

Full Length Article

HSTrans: Homogeneous substructures transformer for predicting frequencies of drug-side effects

Kaiyi Xu^a, Minhui Wang^b, Xin Zou^a, Jingjing Liu^c, Ao Wei^d, Jiajia Chen^{e,*}, Chang Tang^{a,*}

^a School of Computer Science, China University of Geosciences, Wuhan 430074, China

^b Department of Pharmacy, Lianshui People's Hospital Affiliated to Kangda College of Nanjing Medical University, Huai'an 223300, China

^c Department of Cardiac Surgery, Tianjin Chest Hospital, Tianjin 300222, China

^d Department of Cardiology, Tianjin Chest Hospital, Tianjin 300222, China

^e Department of Pharmacy, The Affiliated Huai'an Hospital of Xuzhou Medical University and The Second People's Hospital of Huai'an, Huai'an 223002, China

ARTICLE INFO

Keywords:

Drug-side effects prediction

Transformer encoder

Feature fusion

ABSTRACT

Identifying the frequencies of drug-side effects is crucial for assessing drug risk-benefit. However, accurately determining these frequencies remains challenging due to the limitations of time and scale in clinical randomized controlled trials. As a result, several computational methods have been proposed to address these issues. Nonetheless, two primary problems still persist. Firstly, most of these methods face challenges in generating accurate predictions for novel drugs, as they heavily depend on the interaction graph between drugs and side effects (SEs) within their modeling framework. Secondly, some previous methods often simply concatenate the features of drugs and SEs, which fails to effectively capture their underlying association. In this work, we present HSTrans, a novel approach that treats drugs and SEs as sets of substructures, leveraging a transformer encoder for unified substructure embedding and incorporating an interaction module for association capture. Specifically, HSTrans extracts drug substructures through a specialized algorithm and identifies effective substructures for each SE by employing an indicator that measures the importance of each substructure and SE. Additionally, HSTrans applies convolutional neural network (CNN) in the interaction module to capture complex relationships between drugs and SEs. Experimental results on datasets from Galeano et al.'s study demonstrate that the proposed method outperforms other state-of-the-art approaches. The demo codes for HSTrans are available at <https://github.com/Dtdtxuky/HSTrans/tree/master>.

1. Introduction

Drug-side effects hinder the development of new drugs and pose significant risks to human health. There have been numerous cases of unknown adverse drug reactions resulting in drug delisting and even patient fatalities (Hughes, Rees, Kalindjian, & Philpott, 2011; Pirmohamed, Breckenridge, Kitteringham, & Park, 1998). Therefore, accurate and comprehensive detection of drug-side effects can minimize the risk of drug recalls and provide a safer medication basis for patients (Berry, Knapp, & Raynor, 2002; Chan, Shan, Dahoun, Vogel, & Yuan, 2019).

Currently, the most common method for monitoring frequencies is through clinical randomized controlled trials. In these trials, subjects are randomly assigned to either the experimental group, receiving the drug, or the control group, receiving standard treatment. Conclusions are drawn by comparing the frequencies of drug-side effects between the two groups (Stricker & Psaty, 2004). The frequency of the SE mentioned above is defined in a spatial sense. It is determined by the

proportion of individuals experiencing the SE after taking a specific medication, relative to the total number of individuals taking that medication (Galeano, Li, Gerstein, & Paccanaro, 2020).

Although above approach has been widely utilized, there are certain drawbacks in terms of both (1) time constraints, as clinical trials have a limited duration and may not capture the potential SEs caused by long-term drug use; and (2) sample types, that participants in clinical trials are required to meet specific criteria, which limits the availability of information on drug-side effects in specific populations such as the elderly and children. The mentioned shortcomings lead to incomplete detection of drug-side effects. In addition, the prolonged time and high costs also pose challenges in developing new drugs. To address these limitations, computational methods for predicting frequencies have been proposed in recent years. These methods have the potential to reduce pharmaceutical costs and accelerate the discovery of new drugs (Arshed, Mumtaz, Riaz, Sharif, & Abdullah, 2022; Dimitri & Lió,

* Corresponding authors.

E-mail addresses: jjachen@outlook.com (J. Chen), tangchang@cug.edu.cn (C. Tang).

<https://doi.org/10.1016/j.neunet.2024.106779>

Received 19 March 2024; Received in revised form 29 August 2024; Accepted 1 October 2024

Available online 23 October 2024

0893-6080/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

2017; Ding, Tang, & Guo, 2019; Huang, Wang, Zheng, Chen, & Tang, 2024; Hussain et al., 2020; Lee, Huang, Chang, Lee, & Lai, 2017; Liu, Tong, & Chen, 2023; Muñoz, Nováček, & Vandenbussche, 2019; Wu et al., 2024).

Machine learning techniques have been utilized to predict the frequencies of drug-side effects and demonstrate remarkable performance. This success can be attributed to selecting features closely correlated with the biological information of drugs. To the best of our knowledge, the matrix factorization model introduced by Galeano et al. (2020) is the first method for frequency prediction. This model decomposed the drug-side effect frequency matrix to extract relevant drugs' and SEs' characteristics, ensuring both reproducibility and interpretability of the prediction results. Additionally, Guo et al. (2020) proposed a triple matrix factorization method, which used multiple kernel matrices and kernel target alignment-based multiple kernel learning (KTA-MKL) to fuse multivariate information and obtain new associations. However, handling large-scale data remains a challenge in machine learning due to the complexity of matrix operations.

Motivated by the impressive achievements of deep learning in image processing, natural language processing (NLP), features fusion, and bioinformatics (Lv, Zhou, Yang, He, & Chen, 2023; Si et al., 2023; Tang, Wang, et al., 2023; Tang, Zheng, et al., 2023; Wang, Wu, Ota, Dong, & Li, 2023), an increasing number of researchers are incorporating it into the prediction of associations between drugs and SEs. This is primarily done through similarity-based approaches and graph neural networks (GNNs) (Scarselli, Gori, Tsoi, Hagenbuchner, & Monfardini, 2008).

In respect of the first method, MGPred (Zhao, Zheng, Li, & Wang, 2021) employed drug-drug similarity, calculated through the "Combined score" in the STITCH database, and semantic similarity between SE-SE, obtained from the Directed Acyclic Graph, as components of their encoding. Building upon this, Zhao et al. proposed SDPred (Zhao, Wang, et al., 2022), which represented the drugs and SEs by leveraging multimodal information, including semantic features, structural similarity, and so on. By combining embedding information and interaction information processed through CNNs, SDPred obtained the final predicted frequencies.

Regarding the second approach, GNNs can learn topological graph structure and transfer features between nodes and edges. Based on GNNs, Xu et al. (2022) developed DSGAT, which employed Graph Attention Networks (GATs) (Veličković et al., 2017) to extract drug features from molecular graphs and SE features from similarity graphs. By computing the dot product of drug and SE feature vectors, DSGAT predicted the frequencies of the drug-side effect pairs. In addition, Yu, Cheng, Qiu, Xiao, and Lin (2022) developed a hybrid embedding graph neural network, integrating a graph embedding module that extracts medicinal chemistry information from drug molecular structures with a node embedding module that extracts drug entity associations from biological entity networks.

While the mentioned methods have made progress in predicting drug-side effects, they still exhibit limitations. These drawbacks can be summarized as follows:

- Some models (Galeano et al., 2020; Guo et al., 2020; Yu, Cheng, et al., 2022; Zhang et al., 2016; Zhao et al., 2021), overly rely on known frequency information during training, leading to inadequate predictions or poor predictive performance when it comes to new drugs lacking known frequency information.
- Models based on GNNs mostly adopted shallow GNNs to extract drug information, as deep GNNs can lead to over-smoothing and even vanishing gradient problems. Consequently, these models only extracted information from nearby neighbors, neglecting nodes that are multiple hops away in the molecular graph but physically close to the central node in 3D space.
- The integration of drugs and SEs is often accomplished through simple concatenation or dot product, which fails to fully capture the intricate interactions between heterogeneous features (Uner, Kuru, Cinbis, Tastan, & Cicek, 2022; Xu et al., 2022; Zhao et al., 2021).

To overcome the above limitations, in this paper, we introduce a novel homogeneous substructures transformer network with a refined interaction module, named HSTrans. Its contributions are as follows:

- By creatively utilizing substructures instead of relying solely on known frequency information to represent drugs and SEs, HSTrans ensures the unified embedding spaces and semantic consistency between drugs and SEs, enhancing the model's ability to understand and capture their complex relationships.
- HSTrans establishes a comprehensive interaction module, integrating scalar projection layer and CNN layer. It not only considers the influence of interactions between pair-pair substructures but also takes into account the impact of small region substructures on the final prediction.
- HSTrans extracts drug substructures directly and employs the transformer architecture to encode them, enabling the capture of associations between substructures at any distance within the same drug.
- HSTrans achieves state-of-the-art performance in Galeano et al.'s datasets and demonstrates excellent performance on the independent test set of nine new drugs.

2. Related works

In this section, we will review the development of computational methods for predicting drug-side effects and the related techniques used in this research.

2.1. Drug-side effect prediction

Drug-side effect prediction is a crucial area in bioinformatics aimed at identifying associations between drugs and specific adverse effects. Accurate prediction of SEs for new drugs enables researchers to identify potential safety issues earlier, optimize drug design and improve the efficiency of drug development. With the continuous accumulation of clinical trial data, data-driven computational methods have become increasingly important in understanding the mechanisms behind drug-side effects (Arshed et al., 2022; Ding, Zhou, Zou, & Yuan, 2023; Lee & Chen, 2021), and deep learning has emerged as a significant method for predicting these effects.

Computational methods largely focus on the feature extraction stage. For example, MGPred (Zhao et al., 2021) integrates semantic, similarity, and frequency features of drugs and SEs, using an attention mechanism to extract association information. Mv3SM (Ding, Guo, Tiwari, & Zou, 2023) combines multi-view learning with sparse regularization to effectively integrate drug and SE data while reducing noise. Qian, Ding, Zou, and Guo (2022) provide more accurate and reliable drug-side effect predictions by integrating similarity matrices, weighted K-nearest neighbors preprocessing, Restricted Boltzmann Machines model training, and average decision rules. In addition to multi-view feature integration, graph structures also be used in feature extraction stage. DSGAT (Xu et al., 2022) represents drugs and multiple SEs as atom and neighbor graphs, using graph attention networks for feature extraction and encoding. Some methods also focus on drug-side effect relationship networks, such as LAGCN (Yu, Huang, Zhao, Xiao, & Zhang, 2021), which introduces layer attention graph convolutional networks and integrates feature embeddings from different convolutional layers to predict drug-side effect interactions. HINGR (Zhao, Hu, You, Wang, & Su, 2022) integrates heterogeneous information networks of drug-side effect, drug-protein, and protein-side effect interactions, learning node features and relationships through random walks. In addition, AMDGT (Liu et al., 2024) utilizes transformers to handle multimodal information, including network relationships and similarities between drugs and SEs.

2.2. Substructure-based interaction prediction

Substructures lie between the entire substance and the monomers that constitute it, and they can reflect the physicochemical properties of the whole substance. Therefore, in studies such as Drug–Target Interaction (DTI), Cancer Drug Response (CDR) prediction, and Drug–Drug Interaction (DDI), decomposing compounds into substructures is a crucial and effective strategy. For example, MolTrans (Huang, Xiao, Glass, & Sun, 2021) uses the most frequently occurring n-grams as substructures and decomposes the SMILES of drugs and the amino acid sequences of proteins accordingly. In the CDR task, DeepTTA (Jiang et al., 2022) and iBT-Net (Zhan, Guo, Philip Chen, & Meng, 2023) also decompose drugs into substructures and utilize transformer for encoding. Additionally, STNN-DDI (Yu, Zhao, & Shi, 2022) provides detailed substructure representation of drugs by using a predefined list of chemical substructures, such as PubChem fingerprints and establishes a Substructure–Substructure Interaction (SSI) tensor to capture the interactions between these substructures.

Unlike the above tasks, in the drug-side effect prediction task, SEs do not naturally have decomposable substructures. However, since capturing the interactions between heterogeneous substances needs to occur in a unified encoding space, we choose to employ probabilistic methods to extract effective substructure representations from the drugs to model the SEs. This approach allows us to effectively integrate SEs into the same encoding framework as the drugs, facilitating accurate predictions and analyses.

3. Method

We regard the task of predicting drug-side effect frequencies as a regression problem, and our model learns to predict these frequencies as follows: Initially, given all drugs' SMILES strings, HSTrans employs the FCS algorithm (Huang et al., 2021) to extract their substructures. Building upon this and drawing from the drug-side effect frequency table, HSTrans identifies effective substructures for each SE through probabilistic statistical methods. The substructures of drugs and SEs are then inputted into a transformer for encoding. Within the interaction module, a scalar projection layer mapped substructure vectors of drugs and SEs to an "interaction value" individually. Subsequently, a CNN layer is utilized to capture neighborhood associations within the generated interaction map. Ultimately, the predictor generates frequency prediction scores between drugs and SEs. The overall flowchart of HSTrans is illustrated in Fig. 1.

3.1. Obtain drug substructure

Initially, the FCS algorithm defines a set \mathcal{V} that contains various initial tokens from the drugs' SMILES strings and tokenizes the provided drug corpus and stores the results in set \mathcal{W} . Afterward, it scans set \mathcal{W} to identify the most frequent combinations, labeled as A and B . These combinations are then merged into a single entity, AB , which is added to set \mathcal{V} . This process continues until either no combination exceeds the threshold γ , or set \mathcal{V} reaches its maximum length l . Ultimately, it obtains set $\mathcal{U} = \{sub_1, sub_2, \dots, sub_n\}$ with a total length of n , where each $sub_i \in \mathcal{V}$.

3.2. Identify effective substructure for each SE

To determine the effective substructures corresponding to each SE, we need to establish the relationship between SEs and substructures based on the known drug-side effect frequencies. Specifically, we construct matrix \mathbf{A} to represent the frequencies between drugs and SEs, where element a_{ij} denotes the frequency of drug d_i and SE s_j . Each drug d_i can be regarded as a set of substructures, denoted as

$d_i = \{sub_{i1}, sub_{i2}, \dots, sub_{id}\}$. We establish matrix \mathbf{O} to describe the relationship between substructures and SEs:

$$o(s_j, sub_k) = \sum_{d_i \in \mathcal{N}(sub_k)} a_{ij}, \quad (1)$$

where $\mathcal{N}(sub_k)$ is the set of drugs containing the substructure sub_k .

Afterward, we turn our attention to employing probability statistical methods to identify meaningful correlations in the matrix \mathbf{O} . This approach involves evaluating the variance between the observed co-occurrence probability of substructures and SEs and the expected probability. This comparison enables the assessment of non-random associations between SEs and substructures. A greater deviation indicates a potentially significant relationship between certain substructures and specific SEs. This approach is defined as follows:

$$obs(s_j, sub_k) = \frac{o(s_j, sub_k)}{N}, \quad (2)$$

$$exp(s_j, sub_k) = p(s_j)p(sub_k), \quad (3)$$

where

$$p(s_j) = \frac{\sum_{k=1}^{n_{sub}} o(s_j, sub_k)}{N}, \quad (4)$$

$$p(sub_k) = \frac{\sum_{j=1}^{n_s} o(s_j, sub_k)}{N}, \quad (5)$$

$$N = \sum_k^{n_{sub}} \sum_j^{n_s} o(s_j, sub_k). \quad (6)$$

In the context provided, $obs(s_j, sub_k)$ and $exp(s_j, sub_k)$ represent the observed co-occurrence and expected probability between s_j and sub_k . $p(s_j)$ and $p(sub_k)$ denote the marginal probabilities of s_j and sub_k , respectively. n_{sub} and n_s represent the number of substructure and SE. To evaluate the significance level between s_j and sub_k through hypothesis testing, we construct a statistic R as follows:

$$R(s_j, sub_k) = \frac{obs(s_j, sub_k) - exp(s_j, sub_k)}{\sqrt{\frac{exp(s_j, sub_k)}{N}(1 - p(s_j))(1 - p(sub_k))}}. \quad (7)$$

In previous studies (Chan, Wong, & Chiu, 1994; Ching, Wong, & Chan, 1995), the $R(s_j, sub_k)$ measure has been demonstrated to generally follow a standard normal distribution. The numerator of the formula represents the difference between two probabilities: the observed co-occurrence probability of SE j and substructure k in the dataset, and the expected co-occurrence probability assuming that SE j and substructure k are statistically independent, which is calculated by multiplying their individual marginal probabilities. And the denominator serves as a normalization factor, adjusting the numerator value to a standard normal distribution score. A higher $R(s_j, sub_k)$ value indicates a more significant association between s_j and sub_k . Considering the significance of the 95th percentile in statistical associations and the number of effective substructures obtained for each SE, we use the 95th percentile of R in each fold of the experiment as the threshold. Therefore, we can identify the substructures significantly associated with each SE. These SEs are then considered as the set of above effective substructures, denoted as $s_j = \{sub_{j1}, sub_{j2}, \dots, sub_{js}\}$.

3.3. Encoding module

Functional groups, integral to drug substructures, significantly influence drug efficacy by shaping the pharmacokinetic and pharmacodynamic properties (Tang, Chen, Yang, Zhong, & Chen, 2023). Moreover, interactions among different substructures within a single drug compound also influence the properties of the drug. To achieve an efficient encoding that reflects the characteristics of substructures, we introduce an enhanced encoding module. Similar to acquiring word embedding vectors in NLP, this module initiates by initializing a trainable substructure lookup dictionary. Subsequently, it utilizes a transformer

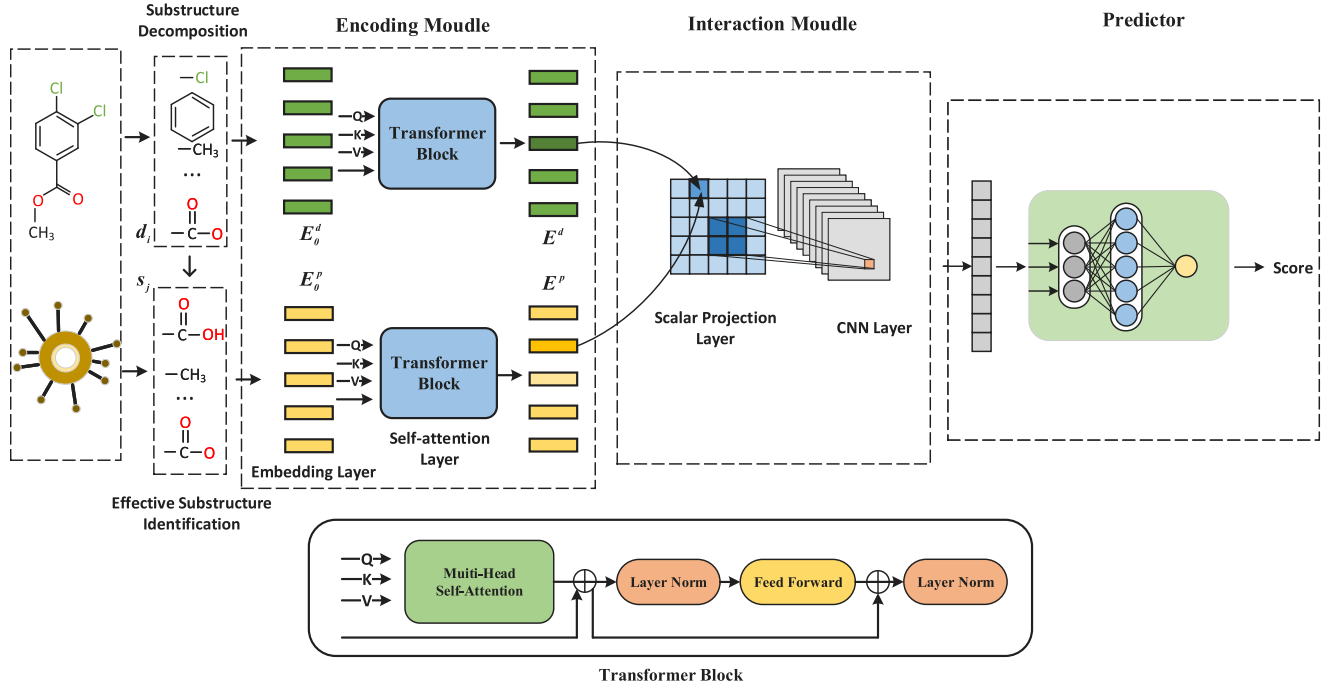


Fig. 1. The overall flowchart of HSTrans. (a) We use the FCS algorithm to decompose all drugs, obtaining a set of all substructures. Then, for drug i , we extract its contained substructures from the previously obtained set. (b) Based on the drug-side effects frequency table and substructure set, we employ probabilistic methods to derive the effective substructure set s_j for SE j . (c) Through an embedding layer, we obtain encoding vectors E_0^d and E_0^p for the drug's and SE's substructures, respectively. Following this, the embedded encoding E_0^d and E_0^p are fed into a self-attention transformer, yielding contextually adjusted vectors E^d and E^p . (d) In the interaction module, firstly, E^d and E^p pass through a scalar projection layer, and we obtain an interaction graph I measuring the strength of interaction. Then, we utilize a CNN layer to capture neighborhood connections in I . (e) The predictor receives the output from the interaction module and feeds it into a multilayer perceptron to obtain the final frequency prediction scores.

encoder (Vaswani et al., 2017) to capture contextual substructural information, thereby enriching the embedding vectors. The transformer with a self-attention mechanism has achieved great success in the field of NLP. In our framework, by employing the aforementioned transformer to capture both short-term and long-term dependencies among substructures, we are able to comprehensively analyze the intricate chemical relationships among all substructures, thus obtaining refined representations for substructures.

For drugs and SEs, we construct two substructure index matrices called $\mathbf{M}^d \in \mathbb{R}^{d \times k}$ and $\mathbf{M}^s \in \mathbb{R}^{s \times k}$, where d and s represent the number of maximum length substructures present in a drug and SE, respectively, and k represents the total number of substructures. Specifically, in the a th row of \mathbf{M}^d , if $M_{(a,i)}^d = 1$, and $M_{(a,j)}^d = 0$ ($j \neq i$), it indicates the index of a th substructure in the drug is i . Then, we look up substructure dictionary matrix $\mathbf{T}_c \in \mathbb{R}^{k \times h}$, where h denotes the dimension of embedding vector, to generate context embedding \mathbf{E}_c^d and \mathbf{E}_c^s :

$$\mathbf{E}_c^d = \mathbf{M}^d \mathbf{T}_c, \mathbf{E}_c^s = \mathbf{M}^s \mathbf{T}_c. \quad (8)$$

To encode position vector \mathbf{E}_p^d and \mathbf{E}_p^s , we similarly introduce two position look up dictionary matrices $\mathbf{T}_p^d \in \mathbb{R}^{d \times h}$ and $\mathbf{T}_{pos}^s \in \mathbb{R}^{s \times h}$:

$$\mathbf{E}_p^d = \mathbf{I}^d \mathbf{T}_p^d, \mathbf{E}_p^s = \mathbf{I}^s \mathbf{T}_{pos}^s, \quad (9)$$

where $\mathbf{I}^d \in \mathbb{R}^{d \times d}$ and $\mathbf{I}^s \in \mathbb{R}^{s \times s}$ are position index matrices. In these matrices, if the element in the i th row and j th column equals 1, it indicates that the index of the i th substructure of a drug or SE is j . By combining content encoding $\mathbf{E}_c^d, \mathbf{E}_c^s$ and positional encoding $\mathbf{E}_p^d, \mathbf{E}_p^s$, we obtain the preliminary representation \mathbf{E}_0^d and \mathbf{E}_0^s :

$$\mathbf{E}_0^d = \mathbf{E}_c^d + \mathbf{E}_p^d, \mathbf{E}_0^s = \mathbf{E}_c^s + \mathbf{E}_p^s. \quad (10)$$

Finally, we use a self-attention mechanism to obtain contextual embeddings for the substructure:

$$\begin{aligned} \mathbf{E}^d &= \text{Softmax} \left(\frac{(\mathbf{E}_0^d \mathbf{Q}_0^d)(\mathbf{E}_0^d \mathbf{K}_0^d)^T}{\sqrt{c/h_d}} \right) (\mathbf{E}_0^d \mathbf{V}_0^d), \\ \mathbf{E}^s &= \text{Softmax} \left(\frac{(\mathbf{E}_0^s \mathbf{Q}_0^s)(\mathbf{E}_0^s \mathbf{K}_0^s)^T}{\sqrt{c/h_d}} \right) (\mathbf{E}_0^s \mathbf{V}_0^s), \end{aligned} \quad (11)$$

where $\mathbf{Q}_0^d, \mathbf{K}_0^d, \mathbf{V}_0^d \in \mathbb{R}^{h \times c}$, $\mathbf{Q}_0^s, \mathbf{K}_0^s, \mathbf{V}_0^s \in \mathbb{R}^{s \times c}$ are weight parameters. c and h_d are the embedding dimensions and number of heads, respectively.

3.4. Interaction module

So far, we have obtained embeddings for both drugs' and SEs' substructures. To further explore the relationship between drugs and SEs, we introduce an interaction module. Our interaction module includes two layers, (1) a scalar projection layer to quantify the one-to-one interaction strength between drugs' and SEs' substructures, and (2) a CNN layer to capture local region interaction patterns.

3.4.1. Scalar projection layer

To quantify the interaction between substructures in each pair of drug and SE, we employ the dot product, transforming their encoding vectors into a singular value. A higher value, reflecting a strong association among responding substructures, is more likely to be activated during the prediction process, and vice versa. Consequently, we derive an interaction map \mathbf{I} as follows:

$$\mathbf{I} = \mathbf{E}^d \cdot \mathbf{E}^s. \quad (12)$$

3.4.2. CNN layer

The interaction between pairs of substructures is not isolated: they are influenced by neighboring substructures, and in turn, can also af-

fect surrounding substructures. Based on this understanding, we believe that **I** exhibits local correlations. Similar to the local correlations in images, we employ CNN with small order-invariant convolution filters to capture the interactions among neighborhoods in the interaction map **I**. Consequently, we obtain the output **M** for the drug-side effect pair.

$$\mathbf{M} = \text{CNN}(\mathbf{I}). \quad (13)$$

Finally, we flatten the **M** into a vector and feed it into a multilayer perceptron consisting of three fully connected layers to obtain the final predicted frequency scores:

$$\mathbf{O}_1 = \text{ReLU}(\mathbf{W}_1 \text{Flatten}(\mathbf{M}) \mathbf{b}_1), \quad (14)$$

$$\text{Score} = \mathbf{W}_4 \text{ReLU}(\mathbf{W}_3 \text{ReLU}(\mathbf{W}_2 \mathbf{O}_1 + \mathbf{b}_2) + \mathbf{b}_3) + \mathbf{b}_4, \quad (15)$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4$ are weight parameters and $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4$ are biases.

3.5. Loss function

We regard the task of predicting drug-side effect frequencies as a regression problem, aiming to minimize the discrepancy between the prediction and the true frequency. Therefore, we employ the mean squared error as the loss function as follows:

$$\text{Loss} = \sum_{(i,j) \in \mathcal{N}} (\hat{y}_{ij} - y_{ij})^2, \quad (16)$$

where \mathcal{N} represents the set of drug-side effect pairs in the training dataset. \hat{y}_{ij} and y_{ij} denotes the predicted and true frequency of i th drug- j th SE, respectively.

4. Results and discussion

4.1. Implementation details

HSTrans is implemented in the PyTorch framework and runs on an RTX 3090 GPU. During the substructure extraction, we extract 50 substructures for each drug and SE, adjusting for this consistent count through truncation or filling. In the encoding module, each substructure is encoded as a 300-dimensional vector. Additionally, for the self-attention transformer, we set up 8 transformer layers, each with 8 attention heads. And we set the hidden layer dimension and output feature dimension to 512 and 300, respectively. In the interaction module, we utilize 10 convolutional kernels with a size of 3 for the CNN layer. For hyperparameter optimization, we employ the Adam optimizer with a learning rate of $1e-4$. Additionally, we set the batch size to 128, the dropout rate to 0.1, and conduct training for 300 epochs.

4.2. Datasets

We utilize the benchmark dataset from Xu et al. (2022), Galeano et al. (2020), and Zhao et al. (2021) to assess the performance of HSTrans. The dataset comprises 750 drugs and 994 SEs, totaling 37,071 known frequency entries. The frequency annotations for drug-side effect pairs are classified into five categories based on $p = \frac{n_1}{n_2}$, where n_1 is the number of people who experience a specific SE after taking a certain drug, and n_2 is the total number of people taking that drug. Specifically, the categories are: very rare ($p < \frac{1}{10000}$ and frequency = 1), rare ($\frac{1}{10000} \leq p < \frac{1}{1000}$ and frequency = 2), infrequent ($\frac{1}{1000} \leq p < \frac{1}{100}$ and frequency = 3), frequent ($\frac{1}{100} \leq p < \frac{1}{10}$ and frequency = 4), and very frequent ($p \geq \frac{1}{10}$ and frequency = 5). And the frequency labels among the pairs are distributed as follows: very rare (3.21%), rare (11.29%), infrequent (26.92%), frequent (47.46%), and very frequent (11.12%). The remaining unannotated drug-side effects are considered as category 0 (i.e. not discovered in clinical trials) (Fig. 2).

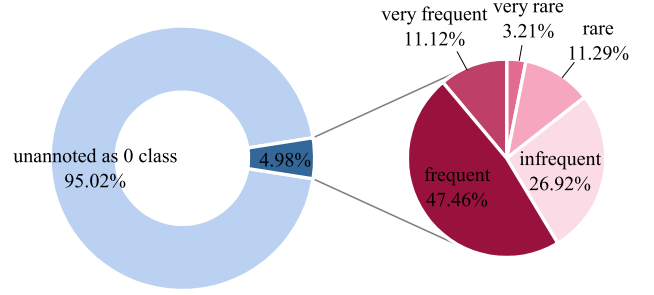


Fig. 2. The distribution of drug-side effects frequencies in the dataset.

4.3. Evaluation framework and metrics

To ensure a balance between positive and negative samples, we randomly selected 37,071 instances of class 0 data as negative samples. Following this, we assess the performance of HSTrans using a 5-fold cross-validation on the drug-side effect frequency dataset mentioned above. Pairs of drugs-side effects are divided into 5 distinct subsets randomly. Each subset is then utilized as the test set, while the remaining nine subsets are consecutively employed as the training set. The final evaluation of HSTrans's performance is conducted by averaging the results obtained across all 5 folds.

We evaluate HSTrans's performance from two perspectives: accuracy in frequency prediction and association prediction.

For frequency prediction, we utilize Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as evaluation metrics:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (17)$$

and

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (18)$$

where y_i and \hat{y}_i is observed value and predicted value, respectively. And n is the number of drug-side effect pairs.

Regarding the association prediction of drug-side effects, we employ Spearman's rank correlation coefficient (SCC) for the comparison of distinct models' performance as follows:

$$SCC = 1 - \frac{6 \sum_{(i,j) \in \mathcal{N}} d_{ij}^2}{|\mathcal{N}|(|\mathcal{N}|^2 - 1)}, \quad (19)$$

where \mathcal{N} represents the set of drug-side effect pairs in the test dataset. And d_{ij} represents the difference in ranks between the predicted frequency \hat{y}_{ij} and the true frequency y_{ij} . Considering that the high-frequency SEs of drugs have a greater impact on patients' quality of life and treatment outcomes compared to less common SEs, we referred evaluation metrics from recommendation systems and introduced Overlap@N% (N is set to 1, 5, 10, and 20) to evaluate the recommendation performance for the top N%-ranked drug-side effects as follows:

$$\text{Overlap@N\%} = \frac{TP}{T \times N\%}, \quad (20)$$

where TP is the number of positive samples in the top N% of the predicted results, and T is the total number of samples in the test set.

4.4. Comparing the performance of HSTrans with existing approaches

To assess our model's effectiveness, we compare it with the top-performing models in the benchmark dataset.

- **MGPred** (Zhao et al., 2021), based on a deep learning architecture, utilizes a graph attention network to integrate features of similarity, frequency, and word embedding, to predict frequencies

Table 1
Comparison results for forecasting drug-side effects frequencies.

Method	RMSE	MAE	SCC	Overlap@1%	Overlap@5%	Overlap@10%	Overlap@20%
Ridge Regression	1.576 ± 0.004	1.328 ± 0.004	0.352 ± 0.009	0.054 ± 0.018	0.232 ± 0.003	0.317 ± 0.009	0.480 ± 0.006
Random Forest	1.430 ± 0.005	1.123 ± 0.004	0.405 ± 0.005	0.084 ± 0.007	0.300 ± 0.006	0.346 ± 0.008	0.506 ± 0.006
MGPred	1.389 ± 0.012	0.975 ± 0.007	0.412 ± 0.009	0.108 ± 0.005	0.351 ± 0.011	0.391 ± 0.005	0.533 ± 0.007
A ³ Net	1.437 ± 0.011	1.053 ± 0.009	0.423 ± 0.015	0.092 ± 0.026	0.305 ± 0.004	0.368 ± 0.011	0.527 ± 0.007
DSGAT	1.300 ± 0.006	0.959 ± 0.028	0.485 ± 0.008	0.114 ± 0.018	0.383 ± 0.021	0.411 ± 0.010	0.567 ± 0.004
HSTrans	1.390 ± 0.009	0.725 ± 0.008	0.527 ± 0.012	0.129 ± 0.025	0.463 ± 0.012	0.436 ± 0.012	0.579 ± 0.003

Table 2
The frequency prediction scores for the SEs of atorvastatin and telithromycin post-market release.

Drug	SE	Predicted score
atorvastatin	acute coronary syndrome	0.669
	renal impairment	1.780
	speech disorder	1.888
	ventricular fibrillation	2.058
	hypokalaemia	2.923
	arrhythmia	2.938
	angioedema	3.345
	infection	3.749
telithromycin	myalgia	0.556
	loss of consciousness	0.786
	hypersensitivity	1.010
	anaemia	2.345
	ascites	2.492
	abdominal distension	2.984
	hepatic function abnormal	3.245
	pain	3.454

of drug-side effects. We used grid search to determine the optimal parameters for MGPred. Specifically, we tuned the dimensionalities of projection Q with values [32, 64, 128, 256] and the weight decay w with values [1×10^{-3} , 5×10^{-3} , 1×10^{-2} , 5×10^{-2}]. Based on the results from the test set, we selected $Q = 64$ and $w = 5 \times 10^{-3}$ as the optimal parameters.

- **DSGAT** (Xu et al., 2022) is an encoder–decoder framework for predicting drug-side effect frequencies. It employs GATs to encode drug molecule graphs and the similarity graph of SEs, with matrix factorization serving as the decoder. Meanwhile, similar to MGPred, we also employed grid search to tune the parameters for DSGAT. The parameters we adjusted include the number of GAT layers N with values [1, 2, 3, 4] and the dimensionalities of projection Q with values [100, 200, 300]. Based on the results from the test set, we selected $N = 3$ and $Q = 200$ as the optimal parameters.
- **A³ Net** (Jin, Wang, Zheng, Chen, & Tang, 2024) is a graph neural network-based model for predicting the frequencies of drug-side effects. It encodes drugs and SEs by GAT, utilizes a cross-attention module to learn interaction features between drugs’ atoms and SEs, and finally aggregates graph features with interaction features to obtain prediction results. Similarly, based on the grid search results, we tuned the number of attention heads in CAM h_1 with values [1, 4, 8, 20], the number of attention heads in the GAT layer h_2 with values [1, 5, 10, 20], and the embedding size Q with values [100, 200, 300, 400]. We selected $h_1 = 8$, $h_2 = 10$, and $Q = 200$ as the optimal parameters.
- **Ridge Regression** (Hoerl & Kennard, 1970), a variant of linear regression, enhances model generalization by introducing L_2 regularization. In the process of constructing drug features, we used Mol2vec (Jaeger, Fulle, & Turk, 2018), an unsupervised learning method, to represent drugs as 100-dimensional vectors based on their structural characteristics. So drugs with similar structures are closer in this vector space. For SEs, we employed Glove (Pennington, Socher, & Manning, 2014) to obtain their word vectors. Subsequently, we concatenated the drug vectors

with the SE vectors and fed the results into a ridge regression model for learning. The model was built using the sklearn library (Pedregosa et al., 2011). During training, we selected the optimal regularization coefficient, denoted as β , from the set [0.1, 0.5, 1.0, 3.0]. Based on a comparison of model performance, we set $\beta = 1.0$.

- **Random Forest** (Breiman, 2001) is an ensemble learning method that improves overall model performance by constructing multiple decision trees. The feature inputs are identical to that of the Ridge Regression model. During the training process, we selected the number of decision trees n_e from the set [100, 200, 300], and the maximum depth of the decision trees m_d from the set [10, 20, 30]. Through grid search, we set $n_e = 200$, $m_d = 20$.

Table 1 presents the experiment results of different models. It shows HSTrans outperforms other models with a relatively large margin by achieving an Overlap@1% value of 0.129 as compared to 0.054 of Ridge Regression and 0.084 of Random Forest, 0.108 of MGPred, 0.092 of A³ Net and 0.114 of DSGAT. This implies that HSTrans could discern the intrinsic associations between drugs and SEs effectively. This conclusion is also consistent considering other metrics such as SCC, Overlap@5% and so on. It is worth noting that in the HSTrans prediction model, although the MAE is significantly lower compared to other models, the RMSE does not show a similar downward trend. We speculate that this may be because HSTrans predicts some zero-frequency samples as having higher frequencies. While this leads to an increase in RMSE, it could actually serve as a warning. These drug-side effects, although not currently observed in the population, may pose a risk over time or under certain conditions. To verify our hypothesis, we randomly selected two drugs: atorvastatin and telithromycin, and collected post-market SE associations from the SIDER and OFFSIDES databases. These associations, not observed during clinical trials, are labeled as class 0 in the training dataset. Table 2 illustrates frequency prediction scores for the SEs of two drugs post-market release. As we suspected, HSTrans predicts some zero-frequency samples as having higher frequencies, such as 2, 3, or 4. This prediction approach of the HSTrans model can provide additional value in drug safety assessments by helping to identify and prevent potential drug-side effects in advance, ensuring patient safety.

To further illustrate our model’s effectiveness, we randomly selected cladribine and estrone as case studies, limiting each frequency category’s samples to no more than 6. Fig. 3 compares the HSTrans’s prediction scores with the true frequencies of these drugs.

4.5. Ablation experiments

In the same setup, we conduct ablation experiments to assess the individual impacts of the self-attention layer, and interaction module. The specific treatments are as follows:

- **HSTrans/ SAL**: We exclude the Self Attention Layer from the encoding module, and use the embeddings of substructures obtained from the embedding layer, E_0^d and E_0^s , as the input of the interaction module.
- **HSTrans/ IM**: We completely remove the Interaction Module from HSTrans. Instead, we concatenate the embeddings of drugs and SEs obtained from the encoding module and then input them into a multilayer perceptron to generate prediction frequency scores.

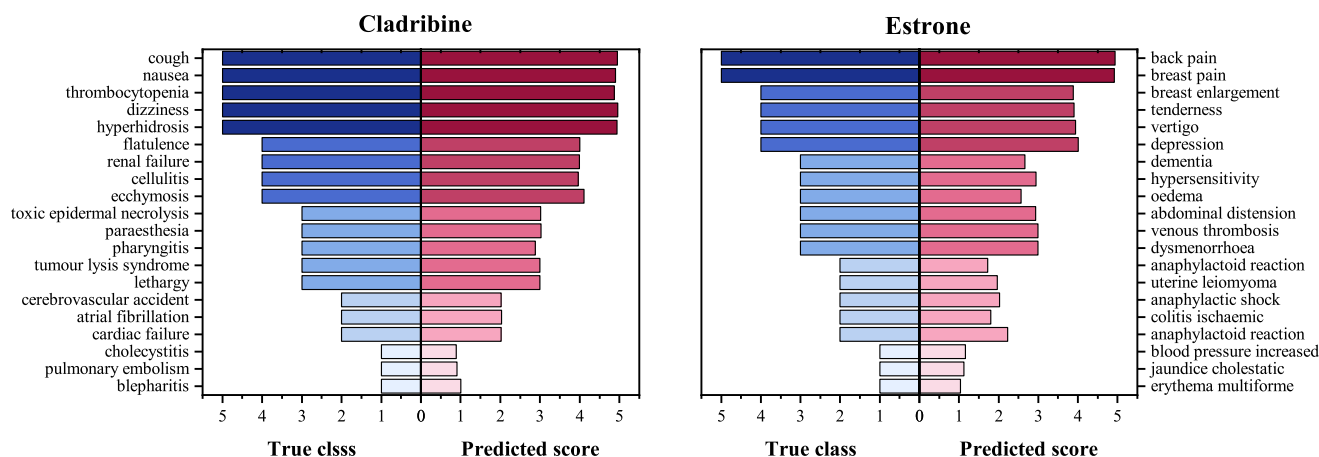


Fig. 3. Comparing the predicted scores and true frequencies of cladribine and estrone under the HSTrans model.

Table 3

Ablation experiment results.

Method	RMSE	MAE	SCC	Overlap@1%	Overlap@5%	Overlap@10%	Overlap@20%
HSTrans/ SAL	1.419 \pm 0.007	0.842 \pm 0.025	0.440 \pm 0.012	0.113 \pm 0.009	0.396 \pm 0.013	0.402 \pm 0.008	0.554 \pm 0.009
HSTrans/ IM	1.402 \pm 0.008	0.767 \pm 0.007	0.505 \pm 0.011	0.128 \pm 0.016	0.451 \pm 0.011	0.429 \pm 0.005	0.568 \pm 0.008
HSTrans	1.390 \pm 0.009	0.725 \pm 0.008	0.527 \pm 0.012	0.129 \pm 0.025	0.463 \pm 0.012	0.436 \pm 0.012	0.579 \pm 0.003

In Table 3, the MAE of HSTrans/SAL increases significantly. We believe that the Self-Attention Layer is a crucial component for capturing the associations between substructures of the same drug. SAL effectively establishes a fully connected graph of drug substructures, enabling the model to learn the relationships and topological structures between these substructures. By learning from other substructures, SAL refines the embedding representations of its own substructures, resulting in more precise embeddings from the perspective of drug components. Removing the self-attention layer leads to less accurate embeddings of drug substructures, failing to capture the internal associations and topological information of the drug, which in turn causes an increase in the final MAE. Additionally, the decreased accuracy of individual drug-side effect predictions further affects the association prediction results. We observed that the recommendation performance for the top 5% of drug-side effects significantly declines in HSTrans/SAL. However, as the value of N increases, the tolerance range for Overlap@N% also expands, leading to some improvement in the performance of HSTrans/SAL. On the contrary, due to the more refined embedding encoding of drug and SE substructures, the results of HSTrans/IM are slightly better than those of HSTrans/SAL. The interaction module converts the final frequency prediction into a combination of multiple substructure interaction values, while also using CNN to consider the interactions among neighboring substructures. This module, through decomposition and combination, allows for a more flexible prediction of frequency values. According to the RMSE and MAE results, this approach is more accurate compared to directly concatenating feature vectors to obtain frequency values. Additionally, by fully considering the interactions between drug and SE substructures, the interaction module also shows improvement in the association metrics.

4.6. Hyperparameter analysis

In this section, we will examine the impact of various hyperparameters on the model's performance and assess its effectiveness based on the settings and metrics from the first experiment mentioned above.

4.6.1. Kernel size

The convolutional kernel size k is a significant factor in the interaction module, directly affecting the model's capacity to capture features of substructure interactions. Fig. 4A illustrates that as the kernel size increases, there is a pattern of AUC initially rising before declining, and MAE initially decreasing before rising. This occurs because the model overlooks the influence of surrounding substructure pairs when k is too small. Conversely, excessively large k may introduce noise despite covering more information, thus affecting model performance. Therefore, we set k to 3.

4.6.2. Embedding size

In HSTrans, embedding size is a critical parameter for representing features of drugs and SEs, which impacts the quality of embeddings. Specifically, a low embedding dimension in HSTrans may fail to adequately capture the substructure features within the data, resulting in underfitting. Conversely, a large embedding size would increase the risk of overfitting. Fig. 4B illustrates the model's AUC and MAE metrics across embedding dimensions of 100, 200, 300, and 400. We observe that the model achieved optimal performance at an embedding dimension of 300.

4.7. Independent test

Predicting the SEs of new drugs is crucial for drug development, and it is also one of the core applications of computational methods in pharmacology. Considering the outstanding performance of HSTrans in the aforementioned experiments, we now employ the nine new drugs from Galeano et al. as an independent test set. To further investigate our model's capability, we included an additional 93 pairs of drug-side effects discovered after market release, supplementing the existing 370 pairs. These SEs are classified as very rare (frequency = 1), as they were not detected during the initial drug trials. Fig. 5 illustrates the prediction scores for 994 SEs across the nine new drugs. The results show that HSTrans excels in predicting correlations and frequencies, particularly in high-frequency adverse event tasks. Not only that, it can still detect correlations for low-frequency adverse events. However, due to the limited number of samples in the dataset with frequency = 1, there is a slight deviation in the predicted frequency values.

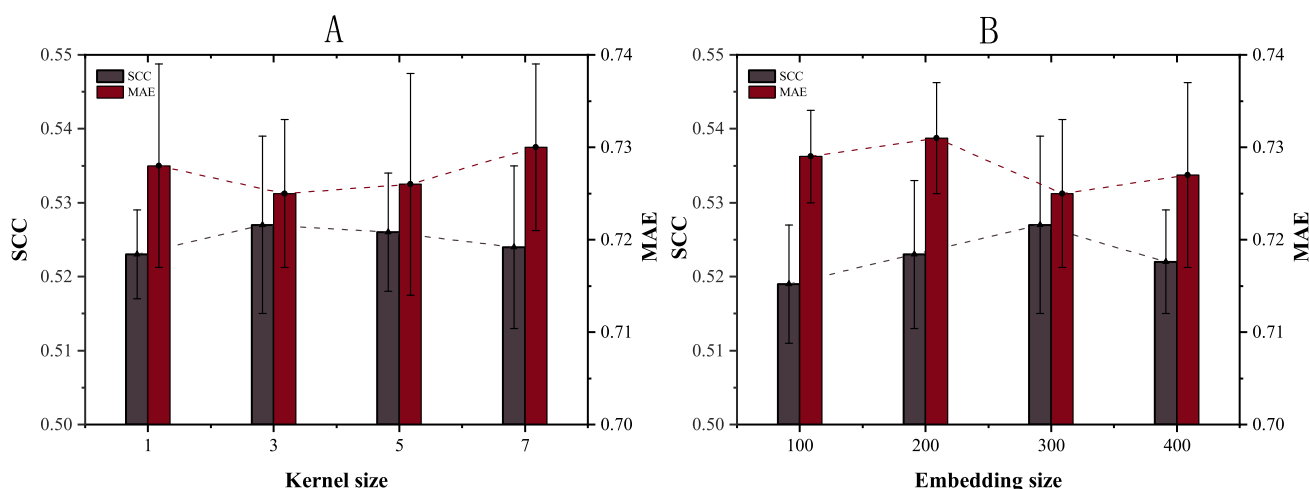


Fig. 4. Hyperparameter Analysis. The trends of SCC and MAE values with varying kernel sizes and embedding sizes for the test set. A illustrates the scenarios with kernel sizes of 1, 3, 5, and 7, while B presents the cases with embedding sizes of 100, 200, 300, and 400.

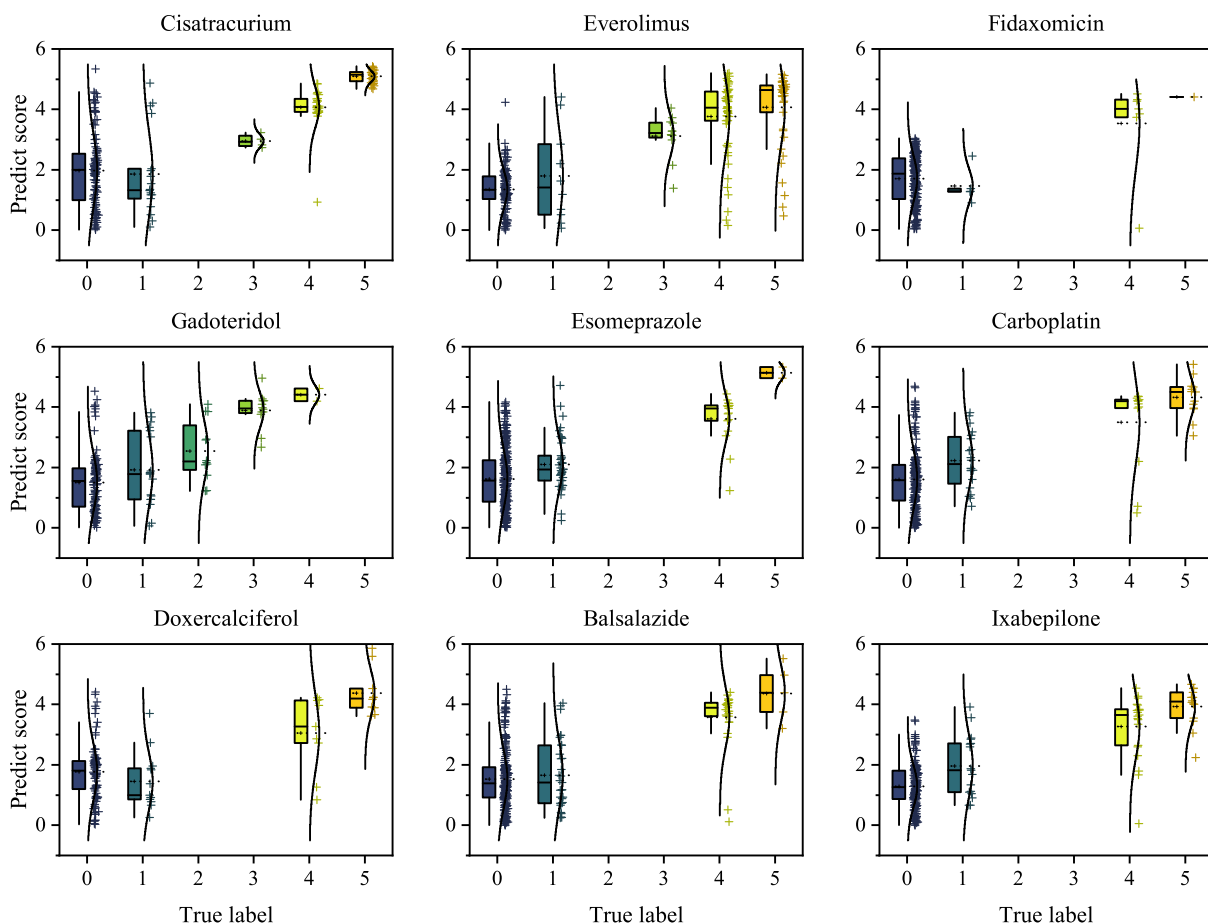


Fig. 5. Box plots accompanied by scatter plots illustrate predicted scores for new nine drugs' items. Class 0, representing unknown associations, is excluded from the scatter plot due to the abundance of unknown SEs linked to a given drug.

4.8. Visualization analysis

In the interaction module, the values of interaction map I between drug and SE substructures directly reflect their correlation. To verify whether HSTrans truly captures this correlation, we visualize the interaction map as a heatmap. We select the following drug-side effect pairs: doxorubicin and vomiting (frequency = 5), doxorubicin and abnormal behavior (frequency = 0), and cladribine and vomiting

(frequency = 5). For drugs: doxorubicin and cladribine, we extracted 33 and 16 substructures respectively, and padded them with 0 to reach 50. Similarly, for SEs: vomiting and abnormal behavior, we obtained 57 and 29 effective substructures, which were also truncated or padded to 50. Fig. 6 displays their interaction maps, with highlighted areas (purple) representing the interaction values between self-substructures. We observe that for pairs with a frequency of 5 (Fig. 6A, 6C), the interaction values among their self-substructures are generally high.

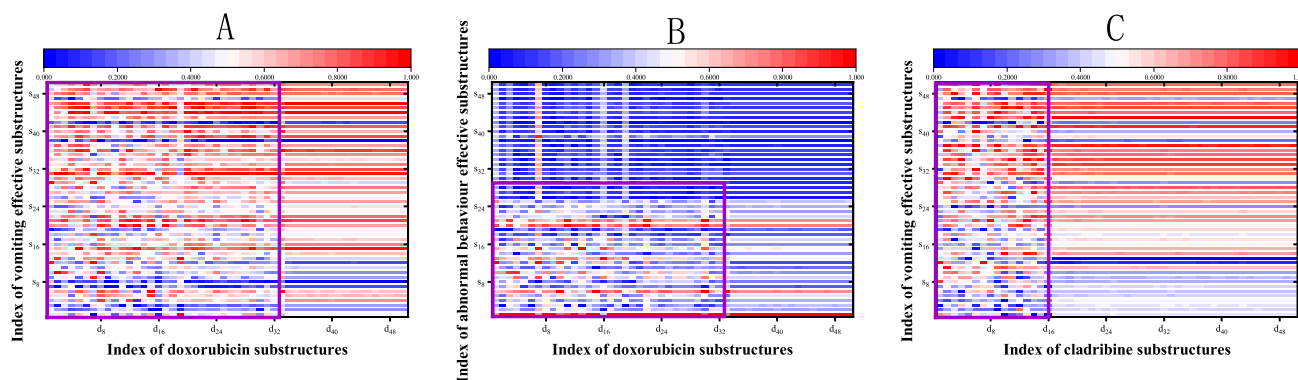


Fig. 6. The interaction value heatmap between drug and SE substructures, with highlighted areas (purple) indicating the values between self-substructures. Specifically, A represents the interaction between doxorubicin and vomiting (frequency = 5), B represents the interaction between doxorubicin and abnormal behavior (frequency = 0), and C represents the interaction between cladribine and vomiting (frequency = 5).

In contrast, for pair with a frequency of 0 (Fig. 6B), the interaction values are generally low. This indicates that our model effectively identifies and activates the association between drug and SE substructures, achieving good correlation prediction performance.

5. Conclusion

Rapid and accurate assessment of drug-side effects is crucial in drug discovery. Computational approaches have gained attention for their efficiency and cost-effectiveness compared to lengthy and costly clinical trial methods.

In this paper, we introduce HSTrans, an end-to-end homogeneous substructures transformer network designed for predicting drug-side effect frequencies. To our knowledge, this is the first study to utilize the transformer framework to fully represent both drug and SE features in this domain.

Initially, drugs are decomposed into sets of substructures, and based on the drug-side effect frequency table, effective substructures are identified for each SE using probabilistic statistical methods. Next, a transformer framework is utilized in the encoding module to learn the embeddings of substructures. Subsequently, in the interaction module, we capture the interactions between individual drugs and SE substructures, as well as the interactions among substructure regions, using the scalar projection layer and CNN layer. Finally, prediction frequency scores are obtained through a multilayer perceptron.

In previous studies, embeddings of SEs typically focused on their similarity or frequency values, overlooking the unique characteristics of SEs themselves. However, in HSTrans, we use substructures to represent both drugs and SEs, thus unifying their feature representation frameworks and avoiding the complexity of interactions between different feature dimensions in subsequent processing.

The results from the benchmark dataset demonstrate that HSTrans achieves significant improvements and outperforms the current state-of-the-art methods. Independent test also demonstrates that HSTrans accurately predicts the frequencies of SEs for new drugs, offering more precise guidance for drug development. Nonetheless, our current focus on drug structure neglects considerations of drug targets and in vivo responses. Additionally, our model lacks robust biological interpretability, which we aim to improve in future iterations.

Funding

This work was supported by the National Natural Science Foundation of China under Grant 62476258, and in part by the Tianjin Key Medical Discipline (Specialty) Construction under Grant TJYXZDXK-055B.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Arshed, M. A., Mumtaz, S., Riaz, O., Sharif, W., & Abdullah, S. (2022). A deep learning framework for multi-drug side effects prediction with drug chemical substructure. *International Journal of Innovations in Science & Technology*, 4(1), 19–31.
- Berry, D. C., Knapp, P., & Raynor, D. (2002). Provision of information about drug side-effects to patients. *The Lancet*, 359(9309), 853–854.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Chan, H. S., Shan, H., Dahoun, T., Vogel, H., & Yuan, S. (2019). Advancing drug discovery via artificial intelligence. *Trends in Pharmacological Sciences*, 40(8), 592–604.
- Chan, K. C., Wong, A. K., & Chiu, D. K. (1994). Learning sequential patterns for probabilistic inductive prediction. *IEEE Transactions on Systems, Man and Cybernetics*, 24(10), 1532–1547.
- Ching, J. Y., Wong, A. K. C., & Chan, K. C. C. (1995). Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7), 641–651.
- Dimitri, G. M., & Lió, P. (2017). DrugClust: a machine learning approach for drugs side effects prediction. *Computational Biology and Chemistry*, 68, 204–210.
- Ding, Y., Guo, F., Tiwari, P., & Zou, Q. (2023). Identification of drug-side effect association via multi-view semi-supervised sparse model. *IEEE Transactions on Artificial Intelligence*.
- Ding, Y., Tang, J., & Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing*, 325, 211–224.
- Ding, Y., Zhou, H., Zou, Q., & Yuan, L. (2023). Identification of drug-side effect association via coreentropy-loss based matrix factorization with neural tangent kernel. *Methods*, 219, 73–81.
- Galeano, D., Li, S., Gerstein, M., & Paccanaro, A. (2020). Predicting the frequencies of drug side effects. *Nature Communications*, 11(1), 4575.
- Guo, X., Zhou, W., Yu, Y., Ding, Y., Tang, J., & Guo, F. (2020). A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment. *BioMed Research International*, 2020.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1), 69–82.
- Huang, S., Wang, M., Zheng, X., Chen, J., & Tang, C. (2024). Hierarchical and dynamic graph attention network for drug-disease association prediction. *IEEE Journal of Biomedical and Health Informatics*.
- Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2021). MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, 37(6), 830–836.
- Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, 162(6), 1239–1249.
- Hussain, S., Anees, A., Das, A., Nguyen, B. P., Marzuki, M., Lin, S., et al. (2020). High-content image generation for drug discovery using generative adversarial networks. *Neural Networks*, 132, 353–363.
- Jaeger, S., Fulle, S., & Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 58(1), 27–35.
- Jiang, L., Jiang, C., Yu, X., Fu, R., Jin, S., & Liu, X. (2022). DeepTTA: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics*, 23(3), bbac100.

- Jin, Z., Wang, M., Zheng, X., Chen, J., & Tang, C. (2024). Drug side effects prediction via cross attention learning and feature aggregation. *Expert Systems with Applications*, 248, Article 123346.
- Lee, C. Y., & Chen, Y. P. P. (2021). Descriptive prediction of drug side-effects using a hybrid deep learning model. *International Journal of Intelligent Systems*, 36(6), 2491–2510.
- Lee, W. P., Huang, J. Y., Chang, H. H., Lee, K. T., & Lai, C. T. (2017). Predicting drug side effects using data analytics and the integration of multiple data sources. *IEEE Access*, 5, 20449–20462.
- Liu, J., Guan, S., Zou, Q., Wu, H., Tiwari, P., & Ding, Y. (2024). AMDGT: Attention aware multi-modal fusion using a dual graph transformer for drug-disease associations prediction. *Knowledge-Based Systems*, 284, Article 111329.
- Liu, Y., Tong, S., & Chen, Y. (2023). HMM-GDAN: Hybrid multi-view and multi-scale graph duplex-attention networks for drug response prediction in cancer. *Neural Networks*, 167, 213–222.
- Lv, Q., Zhou, J., Yang, Z., He, H., & Chen, C. Y. C. (2023). 3D graph neural network with few-shot learning for predicting drug-drug interactions in scaffold-based cold start scenario. *Neural Networks*, 165, 94–105.
- Muñoz, E., Nováček, V., & Vandenbussche, P. Y. (2019). Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Briefings in Bioinformatics*, 20(1), 190–202.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).
- Pirmohamed, M., Breckenridge, A. M., Kitteringham, N. R., & Park, B. K. (1998). Adverse drug reactions. *Bmj*, 316(7140), 1295–1298.
- Qian, Y., Ding, Y., Zou, Q., & Guo, F. (2022). Identification of drug-side effect association via restricted Boltzmann machines with penalized term. *Briefings in Bioinformatics*, 23(6), bbac458.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80.
- Si, J., Tian, Z., Li, D., Zhang, L., Yao, L., Jiang, W., et al. (2023). A multi-modal clustering method for traditional Chinese medicine clinical data via media convergence. *CAAI Transactions on Intelligence Technology*, 8(2), 390–400.
- Stricker, B. H., & Psaty, B. M. (2004). Detection, verification, and quantification of adverse drug reactions. *Bmj*, 329(7456), 44–47.
- Tang, Z., Chen, G., Yang, H., Zhong, W., & Chen, C. Y. C. (2023). DSIL-DDI: A domain-invariant substructure interaction learning for generalizable drug-drug interaction prediction. *IEEE Transactions on Neural Networks and Learning Systems*.
- Tang, C., Wang, J., Zheng, X., Liu, X., Xie, W., Li, X., et al. (2023). Spatial and spectral structure preserved self-representation for unsupervised hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13.
- Tang, C., Zheng, X., Zhang, W., Liu, X., Zhu, X., & Zhu, E. (2023). Unsupervised feature selection via multiple graph fusion and feature weight learning. *Science China. Information Sciences*, 66(5), Article 152101.
- Uner, O. C., Kuru, H. I., Cinbis, R. G., Tastan, O., & Cicek, A. E. (2022). DeepSide: a deep learning approach for drug side effect prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1), 330–339.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Wang, Z., Wu, B., Ota, K., Dong, M., & Li, H. (2023). A multi-scale self-supervised hypergraph contrastive learning framework for video question answering. *Neural Networks*, 168, 272–286.
- Wu, H., Liu, J., Jiang, T., Zou, Q., Qi, S., Cui, Z., et al. (2024). AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, 169, 623–636.
- Xu, X., Yue, L., Li, B., Liu, Y., Wang, Y., Zhang, W., et al. (2022). DSGAT: predicting frequencies of drug side effects by graph attention networks. *Briefings in Bioinformatics*, 23(2), bbab586.
- Yu, L., Cheng, M., Qiu, W., Xiao, X., & Lin, W. (2022). idse-HE: Hybrid embedding graph neural network for drug side effects prediction. *Journal of Biomedical Informatics*, 131, Article 104098.
- Yu, Z., Huang, F., Zhao, X., Xiao, W., & Zhang, W. (2021). Predicting drug-disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics*, 22(4), bbab243.
- Yu, H., Zhao, S., & Shi, J. (2022). Stnn-ddi: a substructure-aware tensor neural network to predict drug-drug interactions. *Briefings in Bioinformatics*, 23(4), bbac209.
- Zhan, Y., Guo, J., Philip Chen, C., & Meng, X. B. (2023). iBT-Net: an incremental broad transformer network for cancer drug response prediction. *Briefings in Bioinformatics*, 24(4), bbab256.
- Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., & Xiao, W. (2016). Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing*, 173, 979–987.
- Zhao, B. W., Hu, L., You, Z. H., Wang, L., & Su, X. R. (2022). HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Briefings in Bioinformatics*, 23(1), bbab515.
- Zhao, H., Wang, S., Zheng, K., Zhao, Q., Zhu, F., & Wang, J. (2022). A similarity-based deep learning approach for determining the frequencies of drug side effects. *Briefings in Bioinformatics*, 23(1), bbab449.
- Zhao, H., Zheng, K., Li, Y., & Wang, J. (2021). A novel graph attention model for predicting frequencies of drug-side effects from multi-view data. *Briefings in Bioinformatics*, 22(6), bbab239.