De novo molecular design with deep molecular generative models for PPI inhibitors

Jianmin Wang[1], Yanyi Chu, Jiashun Mao, Hyeon-Nae Jeon, Haiyan Jin, Amir Zeb, Yuil Jang, Kwang-Hwi Cho, Tao Song, Kyoung Tai No[1]

[1]The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University,
Incheon 21983, Republic of Korea

Jianmin Wang: https://jianmin2drugai.github.io/ E-mail: drugai@hotmail.com

Volume 23 • Number 4 • July 2022
Briefings in Bioinformatics
https://academic.oup.com/bib
OXFORD

Introduction

Molecular design strategies are essential and revolutionized the therapeutic advances in drug discovery. In parallel, computational methods for de novo molecular design have been developed over the past three decades, which is recently accelerated by the incorporation of artificial intelligence (AI). To this end, deep generative models are gaining interest in a variety of AI approaches. Researchers are applying deep generative models to accelerate efficient, effective and selective drug designing avenues.

Protein–protein interactions (PPIs) play a vital role in a diverse range of biological processes and are therefore critical for the development of human health and disease states. The research shows that abnormal PPIs are involved in a wide spectrum of diseases, including cancer, infectious diseases and neurodegenerative diseases. Due to their critical impact, PPIs have been considered promising drug targets of therapeutic interest. However, previous attempts to target PPIs have faced serious challenges because of their general properties, such as flat surfaces, featureless conformations, complex topologies and shallow pockets. Significant progress has been made in the design of conventional computer-guided PPIs inhibitors, and molecular generative models have been rapidly developed in recent years. Unfortunately, to date, no deep generative model-based approach has been applied to the design of PPIs inhibitors. Furthermore, quantitative estimation of drug-likeness (QED) similarity is commonly used to assess quantitative drug similarity, but not for the evaluation of compounds targeting PPI. Kosugi and Ohue developed a quantitative estimation index for compounds targeting PPI (QEPPI) specifically for the evaluation of PPI targeting compounds. QEPPI is an extension of the QED method for PPI-targeted drugs, developed using the QED concept, involving the modeling of physicochemical properties based on information available for approved drugs. QEPPI models the physicochemical properties of compounds reported in the literature to act for PPI, and the results show that QEPPI is more suitable than QED for quantifying drug similarity for early PPI drug discovery. PPI inhibitors harbor two essential molecular characteristics: molecular shape and aromatic bonding, so we wanted to use 3D features such as molecular shape as input into the model. Although the shape-based VAE molecular generation model enables the design of novel compounds with desired properties, it cannot generate a library of compounds with diversity. To accelerate the process of PPI-targeted compound design, we introduced a GAN model to improve the shape-based VAE molecule generation model for molecular generation of PPI-targeted inhibitors.

In this study, we construct a PPI targeted drug-likeness dataset and propose a deep molecular generative framework to generate new PPI-targeted drug-likeness molecules from the features of seed compounds. The research explores for the first time the de novo molecule design of molecular generation models for PPI inhibitors. Our model exhibited comparable performance to various state-of-the-art molecule generative models. This is the first time that QEPPI was applied to the PPI drug-likeness assessment of the generated molecules in the molecular generation model. The results show that the generated molecules have better PPI-targeted drug-likeness and drug-likeness. The generated molecules share chemical space with iPPI-DB inhibitors as demonstrated by chemical space analysis and the generated novel molecules bridge the gap to extend it. Peptide characterization-oriented design of PPI inhibitors and the ligand-based design of PPI inhibitors are explored.

Data and Code Availability

GitHub : https://github.com/AspirinCode/iPPIGAN

DRUGAI

Methodology

Model architecture

The deep molecular generation model pipeline for PPI inhibitors is inspired by shape-based compound generation, which consists of two main steps: (i) GANs using 3D convolutional neural networks (CNNs) to capture molecular representation and (ii) a combination of CNNs and long short-term memory (LSTM) networks parsed SMILES strings from molecular representations. A schematic of the process is shown in Figure 1. The molecular shape and pharmacophore representations are used as inputs to the GANs, followed by a caption network that decodes the molecular shape and pharmacophore representations into SMILES strings, which generate molecules that matches the ligand representation.

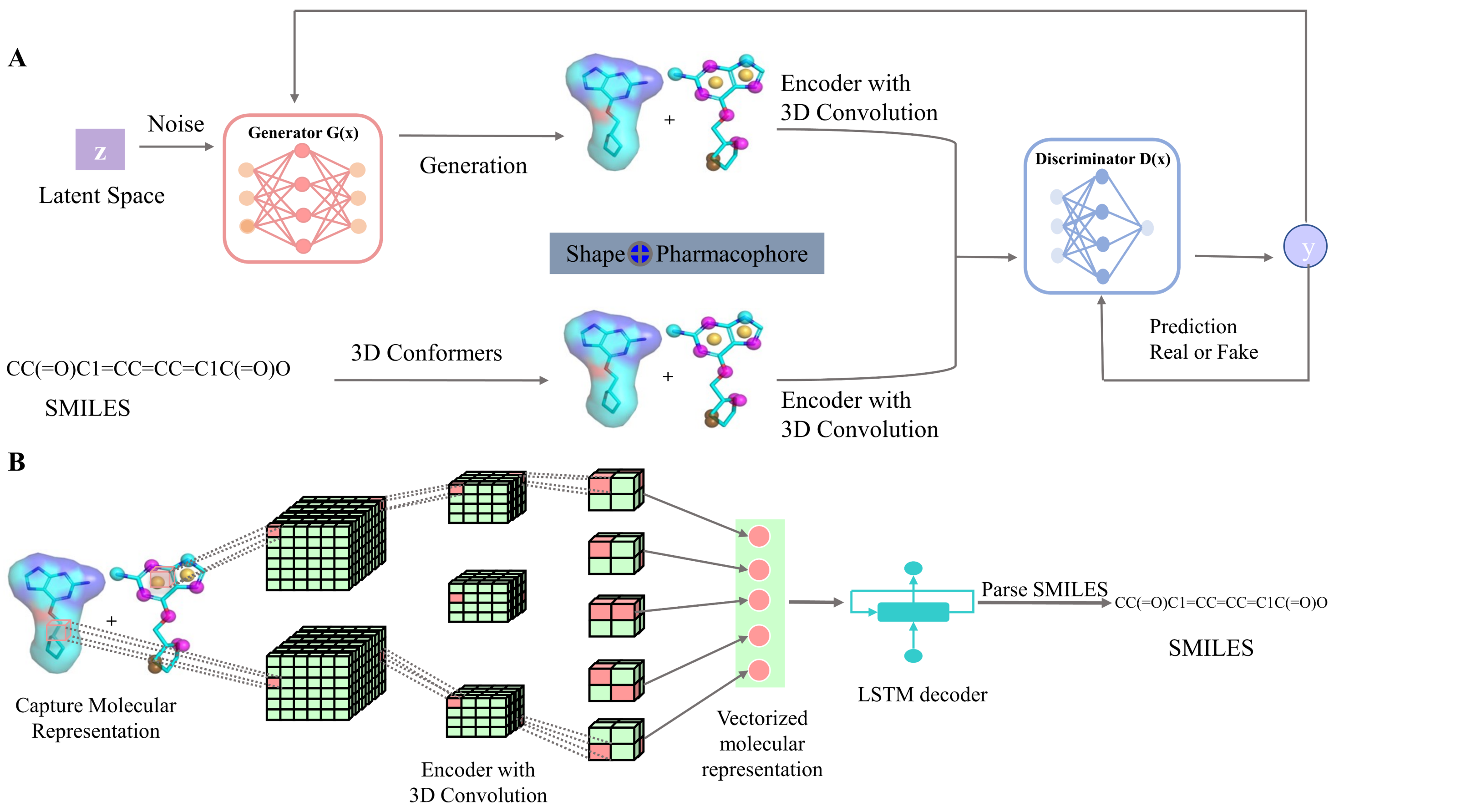


Figure 1. iPPIGAN pipeline. (A) First, each molecule is voxelized. Second, the molecule's voxel array and pharmacophore characteristics are entered into the generator network along with the latent code z to generate the molecular representations. The generated compounds can be modified by changing the latent code z. (B) Generated molecular shapes then processed by a shape-captioning network, outputting a SMILES string.

Evaluation metrics

To provide insight into the model performance and identify potential strengths and weaknesses compared with other generative models, the method was benchmarked on the MOSES platform. MOSES provide three basic metrics to evaluate the quality of generated molecules: validity, uniqueness and novelty. Validity is the fraction of valid molecules among generated molecules. Uniqueness is the fraction of unique molecules among valid molecules. Novelty is the fraction of unique valid generated molecules not present in the training set.

Results

Model performance

To provide insight into the model performance and identify potential strengths and limitations compared with other generative models, the method was benchmarked on the MOSES platform.

Following the proposed pipeline, models were trained on 300 000 preprocessed leads from the training data set and tested on 10 000 test and scaffold split test sets. We used the models and hyperparameters available in the MOSES platform, such as the adversarial autoencoder (AAE), a VAE, a character-level recurrent neural network (CharRNN) and a latent vector-based generative adversarial network (LatentGAN) . We evaluated the model by sampling 30 000 SMILES at a time in five independent runs.

Table 1. Valid, unique, novelty and FCD (mean ± SD) of sampling SMILES after training. We sampled 30 000 SMILES each time (in five independent runs)

Model	Valid	Unique@1 k	Unique@10 k	Novelty	FCD	
					Test	Test SF
AAE	0.881 ± 0.032	1.000 ± 0.114	0.990 ± 0.103	0.732 ± 0.001	2.018 ± 0.010	8.68 ± 0.028
CharRNN	0.987 ± 0.025	0.998 ± 0.003	0.990 ± 0.012	0.998 ± 0.008	8.727 ± 0.022	9.084 ± 0.029
VAE	0.858 ± 0.006	0.998 ± 0.001	0.995 ± 0.003	0.998 ± 0.007	8.007 ± 0.036	8.495 ± 0.054
LatentGAN	0.737 ± 0.002	1.000 ± 0.000	0.999 ± 0.004	0.999 ± 0.003	8.195 ± 0.023	8.675 ± 0.037
iPPIGAN	0.989 ± 0.005	1.000 ± 0.000	0.999 ± 0.006	0.990 ± 0.005	5.879 ± 0.034	6.171 ± 0.028

Results

Distribution of properties of the generated molecules

Nevertheless, in the realms of drug design and drug screening, n-octanol/water partition coefficient (LogP), synthetic accessibility score (SA score), natural-product-likeness (NP-likeness) and QED play a fundamental role. But the QEPPI is more suitable than QED for quantitative estimation of PPI-targeted compounds. We describe the properties' distribution of iPPI-DB inhibitors in Figure S1 (see Supplementary Data available online at http://bib.oxfordjournals.org/); the mean values of QED and QEPPI of iPPI-DB inhibitors were 0.43 and 0.61, respectively.

We compared the six molecular properties' distribution in the test set, the iPPI-DB inhibitors and the generated sets of AAE, CharRNN, VAE, LatentGAN and iPPIGAN. As shown in Figure 2, the properties' distribution of molecules generated by iPPIGAN is close to the properties' distribution of the test set. The properties distribution of the molecules generated by the iPPIGAN model is distinct from that of the iPPI-DB inhibitors, mainly because the properties' distribution of the training dataset is dissimilar to that of the iPPI-DB inhibitors. Moreover, the iPPIGAN generated molecules with higher QED values, higher QEPPI values and lower SA scores than other models. As shown in Figure 2, the results demonstrate that the iPPIGAN generates molecules that are easy to synthesize and have better drug-likeness and PPI-targeted drug-likeness properties.

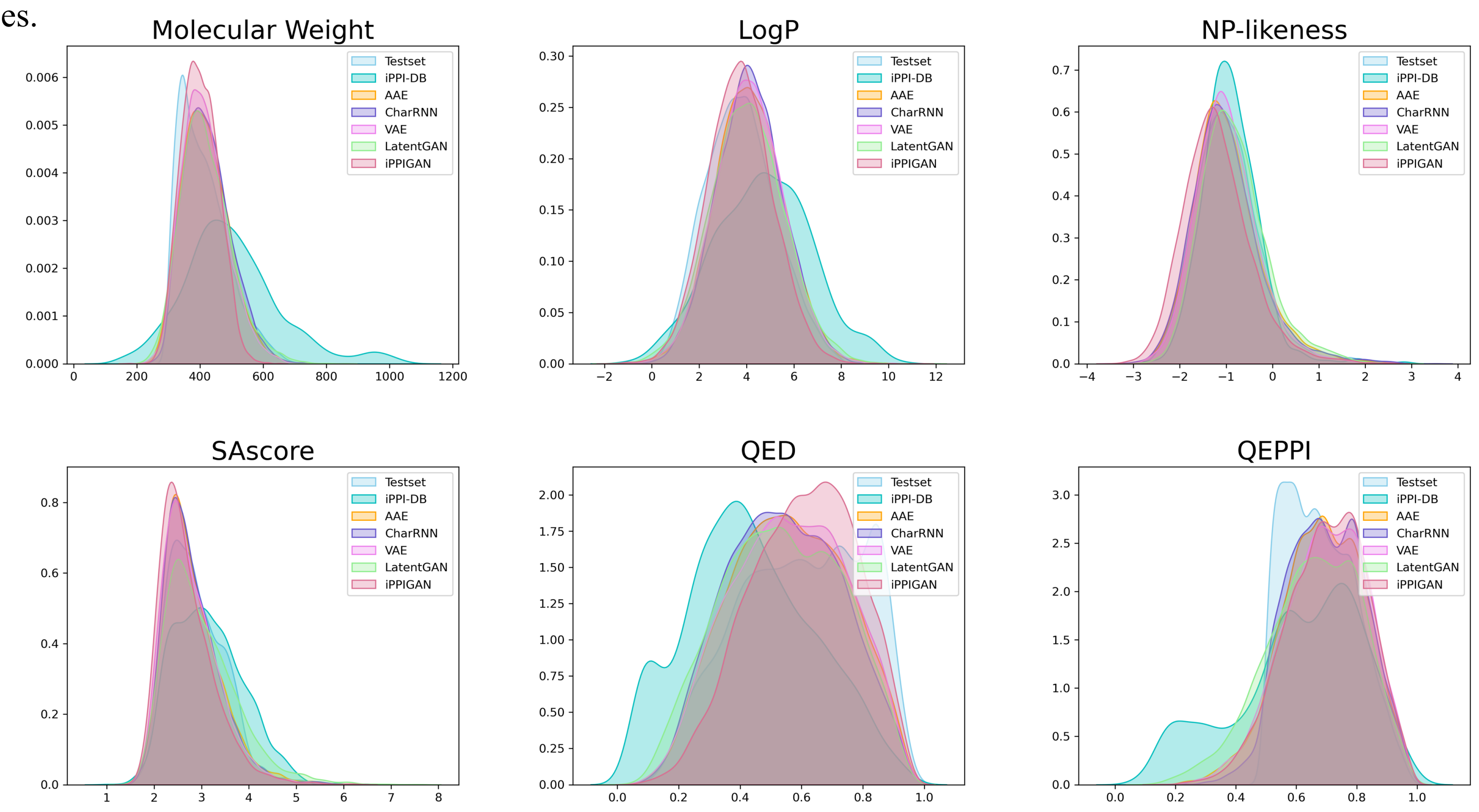
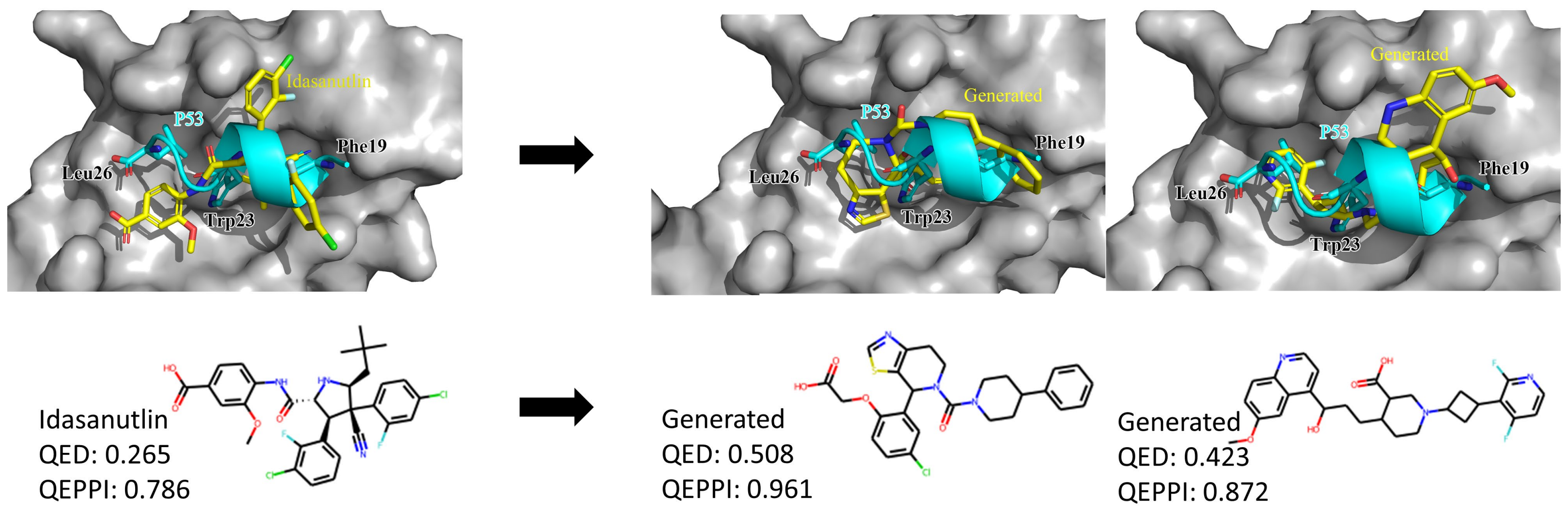


Figure 2. Distributions of properties for the test set, the iPPI-DB inhibitors and the generated compounds. Properties include molecular weight, lipophilicity (LogP), natural product-likeness (NP-likeness), SA score, QED and QEPPI.

A case of peptide-based generate molecules

To evaluate whether peptide-based generated compounds could be potent molecular candidates, we selected p53(peptide)-based generate potential candidate inhibitors of MDM2-p53 targets. A LightGBM regression model was then used to predict whether the generated molecules were biologically active against the MDM2-p53 protein-protein interaction target. In addition, compounds were filtered by QEDPPI values >0.5 and SAscore values <4. To further prioritize molecules for bioassay, we used Idasanutlin (RG-7388) a potent selective p53-MDM2 inhibitor as a reference compound, during the molecular docking by DOCK6.8 to predict binding affinity and the results were plotted by Pymol. The results shown in Figure, are the docking poses of Idasanutlin and the generated hit compounds which suitably occupied the binding site of p53 to MDM2. Furthermore, the generated candidate hits showed better binding affinity, drug-likeness and PPI drug-likeness than Idasanutlin.



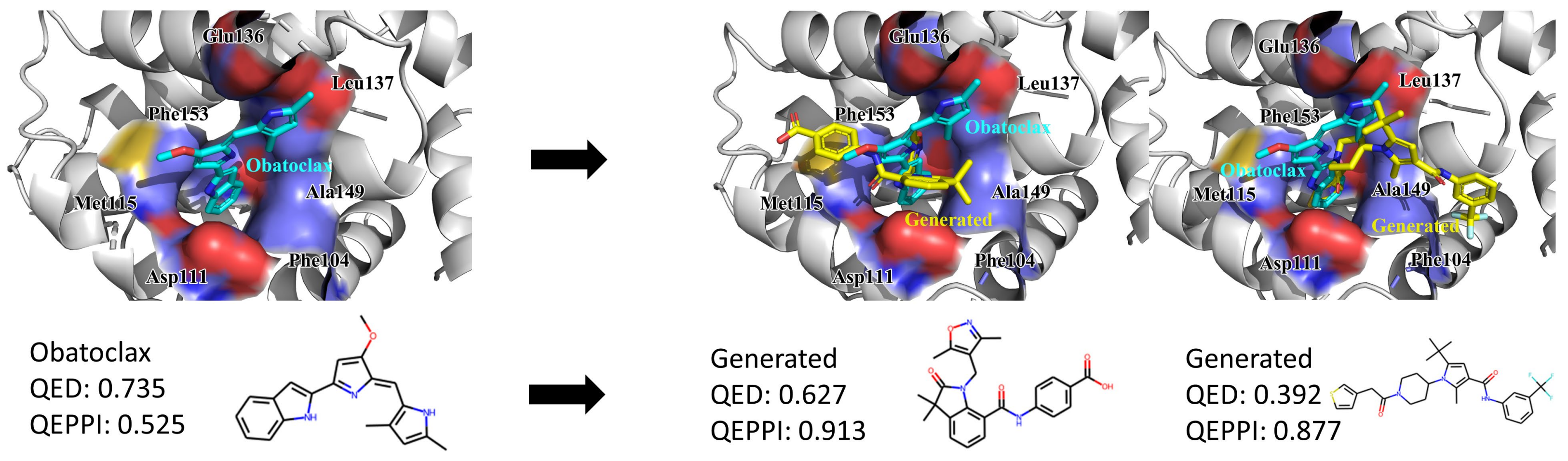
Idasanutlin
QED: 0.265
QEPPI: 0.786

Generated
QED: 0.508
QEPPI: 0.961

Generated
QED: 0.423
QEPPI: 0.872

A case of ligand-based generate molecules

In order to evaluate whether compounds generated on the ligand-based model could be good molecular candidates, we selected obatoclax-based generate potential candidate inhibitor of Bcl2 targets. A LightGBM regression model was then used to estimate whether the generated molecules were biologically active against the Bcl2 target. In addition, compounds were filtered by QEPPI values >0.5 and SAscore values <4. To further prioritize molecules for bioassay, we used Obatoclax a selective Bcl-2 inhibitor as a reference compound, and the binding affinity was predicted by docking which was performed by DOCK6.8 and plotted with Pymol. Figure shows the docked poses of the generated hit compounds. Our findings suggested strong binding of the hit compounds over the Obatoclax in the Bcl2 binding site and high drug-likeness and PPI drug-likeness scores.



Obatoclax
QED: 0.735
QEPPI: 0.525

Generated
QED: 0.627
QEPPI: 0.913

Generated
QED: 0.392
QEPPI: 0.877

Conclusion

In this work, we have constructed a PPI targeted drug-likeness dataset and have developed a novel shape-based framework for generating novel and potent drug-likeness molecules to target PPI. Our strategy exploits GANs and caption networks to enterprise diverse molecules targeting PPI from the 3D features of seed molecules. This method relies on the spatial orientation of a molecule or peptide as a seed molecule that can be variably designed for multiple molecules starting from a single 3D characterization. To our knowledge, this is the first time that a deep molecular generation model has been applied to the de novo design of PPI inhibitors. In addition, we applied QEPPI for the first time as an evaluation metrics for molecular generation models for de novo molecular design of PPI-targeted compounds. Our model shows performance comparable to several other state-of-the-art molecular generation models. The chemical space analysis demonstrates that the generated molecules share a similar chemical space with iPPI-DB inhibitors. We explored the peptide-based design of PPI inhibitors and the ligand-based design of PPI inhibitors. Our results show that the generated molecules have better PPI-targeted drug-likeness and drug-likeness. PPIs are pervasive in life, and their study and understanding are critical to drug discovery and bioengineering efforts. The molecular generative models of PPI inhibitors are still only a small step forward.

Acknowledgements

We thank anonymous reviewers for valuable suggestions.

Jianmin Wang, Yanyi Chu, Jiashun Mao, Hyeon-Nae Jeon, Haiyan Jin, Amir Zeb, Yuil Jang, Kwang-Hwi Cho, Tao Song, Kyoung Tai No, De novo molecular design with deep molecular generative models for PPI inhibitors, Briefings in Bioinformatics, Volume 23, Issue 4, July 2022, bbac285, https://doi.org/10.1093/bib/bbac285