Modern Health Science; Vol. 8, No. 4; 2025
ISSN 2576-7291 E-ISSN 2576-7305
https://doi.org/10.30560/mhs.v8n4p1

An AI Fusion Model Based on Molecular Docking and Large Language Models—— Used to Predict the Interaction Between Traditional Chinese Medicine and Disease

Tan Chit Fong[1a], Naiwen Zhang[2b], Jiale Zhang[1c], Ma Chon Hou[4d], Nanxin Ruan[3e], Zixuan Tang[3f], Yunfan Fan[5g], Yutao Xia[6h], Ruizi Li[3i]

[1] Macao Polytechnic University, Faculty of Health Sciences and Sports, Macao

[2] Macao Polytechnic University, Faculty of Business, Macao

[3] Macau University of Science and Technology, Faculty of Hospitality and Tourism Management, Macao

[4] Nanjing Medical University, The Second School of Clinical Medicine, China

[5] Beijing Wodong Tianjun Information Technology Co., Ltd., China

[6] The Hong Kong Polytechnic University, School of Hotel and Tourism Management, Hong Kong

Correspondence: Naiwen Zhang, Macao Polytechnic University, Faculty of Business, Macao. E-mail: [b]p2411161@mpu.edu.mo; [a]p2306295@mpu.edu.mo; [c]p2321480@mpu.edu.mo; [d]William20304@njmu.edu.cn; [e]1230033551@student.must.edu.mo; [f]1230033607@student.must.edu.mo; [g]fanyunfanl@jd.com; [h]yutao.xia@connect.polyu.hk; [i]1220027851@student.must.edu.mo

Received: October 3, 2025; Accepted: October 17, 2025; Published: October 18, 2025

Abstract

Traditional Chinese Medicine (TCM) is a profound and sophisticated medical system. However, the complexity of multi-component interactions and limited computational methods pose significant challenges for optimizing TCM formulations and developing new drugs. This study aims to develop an integrated model that combines molecular docking technology with large language models to achieve high-throughput prediction of interactions between TCM herbs and diseases. We first collected active components from 500 TCM herbs in the TCMSP database and obtained 3D structural information on 100 disease-related targets from the RCSB PDB database. Using OpenBabel to convert SMILES strings into 3D molecular structures, we performed molecular docking calculations with AutoDock Vina. The study defined effective binding as interactions with binding energies ≤-7.0 kcal/mol, yielding 12,408 valid herb-target pairs. On the basis of these data, we trained a Transformer-based neural network model for predicting new TCM–disease interactions. The experimental results demonstrated that the integrated model achieved excellent performance, with an AUC of 0.984 and an AUPR of 0.982 on the test set, significantly outperforming standalone molecular docking or machine learning methods. This proposed integrated model can substantially accelerate modernization research in TCM, providing a powerful tool for elucidating TCM mechanisms and advancing drug development.

Keywords: Traditional Chinese Medicine, Molecular Docking, TCM–Disease interaction, CADD, Transformer

1. Introduction

Traditional Chinese medicine (TCM), a valuable cultural heritage of the Chinese nation, has a history of clinical application spanning thousands of years and plays a vital role in disease prevention and treatment. However, TCM relies on complex herbal formulations, with therapeutic effects typically achieved through multiple bioactive compounds and targets (Zhang et al., 2014). [1] The multi-component and multi-target characteristics of TCM make its mechanisms of action exceptionally complex. Moreover, the lack of effective computational tools to predict interactions between TCM herbs and diseases has hindered modern TCM drug development.

While machine learning methods demonstrate strong pattern recognition capabilities, their predictions often lack structural medical-level explanations, making it difficult to provide mechanistic insights. Recent advancements in artificial intelligence have revealed that large language models (LLMs) have potential for simulating biological sequences and interactions (Yuan et al., 2025). [2] The maturation of computer-aided drug design (CADD) technology has provided new technical approaches for deciphering complex systems in traditional Chinese medicine. Numerous researchers have attempted to apply computational biology methods to TCM research.

Currently, molecular docking remains the gold standard for predicting the binding affinity between ligands and targets (Trotter and Olsen, 2010). [3] Researchers have utilized Surflex-Dock to conduct molecular docking studies on xanthine oxidase inhibitors and epoxy enzymes, identifying potential therapeutic agents for gout. However, most existing research methods lack comprehensive consideration of the holistic nature and systematic approach inherent in traditional Chinese medicine.

To overcome these limitations, this study proposes an integrated model that combines molecular docking technology with large language models to achieve high-precision prediction of interactions between traditional Chinese medicine (TCM) components and disease targets. The framework employs SMILES-based 3D structure generation, AutoDock Vina for docking, and Transformer models for interaction prediction. This approach not only considers the energy characteristics of interactions but also integrates the sequence and structural information of TCM components and disease targets, enabling a more comprehensive evaluation of the pharmacological effects and potential applications of TCM formulations.

2. Related Studies

2.1 Application of Computer Aided Drug Design in Research of Traditional Chinese Medicine Research

Computer-Aided Drug Design (CADD) technology has become a vital component in modern pharmaceutical development, significantly shortening R&D cycles and reducing costs through computational simulation methods. In traditional Chinese medicine (TCM) research, CADD is applied primarily to screen for active components, predict target molecules, and investigate pharmacological mechanisms (Xia et al., 2021). [4]

Feng Rongkai and colleagues conducted flexible molecular docking studies on xanthine oxidase inhibitors that target cyclooxygenases (COX-1 and COX-2) via tools such as Surflex-Dock, FAF-Drugs2, Toxtree, and PharmMapper. They performed ADMET predictions, carcinogenicity assessments, and potential drug target identification. From 403 compounds, they identified two highly promising lead compounds (ChEMBL 170303 and 460160) that could treat hyperuricemia and inflammation simultaneously (Feng et al., 2013). [5] These studies highlight the significant role of CADD technology in discovering active components in traditional Chinese medicine.

2.2 Application of Molecular Docking Technology in Drug Discovery

Molecular docking technology, a core component of Chemical Structure Activity Discovery (CADD), is employed to predict binding patterns and affinities between small-molecule ligands and biomolecular receptors (Li et al., 2020). [6] Traditional docking methods such as AutoDock Vina utilize empirical scoring functions to evaluate binding energies, with these functions typically accounting for key factors such as hydrogen bonds, hydrophobic interactions, and electrostatic interactions.

However, traditional molecular docking methods have inherent limitations. Classical scoring functions typically employ linear summation models, which fail to accurately characterize the intricate relationships between biological activity and protein–ligand interactions (Shen Chao, 2022). [7] In recent years, machine learning-based scoring functions (MLSF) have emerged as alternative approaches. These methods do not require predefined functional forms; instead, they leverage machine learning algorithms to identify complex interaction patterns from data, demonstrating greater flexibility than classical approaches do.

2.3 Application of Large Language Models in Drug Discovery

Large language model technology is increasingly being applied in drug discovery, particularly in areas such as compound activity prediction, drug repositioning, and reaction prediction. A research team proposed a benchmarking method based on molecular docking to evaluate the performance of molecular generation models. By combining docking scores of generated molecules with structural diversity analysis, they effectively assessed molecular generation models and identified current limitations in these models (Tobiasz Cieplinski et al., 2023). [8]

In the field of traditional Chinese medicine (TCM) research, large language model technology is employed primarily for syndrome classification, formula optimization, and pharmacological effect prediction. However, existing approaches predominantly rely on one-dimensional or two-dimensional compound information, with insufficient consideration of three-dimensional structural data. This limitation significantly compromises the accuracy and reliability of predictions.

3. Methodology

3.1 Data Collection and Processing

The data collected in this study covered a variety of aspects, such as TCM components, disease targets and known interactions, as shown in Table 1. The diversity of data sources and strict quality control were the basic guarantees of this study.

Table 1. Data sources and statistical information

Data type	Source	Quantity	Edition	Remarks
Chinese medicinal ingredients	TCMSP	500 kinds of Chinese medicine	2014	Screening OB≥30%, DL≥0.18
Active ingredient	TCMSP	23,294 compounds	2014	-
Targeting structures	RCSB PDB	100 targets	2023	Including kinases, GPCRs and so on
Known interactions	STITCH、BindingDB	7591 pairs	2023	Used as a positive sample

The TCMSP database was utilized to collect information on the active components of 500 commonly used traditional Chinese medicines, including physicochemical parameters such as SMILES strings, molecular weights, logP (logarithmic partition coefficient), and oral bioavailability (OB) (Ru et al., 2014). [9] This study selected the top five active components from each herbal material on the basis of the criteria of ≥30% oral bioavailability (OB%) and ≥18% drug likeness (DL%), [10] ultimately identifying 5,738 potential active components for subsequent analysis.

The RCSB PDB database was utilized to obtain 100 three-dimensional structural datasets of target proteins associated with common diseases. The selected targets encompassed multiple crucial drug target families, including kinases, G protein-coupled receptors (GPCRs), nuclear receptors, and ion channels (Berman et al., 2000). [11] For each target, this study extracted binding pocket information and prepared receptor files for molecular docking.

The STITCH and BindingDB databases were used to collect 7,591 pairs of known component–target interactions. These interactions served as positive samples for model training, whereas randomly paired components that did not exist in the known interactions were designated negative samples.

3.2 Molecular Docking Process

Molecular docking is one of the core aspects of this study. This study adopted a high-throughput docking strategy to evaluate the interaction potential between TCM components and disease targets. The docking process mainly includes the following steps:

(1) Molecular preparation: The SMILES strings of the components were converted into 3D molecular structures in mol2 format via the OpenBabel tool, followed by energy minimization with the MMFF94 force field. For target proteins, water molecules and original ligands were removed, hydrogen atoms were added, and Gasteiger charges were assigned (O'Boyle et al., 2011). [12]

(2) Binding pocket definition: For each target, if there is ligand information, the docking frame is defined with the original ligand as the center; if there is no original ligand, the potential binding site is predicted by the binding pocket prediction software FPocket.

(3) Molecular docking: Flexible docking was performed via AutoDock Vina with the following docking parameters: exhaustiveness=32, energy range=4, and maximum output mode=20. The scoring function of Vina is as follows:

$$E_{vina} = w_{gauss1} \times E_{gauss1} + w_{gauss2} \times E_{gauss2} + w_{repulsion} \times E_{repulsion} + w_{hydrophobic} \times E_{hydrophobic} + w_{hydrogen} \times E_{hydrogen} + E_{tors}$$

Among them, E_gauss1, E_gauss2, E_repulsion, E_hydrophobic and E_hydrogen correspond to the interaction energy of hydrogen bond, hydrophobic interaction and electrostatic interaction, respectively, w is the corresponding weight parameter, and E_tors represents the torsional free energy penalty term. [13]

(4) Results analysis: The optimal docking conformation and binding energy of each TCM component–disease target pair were extracted, and strongly binding pairs with a Vina binding energy ≤-7.0 kcal/mol were retained as the threshold. A total of 12,408 effective component–target pair combinations were obtained for subsequent training of the large language model.

3.3 Large Language Model Construction

To fully leverage molecular docking results and compositional-target characteristic information, this study developed a Transformer-based neural network model for predicting interactions between traditional Chinese medicine components and disease targets. The model architecture consists of three main components: a feature encoder, a cross-attention mechanism, and a prediction head.

(1) Feature Extraction:

For Chinese herbal ingredients, this study uses a SMILES based Transformer encoder to obtain the feature representations of the ingredients:

$$H_{compound} = TransformerEncoder(X_{compound})$$

For the target protein, this study used a pre-trained protein language model (ESM-2) to obtain the feature representation of the target:

$$H_{target} = TransformerEncoder(X_{target})$$

(2) Cross-attention mechanism: To capture the interaction information between components and targets, this study introduces a cross-attention mechanism:

$$H_{cross} = CrossAttention(H_{compound}, H_{target})$$

(3) Prediction head: The output of the cross-attention mechanism is input into a multi-layer perceptron (MLP) to predict the interaction probability between components and targets:

$$y_{pred} = Sigmoid\big(MLP(H_{cross})\big)$$

(4) Model training: The model is trained with a binary cross entropy loss function and the Adam optimizer for parameter optimization. The learning rate is set to 0.001, the batch size is 64, and the number of training rounds is 100.

This study used five-fold cross-validation to evaluate model performance and ensure the stability and reliability of the evaluation results. [14] At the same time, this study adopted an early stopping strategy to prevent overfitting.

4. Experimental Results

4.1 Model Performance Evaluation

To comprehensively evaluate the performance of the fusion computing framework proposed in this study, we conducted comparative experiments and ablation experiments, evaluating it against multiple existing methods. The assessment metrics included Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPR), Accuracy (ACC), and F1 score, as detailed in Table 2.

Table 2. Comparison of the performance of different methods on the test set

Method	AUC	AUPR	ACC	F1-Score
Molecular docking only	0.872	0.851	0.821	0.835
Pure machine learning	0.901	0.883	0.862	0.871
Tradition MLSF	0.932	0.915	0.894	0.901
Integration model	0.984	0.982	0.952	0.951

The experimental results demonstrate that the proposed fusion model (Our) significantly outperforms other comparison methods in all evaluation metrics, including those using only molecular docking (Docking Only), pure machine learning (ML Only), and traditional machine learning scoring functions (Traditional MLSF). The relatively poor performance of standalone molecular docking methods is due primarily to their inability to adequately capture the complexity of traditional Chinese medicine–disease interactions. While machine learning approaches outperform molecular docking when used alone, their effectiveness remains constrained by the quantity and quality of training data. Although the traditional MLSF method outperforms both approaches, it still falls short compared with our proposed fusion method.

The superior performance of this research methodology stems from the complementary synergy between structural information derived from molecular docking and the robust pattern recognition capabilities of large language models. Our framework effectively integrates structural biology with sequence-based artificial intelligence to predict interactions between traditional Chinese medicine (TCM) compounds and diseases. Molecular docking provides physicochemical binding patterns and biological validation, whereas large language models enable the

extraction of complex interaction patterns from extensive SMILES sequences. This scalable approach demonstrates broad applicability across various multi-component drug systems.

4.2 Case Analysis

To further verify the practical value of this research method, two case studies were conducted to explore the potential mechanism of TCM formulas in treating specific diseases.

Case 1: Analysis of the mechanism by which danshen treats cardiovascular diseases

This study analyzed the interactions between active components in Salvia miltiorrhiza and cardiovascular disease-related targets. Salvia miltiorrhiza is a commonly used traditional Chinese medicine for the treatment of cardiovascular diseases, and its main active components include salvianolic acid IIA, salvianolic acid B, and cryptosalviolic acid.

The analysis revealed that Danshenone IIA has strong binding affinity for multiple cardiovascular disease targets, with binding energies of -9.2 kcal/mol for angiotensin-converting enzyme (ACE) and -8.7 kcal/mol for clotting factor Xa. These findings align with existing pharmacological studies, confirming the molecular mechanisms underlying the antithrombotic and blood pressure-lowering effects of Danshenone IIA.

In addition, the results of this study revealed several new potential targets, such as prostaglandin H2 synthase (PTGS2) and phosphodiesterase 5A (PDE5A), which are closely related to cardiovascular function regulation and may constitute a new mechanism for the treatment of cardiovascular diseases.

Case 2: Analysis of the mechanism of the antibacterial action of Coptis chinensis

This study also analyzed the interactions between alkaloid components in Coptis chinensis and bacterial targets. Coptis chinensis is a commonly used heat-clearing and dampness-drying drug, and its main active ingredients include berberine, berberine, and bamaatin.

The results revealed that berberine strongly bound to bacterial DNA gyrase and topoisomerase IV, with binding energies of -9.5 kcal/mol and -9.1 kcal/mol, respectively. This finding is consistent with the known mechanism of antibiotics, and the reliability of this study method was confirmed.

In addition, the results of this study revealed that berberine may strongly interact with bacterial quorum sensing system-related proteins, which indicates that Coptis chinensis may exert antibacterial effects by interfering with bacterial quorum sensing, providing a new perspective for understanding the antibacterial mechanism of Chinese medicine.

4.3 Ablation Experiments

To evaluate the contribution of each component in this study method, ablation experiments were conducted to observe the change in model performance by removing or replacing some components, as shown in Table 3.

Table 3. Ablation experimental results

Model variants	AUC	AUPR	ACC	F1-Score	Number of parameters (M)
Complete pattern	**0.984**	**0.982**	**0.952**	**0.951**	48.5
Nonmolecular docking features	0.934	0.921	0.897	0.903	45.2
No cross attention	0.961	0.953	0.925	0.928	41.7
Use the CNN encoder	0.972	0.966	0.938	0.941	47.1
Use a RNN encoder	0.968	0.962	0.932	0.936	46.3

The experimental results show that molecular docking features contribute significantly to the performance of the model, and the performance of the model decreases significantly after removal (the AUC decreases from 0.984 to 0.934), which confirms the importance of the structural information provided by molecular docking for predicting component–target interactions.

The cross attention mechanism also has an important impact on the model performance, and the performance of the model decreases significantly after removal (the AUC decreases from 0.984 to 0.961), which indicates that the fine-grained interaction information between components and targets is crucial for prediction accuracy.

In addition, this study also attempts to use CNN and RNN instead of Transformer as feature encoders and finds that the performance of Transformer architecture is better than CNN and RNN, which is due to its powerful sequence modeling ability and long-distance dependence capture ability.

5. Conclusions and Prospects

This study proposes an integrated model that combines molecular docking technology with large language models. The experimental results demonstrate that this approach significantly outperforms existing single-method approaches, enabling efficient and accurate prediction of interactions between active components in traditional Chinese medicine (TCM) and disease targets. The practical value of this research lies in accelerating the elucidation of TCM mechanisms of action and advancing drug development processes, providing a powerful tool for modernizing traditional Chinese medicine and pharmaceutical innovation. Through computational simulation methods, studies can rapidly screen active components in TCM, predict their target molecules, and evaluate their therapeutic potential, thereby reducing the time and costs associated with experimental screening.

The main contributions of this study include the following:

(1) A large-scale standardized dataset of interactions between TCM components and disease targets was constructed;

(2) A fusion model based on molecular docking and a large language model was developed;

(3) The excellent performance of this method in predicting the interaction between Chinese medicine components and disease targets was verified;

(4) The practical application value of this method in TCM modernization research is demonstrated via a case study.

However, this study has several limitations. First, the accuracy of molecular docking is constrained by the scoring function's precision, although we employed large language models to compensate for this shortcoming. Second, our model primarily considers single-component-target interactions, failing to adequately reflect the holistic regulatory characteristics of multi-component, multi-target compounds in traditional Chinese medicine. Finally, model predictions require experimental validation before being applicable to actual drug development.
Future research directions include the following:

(1) Build a more accurate molecular docking scoring function;

(2) Develop a multi-component and multi-target network pharmacological model considering the integrity of TCM;

(3) Integrate pharmacokinetic and pharmacodynamic parameters to improve the physiological relevance of the prediction;

(4) Experimental verification research is carried out to verify the reliability of the model predictions.

In conclusion, the fusion computing framework proposed in this study provides a new idea and method for the modernization of TCM research, which is expected to accelerate the analysis of TCM mechanism and the process of new drug development and promote the international development of TCM.

References

[1] Zhang, A., Sun, H., & Wang, X. (2014). Potentiating therapeutic effects by enhancing synergism based on active constituents from traditional medicine. *Phytotherapy Research, 28*(4), 526–533. https://doi.org/10.1002/ptr.5032

[2] Yuan, S., Zhou, Z., Jin, X., Zhuo, L., & Li, K. (2025). Enhancing herbal medicine-drug interaction prediction using large language models. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2025.3558667

[3] Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455–461. https://doi.org/10.1002/jcc.21334

[4] Xia, Y. Y., He, Y. Y., Zhang, H. Y., Liu, S. H., Yu, L. Z., & Qin, R. A. (2021). Network pharmacology-based investigation on the targets and mechanisms of Ganmao Qingre Granules against common cold. *Studies of Trace Elements and Health, 38*(4), 1–3.

[5] Feng, R. K., Meng, L. R., Yue, Z. Q., Lin, S. L., & Wang, Q. T. (2013). Molecular docking and drug-likeness evaluation of xanthine oxidase inhibitors targeting cyclooxygenase. *Computers and Applied Chemistry,*

30(12), 1383–1388.

[6] Li, B. X., Zhu, X. L., Sun, G. Y., & Dong, Y. (2020). Molecular docking study on Mongolian medicine compound Sendeng-4. *China Journal of Traditional Chinese Medicine and Pharmacy, 35*(12), 5293–5296.

[7] Shen, C. (2022). *Research on protein-ligand scoring methods based on machine learning* [Doctoral dissertation, Zhejiang University].

[8] Cieplinski, T., Danel, T., Podlewska, S., & Jastrzebski, S. (2023). Generative models should at least be able to design molecules that dock well: A new benchmark. *Journal of Chemical Information and Modeling, 63*(11), 3238–3247. https://doi.org/10.1021/acs.jcim.2c01355

[9] Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Guo, J., Li, L., Ding, Y., & Yang, L. (2014). TCMSP: A database of systems pharmacology for drug discovery from herbal medicines. *Journal of Cheminformatics, 6*(1), 13. https://doi.org/10.1186/1758-2946-6-13

[10] Gu, Y. Y., Zhang, C. F., Zhang, H. H., Wu, Z., Zou, N. T., Mo, Q. Y., Zhao, S. B., & Wan, C. P. (2023). Network pharmacology-based study on the effective mechanism of Guaizijin against cerebral ischemia-reperfusion injury. *Yunnan Journal of Traditional Chinese Medicine and Materia Medica, 44*(12), 38–47. https://doi.org/10.16254/j.cnki.53-1120/r.2023.12.019

[11] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research, 28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

[12] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics, 3*(1), 33. https://doi.org/10.1186/1758-2946-3-33

[13] Cieplinski, T., Danel, T., Podlewska, S., & Jastrzebski, S. (2023). Generative models should at least be able to design molecules that dock well: A new benchmark. *Journal of Chemical Information and Modeling, 63*(11), 3238–3247. https://doi.org/10.1021/acs.jcim.2c01355

[14] Liu, X. W. (2024). *Construction of a drug prediction model for diabetes treatment based on machine learning* [Master's thesis, University of Electronic Science and Technology of China]. https://doi.org/10.27005/d.cnki.gdzku.2024.003370