Fragment Linker Prediction Using Deep Encoder-Decoder Network for PROTAC Drug Design

Chien-Ting Kao,[†] Chieh-Te Lin,[‡] Cheng-Li Chou,[†] and Chu-Chung Lin[*,†]

[†]*AnHorn Medicines Co., Ltd., Taipei, Taiwan*

[‡]*Department of Biomedical Engineering, University of California Davis, Davis, CA, USA*

E-mail: cclin@anhornmed.com

Abstract

Drug discovery and development pipeline is a prolonged and complex process and remains challenging for both computational methods and medicinal chemists. Deep learning has shed lights in various fields and achieved tremendous success in designing novel molecules in pharmaceutical industry. We utilize state-of-the-art techniques to propose a deep neural network for rapid designing and generating meaningful drug-like Proteolysis Targeting Chimeras (PROTAC) analogs. Our method, AIMLinker, takes the structural information from the corresponding fragments and generates linkers to incorporate them. In this model, we integrate filters for excluding non-druggable structures guided by protein-protein complexes while retaining molecules with potent chemical properties. The novel PROTACs subsequently pass through molecular docking, taking root-mean square deviation (RMSD), the change of Gibbs free energy (ΔG), and the "Rule of Three" as the measurement criteria for testing the robustness and feasibility of the model. The generated novel PROTAC molecules possess similar structural information with superior binding affinity to the binding pockets in comparison to existing CRBN-dBET6-BRD4 ternary complexes. We demonstrate the effectiveness of AIMLinker having the power to design compounds for PROTAC molecules with better chemical properties.

Keywords

Drug design, PROTACs, Deep learning, Ligand binding

Introduction

Drug target design is an iterative process with accounts of binding affinity, pharmacokinetics, and molecular structures to involve multiple cycles before optimizing a lead drug for trials.[1] Current challenge in structure-based drug design remains due to the size constraint of the search space and the logical drug-like molecules for chemical synthesis.[2] Kick et al.[3] demonstrates the ability of coupling the complementary methods of combinatorial chemistry and structure-based design in a nanomolar range. Structure-based design also directs the discovery of a drug lead, which is not a drug product but, specifically, a compound with at least micromolar affinity for a target.[4] Another challenge lies in the discontinuous of the molecule landscape and the size of the search space, affecting the binding pocket affinity and selection.[5,6] A high-throughput screening according to binding site and ligand binding affinity to structurally optimize the lead compounds and demonstrates the varying affinity from nanomolar to micromolar is needed.[7] These screening and prediction process are computational expensive and was highly manipulated by experts to achieve the best optimization and development. Therefore, designing novel molecule for drug has been a prolonged process, in particular, a human-influenced task by experience.

Current studies have leveraged the aid of rapid simulation and state-of-the-art deep learning for discovering novel structures, demonstrating the feasibility in rapid screening to the potential targets.[8] There are many studies have been done to utilize DL methods, such as multi-task neural network, and ligand-based virtual screening. Multi-task Dense Neural Network (DNN) has been proven to be an effective way to boost inferring the properties and activities of small molecules.[9] With the limited data, the one-shot learning technique can substantially reduce the amount of data required to make meaningful predictions of designing a new molecule in a new experimental setup. Aliper et al.[10] built DNN models for predicting pharmacological properties of drugs and for drug repurposing transcriptomic data. The benefits of using multi-task models over single-task models are, however, dependent to certain datasets. To address this, a benchmark machine learning (ML) algorithm from Wu et al.,[11] complied a large benchmark dataset, MoleculeNet. It was used for comparing between different ML algorithms, containing data on the properties of over 700,000 compounds, and the datasets have been integrated into the open-source DeepChem package.

Graph Neural Network (GNN) is another technique gaining attention in drug discovery as it automatically learns task-specific representations using graph convolutions that the graph information is conserved as the atom-bond interactions.[12,13] GNN learns the representations of each atom by aggregating the information from its surrounding atoms encoded by the atom feature vector, and recursively encode the connected bond feature vector through the message passing across the molecular graph, followed by a read-out operation that forms corresponding atoms and bonds.[14–16] The state-of-the-art GNN models in predicting properties has been well demonstrated, and typically superior or comparable to traditional descriptor-based models.[17,18] As such, GNN has been proven to be a potential model for designing and generating novel structures for drug discovery and investigating drug-like candidates. In the study from Wu et al.[11] again showed evaluation results that GNN outperformed on most dataset, giving the network the feasibility of predicting various chemical endpoints.

Inducing degradation of target proteins by bifunctional small molecules, which is also known as proteolysis-targeting chimeras (PROTACs), is one of the newly drawing attention modalities. PROTACs consist of ligand for recruiting a target protein of interest (POI) and a ligand for an ubiquitin–protein ligases (E3), joint with an appropriate linker that further induces degradation of an target protein.[19] Compared with classical inhibition by small molecules, PROTACs show potential advantages: 1) PROTACs are expected to exert similar phenotypes via knockdowns using genetic tools, such as small interfering RNA (siRNA), short hairpin RNA (shRNA), or clustered regularly interspaced short palindromic repeats (CRISPR), where the downstream of these are the same to deplete the intercellular protein levels.[20] Eliminating POI gives additional effect to disrupt the formation of biologically functional complexes, 2) PROTACs can work catalytically, i.e., they can be recycled for degrading same POI with the same molecule. They can show more binding affinity to POI alone in comparison to small molecule drugs,[21] 3) the degradation by PROTACs can suppress the target protein from mutation and/or upregulation.[22] Multiple E3 ligases have been targeted for PROTACs development,[23] however, here we focus on PROTACs of the substrate recognition protein ($CRL4^{CRBN}$)[20] ring in E3 ligase. The degradation interaction exhibits the traits of second point above that Bromodomain-containing protein 4 (BRD4) degraders, dBET23, recruit $CRL4^{CRBN}$. Moreover, the potency and isozyme selectivity of PROTACs can be optimized through structure-activity relationships (SARs) with the linkers. The length and chemical property has been proven to affect the structural rigidity, hydrophobicity, and solubility of PORTACs molecules.[24,25] While studies have been made toward rational PROTACs design through structural biological and computational studies, linker design and generation still present a significant synthetic burden.

Recent studies have shown advances in deep generative models to demonstrate the structure generation in drug design.[12,26,27] Graph-based variational autoencoder (GVAE) is one of the most famous networks for molecular generation and design. Many approaches use 2D SMILES-based chemical graphs embedded in low dimensional space, and generate new molecule by perturbing the hidden values of the sampled atoms.[28–30] These studies are missing the nature of the molecular shape and the 3D information, which may greatly differ from the starting point of structure design. Another recent popular deep neural network drug design is in fragment linking technique. DeLinker,[31] adapted from Liu et al.,[29] is the first attempt to apply GNN in linker design, particularly retaining the 3D structural information and generating linkers by giving two input fragments. 3DLinker[32] puts the step forward with predicting the fragment nodes and sampling linker molecules simultaneously. However, none of these works have demonstrated an effective methodology of refining the generated molecule nor further considering validation in molecular conformations. The integration pipeline of adapting deep neural networks as the core technique in drug discovery and substantial validation process are still lacking in investigation.

In this work, we propose a novel deep learning based neural network, Artificial Intelligent Molecule Linker (AIMLinker). This network integrates designing, generating, and screening novel small molecular structures for PROTAC linkers, demonstrating a highly effective methodology of creating neo-structures to address the current difficulties in drug discovery. Our network considers the structural 3D information that first takes two fragments, with pre-defined anchors on both sides, and its structural information of angle and distance to represent the spatial positions between the input fragments. The core architecture of the network is gated graph neural network (GGNN)[33] with atoms and bonds denoting as nodes and edges, respectively. In addition, the iterative process of adding atoms and forming bonds is repeated until termination, followed by a read-out step for returning a 2D structural compound and subsequently screening with the postprocess step. Outputs from the network are docked back to CRBN-BRD4 complex through AutoDock4 and validated with the measurement of the root-mean-square deviation (RMSD), the change of Gibbs free energy (ΔG), and the "Rule of Three" criteria to test the robustness and feasibility as a drug-like molecule. This end-to-end pipeline demonstrates a novel methodology for using state-of-the-art deep learning technique for drug discovery and shows the viability of designing novel PROTACs linker molecules.

Methods

In this section, we first provide the details on processing the PROTAC structure selected for our encoder-decoder network. Next, we present the network architecture and designs to generate the linker molecular structure with good viability and reasonable for drug synthesis. The postprocessing algorithm is provided for validating the predicted molecules and conserving drug-like PROTAC molecules. Finally, the robustness of our predicted molecules is evaluated by the docking tool and the calculation of binding energy. The overall pipeline of our study is provided in Figure 1.

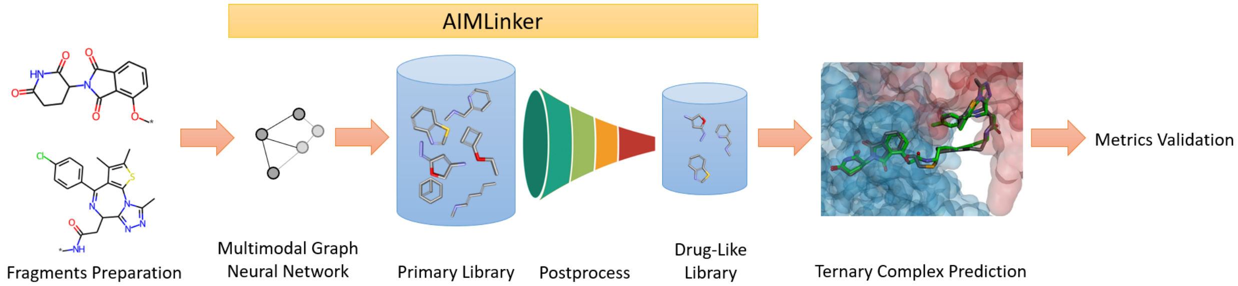


Figure 1: Scheme of the pipeline. The starting input data is two fragments, which are preprocessed and having their relative structural information. Next, AIMLinker takes the input fragments and generates the linker to form a unit compound. The molecules are postprocessed with our algorithm. We then take the best four molecules to validate their robustness of recognition as drug-like molecules.

The study from Nowak[34] showed that dBET6 exhibited one of the most potent PROTACs property on BD1 domain of BRD4 protein levels. For validating our proposed pipeline, we choose this famous and proven effective PROTAC molecule as the testing target. The following details demonstrate our workflow executing on the crystal structure of CRBN-BRD4 complex bounded with dBET6.

Data Processing

The plastic binding between the ligase and the substrate adapts distinct conformations depending on the linker length and position. Nowak et al.[34] showed dBET6, a PROTAC molecule, exhibiting high selectivity property with the structure. Its number of possible accessible binding conformations is proved to be limited with a relatively short linker compound. The integration of structural, biochemical, and cellular properties of CRBN-dBET6-BRD4 ternary complex are designed to be a neo-degrader-mediated PROTAC structure. The crystal structure of DDB1-CRBN-BRD4 complex bounding to dBET6 is available in Protein Data Bank (PDB)[35] (pdb: 6BOY). Figure 2A illustrates the relative spatial pose of CRBN-dBET6-BRD4 ternary complex via Discovery Studio Visualizer (DSV)[36] that red labeled and blue labeled structures are BRD4 and CRBN, respectively. dBET6 molecule is bounding between the two proteins in their corresponding binding pockets.

The input data of the network are two fragments composed of two ligands extracted from the PROTAC excluding the linear linker structure. To prepare the input data from dBET6, we first retrieve the PROTAC molecule of 6BOY structure. With consideration of a potent inhibitor for BRD4 examined by Filippakopoulos et al.,[37] the fragment of BRD4 ligand is defined as illustration in Figure 2B; conversely, the CRBN ligand is defined as a pomalidomide-like structure. The linking anchors on each ligand are labeled with $R*$, and the linker between these two anchors is removed. Since the co-crystal structure retrieving from PDB is spatially predefined and fixed, the anchors provide the relative spatial angle and distance information between the two fragments. The network further takes the two fragments and the corresponding spatial information as the input to generate and design a linker library with the constraint of the space between the anchors.

Multimodal Encoder-Decoder Network

We propose a novel network, AIMLinker, generating and designing novel structural linkers between fragments and postprocessing the predicted structures. The network is inspired by Imrie et al.[31] and Liu et al.,[29] taking two fragments, where containing their relative spatial position and orientation information, to generate the linker structures binding to the anchors on both fragments and form a novel molecular structure. This process is achieved by edge-to-edge generation in a breadth-first manner with iteratively adding or replacing the atoms

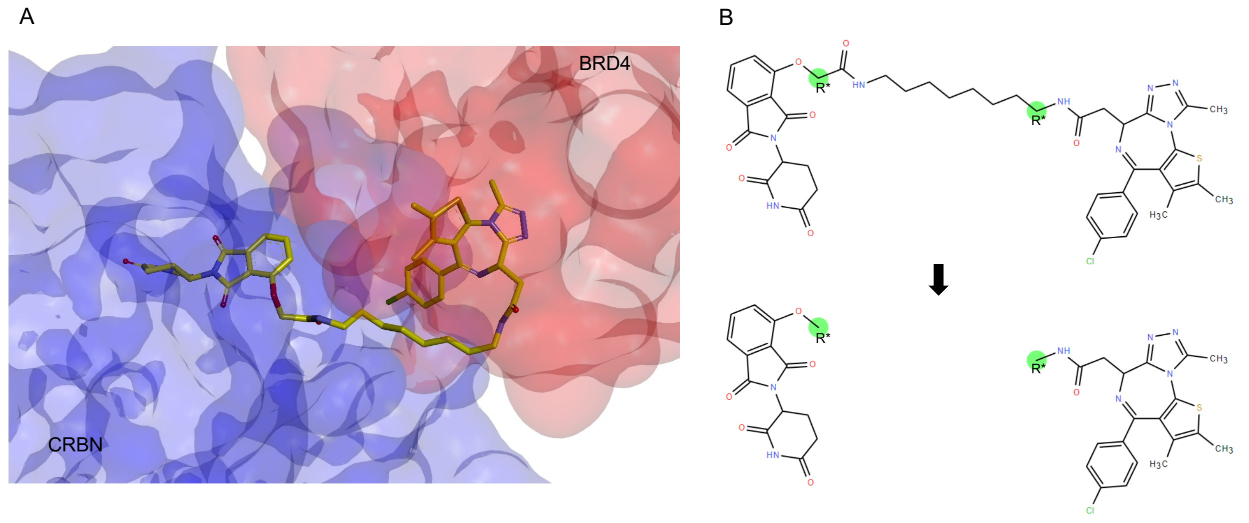


Figure 2: Scheme of the CRBN-dBET6-BRD4 ternary structure and processing protocol of dBET6. (A) The structure and the positional information of 6BOY with the dBET6 binding to BRD4 and CRBN.[34] The red and blue label protein represent BRD4 and CRBN, respectively. (B) The 2D illustration of the dBET6 molecule. The anchors are highlighted and labeled with $R*$ to indicate the network with the start/end positions of the generated linkers. Next, the molecule between the anchors are removed and the remaining two fragments are considered as the input data for the network.

from the selected pool, specifically 14 permitted atom types. In addition, the network allows users to define the number of atoms between the anchors for maximizing the variations in generating the new linker molecules and provide the validity size of the two fragments corresponding to their distances. The other selection is the number of molecules to be generated that the network includes a postprocessing step for removing the molecules not subject to basic chemistry rule, duplicates, and illogical structures.

Figure 3 illustrates the iteration process, where the network uses an encoder network, a standard gated graph neural network (GGNN), and the hidden states of the nodes are updated to incorporate their local environment. The angles and distances information of two fragments prepared from the data processing step are further calculated by the selection of the anchors. After AIMLinker initialization, the fragments are converted into graph representation, where atoms and bonds are respectively denoting as nodes and edges. Each node is associated with a hidden state z_v, and labeled with l_v to represent the atom type of the node. The graph information is passed through AIMLinker, which utilizes GGNN as the core encoder structure, and the hidden state of nodes and edges are updated to integrate the learning process (Figure 3a).

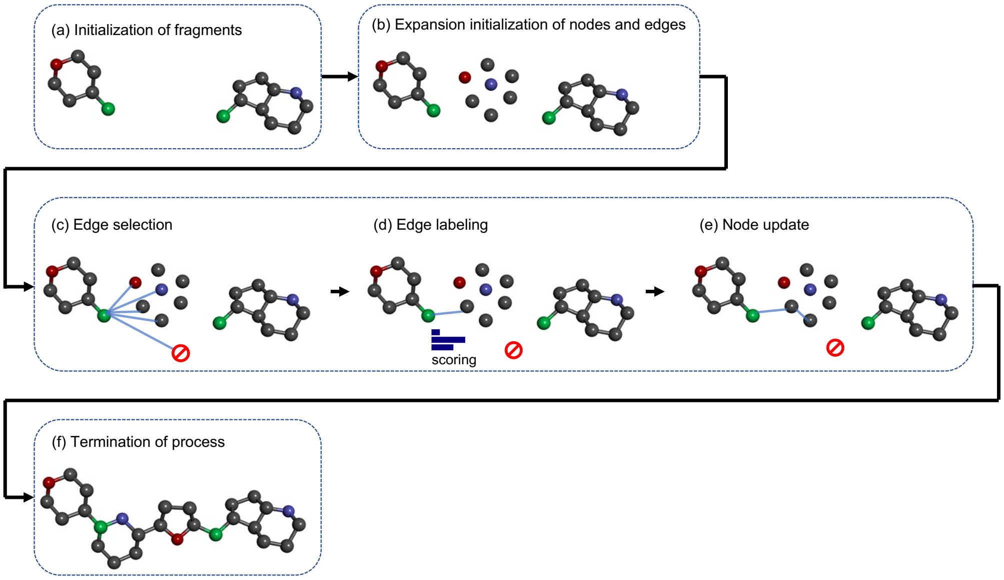

(a) Initialization of fragments
(b) Expansion initialization of nodes and edges
(c) Edge selection
(d) Edge labeling
scoring
(e) Node update
(f) Termination of process

Figure 3: Network generation process. Two fragments in (a) are generated with the data processing steps, providing the spatial information, and the angle and distance between the anchors are calculated accordingly. Initialization of the nodes and edges in the network is illustrated in (b), where the 14 permitted atoms are randomly selected between the space. From (c) to (e) are the steps to process edge, selection, edge labeling, and node updates, respectively. In particular, the three steps are sequentially repeated operation until the atom number reaches the maximum setting or all the edges and nodes are generated to cause (f) termination of the process.

It follows by a set of expansion nodes initializing at random with hidden states z_v drawn from the d-dimensional standard normal distribution, $\mathcal{N}(0, I)$, where d is the length of the hidden state (Figure 3b) and $h_v^{t=0}$ is the concatenation of $[z_v^t, l_v]$. The nodes are then labeled with an atom type according to their hidden state of z_v, and the structural information learned from the softmax output w, where f is implemented as a linear classifier but could be any function mapping a node's hidden state to an atom type. The number of expansion nodes determines the maximum length of the linker and is a parameter chosen by the user.

The new molecule is constructed from this set of nodes via an iterative process consisting of edge selection, edge labeling, and node update (Figure 3c-e). The process is repeated until the number of connected nodes fall between the number we set, and no further nodes can be formed (i.e. termination of process in Figure 3f). At each step, we consider adding edges on each node, v, and other nodes in the graph are randomly scattered for the following connection process. At each iterative process, an edge and node will be constructed and link to the prior atom. v is chosen according to a deterministic first-in-first-out queue that is initialized with the exit vectors of each fragment. Each node is connected to the graph with no repetition, and new edge is added to the node v until an edge to the halt node is selected or the length of the linker is filled. Whenever obtaining a new graph $\mathcal{G}^{(t+1)}$, the h_v^t is discarded, and the representation of h_v^{t+1} is computed. The node becomes closed with no additional edges permitted to connected with it.

We consider all possible edges between the node v and other nodes in the graph (Figure 3c), subject to basic valency constraints. The f, representing the core network GGNN, constructs a feature vector with the subsequent node u with such $l_v \sim f(z_v^t)$. The feature vector $\phi_{v,u}^t$ for the edge between node v and candidate node u is given by

$$\phi_{v,u}^t = [h_v^t, h_u^t, d_{v,u}, H^0, H^t, I]$$

where h_v^t is the hidden state of initial node v after t steps update and its atomic label l and h_u^t represents the target u updated with t steps. $d_{v,u}$ is the graph distance between v and u. Note that $h_{\mathcal{G}}^{t+1}$ is computed from $h_{\mathcal{G}}^0$ rather than $h_{\mathcal{G}}^t$. This means that the state of each node is independent of the generation history of the graph and depends only on the current state of the graph. H^0 is the average initial representation of all nodes, while H^t is the average representation of nodes at generation step t, and I represents the 3D structural information, which contains the angle and distance data. The following representation is used to produce a distribution over candidate edge:

$$p(v \leftrightarrow u \mid \phi_{v,u}^t) = p(l \mid \phi_{v,u}^t, v \leftrightarrow u) \cdot p(v \leftrightarrow u \mid \phi_{v,u}^t)$$

The network further chooses which edge to add to the graph, in which it utilizes local information of the nodes, global information regarding the unlinked initial side of the fragment and the current graph state. Once a node u has been selected, the edge between v and u is constructed as either a single, double, or triple bond (subject to valency constraints) by the neural network taking as input the same feature vector $\phi^t_{v,u}$ (Figure 3d). Finally, the hidden states of all nodes are updated according to a GGNN (Figure 3e). Steps c - e in Figure 3 are repeated for each node in the queue, until the queue is empty, at which point the generation process terminates. At termination step (Figure 3f), all unconnected scattered nodes are removed and the largest connected component is returned as the generated molecule. The stereochemistry information of the generated molecules is not assigned during the generative process. As such, a postprocessing step is needed for screening the predicted molecules and test the robustness.

Model Training

To train the network with relevant molecular structures, we prepare a dataset containing conventional ZINC dataset[38,39] and PROTAC-DB, and train under variational autoencoder (VAE) framework. We construct the training dataset of 160,491 molecules with 157,221 and 3,270 from ZINC and PROTAC-DB,[40] respectively. In the ZINC dataset, we select the chemical compounds with heavier and more complex structures. Meanwhile, in the PROTAC-DB dataset, all compounds to date are selected into our training dataset. Next, we split the dataset into 80% for training and 20% for validation process with adapting 10-fold cross validation to overcome the overfitting issue. We tune the hyperparameter as reflected in Table S1 to retrieve the best performance.

The model trains on the dataset focusing on the fragment-molecule pairs. Given a fragment A and linked molecule B, the model reconstructs B from (X, z), where z is the latent code, derived via a learnt mapping from the mean of the node embedding of the training label B. The objective is similar to standard VAE loss including a reconstruction terms on a

Kullback-Leibler regularization, and a regression loss: $\mathcal{L} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{latent} + \lambda_2 \mathcal{L}_{regression}$. Note that we allow variations from the pure VAE loss ($\lambda_1 = 1$) by Yeung et al.[41]

Postprocessing

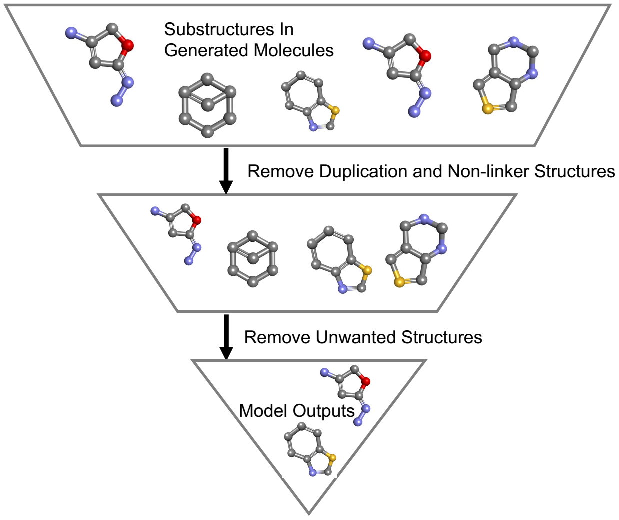

Substructures In
Generated Molecules
Remove Duplication and Non-linker Structures
Remove Unwanted Structures
Model Outputs

Figure 4: Workflow of postprocess. The generated molecules pass through multiple steps of filters, specifically, removing duplicates, non-linker, and unwanted substructures. This postprocess step is integrated in AIMLinker.

The raw outputs from the model are PROTACs library in 2D chemical structure and are further postprocessed with our screening procedures to remove the unwanted targets. Figure 4 shows the filters of our proposed method that is integrated in the AIMLinker. Due to the constraint of the graph computational process and the linked substructures, the model generates the molecules by choice, which includes duplicated predictions, undruggable targets and the structures violating the basic chemistry law. The first filter rules out the duplicates, same structures predicted by the model, leaving every unique molecule after this process. This process also includes the non-linker substructures, i.e., two fragments are not formed into one compound with the linker structure. Next, we have a library with the unfavorable substructures that are not feasible for chemical synthesis, or unable to be a druggable target. This library includes the substructures such as acid halide, disulfide bond, peroxide bond, and small-number cyclic rings with double bonds, and additionally by the model to predict whether the newly generated substructures is feasible as a drug lead. This step is significantly important for screening the target pool to reduce the molecule numbers while retaining the candidates potentially having high binding affinity and chemical activity. Finally, the molecules that violate the Bredt's Rule are removed from the target pool. From steric structure point of view, Bredt concluded that certain bicycle atomic-bridged-ring structures with a carbon-carbon double bond at a bridgehead atom should not be capable of existence.[42] These steps remove the unwanted molecules from the target pool and the remaining targets remarkably reduce the needed computational resources and the time span for simulation.

To further validate the merit of using postprocess steps, we utilize "Rule of Three" to measure the effectiveness of our generated molecules. "Rule of Three" refers to molecular weight (MW) of a fragment is $<$300 Dalton, the calculated logarithm of the 1-octanol–water partition coefficient of the non-ionized molecule (cLogP) is ≤ 3, the number of hydrogen bond donors (HBD) is ≤ 3, and the number of hydrogen bond acceptors (HBA) is ≤ 3, and we also include the polar surface area (PSA) is ≤ 60 Å in addition to the standard setting of the rule.[43,44] We apply this rule to the linker structure to show our generated molecules properties, specifically, we measure and compare these metrics on the molecules generated after the first and second filters, respectively.

Docking Validations

Before applying docking methods, 3D protein-protein interaction poses and 3D conformations of postprocessed molecules need to be constructed first. With the aim of 6BOY structure, released by Nowak et al.,[34] co-crystal structure of CRBN-dBET6-BRD4 can be easily retrieved via DSV; thus, the spatial pose of CRBN-BRD4 conformation is defined. Meanwhile, all postprocessed molecules, initially sketched as 2D chemical structures, are converted into 3D PROTAC conformations through DSV. The reference compound dBET6 is also re-constructed into series of 3D conformations for validating the consistency of our docking methodology. These 3D compounds are subsequently minimized using energy minimization method.[45] After protein-protein interaction poses and 3D conformations of PROTAC candidates are well prepared, AutoDock4[46] is applied to predict the best PROTAC binding pose with labeling the binding pocket as grid. During the docking procedure, each 3D PROTAC freely bind to CRBN and BRD4 with consideration of the binding energy, biochemistry property and entropy. We allow 10 binding poses of each PROTAC from the network and form a 10-time dataset as the initial docking inputs. As such, our PROTAC library and dBET6 can freely rotate, fold, and bind to the pocket to form a best pose in DSV with highest binding affinity and lowest entropy energy.

To validate the robustness of generated molecules from AIMLinker and dBET6 provided in PDB, we measure the metrics including structural information and binding affinity. For measuring the alignment with dBET6, we use RMSD, which shows the structural similarity level between two molecules, and demonstrates a certain crystal pose possessing symmetric structure based on default atomic ordering. RMSD was introduced by Bell et al.[47] for measuring between respective atoms in two molecules,

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

where N is the number of atoms in the ligand, and d_i is the Euclidean distance between the i^{th} pair of corresponding atoms.

Another metric taken into consideration for validating the molecules is ΔG. ΔG of the protein–ligand complexes is calculated from molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) method.[48] The MM-PBSA approach is one of the most widely used methods to compute interaction energies among the biomolecular complexes. In general, ΔG between a protein and a ligand in solvent can be expressed as:

$$\Delta G = G_{Complex} - (G_{Protein} + G_{Ligand})$$

where $G_{Complex}$ is the total free energy of the protein-ligand complex and $G_{Protein}$ and G_{Ligand} are total free energies of the separated protein and ligand in solvent, respectively.

Results and Discussions

We develop a novel network, AIMLinker, generating neo-structure of small molecule linker for PROTAC degradation protein. AIMLinker takes two fragments with structural information as the input data and process with deep learning network to create linker molecule. We first provide the details in the generated molecules and their chemical properties and structural statistics. Next, we take four molecules with highest binding affinity to compare dBET6 with their RMSD and ΔG.

AIMLinker demonstrates a robust and rapid pipeline for generating and designing new PROTAC linkers. Our network combines two processes into one end-to-end pipeline: 1) takes two unlinked fragments as input and uses an encoder-decoder deep learning network to generate the substructures forming a new PROTAC molecule. We consider the structural information in the form of angle and distance between the two initial fragments, and iteratively adding or replacing atoms between the space until filled or reaching limitation of maximum atom setting, 2) postprocesses the generated molecules to extract the potential drug-like molecules. In particular, we screen with duplicates, and exclude structures violating basic chemistry rules and unwanted substructures. This rapid pipeline, taking less than two days to generate postprocessed PROTAC library and docking validation, provides the viability of generating novel small molecules to possess high binding affinity to CRBN-BRD4 binding pockets retrieved from 6BOY, and the potential to translate the work to other PROTAC proteins.

Generated Molecules

We generate the molecules with the input of angle and distance between the fragments and set the minimum and maximum of the atoms allowed between the spaces. The allowed atom range is calculated by the distance between the fragments, giving the flexibility to the network designing a more linear or ring-like structure. The raw output from the multimodal neural network is 2,000 structures, including illogical molecules and unwanted substructures. Therefore, we then take these outputs to undergo postprocess procedures with two filters applied: 1) first filter removes the duplicated molecules and non-linker structure, i.e., two fragments are not formed into one compound with the linker structure. After this filter, the remaining number of molecules is 1,175, 2) filters out the unwanted substructures that are not applicable drug-like molecules. The final output number from AIMLinker is 524 molecules, yielding 26.4% compared to the raw output. This remaining portion indicates the effectiveness of adapting postprocess steps to exclude the non-target molecules.

Table 1: Results of "Rule of Three" parameters analysis. Abbreviations: molecular weight, MW; the calculated logarithm of the 1-octanol–water partition coefficient of the non-ionized molecule, cLogP; number of hydrogen bond donors, HBD; number of hydrogen bond acceptors, HBA; polar surface area, PSA.

Parameters	Preprocess (%)	Postprocess (%)
MW (<300 Da)	91	**93**
cLogP (≤ 3)	97	95
HBD (≤ 3)	90	**95**
HBA (≤ 3)	49	**60**
PSA ($\leq 60\text{Å}^2$)	36	**48**

Next, we utilize "Rule of Three" to validate the effectiveness of applying postprocess steps in the linker structure. Table 1 shows the rule of three metrics of MW, cLogP, HBD, and HBA, and we additionally include PSA here. The generated pool of molecules applied with postprocess step outperformed that of preprocess step except cLogP. Specifically, our proposed method has 93%, 95%, 95%, 60%, and 48% of the molecules that pass the rules in MW, cLogP, HBD, HBA, and PSA, respectively. For preprocessed molecular pool, it achieves 91%, 97%, 90%, 49%, and 36% in the corresponding metrics. In addition, this method surpasses the preprocessed data with as high as 12% in PSA, while the lowest surpassed percentage compared to the preprocessed data is 2% in MW. We show better performance with this additional postprocess step in 4 out of 5 metrics, demonstrating the robustness of the linker molecules possessing better chemical properties.

Table 2 provides the structural statistics of the final output from AIMLinker. In particular, the generated molecules from AIMLinker provides ring-shape structures, while dBET6 is a linear structure, which gives the compound more flexible to freely rotate inside the binding pockets and the opportunities of binding to other positions to reduce the compound potency and the pharmacokinetics property. Our generated linker structures between the two input fragments provide 229 ring-like substructures out of 524 molecules, 43% of the total number. Of the 229 compounds, there are 32 having bicyclic rings, and one compound having tricyclic rings. The incidences of ring-shape structures of the designed molecules are demonstrated in Table 2 that the numbers of three-membered, four-membered, five-membered, six-membered rings, and the number of atoms in the ring structure above 6 are 24, 30, 90, 112, and 6, respectively. These cyclic compounds restrict the rotational angles and the possibilities binding to non-target binding positions. In addition, the ring-link structures generated by AIMLinker provide more stability of the compound and possess the ability to form strong π bond increasing the binding affinity in the binding pockets.

Table 2: Ring-structure statistics of the generated molecules. Here we provide the number of incidences of different numbers of membered rings.

Ring Structures	Number of Molecules
3-membered Ring	24
4-membered Ring	30
5-membered Ring	90
6-membered Ring	112
Above 6-membered Ring	6

Docking Performance

To assess the generated molecules and compare with the existing dBET6 structure, we use AutoDock4 for docking and validation. In order to consolidate AutoDock4 having the ability to be a reference tool for measuring the generated molecules, we re-dock the compound to the binding pockets of CRBN and BRD4. We constrain the free energy of dBET6 for mostly generating the similar spatial position as provided in the 6BOY structure. The final output of 524 molecules from AIMLinker are then passed through AutoDock4. In the standard protocol and matching with the biological interaction, we allow a maximum of 10 binding poses for each molecule. Since not every molecule is feasible to bind within the pocket, the total are 5,095 poses generated from docking. We use RMSD as the metric to measure the similarity between a certain docked molecule with initial dBET6 conformation. We set the RMSD threshold value of less than 1 Å to be considered as drug-like molecules. There are four molecules, displayed in Figure 5, extracted from this threshold.

Table 3: Docking performance of redocked dBET6 and the generated molecules.

Linker Molecules	RMSD (Å)	ΔG (kcal/mol)
dBET6 (crystal pose)	–	−22.51
dBET6 (redocked pose)	0.58	−37.36
6BOY_1268	**0.36**	**-52.72**
6BOY_1974	0.55	−48.98
6BOY_1854	0.68	−44.75
6BOY_0518	0.65	−42.57

In Table 3, we show the alignment RMSD and ΔG values. The spacial structural information is measure with RMSD values. Of the re-docked dBET6 and four generated molecules, 6BOY_1268 has the best RMSD of 0.36 Å, while re-docked dBET6 has the second-best value of 0.58 Å. For the change of ΔG, the lower value indicates better performance. In this situation, all four generated molecules possess lower ΔG than re-docked dBET6 with −52.72, −48.98, −44.75, and −42.57 for 6BOY_1268, 6BOY_1974, 6BOY_1854 and 6BOY_0518, respectively. Figure 5 shows the four molecules designed by AIMLinker with the best similarity to dBET6 structure and possess drug-like property.

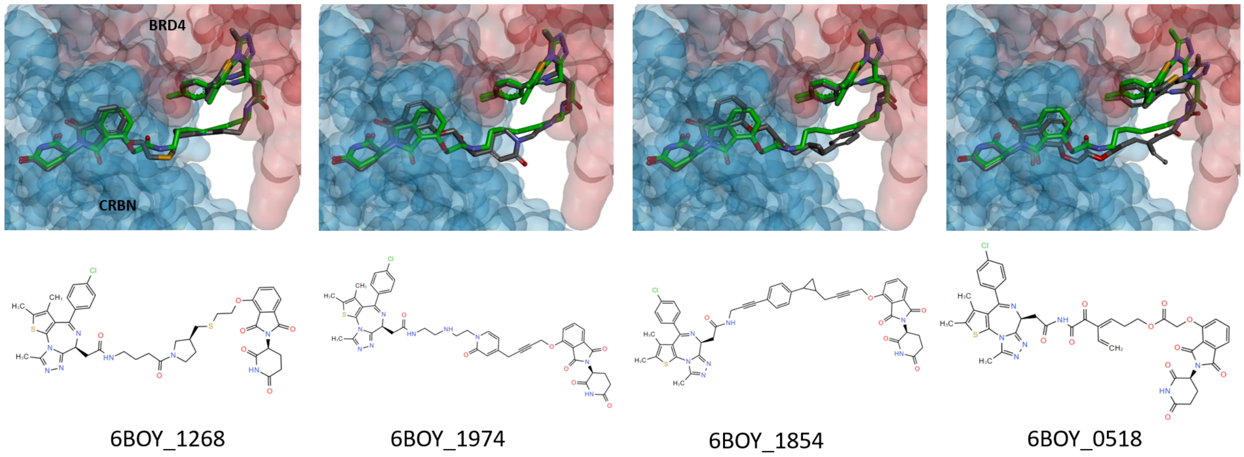


Figure 5: Crystal poses of the generated molecules that the RMSD value less than 1 Å, which takes dBET6 conformation as the reference compound.

Conclusion

In this work, we propose a deep neural network to generate and design novel PROTAC molecules. We integrate sampling and postprocessing steps to extract the potent drug-like molecules and demonstrate the robustness of the generated molecules. We show that our generated structures possess superior chemical properties than the existing compound. Our model can perform virtual high-throughput screening for rapid generation and reducing manual labors. We focus on single PROTACs target for testing and validating our proposed model. In the next stage, we aim to expand the utilization of the model and apply to more PROTACs targets, and further investigate the applications in other fields of drug discovery.

Data Availability Statement

All data mentioned in this study are publicly open at ZINC dataset, PROTAC-DB, and Protein Data Bank (PDB). We retrieved the training and validation data from above databanks. All the data we applied can be found at the supplementary document and https://github.com/AnHorn/AIMLinker.

References

(1) Anderson, A. C. The Process of Strcuture-Based Drug Design. *Chemistry Biology* **2003**, *10*, 787–797.

(2) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-aided Molecular Design* **2013**, *27*, 67–679.

(3) Kick, E. K.; Roe, D. C.; Skillman, G.; et al., Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chemistry Biology* **1997**, *4*, 297–307.

(4) Verlinde, C. L.; Hol, W. G. Structure-based drug design: progress, results and challenges. *Structure* **1994**, *2*, 577–587.

(5) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675–679.

(6) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942.

(7) Annis, D. A.; Nazef, N.; Chuang, C. C. A General Technique To Rank Protein-Ligand Binding Affinities and Determine Allosteric versus Direct Binding Site Competition in Compound Mixtures. *J. AM. CHEM. SOC* **2004**, *126*, 15495–15503.

(8) Vamathevan, J.; Clark, D.; Czodrowski, P.; et al., Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **2019**, *18*, 463–477.

(9) Ramsundar, B.; Liu, B.; Wu, Z.; et al., Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068–2076.

(10) Aliper, A.; Plis, S.; Artemov, A.; et al., Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular Pharmaceutics* **2016**, *13*, 2524–2530.

(11) Wu, Z.; Ramsundar, B.; Feinberg,; N., E.; et al., MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

(12) Aliper, A.; Plis, S.; Artemov, A.; et al., The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *123*, 1241–1250.

(13) Zhou, J.; Cui, G.; Hu, S.; et al., Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81.

(14) Jiang, D.; Wu, Z.; Hsieh, C.; et al., Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.* **2021**, *13*, 758–780.

(15) Sun, M.; Zhao, S.; Gilvary, C.; et al., Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics* **2020**, *21*, 919–935.

(16) Flam-Shepherd, D.; Wu, T. C.; Friederich, P.; Aspuru-Guzik, A. Neural message passing on high order paths. *Mach. Learn.: Sci. Technol.* **2021**, *2*.

(17) Li, J.; Cai, D.; He, X. Learning Graph-Level Representation for Drug Discovery. *arXiv* **2017**, *03741*.

(18) Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J. Cheminform.* **2020**, *12*.

(19) Sakamoto,; M., K.; Kim, K. B.; Kumagai, A.; et al., Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Biochemistry* **2001**, *98*, 8554–8559.

(20) Winter, G. E.; Buckley, D. L.; Paulk, J.; et al., DRUG DEVELOPMENT. Phthalimide conjugation as a strategy for in vivo target protein degradation. *Science* **2015**, *348*, 1376–1381.

(21) Sun, X.; Gao, H.; Yang, Y.; et al., PROTACs: great opportunities for academia and industry. *Signal Transduction and Targeted Therapy* **2019**, *4*.

(22) Ishida, T.; Ciulli, A. E3 Ligase Ligands for PROTACs: How They Were Found and How to Discover New Ones. *SLAS Discovery* **2021**, *26*, 484–502.

(23) Konstantinidoua, M.; Li, J.; Zhang, B.; et al., PROTACs – a game-changing technology. *Expert Opinion on Drug Discovery* **2019**, *13*, 1255–1268.

(24) Cyrus, K.; Wehenkel, C. E.-Y., Marie; et al., Impact of linker length on the activity of PROTACs. *Molecular BioSystems* **2011**, *7*, 359–364.

(25) Troup, R. I.; Fallan, C.; Baud, M. G. J. Current strategies for the design of PROTAC linkers: a critical review. *Exploration of Targeted Anti-tumor Therapy* **2020**, *1*, 273–312.

(26) Xu, Y.; Lin, K.; Wang, S.; et al., Deep learning for molecular generation. *Future Med. Chem.* **2019**, *11*, 567–597.

(27) Elton, B.-Z., Daniel C.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *arXiv* **2019**,

(28) Simonovsky, M.; Komodakis, N. Towards generation of small graphs using variational autoencoders. *FInternational conference on artificial neural networks* **2018**, 412–422.

(29) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. L. Constrained Graph Variational Autoencoders for Molecule Design. *Advances in Neural Information Processing Systems 31* **2018**, 7795–7804.

(30) Yang, Y.; Zheng, S.; Su, S.; et al., SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.* **2020**, *11*, 8312–8322.

(31) Imrie, F.; Bradley, A. R.; Van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60*, 1983–1995.

(32) Huang, Y.; Peng, X.; Ma, J.; Zhang, M. 3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design. *International Conference on Machine Learning* **2022**,

(33) Li, Y.; Zeme, R.; et al., Gated Graph Sequence Neural Networks. *arXiv* **2016**,

(34) Nowak, R. P.; DeAngelo, S. L.; Buckley, D.; et al., Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nature Chemical Biology* **2018**, *14*, 706–714.

(35) Berman, H. M.; Westbrook, J.; Feng, Z.; et al., The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(36) BIOVIA; Dassault Systèmes, Discovery Studio Visualizer, v21.1.0.20298. *Dassault Systèmes: San Diego, CA, USA, 2021*

(37) Filippakopoulos, P.; Qi, J.; et al., Selective inhibition of BET bromodomains. *Nature* **2010**, *468*, 1067–1073.

(38) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 177–182.

(39) Irwin, J. J.; Tang, K. G.; Young, J.; et al., ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling* **2020**, *60*, 6065–6073.

(40) Weng, G.; Shen, C.; Cao, D. PROTAC-DB: an online database of PROTACs. *Nucleic Acids Research* **2021**, *49*, 1381–1387.

(41) Yeung, S.; Kannan, A.; Dauphin, Y.; Li, F.-F. Tackling Over-pruning in Variational Autoencoders. *arXiv preprint arXiv:1706.03643* **2017**,

(42) Fawcett, F. S. Bredt's Rule of Double Bonds in Atomic-bridged-ring Structures. *Chemical Reviews* **1950**, *126*, 219–274.

(43) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.

(44) Jhoti, H.; Williams, G.; Rees, D. C.; Murray, C. W. The 'rule of three' for fragment-based drug discovery: where are we now? *Nature Reviews Drug Discovery* **2013**, *12*.

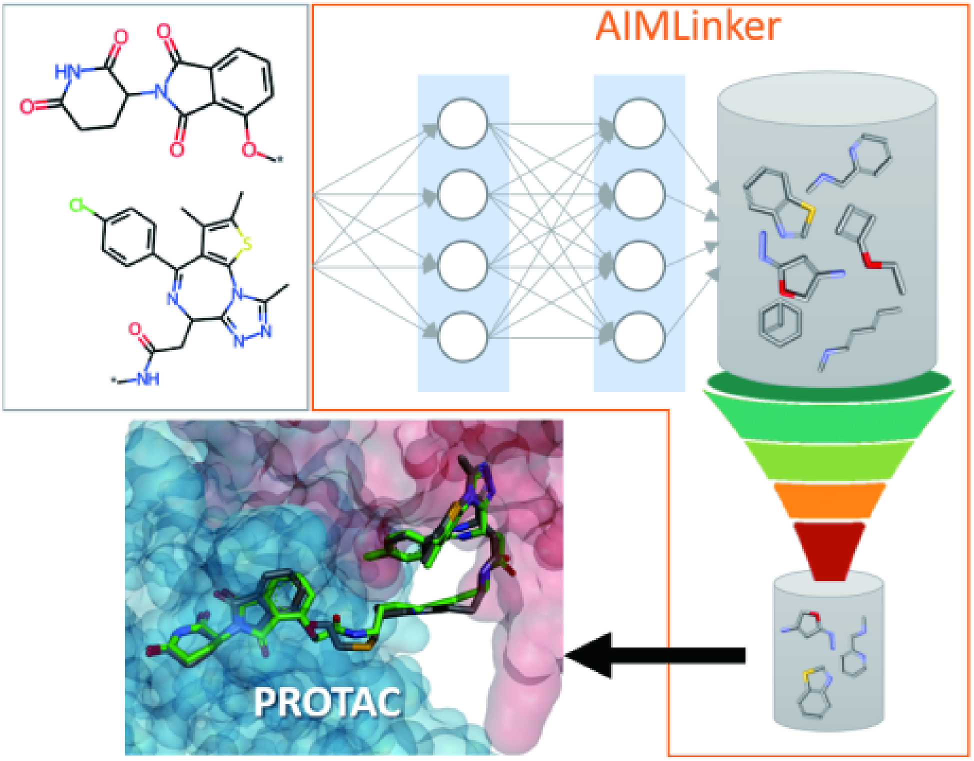
(45) Hahn, M. Receptor surface models. 1. Definition and construction. *Journal of Medicinal Chemistry* **1995**, *38*, 2080–2090.

(46) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **2009**, *30*, 2785–2791.

(47) Bell, E. W.; Zhang, Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *Journal of Cheminformatics* **2019**, *11*, 9–18.

(48) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W., et al. Calculating structures and free energies of com-

plex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research* **2000**, *33*, 889–897.


AIMLinker
PROTAC