A deep learning based scaffold hopping strategy for the design of kinase inhibitors

Lizhao Hu[a,c], Yuyao Yang[b,c], Shuangjia Zheng[d], Jun Xu[a,c,], Ting Ran[b,*], Hongming Chen[b,*]*

[a]School of Biotechnology and Health Sciences, Wuyi University, Jiangmen 529020, China.

[b]Bioland Laboratory (Guangzhou Regenerative Medicine and Health - Guangdong Laboratory)，Guangzhou 510530, China

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

[d]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

[*]Corresponding authors:

ran_ting@grmh-gdl.cn; junxu@biochemomes.com; chen_hongming@grmh-gdl.cn

KEYWORDS: Scaffold hopping, Kinase inhibitor, Chemical informatics, Deep learning

ABSTRACT: Protein kinase family has become the hot spot for drug discovery especially in the area of cancer therapy. Many kinase inhibitors targeting the ATP-binding pocket of kinase domain have been approved in past decades. Scaffold hopping is a widely used strategy for drug design towards kinase inhibitors. Particularly for kinase targets, the high conservation of the ATP-binding pocket across all kinases provides abundant opportunities for scaffold hopping. In this study, as an extension of previously reported SyntaLinker generative model, we developed a fragment-based deep learning workflow, named SyntaLinker-Hybrid, for scaffold hopping purpose specifically for replacing the molecular segments bound at the conserved kinase hinge region. Through this automated workflow, new chemical structures can be generated by hybridizing novel hinge binding fragments with motifs of existing kinase inhibitors bound at the sub-pockets of non-hinge regions. Our study showed that this strategy could effectively generate novel kinase inhibitors, while to a great extent retaining the binding characteristics of existing kinase inhibitors. We expect that it could be a useful tool for making 'me-too' design especially for those heavily patented kinase targets.

Introduction

Kinase is a large protein family that plays a key role in the regulation of cellular signal transduction[1]. Aberrantly high expression and gene mutations of kinases usually cause the disorder of cell signaling pathways, which has been acknowledged to be responsible for

several human pathological conditions, such as cancer and inflammation[2-6]. Kinase family is a highly druggable target family. Up till now, tens of kinase inhibitors have been successfully developed as marketed drugs for oncology and inflammation diseases[7]. Majority of approved kinase inhibitors are bound to the ATP-binding pocket of kinase domain and as a matter of fact, many kinase inhibitors are actually targeting multiple kinases due to the high structural conservation of the ATP-binding pockets among kinases[8-12]. Due to its druggable nature, development of kinase inhibitor is a highly competitive area and thousands of chemical patents were filed for various kinase targets[10]. The high structural conservation among the kinase ATP-binding pockets also contributes to the fact that the structural novelty of hinge binding motif of kinase inhibitor is actually quite limited and some known hinge binding scaffolds are very crowded spaces for patenting. Therefore, achieving structural novelty is quite crucial in the development of kinase inhibitors due to the intensive competition. In this study, our interest is to explore the chemical space of kinase inhibitors from deep learning perspective, and provide a new strategy for scaffold hopping in the design of novel kinase inhibitors.

Many reviews have summarized the structural characteristics of kinase inhibitors, which showed that the binding of kinase inhibitors are dependent on the interactions occurring at several key residues in the ATP-binding pocket[10, 13, 14]. The highly conserved hinge loop linking the N-terminal lobe and C-terminal lobe of the kinase domain seems to be essential for the binding of kinase inhibitors. Thus, scaffold hopping at the hinge region as a strategy was widely used in the design of novel kinase inhibitors.

In general, scaffold hopping, at a large extent, was achieved by designing bioisosteric replacement and this was usually done by medicinal chemist manually[15]. Some computational methods were also proposed to carry out automatic replacement in a large scale. [16] For example, a scaffold specific shape descriptor was defined in the CAVEAT program to replace the core fragment in active compounds. Ligand-based pharmacophore and molecular fingerprints have also been applied for scaffold hopping[17, 18]. Biological fingerprints encoding the biological profiles of compounds were emerging for scaffold hopping despite the lack of structural and chemical information[19]. In addition, various machine learning methods, for example self-organizing maps, can visualize the similarity distribution of compounds and have been used for scaffold hopping[20-22]. Fragment centric scaffold hopping methods were also proposed previously. Mukherjee et al. extracted the smallest fragments forming hydrogen bonds to the hinge residues in each kinase-ligand crystal structure and applied these fragments for substructure searching, which successfully led to the identification of novel kinase inhibitors[23]. Fragment linking or growing based on known binding fragments is also a way to design novel compounds. Dey *et al* developed a tool named GANDI (Genetic Algorithm-based de Novo Design of Inhibitors), which automatically linked the fragments docked at different sub-pockets with high two-dimensional (2D) or three-dimensional (3D) similarity to known active compounds[19]. Recently, Sydow *et al* built a KinFragLib fragment library by decomposing known kinase inhibitors in the KLIFS database according to their spacial positions in the ATP-binding pockets[24]. More specifically, these compounds were broken down according to the sub-

pockets composing the ATP-binding site so that the resulted fragments can be classified into groups with sub-pocket ID as class label. Recombination of fragments across sub-pockets can then generate millions of virtual compounds with high chemical novelty and diversity, while their substructures were still originated from known kinase inhibitors. This method showed good performance on target specificity via automatic fragment linking.

Recent years, advances in the development of deep generative models have spawned a mass of promising methods to address the structure generation issue in drug design[25]. The deep generative models have been applied in de novo molecular design[26-28] and lead optimization[29-31]. Methods for fragment linking based on generative model have also been proposed[32, 33]. We previously proposed a deep learning based method SyntaLinker to automatically assembly given fragments without using any predefined chemical rules or motifs[35]. Inspired by Sydow's work, here we extended the SyntaLinker methodology and proposed a workflow, SyntaLinker-Hybrid, to automatically connect two terminal fragments designated as the binding motifs of specific ATP sub-pockets with novel hinge binding fragments.

Firstly, a kinase-focused deep generative model was constructed by SyntaLinker algorithm using known kinase inhibitor data set; Then, the fragments bound at the non-hinge sub-pockets extracted from the KinFragLib data set were used as the input terminal fragment pairs of SyntaLinker model and subsequently hinge binding linkers were generated by the SyntaLinker model to assemble into new molecules. The quality of the generated molecules was investigated regarding chemical space and structural similarity to

known inhibitors. In order to further evaluate the capability of SyntaLinker-Hybrid on scaffold hopping for hinge binding motif, chemical space of generated linker fragments by SyntaLinker-Hybrid was compared to that of the fragments bound at the hinge region in KinFragLib. Additionally, the binding features of the linkers were studied by molecular docking. The method was finally applied to the scaffold hopping of CDK9 inhibitors. Our results showed that the proposed workflow is able to generate novel and diverse kinase scaffolds and represents a new strategy for designing kinase inhibitors.

Materials and Methods

Data collection and preprocessing

To collect a kinase data set for training the SyntaLinker model, 429,764 kinase inhibitors were retrieved from the ChEMBL database[34], which in total consists more than 2 million bioactive compounds collected from various sources. They were downloaded as a SDF file and for molecules with multiple components, the largest component was retained as molecule structure. Then, the compounds with less than 2 rotatable single bonds were discarded to ensure successful compound decomposition in the following preparation steps. The compounds whose kinase activity (IC_{50}, Ki, Kd) larger than 1 μM were also filtered out, which resulted to 101,520 compounds targeting 366 kinases. In addition, duplicate structures were removed as multiple activities can exist for the ChEMBL compounds. Finally, 50,420 unique compounds were collected and named as the KID data set. As described in the previous report[33], each compound was then decomposed into three parts, i.e. a linker and two terminal fragments, using the MMPs cutting algorithm[35], and was transformed into a quadruple form like "fragment 1, linker, fragment 2, molecule", which was used as the input form for model construction.

On the other hand, following Sydow's protocol, kinase terminal fragments were extracted from the KinFragLib fragment library, in which the fragments were generated by applying sub-pocket based fragmentation protocol on available kinase crystal structures collected in the KLIFS database. The KinFragLib fragmentation method was carried out

as following: the compounds bound with the DFG-in conformations of kinases were split into fragments with respect to their 3D proximity to the six predefined sub-pockets in the ATP binding site such as adenine pocket (AP), solvent exposed pocket (SE), front pocket (FP), gate area (GA), back pocket 1 (B1) and back pocket 2 (B2) (**Figure 1A**). The fragments were classified according to their belonging sub-pocket ID and dummy atoms were added to the anchor points of each fragment. Here, only the fragments bound at non-AP sub-pockets were regarded as terminal fragments, because the AP sub-pocket is located at the hinge region which the scaffold hopping operation should focus on. In total, we chose 1560 fragments bound at the SE sub-pocket, 1024 GA sub-pocket fragments, 98 B1 sub-pocket fragments and 92 B2 sub-pocket fragments for fragment hybridization (**Figure 1B**). For fragments with multiple anchoring sites, all sites were used for hybridization. The hybridization scripts of Sydow's work was used to make library enumeration and the generated structures were named as the K-MOL set.

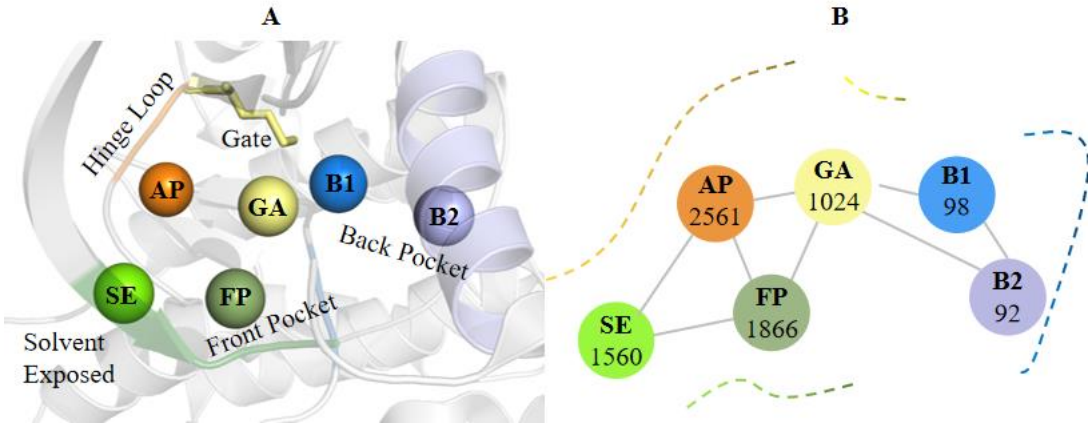

Figure 1. (**A**) 3D spatial locations of AP, GA, SE, FP, B1, B2 sub-pockets in the ATP-binding pocket. The protein structure was shown as gray cartoon, while the local

structures corresponding to the AP, FP, B2 sub-pockets were colored in green, orange and purple, respectively. The gate residue was shown as yellow stick. (**B**) 2D distribution of the six sub-pockets labeled by the number of fragments in the KinFragLib library.

In addition, 11 highly active CDK9 inhibitors were chosen as a test case to evaluate the effectiveness of the deep learning model on target-specific scaffold hopping[36]. These inhibitors were manually decomposed into a terminal fragment pair and one linker fragment according to the sub-pockets defined in the KinFragLib method (**Table S1**). The terminal fragment pairs were used for sampling the SyntaLinker model.

SyntaLinker model

Recently, we proposed the SyntaLinker algorithm for fragment linking (**Figure 2**). Here we provide a brief introduction on the method. The details should be refer to the original literature[33]. SyntaLinker is a generative model based on the syntactic pattern recognition using deep conditional transformer neural network. The implicit rules of linking fragments in known molecules were learnt by recognizing the syntax patterns embedded in the SMILES notation. Firstly, a large amount of compounds were decomposed into terminal fragments and linkers, which were then used to train transformer models to learn fragment linking rules. In the process, each molecule was split into a terminal fragment pair and a linker fragment and dummy atoms were used to label anchor points. Additionally, the shortest 2D connectivity distance between the two anchor points of the linker can also be

used as a constraint for liner generation. Once the SyntaLinker Model was trained, it can be sampled to output full compound which comprises the input terminal fragments and a linker.

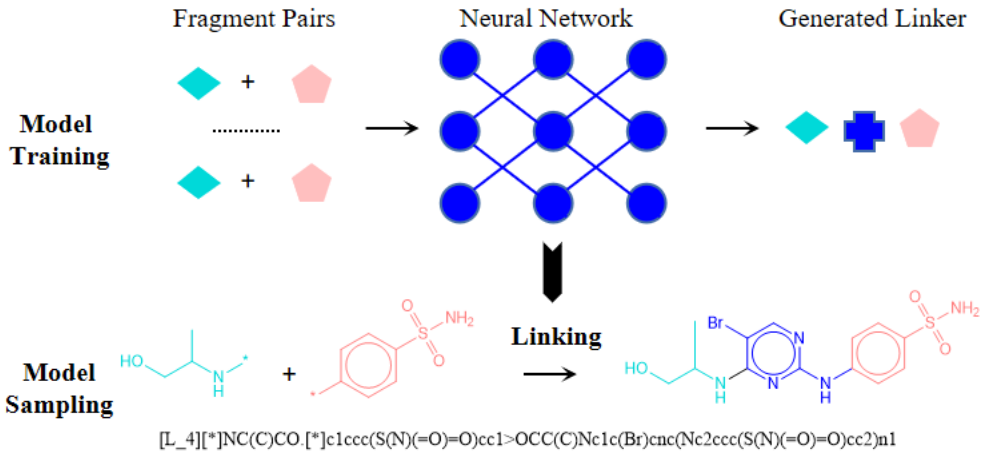

Figure 2. Illustration of training and sampling of a SyntaLinker model. An example of SMILES transformation from the source sequence to the target sequence with a constraint of L_4 was shown at the bottom of the picture.

SyntaLinker-Hybrid workflow

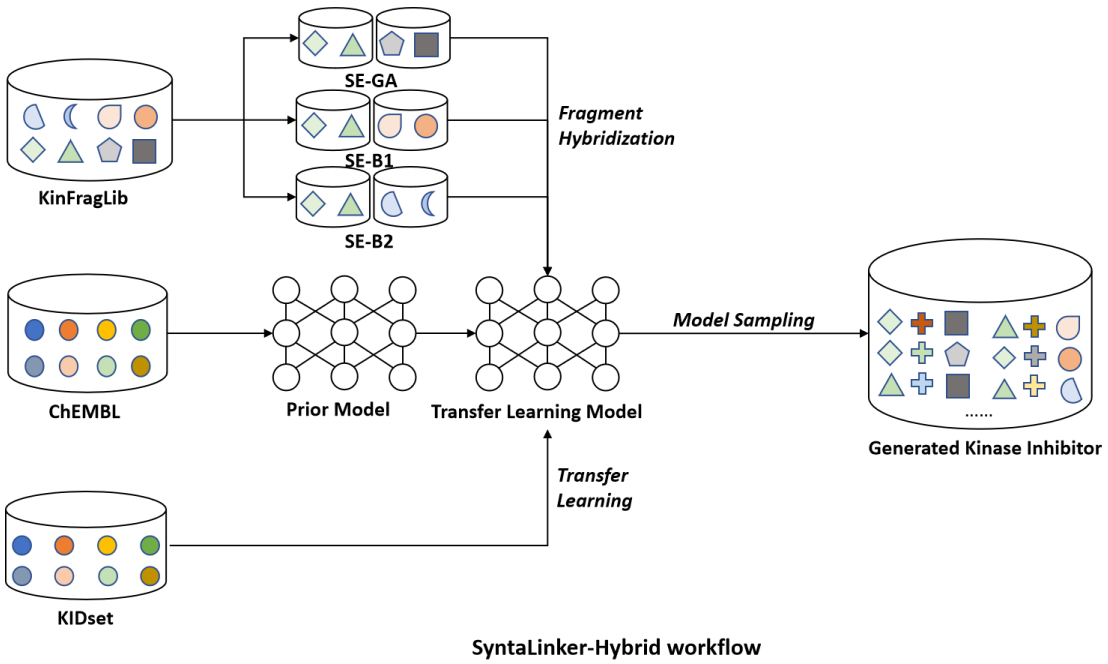

Figure 3. SyntaLinker-Hybrid workflow for generation of kinase inhibitors.

As shown in **Figure 3**, SyntaLinker-Hybrid is a workflow based on SyntaLinker method and its main task is to make automatic fragment hybridization to form multiple terminal fragment pairs and sample SyntaLinker model with a pool of terminal fragment pairs to generate diverse solutions. Comparing with normal SyntaLinker method, this workflow embedded with a fragment hybridization process generating a lot more input fragment pairs, thus provide more diverse solutions. In current study, fragment hybridization was carried out on a set of the kinase fragments binding at non-AP sub-pockets to create a pool of terminal fragment pairs. Based on the positions of the sub-pockets in ATP-binding pocket of kinase (as shown in **Figure 1**), three sub-pocket combinations were considered for linker generation: The first one was the combination of the SE and GA sub-pockets, these two

sub-pockets were direct neighbours of the AP sub-pocket and their distance constraint was set to 4~5 bond distance (**Figure S1A**); The other two alternatives referred to the combinations of the SE/B1 sub-pockets and the SE/B2 sub-pockets. They were utilized to explore the possibility of generating longer linkers that can cover the AP and GA sub-pockets. The distance constraints were set to 8~10 bond distance (**Figure S1B**), which is roughly equal to the maximal bond length of the fragments spanning both AP and GA sub-pockets. Exhaustive enumeration was employed for fragment hybridization, which meant all possible combination between fragments of the SE sub-pocket and the fragments of GA, B1 and B2 sub-pockets were considered. After fragment hybridization was finished, a pool of terminal fragment pairs can be generated.

The same ChEMBL fragment set and training protocol which were employed in our previous study[33] were again used to train a prior SyntaLinker model. Based on the prior model, the KID fragment set was then used to make a transfer learning to ensure the final model learning how to generate kinase specific linkers. The KID set was randomly split into training, validation and test sets with a ratio of 8:1:1. The first two sets were used for training a SyntaLinker model, while the last one was used as an internal test for model evaluation. During the transfer learning, the same type of neural network was created by initiating its parameters with those of the prior ChEMBL model. A small learning rate of 0.0001 was used for transfer learning. Meanwhile, the training step was empirically set to 50000 steps to avoid over-fitting. The rest settings were the same as described in our

previous study. The model checkpoints were saved every 100 steps, and the checkpoint at the final step was used as the model for subsequent sampling.

For model sampling, terminal fragment pairs were used as the sampling input together with bond distance constraints, and 10 molecules were generated for each fragment pair. Invalid SMILES were always removed from the final list. In addition, the molecules whose terminal fragments cannot exactly match the input terminal fragments were also discarded. The structures generated from SyntaLinker-Hybrid were named as the S-MOL set.

Model evaluation

The SyntaLinker model was evaluated in terms of molecular validity, uniqueness, novelty and recovery based on the terminal fragments extracted from the compounds in the test set[37, 38]. Their calculation was illustrated as formula 1-4. Specifically, validity refers to the percentage of chemically valid molecules among generated molecules. Uniqueness refers to the percentage of the unique molecules among valid structures. Novelty, as the name implies, refers to the percentage of unique novel molecules in the generated set, i.e. the structures which doesn't exist in the test set. Recovery refers to the percentage of generated molecules that can be found in the test set. To be noted, only 2D structure was used for above comparison.

$$\text{Validity} = \frac{\#\ \text{of chemically valid SMILES with fragments}}{\#\ \text{of generated SMILES}} \quad (1)$$

$$\text{Uniqueness} = \frac{\#\ \text{of non-duplicate,valid structures}}{\#\ \text{of valid structures}} \quad (2)$$

$$\text{Novelty} = \frac{\#\ \text{of novel structures not in test set}}{\#\ \text{of unique valid structures}} \tag{3}$$

$$\text{Recovery} = \frac{\#\ \text{of valid strectures in test set}}{\#\ \text{of test set}} \tag{4}$$

Coverage of chemical space

Analysis on chemical space of S-MOL set and K-MOL set was carried out. Here, the chemical space was characterized either by structural fingerprints or physico-chemical properties. In current study, the MACCS fingerprint[39] was used as structural descriptors and it is a 166-dimensional binary vector, where each dimension corresponds to a predefined seed structure. Principle component analysis (PCA) was done to map a high dimension space to a low dimension space constructed by a few orthogonal principle components that represent the implicit chemical feature. Besides the structural fingerprints, a set of physico-chemical descriptors was also used for the PCA analysis, including molecular weight (MW), number of rotatable bond (RTB), number of hydrogen bond acceptor (HBA), number of hydrogen bond donor (HBD) and lipid-water distribution coefficient (LogP). These descriptors were calculated with the RDKit package[40].

The oetoolkit[41] based program Flush[41] was used for clustering, it was developed based on Taylor clustering algorithm[42]. The Tanimoto similarity was calculated based on Foyfi fingerprints[41] and two similarity thresholds (0.5 and 0.7) were used for clustering.

Molecular Docking

Molecular docking was employed to evaluate the quality of molecules generated by SyntaLinker in terms of their binding affinity to a specific target. BRAF kinase was a well-known anti-cancer drug target. We docked the generated molecules to the BRAF crystal

structure (PDB ID: **6U2V**), which adopted the DFG-in conformation. CDK9 kinase was chosen as the other example and the crystal structure **3BLR** (PDB ID) was selected as the protein model for docking. The Glide docking module in the Schrodinger software (version 2020)[43] was employed for docking study. All protein structures were prepared with the Protein Preparation Wizard in maestro module and ligands were prepared with the LigPrep module, in which all tautomers and isomers were enumerated. Their ionized states at the target pH 7.0 were determined using the Epik algorithm. Flexible docking was carried out for each compound and the docking precision was set to Glide-SP. For each compound, 10 docking poses were generated and the best scored pose (based on Glidescore) for each ligand was saved for further analysis. Kernel density estimation (KDE)[44] module implemented in python was used to calculate statistical probability distribution of the docking score.

Results and Discussion

Kinase model evaluation

A transfer learning model for kinase inhibitors was constructed by using the SyntaLinker algorithm. It was evaluated in terms of validity, uniqueness, novelty and recovery based on the KID test set. There are 13,964 compounds in the test set and after fragmentation, in total 19,364 terminal fragment pairs were obtained for sampling the model. In the end 193,640 molecules were generated, and 93.1% molecules were chemically valid. This percentage was larger than that of the prior ChEMBL model (**Table 1**). Moreover, 98.3% of valid molecules were unique. As for novelty, 92.4% of unique compounds cannot be

found in the test set. On the other hand, it was found that 26,673 linkers of 694,584 unique compounds were novel to the test set, indicating decent structural novelty for generated linkers. There were 4221 linkers of 71756 unique compounds appeared in the test set, which means 70% of the linkers in the test set were reproduced. In other words, many linkers in the test set can be reproduced by the model, if not considering the consistence of terminal fragment pairs. In comparison, only 1272 linkers were reproduced by sampling the ChEMBL model with the same set of terminal fragments, suggesting that the transfer learning kinase model was quite effective on generating kinase specific linkers. This was also supported by the fact that around 76.8% of test set molecules were reproduced by the transfer learning model, while the percentage was only 27.3% for the ChEMBL model.

Table 1. Statistics for model evaluation based on test set

Metrics	TL[a]	PR[b]
# of compounds in test set	13964	13964
# of fragment pairs in test set	19364	19364
# of linkers in test set	6045	6045
# of sampled molecules	193640	193640
# of valid molecules	180323	168709
# of unique molecules	177257	165449
# of novel molecules	166527	162028
# of recovered molecules	10730	3821

Notes:

a) the transfer learning model based on KID data set

b) prior model based on ChEMBL database

Molecular Generation Based on the SyntaLinker-Hybrid Workflow

Sydow et al constructed a kinase virtual library by making exhaustive enumeration on sub-pocket specific fragments extracted from KLIFS database. In their work, a set of fixed

molecule enumeration rules needed to be defined and the hinge binding motifs are strictly limited to the ones defined in the KLIFS. However, one attractive feature for deep generative model is that the fixed enumeration rules are no longer needed for structure generation and model can learn assembling rules implicitly. In current study, we proposed the SyntaLinker-Hybrid workflow to automatically make terminal fragment pair hybridization to sample the kinase focused generative model for structure generation. In contrast to Sydow's work, our methodology had two advantages: Firstly, we don't need the predefined reaction templates for combining fragments into full structure, and it could be imagined that some constructed compounds could be unfeasible to synthesis due to the rigid rules; Secondly, in our method, the hinge motifs were generated via a kinase focused generative model constructed on a much larger kinase compound set, while the hinge binding fragments in Sydow's method can only use a limited fragment set from KLIFS.

In SyntaLinker-Hybrid workflow, the terminal fragment hybridization was carried out using the fragments extracted from KinFragLib fragment library. In total, 2,392,000, 193,200 and 191,360 fragment pairs were generated for SE/GA, SE/B1 and SE/B2 sub-pocket combinations, respectively. Sampling the kinase model by these fragment pairs generated 7,387,968, 1,114,506, 1,094,998 molecules respectively. After validity check, 5,077,578, 770,123, 790,762 molecules were remained for these three sub-pocket combinations. The percentage of valid molecules was obviously smaller than that obtained in model evaluation. This difference was probably attributed to the fact that during fragment hybridization, a lot of terminal fragment pairs unseen to the known kinase

inhibitors were generated, while in kinase transfer learning stage a lot of terminal fragment pairs actually came from the same molecule and part of their structural information had already been included in the training set. After removing duplicates and the molecules whose terminal fragments didn't match to the ones of KinFragLib fragment library, eventually, 232,705, 42,117, 37,246 molecules were collected respectively for above mentioned three sub-pocket combinations. The hybridization scripts of Sydow's work was used to make library enumeration and 102,733, 255,096, 359,554 molecules were generated for above sub-pocket combinations respectively. It can be seen that the transfer learning model generated more molecules for the SE/GA sub-pocket combination than the KinFragLib method, but much less molecules for the other two combinations. This indicated that it was still a challenge for the deep learning model to generate large number of molecules with long linkers. This's probably because the linking rules for those long linkers was not learnt by the transfer learning model properly. For simplicity, only the molecules corresponding to the SE/GA combination were used for subsequent analysis.

Structural Analysis for Generated Molecules

The structural fingerprint based PCA analysis on generated sets of kinase model and ChEMBL prior model were carried out and their chemical spaces were compared. As shown in **Figure 4A**, the KID set covered only part of the ChEMBL space. This was understandable as the KID set was extracted from the ChEMBL database. The chemical space of generated set of kinase transfer learning model pretty much covered most of the

chemical space of the whole KID set (**Figure 4C**), the same trend was observed for that of the ChEMBL model (**Figure 4D**). This means that the chemical space of sampling sets from both generative models can more of less mimic that of the whole training sets. This demonstrated that the kinase transfer learning model did learn the chemical features of general kinase inhibitors.

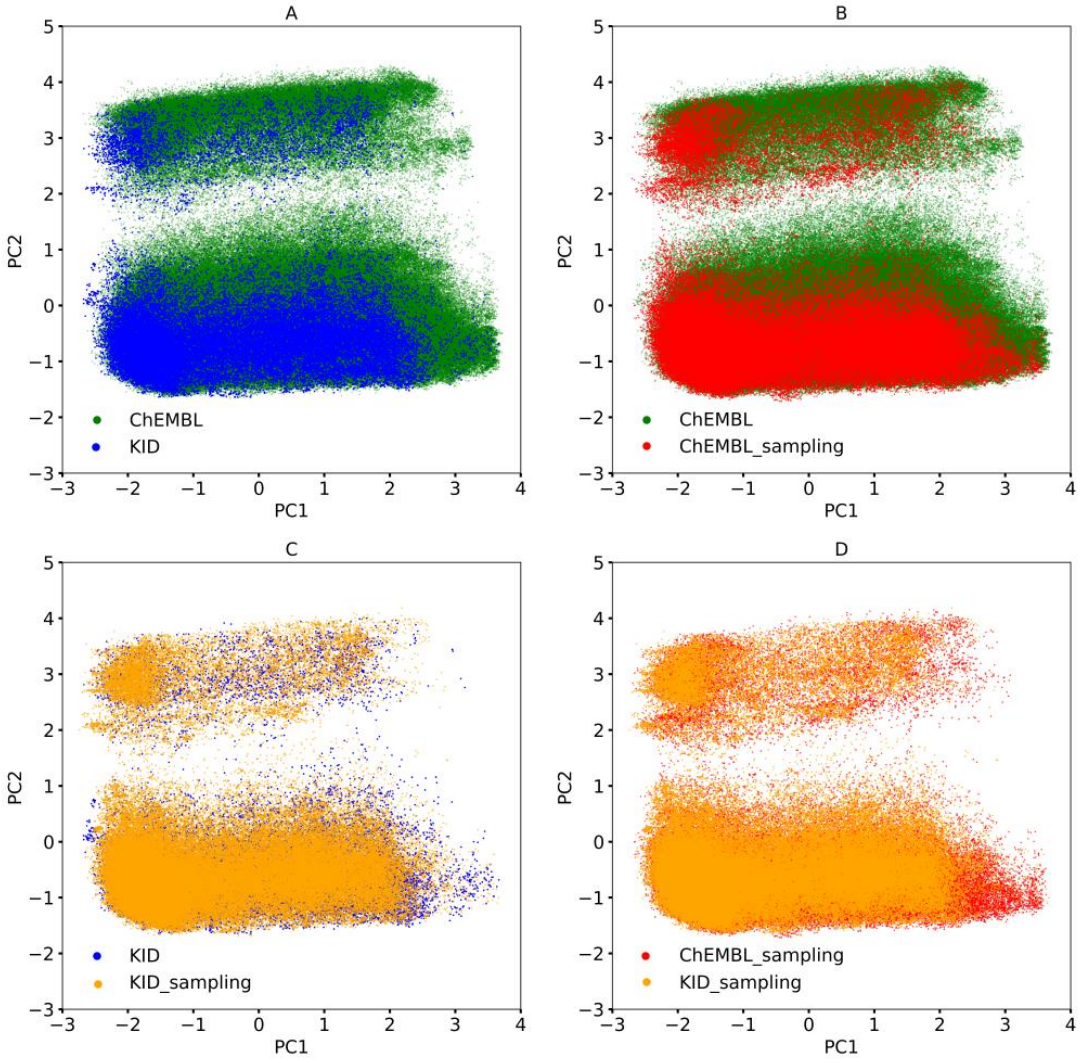

Figure 4. Chemical space comparison (**A**) between the ChEMBL and KID data sets, (**B**) between ChEMBL and the generated set by sampling the ChEMBL prior model, (**C**) between KID and the generated set by sampling the kinase transfer learning model, (**D**) between the two sampled molecule sets.

The fingerprint based PCA analysis were also done on the S-MOL and K-MOL sets to investigate the corresponding chemical spaces of these two sets. As it was described before, the S-MOL set was generated through the kinase generative model in SyntaLinker-Hybrid workflow and the K-MOL set was generated via fragment hybridization, and both set used the same set of non-AP fragments in KLIFS database. First of all, the chemical space of the S-MOL and KID sets were shown in **Figure 5A**, not surprisingly, the generated set S-MOL pretty much fully covered the KID set, which means the transfer learning kinase model did a good job in learning the chemical features in KID set and the generated compounds largely overlapped with that of the known kinase compounds. Regarding to the comparison between the K-MOL and KLIFS sets (as shown in **Figure 5B**), it seems that their spatial distribution was somewhat different. The KLIFS set was more or less evenly distributed, while the K-MOL set was more enriched on the two side regions and the density in the middle region was lower. **Figure 5C** demonstrated the comparison between the K-MOL and S-MOL sets. It can be seen that the S-MOL compounds distributed evenly in the space just like the KLIFS set, while the K-MOL set was more enriched in the side regions as shown in **Figure 5B**. It seems that the deep learning method was more effective on exploring the whole chemical space of kinase inhibitors than the direct enumeration

method. S-MOL was focused on the linkers with the binding features at the SE, AP and GA sub-pockets. Its relatively balanced distribution in the chemical space indicated that a majority of kinase inhibitors in the KID set and KLIFS datasets were bound at these three sub-pockets. In fact, there were really very few fragments bound at the back pockets for kinase inhibitors in the KLIFS data set. Interestingly, K-MOL, which focused on the same sub-pocket combination, exhibited significantly biased distribution. The possible reason is that the deep learning model was able to bring more chemical diversity than simply fragment linking based on available kinase inhibitors.

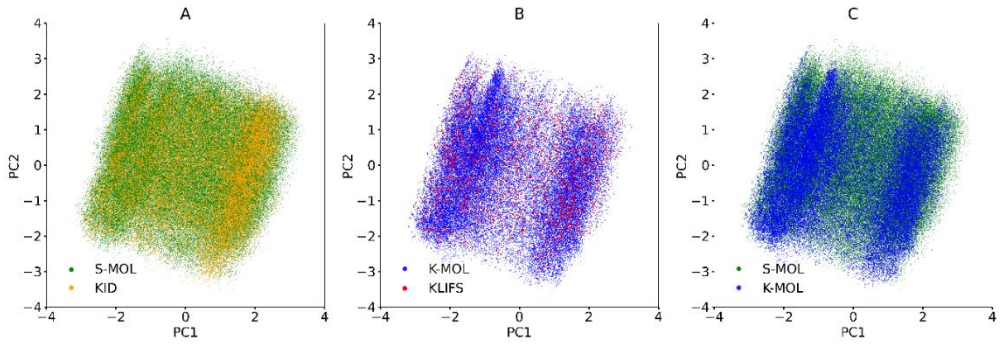

Figure 5. Chemical space comparison between the data sets of S-MOL (green), K-MOL (blue), KID (yellow), KLIFS (read).

Six physico-chemical features of molecules of S-MOL, K-MOL, KLIFS and KID sets were calculated (**Figure 6**). Overall, the distribution of the six features for S-MOL was similar to that for KLIFS and KID sets. A majority of molecules in S-MOL were concentrated in the MW range from 300 to 600 that is usually corresponding to drug-like molecules. But, the MW distribution of K-MOL, KID and KLIFS was wider than S-MOL.

The S-MOL doesn't have extremely low or high MW as in K-MOL, KLIFS and KID set. As for LogP, that of S-MOL was on average higher than K-MOL, but its mean LogP values was still in the range of drug-like compound. Similar results were observed for nTotB, TPSA， HBD and HBA descriptors. PCA analysis based on the six physico-chemical features showed that the physico-chemical space of KID set was well aligned with that of KLIFS (**Figure S2)**. most region of the S-MOL set were aligned well with that of other sets, while it still had some compounds stretching out to some unoccupied region.

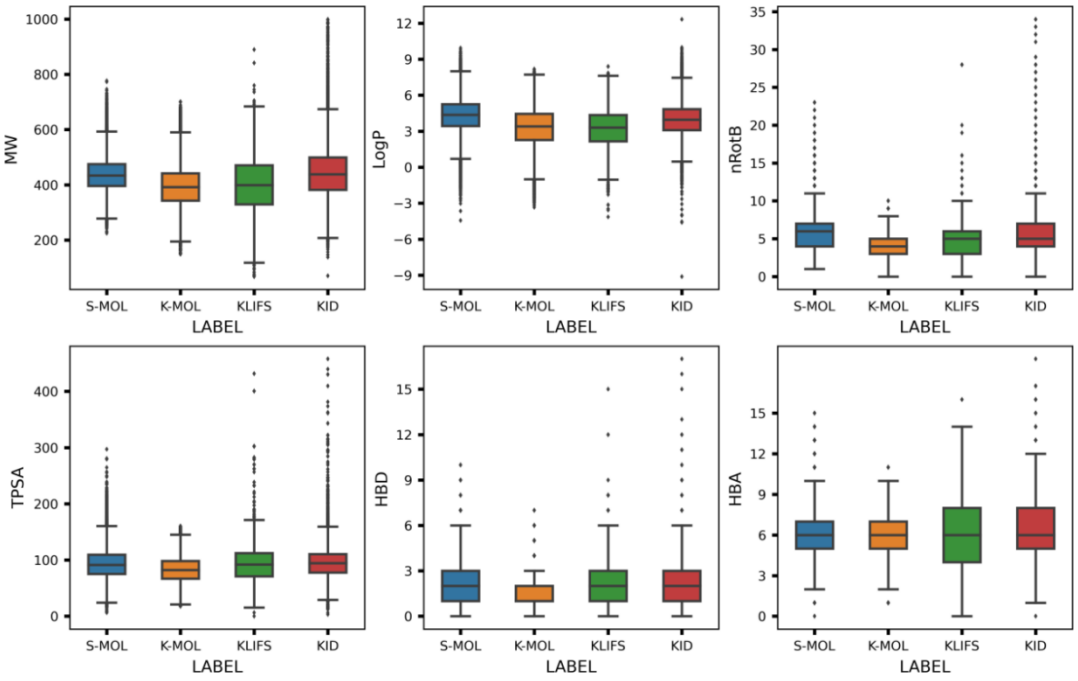

Figure 6. Boxplot of the features such as molecular weight (MW), LogP, number of rotatable bonds (nTotB), TPSA, number of hydrogen bond donors (HBD), and number of hydrogen bond acceptors (HBA).

We further analyzed the structural diversity of generated inhibitors based on clustering results. The molecules in S-MOL and K-MOL were clustered at full compound level as well as molecular scaffold level and their clustering results are shown in **Figure 7**. The S-MOL set had much more clusters than K-MOL at two different similarity cut-off levels. To verify this, we carried out molecular skeleton analysis using Murcko scaffold, which produced 137057 and 44092 unique skeletons for S-MOL and K-MOL respectively, the clustering analysis results on scaffold were shown in **Figure 7**.

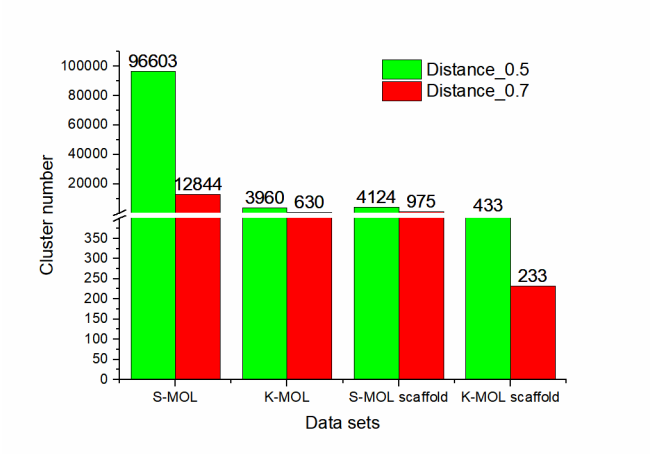

Figure 7. Histogram for the cluster numbers of S-MOL, K-MOL and their Murcko scaffold at two similarity cut-offs (0.5 and 0.7). The cluster numbers were labeled on the bars.

Structural analysis for linker fragments

The core idea of the SyntaLinker algorithm was to generate linker fragments by sampling a deep learning model with given terminal fragment pairs, while the enumeration method is to hybridize the linker fragments of AP sub-pocket with other components to form a

virtual library. It would be interesting to make a structural comparison on the linker fragments of the AP sub-pocket between the S-MOL and K-MOL sets.

In total 43,784 linkers were generated by SyntaLinker-Hybrid workflow, which was roughly 7 times more than the linkers of compounds in K-MOL extracted from the AP fragments in the KinFragLib library, and there are 342 K-MOL linkers were reproduced (**Figure 8A**). Compared to K-MOL linkers, structural classification of linkers in S-MOL resulted to more clusters (**Figure 8B**), suggesting the higher structural diversity of the molecules in S-MOL. The PCA analysis based on fingerprint and physico-chemical features were shown in **Figure 8C/D**. **Figure 8C** shows that the linkers generated by SyntaLinker model not only covered the entire space of the K-MOL linkers but also had a large extra space exploited, which corresponds to novel linker structures. The overall physico-chemical properties were similar to that of the K-MOL linkers (as shown in **Figure 8D**). The detailed analysis among the physico-chemical properties are shown in **Figure S3**. It seems that the MW of SyntaLinker generated linkers were distributed in a larger range comparing to the K-MOL linkers. But their LogP and TPSA distribution are rather similar. Moreover, over 90% SyntaLinker generated fragments had less than 4 rotatable bonds, which was similar to the K-MOL linkers. The number of nitrogen atoms contained in linkers were also compared, as this element was often employed as hydrogen bond related interaction sites when interacting with the key residues at the hinge region. As a result, the distributions of nitrogen atom contained in the two sets were similar, suggesting the potential of the SyntaLinker generated linkers mimicking the hinge hydrogen bonds formed

in the available kinase inhibitors. This was also confirmed by the analysis on distribution of hydrogen bonds (**Figure S4**), which demonstrated that the SyntaLinker generated linkers had similar distribution to the K-MOL linkers regarding to the number of hydrogen bonds.

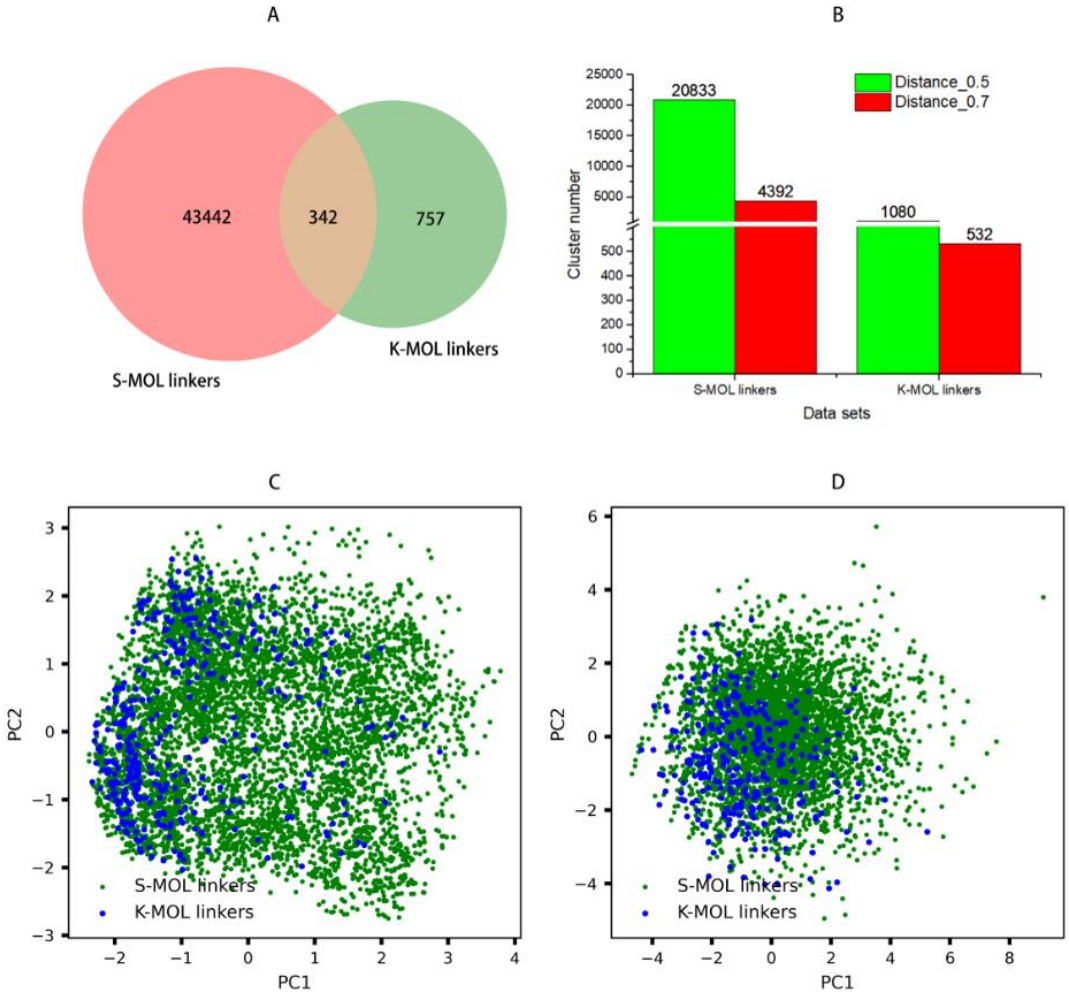

Figure 8. (**A**) Venn diagram for overlapping between the S-MOL and K-MOL linker sets. The linkers were anchored at the AP sub-pocket of the hinge region. (**B**) Histogram for the cluster numbers of the S-MOL and K-MOL linker sets at two similarity thresholds (0.5 and 0.7). The cluster numbers were labeled on the bars. (**C**) Chemical and (**D**) Physico-chemical space comparison between the S-MOL and K-MOL linker sets.

Structural similarity was computed between the generated linkers and the K-MOL linker fragments. According to the heatmap of similarity, we could see that there was always a portion of generated linkers having relatively high similarity to some K-MOL linkers (**Figure 9A**). This indicated that it was possible to find some SyntaLinker generated linkers capable of recovering the interactions of kinase inhibitors at the hinge region. In other word, scaffold hopping at the hinge region was feasible based on the deep learning model built in this study. In the meanwhile, the similarity was centered on the value of 0.4 while varying in a range from 0.2 to 0.6 in most cases (**Figure 9B**), which implied that most generated linkers had relatively high structural novelty. Examples of the linkers with high and low similarity were shown in **Figure 9C**.

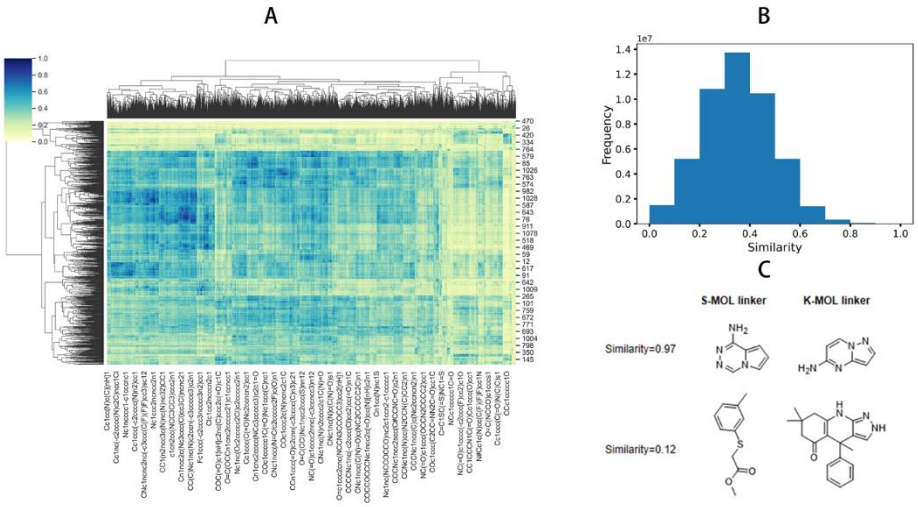

Figure 9. (**A**) Heatmap of similarity between the S-MOL and K-MOL linkers. (**B**) Frequency distribution of the similarity in the range from 0 to 1 using 10 equally spaced bins. (**C**) Examples for the fragment pairs with high and low similarity.

Case Study on Scaffold Hopping

In order to validate the effectiveness of the SyntaLinker-Hybrid workflow on scaffold hopping, BRAF kinase was selected as the target for the case study and SyntaLinker generated molecules in S-MOL were docked into the BRAF ATP-binding pocket. The crystal structure **6U2V** was used for docking and the best docking pose (the pose selection was described in the method section) was reserved for each molecule. **Figure 10A** shows that the distribution of docking score of S-MOL set was left shifted comparing to the K-MOL and KLIFS sets, which indicated that the S-MOL molecules in general had better docking score than the two other sets. This was further evidenced by the distribution of docking scores of the top 100 molecule set and top 1000 molecule set in each data set (**Figure S5A/B**). Moreover, randomly selected 2000 compounds from S-MOL also had better scores than the non-kinase bioactive compounds randomly selected from the ChEMBL database (**Figure 10B**). Since usually docking score is not necessarily correlated with binding potency, we examined the docking poses of S-MOL molecules to check if they can interact with the backbone atoms in Cys532 and Gln530 residues at the hinge region of BRAF kinase. As a result, 1,386 molecules of the 2,000 selected compounds can form at least one interaction with the hinge residues. Of the top 10 best scored molecules, six molecules can form two hydrogen bonds, while the other four molecules formed only one hydrogen bond. Their molecular structures and docking poses were shown in **Figure S6**.

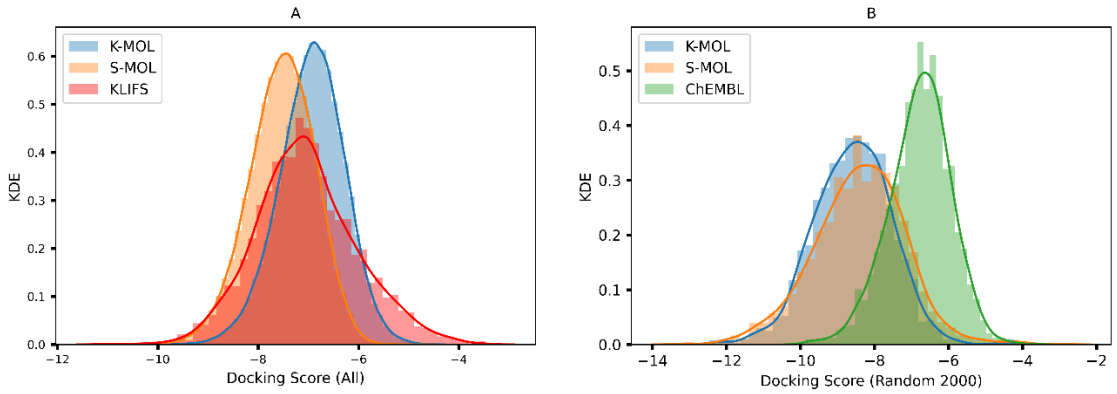

Figure 10. Statistical probability distribution of the docking scores for (**A**) all molecules in K-MOL, S-MOL, KLIFS data sets, and for (**B**) 2000 randomly selected molecules in K-MOL, S-MOL and ChEMBL data sets.

De novo design of CDK9 inhibitors

In order to further evaluate the performance of the trained SyntaLinker kinase model, CDK9 was also selected for doing scaffold hopping with the method. 11 known CDK9 inhibitor structures were selected from ChEMBL and 11 terminal fragment pairs (shown in **Table S1**) were extracted from these CDK9 inhibitors for doing scaffold hopping at the hinge motif. The SMILES strings of the terminal fragments were randomized to form 63,273 pairs to augment the input query for sampling. As a result, in total 5,146 new molecules containing the input fragments were generated from the model. Among these molecules, 1934 unique linkers were generated and they were all novel linkers to the KID set. The linkers were clustered to into 29 clusters according to the scaffolds of the generated

linkers. The results showed that N -heterocycles was most frequent functional group in the generated linkers (**Table S2**).

Subsequently, the generated molecules were docked to the ATP-binding pocket of CDK9. Eventually, 2,457 molecules were able to make hydrogen bond interactions with the residue Cys106 that is the key anchor site at the hinge region of CDK9. The docking poses of three representative molecules demonstrated that not only the hydrogen bond interactions at the hinge region were preserved by the new linkers (**Compound 1**, **Compound 2**, **Compound 3** in **Figure 11A/B/C**), but also the binding mode of non-AP sub-pocket fragments was the same as that of known CDK9 inhibitors. It was interesting to see that some known hinge fragments were partially reproduced by the deep learning model, although the corresponding terminal fragments was not the same to the original inhibitor. One example was **SNS-032** (**Table S1**), whose linker fragment was quite similar to the linker of generated **Compound 4**. In the meanwhile, for **Compound 4**, the hydrogen bond of its hinge motif and the positions of terminal fragments were very similar to the known CDK9 inhibitor **BAY-114357** (**Figure 11D**). This examples highlighted that the generated molecules from SyntaLinker model are structurally novel and potentially interesting compounds for synthesis. Accordingly, the SyntaLinker-Hybrid workflow can be a useful tool for scaffold hopping on kinase inhibitors by generating structures with considerable structural novelty.

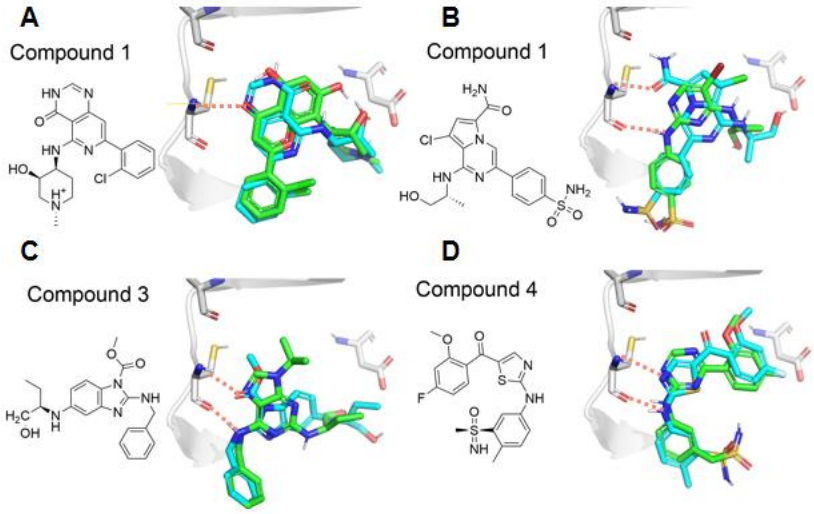

Figure 11. Docking poses of four representative compounds. The docking poses of **Compound 1** (**A**), **Compound 2** (**B**), **Compound 3** (**C**) and **Compound 4** (**D**) were shown as sky-blue sticks, and, correspondingly, the binding poses of CDK9 inhibitor **Flavopiridol** (**A**), **ZK-304709** (**B**), **(R)-Roscovitine** (**C**) and **SNS-032** (**D**) were shown in green sticks. The hydrogen bonds at the hinge region were shown as orange dashes.

Conclusions

In current study, the SyntaLinker-Hybrid workflow was utilized to carry out scaffold hopping at the hinge region for kinase inhibitors. A kinase-focused transfer learning model was first built using a diverse kinase inhibitor library. Then, large number of terminal fragments, extracted from the KinFragLib database, were hybridized to form terminal fragment pairs for sampling the kinase transfer learning model. More specifically, two fragments were hybridized into a fragment pair only if they were respectively bound at two different sub-pockets neighboring the hinge region. Through this workflow, a large number

of kinase specific novel structures can be generated. Among the three selected sub-pocket combinations for hybridization, the SE/GA combination resulted to the most number of molecules, in which over 80% molecules were chemically valid. Further analysis showed that the generated molecules from SyntaLinker-Hybrid workflow occupied the same chemical space as known kinase inhibitors and achieved larger diversity comparing with the molecules generated from the library enumeration manner (i.e. by systematically combining the KinFragLib fragments). To illustrate the effectiveness of the workflow, SyntaLinker-Hybrid generated compounds and compounds generated by library enumeration were docked to a BRAF kinase structure and SyntaLinker generated compounds have better docking scores. A scaffold hopping exercise towards CDK9 inhibitors was also carried out by using terminal fragment pairs from known CDK9 inhibitors for sampling the kinase transfer learning model. Our results show that in the docking pose, about 70% of generated molecules can form at least one hydrogen bond with the key residues at the hinge region. At the same time, the binding conformations of some generated structures have high similarity to known CDK9 inhibitors while containing novel hinge binding motifs. We expect that this method could also be expanded to do scaffold hopping for other target families given enough structural information existed for transfer learning.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge via the Internet at http://pubs.acs.org. Simulation details including splitting of 11 CDK9 inhibitors in Table S1, clustering results of linker scaffolds extracted from the CDK9-targeted molecules in Table S2, length distribution of fragments in Figure S1, PCA analysis of physico-chemical space for molecules in Figure S2, distribution of physico-chemical properties for linker fragments in Figure S3, frequency of hydrogen bonds for linker fragments in Figure S4, Statistical probability distribution of the docking scores in Figure S5, and docking poses of the top best scored 10 molecules for BRAF in Figure S6.

AUTHOR INFORMATION

Corresponding Author

* E-mail: chen_hongming@grmh-gdl.cn; Tel: +86-20-62726115

* E-mail: ran_ting@grmh-gdl.cn; Tel: +86-20-62726115

* E-mail: junxu@biochemomes.com; Tel: +86-20-39943074

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

REFERENCES

1. Manning; G., The protein kinase complement of the human genome. *Science* **2002,** *298* (5600), 1912-1934.
2. Cohen, P.; Alessi, D. R., Kinase drug discovery - What's next in the field? *ACS Chem. Biol.* **2012,** *8* (1), 96-104.
3. Lahiry, P.; Torkamani, A.; Schork, N. J.; Hegele, R. A., Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nat. Rev. Genet.* **2010,** *11* (1), 60-74.
4. Wang, P. F.; Qiu, H. Y.; Zhu, H. L., A patent review of BRAF inhibitors: 2013-2018. *Expert Opinion on Therapeutic Patents* **2019,** *29* (8), 595-603.
5. Cohen, P., Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **2002,** *1* (4), 309-315.
6. Fabbro, D.; Cowan-Jacob, S. W.; Moebitz, H., Ten things you should know about protein kinases: IUPHAR Review 14. *Br. J. Pharmacol.* **2015,** *172* (11), 2675-2700.
7. Wu, P.; Nielsen, T. E.; Clausen, M. H., FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **2015,** *36* (7), 422-439.
8. Beilei, W.; Jiaxin, W.; Yun, W.; Cheng, C.; Fengming, Z.; Aoli, W.; Hong, W.; Zhenquan, H.; Zongru, J.; Qingwang, L., Discovery of 4-(((4-(5-chloro-2-(((1s,4s)-4-((2-methoxyethyl)amino)cyclohexyl)amino)pyridin-4-yl)thiazol-2-yl)amino)methyl)tetrahydro-2 H -pyran-4-carbonitrile (JSH-150) as a novel highly selective and potent CDK9 kinase inhibitor. *Eur. J. Med. Chem.* **2018,** *158*, 896-916.
9. Asghar, U.; Witkiewicz, A. K.; Turner, N. C.; Knudsen, E. S., The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nat. Rev. Drug Discov.* **2015,** *14* (2), 130-146.
10. Mortenson; Paul, N., Fragment-based approaches to the discovery of kinase inhibitors. *Methods Enzymol.* **2014,** *548*, 69-92.
11. Reck, M.; Horn, L.; Novello, S.; Barlesi, F.; Albert, I.; Juhász, E.; Kowalski, D.; Robinet, G.; Cadranel, J.; Bidoli, P.; Chung, J.; Fritsch, A.; Drews, U.; Wagner, A.; Govindan, R., Phase II study of roniciclib in combination with cisplatin/etoposide or carboplatin/etoposide as first-line therapy in subjects with extensive-disease small cell lung cancer. *J. Thorac. Oncol.* **2019,** *14* (4), 701-711.
12. Whittaker, S. R.; Mallinger, A.; Workman, P.; Clarke, P. A., Inhibitors of cyclin-dependent kinases as cancer therapeutics. *Pharmacol. Ther.* **2017,** *173*, 83-105.

13. Erickson, J. A., Fragment-based design of kinase inhibitors: A practical guide. *Methods Mol. Biol.* **2015,** *1289*, 157-183.
14. Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisà, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J., Structural basis of inhibitor selectivity in MAP kinases. *Structure* **1998,** *6* (9), 1117-1128.
15. Paul; Czodrowski; Günter; Hlzemann; Gerhard; Barnickel; Hartmut; Greiner; Djordje; Musil, Selection of fragments for kinase inhibitor design: Decoration is key. *J. of Med. Chem.* **2014,** *58* (1), 457-465.
16. Kumar, A.; Voet, A.; Zhang, K. Y. J., Fragment based drug design: From experimental to computational approaches. *Curr. Med. Chem.* **2012,** *19* (30), 5128-5147.
17. Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P., Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **2011,** *3* (4), 405-414.
18. Renner, S.; Schneider, G., Scaffold - hopping potential of ligand - based similarity concepts. *ChemMedChem* **2010,** *1* (2), 181-185.
19. Dey, F.; Caflisch, A., Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **2008,** *48* (3), 679-690.
20. Geppert, H.; Vogt, M.; Bajorath, J. R., Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010,** *50* (2), 205-216.
21. Hu, Y.; Stumpfe, D.; Bajorath, J., Recent advances in scaffold hopping. *J. Med. Chem.* **2016,** *60* (4), 1238-1246.
22. P. Schneider, Y. T., G. Schneider, Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. *Curr. Med. Chem.* **2009,** *16* (3), 258-266.
23. Mukherjee, P.; Bentzien, J.; Bosanac, T.; Mao, W.; Burke, M.; Muegge, I., Kinase crystal miner: A powerful approach to repurposing 3D hinge binding fragments and its application to finding novel Bruton tyrosine kinase inhibitors. *J. Chem. Inf. Model.* **2017,** *59* (9), 2152-2160.
24. Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A., KinFragLib: Exploring the kinase inhibitor space using subpocket-focused fragmentation and recombination. *J. Chem. Inf. Model.* **2020,** *60* (12), 6081-6094.
25. Elton, D. C.; Boukouvalas, Z.; Fuge, M.; Chung, P. W., Deep learning for molecular generation and optimization - a review of the state of the art. *Mol. Syst. Des. Eng.* **2019,** *4* (4), 828-849.
26. Aspuru-Guzik; Alan; Markopoulos; Georgios; Chae; Sik, H.; Duvenaud; David; Maclaurin; Dougal, Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016,** *15* (10), 1120-1127.
27. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017,** *9* (1), 48-61.
28. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018** *4*(1), 120-131.
29. Fu, T.; Xiao, C.; Sun, J., CORE: Automatic molecule optimization using copy & refine strategy. *Analyst* **2019,** *144* (16), 4757-4771.
30. Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P., Optimization of molecules via deep

reinforcement learning. *Sci. Rep.* **2018,** *9* (1), 10752-10761.
31. Jin, W.; Barzilay, R.; Jaakkola, T., Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer, D.; Andreas, K., Eds. PMLR: Proceedings of Machine Learning Research, 2018; Vol. 80, pp 2323--2332.
32. Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J., Deep learning for molecular generation. *Future Med. Chem.* **2019,** *11* (6), 567-597.
33. Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H., SyntaLinker: Automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.* **2020,** *11* (31), 8312-8322.
34. Anna, G.; Bellis, L. J.; Patricia, B. A.; Jon, C.; Mark, D.; Anne, H.; Yvonne, L.; Shaun, M. G.; David, M.; Bissan, A. L., ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.
35. Hussain; Rea, Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010,** *50* (3), 339-348.
36. Wu, T.; Qin, Z.; Tian, Y.; Wang, J.; Xu, C.; Li, Z.; Bian, J., Recent developments in the biology and medicinal chemistry of CDK9 inhibitors: An update. *J. Med. Chem.* **2020,** *63* (22), 13228-13257.
37. Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C., GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **2019,** *59* (3), 1096-1108.
38. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A., Molecular Sets (MOSES): A benchmarking platform for molecular generation models. *Front. Pharmacol.* **2020,** *11*, 1931-1940.
39. Joseph L Durant 1, B. A. L., Douglas R Henry, James G Nourse, Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (6), 1273-1280.
40. Landrum, G.; Kelley, B.; Tosco, P.; sriniker; gedeck; NadineSchneider; Vianello, R.; Dalke, A.; AlexanderSavelyev; Turk, S. *rdkit/rdkit: 2017_09_2 (Q3 2017) Release*, 2017.
41. Butina, D., Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Model.* **1999,** *39* (4), 747-750.
42. Huang, Z. H.; Xiang, Y.; Zhang, B.; Wang, D.; Liu, X. L., An efficient method for K-means clustering. *PRAI* **2010,** *23* (4), 516-521.
43. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004,** *47* (7), 1739-1749.
44. Jones, M. C.; Marron, J. S.; Sheather, S. J., A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* **1996,** *91* (433), 401–407.