SyntaLinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks

Yuyao Yang[§1,2], Shuangjia Zheng[§1], Shimin Su[1,2], Jun Xu[1,*], Hongming Chen[2,*]

[1]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou 510006, China

[2]Centre of Chemistry and Chemical Biology, Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou 510530, China

[§]Equal contributors. *To whom correspondence should be addressed.

Contact: chen_hongming@grmh-gdl.cn, junxu@biochemomes.com.

Abstract

Fragment based drug design represents a promising drug discovery paradigm complimentary to the traditional HTS based lead generation strategy. How to link fragment structures to increase compound affinity is remaining a challenge task in this paradigm. Hereby a novel deep generative model (SyntaLinker) for linking fragments is developed with the potential for applying in the fragment-based lead generation scenario. The state-of-the-art transformer architecture was employed to learn the linker grammar in ChEMBL compounds and generate novel linker without relying on any predefined chemical rules. Our results show that, given starting fragments and user customized linker constraints, SyntaLinker model can design abundant drug-like molecules fulfilling these constraints and its performance was superior to other reference models. Moreover, several examples were showcased that SyntaLinker can be useful tools for carrying out drug design tasks such as fragment linking, lead optimization and scaffold hopping.

Introduction

Over the past two decades, the fast development of gene sequencing technologies, together with high-throughput screening[1] (HTS) and combinatorial chemistry[2] for library synthesis have largely changed the drug discovery paradigm from a phenotypic centric approach to a target centric approach.[3-4] Lead identification by screening large compound collection has become standard exercise among large pharmaceutical companies.[5] Albeit its success in drug discovery, the high cost for maintaining the large compound collection and launching a screening campaign is a big hurdle for drug developers in academics and small biotech companies.[6] Also, there are many factors influencing the quality of HTS hits such as technology hitter, sample purity, and sample aggregation etc.[7-8]

In recent years fragment-based drug design (FBDD) has gained considerable attention as an alternative drug discovery strategy due to its relatively low cost in running the assay and potential advantages in identifying hits for difficult targets.[9-11] The concept of FBDD can date back to the pioneering work of William Jencks in mid-1990s.[12] It usually starts from screening low molecular weight molecules which have weak, but efficient interactions with a target protein (for example, MW<300 Daltons; binding affinity of the order of mM).[13] The fragment screening is usually carried out at high concentration and a typical fragment collection is around a few thousands of compound in contrast to millions of compound in HTS which are routinely run in “Big Pharma”.[14] The effective use of fragments as starting points for step-wise optimizations has shown capability to overcome the major obstacles for further drug development, such as limited chemical space, low structural diversity, and unfavorable drug absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties.[15] Therefore, the popularity of fragment-based drug design has grown at a remarkable rate in both industry and academic institutions.[16]

In practice, there are still two key factors for successfully utilizing FBDD in drug discovery: (i) identify suitable fragments (ii) grow and optimize these fragments to develop drug-like molecules. Actually, finding desirable fragments is relatively straightforward. Many experimental and computational efforts have accelerated the identification process in the past decade[17], including nuclear magnetic resonance (NMR), X-ray crystallography, surface plasmon resonance (SPR) and virtual screening[18]. After obtaining the initial fragment hits, the next key step is to expand these fragments into molecules with larger size for lead optimization. Fragment growing, merging, and linking are three main strategies commonly used by medicinal chemists.[19,20] Although there are some fragment linking examples reported in literature[21-22], unlike the other two popular strategies, fragment linking is generally regarded as a challenging task due to the stringent requirements on keeping original binding modes for fragment moieties after linking. Mismatch of length or geometry in the linker can have a dramatically negative effect on binding of full compound. Nevertheless, fragment linking is still attractive because of the promise of additive binding potency (an improvement in ligand efficiency (LE) rather than mere maintenance of LE).[23-25]

How to identify a linker featuring an optimal fit has stimulated research enthusiasm among drug developers. Traditional computational tools for fragment linking are mainly two types[26]: (i) database search[27-28], (ii) quantum mechanical (QM) calculation (such as fragment molecular orbital (FMO)).[29-31] However, these methods have been significantly limited by the size of database or complex computations.

Recently, advances in the development of deep generative models have spawned a mass of promising methods to address the structure generation issue in drug design.[32-34] The applications of deep generative models in pharmaceutical research has covered de novo molecular design[35-38], molecular optimization[39-41]. Various generative architectures including RNNs[42], autoencoders[43-44], and generative adversarial networks[45] (GANs) have been proven to be effective data-driven methods for creating desirable molecules, which are either represented by simplified molecular input line entry specification[46] (SMILES) or molecular graph. Recently, a fragment linking methodology based on generative model has been reported by Fergus Imrie and co-workers.[47] They put forward a molecule graph based deep generative model (DeLinker), which can join fragments through restraining their relative positions. In their method the relative position between the starting fragments is merely represented by the distance and angle between the two bonds in the vicinity of the linker sites. This model requires pre-generated three dimensional conformations of a molecule for selecting training set compounds, and the algorithm to seek the biologically active conformation[48] could be computationally expensive and challenging.

Herein, we developed a novel fragment linking methodology, SyntaLinker, to link fragments efficiently and rapidly based on a SMILES based generative model and the 2D bond distance among linked fragments can be added as a constraint for linker generation. Inspired by machine translation task in natural language processing (NLP), the transformer architecture[49] has recently been successfully used for reaction prediction and retrosynthesis planning.[50-52] Motivated by these applications, we regard the fragment linking as a common task, called sentence completion[53], in NLP, and develop a novel conditional transformer architecture (SyntaLinker) for linker generation in a controllable manner. In current study, ChEMBL compounds were broken down into terminal fragment and linker components and transformer models were trained to learn the compound synthetic grammar and generate linker without relying on any predefined rigid transformation rules.

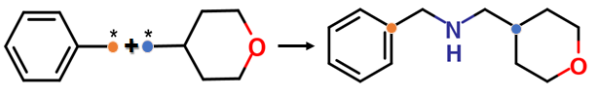
Our model takes terminal fragments and linker constraints, such as the shortest linker bond distance

(SLBD), the existence of the hydrogen bond donor, hydrogen bond acceptor, rotatable bond and ring etc., as inputs and generate product compounds containing input fragments. Compared to DeLinker, our approach achieved higher recover rate in terms of rational linker prediction. Finally, through a few case study examples, we also demonstrated the effectiveness of our approach on some common drug design tasks such as fragment linking, lead optimization and scaffold hopping.

Methods

Task Definition. Our goal is to generate a drug-like molecule by connecting two given starting fragments under specified constraints of the linker as shown in Figure 1. There are two types of linker constraints used as control codes during training. One is the SLBD between two linker sites, which is used to maintain relative position of two starting fragments. The other one is the constraint with multiple features, besides the SLBD, including also the presence of hydrogen bond donor (HBD), hydrogen bond receptor (HBA), rotatable bond (RB) and ring, which can be regarded as additional pharmacophoric constraints.

This process can be regarded as an end-to-end sentence completion process, where the starting fragments is the input signal and the full compound is the output signal. Swaller et al[50] used encoder-decoder neutral network architecture to predict reaction product in an end-to-end manner, where the reactants serve as source sequence and reaction product as target sequence. Whereas in our case, the two starting fragments and constraints of the SLBD (as prepended token) are defined as the source sequence, and the full molecule as target sequence. Examples of the sequence expression are shown in Figure 1. A training example with SLBD as constraint is described as "[L_4]c1ccccc1[*].[*]C1CCOCC1>>c1ccc(CNCC2CCOCC2)cc1", where "[L_4]c1ccccc1[*].[*]C1CCOCC1" is the source sequence, "c1ccc(CNCC2CCOCC2)cc1" is the target sequence, and “[L_4]” is the SLBD (equal to four bond distance) as the control code. In the other example, “[L_4 1 1 1 0]” represents the multiple constraints, where the “1 1 1 0” represents the presence of HBD, HBA, RB and the absence of ring.



SLBD constraint	[L_4] c1ccccc1[*].[*]C1CCOCC1>>c1ccc(CNCC2CCOCC2)cc1
Multiple constraints	[L_4 1 1 1 0] c1ccccc1[*].[*]C1CCOCC1>>c1ccc(CNCC2CCOCC2)cc1

Figure 1. An example of the source sequence and target sequence SMILES with different constraints.

Model Architecture. A novel conditional transformer model (SyntaLinker), based on transformer architecture, was proposed for generating structures with the customized conditions. Compared to the original transformer model[49], SyntaLinker (shown in Figure 2) introduces a variety of prepended control codes[54-55] to ensure generated molecules to satisfy explicit criterion.

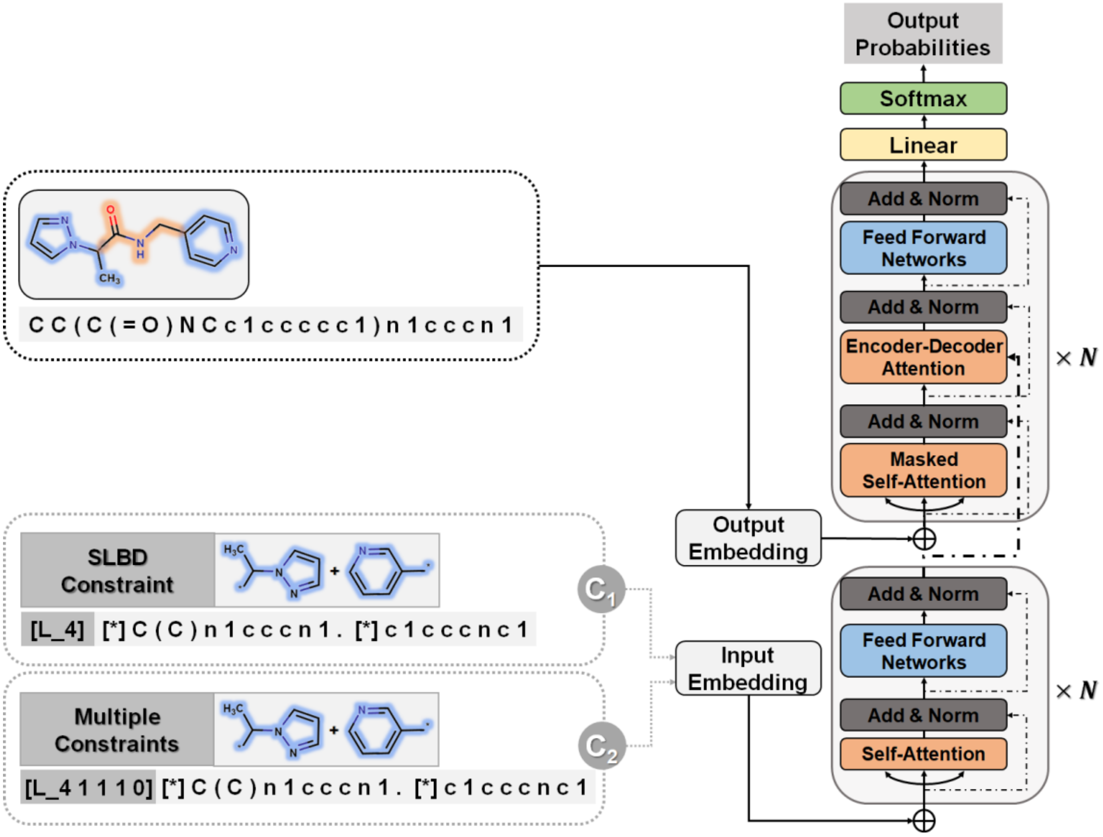


Figure 2. Basic architecture of the controllable transformer model (SyntaLinker). In the input embedding layer, we embed each fragment pairs with SLBD constraints (C_1) or multiple constraints (C_2).

All source and target sequences of our data set were first tokenized to construct a vocabulary. For each single example sequence containing n tokens, it was first encoded into a one-hot matrix by the vocabulary and then transformed into a embedding matrix $M_s = (e_1, \ldots, e_n)$ by a word embedding algorithm[56] following our previous work[57]. M_s was composed of n corresponding vectors in R^d, where d is embedding dimension.

The core architecture of SyntaLinker contains multiple encoder-decoder stacks. The encoder and decoder consist of six identical layers, respectively. Each encoder layer has a multihead self-attention sub-layer and a position-wise feedforward network (FFN) sub-layer. A multihead attention consists of several scaled dot-product attention functions running in parallel and concatenate their outputs into final values, which allows the model to focus on information from different subspaces at different positions. An attention mechanism computes the dot products of the query (*Q*) with all keys (*K*), introduces a scaling factor d_k (equal to the size of weight matrices) to avoid excessive dot products, and then apply a softmax function to obtain the weights on the values (*V*). Formally

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

The FFN sub-layer adopt ReLU activation.[58] Then, layer normalization[59-60], and a residual connection[61] introduced to integrate above two core sub-layers. Each decoder layer has three sub-layers, including two attention sub-layer and an FFN sub-layer. The decoder self-attention sub-layer uses a mask to preclude attending to future tokens. While the encoder-decoder attention sub-layer helps the decoder to focus on important prats in the source sequence, and capture the relationship between the encoder and decoder.

For a given source sequence, its input embedding was processed by encoder layers into a latent representation $L = (l_1, ..., l_n)$. Given *L*, the decoder output was normalized with a softmax, yielding a probability distribution for sampling a token, and then generated an output sequence $Y = (y_1, ..., y_m)$ of one token at a time until the ending token "⟨/s⟩" was generated. Finally, model calculate and minimize the cross-entropy loss between the target sequence $M_t = (e_1, ..., e_k)$ and the output sequence *Y* during training.

$$\mathcal{L}(Y,M) = -\sum_{i=1}^{k} y_i \, log \, m_i \qquad (2)$$

Data set preparation. Our data set were derived from the ChEMBL[62] database and preprocessed in the same way as in previous study[35]. After the preprocessing, remaining molecules were further filtered through Lipinski's "Rules of Five"[63], pan assay interference compounds[7] (PAINS) substructures and synthetic accessibility score[64] (SAscore, the cut-off value set to 6.5) to make sure the generated molecules are drug-like and likely to be synthesizable. In the end 718,652 unique molecules were kept for further processing.

To mimic the fragment linking scenario, we constructed our data set using matched molecular pairs (MMPs) cutting algorithm[65] proposed by Hussain et al. Firstly, each molecule was deconstructed using MMPs cutting algorithm, which executed double cuts of non-functional group, acyclic single bonds in every compound and this will transform the compound into a quadruple form like "fragment1, linker, fragment2, molecule", which corresponds to two terminal fragments, a linker and the original compound. In total 5,873,503 fragment molecule quadruples (FMQ) were enumerated; Secondly, the FMQs were further filtered using "Rule of three"[66] criteria, i.e. a FMQ will be removed if any of its terminal fragment violate the "Rule of three" criteria; Lastly, considering that the requirement of linking fragments in reality is to connect two close fragments using linker as simple as possible, the remaining FMQs were then filtered by SLBD of linker (SLBD less than 15) and SAscore according to Equation 1 to ensure the terminal fragments have reasonable synthesis feasibility (SAscore is less than 5) and SAscore of linker is lower than the sum of fragments.

$$SAscore\_filter = \begin{cases} SAscore_{fragment1} < 5 \\ SAscore_{fragment2} < 5 \\ SAscore_{linker} < \left(SAscore_{fragment1} + SAscore_{fragment2}\right) \end{cases} \tag{3}$$

In the end, 784,728 FMQs were obtained. Only terminal fragments and original compounds were kept as fragment molecule triplet (FMT) for model training and all chemical structures in the FMTs were translated to canonicalized SMILES format[46, 67] with RDKit[68].

Our ChEMBL data set were further divided into three sets with a ratio of 8:1:1 for training, validating, and testing, respectively. All FMTs were grouped by corresponding SLBD. When splitting the ChEMBL set into those three sets, a random sampling strategy was adopted to make sure the distribution of SLBD is similar among all three sets.

In addition, for further evaluating the generalization capability of our model, we also considered an external validation set derived from CASF-2016 data set[69], which consists of 285 protein-ligand complexes with high-quality crystal structures. The same data set was also employed by Imrie et al. for their work on DeLinker.

Lastly, our SyntaLinker method was applied on three case study examples from literature, which were also reported by Imrie et al, to demonstrate the capability of the model for doing fragment linking, lead optimization, and scaffold hopping. It is worthwhile to mention that the ground truth compounds in these examples were not included in the training set of our SyntaLinker models.

Evaluation metrics for SyntaLinker model. The ultimate goal of our model is to generate various molecules, containing two starting fragments. Therefore, four different metrics on 2D level, validity, uniqueness, recovery and novelty, were employed to compare generated molecules and their ground truth in test set.[70-71] Here, validity refers to the percentage of generated chemically valid molecules with two starting fragments; Novelty is the percentage of generated chemically valid molecules with novel linkers (not present in the training set); Uniqueness is the number of unique structures generated and recovery means the percentage of ground truth is generated among test set compounds. Formally,

$$Validity = \frac{\#\ of\ chemically\ valid\ SMILES\ with\ starting\ fragments}{\#\ of\ generated\ SMILES} \quad (4)$$

$$Uniqueness = \frac{\# \, of \; non - duplicate, \; valid \; structures}{\# \, of \; valid \; structures} \qquad (5)$$

$$Noverty = \frac{\# \, of \; novel \; linkers \; not \; in \; training \; set}{\# \, of \; unique \; structures} \qquad (6)$$

To mimic the FBDD scenario, the quality of compounds generated by SyntaLinker model at 3D level was also examined. In this case, the 3D conformations of compound are generated and docked into protein binding pocket, the root mean square deviation (RMSD) and shape and color combo-similarity score (SC) to the X-ray bound conformation of actual ligand are generated to evaluate model performance for selected cases. Here, RMSD is merely calculated among the two starting fragments of the X-ray and generated structures. The SC score is calculated by the pharmacophoric feature similarity[72] and the shape similarity[73] between the X-ray conformation of actual ligand and the docking pose of generated structure. The SC score is a float value in the range of [0,1] and the higher value is, the more similar the generated molecule is to the original ligand. For each molecule, the best similarity score among all docking conformers was taken as the SC score. In current study, converting SMILES to 3D conformation and docking were done by using Molecular Operating Environment (MOE) software[74]. The RMSD and SC score for each conformation were calculated via RDkit[68].

Model Training and Optimization of Hyperparameters. The SyntaLinker model was implemented using OpenNMT.[75] All scripts were written in Python[76] (version 3.7). We trained our models on GPU (Nvidia 2080Ti), and saved checkpoint per 1000 steps. The best hyperparameters were obtained based on the recovery metric of the ChEMBL validation set. We built our model with the best hyperparameters (shown in Table S1) and adopted the beam search procedure[77] to generate multiple candidates with a special beam width. All generated candidates were canonicalized using RDkit and compared to the ground-truth molecules.

Results and Discussion

As mentioned in the above section, we trained models with two different constraints, i.e. SLBD only and SLBD plus pharmacophore constraints and named models as SyntaLinker and SyntaLinker_multi respectively. Additionally, a reference model without using any constraints, SyntaLinker_n, was also trained for comparison. The performance of our models on ChEMBL test set and CASF validation set was examined.

Model Performance on ChEMBL Test Set. We first assessed the performance of models on ChEMBL test set using 2D metrics. Top 10 candidates for each pair of starting fragments were generated. The performance is demonstrated in Table 1. All models achieved over 95% validity, 90% linker novelty and recovered over 80% of the original molecules, demonstrating that the conditional transformer model can learn to identify the linker part of the structures, generate the linker accordingly, and also the models are generalized well enough to create new linkers. It seems that the constraint models are better than the model without using constraint in terms of recovery, novelty and validity. Especially, the most detailed constraints model SyntaLinker_multi achieves a recovery rate of 87.1% in the top10 recommendations. This is probably due to the fact that more prior knowledge about the linker are defined in the multiple constraints model than others.

Table 1. Performance comparison of models with different constraints on ChEMBL test set.

Metrics	Models		
	SyntaLinker	SyntaLinker_multi	SyntaLinker_n
Validity	97.2%	97.8%	96.0%
Uniqueness	88.1%	84.9%	86.7%
Recovery	84.7%	87.1%	80.0%
Novelty	91.8%	92.3%	90.3%

Model Performance on CASF. To compare with the DeLinker model (downloaded at https://github.com/oxpig/DeLinker), we further evaluated the models on the external validation set CASF-2016[69] which was used in the DeLinker model. Following the same sampling strategy as DeLinker, we generated 250 molecules for each pair of starting fragments. The detailed performance of various models on the same validation set is demonstrated in Table 2.

Table 2. Performance comparison of models on the CASF-2016 validation set.

Metrics	Models			
	DeLinker	SyntaLinker	SyntaLinker_multi	SyntaLinker_n
Validity	95.5%	96.4%	96.5%	86.8%
Uniqueness	51.9%	69.9%	63.8%	65.6%
Recovery	53.7%	62.7%	60.2%	55.4%
Novelty	51.0%	75.4%	77.2%	71.3%

These metrics indicate that the performance of our models is significantly improved over the DeLinker model on CASF validation set. Especially, our model improves the linker novelty of DeLinker model by a margin of 20% without losing the recovery, which means our model can sample a diverse range of linkers effectively.

Efficiency of Controlling Structure Generation. The aim of building constrained transformer model is to make sure generated structures can fulfill certain criteria. We further calculated the SLBD and pharmacophore properties of linkers in the generated molecules to evaluated whether these constraints achieved desirable control or not. The original linker bond length and pharmacophore properties of compounds in test sets were set as the control criterion when generating structures using constrained model SyntaLinker and SyntaLinker_multi, respectively, while for unconstrained model SyntaLinker_n, no control criterion was used. For structures generated by all three models, the linker length and pharmacophore properties were examined to compare the efficiency in controlling structure generation. Besides evaluating the control efficiency on ChEMBL test set and CASF validation set, the effect of beam search width (top 10 and top 250) was also assessed. The percentage of structures with correct shortest length of linker and structures whose linker length variation is less than one bond distance from all three SyntaLinker models on ChEMBL test set in top10 and top 250 are shown in Figure 3a and Figure 3b. As expected, the model with length constraints (SyntaLinker, SyntaLinker_multi) outperform the model without constraints (SyntaLinker_n), where 79.2%, 78.1% and 39.9% of structures have exact same linker length as the control for SyntaLinker, SyntaLinker_multi and SyntaLinker_n models respectively. If the allowed variation of link bond length is expanded to no more than one bond, the percentage of structures fulfilling the criteria from three models are 96.1%, 96.2% and 69% respectively. For CASF validation set, the same trend was observed in Figure 3c. It is worthwhile to point out that when we increase the beam search width from 10 to 250, the control efficiency of shortest bond length for all three models decrease as shown in Figure 3b and Figure 3d. But the advantage of constrained model versus unconstrained model is still obvious.

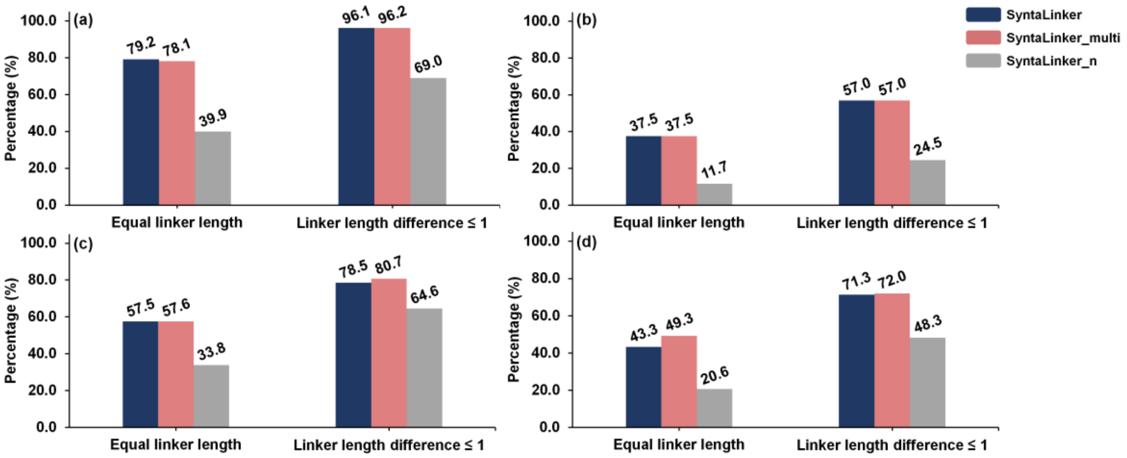


Figure 3. The comparison of efficiency in controlling SLBD for three SyntaLinker models. (a) The percentage of compound among top 10 solutions (beam search width of 10) fulfilling bond length criteria in ChEMBL test set; (b) The percentage of compound among top 250 solutions fulfilling bond length criteria in ChEMBL test set; (c) The percentage of compound among top 10 solutions fulfilling bond length criteria in CASF set; (d) The percentage of compound among top 250 solutions fulfilling bond length criteria in CASF set.

The percentage of structures with exactly equivalent pharmacophore properties to their ground truth of linker from all three SyntaLinker models on ChEMBL test set in top10 and top 250 are shown in Figure 4. As expected, the model with pharmacophore constraints (SyntaLinker_multi) outperform the model without this constraint (SyntaLinker, SyntaLinker_n), where 36.5%, 55.2% and 35.0% of structures have exact same pharmacophore properties as the control for SyntaLinker, SyntaLinker_multi and SyntaLinker_n models on ChEMBL test set respectively. For CASF validation set, the same trend was also observed. Furthermore, when we increased the beam search width from 10 to 250, the control efficiency of pharmacophore for all three models decreased. Due to the combination of multiple constraints, the control efficiency of pharmacophore constraints is lower than the control efficiency of bond length constraint (as shown in Figure 3), but in general the advantage of the constrained model over the unconstrained model remains.

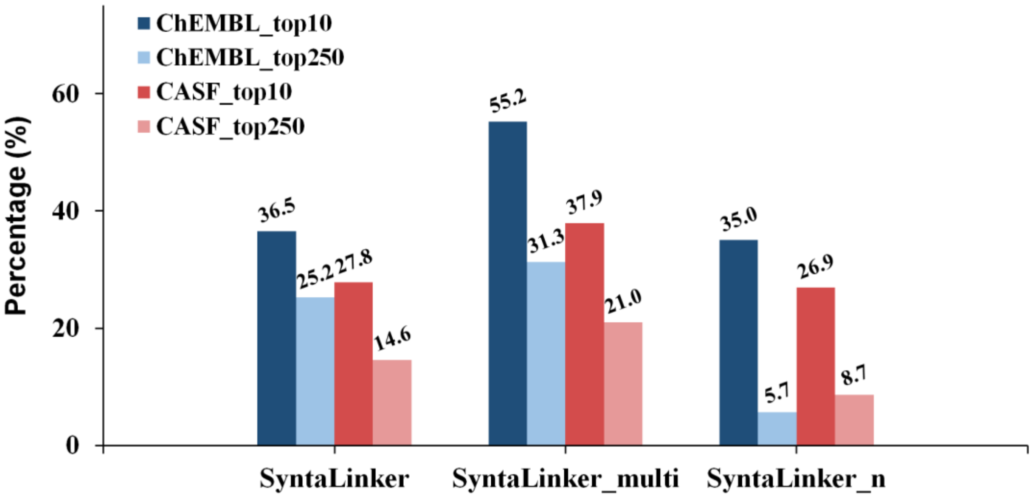


Figure 5. The efficiency in controlling pharmacophore properties of linker for various models in ChEMBL test set and CASF validation set with top10 and 250 solutions, the percentage of structures with exactly the same pharmacophore pattern of linker in generated molecules are compared.

Properties of the Generated Molecules. To evaluate the quality of generated compound from SyntaLinker models, Drug-likeness score (QED score), synthetic accessibility score (SAscore), the calculated water-octanol partition coefficient (logP) and molecule weight (MW) calculated in RDkit was used to characterize the quality of generated molecules. For each pair of starting fragments in ChEMBL test set, the properties of its top 10 and top 250 candidates were calculated and averaged to obtain a final value. A comparison with the properties of the original ChEMBL data (Figure 5) showed that molecules generated from various SyntaLinker models clearly have higher QED and lower SAscore than ChEMBL compounds and difference among various SyntaLinker models is quite small, while the distribution of logP and MW are similar among ChEMBL compounds and compounds generated by SyntaLinker models. It suggests that SyntaLinker generated compounds, at some extent, are more drug-like and have lower complexity for synthesis comparing with ChEMBL compounds. This is probably due to the fact that the chosen starting fragment pairs restraint the properties of generated structures.

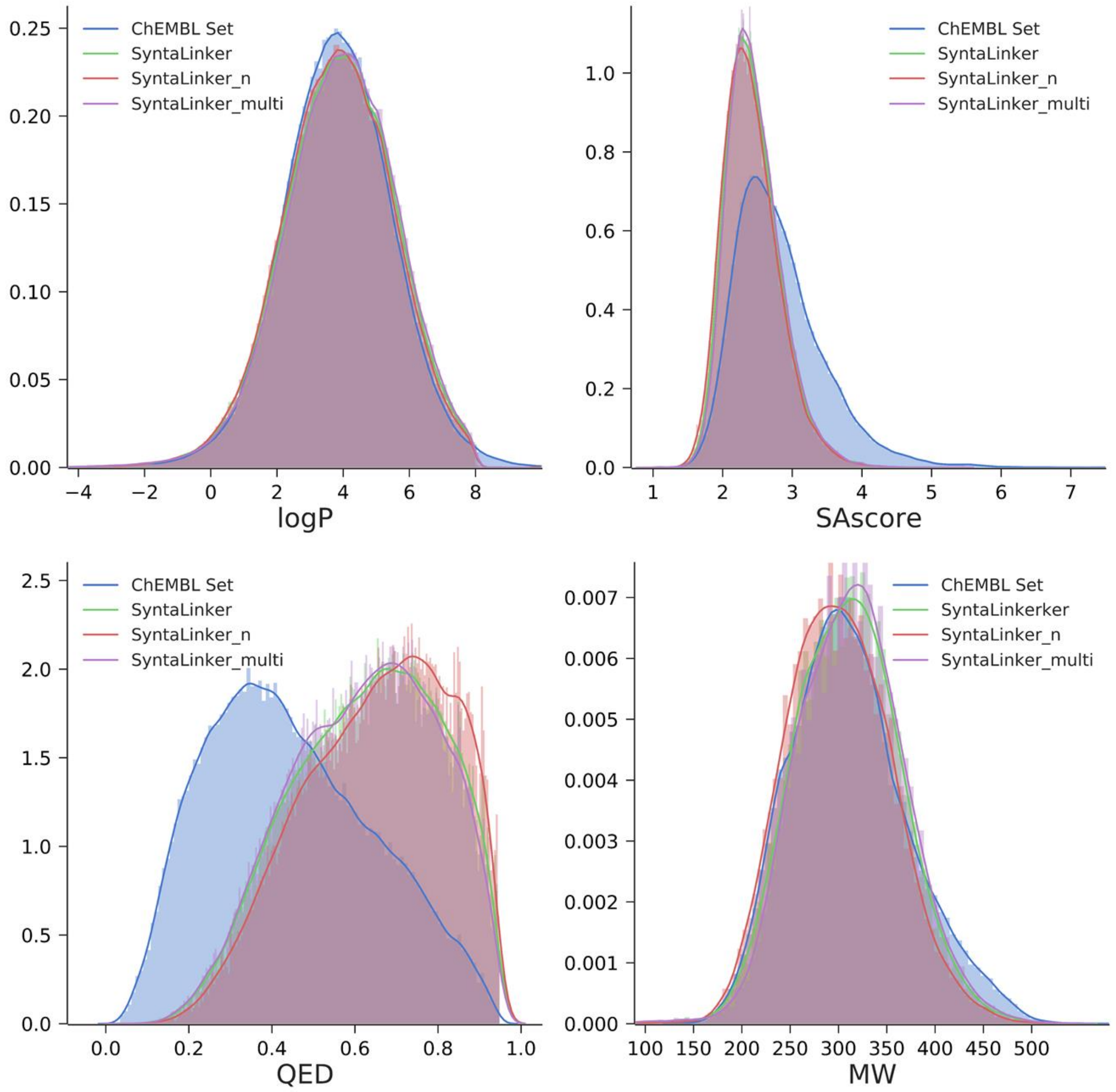


Figure 5. Distribution of chemical properties for ChEMBL set and generated molecules sets from different models.

Fragment Linking Case. In Medicinal chemistry, fragment linking is generally considered as an attractive approach for providing a positive impact on increasing affinity. Trapero et al. reported a successful fragment linking example.[78] They identified a low affinity phenylimidazole derivative through virtual screening target IMPDH, and then linked these fragments to form a larger compound with a more than 1000-fold boost in binding affinity. Inspired by their successful experiments, we employed their low affinity fragments (PDB ID: 5OU2) as our starting fragments and generated the linked structure (PDB ID: 5OU3) through our model (as shown in Figure 6).

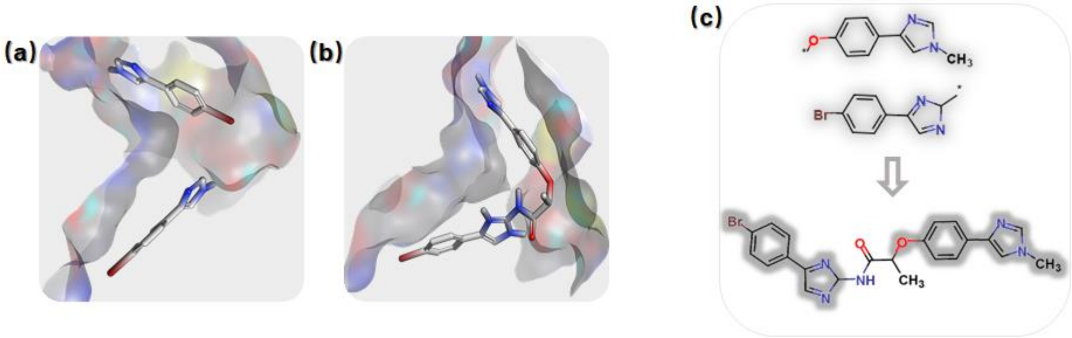


Figure 6. Fragment linking case study. (a) The binding poses of the two starting fragments (PDB 5OU2). (b) The bound conformation of the merged ligand in PDB 5OU3. (c) Chemical structure of starting fragments and active target compound.

Given the starting fragments as in 5OU2, we set the SLBD between 3 to 5 and generated 500 candidates. The ground truth target compound was successfully recovered and after 3D conformer generation and docking, most of the generated structures actually have a better docking score than the target compound (Table 3). Three generated molecules with the highest 3D fragment similarity and favorable MOE docking score are shown in Figure 7, where the docking pose of the recovered original lead by our SyntaLinker model is shown in Figure 7c.

Table 3. MOE docking score and 3D similarity metrics of generated molecules in fragment linking and lead optimization cases.

Metrics	Cases		
	Fragment Linking		**Lead Optimization**
	Top 100	Top 500	Top 100
Unique structures	47	341	808
MOE Score < Lead	35	306	107
MOE Score < -8.0	7	153	462
MOE Score < -9.0	1	22	22
RMSD < 2.0Å	22	152	179
SC > 0.5	9	36	44

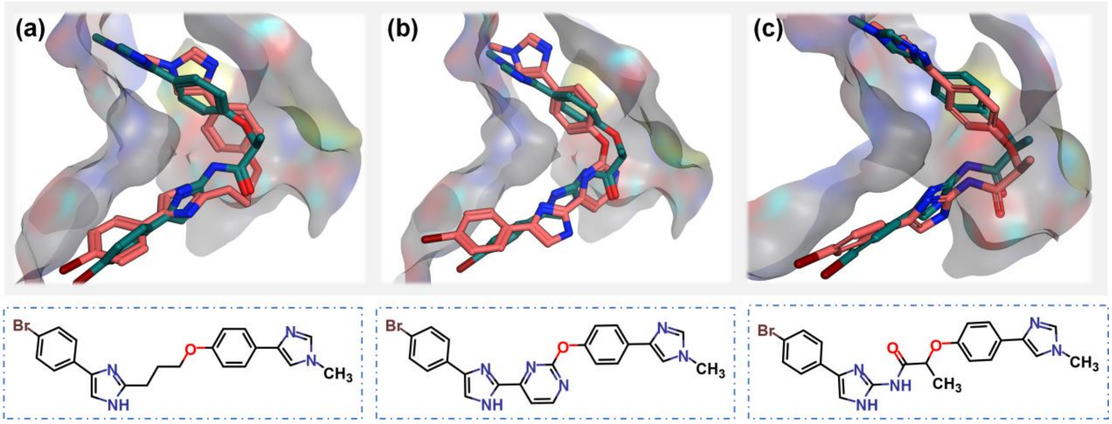


Figure 7. Overlay of the original conformer (PDB 5OU3, green carbons) and three generated molecules (pink carbons, chemical structures shown in blue boxes) with the highest 3D fragment similarity and highly MOE docking score.

Lead Optimization Case. Lead optimization is a iterative process of continuously modifying lead structures to improving potency and ADMET properties after initial lead compounds are identified. Here, we mimicked a typical lead optimization process via our SyntaLinker model. Dequalinium (Figure 8c) has an inhibitory effect of Chitinase A in the low nanomolar range (Ki: 70 nM), which makes it attractive for plausible development of therapeutics against human diseases involving chitinase-mediated pathologies.[79] Previous studies have clearly demonstrated its binding mode (PDB

ID: 3ARP), and proven that the two fragments (Figure 8d) connected by a decane linker is critical, due to occupying the hydrophobic areas of the binding pocked.

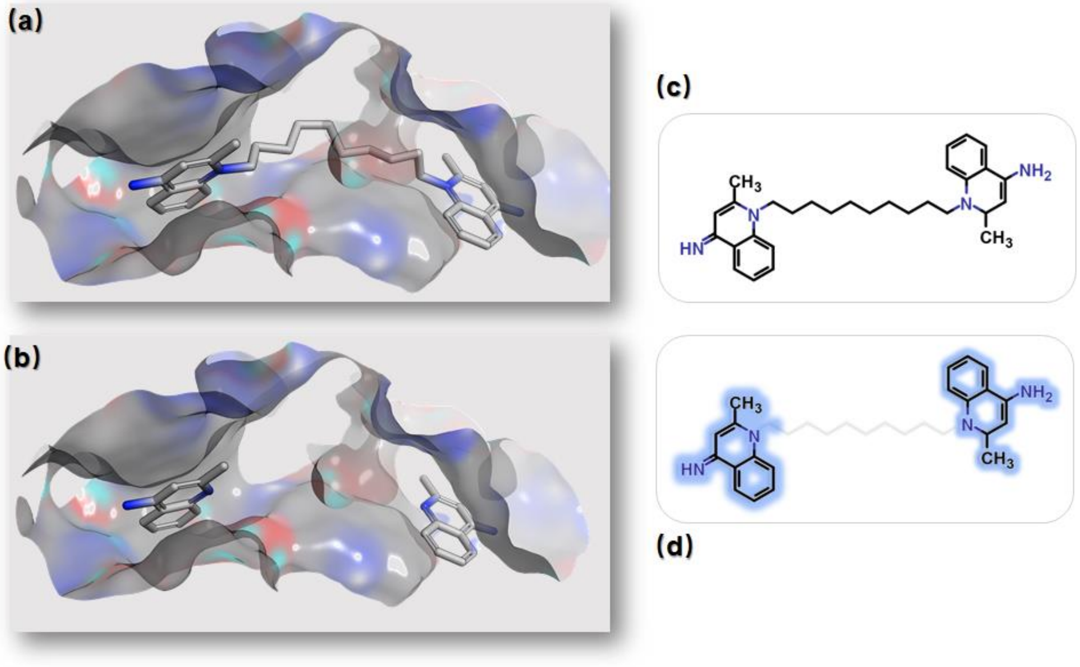


Figure 8. Structure information of lead optimization case study. (a) Surface representations of the complexes of the Chitinase A and dequalinium (PDB ID: 3ARP). (b) The binding poses of the two critical fragments in binding pocket. (c) Chemical structure of dequalinium. (d) Two critical fragments of dequalinium (mapped by blue shadow).

For optimizing the linker of dequalinium, we first broke down its structure using MMPs cutting algorithm and filtered generated fragments pairs as mentioned before, resulting 45 unique fragments pairs. All these pairs were employed as starting fragments to generate 100 candidates using our SyntaLinker model. The original molecule dequalinium was constantly recovered in all trials, and after removing redundant molecules, the RMSD of two critical fragments were calculated. Details are also shown in Table 3.

It seems that SyntaLinker model can generate a large number of novel molecules, where 179 and 30 compounds whose RMSD of starting fragment are less than 2 and 1Å respectively. While, the best RMSD in our hands of top the 5 compounds from DeLinker model is 2.5 Å. The most similar molecules in 3D to the lead compound are shown in Figure 9. Comparing with solutions found by DeLinker, our SyntaLinker model seems give chemically more reasonable compounds (three examples are shown in Figure 10) and also better RMSD and SC metrics. This may due to chemically more attractive linkers existed in our ChEMBL training set and also the ease of learning a sequence generation model comparing with the graph generation model.

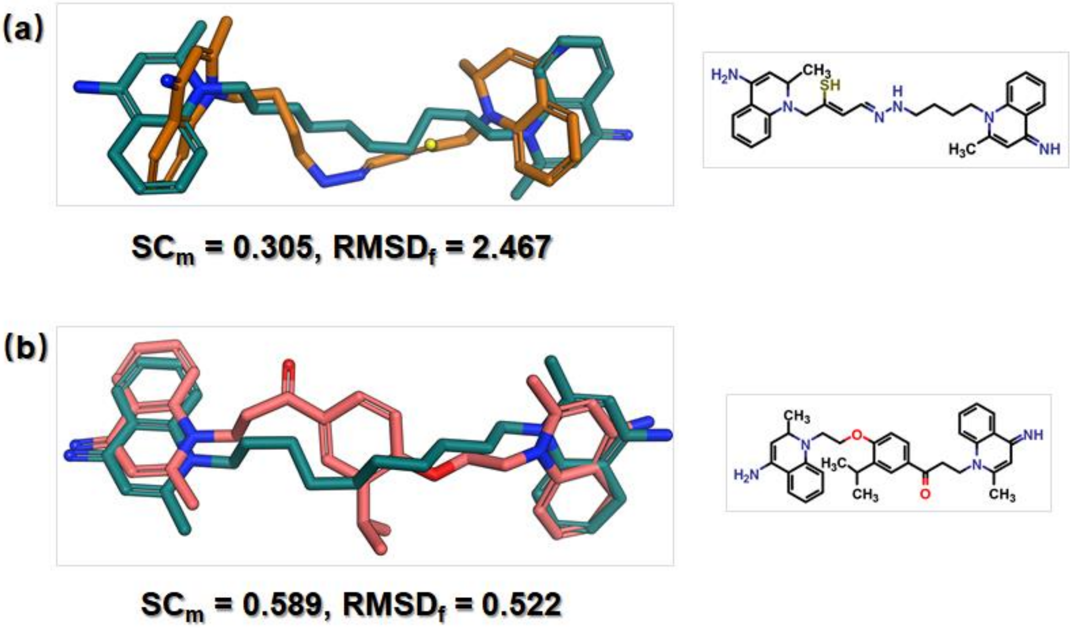


Figure 9. Examples of Overlays (2D chemical structure shown in right) and the 3D metrics of conformers from generated structures. The reference Dequalinium conformer is shown in green, while docked conformers of the generated molecule are shown in orange (DeLinker) or pink (our SyntaLinker model). (a) Overlays of reference conformer with the best DeLinker compound in terms of SC and RMSD score. (b) Overlays of reference conformer with the best SyntaLinker compound in terms of 3D metrics.

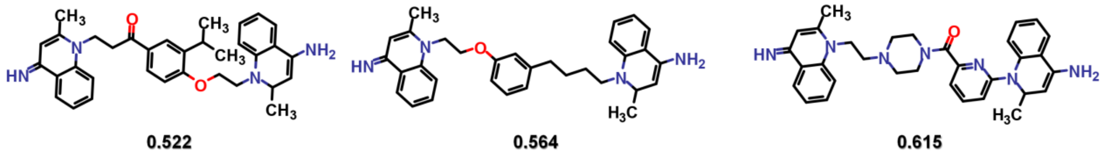


Figure 10. Examples of generated molecules and their RMSD values.

Scaffold Hopping Case. As a commonly used drug design strategy, scaffold hopping is a methodology that focuses explicitly on replacing central core of a template compound, while still keep more or less the same level of potency. Kamenecka et al. have reported an interesting study of scaffold hopping, which significantly improved selectivity toward its specific targets.[80] In this study, they aimed to design JNK3-selective inhibitors that had more than 1000-fold selectivity over p38. Eventually, they obtained aminopyrazole based inhibitors (PDB ID 3FI2) instead of the original indazole based inhibitors (PDB ID 3FI3), and improved the selectivity to 2800 fold. These two scaffolds have nearly identical binding modes to JNK3 (Figure 11). We start with the indazole class inhibitor, and aim to use our SyntaLinker model to reproduce this successful scaffold hopping process. We first set the SLBD range from 6 to 8, and 2500 molecules were generated. There were 2138 unique molecules. Among them, more than 1000 molecules with a RMSD value less than 1Å and SC value greater than 0.5. In total we identified 634 novel linkers not included in the training set, covering 186 unique Murcko scaffolds[81]. Among them, encouragingly, both indazole (Figure 12a) and aminopyrazole class linkers (Figure 12b) were recovered. Additionally, some novel linker scaffolds were also identified. This result highlights that our SyntaLinker model is well generalized to design novel linkers by combining the chemical information of starting fragments, not merely remembering the linkers in the ChEMBL training set. Several of examples are shown in Figure 12. All scaffolds showed good overlap with the original indazole linker, while maintaining the conformation of starting fragments in the molecule.

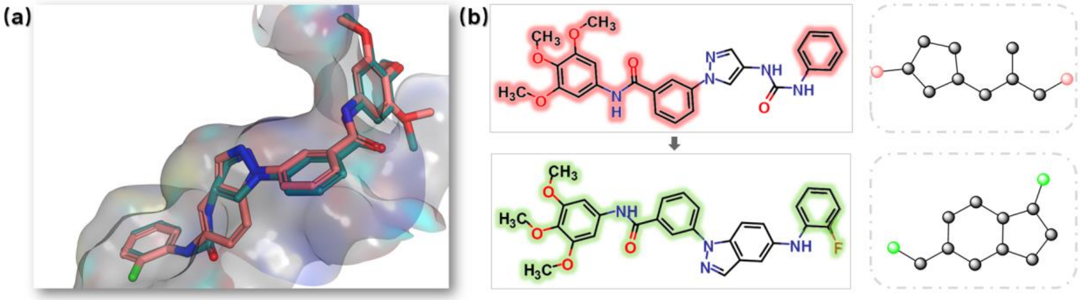


Figure 11. Scaffold hopping case study. (a) Overlay of the indazole (PDB 3FI3, green) and aminopyrazole (PDB 3FI2, pink) structures. (b) Chemical structure and Murcko scaffold of the indazole (upper) and aminopyrazole (down) compounds. The highlighted part of compounds is set as the starting fragments.

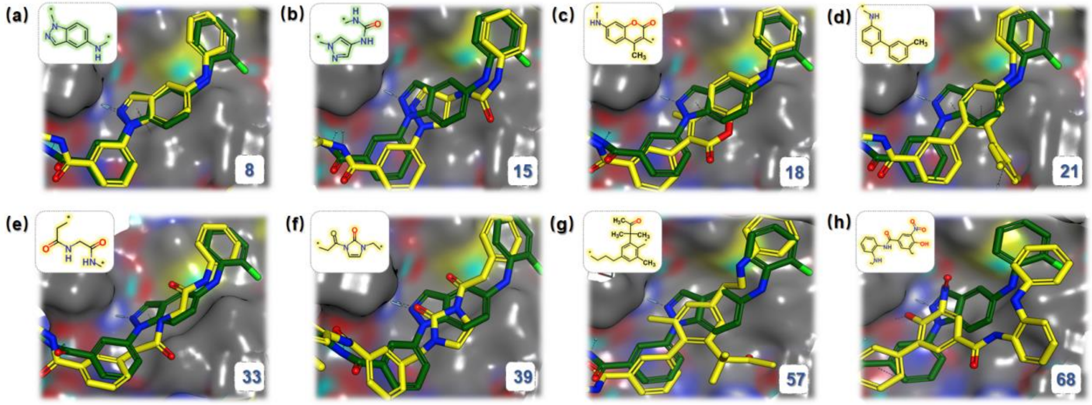


Figure 12. Overlay of the indazole inhibitor (PDB 3FI3, green) and several example structures (yellow) with high 3D similarity. The linker structures are shown (novel scaffolds colored by yellow, the recovered ground truth scaffolds colored by green) in upper left, and the order number (sorted by SC score) are shown in bottom right.

Conclusion

In current study, we have proposed a novel deep generative model, SyntaLinker, for fragment linking. Unlike previous attempts in generating linkers with graph convolution neural network, our method is trained purely at 2D level, and does not need to search 3D conformational databases and make complex conformation analysis. A novel conditional transformer neural network architecture was proposed to generate compounds by learning the linker structure information given the starting fragment pairs. Simply setting the shortest bond distance of linker between the starting fragments as constraint, a large number of novel drug-like candidates satisfying the constraint can be obtained. We have demonstrated that our generative model is able to learn and infer novel linkers that match the pre-defined constraints. In addition, we also build models with multiple pharmacophore constraints for doing more specific linker design. More importantly, through several case study examples, we have shown that our method can be applied on fragment linking, lead optimization and scaffold hopping tasks. Most of the generated molecules have better docking score than the original hits, while maintaining similar binding modes and possessing high 3D similarity to the bound conformation in X-ray structure. It is expected that SyntaLinker can become a useful tool in fragment-based lead generation and scaffold hopping process.

Reference

1. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S., Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery* **2011,** *10* (3), 188-195.

2. Ecker, D. J.; Crooke, S. T., Combinatorial Drug Discovery: Which Methods Will Produce the Greatest Value? *Bio/Technology* **1995,** *13* (4), 351-360.

3. Hajduk, P. J.; Greer, J., A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov* **2007,** *6* (3), 211-219.

4. Fattori, D.; Squarcia, A.; Bartoli, S., Fragment-based approach to drug lead discovery: overview and advances in various techniques. *Drugs R D* **2008,** *9* (4), 217-227.

5. Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I., Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery* **2003,** *2* (5), 369-378.

6. Murray, C. W.; Rees, D. C., The rise of fragment-based drug discovery. *Nature Chemistry* **2009,** *1* (3), 187-192.

7. Baell, J. B.; Holloway, G. A., New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry* **2010,** *53* (7), 2719-2740.

8. Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A., Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *Journal of medicinal chemistry* **2010,** *53* (1), 37-51.

9. Hajduk, P. J., Puzzling through fragment-based drug design. *Nature Chemical Biology* **2006,** *2* (12), 658-659.

10. Hajduk†, P. J., Fragment-Based Drug Design: How Big Is Too Big? *Journal of Medicinal Chemistry* **2006,** *49* (24), 6972-6976.

11. Baker, M., Fragment-based lead discovery grows up. *Nature Reviews Drug Discovery* **2013,** *12* (1), 5-10.

12. Jencks, W. P., On the Attribution and Additivity of Binding Energies. *Proceedings of the National Academy of Sciences of the United States of America* **1981,** *78* (7), 4046-4050.

13. Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H., Twenty years on: the impact of fragments on drug discovery. *Nature Reviews Drug Discovery* **2016,** *15* (9), 605-619.

14. Davies, T. G.; Tickle, I. J., Fragment Screening Using X-Ray Crystallography. In *Fragment-Based Drug Discovery and X-Ray Crystallography*, Davies, T. G.; Hyvönen, M., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, **2012**; pp 33-59.

15. Chen, H.; Zhou, X.; Wang, A.; Zheng, Y.; Gao, Y.; Zhou, J., Evolutions in fragment-based drug design: the deconstruction–reconstruction approach. *Drug Discovery Today* **2015,** *20* (1), 105-113.

16. Zhang, W., Fragment Informatics and Computational Fragment-Based Drug Design: An Overview and Update. *Medicinal Research Reviews* **2013,** *33* (3), 554-598.

17. Joseph-McCarthy, D.; Campbell, A. J.; Kern, G.; Moustakas, D., Fragment-Based Lead Discovery and Design. *Journal of Chemical Information and Modeling* **2014,** *54* (3), 693-704.

18. Chen, H.; Knerr, L.; Åkerud, T.; Hallberg, K.; Öster, L.; Rohman, M.; Österlund, K.; Beisel, H.-G.; Olsson, T.; Brengdhal, J.; Sandmark, J.; Bodin, C., Discovery of a novel pyrazole series of group X secreted phospholipase A2 inhibitor (sPLA2X) via fragment based virtual screening. *Bioorg Med*

Chem Lett **2014,** *24* (22), 5251-5255.

19. Rees, D. C., Fragment-Based Lead Discovery. *Annual Reports in Medicinal Chemistry* **2007,** *42*, 431-448.

20. Möbitz, H.; Machauer, R.; Holzer, P.; Vaupel, A.; Stauffer, F.; Ragot, C.; Caravatti, G.; Scheufler, C.; Fernandez, C.; Hommel, U.; Tiedt, R.; Beyer, K. S.; Chen, C.; Zhu, H.; Gaul, C., Discovery of Potent, Selective, and Structurally Novel Dot1L Inhibitors by a Fragment Linking Approach. *ACS Medicinal Chemistry Letters* **2017,** *8* (3), 338-343.

21. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W., Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996,** *274* (5292), 1531.

22. Medek, A.; Hajduk, P. J.; Mack, J.; Fesik, S. W., The Use of Differential Chemical Shifts for Determining the Binding Site Location and Orientation of Protein-Bound Ligands. *Journal of the American Chemical Society* **2000,** *122* (6), 1241-1242.

23. Mondal, M.; Radeva, N.; Fanlo-Virgós, H.; Otto, S.; Klebe, G.; Hirsch, A. K. H., Fragment Linking and Optimization of Inhibitors of the Aspartic Protease Endothiapepsin: Fragment-Based Drug Design Facilitated by Dynamic Combinatorial Chemistry. *Angew Chem Int Ed Engl* **2016,** *55* (32).

24. Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N., Entropic Contribution to the Linking Coefficient in Fragment Based Drug Design: A Case Study. *Journal of Medicinal Chemistry* **2010,** *53* (10), 4285-4289.

25. Chodera, J. D.; Mobley, D. L., Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annual Review of Biophysics* **2013,** *42* (1), 121-142.

26. Ichihara, O.; Barker, J.; Law, R. J.; Whittaker, M., Compound Design by Fragment-Linking. *Molecular Informatics* **2011,** *30* (4), 298-306.

27. Glick, M., "Virtual Fragment Linking": An Approach To Identify Potent Binders from Low Affinity Fragment Hits. *Journal of Medicinal Chemistry* **2008,** *51* (8), 2481-2491.

28. Chung, S.; Parker, J. B.; Bianchet, M.; Amzel, L. M.; Stivers, J. T., Impact of linker strain and flexibility in the design of a fragment-based inhibitor. *Nature Chemical Biology* **2009,** *5* (6), 407-413.

29. Fedorov, D. G.; Kitaura, K., Pair interaction energy decomposition analysis. *Journal of Computational Chemistry* **2007,** *28* (1), 222-237.

30. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M., Fragment molecular orbital method: an approximate computational method for large molecules. *Chemical Physics Letters* **1999,** *313* (3), 701-706.

31. Fedorov, D. G.; Kitaura, K., Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *The Journal of Physical Chemistry A* **2007,** *111* (30), 6904-6914.

32. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. *Drug Discovery Today* **2018,** *23* (6), 1241-1250.

33. Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J., Deep learning for molecular generation. *Future Medicinal Chemistry* **2019,** *11* (6), 567-597.

34. Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W., Deep learning for molecular generation and optimization - a review of the state of the art. *CoRR* **2019,** *abs/1903.04388*.

35. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* **2017,** *9* (1), 48.

36. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating Focussed Molecule Libraries

for Drug Discovery with Recurrent Neural Networks. *CoRR* **2017,** *abs/1701.01329*.

37. Gómez-Bombarelli, R.; Duvenaud, D.; Hernández-Lobato, J. M.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic chemical design using a data-driven continuous representation of molecules. *CoRR* **2016,** *abs/1610.02415*.

38. Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H., A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics* **2019,** *11* (1), 74.

39. Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. S., Learning Multimodal Graph-to-Graph Translation for Molecular Optimization. *CoRR* **2018,** *abs/1812.01070*.

40. Zhou, Z.; Kearnes, S. M.; Li, L.; Zare, R. N.; Riley, P., Optimization of Molecules via Deep Reinforcement Learning. *CoRR* **2018,** *abs/1810.08678*.

41. Fu, T.; Xiao, C.; Sun, J., CORE: Automatic Molecule Optimization Using Copy & Refine Strategy. *ArXiv* **2020,** *abs/1912.05910*.

42. Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. In *Recurrent neural network based language model*, INTERSPEECH, 2010.

43. Kingma, D. P.; Welling, M., Auto-Encoding Variational Bayes. *CoRR* **2014,** *abs/1312.6114*.

44. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I. J., Adversarial Autoencoders. *ArXiv* **2015,** *abs/1511.05644*.

45. Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; Bengio, Y., Generative Adversarial Networks. *ArXiv* **2014,** *abs/1406.2661*.

46. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988,** *28* (1), 31-36.

47. 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**.

48. Davies, R. H.; Smith, D. A.; McNeillie, D. J.; Morris, T. R., Identification of biologically active conformations in flexible drug molecules. *International Journal of Quantum Chemistry* **1979,** *16* (S6), 203-221.

49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I., Attention Is All You Need. *CoRR* **2017,** *abs/1706.03762*.

50. Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V., Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* **2017,** *3* (10), 1103-1113.

51. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A., Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019,** *5* (9), 1572-1583.

52. Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y., Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *Journal of Chemical Information and Modeling* **2020,** *60* (1), 47-55.

53. Zweig, G.; Platt, J. C.; Meek, C.; Burges, C. J. C.; Yessenalina, A.; Liu, Q., Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Association for Computational Linguistics: Jeju Island, Korea, **2012**; pp 601–610.

54. Pfaff, C. W., Constraints on Language Mixing: Intrasentential Code-Switching and Borrowing in

Spanish/English. *Language* **1979,** *55* (2), 291-318.
55. Shana, P., Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching1. *Linguistics* **1980,** *18* (7-8), 581-618.
56. Chen, T.; Xu, R.; Lu, Q.; Liu, B.; Xu, J.; Yao, L.; He, Z. In *A Sentence Vector Based Over-Sampling Method for Imbalanced Emotion Classification*, Computational Linguistics and Intelligent Text Processing, Berlin, Heidelberg, 2014//; Gelbukh, A., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, **2014**; pp 62-72.
57. Su, S.; Yang, Y.; Gan, H.; Zheng, S.; Gu, F.; Zhao, C.; Xu, J., Predicting the Feasibility of Copper(I)-Catalyzed Alkyne–Azide Cycloaddition Reactions Using a Recurrent Neural Network with a Self-Attention Mechanism. *Journal of Chemical Information and Modeling* **2020,** *60* (3), 1165-1174.
58. Nair, V.; Hinton, G. E., In *ICML*, **2010**.
59. Ba, J.; Kiros, J. R.; Hinton, G. E., Layer Normalization. *ArXiv* **2016,** *abs/1607.06450*.
60. Barrault, L.; Bojar, O. e.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; Monz, C.; Müller, M.; Pal, S.; Post, M.; Zampieri, M. In *Findings of the 2019 Conference on Machine Translation (WMT19)*, Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Florence, Italy, aug; Association for Computational Linguistics: Florence, Italy, **2019**; pp 1-61.
61. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. *CoRR* **2015,** *abs/1512.03385*.
62. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011,** *40* (D1), D1100-D1107.
63. Lipinski, C.; Hopkins, A., Navigating chemical space for biology and medicine. *Nature* **2004,** *432* (7019), 855-861.
64. Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009,** *1* (1), 8.
65. Hussain, J.; Rea, C., Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of chemical information and modeling* **2010,** *50* (3), 339-348.
66. 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* (8), 644-644.
67. Weininger, D.; Weininger, A.; Weininger, J. L., SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989,** *29* (2), 97-101.
68. Landrum, G. RDKit: Open-source cheminformatics, http://www.rdkit.org (accessed December 20, 2018).
69. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **2019,** *59* (2), 895-913.
70. Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C., GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling* **2019,** *59* (3), 1096-1108.
71. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Nikolenko, S. I.; Aspuru-Guzik, A.; Zhavoronkov, A., Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *CoRR* **2018,** *abs/1811.12823*.
72. Putta, S.; Landrum, G. A.; Penzotti, J. E., Conformation Mining: An Algorithm for Finding

Biologically Relevant Conformations. *Journal of Medicinal Chemistry* **2005,** *48* (9), 3313-3318.

73. Landrum, G. A.; Penzotti, J. E.; Putta, S., Feature-map vectors: a new class of informative descriptors for computational drug discovery. *Journal of Computer-Aided Molecular Design* **2006,** *20* (12), 751-762.

74. MOE; Chemical Computing Group: 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7. http://www.chemcomp.com (accessed February 16, 2020) .

75. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M., OpenNMT: Open-Source Toolkit for Neural Machine Translation. *CoRR* **2017,** *abs/1701.02810*.

76. Python Core Team. Python: A dynamic, open source programming language. Python Software Foundation. URL https://www.python.org/.

77. Ow, P. S.; Morton, T. E., Filtered beam search in scheduling†. *International Journal of Production Research* **1988,** *26* (1), 35-62.

78. Trapero, A.; Pacitto, A.; Singh, V.; Sabbah, M.; Coyne, A. G.; Mizrahi, V.; Blundell, T. L.; Ascher, D. B.; Abell, C., Fragment-Based Approach to Targeting Inosine-5'-monophosphate Dehydrogenase (IMPDH) from Mycobacterium tuberculosis. *Journal of medicinal chemistry* **2018,** *61* (7), 2806-2822.

79. Pantoom, S.; Vetter, I. R.; Prinz, H.; Suginta, W., Potent Family-18 Chitinase Inhibitors: X-RAY STRUCTURES, AFFINITIES, AND BINDING MECHANISMS. *Journal of Biological Chemistry* **2011,** *286* (27), 24312-24323.

80. Kamenecka, T.; Habel, J.; Duckett, D.; Chen, W.; Ling, Y. Y.; Frackowiak, B.; Jiang, R.; Shin, Y.; Song, X.; LoGrasso, P., Structure-activity relationships and X-ray structures describing the selectivity of aminopyrazole inhibitors for c-Jun N-terminal kinase 3 (JNK3) over p38. *J Biol Chem* **2009,** *284* (19), 12853-12861.

81. Bemis, G. W.; Murcko, M. A., The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996,** *39* (15), 2887-2893.