Rag2Mol: Structure-based drug design based on Retrieval Augmented Generation

Peidong Zhang[1,2], **Xingang Peng**[3], **Rong Han**[1,2], **Ting Chen**[1,2,*], **and Jianzhu Ma**[3,*]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Institute of Artificial Intelligence, Tsinghua University, Beijing, China
[3]Institute for Artificial Intelligence, Peking University, Beijing, China
[*]Correspondence should be addressed to

ABSTRACT

Artificial intelligence (AI) has brought tremendous progress to drug discovery, yet identifying hit and lead compounds with optimal physicochemical and pharmacological properties remains a significant challenge. Structure-based drug design (SBDD) has emerged as a promising paradigm, but the inherent data biases and ignorance of synthetic accessibility render SBDD models disconnected from practical drug discovery. In this work, we explore two methodologies, Rag2Mol-G and Rag2Mol-R, both based on retrieval-augmented generation (RAG) to design small molecules to fit a 3D pocket. These two methods involve searching for similar small molecules that are purchasable in the database based on the generated ones, or creating new molecules from those in the database that can fit into a 3D pocket. Experimental results demonstrate that Rag2Mol methods consistently produce drug candidates with superior binding affinities and drug-likeness. We find that Rag2Mol-R provides a broader coverage of the chemical landscapes and more precise targeting capability than advanced virtual screening models. Notably, both workflows identified promising inhibitors for the challenging target PTPN2. Our highly extensible framework can integrate diverse SBDD methods, marking a significant advancement in AI-driven SBDD. The codes are available at: https://github.com/CQ-zhang-2016/Rag2Mol.

1 Introduction

Artificial intelligence (AI) has achieved tremendous success in drug discovery, providing profound and inspiring insights across various stages[1,2]. Of these, finding hit/lead compounds, as the foundation for starting a new drug discovery project, remains a significant challenge in biopharmaceutical research, particularly in acquiring favorable physicochemical and pharmacological properties[3]. Researchers have explored various approaches to discover protein pocket-specific molecules[4–6], and structure-based drug design (SBDD) has emerged as a promising paradigm due to its superior performance among these efforts[7]. Unlike virtual screening methods[8–12] limited by the chemical diversity of existing compound libraries or ligand-based drug design (LBDD) methods[13–17] which lacked protein structural information, SBDD models explicitly incorporate the three-dimensional geometric information of target proteins, directly generating ligands with appropriate topological structures and high binding affinity within the protein pocket[18–23].

Recently, several representative methods have been proposed for SBDD tasks, demonstrating promising results. Work[24] is the first approach to generate molecules based on 3D structures with favorable biological properties. It iteratively adds atoms or bonds to construct 3D molecules, making autoregressive models[24–28] as a key category in the SBDD field. Diffusion-based methods[29–32] sample the distribution of atoms from the prior noise distribution and generate entire molecules in iterative feed-forward processes followed by post-processing to assign bonds. Additionally, flow-based[33] and language model-based[34] approaches encode contextual information to obtain rotation-invariant features, offering different perspectives in molecular generation.

However, several challenges have rendered SBDD models disconnected from practical drug discovery. Firstly, SBDD models inherently ignore the synthetic accessibility, resulting in generated structures usually falling outside the synthesizable chemical spaces[33,35]. Secondly, as data-driven methods, SBDD models are intrinsically limited by the quality and coverage of their training data[36], which is undoubtedly a drop in the ocean compared to the potential space of complexes. Finally, pocket-specific molecules designed by SBDD models require extensive downstream validation[37], with cumbersome procedures and accumulated

uncertainties deterring drug developers. Recent studies have proposed various innovative approaches to partially address these challenges, including introducing chemical knowledge[36,38] and relevant features[39], projection into synthesizable chemical space[35], and extracting interaction patterns from retrieved molecules[32]. While these pioneering efforts are promising, much room for improvement remains.

In this work, we have designed two retrieval augmented generation[40] (RAG)-based workflows to enhance the performance of SBDD models in real-world applications (as illustrated in Figure 1). In Rag2Mol-G, we use a pre-trained two-level retriever during both training and sampling. The global retriever is to retrieve database for small molecules that potentially bind to the pocket, and the molecular retriever is to rank and choose them as context information. The advantage of this workflow is that the retrieved molecules have both potential interaction affinity and synthetic accessibility, implicitly assisting the AI model in learning structural knowledge and topological rules. In Rag2Mol-R, we first generate candidate molecules based on Rag2Mol and then search for similar molecules in the public database. Such methodology is similar to the pipeline PocketCrafter developed by Novartis[41], which successfully designed three WDAC inhibitors and found the success rate of detecting similar lead compounds significantly surpasses that of direct virtual screening. In conventional virtual screening, at least three AI models need to be trained for screening, docking, and binding affinity prediction based on docked structures. The first and second type errors generated by each model will accumulate, ultimately leading to a high rate of false positives. The intuition behind the Rag2Mol-R methodology is that it consolidates multiple AI models into one model, greatly improving the success rates.

Extensive experiments provided by this study demonstrate the superiority of our workflows. The vanilla Rag2Mol achieves state-of-the-art (SOTA) performance across multiple evaluation metrics on widely-used dataset, surpassing other advanced SBDD models while maintaining interpretability; Molecules screened by Rag2Mol-R significantly outperform those selected by other advanced virtual screening tools, covering a broader chemical landscape. Thus, Rag2Mol-G fits targets with multiple binding templates, while Rag2Mol-R excels with traditionally "undruggable" targets. Furthermore, both workflows could identify promising drug candidates for the challenging case PTPN2, outperforming current active site inhibitors. Finally, although we employ widely recognized autoregressive-based methods, almost all AI-based SBDD methods could be integrated into our framework. Meanwhile, each module of the protocol exhibits convenient extensibility.

2 Methods

2.1 Overview of Rag2Mol

Rag2Mol is a RAG-based SBDD generative model, consisting of two components: an augmented molecular generator and two-level retrievers (a global retriever and a molecular retriever). As an auto-regressive model, the generator depicts the molecule generation as generating atoms one by one conditioned on the protein pocket and previously generated fragments. The retriever augments the generation process in the following ways, which distinguishes Rag2Mol from other auto-regressive SBDD models. At the beginning of the generation, we use a global retriever to build a pocket-specific molecular database (Figure 1a). This retriever contains a virtual screening model and a docking model to screen and dock potential small molecules from an external molecular database to the given pocket. In the generation step, a molecular retriever is proposed to retrieve the pocket-specific database based on the previous fragment and select reference molecules to assist the generator (Figure 1b), and the generator produces a new atom by sequentially predicting the focal atoms among existing atoms, predicting the new atom's relative position, and determined the new atom's element type and the valence bonds. Based on the Rag2Mol-generated molecules, two application workflows are developed to fit different scenarios (Figure 1c). Finally, Figures 1d and 1e present a simplified framework of Rag2Mol and the detailed information flow within the module.

Formally, let P be the pocket and M^t be the generated molecular fragment at the t-th generation step ($t = 1, 2, \ldots, T$). In the beginning, the pocket-specific molecular database, denoted as $\mathcal{D}$, is derived as:

$$\mathcal{D} = \Phi_{\mathrm{d}}(P, \mathcal{D}_{\mathrm{ext}}; \theta_{\mathrm{d}}) \tag{1}$$

where Φ_{d} is the global retriever with parameters θ_{d}, and $\mathcal{D}_{\mathrm{ext}}$ is the external molecular database. The generation process is defined as:

$$\begin{aligned} C^{t-1} &= P \cup M^{t-1} \\ R &= \Phi_{\mathrm{r}}(C^{t-1}, \mathcal{D}; \theta_{\mathrm{r}}) \\ M^t &= \Phi_{\mathrm{g}}(C^{t-1}, R; \theta_g) \end{aligned} \tag{2}$$

where Φ_{r} and Φ_{g} are the molecular retriever and the generator with parameters θ_{d} and θ_{g}, respectively. R is the reference molecule retrieved from the pocket-specific molecule database.

2.2 Implementation of the two-level retriever

Inspired by previous works, the global retriever Φ_{d} directly utilizes ConPLex[9] and FABind[42] for docking and screening. However, for the molecular retriever, we trained a light network for faster retrieving due to its frequent usage. The molecular retriever essentially needs the capability to measure the similarity between the reference molecule and the final molecules that will be generated from the current intermediate fragment. To train this retriever, we sampled a protein-molecule pair (P, M) from the dataset and randomly masked several atoms of the molecule, denoted as M_{mask}. We then selected reference molecules M_{ref} from its pocket-specific molecule database. The molecular retriever took as input the protein P, the masked molecule M_{mask}, and the reference molecule M_{ref} and was trained to predict the similarity between the reference molecule M_{ref} and the ground truth molecule M.

In the implementation, ProtBert[43] and Morgan fingerprints[44] were used to encode the proteins and the small molecules into vectors with dimensions of 1024 and 2048, respectively. Then multi-layer perceptrons were utilized to process these features and finally predicted the similarity scores. Here, the similarity was defined as the cosine similarity of the molecular fingerprints, and the mean squared error was used as the loss function.

2.3 Encoders

We modeled the binding pocket and the molecules as the k-nearest neighbor (KNN) graph where heavy atoms are nodes and each atom is connected to its k-nearest atoms with edges. To keep the E(3)-equivariance of the generator, we preserved both scalar and vector features for the node and edge features. The input scalar node features were composed of the element type, the amino acid type, the backbone/side-chain identifier of protein atoms, and an identifier to indicate whether the atom belongs to a protein or molecule. The input scalar edge features were composed of the edge lengths, the bond types, and a bool identifier indicating the valence of the bond. The input vector node features were the coordinates of heavy atoms, and vector edge features were the 3D unit directional vector of the edge.

Using the geometric vector perceptrons (GVP[45]) as the basic block, we built two parallel encoders where multiple aggregating layers G_a and updating layers G_u are concatenated and interleaved to learn the local structure representations. We denote the node and edge features as $\boldsymbol{v}$ and $\boldsymbol{e}$, respectively. The features of molecular fragment and retrieved reference molecule are encoded separately according to the following formula:

$$\boldsymbol{h} \leftarrow \sum_{KNN} G_a(\boldsymbol{v}, \boldsymbol{e}), \tag{3}$$

$$\boldsymbol{v} \leftarrow G_u(\boldsymbol{v}, \boldsymbol{h}). \tag{4}$$

2.4 Message passing

We construct a cross-KNN graph to allow messages to flow from the reference molecule to the molecular fragment. Specifically, the reference molecule and molecular fragment are aligned based on the pocket structures. Thus, the cross-KNN graph takes the generated molecular atoms as nodes, and each atom is connected to its k-nearest reference molecular atoms. Figure 1e shows the details of the message-passing module. Let $(\boldsymbol{v}_f, \boldsymbol{e}_f)$ and $(\boldsymbol{v}_r, \boldsymbol{e}_r)$ represent the encoded features of molecular fragment and reference molecule. The edges from the nodes of reference molecules to molecular fragments were represented as $\boldsymbol{e}_{r\to f}$, which were initialized as the relative distances and the unit directional vector of the atom pairs.

A similar aggregating layer is employed in the cross-KNN graph, and the aggregated node features are used to update information through cross-attention layers, ensuring smoothness and robustness. The specific formulas are as follows:

$$\boldsymbol{h}_r \leftarrow \sum_{cross-KNN} G_a(\boldsymbol{v}_r,\ \boldsymbol{e}_{r\to f}), \tag{5}$$

$$\boldsymbol{v}_f \leftarrow Attn(\boldsymbol{v}_f,\ \boldsymbol{h}_r,\ \boldsymbol{h}_r) \tag{6}$$

where *Attn* indicates the cross-attention network. The detailed formulas can be found in Supplementary Information.

2.5 Predictors

The extracted hidden representations capture not only the chemical and geometric attributes within a protein pocket but also the general molecular structural laws and interaction patterns from reference knowledge. These representations are used as the inputs for the predictors including the focal atom predictor, position predictor, and element-and-bond predictor. Following previous works[24,25,36,46], we use GVP-based networks to predict the probability of focal atoms, parameters of multivariate Gaussian mixture distribution modeling interatomic distance, element types, and probable bonding types. Particularly, we use trigonometry self-attention to capture chemical bonding relationships. The detailed explanation can be found in Supplementary Information.

2.6 Training procedure

The generator is trained to recover randomly masked atoms. Specifically, for a protein-molecule pair in the dataset, a random ratio of molecular atoms is masked, where the ratio value is sampled from a uniform distribution. The remaining atoms that have chemical bonds to the masked atoms are labeled as focus atoms. The focal predictor is trained to predict focal atoms and recover the masked atoms. We also sample noise positions from the surrounding environment as negative examples, enhancing the learning capability of Rag2Mol. For each masked fragment, we utilize the *Rank* model to retrieve the top 4 reference molecules, and Rag2Mol randomly selects one reference molecule or disregards reference information during each training step. Besides, a teacher-forcing strategy is employed, with each predictor independently trained with ground truth.

Finally, the focus atom predictor and the position predictor use binary cross-entropy loss and negative log-likelihood loss, respectively. Cross-entropy loss is applied to the multi-classification predictions of element types and bonding relationships. Rag2Mol is optimized through the sum loss of predictors.

2.7 Details of downstream workflows

As shown in Figure 1c, we have developed two workflows for drug discovery, named Rag2Mol-G and Rag2Mol-R. In the Rag2Mol-G workflow, we subject the filtered drug candidates to precise binding affinity calculations and subsequent wet-lab experiments. In the Rag2Mol-R workflow, the representative molecule is randomly selected as a scaffold template from each molecular cluster. Based on these templates, we search for similar molecules within existing synthesizable compounds. These molecules are deduplicated and then subjected to accurate docking software Glide[47], yielding the final set of drug candidates. The detailed implementation, tool explanations, and other set-ups are included in the Supplementary Information.

3 Result and discussion

We evaluate Rag2Mol's performance on the SBDD tasks and the virtual screen tasks. For the SBDD task, we chose eight SBDD models: Pocket2Mol[46], ResGen[46], AR[24], GraphBP[48], FLAG[49], TargetDiff[30], Decomp-o and Decomp-r[29], as baselines for subsequent comparison. Fairly, all models were trained and tested on the CrossDock 2020[50], and the native ligands were also compared. We use widely recognized metrics to assess the common properties: (a) **Vina Dock** and **Vina Score**: Affinity scores before and after docking, respectively. (b) **Affinity^1** and **Affinity^2**: Percentage of molecules with higher **Vina Dock** and **Vina score** to the existing ligands. (c) **PB-Valid**: Percentage of valid molecules checked by the PoseBusters tool[51]. (d) **CNN affinity**: CNN-based predicted affinity[52]. (e) **Clash**: Number of steric clashes. (f) **SE**: Strain energy. (g) **QED**: Quantitative estimation of drug-likeness. (h) **SA**: Synthetic accessibility. (i) **Lipinski**: Number of obeyed rules of Lipinski's rule of five. (j) **LogP**: Partition coefficient. (k) **Diversity**: Average molecular similarities for each pocket. Besides, we use Kullback-Leibler (KL) divergence to analyze the distributions of bond angles and dihedral angles and the ratio of rings with different sizes. The root-mean-square deviation (RMSD) is used to measure pose differences before and after reprocessing.

For the virtual screen task, we use the same metrics to compare the common properties of chosen molecules by Rag2Mol-R and three other advanced virtual screen methods: ConPLex[9], DrugBAN[8], and UdanDTI[11]. Notably, the structures of target proteins are downloaded from PDB (`https://www.rcsb.org`) in real-world cases. Details of the datasets, baselines, and evaluation schemes are included in Supplementary Information.

3.1 Evaluation of common properties for generated molecules

Table 1. The mean binding energies and drug-likeness properties of top 1/3/5/10 molecules in drug generation. (↑)/ (↓) indicates larger / smaller is better. Top 3 results are highlighted with **bold** text, underlined text, and *, respectively.

	Test set	GraphBP	Pocket2Mol	ResGen	AR	FLAG	TargetDiff	Devomp-o	Decomp-r	OurModel
Affinity[1](↑)	-	52.8%	56.8%*	40.6%	35.1%	26.8%	52.9%	57.8%	46.1%	**61.5%**
Affinity[2](↑)	-	39.8%	55.7%	32.3%	42.8%	12.8%	54.8%*	54.0%	40.0%	**60.7%**
Diversity(↑)	-	0.882	0.863	0.847	0.837	**0.900**	0.892*	0.889	0.842	0.893
PB-Valid(↑)	95.0%	31.2%	73.1%	21.1%	55.6%	18.4%	50.8%	48.3%	71.8%*	**79.7%**
CNN affinity(↓)	5.551	5.107	3.714	3.424*	3.819	**2.958**	4.571	5.364	4.778	3.036
Clash(↓)	7.458	19.107	7.032*	29.228	5.923	69.441	13.213	16.947	10.086	**5.765**
SE-75%(↓)	61.184	152.322	49.792*	39.762	432.008	4532.652	589.482	3582.080	13722.683	**39.060**
Top 1										
Vina Dock(↓)	-7.204	-9.332	-9.418	-9.033	-8.391	-8.333	-10.132	-10.031*	-8.387	**-10.636**
Vina Score(↓)	-5.916	-1.170	-8.742	-7.245	-8.154	-6.498	-9.886	-9.422*	-7.562	**-10.160**
QED(↑)	0.476	0.556	0.541	0.566*	0.536	**0.619**	0.466	0.470	0.524	0.570
SA(↑)	0.728	0.468	0.740*	0.743	0.569	0.637	0.498	0.602	0.681	**0.761**
Lipinski(↑)	4.340	4.810	4.910*	4.848	4.691	**4.980**	4.610	4.525	4.680	4.940
LogP	0.894	1.552	2.761	2.711	0.664	2.702	2.448	3.492	2.174	3.486
Top 3										
Vina Dock(↓)	-7.204	-8.809	-9.258	-8.847	-8.190	-8.039	-9.735	-9.695*	-8.272	**-10.450**
Vina Score(↓)	-5.916	2.983	-8.504	-6.958	-7.914	-6.098	-9.232	-9.035*	-7.400	**-9.992**
QED(↑)	0.476	0.496	0.538	0.572	0.525	**0.624**	0.483	0.472	0.525	0.567*
SA(↑)	0.728	0.478	0.738*	0.755	0.571	0.641	0.506	0.600	0.678	**0.768**
Lipinski(↑)	4.340	4.787	4.890*	4.859	4.729	**4.993**	4.587	4.566	4.643	4.933
LogP	0.894	1.470	2.656	2.718	0.649	2.438	2.481	3.433	2.073	3.399
Top 5										
Vina Dock(↓)	-7.204	-8.515	-9.144	-8.718	-8.071	-7.849	-9.499*	-9.507	-8.160	**-10.342**
Vina Score(↓)	-5.916	6.551	-8.364	-6.780	-7.735	-5.856	-8.872*	-8.796	-7.246	**-9.886**
QED(↑)	0.476	0.523	0.541	0.571	0.524	**0.635**	0.492	0.471	0.526	0.557*
SA(↑)	0.728	0.478	0.739*	0.761	0.572	0.643	0.513	0.602	0.674	**0.780**
Lipinski(↑)	4.340	4.776	4.898*	4.870	4.732	**4.994**	4.598	4.560	4.628	4.918
LogP	0.894	1.430	2.608	2.706	0.629	2.306	2.409	3.381	2.016	3.367
Top 10										
Vina Dock(↓)	-7.204	-8.091	-8.954	-8.510	-7.880	-7.544	-9.131*	-9.207	-7.957	**-10.171**
Vina Score(↓)	-5.916	9.120	-8.136	-6.478	-7.442	-5.473	-8.316*	-8.400	-6.987	**-9.702**
QED(↑)	0.476	0.529	0.553	0.571	0.516	**0.640**	0.490	0.473	0.529	0.560*
SA(↑)	0.728	0.485	0.741*	0.763	0.574	0.644	0.525	0.604	0.670	**0.776**
Lipinski(↑)	4.340	4.778	4.901*	4.861	4.731	**4.993**	4.609	4.518	4.609	4.919
LogP	0.894	1.366	2.572	2.584	0.579	2.119	2.253	3.234	1.891	3.305

Table 1 shows that Rag2Mol exhibits optimal or near-optimal results across almost all metrics covering binding energies and drug-likenesses. Docking-related scores evaluate both the binding strength of generated molecules and the SBDD model's ability to position molecular poses. Some generated molecules with distorted structures (observed in GraphBP) would be reprocessed during docking, resulting in false-positive results. In contrast, Rag2Mol's top 1/3/5/10 molecules consistently demonstrate the lowest binding energies across both affinity evaluations, exceeding the affinity of native ligands by about 3 kcal/mol (around 125-fold activity). Affinity metrics indicate that more than 60% of the Rag2Mol-generated molecules have higher binding affinities than natural ligands, outperforming the closest SBDD method by about 5%. These results indicate that Rag2Mol is more likely to produce tightly binding molecules.

Moreover, SA score, SE, Clash and PB-Valid metrics suggest that Rag2Mol demonstrates a greater ratio of generating valid molecules under equivalent conditions. The QED and Lipinski scores of Rag2Mol-generated molecules are slightly lower than that of FLAG, which can be attributed to the fact that FLAG assembles predefined molecular fragments rather than individual atoms, reflecting a methodological bias. The most detected clashes suggest inherent thermodynamic instability in the FLAG method. Besides, the average LogP values for all generated molecules range between 0.58 and 3.5, within the commonly accepted range. Finally, Rag2Mol also shows high diversity.

3.2 Quality of generated conformation

The substructure distribution in generated molecules should align with those in natural molecules, exhibiting appropriate docking poses and thermodynamic stability. As illustrated in Figures 2a-f, Rag2Mol closely mirrors the test set's distribution of ring sizes, particularly favoring stable 6-membered rings while disfavoring uncommon 3-, 4-, and 7-membered rings, reflecting its effective learning of structural features via retrieval-

augmented learning. Diffusion-based methods, however, display lower reliability due to denoising challenges in mixed spaces, while FLAG explicitly avoids larger rings but ignores fused rings (e.g., 6+6, 6+5).

Figure 2g shows that Rag2Mol generates smoother conformations, reflected by the lowest RMSD values, while Figure 2h indicates Rag2Mol captures complex geometric distributions within pockets more effectively than other methods. Statistically significant p-values further support these findings.

We also evaluated 9 common bond angles and dihedral angles using KL divergence, where lower values signify higher consistency with the test set. As shown in Figure 3, Rag2Mol achieved the best performance in 5 metrics and competitive results in the remaining ones, particularly excelling in dihedral angles, a challenging metric for other SBDD methods. We believe that the introduction of retrieval augmentation has mitigated SBDD's limitations in modeling complex substructures.

3.3 Interaction pattern analysis

Figure 4b-d demonstrates Rag2Mol's ability to capture microscopic interaction patterns on three therapeutic targets: AKT1 (PDB id: 4gv1), CDK2 (PDB id: 1h00), and AROK (PDB id: 1zyu). Rag2Mol-generated molecules display reasonable binding poses, suggesting effective inference of hit positioning within the protein pocket. Using PLIP[55], we analyze interactions between these targets and Rag2Mol-generated ligands, comparing them to experimentally validated active ligands. Rag2Mol reproduces most key interactions observed in experimental ligands (6/7 for 4gv1, 4/6 for 1h00, 3/6 for 1zyu) and predicts additional, physically plausible interactions (e.g., 4 and 5 extra interactions for 4gv1 and 1zyu, respectively), enhancing binding potential. Detailed interaction statistics are provided in Supplementary Tables.

For the CDK2 target, Rag2Mol preserves critical interactions, including hydrophobic contacts and hydrogen bonds with ASP145 and LEU83, while capturing an extra water-mediated bridge involving LYS20. For AROK, Rag2Mol retains all water bridges (ARG110, ARG117) and π-cation interaction (ARG110), and also favors phosphorus atom generation. While it generates five additional hydrogen bonds. Similarly, for AKT1, Rag2Mol accurately reproduced hydrophobic interactions with ALA177 and LYS179 and hydrogen bonds with ALA230, GLU234, and ASP292, while introducing extra interactions that further enhanced binding potential. These results highlight Rag2Mol's ability to learn advanced energy distributions and interaction rules.

3.4 Detecting similar lead compounds

In Table 2, we compared the molecules identified by Rag2Mol-R with those by advanced virtual screening models. Overall, Rag2Mol-R demonstrated superior performance across nearly all metrics related to binding affinity and drug-likeness. It not only inherits the high affinity of designed molecules of Rag2Mol (we demonstrate the connection in Supplementary Figure) but also ensures the identified drug candidates are readily synthesizable. Note that a significant percentage of molecules identified by Rag2Mol-R (approximately 74%) showed higher affinity scores than the native ligands on the target, obviously outperforming other virtual screening models.

Furthermore, we used three baseline models to search approximately 10,000 molecules for a randomly selected target protein (PDB: 3zkg) and visualized the chemical space of those molecules using the t-SNE algorithm. As illustrated in Figure 4a, even though the broad chemical spaces are given, a considerable ratio of the 115 drug candidates identified by Rag2Mol-R are positioned at or beyond the boundaries of these spaces. This effectively highlights a broader coverage of the chemical landscape and the precise targeting capability of Rag2Mol-R. Finally, the retrieval database used by Rag2Mol and the similarity search database utilized by Rag2Mol-R are both scalable.

3.5 Case study: PTPN2

Protein tyrosine phosphatases PTPN2 and PTPN1 are central regulators of inflammation, and their deletion in tumor or immune cells enhances anti-tumor immunity. However, phosphatases have long been considered undruggable due to the challenges of targeting their active sites. Recently, AC484[56] was identified as a promising monotherapy and is currently under evaluation in advanced solid tumors.

In response to this challenge, we apply Rag2Mol's two workflows to this complex problem. In Glide's precise docking evaluation, the drug candidates G1 and R1 identified by Rag2Mol-G and Rag2Mol-R exhibit affinity scores of -13.8 and -12.1, respectively, surpassing AC484's score of -11.2. Additionally, G1 and R1 both demonstrate higher synthetic accessibility, while maintaining comparable molecular weights and logP

Table 2. The mean binding energies and drug-likeness properties of top 1/3/5/10 molecules in drug repurposing. (↑)/ (↓) indicates larger / smaller is better. Top 1 results are highlighted with **bold**.

	Test set	UdanDTI	DrugBAN	ConPLex	Our model
Affinity[1](↑)	-	52.8%	53.0%	57.4%	**74.2%**
			Top 1 / Top 3		
Vina Dock(↓)	-7.204 / -7.204	-9.001 / -8.384	-8.100 / -7.567	-10.141 / -9.837	**-10.625 / -10.411**
QED(↑)	0.476 / 0.476	0.351 / 0.355	0.272 / 0.282	0.365 / 0.393	**0.509 / 0.503**
SA(↑)	0.728 / 0.728	0.610 / 0.604	0.591 / 0.594	0.714 / 0.727	**0.814 / 0.823**
Lipinski(↑)	4.340 / 4.340	3.990 / 4.020	3.910 / 3.853	4.240 / 4.300	**4.990 / 4.963**
LogP	0.894 / 0.894	4.520 / 4.618	2.269 / 4.026	5.179 / 4.766	4.463 / 4.313
			Top 5 / Top 10		
Vina Dock(↓)	-7.204 / -7.204	-8.312 / -8.215	-7.252 / -6.326	-9.635 / -9.281	**-10.273 / -10.047**
QED(↑)	0.476 / 0.476	0.352 / 0.361	0.290 / 0.306	0.407 / 0.438	**0.500 / 0.507**
SA(↑)	0.728 / 0.728	0.622 / 0.638	0.598 / 0.597	0.733 / 0.741	**0.822 / 0.825**
Lipinski(↑)	4.340 / 4.340	4.040 / 4.124	3.830 / 3.823	4.336 / 4.379	**4.962 / 4.962**
LogP	0.894 / 0.894	4.585 / 4.510	3.761 / 3.695	4.643 / 4.394	4.174 / 4.060

values. As illustrated in Figure 4e, G1 shares structural similarities with AC484, while R1 presents a distinct molecular scaffold. All three molecules exhibit significant overlap within the target pocket.

Interaction analysis shows that R1 and G1 capture most of the key hydrogen bonds formed with LYS122, ASP182-PHE183, SER217-ARG222, and GLN264, which play a crucial role in AC484's inhibition of phosphatases. The hydrophobic interactions of R1 and G1 closely match those of AC484. Additionally, both R1 and G1 form extra pi-stacking interactions with the TYR48's side chain, further strengthening their binding. Notably, G1 enhances stability by converting the original hydrophobic interaction between TYR48 and AC484 into a hydrogen bond. Complete statistical metrics are provided in the Supplementary Information.

4 Conclusion

This paper introduces a drug discovery protocol with two distinct workflows. Inspired by the concept of RAG, we developed Rag2Mol, a RAG-based E(3)-equivariant GNN generative model. Rag2Mol actively integrates retrieved reference knowledge during both training and generation procedures to generate 3D drug-like molecules targeting protein pockets. Experimental results demonstrate that introducing retrieved knowledge enables the SBDD model to better comprehend biochemical rules and perceive the geometric environment, thereby meeting the demands of real-world drug design projects. The molecules identified through the Rag2Mol-R workflow significantly outperform those found by current SOTA virtual screening models across various evaluation metrics, meanwhile covering a broader chemical landscape. The components of our protocol, including the two-level retriever, docking module, databases, similarity search module, and filtering mechanisms, all exhibit strong scalability. We believe this work offers valuable new perspectives for the drug discovery field.

References

1. Mak, K.-K., Wong, Y.-H. & Pichika, M. R. Artificial intelligence in drug discovery and development. *Drug Discovery Evaluation: Safety Pharmacokinetic Assays* 1–38 (2023).
2. Blanco-Gonzalez, A. *et al.* The role of ai in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals* **16**, 891 (2023).
3. Moret, M. *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications* **14**, 114 (2023).
4. Böhm, H.-J. The computer program ludi: a new method for the de novo design of enzyme inhibitors. *Journal computer-aided molecular design* **6**, 61–78 (1992).
5. Wang, R., Gao, Y. & Lai, L. Ligbuilder: a multi-purpose program for structure-based drug design. *Molecular modeling annual* **6**, 498–516 (2000).
6. Li, Y., Pei, J. & Lai, L. Structure-based de novo drug design using 3d deep generative models. *Chemical science* **12**, 13664–13675 (2021).
7. Xie, W., Wang, F., Li, Y., Lai, L. & Pei, J. Advances and challenges in de novo drug design using three-dimensional deep generative models. *Journal Chemical Information Modeling* **62**, 2269–2279 (2022).
8. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence* **5**, 126–136 (2023).
9. Singh, R., Sledzieski, S., Bryson, B., Cowen, L. & Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings National Academy Sciences* **120**, e2220778120 (2023).
10. Huang, K., Xiao, C., Glass, L. M. & Sun, J. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* **37**, 830–836 (2021).
11. Zhang, P.-D., Ma, J. & Chen, T. Escaping the drug-bias trap: using debiasing design to improve interpretability and generalization of drug-target interaction prediction. *bioRxiv* 2024–09 (2024).
12. Gao, B. *et al.* Drugclip: Contrasive protein-molecule representation learning for virtual screening. *Advances Neural Information Processing Systems* **36** (2024).
13. Shi, C. *et al.* Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382* (2020).
14. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332 (PMLR, 2018).
15. Wang, Z. *et al.* Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126* (2022).
16. Blaschke, T. *et al.* Reinvent 2.0: an ai tool for de novo drug design. *Journal chemical information modeling* **60**, 5918–5922 (2020).
17. Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572–1583 (2019).
18. Weller, J. A. & Rohs, R. Structure-based drug design with a deep hierarchical generative model. *Journal Chemical Information Modeling* (2024).
19. Skalic, M., Varela-Rial, A., Jiménez, J., Martínez-Rosell, G. & De Fabritiis, G. Ligvoxel: inpainting binding pockets using 3d-convolutional neural networks. *Bioinformatics* **35**, 243–250 (2019).
20. Aumentado-Armstrong, T. Latent molecular optimization for targeted therapeutic design. *arXiv preprint arXiv:1809.02032* (2018).
21. Luo, S. *et al.* One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations* (2022).
22. Xu, M., Ran, T. & Chen, H. De novo molecule design through the molecular generative model conditioned by 3d information of protein binding sites. *Journal Chemical Information Modeling* **61**, 3240–3254 (2021).

23. Zhang, J. & Chen, H. De novo molecule design using molecular generative models constrained by ligand–protein interactions. *Journal chemical information modeling* **62**, 3291–3306 (2022).

24. Luo, S., Guan, J., Ma, J. & Peng, J. A 3d generative model for structure-based drug design. *Advances Neural Information Processing Systems* **34**, 6229–6239 (2021).

25. Zhang, O. *et al.* Resgen is a pocket-aware 3d molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence* **5**, 1020–1030 (2023).

26. Qian, H., Lin, C., Zhao, D., Tu, S. & Xu, L. Alphadrug: protein target specific de novo molecular generation. *PNAS nexus* **1**, pgac227 (2022).

27. Ragoza, M., Masuda, T. & Koes, D. R. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science* **13**, 2701–2713 (2022).

28. Yang, Y. *et al.* Enabling target-aware molecule generation to follow multi objectives with pareto mcts. *Communications Biology* **7**, 1074 (2024).

29. Guan, J. *et al.* Decompdiff: diffusion models with decomposed priors for structure-based drug design. *arXiv preprint arXiv:2403.07902* (2024).

30. Guan, J. *et al.* 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543* (2023).

31. Lin, H. *et al.* Diffbp: Generative diffusion of 3d molecules for target protein binding. *arXiv preprint arXiv:2211.11214* (2022).

32. Huang, Z. *et al.* Interaction-based retrieval-augmented diffusion models for protein-specific 3d molecule generation. In *Forty-first International Conference on Machine Learning*.

33. Qu, Y. *et al.* Molcraft: Structure-based drug design in continuous parameter space. *arXiv preprint arXiv:2404.12141* (2024).

34. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific reports* **11**, 321 (2021).

35. Luo, S. *et al.* Projecting molecules into synthesizable chemical spaces. *arXiv preprint arXiv:2406.04628* (2024).

36. Jiang, Y. *et al.* Pocketflow is a data-and-knowledge-driven structure-based molecular generative model. *Nature Machine Intelligence* **6**, 326–337 (2024).

37. Zhang, Z. & Liu, Q. Learning subpocket prototypes for generalizable structure-based drug design. In *International Conference on Machine Learning*, 41382–41398 (PMLR, 2023).

38. Zhu, H., Zhou, R., Cao, D., Tang, J. & Li, M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications* **14**, 6234 (2023).

39. Zhang, O. *et al.* Learning on topological surface and geometric structure for 3d molecular generation. *Nature Computational Science* **3**, 849–859 (2023).

40. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances Neural Information Processing Systems* **33**, 9459–9474 (2020).

41. Shen, L. *et al.* Pocket crafter: a 3d generative modeling based workflow for the rapid generation of hit molecules in drug discovery. *Journal Cheminformatics* **16**, 33 (2024).

42. Pei, Q. *et al.* Fabind: Fast and accurate protein-ligand binding. *Advances Neural Information Processing Systems* **36** (2024).

43. Elnaggar, A. *et al.* Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing (2021). 2007.06225.

44. Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal chemical documentation* **5**, 107–113 (1965).

45. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations* (2020).

46. Peng, X. *et al.* Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, 17644–17655 (PMLR, 2022).

47. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal medicinal chemistry* **47**, 1739–1749 (2004).

48. Liu, M., Luo, Y., Uchino, K., Maruhashi, K. & Ji, S. Generating 3d molecules for target protein binding. *arXiv preprint arXiv:2204.09410* (2022).

49. Zhang, Z., Min, Y., Zheng, S. & Liu, Q. Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations* (2023).

50. Francoeur, P. G. *et al.* Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal chemical information modeling* **60**, 4200–4215 (2020).

51. Buttenschoen, M., Morris, G. M. & Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science* **15**, 3130–3139 (2024).

52. McNutt, A. T. *et al.* Gnina 1.0: molecular docking with deep learning. *Journal cheminformatics* **13**, 43 (2021).

53. RDKit: Open-source cheminformatics. http://www.rdkit.org. [Online; accessed 11-April-2013].

54. Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C.-K. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics* **31**, 2214–2216 (2015).

55. Adasme, M. F. *et al.* PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Research* **49**, W530–W534, DOI: 10.1093/nar/gkab294 (2021).

56. Baumgartner, C. K. *et al.* The ptpn2/ptpn1 inhibitor abbv-cls-484 unleashes potent anti-tumour immunity. *Nature* **622**, 850–862 (2023).

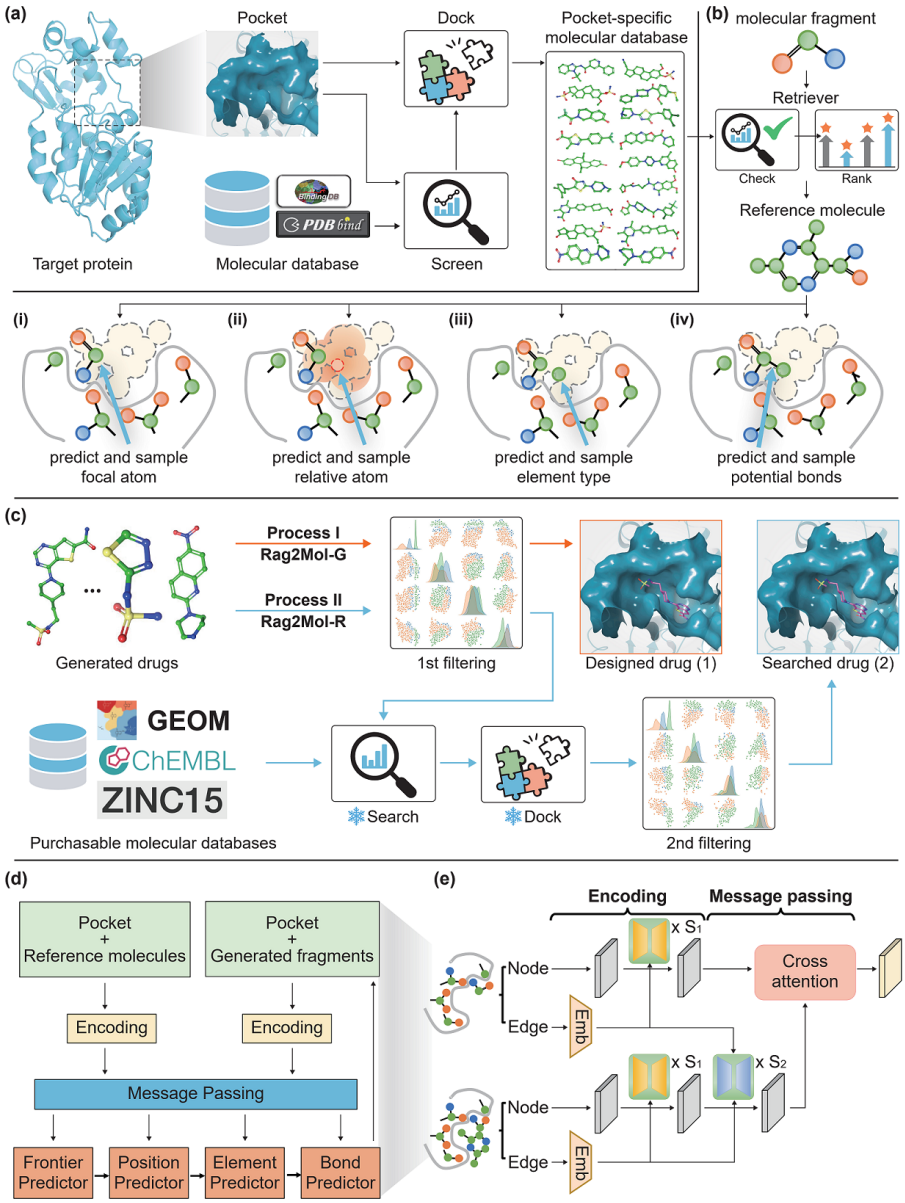

Figure 1. The Rag2Mol pipeline. (a) **Global retriever.** Constructing pocket-specific molecular database for each given target protein. (b) **One step of the Rag2Mol.** Using the reference molecule selected by the molecular retriever, the autoregressive model sequentially predicts various information for the next atom based on the generated molecular fragment. (c) **Two workflows for applying Rag2Mol.** In the Rag2Mol-G workflow, we filter drug candidates following the widely accepted threshold settings. In the Rag2Mol-R workflow, filtered molecules are then subjected to clustering and sampling. These sampled scaffolds are employed for similarity searches within synthesizable compounds. (d) **Simplified architecture of Rag2Mol.** (e) **Information transfer mechanism in hidden space.**

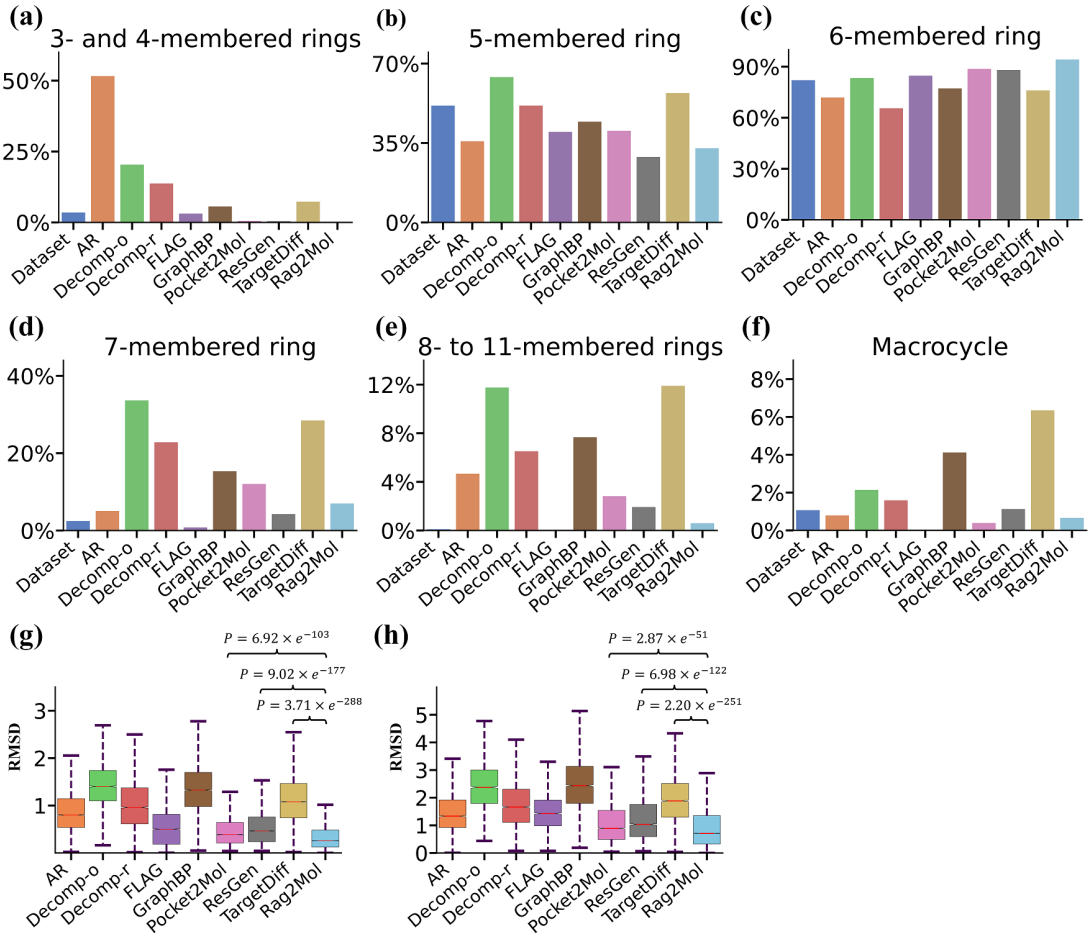

Figure 2. Conformational quality assessment. (a-f) **Rings analysis.** We compare the ratio of rings of different sizes in generated molecules by different SBDD models. The statistics of native molecules in the dataset are shown in the first column. (g-h) **SBDD models' conformational capability.** The RMSD offsets between the generated molecular structures and the structures calculated by RDKit[53] and QVina[54] are shown, respectively.

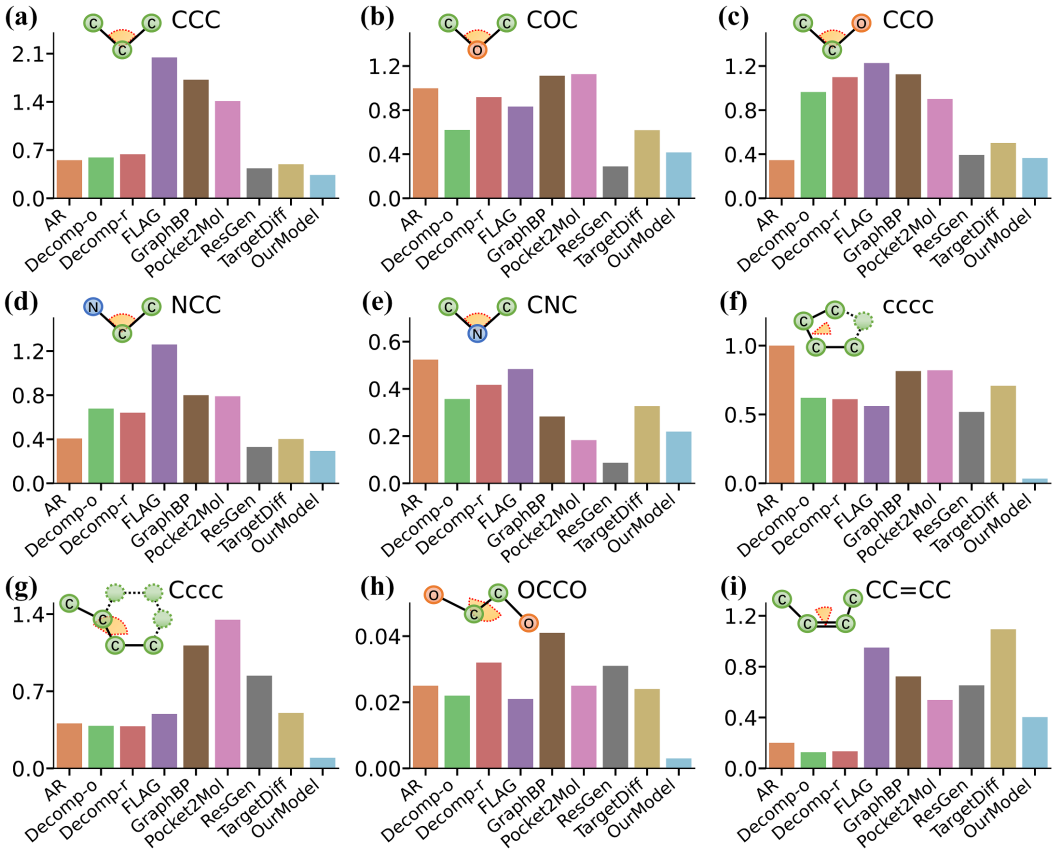

Figure 3. Bond angles and dihedral angles analysis. The distributions of the bond angles and dihedral angles of the generated molecules agree with the test set by using the KL divergence. We provide a simple diagram for each type of angles.

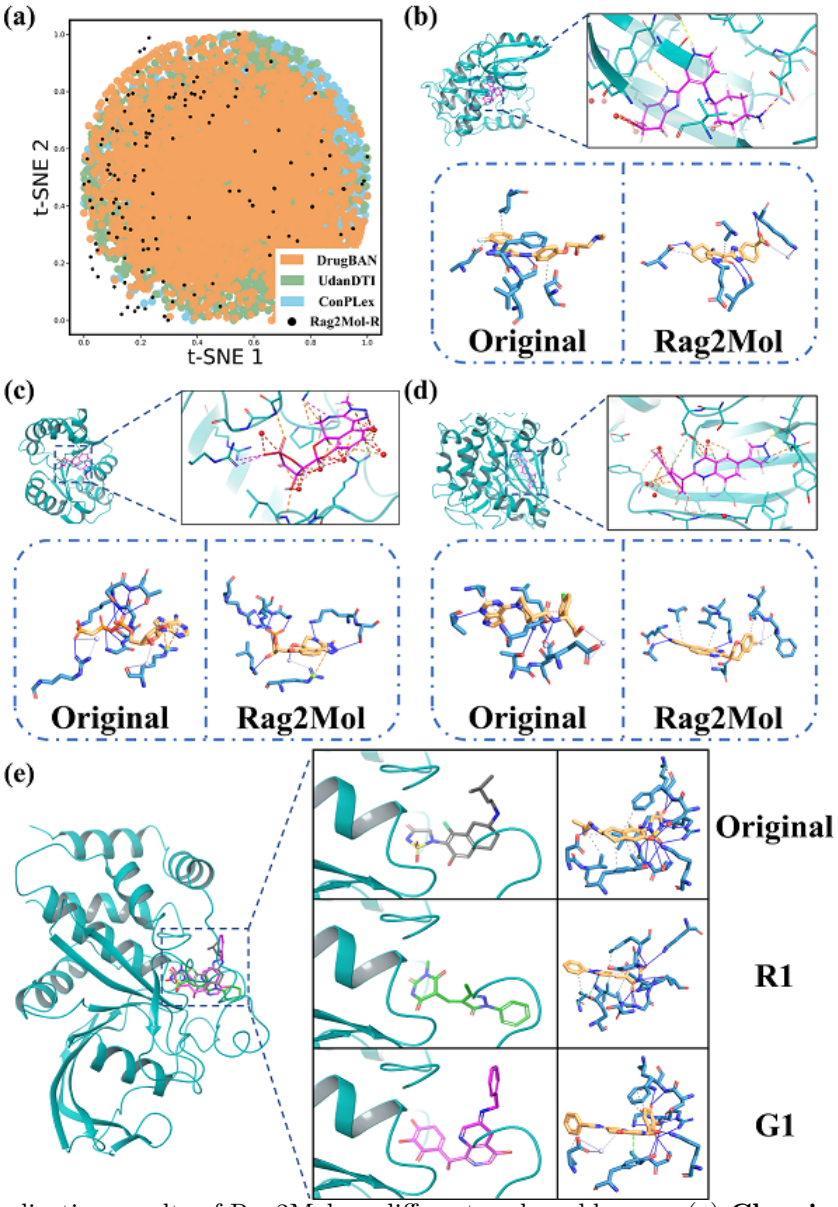

Figure 4. The application results of Rag2Mol on different real-world cases. (a) **Chemical space map representations for compounds screened by different models.** t-SNE is used for visualizing data by giving each data point a location in this two-dimensional map. The screening hits from Rag2Mol are represented in black. In comparison, compounds screened by DrugBAN, UdanDTI, and ConPLex are depicted in orange, green, and blue, respectively. (b-d) **Protein–ligand interaction analysis.** The top quadrants show the poses of Rag2Mol's generated ligands within the protein pocket, whereas the below quadrants denote the protein-ligand interaction patterns for the original and generated ligands, respectively. (e) **Rag2Mol on PTPN2.** The left image shows the overlap of the three drugs within the protein pocket, while the two columns on the right display their binding conformations and binding interaction analysis.