Journal of Cheminformatics

RESEARCH Open Access

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

Jianmin Wang[1], Jiashun Mao[1], Chunyan Li[4], Hongxin Xiang[2], Xun Wang[3,5], Shuang Wang[3], Zixu Wang[6], Yangyang Chen[6], Yuquan Li[7], Kyoung Tai No[1*], Tao Song[3*] and Xiangxiang Zeng[2*]

Abstract

Protein–protein interactions (PPIs) play a crucial role in numerous biochemical and biological processes. Although several structure-based molecular generative models have been developed, PPI interfaces and compounds targeting PPIs exhibit distinct physicochemical properties compared to traditional binding pockets and small-molecule drugs. As a result, generating compounds that effectively target PPIs, particularly by considering PPI complexes or interface hotspot residues, remains a significant challenge. In this work, we constructed a comprehensive dataset of PPI interfaces with active and inactive compound pairs. Based on this, we propose a novel molecular generative framework tailored to PPI interfaces, named GENiPPI. Our evaluation demonstrates that GENiPPI captures the implicit relationships between the PPI interfaces and the active molecules, and can generate novel compounds that target these interfaces. Moreover, GENiPPI can generate structurally diverse novel compounds with limited PPI interface modulators. To the best of our knowledge, this is the first exploration of a structure-based molecular generative model focused on PPI interfaces, which could facilitate the design of PPI modulators. The PPI interface-based molecular generative model enriches the existing landscape of structure-based (pocket/interface) molecular generative model.

Scientific contribution This study introduces GENiPPI, a protein-protein interaction (PPI) interface-aware molecular generative framework. The framework first employs Graph Attention Networks to capture atomic-level interaction features at the protein complex interface. Subsequently, Convolutional Neural Networks extract compound representations in voxel and electron density spaces. These features are integrated into a Conditional Wasserstein Generative AdversarialNetwork, which trains the model to generate compound representations targeting PPI interfaces. GENiPPI effectively captures the relationship between PPI interfaces and active/inactive compounds. Furthermore, in fewshot molecular generation, GENiPPI successfully generates compounds comparable to known disruptors. GENiPPI provides an efficient tool for structure-based design of PPI modulators.

Keywords Protein–protein interaction modulators, Molecular generative model, Geometric deep learning, GAT, Conditional WGAN

*Correspondence:
Kyoung Tai No
ktno@yonsei.ac.kr
Tao Song
tsong@upc.edu.cn
Xiangxiang Zeng
xzeng@hnu.edu.cn
Full list of author information is available at the end of the article

Introduction

A vast network of genes is interconnected through protein–protein interactions (PPIs), which are critical components of nearly every biological process under physiological conditions and are ubiquitous in various organisms and biological pathways [1–4]. Modulating PPIs broadens the drug target space and holds significant potential in drug discovery. In humans, the estimated size of the interactome ranges from 130,000 to 930,000 binary PPIs [5–7]. Despite considerable efforts, developing modulators of PPI targets, particularly those targeting PPI interfaces, remains challenging [6, 8–12]. Structure-based rational design plays a vital role in identifying lead compounds for drug discovery [13–17]. Traditional drug targets and PPI targets exhibit distinct biochemical characteristics (Table 1) [11, 18–22], further exhibiting differences in the physicochemical and drug-like properties of conventional drugs and PPI modulators (Table 1) [11, 23–30]. Given the differences outlined in Table 1, developing molecular generative models tailored to different paradigms is crucial for designing drugs for various target types [10, 19, 31].

Generative artificial intelligence (AI) is capable of modeling the distribution of training samples and generating novel samples [32, 33]. In drug discovery, generative AI can accelerate the process of drug discovery by generating novel molecules with desired properties. Several excellent review articles have summarized development in this field [16, 17, 34–41]. Molecular generative models in drug design can be broadly categorized into three categories: ligand-based molecular generative (LBMG) models, structure-based molecular generative (SBMG) models (focusing on pockets or binding sites), and fragment-based molecular generative (FBMG) models. Among these, SBMG models have garnered significant attention [17, 39, 42]. While key methods for structure-based molecular generative models are well-documented [43–51], molecular generative models specifically targeting PPI structures or interfaces are rarely reported in the literature. In recent years, classical machine learning [52–54], active learning [55], and deep learning-assisted approaches have been explored to improve the screening and design of PPI modulators [56], and some ligand-based molecular generative models for PPI modulators have been reported [57]. However, structure-based molecular generative models for PPI targets remain underexplored.

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

In this study, we developed GENiPPI, a structure-based conditional molecular generative framework designed for the generation of protein–protein interaction (PPI) interface modulators. The framework begins by utilizing Graph Attention Networks (GATs) to capture the subtle atomic-level interaction features present at the protein complex interface. Convolutional Neural Networks (CNNs) are then employed to extract compound representations in voxel and electron density space. Following this, a Conditional Wasserstein Generative Adversarial Network (cWGAN) integrates these features to train a model that generates compound representations targeting PPI interfaces. Finally, the CNN module and LSTM network decodes the molecular embeddings into SMILES strings. The framework is designed to capture the relationship between PPI interface with active/inactive compounds, enabling the training of conditional molecular generative models specifically tailored to PPI interfaces (Fig. 1). Conditional model evaluation shows that the GENiPPI framework effectively captures the implicit relationships between PPI interfaces and active compounds, generating compounds with drug-like properties that resemble those of active compounds targeting specific PPI sites. In terms of performance, GENiPPI outperforms other generative models such as LatentGAN and ORGAN, demonstrating superiority in the novelty, diversity, and validity of the generated molecules. Additionally, in few-shot molecular generation, GENiPPI successfully generated compounds targeting the Hsp90-Cdc37 interaction with chemical properties similar to known disruptors, proving effective even with limited labeled data. In conclusion, GENiPPI represents a potent deep learning framework for structure-based design of PPI modulators.

Results and discussion

Generation of molecules targeting the PPI interface

In this study, we introduce GENiPPI, a modular deep learning framework for the design of structure-based

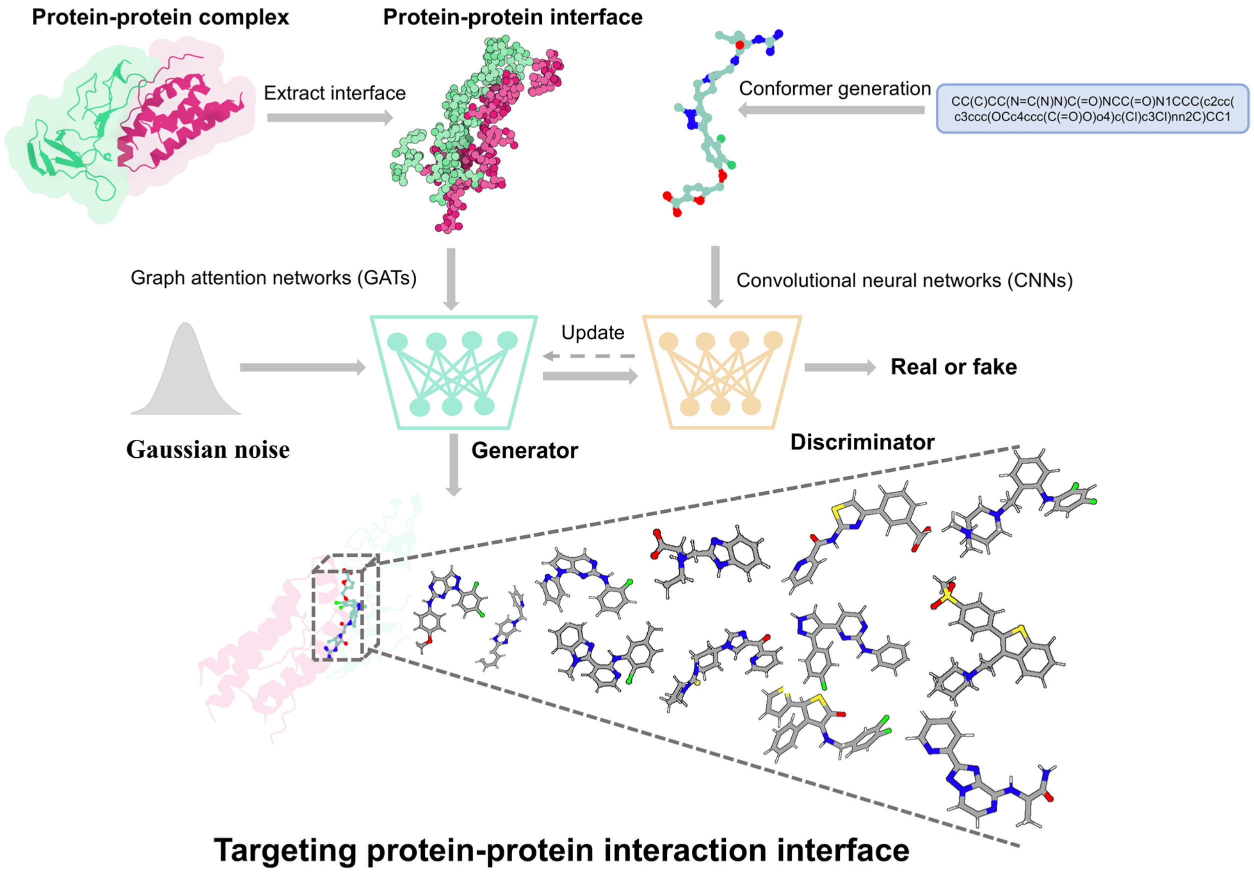
Protein-protein complex
Protein-protein interface
Extract interface
Conformer generation
CC(C)CC(N=C(N)N)C(=O)NCC(=O)N1CCC(c2cc(
c3ccc(OCc4ccc(C(=O)O)o4)c(Cl)c3Cl)nn2C)CC1
Graph attention networks (GATs)
Convolutional neural networks (CNNs)
Update
Gaussian noise
Generator
Discriminator
Real or fake
Targeting protein-protein interaction interface

Fig. 1 The generation of molecules targeting PPI. The 3D structural information of the protein–protein complex interface is represented as a graph, with feature representation of the interface region captured using a graph attention neural networks (GATs). The voxel and electron density of the compound are encoded by 3D convolutional neural networks (CNNs). Conditional Wasserstein generative adversarial networks (cWGAN) is trained to generate molecular embeddings conditioned on interface features. The generator takes interface features and random noise vectors to generate molecular embeddings, while the discriminator evaluates the probability that a molecule is real or generated. The condition regulates the generation of molecules constrained by specific protein–protein interfaces. Finally, long short-term memory (LSTM) networks decode the molecular representations into SMILES strings

PPI modulators (Fig. 1). GENiPPI consists of four main modules: a Graph Attention Networks (GATs) module for representation learning of the protein complex interface as the conditional vector [58–60], a Convolutional Neural Networks (CNNs) module is used to capture the molecular features of modulators as the primary input, Conditional Wasserstein GAN (cWGAN) module for conditional molecular generation takes the conditional vector and main input [61], and a molecular captioning network module for decoding molecular embeddings into SMILES strings (as shown in Supplementary Figs. 1, 2, 3, and 4, respectively).

Our framework follows four steps to generate molecules targeting PPI interfaces. First, the GATs module extracts atomic-level interaction characteristics of the protein complex interface region, effectively capturing the nuanced structural characteristics crucial for interaction. Then, the CNN module encodes the molecular features in a three-dimensional space, incorporating voxel and electronic density information [62]. This ensures a robust representation of the molecular structure that is suitable for generational tasks.

Next, the cWGAN module then generates compounds targeting PPI interfaces by utilizing features from the protein complex interface to regulate the inputs [63]. This cWGAN module consists of three components: the generator, the discriminator, and 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 determines whether the generated molecular embedding corresponds to a real or generated molecule, while the conditional network assesses whether the molecular embedding matches the protein complex interface features. Finally, the molecular captioning network decodes the molecular embeddings into SMILES strings. This network comprises a 3D CNN that processes the molecular embedding followed by an LSTM (Long Short-Term Memory) network, which sequentially decodes the learned embeddings into valid molecular structures in SMILES strings [64]. This step ensures that the generated molecular designs are usable in further drug design studies.

Conditional evaluation

To comprehensively assess the validity of the conditions used as conditional molecular generative models targeting the protein complex interfaces. To do this, we conducted a detailed analysis using three distinct 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). These targets were selected due to the availability of high-quality labeled data and their significance in cancer biology.

For each PPI target, we used the GENiPPI framework to generate 10,000 validated molecules and calculated the key drug-like metrics of the generated compounds: QED [27], QEPPI [28, 29] and Fsp3(fraction of sp3 carbon atoms) [65]. These metrics are essential indicators of drug-likeness, PPI-targeting drug-likeness, and molecular complexity, respectively. The aim of these calculations was to determine how well the generated molecules align with the drug-like properties of known active compounds and to evaluate the influence of conditional input on the generative process.

We then compared the QED, QEPPI, and Fsp3 distributions of the active compounds and generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4 (Fig. 2). The results show that the drug-like property distributions of the generated compounds closely resemble those of the active compounds for all three PPI interface targets (Fig. 2a, b, c). This suggests that the conditional input derived from the PPI interface features plays a critical role in guiding the generation process toward biologically relevant compounds.

Interestingly, we observed differences in drug-like property distributions between the generated compounds across different PPI targets (Fig. 2d, e, and f). These findings demonstrate the effectiveness of the PPI interface in conditioning the molecular generative model. For instance, MDM2/p53 and Bcl-2/Bax have notably different interface architectures and binding hot spots, which likely result in the generation of compounds with distinct QEPPI and Fsp3 profiles. These findings underscore the specificity of the conditioning framework, which adapts the molecular generation process to the target PPI interface, thus ensuring that the generated molecules are tailored to the unique features of each PPI target.

We performed independent t-tests to compare the mean QED, QEPPI, and Fsp^3 values between the active and generated compounds for each PPI target. The results indicated statistically significant differences between the generated and active compounds across various metrics (Fig. 3). For example, significant differences in QED and QEPPI were observed for MDM2/p53 and Bcl-2/Bax, suggesting that active compounds exhibit better drug-likeness and PPI-targeting drug-likeness, while Fsp^3 distributions remain comparable. However, in the case of BAZ2B/H4, the QED and QEPPI values were similar between active and generated compounds, with notable differences observed only in the Fsp^3 metric. The overall analysis confirms that while generated compounds can closely mimic the drug-like properties of active compounds, especially in QED and QEPPI metrics for specific targets, significant differences exist, particularly in molecular complexity, depending on the PPI target.

Moreover, the drug-like properties of the generated compounds shifted relative to those in the training dataset, indicating that the GENiPPI framework does more than merely reproduce the distributions of known molecules; it generates novel compounds that maintain drug-likeness while exploring new regions of chemical space. This capacity to innovate within the bounds of known drug-likeness properties is a hallmark of successful generative models, as it enables the discovery of potentially more effective or optimized PPI modulators.

Model performance

To assess the performance of the GENiPPI framework and compare it with other molecular generative models. We benchmarked our method using 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. The models and hyperparameters provided by the MOSES platform were used, including the Adversarial Autoencoder (AAE) [67], character-level recurrent neural networks (Char-RNN) [68], Variational Autoencoder (VAE) [69], Latent-GAN [70] and ORGAN [71]. To validate the quality of the molecules generated by the conditioned model, we compared them to molecules generated by the GENiPPI framework and the GENiPPI-noninterface framework, which lacks the conditioned module. Our results showed that molecules generated by the conditioned GENiPPI framework outperformed those generated by other models in terms of novelty and diversity. This improvement can be attributed to the conditioning module, which directs the model to focus on specific chemical spaces associated with PPI interfaces. By conditioning molecular generation on relevant PPI features, the GENiPPI framework ensures that the generated molecules possess characteristics desirable for PPI inhibition. In contrast, models lacking this module, such as the GENiPPI-noninterface, fail to maintain this level of focus, resulting in lower performance across novelty and diversity metrics.

As shown in Table 2, the GENiPPI framework demonstrated advantages in uniqueness, novelty, and diversity over the GENiPPI-noninterface framework. Overall, the GENiPPI framework performed better in molecular generation. Compared to LatentGAN and ORGAN, GENiPPI offered superior results in terms of validity and diversity. While each molecular generative model has its strengths across various performance metrics, models tailored to specific tasks, such as those based on PPI structures, show advantages and inspirations from the GENiPPI framework.

To further understand the similarities and differences between the molecular distributions generated by the GENiPPI framework and other models, we compared the distribution of molecular properties in the Testset, iPPI-DB inhibitor [72], and the generated molecular datasets from AAE, CharRNN, VAE, LatentGAN, GENiPPI-noninterface and GENiPPI (Supplementary Fig. 5). The generated compounds showed similar distributions of physicochemical properties to those in the training set. The QED values of the generated molecules were particularly close to those of the iPPI-DB inhibitors, indicating that the GENiPPI framework effectively learns and applies the desired molecular characteristics. Notably, while most iPPI-DB inhibitors have QED values lower than 0.5, the majority of generated molecules from the GENiPPI framework exhibited QEPPI values higher than 0.5. This suggests that the GENiPPI framework not only captures the drug-likeness of molecules but also their PPI-targeting potential, which is essential for PPI-related drug discovery tasks. This consistency in property distribution underscores the framework's ability to model complex chemical spaces and generate biologically relevant, novel compounds.

Chemical space exploration

To more comprehensively estimate the chemical space distribution of the model generated molecules in comparison to the active compounds from the training datasets, we evaluated the chemical drug-like space of the generated compounds by calculating t-distributed stochastic neighbourhood embedding (t-SNE) maps of MACCS fingerprints [73]. t-SNE is a widely-used dimensionality reduction method used for visualizing data points in two or three-dimensional space by mapping high-dimensional data to lower dimension [74, 75]. This method clusters similar compounds, allowing for a clear understanding of how generated compounds occupy chemical space in comparison to active, known compounds.

The distribution of both generated and active compounds in the chemical drug-like space was visualized using t-SNE visualization (Fig. 4a–c). Our findings reveal that 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. This observation indicates that the GENiPPI framework successfully generates compounds that occupy the same drug-like space as known active modulators, reinforcing the model's capacity to produce viable drug candidates. Moreover, under the 2D topological fingerprints, the generated compounds exhibit a similar chemical drug-like space to that of the active. This similarity suggests that the generative model is effective in capturing key topological features of molecules that are critical for drug-likeness. However, relying solely on two-dimensional representations may not be sufficient to fully assess drug-likeness, particularly for PPI modulators, which often require more complex three-dimensional features for effective binding. Incorporating three-dimensional features into the compounds contributes to the design of promising drug-like compounds [30, 76]. To address this, we conducted principal moments of inertia (PMI) shape analysis on the generated compounds and compared them with drug-like compounds from DrugBank and iPPI-DB (Fig. 4d). This analysis revealed that many approved compounds are either rod or disk shaped, and the generated drug-like compounds display a similar three-dimensional shape distribution. Such shapes are often

(See figure on next page.)

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

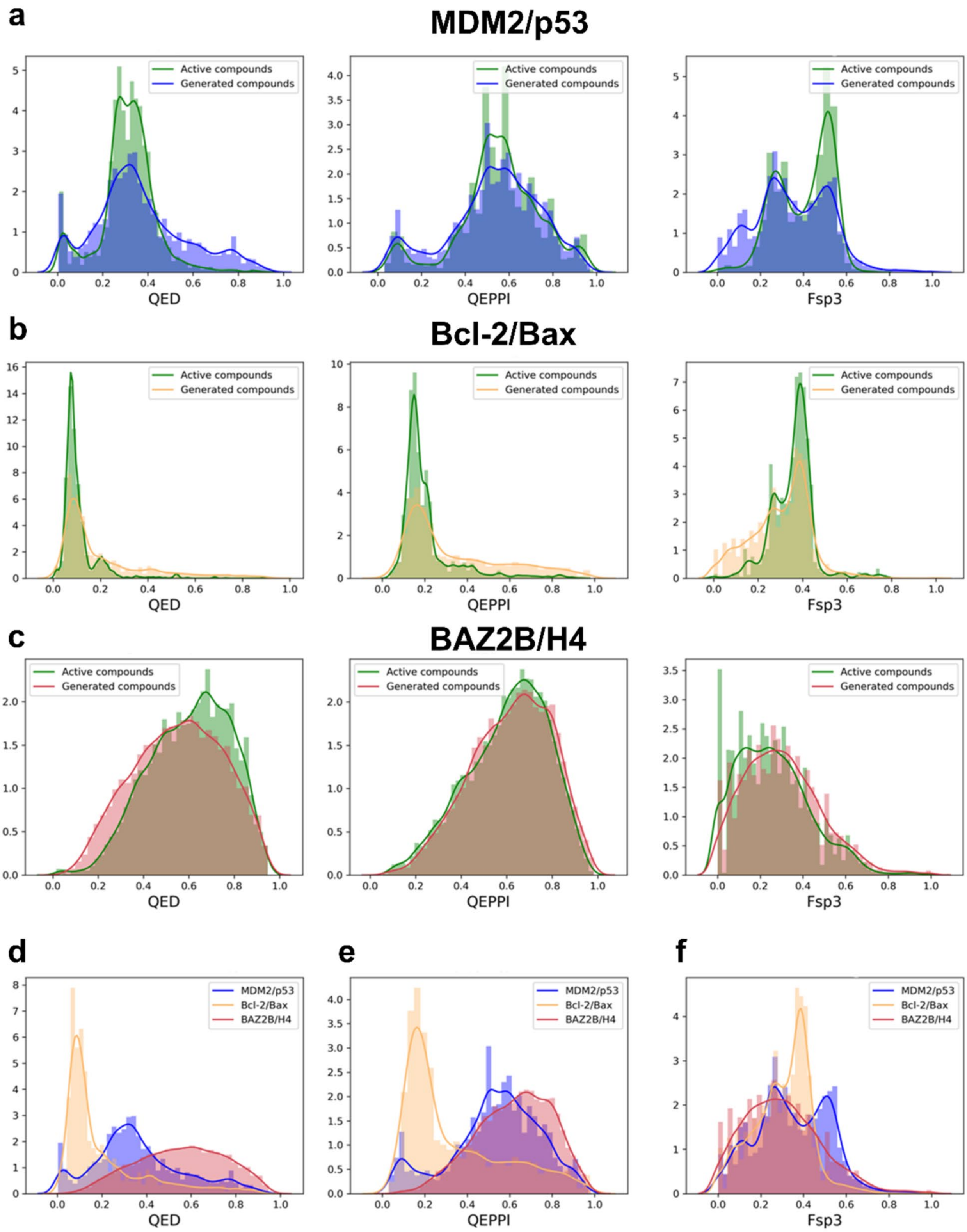

Fig. 2 (See legend on previous page.)

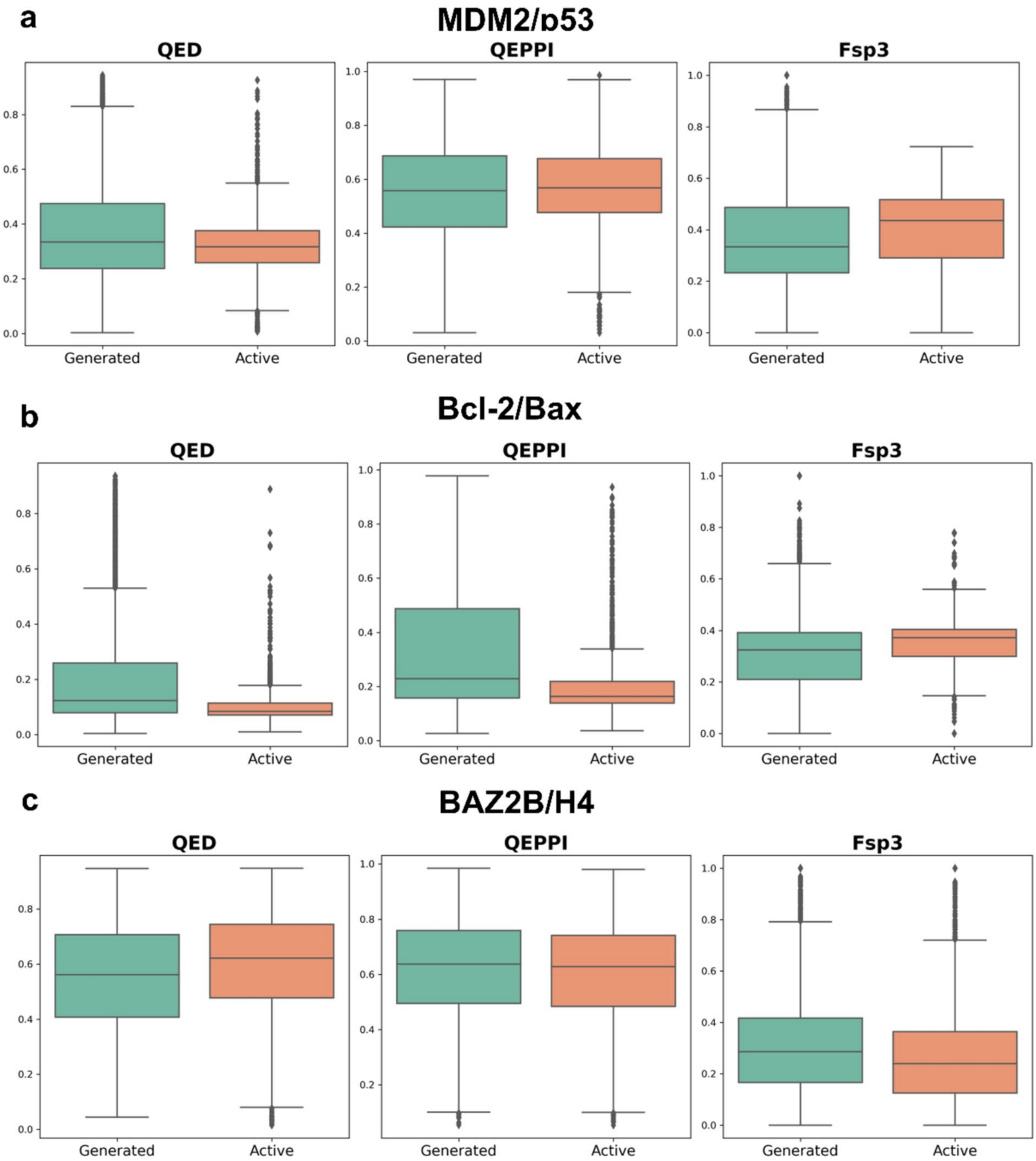

Fig. 3 Independent t-test comparisons of QED, QEPP, and Fsp.[3] values between generated and active compounds across three PPI targets: **a** MDM2/p53, **b** Bcl-2/Bax, **c** BAZ2B/H4

crucial for the spatial complementarity required for targeting PPI interfaces. Additionally, the PMI shape analysis suggests that the model is generating compounds with appropriate three-dimensional features that align with known drug-like shapes, further validating the robustness of the generation process. We also evaluated the plane of best fit (PBF) score of the generated drug-like compounds, a parameter that describes the extent to which molecular scaffolds deviate from planarity. The PBF distribution of the generated library ranges from 0 to 2 Å (Fig. 4e), indicating that many generated

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

Model	Valid	Unique@1 k	Unique@10 k	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

drug-like compounds are derived from relatively planar molecular scaffolds.

Additionally, we assessed the ability of the model to generate PPI target-specific compounds using chemical space maps. To evaluate the overlap of drug-like chemical space, we utilized Tree MAP (TMAP) to create a 2D projection (Fig. 4f) [77]. Each point represents a compound, colored by its target label, with dark and light colors denoting generated and active compounds, respectively. These results suggest that the GENiPPI model can generate compounds similar to the active compounds in the training set while introducing novel structures. Overall, the framework enriches and expands the chemical space of PPI-targeted drug-like compounds.

Few-shot molecular generation

Due to the high associated with data collection, only a small amount of labeled biomedical data is typically available, presenting challenges for drug design and optimization, which often faces the problem of limited data [78]. This scarcity of labeled data can diminish the practical performance of most deep learning frameworks in drug design. Addressing the challenge of generating molecular designs with limited labeled data has become a significant focus in the few-shot generative community [79, 80]. Few-shot learning aims to train models using only a small number of examples while still enabling them to generalize effectively to novel tasks. This capability is critical for drug discovery, where only a few experimentally verified molecules are often available for new PPI targets. The GENiPPI model was applied to generate a virtual compound library targeting the interaction interface between heat shock protein 90 (Hsp90) and cell division cycle 37(Cdc37). By training the model on the PPI structure of Hsp90-Cdc37 (PDB ID: 1US7) and using data from seven disruptors, we sampled 500 valid compounds. The chemical space similarity between active disruptors and generated compounds for Hsp90/Cdc37 was visualized using t-SNE projection maps (Fig. 5a), which revealed that the generated molecules were largely clustered around the active disruptors in chemical space. This result demonstrates the effectiveness of few-shot learning in navigating through the targeted chemical space and generating compounds that are structurally similar to known active disruptors despite limited training data.

In order to further assess the chemical relevance of the generated compounds, we performed pharmacophore-based matching using DCZ3112, a novel triazine derivative that disrupts Hsp90-Cdc37 interactions, as a reference molecule [81]. The top 5 generated molecules showed similar pharmacophore and shape features to DCZ3112 (Fig. 5b). The similarity of these features between DCZ3112 and the generated molecules indicates that the model was able to successfully learn the essential characteristics required for disrupting the Hsp90-Cdc37 interaction, even with limited input data.

To further validate the model s performance, we examined the interaction patterns between the generated compounds and the PPI interface by performing molecular docking. Previous studies have identified key hot spot amino acid residues at the PPI interface (PDB ID: 1US7) [81, 82], as shown in Fig. 5c. We performed molecular docking for prediction of the binding poses (Fig. 5d) of DCZ3112 with the Hsp90-Cdc37 complex using the UCSF DOCK6.9 program [83]. 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. 5d), which are major contributors to protein–ligand interactions. Subsequently, molecular docking was also performed on the GENiPPI-generated compounds, alongside DCZ3112, and compounds with reasonable binding modes and higher binding affinity were selected for interaction pattern analysis. The GENiPPI-generated compounds not only achieved a better docking score than the active compounds but also reproduced the key interactions with the crucial residues of the PPI interface. This indicates that the model

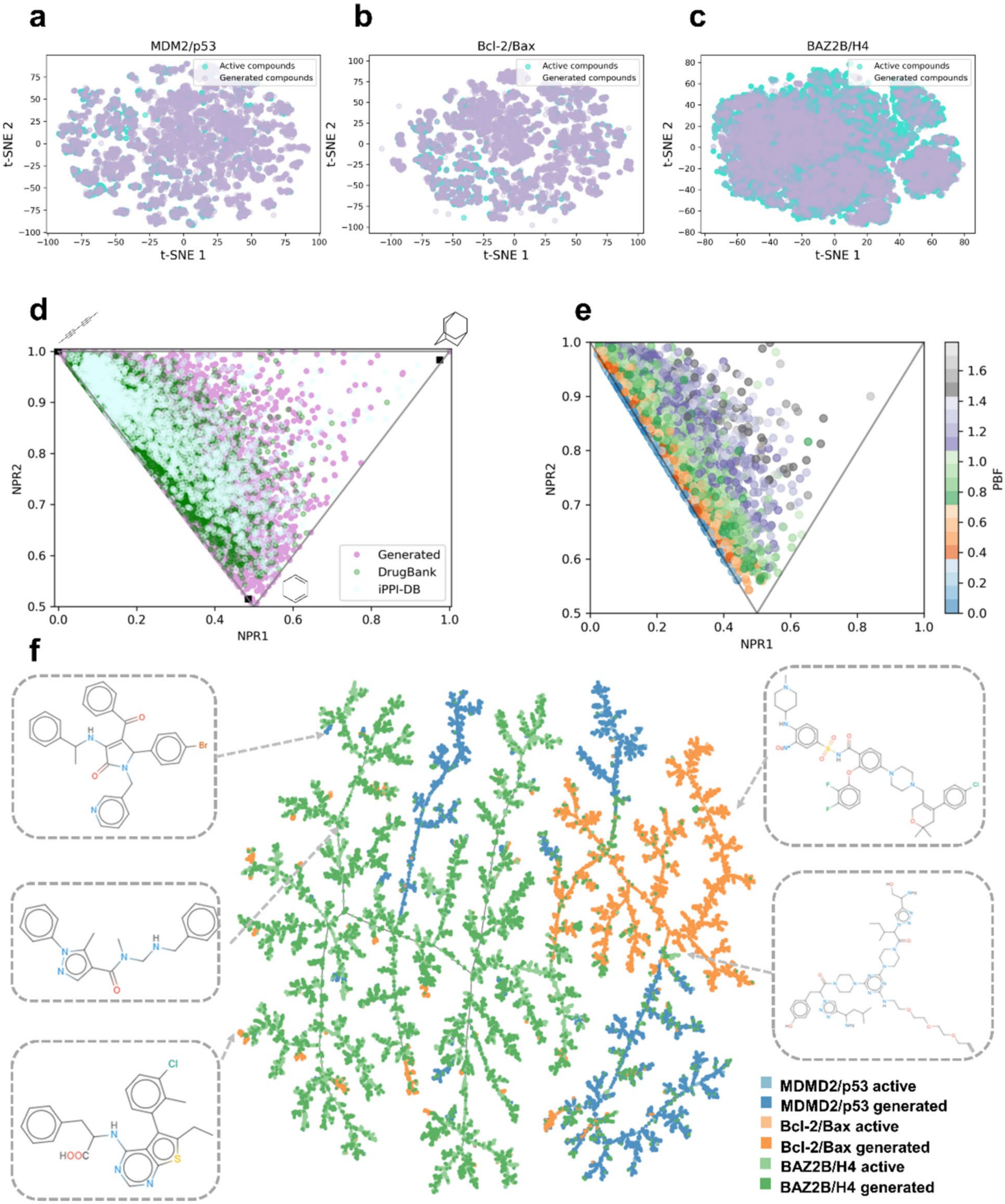

Fig. 4 Chemical space exploration. **a** t-SNE visualization of active and generated compounds for MDM2/p53. **b** t-SNE visualization of active and generated compounds for Bcl-2/Bax. **c** t-SNE visualization of active and generated compounds for BAZ2B/H4. **d** PMI ternary density plots of generated compounds, small molecule drugs from DrugBank, and iPPI-DB inhibitors. Top left: propyne, bottom: benzene, and the top right: adamantane. **e** Molecular three-dimensionality distribution of generated molecules visualized using NPR and PBF descriptors. **f** TMAP visualization of active and generated compounds for MDM2/p53, Bcl-2/Bax and BAZ2B/H4

was able to accurately capture the binding preferences of the target PPI interface based on limited input data. The generated compounds formed additional halogen bonds, salt bridges and π-cation interactions, which were not observed in the reference disruptor. These additional interactions likely contributed to the improved binding affinity to the target PPI interface (Fig. 5e). In conclusion, by analyzing the interaction patterns between the generated compounds and the PPI interface, GENiPPI successfully learned the implicit interaction rules between active compounds and the PPI interface. The ability to generate molecules that introduce novel interaction patterns, while retaining key interactions, demonstrates the model s capability to innovate within the chemical space and design compounds with enhanced binding potential.

Conclusion

In this work, we developed the GENiPPI framework, which combines PPI interface features with a conditional molecular generative model to generate novel modulators for PPI interfaces. Through extensive conditional evaluation experiments, we validated the ability of the GENiPPI framework to learn the implicit relationship between PPI interfaces and active molecules, demonstrating its capacity to generate chemically diverse and biologically relevant molecules. One of the key innovations of GENiPPI is its use of GATs to extract fine-grained, atomic-level interaction features from PPI interfaces. This allows the model to focus on critical interaction "hot spots" that are often difficult to target using conventional drug design methods. Additionally, by incorporating a conditional wGAN, the model is able to impose specific constraints on molecular generation, ensuring that the generated molecules are not only structurally novel but also align with the required PPI-targeting drug-likeness. Our comparative benchmarks and evaluation experiments across various settings demonstrate the practical potential of GENiPPI for rational PPIs drug design.

Despite these promising results, our framework has some limitations that can be addressed to improve its performance and applicability. First, the model has not been extensively tested across a large number of receptor-ligand pairs of PPIs, which may affect its generalization ability. This limitation arises from the relatively scarce data on drug-PPI target complexes compared to traditional drug-target dataset. The limited number of high-quality datasets for drug-PPI target complexes compared to traditional drug-target datasets poses a challenge. This scarcity of data remains a common issue in PPI-targeted drug discovery, largely due to the complexity of characterizing PPIs experimentally, and underscores the need for more comprehensive and curated datasets. Furthermore, the current framework does not incorporate the 3D structural information of ligand-receptor interactions in PPIs. Additionally, improvements can be made in representation learning, balancing training speed, and enhancing the diversity of generated molecules.

Several potential directions could further improve GENiPPI and its application to PPI-targeted drug discovery: (1) collecting and cleaning higher quality data pairs on receptor-ligand PPI complexes is essential. Improving the diversity and accuracy of the data used for model training can significantly enhance the model s ability to generalize to new, unseen targets. (2) integrating molecular chemical language models and pre-trained models of protein–protein structural features to fine-tune receptor-ligand PPI datasets, thereby enhancing model generalization, novelty and diversity of generated compounds. (3) The current framework could be further enhanced by integrating fragment-based molecular generative models with 3D structural information. (4) Modifying the model architecture or combining it with deep reinforcement learning to optimize the generated compounds [84, 85]. By defining specific objectives such as maximizing binding affinity or improving pharmacokinetic properties, reinforcement learning agents could iteratively refine the generated molecules to achieve more desirable drug-like characteristics.

In light of these potential improvements, our future work will focus on enhancing the GENiPPI framework by combining advanced representation learning methods and deep generative approaches.

In summary, the GENiPPI framework represents a significant advance in the field of PPI structure-based molecular generation. Its ability to integrate PPI interface features into a generative framework, combined with its performance in both few-shot learning and conditional evaluation experiments, highlights its potential as a powerful tool for rational PPI drug design.

(See figure on next page.)

Fig. 5 Few shot molecular generation analysis. **a** t-SNE visualization of the distribution of active and generated compounds for Hsp90/Cdc37. **b** Comparison of the pharmacophore of the generated molecules with the reference molecule(DCZ3112). **c** PPI interface region(green) of the Hsp90(palecyan)/Cdc37(lightpink) complex. **d** The complex structure of DCZ3112(green) and Hsp90(palecyan)-CDC37(lightpink) modeled by molecular docking (PDB ID: 1US7). **e** The binding poses of generated compounds(green) and Hsp90(palecyan)-CDC37(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

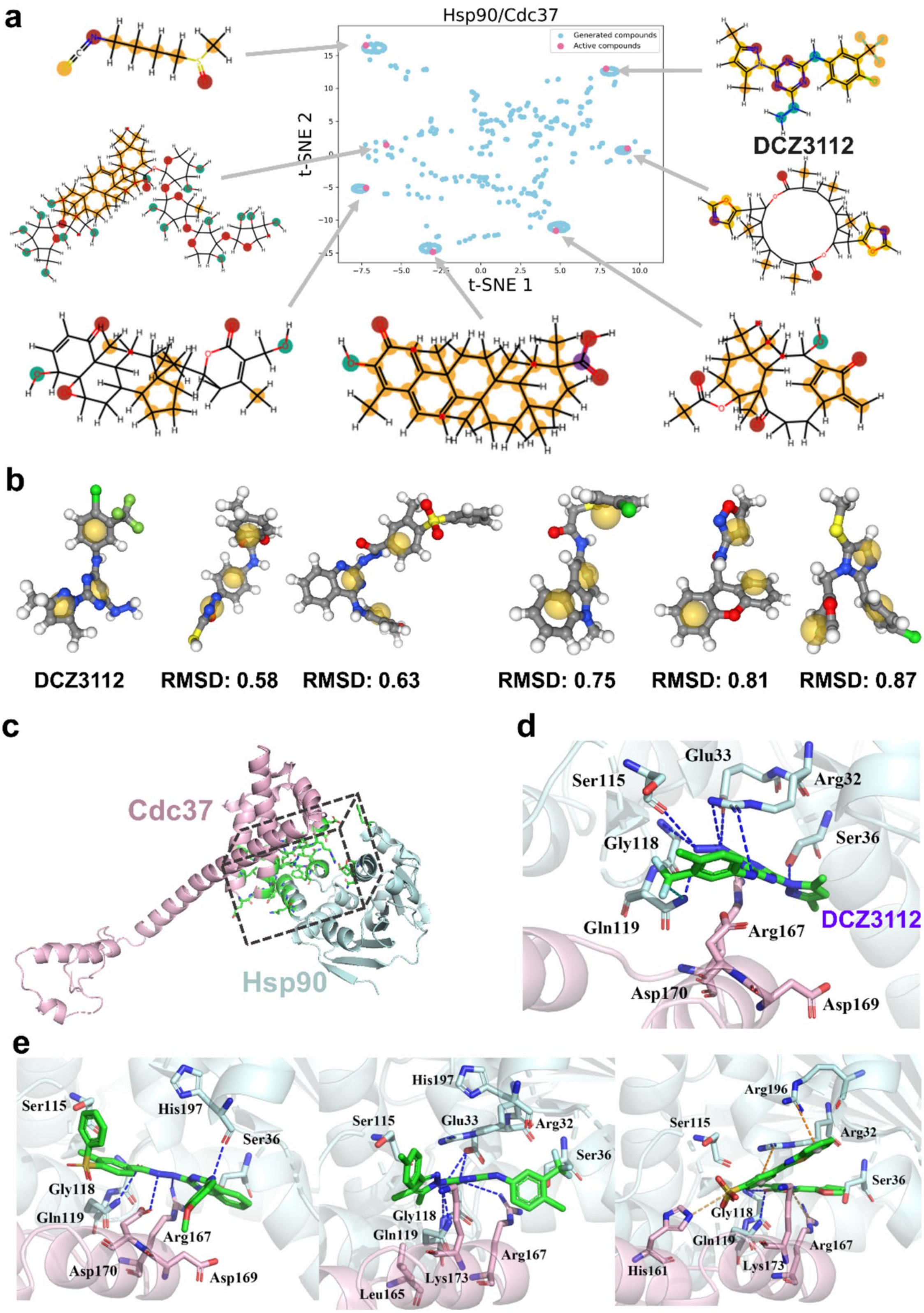

Fig. 5 (See legend on previous page.)

Methods

Datasets

We first investigated PPI targets with sufficient compound bioactivity data for training and evaluation of our model [86]. For this study, we selected 10 validated PPI drug targets that cover the binding interface (Supplementary Tables S1). These targets include 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. Additionally, we randomly selected a subset of 250,000 compounds as additional inactive compounds from the ChEMBL dataset that was used as part of the training datasets [87]. 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 complex interfaces is inspired by previous work on protein docking model evaluation [60], which introduced a double-graph representation to capture the interface features and interactions of protein–protein complexes (Supplementary Fig. 1). The extracted interface region is modeled as two graphs (G^1 and G^2), representing the interfacial information and the residues involved in the two interacting proteins. A graph G is 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, which maps the association between the nodes, numerically representing the connectivity of the graph. If the graph G has N nodes, the dimension of the adjacency matrix A 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 $\boldsymbol{G}^1$ encodes the atomic types of all residues in the interface region, and its adjacency matrix $\boldsymbol{A}^1$ classifies the interatomic bonding types for all residues at the interface region. This representation only considers the covalent bonds between atoms of interface residues within each subunit as edges. The definition follows as:

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

The graph $\boldsymbol{G}^2$ represents both covalent bonds (including those captured $\boldsymbol{G}^1$) and non-covalent residue interactions as edges. The adjacency matrix $\boldsymbol{A}^2$ for $\boldsymbol{G}^2$ accounts for both covalent bonds and non-covalent interactions between atoms that are within 10.0 Å of each other. The non-covalent atom pairs are defined as those whose distance is less than 10.0 Å. The definition follows as:

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

Here, $\boldsymbol{d_{ij}}$ represents the distance between the i-th and the j-th atoms of all residues in the interaction region. μ and σ are learnable, with initial values of 0.0 and 1.0, respectively. The function $\boldsymbol{e}^{-(\boldsymbol{d}_i\boldsymbol{j}-\frac{\boldsymbol{\mu})^2}{\boldsymbol{\sigma}}}$ decays as the distance between atoms increases.

The graph representation provides a flexible and intuitive way to encode interactive information and adjacent(local) relationships. For the node features, we considered the physicochemical properties of the atoms, using the same features as in previous work [60, 88, 89]. The initial feature vector of each node, with a length of

23, was then embedded into 140 features using a one-layer fully connected (FC) network.

The constructed graphs are used as the inputs for the Graph Attention Networks (GATs). Each 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, $\boldsymbol{x}^{in} = \{\boldsymbol{x}_1^{in}, \boldsymbol{x}_2^{in}, \cdots, \boldsymbol{x}_N^{in}\}$ 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, representing the relative importance between the i-th and j-th nodes:

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

Here, $\boldsymbol{x}_i^{'}$ and $\boldsymbol{x}_j^{'}$ are the transformed feature representations, defined as $\boldsymbol{x}_i^{'} = \boldsymbol{W}\boldsymbol{x}_i^{in}$ 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. To satisfy the symmetrical property of the graph, $\boldsymbol{e}_{ij}$ and $\boldsymbol{e}_{ji}$ become identical by adding $\boldsymbol{x}_i^T\boldsymbol{E}\boldsymbol{x}_j^T$ and $\boldsymbol{x}_i^T\boldsymbol{E}\boldsymbol{x}_i^{'}$. The attention coefficient will only be computed for i and j where $\boldsymbol{A}_{ij} > 0$.

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

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

Here, $\boldsymbol{a}_{ij}$ 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 $A_{ij} > 0$. The goal is to define attention by simultaneously considering both the physical structure $\boldsymbol{A}_{ij}$ and the normalized attention coefficient $\boldsymbol{e}_{ij}$ of the interactions.

Based on the attention mechanism, the new features of each node are updated by considering its neighboring nodes. This update is a linear combination of the neighboring node features and the final attention coefficient $\boldsymbol{a}_{ij}$:

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

Using the previously described GATs mechanism, we applied four layers of GATs to process the node embedding information from the neighboring nodes and output the updated node embedding. For the two adjacency matrices $\boldsymbol{A}^1$ and $\boldsymbol{A}^2$, we use a shared GAT. the initial input to the network consists of the atomic feature. With two matrices $\boldsymbol{A}^1$ and $\boldsymbol{A}^2$, we compute $\boldsymbol{x}_1 = \boldsymbol{GAT}\left(\boldsymbol{x}^{in}, \boldsymbol{A}^1\right)$ and $\boldsymbol{x}_2 = \boldsymbol{GAT}\left(\boldsymbol{x}^{in}, \boldsymbol{A}^2\right)$.

To focus exclusively 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 about intermolecular interactions from the other nodes at the protein–protein complex interface. The output node feature is therefore defined as:

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

Afterward, the updated x^{out} becomes x^{in} and iteratively passes through the subsequent three GAT layers to further increase the information. After all four GAT layers updated the node embeddings, the embedding of all nodes in the graph are summed to represent the overall intermolecular interaction of the protein–protein complex:

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

Finally, fully connected (FC) layers were applied to the $\boldsymbol{x}_{graph}$ to obtain a [4, 4, 4] features vector representing the protein–protein interface.

Molecular representation

For each SMILES string, a 3D conformer is generated using RDKit [90] and optimized with the default settings of the MMFF94 force field. The molecular structure is then extracted into a 35 Å grid centered at the geometric center of the molecule using the HTMD package [91]. The atoms are discretized into a 1 Å cubic grid, and eight channels are used to compute voxelized information. Finally, the electronic density for the 9th channel is calculated using the original molecule method in Multiwfn (Supplementary Fig. 2) [92].

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 [4, 4, 4, 9] and input into a 4-layer transposed convolutional neural network (CNN) with 256, 512, 1024, and 1024 filters, respectively. The first three layers downsample the array size using strided convolution (stride = 2). For all convolutions, a kernel size of 4 is used, and the Leaky ReLU is applied as the activation function after each convolution. BatchNorm3d is applied between the convolution and activation operations to normalize the values across each channel for each sample.

The discriminator consists of a 4-layer sequential convolutional neural network (CNN) with 256, 512, 1024, and 1024 filters, respectively. The first three layers downsample the array size using strided convolution (strided = 2). As with the generator,, a kernel size of 4 is used for all convolutions, and Leaky ReLU (α = 0.2) is applied as the activation. InstanceNorm3d is applied between the convolution and activation steps to normalize the values across each channel for each sample.

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

Molecular captioning network

In this section, we describe the process of decoding the generated molecular representation into SMILES strings. Our approach is inspired by shape-based molecular generation [93, 94], which utilize a combination of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks to generate SMILES strings. Briefly, the molecular captioning network consists of a 3D CNN and a recurrent LSTM networks. The molecular representation generated by the generator is first fed into the 3D CNN, and the output is then passed into the LSTM to decode the SMILES strings (Supplementary Fig. 4).

Model training

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

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

Here, x and ⌋ represent molecular representations and PPI interface features, respectively, sampled from the true data distribution $\mathrm{p}_{\mathrm{real}}$. The variable ‡ 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. The terms λ and α are regularization parameters, both empirically set to 10. The λ term controls the effect of the gradient penalty on discriminator loss, while the α term controls the influence of f_0 on the generator's loss.

The model was trained for 50,000 iterations with a batch size of 8 (65 steps per iteration). The discriminator was updated after each step, while the generator was updated every 30 steps. The network was trained using the RMSprop optimizer, with a learning rate of 1×10^{-4} for both 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 using binary cross-entropy loss and were 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 [95] and Tensorflow 2.5 [96].

Molecular generation

After training, the embedding information of the protein–protein complex interface is used to guide the model in generating novel molecules from the latent space. The maximum sampling strategy was applied in the LSTM, where the next token in the SMILES string is generated by selecting the one with the highest prediction probability [93].

Evaluation settings

Conditional evaluation metrics

In this study, the primary objective was to evaluate the effectiveness of the proposed framework for protein–protein interaction (PPI) interface-based conditional molecular generation. We sampled the same number of valid molecules for three PPI targets. For the generated compounds and comparison sets, we calculated the QED and Fsp3 values using RDKit, and the QEPPI values using the QEPPI package (https://github.com/ohuelab/QEPPI). The density distribution of these drug-likeness metrics was then plotted to compare the differences.

To assess the differences between generated and active compounds in terms of molecular properties (QED, QEPPI, and Fsp^3) across different PPI targets, we employed independent t-tests. The t-tests were used to determine whether there were statistically significant differences in the means of these properties between the two groups for each PPI target. For each comparison, the significance threshold was set at $p < 0.05$. When data met the normality assumption, independent t-tests were used. The t-tests provided insights into whether generated compounds successfully mimicked the molecular characteristics of active compounds, or if significant differences remained, particularly in drug-likeness and molecular complexity.

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, 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 [97].

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) [98] and the best-fit plane (PBF) [99]. The PMI descriptor classifies the geometric shape of molecules based on the degree to which they are rod-shaped (linear shape, such as acetylene), disk-shaped (planar shape, such as benzene), or sphere (spherical shape, such as adamantane). The normalized PMI ratios (NPRs) are plotted on a two-dimensional triangle to compare the shape space covered by different sets of molecules, allowing for the evaluation and visualization of the diversity of molecular shape with a given set. The PBF descriptor is a three-dimensional measure that represents the deviation of a molecule from a plane. It is defined as the mean distance of each heavy atom from the best-fit plane passing through all heavy atoms.

Tree MAP

To explore and visualize the chemical space through unsupervised visualization of high-dimensional data, we calculated MinHash fingerprint vectors for both active and generated compounds [100]. We then used tmap and faerun to construct two-dimensional projections using Tree MAP (TMAP) [101].

Protocol for few-shot generation

Targeting the Hsp90-Cdc37 PPI interface is recognized as an important strategy for cancer therapy. The crystal structure of the Hsp90-Cdc37 protein complex (PDB ID: 1US7) is available for molecular docking [102]. In addition, known Hsp90-Cdc37 PPI disruptors were collected for training in few-shot generative tasks. These disruptors include DCZ3112, Celastrol, FW-04–804, Sulforaphane, Withaferin A, Platycodin D, Kongensin A [103].

OpenPharmacophore(https://github.com/uibcdf/OpenPharmacophore) was utilized to create pharmacophore models and perform virtual screening. The protein structures were processed using UCSF Chimera [104], the program DOCK6.9 was used for semiflexible docking. Figures were generated using PyMOL [105]. A detailed docking protocol is provided in Supplementary Note B.

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
t-SNE	T-distributed stochastic neighbor embedding

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00930-0.

Supplementary material 1: Method details: Supplementary A, Dataset and preprocessing; Supplementary B, Docking protocol; Supplementary C, Model evaluation; Supplementary Table S1. Overview of the 10 PPI interface datasets used for training; Supplementary Figure S1. GATs module for representation learning of the protein complex interface; Supplementary Figure S2. CNNs module for molecular representation learning; Supplementary Figure S3. cWGAN module for conditional molecular generation; Supplementary Figure S4. molecular captioning network module for SMILES strings decoding; Supplementary Figure S5. Distribution of molecular properties.

Acknowledgements

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

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 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.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62425204, 62122025, U22A2037, 62450002, 62432011, 62272479, 62372469, 62202498), National Key Research and Development Project of China (2021YFA1000103, 2021YFA1000102), Taishan Scholarship (tstq20240506), and Shandong Provincial Natural Science Foundation (ZR2021QF023).

Data availability

The original data downloaded from ChEMBL and processed datasets are available on Zenodo (https://zenodo.org/records/13968592). All the source code and datasets are available at Github (https://github.com/AspirinCode/GENiPPI).

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea. [2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, Hunan, China. [3]School of Computer Science and Technology, China University of Petroleum, Qingdao 266580, Shandong, China. [4]School of Informatics, Yunnan Normal University, Kunming, China. [5]High Performance Computer Research Center, University of Chinese Academy of Sciences, Beijing 100190, China. [6]Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan. [7]College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China.

Received: 15 March 2024 Accepted: 11 November 2024
Published online: 20 December 2024

References

1. Stelzl U, Worm U, Lalowski M et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–968
2. Rual J-F, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437:1173–1178
3. Titeca K, Lemmens I, Tavernier J, Eyckerman S (2019) Discovering cellular protein-protein interactions: technological strategies and opportunities. Mass Spectrom Rev 38:79–111
4. Rhys GG, Cross JA, Dawson WM et al (2022) De novo designed peptides for cellular delivery and subcellular localisation. Nat Chem Biol 18:999–1004
5. Venkatesan K, Rual J-F, Vazquez A et al (2009) An empirical framework for binary interactome mapping. Nat Methods 6:83–90
6. Nero TL, Morton CJ, Holien JK et al (2014) Oncogenic protein interfaces: small molecules, big challenges. Nat Rev Cancer 14:248–262
7. Oughtred R, Rust J, Chang C et al (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 30:187–200
8. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature 450:1001–1009
9. Ivanov AA, Khuri FR, Fu H (2013) Targeting protein–protein interactions as an anticancer strategy. Trends Pharmacol Sci 34:393–400
10. Ashkenazi A, Fairbrother WJ, Leverson JD, Souers AJ (2017) From basic apoptosis discoveries to advanced selective BCL-2 family inhibitors. Nat Rev Drug Discov 16:273–284
11. Shin W-H, Christoffer CW, Kihara D (2017) In silico structure-based approaches to discover protein-protein interaction-targeting drugs. Methods 131:22–32
12. Shin W-H, Kumazawa K, Imai K et al (2020) Current challenges and opportunities in designing protein–protein interaction targeted drugs. Adv Appl Bioinf Chem 12:11–25
13. Anderson AC (2003) The process of structure-based drug design. Chem Biol 10:787–797
14. Wang X, Song K, Li L, Chen L (2018) Structure-based drug design strategies and challenges. Curr Top Med Chem 18:998–1006
15. Batool M, Ahmad B, Choi S (2019) A structure-based drug discovery paradigm. Int J Mol Sci 20:2783
16. Danel T, Łęski J, Podlewska S, Podolak IT (2022) Docking-based generative approaches in the search for new drug candidates. Drug Discov Today 28:103439
17. Isert C, Atz K, Schneider G (2023) Structure-based drug design with geometric deep learning. Curr Opin Struct Biol 79:102548
18. Rakers C, Bermudez M, Keller BG et al (2015) Computational close up on protein–protein interactions: how to unravel the invisible using molecular dynamics simulations? Wiley Interdiscip Rev Comput Mol Sci 5:345–359
19. Ni D, Lu S, Zhang J (2019) Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): a new paradigm for PPI drug discovery. Med Res Rev 39:2314–2342
20. Janin J, Chotia C (1990) The structure of protein-protein recognition sites. J Biol Chem 265:16027–16030
21. Smith MC, Gestwicki JE (2012) Features of protein–protein interactions that translate into potent inhibitors: topology, surface area and affinity. Expert Rev Mol Med 14:e16
22. Wang Z-Z, Shi X-X, Huang G-Y et al (2023) Fragment-based drug discovery supports drugging 'undruggable protein–protein interactions. Trends Biochem Sci. https://doi.org/10.1016/j.tibs.2023.01.008
23. Mignani S, Rodrigues J, Tomas H et al (2018) Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? Drug Discov Today 23:605–615
24. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) 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 23:3–25
25. Lipinski CA (2004) Lead-and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1:337–341
26. Morelli X, Bourgeas R, Roche P (2011) Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). Curr Opin Chem Biol 15:475–481
27. Bickerton GR, Paolini GV, Besnard J et al (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
28. Kosugi T, Ohue M (2021) Quantitative Estimate of Protein-Protein Interaction Targeting Drug-likeness. In: 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp 1–8
29. Kosugi T, Ohue M (2021) Quantitative estimate index for early-stage screening of compounds targeting protein-protein interactions. Int J Mol Sci 22:10925
30. Wang J, Mao J, Wang M et al (2023) Explore drug-like space with deep generative models. Methods. https://doi.org/10.1016/j.ymeth.2023.01.004
31. Qiu Y, Li X, He X et al (2020) Computational methods-guided design of modulators targeting protein-protein interactions (PPIs). Eur J Med Chem 207:112764
32. Stokel-Walker C, Van Noorden R (2023) What ChatGPT and generative AI mean for science. Nature 614:214–216
33. Urbina F, Lentzos F, Invernizzi C, Ekins S (2022) Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell 4:189–191
34. Bilodeau C, Jin W, Jaakkola T et al (2022) Generative models for molecular discovery: recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci 12:e1608
35. Cheng Y, Gong Y, Liu Y et al (2021) Molecular design in drug discovery: a comprehensive review of deep generative models. Brief Bioinform 22:bbab344
36. Tong X, Liu X, Tan X et al (2021) Generative models for De Novo drug design. J Med Chem 64:14011–14027
37. Wang M, Wang Z, Sun H et al (2022) Deep learning approaches for de novo drug design: an overview. Curr Opin Struct Biol 72:135–144
38. Meyers J, Fabian B, Brown N (2021) De novo molecular design and generative models. Drug Discov Today 26:2707

39. Thomas M, Bender A, de Graaf C (2023) Integrating structure-based approaches in generative molecular design. Curr Opin Struct Biol 79:102559
40. Zeng X, Wang F, Luo Y et al (2022) Deep generative molecular design reshapes drug discovery. Cell Rep Med 145:100794
41. Martinelli D (2022) Generative machine learning for de novo drug discovery: A systematic review. Comput Biol Med 105403
42. Özçelik R, van Tilborg D, Jiménez-Luna J, Grisoni F (2022) Structure-based drug discovery with deep learning. arXiv preprint arXiv:221213295
43. Ma B, Terayama K, Matsumoto S et al (2021) Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. J Chem Inf Model 61:3304–3313
44. Luo S, Guan J, Ma J, Peng J (2021) A 3D generative model for structure-based drug design. Adv Neural Inf Process Syst 34:6229–6239
45. Drotár P, Jamasb AR, Day B, et al (2021) Structure-aware generation of drug-like molecules. arXiv preprint arXiv:211104107
46. Li Y, Pei J, Lai L (2021) Structure-based de novo drug design using 3D deep generative models. Chem Sci 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, et al (2022) Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In: International Conference on Machine Learning. PMLR, pp 17644–17655
49. Wang M, Hsieh C-Y, Wang J et al (2022) Relation: a deep generative model for structure-based de novo drug design. J Med Chem 65:9478–9492
50. Chan L, Kumar R, Verdonk M, Poelking C (2022) A multilevel generative framework with hierarchical self-contrasting for bias control and transparency in structure-based ligand design. Nat Mach Intell. https://doi.org/10.1038/s42256-023-00712-7
51. Zhang O, Zhang J, Jin J, et al (2023) ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. Nat Mach Intell 1–11
52. Neugebauer A, Hartmann RW, Klein CD (2007) Prediction of protein–protein interaction inhibitors by chemoinformatics and machine learning methods. J Med Chem 50:4665–4668
53. Gupta P, Mohanty D (2021) 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. Brief Bioinform 22:bbab111
54. Díaz-Eufracio BI, Medina-Franco JL (2022) Machine learning models to predict protein-protein interaction inhibitors. Molecules 27:7986
55. Reker D, Schneider P, Schneider G (2016) Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. Chem Sci 7:3919–3927
56. Mallet V, Checa Ruano L, Moine Franel A et al (2022) InDeep: 3D fully convolutional neural networks to assist in silico drug design on protein–protein interactions. Bioinformatics 38:1261–1268
57. Wang J, Chu Y, Mao J et al (2022) De novo molecular design with deep molecular generative models for PPI inhibitors. Brief Bioinform. https://doi.org/10.1093/bib/bbac285
58. Scarselli F, Gori M, Tsoi AC et al (2008) The graph neural network model. IEEE Trans Neural Netw 20:61–80
59. Velickovic P, Cucurull G, Casanova A et al (2017) Graph attention networks. Stat 1050(10):48550
60. Wang X, Flannery ST, Kihara D (2021) Protein docking model evaluation by graph neural networks. Front Mol Biosci 8:647915
61. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:14111784
62. Wang Y, Wu S, Duan Y, Huang Y (2021) ResAtom system: protein and ligand affinity prediction model based on deep learning. arXiv preprint arXiv:210505125
63. Zapata PAM, Méndez-Lucio O, Le T et al (2023) Cell morphology-guided de novo hit design by conditioning GANs on phenotypic image features. Dig Dis 2:91
64. Wang F, Feng X, Guo X et al (2021) Improving de novo molecule generation by embedding LSTM and attention mechanism in CycleGAN. Front Genet 12:709500
65. Wei W, Cherukupalli S, Jing L et al (2020) Fsp3: a new parameter for drug-likeness. Drug Discov Today 25:1839–1845
66. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B et al (2020) Molecular sets (MOSES): a benchmarking platform for molecular generation models. Front Pharmacol 11:1931
67. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13:1445–1454
68. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 4:120–131
69. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
70. Prykhodko O, Johansson SV, Kotsias P-C et al (2019) A de novo molecular generation method using latent vector based generative adversarial network. J Cheminform 11:1–13
71. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, et al (2017) Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint arXiv: 170510843
72. Torchet R, Druart K, Ruano LC et al (2021) The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators. Bioinformatics 37:89–96
73. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
74. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learning Res 9:11
75. Hinton GE, Roweis S (2002) Stochastic neighbor embedding. Adv Neural Inf Process Syst 15:963
76. Meyers J, Carter M, Mok NY, Brown N (2016) On the origins of three-dimensionality in drug-like molecules. Future Med Chem 8:1753–1767
77. Probst D, Reymond J-L (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform 12:1–13
78. Altae-Tran H, Ramsundar B, Pappu AS, Pande V (2017) Low data drug discovery with one-shot learning. ACS Cent Sci 3:283–293
79. Moret M, Friedrich L, Grisoni F et al (2020) Generative molecular design in low data regimes. Nat Mach Intell 2:171–180
80. Wang J, Zheng S, Chen J, Yang Y (2021) Meta learning for low-resource molecular optimization. J Chem Inf Model 61:1627–1636
81. Chen X, Liu P, Wang Q et al (2018) DCZ3112, a novel Hsp90 inhibitor, exerts potent antitumor activity against HER2-positive breast cancer through disruption of Hsp90-Cdc37 interaction. Cancer Lett 434:70–80
82. Wang L, Zhang L, Li L et al (2019) Small-molecule inhibitor targeting the Hsp90-Cdc37 protein-protein interaction in colorectal cancer. Sci Adv 5:eaax2277
83. Allen WJ, Balius TE, Mukherjee S et al (2015) DOCK 6: Impact of new features and current docking performance. J Comput Chem 36:1132–1156
84. Sun H, Wang J, Wu H et al (2023) A multimodal deep learning framework for predicting PPI-modulator interactions. J Chem Inf Model 111:197
85. Shen L, Feng H, Qiu Y, Wei GW (2023) SVSBI: sequence-based virtual screening of biomolecular interactions. Commun Biol 6:1–12. https://doi.org/10.1038/s42003-023-04866-3
86. Singh N, Chaput L, Villoutreix BO (2020) Fast rescoring protocols to improve the performance of structure-based virtual screening performed on protein–protein interfaces. J Chem Inf Model 60:3910–3934
87. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
88. Torng W, Altman RB (2019) Graph convolutional neural networks for predicting drug-target interactions. J Chem Inf Model 59:4131–4149
89. Lim J, Ryu S, Park K et al (2019) Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. J Chem Inf Model 59:3981–3988
90. Landrum G (2006) RDKit: Open-source cheminformatics. https://www.rdkit.org
91. Doerr S, Harvey MJ, Noé F, De Fabritiis G (2016) HTMD: high-throughput molecular dynamics for molecular discovery. J Chem Theory Comput 12:1845–1852
92. Lu T, Chen F (2012) Multiwfn: a multifunctional wavefunction analyzer. J Comput Chem 33:580–592

93. Skalic M, Jiménez J, Sabbadin D, De Fabritiis G (2019) Shape-based generative modeling for de novo drug design. J Chem Inf Model 59:1205–1214
94. Gaulton A, Kale N, van Westen GJP et al (2015) A large-scale crop protection bioassay data set. Sci Data 2:150032
95. Paszke A, Gross S, Massa F et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:1
96. Abadi M, Agarwal A, Barham P, et al (2015) TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow. org (2015).
97. Preuer K, Renz P, Unterthiner T et al (2018) Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. J Chem Inf Model 58:1736–1741
98. Sauer WHB, Schwarz MK (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. J Chem Inf Comput Sci 43:987–1003
99. Firth NC, Brown N, Blagg J (2012) Plane of best fit: a novel method to characterize the three-dimensionality of molecules. J Chem Inf Model 52:2516–2525
100. Probst D, Reymond J-L (2018) A probabilistic molecular fingerprint for big data settings. J Cheminf 10:1–12
101. Probst D, Reymond J-L (2018) FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. Bioinformatics 34:1433–1435
102. Roe SM, Ali MMU, Meyer P et al (2004) The mechanism of Hsp90 regulation by the protein kinase-specific cochaperone p50cdc37. Cell 116:87–98
103. Dike PP, Bhowmick S, Eldesoky GE et al (2022) In silico identification of small molecule modulators for disruption of Hsp90–Cdc37 protein–protein interaction interface for cancer therapeutic application. J Biomol Struct Dyn 40:2082–2098
104. Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
105. Schrödinger LLC (2015) The PyMOL molecular graphics system, version 1.8

Publisher's Note