

RESEARCH ARTICLE

Targeted drug design method based on scaffold-constrained molecular generation and distance matrix

Zhen Qin*

Department of Medicine and Pharmacy, Yangquan Vocational and Technical College, Yangquan, Shanxi, China.

Received: April 2, 2026; accepted: August 1, 2026.

Targeted drug design is a core technology for developing small-molecule drugs against specific disease targets. However, traditional methods suffer from issues such as low screening efficiency and limited docking accuracy, which makes it difficult for existing models to balance conformation prediction speed and affinity prediction accuracy. This research proposed a targeted drug design method based on skeleton constrained molecule generation and distance matrix by building a goal-oriented molecular generation model based on the scaffold-aware molecular optimization algorithm and combining the structure/ligand drug design scoring function to achieve multi-dimensional constraints. A conformation prediction model that integrated the edge feature map attention network and the distance matrix was designed. Geometric constraints were introduced to solve the conformational self-intersection problem. The results showed that the produced molecules all complied with the five rules of drug class. The single sample running time of conformation prediction was 0.5 s. The average absolute error of drug target affinity prediction was 0.702. The root mean square error was 1.005. The Pearson and Spearman correlation coefficients were better than those of the mainstream comparison models. The results confirmed that this study built a complete targeted drug design system from molecular generation to conformational prediction, significantly improved design efficiency and prediction accuracy, and provided a new technical path for data-driven targeted drug research and development.

Keywords: targeted drug design; scaffold-constrained molecular generation; distance matrix; edge-featured graph attention network; geometric constraints.

*Corresponding author: Zhen Qin, Department of Medicine and Pharmacy, Yangquan Vocational and Technical College, Yangquan, Shanxi 045000, China. Email: QZqin198511@163.com.

Introduction

Traditional targeted drug discovery relies on high-throughput screening technology, which requires *in vitro* experimental verification of a large number of compounds. It faces not only challenges such as insufficient exploration of chemical space, long R&D cycles, and high costs, but also difficulties in accurately identifying candidate molecules that combine high activity

with favorable drug-like properties. It can no longer meet the demands for high efficiency in modern drug discovery and development [1, 2]. Influenced by artificial intelligence (AI) and deep learning technology, data-driven computational drug design methods provide a new way to break through the above bottlenecks, which can quickly explore chemical space and predict drug-target interactions through algorithmic models, significantly improving the efficiency of

molecular design and screening [3, 4]. Drug development cycle is long, cost is high, and failure rate is high. Traditional calculation methods rely on empirical scoring functions and have limited accuracy. Through end-to-end learning of molecular representations, deep learning has demonstrated efficient and accurate advantages in tasks such as virtual screening, affinity prediction, and molecular generation, and has become an important tool for accelerating drug discovery.

Sako *et al.* proposed a diffusion model to address hydrogen bond interactions in structure-based drug design (SBDD) by treating proteins and ligands as pseudo-particles to explicitly model hydrogen-bonding interactions, significantly outperforming existing models in hydrogen-bonding reproduction and binding energies [5]. Zheng *et al.* developed an equivariant 3D conditional diffusion model to address the insufficient 3D geometric information of the target protein in *de novo* drug design and the generation sequence that did not conform to physics by integrating pocket atom information through an equivariant graph neural network and adopting a fragmented diffusion generation strategy to significantly reduce computing resource consumption. The model outperformed existing models in binding affinity, effectiveness, uniqueness, and novelty [6]. Sooram *et al.* designed a pharmacophore-guided drug design method to address the drug resistance in leishmaniasis by targeting Leishmania N-myristoyltransferase, screening out 20 amino acid fragments, constructing a heat map, and combining molecular dynamics simulations to evaluate stability [7]. Further, scientists proposed a k-nearest neighbor quantitative structure-property relationship modeling method for predicting hydroxamic acid lipophilicity and hydrogen bond strength by using the sliding window k-nearest neighbor algorithm and the simulated annealing-optimized k-nearest neighbor algorithm to optimize the prediction model of hydrogen bond donor and acceptor strengths, respectively, combining with electronegativity counting, hydrophobicity, and

other molecular descriptors [8]. Igashov *et al.* developed an E(3) equivariant three-dimensional conditional diffusion model for designing molecular linkers, which could connect any number of fragments and automatically determine the number of connected atoms. The results showed that this method performed better on standard data sets and generated more diverse and synthetic molecules, which was successfully applied to generate linkers under target protein pocket conditions [9]. Raheem *et al.* designed two separate message passing neural network models for drug interaction prediction, focusing on each drug in a pair of drugs separately, capturing its unique characteristics and interactions, and integrating molecular feature information and obtained superior performance [10]. Although the current targeted drug design models based on deep learning show advantages in aspects such as virtual screening and affinity prediction, there are still key problems that the lack of scaffolding structural constraints molecular generation, insufficient drug-likeness, and synthetic feasibility. Molecular docking and conformation prediction are limited by large search spaces, time-consuming calculations, and low accuracy of scoring functions. Most existing models are single links, without forming an integrated solution from generation, screening to prediction, and insufficient 3D geometric information.

This study proposed a targeted drug design method based on scaffold-constrained molecular generation and distance matrix by integrating the three major links of scaffold-constrained molecular generation, multi-dimensional virtual screening, and high-precision conformation prediction and affinity to build a complete design system and comprehensively improve research and development efficiency and quality. By applying scaffold-aware molecular optimization algorithm (SAMOA) and a dual screening of electrostatic fingerprint similarity and structural affinity, the drug-like properties and target binding potential of the molecules were improved. The ligand-protein interaction layer

was introduced into the edge-featured graph attention network (EGAT), and a dual-branch parallel architecture of distance graph and affinity prediction was designed to realize multi-task joint learning. The Cartesian coordinates of the ligand were derived based on the predicted distance matrix, and geometric constraints were combined to solve the conformational self-intersection problem and improve the chemical rationality of the generated structure. The results of this research improved the efficiency and prediction accuracy of drug design, shortened the research and development cycle, and reduced research and development costs. It also expanded the application ideas of skeleton constrained molecule generation and distance matrix combined deep learning, providing a new technological paradigm for AI driven computational drug development, and provided reference for developing similar target drugs and related algorithm optimization.

Materials and methods

Targeted drug molecular generation method based on SAMOA and target-oriented

Several highly active virtual compounds were first generated, and candidate molecules were obtained through multi-dimensional virtual screening such as synthesizability scores, drug-likeness indicators, molecular docking, and multi-dimensional virtual screening like absorption, distribution, metabolism, excretion, and toxicity prediction. SAMOA used recurrent neural network (RNN) to simplify the generation process of scaffold-constrained molecules [11, 12]. The model predicted the next reduced molecular linear input canonical marker for the appropriate open position of the scaffold sequence by minimizing the negative log-likelihood probability of the reduced molecular linear input canonical. The loss function of the model was calculated below.

$$L(\theta) = -\sum_{t=1}^T \log P(x_t | x_0, \dots, x_{t-1}) \quad (1)$$

where $L(\theta)$ was the loss function. θ was the set of trainable parameters of the RNN. T was the total length of the molecular linear input canonical sequence. t was the time step index. x_t was the true mark of the t -th time step. $P(\cdot)$ was conditional probability. To ensure that the generation position of targeted drug molecules was located in the target chemical space, the scoring function needed to be used to constrain the sampling process to adjust the conditional probability density. The scoring function was divided into two parts of SBDD scoring function and ligand-based drug design (LBDD) scoring function [13, 14], which were evaluated based on chemical features extracted from ligands including molecular fingerprints, physical and chemical properties, topological structure features, etc. The feature extraction model contained several stacked graph transformer layers. The input features were first processed through the normalization layer to generate query, key, value, and edge features. After dot product calculation, scaling, and softmax operation, the attention weight was obtained. The value weighted sum was then spliced with the edge features, and the updated node and edge features were output through the normalization and feed-forward network. After extracting chemical features, they were spliced using the feature splicing model based on a cross-attention mechanism to achieve two-way interaction between ligands and protein features. Ligand node features and protein node features were normalized to generate their respective queries, keys, and values. Cross-modal feature fusion was achieved through cross-attention calculations. After splicing, normalization, and feed-forward networks, the final ligand and protein feature representations were output for subsequent affinity prediction. The hidden feature of the model was shown as follows.

$$\begin{cases} h_i^0 = \omega_N^0 F_{N,i} + b_N^0 \\ e_{ij}^0 = \omega_E^0 F_{E,ij} + b_E^0 \end{cases} \quad (2)$$

where h_i^0 was the hidden feature. ω_N^0 and ω_E^0 were the weights of node features and edge features. $F_{N,i}$ and $F_{E,ij}$ were node features and edge features. b_N^0 and b_E^0 were both offsets. e_{ij}^0 was edge hidden features. The LBDD scoring function was calculated through electrostatic fingerprint similarity. The electrostatic fingerprint similarity was shown below.

$$S_E = \max(S_T(fp, query_fp))fp \quad (3)$$

where S_E was the electrostatic fingerprint similarity. S_T was Tanimoto similarity. fp was the electrostatic fingerprint vector of the generated compound. $query_fp$ was the electrostatic fingerprint vector of a known highly active molecule. The SBDD and LDBB scores were then constrained to the 0 - 1 distribution to obtain the final score as follows.

$$S_f = 0.5 * (\bar{S}_E + \bar{S}_S) \quad (4)$$

where S_f was the final score. $\bar{S}_E$ was the normalized electrostatic fingerprint similarity. $\bar{S}_S$ was the normalized SBDD score. Based on the above method, candidate virtual targeting drug molecules were obtained.

Virtual screening and conformational evolution of targeted drug molecules

After obtaining candidate virtual drug molecules, virtual screening was performed. Computer molecular dynamics virtual simulation optimization was conducted on the initial structure. A three-point rectangular virtual box with transferable intermolecular potential energy was constructed for the complex in the computational program, and a sodium ion model was added to achieve electrical neutralization of the system. The calculation process enabled periodic boundary conditions with a cutoff radius of 12 Å, and the particle grid Ewald summation algorithm (National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA) was used to solve long-range electrostatic interactions [15, 16]. The protein

ligand structure was fixed, and 8,000 steps of steepest descent and 2,000 steps of conjugate gradient energy minimization on the water molecule model were performed followed by 10,000 steps of global energy minimization on the entire system. The simulation system was heated to 300 K in the number-volume-temperature (NVT) ensemble for 50 ps followed by a 500 ps equilibrium simulation in the number-pressure-temperature (NPT) ensemble. A 300 ns formal dynamic simulation was eventually conducted. The simulation operation was integrated using the Verlet algorithm (Loup Verlet, Paris, France) with a step size of 2 fs, and the hydrogen bond structure was constrained using the Shake algorithm (European Center for Atomic and Molecular Computing, Osay, France). After screening drug molecules, principal component analysis (PCA) was used to perform conformational evolution. In a polyatomic system, the spatial position of each atom was determined by 3D coordinates. Since a large amount of configuration data generated by molecular dynamics simulations formed a high-dimensional matrix, the original high-dimensional matrix needed to be compressed into a low-dimensional square matrix form by integrating all frame dynamics trajectories using a covariance matrix as follows.

$$C = \begin{pmatrix} c(X_1, X_1) & c(X_1, X_2) & L & c(X_1, X_{3N}) \\ c(X_2, X_1) & c(X_2, X_2) & L & c(X_2, X_{3N}) \\ M & M & O & M \\ c(X_{3N}, X_1) & c(X_{3N}, X_2) & L & c(X_{3N}, X_{3N}) \end{pmatrix} \quad (5)$$

where C was the covariance matrix. c was the covariance between coordinate components. X_i was the Cartesian coordinate component. The covariance between coordinate components was shown below.

$$c(X_i, X_j) = \frac{(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{3N - 1} \quad (6)$$

where $\bar{X}_i$ was the coordinate mean of the trajectory. N was the number of frames. The synthesizability and drug-like properties were

calculated. Important parameters related to drug properties included molecular weight, quantitative estimation of drug-like properties, and the quantity of hydrogen bond acceptors. The simplified molecular linear input standard database was first screened for druggability, and compounds that did not meet the five rules of oral bioavailability and drug-like properties were eliminated. The screened molecules were batch converted into three-dimensional conformations through the OpenBabel (<http://openbabel.org>) toolkit to generate a spatial structure containing atomic coordinates. PCA was used to reduce the dimensionality of the protein conformation set obtained from molecular dynamics simulations, and representative conformations were extracted as the initial receptor structure for molecular docking. The three-dimensional coordinates of the ligand were docked with the PCA-optimized protein conformation to explore its binding mode and interaction mechanism [17]. AutoDockTools (<http://mgltools.scripps.edu>) was used to complete preprocessing. AutoDock Vina was used for molecular docking and verification, and protein-ligand interaction profiler (PLIP) (<https://plip-tool.biotec.tu-dresden.de>) and PyMolecular visualization system (PyMOL) (<https://www.pymol.org>) were used for interaction visualization and drawing, respectively. The drug-like properties were evaluated through absorption, distribution, metabolism, excretion, and toxicity (ADMET) analysis (Central South University, Changsha, Hunan, China).

Conformation prediction method based on EGAT and distance matrix

The study proposed a conformation prediction method based on EGAT, which was divided into two parts including an affinity prediction model and a distance matrix prediction model. Since the ligand atoms interacted with the target protein residue nodes, the study introduced an interaction layer into EGAT. The Dropout layer and the BatchNorm layer were added after each convolution layer of the model. The Dropout layer could randomly discard some neurons to avoid overfitting and improve generalization

capabilities. The BatchNorm layer could normalize the output of each layer, accelerate training and enhance network robustness. Residual connections were applied to speed up convergence and alleviate the gradient disappearance. In the interaction layer, the model first extracted the feature representations of the ligand and target protein, respectively, constructed a dual-channel feature matrix, and built an interaction matrix based on this matrix to capture atomic-level non-bonded interaction information. The interaction feature vector was updated accordingly. Subsequently, the node features were aggregated using the Set2Set readout layer. The mean square error was employed as the loss function to conduct end-to-end regression prediction of the binding affinity of the compound and the target protein to accurately estimate the affinity value. The interaction layer was defined as below.

$$I = \tanh(F_l, F_p^T) \quad (7)$$

where I was the interaction matrix. F_l and F_p were the feature matrices of ligand and target protein, respectively. The model adopted a dual-branch parallel design with the distance map prediction branch on the left and the binding affinity prediction branch on the right. The input layer received the node features and edge features of the ligand and receptor and extracted the ligand features and receptor features through the graph neural network module. The two were jointly input into the interaction layer for information fusion. The fused features were sent all the way to the distance map module to generate an inter-atom distance map, and the other channel output the dissociation constant prediction value through the interaction module and readout layer, realizing multi-task joint learning of distance map prediction and affinity prediction. For any distance prediction matrix, its distance mapping was shown as follows.

$$M_{DP} = 10 \cdot \sigma \left[\text{layernorm}(F_l), \text{layernorm}(F_p)^T \right] \quad (8)$$

where M_{DP} was the distance prediction matrix. σ was the sigmoid function. Based on the predicted distance matrix and protein-compound node coordinates, the ligand heavy atom coordinates could be analytically derived. Considering the prediction noise, numerical optimization methods were used to minimize the interaction loss to obtain the final docking conformation [18-20]. The interaction loss was calculated below.

$$L_I = \sum_i^n \sum_j^m (|M_{ij}^{true} - M_{ij}^{pred}|) \quad (9)$$

where L_I was the interaction loss. n and m were the total number of ligands and protein residues, respectively. M_{ij}^{true} and M_{ij}^{pred} were the true distance and predicted distance matrices, respectively. To speed up the calculation process, the study introduced RDKit (<https://rdkit.org>) conformation as a geometric restriction condition to optimize the solution efficiency by adjusting the flexible bond angle and avoiding three-dimensional collisions between atoms. For any rotatable bond, the dihedral angle was shown below.

$$\alpha = \arctan \frac{v_2}{v_1} \quad (10)$$

where α was the dihedral angle of the rotatable bond. v_1 and v_2 were the first order and second order statistics of the von Mises distribution, respectively. The probability density function of von Mises distribution was shown below.

$$P(\alpha) = \exp(\lambda \cos(\alpha) - \lambda) \quad (11)$$

where $P(\alpha)$ was the probability density of von Mises distribution. λ was the concentration of the distribution. The matrix B was constructed by connecting the transposed vectors of two adjacent dihedral angles in the bond rotation matrix as follows.

$$B = \begin{bmatrix} \cos(\alpha_{ikj}) & \sin(\alpha_{ikj}) & 0 & 0 \\ -\sin(\alpha_{ikj}) & \cos(\alpha_{ikj}) & 0 & 0 \\ 0 & 0 & \cos(\alpha_{ilj}) & \sin(\alpha_{ilj}) \\ 0 & 0 & -\sin(\alpha_{ilj}) & \cos(\alpha_{ilj}) \end{bmatrix} \quad (12)$$

where α_{ijk} was the adjacent dihedral angle composed of atoms i , j , and k . After determining the optimal dihedral angle of each rotatable bond, the Kabsch algorithm (Max Planck Institute for Medical Research, Germany) was used to perform three-dimensional rigid body transformation to precisely superimpose the optimized conformation and active site of the target protein.

Databases and overall workflow

The crystal structure of Ataxia-telangiectasia mutated (ATM) protein was obtained from Protein Data Bank (PDB) (Brookhaven National Laboratory, Upton, New York, USA). Active ATM inhibitor datasets were collected from ChEMBL database (European Bioinformatics Institute, Cambridge, UK). The study compared multiple open-source mainstream deep learning models obtained from GitHub platform (<https://github.com/>) including Deep Affinity, DeepDTA, RTMScore, Pafnucy, TopologyNet, FAST, OnionNet, DGMSMSI, and Token Mol. The candidate molecules were generated *via* scaffold-constrained molecular generation. Multi-dimensional virtual screening and molecular dynamics simulation optimization were then performed before conformation and affinity prediction were conducted, and druggability indicators of screened molecules were evaluated. The CASF-2016 dataset (<https://www.pdbbind-plus.org.cn/>) was selected for model comparison experiments, and EquiBind (<https://github.com/HannesStark/EquiBind>) was used as the control method to uniformly input protein and ligand structures for parallel conformation prediction. Evaluation indicators included ligand root mean square deviation, running time, mean square error, standard deviation, and correlation coefficient. For affinity prediction, additional performance evaluations were conducted using mean absolute

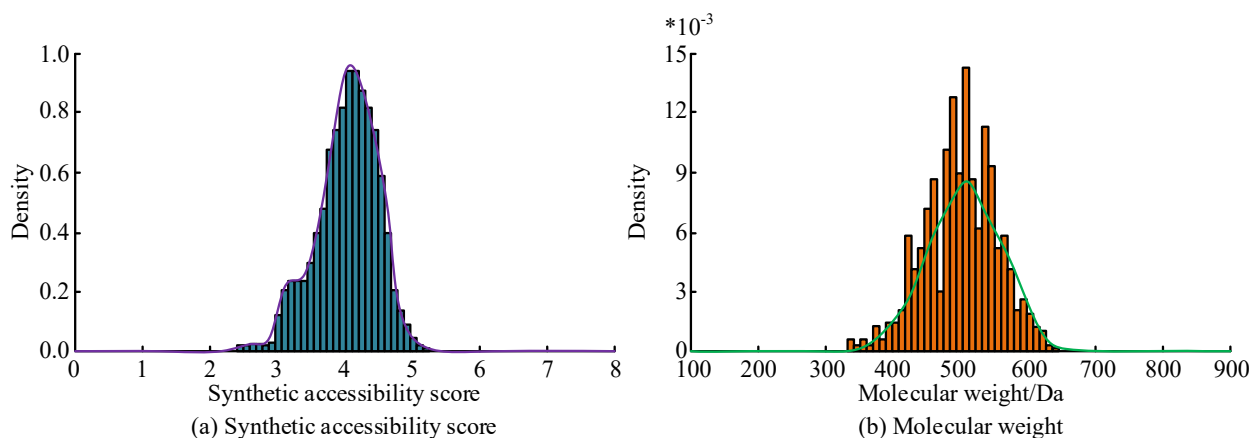


Figure 1. Synthesis accessibility score and molecular weight.

error (MAE), root mean square error (RMSE), Pearson correlation coefficient, and Spearman correlation coefficient.

Statistical analysis

Origin 2023 (OriginLab Corporation, Northampton, Massachusetts, USA) was used for statistical analysis. Independent sample t-test and correlation analysis were used in this study. *P* value less than 0.05 was defined as statistically significant.

Results and discussion

Synthetic accessibility and molecular weight

The synthetic accessibility scores and molecular weights of potential drug molecules screened by this method demonstrated that the synthesis accessibility score had an approximately normal distribution with the peak value at around 4.0, indicating that the synthesis difficulty of most of the generated molecules was moderate. The scores were mainly concentrated in the 3.0 - 5.0 range, and very few molecules had scores lower than 2.0 or higher than 6.0, indicating that the molecules generated by the model had good synthetic feasibility (Figure 1a). The molecular weight distribution was right-skewed with the peak value approximately between 450 Da and 500 Da, which was in line with the typical range of drug-like molecules. Most molecular weights

were concentrated in 400 – 600 Da, and a small number of molecules exceeded 700 Da (Figure 1b). The overall distribution was reasonable and met the molecular weight requirements of oral drugs.

Hydrogen bond acceptors and drug affinity

The number distribution and drug affinity of hydrogen bond receptors showed that the distribution of the number of hydrogen bond receptors demonstrated multi-peak discrete characteristics, mainly concentrated in the range of 6 - 10, of which compounds with 7 and 8 hydrogen bond receptors accounted for the highest proportion. The molecules in this data set were mainly medium polarity, which met the common requirement of no more than 10 hydrogen bond receptors in the five rules of drug-like drugs. The overall distribution was right-skewed, indicating that there were a small number of highly polar compounds (Figure 2a). The drug affinity score presented a unimodal continuous distribution, which was approximately a normal distribution shape with the peak value located in the 0.2 - 0.3 range, indicating that most compounds had medium binding affinity. The score range was mainly concentrated between 0.1 - 0.6 with the tail extending to above 0.8, indicating that there were relatively few high-affinity drugs in the data set, and the overall drug efficacy level was moderate (Figure 2b).

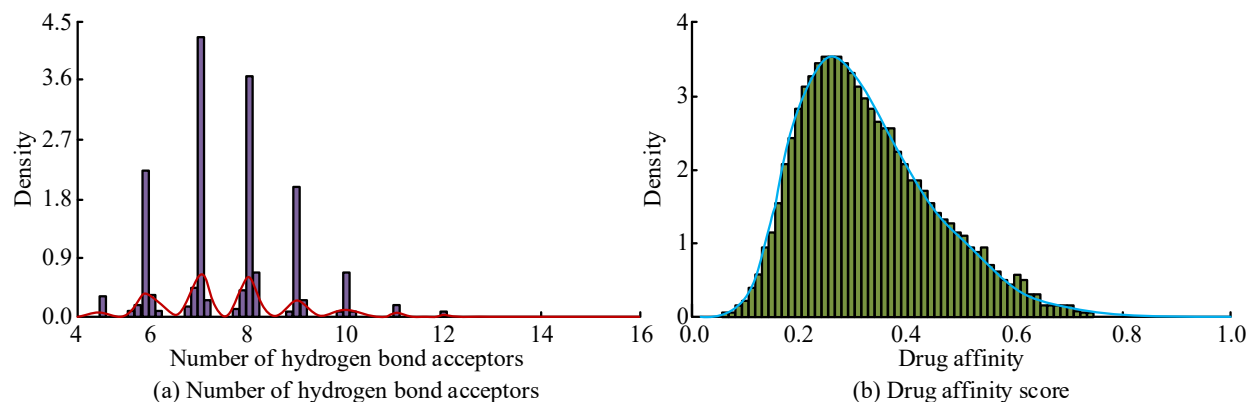


Figure 2. Distribution of hydrogen bond receptors and drug affinity.

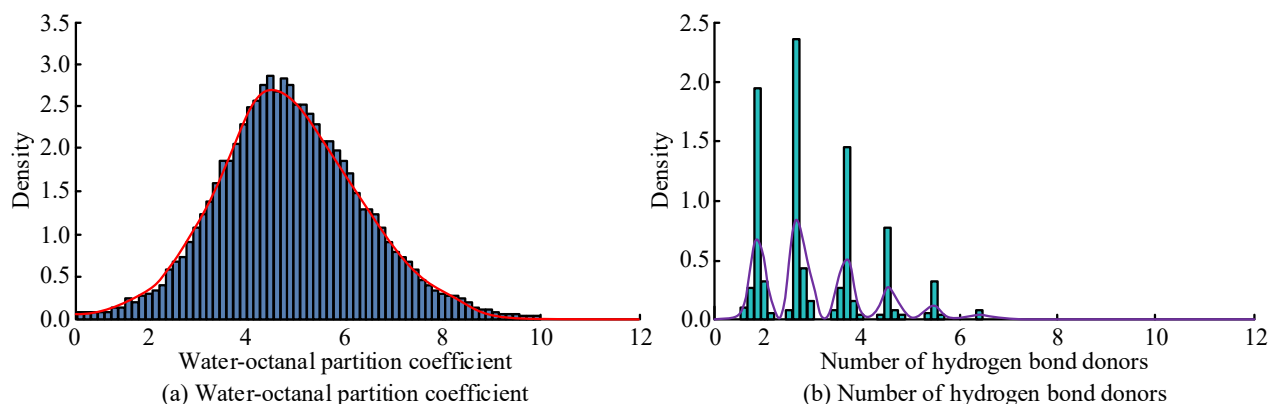


Figure 3. Water octanal partition coefficient and number of hydrogen bond donors.

Octanol-water partition coefficient and hydrogen bond donors

The water-octanal partition coefficient and the number of hydrogen bond donors showed that the water-octanal partition coefficient had an approximately normal distribution with a peak value near 4 (density of about 2.8), indicating that the molecules in this data set were mainly moderately fat-soluble. The distribution range was mainly concentrated between 2 and 6, which was in line with the recommended range of the water-octanol partition coefficient of less than 5 in the five drug-like rules. However, there was a long tail on the right that extended to more than 10, indicating that it contained a small amount of highly fat-soluble compounds (Figure 3a). The number of hydrogen bond donors was discretely distributed, mainly concentrated in the range of

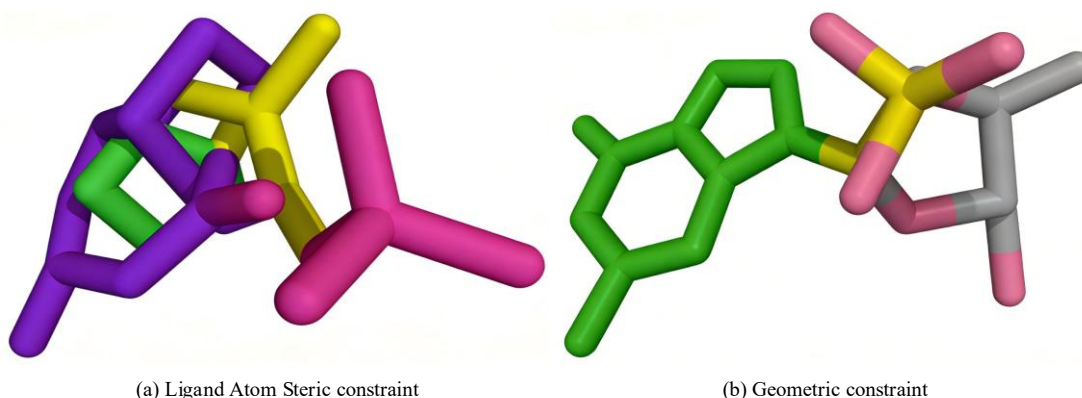
0 - 5, of which compounds with 2 hydrogen bond donors accounted for the highest proportion followed by 1 and 3. The distribution was obviously right-skewed, and there were very few compounds with more than 6 donors (Figure 3b). Overall, it met the requirement of no more than 5 hydrogen bond donors in the five rules for drug-like drugs, indicating that the number of polar groups of molecules in the data set is moderate.

Conformation of prediction generation model

To verify the conformational prediction generation model, the proposed model was compared with EquiBind using CASF-2016 dataset, which contained 285 protein targets, crystal complex structures, and active ligands each, and 7,065 bait molecules including 52 protein families, covering kinases, proteases,

Table 1. Results of ligand structure generation using different methods.

Index		EquiBind		Proposed method		P value
		Pocket	Protein	Pocket	Protein	
Root mean square deviation of ligand (Ligand RMSD)	< 2 Å (%)	0.00	7.97	6.37	5.96	< 0.05
	< 5 Å (%)	6.24	64.28	63.56	51.35	< 0.05
	Mean	8.15	7.31	4.10	4.81	< 0.05
Run time (s)		0.1	2.6	0.5	0.5	< 0.05
Mean squared error (MSE)		1.434	1.386	0.882	0.896	< 0.05
Standard deviation (SD)		1.778	1.571	1.195	1.231	< 0.05
R		1.269	1.107	0.843	0.858	< 0.05

**Figure 4.** Conformation generation results of different constraint modes.

nuclear receptors. The ligand structure generation results showed that the proposed method could rapidly generate ligand structures with an average speed of 0.5 s/piece (GPU time) and good accuracy. The proportion of molecules < 5 Å was relatively high, all above 50%. The average speed of Equibind fluctuated significantly, up to 2.6 s/piece, and the proportion of molecules with EquiBind (pocket) < 5 Å was only 6.24% (Table 1). The conformation generation results of different constraint methods showed that the steric constraints of the ligand atoms lacked effective restrictions on the spatial arrangement of the atoms within the molecule, and the generated ligand conformation was prone to self-intersections, in which non-bonded interactions between atoms were too close or even overlapping, resulting in unreasonable structural chemistry (Figure 4). The geometric constraint generation method could effectively maintain the geometric integrity and

spatial rationality of the molecule by strictly limiting local atomic structures such as bond lengths, bond angles, and dihedral angles, thereby well avoiding the above drawbacks and significantly improving the chemical credibility of the generated conformations. The Pearson correlation coefficient and Spearman correlation coefficient of different methods demonstrated that, in terms of the Pearson correlation coefficient, the proposed model performed best at 0.858 followed by Token-Mol and DGEMST at 0.839 and 0.827, respectively. The traditional method DeepAffinity performed the worst at 0.415 (Figure 5a). The trend of the Spearman correlation coefficient was basically consistent with that of Pearson with proposed model still ranking first at 0.828 followed by Token-Mol and DGEMST. The overall stability was better, but the Spearman coefficients of RTMScore and DeepDTA improved compared to their Pearson coefficients (Figure 5b). The proposed method

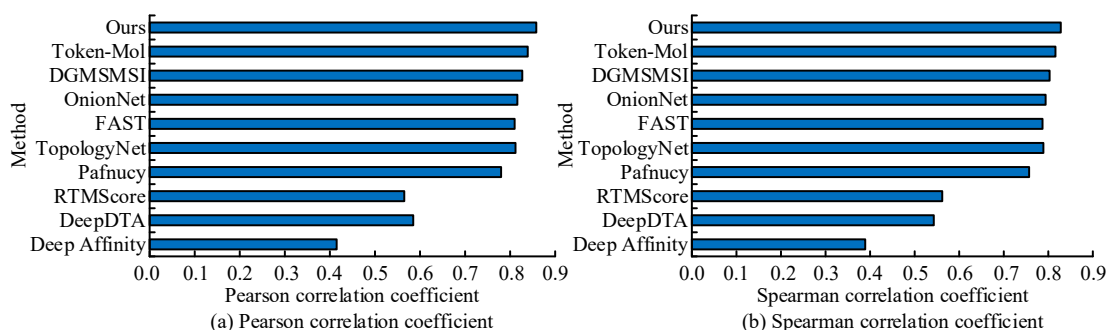


Figure 5. Pearson correlation coefficient and Spearman correlation coefficient of different methods.

Table 2. Comparison results of different methods.

Method	MAE	RMSE	Native ligand conf	Native protein conf
Deep Affinity	1.905	-	-	-
DeepDTA	1.531	-	-	-
RTMScore	1.247	-	-	Yes
Pafnucy	1.138	1.435	Yes	Yes
TopologyNet	-	1.362	Yes	Yes
FAST	1.056	1.311	Yes	Yes
OnionNet	0.971	1.265	Yes	Yes
DGMSMSI	0.848	1.204	Yes	Yes
Token-Mol	0.783	1.157	Yes	Yes
Proposed method	0.702	1.005	-	Yes
P value	<0.05	<0.05	-	-

has excellent conformation prediction generation performance. In the drug target affinity prediction task, the proposed method achieved the best performance in both MAE and RMSE as 0.702 and 1.005, respectively, significantly better than Deep Affinity, DeepDTA, RTMScore, and Pafnucy (Table 2). The results suggested that the proposed method had effectively improved the conformation prediction accuracy and demonstrated stronger affinity prediction ability.

Conclusion

Traditional targeted drug development relies on high-throughput screening techniques, which generally suffer from low efficiency in chemical space exploration, limited molecular docking accuracy, and time-consuming conformational prediction. Existing deep learning models also

struggle to balance prediction speed and accuracy, and there is still room for improvement in target affinity prediction capabilities. This research proposed a targeted drug design method based on skeleton constrained molecule generation and distance matrix, establishing a complete technical system integrating molecule generation, virtual screening, and conformation prediction, which could generate candidate molecules that complied with the five rules of class drugs. The geometric constraint strategy effectively solved the ligand conformation self-intersection and improved the chemical rationality of the molecular structure. Compared with mainstream models, this method had better comprehensive performance in conformation prediction and target affinity prediction tasks, achieving synchronous improvement in prediction efficiency and accuracy and providing a new technical idea for data-driven targeted drug development. This study still had certain

limitations. The experiment was only conducted to validate ATM inhibitors, and the universality of the model on other target proteins needed to be further tested. The ability to capture spatial features of structurally complex protein pockets also needed to be strengthened. Subsequent research should expand experimental targets for different protein families, optimize model generalization ability, integrate protein three-dimensional structure and dynamic conformation information, further improve conformation prediction accuracy in complex scenarios, and promote the practical application of this algorithm in targeted drug development scenarios.

References

- Weller JA, Rohs R. 2024. Structure-based drug design with a deep hierarchical generative model. *J Chem Inf Model*. 64(16):6450-6463.
- Renawala HK, Chandrababu KB, Smith KJ, D'Addio SM, Topp EM. 2024. A model study to assess fibrillation and product stability to support peptide drug design. *Mol Pharm*. 21(5):2223-2237.
- Wang K, Huang Y, Wang Y, You Q, Wang L. 2024. Recent advances from computer-aided drug design to artificial intelligence drug design. *RSC Med Chem*. 15(12):3978-4000.
- van den Maagdenberg HW, de Mol van Otterloo J, van Hasselt JGC, van der Graaf PH, van Westen GJP. 2025. Integrating pharmacokinetics and quantitative systems pharmacology approaches in generative drug design. *J Chem Inf Model*. 65(10):4783-4796.
- Sako M, Yasuo N, Sekijima M. 2025. DiffInt: A diffusion model for structure-based drug design with explicit hydrogen bond interaction guidance. *J Chem Inf Model*. 65(1):71-82.
- Zheng J, Yi HC, You ZH. 2025. Equivariant 3D-conditional diffusion model for *de novo* drug design. *IEEE J Biomed Health Inform*. 29(3):1805-1816.
- Sooram B, Mallikarjunachari U, Uddavesh SSP. 2024. Pharmacophore-guided drug design using LdNMT as a model drug target for leishmaniasis. *J Biomol Struct Dyn*. 42(1):863-875.
- Diwan J, Rajwade RP. 2025. Hydrogen bond descriptors of hydroxamic acids for use in drug design: A 2DQSPR approach using kNN. *Int J Biol Chem Sci*. 19(2):811-821.
- Igashov I, Stark H, Vignac C, Schneuing A, Satorras VG, Frossard P, *et al*. 2024. Equivariant 3D-conditional diffusion model for molecular linker design. *Nat Mach Intell*. 6(4):417-427.
- Raheem AKA, Dhannoon BN. 2024. A novel deep learning model for drug-drug interactions. *Curr Comput Aided Drug Des*. 20(5):666-672.
- Kasture P, Shirsath K. 2024. Enhancing stock market prediction: A hybrid RNN-LSTM framework with sentiment analysis. *Indian J Sci Technol*. 17(18):1880-1888.
- Andabil ZF, Nazemi A, Mirlohi SM, Fang SC, Liu B. 2025. Utilizing RNN based model and bi-objective programming to a new mean-conditional value at risk-entropy for uncertain portfolio optimization with liquidity and diversification. *Fuzzy Optim Decis Mak*. 24(3):397-429.
- Chen L, Wu M, Ohwada T. 2025. Identifying bioactive conformations of flexible polyunsaturated fatty acids for ligand-based drug design. *Chem Pharm Bull*. 73(10):1000-1009.
- Vyas A, Patel PK, Patel S, Dixit N, Sagar S. 2025. Unlocking potential topoisomerase-II inhibitors: Ligand-based drug design pipeline integrating pharmacophore mapping, virtual screening, molecular docking, and ADMET profiling. *Chem Zvesti*. 79(8):5023-5038.
- Vasant M, Ganesan S, Kumar G. 2025. Enhancing e-commerce security: A hybrid machine learning approach to fraud detection. *FinTech Sustain Innov*. 1:A7.
- Khalid M, Alqarni MH, Foudah AI. 2025. Repurposed pharmacotherapy: Targeting cathepsin L with repurposed drugs in virtual screening. *Mol Divers*. 29(5):4117-4129.
- Ma Y, Shih CH, Cheng J, Chen HC, Wang LJ, Tan Y, *et al*. 2025. High-throughput empirical and virtual screening to discover novel inhibitors of polyploid giant cancer cells in breast cancer. *Anal Chem*. 97(10):5498-5506.
- Karade D, Karade V. 2023. AIDrugApp: Artificial intelligence-based Web-App for virtual screening of inhibitors against SARS-COV-2. *J Exp Theor Artif Intell*. 35(3):395-443.
- Li V, Lee Y, Kim H. 2023. Repurposing existing drugs for monkeypox: Applications of virtual screening methods. *Genes Genom*. 45(11):1347-1348.
- Gangopadhyay A, Saha A. 2023. Drug repurposing against the RNA-dependent RNA polymerase domain of dengue serotype 3 by virtual screening and molecular dynamics simulations. *J Biomol Struct Dyn*. 41(11):5152-5165.