Deep Scaffold Hopping with Multi-modal Transformer Neural Networks

Shuangjia Zheng[1#], Zengrong Lei[2#], Haitao Ai[2], Hongming Chen[3], Daiguo Deng[2]*, Yuedong Yang[1]*

[1]School of Data and Computer Science, Sun Yat-sen University, China, 132 East Circle at University City, Guangzhou 510006, China

[2]Fermion Technology Co., Ltd, 1088 Newport East Road, Guangzhou 510335, China

[3]Centre of Chemistry and Chemical Biology, Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou 510530, China

Abstract

Scaffold hopping, aiming to identify molecules with novel scaffolds but share a similar target biological activity toward known hit molecules, has always been a topic of interest in rational drug design. Computer-aided scaffold hopping would be a valuable tool but at present it suffers from limited search space and incomplete expert-defined rules and thus provides results of unsatisfactory quality. To addree the issue, we describe a fully data-driven model that learns to perform target-centric scaffold hopping tasks. Our deep multi-modal model, DeepHop, accepts a hit molecule and an interest target protein sequence as inputs and design bioisosteric molecular structures to the target compound. The model was trained on 50K experimental scaffold hopping pairs curated from the public bioactivity database, which spans 40 kinases commonly investigated by medicinal chemists. Extensive experiments demonstrated that DeepHop could design more than 70% molecules with improved bioactivity, high 3D similarity, while low 2D scaffold similarity to the template molecules. Our method achieves 2.2 times larger efficiency than state-of-the-art deep learning methods and 4.7 times than rule-based methods. Case studies have also shown the advantages and usefulness of DeepHop in practical scaffold hopping scenario.

Introduction

Over the past decades, the rapid developments of both high-throughput screening (HTS)[1] and fragment-based screening technologies[2] have largely facilitated the hit identification process for drug discovery. These screening strategies, together with the combinatorial compound library, discover extensive collections of diverse chemical series as starting points (hits) for further work towards identifying clinical candidates.[3] However, in general these identified hits have weak potency and do not necessarily possess ideal ADMET profile. Usually lead identification (LI), i.e. transform a hit to a more potent lead compound), and lead optimization (LO) process is needed to further optimize their structures to improve the potency and ADMET properties before the drug discovery project can be moved into clinic.

One common strategy in LI is scaffold hopping. The goal of scaffold hopping is to discover structurally distinct compounds starting from a known reference compound for a given target protein by changing the backbone of the compound, while the three-dimensional structure or the pharmacophore of the original compound is largely unchanged to preserve the biological activity[4-5]. This terminology was coined by Schineider and co-workers[5] and has been widely used, as such design can result in novel hit series and improved molecular properties including activity, toxicity, synthetic accessibility, and selectivity. However, when a “hop” is applied to a hit molecule, there is no guarantee that the corresponding transformation will work in an expected way. This can be due to an incomplete understanding of the protein-ligand interaction mechanism[6], unfavorable ADMET properties of hopped structure, or activity cliff.[7] Predicting which hops can be successful is quite challenging even for the best human medicinal chemists.

To this end, a variety of computational methods have been proposed to help chemists to find better scaffold hops.[4, 8-9] Many of them utilized matching algorithms based on two-dimensional (2D) or three-dimensional (3D) molecular representations[10-14]. Others developed fragment replacement techniques, including fingerprints and combining experimental information[15-19]. However, almost all methods published to date relied exclusively on a predefined database to select a molecule or a fragment, with the differences between approaches arising solely from the searching algorithms of the database, the ways to define similarity of compound pairs or the contents of the scanned database. As a result, these methods are inherently restricted to a set of pre-defined rules or examples, limiting the exploration of vast chemical space.

In parallel, upon the call for the more exhaustive and intelligent exploration of chemical space, the field of *de novo* molecule design has been advanced by recent breakthroughs in deep generative models[20-21]. Various generative architectures, including RNNs[22-24], autoencoders[25-27], and generative adversarial networks (GANs)[28] have been proven to be effective methods for generating desirable molecules, which are either represented by simplified molecular input line entry specification (SMILES)[29] or molecular graph. However, all the methods aim for the same goal, which is to design structurally "diverse" compounds from scratch, has the capability to search in the whole drug-like space and doesn't rely on any predefined database or rules, no matter for scaffold hopping or derivative generation.

More recently, two different lines of research are carried out for molecule design under scaffold constraint. One research line is named as scaffold-based molecule design worked by Lim et al.[30] and Li et al.[31], where the graph generative models were applied to extend a given scaffold by sequentially adding atoms and bonds. In this context, the generated derivatives are guaranteed to maintain the scaffold with certainty, and their properties can thus be controlled by conditioning the generation process on desired properties. However, due to the nature of these tasks, the shapes of the suggested molecules often differ significantly from the starting points in the 3D level, and many of the proposed transformations are R-group modifications.[32] The other research line is referred to as fragment linking. The original idea of Imrie and co-workers is to join fragments together with a generated linker while keeping the relative conformations of the fragments[33]. Yang et al. further extended it as a sentence completion problem with transformer neural networks[34]. Although these approaches claim its capability in scaffold hopping since they can generate molecules with high 3D similarity to the original molecule, the 2D similarity is often higher than expected due to the nature of fragment replacement, resulting in unfavorable intellectual property issues. Moreover, all these models were trained in a ligand-based paradigm using a large number of bioactive compounds in a broad sense from public database, regardless of the correspondent relationship between bioactivity and specific target protein. This imposes a limitation on applying these methods into the target-centric drug development process.

In this work, we introduce a novel target-based scaffold hopping framework, DeepHop, to optimize hit/lead compounds based on a multi-modal deep generative model. Our method takes as input a reference molecule X and a specified protein target sequence Z, and designs a scaffold hop Y incorporating 2D and 3D structural information, protein target information, as well as bioactivity information. The model has been trained with over 50K constructed scaffold hopping pairs across 40 kinases. Extensive experiments show that our model is capable of generating bioisosteric molecular structures for seed molecules with novel backbones but improved activity. More importantly, our model could be easily extended to new protein targets outside the training set, which is essential for target-centric drug development.

Methods

Task Definition. An exemplary scaffold hop is shown in Figure 1. In this work, we broadly define a scaffold hopping process as such: given an input reference molecule X and a specified protein target Z, the model predicts the “hopped” molecule Y with improved pharmaceutical activity and dissimilar 2D structure but similar 3D structure.

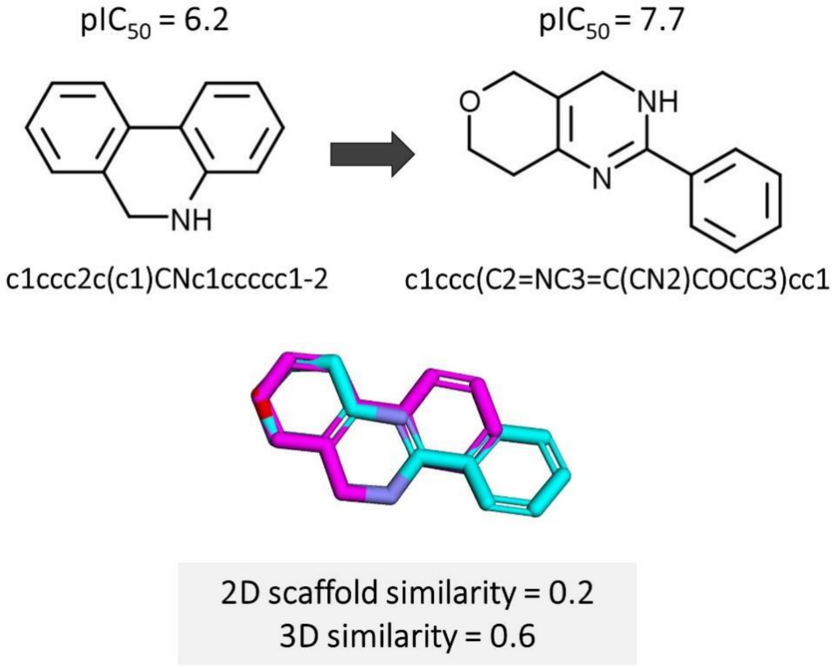

Figure 1. A typical scaffold hop extracted from tankyrase-2 inhibitors.[4] The two compounds have different scaffolds (2D Tanimoto scaffold similarity = 0.2), but similar 3D shapes (3D shape and pharmacophore similarity =0.6).

Data preparation. There have only been a limited number of successfully reported examples for scaffold hopping. As such, for training and large-scale evaluation, we constructed sets of scaffold-hopping pairs using a custom-made similarity scoring function from a subset of ChEMBL20.[35]

Specifically, we first processed the ChEMBL20 dataset by filtering kinase-related target proteins with at least 300 up to 5000 unique bioactivity instances. The scaffold hopping application in the kinase family has always been a topic of interest because the kinase patent literature is notoriously complicated.[36] We further filtered out the SMILES strings containing disconnected ions or fragments. The molecules were then normalized using RDKit, which involved the removal of salt and isotopes, as well as charge neutralization. After the preprocessing, the final data set contained 103,511 bioactivity data points across 152 kinases (Table S1). Note that we used pChEMBL values as the standard activity unit, which were defined as: -Log(molar IC50, XC50, EC50, AC50, Ki, Kd or Potency).

Multi-task QSAR model. Before constructing the scaffold hopping pairs, one important factor required to assess the performance of scaffold hopping is whether the generated molecules have similar bioactivity on the desired targets. To enable a rapid and accurate profiling of generated molecules, virtual profiling models were trained on all the data points in the above-curated kinase datasets. We evaluated several state-of-the-art messages passing neural networks (MPNNs)[37-38] and multi-task deep neural networks (MTDNN)[39] with molecular graphs or molecular fingerprints as the molecular representations. MTDNN was found to obviously outperform other models with an average R^2 of 0.62 and RMSE of 0.61 (pCHEMBL value) on internal test sets. Thus, the MTDNN model was used as the virtual profiling model in the following studies. All the modeling details and results are shown in Supplementary files.

For the quality of the virtual bioactive assessments, we kept only targets that had a 5-fold cross-validation R^2 higher than 0.70, resulting in 40 targets in the end.

Construction of scaffold hopping pairs. The scaffold hopping definition emphasized on two key components: (i) different core structures and (ii) similar topology and pharmacophore that ensure biological activities of the new compounds relative to the parent compounds. To mimic the scaffold hopping scenario, we constructed our data set following the idea of matched molecular pairs (MMPs) cutting algorithm proposed by Hussain et al[40]. More specifically, we sampled **target-based hopping pairs** ((X; Y)|Z) with significant bioactivity improvement (pCHEMBL Value ≥ 1) for new compound Y over original compound X in the context of protein Z and a strict molecular similarity condition of (2D scaffold similarity (X; Y) ≤ 0.6) $\cap$ (3D similarity (X; Y) ≥ 0.6).

The 2D scaffold similarity was measured by the Tanimoto score over Morgan fingerprints[41] of the compound scaffolds and 3D molecular similarity by the shape and color similarity score (SC score) used by Imrie et al.[33]. To provide molecular 3D structural information, we generated 3D conformers for the curated data set using RDKit and took the lowest-energy conformation among 100 samples for each molecule. The SC score was computed by the pharmacophoric feature similarity[42] and the shape similarity[43] between a pair of molecular conformers. The SC score is a float value in the range of [0, 1], and a higher value represents a higher similarity between the generated molecule and its parent compound. Scores above 0.6 indicate a fair structural match, and those above 0.8 indicate an excellent one.

To avoid redundancy of training pairs, we only allowed up to 10 hops for each source molecule. For each target, we first randomly took 10% molecules as the test set. The rest 90% molecules were constructed into scaffold hopping pairs and randomly divided into two sets of a ratio of 9:1 for training and validating. These processing steps resulted in a training set of 57,537 pairs over 40 kinases.

Independent test set. To explore the generalization ability to proteins that have never been observed during the training process, we further used CDhit[44] to retrieve six typical target proteins (the R^2 of QSAR models of these six proteins are higher than 0.65) from the curated database as the independent test set based on the sequence identity. Among them, three proteins (CHEMBL2208, CHEMBL2147, CHEMBL2523) are non-homologous with sequence identity less than 25% to any sequence in the training set, while others (CHEMBL4225, CHEMBL2292, CHEMBL2041) are homologous to the training set with the highest sequence identities of 59%, 63%, 76%, respectively. The compounds in these six proteins would never be observed in the previous training, validating, and testing processes. The details of these six proteins have been shown in Table S2.

Model Architecture. A novel multi-modal graph transformer model was proposed for generating scaffold hops with inputs of a source molecule and a protein sequence based on the transformer architecture[45]. Transformer, as a classical encoder-decoder architecture, has recently shown the state of the art performances in many sequence-to-sequence translation tasks, including machine translation[46], retrosynthesis[47], and fragment assembly[34]. In previous chemical applications like retrosynthesis and fragment assembly, chemical structures were often simply converted into SMILES strings that ignored spatial information naturally embedded in chemical 3D conformers. In addition, none of them considered the protein target information during the transformation of the molecule pairs. Obviously, both of these two features play crucial roles in the scaffold hopping task that need to be considered.

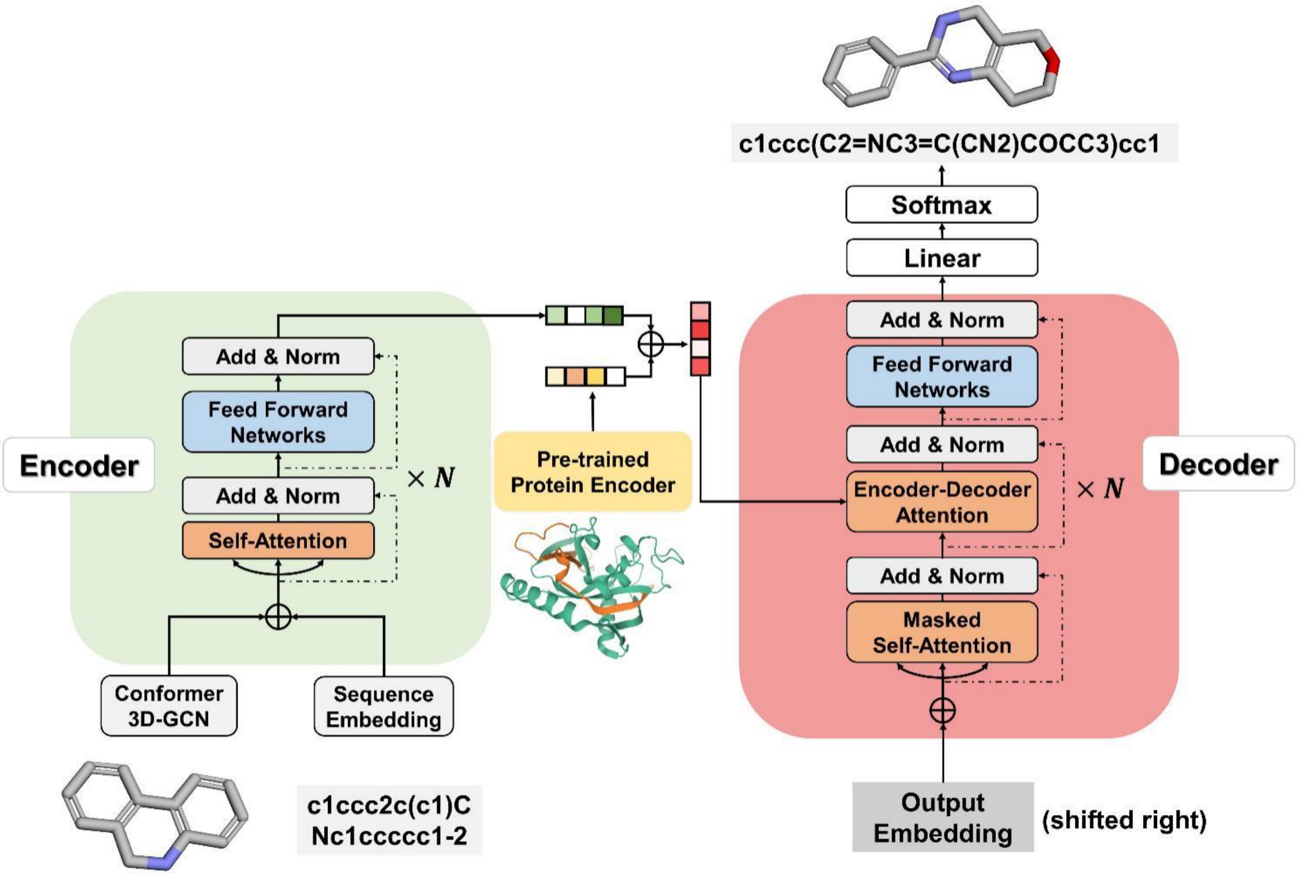

Figure 2. The basic architecture of the multi-modal transformer model DeepHop. The model comprises three main components: (1) a 3D graph neural network for molecular conformer embedding, (2) a pre-trained encoder for the target protein embedding, and (3) a transformer for mapping the scaffold hopping pairs.

To this end, DeepHop (as shown in Figure 2) introduces a graph neural network (GNN)-based molecular conformer encoder and a protein encoder to ensure the conformational information as well as the protein information to satisfy the explicit criterion. Basically, DeepHop comprises three main components: (1) a 3D graph neural network (GNN) for molecular conformer embedding, (2) a pre-trained encoder for target protein embedding, and (3) a transformer for mapping the scaffold hopping pairs.

3D graph conformer encoder. We adopted a simple 3D spatial GNN as the conformer encoder following the strategy of Danel et al.[48], which can learn both the molecular graph representation and spatial distances between atoms in the 3D space. The GNN follows the paradigm of message passing neural networks. The input of the conformer encoder is a 3D molecular graph $G = (V, E)$, where $V = \{ v_1, \dots, v_n\}$ denotes a set of nodes (atoms) and $E = [e_{ij}]_{i,j=1}^{n}$ represents edges (bonds) between atoms *i* and *j*. Each atom v_i is represented by a *d*-dimensional initial feature vector h_i containing the 2D chemical features computed by RDkit (See more details in Table S3). The atom is additionally attached with its 3D coordinates $p_i \in \mathbb{R}^t$ obtained by the molecular conformer. The 3D GNN then updates the atom embedding with message passing operations:

$$h_i^{(l+1)}(U, b) = \sum_{j \in N_i} ReLU\left(U^T (p_j - p_i) + b\right) \odot h_j^{(l)})$$

where $U \in \mathbb{R}^{t \times d}$, $b \in \mathbb{R}^d$ are trainable network parameters, *d* is the dimension of the feature vector $h_j^{(l)}$, *l* denotes the updating at the *l*-th iteration and $\odot$ denotes element-wise multiplication.

Herein, the overall atom embeddings can be described as $H^{(l)} = \{h_1^{(l)}, \dots h_n^{(l)}\}$. In the last iteration of the node embedding updating, we introduced a Gated Recurrent Unit (GRU) network[49] to increase the power of the network and obtained the final atom embedding, as shown as

$$\widehat{H}(v) = GRU(H^{(l)}(v))$$

Protein encoder. Compared to the drug molecules, protein molecules are much bigger, typically containing more than 1,000 heavy atoms. To avoid a bulky model that contains too many parameters, we adopted Tasks Assessing Protein Embeddings (TAPE)[50], a recently proposed semi-supervised protein sequence representation learning method, to generate the protein pre-trained embeddings. TAPE was trained by a large transformer neural network in an unsupervised paradigm with millions of protein sequences. After training, it can generate an information-enriched feature vector for an input protein sequence. Formally, a protein can be described as a linear sequence that consists of a list of amino acid residues $P = (r_1, \ldots r_l)$. After processing through the TAPE, a fixed vector H_p can be obtained, where H_p is a k-dimensional pre-trained feature vector for protein sequence P.

Transformer architecture. The fundamental architecture of DeepHop is a typical Transformer neural network containing multiple encoder-decoder modules. Each encoder layer consists of a multi-head self-attention sub-layer and a position-wise feed forward network (FFN) sub-layer. Multi-head attention has several scaled dot-product attention functions working in parallel, which allows the model to focus on messages from different subspaces at different positions. An attention mechanism computes the dot products of the query (Q) with all keys (K), introduces a scaling factor d_k (equal to the size of weight matrices) to avoid excessive dot products, and then applies a softmax function to obtain the weights on the values (V). Formally

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

The FFN sub-layer adopts the ReLU activation.[51] Then, layer normalization[52-53] and a residual connection[54] were introduced to link the above two sub-layers. Each decoder layer has three sub-layers, including an FFN sub-layer and two attention sub-layers. The decoder self-attention sub-layer utilizes a mask function to hinder attending to unseen future tokens. The encoder-decoder attention layer helps the decoder to focus on essential parts in the source sequence, and to capture the relationship between the encoder and decoder.

For a given source molecule, we concatenate the learned 3D graph representations $\widehat{H}_s$ with SMILES sequence embedding $M_s = (s_1, \ldots, s_m)$ and convert them through a simple linear transformation.

The combined multi-modal molecular representations are then sent to the Transformer encoder to convert into a latent representation $L \in \mathbb{R}^{m \times f}$, where *m* is the sequence length of molecular SMILES and *f* is the hidden state dimension. Afterwards, we concatenate L with target protein embedding $H_p \in \mathbb{R}^k$, resulting in a comprehensive representation $\hat{L} \in \mathbb{R}^{m \times (f+k)}$. Given $\hat{L}$, the decoder iteratively generates an output SMILES sequence $Y = (y_1, ..., y_o)$ until the ending token "⟨/s⟩" is generated.

During training, the model minimizes the cross-entropy loss between the target sequence $M_t = (t_1, ..., t_k)$ and the output sequence *Y*.

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

Baseline Models. We compare our approaches with the following baselines:

1. MMPA. We utilized the implementation by Hussain et al.[55] to perform MMPA as a baseline algorithm. Molecular transformation rules were extracted from the kinase dataset for corresponding tasks. During the test, we translated a source molecule multiple times using different matching transformation rules and selected the translation with the highest average bioactivity as scored by the virtual profiling model.
2. Seq2seq. Our second baseline is a seq2seq model that utilizes SMILES strings to encode molecules. The model has been successfully applied to other molecule transformation tasks[56].
3. G2G. The third baseline is a Graph-to-Graph model[57] that extends the junction variational autoencoder (VAE) via attention mechanism and generative adversarial networks (GAN). The model is capable of translating the current molecule to a similar molecule with pre-defined desired property (e.g., logP).

Notably, these models were not designed for multi-task transformation. For a fair comparison, we choose a single target protein CHEMBL2835, which contained the largest number of scaffold pairs in our training set, as a representative to evaluate the effectiveness of the baselines and our model.

Evaluation Metrics. Our model aims to generate various hops for an input molecule for a specific target. We quantitatively analyze the hopping success rate, bioactivity improvement, validity, uniqueness, diversity, and novelty of different methods.

- ✓ **Success rate** is a metric that considers both similarity and bioactivity improvement. Since the task is to generate a molecule that (i) is similar to the input molecule and (ii) has bioactivity improvement simultaneously. We design criteria to judge whether it satisfies these two requirements: the generated molecule Y should (a) meet the structural condition, i.e., (2D scaffold similarity (X; Y) $\leq$ 0.6)$\cap$(3D similarity (X; Y) $\geq$ 0.6); (b) has a positive bioactivity gain, i.e., pBioactivity(Y) - pBioactivity(X) $\geq$ 0, where the multi-task QSAR models are used to compute the activity of generated molecules. A **constraint success rate** is also accounted by confining the pBioactivity(Y) - pBioactivity(X) $\geq$ 1.
- ✓ **Bioactivity improvement** is the average improvement of biological activity between source molecule and the generated molecule computed as pBioactivity(Y) - pBioactivity(X)
- ✓ **Validity** is the percentage of generated chemically validly generated molecules;
- ✓ **Uniqueness** refers to the number of unique structures generated;
- ✓ **Novelty** refers to the percentage of novel molecules among the chemically validly generated molecules (not present in the training set);
- ✓ **Diversity** is the average pairwise Tanimoto distance between a set of molecules, where Tanimoto distance is defined as dist(X; Y) = 1 − sim(X; Y).

Model Training and Optimization of Hyperparameters. The DeepHop model was implemented based on OpenNMT[58] and all scripts were written in Python[59] (version 3.7). The models were trained on four GPU (Nvidia 2080Ti) and saved checkpoint per epoch. The best hyperparameters were decided based on the loss of the validation set. We adopted the beam search procedure[60] to generate multiple candidates with different beam widths. All generated candidates were canonicalized using RDkit and compared to the source molecules.

Results and Discussion

In this section, we mainly discussed our DeepHop performance from four parts. First, we evaluated our model with different training paradigms on the kinase internal datasets. Then, we compared our methods with other state-of-the-art deep learning models as well as conventional methods on a single target protein. Subsequently, we tested our model in unseen protein sets and performed few-shot transfer learning on proteins with low performance. Lastly, our DeepHop method was applied to several case study examples to demonstrate the capability of the model for practical scaffold hopping.

Model Performance on the multi-kinase dataset. We first assessed the performance of models on the internal test set with different training paradigms (single-task, multi-task and DeepHop). Note that both the single-task and multi-task models did not integrate the protein information. The top 10 candidate sequences for each reference compound were generated (note that the generated sequences are not guaranteed to be valid or unique SMILES strings). Table 1 lists the average results on a total of 40 targets. Our multi-modal DeepHops achieved a success rate of 65.2 ±17.5 and constraint success rate of 43.7 ±21.0, significantly outperforming those by Multi-task model (success rate = 58.9±20.9, constraint success rate = 34.6±19.7) and Single-Task models (success rate = 27.5±15.9, constraint success rate = 15.5±14.7). The poor performance in most of the metrics by the single-task model should be due to the relatively small number of data points for each single kinase task, leading to a fragile model that is difficult to learn the transformation between scaffold pairs. When integrating all the pairs from different kinase sets for multi-task learning, the model is capable of capturing key structural information in molecular translation, achieving a top-10 success rate of 58.9%. However, as there is no protein target information given to the model, its bioactivity improvement is much lower than multi-modal DeepHop, with an average bioactivity improvement of 0.64 versus 0.97. When simultaneously considering the structural condition and bioactivity condition, DeepHop vastly outperforms the multi-task model with an increase of 9.6% on the Constraint Success Rate, indicating that the protein information can help the model learn specific bioactivity information for each target.

Table 1. Performance comparison of different training settings on curated internal test set.

Metrics	**Models**		
	Single-task	Multi-task	DeepHop
Success Rate (%)	27.5±15.9	58.9±20.9	**65.2±17.5**
Constraint Success (%)	15.5±14.7	34.6±19.7	**43.7±21.0**
Improvement	0.53±0.31	0.64±0.28	**0.97±0.24**
Validity (%)	12.9±6.3	92.7±3.8	**95.7±3.8**
Uniqueness (%)	8.7±5.5	**88.2±8.8**	76.4±9.3
Novelty (%)	99.0±5.1	**99.6±0.3**	99.4±0.5

Comparison of other Methods. Since all the previous methods were performed on a single target or property, it is impractical to evaluate them in the same setting. For a relatively fair comparison, we choose the Tyrosine-protein kinase JAK1 (CHEMBL2835), which has the largest number of scaffold hopping pairs in our training set, as a representative. Table 2 summarizes the validity, success rate, and improvement of the molecules generated by different models for comparison. Despite the strict restriction imposed in the scaffolds, DeepHop achieved an order-of-magnitude improvement over the current state-of-the-art baseline G2G model with 2.6 times of increases in both the success rate and the constraint success rate. Though our validity is slightly lower than G2G (95.8% vs 99.5), the 66.2% uniqueness (4.8 times higher than the one by G2G) indicates the quality and diversity of our generated molecules. The low success rates and uniqueness by G2G are probably because the model cannot process the transformation between two 2D dissimilar graphs, as it only focused on the 2D topological information while ignored the 3D spatial information. In addition, the G2G model suffered from a limited chemical space because of the small size of substructure vocabulary derived from the small training set, resulting in poor uniqueness value. On the other hand, though MMPA had higher uniqueness than DeepHop (98.6 vs. 66.2%), it achieved a poor success rate and constraint success rate that are 4.7 and 5.1 times lower than our DeepHop model. In addition, the lower bioactivity improvement indicates that the simple rule-based method cannot handle scaffold hopping well. Compared to these two baselines, seq2seq performed relatively well with a success rate of 32.6%, indicating that the SMILES sequence transformation model is better suited for scaffold hopping tasks. However, the low uniqueness by the seq2seq model suggests that the generated hops are also stuck into a limited chemical space. In summary, DeepHop has a balanced performance in all metrics.

Table 2. Performance comparison of models on the Tyrosine-protein kinase JAK1 (CHEMBL2835).

Metrics	**Models**				
	Validity (%)	**Success rate (%)**	**Constraint success rate (%)**	**Improvement**	**Uniqueness (%)**
MMPA	**100**	15.7	6.6	0.29	**98.6**
Seq2seq	34.6	32.6	13.4	0.28	16.9
G2G	99.5	27.7	13.1	**1.11**	13.9
DeepHop	95.8	**73.2**	**33.8**	0.41	66.2

Properties of the Generated Molecules. We further dug into the property of generated molecules by comparing Drug-likeness score (QED score), the calculated water-octanol partition coefficient (logP), molecule weight (MW), bioactivity improvement, and similarity with source ones.

For each source compound in the test set, the properties of its top 10 candidates generated by different models were calculated. As shown in Fig 3a-c, the hopping molecules generated by DeepHop are similar to the source molecules in the distribution of QED, logP, and MW. In contrast, there is a large deviation in the distribution among source molecules and molecules generated by G2G and MMPA, suggesting that both methods cannot maintain the properties of source molecules when performing the molecular transformation. Though Seq2seq also generated a similar distribution to the source molecules, it has much smaller diversity and scale due to its much lower validity and uniqueness.

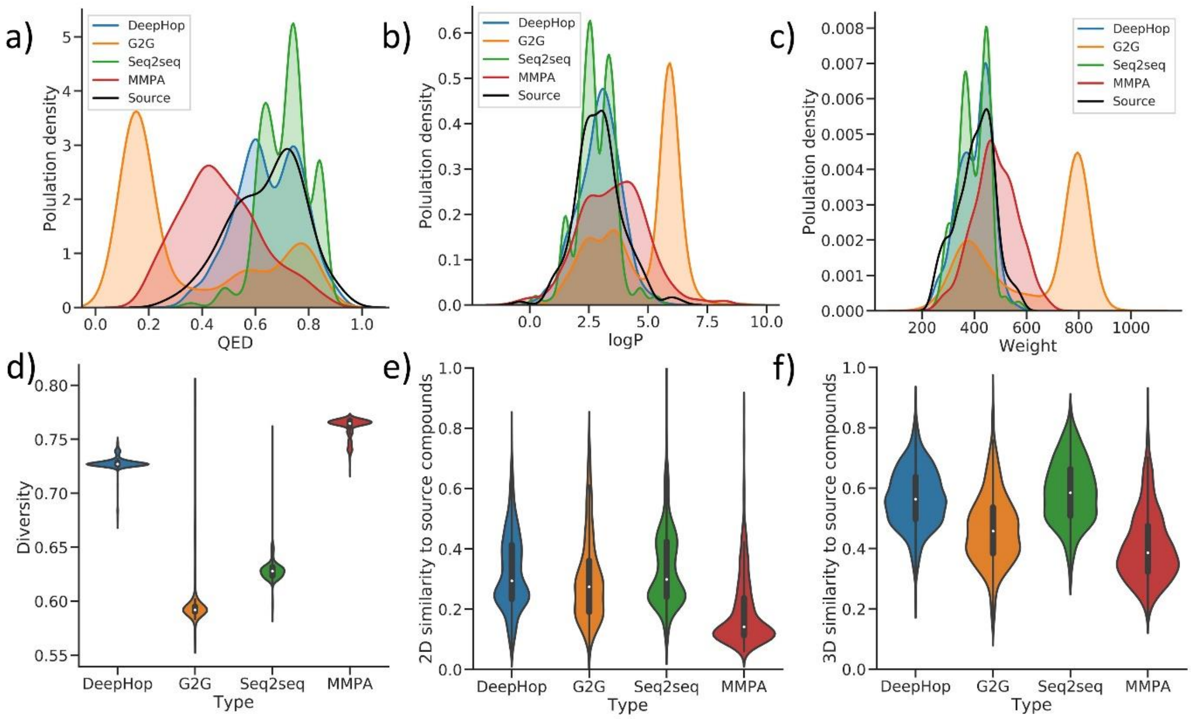

Figure 3. Distribution of chemical properties (a) QED, (b) logP, (c) Weight, (d) diversity, (e) 2D scaffold and (f) 3D structural similarity for source molecules and generated molecules from different models.

When comparing the diversity of different models, MMPA achieved the highest diversity score (Figure 3d), and DeepHop had slightly lower diversity. The seq2seq and G2G models kept much lower diversity, indicating that both models are prone to overfit the training set and get stuck in small chemical space. Besides, the compounds generated by G2G and MMPA are less similar than the ones generated by DeepHop, especially the 3D similarity. These results again proved that DeepHop overall outperformed state-of-the-art methods even in a single protein task and can generate high-quality scaffold hops for seed molecules efficiently.

Figure 4 shows two example of the top-predicted molecules generated by DeepHop. The modified groups lead to big changes in 2D while small changes in 3D. More cases are shown in the Supplementary File.

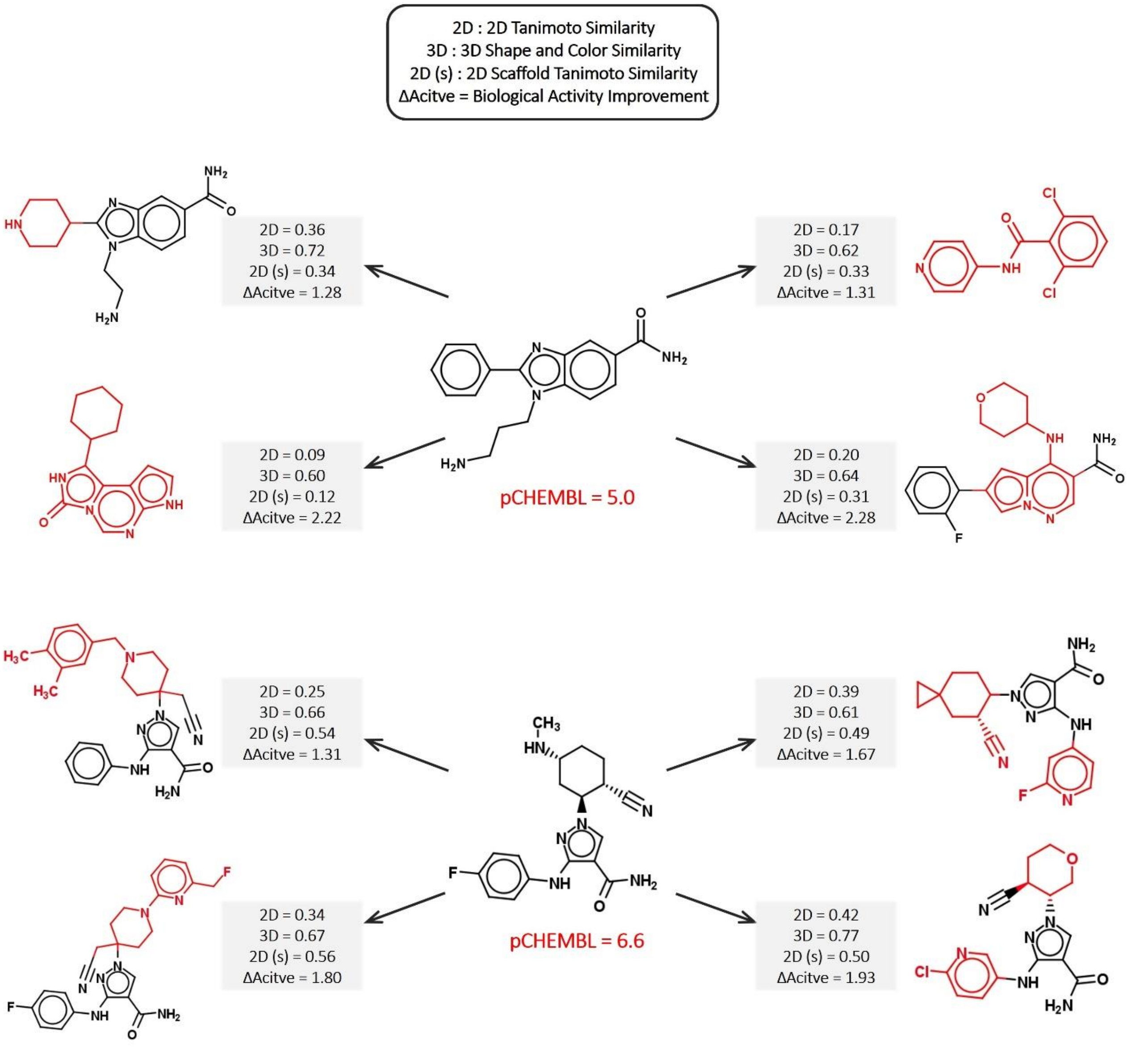

Figure 4. Example of top-4 successful hops ((2D scaffold similarity ≤ 0.6) ∩ (3D similarity (X; Y) ≥ 0.6) ∩ Activity Improvement ≥ 1) with two test molecules generated by DeepHop. The changes in the generated molecules compared with starting molecule are highlighted in red.

Model Performance on External Targets. We have shown that DeepHop achieves good performance in the internal test set. However, in real-world cases, scaffold hopping is often required for target proteins that have only a few known active compounds, and thus it is unable to construct sufficient scaffold hopping pairs for training. To mimic this scenario, we further examined whether DeepHop can be generalized to external targets that have never been observed in the training set. Following the same sampling strategy as above, we generated ten molecules for each parent molecule on six unseen targets.

Table 3. The independent tests of three heterogeneous proteins without homologs and three homogeneous proteins with homologs to the proteins in the training set.

Metrics	**ChEMBL Target**					
	Homogeneous			Heterogeneous		
	CHEMBL 4225	CHEMBL 2041	CHEMBL 2292	CHEMBL 2208	CHEMBL 4523	CHEMBL 2147
Success Rate	0.765	0.630	0.705	0.024	0.055	0.129
Constraint Success Rate	0.471	0.519	0.341	0.024	0.009	0.036
Improvement	0.515	1.259	0.824	-0.378	-1.210	-1.263

As shown in Table 3, the homogeneous targets performed very well in the external test set, even if all the molecular structures and protein sequences in these tasks have never been observed by the model. The results are expected as the deep learning models are often capable of generalizing to similar tasks while insufficient to perform tasks that are out-of-scope. It also suggests that when there are only a few known actives for a specific target protein that has over 60% sequence identity similarity to the training target proteins, DeepHop can be alternatively applied to generate scaffold hops directly without the need of re-training from scratch. We also noticed that the model achieved low success rates on three heterogeneous protein targets that are non-homologous to our training proteins (sequence

ID<25%).

For the heterogeneous target proteins, we wonder how many scaffold pairs are required to achieve a decent hopping. To this end, we equipped the model with the scheme of transfer learning and tested how well it can design inhibitors for the unfamiliar proteins. Specifically, the trained DeepHop were fine-tuned with 5%, 20%, 50%, 80% of scaffold hopping pairs from each unseen target protein, respectively.

As shown in Figure 5, by iteration with 100 epochs, only 5% (around 40~200) scaffold pairs can help unseen proteins to achieve fair success rates. At this point, the uniqueness of the generated molecules is poor because of the overfitting of limited data points. Thereafter, with the increase of scaffold hopping pairs, the model can gradually achieve a decent level of success rates and uniqueness. Note that the improvements are stable after fine-tuning 5% pairs, suggesting that the bioactivity feature is easy to capture compared to structural ones.

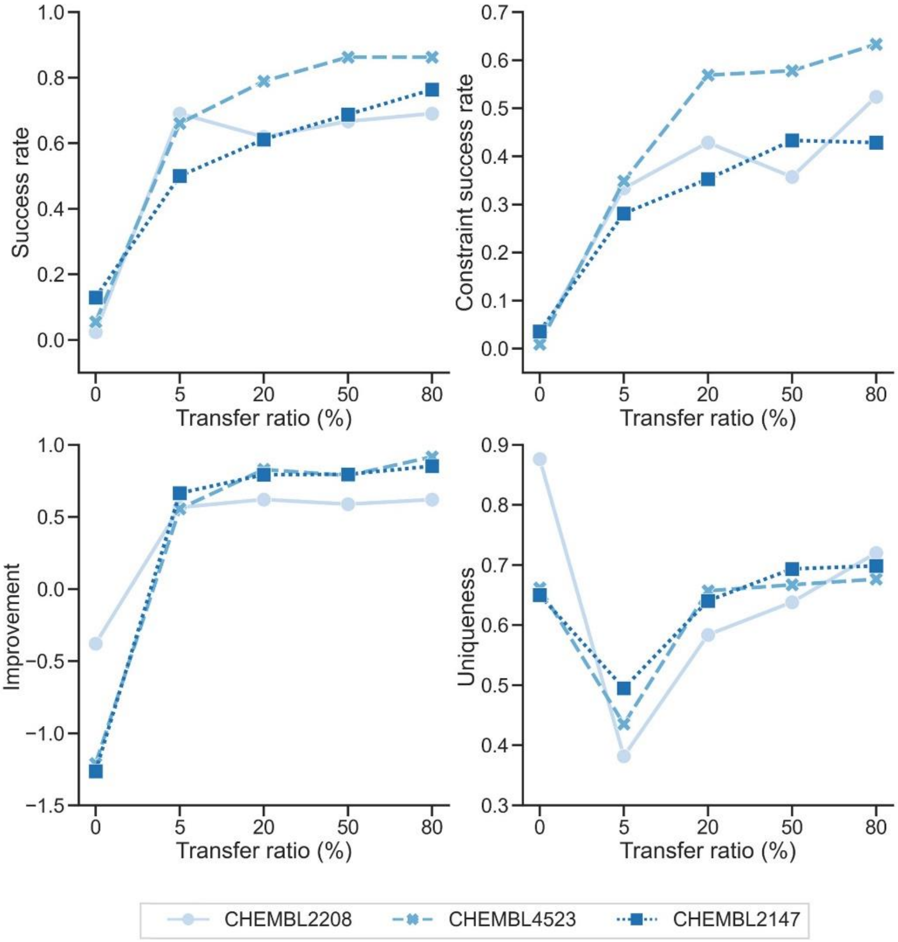

Figure 5. Transfer learning with different ratios of scaffold hopping pairs on the heterogeneous unseen protein targets.

Scaffold Hopping Case Study. Next, we chose PIM-1 kinase (CHEMBL2147), a well-studied target for antitumor drugs, as a representative to mimic a real-world scaffold hopping process. To search for novel inhibitors of the PIM-1 kinase, Saluste and co-workers once reported a typical fragment hopping by replacing imidazopyridazine scaffold with triazolopyridine, which maintained the primary activity and significantly improved off-target selectivity as well as ADME property[16].

We started with one lead inhibitor (seed 1, IC_{50} = 0.024nM) and two hit inhibitors (seed 2, IC_{50} =

155nM; seed 3, IC_{50} = 130nM), and aimed to generate potential scaffold hopping candidates with the improved pharmaceutical property. We used the trained model to generate 500 candidates for three seed compounds, respectively. All the generated candidates were then carried out with the docking process using AutoDock Vina[61].

Table 4. Scaffold hopping case study on PIM-1 kinase with three seed compounds.

Metrics	Scaffold Hopping		
	Seed 1	Seed 2	Seed 3
Unique structures	387	264	332
Structurally successful hops	51	66	40
Predicted activity < Lead	11	138	167
Docking score < Lead	102	184	202

As shown in Table 4, DeepHops can generate a large number of novel hops for each molecule by simply increasing the beam search width. The uniqueness values for seeds 1-3 are 77.4%, 52.8% and 66.4%, respectively. Among them, there are 51, 66, and 40 structurally successful hops generated for seed 1,2 and 3, meeting the requirements of (2D scaffold similarity $\leq$ 0.6) $\cap$ (3D similarity (X; Y) $\geq$ 0.6). In terms of bioactivity, we found that 26.4%, 69.7% and 60.8% of generated hops have a better docking scores than the seed compounds, demonstrating the effectiveness of our model. It is worth to note that even though the seed 1 has extremely high activity (IC_{50} = 0.024nM), there are 11 molecules to have better-predicted activities and 102 molecules that have better docking scores, suggesting that DeepHop could be a powerful tool in developing Me-too or Me-better molecules.

Several examples are shown in Figure 6. All scaffolds hops meet the condition of structure while obtaining a similar or improved activities compared to the starting seeds.

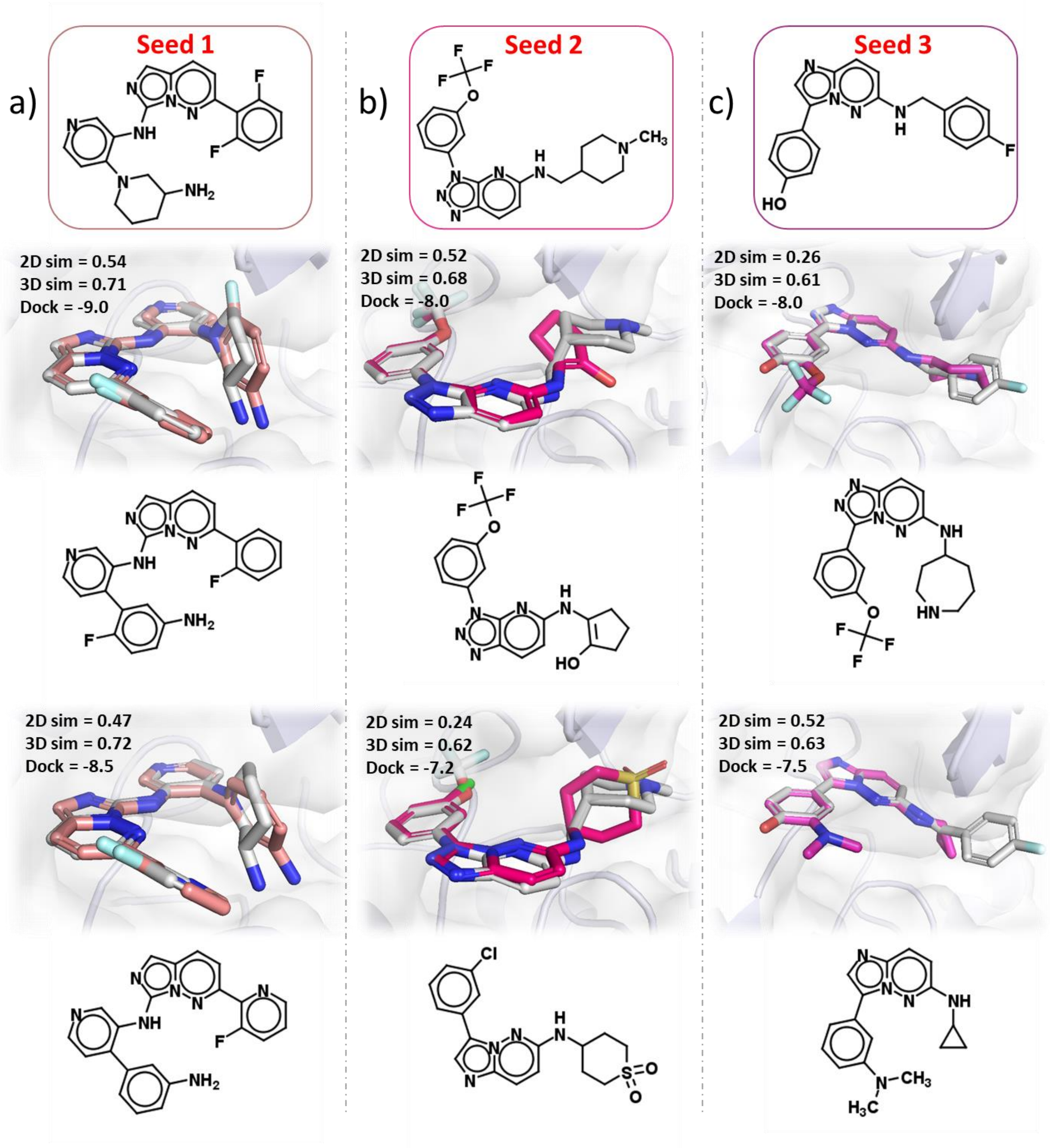

Figure 6. Overlay of the seed inhibitors (sliver) and top-predicted hops (colors). The 2D structures are shown below, and the structural similarity (2D Scaffold Tanimoto Similarity and 3D Shape and Color Similarity) and docking scores are attached in the upper left. Protein structure is retrieved from 5KZI[62].

Conclusions

In the current study, we have proposed a novel multi-modal deep generative model, DeepHop, for scaffold hopping, which is a critical task in rational drug design. The model can generate large sets of potential hops with novel backbones and improved bioactivities. This can be used in not only early drug discovery phases like hit-to-lead or lead optimization, but also patents busting for Me-Too and Me-Better molecules. Furthermore, we demonstrated that the model could generalize to new target proteins if fine-tuned with a small set of active compounds. This enables the generation of scaffold hops in low source scenarios. Through several case examples, we have shown that our method can be applied to practical scaffold hopping tasks, where most of the generated molecules have better docking scores than the original seeds while maintaining 3D similar but 2D dissimilar structure.

In the following work, we will seek to classify the types of scaffold hopping by deeply analyzing the molecular transformation paradigm. The hopping mode, like heterocycle replacement, ring-opening and ring closure, and topology transformation, should become a controlled condition. Additionally, a broad-spectrum target proteins dataset will be constructed and tested in order to enhance its scope of use. Furthermore, an interesting extension to the DeepHop models would be to use multi-objective reinforcement learning to allow our generated hops to match the comprehensive expectation (e.g., ADMET, synthesizability) of medicinal chemists.

The use of DeepHop in chemistry can only be validated experimentally, but we believe that the pipeline and model architecture described in our work constitutes an important early step toward solving the structure based rational drug design.

Corresponding Authors

*Email: yangyd25@mail.sysu.edu.cn (Y.Y).

*Email: deco@fulmz.com (D.D).

Authors' contributions

#S.Z. and Z.L. contributed equally to this work.

Authors' contributions

S.Z., Z.L., and H.A. contributed concept and implementation. S.Z., H. C. and Y.Y. wrote the manuscript. All authors contributed to the interpretation of results. All authors reviewed and approved the final manuscript.

Notes

Z.L, H.A and D.D declare a financial interest in that they are employees of and shareholders in Fermion Technology Co., Ltd.

Reference

1. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S., Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011,** *10* (3), 188-95.
2. Ecker, D. J.; Crooke, S. T., Combinatorial drug discovery: which methods will produce the greatest value? *Biotechnology (N Y)* **1995,** *13* (4), 351-60.
3. Fattori, D.; Squarcia, A.; Bartoli, S., Fragment-based approach to drug lead discovery: overview and advances in various techniques. *Drugs R D* **2008,** *9* (4), 217-27.
4. Hu, Y.; Stumpfe, D.; Bajorath, J., Recent Advances in Scaffold Hopping. *J Med Chem* **2017,** *60* (4), 1238-1246.
5. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G., "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew Chem Int Ed Engl* **1999,** *38* (19), 2894-2896.
6. Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y., Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence* **2020,** *2* (2), 134-140.
7. Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **2005,** *48* (5), 1489-95.
8. Hu, Y.; Stumpfe, D.; Bajorath, J., Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J Med Chem* **2016,** *59* (9), 4062-76.
9. Sun, H.; Tawa, G.; Wallqvist, A., Classification of scaffold-hopping approaches. *Drug Discov Today* **2012,** *17* (7-8), 310-24.
10. Nakano, H.; Miyao, T.; Funatsu, K., Exploring Topological Pharmacophore Graphs for Scaffold Hopping. *J Chem Inf Model* **2020,** *60* (4), 2073-2081.
11. Laufkotter, O.; Sturm, N.; Bajorath, J.; Chen, H.; Engkvist, O., Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. *J Cheminform* **2019,** *11* (1), 54.
12. Renner, S.; Schneider, G., Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **2006,** *1* (2), 181-5.
13. Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G., Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci Rep* **2018,** *8* (1), 16469.
14. Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G., Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol Inform* **2013,** *32* (2), 133-138.
15. Floresta, G.; Amata, E.; Dichiara, M.; Marrazzo, A.; Salerno, L.; Romeo, G.; Prezzavento, O.; Pittala, V.; Rescifina, A., Identification of Potentially Potent Heme Oxygenase 1 Inhibitors through 3D-QSAR Coupled to Scaffold-Hopping Analysis. *ChemMedChem* **2018,** *13* (13), 1336-1342.
16. Saluste, G.; Albarran, M. I.; Alvarez, R. M.; Rabal, O.; Ortega, M. A.; Blanco, C.; Kurz, G.; Salgado, A.; Pevarello, P.; Bischoff, J. R.; Pastor, J.; Oyarzabal, J., Fragment-hopping-based discovery of a novel chemical series of proto-oncogene PIM-1 kinase inhibitors. *PLoS One* **2012,** *7* (10), e45964.
17. Stahura, F. L.; Xue, L.; Godden, J. W.; Bajorath, J., Molecular scaffold-based design and comparison of combinatorial libraries focused on the ATP-binding site of protein kinases. *J Mol Graph Model* **1999,** *17* (1), 1-9, 51-2.
18. Vainio, M. J.; Kogej, T.; Raubacher, F.; Sadowski, J., Scaffold hopping by fragment replacement. *J Chem Inf Model* **2013,** *53* (7), 1825-35.

19. Rabal, O.; Amr, F. I.; Oyarzabal, J., Novel Scaffold FingerPrint (SFP): applications in scaffold hopping and scaffold-based selection of diverse compounds. *J Chem Inf Model* **2015,** *55* (1), 1-18.

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

21. 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.

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

23. 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.

24. Zheng, S.; Yan, X.; Gu, Q.; Yang, Y.; Du, Y.; Lu, Y.; Xu, J., QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J Cheminform* **2019,** *11* (1), 5.

25. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **2018,** *4* (2), 268-276.

26. Skalic, M.; Jimenez, J.; Sabbadin, D.; De Fabritiis, G., Shape-Based Generative Modeling for de Novo Drug Design. *J Chem Inf Model* **2019,** *59* (3), 1205-1214.

27. Stahl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Bostrom, J., Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. *J Chem Inf Model* **2019,** *59* (7), 3166-3176.

28. De Cao, N.; Kipf, T., MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973* **2018**.

29. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988,** *28* (1), 31-36.

30. Lim, J.; Hwang, S.-Y.; Moon, S.; Kim, S.; Kim, W. Y., Scaffold-based molecular design with a graph generative model. *Chemical Science* **2020,** *11* (4), 1153-1164.

31. Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z., DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *J Chem Inf Model* **2020,** *60* (1), 77-91.

32. Arús-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O., SMILES-based deep generative scaffold decorator for de-novo drug design. *Journal of Cheminformatics* **2020,** *12* (1), 1-18.

33. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M., Deep Generative Models for 3D Linker Design. *J Chem Inf Model* **2020,** *60* (4), 1983-1995.

34. Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H., SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical Science* **2020,** *11* (31), 8312-8322.

35. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012,** *40* (Database issue), D1100-7.

36. Southall, N. T.; Ajay, Kinase patent space visualization using chemical replacements. *J Med Chem* **2006,** *49* (6), 2103-9.

37. Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. In *Communicative Representation Learning on Attributed Molecular Graphs*, IJCAI: 2020.

38. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **2019,** *59* (8), 3370-3388.

39. Li, X.; Li, Z.; Wu, X.; Xiong, Z.; Yang, T.; Fu, Z.; Liu, X.; Tan, X.; Zhong, F.; Wan, X.; Wang, D.; Ding, X.; Yang, R.; Hou, H.; Li, C.; Liu, H.; Chen, K.; Jiang, H.; Zheng, M., Deep Learning Enhancing Kinome-Wide Polypharmacology Profiling: Model Construction and Experiment Validation. *J Med Chem* **2020,** *63* (16), 8723-8737.

40. Hussain, J.; Rea, C., Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of chemical information and modeling* **2010,** *50* (3), 339-348.

41. Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010,** *50* (5), 742-754.

42. Landrum, G. A.; Penzotti, J. E.; Putta, S., Feature-map vectors: a new class of informative descriptors for computational drug discovery. *Journal of computer-aided molecular design* **2006,** *20* (12), 751-762.

43. Putta, S.; Landrum, G. A.; Penzotti, J. E., Conformation mining: an algorithm for finding biologically relevant conformations. *Journal of medicinal chemistry* **2005,** *48* (9), 3313-3318.

44. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006,** *22* (13), 1658-1659.

45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. In *Attention is all you need*, Advances in neural information processing systems, 2017; pp 5998-6008.

46. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D. F.; Chao, L. S., Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787* **2019**.

47. Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y., Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J Chem Inf Model* **2020,** *60* (1), 47-55.

48. Danel, T.; Spurek, P.; Tabor, J.; Śmieja, M.; Struski, Ł.; Słowik, A.; Maziarka, Ł., Spatial Graph Convolutional Networks. *arXiv preprint arXiv:1909.05310* **2019**.

49. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y., Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* **2014**.

50. Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; Song, Y. In *Evaluating protein transfer learning with TAPE*, Advances in Neural Information Processing Systems, 2019; pp 9689-9701.

51. Nair, V.; Hinton, G. E., In *ICML*, 2010.

52. Ba, J.; Kiros, J. R.; Hinton, G. E., Layer Normalization. *ArXiv* **2016,** *abs/1607.06450*.

53. Barrault, L.; Bojar, O. e.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; Monz, C.; Müller, M.; Pal, S.; Post, M.; Zampieri, M. In *Findings of the 2019 Conference on Machine Translation (WMT19)*, Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Florence, Italy, aug; Association for Computational Linguistics: Florence, Italy, 2019; pp 1-61.

54. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. *CoRR* **2015,** *abs/1512.03385*.

55. Hussain, J.; Rea, C., Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model* **2010,** *50* (3), 339-48.

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

57. Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T., Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070* **2018**.

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

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

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

61. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010,** *31* (2), 455-461.

62. Wurz, R. P.; Sastri, C.; D'Amico, D. C.; Herberich, B.; Jackson, C. L. M.; Pettus, L. H.; Tasker, A. S.; Wu, B.; Guerrero, N.; Lipford, J. R.; Winston, J. T.; Yang, Y.; Wang, P.; Nguyen, Y.; Andrews, K. L.; Huang, X.; Lee, M. R.; Mohr, C.; Zhang, J. D.; Reid, D. L.; Xu, Y.; Zhou, Y.; Wang, H. L., Discovery of imidazopyridazines as potent Pim-1/2 kinase inhibitors. *Bioorg Med Chem Lett* **2016,** *26* (22), 5580-5590.