nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00808-8

PocketFlow is a data-and-knowledge-driven structure-based molecular generative model

Received: 18 June 2023

Accepted: 8 February 2024

Published online: 11 March 2024

Check for updates

Yuanyuan Jiang[1,2,6], Guo Zhang[1,2,6], Jing You[1,2,6], Hailin Zhang[1,2,6], Rui Yao[1,2,6], Huanzhang Xie[3], Liyun Zhang[4], Ziyi Xia[1,2], Mengzhe Dai[1,2], Yunjie Wu[5], Linli Li[5] & Shengyong Yang[1,2]

Deep learning-based molecular generation has extensive applications in many fields, particularly drug discovery. However, the majority of current deep generative models are ligand-based and do not consider chemical knowledge in the molecular generation process, often resulting in a relatively low success rate. We herein propose a structure-based molecular generative framework with chemical knowledge explicitly considered (named PocketFlow), which generates novel ligand molecules inside protein binding pockets. In various computational evaluations, PocketFlow showed state-of-the-art performance, with generated molecules being 100% chemically valid and highly drug-like. Ablation experiments prove the critical role of chemical knowledge in ensuring the validity and drug-likeness of the generated molecules. We applied PocketFlow to two new target proteins that are related to epigenetic regulation, HAT1 and YTHDC1, and successfully obtained wet-lab validated bioactive compounds. The binding modes of the active compounds with target proteins are close to those predicted by molecular docking and further confirmed by the X-ray crystal structure. All the results suggest that PocketFlow is a useful deep generative model, capable of generating innovative bioactive molecules from scratch given a protein binding pocket.

Innovative drug discovery is an extremely complex and expensive process that mainly includes retrieval of an active seed compound (often called a hit or lead compound), hit/lead optimization, preclinical evaluations and clinical trials. Of these, the retrieval of hit/lead compounds is the first and critical step because it is the foundation for starting a new drug development project and can substantially impact subsequent drug development steps[1–3]. Conventionally, hit/lead discovery is accomplished through high-throughput screening against known compound libraries. However, the limited structural diversity of the existing compound libraries and long-term continuous screening by various drug development institutes or companies make it increasingly difficult to retrieve new active compounds and establish intellectual property rights.

Deep generative models (DGMs), which have achieved great success in producing images, texts and voices[4], offer an efficient approach to generating completely new seed compounds[5–9]. Commonly used DGMs for molecular generation are mainly ligand-based[10–14], which first employ neural networks to learn a probability distribution of structural information from a large number of known active compounds and then generate new molecular structures by sampling the learned distribution. However, a majority of these ligand-based DGMs do not consider the structural information of target proteins, and until recently such

[1]Department of Biotherapy, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, China. [2]New Cornerstone Science Laboratory, West China Hospital, Sichuan University, Chengdu, China. [3]College of Materials and Chemical Engineering, Minjiang University, Fuzhou, China. [4]Lead Generation Unit, HitGen Inc., Chengdu, China. [5]Key Laboratory of Drug Targeting and Drug Delivery System of Ministry of Education, West China School of Pharmacy, Sichuan University, Chengdu, China. [6]These authors contributed equally: Yuanyuan Jiang, Guo Zhang, Jing You, Hailin Zhang, Rui Yao. e-mail: yangsy@scu.edu.cn

ligand-based DGMs have been extended to structure-based approaches that incorporate explicit information of the targeted protein[2]; considering the receptor's structural information is thought to be beneficial for improving the accuracy of drug design because a drug molecule must precisely bind to its corresponding target protein in the body to exert its specific therapeutic effect. Furthermore, many proteins have only very few or no known ligand molecules, for which ligand-based DGMs cannot be adopted to generate molecules. Structure-based DGMs that generate novel ligand molecules inside protein binding pockets are expected to be able to overcome the drawbacks of ligand-based ones and are attracting more and more attention[15–17].

Although structure-based DGMs can generate new molecular structures conditional on protein pockets, many challenging problems remain, which are summarized as follows. (1) The existing dataset, composed of known experimental protein–ligand complex structures, is small and insufficient for training a generative model. (2) Currently, structure-based DGMs are still a data-driven approach; there is a growing view that incorporating domain knowledge or rules into deep learning models can effectively solve the problems of lack of data, robustness and poor interpretability[18–21]. (3) Presently, most deep molecular generative models do not consider chemical bond information in the training and generation process. Instead, they output a set of discrete atoms without connectivity and then assemble these atoms to form molecules by third-party methods, such as OpenBabel[22]. This strategy may result in many undesirable substructures that cause difficulty in chemical synthesis or low drug-likeness. (4) Although the generated molecules have been validated theoretically, their bioactivity and binding modes are not verified by wet-lab experiments.

Motivated by the challenges mentioned above, we propose a structure-based molecular generative framework driven by data and chemical knowledge, named PocketFlow. The main contributions of this work are as follows. (1) We explicitly incorporate chemical domain knowledge in three-dimensional (3D) molecular generation, which provides knowledge guidance and ensures the validity and rationality of the generated molecules. (2) Inspired by molecular dynamic simulation, we not only consider atoms particles that have properties and positions but also inject the topology knowledge of protein and ligands into the model to capture subtle intermolecular interactions. (3) The covalent bond distribution is explicitly modelled by the model to enhance structural inference. Triangular self-attention is also introduced, which adds a geometric constraint to the molecular generation. (4) Transfer learning is used in model training, which alleviates the lack of complex data and achieves a larger search scope in chemical space during generation. (5) We generate molecules in any continuous position without discretizing the 3D space, thereby enabling more flexible atom placement. (6) Finally, the proposed model is applied to two new target proteins that are related to epigenetic regulation, HAT1 and YTHDC1, and we successfully obtain wet-lab-validated active hit compounds.

Results

An overview of the framework of PocketFlow

PocketFlow is an autoregressive flow-based generative model that generates small organic molecules inside protein pockets in a stepwise manner. Essentially, it establishes an invertible mapping between a base distribution (such as a normal distribution) and the distributions of atom types and covalent bonds in a given 3D pocket by model training. At each step of the generation process, the model samples from the base distribution to generate a new atom along with the corresponding coordinates and covalent bonds. The molecular generation process for one step (for example, the tth step) is briefly described as follows (for details, see Methods). The model takes the binding site and the generated molecular fragment as environment information (Fig. 1a). The Context Encoder module encodes the environment information to extract the environment features $C^{(t-1)}$ (Fig. 1b). Before generating new components, an auxiliary network, Focal Net, is used to select a focal atom from the environment features as a reference point for generating a new atom (Fig. 1c). Based on these environmental features, PocketFlow employs a strategy of sequence dependency[15] to generate a new atom: atom type $a^{(t)}$, coordinate $r^{(t)}$ and covalent bonds $e^{(t)}$ (Fig. 1d–f), that is, $a^{(t)} \rightarrow r^{(t)} \rightarrow e^{(t)}$.

$$a^{(t)} = \text{AtomFlow}\left(C^{(t-1)}\right) \quad (1)$$

$$r^{(t)} = \text{PosPred}\left(C^{(t-1)}, a^{(t)}\right) \quad (2)$$

$$e^{(t)} = \text{BondFlow}\left(C^{(t-1)}, a^{(t)}, r^{(t)}\right) \quad (3)$$

During the generation process, we explicitly include chemical knowledge constraints to ensure that the topology (Fig. 1f) and conformation (Fig. 1e) of the generated molecules are reasonable (Methods and Supplementary Information). If the newly generated components do not meet the chemical knowledge constraints, the model will resample and generate other components to meet these constraints. At the end of the generation step, $a^{(t)}$, $r^{(t)}$ and $e^{(t)}$ are added to the environment as inputs for the next generation step (Fig. 1g). The generation process will be terminated if any of the follow conditions are met: (1) no atom can be predicted to be a focal atom, (2) the number of generated atoms reaches the predefined maximum value and (3) the resampling times reach the predefined maximum number.

Evaluation of common properties for generated molecules

We first evaluated common properties of molecules generated by PocketFlow, including quantitative estimate of drug-likeness (QED)[23], LogP (the octanol-water partition coefficient), synthetic accessibility (SA)[24], diversity[25] and validity. To make a comparison, we also calculated the five properties of molecules generated by the current state-of-the-art (SOTA) structure-based DGMs: LiGAN[17], GraphBP[15] and Pocket2Mol[16]; these models will be taken as the baselines in subsequent evaluations as well. Here, the same test set as that used in the evaluation of LiGAN was employed[17], which includes ten different types of target proteins with Protein Data Bank (PDB) IDs: 2ah9, 5lvq, 5g3n, 1u0f, 4bnw, 2ati, 2hw1, 1bvr, 1zyu and 4i91. For each protein pocket, we generated 10,000 molecules by using PocketFlow and the three baselines separately (see Supplementary Information for details). For comparison purposes, we also calculated the properties of molecules from the CrossDocked2020 dataset[26], which contains approximately 13,000 real and drug-like small molecular ligands binding to protein pockets.

The calculated average values of the five properties for molecules generated by different DGMs and for the CrossDocked2020 molecules are shown in Table 1 and Supplementary Tables 1–4. For QED, only PocketFlow has an average value above 0.5 among all models, which is very close to the QED value of CrossDocked2020 molecules, indicating that PocketFlow has a better ability to generate drug-like molecules. The average LogP values of molecules generated by PocketFlow and the three baselines are between 0.552 and 3.719, which are within the commonly recognized LogP range of drug-like molecules. The SA score for PocketFlow is 2.927, which is close to that of CrossDocked2020 molecules (3.246). In contrast, the SA scores for the three baselines are approximately 5. These results imply that molecules generated by PocketFlow are likely easier to chemically synthesize than those generated by other models. Consistent with the average values of QED, SA and LogP, the distributions of these properties also indicate that molecules generated by PocketFlow are closer to the CrossDocked2020 molecules than those generated by the three baselines (Supplementary Fig. 1 and Supplementary Table 5). Regarding diversity, molecules generated by PocketFlow show comparable diversity to those generated by the three baselines. Of special note is that all molecules generated by PocketFlow are chemically valid (validity: 100%). GraphBP also generated highly valid molecules (validity: 99.6%), whereas LiGAN showed poor performance in this respect.

https://doi.org/10.1038/s42256-024-00808-8

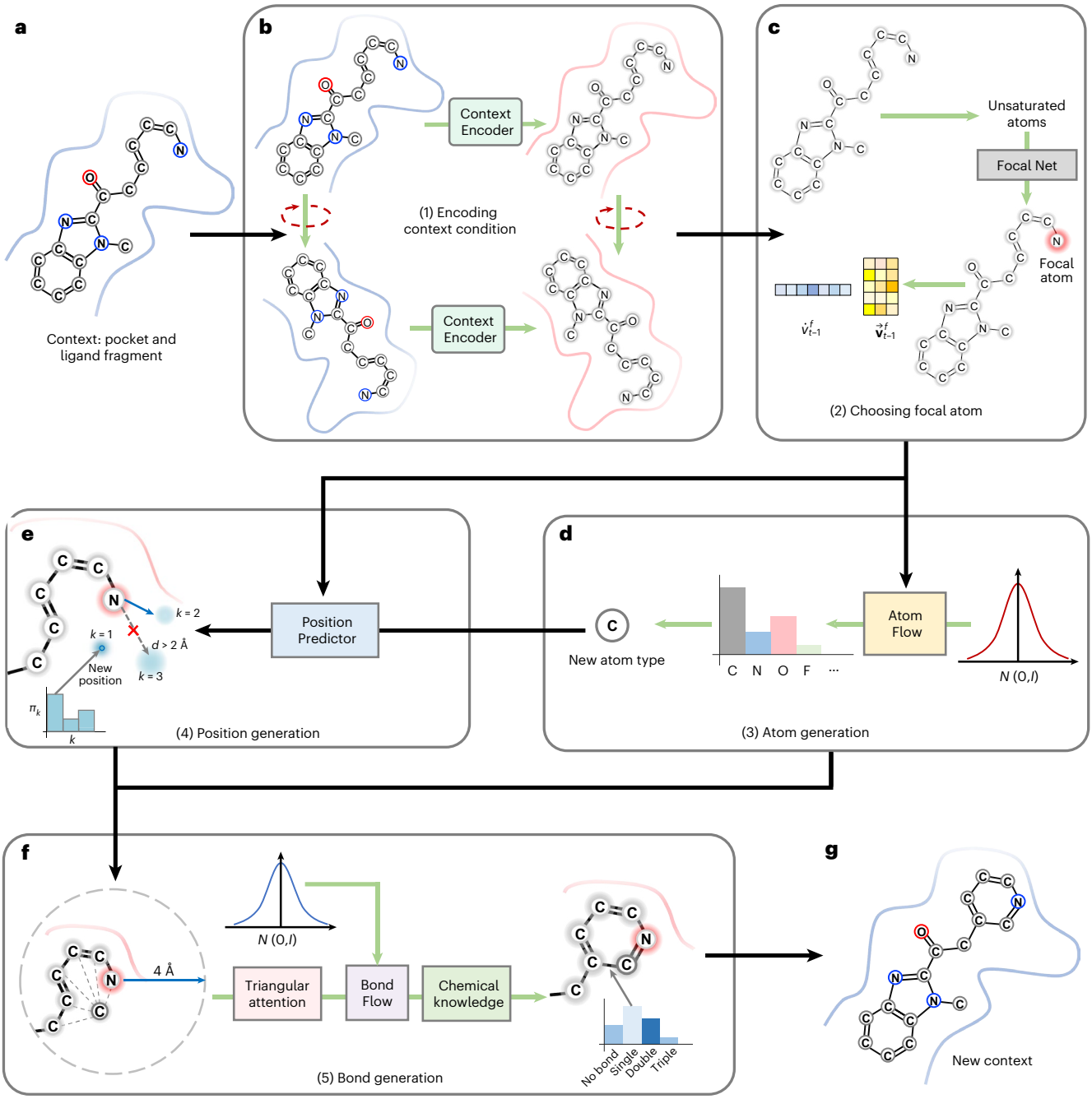


Fig. 1 | Architecture and generative process of PocketFlow. a, The context (step t) includes the protein pocket (blue curve) and the ligand fragment generated in previous $t-1$ steps. **b**, The Context Encoder module is used to extract features of the context as conditional information for generating new atoms. The dashed circles with arrows denote rotation operations. The hidden states of pockets and molecular fragments are indicated by red and grey curves, respectively. **c**, The Focal Net module is used to choose focal atoms. The embedding of the focal atom is $(v^f_{t-1}, \mathbf{v}^f_{t-1})$, where v and $\mathbf{v}$ denote the scalar features and vector features, respectively. **d**, Atom generation. The type of the new atom is generated by sampling from the standard normal distribution $N(0,I)$ and combining the embedding of the focal atom. **e**, Position generation. The Position Predictor module is an MDN[16,56], which uses a neural network to model the parameters of Gaussian mixture models. The MDN combines the embedding of the focal atom $(v^f_{t-1}, \mathbf{v}^f_{t-1})$ and the new atom type to generate candidate coordinates for the new atom. We only consider coordinates that are less than 2 Å from the focal atom. k is the kth component in the Gaussian mixture model built by the MDN, and π_k is the mixture weight of component k. **f**, Bond generation. Only atoms less than 4 Å from the focal atom are selected as candidates for generating covalent bonds with new atoms. The triangular attention mechanism and chemical knowledge are involved to add geometric constraints to the generation of covalent bonds. **g**, The newly generated atoms, coordinates and covalent bonds are added to the current context and used as the context for the next generation step.

Rationality of chemical structures for generated molecules

We then evaluated the rationality of chemical structures for the generated molecules, including bond lengths, bond angles and ring structures. Nine kinds of common covalent bonds are analysed, including C–C, C=C, C–O, C=O, C–N, C–F, C=N, C–Cl and C–S. As shown in Fig. 2a–i and Extended Data Table 1, for all nine covalent bonds, the bond length distribution of molecules generated by PocketFlow is closer to that of the CrossDocked2020 molecules than that of

Table 1 | Five common properties (averages) of molecules generated by different DGMs and molecules from the CrossDocked2020 dataset

Properties	CrossDocked2020	LiGAN	Pocket2Mol[a]	GraphBP	PocketFlow
QED[b]	0.531±0.210	0.480±0.083	0.397±0.121	0.415±0.004	0.507±0.051
SA[c]	3.246±0.981	4.809±0.160	4.837±0.646	5.814±0.038	2.927±0.372
LogP[d]	2.286±2.518	0.552±0.893	2.469±2.628	3.618±0.122	3.719±1.074
Diversity[e]	–	0.889±0.005	0.853±0.017	0.888±0.002	0.871±0.014
Validity (%)[f]	–	81.4±4.7	–	99.7±0.1	100.0±0.0

Data are shown as mean±s.d. [a]Because Pocket2Mol failed to generate a specified number of molecules for some target proteins for unknown reasons, we had to use the actually generated molecules for statistical analysis. [b]QED, an index of drug-likeness, with a value between 0 (drug-unlike) and 1 (drug-like)[23]. [c]SA, the difficulty of chemical synthesis for molecules, with a value between 0 (easy to synthesize) and 10 (very difficult to synthesize)[24]. [d]LogP, an important parameter to characterize the overall hydrophobicity of organic compounds. [e]Diversity, a measurement of the chemical structural similarity among a set of molecules, with a value between 0 (poor diversity) and 1 (excellent diversity)[25]. [f]Validity, the percentage of the number of chemically valid molecules to the total number of generated molecules.

molecules generated by the three baselines. Despite better performance of Pocket2Mol for some covalent bond types, the overall performance of PocketFlow is still superior to that of Pocket2Mol. Eight common bond angles are evaluated, including CCC, CC=C, OC=O, NC=O, CNC, COC, CN=C and CSC. Again, the bond angle distributions of molecules generated by PocketFlow are closer to those of the CrossDocked2020 molecules than those of molecules generated by the three baselines (Extended Data Fig. 1 and Extended Data Table 2).

For the ring structures, five- and six-membered rings and their fused rings (such as 6 + 6, 6 + 5 and 5 + 5) are the most common in drug-like molecules, whereas rings such as three-, four-, seven-membered and more larger rings or polycyclic rings (≥3 rings fused; Supplementary Fig. 2a–g) are uncommon or disfavoured in drug-like molecules due to their poor synthesis accessibility, low chemical instability, high toxicity and metabolic instability. As shown in Fig. 2j–o, the percentages of molecules containing the uncommon or disfavoured rings mentioned above are very low among molecules generated by PocketFlow, very similar to those of CrossDocked2020 molecules (cyan in Fig. 2j–o). In contrast, more molecules generated by the three baselines contain the uncommon or disfavoured rings (Supplementary Fig. 2).

Binding sites and ligand efficiency of generated molecules

We next evaluated the detailed binding sites and binding affinities/ligand efficiency (LE) of generated molecules. To analyse the detailed binding sites of generated molecules, we randomly selected 1,000 molecules for each target from the 10,000 molecules generated above to do statistical analysis due to the limited capacity of the program (PyMol) we used. As shown in Fig. 3 and Extended Data Fig. 2, for all ten proteins with different pocket shapes, molecules generated by PocketFlow are mainly located at the inner protein pockets, whereas molecules generated by GraphBP are distributed dispersedly around the protein pockets, with a large number of molecules located outside the pockets. Molecules generated by Pocket2Mol are also mainly located inside the pocket, but the atomic distribution of Pocket2Mol is likely sparser than that of the other three models, which could be due to the low diversity of molecules generated by Pocket2Mol and the close proximity of the atomic coordinates between different molecules, leading to many atoms overlapping or clustering with each other in space. Compared with GraphBP and Pocket2Mol, LiGAN shows better performance in some target proteins (for example, 1bvr, 2ati), and a possible reason could be that LiGAN forcibly introduces an additional generation boundary that constrains the generation scope into a box[17].

To evaluate binding affinities and LE of generated molecules, ChemScore[27,28] was used to directly estimate the binding affinities between generated molecules and corresponding target proteins. PocketFlow and Pocket2Mol show the best performance in terms of the mean ChemScore values, with Pocket2Mol being slightly better than PocketFlow (Supplementary Table 6 and Supplementary Fig. 3). However, PocketFlow is better than Pocket2Mol in terms of the LE values (Table 2), which were obtained by dividing the ChemScore[27,28] values by the numbers of heavy atoms. In general, selecting a molecule with a large LE value as a hit/lead compound may imply a large space for subsequent structural optimization[29–32]. It is worth noting that GraphBP offered large negative ChemScore and LE values, which are abnormal. A careful inspection showed that molecules generated by GraphBP have many atoms distributed outside the pockets, and some atoms had a clear clash with receptor atoms, leading to a penalty score (Fig. 3 and Extended Data Fig. 2). Distributions of ChemScore values show that PocketFlow and Pocket2Mol have a higher probability of generating molecules with high binding affinity (Supplementary Fig. 3). And distributions of LE indicate that PocketFlow tends to generate molecules with higher LE (Supplementary Fig. 4).

Ablation analyses

In all the above computational experiments, PocketFlow showed SOTA performance, largely due to the adoption of many advanced techniques, including the constraint of chemical knowledge, pretraining and topological knowledge of proteins. To analyse the impact of each technique on model performance, we carried out ablation experiments. As a result, we obtained three new models: the no-knowledge model, the no-pretraining model and the no-protein-topo model. The no-knowledge model is the model with the chemical knowledge constraint removed. The no-pretraining model is the model trained directly by using the CrossDocked2020 dataset. The no-protein-topo model is the model with no protein topology knowledge injected during the training process. Again, we used the ten protein targets mentioned before as a test set and generated 10,000 molecules by using each model for each target. Various properties were then calculated for the generated molecules as before.

The results (Supplementary Table 8 and Supplementary Figs. 5–7) showed that ignoring the constraint of chemical knowledge substantially decreased model performance, particularly with the worst validity (34.0%) and QED (0.324), indicating that most of the generated molecules are chemically invalid and their drug-likeness becomes poorer. Further, the mean SA value of the molecules generated by the no-knowledge model is close to 5, implying that, on average, they are more difficult to synthesize compared with molecules generated by the full PocketFlow model. Unexpectedly, molecules generated by the no-knowledge model have the largest mean ChemScore value (19.032), which is even larger than that of the full model (17.117). However, in terms of the LE value, the full model is still the best. Both the no-pretraining model and the no-protein-topo model generated 100% chemically valid molecules. Nevertheless, many properties of generated molecules related to drug-likeness or druggability

https://doi.org/10.1038/s42256-024-00808-8

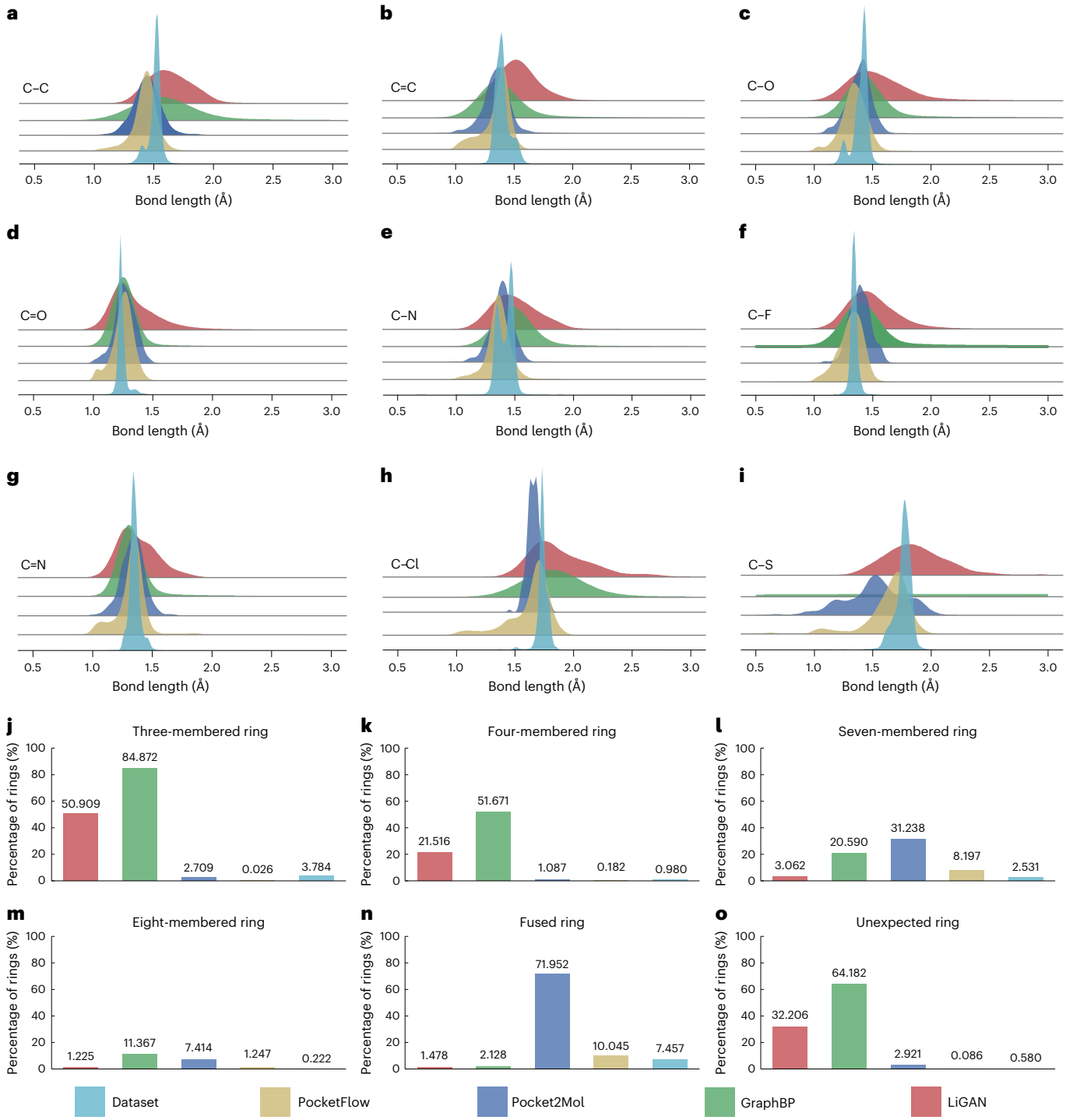


Fig. 2 | Evaluation of the geometry for generated molecules. a–i, The bond length distributions of molecules generated by different DGMs and molecules in CrossDocked2020. Nine kinds of chemical bonds are analysed, including C–C (**a**), C=C (**b**), C–O (**c**), C=O (**d**), C–N (**e**), C–F (**f**), C=N (**g**), C–Cl (**h**) and C–S (**i**). **j–o**, The proportion of molecules containing different ring structures, including a three-membered ring (**j**), a four-membered ring (**k**), a seven-membered ring (**l**), an eight-membered ring (**m**), a fused ring (**n**) (Supplementary Fig. 2a–f) and an unexpected ring (**o**) (Supplementary Fig. 2g). The results of PocketFlow, Pocket2Mol, GraphBP and LiGAN are indicated in yellow, blue, green and red, respectively, and those of molecules from the CrossDocked2020 dataset are depicted in cyan.

were worse than those of the full model but likely better than those of the no-knowledge model. Overall, the data obtained here demonstrate that the constraint of chemical knowledge, pretraining and topological knowledge of proteins have a notable impact on model performance, and the constraint of chemical knowledge is the most critical to ensure the validity and drug-likeness of the generated molecules.

Wet-lab validation of PocketFlow

Finally, to validate PocketFlow by wet-lab experiments, we applied PocketFlow to two target proteins: histone acetyltransferase 1 (HAT1) and YTH domain-containing protein 1 (YTHDC1). The two proteins are newly recognized potential targets for disease treatment and have recently attracted much attention. However, there is still a lack of small molecule inhibitors targeting these proteins.

https://doi.org/10.1038/s42256-024-00808-8

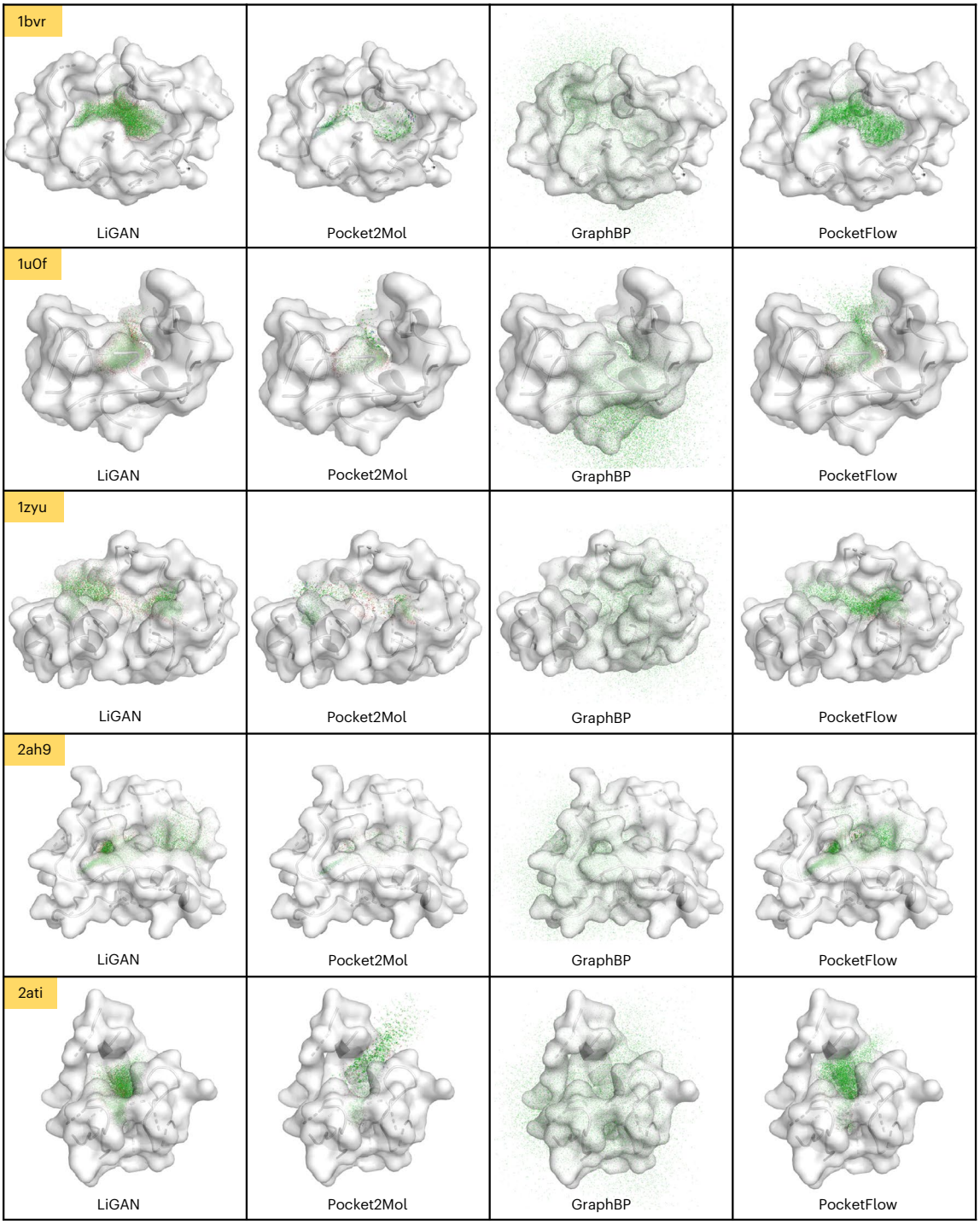


Fig. 3 | Distributions of atom positions for 1,000 molecules randomly selected from molecules generated by different DGMs. Target proteins and their active pockets are displayed as protein surfaces (white). Green points indicate heavy atoms of molecules. Pocket2Mol failed to generate a sufficient number of molecules against the target protein 2ah9, leading to an obvious sparse distribution of heavy atoms.

Case 1: HAT1

HAT1 is an enzyme that catalyses the acetylation of histone H4 on lysines 5 and 12 during the process of chromatin assembly[33]. HAT1 plays an important role in many physiological processes. Dysregulation of HAT1 has been linked to various human diseases, particularly cancer[34–36]. For example, HAT1 has been demonstrated to be an oncogene in glioma, and silencing HAT1 reduces proliferation and migration and increases apoptosis[37]. HAT1 has thus been thought of as a potential target for the treatment of related diseases.

Herein, PocketFlow was applied to generate new active seed inhibitors of HAT1. The 3D structure of HAT1 was taken from the crystal structure of HAT1 in complex with a coenzyme (PDB ID 6vo5), and the catalytic site of HAT1 was chosen as the active pocket. We first generated 100,000 molecules by using PocketFlow (50,000 each with two Tesla P100 GPUs, running for 2 days). To narrow the choices, we only considered molecules (1) with 600 > molecular weight > 400, (2) with QED > 0.4, (3) with SA score < 3.0 and (4) not containing alert substructures[38], uncommon rings (Supplementary Fig. 2b–g) or big rings (ring

Table 2 | The average LE values of molecules generated by various DGMs

DGM	LE values
LiGAN	0.652±0.463
GraphBP[a]	−29.492±53.739
Pocket2Mol	0.649±0.442
PocketFlow	0.980±0.372

Data are shown as mean±s.d. These values are the mean of the mean of LE of the ten targets. LE for each target is shown in Supplementary Table 7. [a]GraphBP generates many molecules outside the pocket or clashing with the atoms of proteins, resulting in abnormal scores.

size > 7). A total of 125 molecules met all the conditions and were then sorted by their QED values. From the top ten molecules, we selected two molecules (H1 and H9, Supplementary Fig. 8) that can be easily and quickly synthesized. We then chemically prepared H1 and H9 and tested their bioactivity. Synthetic routes of H1 and H9 are described in the Supplementary Information. H9 (Fig. 4a) showed bioactivity, with a half maximal inhibitory concentration (IC_{50}) value of 72.36 μM (Fig. 4b). The bioactivity of H9 was validated by a differential scanning fluorimetry (DSF) assay, which gave a thermal shift (ΔT_m, melting temperature) of 1.83 °C (Supplementary Table 16). We next docked H9 to the binding pocket of HAT1 and superimposed the predicted HAT1–H9 complex structure onto the generated HAT1–H9 complex structure, which showed that the binding pose of H9 generated by PocketFlow was very close to that predicted by molecular docking (Fig. 4c). We finally conducted a preliminary evaluation of the drug-likeness of H9, and the results showed that H9 could be used as a starting hit compound for further studies (Supplementary Table 17).

Case 2: YTHDC1

YTHDC1 is a kind of epigenetic regulation protein that specifically recognizes the N^6-methyladenosine (m^6A) RNA modification; m^6A is the most abundant chemical modification mark in eukaryotic RNAs and is recognized by m^6A-recognition proteins ('readers'), thereby exerting its biological functions[39]. YTHDC1 is the only nuclear RNA m^6A reader of YTH domain-containing proteins. It has distinct roles in regulating nuclear RNA splicing, alternative polyadenylation, nuclear export and decay[40]. Dysregulation of YTHDC1 is associated with a number of pathologies, particularly acute myeloid leukaemia[41]. YTHDC1 has thus been considered a potential target for the treatment of acute myeloid leukaemia, and small molecule inhibitors of YTHDC1 could be potential disease intervention agents.

Again, PocketFlow was adopted to design active seed inhibitors of YTHDC1. The 3D structure was taken from the crystal structure of YTHDC1 in complex with a m^6A RNA substrate (PDB ID 4r3i), and the substrate binding site of YTHDC1 was chosen as the active pocket. Similar to before, we generated 100,000 molecules. To reduce the range of choices, we only kept molecules (1) with 400 > molecular weight > 250, (2) with QED > 0.9, and (3) with SA score < 3.0 (4) not containing alert substructures[38], uncommon rings (Supplementary Fig. 2b–g) and big rings (ring size > 7); the molecular weight was set to between 400 and 250 because the active pocket of YTHDC1 is small. The conditions were met by 235 molecules, which were then subjected to molecular docking and sorted by predicted LE values. From the top five molecules, we chose three (Y1, Y3 and Y5; Supplementary Fig. 11) that look much easier to synthesize. We then prepared the three compounds by chemical methods and tested their bioactivity. Synthetic routes of Y1, Y3 and Y5 are described in the Supplementary Information. Two compounds, Y3 and Y5, displayed activity against YTHDC1 (Supplementary Fig. 11). Y3 (Fig. 4d) is the most active, with an IC_{50} value of 32.6 μM (Fig. 4e). The bioactivity of Y3 was further verified by DSF and isothermal titration calorimetry assays, which offered a thermal shift (ΔT_m) of 1.08 °C in the DSF assay and an equilibrium dissociation constant (K_D) value of 108 μM in the isothermal titration calorimetry assay (Supplementary Table 16). Figure 4f displays the superimposition of the binding mode generated by PocketFlow and that predicted by molecular docking, indicating a very similar binding mode.

To verify whether the generated molecules truly bind to the target protein pockets, we solved the cocrystal structure of YTHDC1 in complex with Y3 by X-ray crystallography at a resolution of 1.6 Å (PDB ID 8k2e; Supplementary Table 18); we failed to obtain the cocrystal structure of HAT1 in complex with H9, which may be due to the low bioactivity of H9. As shown in Fig. 4g, Y3 binds to a pocket of YTHDC1, which is the same pocket in which we generated small molecular ligands. The 2-chlorobenzoic acid group of Y3 occupies a hydrophobic region formed by N363, N364, N367, W377, L380, W428, L430, P431, M434, M438 and L439, which is the m^6A binding site. The carboxyl forms four hydrogen bonds with residues N363, N364 and N367. These results clearly demonstrate the ability and effectiveness of our model in generating bioactive ligands inside a designated binding pocket of a target protein.

We further tested the activity of Y3 against other subfamily members of YTHDC1, including YTHDC2 and YTHDF1–3. Y3 did not show activity against these subfamily members (IC_{50} > 1,000 μM; Supplementary Fig. 12), implying a good selectivity for YTHDC1. Again, a preliminary evaluation of the drug-likeness showed that Y3 could be taken as a starting hit compound for later studies (Supplementary Table 17).

Discussion

In this study, we establish a data-and-knowledge dual-driven DGM, PocketFlow, for structure-based de novo drug design. In PocketFlow, an optimal vector-based equivariant graph neural network, a geometric double bottleneck perceptron (GDBP; Methods), is proposed to model the geometry of the protein–ligand complex. To capture the interaction information between protein and ligand, the topology knowledge of proteins and ligands was introduced into the model. Many technologies, such as triangular self-attention mechanisms and transfer learning, were also adopted to enhance the model's ability to learn geometric constraints and chemical structures. It is particularly worth mentioning that chemical knowledge is deeply incorporated in molecular generation. In various computational validations, PocketFlow shows the best ability to generate drug-like molecules compared with the baselines. Molecules generated by PocketFlow are much closer to real drug-like molecules (CrossDocked2020) than those generated by the baselines in many aspects, including synthetic accessibility, bond length distribution, bond angle distribution and ring structures. Furthermore, compared with the baselines, PocketFlow can generate molecules with better binding sites (inside the pocket) and higher ligand efficiency. It is also worth noting that we used relatively few parameters (about 210,000) in PocketFow, indicating less demand for computing resources; models with more parameters are expected to have a better performance than those with fewer parameters[42,43], but more computational resources are needed.

Of significance is that the effectiveness of PocketFlow was validated by wet-lab experiments. We applied PocketFlow to HAT1 and YTHDC1, which are thought to be promising targets for the treatment of various diseases, particularly cancer. First PocketFlow was used to generate small molecules inside the active pockets of the two proteins. Then, from the generated molecules, we selected and synthesized only two and three very simple molecules for HAT1 and YTHDC1, respectively; these molecules were chosen because they can be easily and quickly prepared, in addition to having good QED and/or LE values. One active compound for HAT1 and two for YTHDC1 were obtained. It is necessary to mention that the binding sites and binding poses of the obtained active compounds generated by PocketFlow are very similar to those predicted by molecular docking. Experimental X-ray cocrystal

https://doi.org/10.1038/s42256-024-00808-8

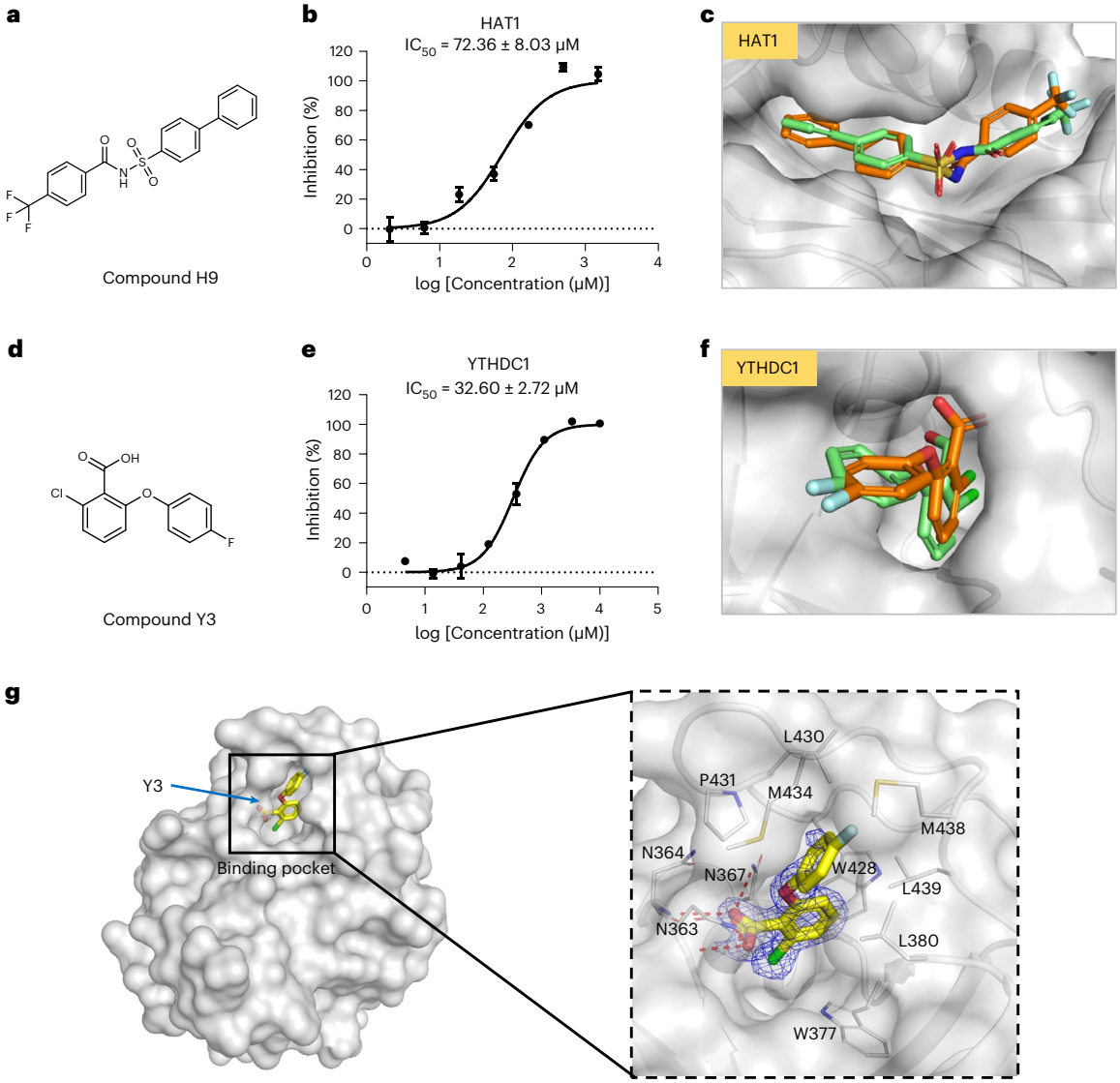


Fig. 4 | Application of PocketFlow leads to the discovery of new small molecule inhibitors of HAT1 and YTHDC1. a, Chemical structure of HAT1 inhibitor H9. **b**, Dose–activity curve of H9. The experiments were performed independently three times. All data are shown as mean ± s.d. **c**, Superimposition of the binding mode of H9 with HAT1 generated by PocketFlow (orange) and that predicted by molecular docking (light green). **d**, Chemical structure of the YTHDC1 inhibitor Y3. **e**, Dose–activity curve of Y3. The experiments were performed independently three times. All data are shown as mean ± s.d. **f**, Superimposition of the binding mode of Y3 with YTHDC1 generated by PocketFlow (orange) and that predicted by molecular docking (light green). **g**, The cocrystal structure of YTHDC1 in complex with Y3 (PDB ID 8k2e). Left, overview of the cocrystal structure. Right, interaction mode between Y3 and YTHDC1. Side chains involved in intermolecular interactions are represented by stick. YTHDC1 and Y3 are shown in grey and yellow, respectively. Red dashes indicate hydrogen bonds. The $2F_o$–F_c density map of Y3 ($\sigma = 0.8$) is shown in blue. F_o is F-observe, the measured structure factor in the experiment, which is the true structure factor; F_c is F-calculate, which is the structural factor calculated based on a given structural model.

structure further confirms that the obtained active compound truly binds to the designated protein pocket.

In summary, PocketFlow, developed here, is a data-and-knowledge dual-driven DGM that shows SOTA performance among all the tested DGMs. However, there is still room for improvement in some aspects. For example, the binding affinities of generated molecules can be further improved, which may be achieved by introducing reinforcement learning[44–47]. Other areas for improvement include considering protein flexibility and taking into account the pharmacokinetic properties and toxicity of generated molecules.

Methods

Details of PocketFlow

PocketFlow is composed of five modules: Context Encoder, Focal Net, Atom Flow, Position Predictor and Bond Flow. Context Encoder is designed to encode the environmental information for the subsequent generation task. Focal Net is for predicting a focal atom that is an origin of the local coordinate system for the generation of atoms, positions and bonds. Atom Flow, Position Predictor and Bond Flow are designed to generate new atoms, positions and bonds, respectively. To keep the equivariance of rotation and translation operations in 3D space, we adopt a new equivariant graph neural network, GDBP (Supplementary Fig. 13a), as the basic component of PocketFlow (an upgraded version of a geometric vector perceptron (GVP) and a geometric bottleneck perceptron (GBP)[48–50]) by adding bottleneck layers for scalar and vector features; the addition of bottleneck layers improves the model speed and enhances information integration[51–53]. Furthermore, autoregressive models usually face the problem of exposure bias[54,55], the accumulation of which can lead to unrealistic structures in the generated molecules, especially when generating drug-scale molecules:

for example, molecular weight > 300. To reduce the impact of exposure bias, we use chemical knowledge to constrain and modify the molecular structures during and after generation.

PocketFlow generates small molecules step by step inside a given protein pocket. Without loss of generality, in the t-th step, PocketFlow sequentially generates atom type $a^{(t)}$, coordinate $r^{(t)}$ and covalent bond type $e^{(t)}$ by using $C^{(t-1)}$ as the contextual environment that contains the binding pocket as well as the molecular fragment generated in the previous $t-1$ steps. Each step can be further divided into five substeps (Fig. 1).

(1) The Context Encoder module extracts the features of the complexes composed of proteins and existing molecular fragments to form a contextual environment for generating the next atom (Fig. 1b). Context Encoder is an equivariant graph attention network with multiheads stacked by multiple interaction blocks that consist of a message module and an attention module (Supplementary Fig. 13b–e).
(2) Based on the contextual environment ($C^{(t-1)}$), the Focal Net module is adopted to predict the focal atom from the current molecular fragment. The focal atom is defined as the origin for establishing the local coordinate system and should have a covalent bond with the new atom that will be generated. To reduce the chance of violating the basic rules of covalent bonds (Fig. 1c), atoms whose valence bonds are saturated are not considered as focal atoms. For the generation of the first atom, we select the focal atom from the pocket wall of the protein and generate a new atom type and coordinates sequentially, without considering covalent bonds.
(3) The Atom Flow module converts the z randomly sampled from the normal distribution into a distribution of the new atom type $P(a^{(t)}|C^{(t-1)})$ according to the hidden features of focal atom $(v^f_{t-1}, \mathbf{v}^f_{t-1})$ (Fig. 1d).
(4) A local coordinate system is established with the focal atom as the origin, and the Position Predictor module calculates the relative position of the new atom $\Delta r^{(t)}$ using the features of the focal atom and the embedding of the new atom type as input, which in turn yields the coordinates $r^{(t)}$. The Position Predictor module is a mixture density network (MDN)[16,56] consisting of GDBP. Thanks to the vector feature-based equivariant graph neural network, the xyz coordinates of the new atoms can be generated directly. To reduce the possibility of unrealistic bond lengths, the distance between the new atom and the focal atom is restricted to within 2 Å (Fig. 1e).
(5) Covalent bonds are generated, into which much chemical knowledge is injected to guide the generation process. To avoid unrealistic bond lengths in generated molecules, we only consider atoms with a distance of less than 4 Å from the focal atom as candidate atoms to form covalent bonds with the newly generated atom (Fig. 1f). Then, the Bond Flow module samples z from the normal distribution and performs the feedforward transformation of the normalizing flow to calculate the distribution of the bond type. Of note is that the distribution of bond type depends on $C^{(t-1)}$, new atom type $a^{(t)}$ and coordinates $r^{(t)}$: that is, $P(e^{(t)}|C^{(t-1)},a^{(t)},r^{(t)})$ (Fig. 1f). Inspired by AlphaFold2[57] and Pocket2Mol[16], we apply a triangular equivariant attention module to the Bond Flow module to capture the geometric constraints. We explicitly apply a chemical valence constraint during sampling to check whether the current covalent bond exceeds the allowed chemical valence of the atom, which is widely used in two-dimensional molecular generation models[58–61]. If a newly generated covalent bond exceeds the maximum chemical valence of the atom, we delete the covalent bond and resample to generate another type of bond. PocketFlow extends this approach to 3D molecular generation.

Furthermore, if alert structures are formed by new atoms, such as O–O, O–N, C=C=C and three-membered rings, we also modify the new bond type or delete it and then resample the new atom or bond type.

After molecular generation, we further checked the rationality of bond types within the generated molecules and made modifications if necessary. For example, if a six-membered ring composed of C or N contains two double bonds, we modify it to an aromatic ring. More details are described in the Supplementary Information.

Data preparation

We randomly selected approximately 8 million molecules from the ZINC database[62] that met the following conditions as pretraining samples: (1) 300 < molecular weight < 600; (2) QED > 0.6; (3) the smallest set of smallest rings ≤5; (4) not containing eight or more membered rings; and (5) containing only C, N, O, F, P, S, Cl, Br and I. The CrossDocked2020 dataset contains approximately 22 million docked protein–ligand complexes, from which we selected samples with root mean square deviation (RMSD) of binding pose less than 1 Å and obtained approximately 180,000 data points (complexes). From the 180,000 complexes, we further chose samples that contained only C, N, O, F, P, S, Cl, Br and I and in which the number of heavy atoms was no more than 35. The final dataset obtained contains approximately 150,000 samples of complexes and was used to fine-tune PocketFlow.

The training input of PocketFlow consists of the pockets and trajectories of ligands. As shown in Supplementary Fig. 14, we mask an atom and its covalent bond at each step (red dashed circles and grey dashed lines in Supplementary Fig. 14), allowing the model to learn how to generate the masked part based on the existing part. In each step, pockets and existing molecular fragments are represented by a K-nearest neighbour graph: that is, all atoms are connected to the nearest K atoms. The atom features selected in this work are shown in Supplementary Table 19. For the edge features, except for distance, we include the topology of protein and ligand: that is, covalent bonding information of the ligand is included, as well as that of the proteins. All molecules (including ligands and receptors) are represented by the Kekule formula (Supplementary Table 19). The Euclidean distances between atoms are expanded into vectors by Gaussian RBF kernels. We further improve the existing K-nearest neighbour-based representation by adding the covalent bonding information of protein and ligand, which are usually closely related to the mode of intermolecular interaction. This representation not only enables the model to learn how to generate new atoms, coordinates and covalent bonds based on pockets and molecular fragments but also enables the model to evaluate the probability density of generating complete molecules from scratch by learning the whole trajectory, thus reducing the ratio of unrealistic substructures appearing. In addition, this masked autoregressive approach is independent between each step, so the trajectory can be computed in parallel, thus improving the training efficiency. The training trajectories of ligands can be computed by a graph search algorithm. In this study, we adopt a mixed strategy where the training trajectories are computed by randomly selecting from breadth-first search and ring-first search during training. Fing-first search is a variant of depth-first search, and it has been shown that using it as a trajectory generation algorithm can improve the performance of molecular generation[7].

Training

We used PyTorch and PyTorch Geometric to build the model. RDKit and PyMOL were used to process the ligand and protein files. The Adam[63] optimizer was used to train the model with an initial learning rate of 2×10^{-4}, a decay rate of 0.6, patience of 10 and a minimum learning rate of 1×10^{-5}. Gradient clipping was also applied to the model training process to ensure training stability. The batch sizes were 256 and 4 for

pretraining and fine-tuning, respectively. The total number of training steps was 1 million, and validation was performed every 5,000 steps. One hundred complex samples were randomly divided as the test set to check whether the loss converges; the remaining dataset was used as the training set. The model was trained using an NVIDIA A100 40 GB GPU. Training for PocketFlow requires optimizing multiple loss functions at the same time. The loss of the focal atom prediction, L_{focal}, is the binary cross entropy loss of the predicted focal atom, and the loss of the Position Predictor, L_{pos}, is the negative log likelihood of the masked atom positions. The losses of the atom type and bond type are also the negative log likelihood of the masked atom and bond, denoted as L_{atom} and L_{bond}, respectively. The overall loss function L_{all} is the summation of the above four loss functions:

$$L_{all} = L_{focal} + L_{atom} + L_{pos} + L_{bond} \quad (4)$$

Molecular docking

Molecular docking was carried out by Glide (Schrödinger). The docking score incorporated in Glide was used for the calculation of ligand efficiency (for screening the generated molecules of YTHDC1). The size of the docking box was defined as (13 Å, 13 Å, 19 Å) and (27 Å, 27 Å, 33 Å) around the ligand for the inner and outer boxes, respectively. Other parameters were set to their default values.

ChemScore and ligand efficiency

ChemScore is computed by Gold Suite v.5.3.0, and ligand efficiency is obtained by dividing the ChemScore values by the numbers of heavy atoms (for evaluating affinity of molecules generated by each model).

Expression and purification of target proteins

Human HAT1 (residues 1–419) was subcloned into the pET-28a vector. The recombinant protein was overexpressed in *Escherichia coli* BL21(DE3) induced with 0.3 mM isopropyl-1-thio-d-galactopyranoside at 16 °C overnight. Cells were harvested and lysed in buffer containing 20 mM Tris-HCl pH 8.0, 300 mM NaCl and 5% glycerol. The lysate was centrifuged at 15,000 rpm for 45 min. Supernatants were loaded on to a Ni-NTA column (GE Healthcare) equilibrated with lysis buffer and then eluted with 300 mM iminazole in the lysis buffer. Eluted protein was further purified by HiTrap Q HP column (GE healthcare) and Superdex 200 gel filtration (GE Healthcare) in a buffer containing 20 mM Tris-HCl pH 8.0, 150 mM NaCl and 3 mM dithiothreitol (DTT).

Human YTHDC1 (residues 345–509) was subcloned into the pGEX-6P-1 vector. The recombinant protein was overexpressed in *Escherichia coli* BL21(DE3) induced with 0.2 mM isopropyl -1-thio-d-galactopyranoside at 16 °C overnight. The cell pellet was dissolved and further lysed in a buffer containing 20 mM Tris-HCl pH 8.0, 500 mM NaCl and 1 mM DTT. Supernatants were collected after centrifugation at 16,000 g for 1 h and loaded on to a Glutathione Sepharose column (GE Healthcare) equilibrated with lysis buffer, washed with lysis buffer and then eluted with 30 mM reduced glutathione in the lysis buffer. Purified protein was treated with human rhinovirus (HRV 3 C) protease to remove the GST-tag. The treated sample was further analysed by gel filtration column (GE Healthcare). Finally, the pure protein was concentrated to 10 mg/ml in a buffer containing 20 mM Tris-HCl pH 8.0, 150 mM NaCl and 1 mM DTT. The expression and purification methods of YTHDC2 and YTHDF1–3 are similar to those of YTHDC1.

Fluorescence assay for HAT1 inhibitors

The fluorescence assay was carried out with a final concentration of 0.04 μM recombinant human HAT1, 30 μM H4-20 peptide (Ac-SGRGKGGKGLGKGGAKRHRK) and 30 μM acetyl-CoA in reaction buffer (20 mM Tris-HCl pH 8.0, 0.01% alkylated bovine serum albumin, 0.1 mM ethylenediaminetetraacetic acid). For the determination of IC_{50} values, the recombinant human HAT1 was pre-incubated with various concentrations of compounds for 10 min, and then H4-20 peptide and acetyl-CoA were added and incubated for 3 h at room temperature. Reactions were quenched by adding ice-cold isopropanol and then incubated with 25 μM 7-diethylamino-3-(4'-maleimidylphenyl)-4-methylcoumarin in darkness for 10 min. The fluorescence signal was measured using a CLARIOstar Plus Plate Reader (BMG Labtech) at an excitation wavelength of 390 nm and an emission wavelength of 469 nm. The data were fitted using a four-parameter logistic equation in GraphPad Prism v.8.0 to determine IC_{50} values.

Fluorescence polarization (FP) assay for YTHDC1 inhibitors

The fluorescein amidites (FAM)-labelled RNA oligo (FAM-AAGA ACCGGm^{6}ACUAAG) was used at the final concentration of 3 nM and recombinant GST-YTHDC1 at 30 nM in FP buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM tris(2-carboxyethyl)phosphine and 0.01% bovine serum albumin). Compounds were added and incubated in the dark for 1 h at 16 °C before adding FAM-labelled RNA oligo. FP was measured using a CLARIOstar Plus Plate Reader (BMG Labtech) at an excitation wavelength of 480 nm and an emission wavelength of 520 nm. The data were fitted using a four-parameter logistic equation in GraphPad Prism v.8.0 to determine IC_{50} values.

The bioactivities of compounds against YTHDC2 and YTHDF1–3 were also measured by the FP competitive binding assay. The test conditions are the same as those for YTHDC1.

Crystallization and structure determination

Crystals of YTHDC1 (residues 345–509) were obtained by mixing 1 μl protein solution at 10 mg ml^{-1} with reservoir solution containing 0.1 M Bis-Tris pH 6.5, 0.2 M ammonium sulfate and 20% polyethylene glycol 3350 at 18 °C in a hanging drop vapour diffusion setup. To obtain crystals of protein complexed with Y3, the crystals were transferred to a 1 μl drop containing 50 mM Y3 directly dissolved in 0.1 M Bis-Tris at pH 6.5, 0.2 M ammonium sulfate and 30% polyethylene glycol 3350, soaked overnight at 18 °C, harvested and frozen in liquid nitrogen.

Diffraction data were collected at beamlines BL18U1 and BL19U1 of the National Facility for Protein Science in Shanghai at Shanghai Synchrotron Radiation Facility and processed with the HKL3000 program suite and XDS packages. The structures were solved by molecular replacement using the Phaser program from the Phenix package. The data collection and structure refinement statistics are summarized in Supplementary Table 18. All figures representing structures were prepared with PyMOL.

Reporting summary

Further information on the research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The pretrain dataset of this study was randomly selected from ZINC database: https://zinc.docking.org. The fine-tuning dataset of this study was extracted from CrosDocked2020: https://bits.csb.pitt.edu/files/crossdock2020/. The pretrain and fine-tuning data of this study are available at Zenodo (https://doi.org/10.5281/zenodo.10142813).

Code availability

Computer codes of PocketFlow are available at https://github.com/Saoge123/PocketFlow (https://doi.org/10.5281/zenodo.10460455)[64].

References

1. Li, Y. et al. Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor. *Nat. Commun.* **13**, 6891 (2022).
2. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Opin. Struct. Biol.* **79**, 102548 (2023).

3. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 114 (2023).
4. Ramesh, A. et al. Hierarchical text-conditional image generation with clip latents. Preprint at https://doi.org/10.48550/arXiv.2204.06125 (2022).
5. Tong, X. et al. Generative models for de novo drug design. *J. Med. Chem.* **64**, 14011–14027 (2021).
6. Wang, J. et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach. Intell.* **3**, 914–922 (2021).
7. Li, Y., Pei, J. & Lai, L. Structure-based de novo drug design using 3D deep generative models. *Chem. Sci.* **12**, 13664–13675 (2021).
8. Zheng, S. et al. Accelerated rational PROTAC design via deep learning and molecular simulations. *Nat. Mach. Intell.* **4**, 739–748 (2022).
9. Zhang, J. & Chen, H. De novo molecule design using molecular generative models constrained by ligand–protein interactions. *J. Chem. Inf. Model.* **62**, 3291–3306 (2022).
10. Godinez, W. J. et al. Design of potent antimalarials with generative chemistry. *Nat. Mach. Intell.* **4**, 180–186 (2022).
11. Bagal, V. et al. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **62**, 2064–2076 (2022).
12. Blaschke, T. et al. REINVENT 2.0: An AI tool for de novo drug design. *J. Chem. Inf. Model.* **60**, 5918–5922 (2020).
13. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
14. Moret, M. et al. Beam search for automated design and scoring of novel ROR ligands with machine intelligence. *Angew. Chem. Int. Ed.* **60**, 19477–19482 (2021).
15. Liu, M. et al. Generating 3d molecules for target protein binding. Preprint at https://doi.org/10.48550/arXiv.2204.09410 (2022).
16. Peng, X., et al. Pocket2mol: efficient molecular sampling based on 3d protein pockets. In *Proceedings of the International Conference on Machine Learning* **162**, 17644–17655 (2022).
17. Ragoza, M., Masuda, T. & Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **13**, 2701–2713 (2022).
18. Pearl, J. Radical empiricism and machine learning research. *J. Causal Inference* **9**, 78–82 (2021).
19. Pan, Y. Heading toward artificial intelligence 2.0. *Engineering* **2**, 409–413 (2016).
20. Cheng, G., Gong, X.-G. & Yin, W.-J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **13**, 1492 (2022).
21. Jiang, Y. et al. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **12**, 5950 (2021).
22. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
23. Bickerton, G. R. et al. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
24. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
25. Polykovskiy, D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 565644 (2020).
26. Francoeur, P. G. et al. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.* **60**, 4200–4215 (2020).
27. Eldridge, M. D. et al. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **11**, 425–445 (1997).
28. Hartshorn, M. J. et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **50**, 726–741 (2007).
29. Hopkins, A. L., Groom, C. R. & Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **9**, 430–431 (2004).
30. Kenny, P. W. The nature of ligand efficiency. *J. Cheminform.* **11**, 8 (2019).
31. Chen, H. et al. in *Comprehensive Medicinal Chemistry III* (eds Chackalamannil, S. et al.) Ch. 2.08 (Elsevier, 2017).
32. Verdonk, M. L. et al. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **54**, 5422–5431 (2011).
33. Wu, H. et al. Structural basis for substrate specificity and catalysis of human histone acetyltransferase 1. *Proc. Natl Acad. Sci. USA* **109**, 8925–8930 (2012).
34. Fan, P. et al. Overexpressed histone acetyltransferase 1 regulates cancer immunity by increasing programmed death-ligand 1 expression in pancreatic cancer. *J. Exp. Clin. Cancer Res.* **38**, 47 (2019).
35. Xue, L. et al. RNAi screening identifies HAT1 as a potential drug target in esophageal squamous cell carcinoma. *Int. J. Clin. Exp. Pathol.* **7**, 3898–3907 (2014).
36. Xia, P. et al. MicroRNA-377 exerts a potent suppressive role in osteosarcoma through the involvement of the histone acetyltransferase 1-mediated Wnt axis. *J. Cell. Physiol.* **234**, 22787–22798 (2019).
37. Kumar, N. et al. Histone acetyltransferase 1 (HAT1) acetylates hypoxia-inducible factor 2 alpha (HIF2A) to execute hypoxia response. *Biochim. Biophys. Acta Gene Regul. Mech.* **194900**, 2023 (**1866**).
38. Lahue, B. R. et al. Diversity & tractability revisited in collaborative small molecule phenotypic screening library design. *Bioorg. Med. Chem.* **28**, 115192 (2020).
39. Roundtree, I. A. et al. YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs. *eLife* **6**, e31311 (2017).
40. Xiao, W. et al. Nuclear m6A reader YTHDC1 regulates mRNA splicing. *Mol. Cell* **61**, 507–519 (2016).
41. Sheng, Y. et al. A critical role of nuclear m6A reader YTHDC1 in leukemogenesis by regulating MCM complex–mediated DNA replication. *Blood* **138**, 2838–2852 (2021).
42. Bubeck, S. & Sellke, M. A universal law of robustness via isoperimetry. *J. ACM* **70**, 1–18 (2023).
43. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. *J. Stat. Mech.: Theory Exp.* **2021**, 124003 (2021).
44. Schulman, J. et al. Proximal policy optimization algorithms. Preprint at https://doi.org/10.48550/arXiv.1707.06347 (2017).
45. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
46. Sutton, R. S. et al. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems* https://proceedings.neurips.cc/paper_files/paper/2018/hash/d60678e8f2ba9c540798ebbde31177e8-Abstract.html (1999).
47. Haarnoja, T. et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning* **80**, 1861–1870 (2018).
48. Jing, B. et al. Learning from protein structure with geometric vector perceptrons. Preprint at https://doi.org/10.48550/arXiv.2009.01411 (2020).
49. Aykent S. and T. Xia. Gbpnet: Universal geometric representation learning on protein structures. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* https://doi.org/10.1145/3534678.3539441 (2022).

50. Deng, C. et al. Vector neurons: a general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* https://openaccess.thecvf.com/content/ICCV2021/html/Deng_Vector_Neurons_A_General_Framework_for_SO3-Equivariant_Networks_ICCV_2021_paper.html (2021).
51. He, K. et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (2016).
52. Gasteiger, J. et al. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. Preprint at https://doi.org/10.48550/arXiv.2011.14115 (2020).
53. Yu, D. & Seltzer, M. L. Improved bottleneck features using pretrained deep neural networks. In *Twelfth Annual Conference of the International Speech Communication Association* https://jackyguo624.github.io/img/2020-02-12-bottle-feature-for-asr/Bottleneck-Interspeech2011-pub.pdf (2011).
54. Ranzato, M. A. et al. Sequence level training with recurrent neural networks. Preprint at https://doi.org/10.48550/arXiv.1511.06732 (2015).
55. Schmidt, F. J. Generalization in generation: a closer look at exposure bias. Preprint at https://doi.org/10.48550/arXiv.1910.00292 (2019).
56. Bishop, C. M. Mixture density networks. Technical Report. https://publications.aston.ac.uk/id/eprint/373/ (Aston University, 1994).
57. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
58. Luo, Y., Yan, K. & Ji, S. Graphdf: a discrete flow model for molecular graph generation. In *Proceedings of the 38th International Conference on Machine Learning* **139**, 7192–7203 (2021).
59. Shi, C. et al. Graphaf: a flow-based autoregressive model for molecular graph generation. Preprint at https://doi.org/10.48550/arXiv.2001.09382 (2020).
60. You, J. et al. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems* https://proceedings.neurips.cc/paper_files/paper/2018/hash/d60678e8f2ba9c540798ebbde31177e8-Abstract.html (2018).
61. Popova, M. et al. MolecularRNN: generating realistic molecular graphs with optimized properties. Preprint at https://doi.org/10.48550/arXiv.1905.13372 (2019).
62. Irwin, J. J. et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
63. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2014).
64. Jiang, Y. et al. PocketFlow is a data-and-knowledge driven structure-based molecular generative model. *Zenodo* https://doi.org/10.5281/zenodo.10460455 (2024).

Acknowledgements

This work was supported by National Key R&D Program of China (grant no. 2023YFF1204905, S.Y.); the National Natural Science Foundation of China (grant nos. T2221004, S.Y.; 81930125, S.Y.; and 82273787, L.L.); the New Cornerstone Science Foundation; Major Project of Guangzhou National Laboratory (grant no. GZNL2024A01005); 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (grant nos. ZYXY21001, S.Y.; ZYGD23006, S.Y.); the Frontiers Medical Center, Tianfu Jincheng Laboratory Foundation (grant no. TFJC2023010009, S.Y.); the Natural Science Foundation of Sichuan Province (grant no. 24NSFSC6411) and Sichuan University Postdoctoral Interdisciplinary Innovation Fund (grant no. JCXK2227). We also thank the staff from beamlines BL18U1 and BL19U1 at Shanghai Synchrotron Radiation Facility of the National Facility for Protein Science (Shanghai, China) for great support.

Author contributions

S.Y. conceived and supervised the research and designed the experiments. S.Y. and Y.J. established and validated the DGM model. Y.J., H.X. and M.D. performed molecular docking. G.Z., J.Y., Z.X. and Y.W. carried out chemical synthesis. H.Z. and R.Y. performed the bioactivity assays. S.Y., Y.J. G.Z., J.Y., H.Z., L.Z., R.Y. and L.L. analysed the data. S.Y. and Y.J. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

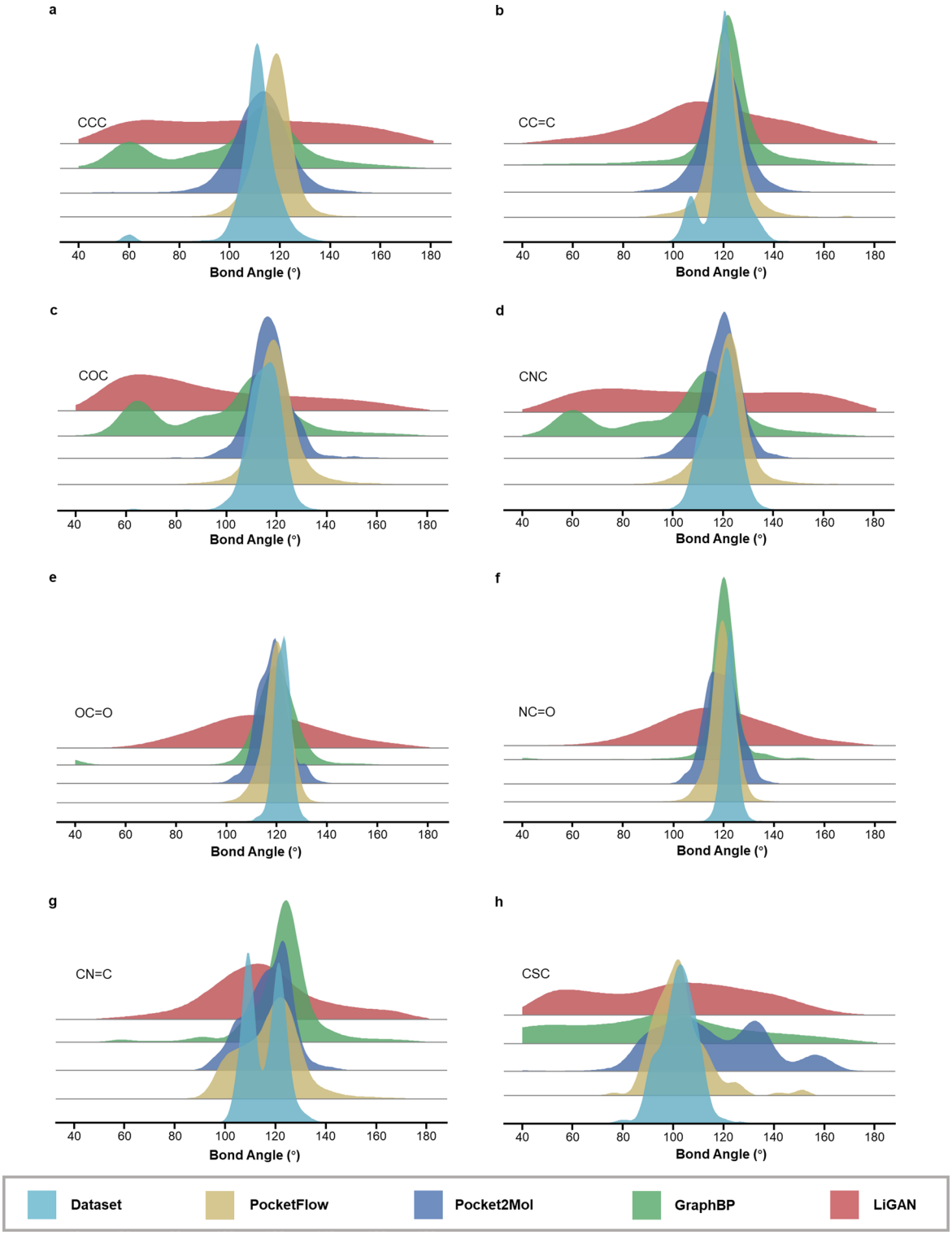
Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00808-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00808-8.

Correspondence and requests for materials should be addressed to Shengyong Yang.

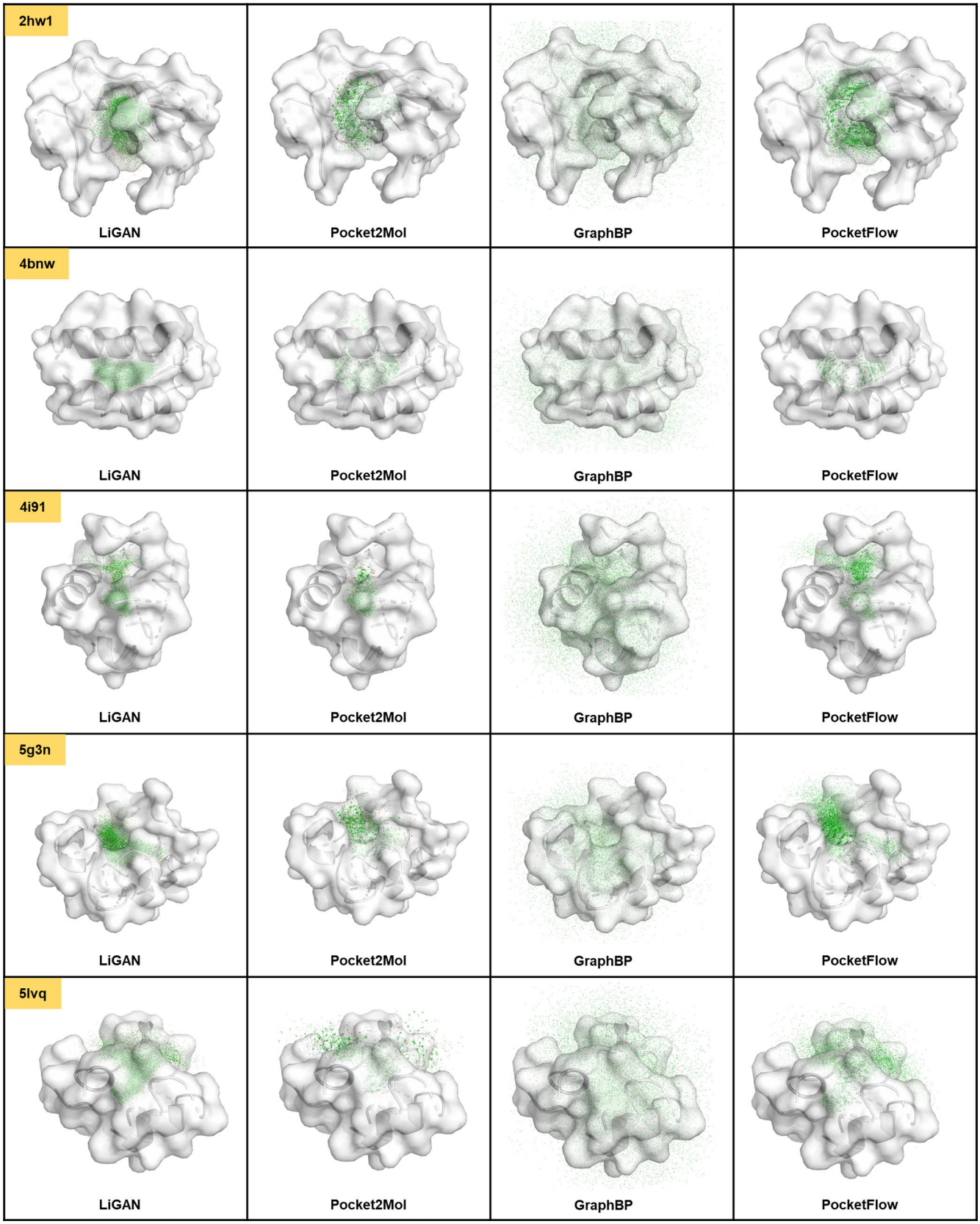
Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-024-00808-8



Extended Data Fig. 1 | Bond angle distributions of molecules generated by different generative models and molecules in the CrossDocked2020 dataset. (**a**) CCC, (**b**) CC = C, (**c**) COC, (**d**) CNC, (**e**) OC = O, (**f**) CN = O, (**g**) CN = C, (**h**) CSC. Results for molecules generated by PocketFlow, Pocket2Mol, GraphBP, and LiGAN, and molecules in the CrossDocked2020 dataset are shown in yellow, blue, green, red, and cyan, respectively.

https://doi.org/10.1038/s42256-024-00808-8



Extended Data Fig. 2 | Distributions of atom positions for 1000 molecules randomly selected from molecules generated by different DGMs. Target proteins and their active pockets are displayed as protein surfaces (white). Green points indicate heavy atoms of molecules.

https://doi.org/10.1038/s42256-024-00808-8

Extended Data Table 1 | KL divergence between the bond length distribution of molecules generated by each DGM and that of molecules in CrossDocked2020

	LiGAN	GraphBP	Pocket2Mol	PocketFlow
C-C	1.098	1.28	0.705	0.787
C=C	0.897	0.701	0.361	0.297
C-O	1.031	0.68	0.276	0.471
C=O	1.176	0.819	0.705	0.703
C-N	0.811	0.742	0.292	0.433
C-Cl	1.33	1.543	1.244	0.76
C-F	1.494	1.251	0.881	0.765
C=N	0.968	0.661	0.462	0.366
C-S	0.949	2.189	1.426	0.546
Mean	1.084±0.206	1.096±0.491	0.706±0.391	0.570±0.178

https://doi.org/10.1038/s42256-024-00808-8

Extended Data Table 2 | KL divergence between the bond angle distribution of different DGM and that of molecules in CrossDocked2020

	LiGAN	GraphBP	Pocket2Mol	PocketFlow
CCC	1.375	0.609	0.21	0.442
CC=C	1.029	0.188	0.13	0.095
OC=O	1.562	0.565	0.523	0.228
NC=O	1.574	0.325	0.542	0.335
CNC	1.529	0.662	0.042	0.094
COC	1.737	0.668	0.086	0.197
CN=C	0.714	0.605	0.196	0.332
CSC	1.113	1.209	0.683	0.104
Mean	1.329±0.324	0.604±0.280	0.301±0.228	0.228±0.122

nature portfolio

Corresponding author(s): Shengyong Yang

Last updated by author(s): Jan 5, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data used in our manuscript was generated using Python (3.8.12), Pytorch (1.13.0), RDkit (2019.09.3) and Pytorch Geometric (2.3.1). These codes can be found in our GitHub repository: https://github.com/Saoge123/PocketFlow.
Data analysis	Our data analysis was based on Python (3.8.12), Pymol-2.3.2, Schrodinger-2023-1, goldsuite-5.3.0, RDkit (2019.09.3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The pretrain dataset of this study was randomly selected from ZINC database: https://zinc.docking.org. The finetuning dataset of this study was extracted from CrossDocked2020: https://bits.csb.pitt.edu/files/crossdock2020. The two datasets of this study are available at Zenodo (10.5281/zenodo.10142813).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For each target, we generated 10,000 molecules by using PocketFlow, baseline models and ablation models. Because Pocket2Mol failed to generated a specified number of molecules for some targets due to unkonwn reason, we had to use the actually generated molecules for statistical analysis. IC50 measurements were carried out with three independent experiments.
Data exclusions	NA
Replication	Experiments of bioactivity were carried out with three independent experiments and these data were used to calculate mean values. All attempts at replication were successful.
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging