中国人民大学
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Generalized Protein Pocket Generation with Prior-Informed Flow Matching

Zaixi Zhang, Marinka Zitnik, Qi Liu

(NeurIPS 2024 Spotlight)

Fanmeng Wang

2024-10-15

Outline

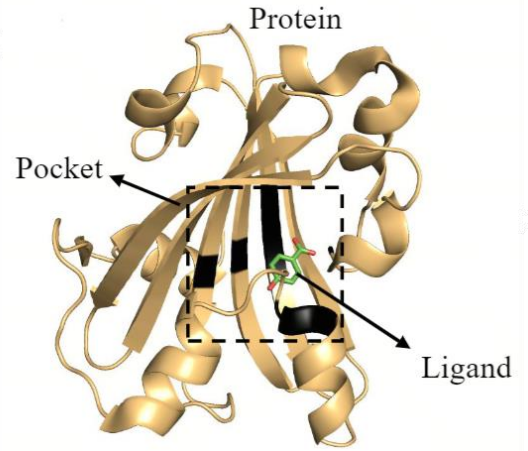
- Introduction
- Method
- Experiment
- Conclusion

- Introduction
- Method
- Experiment
- Conclusion

Introduction

■ Protein Pocket Designing

- **Proteins** are the fundamental building blocks of living organisms, and they often interact with **ligands** (e.g., small molecules, nucleic acids, and peptides) to execute their functions.
- Therefore, one critical step in functional protein design is to **design protein pockets**, which refers to the protein interface binding with the ligand.
- However, the complexity of ligand-protein interactions, the variability of protein sidechains, and sequence-structure relationships pose great challenges for pocket design.



Introduction

■ Existing works

- **Traditional methods** for protein pocket design mainly focus on physics modeling or template-matching. However, the involved physical energy calculation or substructure enumeration could be quite time-consuming.
- Recent advancements in pocket design have benefited a lot from **deep learning-based methods**.
  - On the one hand, these methods often overlook essential domain knowledge, such as the protein-ligand interactions and the geometric constraints governing them.
  - On the other hand, these methods are restricted to small molecule ligands, omitting other important ligand types such as nucleic acids and peptides.

Introduction

■ PocketFlow

➢ In this work, we propose **PocketFlow**, a protein-ligand interaction prior-informed flow matching model for protein pocket generation.

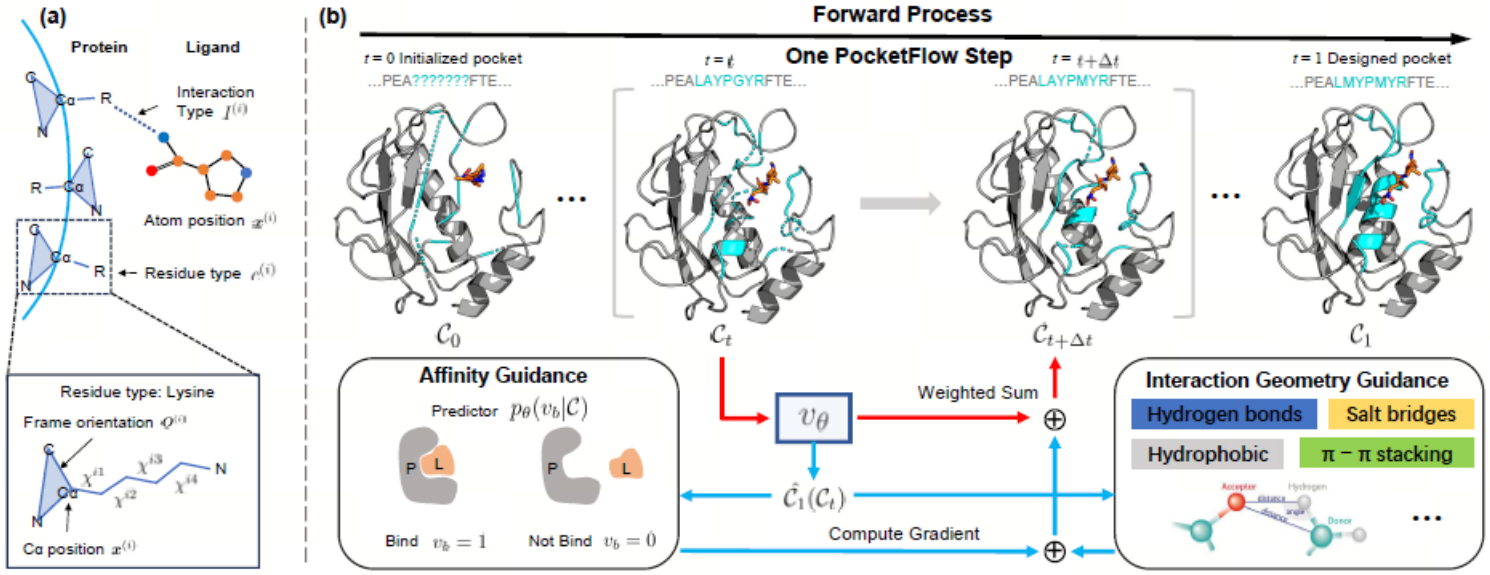


Figure 1: (a) Parameterization of protein-ligand complex. (b) Illustration of PocketFlow forward process. The affinity and interaction geometry guidance are proposed to improve affinity and structural validity. The red/blue lines denote the unconditional/guidance paths respectively.

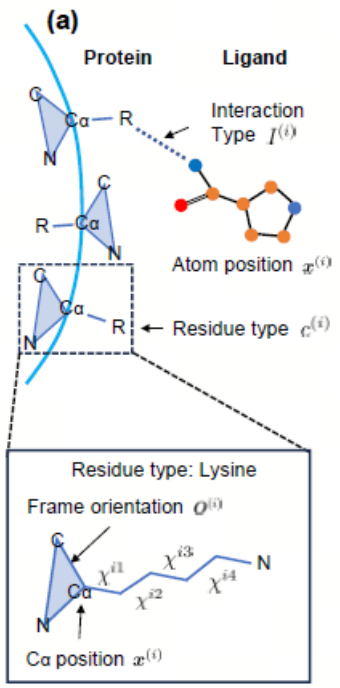
Generalized tasks

Strong performance

- Introduction
- **Method**
- Experiment
- Conclusion

Method

■ Notations

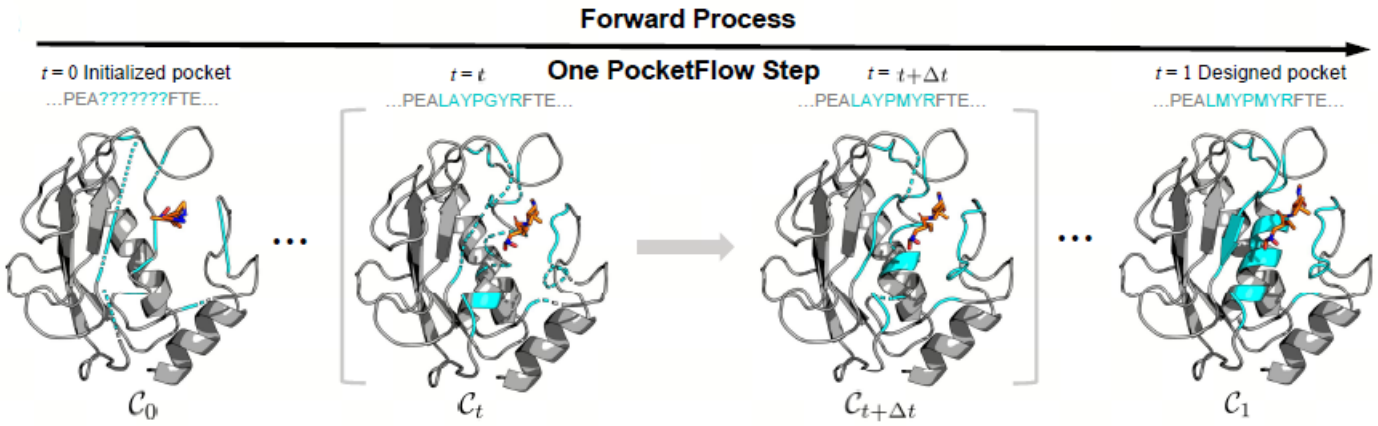


Notations. As shown in Figure 1(a), we model protein-ligand complex as $\mathcal{C} = \{\mathcal{P}, \mathcal{G}\}$ consisting of protein $\mathcal{P}$ and ligand $\mathcal{G}$ (small molecule as an example). Protein $\mathcal{P}$ is composed of a sequence of residues (amino acids) with residue types denoted $\boldsymbol{c}^{(i)} \in \mathbb{R}^{20}$. Consistent with [92, 87], the protein pocket $\mathcal{R} \subset \mathcal{P}$ is defined as the subset of residues closest to the ligand atoms under a threshold δ (e.g., 3.5 Å). In a residue, the backbone structure (consisting of C_α, N, C, O) is parameterized with C_α position $\boldsymbol{x}^{(i)} \in \mathbb{R}^3$ and a frame orientation matrix $\boldsymbol{O}^{(i)} \in SO(3)$ following [43, 84]. The sidechain is parameterized with maximal 4 torsion angles $\boldsymbol{\chi}^{(i)} = \{\chi^{i1}, \chi^{i2}, \chi^{i3}, \chi^{i4}\} \in [0, 2\pi)^4$. Given these key parameters, the full atom protein structure can be derived with the ideal frame coordinates and the sidechain bond length/angles [43]. The protein-ligand interaction type for each residue is marked as $I^{(i)} \in \mathbb{R}^5$ (Hydrogen bond, Salt bridge, Hydrophobic, π-π stacking, no interaction). A pocket with N_r residues can be compactly represented as $\mathcal{R} = \{\boldsymbol{c}^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{O}^{(i)}, \boldsymbol{\chi}^{(i)}, I^{(i)}\}_{i=1}^{N_r}$. As for the ligand, we use a generalized atom-level representation that accommodates various modalities including small molecules, peptides, and RNA. The atom types and bonding information between atoms are given and PocketFlow predicts the N_l ligand atom coordinates (also denoted as $\boldsymbol{x}^{(i)}$ for conciseness).

Method

■ Problem Definition

- PocketFlow co-designs **residue types, 3D structures of the protein pocket, 3D structures of the ligand** conditioned on the ligand (could be small molecules, nucleic acids, peptides, etc.) and protein scaffold (the other parts of protein besides the pocket region.
- Here, each atom in the ligand is treated as an individual residue, but only coordinate (similar to C_α coordinates) need to be predicted.



Method

■ PocketFlow

- For conditional flow matching, the key point is to learn the conditional vector field, which can be calculated based on **prior distribution**, sample distribution and the **conditional flow**.

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],p_1(x_1),p_t(x|x_1)}\|v_\theta(x,t) - u_t(x|x_1)\|_g^2$$

$$\frac{d}{dt}\psi_t(x) = u_t\left(\psi_t(x) \mid x_1\right)$$

- Here, considering that protein-ligand complex has **many different components**, we need to define PocketFlow for these different components (backbone, sidechain, and residue/interaction types).

$$\mathcal{R} = \{\boldsymbol{c}^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{O}^{(i)}, \boldsymbol{\chi}^{(i)}, I^{(i)}\}_{i=1}^{N_r}$$

Define prior distribution and conditional flow for these different components

Method

■ PocketFlow on Backbone

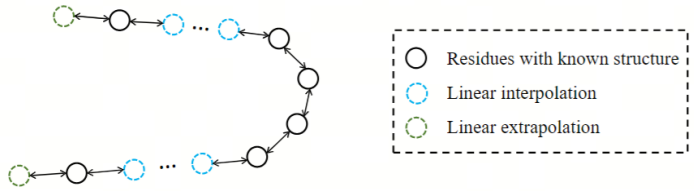


Figure 5: Structure initialization based on interpolation and extrapolation

➢ Prior Distirbution:

- For C_α **coordinates**: the linear interpolation and extrapolation based on the known coordinates of neighboring scaffold residues
- For **frame orientation matrix**: the uniform distribution on SO(3)

➢ Conditional flow:

$$\boldsymbol{x}_t^{(i)} = (1-t)\boldsymbol{x}_0^{(i)} + t\boldsymbol{x}_1^{(i)} \qquad \boldsymbol{O}_t^{(i)} = \exp_{\boldsymbol{O}_0^{(i)}}\left(t \log_{\boldsymbol{O}_0^{(i)}}(\boldsymbol{O}_1^{(i)})\right)$$

➢ Loss:

$$\mathcal{L}_{coord}(\theta) = \mathbb{E}_{t,p_1(\boldsymbol{x}_1),p_0(\boldsymbol{x}_0),p_t(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{x}_1)} \frac{1}{N_r+N_l} \sum_{i=1}^{N_r+N_l} \left\| v_\theta^{(i)}(\boldsymbol{x}_t^{(i)}, t) - \boldsymbol{x}_1^{(i)} + \boldsymbol{x}_0^{(i)} \right\|_2^2, \quad (1)$$

$$\mathcal{L}_{ori}(\theta) = \mathbb{E}_{t,p_1(\boldsymbol{O}_1),p_0(\boldsymbol{O}_0),p_t(\boldsymbol{O}_t|\boldsymbol{O}_0,\boldsymbol{O}_1)} \frac{1}{N_r} \sum_{i=1}^{N_r} \left\| v_\theta^{(i)}(\boldsymbol{O}_t^{(i)}, t) - \frac{\log_{\boldsymbol{O}_t^{(i)}}(\boldsymbol{O}_1^{(i)})}{1-t} \right\|_{\mathrm{SO}(3)}^2, \quad (2)$$

where we additionally consider N_l ligand atom coordinates in $\mathcal{L}_{coord}(\theta)$, for which we use Gaussian distribution at the center of ligand mass as the prior distribution.

Method

PocketFlow on Sidechain

As described in Sec. 3.1, the sidechain conformation of each residue can be represented as maximally four torsion angles $\boldsymbol{\chi}^{(i)} = \{\chi^{i1}, \chi^{i2}, \chi^{i3}, \chi^{i4}\} \in [0, 2\pi)^4$. In a pocket with N_r residues, the sidechain torsion angles form a hypertorus $\mathbb{T}^{4N_r}$, which is the quotient space $\mathbb{R}^{4N_r}/2\pi\mathbb{Z}^{4N_r}$ with the equivalence relation: $\boldsymbol{\chi} = (\chi^1, \ldots, \chi^{4N_r}) \sim (\chi^1 + 2\pi, \ldots, \chi^{4N_r}) \sim (\chi^1, \ldots, \chi^{4N_r} + 2\pi)$ [41, 90]. Following [42], the prior distribution is chosen as a uniform distribution over $\mathbb{T}^{4N_r}$. We regard the torsion angles as mutually independent and use interpolation paths as: $\boldsymbol{\chi}_t = (1-t)\boldsymbol{\chi}_0 + t(\boldsymbol{\chi}'_1 - \boldsymbol{\chi}_0)$ where $\boldsymbol{\chi}'_1 = (\boldsymbol{\chi}_1 - \boldsymbol{\chi}_0 + \pi) \mod (2\pi) - \pi + \boldsymbol{\chi}_0$. The loss for the torsion angles is defined as:

$$\mathcal{L}_{tor}(\theta) = \mathbb{E}_{t, p_1(\boldsymbol{\chi}_1), p_0(\boldsymbol{\chi}_0), p_t(\boldsymbol{\chi}_t|\boldsymbol{\chi}_0, \boldsymbol{\chi}_1)} \frac{1}{N_r} \sum_{i=1}^{N_r} \left\| v_\theta^{(i)}(\boldsymbol{\chi}_t^{(i)}, t) - \boldsymbol{\chi}_1'^{(i)} + \boldsymbol{\chi}_0^{(i)} \right\|_2^2. \tag{3}$$

PocketFlow on Residue Types and Interaction Types

- The prior distribution is set as the uniform distribution and the conditional flow is defined as the Euclidean interpolation between initial data and sample data.

$$\mathcal{L}_{res} = \mathbb{E}_{t\sim\mathcal{U}(0,1), p_1(\boldsymbol{c}_1), p_0(\boldsymbol{c}_0), p_t(\boldsymbol{c}|\boldsymbol{c}_0, \boldsymbol{c}_1)} \sum_{i=1}^{N_r} \mathrm{CE}\left(\boldsymbol{c}_t^{(i)} + (1-t)v_\theta^{(i)}(\boldsymbol{c}_t^{(i)}, t), \boldsymbol{c}_1^{(i)}\right),$$

$$\mathcal{L}_{inter} = \mathbb{E}_{t\sim\mathcal{U}(0,1), p_1(I_1), p_0(I_0), p_t(I|I_0, I_1)} \sum_{i=1}^{N_r} \mathrm{CE}\left(I_t^{(i)} + (1-t)v_\theta^{(i)}(I_t^{(i)}, t), I_1^{(i)}\right).$$

Method

■ Model Architecture

➢ PocketFlow adopt the neural network architecture from the FrameDiff , which incorporates Invariant Point Attention from AF2 to encode spatial features combined with transformer layers to encode sequence-level features.

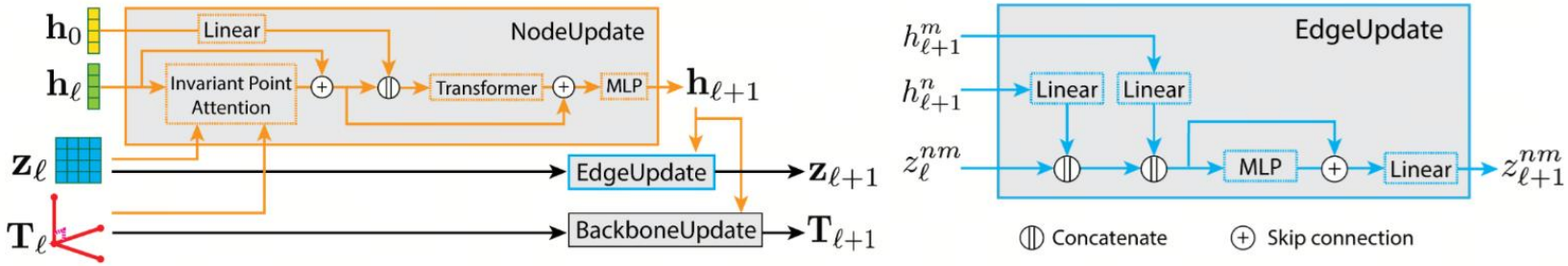


Figure 2: Framework of FrameDiff

$$\mathcal{R} = \{\boldsymbol{c}^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{O}^{(i)}, \boldsymbol{\chi}^{(i)}, I^{(i)}\}_{i=1}^{N_r}$$

Residue/Interaction Type and Torsion angle Prediction. We predict the residue/interaction types and sidechain torsion angles based on node embeddings.

$$\boldsymbol{h_c} = \text{MLP}(\boldsymbol{h}^L), \quad \boldsymbol{h}_I = \text{MLP}(\boldsymbol{h}^L), \quad \boldsymbol{h_\chi} = \text{MLP}(\boldsymbol{h}^L), \tag{31}$$

$$\boldsymbol{c} = \text{softmax}(\text{Linear}(\boldsymbol{h_c} + \boldsymbol{h}^L)), I = \text{softmax}(\text{Linear}(\boldsymbol{h}_\psi + \boldsymbol{h}^L)), \tag{32}$$

$$\boldsymbol{\chi} = \text{Linear}(\boldsymbol{h_\chi} + \boldsymbol{h}^L) \text{mod} 2\pi \tag{33}$$

Method

unconditional vector field

guidance term

$$\nabla_{\mathcal{C}_t} \log p(\mathcal{C}_t|y) = \nabla_{\mathcal{C}_t} \log p(\mathcal{C}_t) + \nabla_{\mathcal{C}_t} \log p(y|\mathcal{C}_t),$$

■ Sampling

➢ Generally, we use **classifier-guided sampling** and consider overall binding affinity guidance and interaction geometry guidance.

- For binding affinity guidance, we just train a separate lightweight affinity predictor for guidance.
- For interaction geometry guidance, we make the local geometries satisfy a series of distance/angle constraints.

Sampling. With the initialized data, the sampling process is the integration of the ODE $\frac{d\mathcal{C}_t}{dt} = v_\theta(\mathcal{C}_t, t)$ from $t = 0$ to $t = 1$ with an Euler solver [14]. γ, ξ_1, ξ_2, and ξ_3 are set as 1 in the default setting. To apply the guidance, we use $\tilde{v}_\theta$ which is v_θ plus guidance terms (Equ. 7, 9, and 10):

$$\chi_{t+\Delta t}^{(i)} = \text{reg}\left(\chi_t^{(i)} + \tilde{v}_\theta(\chi_t^{(i)}, t)\Delta t\right); \tag{11}$$

$$\boldsymbol{x}_{t+\Delta t}^{(i)} = \boldsymbol{x}_t^{(i)} + \tilde{v}_\theta(\boldsymbol{x}_t^{(i)}, t)\Delta t; \quad \boldsymbol{O}_{t+\Delta t}^{(i)} = \boldsymbol{O}_t^{(i)} \exp\left(\tilde{v}_\theta(\boldsymbol{O}_t^{(i)}, t)\Delta t\right); \tag{12}$$

$$\boldsymbol{c}_{t+\Delta t}^{(i)} = \text{norm}\left(\boldsymbol{c}_t^{(i)} + \tilde{v}_\theta(\boldsymbol{c}_t^{(i)}, t)\Delta t\right); \quad I_{t+\Delta t}^{(i)} = \text{norm}\left(I_t^{(i)} + \tilde{v}_\theta(I_t^{(i)}, t)\Delta t\right); \tag{13}$$

- Introduction
- Method
- Experiment
- Conclusion

Experiment

■ Datasets

- Following previous works, we consider two widely used protein-small molecule binding datasets for experimental evaluations:
  - **CrossDocked dataset:** This dataset is generated through crossdocking and is split with mmseqs2 at 30% sequence identity, leading to train/val/test set of 100k/100/100 complexes.
  - **Binding MOAD dataset:** This dataset contains experimentally determined protein-small molecule complexes and is split based on the proteins' enzyme commission number, leading to train/val/test set of 40k/100/100.
- Besides, to test the generalizability of PocketFlow to other ligand modalities, we further consider **PPDBench**, which contains 133 non-redundant complexes of protein-peptides and **PDBBind RNA**, which contains 56 protein-RNA pairs by filtering the PDBBind nucleic acid subset.

Experiment

■ Performance Metrics

- **Amino Acid Recovery (AAR):** the overlapping ratio between the predicted and ground truth residue types.
- **scRMSD:** the self-consistency Root Mean Squared Deviation between the generated and the predicted pocket's backbone atoms to reflect structural validity.
- Binding affinity: we choose different binding affinity metrics for different ligands.

predicted structures. To measure the binding affinity for protein-small molecule pairs, we calculate **Vina Score** with AutoDock Vina [78] following [64, 92]. For protein-peptide and protein-RNA pairs, we calculate **Rosetta $\Delta\Delta G$** [5] and **Rosetta-Vienna RNP-$\Delta\Delta G$** [44] respectively that measure the binding affinity change. The unit is kcal/mol and a lower Vina score/$\Delta\Delta G$ indicates higher affinity.

Experiment

■ Small-molecule-binding Pocket Design

Table 1: Evaluation of different models on **small-molecule-binding** protein pocket design. We report the average and standard deviation values across three independent runs. We highlight the best two results with **bold text** and underlined text, respectively.

Model	CrossDocked			Binding MOAD		
	AAR (↑)	scRMSD (↓)	Vina (↓)	AAR (↑)	scRMSD (↓)	Vina (↓)
Test set	-	0.65	-7.016	-	0.67	-8.076
DEPACT	31.52±3.26%	0.73±0.06	-6.632±0.18	35.30±2.19%	0.77±0.08	-7.571±0.15
dyMEAN	38.71±2.16%	0.79±0.09	-6.855±0.06	41.22±1.40%	0.80±0.12	-7.675±0.09
FAIR	40.16±1.17%	0.75±0.03	-7.015±0.12	43.68±0.92%	0.72±0.04	-7.930±0.15
RFDiffusionAA	50.85±1.85%	0.68±0.07	-7.012±0.09	49.09±2.49%	0.70±0.04	-8.020±0.11
PocketFlow	**52.19±1.34%**	0.67±0.04	**-8.236±0.16**	**54.30±1.70%**	0.68±0.03	**-9.370±0.24**
w/o Aff Guide	50.94±1.37%	**0.65±0.04**	-7.375±0.10	51.43±1.52%	**0.64±0.04**	-8.380±0.19
w/o Geo Guide	49.80±1.41%	0.68±0.03	-8.120±0.14	53.49±1.53%	0.71±0.05	-9.197±0.22
w/o Geo & Aff Guide	48.50±1.66%	0.71±0.06	-7.135±0.13	49.71±1.68%	0.69±0.03	-8.241±0.18
w/o Inter Learning	50.72±1.20%	0.66±0.03	-7.968±0.15	52.25±1.74%	0.68±0.05	-9.031±0.17

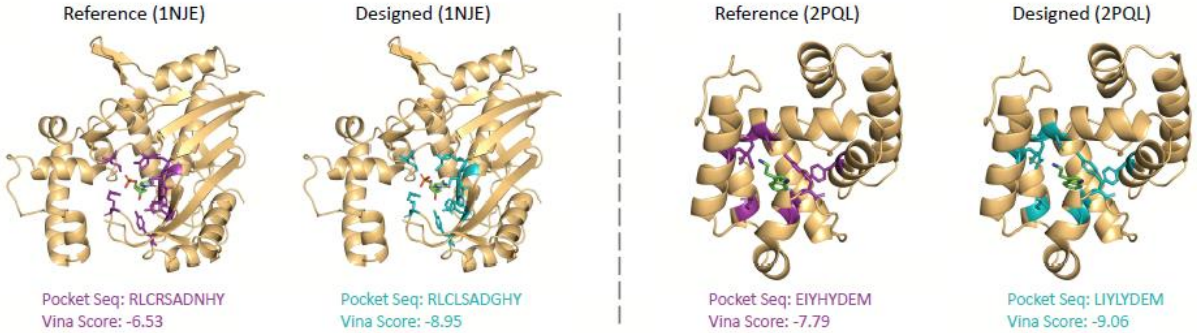
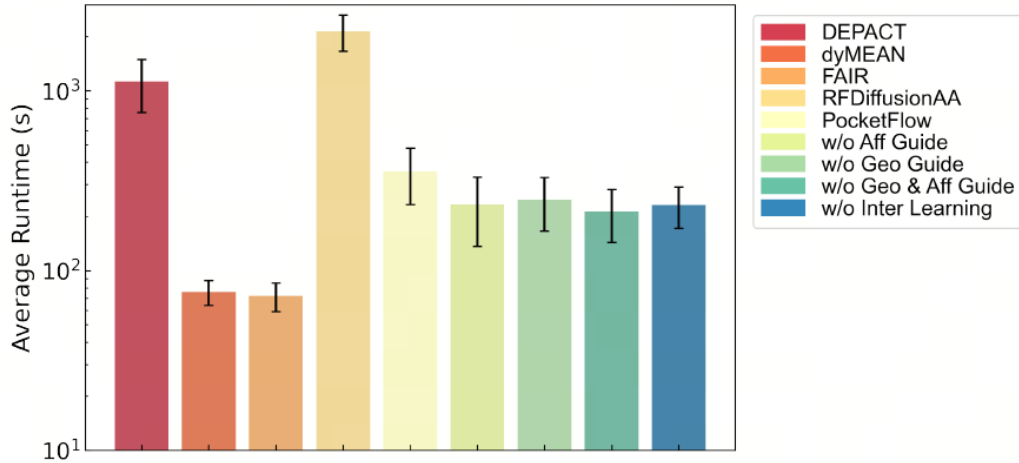


Figure 2: Case studies of small-molecule-binding protein pocket design. We show the reference and designed structures/sequences of two protein pockets from the CrossDocked (PDB ID: 1NJE) and Binding MOAD (PDB ID: 2PQL) datasets respectively.



State-of-the-art performance with acceptable generation efficiency

Experiment

■ Generalization to Other Ligand Domains

➢ It explores whether the pretrained PocketFlow on the combination of CrossDocked and Binding MOAD can generalize to peptide and RNA binding pocket design.

Table 2: Evaluation of different approaches on the **peptide** and **RNA** datasets. DEPACT is not reported here because it is specially designed for small molecules. dyMEAN, FAIR, and PocketFlow are pretrained on protein-small molecule datasets and we use the checkpoint of RFDiffusionAA [1].

Model	PPDBench			PDBBind RNA		
	AAR (↑)	scRMSD (↓)	$\Delta\Delta G$ (↓)	AAR (↑)	scRMSD (↓)	$\Delta\Delta G$ (↓)
Test set	-	0.64	-	-	0.59	-
dyMEAN	26.29±1.05%	0.71±0.05	-0.23±0.04	25.90±1.22%	0.71±0.04	-0.18±0.03
FAIR	32.53±0.89%	0.86±0.04	0.05±0.07	24.90±0.92%	0.80±0.05	0.13±0.05
RFDiffusionAA	46.85±1.45%	**0.65±0.06**	-0.62±0.05	**44.69±1.90%**	**0.65±0.03**	-0.45±0.07
PocketFlow	**48.19±1.34%**	0.67±0.04	**-1.06±0.04**	44.34±1.16%	0.69±0.01	**-0.78±0.07**
w/o Aff Guide	47.78±1.18%	0.70±0.02	-0.47±0.10	42.15±1.56%	0.68±0.04	-0.35±0.11
w/o Geo Guide	47.30±1.94%	0.72±0.05	-0.96±0.08	41.73±2.34%	0.77±0.09	-0.65±0.15
w/o Geo & Aff Guide	44.63±1.79%	0.78±0.05	-0.31±0.05	39.70±1.24%	0.78±0.06	-0.26±0.08
w/o Inter Learning	36.41±1.38%	0.74±0.06	-0.34±0.05	36.27±1.47%	0.82±0.13	-0.23±0.06

Comparable performance to the state-of-the-art baseline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- In this paper, we proposed PocketFlow, a protein-ligand interaction prior-informed flow matching model for **protein pocket generation**.
- We define multimodal flow matching for **protein backbone frames, sidechain torsion angles**, and **residue/interaction types** to appropriately represent the protein-ligand complex.
- The binding affinity and interaction geometry guidance effectively improve the validity and affinity of the generated pockets.
- Moreover, PocketFlow offers a unified framework covering small-molecule, nucleic acids, and peptides-binding protein pocket generation.

Thank You for listening!

Fanmeng Wang

2024-10-15