking process, an outstanding capability of direct generation with physical conformation, which is absent in other models.

Structure-based drug design for specific targets

Many reported pocket-based molecular generation methods lack testing on real targets outside the training set, raising doubt on their practical efficacy in real drug design tasks. To address this, we selected four protein targets independent from the training and testing sets: AKT1 (4gv1), SARS-COV2 3CL proteinase (7d3i), CDK2 (1h00) and DDR1 (5bvk).These targets have been utilized in virtual screening[46, 47] and molecular generation tasks[33, 34, 48], allowing us to simulate real drug discovery scenarios. As shown in **Fig. 4a**, the Vina scores of the molecules generated by 3DSMILES-GPT in the target pockets of these four different protein families predominantly fall within range of -10 to -5 kcal·mol^{-1}. This distribution is closer to the left side of the horizontal axis compared to both Pocket2Mol and TargetDiff, indicating higher binding energy

and better affinity. Compared to Lingo3DMol, the affinity distribution of the molecules generated by Lingo3DMol results is concentrated between -10 and -7, with a greater focus on the high affinity range than 3DSMILES-GPT. However, our model generates more molecules in the high affinity range (Vina score $\leqslant$-10) for targets other than DDR1. In practical drug discovery scenarios, high affinity is not the only pursuit, and molecules with better drug-likeness are also needed. Therefore, we filtered all molecules and re-examined the distribution of the Vina scores for the filtered molecules (**Fig. 4b**). It can be observed that the distribution of the Vina scores for the molecules generated by 3DSMILES-GPT does not change significantly compared to before filtering, indicating that the vast majority of the generated molecules meet our drug-likeness criteria. Additionally, most high-affinity molecules with Vina scores smaller or equal than -10 generated by other baseline models were filtered out, with the majority of molecules in this range being generated by 3DSMILES-GPT. In the virtual screening process, due to the limited number of molecules that can undergo activity validation, medicinal chemists typically select a few molecules with better docking scores for testing. 3DSMILES-GPT can generate molecules with both high affinity and better drug-likeness for specific targets, implying that the model has a promising potential to discover lead compounds that may achieve activity validation at the molecular or cellular level in practical drug discovery applications.

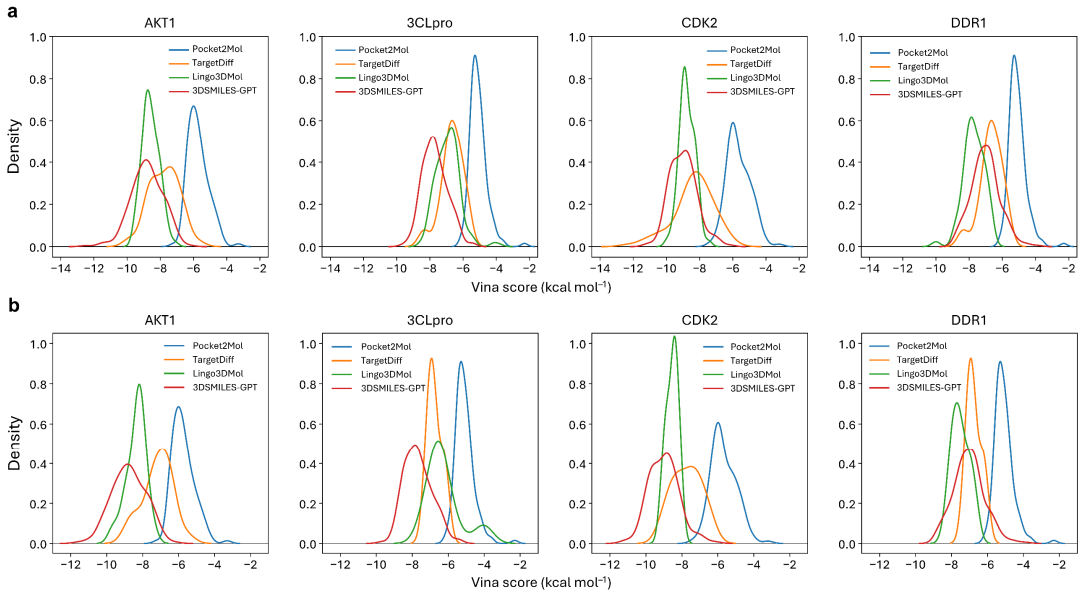


Fig. 4 | The distributions of Vina scores for the specific targets. (a) The distribution of all generated molecules for each model. **(b)** The distribution after drug-likeness screening (QED⩾0.3, SAS⩽5). The horizontal axis to the left represents better affinity.

Beyond affinity, the purpose of generating molecules within specific target pockets also involves comparing the structural and binding mode similarity of the generated molecules to known ligands. As shown in **Fig. 5**, we selected the highest affinity molecules generated by each model to demonstrate their binding modes. Among the baseline models, Pocket2Mol and TargetDiff tend to generate either simple aromatic ring derivatives or complex macrocyclic compounds, lacking distinct target specificity. The molecules generated by Lingo3DMol are excessively large compared to the original ligands, consistent with the average molecular weight of 480 Daltons obtained in the test set results. Furthermore, in the case of molecule within the CDK2 pocket, there are issues of fragmentation and clash with the protein pocket. Compared to other baseline models, the molecules generated by 3DSMILES-GPT exhibit more reasonable structures, and their binding modes within the pocket are relatively close to those of the original ligands. However, in the DDR1 pocket, the binding modes of the molecules are more exterior compared to that of the original

ligand, which may explain why 3DSMILES-GPT does not show a significant advantage in affinity for the DDR1 target over other baseline models.

Overall, the tests on specific target pockets demonstrated that 3DSMILES-GPT, compared to previous baseline models, can generate more molecules that meet drug-likeness criteria with high affinity for the target, while also exhibiting more reasonable structures and higher structural specificity for different targets.

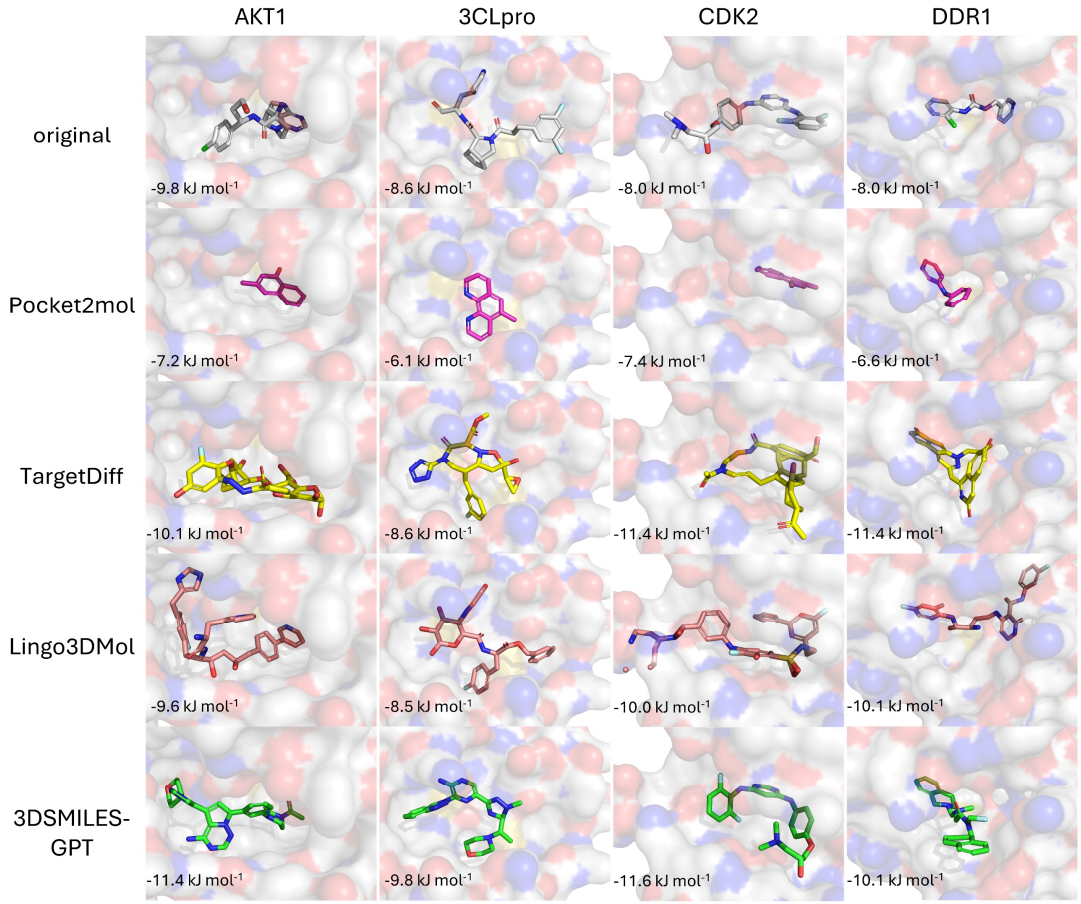


Fig. 5 | The binding modes of the molecules with optimal affinity generated by each model for specific targets, and the comparison with the binding modes of the original ligands.

Conclusion

In this study, we present 3DSMILES-GPT, an innovative token-only pocket-aware molecular generation method for creating 3D molecular structures within protein pockets. This method leverages the robust capabilities of large language models to

perceive and generate molecular structures that are not only chemically valid but also exhibit optimal biophysical and chemical properties. 3DSMILES-GPT exhibits exceptional performance across various benchmark metrics, outperforming existing methods in generating molecules with superior Vina docking scores and enhanced drug-likeness. Notably, the QED of the generated molecules improves by 33%, indicating that our model produces molecules more closely aligned with the pharmaceutical industry's criteria for drug candidates. Furthermore, 3DSMILES-GPT achieves a threefold increase in generation speed compared to the fastest existing methods, fulfilling the demand for rapid identification of drug candidates. Out-of-dataset evaluations validate the model's capability to generate drug-like molecules with strong binding affinities to specific targets, highlighting its potential in real-world drug discovery applications. Leveraging the foundation of 3DSMILES-GPT, future efforts will focus on developing a universal drug design language model by integrating advanced large model techniques with comprehensive training data. In summary, 3DSMILES-GPT signifies a paradigm shift in molecular generation for drug discovery, leveraging the capabilities of large language models to tackle complex biological challenges.

Methods

Backbone

The architecture of 3DSMILES-GPT comprises an 8-layer transformer decoder with 12 attention heads, facilitating the autoregressive prediction of both 2D and 3D molecular structures while explicitly expressing them. The multi-head attention mechanism serves as a cornerstone of the Transformer model, allowing the model to attend to different subspaces of the input simultaneously, thus capturing richer information. Within the multi-head attention mechanism, each attention head learns a set of weights to compute attention weights for different positions in the input sequence, which are then used to weigh the input sequence representations. By performing parallel computation across multiple attention heads, the model gains the ability to interpret input sequences from various perspectives, thereby enhancing its

representational capacity and generalization performance. The attention mechanism is shown in Equation 1:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k is the dimension of K.

Detachable pocket encoder

To augment the extraction of protein pocket information, a detachable pocket encoder has been devised. We implemented a spatial positional encoding strategy proposed by Zhou et al.[49], which based Gaussian kernel to describe the atomic relative positions. The D-dimensional positional encoding between atom pairs can be expressed by the following equation:

$$p_{ij} = \left\{G\left(A\left(d_{ij}, t_{ij}; u, v\right), \mu^s, \sigma^s\right) \middle| s \in [1, D]\right\}. \quad (2)$$

where G denotes Gaussian density function :

$$G(d, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}}. \quad (3)$$

A represents the affine transformation parametrized with u, v:

$$A(d, t; u, v) = u_t d + v_t. \quad (4)$$

Therefore, the information of i and j can be computed as:

$$q_{ij} = W_1 \text{GELU}\left(W_2 p_{ij}\right), \quad (5)$$

where W_1 and W_2 are learnable parameters. Ultimately, the integration between the pocket encoder and the backbone is achieved through cross-attention.

Reinforcement learning

The utilization of reinforcement learning to enhance sequence-based molecular generation models, such as REINVENT[13], is a well-established practice. However, its application to the specific task of generating 3D molecules in protein pockets remains relatively uncommon. By simplifying the intricacies of 3D molecular generation and

representing molecular coordinates as tokens, we facilitate the direct adoption of methodologies akin to REINVENT for refining the generated molecules. Nevertheless, within this context, we opt for a more explicit strategy entailing multiple iterations employing policy gradient[50] techniques to refine the model.

In the present context, $P_\theta(M|C)$ denotes the initial policy governing our model, with C representing the protein pocket and M signifying the molecule. Simultaneously, D_0 stands for the initial fine-tuning dataset. At each iteration, K molecules are sampled for every protein pocket, with those demonstrating favorable properties being incorporated into the fine-tuning dataset D_t at the t time step for subsequent iterations. Throughout each fine-tuning iteration, policy gradient methods are employed to iteratively refine the policy $P_\theta(M|C)$ of the model.

Dataset and data preprocessing

The pretraining phase encompassed a selection of 10 million molecules from the PubChem drug dataset[51], excluding those exceeding 48 atoms or containing elements beyond 'C', 'O', 'N', 'S', 'P', 'F', 'Cl', 'Br', and 'I'. Each molecule underwent stereoisomer enumeration using RDKit, followed by the generation of two conformations per stereoisomer, subsequently minimized using the MMFF94 force field. Conformational centering involved the subtraction of the coordinate center from each conformation.

For fine-tuning, the Crossdocked2020[52] dataset was employed following the Pocket2mol methodology, with poses featuring rmsd greater than 2 Å discarded. A 6 Å residue perimeter surrounding the ligand was isolated as pocket data, and the coordinates of the pocket surface were computed utilizing the MSMS[53]. The resultant coordinates underwent sparsification via pymesh.

Further data processing was conducted to meet the model's input requirements. Initially, the QED and logP (partition coefficient) values were computed. Molecules with QED exceeding 0.5 or logP values falling between -1 and 3 were assigned a label of 1, while those outside these ranges were labeled as 0. During the fine-tuning phase, we employed a similar approach to label the training data and augmented it with Vina score labels. Molecules with a Vina score less than -0.75 were labeled as 1, while those

with a score equal to or greater than -0.75 were labeled as 0.

For the 2D molecular structure representation, SMILES notation was utilized, with SMILES sequences encoded at the character byte level instead of byte-level tokenization[54]. The initial vocabulary comprised 72 characters extracted from the SMILES alphabet. Following tokenization, it was segmented into 1000 most commonly encountered tokens.

To address 3D molecular structures, data augmentation techniques were employed to instill three-dimensional equivariance into the model. This entailed random translation and rotation of the 3D structures, with each coordinate represented by a distinct token.

Baseline

We have selected SOTA models that represent three distinct approaches for pocket generation: Pocket2Mol[31], an autoregressive model based on GNN. TargetDiff, which adopts diffusion-based methodologies for one-shot generation. Lingo3DMol[42], rooted in language models. Pocket2Mol and TargetDiff underwent training utilizing the CrossDocked2020 dataset, whereas Lingo3DMol was initially pretrained on a dataset consisting of 12 million drug-like molecules, followed by fine-tuning on the PDBbind2020[55] dataset. For the evaluation in this study, we directly employed the code and pretrained models provided by the respective works.

Training and generation protocol

In the training phase, we prefixed the molecular QED and logP labels to the SMILES string, and appended the corresponding atoms' coordinates to smiles tail. Coordinates between identical atoms were delineated by commas, while those between distinct atoms were enclosed within curly braces. In the fine-tuning stage, we converted the processed coordinates of the protein pocket surface into a prefix-form input string, allowing the model to comprehend the ligand coordinate boundaries. The sequence's start and end were indicated by '<s>' and '</s>' tokens, respectively. Throughout the training regimen, we adopted a self-supervised methodology to acquaint the model

with the SMILES and coordinates strings of molecules. The primary optimization objective entailed minimizing the negative log-likelihood, as described in Equation 6:

$$\mathcal{L} = -\sum_{i=1}^{n} \log p\,(x_i|x_{<i}). \tag{6}$$

This objective was accomplished by iteratively refining the loss function through gradient descent until convergence.

In the generation stage, we combined the processed protein pocket information with the specified molecular properties, forming the input for the model. The autoregressive process for generating smiles and coordinates strings follows Equation 7:

$$p(x) = \prod_{i=1}^{n} p(x_i|x_{<i})\,. \tag{7}$$

Data availability

The datasets utilized in our study are as follows: PubChem dataset is available at https://pubchem.ncbi.nlm.nih.gov/. For pocket-based molecular generation dataset is provided at https://drive.google.com/drive/folders/1CzwxmTpjbrt83z_wBzcQncq84OVDPurM.

Code availability

The code used in the study is publicly available from the GitHub repository: https://github.com/ashipiling/GPT_3DSMILES.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (22303083), China Postdoctoral Science Foundation (2023M733128, 2023TQ0285), Postdoctoral Fellowship Program of CPSF (GZB20230657)

Author Contributions

T.J.H., Y.K. and C.Y.H. designed the research study. H. L., J.K.W. and R.Q. developed

the method and wrote the code. J.K.W., H. L., R.Q., M.Y.W. performed the analysis. J.K.W., R.Q., M.Y.W., G.Q.L., Y.K., C.Y.H. and T.J.H. wrote the paper. All authors read and approved the manuscript. Zhejiang University support in all GPU-related technologies and work content for this work.

Competing interests

The authors declare that they have no competing interests.

References

1. Brown, N., McKay, B., Gilardoni, F. & Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of Chemical Information and Computer Sciences* **44**, 1079-1087 (2004).
2. Virshup, A.M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D.N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **135**, 7296-7303 (2013).
3. De Cao, N. & Kipf, T. MolGAN: an implicit generative model for small molecular graphs. arXiv preprint, arXiv:1805.11973, 2018
4. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint*, arXiv:1802.04364 (2018).
5. Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A.L. Constrained graph variational autoencoders for molecule design. arXiv preprint, arXiv:1805.09076, 2018
6. Samanta, B. et al. NeVAE: a deep generative model for molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2019: 1110-1117.
7. Mnih, V. et al. Playing atari with deep reinforcement learning. arXiv preprint, 2013
8. Zang, C. & Wang, F. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2020: 617–626.
9. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31-36 (1988).
10. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024 (2020).
11. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., 2017: 6000–6010.
12. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **4**, 268-276 (2018).
13. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* **9**, 48 (2017).

14. Wang, J. et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence* **3**, 914-922 (2021).
15. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Molecular Informatics* **37**, 1700111 (2018).
16. Jannik Bjerrum, E. & Threlfall, R. Molecular generation with recurrent neural networks (RNNs). arXiv preprint, arXiv:1705.04612, 2017
17. Pogány, P., Arad, N., Genway, S. & Pickett, S.D. De novo molecule design by translating from reduced graphs to SMILES. *Journal of Chemical Information and Modeling* **59**, 1136-1146 (2019).
18. Liu, X., Ye, K., van Vlijmen, H.W.T., Ijzerman, A.P. & van Westen, G.J.P. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor. *Journal of Cheminformatics* **11**, 35 (2019).
19. Segler, M.H.S., Kogej, T., Tyrchan, C. & Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science* **4**, 120-131 (2018).
20. Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. & Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Science and Technology of Advanced Materials* **18**, 972-976 (2017).
21. Grisoni, F., Moret, M., Lingwood, R. & Schneider, G. Bidirectional molecule generation with recurrent neural networks. *Journal of Chemical Information and Modeling* **60**, 1175-1183 (2020).
22. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Molecular Informatics* **37**, 1700153 (2018).
23. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **4**, eaap7885 (2018).
24. Wang, M. et al. RELATION: A Deep Generative Model for Structure-Based De Novo Drug Design. *Journal of Medicinal Chemistry* **65**, 9478-9492 (2022).
25. Wang, J. et al. ChemistGA: A Chemical Synthesizable Accessible Molecular Generation Algorithm for Real-World Drug Discovery. *Journal of Medicinal Chemistry* **65**, 12482-12496 (2022).
26. Wang, J. et al. Molecular Generation with Reduced Labeling through Constraint Architecture. *Journal of Chemical Information and Modeling* **63**, 3319-3327 (2023).
27. Gebauer, N.W.A., Gastegger, M. & Schütt, K.T. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. arXiv preprint, arXiv:1906.00957, 2019
28. Ragoza, M., Masuda, T. & Koes, D.R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chemical Science* **13**, 2701-2713 (2022).
29. Liu, M., Luo, Y., Uchino, K., Maruhashi, K. & Ji, S. Generating 3D molecules for target protein binding. arXiv preprint, arXiv:2204.09410, 2022
30. Luo, S., Guan, J., Ma, J. & Peng, J. A 3D Generative Model for Structure-Based Drug Design. arXiv preprint, arXiv:2203.10446, 2022
31. Peng, X. et al. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. arXiv preprint, arXiv:2205.07249, 2022
32. Li, Y., Pei, J. & Lai, L. Structure-based de novo drug design using 3D deep generative models. *Chemical Science* **12**, 13664-13675 (2021).

33. Zhang, O. et al. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence* **5**, 1020-1030 (2023).
34. Zhang, O. et al. Learning on topological surface and geometric structure for 3D molecular generation. *Nature Computational Science* **3**, 849-859 (2023).
35. Du, H. et al. A flexible data-free framework for structure-based de novo drug design with reinforcement learning. *Chemical Science* **14**, 12166-12181 (2023).
36. Guan, J. et al. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. arXiv preprint, arXiv:2303.03543, 2023
37. Hoogeboom, E., Satorras, V.c.G., Vignac, C. & Welling, M. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022: 8867--8887.
38. Huang, L., Zhang, H., Xu, T. & Wong, K.-C. MDM: Molecular Diffusion Model for 3D Molecule Generation. arXiv preprint, arXiv:2209.05710, 2022
39. Xu, M., Powers, A.S., Dror, R.O., Ermon, S. & Leskovec, J. Geometric Latent Diffusion Models for 3D Molecule Generation. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023: 38592--38610.
40. Born, J. & Manica, M. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence* **5**, 432-444 (2023).
41. Flam-Shepherd, D. & Aspuru-Guzik, A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files. arXiv preprint, arXiv:2305.05708, 2023
42. Feng, W. et al. Generation of 3D molecules in pockets via a language model. *Nature Machine Intelligence* **6**, 62-73 (2024).
43. Zholus, A. et al. BindGPT: A Scalable Framework for 3D Molecular Design via Language Modeling and Reinforcement Learning. *arXiv preprint arXiv:2406.03686* (2024).
44. Menéndez, M.L., Pardo, J.A., Pardo, L. & Pardo, M.C. The Jensen-Shannon divergence. *Journal of the Franklin Institute* **334**, 307-318 (1997).
45. 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* **15**, 3130-3139 (2024).
46. Qiao, J. et al. SARS-CoV-2 M^{pro} inhibitors with antiviral activity in a transgenic mouse model. *Science* **371**, 1374-1378 (2021).
47. Clyde, A. et al. High-Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Noncovalent Inhibitor. *Journal of Chemical Information and Modeling* **62**, 116-128 (2022).
48. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology* **37**, 1038-1040 (2019).
49. Zhou, G. et al. Uni-mol: A universal 3d molecular representation learning framework. (2023).
50. Sutton, R.S., McAllester, D., Singh, S. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*. MIT Press, 1999: 1057–1063.
51. Kim, S. et al. PubChem Substance and Compound databases. *Nucleic Acids Research* **44**, D1202-D1213 (2015).
52. Francoeur, P.G. et al. Three-Dimensional Convolutional Neural Networks and a Cross-Docked

Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling* **60**, 4200-4215 (2020).

53. Sanner, M.F., Olson, A.J. & Spehner, J.-C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **38**, 305-320 (1996).

54. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Association for Computational Linguistics, 2016: 1715-1725.

55. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **47**, 2977-2980 (2004).

Supplementary Information

Table S1 | Jensen–Shannon divergence ($\times 10^{-3}$) between the bond length

Bond	Pocket2Mol	TargetDiff	Lingo3DMol	3DSMILES-GPT
C-C	0.4422	**0.3035**	0.4062	0.4285
C=C	0.3091	**0.2014**	0.5038	0.3994
C#C	0.7276	0.6726	0.7048	**0.577**
C:C	0.4273	**0.199**	0.411	0.4711
C-N	0.3251	**0.2419**	0.4	0.4183
C=N	0.3776	**0.1597**	0.6044	0.4086
C#N	0.7487	0.6514	0.5625	**0.4789**
C:N	0.4186	**0.1315**	0.4128	0.4596
N-O	0.3743	**0.3997**	0.6086	0.4234
N=O	0.5625	0.6684	0.7832	**0.5534**
C-O	0.3272	**0.2997**	0.4084	0.409
C=O	0.4604	**0.4006**	0.4992	0.4796
C:O	0.5066	**0.4483**	0.486	0.4805
C-F	0.4775	**0.3141**	0.4882	0.4769
C-S	0.5076	0.3349	0.2188	**0.2164**
C=S	0.5684	0.5845	0.8325	**0.5128**
C:S	0.4105	0.388	0.3368	**0.3084**
O-S	0.8054	**0.3567**	0.7434	0.6134

O=S	0.6534	0.6987	0.7126	**0.4556**
C-Cl	0.3652	0.3433	**0.1766**	0.2284

Table S2 | The performance in PoseBusters benchmark.

Metrics	All atoms connected	Bond lengths	Bond angles	Internal steric clash	Aromatic ring flatness	Double bond flatness	Internal energy	Overall pass Rate
Ref.	≈1.00	1	1	0.95	1.00	≈1.00	1	0.95
Pocket2Mol	1	0.92	**0.99**	**0.99**	0.99	0.99	**0.97**	**0.86**
TargetDiff	1	0.99	0.75	0.92	1	**1**	0.74	0.50
Lingo3DMol	1	0.91	0.55	0.59	0.99	≈1.00	0.58	0.17
3DSMILES-GPT	1	**0.98**	0.98	0.97	**≈1.00**	**1**	0.92	**0.86**