3D Based Generative PROTAC Linker Design with Reinforcement Learning

Baiqing Li [1], **Ting Ran**[1], **Hongming Chen**[1, *]

[1] Guangzhou Laboratory, Guangzhou 510005, Guangdong Province, China

[*]Correspondence e-mail: chen_hongming@gzlab.ac.cn

Abstract

Proteolysis targeting chimeras (PROTACs), have emerged as an effective therapeutic modality by harnessing the ubiquitin-proteasome system to selectively induce targeted protein degradation, with the potential to modulate traditional undruggable targets. Due to its hetero-bifunctional characteristics, in which a linker joins warhead binding to a protein of interest, conferring specificity, and E3-ligand binding to an E3 ubiquitin ligase, a PROTAC molecule can form a PROTAC ternary structure for bring the protein of interest to the vicinity of the E3 ligase. The rational PROTAC linker design is challenging due to its relatively large molecular weight and the complexity of maintaining the binding mode of warhead and E3-ligand in the binding pockets of counterpart. Conventional linker generation method can only generate linkers in either 1D SMILES or 2D graph, without taking into account the information of ternary structures. Here we propose a novel 3D linker generative model PROTAC-INVENT which can not only generate SMILES of PROTAC but also its 3D putative binding conformation coupled with the target protein and the E3 ligase. The model is trained jointly with the RL approach to bias the generation of PROTAC structures toward pre-defined 2D and 3D based properties. Examples were provided to demonstrate the utility of the model for generating reasonable 3D conformation of PROTACs. On the other hand, our results show that the associated workflow for 3D PROTAC conformation generation can also be used as an efficient docking protocol for PROTACs.

Introduction

Protein hydrolysis targeting chimeric (PROTAC) is emerging as a promising technology in the realm of drug discovery. Unlike traditional occupancy-driven drug design, it aims to degrade drug target protein through hijacking the ubiquitin proteasome system (UPS) in eukaryotic cells[1,2]. A standard PROTAC molecule consists of three parts, namely, a warhead fragment that binds to the protein of interest (POI), a warhead fragment that binds to the E3 ubiquitin ligase (E3 ligand), and a linker segment that connects the two fragments. This assembly method allows the PROTAC molecule to bring ubiquitinase to the vicinity of POI to ubiquitinate the POI and the ubiquitinated POI is then degraded by the UPS. It has exhibited some unique advantages over traditional small-molecule drugs. For example, as UPS is involved in degradation of more than 80% of proteins in cells[3], the PROTAC technology is theoretically able to degrade most proteins in cells. Secondly, PROTAC mediated degradation process does not rely on high affinity of the warhead to POI[4], so there is less restriction on the binding site of POI. This greatly expands drug target space so that many traditionally undruggable proteins may become druggable via PROTAC modalities[5,6]. In addition, PROTAC molecule can eliminate all functions of POI through protein degradation, rather than just blocking warhead mediated biological functions, which may lead to drug resistance caused by mutations at the binding site of POI[4,5]. Since the concept of PROTAC was proposed by Crews and coworkers[1,7], the great potential of the technology has attracted huge interests among researchers from academics as well as pharmaceutical industry. Currently a few PROTAC drugs have moved into clinical stage[8]. Some recent reviews have comprehensively summarized the characteristics of PROTACs[4,5].

So far, thousands of PROTAC molecule have been reported and the PROTAC-DB database[9] curated more than 2000 published PROTAC molecules. Compared with the chemical space of traditional drug-like small molecule, the PROTAC space is largely different due to their relatively large molecular size and still under-represented. So far, up to 60 E3 ligases have been found, but existing E3 ligands of PROTAC are mainly

focused on the ligand set of von Hippel-Lindau (VHL) and cereblon (CRBN) ligases[10]. The degradation ability of most E3 ligases has not been fully exploited. On the other hand, studies show that the linker segment is also critical for PROTAC design. It has been shown that the linker segment is closely associated with the pharmacokinetic properties and protein degradation ability of PROTACs[10,11]. Especially, it plays a pivotal role in maintaining the conformational stability of PROTAC ternary structure (PTS) (ie. the POI-PROTAC-E3 complex), since an appropriate linker fragment can ensure the energetically stable binding conformation of E3 ligand and warhead of POI[11], thus forming a stable ternary complex. However, available experience for linker design is still limited. Most of earlier linker fragments come from derived PEG chain due to its high flexibility and ease of synthesis. On this basis, various flexible linker fragments have been developed for PROTAC. Recently, relatively rigid linker fragments have also been proposed exemplified by the AR (androgen receptor) targeted PROTAC compounds in the clinical stage[12]. Thus, developing methodology for PROTAC linker design is urgently needed for PROTAC discovery.

Recently, deep learning based molecular generative model has shown significant advancement[13–18] and have been applied on fragment linking which could be used for fragment-based drug design (FBDD) and PROTAC design. For example, SyntaLinker[19], a SMILES based linker generative model based on the well-known transformer architecture, was recently been proposed to design linker to fuse two terminal fragments into a complete molecule. Zheng et al.[20] recently has applied a similar methodology on PROTAC design. To overcome the deficiency of PROTAC molecules for model training, a large collection of quasi-PROTAC molecules that have similar characteristics to PROTACs was used to pre-train a prior model, which was then fine-tuned by using actual PROTACs augmented with randomized SMILES of fragments. Moreover, they utilized reinforcement learning (RL) to optimize the model to generate PROTACs with potentially better pharmacokinetic properties. This model showed superior performance on PROTAC generation over other linker generative models. In contrast, Delinker[21] is a graph-based 2D linker generative model and it was claimed that it could be used to design PROTAC molecules. Igashov et al[22] proposed Difflinker

model by integrating diffusion model with information of protein pocket for linker generation, and multiple fragments can be assembled into a complete molecule. Guo *et al.*[23] reported Link-INVENT, a 2D linker generative model which demonstrated potential to generate both small molecules that conform to Lipinski "rule of five"[24] and PROTACs. The algorithm was essentially an expansion of the widely used REINVENT model[25] incorporating a reward function to optimize the length, linearity and flexibility of generated linkers.

Although these progresses in the development of PROTAC generative model, available methods are mainly focused on the generation of 2D PROTAC-like structures and did not consider the feasibility of a PROTAC fitting in the binding site of the PTS complex. To address this concern, we proposed a novel linker generative model, named as PROTAC-INVENT, utilizing three dimensional PTS data as constraints. The workflow consists of two stages, i.e. generation of PROTAC linker structures (i.e. SMILES) from normal generative model and then creation of 3D binding conformation of PROTAC under the constraint of the binding pocket of a reference PTS. The model was fine-tuned via RL and generated PROTACs were scored based on their docking poses in the PTS binding site. Compared with other linker generative models, one unique feature of PRO-INVENT is that a practical workflow was developed to enable generation of 3D binding conformation of PROTAC in the binding site with relatively low computation cost, so that generation of PROTAC molecules is optimized by considering its fitness with the PTS binding site. Thus, this method is expected to be more effective in the generation of PROTAC molecules which need to satisfy the 3D requirement of PTS binding site.

Methods

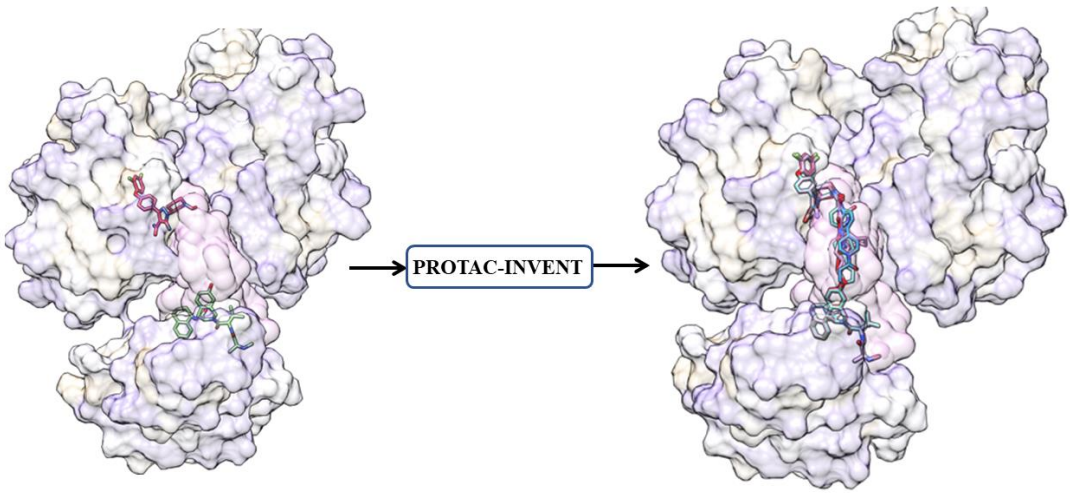


Figure 1 Illustration of PROTAC-INVENT model. The model is intended to generate PROTACs whose E3-ligand and POI warhead motifs can adopt similar binding conformation as that of the PROTAC in the reference PTS.

Model Overview

PROTAC-INVENT takes a pair of fragment (E3-ligand and warhead for POI) and a reference PTS as input, and returns generated linkers and the predicted binding conformations of the full PROTAC compounds in the binding site of reference PTS (as shown in Figure1). Two modules are integrated in PROTAC-INVENT. Firstly, a generative model is trained to produce chemically feasible linker SMILES strings, following by forming complete PROTAC molecules together with the specified warhead and E3-ligand. REINVENT was used as the core generative engine for linker design and it was previously developed as an efficient generative model for *de novo* structure design. RL was utilized to search chemical space for optimization of molecular properties; Secondly, a 3D structure-based workflow was implemented to generate docking conformations of PROTACs and provide scores accordingly. The 3D based scores were then utilized to drive the RL (as shown in Figure 2).

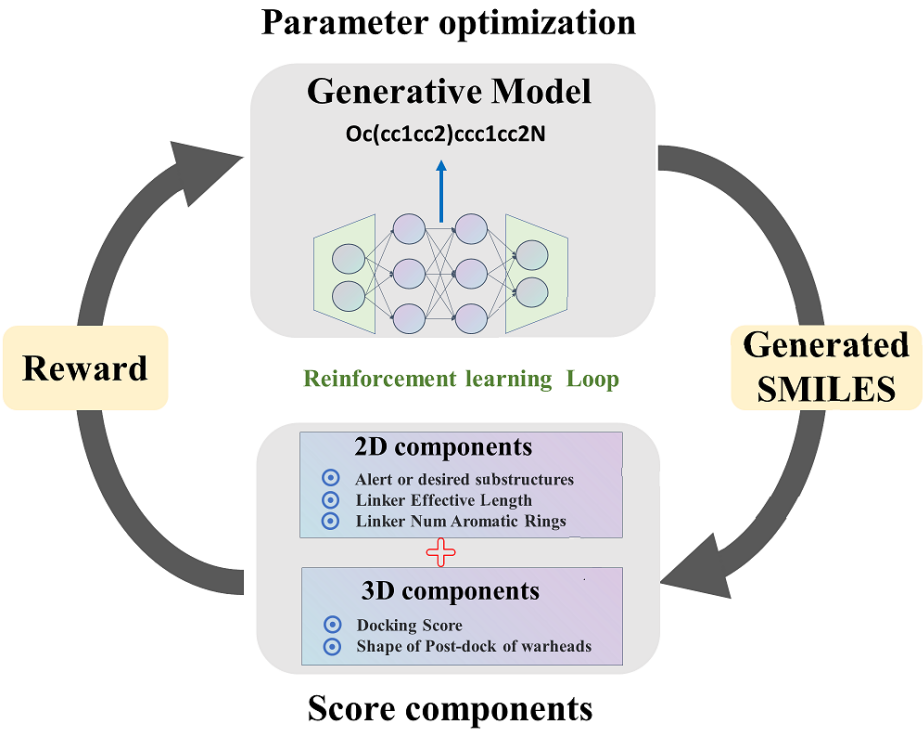


Figure 2 The general workflow of PROTAC-INVENT

Generation of PROTAC 3D conformations

Previously Link-Invent was designed to generate PROTAC linkers via RL, but their scoring function was based on purely 2D related metrics. Here we implemented a practical workflow to enable predicting 3D binding conformations of generated PROTAC molecules and score them by taking into account the structure information of a reference PTS. It takes four steps for the protocol to convert the SMILES of PROTAC to 3D binding conformation given a reference PTS as shown in Figure 3:

(1) Generate initial conformers of PROTAC in the vicinity of reference ligand

First of all, the SMILES of linker was generated by REINVENT prior model and then merged with the supplied warhead pair (as part of the input) to form the SMILES of full PROTAC. *Omega*[26] program was then utilized to convert SMILES to a set of initial 3D conformations. The 3D conformations were then superimposed with the bound conformation of the reference PROTAC by using commercially available *ROCS* program[27]. The ComboScore which is a weighted sum of shape and pharmacophore similarity was used to measure similarity between the reference conformer and the aligned conformation. Due to the large size of PROTAC, this

superimposing is usually not able to create good alignment but still can serve the purpose of bringing the linker part of the PROTAC to the vicinity of the reference linker.

(2) Merge the linker of PROTAC with the warheads of reference ligand

Once the alignment was done, the warheads of the generated PROTAC were removed and only the linker atoms were remained. The warheads of reference ligand were then copied and merged with the linker fragment of the generated PROTAC to form a re-combined PROTAC conformation (RPC). After this operation, the coordinates of warhead atoms in the RPC were exactly the same to that of reference ligand, but the bond lengths, dihedral angles between terminal atoms of the linker and warheads were not corrected yet and need to be fixed.

(3) Optimize the conformation of RPC

In order to optimize the conformation of RPC while keeping its warhead parts as close to that of the reference PROTAC conformation as possible, a constrained molecular mechanics optimization was then carried out. Here we employed the MacroModel module of *Schrödinger* software[28] to achieve the constrained minimization by imposing constraints on warhead atoms of RPC. The specific implementation, parameter setting and precautions can be found in supporting materials. After 200 iterations, the minimized RPC was saved as the optimized PROTAC conformation (OPC).

(4) Docking OPC into the binding pocket of PTS

To further evaluate generated PROTAC, its OPCs were docked into the binding pocket of PTS, which is composed by E3 ligase and POI, to obtain their docking scores. The docking procedure usually contains two parts, i.e. initial conformation/pose sampling to obtain multiple starting points and the following conformation minimization from the sampled starting points. For PROTAC molecule, its large size will result in large conformational search space and accordingly lead to long computation time. Searching for global minimum of PROTAC also becomes difficult due to the large search space, therefore it is not guaranteed that, in docking pose, the POI warhead and E3-ligand motifs can adopt

similar poses as in the reference ligand. To address this problem, we choose to use the "local-only" mode (or refine mode) to do docking, in which the initial pose sampling stage was skipped and the input conformation was taken as the sole starting point for optimization. The "local-only" docking solution was taken as the final conformation of PROTAC molecule and its docking score was used as a scoring component for RL. The docking component from DockStream[29] that is fully compatible with REINVENT[25] was served as the interface between REINVENT and docking software and DockStream supports various docking programs. In this work, we chose *AutoDock Vina*[30–32] for doing "local-only" docking, its docking score was used as a scoring component to evaluate the PROTAC molecule in RL loops.

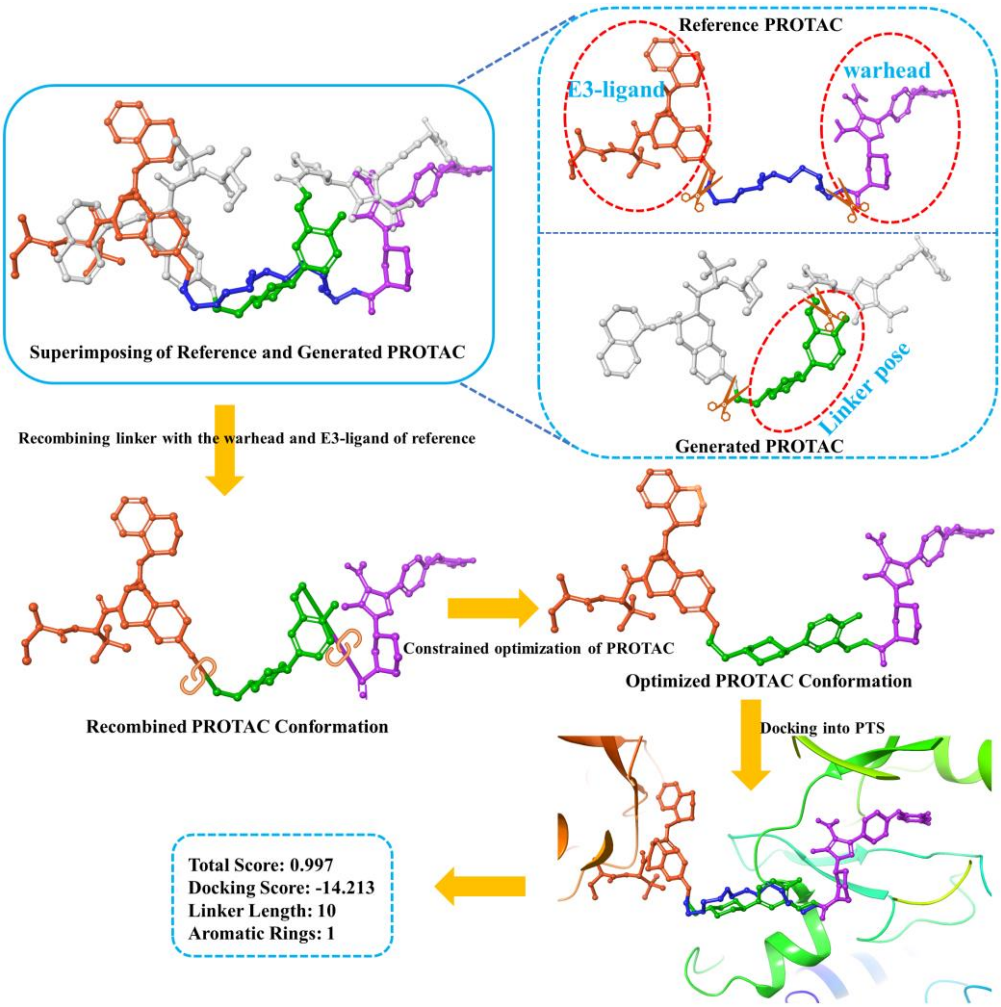


Figure 3 The workflow of generating and scoring 3D binding conformation of PROTAC

Scoring of PROTACs

Molecular generative models such as REINVENT utilize RL to guide structure generation. In RL, compounds are first scored and the scores are then fed back to neural networks to update their parameters to increase the probability of structures with better score. In REINVENT, two types of multi-component scoring function were defined, in which individual components of the scoring function can be either combined as a weighted sum or as a weighted product as shown in Equation 1 and 2. Given a sequence x, the weight coefficient of individual score component $p_i(x)$ reflects its importance in the overall score $S(x)$. For the weighted sum score (Eq. 1), the weight for each component should be a floating number in the [0,1] range and the sum of all weights should be 1.0. For the weighted product score, the overall score is defined as Eq. 2 and the weights can be any floating number.

$$S(x) = \frac{\sum_i w_i \times p_i(x)}{\sum_i w_i} \quad (1)$$

$$S(x) = [\prod_i p_i(x)^{w_i}]^{1/\sum_i w_i} \quad (2)$$

Component of docking score. In current study, several scoring components were used for constructing the over-all score of PROTACs. One component P_d is the docking score of PROTAC and it was transformed to a floating number the [0,1] range according to Eq 3, which is a reversed sigmoid function:

$$P_d(x) = 1/(1 + 10 ** \left(k * v - \frac{high+low}{2}\right) * \frac{10}{high-low})) \quad (3)$$

Here *v* is the docking score (more negative is better). Parameters *low, high* refer to the low end and high end cut-off values of docking score and can be adjusted by user. The parameter *k* is set to 0.25 in current study.

Component of Post-Dock shape similarity. Another component P_s, a value in the range of [0,1], was set to measure the shape similarity between warheads (including the

ligands for E3 ligase and POI) of docking conformation and that of the reference ligand. After "local only" docking of DPC in the binding site of PTS, sometimes there is a large positional shift for the warheads of PROTACs comparing to the warheads of reference PROTAC. The larger the P_S value is, the more similar the warheads of PROTAC is to the reference. Here, *ROCS* was used to only calculate the shape similarity on site and no alignment optimization was done. For example in Figure 4, two PROTAC molecules ***a*** and ***b*** were generated by prior model. After docking procedure, the P_S scores of molecule ***a*** and ***b*** were 0.903 and 0.687 respectively, while their docking scores were roughly equal (ie. -11.37 vs -11.29). Clearly, in the docking poses, the warheads of molecule ***a*** had less deviation to those of reference ligand than molecule ***b***.

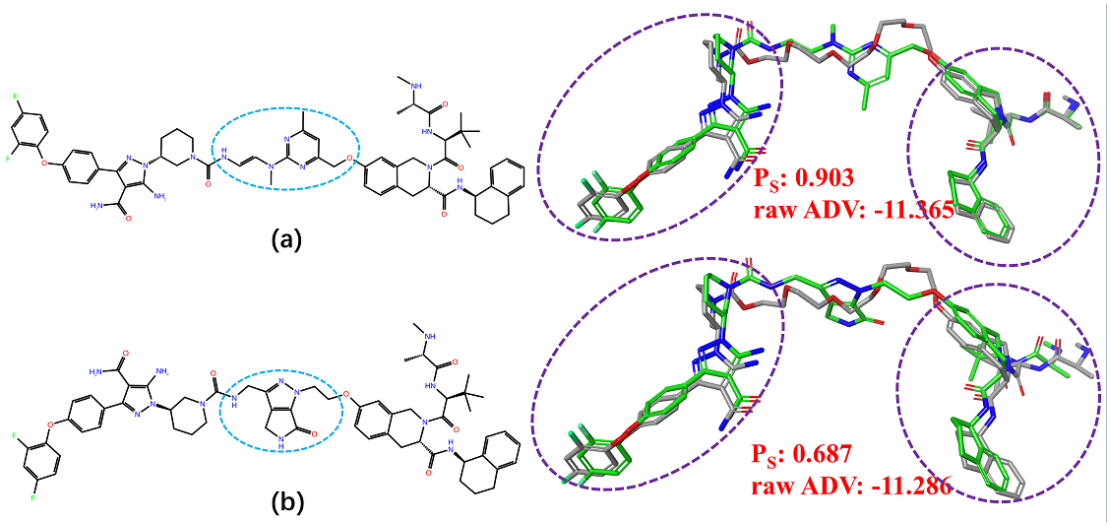


Figure 4 The Ps and raw ADV score of two generated PROTAC examples. The green structures are docking poses and the gray structures are reference poses

Component for substructure alert. To clean the generated structures from generative model, a set of 24 unwanted substructures (as shown in Table S1 in supporting information) was defined. If the generated linker contains these substructures, the alert score is 0, otherwise the score is 1. The alert substructures can be user specified.

Component for linker length. The linker length is critical for maintaining proper conformation of PROTACs in PTS. PROTAC-INVENT can customize the range of linker length of PROTAC through a score component of linker length which is the

shortest 2D bond distance between warheads. If the linker length of generated structure is within the range, the score is 1.0. Otherwise, the score is 0.0.

Component for the number of aromatic ring. Studies have shown[33,34] that multiple aromatic rings in linker of PROTAC molecules could be detrimental to the PK properties of PROTACs. PROTAC-INVENT can customize the number of aromatic ring in the generated linker. Here, the default allowed number of aromatic ring is 1. If the number of aromatic ring of linker is less than or equal to the cut-off value, this score is 1. Otherwise, the score is 0.

Computational details and datasets

REINVENT model was used as the core engine of PROTAC-INVENT to generate linkers in SMILES format. The relevant source code was mainly adapted from Link-INVENT[23], an extension of REINVENT model for linker design by *Guo et al*. Details should be referred to the original literature. In this work, the prior model of Link-INVENT was used as our prior model. This model was originally trained on a dataset processed from ChEMBL database, which comprises 10,313,109 structure triplets which are formed as ("Warhead1 | Warhead2, Linker, Full Molecule"). The processing details should be referred to the original paper[23]. During RL, the training epoch was set to 200 and batch size was set to 100. Other parameters and high-parameters of training process can be found in supporting materials. In current study, Bruton's tyrosine kinase (BTK) PROTAC was chosen as the test case. Published PTS crystal structure 6W8I was chosen as the reference structure for structure generation. For the diversity filter setting[23,25,35,36] in the scoring function of RL, the 'IdenticalMurckoScaffold' option was used to increase the diversity of generated linkers. The "minscore" and "minsimilarity" paremeters were all set as 0.0, so that all generated PROTACs during RL run will be kept.

Result and Discussion

Design of BTK PROTAC through PROTAC-INVENT model

To validate our methodology, design of Bruton's tyrosine kinase (BTK) PROTAC was selected for our case study. BTK is non-receptor tyrosine kinase that plays a role in the maturation of B cells[37]. It represents a validated drug target in leukemia and lymphoma[38–40] with well-described potent and selective ligands[41–43]. Within the protein degrader field, development of BTK PROTAC was reported, owing in part to observations related to its high apparent degradability[44]. *Schiemer et al*[45] recently reported two BTK PTSs (PDB code: 6W8I/6W7O), demonstrating two distinct binding poses between the target and E3 ligase. In one PTS, the linker part is five repeated PEG units and the effective linker length is 15, and, for the other PTS, a pyrazine ring was used instead (the linker effective length is 7).

We took one of the published BTK PTS (6W8I) as the reference structure for linker generation. To explore the effect of the linker length on the generation of PROTAC molecules, we imposed constraint on linker length which refers to the topological bond distance between POI warhead and E3 ligand. Given the published BTK PTSs, we set several cut-offs for linker length, i.e. linker length cut-off [7-9], [7-11], [7-13] and [7-15], and the definition of linker score can be seen in previous section.

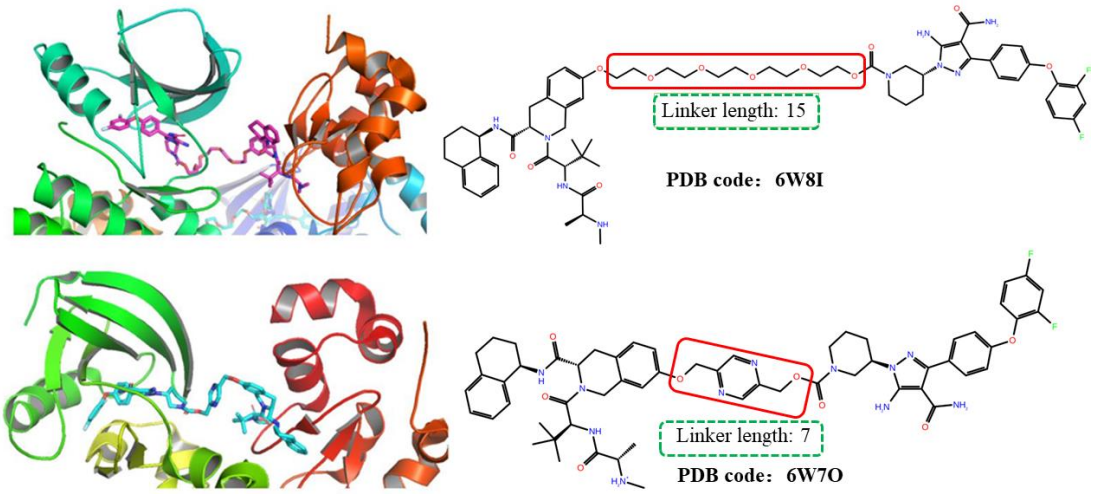
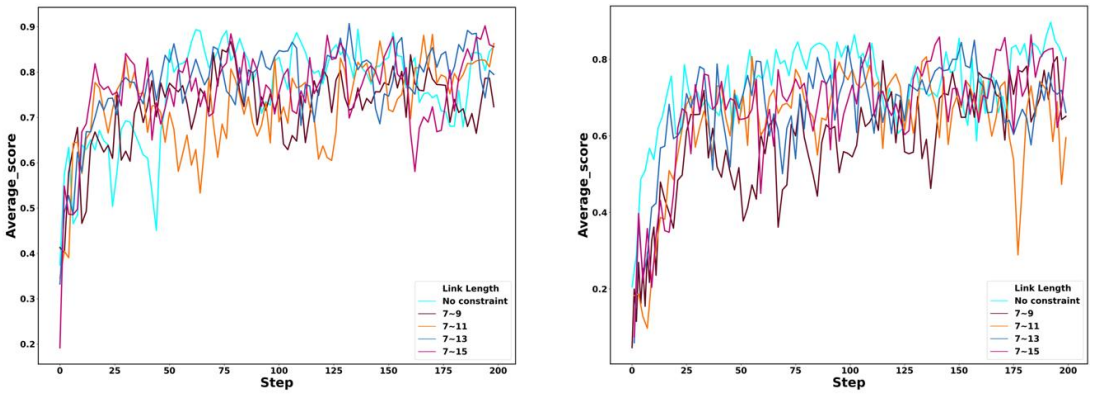


Figure 5 The BTK-degrader-cIAP1 ternary complex crystal structure. The red box of PROTAC molecules represents the linker part and the two terminal fragments are POI warhead and E3-ligand respectively.

Analysis of different aggregation methods

As described in previous section, two aggregation methods, i.e. production and summation form, were used for scoring molecules and their influence on training process were examined. Experiments with different aggregation method under different linker length constraints were carried out and their training curves were shown in Figure 6. When summation score was used, as shown in Figure 6b, during the training of 200 epochs, the curves of average total score were largely fluctuated in all models constrained with linker length, while the curves for production score (as shown in Figure 6a) behaved much better. In most of cases, the total score increased to around 0.8 in 50 epochs and can be stabilized more or less at the level for the rest of training. The results show that the training of production based aggregation method is easier than the summation based method. So in the following experiments, only production scoring models were discussed.



(a). Custom Product: Average Score (b). Custom Sum: Average Score

Figure 6 The training curves based on two aggregation method of score components at different link length.

The changes of average total score using (a) product aggregation and (b) additive aggregation.

As the total score contains multiple components, the weights of component might also influence the training. Among all the score components, ADV score (i.e. the docking score of *Autodock Vina*) was regarded as the most important component for evaluating generated molecules in PTS. We paid more attention on this component and several weight values of ADV score were explored, while the weight values for other components were set as 1.0. The training curves are shown in Figure 7. It can be seen

in Figure 7a that, for all ADV weights, the average total score can quickly converge to 0.8. In the case of ADV weight of 4.0, the fluctuation of total score become larger than other weight level. The actual ADV scores during training can be seen in Figure 7b. It seems that the average docking score quickly reaches 0.7 and fluctuates between 0.7 and 0.9 for the rest of training epochs. There is no clear trend on how increasing weight of docking score would influence the docking score itself. This might due to the fact that slight change of the linker structure could change the docking score of the structure drastically. However, it can be seen that, in general, the total score can be improved during the RL.

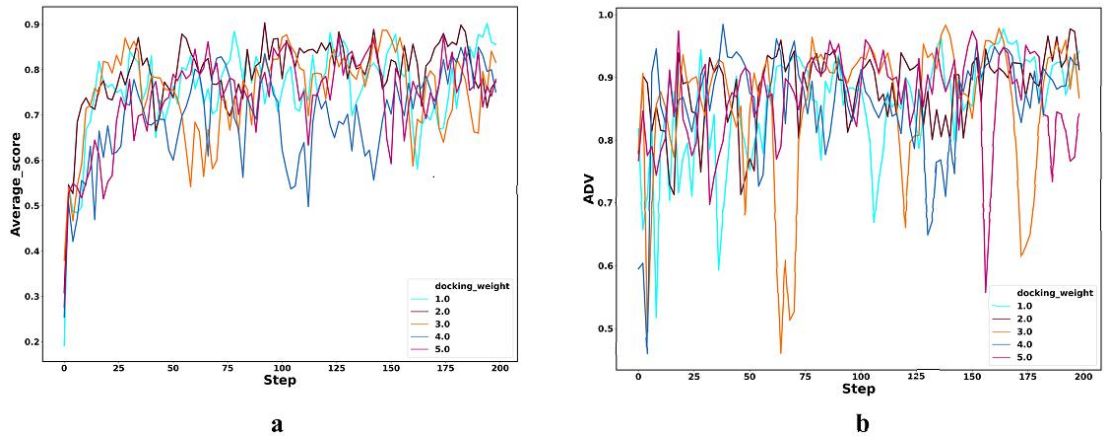


Figure 7 The training curves at different docking weight. The change of (a) average total score and (b) ADV score.

During RL, all generated PROTACs will be saved as potential solutions for later inspection. We analyzed the number of molecules in the collection under different constraints of linker length and the result are shown in Figure 8. In this case, during the RL run, only constraint of link length were varied and all other parameters were the same. In Figure 8, if no constraint of link length was imposed, the link length of generated solutions covers a very wide range (2~35), while adding constraint on linker length, the linker length of generated solutions is in the range of 5~15. As the range of constraint of link length becomes wider, the distribution of linker length of collected solutions also become wider, although it is not dramatically. This suggested that in RL framework, the restraining of link length does work.

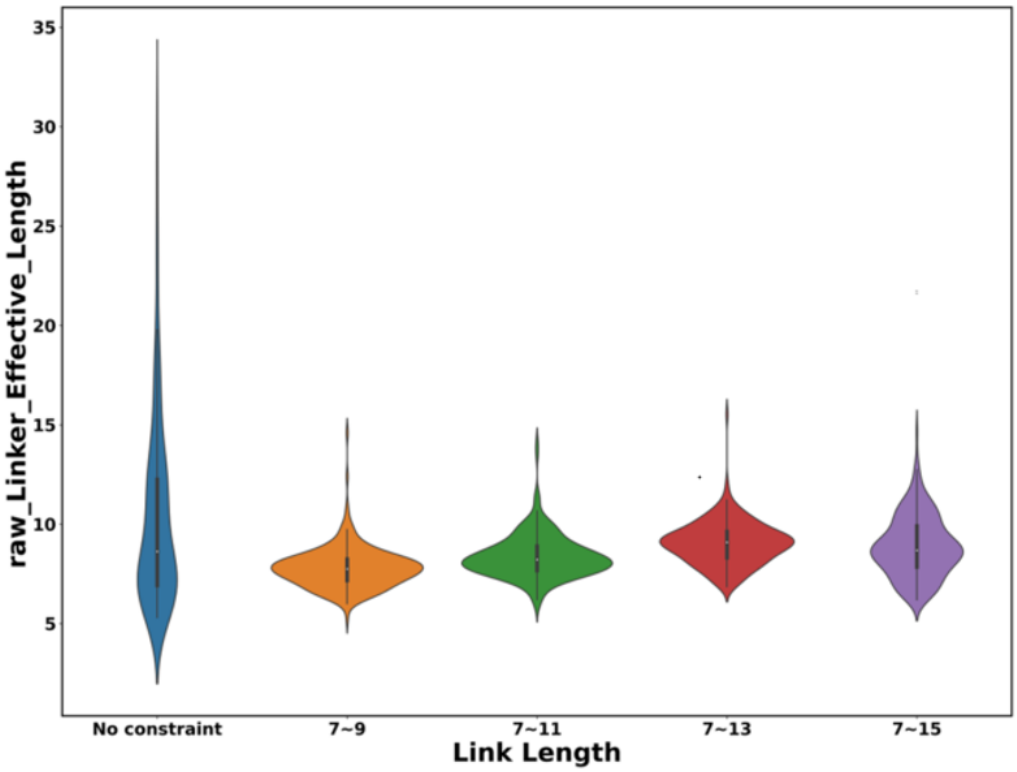


Figure 8 Distribution of linker length of generated PROTACs in various RL runs.

The effect of RL was also examined by looking at ratio of compounds which have good scores as shown in Table 1. As a comparison, without running RL, we randomly sampled the prior model (ie. the model from Link-INVENT) to generate 20K PROTACs as the baseline model and removed duplicates. For the ratio of compound whose total score is higher than 0.8, the "No RL" procedure is roughly 0.51, which is lower than that of other RL procedures. If we look into the ratio of *Vina* docking score (ie. ratio for ADV score < -10.9, which is the docking score of reference PROTAC), a component of the total score, the difference is more significant. For PROTACs generated from the "No RL" procedure, its ratio is about 0.35, while that of other RL procedures are above 0.54. Similar trend was observed for the ratio of P_s> 0.9. This result shows that the RL approach can generated more desirable solutions than sampling from prior model alone without running RL. Some example BTK PROTAC compounds and their potential docking conformations were illustrated in Figure 9.

Table 1 Comparison on the ratio of compound having desirable properties

	Model	Total unique count	Average Score > 0.8	Ratio
Score Component: Average Score				
Link Length	**No Constraint**	17673	12441	**0.704**
	7~9	17869	9692	**0.542**
	7~11	17326	11574	**0.668**
	7~13	18108	12953	**0.715**
	7~15	17714	12164	**0.687**
	No RL	5124	2586	**0.505**
	Model	**Total unique count**	**Docking Score < -10.9**	**Ratio**
Score Component: Docking Score				
Link Length	**No Constraint**	17673	9470	**0.536**
	7~9	17869	13005	**0.728**
	7~11	17326	11831	**0.683**
	7~13	18108	11610	**0.641**
	7~15	17714	11738	**0.663**
	No RL	5124	1782	**0.348**
	Model	**Total unique count**	**P_S > 0.9**	**Ratio**
Score Component: P_S				
Link Length	**No Constraint**	17673	10556	**0.597**
	7~9	17869	12891	**0.721**
	7~11	17326	12705	**0.733**
	7~13	18108	12390	**0.684**
	7~15	17714	12175	**0.687**
	No RL	5124	2554	**0.498**

Note: a) "**No RL**" refers to random sampling from the prior model

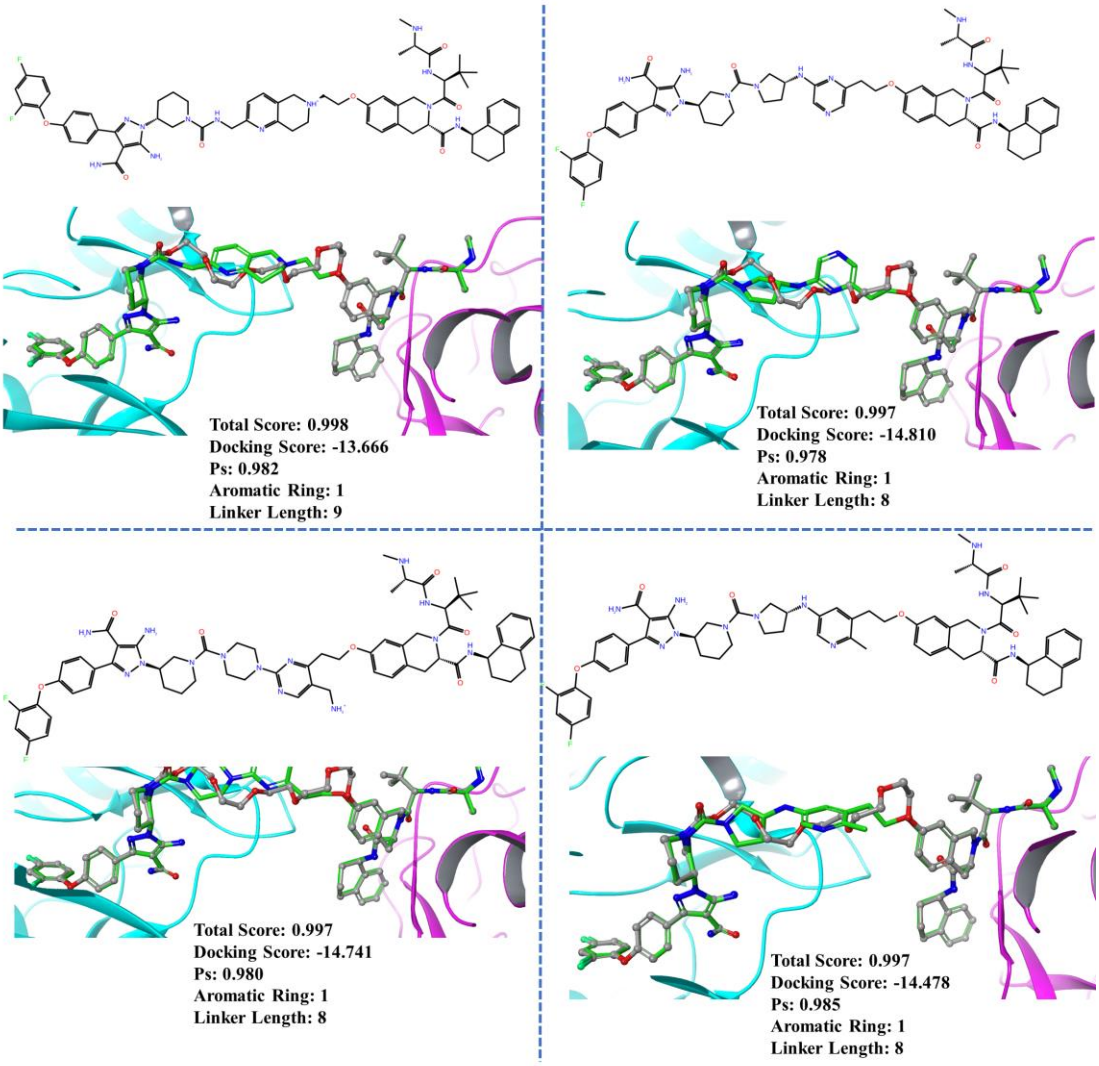


Figure 9 four BTK PROTAC compounds and their docking conformations. The gray molecule is the reference PROTAC conformation, and the green molecule is the docking conformation of generated PROTACs

PROTAC docking protocol

Besides the generation of linker structures, the workflow for generation of PROTAC docking pose can also be used as a practical docking protocol. To the best of our knowledge, there are still no docking software dedicated for PROTAC molecules. Given the large size of PROTAC compound, it would be challenge to properly dock designed PROTAC compounds into the binding site of PTS. The PROTAC docking protocol used in PROTAC-INVENT can quickly generate docking pose of new PROTACs by mimicking the pose of reference PROTAC structure. To evaluate the docking performance of PROTAC-INVENT, we compared PROTAC-INVENT

alongside with two widely used docking programs *Glide*[46] and *Autodock Vina*[47] on re-docking of PROTAC into its respective PTS. Fourteen published PROTAC PTS crystal structures were used as our test set. For *Glide* and *Vina* docking, the initial 3D conformation of 14 PROTAC molecules were converted from SMILES by using the *LigPrep* module of *Schrodinger* and then docked into the complex structure composed by E3 ligase and POI. Initial 3D conformations of PROTAC in the workflow of PROTAC-INVENT were optimized by *ROCS*, which serves the purpose of bringing the generated PROTAC to the vicinity of the reference ligand. As a rule of thumb, initial conformation and starting pose of a compound will certainly affect the subsequent docking process (such as elapsed time, docking score *etc*). For making a fair comparison, we would like to eliminate the effect of the initial conformation on docking results. Therefore, in addition to using *LigPrep* for conformation generation, we also applied *Omega*[26] and *ROCS* software[27] to generate multiple conformations and obtain best aligned 3D conformation as the starting conformation of docking.

The RMSD and elapsed time of re-docking were evaluated and the results can be seen in Figure 10 and Table 2. As shown in Figure 10, for *Glide* docking using *ROCS* results as starting conformation could slightly improve performance, in which, for eight structures, the RMSD of docked conformations of using *ROCS* is better than using *LigPrep* alone. While for *Vina*, no advantages of using *ROCS* in docking protocol was observed. Overall, best docking solution in 12 of 14 structures comes from PROTAC-INVENT, which clearly shows the efficiency of the constrained docking protocol used in PROTAC-INVENT. In terms of computation speed (as shown in Table 2), the docking protocol in PROTAC-INVENT also out-performed all other methods, regardless of using *ROCS* or not in the docking protocol. The elapsed time of our approach is basically less than one minute, while the longest time of *Glide* and *Vina* docking could reach 10.38 and 9.89 minutes respectively.

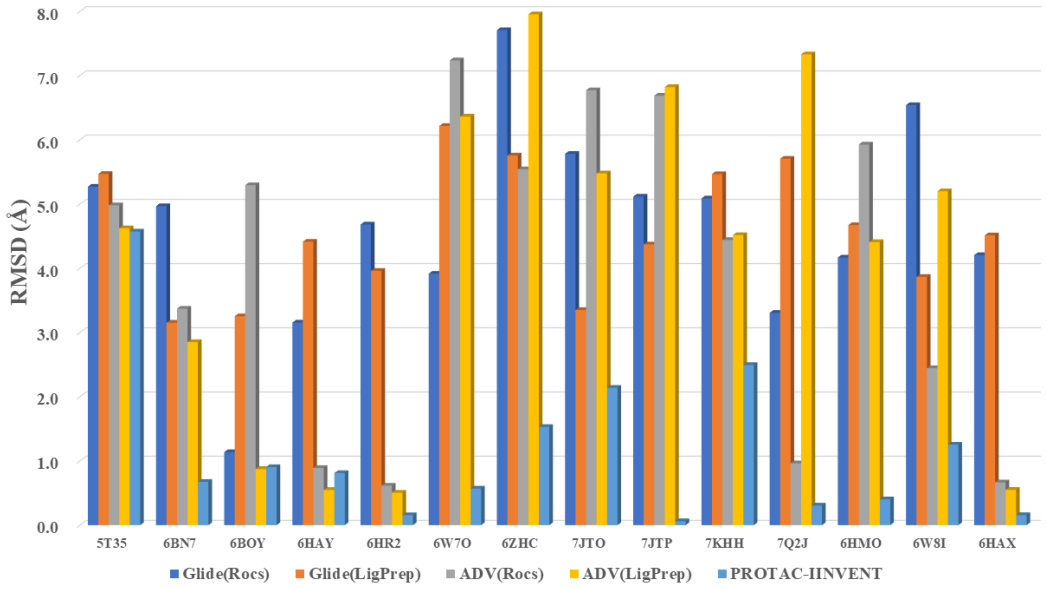


Figure 10 The RMSDs of re-docking PROTAC for various methods.

Table 2 The computation speed for various methods (minute)

PDB	*I*nitial ligand poses from ROCS		Initial ligand poses from LigPrep		PROTAC-INVENT
code	*Glide*(SP)	*ADV*	*Glide*(SP)	*ADV*	
5T35	2.86	3.41	3.07	3.36	0.72
6BN7	1.07	1.42	1.48	1.27	0.69
6BOY	1.16	1.43	0.97	1.70	0.66
6HAY	1.92	2.34	2.28	2.07	0.54
6HR2	1.86	2.15	1.67	1.38	0.59
6W7O	2.61	2.44	3.35	2.13	0.80
6ZHC	8.34	**9.89**	8.05	9.41	0.68
7JTO	5.82	4.88	5.77	4.69	0.70
7JTP	2.25	1.93	2.03	1.48	0.61
7KHH	2.55	3.57	2.78	2.64	0.60
7Q2J	2.92	2.10	3.00	1.84	0.61
6HMO	5.21	6.10	**10.38**	3.81	0.65
6W8I	7.93	5.84	8.27	5.99	0.65
6HAX	1.91	2.02	1.78	1.89	0.55

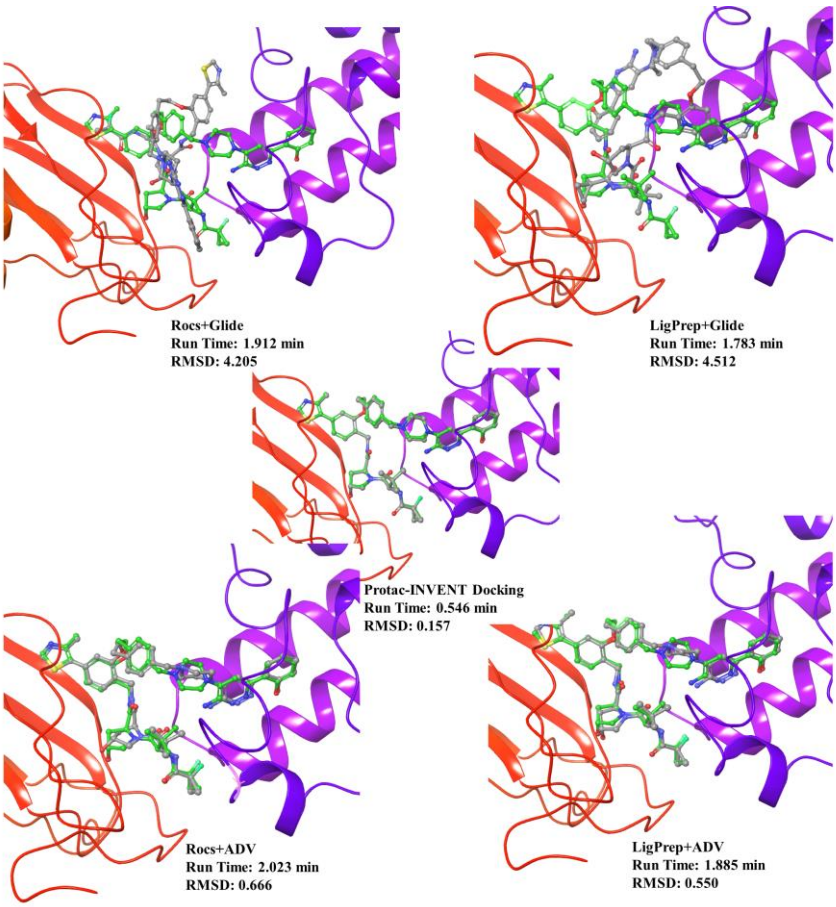


Figure 11 The run time and RMSD of different docking methods in re-docking of structure 6HAX. The gray structures are reference conformation in the crystal structure and the blue ones are docking solutions.

The 6HAX (crystal structure of BAF[48]) system was taken as the example (as shown in Figure 11). The running time of *Glide* and *Vina* were around 2 minutes. While the running time of PROTAC-INVENT is only 0.55 minutes. The RMSD of PROTAC-INVENT is 0.157, while those of Glide and Vina are much higher. These results demonstrate that the constrained docking protocol of PROTAC-INVENT is a better choice for doing PROTAC docking when it was hypothesized that a PROTAC may mimic a known PROTAC binding pose. Our approach performs better in this scenario due to two reasons: Firstly, it adopts a linker conformation to make sure two terminal warheads mimics those of template as close as possible; Secondly, the "local only" mode of *Vina* docking was used to fine tune the initial conformation so that the docking speed was greatly improved while the docking pose of PROTAC molecule can still fit with the binding pocket.

Conclusion

In current work, a novel 3D linker generative model, PROTAC-INVENT, was proposed to rationally design PROTAC molecule. Previous analysis has shown that linker structure can largely affect the degradation efficacy of PROTAC molecules. So far, most of existing linker generation method of PROTAC can only design linker on either 1D SMILES or 2D graph level, which doesn't take into account the 3D information of E3 ligase/POI complex structure. In PROTAC-INVENT model, the protein complex structures were introduced and for the first time, the design of PROTAC linker was done at 3D level in which the 3D binding conformation PROTAC can be generated and the conformation of POI warhead and E3-ligand can be adjusted to the PTS pocket to mimic those of reference structure. The re-docking study of PROTAC-INVENT on known PROTAC crystal structures was also carried out and the performance was compared with Glide and *Vina*, our model achieved best RMSD and computation speed among these methods.

Data and Software Available

The detailed hyper-parameters of all models can be found in supplementary materials. The scripts and model building can be found in the GitHub repository. The source code will be available once the manuscript is accepted for publication,

Author Information

Corresponding Author

Hongming Chen, E-mail: chen_hongming@gzlab.ac.cn

Competing interests

The authors declare that they have no competing interests.

Abbreviations

PTS, PROTAC ternary structure; ADV, *Autodock Vina*; P_S, Post-Dock shape similarity of warheads; Warheads, warhead and E3-ligand; RPC, re-combined PROTAC conformation; DPC, desired PROTAC conformation. RL, Reinforcement Learning.

Rerefences

(1) Sakamoto, K. M.; Kim, K. B.; Kumagai, A.; Mercurio, F.; Crews, C. M.; Deshaies, R. J. Protacs: Chimeric Molecules That Target Proteins to the Skp1-Cullin-F Box Complex for Ubiquitination and Degradation. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98* (15), 8554–8559. https://doi.org/10.1073/pnas.141230798.

(2) Deshaies, R. J. Protein Degradation: Prime Time for PROTACs. *Nature chemical biology* **2015**, *11* (9), 634–635. https://doi.org/10.1038/nchembio.1887.

(3) Dale, B.; Cheng, M.; Park, K. S.; Kaniskan, H. Ü.; Xiong, Y.; Jin, J. Advancing Targeted Protein Degradation for Cancer Therapy. *Nature Reviews Cancer* **2021**, *21* (10), 638–654. https://doi.org/10.1038/s41568-021-00365-x.

(4) Pettersson, M.; Crews, C. M. PROteolysis TArgeting Chimeras (PROTACs) — Past, Present and Future. *Drug Discovery Today: Technologies* **2019**, *31*, 15–27. https://doi.org/10.1016/j.ddtec.2019.01.002.

(5) Lai, A. C.; Crews, C. M. Induced Protein Degradation: An Emerging Drug Discovery Paradigm. *Nature Reviews Drug Discovery* **2017**, *16* (2), 101–114. https://doi.org/10.1038/nrd.2016.211.

(6) Bai, L.; Zhou, H.; Xu, R.; Zhao, Y.; Chinnaswamy, K.; McEachern, D.; Chen, J.; Yang, C.-Y.; Liu, Z.; Wang, M.; Liu, L.; Jiang, H.; Wen, B.; Kumar, P.; Meagher, J. L.; Sun, D.; Stuckey, J. A.; Wang, S. A Potent and Selective Small-Molecule Degrader of STAT3 Achieves Complete Tumor Regression In Vivo. *Cancer cell* **2019**, *36* (5), 498-511.e17. https://doi.org/10.1016/j.ccell.2019.10.002.

(7) Tunjic, T. M.; Weber, N.; Brunsteiner, M. Computer Aided Drug Design in the Development of Proteolysis Targeting Chimeras. *Computational and Structural*

Biotechnology Journal **2023**, *21*, 2058–2067. https://doi.org/10.1016/j.csbj.2023.02.042.

(8) Liu, Z.; Hu, X.; Wang, Q.; Wu, X.; Zhang, Q.; Wei, W.; Su, X.; He, H.; Zhou, S.; Hu, R.; Ye, T.; Zhu, Y.; Wang, N.; Yu, L. Design and Synthesis of EZH2-Based PROTACs to Degrade the PRC2 Complex for Targeting the Noncatalytic Activity of EZH2. *Journal of Medicinal Chemistry* **2021**, *64* (5), 2829–2848. https://doi.org/10.1021/acs.jmedchem.0c02234.

(9) Weng, G.; Cai, X.; Cao, D.; Du, H.; Shen, C.; Deng, Y.; He, Q.; Yang, B.; Li, D.; Hou, T. PROTAC-DB 2.0: An Updated Database of PROTACs. *Nucleic acids research* **2023**, *51* (D1), D1367–D1372. https://doi.org/10.1093/nar/gkac946.

(10) Smith, B. E.; Wang, S. L.; Jaime-Figueroa, S.; Harbin, A.; Wang, J.; Hamman, B. D.; Crews, C. M. Differential PROTAC Substrate Specificity Dictated by Orientation of Recruited E3 Ligase. *Nature Communications* **2019**, *10* (1), 1–13. https://doi.org/10.1038/s41467-018-08027-7.

(11) Bemis, T. A.; Clair, J. J. La; Burkart, M. D. Unraveling the Role of Linker Design in Proteolysis Targeting Chimeras. *Journal of Medicinal Chemistry* **2021**, *64* (12), 8042–8052. https://doi.org/10.1021/acs.jmedchem.1c00482.

(12) Mohler, M. L.; Sikdar, A.; Ponnusamy, S.; Hwang, D.-J.; He, Y.; Miller, D. D.; Narayanan, R. An Overview of Next-Generation Androgen Receptor-Targeted Therapeutics in Development for the Treatment of Prostate Cancer. *International journal of molecular sciences* **2021**, *22* (4). https://doi.org/10.3390/ijms22042124.

(13) Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *Journal of Chemical Information and Modeling* **2020**, *60* (1), 77–91. https://doi.org/10.1021/acs.jcim.9b00727.

(14) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective de Novo Drug Design with Conditional Graph Generative Model. *Journal of Cheminformatics* **2018**, *10* (1), 33. https://doi.org/10.1186/s13321-018-0287-6.

(15) Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *Journal of Chemical Information and Modeling* **2019**, *59* (3), 1205–1214. https://doi.org/10.1021/acs.jcim.8b00706.

(16) Sturm, N.; Sun, J.; Vandriessche, Y.; Mayr, A.; Klambauer, G.; Carlsson, L.; Engkvist, O.; Chen, H. Application of Bioactivity Profile-Based Fingerprints for Building Machine Learning Models. *Journal of Chemical Information and Modeling* **2019**, *59* (3), 962–972. https://doi.org/10.1021/acs.jcim.8b00550.

(17) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular Optimization by Capturing Chemist's Intuition Using Deep Neural Networks. *Journal of Cheminformatics* **2021**, *13* (1), 1–17. https://doi.org/10.1186/s13321-021-00497-0.

(18) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J. L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *Journal of Cheminformatics* **2019**, *11* (1), 1–14. https://doi.org/10.1186/s13321-019-0341-z.

(19) Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H. SyntaLinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chemical Science* **2020**, *11* (31), 8312–8322. https://doi.org/10.1039/d0sc03126g.

(20) Zheng, S.; Tan, Y.; Wang, Z.; Li, C.; Zhang, Z.; Sang, X.; Chen, H.; Yang, Y. Accelerated Rational PROTAC Design via Deep Learning and Molecular Simulations. *Nature Machine Intelligence* **2022**, *4* (9), 739–748. https://doi.org/10.1038/s42256-022-00527-y.

(21) Imrie, F.; Bradley, A. R.; Van Der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *Journal of Chemical Information and Modeling* **2020**, *60* (4), 1983–1995. https://doi.org/10.1021/acs.jcim.9b01120.

(22) Igashov, I.; Stärk, H.; Vignac, C.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3D-Conditional Diffusion Models for Molecular Linker Design. **2022**, 1–21.

(23) Guo, J.; Knuth, F.; Margreitter, C.; Janet, J. P.; Papadopoulos, K. Link-

INVENT : Generative Linker Design with Reinforcement Learning.

(24) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews* **2001**, *46* (1–3), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(25) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for de Novo Drug Design. *Journal of Chemical Information and Modeling* **2020**, *60* (12), 5918–5922. https://doi.org/10.1021/acs.jcim.0c00915.

(26) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2010**, *50* (4), 572–584. https://doi.org/10.1021/ci100031x.

(27) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *Journal of Medicinal Chemistry* **2007**, *50* (1), 74–82. https://doi.org/10.1021/jm0603365.

(28) Schrödinger Release 2022-4. *MacroModel, Schrödinger, LLC, New York, NY, 2021*.

(29) Guo, J.; Janet, J. P.; Bauer, M. R.; Nittinger, E.; Giblin, K. A.; Papadopoulos, K.; Voronov, A.; Patronov, A.; Engkvist, O.; Margreitter, C. DockStream: A Docking Wrapper to Enhance de Novo Molecular Design. *Journal of Cheminformatics* **2021**, *13* (1), 1–21. https://doi.org/10.1186/s13321-021-00563-7.

(30) Tang, S.; Chen, R.; Lin, M.; Lin, Q.; Zhu, Y. Accelerating AutoDock VINA with GPUs. *ChemRxiv* **2022**, 1–32.

(31) Huey, R.; Morris, G. M.; Forli, S. Using AutoDock 4 and AutoDock Vina with AutoDockTools: A Tutorial. **2012**, 32.

(32) Morris, G. M.; Ruth, H.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Software News and Updates AutoDock4 and

AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry* **2009**. https://doi.org/10.1002/jcc.21256.

(33) Pike, A.; Williamson, B.; Harlfinger, S.; Martin, S.; McGinnity, D. F. Optimising Proteolysis-Targeting Chimeras (PROTACs) for Oral Drug Delivery: A Drug Metabolism and Pharmacokinetics Perspective. *Drug Discovery Today* **2020**, *25* (10), 1793–1800. https://doi.org/10.1016/j.drudis.2020.07.013.

(34) Cantrill, C.; Chaturvedi, P.; Rynn, C.; Petrig Schaffland, J.; Walter, I.; Wittwer, M. B. Fundamental Aspects of DMPK Optimization of Targeted Protein Degraders. *Drug Discovery Today* **2020**, *25* (6), 969–982. https://doi.org/10.1016/j.drudis.2020.03.012.

(35) Cummins, D. J.; Bell, M. A. Integrating Everything: The Molecule Selection Toolkit, a System for Compound Prioritization in Drug Discovery. *Journal of Medicinal Chemistry* **2016**, *59* (15), 6999–7010. https://doi.org/10.1021/acs.jmedchem.5b01338.

(36) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-Assisted Reinforcement Learning for Diverse Molecular de Novo Design. *Journal of Cheminformatics* **2020**, *12* (1), 1–17. https://doi.org/10.1186/s13321-020-00473-0.

(37) Hendriks, R. W.; Yuvaraj, S.; Kil, L. P. Targeting Bruton's Tyrosine Kinase in B Cell Malignancies. *Nature Reviews Cancer* **2014**, *14* (4), 219–232. https://doi.org/10.1038/nrc3702.

(38) Dunleavy, K.; Erdmann, T.; Lenz, G. Targeting the B-Cell Receptor Pathway in Diffuse Large B-Cell Lymphoma. *Cancer Treatment Reviews* **2018**, *65*, 41–46. https://doi.org/10.1016/j.ctrv.2018.01.002.

(39) Byrd, J. C.; Furman, R. R.; Coutre, S. E.; Flinn, I. W.; Burger, J. A.; Blum, K. A.; Grant, B.; Sharman, J. P.; Coleman, M.; Wierda, W. G.; Jones, J. A.; Zhao, W.; Heerema, N. A.; Johnson, A. J.; Sukbuntherng, J.; Chang, B. Y.; Clow, F.; Hedrick, E.; Buggy, J. J.; James, D. F.; O'Brien, S. Targeting BTK with Ibrutinib in Relapsed Chronic Lymphocytic Leukemia. *New England Journal*

of Medicine **2013**, *369* (1), 32–42. https://doi.org/10.1056/NEJMoa1215637.

(40) Wang, M. L.; Rule, S.; Martin, P.; Goy, A.; Auer, R.; Kahl, B. S.; Jurczak, W.; Advani, R. H.; Romaguera, J. E.; Williams, M. E.; Barrientos, J. C.; Chmielowska, E.; Radford, J.; Stilgenbauer, S.; Dreyling, M.; Jedrzejczak, W. W.; Johnson, P.; Spurgeon, S. E.; Li, L.; Zhang, L.; Newberry, K.; Ou, Z.; Cheng, N.; Fang, B.; McGreivy, J.; Clow, F.; Buggy, J. J.; Chang, B. Y.; Beaupre, D. M.; Kunkel, L. A.; Blum, K. A. Targeting BTK with Ibrutinib in Relapsed or Refractory Mantle-Cell Lymphoma. *New England Journal of Medicine* **2013**, *369* (6), 507–516. https://doi.org/10.1056/NEJMoa1306220.

(41) Qiu, H.; Liu-Bujalski, L.; Caldwell, R. D.; Follis, A. V.; Gardberg, A.; Goutopoulos, A.; Grenningloh, R.; Head, J.; Johnson, T.; Jones, R.; Mochalkin, I.; Morandi, F.; Neagu, C.; Sherer, B. Discovery of Potent, Highly Selective Covalent Irreversible BTK Inhibitors from a Fragment Hit. *Bioorganic and Medicinal Chemistry Letters* **2018**, *28* (17), 2939–2944. https://doi.org/10.1016/j.bmcl.2018.07.008.

(42) Pan, Z.; Scheerens, H.; Li, J.; Schultz, B. E.; Sprengeler, P. A.; Burrill, L. C.; Mendonca, R. V; Sweeney, M. D.; Scott, K. C. K.; Grothaus, P. G.; Jeffery, D. A.; Spoerke, J. M.; Honigberg, L. A.; Young, P. R.; Dalrymple, S. A.; Palmer, J. T. Discovery of Selective Irreversible Inhibitors for Bruton's Tyrosine Kinase. **2007**, 58–61. https://doi.org/10.1002/cmdc.200600221.

(43) Wu, H.; Huang, Q.; Qi, Z.; Chen, Y.; Wang, A.; Chen, C.; Liang, Q.; Wang, J.; Chen, W.; Dong, J.; Yu, K.; Hu, C.; Wang, W.; Liu, X.; Deng, Y.; Wang, L.; Wang, B.; Li, X.; Gray, N. S.; Liu, J.; Wei, W.; Liu, Q. Irreversible Inhibition of BTK Kinase by a Novel Highly Selective Inhibitor CHMFL-BTK-11 Suppresses Inflammatory Response in Rheumatoid Arthritis Model. *Scientific Reports* **2017**, *7* (1), 1–10. https://doi.org/10.1038/s41598-017-00482-4.

(44) Huang, H.-T.; Dobrovolsky, D.; Paulk, J.; Yang, G.; Weisberg, E. L.; Doctor, Z. M.; Buckley, D. L.; Cho, J.-H.; Ko, E.; Jang, J.; Shi, K.; Choi, H. G.; Griffin, J. D.; Li, Y.; Treon, S. P.; Fischer, E. S.; Bradner, J. E.; Tan, L.; Gray, N. S. A Chemoproteomic Approach to Query the Degradable Kinome Using a

Multi-Kinase Degrader. *Cell Chemical Biology* **2018**, *25* (1), 88-99.e6. https://doi.org/10.1016/j.chembiol.2017.10.005.

(45) Schiemer, J.; Horst, R.; Meng, Y.; Montgomery, J. I.; Xu, Y.; Feng, X.; Borzilleri, K.; Uccello, D. P.; Leverett, C.; Brown, S.; Che, Y.; Brown, M. F.; Hayward, M. M.; Gilbert, A. M.; Noe, M. C.; Calabrese, M. F. Snapshots and Ensembles of BTK and CIAP1 Protein Degrader Ternary Complexes. *Nature Chemical Biology* **2021**, *17* (2), 152–160. https://doi.org/10.1038/s41589-020-00686-2.

(46) Friesner, R. a; Banks, J. L.; Murphy, R. B.; Halgren, T. a; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of medicinal chemistry* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(47) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2009**, *29* (6), NA-NA. https://doi.org/10.1002/jcc.21334.

(48) Farnaby, W.; Koegl, M.; Roy, M. J.; Whitworth, C.; Diers, E.; Trainor, N.; Zollman, D.; Steurer, S.; Karolyi-Oezguer, J.; Riedmueller, C.; Gmaschitz, T.; Wachter, J.; Dank, C.; Galant, M.; Sharps, B.; Rumpel, K.; Traxler, E.; Gerstberger, T.; Schnitzer, R.; Petermann, O.; Greb, P.; Weinstabl, H.; Bader, G.; Zoephel, A.; Weiss-Puxbaum, A.; Ehrenhöfer-Wölfer, K.; Wöhrle, S.; Boehmelt, G.; Rinnenthal, J.; Arnhof, H.; Wiechens, N.; Wu, M.-Y.; Owen-Hughes, T.; Ettmayer, P.; Pearson, M.; McConnell, D. B.; Ciulli, A. BAF Complex Vulnerabilities in Cancer Demonstrated via Structure-Based PROTAC Design. *Nature Chemical Biology* **2019**, *15* (7), 672–680. https://doi.org/10.1038/s41589-019-0294-6.