PAMol: Pocket-Aware Drug Design Method with Hypergraph Representation of Protein Pocket Structure and Feature Fusion

Xiaoli Lin[*], **Xiongwei Liao**, **Jun Pang**, **Bo Li** and **Xiaolong Zhang**[†]
School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, China
{linxiaoli, liaoxiongwei, pangjun, libo, xiaolong.zhang}@wust.edu.cn

Abstract

Efficient generation of targeted drug molecules is crucial in the field of drug discovery. Most existing methods neglect the high-order information in the structure of protein pockets, limiting the performance of generated drug molecules. This paper proposes a pocket-aware drug design framework, namely PAMol, constructing the hypergraph to represent the spatial structure of protein pockets, effectively capturing high-order relations and neighborhood information within the pocket structures. This framework also fuses different modal embeddings from proteins and molecules, to generate high-quality molecules. In addition, a conditional molecule generation module uses the high-order structural information in protein pockets as constraints to more accurately generate molecules for specific targets. The performance of PAMol has been assessed by analyzing generated molecules in terms of vina score, high affinity, QED, SA, LogP, Lipinski, diversity, and time. Experimental results demonstrate the potential of PAMol for targeted drug design. The source code is available at https://github.com/YICHUANSYQ/PAMol.git.

1 Introduction

Drug design aims to efficiently generate molecules that have both significant potential for clinical application and precise treatment of disease [Wong *et al.*, 2024]. It relies on in-depth analysis of structures and biochemical properties of existing drugs or target proteins. Traditional drug design is a complex process with high costs, long cycles and high risks. It costs about 2.5 billion dollars to design a new drug, and the development process can take up to 10 to 15 years [Bano *et al.*, 2023]. The chemical space for drugs ranges between approximately 10^{23} and 10^{60} molecules [Medina-Franco and López-López, 2024]. There will be more than 10^{15} kinds of diverse and new compounds that can be synthesized [Sadybekov and Katritch, 2023]. In such a large, discrete and disorganised chemistry space, it is a very difficult task to find molecules that interact with disease targets and conform to specific physicochemical properties. The application of deep learning in the field of drug design has received increasing attention [Zhang and Chen, 2022]. Compared with traditional methods, deep learning can learn molecular and protein features from massive data, accelerating the drug discovery process. Currently, drug design methods are usually divided into ligand-based and structure-based methods.

Ligand-based methods are based on the fact that compounds with the same physicochemical properties or structures should have the same activity or similar targets [Fenglei *et al.*, 2021]. [Wang *et al.*, 2021] proposed a generation model that satisfies multiple constraints by combining the knowledge distillation, conditional Transformer and reinforcement learning. [Iwata *et al.*, 2023] combined variation graph autoencoder and Monte Carlo Tree Search to capture structural features of molecules. [Mao *et al.*, 2023] proposed a novel data-driven self-supervised pre-trained model to generate molecules, which extends the SMILES molecule generation space to optimize the generated molecules from a chemical semantic perspective. These methods have limitations in prediction accuracy and reliability due to ignoring the structural information of the target protein.

Structure-based methods are currently dominated by protein pocket-based drug design [Zhang *et al.*, 2024], which relies on known structures of protein pockets. Currently, the structure of protein pockets is mainly represented in the form of graphs. [Peng *et al.*, 2022] used the graph neural network (GNN) to capture the spatial relations of binding pockets, and generated molecules that satisfy geometric and chemical constraints. [Guan *et al.*, 2023] represented protein pockets as sets of atomic points in 3D space, and used GNN to generate target-aware molecules in continuous space. [Zhang *et al.*, 2023] proposed a fragment-based generation framework that encodes contextual information and used GNN to generate molecules. [Zhang and Liu, 2023] captured the interactions between sub-pockets and molecular motifs by learning sub-pocket prototypes, and constructed a global interaction graph to generate molecules. [Lin *et al.*, 2024] represented pocket amino acids and molecular functional groups as fragments to generate molecules. [Qian *et al.*, 2024] introduced a scoring function for binding affinity to generate molecules that bind with high affinity to specific targets. [Huang *et al.*, 2024b] incorporated protein-ligand interaction priors to

[*]Corresponding author
[†]Corresponding author

generate molecules with high affinity. [Huang *et al.*, 2024a] proposed a pocket-based molecular diffusion model that incorporates protein pocket information to generate drug-like molecules. The limitations of structure-based methods are mainly in two aspects: (1) 3D structures of target proteins are often difficult to be obtained directly by experimental methods, which are demanding and time-consuming in terms of computational resources. (2) Even though some proteins have been experimentally resolved, their critical pocket structure information may not be fully annotated or compiled into protein databases, which restricts their applications.

These methods have made great progress in generating molecules, but there are still challenges. Ligand-based methods are limited by the available chemical space, making it difficult to generate molecules with novel structures. Structure-based methods require in-depth knowledge of protein structures, but obtaining 3D structures is expensive, and still has much to be explored. Although the protein structure prediction can be performed using AlphaFold [Jumper *et al.*, 2021], the accuracy of the prediction cannot be fully guaranteed. In addition, most methods neglect the high-order information in the structure of protein pockets, which leads to an incomplete understanding of the properties of protein pockets. This limitation may restrict the model's ability to provide a comprehensive understanding of complex biological systems.

To address these issues, we propose a pocket-aware drug design framework (PAMol) to generate molecules, which constructs the hypergraph of protein pockets to represent the spatial structure. It can capture high-order relations and neighborhood information of protein pockets. This framework also fuses multi-modal embeddings from proteins and molecules, including the structure and sequence of protein pockets, fingerprint features and physicochemical properties of molecules. A multi-level cross fusion module integrates the structure and sequence of protein pockets to obtain fused features, which contain high-order structural information. The fused features serve as constraints for the conditional molecule generation module, helping to improve the quality of generated molecules for specific targets. In addition, the fused features of protein pockets and molecules provide more comprehensive information for the supervised discriminator, enabling it to optimize the quality of the generated molecules. The contribution of this work can be concluded as follows:

- To capture the high-order structural information in protein pockets, we proposed a pocket-aware drug design method with hypergraph representation of protein pocket structure. The proposed method helps to improve the performance of targeted drug design.
- We fused different modal embeddings from structure and sequence of proteins, fingerprint and physicochemical properties of molecules. We also developed a conditional molecule generation module that incorporates an unsupervised discriminator and a supervised discriminator. It uses the fused features of pockets that contain high-order structural information, as constraints to guide and optimize the process of molecule generation.
- We demonstrated the effectiveness of PAMol on CrossDocked dataset. PAMol outperforms related state-of-the-art methods in terms of vina score, QED, Lipinski, diversity, and time, showing the feasibility for targeted drug design.

2 Methods

Figure 1 shows the framework of PAMol model. First, for a given protein pocket, the spacial structure and sequence are represented by HGNN and ProteinBERT, respectively, as shown in Figure 1 (a). Hypergraph structure and sequence features of protein pockets are fused by multi-level cross fusion module, as shown in Figure 1 (b). For a given molecule, fingerprint features and physicochemical properties are represented separately. Figure 1 (c) illustrates the process of obtaining and fusing these molecular features. Finally, the Conditional Molecule Generation module (Figure 1 (d)) uses fused features of pockets as conditions to guide the molecule generation.

2.1 Hypergraph Representation of Protein Pocket Structure

The spatial structure of proteins is the basis of their functions. We construct a hypergraph that contains the structural hyperedges of multiple amino acids. It can reflect the spatial relation between amino acids in protein pocket and thus represent the higher-order structural information of protein pocket more accurately.

The hypergraph of a protein pocket can be defined as $G = (V, E, W)$, where $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of nodes, with each node representing an amino acid in the protein pocket. $E = \{e_1, e_2, \ldots, e_m\}$ denotes the spatial structure hyperedge set. Suppose the coordinates of the central carbon atoms of the i-th and j-th amino acid are (x_i, y_i, z_i) and (x_j, y_j, z_j), respectively. The distance between them can be calculated by:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

where $j \in \{1, 2, \ldots, n\}$ *and* $i \neq j$. If $D_{ij} < 5\mathring{A}$, the j-th amino acid is added to the hyperedge e_i.

In a hypergraph G, each hyperedge $e_i \in E$ is assigned a weight $w(e_i)$ that indicates the importance of its connectivity relations within the entire hypergraph. These weights are organized into a diagonal matrix W, defined as follows:

$$diag(W) = \left[w(e_1), w(e_2), \ldots, w(e_{|E|})\right]$$

where $diag(W)$ denotes the diagonal of matrix. Each diagonal element $w(e_i)$ corresponds to the weight of the i-th hyperedge e_i in the hyperedge set E. This represents the individual importance of each hyperedge in the hypergraph.

To specifically describe the relation between nodes and hyperedges, the hypergraph of a protein pocket can be further represented as an association matrix $H_p \in \{0, 1\}^{|V| \times |E|}$, which is defined as:

$$H_p(v, e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{if } v \notin e \end{cases} \quad (2)$$

where $H_p(v, e) = 1$ indicates that node v is a member of hyperedge e. Conversely, $H_p(v, e) = 0$ indicates that node v does not belong to hyperedge e.

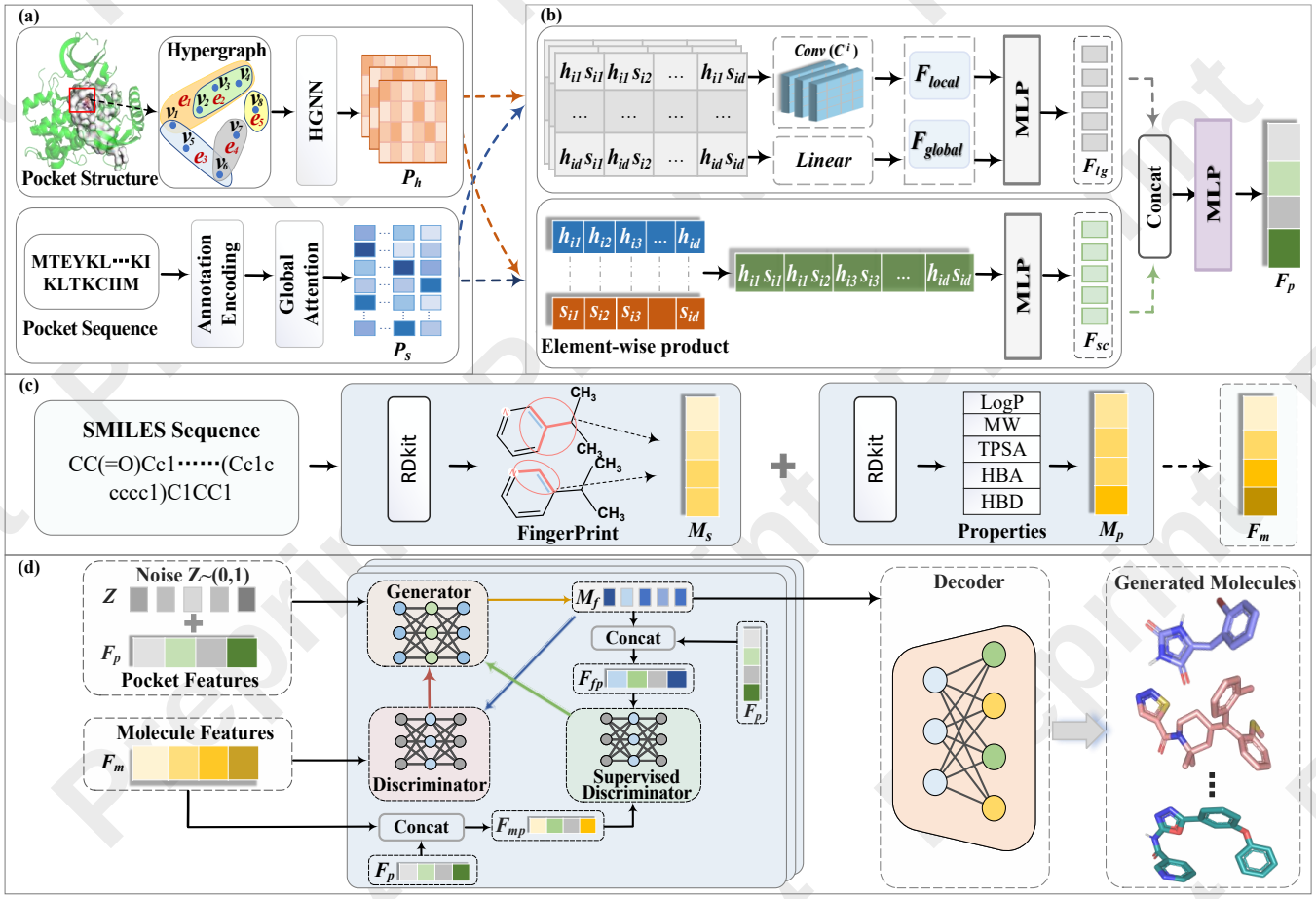


Figure 1: Framework of the proposed PAMol. (a) Representations of the structural and sequence features of protein pockets. PAMol constructs the hypergraph to represent the spacial structure of protein pockets. HGNN is used to capture the high-order relation in pocket structures. ProteinBERT is used to obtain sequence features. (b) Hypergraph structure features and sequence features of protein pockets are fused by Multi-level Cross Fusion (MCF) module. (c) Representation of molecules, including fingerprint features and physicochemical properties of molecules. (d) Conditional Molecule Generation (CMG) module, consisting of a generator, an unsupervised discriminator, a supervised discriminator and a decoder, which uses the fused features F_p of pockets as constraints to guide and optimize the molecule generation.

To learn the high-order structural information in protein pockets, we apply hypergraph neural network [Feng *et al.*, 2019] to encode the protein pocket hypergraph through its unique hypergraph convolution layer. The hypergraph of protein pocket is first mapped to a feature matrix $X_h \in \mathbb{R}^{|V|*|E|}$. In the hypergraph convolution process, the representation of each node is updated based on the features of its connected hyperedges and neighboring nodes. The hypergraph convolution operation is defined as:

$$P_h^{(l+1)} = \sigma(\lambda_v^{-\frac{1}{2}} H_p W \lambda_e^{-1} H_p^T \lambda_v^{-\frac{1}{2}} P_h^{(l)} \Theta^{(l)}) \quad (3)$$

where $\lambda_v \in \mathbb{R}^{|V|*|V|}$ and $\lambda_e \in \mathbb{R}^{|E|*|E|}$ denote the diagonal matrices of the node degree and hyperedge degree, respectively. H_p is the association matrix. W is the weight matrix of hyperedges for protein pockets. $\Theta^{(l)}$ is the trainable parameter matrix at layer l, which is used to capture the complex structural and attribute relations of protein pockets. $P_h^{(l)}$ is the feature matrix of protein pocket nodes at layer l with $P_h^{(0)} = X_h$. This initial feature matrix X_h passes through two convolutional layers and one average pooling layer to obtain a final representation $P_h \in \mathbb{R}^{|V|*d_s}$, where d_s =768.

By multi-layer aggregation and propagating mechanisms, hypergraph neural network is able to learn high-order relations within the protein pocket, capturing both local and global structural information.

2.2 Embedding Representation of Protein Pocket Sequence

The protein pocket sequences can map discrete amino acid sequences into a low-dimensional continuous vector space, generating embedding vectors. These embedding vectors are used as feature representations of pocket sequences and can capture key information in the sequence, such as the relative positions between amino acids, thus improving the performance of the model in drug design tasks. A pre-trained ProteinBERT model [Rao *et al.*, 2019] is used to obtain the sequence features. First, the protein pocket sequences are encoded by IUPAC Tokenizer to get the token sequences.

The resulting token sequences are passed through a 12-layer Transformer with a hidden layer size of 512 units and 8 attention heads, to obtain the final representation $P_s \in \mathbb{R}^{n_p * d_s}$, where n_p is the number of amino acids in the protein pocket and d_s is the embedding dimension of 768.

2.3 Multi-Level Cross Fusion

In this work, we fuse features of different modalities and different scales, improving the model's capability to represent features. Figure 1 (b) shows the process of feature fusion across the structure and sequence of protein pockets based on Multi-level Cross Fusion (MCF).

Before cross fusion operation, a fully connected layer maps the feature vectors into a unified embedding space. Specifically, P_h is the feature of protein pocket hypergraph structures, and P_s is the feature of protein pocket sequences. Then, the feature vectors are represented as h and s, corresponding to P_h and P_s, respectively.

$$h = wP_h + b \tag{4}$$

$$s = wP_s + b \tag{5}$$

where w is trainable weight, and b is bias, respectively.

Multi-Scale Feature Fusion. This block first constructs a cross matrix by the cross-product operation [Chen *et al.*, 2021]. Let the structure representation and sequence representation of the protein pocket after transformation be denoted as vectors $h_i = [h_{i1}, h_{i2}, \ldots, h_{id}]$ and $s_i = [s_{i1}, s_{i2}, \ldots, s_{id}]$, respectively, where d=768. h_i and s_i denote the i-th row in vectors h and s. The cross matrix $C_i \in \mathbb{R}^{d \times d}$ represents the interaction between h_i and s_i, which is defined as:

$$C_i = CrossProduct(h_i, s_i) \tag{6}$$

Then, it extracts local and global features from the cross matrix at different scales for comprehensive understanding. The CNN model incorporates a pooling layer to capture localized interactive patterns, denoted as feature F_{local}:

$$F_{local} = ReLU(Pooling(Conv(C_i)))$$

The flatten operation on the cross matrix C_i allows learning global features.

$$F_{global} = Linear(flatten(C_i))$$

F_{local} and F_{global} are passed through a MLP to obtain the final multi-scale fused feature representation F_{lg} of protein pocket.

$$F_{lg} = MLP(Concat(F_{local}, F_{global})) \tag{7}$$

Scalar-Based Multi-Feature Fusion. First, the feature interaction between h_i and s_i obtained from structure and sequence representations of the protein pocket is encoded by an element-wise product operation. Then, the element-wise vector is passed through MLP to obtain the scalar fusion feature F_{sc}, which is defined as:

$$F_{sc} = MLP(h_i \odot s_i) \tag{8}$$

F_{lg} and F_{sc} are concatenated to obtain the final representation F_p with the embedding dimension of 512, which fuses the hypergraph structure features and sequence features of the protein pocket.

$$F_p = MLP(Concat(F_{lg}, F_{sc})) \tag{9}$$

2.4 Embedding Representation of Molecule

This module can extract fingerprint features and physicochemical properties of molecules [Kotsias *et al.*, 2020], which is useful for optimizing the performance of generated molecules. The embedding vectors of structural features are obtained by Morgan fingerprint, which are generated by considering the topology structure of molecules. In the generation process, each atom and its neighboring atoms are iteratively considered until a predetermined radius is reached. This iterative process captures the connection patterns and distances between atoms, thereby reflecting the structure of molecules. Based on this structural data, we obtain a one-dimensional vector, denoted as M_s. The physicochemical properties obtained by RDKit [Landrum and others, 2013], including Octanol-Water Partition coefficient (LogP), Topological Polarity Surface Area (TPSA), Molecular Weight (MW), Number of Hydrogen Bond Acceptors (HBA) and Number of Hydrogen Bond Donors (HBD). The embedding representation of physicochemical properties is denoted as a one-dimensional vector M_p. The fingerprint features M_s and physicochemical features M_p are concatenated to obtain F_m with the embedding dimension of 512.

2.5 Conditional Molecule Generation (CMG) Module

The Conditional Molecule Generation (CMG) module uses the fusion features F_p of protein pockets that contain high-order structural information as generative conditions, to facilitate the discovery of potential drug molecules against specific targets. CMG module includes two discriminators, one generator, and one decoder, as shown in Figure 1 (d).

First, the 512-dimensional noise vector z sampled from a normal distribution [Chen *et al.*, 2023] is concatenated with the fused features F_p of the protein pocket, serving as the input to the generator. It can enhance the expressive power of the generator, enabling it to produce molecules that are both diverse and conform to specific biological properties. Through the processing of a 3-layer neural network, the latent feature vectors M_f of the molecules are obtained.

$$M_f = network([z, F_p]) \tag{10}$$

In the process of molecule generation, to more accurately optimize the fit between the generated molecules and the protein pockets, the real molecule features F_m and the protein pocket features F_p are concatenated as fused feature F_{mp}:

$$F_{mp} = Concat(F_m, F_p) \tag{11}$$

We also fuse the latent feature vector M_f of generated molecules and protein pocket features F_p, obtaining the fusion feature F_{fp}:

$$F_{fp} = Concat(M_f, F_p) \tag{12}$$

Then, an unsupervised discriminator captures global features of real molecules to refine the overall quality of the generated ones. The features F_m of real molecules and the latent feature vectors M_f of generated molecules are passed through an unsupervised discriminator, to compute the unsupervised loss. This loss measures the difference or similarity between the features of generated molecules and real

molecules. The quality of generated molecules is optimized in the back propagation, which motivates the generator to produce molecules that are closer to the real data distribution.

Meanwhile, a supervised discriminator, by focusing on protein pocket features including the high-order structural information, learns and optimizes relevant features of generated molecules to ensure compatibility with those pockets. The fusion features F_{mp} and F_{fp} are served as the inputs to the supervised discriminator. During the back propagation process, the discriminator learns the relevant features between the generated molecules and protein pockets. By calculating the supervised loss, it performs back propagation and updates its internal neural network parameters based on the loss gradients. This optimization process aims to improve the quality of the generated molecules for targeting the protein pockets.

An integrated loss function combines the discriminatory capabilities of the two discriminators, enabling the generative model to integrate the information of molecules and protein pockets during the training process. The loss is defined as:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{real}\left[\frac{D(F_m) + SD(F_{mp})}{2}\right] \\ &+ \mathbb{E}_{fake}\left[\frac{D(M_f) + SD(F_{fp})}{2}\right] + \lambda_{gp} \cdot gp\end{aligned} \tag{13}$$

where $\mathbb{E}_{real}$ represents the expectation over the real data distribution. $\mathbb{E}_{fake}$ represents the expectation over the generated data distribution. gp is a gradient penalty. λ_{gp} represents the weight coefficient of the gradient penalty. This loss function effectively balances the learning process of two discriminators, thus improving the overall performance of the model.

3 Experiments and Results

3.1 Dataset and Preprocessing

We used the same dataset CrossDocked with [Luo *et al.*, 2021]. This dataset removes the protein-ligand pairs with a binding pose RMSD of less than 1Å, leading to a total of 183,468 pairs. To avoid overlap between training and test sets, [Luo *et al.*, 2021] first clustered the data based on protein sequence similarities. Then, 100,000 protein-ligand pairs were randomly selected from the clustered data for training. For the test set, 100 proteins were randomly selected from the remaining clusters, ensuring no overlap with the training set. In this work, to construct the structural hypergraph of protein pockets, we parsed the PDB files containing information about protein pockets. The files that could not be parsed were removed from both the training and test sets. Finally, we obtained 53,268 protein-ligand pairs for training and 47 protein-ligand pairs for testing.

3.2 Implementation Details

PAMol has been performed based on Python 3.8, Tensorflow and Keras. The hardware setup consisted of an NVIDIA GeForce RTX 3090 with CUDA and cuDNN. We set the training process to run for 2000 epochs, with a batch size of 64, and a low learning rate of 0.0001 to ensure smooth and stable convergence. Adam optimizer was used to optimize the training process of PAMol, ensuring an efficient adjustment of learning rates and a well-behaved convergence.

3.3 Evaluation Metrics

We evaluated the performance of PAMol, with common metrics [Luo *et al.*, 2021; Polykovskiy *et al.*, 2020] including: (1) Vina Score, it estimates the binding affinity between the ligand and target protein, which is a crucial measure to evaluate how well the generated molecule fits into the target protein pocket. (2) High Affinity, it represents the percentage of molecules whose Vina Score is higher than that of the ground truth molecule in the test set. (3) QED, it evaluates the drug-likeness of a molecule by combining multiple desirable molecular properties. (4) SA (Synthetic Accessibility), it measures the synthetic difficulty of the molecule. (5) LogP, it indicates the octanol-water partition coefficient, which should be between -0.4 and 5.6 [Ghose *et al.*, 1999] for a good drug candidate. (6) Lipinski, it measures how well the molecule complies with Lipinski's five rules. (7) Diversity, it quantifies the average pairwise Tanimoto dissimilarity of the generated molecules for each target pocket. (8) Time, it represents the average time required to generate 100 samples for each pocket across all targets.

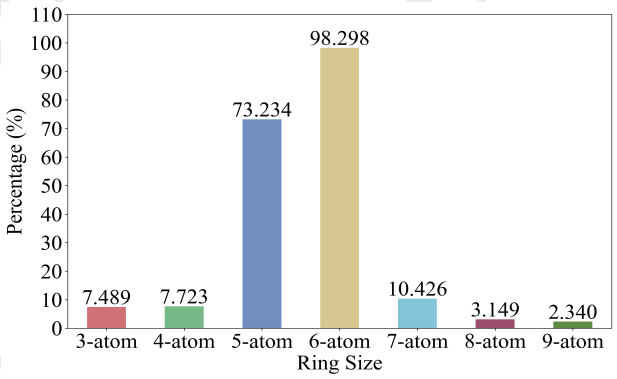


Figure 2: The proportion of different ring sizes among molecules generated by PAMol.

3.4 Ablation Study

The ablation studies were performed with different feature combinations, including the molecule features (MolF), the pocket sequence features (SeqF), the pocket structure features (StruF), and the pocket fused features (CrossF), to investigate their impacts. As shown in Table 1, MolF+CrossF achieves the best vina score of -7.646 among all models, indicating that it can generate molecules with higher binding affinity. MolF+CrossF performs second-best on the high affinity (0.840), with only a 0.001 gap from MolF+SeqF. In addition, MolF+CrossF obtains the highest score on both QED (0.778) and diversity (0.823), and has a logP value of 3.149 within the acceptable range, which indicates that it can improve the drug-likeness and diversity of generated molecules. SA of MolF+CrossF is higher than that of MolF+SeqF, but lower than that of two other combinations. MolF+CrossF and MolF+SeqF both score 5.000, complying with Lipinski's Rule of Five. Despite MolF+CrossF slightly longer runtime compared to the second best model, it is a promising model due to its superior performance on several key metrics.

Models	Vina Score(↓)	High Affinity(↑)	QED(↑)	SA(↑)	LogP	Lipinski(↑)	Diversity(↑)	Time(↓)
MolF+SeqF	-7.628	**0.841**	0.777	0.654	2.610	**5.000**	0.778	368.59
MolF+StruF	-7.556	0.817	0.744	**0.670**	2.022	4.993	0.756	334.21
MolF+SeqF+StruF	-7.547	0.823	0.634	0.666	2.769	4.935	0.766	**142.94**
MolF+CrossF (PAMol)	**-7.646**	0.840	**0.778**	0.659	3.149	**5.000**	**0.823**	341.08

Table 1: Ablation experiment results of different features. (**Best**, Second Best)

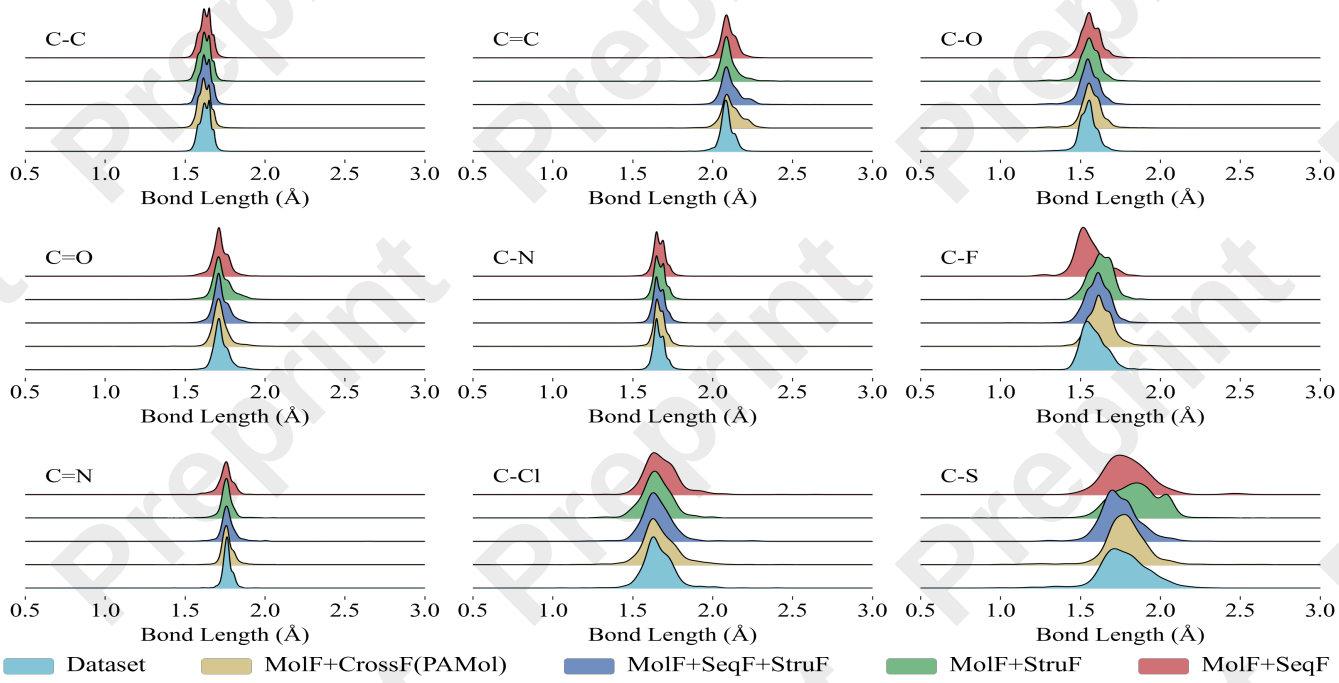
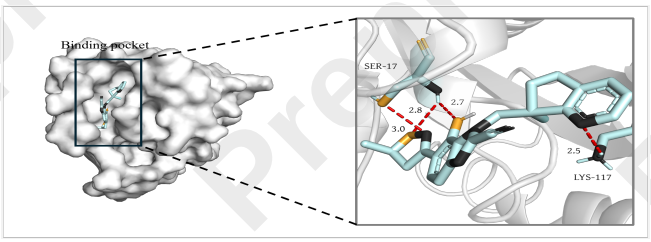


Figure 3: The distribution of the nine common covalent bonds in the dataset and the generated molecules, including C–C, C=C, C–O, C=O, C–N, C–F, C=N, C–Cl and C–S.

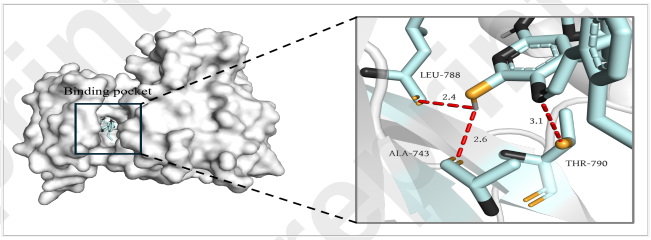
To evaluate the global view of the chemical structures of generated molecules, Figure 2 shows the proportion of different ring sizes among molecules generated by PAMol. We notice that PAMol tends to generate molecules containing relatively more stable rings (5-atom ring and 6-atom ring), and few unstable rings. This is consistent with the regular principles of drug design, indicating that PAMol has high effectiveness in generating molecules with drug potential. In addition, we present the distribution of the nine common covalent bonds in the dataset and the generated molecules, including C–C, C=C, C–O, C=O, C–N, C–F, C=N, C–Cl and C–S. As shown in Figure 3, for all nine covalent bonds, the bond distributions of molecules generated by MolF+CrossF (PAMol) are closer to those of dataset (CrossDocked2020).

3.5 Performance Comparison and Analysis

Table 2 shows the comparison results of different models. PAMol outperforms other models on general metrics except vina score and SA. PAMol can generate the molecules with high affinity (0.840), which has improved by 6.06% compared to the second-best method. PAMol improves by 26.92% over the second-best method in QED, indicating significantly enhanced drug-likeness of generated molecules.



(a) Mol_K1 with target KRAS (Vina Score: -10.8 kcal/mol)



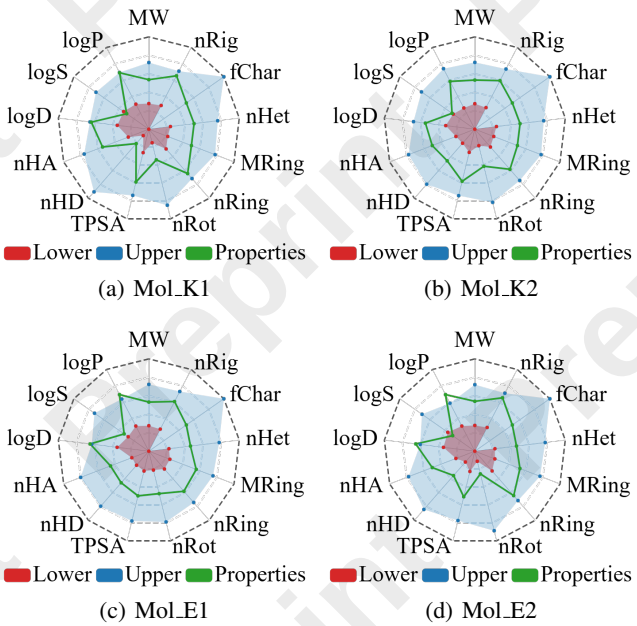
(b) Mol_E1 with target EGFR (Vina Score: -10.9 kcal/mol)

Figure 4: Docking results of the generated molecules with targets KRAS and EGFR.

Models	Vina Score(↓)	High Affinity(↑)	QED(↑)	SA(↑)	LogP	Lipinski(↑)	Diversity(↑)	Time(↓)
Ref. (Test)	-6.277	-	0.372	0.654	0.777	3.723	-	-
Pocket2Mol [Peng *et al.*, 2022]	-7.288	0.542	0.563	<u>0.765</u>	1.586	4.902	0.688	2503.51
Targetdiff [Guan *et al.*, 2023]	-7.800	0.581	0.480	0.580	-	-	0.720	-
FLAG [Zhang *et al.*, 2023]	-7.247	0.580	0.495	0.745	0.630	4.943	0.704	1047.60
DrugGPS [Zhang and Liu, 2023]	-7.276	0.565	<u>0.613</u>	0.743	0.913	4.917	0.681	1007.8
D3FG [Lin *et al.*, 2024]	-6.960	0.459	0.501	**0.840**	2.821	4.965	-	-
KGDiff [Qian *et al.*, 2024]	**-9.430**	<u>0.792</u>	0.510	0.540	-	-	-	-
IPDiff [Huang *et al.*, 2024b]	<u>-8.570</u>	0.695	0.520	0.610	-	-	<u>0.740</u>	-
PMDM [Huang *et al.*, 2024a]	-7.572	0.628	0.594	0.611	0.301	<u>4.975</u>	0.709	<u>906</u>
PAMol(Ours)	-7.646	**0.840**	**0.778**	0.659	3.149	**5.000**	**0.823**	**341.08**

Table 2: Performance comparison of PAMol and other different models. (**Best**, <u>Second Best</u>)

The logP value (3.149) of PAMol within the acceptable range (-0.4 to 5.6) indicates that the generated molecules are more potential as drug candidates. PAMol scores 5.000 on the Lipinski criteria, indicating that generated molecules meet all the conditions of the Rule of Five. PAMol improves diversity by at least 11.22%, showing its ability to generate diverse and novel molecular structures. As in Table 2, vina score and SA of PAMol are not optimal among all models but higher than those in the test set, indicated that PAMol has certain potential. In addition, PAMol achieves a significant reduction in time, thereby enhancing efficiency and saving time costs in molecule generation, and showing competitiveness.



(a) Mol_K1 (b) Mol_K2

(c) Mol_E1 (d) Mol_E2

Figure 5: Radar charts of the basic properties about targeted molecules generated by PAMol.

3.6 Case Study

To further validate the capability of PAMol model to generate molecules, the case experiments were conducted. We selected two key targets in pancreatic cancer, KRAS (UniProt ID: P01116, PDB ID: 8onv) and EGFR (UniProt ID: P00533, PDB ID: 8a27). Figure 4 shows the docking results of the generated molecule Mol_K1 with target KRAS and the generated molecule Mol_E1 with target EGFR. Light cyan indicates carbon, sulfur, hydrogen, and fluorine atoms. Black indicates nitrogen atoms, and yellow indicates oxygen atoms. It can be seen that the generated molecules bind to the target at distances between 2.4 $\mathring{A}$ and 3.1 $\mathring{A}$, with the promising vina scores. It indicates that there are the strong interactions between the molecules and the targets. In addition, it is clear that the shapes of the generated molecules are ideally suited to the shapes of the active pockets.

To evaluate physicochemical properties of targeted molecules generated by PAMol, we used ADMET [Fu *et al.*, 2024] to obtain radar charts of some basic properties, including MW, nRig, fChar, nHet, MRing, nRing, nRot, TPSA, nHD, nHA, logD, logS and logP, as shown in Figure 5. The green line represents the physicochemical property scores of the generated molecule, the blue outline indicates the upper limit, and the red outline indicates the lower limit. It can be seen that four targeted molecules (Mol_K1, Mol_K2, Mol_E1 and Mol_E2) meet most of the physicochemical property standards.

4 Conclusion

This paper proposes a pocket-aware drug design framework, namely PAMol, which captures the high-order structural information of protein pockets to generate molecules for specific targets. We constructed the hypergraph to represent the intricate spatial structure of protein pockets, aiming to capture high-order relations among residues and detailed neighborhood information that reflects the local environment within the pocket. We also fused the cross-modal embeddings from protein pockets and molecules to guide and optimize the process of molecule generation. In addition, we designed a Conditional Molecule Generation (CMG) module that focuses on the features of protein pockets including the high-order structural information. It learns the latent features and distribution patterns of molecules through an unsupervised discriminator, and uses a supervised discriminator to optimize relevant features of generated molecules to ensure compatibility with specific pockets. Experiments show that PAMol can efficiently generate molecules for specific targets. In the future, we will consider incorporating the molecule hypergraph and further optimize the quality of generated molecules.

Acknowledgments

The authors thank the members of Machine Learning and Artificial Intelligence Laboratory, School of Computer Science and Technology, Wuhan University of Science and Technology, for their helpful discussion within seminars. This work was supported by Hubei Province Natural Science Foundation of China (No.2024AFB865) and National Natural Science Foundation of China (No.61972299, 62372342).

References

[Bano *et al.*, 2023] Iqra Bano, Usman Dawood Butt, and Syed Agha Hassnain Mohsan. New challenges in drug discovery. In *Novel Platforms for Drug Delivery Applications*, pages 619–643. Elsevier, 2023.

[Chen *et al.*, 2021] Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. Muffin: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*, 37(17):2651–2658, 2021.

[Chen *et al.*, 2023] Yangyang Chen, Zixu Wang, Lei Wang, Jianmin Wang, Pengyong Li, Dongsheng Cao, Xiangxiang Zeng, Xiucai Ye, and Tetsuya Sakurai. Deep generative model for drug design from protein target sequence. *Journal of Cheminformatics*, 15(1):38, 2023.

[Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3558–3565, 2019.

[Fenglei *et al.*, 2021] LI Fenglei, HU Qiaoyu, XIONG Ruofan, and BAI Fang. Computational drug design methods by deep learning algorithms. *Chinese Journal of Nature*, 43(5):383–390, 2021.

[Fu *et al.*, 2024] Li Fu, Shaohua Shi, Jiacai Yi, Ningning Wang, Yuanhang He, Zhenxing Wu, Jinfu Peng, Youchao Deng, Wenxuan Wang, Chengkun Wu, et al. Admetlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. *Nucleic Acids Research*, page gkae236, 2024.

[Ghose *et al.*, 1999] Arup K Ghose, Vellarkad N Viswanadhan, and John J Wendoloski. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *Journal of combinatorial chemistry*, 1(1):55–68, 1999.

[Guan *et al.*, 2023] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543*, 2023.

[Huang *et al.*, 2024a] Lei Huang, Tingyang Xu, Yang Yu, Peilin Zhao, Xingjian Chen, Jing Han, Zhi Xie, Hailong Li, Wenge Zhong, Ka-Chun Wong, et al. A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications*, 15(1):2657, 2024.

[Huang *et al.*, 2024b] Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, CUI Bin, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Iwata *et al.*, 2023] Hiroaki Iwata, Taichi Nakai, Takuto Koyama, Shigeyuki Matsumoto, Ryosuke Kojima, and Yasushi Okuno. Vgae-mcts: A new molecular generative model combining the variational graph auto-encoder and monte carlo tree search. *Journal of Chemical Information and Modeling*, 63(23):7392–7400, 2023.

[Jumper *et al.*, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[Kotsias *et al.*, 2020] Panagiotis-Christos Kotsias, Josep Arús-Pous, Hongming Chen, Ola Engkvist, Christian Tyrchan, and Esben Jannik Bjerrum. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, 2(5):254–265, 2020.

[Landrum and others, 2013] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

[Lin *et al.*, 2024] Haitao Lin, Yufei Huang, Odin Zhang, Yunfan Liu, Lirong Wu, Siyuan Li, Zhiyuan Chen, and Stan Z Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[Mao *et al.*, 2023] Jiashun Mao, Jianmin Wang, Amir Zeb, Kwang-Hwi Cho, Haiyan Jin, Jongwan Kim, Onju Lee, Yunyun Wang, and Kyoung Tai No. Transformer-based molecular generative model for antiviral drug design. *Journal of chemical information and modeling*, 64(7):2733–2745, 2023.

[Medina-Franco and López-López, 2024] José L Medina-Franco and Edgar López-López. What is the plausibility that all drugs will be designed by computers by the end of the decade?, 2024.

[Peng *et al.*, 2022] Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pages 17644–17655. PMLR, 2022.

[Polykovskiy *et al.*, 2020] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov,

Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

[Qian *et al.*, 2024] Hao Qian, Wenjing Huang, Shikui Tu, and Lei Xu. Kgdiff: towards explainable target-aware molecule generation with knowledge guidance. *Briefings in Bioinformatics*, 25(1):bbad435, 2024.

[Rao *et al.*, 2019] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

[Sadybekov and Katritch, 2023] Anastasiia V Sadybekov and Vsevolod Katritch. Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685, 2023.

[Wang *et al.*, 2021] Jike Wang, Chang-Yu Hsieh, Mingyang Wang, Xiaorui Wang, Zhenxing Wu, Dejun Jiang, Benben Liao, Xujun Zhang, Bo Yang, Qiaojun He, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence*, 3(10):914–922, 2021.

[Wong *et al.*, 2024] Felix Wong, Erica J Zheng, Jacqueline A Valeri, Nina M Donghia, Melis N Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, 2024.

[Zhang and Chen, 2022] Jie Zhang and Hongming Chen. De novo molecule design using molecular generative models constrained by ligand–protein interactions. *Journal of chemical information and modeling*, 62(14):3291–3306, 2022.

[Zhang and Liu, 2023] Zaixi Zhang and Qi Liu. Learning subpocket prototypes for generalizable structure-based drug design. In *International Conference on Machine Learning*, pages 41382–41398. PMLR, 2023.

[Zhang *et al.*, 2023] Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations*, 2023.

[Zhang *et al.*, 2024] Zaixi Zhang, Wan Xiang Shen, Qi Liu, and Marinka Zitnik. Efficient generation of protein pockets with pocketgen. *Nature Machine Intelligence*, pages 1–14, 2024.