Artificial Intelligence in the Life Sciences 2 (2022) 100035

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

SyntaLinker-Hybrid: A deep learning approach for target specific drug design

Yu Feng [a,b], Yuyao Yang [b], Wenbin Deng [a,*], Hongming Chen [b,*], Ting Ran [b,*]

[a] School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University, Guangzhou 510006, China
[b] Department of drug and vaccine research, Guangzhou Laboratory, Guangzhou 510530, China

ARTICLE INFO

Keywords:
Deep generative model
Transfer learning
Fragment-based drug design
Target specific drug design

ABSTRACT

Target specific drug design has attracted much attention in drug discovery. But, it is a great challenge to efficiently explore the target-focused chemical space. Fragment-based drug design (FBDD) has shown its potential to do this thing. In this study, we introduced a deep learning-based fragment linking method, namely SyntaLinker-Hybrid, for target specific molecular generation. By carrying out transfer learning and fragment hybridization, this method allows to generate a great number of linker fragments to assemble given terminal fragments into the molecules with target specificity. This work demonstrates that the method has the capacity to generate target specific structures for various targets. We believe that its application could be extended to a broader target scope.

Introduction

Drug discovery is a process to search for compounds which satisfy a plethora of criteria including desirable biological activities, optimal off-target selectivity and ADMET properties etc. among the gigantic 10^{60}~$100,^{100}$ theoretical chemical space [1]. The process is so lengthy and costly that it is often described as finding "a needle in the haystack" [2]. Thus, the way to efficiently explore the chemical space becomes critical. In a long history, the search space is focused on natural products and synthesized compounds on the shelves. In recent decades, combinatorial chemistry (CC), high-throughput screening (HTS) and DNA-encoded chemical library have significantly expanded the real chemical space [3–7], which is still much smaller than the theoretical space. Nowadays with the application of ML/AI technologies on virtual screening, medicinal chemists are increasingly interested in the chemical space of accessible virtual compound libraries [8] which is far more than the real chemical space [9,10]. Numerous computational approaches have been proposed to generate virtual compound libraries [9,11–13], which have been used to increase the likelihood of finding new chemical entities mainly through virtual screening [14,15]. One typical approach is to exploit fragment-based chemical space through fragment hybridization [16,17]. However, this approach often suffers from combinatorial explosion in the size of compound libraries [18], which is unfavorable for virtual screening due to unaffordable computation cost. To address this problem, some fragment-based methods have been made to directly establish target-focused virtual compound libraries. The representative algorithms include BREED [19], Flux [20,21], BROOD [22] and etc. There is no doubt that these approaches have been useful for mining pharmacologically relevant regions through analyzing the crucial and privileged substructures (known as fragments) from known bioactive compounds [23]. Nonetheless, the generated compounds are usually confined to a small chemical space by attaching various chemical moieties or side chains to a limited number of compound scaffolds. Beside, they largely rely on exhaustive search or stochastic search algorithms [11,24,25]. Thus, new fragment-based methods with the capacity to rapidly design large amount of compounds with high structural novelty and target specific characteristics are needed.

In recent years, many deep learning-based molecular generative models [26–31] have been proposed to generate drug-like molecules [32–34]. The deep generative models can learn directly from the structure data without the need to have an explicit enumeration rule to generate molecules, as the traditional computational methods do. Among these models, fragment-based generative models have been particularly developed. For example, Alessio Micheli et al. reported a deep learning model to generate molecules fragment by fragment [35], while a generative algorithm for fragment growing atom by atom was also unveiled recently [36]. Beside, Ola Engkvist et al. trained a deep generative model based on a dataset comprising a large number of molecular scaffolds with their respective decorations for molecular generation [37]. These methods are mostly adaptive to lead optimization. Thus, a few methods concerning fragment linking were proposed for *de novo* drug design. Charlotte M. Deane et al. developed a graph-based deep generative

* Corresponding authors.
E-mail addresses: dengwb5@mail.sysu.edu.cn (W. Deng), chen_hongming@gzlab.ac.cn (H. Chen), ran_ting@gzlab.ac.cn (T. Ran).

https://doi.org/10.1016/j.ailsci.2022.100035
Received 24 December 2021; Received in revised form 9 April 2022; Accepted 14 April 2022
Available online 25 April 2022

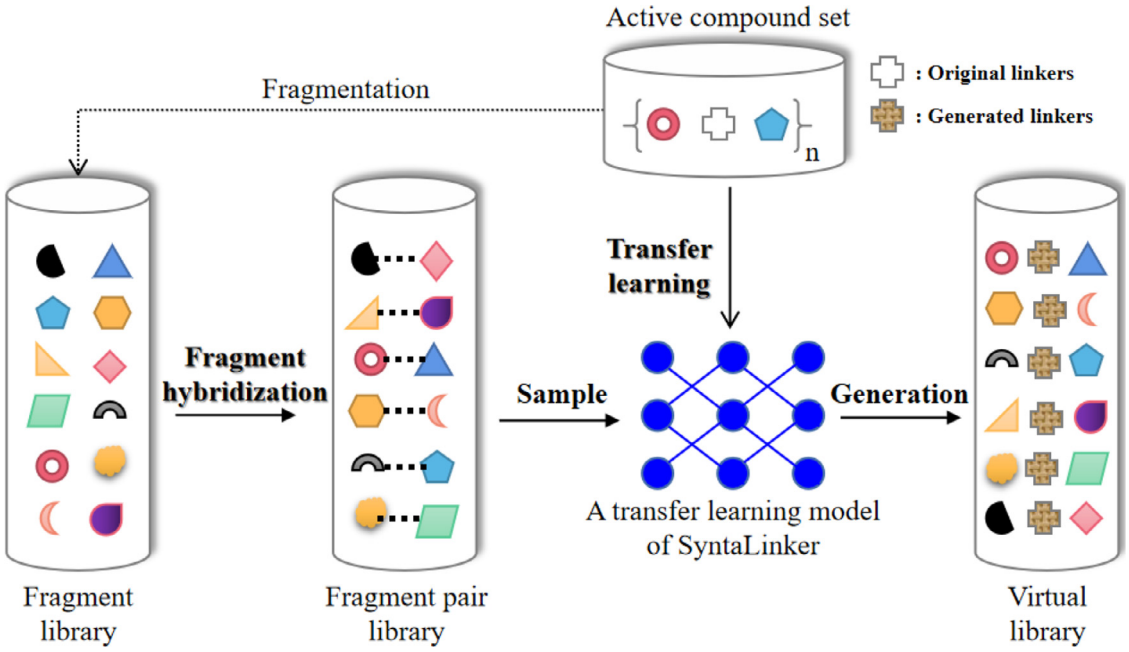

Fig. 1. SyntaLinker-Hybrid workflow for creating virtual compound library.

tool named DeLinker to connect two fragments in a three-dimensional manner [38]. Our group constructed a transformer model named SyntaLinker [39] to link terminal fragment pairs extracted from ChEMBL database [40]. Several recent reports showed that deep generative models can even been tailored to specific targets upon transfer learning [41,42] with focused subsets of bioactive compounds [28,43]. Following this idea, we introduced a SyntaLinker-based workflow, namely SyntaLinker-Hybrid, for target specific molecular generation by carrying out transfer learning and fragment hybridization.

The SyntaLinker-Hybrid workflow begins with learning the "implicit assembly rules" of molecules by training a prior SyntaLinker model using a large set of fragments pairs derived from the ChEMBL database. Then, the prior model is customized on a certain pharmacological target by performing transfer learning on the active compounds of the target to build target-focused SyntaLinker model. With the hope to extend the search space of the model, fragment hybridization is performed to combine the terminal fragments from different active compounds to hybrid terminal fragment pairs for model sampling, which may generate a number of linker fragments to assemble the given terminal fragments into the molecules with target specificity. In fact, this workflow has been successfully applied to scaffold hopping of kinase inhibitors by generating kinase-inhibitor-like structures [44]. In order to evaluate the workflow in a broader scope, we extended its application to other four drug targets that belong to diverse protein families. The capability of the workflow to generate target specific structures was validated by chemical space analysis, quantitative structure and activity relationship (QSAR) study and molecular docking.

Materials and methods

SyntaLinker-Hybrid workflow

SyntaLinker is a deep generative model that can generate a linker fragment to assemble two terminal fragments into a molecule. The model is built on a transformer architecture with multiple encoder-decoder stacks. Each stack is further composed by a multi-head self-attention sub-layer and a position-wise feedforward network (FFN) sub-layer. To prepare model training, each compound in the training set is split into two terminal fragments and a linker fragment. The SMILES pattern of a quadruple form such as "terminal fragment 1, linker fragment, terminal fragment 2, original compound" is then transformed to the input embedding by the one-hot encoding method. With the embedded representation, a SyntaLinker model can be built and sampled to output molecular structures using a terminal fragment pair as the input. More detailed information about SyntaLinker method can be reached at our previous work [39].

SyntaLinker-Hybrid extends above workflow by introducing a transfer learning process to model training and a module of fragment hybridization to model sampling (**Fig. 1**). Transfer learning starts from a prior model of SyntaLinker trained on a large compound set which is not specific for a target. By contrast, the dataset for transfer learning must come from a targeted compound set, and an especially small learning rate should be set (default is 0.0001). Meanwhile, the training step is empirically set to 50,000 steps to avoid overfitting. Model checkpoints are saved per 1000 steps, and the final one is used for model sampling. The rest of settings for SyntaLinker model construction such as batch size are kept as the same as those used in our previous study. As for fragment hybridization, the terminal fragments belonging to different compounds are randomly combined to form a number of hybrid fragment pairs. Then, the hybrid fragment pairs are assembled into molecules by sampling the transfer learning model. To ensure target specificity of generated molecules, the terminal fragments for hybridization are derived from the fragmentation of active compounds of the corresponding targets. Moreover, we used the bond length of the linker as constraint to sample structures. The range of bond length was decided based on the linker length in the original structures which the paired terminal fragments belongs to.

Data preparation

As a prior model must be built before transfer learning, the same ChEMBL fragment set employed in our previous study was used for the prior model training. These fragments were derived from the fragmentation of compounds in the large-scale ChEMBL database, which would ensure the prior model to sufficiently learn fragment assembly in known molecules. Then, we selected the bioactive compounds against four tar-

Table 1
Datasets for model construction and evaluation.

target name	# of actives	# of TFPs in training set	# of TFPs in validation set	# of TFPs in test set
BRD4	1324	2234	277	273
FXR	714	2334	291	285
HDAC1	3544	7302	912	902
DRD2	4311	16,168	2018	2007

TFPs: terminal fragment pairs.

gets in clinical trials, such as BRD4, HDAC1, FXR and DRD2, to build transfer learning models of SyntaLinker. Each target has reported thousands of active compounds, which seems a scale large enough for transfer learning. We retrieved the active compounds of BRD4, HDAC1 and FXR from the ChEMBL database, while the DRD2-targeted compounds were found from the EXCAPE-DB database [45]. The compounds with Ki/KD or IC_{50}/EC_{50} smaller than 1 μM were just reserved for model construction. The number of compounds for the targets were shown in Table 1. MMP cutting algorithm [46] was applied to decompose the compound structures as described above while the bond length of linker fragment was calculated. This cutting method may produce multiple cutting paths for a compound, which meant more than one terminal fragment pairs can be generated for each compound. Therefore, the size of terminal fragment pair sets generally increases several folds over the corresponding compound sets. The terminal fragment pair set of each target was divided into training, validation and test subsets in a ratio of 8:1:1 for model construction and evaluation (Table 1). It should be noted that all terminal fragments in the three subsets were used for fragment hybridization.

Model evaluation

Deep generative models were evaluated in terms of validity, uniqueness, novelty, and recovery of generated molecules based on the terminal fragment pairs in the test set. These metrics were calculated with Eqs. (1)–(4), respectively. Validity refers to the percentage of chemically valid structures among generated molecules, which were checked with the RDKit package [47]. Uniqueness refers to the percentage of non-duplicate molecules among valid structures. Novelty refers to the percentage of unique valid molecular structures that could not be found in test set. Recovery refers to the percentage of test set structures that were recovered in the generated set.

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

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

$$\text{Novelty} = \frac{\#\text{ of unique valid structures in generated set that not in test set}}{\#\text{ of unique valid structures in generated set}} \quad (3)$$

$$\text{Recovery} = \frac{\#\text{ of recovered test set structures in generated set}}{\#\text{ of test set structures}} \quad (4)$$

Chemical space analysis

Chemical space can be characterized by a variety of descriptors, such as structural and physicochemical properties. In this study, the MACCS fingerprint [48] was used as the structural descriptor to represent chemical space. It is formed by a 166-dimensional binary vector, and each dimension corresponds to a predefined seed structure. Then, principle component analysis (PCA) was performed on the molecular fingerprint to analyze chemical space coverage. Structural diversity was also analyzed based on Taylor clustering algorithm [49] using the oetoolkit based program Flush. The clustering was derived from the Tanimoto similarity [50] calculated based on Foyfi fingerprints [51] and the similarity threshold 0.5 were used for clustering. In addition, the physicochemical properties, such as molecular weight (MW), lipid-water distribution coefficient (LogP(o/w)), solubility to water (logS) and topological polar surface area (TPSA) were calculated using the MOE software (version 2020).

QSAR study

QSAR study was carried out for each target based on non-linear SVM classification model [52]. The model was constructed using the scikit-learn module with the Python programming environment [53]. The active compound set was divided into training and test sets in a ratio of 3:1. Beside, 33,500 decoy molecules for each target were generated using the methodology implemented in the Database of Useful Decoys (DUD) [54] based on randomly selected 600 active compounds. The decoy compound set was also divided into training and test sets using the same ratio (**Table S1**). The training sets of active compounds and decoy molecules were combined to train the model, and the combination of test sets was used for model evaluation. All compounds in the training and test sets were encoded as count-based ECFP6 fingerprints. A grid search for the highest Matthews Correlation Coefficient (MCC) [55] was performed to obtain the optimal hyperparameters such as c value and class weights in the final model. In addition, the Min-Max kernel was used as a measure of data similarity [56]. Class imbalance was dealt with by assigning class weights inversely proportional to the class frequency.

Molecular docking

Molecular docking was performed to evaluate the binding potential of generated molecules. Glide docking algorithm [57] was employed for the calculation using the Schrödinger software (version 2020). Protein structures for docking were downloaded from the RCSB PDB database. The PDB codes for BRD4, DRD2, FXR and HDAC1 were 3MXF, 6LUQ, 1OSV and 4BKX, respectively. In the former three structures, highly active compounds are co-crystallized. For HDAC1, the crystal structure is in complex with its substrate peptide as a ligand at the active site. At the beginning, the protein structures were prepared with the Protein Preparation Wizard in maestro module. Low energy conformations of ligands were generated by the LigPrep module, in which all tautomers and isomers were enumerated. The Epik algorithm [58] was employed to determine the ionized states at *pH* 7.0. The docking site was defined by a box centered on the crystal ligand using the Receptor Grid Generation module. Then, docking at the Glide-SP precision was carried out with 10 docking poses saved for each compound, and GlideScore was used as the scoring function. The ROCS method [59] in the OpenEye software (version 2020) was applied to measure shape similarity between different binding poses. Kernel density estimation (KDE) [60] analysis was implemented to calculate the statistical probability distribution of docking scores and shape similarity.

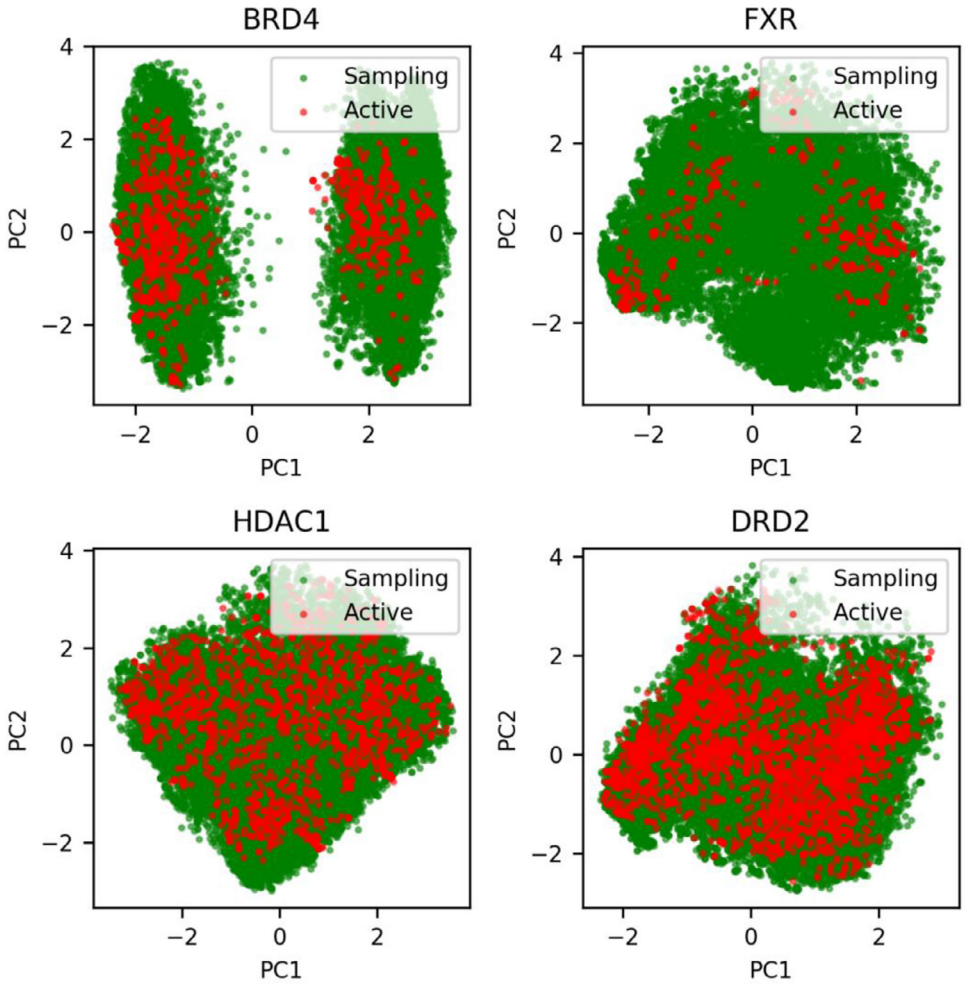

Fig. 2. Chemical space comparison between generated molecules by sampling the transfer learning model (labeled as Sampling) and the active compounds (labeled as Active).

Table 2
Statistics for model evaluation.

Statistics	BRD4		FXR		HDAC1		DRD2	
	TL[a]	GM[b]	TL[a]	GM[b]	TL[a]	GM[b]	TL[a]	GM[b]
generated molecules	2730	2730	2850	2850	9020	9020	20,070	20,070
Validity	76.74%	86.12%	45.16%	90.77%	64.59%	88.08%	73.39%	86.73%
Uniqueness	67.02%	81.62%	72.80%	89.02%	66.34%	86.75%	69.65%	84.95%
Novelty	85.33%	99.53%	87.51%	100.00%	87.19%	100.00%	87.97%	99.53%
Recovery	88.41%	3.86%	60.94%	0.00%	84.18%	0.00%	92.50%	5.23%

Notes: a) transfer learning model built on target specific active compounds.
b) prior model built on the ChEMBL database.

Results and discussion

Transfer learning model

In this study, we applied the SyntaLinker-Hybrid workflow to four targets. For each target, a transfer learning model was constructed. The models were evaluated in terms of validity, uniqueness, novelty and recovery using the test set of terminal fragment pairs. These statistics were also compared with those derived from sampling the prior model using the same fragment set, as the prior model was built for general targets. As shown in Table 2, three targets have more than 60% chemically valid structures among the generated molecules of transfer learning. For FXR, only 45% generated molecules are valid probably due to its smaller training set, although the size of test set is comparable to that for BRD4 (Table 1). In other word, the size of training set may influence the validity of generated molecules. This is also validated by the high validity of generated molecules under the prior model built on a large-scale compound library. The uniqueness of generated molecules for transfer learning models is smaller than that for the prior model for all targets. This is understandable because transfer learning is theoretically focused on the chemical space associated with a target, which may lead to redundant sampling. Thus, the novelty of generated molecules for the transfer learning models is also lower than that for the prior model, although vast majority of them have novel structures. The particular high recovery in-

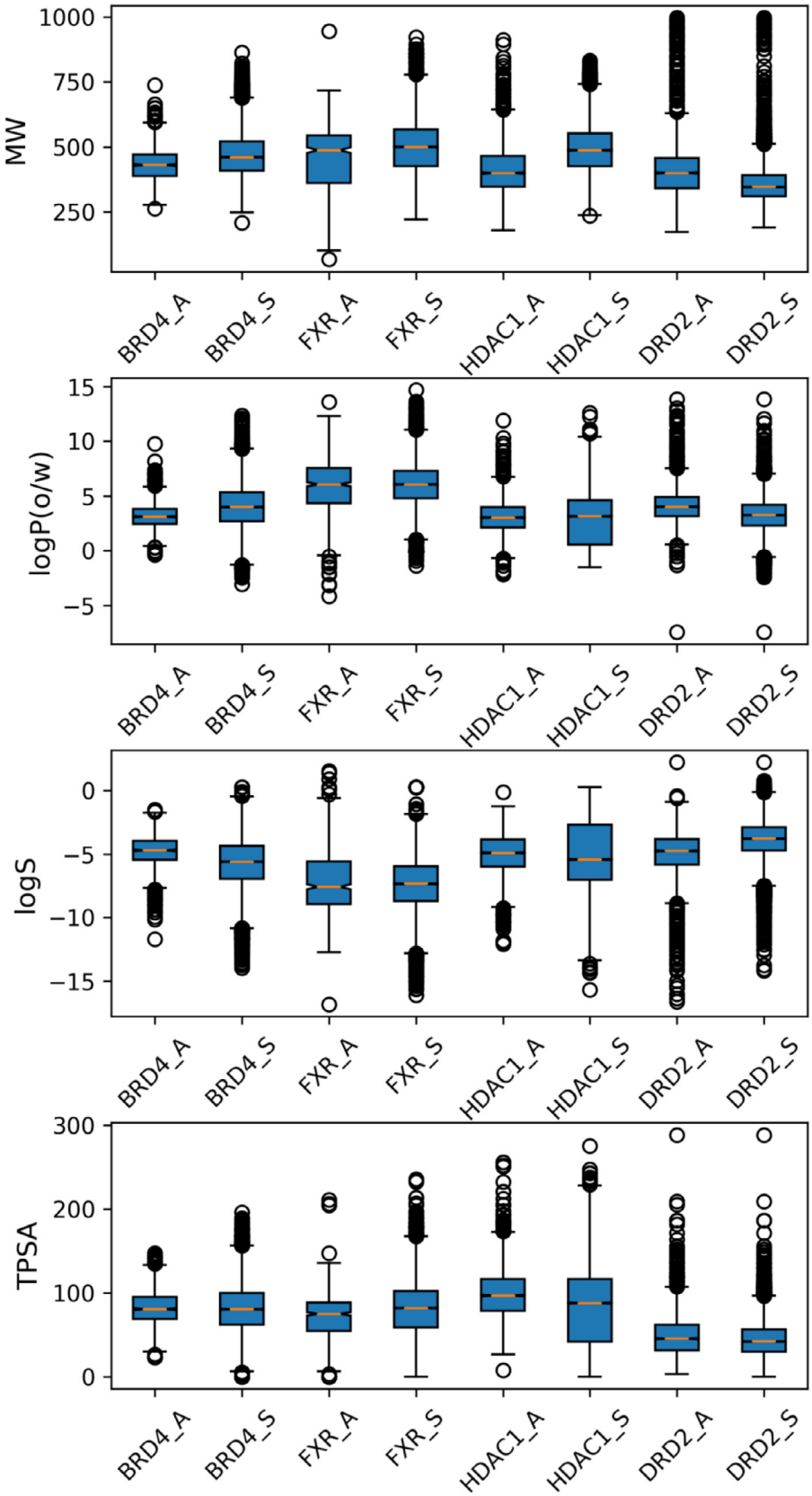

Fig. 3. Boxplots for the physicochemical properties of generated molecules and active compounds. The label S represents sampling molecules, while the label A represents active compounds.

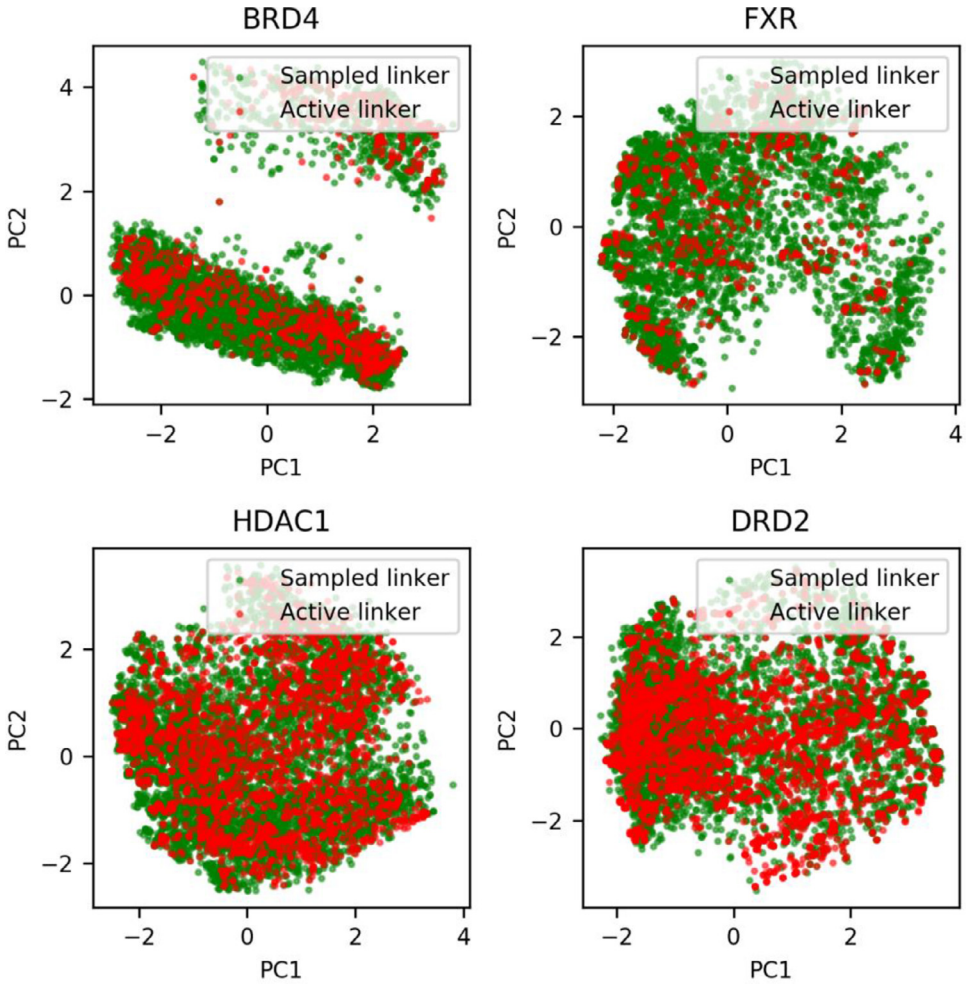

Fig. 4. Chemical space comparison between generated linkers by sampling the transfer learning model (labeled as Sampled linker) and the linkers of active compounds (labeled as Active linker).

Table 3
Molecular generation by sampling the transfer learning model based on fragment hybridization.

Parameters	BRD4	FXR	HDAC1	DRD2
# of 1% hybrid fragment pairs	16,191	26,993	194,719	879,336
# of selected fragment pairs	10,000	10,000	10,000	10,000
# of generated molecules	100,000	100,000	100,000	100,000
# of valid molecules	75,329	63,143	66,468	90,736
# of unique molecules	67,993	56,323	62,618	75,007
# of matched molecules	30,920	23,892	25,705	34,302

dicates that the transfer learning models did learn the assembly rules of the molecules that are located in the target-focused chemical space. Moreover, the results for the last 10 checkpoint models seem to have no big difference (**Table S2**), so the last one will be used for following calculations. In addition, learning rate, batch size and less training steps obviously influence the validity of generated molecules, but have less impact on novelty and recovery (**Table S3**).

Molecular generation

As mentioned, fragment hybridization was introduced to pair any two terminal fragments from different active compounds. As shown in Table 3, the hybridization procedure could produce extremely large amount of fragment pairs through exhaustive enumeration. The number of 1% fragment pairs is as high as tens of thousands. Considering computational cost, we randomly selected 10,000 fragment pairs for model sampling, and ten molecules were generated for each pair. In such a situation, the percentage of valid molecules among the generated molecules is comparable to that obtained by previous model evaluation, in which terminal fragments in a pair came from the same compound. The validity for FXR even increases although its percentage is still lower than other targets. Interestingly, the uniqueness values are improved for all targets, which is probably attributed to the fact that most hybrid fragment pairs did not exist in the training set of transfer learning models. In particular, most targets have over 40% unique valid molecules containing the input terminal fragments (we named them matched molecules).

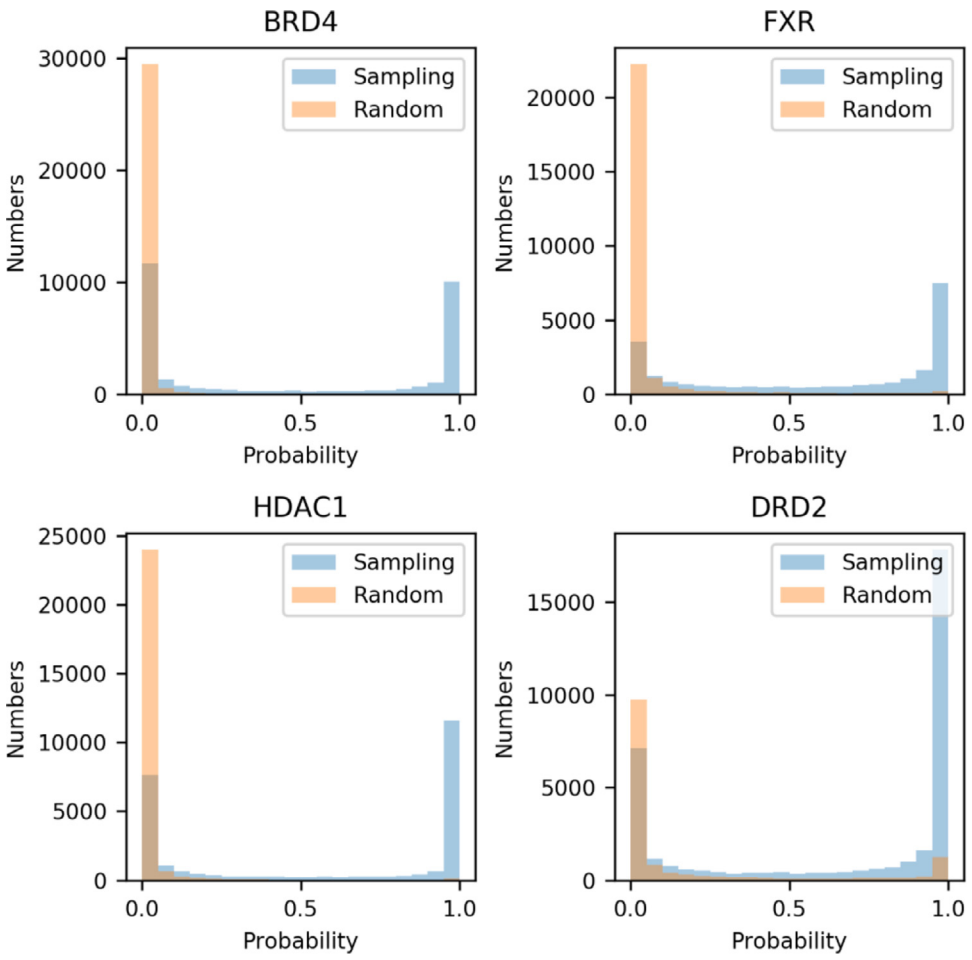

Fig. 5. Histogram distribution of probability to be active for the molecules generated by sampling the transfer learning model (labeled as Sampling) and the randomly selected compounds (labeled as Random).

In summary, it is feasible to utilize the hybrid fragment pairs to sample the transfer learning models even if most fragment pairs never appear in the training set. It is worth noting that some fragments might not coexist in real-life situations, we carefully examined these generated molecules by chemical space analysis, activity prediction and molecular docking.

Chemical space coverage of generated molecules

We particularly focused on the chemical space of matched molecules, because these molecules fit the spirit of SyntaLinker-based fragment linking algorithm. As shown in Fig. 2, these molecules perfectly cover the whole chemical space of available active compounds for all targets, which indicates that transfer learning is effective to locate the target specific chemical space. Among the four targets, the molecular set of DRD2 shows more chemical space overlapping with available active compounds, although the matched molecules occupy only a small portion of valid molecules. This is probably because the active compounds of DRD2 have higher structural diversity (**Fig. S1**). Similar situation exists for HDAC1. BRD4 and FXR have relatively less diverse active compounds, so the overlapping regions are smaller. However, considerably large latent chemical space was explored by the transfer learning models of these two targets, and the structural diversity of generated molecules also increases (**Fig. S1**). This suggests that the transfer learning models can effectively explore the target-focused chemical space. This is also reflected in the uniform sampling in the chemical space of the four targets. Structural similarity between the generated molecules and active compounds showed that many novel structures were generated (**Fig. S2**).

The physicochemical properties of generated molecules were compared to those of active compounds, such as MW, logP(o/w), logS and TPSA. Overall, generated sets have similar range of values with their corresponding active sets (Fig. 3**).** Some minor differences still exist. Moreover, the difference of distribution is not always consistent among all data sets. For example, the average molecular weight of generated molecules seems to be slightly larger than that of active compounds except for the DRD2 data set. However, the molecular weights of linkers and hybrid fragment pairs are individually comparable to those of active compounds despite larger linker size (**Fig. S3**). We assumed that the molecular weights of generated molecules are decided by the combination of linker and terminal fragments. Similarly, logP(o/w) and logS values have wider range for BRD4, HDAC1 and DRD2, but the generated set of FXR shows a narrow range compared to its active compound set. The TPSA scope was expanded for all targets, which indicates generated molecules may have more heteroatoms. In addition, we did synthesizability analysis and compared it with that of known active compounds. We can see that generation set and active set exhibited the same range of SA score (**Fig. S4**). The low average SA score demonstrated that generated molecules may have considerable synthesizability.

Chemical space coverage of linker fragments

As the SyntaLinker model functions as a machine to learn the linking rule between two terminal fragments in a molecule, we further com-

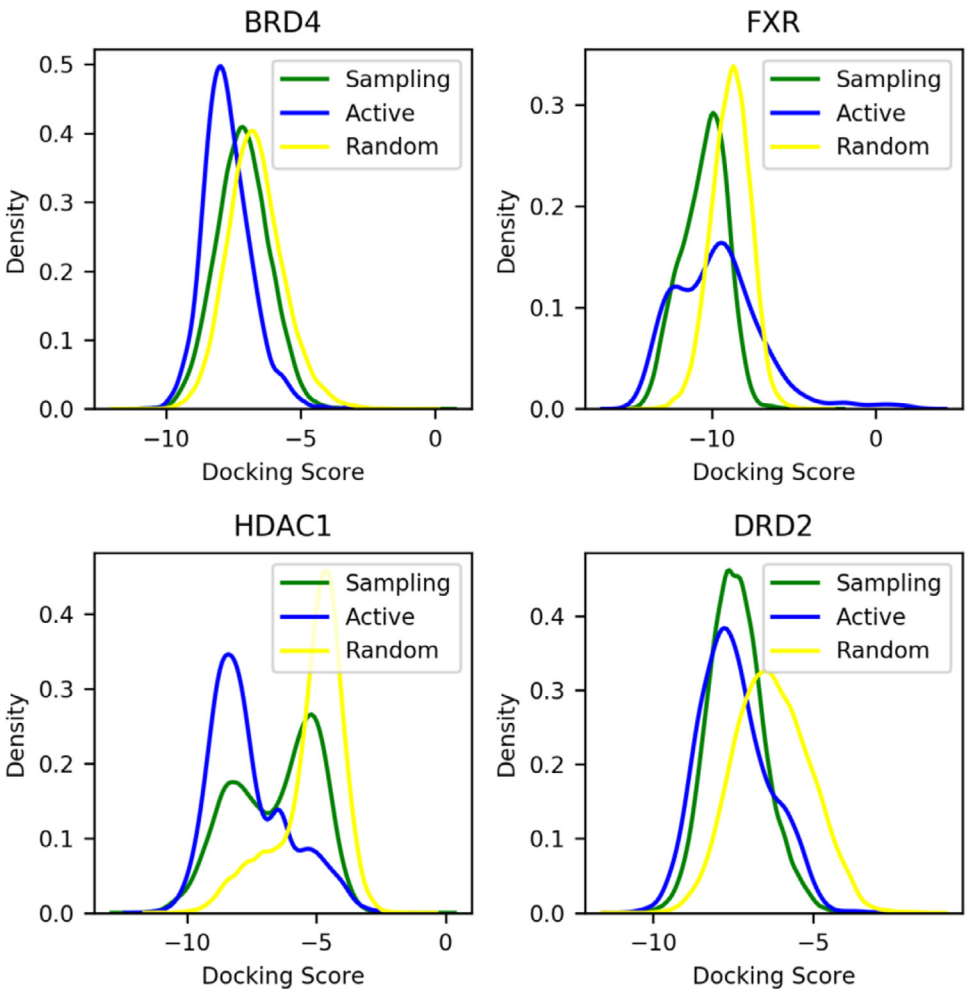

Fig. 6. KDE distribution of docking scores for the molecules generated by sampling the transfer learning model (labeled as Sampling), available active compounds (labeled as Actives) and the randomly selected compounds (labeled as Random).

Table 4
Number of linkers extracted from generated molecules and active compounds.

Target	active compound linkers	generated linkers
BRD4	913	7964
FXR	558	5057
HDAC1	2324	7794
DRD2	2940	6578

pared the structures of linker fragments between generated molecules and active compounds. For the comparison, we specifically extracted the linker fragments of matched molecules. For all targets, several folds of linkers were generated compared to the number of active compound linkers (Table 4). This to some extent explains the expanded chemical space coverage of generated molecules, because HDAC1 and DRD2, which have consistent chemical space coverage between generated molecules and active compounds, only show small increase in number of generated linkers. In fact, generated linkers have the same trend in chemical space coverage as the generated molecules for all targets (**Fig. 4**), although most generated linkers do not exist in available active compounds. Like generated molecules, generated linkers possess small structural similarity with active compound linkers (**Fig. S5**). In other word, structural novelty of generated linkers can be easily achieved by the transfer learning models. Moreover, generated linkers also have higher structural diversity than the linkers extracted from active compounds (**Fig. S6**).

Activity prediction

In order to further evaluate target specificity of generated molecules, we built a SVM classification model for each target. The molecular set comprising of true active compounds and decoy molecules was used to build the SVM model. As shown in **Table S4**, the SVM models for all target exhibit strong predictive ability. The high validation kappa scores indicate that the models even have good quality to predict compounds [61]. Then, we applied the models to predict matched molecules. As shown in **Fig. 5**, a majority of matched molecules for FXR, HDAC1 and DRD2 are predicted to be active compounds. Even for BRD4, there are still about half matched molecules predicted to be active compounds. Moreover, the proportion of active hits in this molecular set is obviously higher than that in the compound set randomly selected from the ChEMBL database. This confirms that the transfer learning models have the capacity to generate target specific molecules. In addition, we extracted the molecules with probability larger than 0.9 as active hits, and found that their chemical space covered active compounds as well and explored some additional space (**Fig. S7**). Moreover, they have similar distribution with the whole generation sets regarding to structural similarity with active compounds (**Fig. S8**).

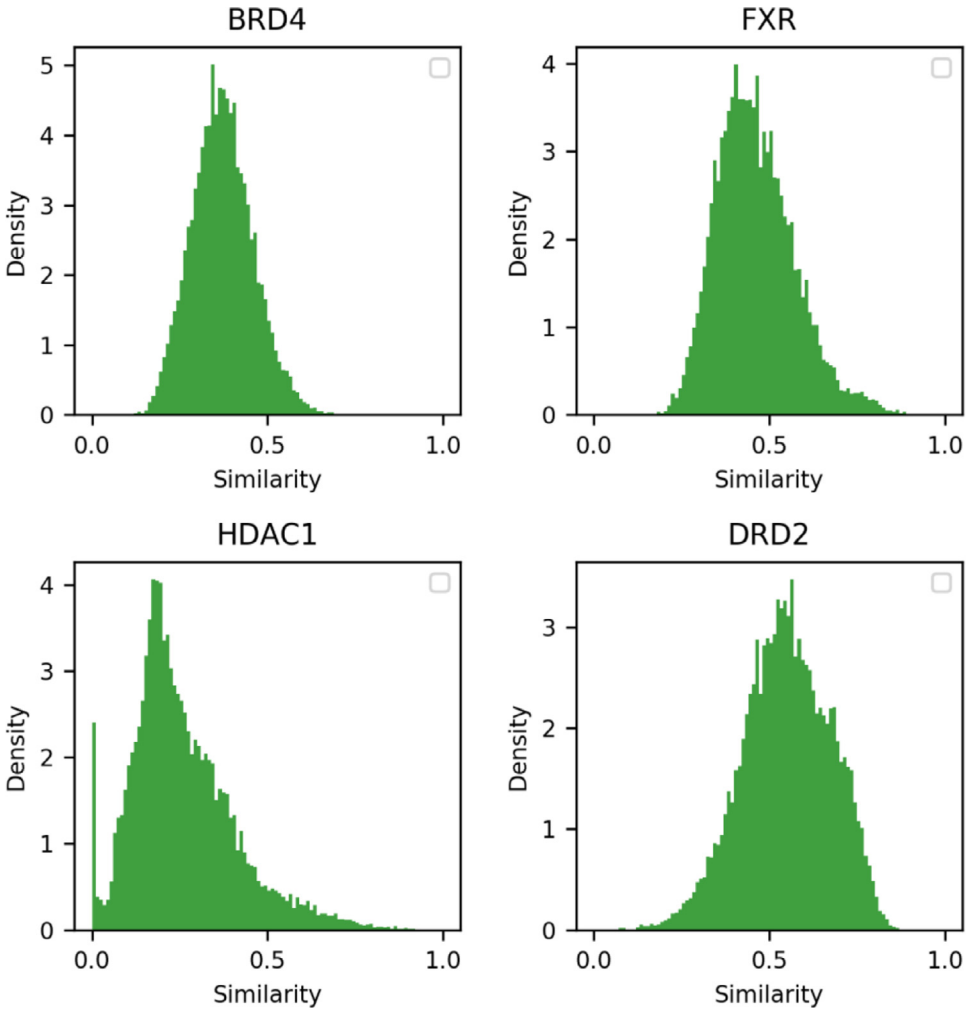

Fig. 7. KDE distribution for the shape similarity of binding poses between the generated molecules and the reference active compounds.

Binding evaluation

We investigated the binding potential of generated molecules by molecular docking. For simplicity, only matched molecules were prepared for docking. Over 90% generated molecules can be successfully docked to the targets (**Table S5**) The docking results were compared with those for active compounds as well as some randomly selected compounds (Fig. 6). With respect to docking score, the generated set of DRD2 shows overall better docking results than those randomly selected compounds (Lower scores represent better docking results), and its score distribution is much similar with known active compounds. For FXR and HDAC1, only a small portion of generated molecules have better scores than randomly selected compounds, and these molecules also exhibit similar score distribution with known active compounds. The generated set of BRD4 does not show significantly better scoring results than the randomly selected compounds. But, this would be acceptable because the active compound set of BRD4 is in the same situation. Moreover, we found that the best 10% scored molecules were mostly predicted to be active compounds (probability > 0.95) (**Table S5**). By contrast, the proportion of active hits in the worst 10% scored molecules is small. This means it is probable to find active compounds from the best scored list.

Subsequently, the binding poses of generated molecules were compared to the active compounds. For BRD4, FXR and DRD2, we selected the crystal binding poses of the ligands co-crystallized in the docking protein structures as the reference. As the ligand in the crystal structure of HDAC1 is a peptide, the compound MS-275 in clinical trials was docked to the active site, and its docking pose was used as the reference. The comparison shows that FXR and DRD2 have a large portion of molecules with shape similarity larger than 0.5, but only a small portion for BRD4 and HDAC1 (Fig. 7). This result seems to be associated with the docking results, as large amount of generated molecules for the latter two targets have worse docking results than the active compounds. Furthermore, the proportion of highly similar molecules (similarity > 0.5) among the best scored 10% molecules is larger than that among the worst 10% score ones. This means better scored molecules usually have similar binding poses with the active compounds.

Finally, we examined the binding mode of generated molecules. One representative molecule with high shape similarity among the best 10% scored molecules was selected for the analysis **(Figs. S9–S12)**. Structures of the selected molecules and corresponding reference compounds were shown in **Table S6**. It can be seen that the molecules for FXR and HDAC1 have highly similar structures with the reference compounds. A linker group in the reference compound was replaced by a similar structure. So, the binding poses were perfectly aligned with each other (Fig. 8). The molecules for BRD4 and DRD2 have totally different structures from the reference compounds, but their binding poses were also superimposed in the same orientation. Moreover, the key pharmacophore features were well matched.

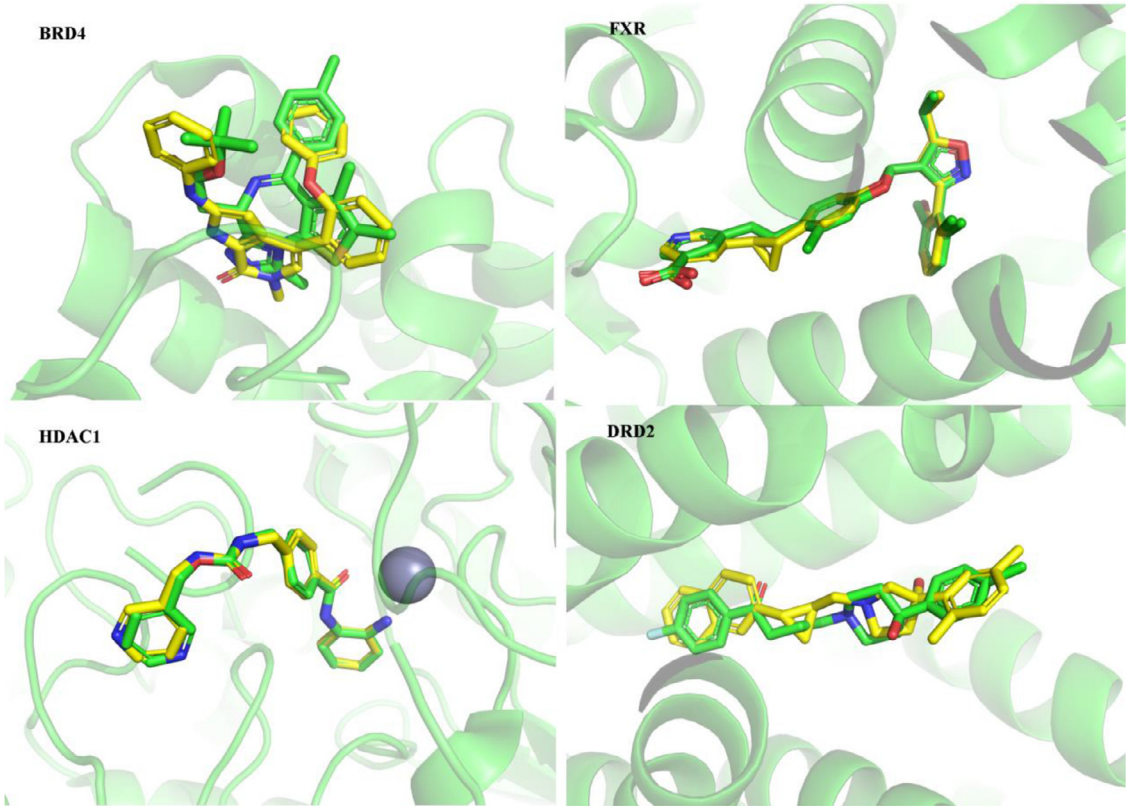

Fig. 8. Binding poses of the representative generated compounds (Yellow sticks) aligned with the reference active compounds (Green sticks). The protein was shown as green cartoons.

Conclusion

In this study, we extended the SyntaLinker-Hybrid workflow to four targets belonging to different protein families. All transfer learning models exhibited strong ability to generate target specific structures. A large portion of chemically valid molecules contained the terminal fragments extracted from active compounds. And, molecular diversity and structural novelty were highly qualified. On the other hand, fragment hybridization of terminal fragments did not influence molecular generation, although almost all fragment combinations have ever been unseen in available active compounds. Oppositely, the validity of generated molecules under fragment hybridization was comparable to the case where the terminal fragments to be assembled came from the same compound. More importantly, the generated molecules covered the entire chemical space of active compounds, and the latent target-focused space was explored. This was also confirmed by the chemical space analysis of generated linkers. Furthermore, we employed QSAR model and molecular docking to evaluate target specificity. The results demonstrated that a majority of generated molecules had the potential to be active compounds, and they not only had good docking scores but also exhibited similar binding poses with the active compounds. Especially, the ones with highly similar binding poses among the best scored molecules had similar structures or pharmacophore features. Therefore, it is very likely to find new active compounds from the generated molecules using this deep learning method. We believe that the SyntaLinker-Hybrid workflow can be extended to more drug targets for target specific drug design.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Dr. Miru Tang for her assistance in language revision.

Funding

This work was supported by Guangzhou Basic and Applied Basic Research Foundation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2022.100035.

References

[1] Satyanarayanajois SD, Hill RA. Medicinal chemistry for 2020. Future Med Chem 2011;3:1765–86.
[2] Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature 2004;432:855–61.
[3] Munos B. Can open-source drug R&D repower pharmaceutical innovation? Clin Pharmacol Ther 2010;87:534–6.
[4] Fabian LV, Thomas C, Karina MM, Marc AG, Adel N, Richard AH, Jose LMF. Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. Comb Chem High Throughput Screen 2011;14:475–87.
[5] Appleton T. Combinatorial chemistry and HTS – feeding a voracious process. Drug Discov Today 1999;4:398–400.
[6] Homon CA, Nelson RM. High-throughput screening: enabling and influencing the process of drug discovery. In: The process of new drug discovery and development. CRC Press; 2006. p. 97–120.
[7] Rienzo M, Jackson SJ, Chao LK, Leaf T, Schmidt TJ, Navidi AH, Nadler DC, Ohler M, Leavell MD. High-throughput screening for high-efficiency small-molecule biosynthesis. Metab Eng 2020.
[8] Walters WP. Virtual chemical libraries. J Med Chem 2019;62:1116–24.
[9] van Hilten N, Chevillard F, Kolb P. Virtual compound libraries in computer-assisted drug discovery. J Chem Inf Model 2019;59:644–51.

[10] Kodadek T. The rise, fall and reinvention of combinatorial chemistry. Chem Commun 2011;47:9757–63.
[11] Saldívar-González FI, Huerta-García CS, Medina-Franco JL. Chemoinformatics-based enumeration of chemical libraries: a tutorial. J Cheminform 2020;12:64.
[12] Hoffmann T, Gastreich M. The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today 2019;24:1148–56.
[13] Gong Z, Hu G, Li Q, Liu Z, Wang F, Zhang X, Xiong J, Li P, Xu Y, Ma R. Compound libraries: recent advances and their applications in drug discovery. Curr Drug Discov Technol 2017;14:216–28.
[14] Claudio NC, Andrew JWO. Ligand docking and structure-based virtual screening in drug discovery. Curr Top Med Chem 2007;7:1006–14.
[15] Bruno OV, Richard E, Maria AM. Structure-based virtual ligand screening: recent success stories. Comb Chem High Throughput Screen 2009;12:1000–16.
[16] 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;60:6081–94.
[17] Yang JF, Wang F, Jiang W, Zhou GY, Li CZ, Zhu XL, Hao GF, Yang GF. PADFrag: a database built for the exploration of bioactive fragment space for drug discovery. J Chem Inf Model 2018;58:1725–30.
[18] Visini R, Awale M, Reymond JL. Fragment database FDB-17. J Chem Inf Model 2017;57:700–9.
[19] Pierce AC, Rao G, Bemis GW. BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. J Med Chem 2004;47:2768–75.
[20] Fechner U, Schneider G. Flux (1): a virtual synthesis scheme for fragment-based de novo design. J Chem Inf Model 2006;46:699–707.
[21] Fechner U, Schneider G. Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. J Chem Inf Model 2007;47:656–67.
[22] Wang LH, Evers A, Monecke P, Naumann T. Ligand based lead generation - considering chemical accessibility in rescaffolding approaches via BROOD. J Cheminform 2012;4:O20.
[23] Taylor RD, MacCoss M, Lawson ADG. Combining molecular scaffolds from FDA approved drugs: application to drug discovery. J Med Chem 2017;60:1638–47.
[24] Nisius B, Rester U. Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design. J Chem Inf Model 2009;49:1211–22.
[25] Sud M, Fahy E, Subramaniam S. Template-based combinatorial enumeration of virtual compound libraries for lipids. J Cheminform 2012;4:23.
[26] De Cao, N.; Kipf, T. MolGAN: an implicit generative model for small molecular graphs. https://arxiv.org/abs/1805.11973 2018.
[27] Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Mol Pharm 2017;14:3098–104.
[28] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 2018;4:120–31.
[29] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4:268–76.
[30] Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. https://arxiv.org/abs/1511.05644v2 2015.
[31] Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. https://arxiv.org/abs/1802.04364 2018.
[32] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23:1241–50.
[33] Prykhodko O, Johansson SV, Kotsias PC, Arús-Pous J, Bjerrum EJ, Engkvist O, Chen H. A de novo molecular generation method using latent vector based generative adversarial network. J Cheminform 2019;11:74.
[34] Blaschke T, Bajorath J. Compound design using generative neural networks. Artificial Intelligence in Drug Discovery 2020;75:217.
[35] Podda M, Bacciu D, Micheli A, Silvia C, Roberto C. A deep generative model for fragment-based molecule generation. In: Proceedings of the twenty third international conference on artificial intelligence and statistics, 108; 2020. p. 2240–50. editors PMLR: Proceedings of Machine Learning Research.
[36] Hadfield, T.; Imrie, F.; Merritt, A.; Birchall, K.; Deane, C. Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. https://www.biorxiv.org/content/10.1101/2021.10.21.465268v1 2021.
[37] Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O. SMILES-based deep generative scaffold decorator for de-novo drug design. J Cheminform 2020;12:38.
[38] Imrie F, Bradley AR, van der Schaar M, Deane CM. Deep generative models for 3D linker design. J Chem Inf Model 2020;60:1983–95.
[39] 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;11:8312–22.
[40] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014;42:D1083–90.
[41] Yosinski J, Clune J, Bengio Y, Lipson H. In how transferable are features in deep neural networks? Adv Neural Inf Process Syst 2014:3320–8.
[42] Peters, M.E.; Ruder, S.; Smith, N.A. To tune or not to tune? adapting pretrained representations to diverse tasks. https://arxiv.org/abs/1903.05987 2019.
[43] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform 2017;9:48.
[44] Hu L, Yang Y, Zheng S, Xu J, Ran T, Chen H. Kinase inhibitor scaffold hopping with deep learning approaches. J Chem Inf Model 2021;61:4900–12.
[45] Sun J, Jeliazkova N, Chupakin V, Golib-Dzib JF, Engkvist O, Carlsson L, Wegner J, Ceulemans H, Georgiev I, Jeliazkov V, Kochev N, Ashby TJ, Chen H. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. J Cheminform 2017;9:17.
[46] Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 2010;50:339–48.
[47] Landrum, G. RDKit: open-source cheminformatics software, 2018.
[48] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 2002;42:1273–80.
[49] Zhenhua Zheng YX, Zhang Bo, Wang Dong. Xiaoling Liu An efficient method for K-means clustering. Pattern Recognit Artif Intell 2010;23:516–21.
[50] 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 Comput Sci 1999;39:747–50.
[51] Blomberg N, Cosgrove DA, Kenny PW, Kolmodin K. Design of compound libraries for fragment screening. J Comput Aided Mol Des 2009;23:513–25.
[52] Pulikkal PB, Marunnan MS, Bandaru S, Yadav M, Nayarisseri A, Sureshkumar S. Common SAR derived from linear and non-linear QSAR studies on AChE inhibitors used in the treatment of Alzheimer's Disease. Curr Neuropharmacol 2017;15:1093–9.
[53] Swami A, Jain R. Scikit-learn: machine learning in Python. J Mach Learn Res 2013;12:2825–30.
[54] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 2012;55:6582–94.
[55] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[56] Ralaivola L., et al. Graph kernels for chemical informatics. Neural Netw 2005;18:1093–110.
[57] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47:1739–49.
[58] Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 2007;21:681–91.
[59] Hawkins PCD, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. J Med Chem 2007;50:74–82.
[60] Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. J Am Stat Assoc 1996;91:401–7.
[61] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82 (Zagreb).