scientific reports

Check for updates

OPEN

Accurate prediction of drug-target interactions in Chinese and western medicine by the CWI-DTI model

Ying Li[1], Xingyu Zhang[1], Zhuo Chen[1], Hongye Yang[1], Yuhui Liu[1], Huiqing Wang[1], Ting Yan[2], Jie Xiang[1] & Bin Wang[1]✉

Accurate prediction of drug-target interactions (DTIs) is crucial for advancing drug discovery and repurposing. Computational methods have significantly improved the efficiency of experimental predictions for drug-target interactions in Western medicine. However, accurately predicting the complex relationships between Chinese medicine ingredients and targets remains a formidable challenge due to the vast number and high heterogeneity of these ingredients. In this study, we introduce the CWI-DTI method, which achieves high-accuracy prediction of DTIs using a large dataset of interactive relationships of drug ingredients or candidate targets. Moreover, we present a novel dataset to evaluate the prediction accuracy of both Chinese and Western medicine. Through meticulous collection and preprocessing of data on ingredients and targets, we employ an innovative autoencoder framework to fuse multiple drug (target) topological similarity matrices. Additionally, we employ denoising blocks, sparse blocks, and stacked blocks to extract crucial features from the similarity matrix, reducing noise and enhancing accuracy across diverse datasets. Our results indicate that the CWI-DTI model shows improved performance compared to several existing state-of-the-art methods on the datasets tested in both Western and Chinese medicine databases. The findings of this study hold immense promise for advancing DTI prediction in Chinese and Western medicine, thus fostering more efficient drug discovery and repurposing endeavors. Our model is available at https://github.com/WANG-BIN-LAB/CWIDTI.

Keywords Drug-target interactions, Chinese medicine, Western medicine, Autoencoder framework, Drug databases

Drug-target interaction (DTI) prediction plays a crucial role in the drug discovery process[1–3], offering a cost-effective alternative to traditional experimental methods[4]. Computational approaches have gained prominence due to their ability to simulate and predict the binding process between target proteins and drug molecules, enabling the evaluation of drug properties and the discovery of new compounds. With the advent of high-throughput technologies, a wealth of data on drugs, diseases, genes, and proteins has become available, fueling the progress of computer-based methods[5–8].

Traditional structure-based and ligand-based virtual screening techniques have been extensively studied but have limitations when the 3D structure of target proteins is unknown or when there is insufficient data on known active molecules[9,10]. Recently, deep learning-based methods have shown remarkable advancements in DTI prediction by leveraging chemical genomics, which integrates diverse data sources into a unified framework[2,6,11,12]. These models primarily use linear or two-dimensional (2D) structure information of drugs and proteins as input, treating DTI prediction as a binary classification task. Various deep encoding and decoding modules, such as deep neural networks (DNNs)[13,14], graph convolutional networks (GCNs)[15–18], and autoencoder architectures[19–21] have been employed to process the input and output. These models learn data-driven representations from large-scale DTI data, surpassing the limitations of predefined descriptors.

Deep learning-based methods have shown significant progress in drug-target interaction (DTI) prediction, but most of these methods have primarily focused on single Western medicine databases. Unfortunately, these models have proven to be less effective in predicting the interactions between Chinese herbal medicine ingredients and targets, and their predictive capability across different datasets is limited[22–24].This limitation stems from the shallow architectures and underutilization of parameters and features, failing to fully explore

[1]Department of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China. [2]Department of Pathology, Shanxi Key Laboratory of Carcinogenesis and Translational Research on Esophageal Cancer, Shanxi Medical University, Taiyuan, China. ✉email: wangbin01@tyut.edu.cn

the potential interactions between drugs and proteins[19,25]. Moreover, joint prediction using both Chinese and Western medicine databases introduces additional noise due to the increased volume of data, requiring models to possess robust noise resistance for accurate predictions[3,26,27]. To address these challenges, we present the CWI-DTI prediction model, which efficiently captures the similarity between drugs and targets, achieving excellent prediction performance on the combined Chinese and Western medicine dataset. To fuse multiple drug/ingredients and target topological similarity matrices, autoencoders with denoising and sparse blocks are employed, facilitating the learning of compact and sparse feature representations in the drug-target interaction space. Furthermore, a stacking mechanism is introduced to enhance cross-dataset generalization, enabling deep extraction and representation of drug and target features. Ultimately, the extracted low-dimensional features are inputted into a fully connected layer for accurate DTI prediction.

The contributions of this study are threefold: First, we propose a CWI-DTI model, which employs a multi-module fusion mechanism to capture deep interactive features and enhance model representation. Second, we demonstrate the cross-dataset generalization ability of CWI-DTI by applying it to predict traditional Chinese and Western medicine datasets, achieving high-performance prediction results. Third, we analyze and visualize the molecular target networks of both Chinese and Western medicine, providing insights into the prediction outcomes. We evaluate the performance of CWI-DTI using diverse traditional Chinese and Western medicine datasets and compare it with several state-of-the-art models. Our experimental results indicate that CWI-DTI demonstrates strong predictive capabilities. Furthermore, an in-depth analysis of the highest predicted DTIs reveals support from previous studies, reinforcing the effectiveness of our model. In conclusion, our proposed approach offers promising prospects for understanding drug action patterns and facilitating drug repurposing efforts.

Materials and methods

Computational prediction of drug-target interactions (DTIs) plays a crucial role in drug discovery and development. Existing models primarily focused on Western medicines, leveraging single-component and explicit relationships. However, Chinese medicines pose unique challenges due to their numerous and complex components, along with limited known relationships between components and targets. Consequently, there is a pressing need for efficient prediction models capable of extracting the potential features of drug components and accommodating diverse data distributions. To construct a computational model suitable for DTI prediction in homogeneous datasets, we collected and preprocessed data on the components and targets of both Chinese and Western medicines. Homogeneous drugs often exhibit similarities in their target proteins, making the measurement of drug-drug and target-target similarities crucial for computational prediction of DTIs. In this study, we aimed to extract multidimensional and complementary information for drugs or targets by utilizing multiple similarity measures (Table S1), including the kernel function of chemical structure fingerprints and protein sequences. We propose the CWI-DTI model, which leverages a stack hybrid autoencoder to automatically fuse multiple similarities and learn advanced features for predicting drug-target interactions. Our model incorporates a fusion mechanism consisting of three blocks (Fig. 1) to capture the combined representation of the multiple similarity measures. To construct a computational model suitable for DTI prediction in homogeneous datasets, we initially collected and preprocessed data on the components and targets of both Chinese and Western medicines. Subsequently, we employed a deep autoencoder to fuse multiple drug (target) similarity matrices. By hybridizing denoised, sparse, and stacked blocks within our model, we effectively extract low-dimensional features, thereby enhancing the accuracy and generalization of DTI prediction. Finally, we employ fully connected layers to calculate binding scores, serving as indicators of the likelihood of interaction between drugs and targets.

Dataset collection and preparation

To evaluate the performance of the CWI-DTI method in predicting drug-target interactions (DTIs), we conducted evaluations on ten datasets. Each dataset consisted of three types of information: (1) drug-target interaction data, (2) multiple similarity data of drugs, and (3) multiple similarity data of targets. Table 1 provides an overview of the statistics for these ten datasets. It is important to note that the ratio of known (positive) to nonexistent (unknown, negative) DTIs varied across the datasets, reflecting the reality that the number of true DTIs is considerably smaller than the number of non-interacting drug targets. The datasets were divided into three main parts. The first part comprised the Western medicine dataset, which included DRUGBANK[28], TTD[29], and CHEMBL[30]. The second part consisted of the Traditional Chinese Medicine (TCM) dataset, which encompassed HERB[31], TCMIO[32], HIT[33], and NPASS[34]. Lastly, the third part included the summary of all datasets, namely TCM_ALL, WEST_ALL, and TW_ALL. The data for these datasets were obtained from the respective database websites or collected through web crawlers and manual sorting. For instance, the HERB database, accessible at http://herb.ac.cn/, required manual collation of herbal-ingredient-target association data along with associated structural information, such as chemical structure details, for 12,933 targets, 7,263 Chinese herbs, and 49,258 ingredients. Table 1 provides an overview of the Chinese and Western drug-target interactions within the ten datasets. Notably, the data in these datasets are unbalanced, with a higher number of negative interactions compared to positive interactions. This imbalance can negatively impact the predictive performance of classifiers. To address this issue, we applied the Synthetic Minority Oversampling Technique (SMOTE) to the unbalanced datasets. SMOTE generates synthetic positive samples to balance the minority category and improve the predictive efficiency of the classifier.

Problem description

In this study, we address the problem of drug-target interactions (DTIs), which involves a set of drugs $D=\{d_i, i=1\ldots n_d\}$ and a set of targets $T=\{t_j, j=1\ldots ., n_t\}$. Here, n_d represents the number of drugs, and n_t represents

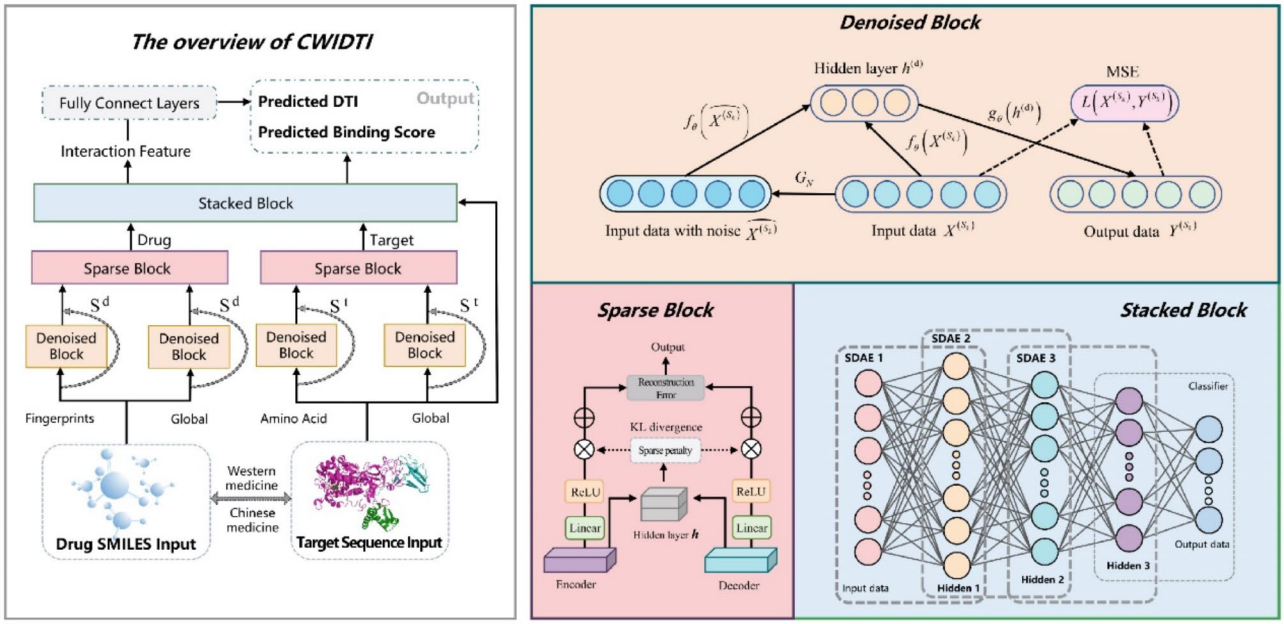

Fig. 1. The figure depicts the workflow of the CWI-IDTI model for predicting drug-target interactions in Chinese and Western medicines. The left-hand side illustrates the overall pipeline of the proposed method. Our approach incorporates three key innovations: the Denoise block, the Spare block, and the Stack block, as demonstrated on the right-hand side of the figure. These blocks represent novel components of our model that contribute to its effectiveness in capturing essential features and improving the accuracy of DTI prediction.

Datasets		Known interactions	Unknown interactions	Drugs/ Ingredients	Targets
Western Medicine	DRUGBANK	10,127(0.19%)	5,364,505(99.81%)	1874	2868
	TTD	1920(0.23%)	833,802(99.77%)	1601	522
	CHEMBL	12,069(0.55%)	2,185,623(99.45%)	2052	1071
	CWI-WM	23,969(0.14%)	17,208,371(99.86%)	5455	3159
Chinese Medicine	TCMIO	41,527(1.79%)	2,281,673(98.21%)	5808	400
	HIT	9796(0.40%)	2,396,828(99.60%)	1166	2064
	NPASS	2444(0.62%)	393,927(99.38%)	607	653
	HERB	109,709(0.18%)	59,327,089(99.82%)	6806	8733
	CWI-CM	131,783(0.12%)	110,787,724(99.88%)	11,571	9586
Chinese and Western Medicine	CWI	155,622(0.09%)	182,423,104(99.91%)	16,882	10,815

Table 1. Summary of samples on the ten data sets.

the number of targets. We represent the interaction between D and T as a binary matrix P with elements that can take the values 0 or 1, denoted as $P \in \mathbb{R}^{n_d \times n_t}$. The matrix P consists of n_d drugs as rows and n_t targets as columns, where P_{ij}=1 indicates an interaction between drug di and target t_j, while P_{ij}=0 indicates no interaction. Additionally, we define the similarity matrix set between drugs in D as S_D, represented as $S_D \in \mathbb{R}^{n_d \times n_d}$. Similarly, we define the set of similarity matrices between targets in T as S_T, denoted as $S_T \in \mathbb{R}^{n_t \times n_t}$. The values within the similarity matrices reflect the degree of similarity between drugs or targets based on various measures. All elements within the matrices fall within the range of [0,1]. The objective of our study is to uncover the underlying factors associated with drug-target pairs $[d_i,t_j]$ and predict new interactions in P (i.e., unknown interactions) based on the similarity matrices of drugs in S_D and targets in S_T.

Preprocessing of multiple similarity measures

To account for the distinctive characteristics of Western medicine (single relationship, limited targets) and traditional Chinese medicine (diverse relationships, multiple targets), we employ molecular SMILES structures and protein amino acid sequences for preprocessing. By assessing their relative similarity across multiple dimensions, we derive nuclear structural representations for the molecular fingerprints and targets of Chinese and Western medicines. These representations incorporate local and global features, including atom characteristics

and global topological information. Subsequently, an efficient association strategy is developed based on these representations. We obtain drug target information from Chinese and Western drug databases through manual collation or crawling. The data undergoes quality control, correcting missing or non-standard molecules or sequences by searching additional databases. The similarity of drug and target structures is calculated by hashing molecular paths and forming molecular fingerprints based on atomic type, aromaticity, and bond type. The Tanimoto coefficient is then computed from the molecular fingerprints to determine the similarity between drugs:

$$T_{d_i d_j} = \frac{c}{a+b-c}$$

Where, $T_{d_i d_j}$ represents the T fraction of drug i and drug j, a and b are the sum of the number of binary bits located at any position in the two fingerprints respectively, and c is the sum of the number of these binary bits are all 1.

The similarity matrix of eight structural fingerprints, including RDK, MACCS, EC4, FC4, EC6, FC6, TOPTOR, and AP, is calculated using the Fingerprint function of the Rdkit package in Python. For targets, similarity measures are extracted by comparing differences between amino acid sequences and mapping them to a high-dimensional vector space. Mismatch and Spectrum kernels are used to compute the similarity between sequences based on k-mers and occurrence frequencies.

The global structure of drugs and targets is processed using Restart Random Walk (RWR) and Positive Pointwise Mutual Information (PPMI)[35] to calculate the topological similarity of drugs in each similarity network. This yields the global structural information of the similarity network, which serves as input for the model.

Feature learning with SDSAE

The feature learning process for drugs and targets in the CWI-DTI model involves Stacked Denoising Sparse Autoencoders (SDSAE). The denoise block introduces Gaussian noise to the input data, ensuring robust feature learning. The model aims to minimize reconstruction error, learning a mapping function from noisy data to original data. Sparsity constraints are incorporated through sparse blocks, preventing overfitting and generating more explanatory and generalizable sparse representations (detailed descriptions in the supplementary materials).

The model's architecture includes stack blocks, combining multiple SDSAEs to form a multi-layer deep neural network for abstract feature extraction. Combining stack denoising sparse autoencoders with CNN, the model becomes an end-to-end deep learning framework for Drug-Target Interaction (DTI) prediction. CNN is utilized for local feature extraction, down-sampling, feature selection, and classification.

The CWI-DTI model is developed by integrating noise reduction, sparse, and stack blocks. The training process involves denoising noisy input, applying sparse blocks for encoding and reconstruction, and stacking blocks for subsequent layers. The loss function of the CWI-DTI model incorporates logistic regression loss, weight attenuation, and L1 norm to enforce sparsity in features, contributing to effective DTI prediction.

Improved convolutional neural network for DTIs prediction

To predict drug-target interactions, we constructed an end-to-end deep learning model by combining the stack denoising sparse autoencoder with a Convolutional Neural Network (CNN). We treated the feature matrices of drugs and targets as two-dimensional images or convolution kernels for extract feature extraction and classification predictions. By combining the stacked denoising sparse autoencoder with the CNN model, we enhance the local feature extraction ability of CNN, thereby improving the classification performance and generalization ability of the model.

CWI-DTI model training

CWI-DTI utilizes a model combined with a stacked denoising sparse autoencoder and an improved CNN, which is used to predict DTI. The training process of the model can be summarized in the following steps:

1. Pre-training of the stack denoising sparse autoencoder: The feature matrices of drugs and targets served as input data. Pre-training was conducted to obtain the parameters of the encoder and decoder for the stack denoising sparse autoencoder.
2. Training of CNN: The encoder obtained from pre-training was utilized as the input layer of the CNN, along with several convolutional layers, pooling layers, and fully connected layers. The feature matrices of drugs and targets were used as input data. Training of the CNN was performed using the cross-entropy loss function and the backpropagation algorithm.
3. Fine-tuning of the overall model: The parameters of the encoder and CNN obtained from pre-training were fine-tuned. The cross-entropy loss function and the backpropagation algorithm were applied to fine-tune the overall model, further enhancing its prediction performance.

The cross-entropy loss function for the overall model, combining the stack denoising sparse autoencoder and improved CNN, is defined as follows:

$$\mathcal{L}(x,y) = -\sum_{k=1}^{K} y_k \log \widehat{y}^* k + \lambda_1 \sum *l = 1^L \|W_l\|_2^2 + \lambda_2 \sum *l = 1^L \sum_{j=1}^{n_l} KL\left(\rho_l \| \widehat{\rho}_{lj}\right)$$

Here, y represents the true label, $\widehat{y}$ represents the prediction result of the model on sample x, and K represents the number of categories. The second term is the L2 regularization term, which is used to prevent the model from overfitting. Here, W_l denotes the weight matrix of the lth layer, and $\|W_l\|_2^2$ denotes the square of its two-norm. The λ_1 is the regularization coefficient used to balance the importance of the regularization term and the cross-entropy loss term. The third term is a sparse regularization term, which is used to constrain the activation value ρ_l of the encoding layer to be close to a small constant $\widehat{\rho}_{lj}$. Here, l denotes layer l and n_l denotes the number of neurons in layer l. ρ_l is the average activation value of all neurons in layer l, and $\widehat{\rho}_{lj}$. is a preset sparsity target, which generally takes a small value, for example 0.05. The λ_2 is the sparse regularization coefficient, which is used to balance the importance of the sparse regularization term and the cross-entropy loss term.

For the entire dataset, the cross-entropy loss function is computed as the average of the loss functions across all samples:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(x_i, y_i)$$

Here, N represents the number of samples in the dataset.

During model training, the momentum-based Stochastic Gradient Descent (SGD) optimization algorithm was employed to minimize the cross-entropy loss function of the overall model, thereby improving classification accuracy and generalization ability. The training process involved multiple rounds of iteration and parameter tuning to achieve the best prediction performance.

Hyperparameter comparison

In our evaluation, we focused on six key hyperparameters: batch size (BS), noise factor (NF), hidden layer dimension of the autoencoder (HLD), sparsity distribution (SD), learning rate (LRA), and training epochs (TE). The performance of the CWI-DTI model showed no significant changes with varying hyperparameters (Figure S1). Ultimately, we selected a batch size of 1024, a noise factor of 0.8, a hidden layer dimension of 300, a sparsity distribution of 0.04, a learning rate of 1e-4, and 100 training epochs, based on their balanced performance and stability during training.

Experimental setup and model evaluation

In this study, we evaluated the performance of the CWI-DTI model using Receiver Operating Characteristic (ROC) curve area under the curve (AUC) and precision-recall curve area under the curve (AUPR) as evaluation metrics. We employed a 10-fold cross-validation (CV) approach with 5 repetitions to assess the performance of the DTI prediction method. Both AUC and AUPR scores were calculated for each repetition of the CV. The final AUC and AUPR scores were obtained by calculating the mean across the 5 repetitions.

The drug-target interaction matrix Y consists of n_d rows for drugs and n_t columns for targets. We performed CV under three different settings (Table S2):

- CVS1: Testing with random entries (i.e., drug-target pairs) in Y.
- CVS2: Blind testing with CV of random rows of drugs (i.e., drugs) in Y.
- CVS3: Blind testing with CV of random columns of targets (i.e., targets) in Y.

Under CVS1, we applied 5-repeats of stratified 10-fold cross-validation, where each round used 90% of the elements in Y as training data and the remaining 10% as test data. Similarly, under CVS2, we used 90% of the rows in Y as training data and the remaining 10% as test data. For CVS3, we utilized 90% of the columns in Y as training data and the remaining 10% as test data. These settings, CVS1, CVS2, and CVS3, respectively refer to the prediction of DTIs for (1) new (unknown) pairs, (2) new drugs, and (3) new targets.

To determine the optimal configuration of blocks (number of layers and number of neurons per layer) in the CWI-DTI model for the Chinese and Western medicine datasets, we performed five repeated 10-fold cross-validation under CVS1. This allowed us to evaluate the model's performance with different layer configurations. This approach ensured that the category ratio in each fold remained consistent with the overall dataset.

To compare the performance of the CWI-DTI model with other datasets, including GADTI, AutoDTI++, MDADTI, NeoDTI, DDR, and DNILMF, we conducted 10-fold cross-validation over five replicates under three different settings. This enabled us to focus on the differences between CWI-DTI and the aforementioned datasets.

Results

Improving DTI prediction in Western and Chinese medicine through the integration of three innovative blocks

We conducted experiments with various configurations of CWI-DTI, analyzing different innovative elements while utilizing the deep autoencoder (DAE) as our baseline method. In this study, we enhanced the baseline model by incorporating denoising blocks, sparse blocks, and stacked blocks. The results of the baseline model and the new models with different configurations are presented in Fig. 2; Table 2. The denoising block (D-M) showed improved performance on the DRUGBANK Western Medicine dataset, with a 1.04% increase in accuracy. This improvement can be attributed to the introduction of Gaussian noise, which helps the autoencoder learn the data distribution, reduce overfitting, and enhance generalization ability. The denoising block facilitated the acquisition of robust and discriminative feature representations. Spa-M, incorporating the sparse block, exhibited lower performance compared to the fusion method in the baseline model. This can be attributed to a mismatch between the assumptions of the sparse block and the diverse distribution characteristics

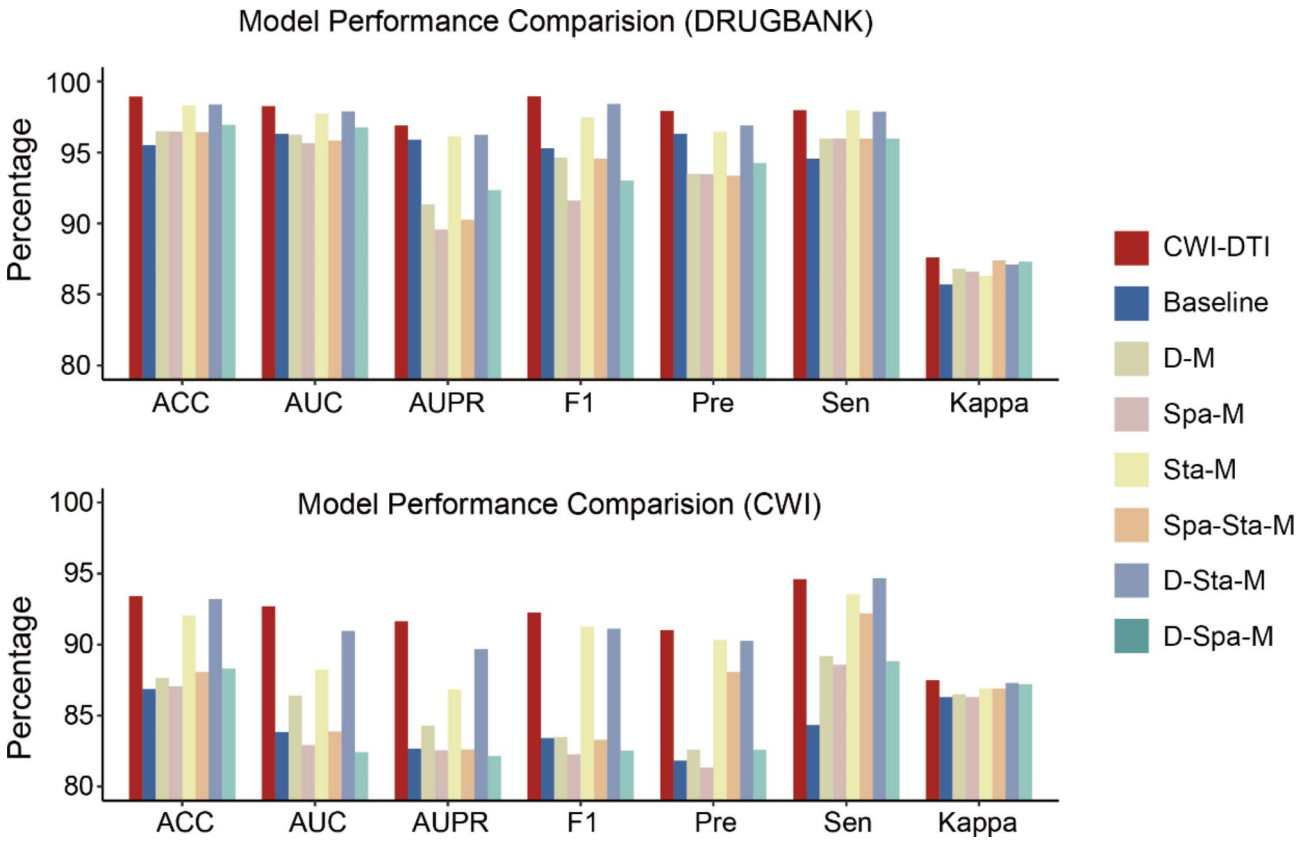

Fig. 2. The performance of DTI with different architectures of CWI-DTI on DRUGBANK and CWI datasets.

Models	Denoise	Sparse	Stack	AUC	AUPR	F1	ACC	Sen	Pre	Kappa
Baseline	-	-	-	96.30	95.90	95.28	95.51	94.55	96.30	0.857
D-M	√	-	-	96.25	91.34	94.62	96.50	95.97	93.48	0.868
Spa-M	-	√	-	95.64	89.56	91.61	96.49	95.98	93.45	0.866
Sta-M	-	-	√	97.71	96.13	97.47	98.31	97.96	96.44	0.863
Spa-Sta-M	-	√	√	95.84	90.27	94.55	96.43	95.97	93.36	0.874
D-Sta-M	√	-	√	97.89	96.24	98.41	98.38	97.87	96.89	0.871
D-Spa-M	√	√	-	96.76	92.34	93.02	96.93	95.97	94.25	0.873
CWIDTI	√	√	√	**98.26**	**96.89**	**98.94**	**98.93**	**97.97**	**97.92**	**0.876**

Table 2. The performance of DTI with different autoencoder model on DRUGBANK dataset.

of the DRUGBANK dataset. To enhance the model's representational capacity, we introduced the stacked block. Sta-M demonstrated significantly improved results in terms of AUC (0.9771) and AUPR (0.9613), providing evidence for the superiority of the stacked block. Spa-Sta-M, combining the sparse and stacked elements, exhibited decreased performance. The effectiveness of the stacked block may be diminished when the input data exhibits significant variations and differences. However, D-Sta-M demonstrated improved performance on the DRUGBANK dataset, with increases of 1.7% and 3.3% in AUC and F1 scores, respectively, supporting the effectiveness of the mixed innovative block. By combining all three innovative elements, we constructed the final CWI-DTI method, which achieved the best results for Drug-Target Interaction (Fig. 2A).

We also conducted experiments on our CWI traditional Chinese medicine dataset (Fig. 2B, Table S3). The stacked block (Sta-M) and the mixed block (D-Sta-M) demonstrated significant improvements, with AUC increasing by 2.1% and 2.8% compared to the CWI dataset. The stack block learning proved capable of extracting more abstract and complex feature representations, which is particularly beneficial for the diverse and complex nature of traditional Chinese medicine. Overall, the denoising block helps in learning data distribution and reducing overfitting, the sparse block captures local features and important structures, and the stacked block enables the learning of more complex feature representations.

Layer configuration optimization for CWI-DTI model

To determine the layer configuration of the CWI-DTI model, we performed cross-validation experiments. Figure 1 illustrates the layer configuration diagram, consisting of three stacked autoencoders. The optimal configuration had a layer structure of the first autoencoder has 400 neurons, the second AE has 200 neurons, and the third stacked layer has 100 neurons, with a layer configuration of [n*n_d, n*400, n*200,100, n*200, n*400, n*n_d] for the drug-related features and [n*n_t, n*400, n*200,100, n*200, n*400, n*n_t] for the target-related features. Table S4 presents the performance of the CWI-DTI model with different layer configurations on the TCM-Western drug dataset. We observed that the highest AUC and AUPR values were achieved when the two CWIs were stacked to three layers for the DRUGBANK dataset. Similarly, for the TCM dataset, the CWI-DTI model achieved the highest AUC and AUPR when the CWI had three layers (0.958 and 0.929 respectively). Together, our study demonstrated the effectiveness of the three innovative blocks in improving DTI prediction in both Western and Chinese medicine datasets. The combination of denoising, sparse, and stacked blocks yielded the best results, with the stacked block particularly advantageous for the complexity of traditional Chinese medicine. The optimal layer configurations were determined through cross-validation experiments, resulting in improved performance metrics for DTI prediction.

Comparative analysis suggested robust performance of CWI-DTI

To comprehensively evaluate the performance of CWI-DTI, we compared the performance of CWI-DTI with several state-of-the-art methods on various datasets. These datasets include DRUGBANK, TTD, and CHEMBL. We also evaluated CWI-DTI on traditional Chinese medicine datasets, namely HERB, TCMIO, HIT, and NPASS. Due to the large amount of data and high complexity of other methods, we exclusively tested our CWI-DTI method under the three CV settings, namely CWI_WM, CWI_CM, and CWI. We compared the performance of CWI-DTI with baseline methods on the DRUGBANK dataset (Fig. 3). CWI-DTI achieved an AUROC value of 0.9826, indicating improved performance compared to GADTI (2.18%), AutoDTI++ (6.33%), MDADTI (2.04%), NeoDTI (2.71%), DDR (6.18%), and NRLMF (35.70%). GADTI employs graph convolutional networks and random walks for interaction discovery, while AutoDTI++ uses denoising autoencoders to address sparsity in interaction matrices. NeoDTI enhances DTI predictions through neighborhood information aggregation, and MDADTI merges topological similarity matrices using deep autoencoders. CWI-DTI also showed superior performance in AUPRC compared to all baseline methods, with a slight improvement over the second-best method (AUROC: 2.04%, AUPRC: 1.03%). Under the CWI dataset, CWI-DTI outperformed GADTI (6.48%), AutoDTI++ (10.96%), MDADTI (8.36%), NeoDTI (7.04%), DDR (20.77%), and NRLMF (34.07%) with an AUROC value of 0.9578. Despite the inherent challenges of traditional Chinese medicine due to its complex molecular structure, CWI-DTI demonstrated remarkable performance. Detailed results are provided in Fig. 4.

In our evaluation using cross-validation, CWI-DTI indicated improved performance over GADTI, AutoDTI++, MDADTI, NeoDTI, DDR, and DNILMF in the CVS1 setting (Fig. 5). Across six datasets representing Chinese and Western medicines, CWI-DTI consistently achieved higher AUC and AUPR values. Specifically, when comparing AUC growth rates for DRUGBANK, TTD, CHEMBL, HIT, TCMIO, and NPASS datasets, CWI-DTI showed significant improvements (ranging from 0.82 to 35.7% of growth rates) compared to other methods. Similarly, CWI-DTI exhibited notable growth rates in AUPR for these datasets (ranging from 1.04–41.87%).We also compared our method with recent self-attention-based transformer models, such as TransformerCPI[36] and TransformerCPI 2.0[37], and found that CWI-DTI still demonstrated superior predictive performance based on similarity matrices (Table S5). While TransformerCPI 2.0 offers unique advantages in protein sequence analysis and drug design, CWI-DTI appears to be more effective in DTI prediction.

Further analysis revealed that CWI-DTI consistently outperformed the other methods, demonstrating its potential in both Western medicine and traditional Chinese medicine datasets. Comparisons under the CVS2 and CVS3 settings confirmed the robustness of CWI-DTI, showcasing improved performance in terms of AUPR across various datasets, with minor limitations observed in the TCMIO dataset due to its small number of targets. The success of CWI-DTI can be attributed to its ability to learn high-level abstract features through the stacking of multiple CWI models, avoiding over-smoothing. By leveraging similarity calculations and capturing local and global molecular information using RWR, CWI-DTI effectively selects important features, introduces multi-layer information, and learns higher-level feature representations to enhance its expressive power.

Enhancing traditional Chinese medicine repositioning through predicting new drug-target interactions with multi-database DTI predictions

In the assessment of the predictive capacity of the CWI-DTI model for identifying novel drug-target interactions, we designated the negative samples within the CWI dataset as uncharacterized drug-target interactions (DTI). Leveraging a pre-trained CWI-DTI model on both Chinese and Western medicine datasets, we derived probability estimations for potential drug-target interactions. Evaluation of new interactions was grounded in high-probability drug-target pairs absent from the initial dataset. Our model successfully predicted 12 out of 15 unknown interactions in the DRUGBANK dataset and 10 out of 15 in the CWI dataset, as highlighted in Tables S6 and S7. For example, our analysis revealed potential interactions between Alcohol and the BCL2 target, which were validated by existing research indicating significant effects on caspase-3 activity, ALDH2 activity, and BCL-2/BAX mRNA expression[38,39]. Another notable prediction involved resveratrol, which was found to result in a two-fold decrease in DNAH8 expression, suggesting its potential role in preventing cardiovascular disease through various signaling pathway[40,41]. Additionally, Niacin was predicted to interact with HIF1A, showing promise in mitigating hypoxia-induced inflammation by upregulating HIF1A expression[42,43].

The model also facilitates drug repositioning across both Chinese and Western medicines. For instance, we identified that NF-κB interacts with the Western drug WM00129 (Ornithine) and Chinese medicine components CM00003 (Deguelin) and CM00199 (Luteolin), with both Deguelin and Luteolin exhibiting similar potencies as

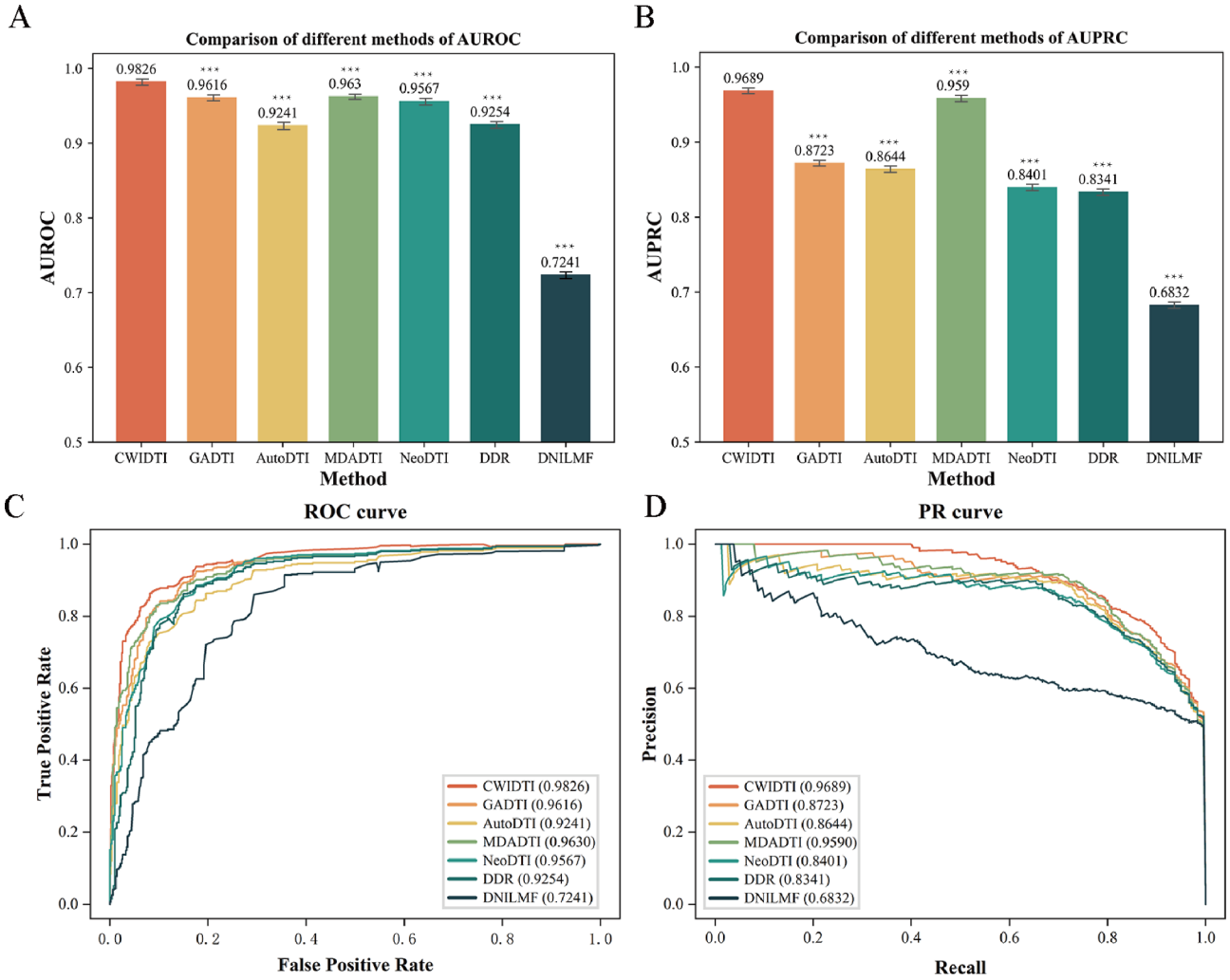

Fig. 3. Comparison of different methods based on 10-fold cross-validation on DRUGBANK dataset. A and B represent AUROC and AUPRC scores; C and D represent ROC curves and PR curves of different methods on DRUGBANK dataset. Statistical differences were evaluated using t-tests. P value: * <0.05; ** <0.01; *** <0.001, $P < 0.05$ was considered as significant for all tests.

anti-angiogenic agents (Fig. 6B)[44,45]. Furthermore, the cancer chemopreventive effects of deguelin analogs are associated with the inhibition of phorbol ester-induced ornithine decarboxylase activity[46]. This indicates that NF-κB serves as a potential convergence point for therapeutic pathways between these two medicinal systems. As illustrated in Fig. 6B, certain target entities interact with both Western drugs and Chinese medicine ingredients, further emphasizing the integration of these medicinal approaches. Given the ongoing updates to the databases we utilized, we anticipate that the number of validated DTIs predicted by our model will continue to grow. The strong predictive performance of the CWI-DTI model across diverse datasets underscores its capability to simultaneously predict drug targets for both Chinese and Western medicines, significantly enhancing the drug-target dataset in our study.

Discussion

We have presented CWI-DTI, a novel approach for drug-target interaction (DTI) prediction that combines traditional Chinese medicine and Western medicine. Our method improved performance compared to several existing approaches by leveraging the similarity between Chinese and Western drug molecules and incorporating both local and global information through random walk with restart (RWR). Additionally, CWI-DTI incorporates denoising, sparsity, and stack blocks to improve cross-dataset generalization capability. Through comprehensive evaluations using cross-validation settings, we have shown that CWI-DTI achieves higher AUC and AUPR compared to GADTI, AutoDTI++, MDADTI, NeoDTI, DDR, and DNILMF across multiple datasets.

Furthermore, we have examined the practical ability of CWI-DTI to predict new interactions by validating the top predictions against reference databases, which supports the effectiveness of CWI-DTI in identifying unknown DTIs. However, our work primarily focuses on chemogenomic-based DTI prediction, utilizing 1D

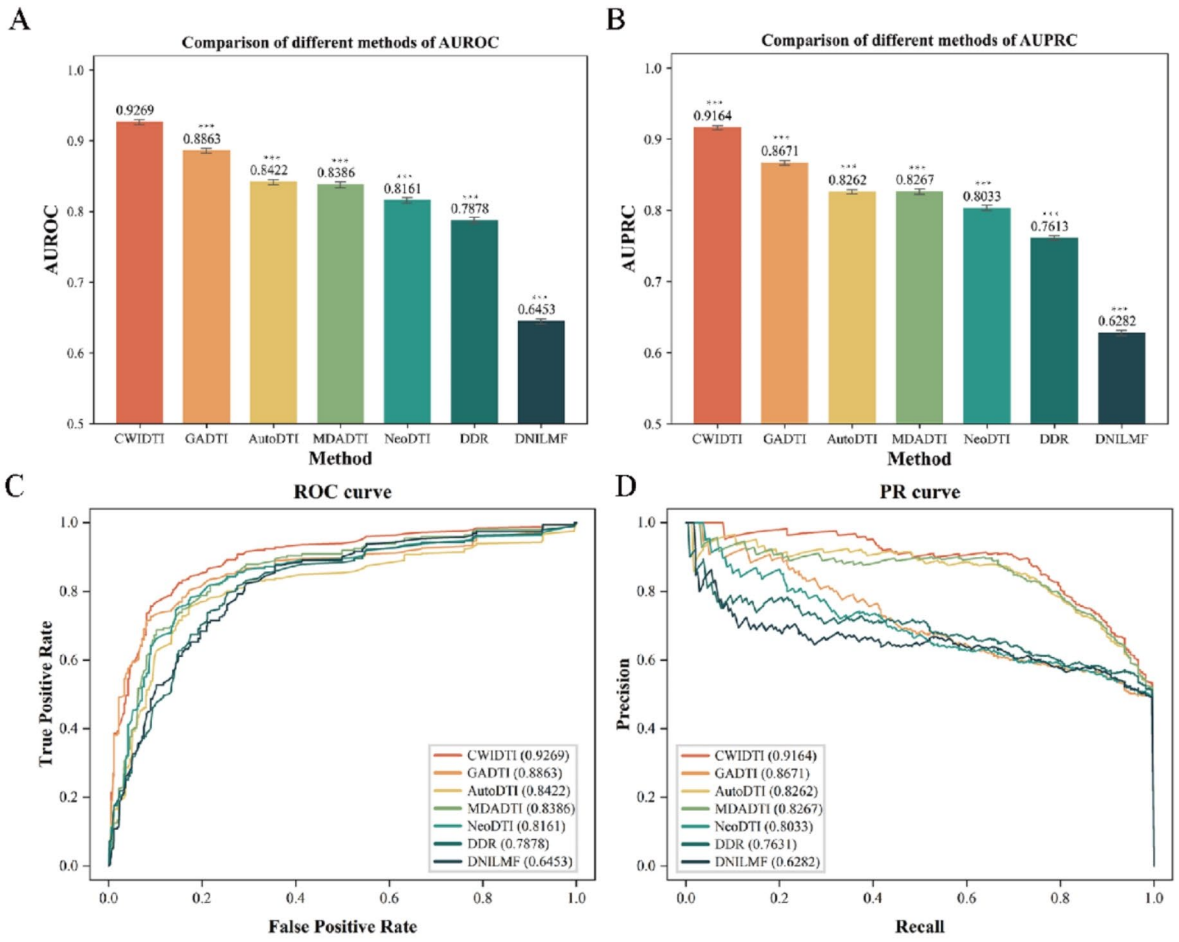

Fig. 4. Comparison of different methods based on 10-fold cross-validation on CWI dataset. A and B represent AUROC and AUPRC scores; C and D represent ROC curves and PR curves of different methods on CWI dataset. Statistical differences were evaluated using t-tests. P value: * <0.05; ** <0.01; *** <0.001, $P < 0.05$ was considered as significant for all tests.

protein sequences and molecular SMILE structures. Incorporating accurate 3D protein structures, such as those predicted by DeepMind's AlphaFold, holds promise for further improving DTI prediction performance and interpretability.

While CWI-DTI has shown promising results, there are areas for future improvement. The model currently requires retraining to predict interactions involving new nodes that were not included in the training process. Additionally, CWI-DTI cannot predict isolated new nodes that are unrelated to known drug or target nodes. To address these limitations, future research should consider several promising approaches. First, the integration of additional node features could enhance the model's capacity to predict new interactions, allowing it to derive insights from partial information. This would enable more robust predictions even when complete data is unavailable. Second, employing transfer learning techniques could significantly enhance the model's generalizability to unseen data by leveraging knowledge from prior training. Such strategies may improve the adaptability and robustness of CWI-DTI, ultimately enhancing its predictive performance in diverse biological contexts. Lastly, this study explored different Chinese and Western medicine datasets independently. To enhance the robustness of our deep learning model, we aim to collect a comprehensive dataset encompassing a wider range of proteins and Chinese and Western drugs. Integrating this dataset with CWI-DTI presents an exciting avenue for future exploration. It is also important to acknowledge potential biases in our datasets, such as underrepresentation of certain drug classes or target families, which may influence the model's performance. A more diverse dataset will allow us to evaluate the model's applicability across various scenarios and enhance its generalizability.

In summary, our study highlights the effectiveness of CWI-DTI in DTI prediction and identifies areas for further development. The combination of Chinese and Western medicine, along with innovative modeling techniques, has the potential to advance drug discovery and promote a deeper understanding of drug-target interactions.

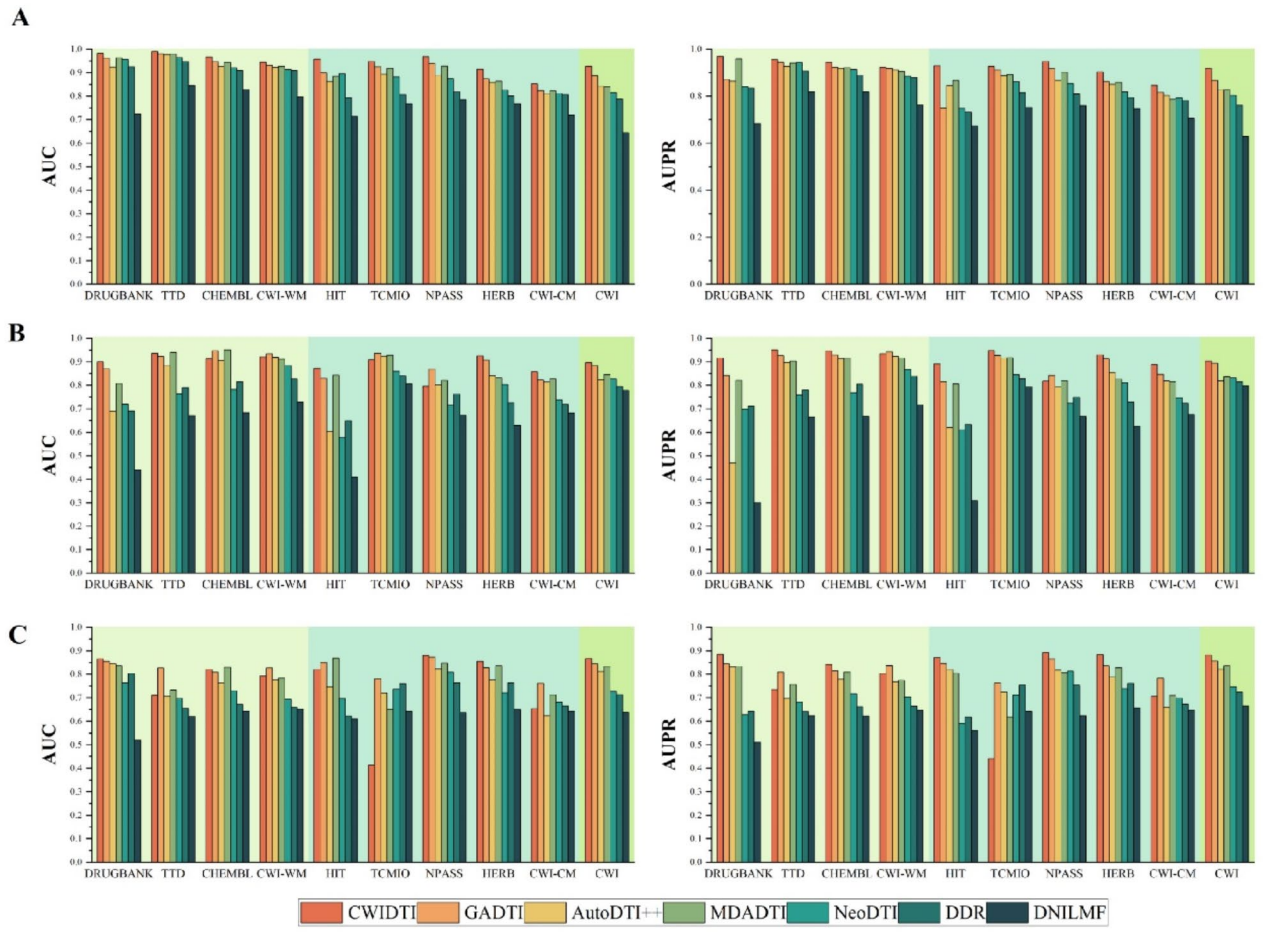

Fig. 5. Comparison of AUC and AUPR among CWI-DTI, GADTI, AutoDTI++, MDADTI, NeoDTI, DDR and DNILMF methods on ten datasets under CVS1, CVS2, and CVS3 setting. (A) Comparison of AUC and AUPR under CVS1 setting; (B) Comparison of AUC and AUPR under CVS2 setting; (C) Comparison of AUC and AUPR under CVS3 setting.

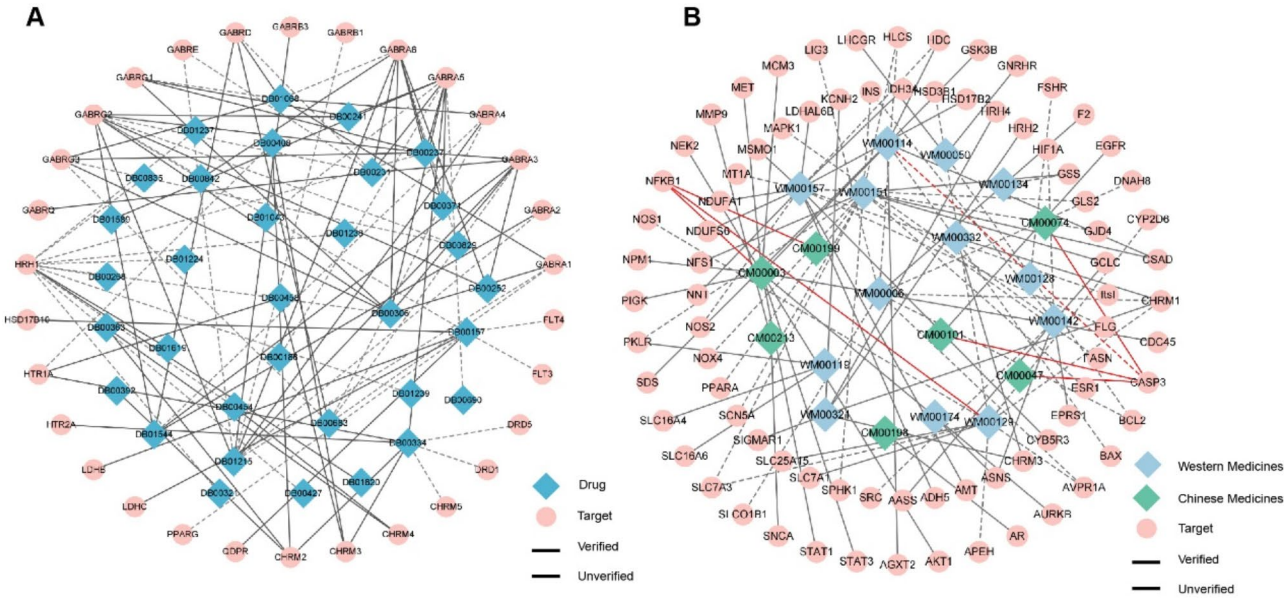

Fig. 6. Network visualization of the top 100 unknown DTIs in DRUGBANK dataset (A) and CWI dataset (B). Blue, green and pink nodes represent drugs (Western medicine drugs, Chinese Medicine ingredients) and targets, respectively. Solid lines represent verified interaction and dashed lines represent unverified interactions.

Data availability

Our model is available at https://github.com/WANG-BIN-LAB/CWIDTI.

Received: 10 July 2024; Accepted: 14 October 2024
Published online: 23 October 2024

References

1. Chu, Y. et al. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinform.* **22**, 451–462 (2019).
2. Shao, K. et al. DTI-HETA: prediction of drug–target interactions based on GCN and GAT on heterogeneous graph. *Brief. Bioinform.* **23**, bbac109 (2022).
3. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat. Mach. Intell.* **5**, 126–136 (2023).
4. Yang, Z., Zhong, W., Zhao, L. & Chen, C. Y. C. ML-DTI: mutual learning mechanism for interpretable drug–target Interaction Prediction. *J. Phys. Chem. Lett.* **12**, 4247–4261 (2021).
5. Chu, Y. et al. DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Brief. Bioinform.* **22**, bbaa205 (2021).
6. Lin, S. et al. MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Brief. Bioinform.* **23**, bbab421 (2022).
7. Shim, J., Hong, Z. Y., Sohn, I. & Hwang, C. Prediction of drug–target binding affinity using similarity-based convolutional neural network. *Sci. Rep.* **11**, 4416 (2021).
8. Singh, R., Sledzieski, S., Bryson, B., Cowen, L. & Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc. Natl. Acad. Sci. U.S.A. **120**, e2220778120 (2023).
9. Hattori, M., Tanaka, N., Kanehisa, M. & Goto, S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.* **38**, W652–W656 (2010).
10. Nascimento, A. C. A., Prudêncio, R. B. C. & Costa I. G. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinform.* **17**, 46 (2016).
11. Cheng, S. et al. GraphMS: drug target prediction using graph representation learning with substructures. *Appl. Sci.* **11**, 3239 (2021).
12. Zhang, P., Wei, Z., Che, C. & Jin, B. DeepMGT-DTI: Transformer network incorporating multilayer graph information for drug–target interaction prediction. *Comput. Biol. Med.* **142**, 105214 (2022).
13. Wan, F., Hong, L., Xiao, A., Jiang, T. & Zeng, J. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics* **35**(1), 104–111 (2019).
14. Wang, H. et al. A Novel Approach for drug-target interactions Prediction based on Multimodal Deep Autoencoder. *Front. Pharmacol.* **10**, 1592 (2020).
15. Peng, J. et al. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Brief. Bioinform.* **22**, bbaa430 (2021).
16. Liu, Z. et al. GADTI: Graph Autoencoder Approach for DTI Prediction from Heterogeneous Network. *Front. Genet.* **12**, 650821 (2021).
17. Sun, C., Xuan, P., Zhang, T. & Ye, Y. Graph Convolutional Autoencoder and Generative Adversarial Network-based method for Predicting Drug-Target interactions. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19**, 455–464 (2022).
18. Wu, Y., Gao, M., Zeng, M., Zhang, J. & Li, M. BridgeDPI: a novel graph neural network for predicting drug–protein interactions. *Bioinformatics*. **38**, 2571–2578 (2022).
19. Sajadi, S. Z., Zare Chahooki, M. A., Gharaghani, S. & Abbasi, K. AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders. *BMC Bioinform.* **22**, 204 (2021).

20. Sun, C., Cao, Y., Wei, J. M. & Liu, J. Autoencoder-based drug–target interaction prediction by preserving the consistency of chemical properties and functions of drugs. *Bioinformatics*. **37**, 3618–3625 (2021).
21. Xuan, P., Fan, M., Cui, H., Zhang, T. & Nakaguchi, T. GVDTI: graph convolutional and variational autoencoders with attribute-level attention for drug–protein interaction prediction. *Brief. Bioinform.* **23**, bbab453 (2022).
22. Lv, Q. et al. TCMBank-the largest TCM database provides deep learning-based chinese-western medicine exclusion prediction. *Sig Transduct. Target. Ther.* **8**, 127 (2023).
23. Yang, K. et al. Heterogeneous network propagation for herb target identification. *BMC Med. Inf. Decis. Mak.* **18**, 17 (2018).
24. Wang, N. et al. Herb Target Prediction based on representation learning of Symptom related Heterogeneous Network. *Comput. Struct. Biotechnol. J.* **17**, 282–290 (2019).
25. Huang, K., Xiao, C., Glass, L. M. & Sun, J. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics*. **37**, 830–836 (2021).
26. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nat. Commun.* **14**, 1989 (2023).
27. Hua, Y., Song, X., Feng, Z. & Wu, X. MFR-DTA: a multi-functional and robust model for predicting drug–target binding affinity and region. *Bioinformatics*. **39**, btad056 (2023).
28. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
29. Wang, Y. et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res. Gkz*. **981** https://doi.org/10.1093/nar/gkz981 (2019).
30. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
31. Fang, S. et al. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. *Nucleic Acids Res.* **49**, D1197–D1206 (2021).
32. Liu, Z. et al. TCMIO: a Comprehensive Database of Traditional Chinese Medicine on Immuno-Oncology. *Front. Pharmacol.* **11**, 439 (2020).
33. Yan, D. et al. HIT 2.0: an enhanced platform for herbal ingredients' targets. *Nucleic Acids Res.* **50**, D1238–D1243 (2022).
34. Zeng, X. et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
35. Fan, X. N., Zhang, S. W., Zhang, S. Y., Zhu, K. & Lu, S. Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. *BMC Bioinform.* **20**, 87 (2019).
36. Chen, L. et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*. **36**, 4406–4414 (2020).
37. Chen, L. et al. Sequence-based drug design as a concept in computational drug design. *Nat. Commun.* **14**, 4217 (2023).
38. Mooney, S. M. & Miller, M. W. Effects of prenatal exposure to ethanol on the expression of bcl-2, bax and caspase 3 in the developing rat cerebral cortex and thalamus. *Brain Res.* **911**, 71–81 (2001).
39. Fish, E. W. et al. The pro-apoptotic bax gene modifies susceptibility to craniofacial dysmorphology following gastrulation-stage alcohol exposure. *Birth Defects Res.* **114**, 1229–1243 (2022).
40. Su, M. et al. Genome-wide transcriptional profiling reveals PHACTR1 as a Novel Molecular Target of Resveratrol in endothelial homeostasis. *Nutrients*. **14**, 4518 (2022).
41. Pinheiro, D. M. L. et al. Resveratrol decreases the expression of genes involved in inflammation through transcriptional regulation. *Free Radic. Biol. Med.* **130**, 8–22 (2019).
42. Tang, Y. Y., Wang, D. C., Wang, Y. Q., Huang, A. F. & Xu, W. D. Emerging role of hypoxia-inducible factor-1α in inflammatory autoimmune diseases: a comprehensive review. *Front. Immunol.* **13**, 1073971 (2023).
43. Curran, C. S. et al. Nicotinamide antagonizes Lipopolysaccharide-Induced hypoxic cell signals in human macrophages. *J. Immunol.* **211**, 261–273 (2023).
44. Wang, Y., Ma, W. & Zheng, W. Deguelin, a novel anti-tumorigenic agent targeting apoptosis, cell cycle arrest and anti-angiogenesis for cancer chemoprevention. *Mol. Clin. Oncol.* **1**, 215–219 (2013).
45. Rocchetti, M. T., Bellanti, F., Zadorozhna, M., Fiocco, D. & Mangieri, D. Multi-faceted role of Luteolin in Cancer Metastasis: EMT, angiogenesis, ECM degradation and apoptosis. *IJMS*. **24**, 8824 (2023).
46. Fang, N. & Casida, J. E. Anticancer action of cubé insecticide: Correlation for rotenoid constituents between inhibition of NADH:ubiquinone oxidoreductase and induced ornithine decarboxylase activities. Proc. Natl. Acad. Sci. U.S.A. **95**, 3380–3384 (1998).

Author contributions

Y.L. and B.W. designed research; Y.L. and X.Z. performed research; Y.L., X.Z., Z.C., H.Y, H.W., and T.Y. analyzed data; Y.L., X.Z., Z.C., H.W., T.Y., J.X, and W.B. wrote the paper.

Funding

This work was supported by the Shanxi Province Science Foundation for Youths (Grant Nos. 20210302123092), the Natural Science Foundation of China (Grant Nos. 62403344, 62176177); the pecial Regional Cooperation Project for Science and Technology Cooperation and Exchange (Grant Nos. 202304041101034); the Natural Science Foundation of Shanxi (20210302123112); the Research Project Supported by Shanxi Scholarship Council of China (2021-039); and the National Key Scientific and Technological Infrastructure project "Earth System Numerical Simulation Facility" (2023-EL-PT-000371, 2023-EL-PT-000374).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76367-0.

Correspondence and requests for materials should be addressed to B.W.