Interface-aware molecular generative framework for protein-protein interaction modulators

Jianmin Wang[a], Jiashun Mao[a], Chunyan Li[d], Hongxin Xiang[b], Xun Wang[c,e], Shuang Wang[c], Zixu Wang[f], Yangyang Chen[f], Yuquan Li[g], Heqi Sun[h], Kyoung Tai No*[a], Tao Song *[c], Xiangxiang Zeng*[b]

[a] The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea

[b] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China

[c] School of Computer Science and Technology, China University of Petroleum, Qingdao, 266580, Shandong, China

[d] School of Informatics, Yunnan Normal University, Kunming, China

[e] High Performance Computer Research Center, University of Chinese Academy of Sciences, Beijing, 100190, China

[f] Department of Computer Science, University of Tsukuba, Tsukuba, 3058577, Japan

[g] College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China

[h] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

ABSTRACT

Protein-protein interactions (PPIs) play a crucial role in many biochemical processes and biological processes. Recently, many structure-based molecular generative models have been proposed. However, PPI sites and compounds targeting PPIs have distinguished physicochemical properties compared to traditional binding pockets and drugs, it is still a challenging task to generate compounds targeting PPIs by considering PPI complexes or interface hotspot residues. In this work, we propose a specifically molecular generative framework based on PPI interfaces, named GENiPPI. We evaluated the framework and found it can capture the implicit relationship between the PPI interface and the active molecules, and can generate novel compounds that target the PPI interface. Furthermore, the framework can generate diverse novel compounds with limited PPI interface modulators. According to our knowledge, this is the first exploration of PPI structure-based molecular generative modeling, which can help accelerate the design of PPI structure-based regulators. The PPI interface-based molecular generative model enriches structure(pocket and interface)-based molecular generative.

INTRODUCTION

A vast network of genes is inter-linked through protein-protein interactions and is critical component of almost every biological process under physiological conditions, and can be ubiquitous in many living organisms and biological pathways[1-4]. Modulation of PPIs expands the drug target space and has enormous potential in drug discovery. In homo sapiens, it is estimated that the entire interactome comprises between 130,000 to 930,000 binary PPIs[5-7]. Despite significant efforts in developing modulators of PPIs, drug design and development for PPI targets, especially targeting the PPI interfaces, remains challenging[6, 8-12]. Structure-based rational design serves as an important tool for the discovery of lead compounds in drug discovery[13-17]. Traditional drug targets and PPIs targets have different bio-chemical features (**Table 1**)[11, 18-22], so conventional drugs and PPIs modulators have different physicochemical properties and drug-like properties (**Table 1**)[11, 23-30]. Given their differences, developing molecular generative models of different paradigms are essential for the drug design of different target types[11, 19, 31].

Table 1. Comparisons between PPI interfaces and binding sites

PPI interfaces	**Binding sites**
Target properties	
Large surface area (1000-6000 Å^2)	Small surface (300-1000 Å^2) Hydrophobic
Preference for Trp (W), Tyr (Y), and Arg (R) as PPI hotspot residues; subpockets	Large volume (~260 Å^3)
Shallow, flat, flexible	Pocket, cliff
Hydrophobic, featureless, undruggability	Diverse properties
Chemical space	
MW≥ 400	MW≤ 500
LogP ≥ 4	LogP≤ 500
HBA ≥ 4	HBA ≤ 10
number of rings: ≥ 4	HBD ≤ 5
Ro4 Morelli's rules	Lipinski's Rule of 5 (Ro5)
Quantitative estimate of drug-likeness scores	
QEPPI	QED

Generative artificial intelligence(AI) is enable to model the distribution of training samples and generate novel samples[32, 33]. In drug discovery, generative AI can accelerate drug discovery by generating novel molecules with desired properties. Numerous excellent review articles have summarized the development in this field[16, 17, 34-41]. Molecular generative models in drug design can be roughly divided into three categories: ligand-based molecular generative (LBMG) models, structure-based(pockets or binding sites) molecular generative models (SBMG), and fragment-based molecular generative models (FBMG), among which SBMG models have received much attention. [17, 39, 42]. Currently, some significant methods in structure-based molecular generative models can be found in [43-51], molecular generative models for PPI structures or PPI interfaces have been rarely reported in the literature. In recent years, classical machine learning[52-54], active learning[55], and deep learning-assisted methods [56] is better screening and design of PPIs modulators have been explored, and ligand-based molecular generative models of PPI modulators have been reported[57]. There are few structure-based molecular generative models for PPIs targets have not been sufficiently explored.

In this study, we developed a conditional molecular generative framework based on protein-protein interaction interfaces (named GENiPPI) for the design of PPI interface modulators. The framework was developed by a conditional Wasserstein generative adversarial network (cWGAN) with convolutional neural networks (CNNs), integrated graph attention networks (GATs) and long short-term memory (LSTM). It was designed to efficiently capture the relationship between PPI interface with active/inactive compounds to train conditional molecular generative models (**Fig. 1**). As demonstrated by the conditional evaluation, GENiPPI is an effective architecture for capturing the implicit relationships between the PPI interface and active compounds. In summary, GENiPPI represents a potent deep learning framework for structure-based design of PPI modulators.

RESULTS AND DISCUSSION

Generation of molecules targeting the PPI interface

Here, we introduce GENiPPI a modular deep learning framework for the design of structure-based PPIs modulators (**Fig. 1**). GENiPPI is composed of four main modules: GATs module [58-60] for representation learning of the protein complex interface, CNNs module for molecular representation learning, cWGAN module [61] for conditional molecular generation, and molecular captioning network module for SMILES strings decoding (**as shown in Supplementary Figs.1, Figs.2, Figs.3 and Figs.4, respectively**).

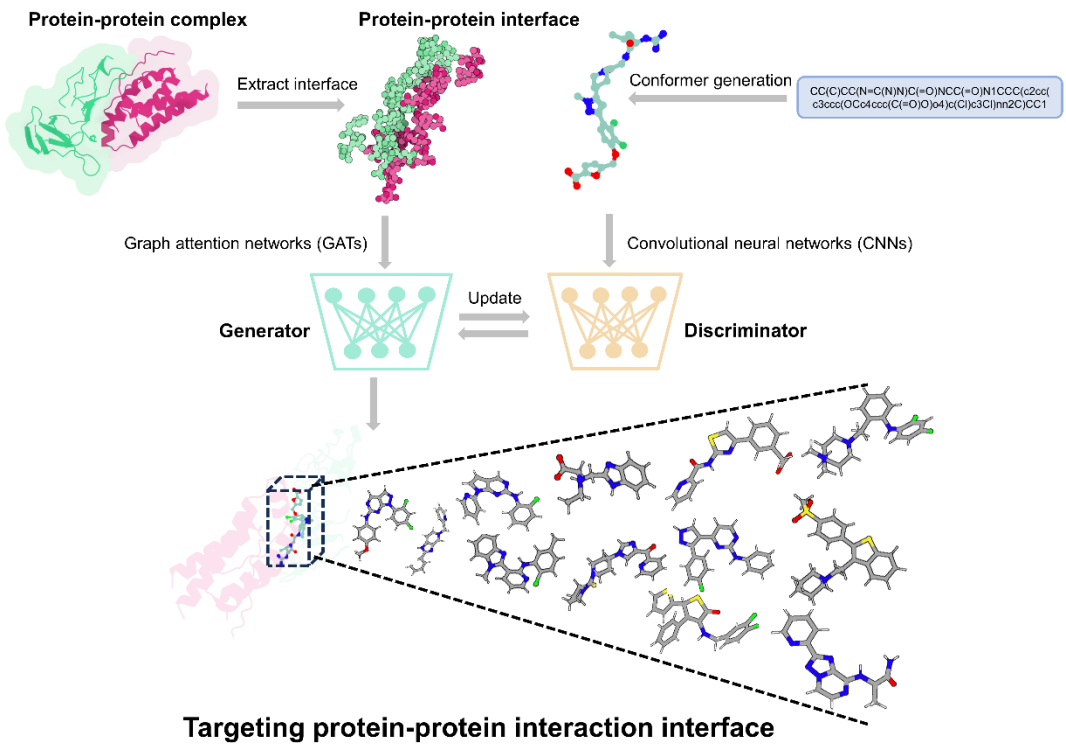

Figure 1. Generation of molecules targeting PPI. 3D structural information of the protein-protein complex interface is represented as a graph. Feature representation of the interface region is captured by using a graph attention neural networks. The representation of the voxel and electron density of the compound is encoded by a 3D convolutional neural networks (CNNs). A conditional Wasserstein generative adversarial networks is trained to generate molecular embeddings with interface features as conditions. Generator: takes interface features and random noise vectors to generate molecular embeddings for the input features. Discriminator: calculates the probability that a molecule is from a real

or a fake molecule. Condition: controls or regulates the generation of molecules constrained by a specific protein-protein interface. Finally, a long short-term memory (LSTM) networks parse SMILES strings from molecular representation.

Our framework undergoes four steps to accomplish the generation of molecules targeting the PPI interface. In the first step, we use GATs module designed for the protein complex interface is to effectively capture the nuanced atomic-level interaction characteristics inherent to the protein complex interface region. Next, we use CNN module to provide a representation of the compound that contains voxel and electronic density information in three-dimension space [62]. And, the cGAN module is designed to generate compounds that target PPI interfaces using features from the protein complex interface region to regulate the inputs [63]. The cGAN module consists of a generator, a discriminator, and a conditional network. The generator takes a Gaussian random noise vector, and the protein complex interface features to generate a vector in the molecular embedding space, the discriminator evaluates whether the generated molecule embedding corresponds to a real or generated molecule, and the conditional network evaluates whether the molecule embedding matches the protein complex interface features. Finally, we use the molecular captioning network, which is made by a 3D CNNs and a recurrent LSTM [64]to decode molecular representations. The molecular representation generated by the generator is fed as input to the 3D convolutional network with the LSTM subsequently decoding the SMILES strings.

Conditional evaluation

First, we verified the validity of the conditions that act as conditional molecular generative models for the protein complex interfaces. For this purpose, we selected three PPI targets: MDM2(mouse double minute 2)/p53, Bcl-2(B-cell lymphoma 2)/Bax (Bcl-2 associated X), and BAZ2B(Bromodomain adjacent to zinc finger domain protein 2B)/H4(histone) for conditional evaluation. We generated 10,000 validated molecules

each by the GENiPPI framework and calculated the drug-like metrics of the generated compounds: QED[27], QEPPI[28, 29] and Fsp3(fraction of sp3 carbon atoms)[65].

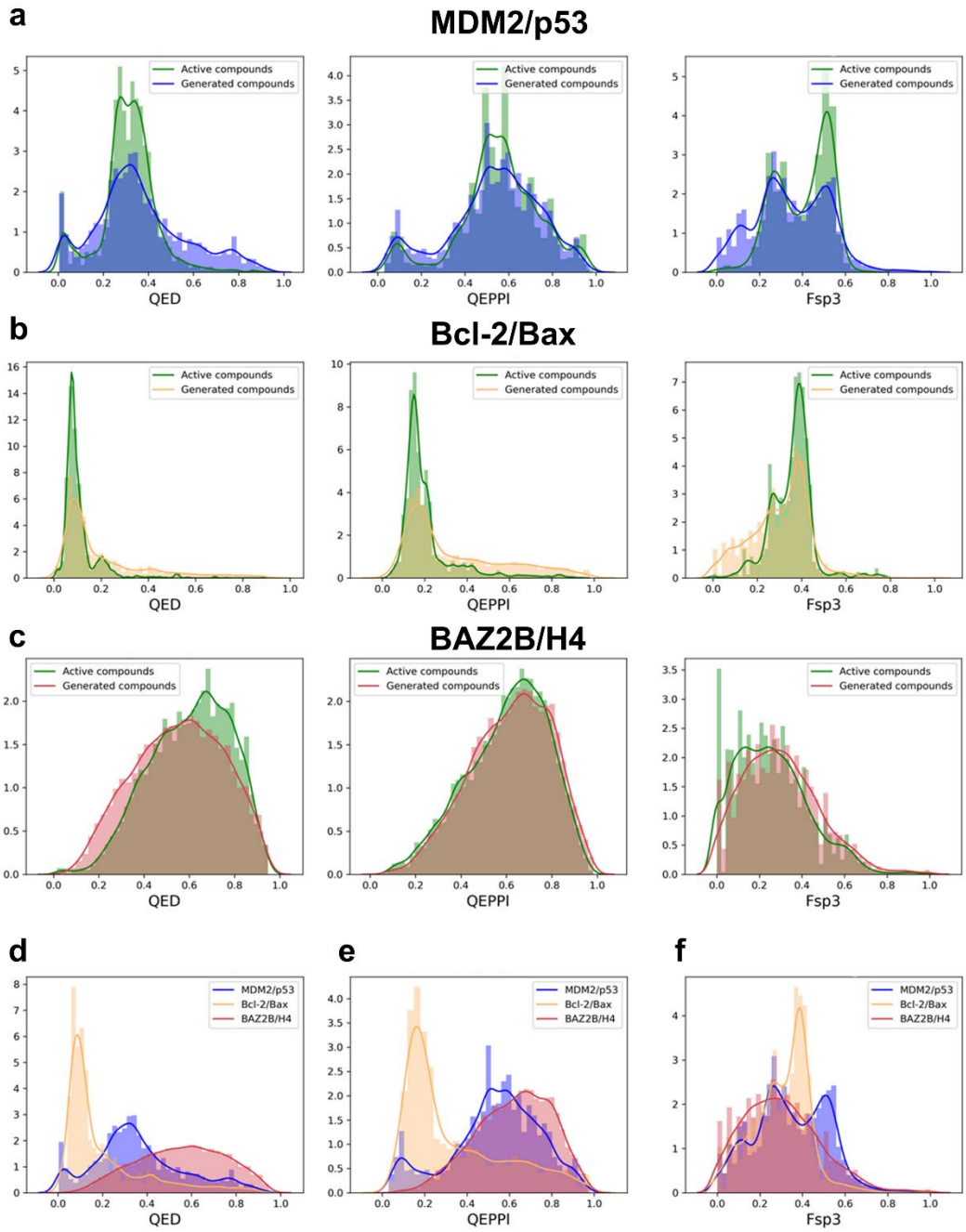

Figure 2. Results of conditional evaluation. (a) The QED, QEPPI and Fsp3 distribution of active compounds and compounds generated by the GENiPPI framework for MDM2/p53; (b) The QED, QEPPI and Fsp3 distribution of active compounds and compounds generated by the GENiPPI framework for Bcl-2/Bax; (c) The QED, QEPPI and Fsp3 distribution of active compounds and compounds generated by the GENiPPI framework for BAZ2B/H4; (d) The QED distribution of generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4; (e) The QEPPI distribution of generated compounds for

MDM2/p53, Bcl-2/Bax and BAZ2B/H4; (f) The Fsp3 distribution of generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4.

We compared the QED, QEPPI, and Fsp3 distributions of the active compounds and the generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4 **(Fig. 2)**. As shown, the distributions of drug-like properties were similar between the generated compounds and the active compounds for the three PPI interface targets (**Fig.2a, Fig.2b, and Fig.2c**), while different distributions of drug-like properties were observed between the generated compounds based on different targets (**Fig.2d, Fig.2e, and Fig.2f**). The results demonstrate the effectiveness of the PPI interface in conditioning the molecular generative model. The drug-like properties of the framework generated compounds migrate relative to those of the compounds in the training dataset, indicating that the framework captures the distributions of the training dataset and generates novel compounds.

Model performance

Table 2. Valid, unique, novelty and FCD of sampling SMILES after training. We sampled 30,000 SMILES each time.

Model	valid	Unique@1k	Unique@10k	novelty	FCD	
					Test	TestSF
AAE	0.881	1.000	0.995	0.995	8.573	9.117
CharRNN	0.985	0.999	0.988	0.994	8.7564	8.952
VAE	0.834	1.000	0.996	0.994	7.703	8.141
LatentGAN	0.724	1.000	0.999	0.998	7.595	8.160
ORGAN	0.609	0.996	0.994	0.999	39.800	41.158
GENiPPI(noninterface)	0.999	0.997	0.975	0.997	7.653	8.132
GENiPPI	0.999	0.998	0.977	0.998	7.450	7.884

In order to gain insight into the performance of the GENiPPI framework and to compare it with other molecular generative models. We benchmarked our method by the MOSES platform[66], a leading benchmark platform of molecular generation. We trained all

models on the full training dataset and randomly sampled 30,000 molecules. We utilized models and hyperparameters provided by the MOSES platform, such as an Adversarial Autoencoder(AAE)[67], character-level recurrent neural networks (CharRNN)[68], Variational Autoencoder(VAE)[69], LatentGAN[70] and ORGAN[71]. To validate the higher quality of the molecules generated by the conditioned model, we compared them with molecules sampled from the GENiPPI framework and the GENiPPI-noninterface framework without the conditioned module. We found that molecules generated by the conditioned GENiPPI framework were superior to other models in novelty and diversity.

As shown in **Table 2**, the GENiPPI framework has advantages in terms of uniqueness, novelty, and diversity over the GENiPPI-noninterface. The GENiPPI framework performs better overall in molecular generation. Compared with LatentGAN and ORGAN, GENiPPI offers more benefits in terms of validity and diversity. While all molecular generative models have their unique advantages in various performance comparison. However, the molecular generative models tailored to specific tasks, especially those based on PPI structure, have more advantages and inspirations from the GENiPPI framework. To understand the similarities and differences between the molecular distributions generated by the GENiPPI framework and other models. We compared the distribution of molecular properties of the Testset, iPPI-DB inhibitor, and the generated molecular datasets of AAE, CharRNN, VAE, LatentGAN, GENiPPI(noninterface) and GENiPPI(**Supplementary Figs.4**). The generated compounds have similar distributions of physicochemical properties to the compounds from the training set. While most of the iPPI-DB inhibitors have QED values lower than 0.5, most of them have QEPPI values higher than 0.5.

Chemical space exploration

To better obtain an estimate of the chemical space distribution of the model generated molecules with the active compounds in the training datasets, we evaluated the

chemical drug-like space of the generated compounds by calculating t-distributed random neighbourhood embedding (t-SNE) maps of MACCS fingerprint [72]. The t-SNE is a dimensionality reduction method used for data points visualization in two or three-dimensional space by mapping high-dimensional data to a lower dimension [73, 74]. By this method, similar compounds are clustered to visualize the high-dimensional chemical space of the compounds.

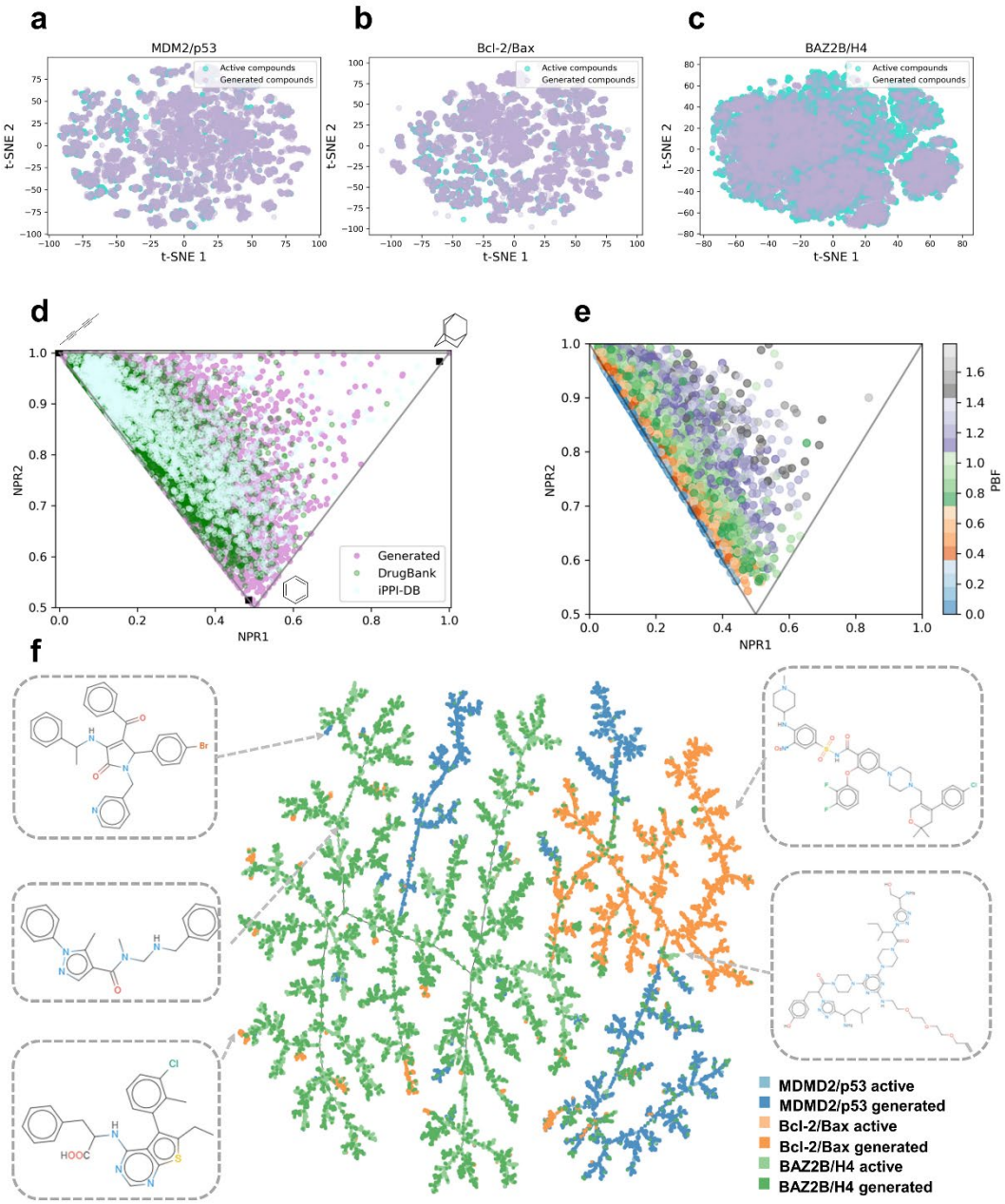

Figure 3. Chemical space exploration. (a) The t-SNE visualization of active compounds and generated compounds for MDM2/p53; (b) The t-SNE visualization of active compounds and generated compounds

for Bcl-2/Bax; (c) The t-SNE visualization of active compounds and generated compounds for BAZ2B/H4; (d) The PMI ternary density plots of generated compounds, small molecule drugs of DrugBank, and iPPI-DB inhibitors. Top left: propyne, bottom: benzene, and the top right: adamantane; (e) The molecular three-dimensionality distribution of the generated molecules was visualized with NPR descriptors and PBF descriptors. (f) TMAP visualization of active compounds and generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4.

The distribution of the generated compounds and active compounds in chemical drug-like space by t-SNE visualization (**Fig.3a, Fig.3b, and Fig.3c**). The generated drug-like compounds not only share the chemical space with the active compounds, but are also homogeneously mixed in the two-dimensional space. The generated compounds show a similar chemical drug-like space to that of the active compounds under 2D topological fingerprint. Adding the three dimensions of compounds contributes to the design of promising drug-like compounds[30, 75]. We performed PMI shape analysis on the generated compounds and compared them with drug-like compounds from DrugBank and iPPI-DB(**Fig.3d**). Many of the approved compounds are either rod or disk shaped, and the generated drug-like compounds library has a similar three-dimensional space. The PBF distribution of the library of generated drug-like compounds is about 0~2 Å(**Fig.3e**). The results show that many of the generated drug-like compounds are derived from relatively planar molecular scaffolds. Moreover, we evaluated the ability of the model to generate target-specific compounds by chemical space maps. To assess the overlapping of drug-like chemical space, we utilized Tree MAP (TMAP)[76] to create the 2D projection(**Fig.3f**). Each point corresponds to a compound and is colored by its target label. The dark and light colors denote the generated compounds and the active compounds in the training set. These results suggest that our GENiPPI model can generate compounds that are similar to the active compounds in the training set and have novel structures. The results show that the framework enriches and expands the chemical space of PPI-targeted drug-likeness compounds.

Few-shot molecular generation

Because of the huge consumptive costs involved in data collection, only a small amount of labeled biomedical data are usually available. The process of drug design and optimization often faces the problem of low data[77]. The lack of effectively labeled data tends to diminish the practical performance of most deep learning frameworks for drug design. To perform generalized molecular generative design with limited labeled data, it has been a trending topic in the few-shot generative community[78, 79]. The GENiPPI model was applied to generate a virtual compound library for the heat shock protein 90 - cell division cycle 37(Hsp90-Cdc37) interaction interface. By training the model on the PPI structure of Hsp90-Cdc37 (PDB ID: 1US7) and seven disruptors, we sampled 500 valid compounds. The similarity between active disruptors and generated compounds of Hsp90/Cdc37 in the chemical space was visualized by t-SNE projection maps(**Fig.4a**). After few-shot learning, the generated compounds were mostly distributed around the active disruptor, which demonstrated the effectiveness of few-shot learning in navigating through the targeted chemical space. We performed pharmacophore-based matching by considering DCZ3112(a novel triazine derivative that disrupts Hsp90-Cdc37 interactions) as a reference molecule[80]. The top 5 generated molecules have similar pharmacophore and shape features with DCZ3112(**Fig.4b**), demonstrating the potential of the model to be applied to low-data PPI targets. **Fig. 4c**, shows the hot spot amino acid residues at the PPI interface of the Hsp90-Cdc37 protein complex(PDB ID: 1US7). We performed molecular docking for prediction of the binding poses(**Fig.4e**) of DCZ3112 with the Hsp90-Cdc37 complex by the UCSF DOCK6.9 program[81]. The structure of the Hsp90-Cdc37 complex with DCZ3112 highlights the hydrogen bond interactions with amino acid residues:Arg32A, Glu33A, Ser36A , Ser115A, Gly118A, Gln119A, and Arg167B(**Fig.4e**), which may be the major energy contributors to protein-ligand interactions. The generated compounds were performed molecular docking together with DCZ3112, and selected compounds with reasonable binding modes and higher binding affinity by visual inspection for

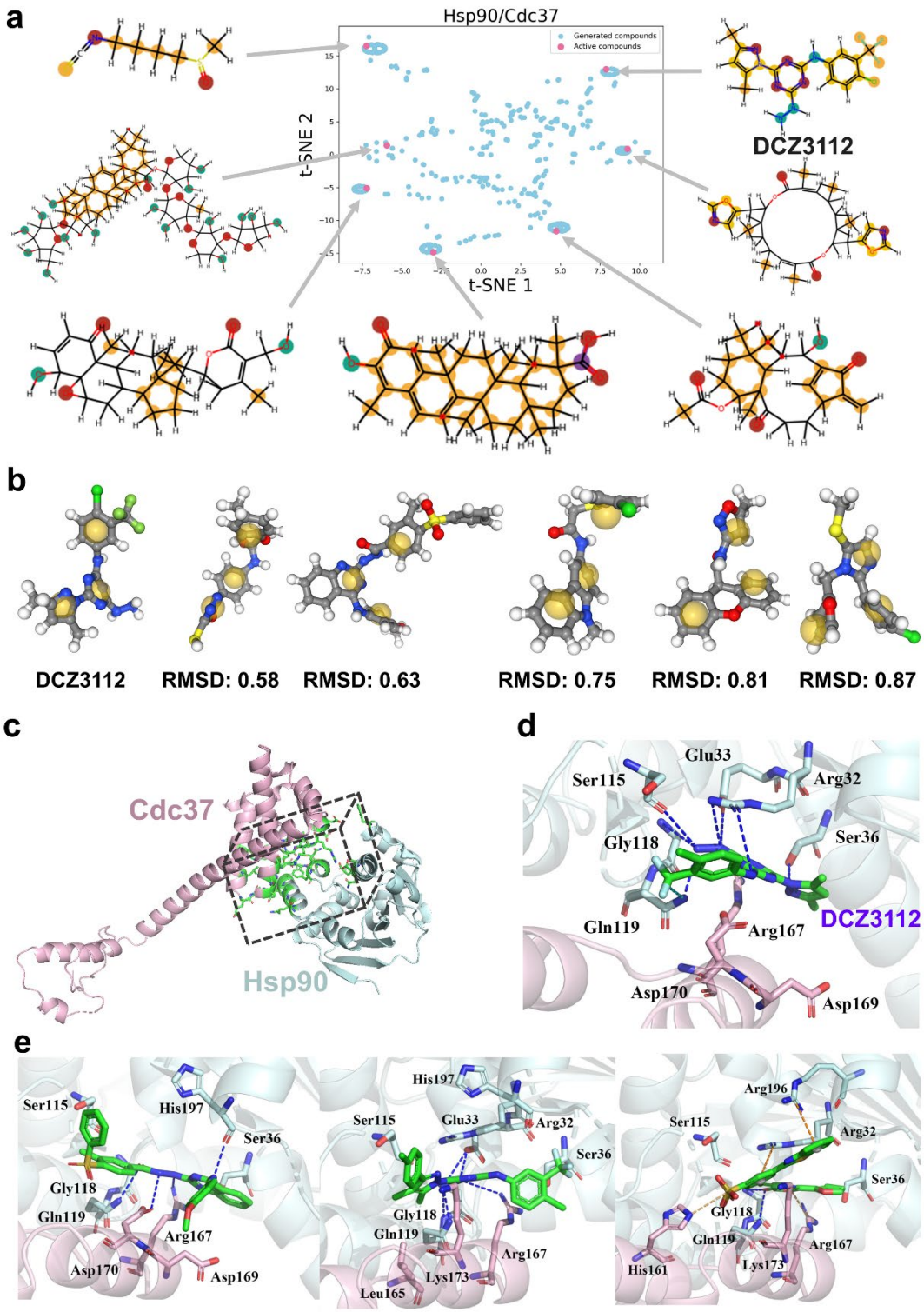

Figure 4. Few shot molecular generation analysis. (a) The t-SNE visualization of the distribution of active compounds and generated compounds for Hsp90/Cdc37; (b) Comparison of the pharmacophore of the generated molecules with the reference molecule(DCZ3112); (c) PPI interface region(in green) of the Hsp90(in palecyan)/Cdc37(in lightpink) complex; (d) The complex structure of DCZ3112(in green) and Hsp90(in palecyan)-CDC37(in lightpink) modeled by molecular docking (PDB ID: 1US7); (e) The

binding poses of generated compounds(in green) and Hsp90(in palecyan)-CDC37(in lightpink) modeled by molecular docking (PDB ID: 1US7). Hydrogen bonds are displayed as blue dotted lines. π-cation Interactions are displayed as orange dotted lines.

interaction pattern analysis. The generated compounds of GENiPPI not only obtained the better docking score than the active compounds, but also reproduced the interactions with the key residues of the PPI interface. The generated compounds also formed halogen bonds, salt bridges and π-cation interactions to improve the binding affinity of the generated compounds to the target interface(**Fig.4f**). In conclusion, by analyzing the interaction patterns between the generated compounds and the PPI interface, GENiPPI learned the implicit interaction rules between the active compounds and the PPI interface.

CONCLUSIONS

We developed the GENiPPI framework, which combines protein-protein interaction (PPI) interfaces features and conditional molecular generative model to generate novel modulators for PPI interfaces. We validated the ability of GENiPPI framework to learn the implicit relationship between PPI interfaces and active molecules through conditional evaluation experiments. GENiPPI used GATs to extract key features of PPI interfaces, and searched molecules by conditional wGAN with specific constraints. We compared GENiPPI with various evaluation settings and benchmarks to demonstrate its practical potential.

Despite the promising results, our framework has some limitations that can be addressed in future work. We have not tested the model on a large number of receptor-ligand pairs of PPIs, which may affect its the generalization ability. The reason is that the current PPI has relatively little data of drug-PPI target complexes than the traditional dataset of drug-target complexes. Furthermore, the current framework does not

incorporate the 3D structural information of ligand-receptor interactions of PPIs. and There are still many ways to improve representation learning, balance training speed of molecular generative models and the diversity of generated molecules. Several potential directions could further improve GENiPPI: (1) collecting and cleaning higher quality data pairs for model development and testing; (2) fusing of molecular chemical language models and pre-trained models of protein-protein structural features to fine-tune the datasets of receptor-ligand of PPIs to enhance the model generalization, novelty and diversity of the generated compounds; (3) incorporating structural information of PPIs into fragment-based molecular generative models is also a promising direction; (4) change the architecture of the model or combining it with deep reinforcement learning to generate novel compounds with better binding affinity[82]. Therefore, we will collect more data to further develop an enhanced version of GENiPPI by combining novel representation learning methods and deep generative approaches. In summary, the GENiPPI framework brings encouraging advances in PPI structure-based molecular generative tasks and presents a tool for rational drug design in finding modulators of macromolecule-macromolecule interactions.

METHODS

Datasets

We first investigated PPI targets that were annotated with sufficient compound bioactivity data for training and evaluation of our model [83]. For this study, we selected 10 validated PPI drug targets that cover the binding interface(**Supplementary Table**). These targets are E3 ubiquitin-protein ligase Mdm2, apoptosis regulator Bcl-2, BAZ2B, apoptosis regulator Bcl-xL, BRD4 bromodomain 1 BRD4-1, CREB-binding protein (CREBBP), ephrin type-A receptor 4 (EphA4), induced myeloid leukemia cell differentiation protein Mcl-1, and menin. In addition, we randomly selected a subset of 250,000 compounds as additional inactive compounds from the ChEMBL [84] dataset that was used as part of the training datasets. A detailed data preprocessing can be found in Supplementary Note A.

Model strategy and training

Graph attention networks of protein-protein interaction interface

In this section, the representation learning of protein-protein complexes interfaces is inspired by the work in protein docking model evaluation[60], which designed a double-graph representation to capture the interface features and interactions of protein-protein complexes (**Supplementary Figs.1**). The extracted interface region is constructed as two graphs (G^1 and G^2) for representing the interfacial information and the residues involved in the two proteins participating in the interaction. A graph G can be defined as $G = (V, E, \text{and } A)$, where V is the set of nodes, and E is the set of edges between them, and A is the adjacency matrix for mapping the association between the nodes of the graph, which numerically denotes the connectivity of the graph. If the graph G has N nodes, the dimension of the adjacency matrix A of the graph is $\boldsymbol{N}^*\boldsymbol{N}$, where $\boldsymbol{A_{ij}} > 0$ if the i-th node is connected to the j-th node, and $\boldsymbol{A_{ij}} = 0$ otherwise. The graph $\boldsymbol{G^1}$ describes the coding of the atomic types of all residues in the interface region, and its adjacency matrix $\boldsymbol{A^1}$ describes the classification of interatomic bonding types for all residues at the interface region, which only considers the covalent bonds between atoms of interface residues within each subunit as edges. Therefore, it is defined as follows:

$$\mathrm{A}^1_{ij} = \begin{cases} 1 & if \text{ atom } i \text{ and atom } j \text{ are connected by a covalent bond or } if\ i{=}j \\ 0 & \text{otherwise} \end{cases}$$

The graph $\boldsymbol{G^2}$ links both covalent bonds (thus including $\boldsymbol{G^1}$) and non-covalent residue interactions as edges. The adjacency matrix $\boldsymbol{A^2}$ for $\boldsymbol{G^2}$ describes both covalent bonds and non-covalent interactions between atoms within the range of 10.0 Å to each other. The non-covalent atom pairs are defined as those which are closer than 10.0 Å to each other. It is defined as follows:

$$A^2_{ij} = \begin{cases} A^1_{ij}, & if\ i,j \in \text{receptor}\ \ or\ i,j \in \text{ligand} \\ e^{\frac{-(d_{ij}-\mu^2)}{\sigma}}, & if\ \ d_{ij}\ \leq\ 10\text{Å}\ \ and\ i \in \text{receptor and } j \in \text{ligand;} \\ & or\ if\ d_j \leq 10\ \text{Å}\ and\ j \in \text{receptor}\ and\ i \in \text{ligand} \\ 0, & otherwise \end{cases}$$

where $\boldsymbol{d_{ij}}$ represents the distance between the i-th and the j-th atoms of all atoms of all

residues in the interaction region. μ and σ are learnable parameter with initial values of 0.0 and 1.0, respectively. The formula $\boldsymbol{e}^{-(\boldsymbol{d_{ij}}-\boldsymbol{\mu})^2/\boldsymbol{\sigma}}$ decays with increasing distance between atoms.

The graph representation is more flexible and natural to encode interactive information and adjacent(local) relationships. For the node features of the graph, we considered the physicochemical properties of the atoms. We used the same features from the previous work[60, 85, 86]. Then, the feature vector of the nodes is 23 in length and was embedded into 140 features by a one-layer fully connected (FC) network.

The constructed graphs are used as the input for GATs. The graph consists of adjacency matrices $\boldsymbol{A^1}$, $\boldsymbol{A^2}$, node matrices $\boldsymbol{N^1_{mn}}$, $\boldsymbol{N^2_{pq}}$, and the node features, $\boldsymbol{x^{in}} = \{\boldsymbol{x^{in}_1}, \boldsymbol{x^{in}_2}, \cdots, \boldsymbol{x^{in}_N}\}$ and $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{F}}$, where F is the dimensionality of the node features. For the input graph of $\boldsymbol{x^{in}}$, the pure graph attention coefficients are defined as follows, which represent the relative importance between the i-th and j-th nodes:

$$e_{ij} = x_i^T E x_j' + x_j^T E x_i',$$

where $\boldsymbol{x_i'}$ and $\boldsymbol{x_j'}$ are the transformed feature representations defined by $\boldsymbol{x_i'} = \boldsymbol{W}\boldsymbol{x_i^{in}}$ and $\boldsymbol{x_j'} = \boldsymbol{W}\boldsymbol{x_j^{in}}.\boldsymbol{W}$, $\boldsymbol{E} \in \mathbb{R}^{\boldsymbol{F}\times\boldsymbol{F}}$ are learnable matrices in the GATs. $\boldsymbol{e_{ij}}$ and $\boldsymbol{e_{ji}}$ become identical to satisfy the symmetrical property of the graph by adding $\boldsymbol{x_i^T E x_j^T}$ and $\boldsymbol{x_i^T E x_i'}$. The coefficient will only be computed for i and j where $\boldsymbol{A_{ij}} > 0$.

The attention coefficients will also be calculated for the elements in the adjacency matrix. For the elements (i, j), they are defined in the following form:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{j\in N_i} \exp(e_{ij})} A_{ij},$$

where $\boldsymbol{a_{ij}}$ represents the normalized attention coefficient between the i-th and j-th node

pairs, while $\boldsymbol{e_{ij}}$ is the computed symmetric graph attention coefficient. $\boldsymbol{N_i}$ denotes the set of neighbors for the i-th node, which includes the interacting node j with $\boldsymbol{A_{ij}} > 0$. The purpose here is to define attention by considering both the physical structure $\boldsymbol{A_{ij}}$ and the normalized attention coefficient $\boldsymbol{e_{ij}}$ of the interactions simultaneously.

Based on the attention mechanism, the new node features of each node are updated in consideration of its neighboring nodes, which is a linear combination of the neighboring node features and the final attention coefficient $\boldsymbol{a_{ij}}$:

$$x_i'' = \sum_{j \in N_i} a_{ij} x_j',$$

Making the use of the GATs mechanism described previously, we applied four layers of GATs to process the node embedding information of the neighboring nodes and output the updated node embedding. For two adjacency matrices $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$, we use a shared GAT. the initial input to the network is the atomic feature. Working with two matrices $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$, we have $\boldsymbol{x_1} = \boldsymbol{GAT}(\boldsymbol{x^{in}}, \boldsymbol{A^1})$ and $\boldsymbol{x_2} = \boldsymbol{GAT}(\boldsymbol{x^{in}}, \boldsymbol{A^2})$**.** In order to focus only on the intermolecular interactions at the interface of the input protein-protein complex, we obtain the final node embedding by subtracting the embeddings of the two graphs. By subtracting the updated embedding $\boldsymbol{x_1}$ from $\boldsymbol{x_2}$, we can capture aggregated information on intermolecular interactions from only the other nodes in the protein-protein complex interface. The output node feature is therefore defined as:

$$x^{out} = x^2 - x^1,$$

After that, the updated x^{out} became $x^{in}$to iteratively increase the information through the three following GATs layers. After the four GATs layers updated the node embeddings, the node embedding of the entire graph was summed up as the overall intermolecular interaction representation of the protein-protein complex:

$$x_{graph} = \sum\nolimits_{k \in G} x_k.$$

Finally, the FC layers were applied to the $\boldsymbol{x_{graph}}$ to obtain a [4,4,4] vector as features

of the protein-protein interface.

Molecular representation

For each SMILES string, a 3D conformer is generated using RDKit [87] and optimized using the default settings of the MMFF94 force field. The molecular structure information is then extracted into a 35Å grid centered at the geometric center of the molecule using the HTMD package[88]. The atoms of the molecule are discretized into a 1 Å cubic grid, and eight channels are considered to compute voxelized information. Finally, the electronic density of the molecules 9th channel is calculated using the original molecule method in Multiwfn(**Supplementary Figs.2**) [89].

Conditional Wasserstein generative adversarial networks

The generator takes a conditional vector and a noise vector sampled from a Gaussian distribution as inputs. The PPI interface features([1,4,4,4], vector shape) are concatenated with a noise vector of size [9, 4, 4, 4] and input to a 4-layer transposed convolutional neural network (CNNs) with 256, 512, 1024, and 1024 filters, respectively. The first three layers downsample the array size using concatenation convolution (s=2). For all convolutions, we use a kernel size of 4, and the Leaky ReLU is used as an activation function after convolution. BatchNorm3d is applied between convolution and activation operations to normalize the values of each channel of each sample.

The discriminator consists of a 4-layer sequential convolutional neural network (CNNs) with 256, 512, 1024, and 1024 filters, respectively. The first three layers downsample the array size using concatenation convolution (s=2). For all convolutions, we use a kernel size of 4, and the Leaky ReLU (α=0.2) is used as an activation function after convolution. InstanceNorm3d is applied between convolution and activation operations to normalize the values of each channel of each sample.

The physical and spatial features of the compounds are derived from the molecular representation learning module, and the PPI interface features are obtained from the GATs module of the protein complex interface. They are used to estimate the matching probability between molecules and PPI interface features(**Supplementary Figs.3**).

Molecular captioning network

In this section, we will describe how to decode the generated molecular representation into a SMILES strings. Our work is inspired by shape-based molecular generation[90, 91], which designs a combination network of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) [64] to generate SMILES strings. In brief, the molecular captioning network consists of a 3D CNNs and a recurrent LSTM. The molecular representation generated by the generator is fed as input to the 3D CNNs, and the output of the 3D CNN is fed into the LSTM to decode the SMILES strings (**Supplementary Figs.4**).

Model training

The conditional generative adversarial network is trained with Wasserstein loss. The loss functions for the generator $\left(\mathrm{G}_{(0(\mathrm{z,c}))}\right)$ and discriminator $\left(\mathrm{D}_0(\mathrm{x})\right)$ are:

$$L_{x_0} = E_{iy_{xx}}[-D_y(x)] + E_{z_{xx},iy_{yx}}[D_{yy_{yx}}(G_{zy}(z,c))] + \lambda E_{iy_1}[(\| \nabla_{zx} D_y(\hat{x}) \|_z - 1)^2],$$
$$I_{x_{xx}} = E_{z_{xx},ixx_{yx}}\left[-D_z\left(G_{zy}(z,c)\right) - \alpha lo\, g\left(f_u(G_u(z,c),c)\right)\right]$$

where x and c are molecular representations and PPI interface features, respectively, sampled from the true data distribution $\mathrm{p}_{\mathrm{real}}$, z is a random noise vector sampled from a Gaussian distribution (p_{z}), and f_0 is a function that evaluates the probability that a PPI interface feature corresponds to a molecular representation. λ and α terms are regularization parameters, both empirically set to 10. λ term weighs the effect of the gradient penalty on discriminator loss. α term weighs the effect of the effect of f_0 on the loss of the generator.

The model was trained for 50,000 calendar hours with a batch size of 8 (65 steps per calendar hour). The discriminators were updated after each step, while the generators were updated every 30 steps. The network was trained using the RMSprop optimizer with a learning rate of 1 × 10-4 for the generator and discriminator. during training, we monitored the similarity between real and generated molecular representations using Fréchet distances. The weights of the conditional networks were pre-trained on a binary cross-entropy loss and frozen during GAN training. Training was performed on a single NVIDIA A40 GPU, and all neural networks were built and trained using Pytorch 1.7.1 [92] and Tensorflow 2.5 [93].

Molecular generation

After the model has been trained, the embedding information of the protein-protein complex interface is used to guide the model to generate novel molecules from the latent space. The maximum sampling strategy was used in the LSTM, meaning that the SMILES strings are generated by selecting the next token based on the highest prediction probability[90].

Evaluation settings

Conditional evaluation metrics

In this study, the key is to evaluate the effectiveness of the proposed framework of protein-protein interaction interface-based conditional molecular generation. Therefore, we sampled the same number of valid molecules for the three PPI targets. Then we calculated the QED values and Fsp3 by RDKit[87] and calculated the QEPPI values by the QEPPI package (https://github.com/ohuelab/QEPPI) for the generated compounds and others, and plotted the density distribution for comparing the differences of drug-likeness.

MOSES evaluation metrics

To evaluate the performance of our proposed conditional molecule generation framework, we used the evaluation metrics of validity, uniqueness, novelty and diversity provided by the MOSES platform[66], which are defined as follows:

Validity: Molecules defined as valid in the generated molecules.

$$\text{Validity} = \frac{N_{\text{valid}}}{N_{\text{generalated}}}$$

Uniqueness: The proportion of unique molecules found among the generated valid molecules.

$$\text{Uniqueness} = \frac{N_{\text{unique}}}{N_{\text{valid}}}$$

Novelty: The generated molecules are not to be covered in the training set.

$$\text{Novelty} = \frac{N_{\text{novel}}}{N_{\text{unique}}}$$

FCD(Fréchet ChemNet Distance): To detect whether the generated molecules are diverse and whether they have chemical and biological properties that are similar with the real molecules[94].

Molecular shape

To evaluate the shape space of molecules, we used two widely adopted molecular descriptors to represent the three dimensions of molecular structure: principal moment of inertia (PMI)[95] and the best-fit plane (PBF)[96]. The PMI descriptor classifies the geometric shape of molecules into the degree of rod-shaped (linear shape, such as acetylene), disk-shaped (planar shape, such as benzene), and sphere (spherical shape, such as adamantane). The normalized PMI ratios (NPRs) are plotted in two-dimensional triangle and then used to compare the shape space covered by different sets of molecules, evaluating and visualizing the diversity of the molecular shape associated with a given set of molecules[30]. PBF is a three-dimensional descriptor that represents the deviation of a molecule from a plane. The PBF descriptor is the mean distance of each heavy atom from the best-fit plane passing via all heavy atoms[96].

Tree MAP

To explore and explain the chemical space by unsupervised visualization of high-dimensional data[76], we calculated MinHash fingerprint[97] vectors for active compounds and generated compounds. Then tmap[76] and faerun[98] were utilized to construct two-dimensional projections of Tree MAP (TMAP).

Protocol for few-shot generation

Targeting the Hsp90-Cdc37 PPI interface is recognized as an important option for cancer therapy. The crystal structure of the Hsp90-Cdc37 protein complex (PDB ID: 1US7) is available for molecular docking[99]. In addition, known Hsp90-Cdc37 PPI disruptors were collected for training of few-shot generative. They are DCZ3112, Celastrol , FW-04-804, Sulforaphane, Withaferin A, Platycodin D, Kongensin A [100]. OpenPharmacophore(https://github.com/uibcdf/OpenPharmacophore) was utilized to create pharmacophore models and virtual screening. The protein structures were processed by using UCSF Chimera[101], the program DOCK6.9 was used for semiflexible docking[81], and PyMOL[102] was used to create the figures. A detailed docking protocol can be found in Supplementary Note B.

ASSOCIATED CONTENT

Data and code Availability

The datasets are available at Github (https://github.com/AspirinCode/GENiPPI).

Supporting Information

Method details, Supplementary Tables 1, and Supplementary Figures 1−5.

AUTHOR INFORMATION

Corresponding Author

Xiangxiang Zeng − College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China; orcid.org/ 0000-0003-1081-7658; Email: xzeng@hnu.edu.cn

Tao Song – School of Computer Science and Technology, China University of Petroleum, Qingdao, 266580, Shandong, China; Email: tsong@upc.edu.cn

Kyoung Tai No – The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea; Email: ktno@yonsei.ac.kr

Authors

Jianmin Wang – The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea; Email: jmwang113@hotmail.com

Jiashun Mao – The Interdisciplinary Graduate Program in Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea; Email: jiashun_mao@yonsei.ac.kr

Chunyan Li – School of Informatics, Yunnan Normal University, Kunming, China; Email: lchy0316@gmail.com

Hongxin Xiang − College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China; Email: xianghx@hnu.edu.cn

Xun Wang – School of Computer Science and Technology, China University of Petroleum, Qingdao, 266580, Shandong, China; High Performance Computer Research Center, University of Chinese Academy of Sciences, Beijing, 100190, China; Email: wangsyun@upc.edu.cnn

Shuang Wang – School of Computer Science and Technology, China University of Petroleum, Qingdao, 266580, Shandong, China; Email: wangshuang@upc.edu.cn

Zixu Wang – Department of Computer Science, University of Tsukuba, Tsukuba, 3058577, Japan; Email: s2230167@u.tsukuba.ac.jp

Yangyang Chen – Department of Computer Science, University of Tsukuba, Tsukuba, 3058577, Japan; Email: chen.yangyang.xp@alumni.tsukuba.ac.jp

Yuquan Li – College of chemistry and chemical engineering, Lanzhou University, Lanzhou 730000, China; Email: yvquan.li@gmail.com

Heqi Sun – School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China; Email: heqi.sun@sjtu.edu.cn

Author Contributions

J. Wang collected data, developed the model, analyzed the data, and wrote the manuscript; J. Mao and C. Li developed the model and analyzed the data; H. Xiang, X. Wang, S. Wang, Z. Wang, Y. Chen, Y. Li and H. Sun helped to refine the research through constructive discussions and revised the manuscript; K. No, T. Song, X. Zeng supported and supervised the project, interpreted the results, and wrote revisions to the manuscript.

Notes

Any additional relevant notes should be placed here.

ACKNOWLEDGMENTS

This research was supported by the Yonsei University graduate school "Integrative Biotechnology".

ABBREVIATIONS

PPIs, protein-protein interactions; cWGAN, conditional wasserstein generative adversarial network; CNNs, convolutional neural networks; GATs, graph attention networks; LSTM, long short-term memory; QED, quantitative estimation of drug-likeness; QEPPI, quantitative estimate of protein-protein interaction targeting drug-likeness; Fsp3, fraction of sp3 carbon atoms; AAE, adversarial autoencoder; CharRNN ,

character-level recurrent neural networks; VAE, variational autoencoder; t-SNE, t-distributed stochastic neighbor embedding.

REFERENCES

1. Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **2005,** 122, 957-968.
2. Rual, J.-F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **2005,** 437, 1173-1178.
3. Titeca, K.; Lemmens, I.; Tavernier, J.; Eyckerman, S. Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass spectrometry reviews* **2019,** 38, 79-111.
4. Rhys, G. G.; Cross, J. A.; Dawson, W. M.; Thompson, H. F.; Shanmugaratnam, S.; Savery, N. J.; Dodding, M. P.; Höcker, B.; Woolfson, D. N. De novo designed peptides for cellular delivery and subcellular localisation. *Nature Chemical Biology* **2022,** 18, 999-1004.
5. Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.-I. An empirical framework for binary interactome mapping. *Nature methods* **2009,** 6, 83-90.
6. Nero, T. L.; Morton, C. J.; Holien, J. K.; Wielens, J.; Parker, M. W. Oncogenic protein interfaces: small molecules, big challenges. *Nature Reviews Cancer* **2014,** 14, 248-262.
7. Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B. J.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **2021,** 30, 187-200.
8. Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **2007,** 450, 1001-1009.
9. Ivanov, A. A.; Khuri, F. R.; Fu, H. Targeting protein–protein interactions as an anticancer strategy. *Trends in pharmacological sciences* **2013,** 34, 393-400.
10. Ashkenazi, A.; Fairbrother, W. J.; Leverson, J. D.; Souers, A. J. From basic apoptosis discoveries to advanced selective BCL-2 family inhibitors. *Nature reviews drug discovery* **2017,** 16, 273-284.
11. Shin, W.-H.; Christoffer, C. W.; Kihara, D. In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods* **2017,** 131, 22-32.
12. Shin, W.-H.; Kumazawa, K.; Imai, K.; Hirokawa, T.; Kihara, D. Current challenges and opportunities in designing protein–protein interaction targeted drugs. *Advances and Applications in Bioinformatics and Chemistry* **2020**, 11-25.
13. Anderson, A. C. The process of structure-based drug design. *Chemistry & biology* **2003,** 10, 787-797.
14. Wang, X.; Song, K.; Li, L.; Chen, L. Structure-based drug design strategies and challenges. *Current Topics in Medicinal Chemistry* **2018,** 18, 998-1006.
15. Batool, M.; Ahmad, B.; Choi, S. A structure-based drug discovery paradigm. *International journal of molecular sciences* **2019,** 20, 2783.
16. Danel, T.; Łęski, J.; Podlewska, S.; Podolak, I. T. Docking-based generative approaches in the

search for new drug candidates. *Drug Discovery Today* **2022**, 103439.
17. Isert, C.; Atz, K.; Schneider, G. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology* **2023,** 79, 102548.
18. Rakers, C.; Bermudez, M.; Keller, B. G.; Mortier, J.; Wolber, G. Computational close up on protein–protein interactions: how to unravel the invisible using molecular dynamics simulations? *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015,** 5, 345-359.
19. Ni, D.; Lu, S.; Zhang, J. Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug discovery. *Medicinal research reviews* **2019,** 39, 2314-2342.
20. Janin, J.; Chotia, C. The structure of protein-protein recognition sites. *The Journal of biological chemistry (Print)* **1990,** 265, 16027-16030.
21. Smith, M. C.; Gestwicki, J. E. Features of protein–protein interactions that translate into potent inhibitors: topology, surface area and affinity. *Expert reviews in molecular medicine* **2012,** 14, e16.
22. Wang, Z.-Z.; Shi, X.-X.; Huang, G.-Y.; Hao, G.-F.; Yang, G.-F. Fragment-based drug discovery supports drugging 'undruggable'protein–protein interactions. *Trends in Biochemical Sciences* **2023**.
23. Mignani, S.; Rodrigues, J.; Tomas, H.; Jalal, R.; Singh, P. P.; Majoral, J.-P.; Vishwakarma, R. A. Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? *Drug discovery today* **2018,** 23, 605-615.
24. Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. In vitro models for selection of development candidatesexperimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **1997,** 23, 3-25.
25. Lipinski, C. A. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies* **2004,** 1, 337-341.
26. Morelli, X.; Bourgeas, R.; Roche, P. Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Current opinion in chemical biology* **2011,** 15, 475-481.
27. Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **2012,** 4, 90-98.
28. Kosugi, T.; Ohue, M. In *Quantitative Estimate of Protein-Protein Interaction Targeting Drug-likeness*, 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 13-15 Oct. 2021, 2021; 2021; pp 1-8.
29. Kosugi, T.; Ohue, M. Quantitative Estimate Index for Early-Stage Screening of Compounds Targeting Protein-Protein Interactions. *International Journal of Molecular Sciences* **2021,** 22, 10925.
30. Wang, J.; Mao, J.; Wang, M.; Le, X.; Wang, Y. Explore drug-like space with deep generative models. *Methods* **2023**.
31. Qiu, Y.; Li, X.; He, X.; Pu, J.; Zhang, J.; Lu, S. Computational methods-guided design of modulators targeting protein-protein interactions (PPIs). *European Journal of Medicinal Chemistry* **2020,** 207, 112764.
32. Stokel-Walker, C.; Van Noorden, R. What ChatGPT and generative AI mean for science. *Nature* **2023,** 614, 214-216.
33. Urbina, F.; Lentzos, F.; Invernizzi, C.; Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* **2022,** 4, 189-191.

34. Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022,** 12, e1608.
35. Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics* **2021,** 22, bbab344.
36. Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative models for De Novo drug design. *Journal of Medicinal Chemistry* **2021,** 64, 14011-14027.
37. Wang, M.; Wang, Z.; Sun, H.; Wang, J.; Shen, C.; Weng, G.; Chai, X.; Li, H.; Cao, D.; Hou, T. Deep learning approaches for de novo drug design: An overview. *Current opinion in structural biology* **2022,** 72, 135-144.
38. Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discovery Today* **2021**.
39. Thomas, M.; Bender, A.; de Graaf, C. Integrating structure-based approaches in generative molecular design. *Current Opinion in Structural Biology* **2023,** 79, 102559.
40. Zeng, X.; Wang, F.; Luo, Y.; Kang, S.-g.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine* **2022**, 100794.
41. Martinelli, D. Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine* **2022**, 105403.
42. Özçelik, R.; van Tilborg, D.; Jiménez-Luna, J.; Grisoni, F. Structure-based drug discovery with deep learning. *arXiv preprint arXiv:2212.13295* **2022**.
43. Ma, B.; Terayama, K.; Matsumoto, S.; Isaka, Y.; Sasakura, Y.; Iwata, H.; Araki, M.; Okuno, Y. Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *Journal of Chemical Information and Modeling* **2021,** 61, 3304-3313.
44. Luo, S.; Guan, J.; Ma, J.; Peng, J. A 3D generative model for structure-based drug design. *Advances in Neural Information Processing Systems* **2021,** 34, 6229-6239.
45. Drotár, P.; Jamasb, A. R.; Day, B.; Cangea, C.; Liò, P. Structure-aware generation of drug-like molecules. *arXiv preprint arXiv:2111.04107* **2021**.
46. Li, Y.; Pei, J.; Lai, L. Structure-based de novo drug design using 3D deep generative models. *Chemical science* **2021,** 12, 13664-13675.
47. Long, S.; Zhou, Y.; Dai, X.; Zhou, H. Zero-Shot 3D Drug Design by Sketching and Generating.
48. Peng, X.; Luo, S.; Guan, J.; Xie, Q.; Peng, J.; Ma, J. In *Pocket2mol: Efficient molecular sampling based on 3d protein pockets*, International Conference on Machine Learning, 2022; PMLR: 2022; pp 17644-17655.
49. Wang, M.; Hsieh, C.-Y.; Wang, J.; Wang, D.; Weng, G.; Shen, C.; Yao, X.; Bing, Z.; Li, H.; Cao, D. Relation: A deep generative model for structure-based de novo drug design. *Journal of Medicinal Chemistry* **2022,** 65, 9478-9492.
50. Chan, L.; Kumar, R.; Verdonk, M.; Poelking, C. A multilevel generative framework with hierarchical self-contrasting for bias control and transparency in structure-based ligand design. *Nature Machine Intelligence* **2022**, 1-13.
51. Zhang, O.; Zhang, J.; Jin, J.; Zhang, X.; Hu, R.; Shen, C.; Cao, H.; Du, H.; Kang, Y.; Deng, Y. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence* **2023**, 1-11.

52. Neugebauer, A.; Hartmann, R. W.; Klein, C. D. Prediction of protein– protein interaction inhibitors by chemoinformatics and machine learning methods. *Journal of medicinal chemistry* **2007,** 50, 4665-4668.
53. Gupta, P.; Mohanty, D. SMMPPI: a machine learning-based approach for prediction of modulators of protein–protein interactions and its application for identification of novel inhibitors for RBD: hACE2 interactions in SARS-CoV-2. *Briefings in Bioinformatics* **2021,** 22, bbab111.
54. Díaz-Eufracio, B. I.; Medina-Franco, J. L. Machine Learning Models to Predict Protein–Protein Interaction Inhibitors. *Molecules* **2022,** 27, 7986.
55. Reker, D.; Schneider, P.; Schneider, G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chemical science* **2016,** 7, 3919-3927.
56. Mallet, V.; Checa Ruano, L.; Moine Franel, A.; Nilges, M.; Druart, K.; Bouvier, G.; Sperandio, O. InDeep: 3D fully convolutional neural networks to assist in silico drug design on protein–protein interactions. *Bioinformatics* **2022,** 38, 1261-1268.
57. Wang, J.; Chu, Y.; Mao, J.; Jeon, H.-N.; Jin, H.; Zeb, A.; Jang, Y.; Cho, K.-H.; Song, T.; No, K. T. De novo molecular design with deep molecular generative models for PPI inhibitors. *Briefings in Bioinformatics* **2022,** 23.
58. Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE transactions on neural networks* **2008,** 20, 61-80.
59. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *stat* **2017,** 1050, 10.48550.
60. Wang, X.; Flannery, S. T.; Kihara, D. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences* **2021,** 8, 647915.
61. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* **2014**.
62. Wang, Y.; Wu, S.; Duan, Y.; Huang, Y. ResAtom system: protein and ligand affinity prediction model based on deep learning. *arXiv preprint arXiv:2105.05125* **2021**.
63. Zapata, P. A. M.; Méndez-Lucio, O.; Le, T.; Beese, C. J.; Wichard, J.; Rouquié, D.; Clevert, D.-A. Cell morphology-guided de novo hit design by conditioning GANs on phenotypic image features. *Digital Discovery* **2023**.
64. Schmidhuber, J.; Hochreiter, S. Long short-term memory. *Neural Comput* **1997,** 9, 1735-1780.
65. Wei, W.; Cherukupalli, S.; Jing, L.; Liu, X.; Zhan, P. Fsp3: A new parameter for drug-likeness. *Drug Discovery Today* **2020,** 25, 1839-1845.
66. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in pharmacology* **2020,** 11, 1931.
67. Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A. Applications of deep learning in biomedicine. *Molecular pharmaceutics* **2016,** 13, 1445-1454.
68. Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* **2018,** 4, 120-131.
69. Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
70. Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de novo molecular generation method using latent vector based generative adversarial

network. *Journal of Cheminformatics* **2019,** 11, 1-13.
71. Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843* **2017**.
72. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002,** 42, 1273-1280.
73. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008,** 9.
74. Hinton, G. E.; Roweis, S. Stochastic neighbor embedding. *Advances in neural information processing systems* **2002,** 15.
75. Meyers, J.; Carter, M.; Mok, N. Y.; Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future medicinal chemistry* **2016,** 8, 1753-1767.
76. Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **2020,** 12, 1-13.
77. Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low data drug discovery with one-shot learning. *ACS central science* **2017,** 3, 283-293.
78. Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative molecular design in low data regimes. *Nature Machine Intelligence* **2020,** 2, 171-180.
79. Wang, J.; Zheng, S.; Chen, J.; Yang, Y. Meta learning for low-resource molecular optimization. *Journal of Chemical Information and Modeling* **2021,** 61, 1627-1636.
80. Chen, X.; Liu, P.; Wang, Q.; Li, Y.; Fu, L.; Fu, H.; Zhu, J.; Chen, Z.; Zhu, W.; Xie, C. DCZ3112, a novel Hsp90 inhibitor, exerts potent antitumor activity against HER2-positive breast cancer through disruption of Hsp90-Cdc37 interaction. *Cancer Letters* **2018,** 434, 70-80.
81. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of new features and current docking performance. *Journal of computational chemistry* **2015,** 36, 1132-1156.
82. Sun, H.; Wang, J.; Wu, H.; Lin, S.; Chen, J.; Wei, J.; Lv, S.; Xiong, Y.; Wei, D.-Q. A Multimodal Deep Learning Framework for Predicting PPI-Modulator Interactions. *Journal of Chemical Information and Modeling* **2023**.
83. Singh, N.; Chaput, L.; Villoutreix, B. O. Fast rescoring protocols to improve the performance of structure-based virtual screening performed on protein–protein interfaces. *Journal of chemical information and modeling* **2020,** 60, 3910-3934.
84. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012,** 40, D1100-D1107.
85. Torng, W.; Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *Journal of chemical information and modeling* **2019,** 59, 4131-4149.
86. Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of chemical information and modeling* **2019,** 59, 3981-3988.
87. Landrum, G. RDKit: Open-source cheminformatics. *https://www.rdkit.org/* **2006**.
88. Doerr, S.; Harvey, M.; Noé, F.; De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *Journal of chemical theory and computation* **2016,** 12, 1845-1852.
89. Lu, T.; Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *Journal of computational*

chemistry **2012,** 33, 580-592.
90. Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling* **2019,** 59, 1205-1214.
91. Gaulton, A.; Kale, N.; van Westen, G. J. P.; Bellis, L. J.; Bento, A. P.; Davies, M.; Hersey, A.; Papadatos, G.; Forster, M.; Wege, P.; Overington, J. P. A large-scale crop protection bioassay data set. *Scientific Data* **2015,** 2, 150032.
92. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019,** 32.
93. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow. org (2015). *URL https://www. tensorflow. org* **2015**.
94. Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. n. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling* **2018,** 58, 1736-1741.
95. Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *Journal of chemical information and computer sciences* **2003,** 43, 987-1003.
96. Firth, N. C.; Brown, N.; Blagg, J. Plane of best fit: a novel method to characterize the three-dimensionality of molecules. *Journal of chemical information and modeling* **2012,** 52, 2516-2525.
97. Probst, D.; Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *Journal of cheminformatics* **2018,** 10, 1-12.
98. Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2018,** 34, 1433-1435.
99. Roe, S. M.; Ali, M. M.; Meyer, P.; Vaughan, C. K.; Panaretou, B.; Piper, P. W.; Prodromou, C.; Pearl, L. H. The mechanism of Hsp90 regulation by the protein kinase-specific cochaperone p50cdc37. *Cell* **2004,** 116, 87-98.
100. Dike, P. P.; Bhowmick, S.; Eldesoky, G. E.; Wabaidur, S. M.; Patil, P. C.; Islam, M. A. In silico identification of small molecule modulators for disruption of Hsp90–Cdc37 protein–protein interaction interface for cancer therapeutic application. *Journal of Biomolecular Structure and Dynamics* **2022,** 40, 2082-2098.
101. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** 25, 1605-1612.
102. Schrödinger, L. The PyMOL molecular graphics system, version 1.8. In November: 2015.