ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Optimization of drug synthesis routes and molecular design of new drugs based on molecular simulation

Jinchao Mu[1], *Chunfen Liu*[1,*], *Lianxin Liu*[2], *Jun Jiang*[3] *and Yu Liu*[1]

[1] College of Chemical Engineering, Xuzhou College of Industrial Technology, Xuzhou 221140, Jiangsu, China
[2] Basic Course Teaching Department, Xuzhou College of Industrial Technology, Xuzhou 221140, Jiangsu, China
[3] Quality Assurance Department, Wisdom Zenith Pharmaceutical Co., Ltd, Suqian 223800, Jiangsu, China

SUMMARY: *To solve the problem of multiple factors limiting the prediction of organic reactions in drug synthesis, machine learning is used to achieve more accurate and efficient reaction prediction and catalyst screening. Firstly, using computers as tools and combining multidisciplinary knowledge, theoretical simulations and other methods are used to guide and assist in the design, discovery, and synthesis of new drug molecules. These methods are mainly divided into three categories based on drug small molecule structure, receptor, and molecular dynamics. Protein structure determination and docking methods are also introduced. Secondly, an improved graph convolutional network model is proposed to address the existing model issues. Firstly, the reactants are transformed into feature matrices, and the improved graph convolutional network and attention mechanism are used to predict candidate reaction centers. Then, candidate products are generated by enumerating chemical constraints. Finally, the improved graph convolutional differential network is used to evaluate and rank the candidate products. The experimental results show that the accuracy of drug synthesis prediction: on the USPTO test set, the proposed method model outperforms the WLDN model in any number of template matches, and the advantage is more pronounced when the number of template matches is low. In terms of the accuracy index of reaction product prediction, the method proposed in this paper has the smallest parameter scale, but it is significantly better than other models in Top-1, Top-2, Top-3, and Top-5 indicators. In terms of protein mechanism analysis, compounds 1 and 2 form multiple hydrogen bonds with the target, and compound 2 has a higher binding energy than compound 1, but the ICso also increases, indicating that changing the substituents on the branched chain connected to the benzene ring in the structure of fluoroquinolone drugs can affect drug activity. The above results indicate that computer-aided drug synthesis design and prediction methods based on graph convolutional neural networks are effective, and the improved graph convolutional network model performs well in drug synthesis prediction, providing new ideas and methods for drug synthesis reaction prediction.*

KEYWORDS: *molecular simulation; Drug synthesis; Graph Convolutional Network; Protein structure; Differential network; molecular structure*

*17351732673@163.com
https://doi.org/10.65102/is20261103
Published: 10 June, 2026

1 Introduction

The discovery of organic chemical reactions in drug synthesis relies on practical experience and the "chemical intuition" dominated by chemical mechanisms. Experimenters attempt to qualitatively identify patterns in organic chemical reactions to determine reaction products and reaction efficiency [1, 2]. However, this method is limited by many factors, including the complexity of the reaction, the activity cliff, the lack of understanding of the mechanism, and the difficulty of manually processing big data. Computer based virtual screening has become an important solution that attracts chemists, mainly because it does not require an understanding of the mechanism, and compound structures can be characterized by molecular simulation numerical representations of molecular properties, thereby quantifying the chemical properties of thousands of candidate molecules. Based on experimental and literature data, virtual screening can quantify the results of drug synthesis reactions and the selectivity of catalysts through computer models.

Machine learning has been successfully applied in the field of chemistry for virtual drug screening, molecular generation, organic reaction prediction, catalyst screening, material discovery, computer-aided synthesis design, and reaction condition optimization [3]. Linear regression is a traditional reaction prediction and analysis tool that assumes a linear relationship between the physical characteristics and reactivity of reactants. It can be used to manually select input variables based on the mechanism of the reaction, which is in line with the thinking and statistical approach of data scientists. Hammet's use of linear regression to fit compound descriptors and outputs in the inference of linear free energy relationships is a representative work. For a long time, due to the multidimensionality of molecular features and the complexity of reaction spaces, it has been difficult to generate sufficiently complete and consistent data, which has limited the development of machine learning in the field of drug synthesis. Nowadays, High Throughput Experimentation (HTE) has become an effective means to gradually eliminate this obstacle. Ahneman et al. used methods such as Support Vector Machine (SVM) and Random Forest (RF) to predict the yield of Buchwald Hartwig coupling reaction in over 4000 high-throughput experimental data. In addition, Zahrt et al. predicted the enantioselectivity of chiral phosphoric acid (CPA) catalysts in over 1000 reactions using RF. The process of using computers for virtual screening of chemical reactions involves first extracting simplified molecular input line entry systems (SMILES) or molecular fingerprints from existing chemical reaction databases and literature, or using density functional theory (DFT) tools such as Gaussian to optimize the structure of these molecules and calculate reaction related properties. Then, molecular descriptors are constructed using these physical and chemical properties, and appropriate machine learning methods are selected for drug synthesis molecular modeling; Finally, screen the reactions to be analyzed in the dataset. This method is intuitive and effective for data scientists, without the need to focus on understanding reaction mechanisms, and has become a standard process for predicting chemical drug synthesis. The success of this process depends on two key factors [5, 6]: 1) the selected DFT features or molecular fingerprints, and the accuracy of the descriptors constructed using them; 2) Is the machine learning method effective. After decades of development, the prediction of organic reactions related to drug synthesis is still constrained in these two aspects:

1) For DFT features based on quantum mechanics, selecting features for different reactions has always been a challenge in predicting drug synthesis, especially in terms of feature selection for predicting reaction yield and selectivity, which often varies greatly. If the difficulty of feature selection can be reduced, it will promote the prediction of drug synthesis related reactions [7]. For sequence features using SMILES and molecular fingerprints, insufficient expression of three-dimensional (3D) structural information has always been a challenge. This

is determined by SMILES as a simplified linear representation of molecular structure and the algorithmic nature of molecular fingerprinting.

2) In the prediction of chemical reactions related to drug synthesis, traditional machine learning methods such as SVM and RF, and even linear regression methods have always been mainstreaming. Due to the existence of the 'curse of dimensionality', as the feature dimension increases, the required chemical reaction data sharply increases, greatly exceeding the workload of manual experiments. The emergence of high-throughput experiments has gradually alleviated this problem. Nowadays, how to apply high-throughput chemical reaction data to deep learning has become a challenge for data scientists. However, due to the lack of knowledge in chemistry and relatively scarce reaction data (compared to traditional deep learning applications such as images, videos, audio, and text), research on deep learning methods in this field is still rare [8, 9]. Although some work has extracted SMILES from reaction data in literature and used deep learning methods for prediction, there is still an urgent need to study how to use deep learning methods for virtual screening of the increasingly accumulated DFT reaction data.

2 Computer aided drug synthesis design

Computer aided drug synthesis design is a widely used method for drug synthesis design today. This method is a method that uses computers as tools, based on the research results of life sciences such as medicinal chemistry, biological enzymology, molecular biology, and genetics, as well as knowledge of biological macromolecular targets. It guides and assists in the design, discovery, and synthesis of new drug molecules through theoretical simulation, calculation, and prediction [10, 11]. This method has the characteristics of fast speed and low cost in predicting the biological activity of new drugs and studying the molecular structure of new drugs. Due to the full utilization of the advantages of computers, this method has strong targeting, high efficiency, and designed drug molecules with strong activity. Currently, it has become an important method for developing new drugs and an essential program for pharmaceutical companies.

Combining molecular simulation results with potential drug synthesis design targets revealed in basic research, including enzymes, receptors, ion channels, and nucleic acids, and referencing the chemical structural characteristics of other class derived ligands or natural products, reasonable drug molecular structures are designed, and then these compounds are synthesized using organic drug synthesis methods [12]. This greatly reduces the blindness and randomness of new drug development. Computer assisted drug synthesis design can be mainly divided into three categories: (a) drug molecule design methods based on the small molecule structure of drugs; (b) Receptor based drug molecule design methods; (c) Molecular dynamics is a fundamental method for drug molecular design.

This type of method mainly targets drug molecules with unknown receptor structures, including quantitative structure-activity testing and pharmacophore modeling methods.

(1) Determination of protein structure. It is a method of drug molecule design based on the structure of drug molecules. The first step of this method is to determine the structure of its target. Only by determining the target structure can the relationship between the receptor ligand be understood, and the next step of work can be carried out. Generally speaking, there are two main methods for determining the structure of proteins: X-ray diffraction and nuclear magnetic resonance [13]. For proteins with confirmed structures, we can obtain their 3D configurations from the protein library. For proteins whose structures have not been confirmed or are difficult to separate using various methods, their structural models can be obtained by establishing models, which can then be used for research. There are four main methods to obtain unknown

target protein structures:

(a) The method of homology modeling is a relatively fast way to obtain protein structures. It is not only used in drug synthesis design, but also in protein-protein interactions and site directed mutagenesis [14]. Its principle is that if more than 30% of the amino acid sequence in a protein lacking structural information is the same as that of a homologous protein, then the structure of the protein can be established based on its homologous protein structure. This method has not only been used in the past but is also increasingly being applied in many scientific studies.

(b) The method of identifying folds to determine protein structure was established by Bowie and his colleagues in 1991 when describing the environment in which amino acid residues interact with each other. This method estimates the 3D structure of a given protein sequence. The interaction between amino acid residues and the protein surface region both exhibit characteristics of helical structure.

(c) The method of protein modeling based on Ab, starting from Ab, is a method of establishing target structures based on physical principles, residue interaction centers, and a framework representation of a protein [15]. This method is very useful, especially when other methods and techniques have failed to predict unknown protein structures, but compared to other methods for determining protein structures, it is more accurate and precise.

(d) The method of hotspot prediction is also very important in drug molecule design based on drug molecular structure, which is to determine the active site of the ligand. Although the position of ligands in the crystal framework structure can be determined by X-ray diffraction to identify their active sites, this method is not applicable to proteins that cannot form crystals. Among the current methods for confirming binding sites, FTMAP is a newly invented method [16]. It uses segmented fragments as probes to explore protein surfaces, predicting possible drug action sites by identifying the points where the fragments aggregate. Typical hydrogen bonds and electrostatic interactions between molecules can also be used to predict the relationship between probes and proteins. In addition, the structure of molecular probes can also serve as the basis for designing new drug molecules.

(2) The method of docking. There are five main methods for docking in computer-aided drug synthesis design [17, 18]:

(a) Autodock is a method for docking flexible ligands and rigid 3D structures. It uses a series of coordinate grids to describe the 3D structure based on the AMBER position, and calculates van der Waals and Coulomb forces using data generated by the AutoGrid software package.

(b) CDOCKER, the CDOCKER docking method is a docking calculation method that maintains the flexibility of all ligands from the perspective of CHARMM. The various configurations of ligands are the optimal configurations obtained through high-temperature molecular dynamics simulations based on the original structure of the ligand. During the docking process, CDOCKER defines the active site through a sphere, so the relevant information of the binding site cannot be obtained. It can also be said to be a minor drawback of the CDOCKER docking method.

(c) Flexible docking is different from other methods in that it can maintain the flexibility of the receptor during the docking process with flexible ligands. The first step in the calculation process is to determine the structure of the target receptor side chain using the ChiFlex algorithm. The ChiFlex algorithm generates various protein configurations with different side chain orientation structures. The second step is to provide configurations with low energy during the docking process. The LibDock module is used during the docking process to indicate the most likely binding sites between polar and non-polar groups in the ligand and the protein, and then eliminate similar ligand configurations [19]. The final step is to optimize and improve

the ligand configuration with the highest score: the side chains are quenched through ChiRotor algorithm and CDOCKER structure simulation, and the energy of each ligand is minimized to achieve the optimization effect. In short, flexible docking can optimize the flexibility of each side chain. However, flexible docking generally requires a large amount of computer resources and generates much more data than rigid docking.

(d) LigandFit is a docking method that calculates the interaction energy between receptors and ligands based on coordinate grids. It plays a crucial role in the morphology of the starting ligand that matches the binding site of the receptor. The docking of ligands with designated sites using Li gandFit modules generally requires three key steps [20]: determining the active site and ligand configuration, docking the ligand with the selected binding site, and scoring the predicted morphology of the complex.

(e) The method of modeling transmembrane proteins, although there are already several drugs targeting transmembrane proteins, the bottleneck problem of this method is that transmembrane proteins are generally difficult to crystallize and their protein structures cannot be accurately analyzed. Moreover, in addition to considering transmembrane protein modeling, this method also considers the importance of phospholipid cell membranes. Therefore, the simple position of phospholipid bilayer membrane should also be considered during the docking process [21]. In the Discovery Studio 2.5 module of Accelrys, the CHARM stance is used as the membrane stance.

The main methods of drug discovery include drug synthesis design and drug screening. The most difficult problem to solve in the process of drug discovery is selecting and discovering a suitable drug target. Drug targets are biomolecules that have important physiological and pathological functions, can bind to drug molecules and produce pharmacological effects, and have specific structural sites. So far, there are about 500 biological target molecules among all discovered targets, which is relatively fewer compared to other targets. There are relatively more targets obtained through genetic research techniques, including enzyme targets, ion channels, G protein coupled receptors, and finally a small number of nuclear receptors. If the amino acid sequences and genomic data of these drug targets can be fully utilized, existing new drug development technologies can be used to obtain several times more drugs than the targets. It can be imagined that the economic benefits it brings are incalculable.

3 Drug synthesis prediction based on graph convolutional neural network

In order to solve the problems that arise in existing models for predicting chemical reaction products in drug synthesis, such as the inability to exhaustively enumerate rules in template methods, frequent atomic non conservation in sequence-to-sequence models, and difficulty in effectively obtaining global information in traditional graph convolutional networks, an improved graph convolutional network model has been proposed and has achieved certain results [22]. There are two basic steps for predicting organic chemical reactions in drug synthesis: generation of candidate products and screening of candidate products. Firstly, the chemical reactants are converted into feature matrices and input into the model. By improving the graph convolutional network and attention mechanism, the location of candidate reaction centers is predicted, that is, which atomic pairs may experience chemical bond breakage and combination; Secondly, candidate products are generated based on chemical constraints and enumeration of candidate reaction centers. By improving the graph convolution differential network, the candidate products are evaluated and ranked, and the candidate product with the highest score is the final reaction product.

3.1 Mapping of reactant characteristics

Consider the given drug synthesis chemical reaction as a pair of molecular diagrams (G_r, G_p), which can also be defined as $G=(V,E)$. The drug synthesis reactant is defined as G_r, the drug synthesis reaction product is defined as G_p, and the atomic set is defined as $V=\{a_1,a_2,a_3,\ldots,a_n\}$. The set of chemical bonds is defined as $E=\{b_1,b_2,b_3,\ldots,b_m\}$, which includes types such as single bonds, double bonds, aromatic bonds, etc. The reaction center, which refers to the location where chemical bonds change, is the minimum set required for graph modification from reactants to products [23]. Each atomic pair (a_u, a_v) in G_r is associated with a binary reaction label y_{uv}, and if the relationship between atomic pairs changes, y_{uv} is true. The node features of atomic a_u and atomic a_v are defined as f_u and f_v, and the feature of edge b_{uv} between nodes is defined as f_{uv}. The types of input features for chemical bonds are shown in Table 1.

Table 1: Chemical Bond Input Characteristics

Chemical bond characteristics	Characteristic Length	Feature Description	Feature Type
Chemical bond type	4	Which type of chemical bond belongs to single bond, double bond, triple bond, or aromatic bond [single, double, triple, aromatic]	One hot vector
conjugate	1	Is the chemical bond conjugated [0/1]	One hot vector
ring	1	Is the chemical bond on the ring [0/1]	One hot vector

3.2 Convolutional Neural Network Prediction

The improved graph convolutional network molecular simulation drug synthesis prediction algorithm is shown in Figure 1, which updates the features of nodes through formulas (1) to (5). In the formula, W_r, W_z, W_h is a cross layer shared matrix variable, $1 \le l \le 3$, $h_v^{(0)} = f_v^{(0)}$, $\sigma(\cdot)$ are sigmoid activation functions, and $N(v)$ is the set of all neighboring nodes of node v [24].

$$r_u^{(t)} = h_u^{(t-1)} * \sum_{i \in N(u)} \sigma\left(W_r \cdot \left[h_i^{(t-1)}, f_{in}\right]\right) \tag{1}$$

$$\tilde{z}_v^{(I)} = \sigma\left(W_z \cdot \left[h_u^{(I-1)}, f_{uv}\right]\right) \tag{2}$$

$$z_v^{(I)} = \sum_{u \in N(v)} \sigma\left(W_z \cdot \left[h_u^{(I-1)}, f_{uv}\right]\right) \tag{3}$$

$$\tilde{h}_v^{(t)} = \sum_{u \in N(v)} \tilde{z}_v^{(t)} * \tanh\left(W_h \cdot \left[r_u^{(l-1)}, f_{uv}\right]\right) \tag{4}$$

$$h_v^{(l)} = \left(1 - z_v^{(l)}\right) * h_v^{(l-1)} + \tilde{h}_v^{(l)} \tag{5}$$

Due to the fact that graph convolutional network algorithms are generally only suitable for updating hidden features of nodes within connected graphs, and reactants are generally composed of multiple molecules, this also means that the input of the algorithm is multiple

disconnected subgraphs. To solve the above problems, it is necessary to introduce a global attention mechanism that allows atomic feature information to flow between multiple disconnected subgraphs.

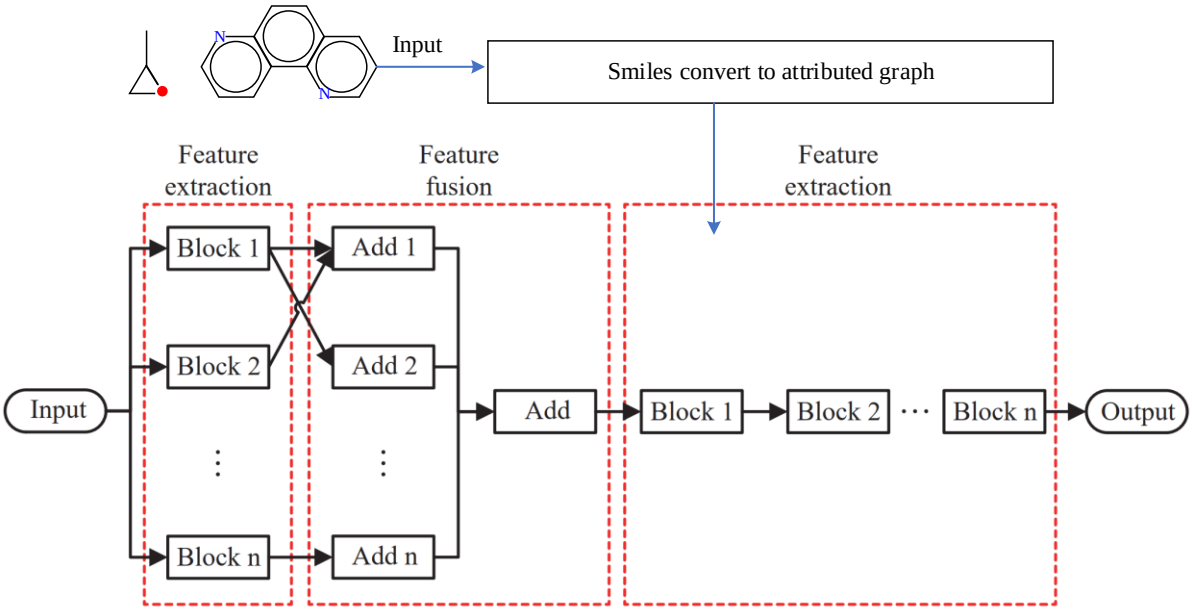

Figure 1: Improved Graph Convolutional Network Drug Synthesis Prediction Algorithm

Define α_{uv} as the attention value assigned to atom *u* by atom *v*, and the higher α_{uv}, the more likely atom v and atom u are to be correlated. c_v and c_u are contextual atomic hidden features obtained by improving graph convolutional networks, $c_v = h_v^{(3)}$, $c_u = h_u^{(3)}$. The reactivity value s_{uv} between atom *u* and atom *v* can be obtained by the following formulas (6) to (8), where $U_1, U_2, P_a, P_b, M_a, M_b, M_c$ is the matrix variable and b_{uv} is an additional eigenvector used to encode auxiliary backup information about the atomic pair, such as whether the atoms in the atomic pair belong to the same molecule or the chemical bond type between them [25].

$$\alpha_{uv} = \sigma\left(U_1^T \sigma\left(P_a c_u + P_a c_v + P_b b_{uv}\right)\right) \tag{6}$$

$$\tilde{c}_u = \sum_v \alpha_{uv} c_v \tag{7}$$

$$s_{uv} = \sigma\left(U_2^T \sigma\left(M_a \tilde{c}_u + M_a \tilde{c}_v + M_b b_{uv} + M_c c_u + M_c c_v\right)\right) \tag{8}$$

The reactivity value s_{uv} can predict the likelihood of chemical bond breakage and connection between atoms. Select atomic pairs with high reactivity values, list possible chemical bond combinations between these pairs, and use them to generate candidate drug synthesis products based on chemical constraints. In step two, the true reaction product is selected from the candidate products by improving the graph convolution differential network.

3.3 List candidate products

When enumerating reaction centers to generate drug synthesis candidate products, up to five chemical bonds can undergo changes in the same combination. In addition, when generating candidate products based on the reaction centers listed, the following chemical constraints must

also be met [26]:

(1) In the same combination, two atoms cannot have multiple types of bond changes directly, that is, only one type of chemical bond change can occur between atoms. For example, the chemical bond between two atoms cannot be both a single bond and a double bond.

(2) When testing the valency, the aromatic carbon and oxygen atoms adjacent to the aromatic nitrogen are considered as ortho hydroxy pyridines when they are connected to form a carbon group. Carbon based hydroxyl treatment is used to calculate the valence.

(3) When a phosphorus atom and an oxygen atom are about to form a double bond, odd number verification is performed on the valence of the phosphorus atom. When the sulfur atom and oxygen atom are about to form a double bond, the valence of the sulfur atom is checked for even numbers. When the nitrogen atom and phosphorus atom are about to form a double bond, odd number verification is performed on the valence of the nitrogen atom and phosphorus atom. When phosphorus atoms and carbon atoms are about to form a double bond, odd number verification is performed on the valence of phosphorus and carbon atoms.

3.4 Screening of candidate products

Step 2: First, input the drug synthesis candidate product p_i into the improved graph convolutional network to obtain the hidden feature $c_v^{(P_i)}$ of atom *v* in the candidate product p_i. Due to atomic conservation, the atoms in reactants and reaction products are mapped one-to-one. Therefore, the differential vector $d_v^{(P_i)}$ is defined by formula (9) to focus on the changes in atomic hidden features [27].

$$d_v^{(P_i)} = c_v^{(P_i)} - c_v^{(r)} \tag{9}$$

The differential vector $c_v^{(P_i)}$ only deviates from zero when it is close to the drug synthesis reaction center, so it focuses on processing the reaction center and its neighboring information. Input the differential vector $d_v^{(P_i)}$ into the improved graph convolutional network, and after *L* cycles, obtain the hidden feature $d_v^{(p_i,L)}$ of the differential vector. Add and pool it to obtain the reaction score, and take the drug synthesis candidate product with the highest reaction score as the final result. The reaction score is calculated using formula (3-10), where U_3 and *M* are matrix variables [28].

$$s\left(p_i\right) = U_3^T \operatorname{relu}\left(M \sum_{v \in p_i} d_v^{(p_i,L)}\right) \tag{10}$$

3.5 Predictive Reaction Process

Firstly, the hidden features of nodes are obtained by improving the graph convolutional network, and then the drug synthesis reaction activity values between atomic pairs are predicted through attention mechanism. The reaction center is located at the atomic pairs with higher reaction activity values. The corresponding pseudocode for improving graph convolutional networks is as follows:

Algorithm: Improved Graph Convolutional Network Algorithm

Input: Atomic adjacency table; Chemical bond adjacency table; Atomic input features; Chemical bond input characteristics

Output: Atomic Hidden Features

1. Map atomic input features through fully connected layer 1 and sigmoid activation function.

2. for i=0; i<3; i++ do

3. Collect adjacent atoms and chemical bond information at a distance of 1st order, and map them using fully connected layer 2 and sigmoid activation function.

4. Collect adjacent atoms and chemical bond information with a distance of 1st order, and update the atomic information through summation after passing through the fully connected layer 3 and sigmoid activation function.

5. Collect adjacent atoms and chemical bond information with a distance of 1st order and map them through fully connected layer 4 and tanh activation function.

6. $\mathbf{a} \leftarrow \mathbf{a} * (1 - \mathrm{sum}(z)) + \mathrm{sum}(z * \mathbf{r})$;

7. end for

8. return $\boldsymbol{a}$

4 Experimental analyses

Molecular simulations were conducted using the Discovery Studio 2.1 (DS2.1) software package from Accelrys in the United States, and all molecular mechanics and molecular dynamics simulations were performed under the CHARMm force field [29]. The initial structure of small molecules of fluoroquinolone drugs is generated by software packages. Target (B) was generated by homology modeling, and based on the information obtained from the experiment, the target Tyr82,179,274,376; Phe87,181,278,375 all have strong interactions with small molecules, so the active pocket is determined by combining experimental information with the active pocket provided by the Binding Site module. According to the software program, small molecules 1-5 are defined as ligands, and molecule B is defined as the receptor. The Flexible Docking module in the DS2.1 software package is used to flexibly dock the ligand with the receptor.

4.1 Prediction accuracy of drug synthesis

In the study of predicting drug synthesis reactions, the method model proposed in this article has unique design and considerations. Specifically, the model focuses only on the first 16 drug synthesis candidate reaction centers for prediction, which is a strategy made after balancing computational efficiency and prediction accuracy. In order to visually demonstrate the performance of the proposed method model, Figure 2 shows the experimental comparison results with the WLDN model on the USPTO test set [30]. Among them, Top-1 accuracy is a key indicator, which represents the probability of successful prediction after one prediction. During the evaluation process, the SMILES string expression of the model's predicted product will be carefully compared with the actual product one by one. If any matching problem occurs, it will be judged as a prediction failure. The horizontal axis in Figure 2 represents the number of template matches for the reaction equation. In the field of drug synthesis, the rarity of reactions is closely related to the number of template matches. The rarer the reaction, the fewer template matches there will be. Moreover, the accuracy of all models decreases as the number of template matches decreases, which is a common rule. However, the method model presented in this article demonstrates strong competitiveness and outperforms the WLDN model in any number of template matching scenarios. Especially when the number of template matches is at a low level, ranging from 5 to 49, or even less than 5, the advantage of our method is more significant. The leading advantages of Top-1 reach 2.79% and 2.98% respectively, which fully demonstrates the excellent performance of our method model in predicting complex and rare drug synthesis reactions.

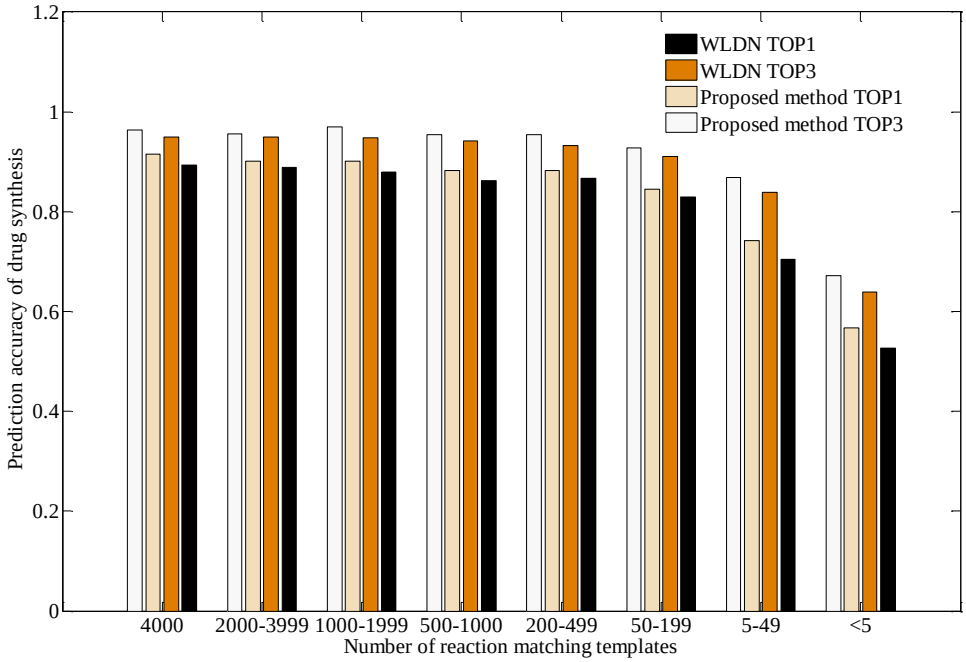

Figure 2: Coverage of reaction centers in the model

Table 2 presents in detail the probability of each model in the US Patent and Trademark Office test set successfully predicting reaction products under different prediction opportunities. In this test set, there were significant differences in the performance of different models, and the method proposed in this paper demonstrated unique advantages. The parameter size of the method model in this article is the smallest among all the compared models. However, its performance is excellent, with four key indicators of Top-1, Top-2, Top-3, and Top-5, all significantly better than other models, reaching 87.18%, 91.65%, 93.07%, and 94.15% respectively. This result fully demonstrates that the method proposed in this paper has strong predictive ability while ensuring the simplicity of the model. Following closely behind is the GGCN model, which did not adopt active learning training methods. Nevertheless, it achieved good results in the Top 1, Top 2, Top 3, and Top 5 indicators, reaching 86.72%, 91.13%, 92.97%, and 93.81%, respectively. However, there is still a certain gap compared to the method presented in this article. This further highlights the effectiveness and progressiveness of this method in the field of reaction product prediction.

Table 2: Accuracy of Reaction Product Prediction

Model name	Parameter scale	Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]
Seq2Seq	3.0×107	80.23	84.65	86.15	87.43
WLDN	2.6×106	85.56	90.48	92.70	93.38
GGCN	2.4×106	86.72	91.13	92.97	93.81
Proposed method	2.4×106	87.18	91.65	93.07	94.15

4.2 Analysis of protein action mechanism

Taking into account the different conformations of all ligand molecules, multiple different small molecule conformations obtained by flexible docking are comprehensively considered, taking into account the Libdockscore, Cdocker Energy and Cdocker Interaction Energy of the ligand receptor complex, the score of the professional scoring function in the software program, and the formation of hydrogen bonds between the two. Finally, the optimal drug small molecule

docking conformation is selected. Among the amino acid residues within the 10A range centered on compound 1, there are only two hydrophilic amino acid residues: aspartic acid (Asn378) and glutamine (Gln384), while the majority are hydrophobic amino acid residues, indicating that the drug molecule of compound 1 is in a stable hydrophobic environment. The results of molecular simulation indicate that the interaction between compound 1 and the target is not only hydrophobic, but there are some polar amino acid residues in the target molecule that can stabilize the position of compound 1 in the active pocket through hydrogen bonding and electrostatic interactions. The specific information of hydrogen bonds is shown in Table 3.

Table 3: Bond length and angle information of compound 1 forming an ammonia bond with the target

X…H-Y	d(X…H)	d(H-Y)	d(X…Y)	(XHY)
Ser247:OG-HG…O28	2.459	0.950	2.673	92.351
Ser247:OG-HG…O29	2.487	0.950	3.294	142.797
Met268:O…H54…O28	2.019	0.947	2.547	113.365
Ser272:OG-HG…O29	2.148	0.951	2.647	111.335
Ser272:N-HN…O28	2.405	1.007	3.002	117.163
Tyr372:OH-H53…N25	2.376	1.015	3.378	169.057

It can be seen that compound 1 forms six hydrogen bonds with the target. From the hydrogen bond parameters, the strong hydrogen bonds formed between the potassium ion channel protein A chain and Ser247, Tyr372 play a significant role in the stability of the conformation. According to theoretical calculations, when considering the implicit solvent model of the entire system as Generalized Born with Molecular Volume (GBMV), the binding energy (i.e. Binding Energy) between the amino acid residues in the active pocket and the small molecule Moxifloxacin of compound 1 is -47.03459 (Kcal/mol).

Compound 2 is a novel broad-spectrum antibacterial drug that modifies the functional group on the original side chain based on the structure of Compound 1, replacing the methyl group on the methoxy group with two fluorides. Following the research method of compound 1, the optimal conformation of compound 2 in the active pocket of potassium ion channel protein was obtained, as shown in Figure 4.3 (a). Both compound 2 and compound 1 can be located in the active pocket of potassium ion channel protein. Figure 4.3 (b) shows the interaction between compound 2 and the amino acid residues surrounding the active pocket. Among the amino acid residues within the 10A range around compound 2, only two are hydrophilic: Asn378 and Gln384. The rest are mostly hydrophobic amino acid residues, indicating that the hydrophobicity of the potassium ion channel protein active pocket where compound 2 and compound 1 are located is good. At the same time, it also indicates that the hydrophobic environment plays a certain role in the efficacy of compound 2 and 1 drug molecules.

When delving into the interaction mechanism between the target protein molecule and compound 2, we discovered a series of interesting and crucial hydrogen bonding phenomena. In addition to some known interaction modes, the hydrogen bonding between specific amino acid residues in the target protein molecule and compound 2 is particularly noteworthy. Ser272, an amino acid residue in the target protein molecule, exhibits unique chemical activity. The hydroxyl O atom contained in it, with its own electronegativity, is easy to establish hydrogen bonds with the fluorine atom on the benzene ring in compound 2. The formation of this hydrogen bond is not accidental, but based on the mutual adaptation of electronic distribution and spatial structure between the two. Fluorine atoms have strong electronegativity and can attract hydrogen on hydroxyl O atoms, forming stable hydrogen bonding connections. This interaction enhances the binding stability between proteins and compounds to a certain extent.

At the same time, the hydroxyl H atom in Ser344 is not to be outdone. It can form hydrogen bonds between F22 substituted methyl groups on complex 2. The special structure of F22 replacing methyl provides suitable binding sites for hydroxyl H atoms, enabling the smooth formation of hydrogen bonds. This hydrogen bonding interaction may play an important role in regulating the binding conformation between proteins and compounds, and affecting biological activity. In order to present the specific information of these hydrogen bonds more clearly, we have recorded them in detail in Table 3. Through Table 3, we can intuitively understand key parameters such as the types of hydrogen bonds, atoms involved in their formation, and bond lengths, providing important data support for further in-depth research on the interaction between target protein molecules and compound 2.

Table 3: Bond length and angle information of compound 2 forming hydrogen bonds with the target

X…H-Y	d(X…H)	d(H-Y)	d(X…Y)	(XHY)
SER344:OG-HG…F22	0.95164	2.4054	2.89496	111.653
SER272:OG-HG…F24	0.95170	2.1089	3.00071	155.464
N27-H53…OG:SER344	1.01161	1.86525	2.5968	126.427

From the table, we can see that the hydrogen bond formed between the amino acid residue SER272 and compound 2, Chinfloracin, is stronger. The formation of hydrogen bonds increases the stability of compound 2 in the active pocket of the target protein. According to theoretical calculations, when considering the implicit solvent model of the entire system as GBMV, the binding energy between the two is -74.60917 (Kcal/mol). The binding energy between compound 1 and the target protein is greater than that of compound 2, indicating that compound 2 is more stable than compound 1. However, the ICso (162.1) of compound 2 is larger than the ICs0 (76.9) of compound 1, which is inconsistent with the experiment. The possible reasons are as follows:

Before replacing with fluorine, the O and H atoms on the carboxyl group of the drug molecule were more likely to form hydrogen bonds with amino acid residues on the target protein, and the number of hydrogen bonds was relatively high. However, after replacing the methyl group with fluorine, the F atom on the benzene ring of the drug molecule and the F atom on the substituted methyl group were more likely to form hydrogen bonds with amino acid residues on the target protein. The position and functional group of hydrogen bonds formed between the compound molecule and the target protein changed, and the number of hydrogen bonds formed was three fewer than that of compound 1. So the combined effect of hydrogen bonding and binding energy resulted in an increase in the ICso (162.1) of compound 2 compared to compound 1 (76.9). Changing the substituents on the branched chain connected to the benzene ring in the structure of fluoroquinolone drugs has a certain impact on the activity of the drug.

5 Conclusion

5.1 Main tasks

This article focuses on key issues in drug synthesis, aiming to use machine learning to achieve more accurate and efficient reaction prediction and catalyst screening. The main work is carried out in the following two aspects: on the one hand, it deeply explores computer-aided drug synthesis design. This method uses computers as tools, integrates multidisciplinary knowledge, and utilizes theoretical simulations to guide and assist in the design, discovery, and synthesis of

new drug molecules. It has the advantages of fast speed, low cost, strong targeting, high efficiency, and highly active drug molecules designed, and has become an important method for new drug development and a necessary procedure for pharmaceutical companies. This method is mainly divided into three categories based on drug small molecule structure, receptor, and molecular dynamics. Multiple methods for determining protein structure and docking are also introduced, providing comprehensive theoretical basis and technical support for new drug design. On the other hand, in response to the problems of existing models in predicting chemical reaction products in drug synthesis, such as the inability to exhaustively use template regulations, atomic non conservation in sequence to sequence model prediction results, and the difficulty of traditional graph convolutional networks in effectively obtaining global information, an improved graph convolutional network model is proposed. The model first converts the reactants into feature matrices, uses an improved graph convolutional network and attention mechanism to predict candidate reaction centers, then generates candidate products based on chemical constraints enumeration, and finally evaluates and ranks the candidate products using an improved graph convolutional differential network. The experimental results show that on the USPTO test set, the proposed method outperforms the WLDN model in any number of template matches, especially when the number of template matches is low; In terms of the accuracy index of reaction product prediction, the method proposed in this paper has the smallest parameter scale, but it is significantly better than other models in the Top 1, Top 2, Top 3, and Top 5 indicators. In addition, the analysis of protein action mechanism indicates that changing the substituents on the branched chain connected to the benzene ring in the structure of fluoroquinolone drugs can affect drug activity.

5.2 Future research directions

In the field of drug synthesis, which is full of challenges and opportunities, although this article has achieved certain phased results, such as breakthroughs in simplifying some drug synthesis routes and exploring preliminary design ideas for new drug molecules, there are still many urgent problems to be solved in the two core directions of optimizing drug synthesis routes and designing new drug molecules. At present, drug synthesis routes often have drawbacks such as cumbersome steps, high costs, low yields, and unfriendly environments, which seriously restrict the large-scale production and widespread application of drugs. In terms of new drug molecular design, there are challenges such as improving drug targeting, reducing toxic side effects, and enhancing drug stability. To overcome these challenges, in-depth research can be conducted from the following two key aspects in the future. On the one hand, it is necessary to continuously optimize and improve graph convolutional network models. Currently, graph convolutional networks have shown great potential in the field of drug synthesis, but their accuracy and generalization ability still need to be improved in predicting the products and selectivity of complex drug synthesis reactions. By introducing more advanced algorithms, increasing data diversity, and optimizing model structure, the predictive accuracy of the model can be further improved, providing a more reliable basis for optimizing drug synthesis routes. On the other hand, it is crucial to strengthen deep cross disciplinary integration with disciplines such as chemistry and biology. Chemistry can provide rich information on reaction types and material structures, while biology helps to deepen the understanding of the interaction mechanisms between drugs and organisms. By fully utilizing interdisciplinary knowledge and experimental data, in-depth exploration of the mechanisms and laws of drug synthesis reactions can provide a more solid theoretical foundation and more effective design strategies for the molecular design of new drugs. Through continuous exploration and innovation, we aim to promote the development of drug synthesis technology to a higher level. We hope to achieve green and efficient drug synthesis routes, design more safe and effective new drugs, provide

stronger support for new drug research and development, and ultimately benefit patients.

About The Author

Jinchao Mu was born in Bozhou, Anhui. China, in 1978. He obtained a master's degree from Nanjing Tech University in China. He currently working at the College of Chemical Engineering, Xuzhou College of Industrial Technology. His main research direction is Drug synthesis and drug analysis.

Chunfen Liu was born in Hangu, Tianjin. China, in 1978. She obtained a master's degree from Southwest University in China. She currently working at the College of Chemical Engineering, Xuzhou College of Industrial Technology. Her main research direction is Drug synthesis and drug analysis.

Lianxin Liu was born in Jilin, Jilin. China, in 1971. She obtained a master's degree from Xi'an Technological University in China. She currently working at the Basic Course Teaching Department, Xuzhou College of Industrial Technology. Her main research direction is Chemical calculation and optimization.

Jun Jiang was born in Suqian, Jiangsu. China, in 1989. He obtained a bachelor's degree from China Pharmaceutical University in China. He currently working at the Quality Assurance Department, Wisdom Zenith Pharmaceutical Co., Ltd. His main research direction is Pharmaceutical Analysis.

Xu Liu was born in Jilin, Jilin. China, in 1972. He obtained a docter's degree from China University of Mining and Technology in China. He currently working at the College of Chemical Engineering, Xuzhou College of Industrial Technology. His main research direction is Drug synthesis.

References

[1] Mei S, Roopashree R, Altalbawy F M A, et al. Synthesis, characterization, and applications of starch-based nano drug delivery systems for breast cancer therapy: A review[J]. International Journal of Biological Macromolecules, 2024: 136058.

[2] Karami M H, Abdouss M, Rahdar A, et al. Graphene quantum dots: Background, synthesis methods, and applications as nanocarrier in drug delivery and cancer treatment: An updated review[J]. Inorganic Chemistry Communications, 2024, 161: 112032.

[3] Fonseca M, Jarak I, Victor F, et al. Polymersomes as the next attractive generation of drug delivery systems: definition, synthesis and applications[J]. Materials, 2024, 17(2): 319.

[4] Mehraji S, DeVoe D L. Microfluidic synthesis of lipid-based nanoparticles for drug delivery: Recent advances and opportunities[J]. Lab on a Chip, 2024, 24(5): 1154-1174.

[5] Rakshit P, Giri T K, Mukherjee K. Research progresses on carboxymethyl xanthan gum: Review of synthesis, physicochemical properties, rheological characterization and applications in drug delivery[J]. International Journal of Biological Macromolecules, 2024: 131122.

[6] Zengin Kurt B, Öztürk Civelek D, Cakmak E B, et al. Synthesis of sorafenib− ruthenium complexes, investigation of biological activities and applications in drug delivery systems as an anticancer agent[J]. Journal of Medicinal Chemistry, 2024, 67(6): 4463-4482.

[7] Yuriy K, Kusdemir G, Volodymyr P, et al. A biochemistry-oriented drug design: synthesis, anticancer activity, enzymes inhibition, molecular docking studies of novel 1, 2, 4-triazole derivatives[J]. Journal of Biomolecular Structure and Dynamics, 2024, 42(3): 1220-1236.

[8] Ghasemi S, Dabirian S, Kariminejad F, et al. Process optimization for green synthesis of silver nanoparticles using Rubus discolor leaves extract and its biological activities against multi-drug resistant bacteria and cancer cells[J]. Scientific reports, 2024, 14(1): 4130.

[9] Bera S, Kabadwal L M, Banerjee D. Harnessing alcohols as sustainable reagents for late-stage functionalisation: synthesis of drugs and bio-inspired compounds[J]. Chemical Society Reviews, 2024, 53(9): 4607-4647.

[10] Ma F, Li Y, Cai M, et al. ML162 derivatives incorporating a naphthoquinone unit as ferroptosis/apoptosis inducers: design, synthesis, anti-cancer activity, and drug-resistance reversal evaluation[J]. European Journal of Medicinal Chemistry, 2024, 270: 116387.

[11] Khan S, Hussain R, Khan Y, et al. Novel bis-thiazole-thiazolidinone hybrid derivatives: Synthesis, structural properties and anticholinesterase bioactive potential as drug competitor based on docking studies[J]. Journal of Molecular Structure, 2024, 1303: 137417.

[12] Bakhshi V, Poursadegh H, Amini-Fazl M S, et al. Synthesis and characterization of bio-nanocomposite hydrogel beads based on magnetic hydroxyapatite and chitosan: a pH-sensitive drug delivery system for potential implantable anticancer platform[J]. Polymer Bulletin, 2024, 81(8): 7499-7518.

[13] El-Saghier A M, Enaili S S, Abdou A, et al. An efficient eco-friendly, simple, and green synthesis of some new spiro-N-(4-sulfamoyl-phenyl)-1, 3, 4-thiadiazole-2-carboxamide derivatives as potential inhibitors of SARS-CoV-2 proteases: drug-likeness, pharmacophore, molecular docking, and DFT exploration[J]. Molecular Diversity, 2024, 28(1): 249-270.

[14] Sahoo R, Pramanik B, Mondal S, et al. A Highly Chemically Robust 3D Interpenetrated MOF Heterogeneous Catalyst for the Synthesis of Hantzsch 1, 4-Dihydropyridines and Drug Molecules[J]. Small, 2024, 20(25): 2309281.

[15] Kheradmandfard M, Sadeghian A, Kouhi M, et al. High Drug Loading Mesoporous Hydroxyapatite Nanoparticles for Periodontal Bone Regeneration: A Facile Microwave-assisted Synthesis Approach[J]. Surfaces and Interfaces, 2025: 106626.

[16] Ogbuagu O O, Mbata A O, Balogun O D, et al. Sustainable pharmaceutical supply chains: Green chemistry approaches to drug production and distribution[J]. IRE Journals, 2024, 8(4): 761-767.

[17] He Y, Wang H, Yan Y, et al. Facile synthesis of nitrogen-doped carbon dots for ultrasensitive detection of anticancer drug gefitinib based on IFE[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024, 310: 123942.

[18] Gul S, Alam A, Assad M, et al. Exploring the synthesis, molecular structure and biological activities of novel Bis-Schiff base derivatives: A combined theoretical and

experimental approach[J]. Journal of Molecular Structure, 2024, 1306: 137828.

[19] Cai Y, Li D, Peng S, et al. Synthesis of dual-responsive carboxymethyl cellulose–based nanogels for drug delivery applications[J]. Colloid and Polymer Science, 2025, 303(2): 287-300.

[20] Altamimi M, Syed S A, Tuzun B, et al. Synthesis biological evaluation and molecular docking of isatin hybrids as anti-cancer and anti-microbial agents[J]. Journal of enzyme inhibition and medicinal chemistry, 2024, 39(1): 2288548.

[21] Akhtar H, Amara U, Mahmood K, et al. Drug carrier wonders: Synthetic strategies of zeolitic imidazolates frameworks (ZIFs) and their applications in drug delivery and anti-cancer activity[J]. Advances in Colloid and Interface Science, 2024, 329: 103184.

[22] Afanasenko A M, Wu X, De Santi A, et al. Clean synthetic strategies to biologically active molecules from lignin: a green path to drug discovery[J]. Angewandte Chemie International Edition, 2024, 63(4): e202308131.

[23] Pal R, Teli G, Akhtar M J, et al. Synthetic product-based approach toward potential antileishmanial drug development[J]. European Journal of Medicinal Chemistry, 2024, 263: 115927.

[24] Vodyashkin A, Stoinova A, Kezimana P. Promising biomedical systems based on copper nanoparticles: synthesis, characterization, and applications[J]. Colloids and Surfaces B: Biointerfaces, 2024: 113861.

[25] Iqbal T, Khan S, Rahim F, et al. Benzothiazole based sulfonamide scaffolds as active inhibitors of alpha-Amylase and alpha-glucosidase; synthesis, structure confirmation, In Silico molecular docking and ADME analysis[J]. Journal of Molecular Structure, 2024, 1309: 138074.

[26] Shukla V, Ahmad M, Siddiqui K A. Synthesis of dual functional Zn (II) MOF for colorimetric detection of norfloxacin and photocatalytic degradation of ornidzole drugs in aqueous medium[J]. Polyhedron, 2024, 260: 117078.

[27] Vodyashkin A, Stoinova A, Kezimana P. Promising biomedical systems based on copper nanoparticles: synthesis, characterization, and applications[J]. Colloids and Surfaces B: Biointerfaces, 2024: 113861.

[28] Xu K, Ren X, Wang J, et al. Clinical development and informatics analysis of natural and semi-synthetic flavonoid drugs: A critical review[J]. Journal of Advanced Research, 2024, 63: 269-284.

[29] Soyler D, Dolgun V, Kurbanoglu S, et al. Synthesis and design of functional fullerenol-based electrochemical nanobiosensor to examine the inhibition effects of anti-Alzheimer drug active pyridostigmine on acetylcholinesterase[J]. Measurement, 2025, 253: 117494.

[30] Banday A H, ul Azha N, Farooq R, et al. Exploring the potential of marine natural products in drug development: A comprehensive review[J]. Phytochemistry Letters, 2024, 59: 124-135.