MEDS Clinical Medicine (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/medsc.2023.040809
ISSN 2616-1907 Vol. 4 Num. 8

Molecular Generator for Multi-objective Optimization Based on the Pareto Algorithm

Dawei Feng[1,*]

[1]*School of Pharmacy, Yantai University, Yantai, 264005, China*
davidvon1998haha@163.com
[*]*Corresponding author*

Keywords: Molecular generator, Multi-objective optimization, Pareto algorithm

Abstract: Designing molecules with certain physicochemical features can promote the discovery and optimization of lead compounds. However, most molecular generation models optimize only one physicochemical property, which is not sufficient to determine the availability of a drug. This is because the availability of a drug substance molecule depends on the combined effect of many physicochemical properties. In this study, the pareto method was conducted to optimize the compounds for multi-target molecular characteristics in close approximation to those of the reference compound. In addition, we similarly used the random SMILES method involving amplification and diversification of molecules. Finally, we further examined the generation ability of the model and also analyzed the probability distribution of the physicochemical attributes and molecular structure of the created compounds. We expect that the model could develop additional molecules for exploring a bigger chemical space for medicinal chemists.

1. Introduction

Discovery and optimization of lead compounds is crucial for improving the success rate of drug research and development (R&D), in which artificial intelligence using deep learning methods can accelerate the discovery of novel compounds with desired properties. The ultimate goal of new drug development is to obtain Lead compound molecules with corresponding physiological activities and ideal physical and chemical properties.[1] And based on the type of input data, these models can be categorized into two main types with graphs as inputs and SMILES as inputs. Nowadays, the most widely utilized deep learning models to produce optimal lead chemical molecules are those that rely on variational auto-decoders (VAE) and recurrent neural networks (RNN), as well as their modifications.[2] These models can learn about databases composed of chemical information through to generate new valid SMILES that are not included in the original dataset.[3] However, RNN-based models have certain drawbacks: due to the structural issues of its model itself, problems such as gradient vanishing and gradient explosion are prone to occur when facing large datasets, and the final output is highly correlated with the last few inputs.[4] So we focused on VAE-based models. The models based on VAE reduct dimensionality to learn latent representations from input data .Therefore they can learn the probability distribution of the dataset and Reproduce the input data as much as possible.[5] CVAE, as a variant of VAE, introduces the concept of conditional

vectors after transforming the input data into potential vectors. After decoding by the decoder, compound molecules with target ideal properties can be generated.[6] For example,Lim et al[7] proposed a model based on CVAE which can generate molecules with ideal properties like Aspirin. These properties are not independent of each other, they always have a certain curve relationship with each other.[8] However, traditional CVAE models tend to lean towards a particular parameter being particularly excellent, without considering the issue of balancing multiple properties. Giuseppe Lamanna et al.[9] incorporated the results of molecular docking as a parameter into the genetic algorithm, resulting in the generation of compound molecules with better scores. They used a genetic algorithm that does balance the problem of one of the parameters to be optimized being too far from the desired value, but calculating the molecular docking scores as a parameter is computationally intensive and the scores do not necessarily match the actual without a single-crystal diffraction conformation comparison. Depending on the type of input, we can additionally categorize the models into those with sequence inputs and those with graphs as inputs. Both models have their merits and weaknesses, but in general, models employing sequences as inputs tend to perform better when huge datasets or complicated chemicals are involved. As part of our team's previous research, we attempted to use molecular graph as input and modeled the configuration of chiral atoms in molecules based on CVAE.[10] Jonghuan Choi et al[11] also mentioned that molecules generated by two-dimensional or three-dimensional input training are not necessarily more consistent with expectations than those generated by SMILES as input training. In this study, we introduced the pareto algorithm to optimize the model. And we used SMILES for data input with the goal of creating compound molecules with better balanced and great properties for diverse physical and chemical attributes. The model was constructed on the CVAE with long short-term memory (LSTM) for the purpose of generating unique compounds with desired properties. This can help us effectively solve the problem of gradient explosion. To ensure the high quality and diversity of data in the database. The randomized SMILES method was used for converting every compound to different SMILES, thus expanding the chemical space, and increasing the diversity of generated molecules.[12] In this model, three parameters, including tPSA, MW and LogP, were used as the controlling labels for generating compounds. Finally, we calculated the novelty and diversity of the generated compounds, and the affinity of compounds for the targets were validated by molecular docking.[13]

2. Materials and Method

2.1 Dataset

Enamine-real is a molecular database containing a huge number of compounds with good drug-likeness properties and is highly searchable. With this huge base of molecular databases, we were able to obtain the molecular structures of a large number of compounds that satisfy the physicochemical properties demand. In this study,a total of 28,896,138 compounds validated by high-throughput screening were collected from Enamine-real 2020[14]. Compounds with metal atoms, isotopes, salts, less than 10 heavy atoms, and unreadable atoms were eliminated by RDkit[15], leaving a total of 28,890,126 compounds at last.

Histone methyltransferase EZH2, the catalytic subunit of polycomb repressive complex 2(PCR2), is recurrently highly expressed in numerous cancers.[16] Several EZH2 inhibitors (EZH2i) have been developed. For the treatment of epithelioid sarcoma and follicular lymphoma,GSK2816126 showed high potency targeted EZH2 (IC50 =12.9 nM) by a hydrogen bond network with Tyr111, Trp624, and Arg685.[17] The data shows that indicating that GSK2816126 [18] is a more promising template for further study. Therefore, we chose GSK2816126 as the reference molecule in this study According to research, extended-connectivity fingerprint 4 (ECFP4) can better characterize the

similarity between two compounds when considering numerous datasets.[19] Compounds with a tanimoto similarity score > 0.3 based on ECFP4 with GSK2816126 were gathered, containing 12,091,652 compounds.

2.2 Pareto algorithm

Every single property in the condition vector c can be efficiently controlled by CVAE, leading to compounds with this attribute that come close to the desired value. The model typically creates one or more preset physicochemical parameters and also sets the structure of the target compound (SMILES format). The similarity to its structure is also used as a criterion for optimization. However, if the parameter values differ significantly from the real physical and chemical properties of the target compounds, one of the parameters of the generated molecule may deviate from the set value to a large extent. To solve this question and develop more compounds with balanced and ideal features, the pareto algorithm [20] was utilized in this study to optimize multi-objective parameters.

The pareto algorithm is a technique used in genetic algorithms for data screening. Between the parameters of chemicals in the database and the predetermined goal values, a loss value is computed. In this study, the MW, tPSA, and LogP balance of three physicochemical properties were optimized using the Pareto method. First, a compound is categorized as the non-dominated solution if there are little variances between it and the target property values. The procedure is repeated until the predetermined threshold is reached for the number of non-dominated solutions. We refer to the remaining compounds as the dominated solutions. The pareto algorithm produced 96,813 non-dominated solutions in total. However, it is challenging to produce molecules with the appropriate features because there aren't as many non-dominated solutions. The pareto technique was used to further identify 100,000 secondary non-dominated solutions (i.e., the non-dominated solutions that meets the conditions is selected again in the set of dominated solutions that were left over the first time) from the dominated set.

2.3 CVAE model

In this study, the conditional variational autoencoder (CVAE) model was used for molecular generation.[21] The model can generate similar but not identical compounds by adding random noise meanwhile considering the physicochemical properties and structure of the training compounds. The formula is described as follows:

$$Loss = E[\log P(X \mid z, c)] - D_{KL}[Q(z \mid X, c) \,||\, P(z \mid c)] \quad (1)$$

Equation (1) represents the loss function of our model, which consists of two parts, one is the difference between the generated features after the model has been trained and the features fed into the neural network, and the other part is the KL divergence. Where E means expectation, P and Q represent the probability distribution: Q represents the distribution learned by the encoder from the input data and condition vectors, while P represents the data distribution learned by the decoder from the latent vector z and the corresponding condition vector c. respectively. X refers to the SMILES of compounds in the database. The latent vector z represents the differential distribution compared with the reference compound, which is used to generate similar but not identical compounds. D_{KL} is the Kullback-Leibler (KL) divergence. The conditional vector c considers three physicochemical properties, including the lipid and water partition coefficient (LogP), molecular weight (MW), and topological polar surface area (tPSA). Therefore, CVAE can generate compounds controlled by the conditional vector.[22]

The model is mainly composed of the encoder and decoder.[23] The SMILES of compounds X and the calculated conditional vector c are imported into the encoder and converted to the latent vector z. Then the latent vector incorporates the random noise learned by the model, in which the default random noise is Gaussian distribution, and is imported into the decoder. The KL divergence is calculated by transforming the difference between the generated and target compounds. Adam optimization algorithm with a quadratic gradient correction function is used to find the globally optimal solutions.[24] Figure 1

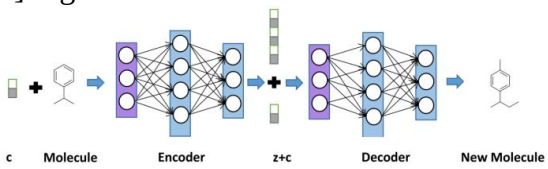

Figure 1: Model architecture. The conditional vector c and structure of compounds imported into the encoder are converted into a latent vector z. Then the information of z and c are transferred to the decoder, and new compounds can be generated by the corresponding parameters.

2.3.1 The Randomized SMILES

In this study, the moiety of GSK2816126, 3-(aminomethyl)-4, 6-dimethylpyridin-2(1H)-one (grape group in Figure 2), was used for showing the reading mode of the randomized SMILES. Canonical SMILES tends to select the longest carbon chain as the backbone, thus traversing the entire structure of the molecule, as shown by the canonical smiles in the figure. In this case, the order in which the atoms are traversed is not only determined by choosing the longest carbon chain first, but also by the constraints imposed by the SMILES string traversal in Rdkit, i.e., when traversing atoms with more than one branch, the side chains are traversed first. This constraint exists to prevent the generation of confusing SMILES strings when converting compound structural formulae to SMILES strings, which can lead to unusable results. Randomized SMILES also generates two types of traversal methods based on this, namely, those that are constrained by Rdkit's read rules (fig. 2) and those that are not. Randomized01 in figure 2) and those that are not restricted by the Rdkit reading rules (Randomized02 in figure 2). Although the likelihood of generating chaotic strings would be higher using an unrestricted approach, the Randomized SMILES model we used can still produce compound molecules with a high degree of diversity and potency by training with a large number of valid SMILES sequences.[12]

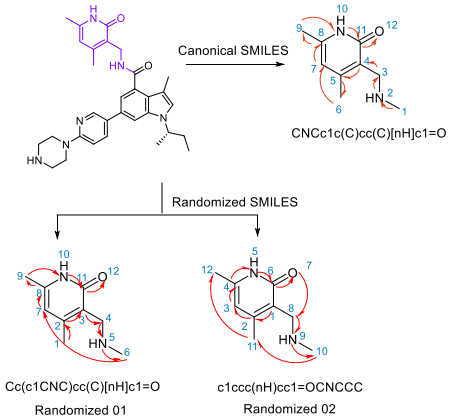

Figure 2: The canonical and randomized SMILES for the substructure of GSK2816126. For clear displaying the difference, 3-(aminomethyl)-4, 6-dimethylpyridin-2(1H)-one in GSK2816126 was selected as a template for showing the trajectory transformed by canonical and randomized SMILES. Blue label and red arrow mean atom ID and traversal order, respectively.

2.3.2 Molecular representation and model construction

In this study, the chemicals in the dataset were represented using the randomized SMILES method.[12] In canonical SMILES, the backbone is the longest carbon chain in the compound, and all side chains are traversed before reading the ring. This approach can prevent the creation of complex and redundant molecules, but it may restrict the chemical space available for the created compounds. By shifting the beginning point of the compounds, the randomized SMILES approach randomly arranges the atoms without affecting the rule that traverses the chemical graph. In unrestricted models, this can result in at most n! different SMILES for a molecule with n heavy atoms. [12] It can produce at least twice as many new compounds, increasing the generating space, by learning various SMILES strings for a single chemical. The randomized SMILES based on 96,813 solutions were added to 298,703 non-dominated solutions. A total of 398,703 non-dominated solutions were gathered and divided into the training and validation sets using the ratio of 4:1 after 100,000 secondary non-dominated solutions based on the pareto algorithm were added.

The letter 'E' was placed at the end of the compound SMILES. The SMILES characters are then represented using a one-hot encoding and transformed into a 300-by-300 embedding vector. The property values of the conditional vector c's MW, LogP, and tPSA are normalized from -1.0 to 1.0. [7] The data are compressed and then fed into the decoder so that the distribution of the various physical and chemical properties of the generated molecules is within a reasonable range rather than the fluctuation ranges of the three parameters being close to the same. The latent vector z is then produced by importing the embedding vectors and conditional vectors into the CVAE encoder. The encoder and decoder of the CVAE are built using RNN with an LSTM cell, which is effective for processing sequences.[25] There are 500 hidden nodes in each of the three layers that make up the RNN. A softmax layer is used to transform the input matrix. The SMILES of every compound are then randomly read by the decoder at each RNN cell's time step. The units of the RNN decoder are unrolled 120 times to produce compounds that roughly match the target property values based on conditional vectors. The probability distribution of the SMILES characters, including the letter 'E', is eventually generated by each LSTM unit of the decoder. The output compound is deemed invalid if there isn't an 'E' among these n characters (including the letter 'E'). The duplicate compounds are eliminated after examining the outputs of each cell in the decoder. This technique produces more molecules with favorable characteristics and significantly enhances the effective recognition of SMILES.

3. Result and discussion

In this study, GSK2816126 (an EZH2 inhibitor) was utilized as a reference to determine the non-dominated and dominated solutions utilizing the pareto algorithm from the database. Then the CVAE model was done for producing new compounds, in which the physicochemical qualities together with the structure of the compounds were regulated attributes.

3.1 Model Learning Capability Assessment

To evaluate the model's capacity for learning, we randomly selected 400 compound molecules from each of the training set as well as the generated molecules set and tabulated their values of tPSA, MW, and LogP, and finally made as graphs. (Figure 3). The distribution of each attribute between the created molecules and the molecules in the training set is consistently close, showing that our model is better able to remember and imitate the referenced attribute distribution molecules. The results show that our model has a strong learning ability to learn the distribution of the

parameters in the dataset and reproduce them well, especially in the distribution of the QED values, which, as a high-level attribute for measuring the drug-like properties, relies on the three parameters of LogP, MW, and tPSA utilized in the model. The density plot shows that the drug-like properties of the generated molecules have been optimized significantly by our model.

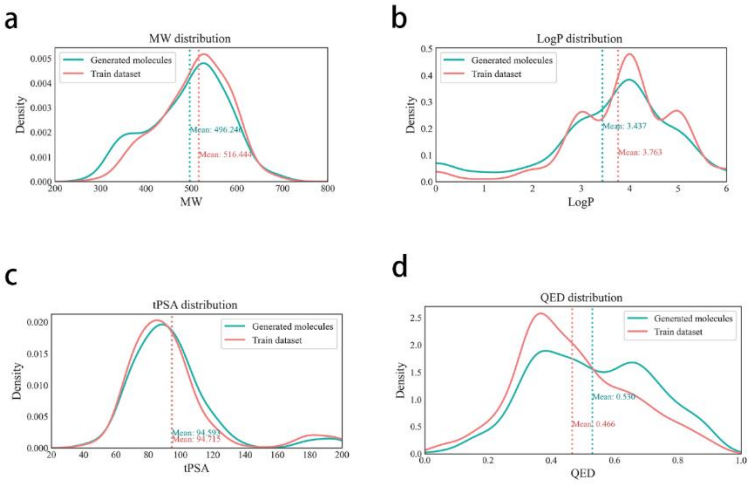

Figure 3: Specific distributions of the different parameters of the training and generator sets of molecules where the red line represents the training set and the green line represents the generated molecule set, and the overall mean is labeled by color on the figure.(a):Distribution of MW (b):Distribution of LogP (c):Distribution of tPSA (d):Distribution of QED

3.2 Validity, Uniqueness, Novelty, and Diversity

Validity, uniqueness, and novelty are important indicators for evaluating the performance of the generative models, and the formulas are defined as follows:

$$P^V = N^V / N^T \tag{2}$$

$$P^U = N^U / N^V \tag{3}$$

$$P^N = N^{U \cap N} / N^U \tag{4}$$

Where P^V represents the ability of the model to generate valid compounds. N^T is the total number of compounds generated by the model, and N^V is the number of the valid compounds checked by RDkit. P^U reflects the ability of the model to generate non-duplicate compounds. N^U is the number of compounds after removing duplicates in the generated set. P^N characterizes the novelty of the generated compounds. $N^{U \cap N}$ is the number of compounds in the unique compound set but not in the training set.

According to the analysis of the data in the table, after using the randomized SMILES method, validity decreased by 0.3%, uniqueness increased by 3.2%, and novelty increased by 2.2%. Analyzing these data together, we can conclude that although there is a decrease in validity, the change is very small and has little impact on the overall effect. However, uniqueness went up by 3.2 %, which is a very significant increase, indicating that the use of the randomized SMILES method can better identify and differentiate between different data samples, thus improving the personalization and uniqueness of the data. In addition, novelty also increased by 2.2 %, which suggests that new and unseen data patterns and regularities can be better identified using the randomized SMILES method, thus increasing the novelty and innovation of the data. Table 1

Table 1: The validity, uniqueness, and novelty of our models

	Canonical Model	Randomized Model
Validity	94.4%	94.1%
Uniqueness	81.4%	84.6%
Novelty	91.2%	93.4%

While the drop in validity may cause some concern in some application scenarios, we should note that validity is only one measure of model accuracy, while uniqueness and novelty are more concerned with model personalization and innovation. It may be related to the fact that we used a randomized SMILES method that is not constrained by the Rdkit read rule. Therefore, as a whole, using the randomized SMILES method is still a great improvement, and we still need to work on refining this modeling approach.

However, in order to further improve the overall performance and effectiveness of the randomized SMILES method, we still need to further improve and optimize the approach. This can be achieved by adjusting the model parameters, increasing the amount of training data, and improving the feature extraction method. At the same time, we can also consider combining other techniques and algorithms to further improve the accuracy, stability and robustness of the model.

To summarize, although the use of the randomized SMILES method decreases in terms of validity, its improvement in terms of uniqueness and novelty is very significant. Through further improvement and optimization, we can further enhance the performance and effectiveness of the randomized SMILES method, making it more suitable for various complex application scenarios.

The concept of molecular diversity is important in drug discovery as it helps to ensure that the screened compound library is sufficiently diverse to increase the success rate in finding potential drug compounds. By selecting compounds on the basis of molecular diversity, duplication can be reduced and the efficiency of research and development can be improved. The molecular diversity formula proposed by Benhenda et al.[26] was used to examine the diversity between the test and produced sets, as follows:

$$\text{Diversity}\ (M)\frac{1}{|M|^2}\sum_{(m_1,\ m_2)\in M\times M} T_d(\mathrm{m}_1,\ \mathrm{m}_2) \tag{5}$$

Where M stands for the dataset (the set of test molecules or the set of generated molecules), m_1, m_2 refer to two molecules, respectively, and $T_d(\mathrm{m}_1,\ \mathrm{m}_2)$ refers to the Tanimoto distance between these two molecules.

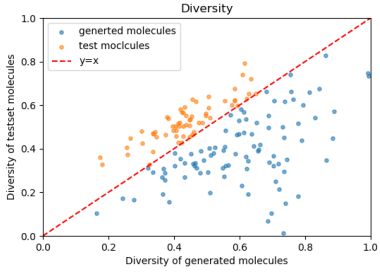

Figure 4: Distribution of molecular diversity in the molecular generation set and molecular test set, where blue dots represent molecules in the generation set and yellow dots represent molecules in the test set.

As shown in Fig. 4, most of the data points are located on the lower side of the diagonal and the distribution is significantly larger than that of the molecules in the test set. This indicates that the structural diversity of the model-generated molecules is higher than that of the molecules in the test set. Therefore, our model using the randomized SMILES method as well as the pareto algorithm is

able to generate diverse molecules more efficiently, which can help us explore a larger chemical space.

3.3 Extraction of non-dominated solutions using Pareto algorithm

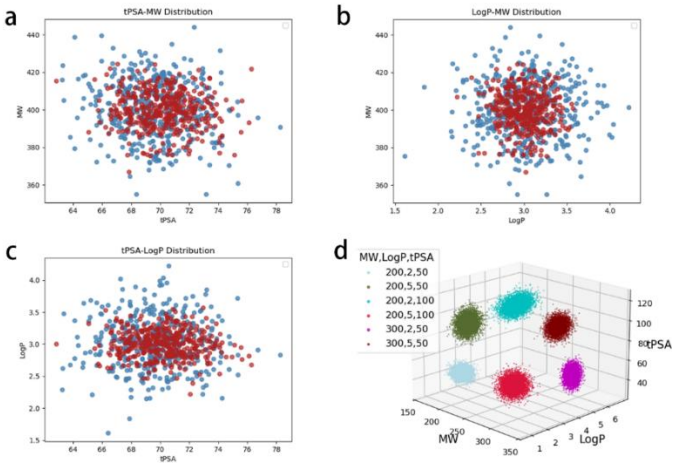

Figure 5: Property distribution for molecules of the generated set under dual (a,b,c) and triple attribute controls(d).

In this study, controls with two and three attributes were assessed. In terms of dual attribute control, we chose three typical physicochemical characteristics of compounds—MW, tPSA, and LogP—for attribute control, as constrained in, and trained the model using compound molecules with higher similarity activity to GSK2816126 to produce new molecules. We used the same preset value (MW=450,tPSA=70,LogP=3) for molecule generation for several iterations, and each time we compared the 400 molecules in the randomly selected set of generated molecules as well as the 400 molecules in the training set, and made some results of the comparison data into Figure 5.Through the graphs comparison, it can be shown that the created molecules have more ideal drug-like properties, and that the random two out of the three values are closer to the ideal range following property limitations. Additionally, we contrast the molecules produced with those produced by the CVAE model, which was trained using the same quantity of databases but without using the pareto approach. The ability of attribute control can be improved by utilizing the pareto algorithm to optimize the training set and comparing the property distribution of the created molecules. The pareto algorithm can make the physical and chemical properties of the created molecules more appropriate for the given parameters when the attribute control parameters are the same. It is clear that using the Pareto approach to optimize databases greatly aids in producing complex molecules with desired physicochemical attributes.

3.4 Molecular Plots Generated (Similarity Comparison)

In the molecular docking part, we used sybyl 2.0 to follow up the generated molecules. For protein molecules, we obtained a single crystal diffraction model of the bound 3D conformation of the EZH2 subunit with the small molecule GSK126 from RCSB. By analyzing the 3D conformation as well as the mode of binding between the protein subunit and the small molecule, the generated molecules were carried out in the same environment as the model. The modeled molecules underwent docking.

Depending on whether or not the produced molecules had three connected aromatic rings, they

were divided into compounds 1–6 and compounds 7–15. There is little difference between these two groups of compounds' similarities to GSK126, and molecules that share the same tricyclic structure as GSK126, i.e., compounds 7 through 15, typically perform better in molecular docking tests. Have generally higher docking ratings. First, in order to more thoroughly examine the binding conformations of GSK126, we exported their 2D binding planes via proteinplus.[27-29] Figure 6

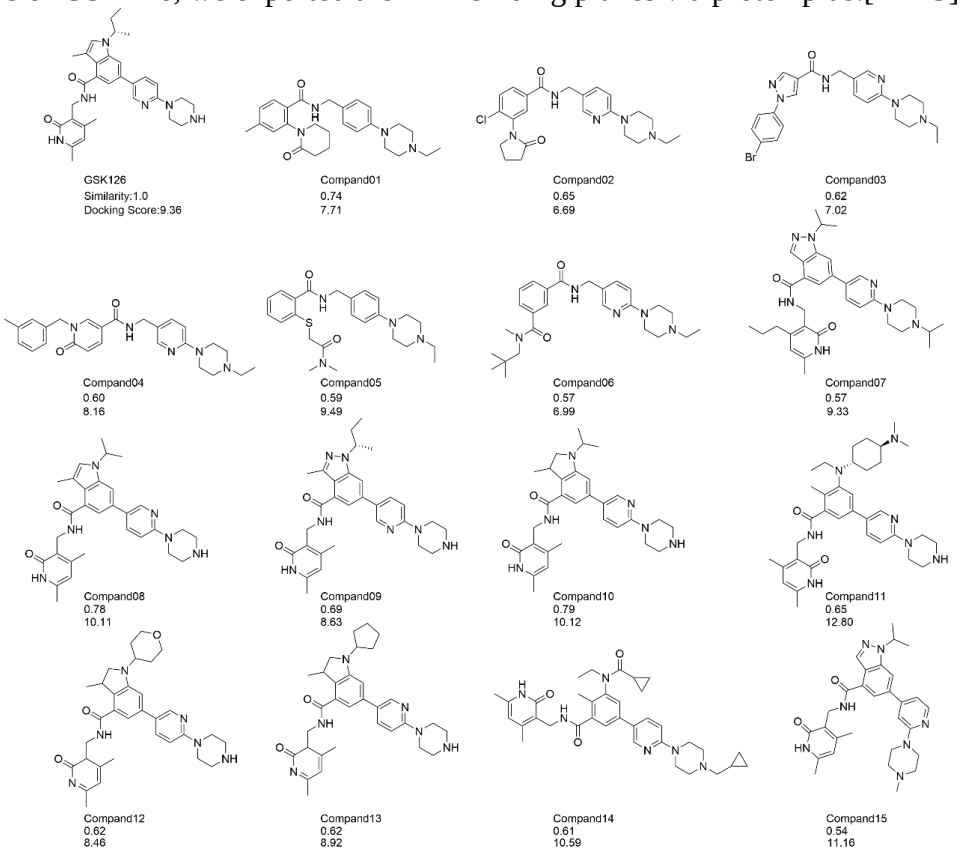

Figure 6: The molecules generated by our model and their similarity. The first line of number below the compound represents the similarity with the reference, and the second line mean the score of molecular docking.

The SET domain and the SET Activation Loop (SAL) make up the split catalytic region of EZH2. It has a variable aromatic "neck" region, a pyridone "head" region, and an amide linker. It has variable aromatic "body" regions (indole, indazole, or another hydrophobic group), a pyridone "head" region, an amide linker "neck" region, and variable "arm" and "tail" structures linked to the top and bottom of the body, respectively. Through Fig. 1, we can see that the residues of the SET structural domains of GSK126 and EZH2 are located on the left side of the binding region, where they form the Tyr661A and Phe665A, which serves as a barrier to the entry and binding of other molecules. In order to compete with SAM for binding to the same residues, the amide nitrogen and carbonyl oxygen of the GSK126 pyridone head mediate twisted hydrogen bonds with the carbonyl oxygen and amide nitrogen of residue Trp624A of EZH2. The side chain of residue Tyr111A connects with the indole body part, while the main chain carbonyl oxygen on the neck area between the pyridone and indole portions attaches to the main chain amide nitrogen of residue Tyr111A. The crystal structures show that EZH2 residues TYR661A and Tyr111A surround the body drug's variable elongation arm area, and that these residues combine to create the gating region, the major site of action of GSK126, in a way similar to other PRC2 structures.

In order to visualize the interactions between our compound molecules and the target EZH2, we

selected several compound molecules and plotted their molecular docking interactions (seen in Figure 7) .The compound with the highest docking score out of all the compounds was compound 15, which had a very similar binding mode and single crystal diffraction imaging with the target protein and major amino acids TYR661A and Tyr111A that also formed a solvent-gated structure, even though its similarity to the target compound was not the highest among the generated molecules. As a result, it received a higher score based on the preservation of its original activity. In contrast to GSK126, GSK126 has a nearby isoimidazole structure in place of the original thiophene structure, and the associated isobutyl group has been changed to an isopropyl group. This is a crucial site for the interaction with residue Tyr661A to sustain activity. The five-membered heterocyclic ring and substituent group alterations did not alter the binding conformation of the molecule and the target site, despite the fact that they altered the electron cloud density and spatial resistance. The binding conformation of the molecules to the target was unaffected by the changes in the five-membered heterocycles and substituents, and the smaller molecules were able to bind to the target more readily as a result of the substituents' decreased spatial resistance. Only compound 11 outperformed compound 15 in terms of molecular docking. It lost the pyrrole's five-membered ring structure when compared to GSK126, and the ring unfurled into a carbon chain with N. We hypothesized that the pyrrole on the ring was not a crucial structure for the binding of small molecules to the EZH2 target protein because, despite the structural change, it had no effect on the molecule's capacity to form a hydrophobic force with Tyr661A, maintaining the activity. Compound 09 among the generated compounds has the highest similarity to the GSK126 molecule, with only the benzopyrrole part replaced by benzisoimidazole. This change in electron cloud density may affect the compound's ability to bind to the EZH2 target protein, but the scoring of the compound and the three-dimensional conformation of its binding to receptor proteins indicate that the binding pattern of the key molecule is not affected.

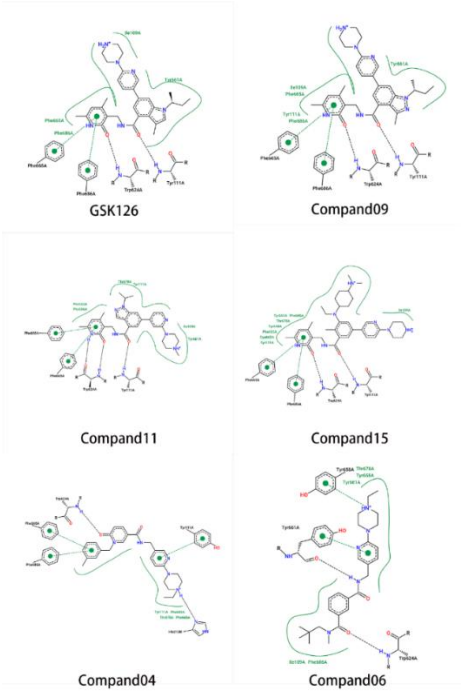

Figure 7: Two-dimensional plot of interaction forces for docking of selected compounds and target EZH2 molecules.

We can see from the binding mode schematic that compounds 04 and 06 still form an interaction with the important amino acid, maintaining the activity of the inhibitor of the EZH2 target.

Compounds 04 and 06 have lost the structure of the tricyclic ring on the basis of GSK126, and the spatial resistance has changed significantly.

4. Conclusions

In this study, a deep molecular generation model incorporating the pareto algorithm is proposed. By learning the molecules in the training set, the model can be oriented into molecules with reasonable properties. It is worth noting that we strive to make the numerical embodiment of the drug-like properties of the generated molecules more in line with the requirements through different methods and perform multi-objective optimization of the molecular generation model for multiple drug-like properties, so that the comprehensive quality of the generated molecular properties is more in line with our requirements. Moreover, by the results of molecular docking, the new molecules generated by our model can have good affinity to the target, which indicates that the resulting new molecules have a chemical space. We believe the model will play an auxiliary role in molecular design, providing new strategies for medicinal chemists to explore biologically relevant chemical spaces.

References

*[1] Walters, W.P. and R. Barzilay, Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. Acc Chem Res, 2021. **54**(2): p. 263-270.*

*[2] Tong, X., et al., Generative Models for De Novo Drug Design. J Med Chem, 2021. **64**(19): p. 14011-14027.*

*[3] D'Souza, S., P. Kv, and S. Balaji, Training recurrent neural networks as generative neural networks for molecular structures: how does it impact drug discovery? Expert Opin Drug Discov, 2022. **17**(10): p. 1071-1079.*

*[4] Kong, W., et al., Application of SMILES-based molecular generative model in new drug design. Front Pharmacol, 2022. **13**: p. 1046524.*

*[5] Bai, X.Y. and Y.X. Yin, Exploration and augmentation of pharmacological space via adversarial auto-encoder model for facilitating kinase-centric drug development. Journal of Cheminformatics, 2021. **13**(1).*

*[6] Joo, S., et al., Generative Model for Proposing Drug Candidates Satisfying Anticancer Properties Using a Conditional Variational Autoencoder. Acs Omega, 2020. **5**(30): p. 18642-18650.*

*[7] Lim, J., et al., Molecular generative model based on conditional variational autoencoder for de novo molecular design. J Cheminform, 2018. **10**(1): p. 31.*

*[8] Cheng, Y., et al., Molecular design in drug discovery: a comprehensive review of deep generative models. Briefings in Bioinformatics, 2021. **22**(6).*

*[9] Lamanna, G., et al., GENERA: A Combined Genetic/Deep-Learning Algorithm for Multiobjective Target-Oriented De Novo Design. J Chem Inf Model, 2023. **63**(16): p. 5107-5119.*

*[10] Xu, T.X., et al., A Scaffold-based Deep Generative Model Considering Molecular Stereochemical Information. Molecular Informatics, 2022. **41**(12).*

*[11] Choi, J., S. Seo, and S. Park, COMA: efficient structure-constrained molecular generation using contractive and margin losses. Journal of Cheminformatics, 2023. **15**(1).*

*[12] Arús-Pous, J., et al., Randomized SMILES strings improve the quality of molecular generative models. J Cheminform, 2019. **11**(1): p. 71.*

*[13] Saikia, S. and M. Bordoloi, Molecular Docking: Challenges, Advances and its Use in Drug Discovery Perspective. Curr Drug Targets, 2019. **20**(5): p. 501-521.*

*[14] Shivanyuk, A.N., et al., Enamine real database: Making chemical diversity real. Chimica Oggi, 2007. **25**(6): p. 58-59.*

[15] RDKit: Open-Source Cheminformatics. http://www.rdkit.org.

*[16] Kim, K.H. and C.W.M. Roberts, Targeting EZH2 in cancer. Nature Medicine, 2016. **22**(2): p. 128-134.*

*[17] Zhou, B., et al., Discovery of IHMT-EZH2-115 as a Potent and Selective Enhancer of Zeste Homolog 2 (EZH2) Inhibitor for the Treatment of B-Cell Lymphomas. Journal of Medicinal Chemistry, 2021. **64**(20): p. 15170-15188.*

*[18] Yap, T.A., et al., Phase I Study of the Novel Enhancer of Zeste Homolog 2 (EZH2) Inhibitor GSK2816126 in Patients with Advanced Hematologic and Solid Tumors. Clinical Cancer Research, 2019. **25**(24): p. 7331-7339.*

*[19] Berenger, F., O. Vu, and J. Meiler, Consensus queries in ligand-based virtual screening experiments. Journal of Cheminformatics, 2017. **9**.*

[20] Liu, X., et al., DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. J Cheminform, 2021. ***13****(1): p. 85.*
[21] Lotfollahi, M., et al., Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics, 2020. ***36****(Suppl_2): p. i610-i617.*
[22] Bjerrum, E.J. and B. Sattarov, Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. Biomolecules, 2018. ***8****(4).*
[23] Jørgensen, P.B., M.N. Schmidt, and O. Winther, Deep Generative Models for Molecular Science. Mol Inform, 2018. ***37****(1-2).*
[24] Kim, K.S. and Y.S. Choi, HyAdamC: A New Adam-Based Hybrid Optimization Algorithm for Convolution Neural Networks. Sensors (Basel), 2021. ***21****(12).*
[25] Arús-Pous, J., et al., SMILES-based deep generative scaffold decorator for de-novo drug design. J Cheminform, 2020. ***12****(1): p. 38.*
[26] Benhenda, M., ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? Arxiv, 2017.
[27] Schöning-Stierand, K., et al., ProteinsPlus: a comprehensive collection of web-based molecular modeling tools. Nucleic Acids Research, 2022. ***50****(W1): p. W611-W615.*
[28] Schöning-Stierand, K., et al., ProteinsPlus: interactive analysis of protein–ligand binding interfaces. Nucleic Acids Research, 2020. ***48****(W1): p. W48-W53.*
[29] Fährrolfes, R., et al., ProteinsPlus: a web portal for structure analysis of macromolecules. Nucleic Acids Research, 2017. ***45****(W1): p. W337-W343.*