BMC Bioinformatics

RESEARCH

Open Access

Crossfeat: a transformer-based cross-feature learning model for predicting drug side effect frequency

Bin Baek[1] and Hyunju Lee[1,2*]

*Correspondence:
hyunjulee@gist.ac.kr

[1] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea
[2] AI Graduate School, Gwangju Institute of Science and Technology, Gwangju 61005, Korea

Abstract

Background: Safe drug treatment requires an understanding of the potential side effects. Identifying the frequency of drug side effects can reduce the risks associated with drug use. However, existing computational methods for predicting drug side effect frequencies heavily depend on known drug side effect frequency information. Consequently, these methods face challenges when predicting the side effect frequencies of new drugs. Although a few methods can predict the side effect frequencies of new drugs, they exhibit unreliable performance owing to the exclusion of drug-side effect relationships.

Results: This study proposed CrossFeat, a model based on convolutional neural network-transformer architecture with cross-feature learning that can predict the occurrence and frequency of drug side effects for new drugs, even in the absence of information regarding drug-side effect relationships. CrossFeat facilitates the concurrent learning of drugs and side effect information within its transformer architecture. This simultaneous exchange of information enables drugs to learn about their associated side effects, while side effects concurrently acquire information about the respective drugs. Such bidirectional learning allows for the comprehensive integration of drug and side effect knowledge. Our five-fold cross-validation experiments demonstrated that CrossFeat outperforms existing studies in predicting side effect frequencies for new drugs without prior knowledge.

Conclusions: Our model offers a promising approach for predicting the drug side effect frequencies, particularly for new drugs where prior information is limited. CrossFeat's superior performance in cross-validation experiments, along with evidence from case studies and ablation experiments, highlights its effectiveness.

Keywords: Drug adverse event frequency prediction, Drug side effect frequency, Deep learning prediction model, Transformer

Background

Most drugs interact with several molecular targets in an organism, thereby showing the complex profiles of human biological activity [1]. Drug side effects, also known as adverse drug reactions (ADRs), are defined as harmful, undesirable, and unintended

secondary effects related to pharmacological properties at normal doses [2]. They have caused notable morbidity and mortality over the centuries [3, 4]. In recent years, significant attention was paid to drug safety concerns arising from side effects [5–7]. The conventional approach for identifying ADRs often encounters issues such as high costs and time consumption owing to the need for rigorous monitoring of side effects in hospitalized patients [8, 9]. Therefore, computational methods and bioinformatics for alternative drug prediction have emerged as prime innovations to ensure safe and reasonable drug use.

Many researchers have proposed machine-learning approaches to predict the occurrence of drug side effects and ensure safe drug use and treatment [10–14]. Although predicting the presence or absence of drug side effects is important, the prediction of side effect frequencies holds greater importance for patient care in clinical practice and pharmaceutical companies. This represents a crucial step towards ensuring safe drug use. The frequency of drug side effects refers to the number of patients who experience side effects caused by the drug. Galeano et al. [15] categorized the occurrence of drug side effects into five frequency classes, ranging from one to five, and then utilized this benchmark dataset to predict the frequency of drug-related adverse effects. MGPred [16], employing a graph attention model, while DSGAT [17], which is a graph attention network that utilizes a new loss function for predicting side effect frequencies, were employed for this prediction. Despite these efforts, current studies have significant limitations. Many models directly utilize known side effect frequency values as input features or construct features based on these known values, rendering them highly reliant on existing data and unsuitable for predicting side effects for new drugs that lack historical side effect information. Methods like SDPred [18] attempt to predict side effect frequencies for new drugs but exhibit low performance metrics, such as the area under the precision-recall curve (AUPRC), indicating their limited ability to generalize to new drugs. This limitation arises from the fact that these methods still rely on derived features from existing side effect data. Additionally, models that heavily depend on pre-existing frequency data can struggle with robustness, as any inaccuracies or biases in the historical data can carry over into the predictions, leading to less reliable results for new drugs.

In contrast, our proposed CrossFeat model addresses these limitations by not relying on pre-existing side effect frequency data for feature construction. Instead, it leverages a cross-feature learning approach that integrates features from drug and side effect encoders through a transformer-based architecture. This method allows the model to capture and fuse information from heterogeneous feature spaces, making it more robust and accurate in predicting side effect frequencies for new drugs. By eliminating the dependency on prior frequency information, CrossFeat significantly enhances the model's ability to predict side effects for drugs with no historical data, thereby improving patient safety and clinical outcomes. This study aimed to predict the occurrence and frequency of side effects for new drugs, even in the absence of any prior knowledge, including side effect frequency information. This goal was achieved through the proposal of CrossFeat, a convolutional neural network (CNN)-transformer-connected model that incorporates cross-feature learning. Specifically, a transformer encoder was designed to provide cross-attention between drugs and their side effects. This design facilitated the exchange

of information between the drugs and their associated side effects. The model not only learns about each drug and its individual side effects but also captures their interdependencies concurrently. In the general transformer architecture, the attention mechanism primarily focuses on self-attention within a single sequence. The queries, keys, and values all originate from the same source, allowing the model to attend to different positions within the same input sequence to capture dependencies and relationships [19]. In contrast, the feature-wise cross-attention mechanism in the CrossFeat architecture is designed to handle and fuse features from two different sources: the drug encoder and the side effect encoder. This approach enables the model to filter and integrate information across heterogeneous feature spaces, resulting in more accurate and robust predictions.

Our approach employs a comprehensive set of inputs, including molecular structure and compound description of drugs, and word embedding and semantic similarity information of side effects. The five-fold cross-validation results demonstrated the performance of CrossFeat in predicting the occurrence and frequency of side effects exceeding that of state-of-the-art models on a benchmark dataset for new drugs. The prediction results were validated using the published literature and drug side effect databases SIDER [20] (http://sideeffects.embl.de/) and OFFSIDES [21] (https://nsides.io/). Additionally, an independent evaluation using the FAERS_SI dataset confirmed the robustness of our model. This novel approach combining dual similarity matrices and vectors with cross-feature learning presents a more effective predictive modeling paradigm in the pharmaceutical domain, particularly noteworthy for its ability to predict side effects of new drugs even in the absence of any prior drug-related information. Moreover, we aimed to demonstrate that our model, by effectively capturing the necessary chemical and side effect information, outperformed existing models in predicting the frequency of side effects for new drugs.

Methods

Benchmark dataset

We downloaded the drug side effect frequencies and unique names of drugs and side effects from Supplementary Data 1 in Galeano et al.'s study [15]. This dataset contains 37,441 frequency-class associations for 759 drugs and 994 side effects. The occurrence of side effects was quantified into side effect frequency classes coded with integers between 1 and 5 (very rare; frequency = 1, rare; frequency = 2, infrequent; frequency = 3, frequent; frequency = 4, very frequent; frequency = 5). A dataset of drug side effect frequencies was used as the target values in this study.

Construction of input features

CrossFeat utilizes two types of drug information and two types of side effect information to generate similarity-embedding matrices and embedding vectors for drugs and side effects. Drug information comprises mol2vec [22] and fingerprint vectors. Mol2vec is a 100- or 300-dimensional vector representing the molecular structure of a drug that is obtained by inputting the drug SMILES into Mol2vec. Drug SMILES sequences were collected from the STITCH [23] database. Meanwhile, the fingerprint is a 2048-dimensional vector obtained by inputting the drug SMILES into RDkit [24], providing

descriptors of the compound. These mol2vec and fingerprint vectors were subsequently employed to create mol2vec and fingerprint similarity vectors representing the similarity between drugs based on cosine similarity [25] and Jaccard similarity [26], respectively. For side effect information, semantic similarities and side effect word vectors were employed. We calculated semantic similarity using the Adverse Drug Reaction Classification System IDs to draw Directed Acyclic Graphs (DAGs). These DAGs represent the hierarchical relationships between side effects [16]. In addition, GloVe [27] was used to generate 300-dimensional side effect word vectors with side effect names, and the word vector similarities between the side effects were calculated using cosine similarity. All drug and side effect information were regenerated using the methods proposed by Zhao et al. [18].

In total, 36,850 frequencies for 736 drugs were obtained after removing 23 drugs with no matching information in the SDPred to ensure consistency in our dataset. This step was necessary to maintain the integrity of our comparisons and to avoid potential ambiguities in the data. Additionally, 994 side effects matched the benchmark dataset. Let n and m be the number of drugs and side effects, respectively. A dataset of drugs can be represented as $D = \{d_1, d_2, \cdots, d_n\}$, where $n = 736$, and a dataset of side effects can be represented as $S = \{s_1, s_2, \cdots, s_m\}$, where $m = 994$. All possible drug-side effect pairs can be $D \times S$ and the number of pairs is $n \times m = 731,584$. As shown in Table 1, we partitioned the samples of drug-side effect pairs into three distinct subsets: PS_1 containing 36,850 drug-side effect pairs with frequency information, PS_2 with 36,850 pairs randomly selected from 694,734 pairs with unknown frequencies, and PS_3 for the remaining 657,884 pairs with unknown frequencies. It was essential to include samples without drug side effect frequency information in the training set to calculate the probability of occurrence of drug side effects. Therefore, we randomly sampled a set of pairs equivalent to the size of PS_1 to form PS_2, assuming a frequency of zero for PS_2. All samples from PS_1 and PS_2 were used for model training and testing in a five-fold cross-validation to predict the probability of drug side effect occurrence and frequency of side effects. PS_3 was used for literature validation to assess the performance of the model.

A transformer-based cross-feature learning model for drug side effect frequency prediction (CrossFeat)

This study introduces the CrossFeat, designed to predict the occurrence probability and frequency of side effects for new drugs based on the molecular structure and compound description information of drugs, along with word embedding and semantic similarity information of side effects. The architecture of the proposed model is

Table 1 Three subsets of drug-side effect pair samples

Dataset	# Of samples
Pairs w/ frequency (PS1)	36,850
Pairs w/o frequency	694,734
Randomly selected pairs w/o frequency (PS2)	36,850
Remaining pairs w/o frequency (PS3)	657,884
Total	731,584

illustrated in Fig. 1. Two critical challenges were addressed during the development of CrossFeat. The first involves creating representations that effectively represent each drug and its side effects, thereby ensuring accurate predictions for new drugs. The second challenge was to integrate the drug and side effect input features into a unified dimension. Samples comprised of pairs of drugs and their side effects; therefore, it was crucial to train features concurrently to enhance their interdependence as the model learned.

The CrossFeat model was trained using the following workflow: (i) Drug and side effect similarities were dimensionally reduced to the same size of vectors. The reduced drug vectors were computed through outer product operations [28] to construct the drug embedding matrix, and the side effect vectors were similarly subjected to outer product operations to generate the side effect embedding matrix; (ii) these embedding matrices were subsequently input to the CNN for feature extraction; (iii) the transformer module was utilized to emphasize crucial information from the drug and side effect features themselves and simultaneously acquire information about each other by facilitating cross-feature learning; (iv) the multi-layer perceptron (MLP) projects the drug mol2vec vector and side effect word vector into the same size of embeddings; and finally, (v) the classifiers predict the occurrence probabilities

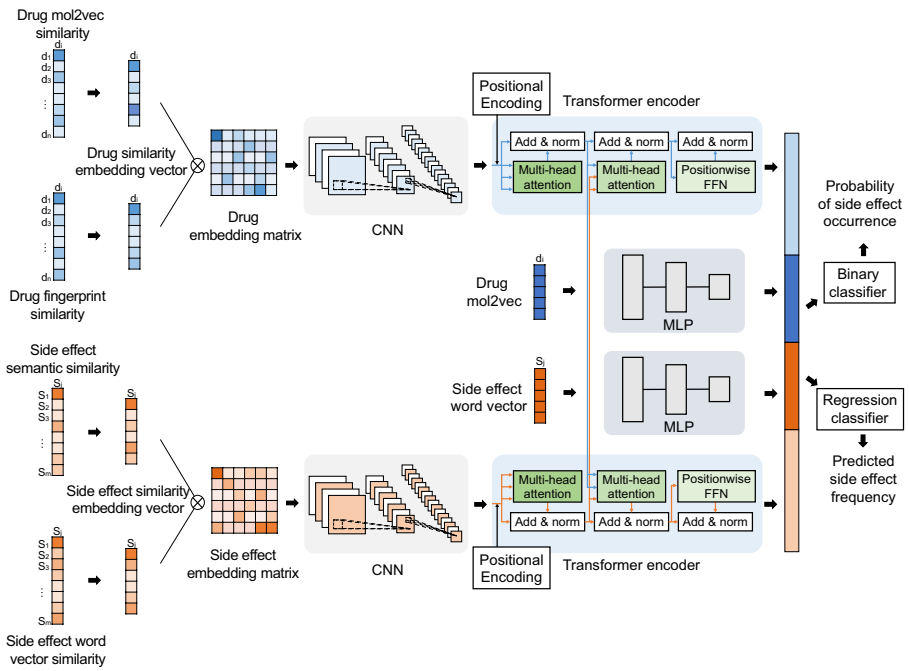

Fig. 1 Workflow of CrossFeat. Drug and side effect similarities are dimensionally reduced and multiplied through outer product operations to generate the drug and side effect embedding matrix (left side of the figure). Subsequently, the CNN architecture extracts features from the drug and side effect embedding matrix (center of the figure). The transformer module learns the representations from individual drug and side effect features and concurrently undergoes cross-learning to acquire the representations of each other (upper and lower right of the figure). Simultaneously, the Multi-Layer Perceptron (MLP) module projects the drug mol2vec vector and side effect word vector into a same-dimensional embedding (right middle of the figure). All output embeddings from MLPs and transformers are concatenated and inputted into two classifiers to predict the occurrence probabilities and frequencies of side effects for the drugs

and frequencies of side effects for the drugs by concatenating all the embeddings from the transformer and MLP. The detailed steps are elaborated in the subsequent subsections.

Generation of the input embedding matrix

The drug similarity vectors of drug d_i, $V^i_{mol} \in \mathbb{R}^n$ and $V^i_{fin} \in \mathbb{R}^n$, represent the cosine similarities of drug molecular substructures with mol2vec and Jaccard scores of chemical substructures by fingerprint, respectively. The side effect similarity vectors of side effect s_j, $V^j_{sem} \in \mathbb{R}^m$ and $V^j_{word} \in \mathbb{R}^m$, represent the side effect semantic similarity and word vector cosine similarity, respectively. The dimensionality of all feature vectors n and m was reduced to l (in this study, $l = 128$); that is, $\left\{ V'^i_{mol}, V'^i_{fin}, V'^j_{sem}, V'^j_{word} \right\} \in \mathbb{R}^l$. Each vector was subsequently multiplied by the others using the outer product operation, denoted by $\bigotimes$. We used the outer product operation expecting that its use between similarity matrices would result in a synergistic effect on the similarity values. Thus, the drug embedding matrix $M_{d_i} \in \mathbb{R}^{l \times l} = V'^i_{mol} \bigotimes V'^i_{fin} = V'^i_{mol}(V'^i_{fin})^\intercal$ and side effect embedding matrix $M_{s_j} \in \mathbb{R}^{l \times l} = V'^j_{sem} \bigotimes V'^j_{word} = V'^j_{sem}(V'^j_{word})^\intercal$ with a size of $l \times l$ were generated and subsequently used as the input to the CNNs. The similarity information for test drugs in V^i_{mol}, V^i_{fin}, V^j_{sem}, and V^j_{word} was uniformly filled to zero to consider the drugs in the test set of each fold in the five-fold cross-validation experiment as new drugs without prior information.

Feature extraction with CNN

A CNN is a type of artificial neural network commonly applied to visual image analysis [29, 30]. It consists of multiple layers, each capable of detecting different features in an image. Our study used two separate CNNs to extract features from the drug and side effect embedding matrices. The structures of both CNNs were identical and comprised four convolutional layers, each consisting of a 2D convolution, batch normalization [31], and a ReLU [32] activation function (see Fig. 2A). Each layer had a channel size of 32, a stride of 2, and a kernel size of 2. In the CNN module, the input is a tensor of the following shape: batch $\times 1 \times l \times l$. After passing through four convolutional layers, the input was abstracted into a feature map with a size of batch x $32 \times 8 \times 8$. Subsequently, mean pooling was applied to each feature map.

Cross-feature learning with transformer encoder through cross-attention

The original transformer [19] is a neural machine translation model consisting of encoder and decoder architectures. The encoder extracts features from an input sentence and the decoder utilizes these features to produce an output sentence for translation. The transformer module in CrossFeat is a variation of the original transformer encoder. The encoders for cross-feature learning are composed of a stack of two identical blocks (or layers), each containing three sub-layers (whereas the original encoder has two sub-layers): two multi-head attention mechanisms and a position-wise fully connected feedforward network. The output embeddings of a sublayer are carried forward to the subsequent layers through residual connections, and layer normalization is

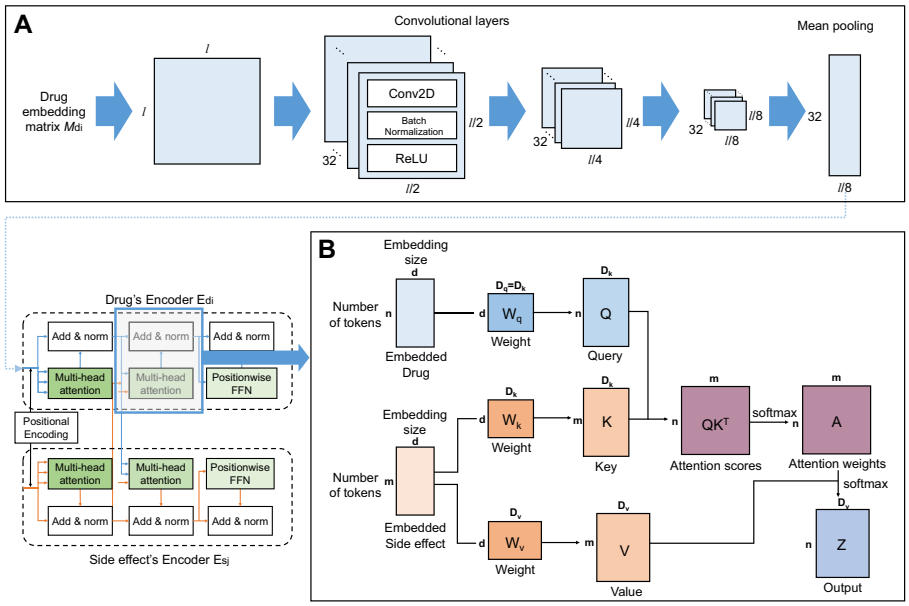

Fig. 2 Schematic of the CNN and the cross-feature learning (feature-wise cross-attention) mechanism in the CrossFeat architecture. **A** An $l \times l$ dimension embedding matrix is passed through four convolutional layers, each consisting of a Conv2D, batch normalization, and ReLU activation function, followed by mean pooling to extract feature matrices. These feature matrices are then input into the transformer encoder. **B** Queries (Q) from the drug encoder and keys (K) from the side effect encoder are used to form the attention scores. Specifically, the queries are derived from the previous sublayer of the drug encoder, while the keys and values (V) are obtained from the first sublayer of the side effect encoder. Attention scores are calculated as the dot product of the queries and keys, which are then passed through a softmax function to generate the attention weights. These weights are subsequently multiplied by the values to produce the output. This cross-attention process enables the effective fusion of features between the drug and side effects. It enhances the ability of the model to capture the complex relationships between drugs and their side effects

applied after each residual connection. The input of the attention function consists of queries and keys with dimensions D_k, and values with dimensions D_v, where the queries, keys, and values are packed together into matrices Q, K, and V, and the output matrix is calculated using the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\mathsf{T}}}{\sqrt{D_k}}V\right). \tag{1}$$

Refer to Fig. 2b for a detailed illustration of the attention mechanism.

In addition to the two sublayers in the original transformer encoder, CrossFeat's encoder includes an additional multi-head attention layer inserted as a second sublayer, which performs cross-feature learning. Cross-feature learning (feature-wise cross-attention) is a module for semantic segmentation used in the CrossFeat architecture. It is employed to fuse features between the drug and side effect encoders. This module guides the filtration of transformer features and eliminates ambiguities in interactions between drugs and side effects. Let us denote the drug's encoder as E_{d_i} and the side effect's encoder as E_{s_j}. The second sublayer of the E_{d_i} performs cross-attention over the output of the first sublayers of E_{d_i} and E_{s_j}. Specifically, the queries are derived from the previous sublayer of E_{d_i}, and the keys and values are obtained from the first sublayer of E_{s_j} as shown in Fig. 2B. Similarly, the second sublayer of E_{s_j} performs cross-attention over the

output of the first sublayers of E_{s_j} and E_{d_i}. Here, the queries come from the previous sublayer of E_{s_j} and the keys and values come from the first sublayer of E_{d_i}. All sublayers of E_{d_i} and E_{s_j} produce outputs with dimensions $D_{E_{d_i}}$ and $D_{E_{s_j}} = p$.

CrossFeat's multi-layer perceptron (MLP)

In the previous steps, we trained CNNs and transformers to generate embedding vectors that described each drug, denoted by d_i, and each side effect, denoted by s_j based on their similarities to other drugs and side effects. In this step, our objective was to learn latent representations for each drug and side effect by directly capturing vectors representing d_i and s_j without relying on similarity information. Mol2vec vectors represent drugs and are projected onto q-dimensional representations for each drug using a two-layer MLP and batch normalization. Similarly, word vectors represent each side effect and are projected onto q-dimensional space using a two-layer MLP and batch normalization to create the corresponding latent representations.

Classifiers

The classifiers consist of a binary classifier to determine whether the side effect s_j occurs owing to the drug d_i and a regression classifier to predict the frequency of s_j occurring. The outputs from steps 3 and 4 were concatenated to create a classifier input vector with $2p + 2q$ dimensions. The binary classifier employs a sigmoid function. The output of the binary classifier was set to one if s_j occurs and zero otherwise. The binary occurrence was determined based on the following thresholds when a predicted score x was obtained:

$$\text{Predictedbinaryoccurrence(x)} = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x = 0 \end{cases}. \tag{2}$$

The output of the regression classifier is a continuous value between zero and five or higher if the output of the binary classifier is one.

Experimental design

We employed a five-fold cross-validation procedure on 73,700 samples (drug-side effect pairs) comprising the PS_1 and PS_2 datasets. We divided the folds by drug rather than by sample to ensure that the drugs in the held-out test fold were not detected in the held-out train folds. Consequently, the folds did not contain the same number of samples. The average numbers of samples in the training and test folds were 58,960, and 14,740, representing approximately 80% and 20% of the total, respectively. The samples in the training fold were further split at a 4:1 ratio based on the samples. This division allocated 80% of the training fold samples (referred to as the training set) for model training and the remaining 20%, referred to as the validation set to set the model hyperparameters and choose the best model per fold.

A grid search was performed to determine optimal hyperparameters for each fold of CrossFeat. The hyperparameter search space for CrossFeat is provided in Supplementary Table S1. We randomly selected 10 hyperparameter combinations and compared their performances on the validation set. During training, early stop endurance was counted if the performance on the validation set deteriorated compared to the previous state. The

training process was concluded when early stop endurance reached 10. Subsequently, the performance of the test fold was evaluated using the best-performing hyperparameter combination determined in the validation set.

We adopted the binary cross-entropy (BCE) loss function for the binary classification of side effects and we applied the L_2 in Eq. 4 to our loss function for side effect frequency prediction. CrossFeat utilizes two Adam optimizers [33] to learn the predicted side effect occurrence probabilities $\hat{y_{i1}}$ and predicted side effect frequency value $\hat{y_{i2}}$ by minimizing the following two loss functions:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} y_{i1} \cdot log(\hat{y_{i1}}) + (1 - y_{i1}) \cdot log(1 - \hat{y_{i1}}) \tag{3}$$

$$L_2 = \sum_{i=1}^{N'} (y_{i2} - \hat{y_{i2}})^2, \tag{4}$$

where N and N' represent the number of training samples in the PS_1 and PS_2 datasets and training samples in the PS_1 dataset, y_{ik} and $\hat{y_{ik}}$ represent the true and predicted values of sample i, respectively. Four metrics were employed to evaluate the performance of the model: area under the receiver operating characteristic curve (AUROC), AUPRC for binary classification, root mean squared error (RMSE) and mean absolute error (MAE) for regression classification.

Independent FAERS_SI dataset

We conducted additional experiments using the FAERS (FDA Adverse Event Reporting System) dataset, which includes reports of actual adverse events and medication errors submitted by patients to the Food and Drug Administration (FDA). The FAERS database relies on voluntary adverse event reports submitted by healthcare professionals, consumers, and manufacturers, including negative placebo effects. In contrast, the SIDER [20] database collects its information on drug side effects from the FAERS dataset; however, it utilizes natural language processing to extract drug-side effect pairs from the drug package insert. For this case study, we collected FAERS reports from the fourth quarter of 2012 to the second quarter of 2023. The original data can be downloaded from https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html. We included reports from healthcare professionals only, including physicians, pharmacists, nurses, dentists, and others. To enhance dataset reliability, we filtered the FAERS dataset to include only the drugs, side effects, and drug-side effect pairs that had frequency information available in the SIDER database. Additionally, we excluded cases involving the simultaneous use of several drugs to ensure clarity in determining the cause of the side effects attributable to a specific drug.

The FAERS dataset we downloaded initially included 10,511,188 samples (drug-side effect pairs), 34,486 drugs, 17,550 side effects, and 1,341,486 distinct drug-side effect pairs. After filtering by 2,962 SIDER side effects, we reduced the sample size to 2,044,255. Further filtering by SIDER's 932 drugs resulted in 808,783 samples. Finally, filtering by 59,333 SIDER drug-side effect pairs resulted in a dataset with 231,464 samples, encompassing 633 drugs, 1,395 side effects, and 19,319 distinct pairs. This dataset

will be referred to as FAERS_SI. Galeano's study [15] used the SIDER 4.1 database, which was released in October 2015, to create frequency classes, covering an earlier period. FAERS_SI includes data from a later period, with additional drugs and side effects absent in the Galeano dataset, making it largely independent, though not completely. Specifically, 61.3% of side effects (855/1,395), 77.4% of drugs (490/633), and 3.6% of distinct pairs (690/19319) in FAERS_SI overlap with those in the Galeano dataset. For a detailed illustration of the process of generating FAERS_SI, see Fig. 3. The frequency of drug side effects was calculated by dividing the number of samples in which a specific side effect occurred with the use of a particular drug by the total number of samples using that specific drug. Finally, we quantified the calculated frequency of drug side effects on a scale of 1 to 5. The frequency value was determined based on the following criteria:

$$\text{Frequency(x)} = \begin{cases} 1\ (\text{Veryrare}) & \text{if } x < 0.0001 \\ 2\ (\text{Rare}) & \text{if } 0.0001 \leq x < 0.001 \\ 3\ (\text{Infrequent}) & \text{if } 0.001 \leq x < 0.01 \\ 4\ (\text{Frequent}) & \text{if } 0.01 \leq x < 0.1 \\ 5\ (\text{Verycommon}) & \text{if } x \geq 0.1. \end{cases} \tag{5}$$

A frequency value of 0 was assigned to cases where there was no information or where no side effects were reported.

Results

Comparison of model performance

Most previously published methods utilize side effect frequency information [14–17, 34] or other drug-related prior knowledge, such as protein targets [18, 35], when creating input features (refer to Supplementary Table S2). However, our method aims to predict side effect frequencies for drugs without any prior information, including side effect

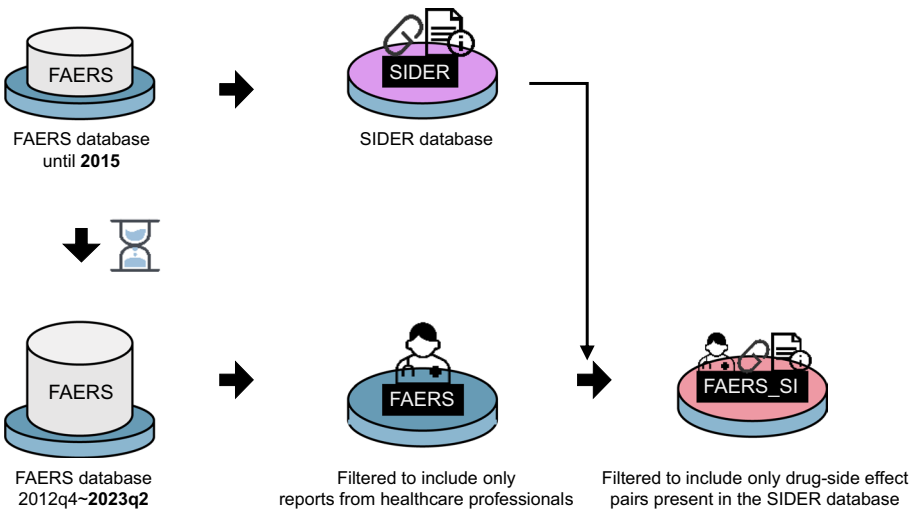

Fig. 3 FAERS_SI dataset creation process. The original FAERS dataset contains data from Q4 2012 to Q2 2023. First, the dataset was filtered to include only reports from healthcare professionals to enhance data reliability. Subsequently, it was further filtered to include only the drugs, side effects, and drug-side effect pairs present in the SIDER database. The SIDER database collects its information on drug side effects from FAERS up to 2015 using natural language processing to extract data from drug package inserts. This resulted in the final FAERS_SI dataset

frequency or other drug-related data. Therefore, the performance comparison focused on models capable of predicting side effect frequencies for new drugs with no prior information. In this comparison, SDPred [18], which utilizes only drug SMILES information and side effect word and semantic information, excluding other features requiring prior drug-related information, was considered as a base method. Additionally, ridge regression [36], XGBoost [37], and CrossFeat's MLP module were compared with CrossFeat. For detailed comparisons between methods, please refer to the Discussion section. Furthermore, SDPred_loss was included as an additional comparison method. SDPred employs the total loss as the sum or product of two L2 losses, whereas SDPred_loss utilizes two optimizations and two losses, BCE loss and L2 loss, similar to CrossFeat. The CrossFeat's MLP specifically denotes the MLP component within the CrossFeat model. All comparing methods were trained using same input features and sample sizes per fold, except for $\text{CrossFeat}_{\text{MLP}}$. $\text{CrossFeat}_{\text{MLP}}$ utilizes only drug mole2vec and side effect word vectors as input, as shown in Fig. 1 and section 'CrossFeat's Multi-Layer Perceptron (MLP)'. The comparison considered the exclusion of features related to unknown information on new drugs.

We performed a five-fold cross-validation on 73,700 samples from the PS_1 and PS_2 datasets. The hyperparameter search space and selected best hyperparameters for each method are provided in Supplementary Table S1 and Supplementary Tables S3-S10, respectively. Table 2 lists the predictive performance of each method. Lower RMSE and MAE values indicate a better prediction of side effect frequency, whereas higher AUROC and AUPRC values indicate a better prediction of side effect occurrence. CrossFeat outperformed all other methods in predicting the occurrence and frequency of the side effects of new drugs. It achieved a 0.06 improvement in AUROC and a 0.04 improvement in AUPRC compared to SDPred in drug side effect occurrence prediction. The drug side effect frequency prediction demonstrates a 0.08 decrease in RMSE and a 0.16 decrease in MAE compared to SDPred.

We also conducted an additional evaluation of our model using five-fold cross-validation with scaffold division. Scaffold division is a method to ensure that structurally similar compounds are grouped together during the training and testing phases, providing a more stringent test of the model's ability to generalize to new chemical structures. For this evaluation, we identified 513 distinct scaffolds in our dataset and used

Table 2 Performance comparison of experimented methods with five-fold cross-validation

Method	Binary classification		Regression	
	AUROC	AUPRC	RMSE	MAE
SDPred	0.76 ± 0.03	0.78 ± 0.02	0.94 ± 0.07	0.80 ± 0.06
SDPred_loss	0.80 ± 0.03	0.81 ± 0.02	0.91 ± 0.03	0.75 ± 0.02
Ridge regression	0.80 ± 0.01	0.81 ± 0.02	1.80 ± 0.08	1.52 ± 0.08
XGBoost	0.71 ± 0.02	0.79 ± 0.02	1.78 ± 0.09	1.52 ± 0.09
MLP	0.81 ± 0.01	0.81 ± 0.02	0.94 ± 0.03	0.77 ± 0.04
$\text{CrossFeat}_{\text{MLP}}$	0.80 ± 0.01	0.81 ± 0.01	1.02 ± 0.03	0.74 ± 0.09
CrossFeat	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.64 \pm 0.03}$
$\text{CrossFeat}_{\text{Scaffold}}$	0.81 ± 0.02	0.81 ± 0.02	0.85 ± 0.03	0.63 ± 0.02

Bold indicates the best result among all models listed in each metric

them as the basis for the scaffold splitting. The results of this evaluation are summarized in Table 2 as $\text{CrossFeat}_{\text{Scaffold}}$. Our findings indicate that CrossFeat maintains robust performance even under scaffold division, further validating its effectiveness in predicting drug side effect frequencies without prior frequency information.

Predictive performance across drug side effect frequencies

Frequency values for drug side effects were assigned on a scale of 1 to 5. Upon examining the distribution of these frequencies, we noted variations in sample sizes of the benchmark dataset across different frequency values. The sample sizes were 3.2% for frequency=1 (1,190/36,850), 11.3% for frequency=2 (4,174/36,850), 27% for

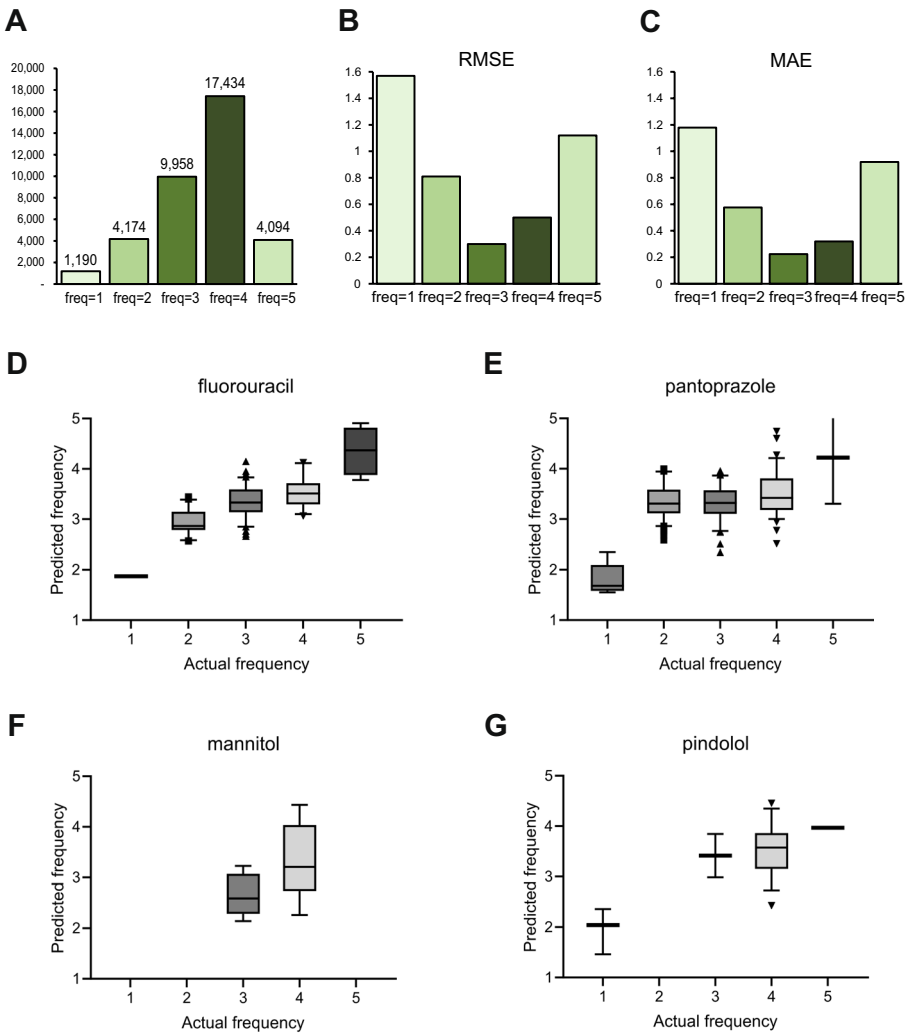

Fig. 4 Freq; frequency. **A** Distribution of drug-side effect pair samples for five frequencies. Freq; frequency. **B** predictive performance of CrossFeat across different drug side effect frequency values is evaluated based on root mean squared error values. **C** predictive performance of CrossFeat across different drug side effect frequency values is evaluated based on mean absolute error values. **D**–**G** Box plots depicting the predicted frequencies of four randomly selected drugs across actual frequency categories. Scatter plots indicate the upper and lower 5% values. **D** Fluorouracil. **E** Pantoprazole. **F** Mannitol. **G** Pindolol

frequency=3 (9,958/36,850), 47.3% for frequency=4 (17,434/36,850), and 11.1% for frequency=5 (4,094/36,850) (Fig. 4A).

We observed strong correlations (Pearson's correlation coefficient [38] > 0.8) between prediction performance and frequency sample size using CrossFeat (see Fig. 4B and C). CrossFeat exhibited the highest prediction performance with RMSE values of 0.33 and 0.52, and MAE values of 0.22 and 0.32, respectively when focusing on frequency=3 and frequency=4, which are characterized by relatively larger sample sizes. In contrast, the smallest sample size observed at frequency=1 resulted in the lowest prediction performance, with an RMSE of 1.57 and MAE of 1.18. This highlights the sensitivity of the model to variations in sample size, particularly when dealing with infrequent side effects. In addition, the predicted frequency values vary across different actual frequency values for four randomly selected drugs (see Fig. 4D–G).

Ablation study

In this section, we performed an ablation study to understand the contributions of both structural components and individual features to the overall performance of our model. The ablation study examined the impact of removing structural components of the model and the effects of eliminating individual features.

First, we conducted experiments to evaluate the performance impact of removing different structural components of the CrossFeat model. By systematically removing components such as the CNN layers, transformer layers, and the MLP, we assessed the resulting changes in performance metrics. The results of these experiments are summarized in Table 3.

The $\text{CrossFeat}_{\text{Transformer}}$, $\text{CrossFeat}_{\text{CNN}}$, and $\text{CrossFeat}_{\text{MLP}}$ models represent variations of the original CrossFeat model without the transformer encoder, CNN, and MLP architectures, respectively. Across all four metrics, $\text{CrossFeat}_{\text{Transformer}}$ exhibited the lowest performance with an RMSE of 0.89, an MAE of 0.75, AUROC of 0.79, and AUPRC of 0.79. The transformer module is expected to significantly contribute to CrossFeat, enabling the representation of drugs to learn the representation of side effects and vice versa, before concatenating the latent representations of drugs and side effects. In $\text{CrossFeat}_{\text{CNN}}$, the CNN module showed the least influence on side effect occurrence prediction with AUROC of 0.80 and AUPRC of 0.81; however, it had the second-highest influence on frequency prediction with RMSE of 0.89 and an MAE of 0.67. In $\text{CrossFeat}_{\text{MLP}}$, the MLP module exhibited a more significant impact on the prediction of side effects, achieving the second-worst binary classification performance with an AUROC of 0.79 and AUPRC of 0.80. The absence of the MLP module resulted in the

Table 3 Results of the structural components ablation study

Method	Binary Classification		Regression	
	AUROC	AUPRC	RMSE	MAE
$\text{CrossFeat}_{\text{Transformer}}$	0.79 ± 0.01	0.79 ± 0.01	0.89 ± 0.06	0.75 ± 0.01
$\text{CrossFeat}_{\text{CNN}}$	0.80 ± 0.02	0.81 ± 0.02	0.89 ± 0.05	0.67 ± 0.03
$\text{CrossFeat}_{\text{MLP}}$	0.79 ± 0.02	0.80 ± 0.01	0.87 ± 0.03	0.65 ± 0.03
CrossFeat	0.82 ± 0.01	0.82 ± 0.01	0.86 ± 0.04	0.64 ± 0.03

unavailability of the mol2vec vector and the side effect word vector, both of which were used as inputs for the MLP. This likely led to a reduction in performance owing to information loss. The optimal hyperparameters per fold, as determined by a grid search of the model are listed in Supplementary Tables S11-S13.

Second, we performed feature ablation experiments where each type of input feature was individually removed in the CrossFeat model. Specifically, if a drug feature was removed, the other drug feature was used twice to maintain the model's structure. The same approach was applied to side effect features. The features examined included drug fingerprints, drug mol2vec embeddings, side effect semantic similarities, and side effect word vectors. The results of these experiments are presented in Table 4, where each value represents the mean performance *metric* $\pm$ *the* standard deviation.

When individual features were removed, the results showed a general decrease in AUROC across all feature ablations. For instance, the removal of drug fingerprints resulted in an AUROC of 0.802 and removing drug mol2vec embeddings resulted in an AUROC of 0.802. Similarly, removing side effect semantic similarity and side effect word vectors resulted in AUROC values of 0.797 and 0.804, respectively. This trend indicates that each feature plays a significant role in the side effect occurrence prediction performance of the model. The AUPRC also decreased consistently when each feature was removed, further highlighting the importance of each feature in maintaining high precision-recall performance. The RMSE and MAE generally increased, indicating a drop in prediction accuracy for side effect frequencies when features were removed. For example, RMSE increased to 0.88 when drug fingerprints were removed and to 0.88 when side effect semantic similarity was removed. The MAE increased to 0.65 when drug fingerprints were removed and to 0.67 when side effect semantic similarity was removed. However, removing the side effect word vectors slightly improved the RMSE to 0.85, although the corresponding MAE increased to 0.65. Removing drug mol2vec embeddings resulted in an RMSE of 0.86 and MAE of 0.64, indicating an overall negative impact. Overall, the ablation study underscores the importance of each feature in contributing to the model's predictive performance and robustness. The optimal hyperparameters per fold are listed in Supplementary Tables S14-S17.

Variation of transformer encoder

CrossFeat's transformer module, comprising the encoders E_{d_i} and E_{s_j} has an additional second sublayer added to the original transformer encoder. In the case of E_{d_i}, this second sublayer takes queries from its first sublayer, and keys and values from the first sublayer of E_{s_j}. This pattern is similarly applied to E_{s_j}, where the second

Table 4 Results of the feature ablation study in CrossFeat

Method	Binary classification		Regression	
	AUROC	AUPRC	RMSE	MAE
$\text{Drug}_{\text{Fingerprint}}$	0.802 ± 0.00	0.806 ± 0.01	0.88 ± 0.08	0.65 ± 0.04
$\text{Drug}_{\text{Mol2vec}}$	0.802 ± 0.00	0.812 ± 0.01	0.86 ± 0.02	0.64 ± 0.02
$\text{Side-Effect}_{\text{Semantic}}$	0.797 ± 0.01	0.805 ± 0.02	0.88 ± 0.06	0.67 ± 0.02
$\text{Side-Effect}_{\text{Word}}$	0.804 ± 0.01	0.811 ± 0.01	0.85 ± 0.05	0.65 ± 0.03

sublayer receives queries from its first sublayer and obtains keys and values from the first sublayer of E_{d_i}. While the previous section 'Predictive performance across drug side effect frequencies' demonstrated the contribution of cross-feature learning to drugs and side effects, this section describes how the model's performance varies when different features are crossed. Figure 4 illustrates the detailed architecture of encoders with diverse cross-feature learning. The cross-feature learning 1_1 (CL1_1) encoder has a second multi-head attention layer, taking queries and keys from its first sub-layer and the values from the input to the first sub-layer of another transformer encoder (Fig. 5A), whereas the CL1_2 encoder has a second sub-layer that takes queries from its first sub-layer and keys and values from the input for the other transformer encoder's first sublayer (Fig. 5B). The CL2_1 encoder involves a second sublayer that takes queries and keys from its first sublayer and values from the output of the first sublayer of another transformer encoder (Fig. 5C). The configuration with no cross-feature learning is illustrated in Fig. 5D.

The performances of each encoder are listed in Table 5. The encoder with cross-feature learning exhibited superior performance when extracting information from the output of the first sublayer of another encoder compared to the information obtained from the input of the first sublayer. In addition, it performed better when taking keys and values from another encoder together, compared with taking values alone. The hyperparameters used in the model are listed in Supplementary Tables S18-S21.

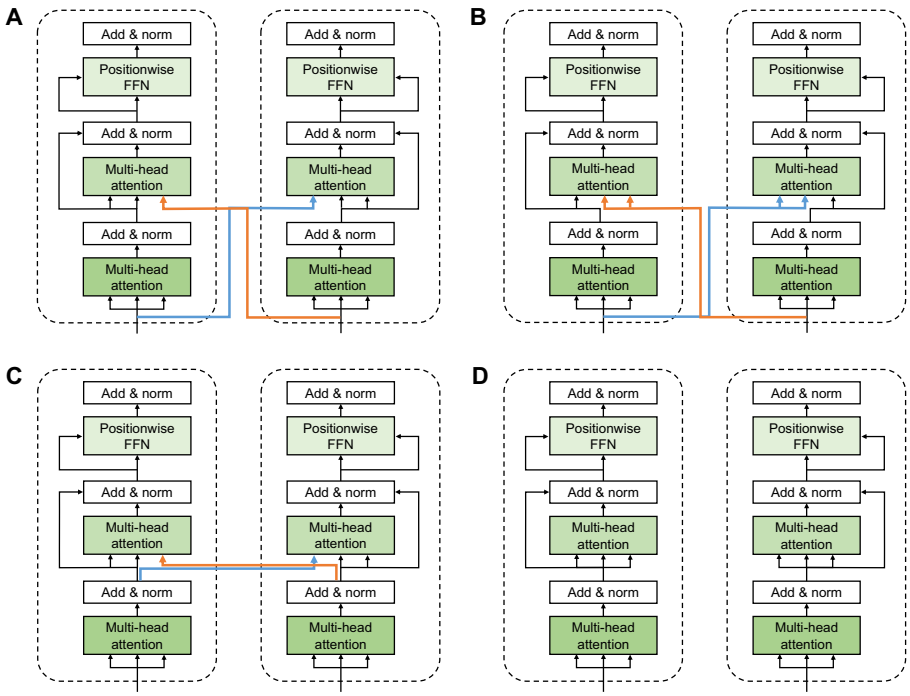

Fig. 5 Encoders with a diverse cross-feature learning **A** CL1_1 encoder has a second multi-head attention layer that takes queries and keys from its first sub-layer and values from the input to the first sub-layer of another transformer encoder. **B** CL1_2 encoder includes a second sub-layer, taking queries from its first sub-layer and keys and values from the input for the other transformer encoder's first sublayer. **C** CL2_1 encoder involves a second sub-layer, taking queries and keys from its first sub-layer and values from the output of the first sub-layer of another transformer encoder. **D** No cross-learning encoder

Table 5 Performance of cross-feature learning

Method	Binary classification		Regression	
	AUROC	AUPRC	RMSE	MAE
CL1_1	0.80 ± 0.01	0.80 ± 0.01	0.95 ± 0.25	0.64 ± 0.03
CL1_2	0.80 ± 0.01	0.81 ± 0.02	0.91 ± 0.14	0.65 ± 0.03
CL2_1	0.80 ± 0.00	0.81 ± 0.02	0.84 ± 0.01	0.64 ± 0.01
No cross	0.80 ± 0.01	0.80 ± 0.02	0.94 ± 0.10	0.67 ± 0.02

Additionally, to assess the impact of varying the number of heads in the transformer's multi-head attention mechanism, we evaluated the performance of CrossFeat with different configurations. The results, shown in Supplementary Table S22, indicate the performance metrics for binary classification (AUROC, AUPRC) and regression (RMSE, MAE) tasks. The results revealed that increasing the number of heads generally improves the RMSE and MAE, suggesting better performance in side effect frequency prediction. However, for side effect occurrence prediction metrics (AUROC and AUPRC), the performance did not show a consistent improvement with an increasing number of heads. Thus, the optimal number of transformer heads appears to be a balance between these metrics. By incorporating multiple heads, the model can capture a richer set of features and relationships, although this does not linearly translate to better performance in all metrics, particularly for the side effect occurrence prediction metrics.

Case studies

To assess the efficacy of CrossFeat in predicting the side effects of new drugs without frequency information, we conducted experiments using the PS_3 dataset. The model was trained on datasets PS_1 and PS_2 to predict the occurrence and frequency of the side effects in PS_3. Subsequently, three drugs were randomly selected: amiloride, nebivolol, and benazepril. We selected the top 10 side effects based on their predicted probabilities to verify the actual occurrence of side effects with high probability for these drugs. We subsequently investigated the evidence of side effects associated with these drugs, drawing information from SIDER [20] (a database containing information on marketed drugs and their reported side effects), OFFSIDES [21] (a database containing the side effects of drugs not listed on the official FDA label but discovered subsequently), and published literature. Table 6 presents the top 10 most probable side effects of the three arbitrarily chosen drugs, along with their predicted frequency values and supporting evidence. Most of the side effects exhibited evidence of their occurrence in the drug, indicating CrossFeat's robust predictive ability for new drugs.

Evaluation with independent dataset

To assess the generalizability of CrossFeat across diverse datasets and evaluate its ability to predict drug side effect frequencies, we utilized the FAERS_SI dataset. The FAERS_SI dataset comprises 19,319 drug-side effect pairs, encompassing 633 drugs and 1395 side effects, as shown in Fig. 6A–C. The distribution of drug-side effect pairs based on frequency values is presented in Fig. 6D. For a side effect to be categorized as "very rare" (frequency=1), it should occur at a frequency less than 0.0001 (Eq. 5). However, even the

Table 6 Top 10 side effects for drugs

Drug	Side effect	Predicted frequency	Occurrence probability	Evidence
Amiloride	Eye irritation	3.6819	0.9873	OFFSIDES
	Mouth ulceration	4.0437	0.9799	OFFSIDES
	Intracranial pressure increased	3.6772	0.9685	N/A
	Torticollis	4.2679	0.9564	[39]
	Hostility	4.3608	0.9537	OFFSIDES
	Asthma	2.7385	0.9509	OFFSIDES
	Hypomagnesaemia	3.7344	0.9327	OFFSIDES
	Cerebral ischaemia	3.7624	0.9279	OFFSIDES
	Renal tubular necrosis	2.6327	0.9236	OFFSIDES
	peripheral ischaemia	3.7264	0.9227	OFFSIDES
Nebivolol	Polyneuropathy	4.2915	0.9976	OFFSIDES
	Hyperaemia	3.5683	0.9972	[40, 41]
	Vaginal haemorrhage	4.0797	0.9961	OFFSIDES
	Pruritus generalised	4.1951	0.9798	OFFSIDES
	Eczema	4.3608	0.9569	OFFSIDES
	Photopsia	3.0045	0.9505	OFFSIDES
	Depressed level of consciousness	2.9673	0.9497	OFFSIDES
	Hypersensitivity	4.2290	0.9479	SIDER/OFFSIDES
	Pain of skin	2.9684	0.9472	OFFSIDES
	basal cell carcinoma	2.6563	0.9411	OFFSIDES
Benazepril	Application site burn	3.2536	0.9967	OFFSIDES
	Impaired healing	3.9402	0.9852	OFFSIDES
	Bradycardia	3.2199	0.9828	OFFSIDES
	Neutropenia	4.8328	0.9817	SIDER/OFFSIDES
	Fatigue	4.0944	0.9767	SIDER/OFFSIDES
	Eythema multiforme	2.6895	0.9596	OFFSIDES
	Aute respiratory distress syndrome	3.9755	0.9494	OFFSIDES
	Hirsutism	3.6562	0.9328	N/A
	Faecalith	3.1588	0.9299	OFFSIDES
	Pyelonephritis	3.5185	0.9236	OFFSIDES

drug-side effect pair with the largest number of samples, 4735, did not meet the required 10,000 samples for a frequency value of 0.0001. Consequently, none of the drug-side effect pairs had a frequency value of 1.

The results of the drug-side effect frequency prediction on the FAERS_SI dataset are presented in Table 7. The selected best hyperparameters for each method are provided in Supplementary Tables S23-S29. Across all methods, CrossFeat demonstrated superior performance compared to the other machine learning and deep learning models, as evidenced by lower RMSE and MAE values, as well as higher AUROC and AUPRC scores. CrossFeat achieved an AUROC of 0.86, an AUPRC of 0.87, an RMSE of 0.72, and an MAE of 0.57. In comparison, the MLP model achieved the highest AUROC of 0.87; however, it also had higher RMSE and MAE values (0.85 and 0.71, respectively) than CrossFeat. This indicates that while MLP had a slightly better AUROC, CrossFeat outperformed it in terms of AUPRC, RMSE, and MAE, highlighting its superior overall prediction accuracy and robustness. Additionally, the second-best model in terms of RMSE and MAE

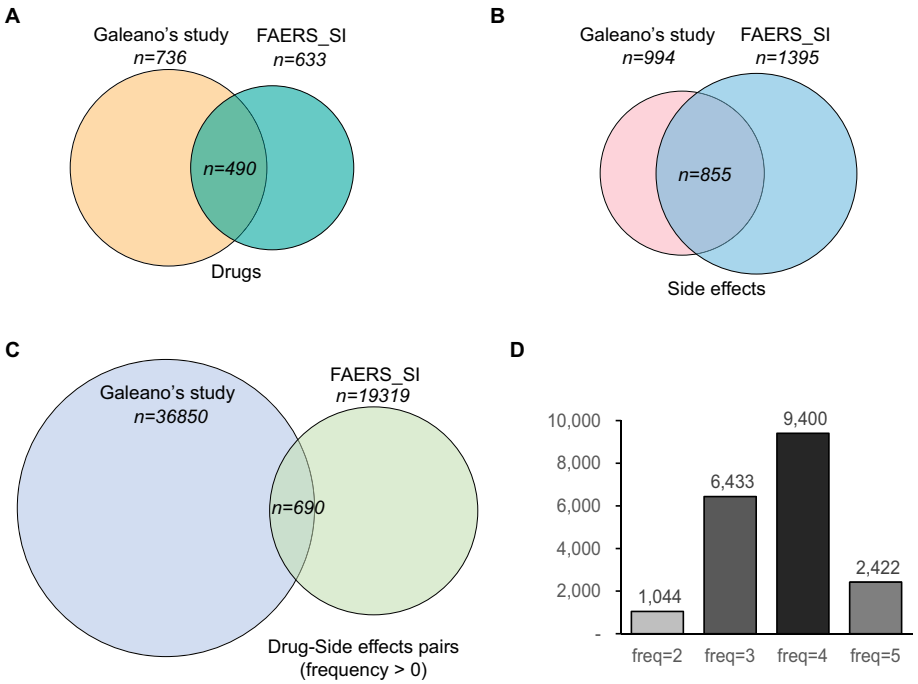

Fig. 6 Comparison between the Galeano and FAERS_SI datasets. **A** Venn diagram showing the overlap of drugs between the Galeano and FAERS_SI datasets. **B** Venn diagram depicting the overlap of side effects between the Galeano and FAERS_SI datasets. **C** Venn diagram illustrating the overlap of drug-side effect pairs between the Galeano and FAERS_SI datasets. **D** Distribution of drug-side effect pairs in the FAERS_SI dataset categorized by frequency values

Table 7 Model performance in FAERS_SI dataset

Method	Binary classification		Regression	
	AUROC	AUPRC	RMSE	MAE
SDPred	0.83 ± 0.03	0.84 ± 0.02	0.81 ± 0.03	0.68 ± 0.02
SDPred_loss	0.85 ± 0.02	0.86 ± 0.01	0.77 ± 0.03	0.63 ± 0.04
Ridge regression	0.86 ± 0.01	0.85 ± 0.02	1.79 ± 0.04	1.49 ± 0.04
XGBoost	0.77 ± 0.01	0.84 ± 0.02	1.76 ± 0.04	1.42 ± 0.04
MLP	**0.87 ± 0.01**	**0.87 ± 0.01**	0.85 ± 0.05	0.71 ± 0.04
$\text{CrossFeat}_{\text{MLP}}$	0.85 ± 0.01	0.86 ± 0.01	0.93 ± 0.06	0.86 ± 0.03
CrossFeat	0.86 ± 0.01	**0.87 ± 0.01**	**0.72 ± 0.04**	**0.57 ± 0.02**

Bold indicates the best result among all models listed in each metric

was SDPred_loss, which achieved values of 0.74 and 0.61, respectively. CrossFeat outperformed SDPred_loss by 0.05 in RMSE and 0.06 in MAE. The evaluation of our model on the FAERS_SI dataset demonstrated its superior performance compared to the base models.

Discussion

This study introduced CrossFeat, a novel model for predicting the occurrence and frequency of drug side effects based on cross-feature learning. The integration of a CNN and transformer architecture coupled with a cross-feature learning mechanism enables

CrossFeat to effectively handle the challenging task of predicting drug side effects. The model demonstrates proficiency in predicting side effect occurrence and excels in estimating the frequency of these occurrences, particularly for new drugs lacking prior frequency information. The machine-learning paradigm adopted in CrossFeat presents a valuable alternative to traditional clinical trials. By leveraging the power of predictive modeling, CrossFeat offers a potential supplement to conventional approaches, providing insights into the occurrence and frequency of side effects without the need for extensive and time-consuming trials. This demonstrates the potential of the model for streamlining drug development processes.

A critical aspect of CrossFeat's design lies in its emphasis on the efficient feature representation of drugs and their side effects. The utilization of suitable drugs and side effects is critical to enhance the model efficacy and prediction precision [42, 43]. The model captures a more holistic understanding of their relationships by incorporating representations of individual drugs and side effects along with their interactions. The utilization of similarity matrices, mol2vec, and word vectors coupled with the CNN-transformer-MLP architecture contributes to a comprehensive feature set that enhances the interpretability and predictive capabilities of CrossFeat. In future work, we aim to explore methods to integrate the heterogeneous information of drugs and side effects into a unified space. Specifically, we are interested in leveraging network structures and attention mechanisms to learn from these relationships. Incorporating methods to effectively combine different types of data into a cohesive framework will provide a more nuanced understanding of drug side effects and improve the overall prediction accuracy. Techniques that utilize network structures to capture complex interactions and dependencies have shown promise in other domains [42, 43], and we believe they can be effectively applied to our research. This integration will allow us to capitalize on the strengths of various data representations, resulting in a more robust and accurate predictive model.

In our study, we observed varying levels of prediction accuracy for different drugs and side effects. For example, the model performed well for drugs like Trovafloxacin (RMSE: 0.72, Pearson correlation: 0.89, p-value < 0.05) and Ecallantide (RMSE: 0.66, Pearson correlation: 0.71, p-value < 0.01), as well as side effects such as infarction (RMSE: 0.30, Pearson correlation: 0.81, p-value < 0.01), impaired glucose tolerance (RMSE: 0.47, Pearson correlation: 0.86, p-value < 0.05), and drug inefficacy (RMSE: 0.58, Pearson correlation: 0.91, p-value < 0.05). These cases demonstrate the model's robustness in accurately predicting the occurrence and frequency of side effects. However, the model showed poor performance for certain drugs and side effects. For instance, Epoprostenol (RMSE: 1.60, Pearson correlation: 0.13, p-value > 0.1) and Clobetasol (RMSE: 1.23, Pearson correlation: 0.24, p-value > 0.1) were among the drugs with low prediction accuracy. Similarly, the side effects "oral pain" and "hepatic necrosis" exhibited high RMSE values (3.03 and 2.47, respectively) and low Pearson correlation coefficients (0.28, p-value > 0.1 and 0.59, p-value < 0.1, respectively), indicating significant discrepancies between the predicted and actual values. Both successful and challenging prediction cases provide a clearer understanding of our model's capabilities and limitations, guiding future improvements and refinements.

Many studies have predicted drug side effect frequencies using known frequency information as input features. For example, Galeano's model [15], MGPred [16], SDPred [14], DSGAT [17], and NRFSE [34] incorporated known side effect frequencies along with various drug properties, employing methods such as non-negative matrix factorization, graph attention networks, and CNNs. In contrast, our study fundamentally differs from these approaches. The key distinction lies in the fact that our model, CrossFeat, does not incorporate known side effect frequency information as part of its feature construction. Instead, CrossFeat is designed to predict side effect frequencies without relying on any prior frequency data, thus enabling a true cold-start scenario. This is achieved by leveraging a cross-feature learning approach that integrates features from drug and side effect encoders through a transformer-based architecture. Moreover, unlike previous methods, CrossFeat can predict side effect frequencies for new drugs without any prior knowledge of their interactions with specific side effects. This means that we not only lack information on the frequency of side effects but also have no data on whether the new drug causes any specific side effects. This level of prediction without pre-existing interaction data sets our model apart from existing approaches.

However, CrossFeat has some limitations. The uneven distribution of side effect frequency values poses a challenge, particularly for less frequent occurrences. Of the 36,850 samples with side effect frequency information, approximately 3% exhibited a frequency of 1, 11% exhibited a frequency of 2, 27% exhibited a frequency of 3, 47% exhibited a frequency of 4, and 11% exhibited a frequency of 5. The model's performance is intricately tied to the sample size, with variations in prediction accuracy across different frequency levels. The RMSE for the smallest sample (frequency $= 1$) was 1.57 and the MAE was 1.18. For samples with frequencies 3 and 4, the RMSE is 0.33 and 0.52 and the MAE is 0.22 and 0.32, respectively. To address the imbalanced data issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE) [44], which has been shown to improve performance in other studies [45, 46]. We oversampled samples with frequency=1 to balance the dataset and improve prediction for the minor class. The application of SMOTE resulted in the following performance metrics: AUROC: 0.79 ± 0.02, AUPRC: 0.79 ± 0.03, RMSE: 0.86 ± 0.03, and MAE: 0.66 ± 0.04. These results indicate that SMOTE did not enhance the performance of the model for AUROC and AUPRC, with both metrics slightly decreasing from 0.82 to 0.79. This suggests that while SMOTE balanced the dataset, it did not improve overall predictive performance in terms of ranking positive cases (AUROC) and precision-recall balance (AUPRC). Additionally, RMSE remained consistent, while MAE slightly increased, indicating a higher average prediction error for individual instances. These results highlight the complexity of addressing imbalanced data and show that SMOTE may not always lead to improved performance across all metrics.

Recently, computational methods to predict drug responses for cancer have been significantly advanced and used for drug repositioning [47–49]. As CrossFeat demonstrated robustness in predicting side effect frequencies of drugs, the sequential application of drug response prediction methods followed by CrossFeat can expedite the drug development process, improving both the efficiency and accuracy of discovering viable treatments.

Conclusion

In this study, we introduced CrossFeat, a novel approach that uses cross-attention to predict the frequency of side effects for drugs without prior information. By integrating the knowledge of both drugs and side effects in the transformer module, CrossFeat achieves superior performance compared to existing prediction models.

The ability to accurately predict the incidence of adverse drug reactions has immense potential for improving drug safety practices for patients and pharmacists. By providing insight into the likelihood and severity of side effects, our model can facilitate informed decision-making during drug prescription and administration, ultimately minimizing the risks associated with drug use.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05915-2.

Additional file 1:Table S1 Hyperparameter search spaces for CrossFeat and comparison models.**Table S2** Comparison of existing methods.**Tables S3–S10** Optimal hyperparameters per fold for CrossFeat and comparison models.**Tables S11–S13** Optimal hyperparameters per fold for CrossFeat structural ablation study models.**Tables S14–S17** Optimal hyperparameters per fold for CrossFeat feature ablation study models.**Tables S18–S21** Optimal hyperparameters per fold for transformer encoders. Table S22, Performance metrics for different # heads configurations.**Tables S23–S29** Optimal hyperparameters per fold for CrossFeat and comparison models in FAERS_SI dataset.

Acknowledgements
Not applicable.

Author contributions
HL initiated and supervised the project. BB collected data and analyzed the results. HL and BB developed the algorithm and wrote the manuscript. BB performed the experiments. All authors reviewed the manuscript.

Funding
This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-00567, Development of Intelligent SW Systems for Uncovering Genetic Variation and Developing Personalized Medicine for Cancer Patients with Unknown Molecular Genetic Mechanisms, No. 2019-0-01842, Artificial Intelligence Graduate School Program [GIST]).

Availability of data and materials
The frequency classes of drug side effects are available in the Supplementary Data 1 of Galeano et al.'s study (DOI:https://doi.org/10.1038/s41467-020-18305-y), and drug SMILES information is available on STITCH (http://stitch.embl.de/). Any other relevant data and model source code are available in https://github.com/DMCB-GIST/CrossFeat.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 8 May 2024 Accepted: 23 August 2024
Published online: 08 October 2024

References

1. Filimonov DA, Rudik AV, Dmitriev AV, Poroikov VV. Computer-aided estimation of biological activity profiles of drug-like compounds taking into account their metabolism in human body. Int J Mol Sci. 2020;21(20):7492.
2. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. The lancet. 2000;356(9237):1255–9.
3. Forman R, Gilmour-White S, Forman N. Drug-induced infertility and sexual dysfunction. Cambridge, New York: Cambridge University Press; 1996.
4. Meltzer HY. Adverse effects of the atypical antipsychotics. J Clin Psychiatry. 1998;59(SUPPL. 12):17–22.

5. Carleton BC, Smith MA. Drug safety: side effects and mistakes or adverse reactions and deadly errors? Br Columbia Med J. 2006;48(7):329.
6. Gandhi TK, Seder D, Bates DW. Methodology matters. identifying drug safety issues: from research to practice. Int J Qual Health Care. 2000;12(1):69–76.
7. Görög S. Drug safety, drug quality, drug analysis. J Pharm Biomed Anal. 2008;48(2):247–53.
8. Niu Y, Zhang W. Quantitative prediction of drug side effects based on drug-related features. Interdiscipl Sci: Comput Life Sci. 2017;9:434–44.
9. Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Associat. 2011;18(Supplementary–1):144–9.
10. Dimitri GM, Lió P. Drugclust: a machine learning approach for drugs side effects prediction. Comput Biol Chem. 2017;68:204–10.
11. Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E. Metabolic network prediction of drug side effects. Cell Syst. 2016;2(3):209–13.
12. Zhang W, Chen Y, Tu S, Liu F, Qu Q. Drug side effect prediction through linear neighborhoods and multiple data source integration. In: 2016 IEEE International conference on bioinformatics and biomedicine (BIBM). IEEE 2016:427–434.
13. Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinform. 2015;16(1):1–11.
14. Zhao X, Chen L, Lu J. A similarity-based method for prediction of drug side effects with heterogeneous information. Math Biosci. 2018;306:136–44.
15. Galeano D, Li S, Gerstein M, Paccanaro A. Predicting the frequencies of drug side effects. Nat Commun. 2020;11(1):4575.
16. Zhao H, Zheng K, Li Y, Wang J. A novel graph attention model for predicting frequencies of drug-side effects from multi-view data. Brief Bioinform. 2021;22(6):239.
17. Xu X, Yue L, Li B, Liu Y, Wang Y, Zhang W, Wang L. Dsgat: predicting frequencies of drug side effects by graph attention networks. Brief Bioinform. 2022;23(2):586.
18. Zhao H, Wang S, Zheng K, Zhao Q, Zhu F, Wang J. A similarity-based deep learning approach for determining the frequencies of drug side effects. Brief Bioinform. 2022;23(1):449.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems 2017;30.
20. Kuhn M, Letunic I, Jensen LJ, Bork P. The sider database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):1075–9.
21. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Translat Med. 2012;4(125):125–3112531.
22. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model. 2018;58(1):27–35.
23. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, Von Mering C, Jensen LJ, Bork P. Stitch 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2014;42(D1):401–7.
24. Landrum G, et al. Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum. 2013;8:31.
25. Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. Inf Sci. 2015;307:39–52.
26. Jaccard P. The distribution of the flora in the alpine zone. 1. New Phytol. 1912;11(2):37–50.
27. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014:1532–1543.
28. Lipschutz S, Lipson ML. Linear algebra. 4th ed. New York: McGraw-Hill; 2001.
29. O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint 2015. arXiv:1511.08458
30. Zarándy Á, Rekeczky C, Szolgay P, Chua LO. Overview of CNN research: 25 years history and the current trends. In: 2015 IEEE International symposium on circuits and systems (ISCAS). IEEE 2015:401–404.
31. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR 2015:448–456.
32. Agarap AF. Deep learning using rectified linear units (relu). arXiv preprint 2018. arXiv:1803.08375
33. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint 2014. arXiv:1412.6980
34. Wang L, Sun C, Xu X, Li J, Zhang W. A neighborhood-regularization method leveraging multiview data for predicting the frequency of drug-side effects. Bioinformatics. 2023;39(9):532.
35. Park S, Lee S, Pak M, Kim S. Dual representation learning for predicting drug-side effect frequency using protein target information. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3350083.
36. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics. 1970;12(1):69–82.
37. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining, 2016 pp. 785–794.
38. Pearson K. Notes on the history of correlation. Biometrika. 1920;13(1):25–45.
39. Cheng SS, Chan PKJ, Luk H-M, Mok MT-S, Lo IF. Adult Chinese twins with Kenny-Caffey syndrome type 2: a potential age-dependent phenotype and review of literature. Am J Med Genet A. 2021;185(2):636–46.
40. Galderisi M, D'Errico A. β-blockers and coronary flow reserve: the importance of a vasodilatory action. Drugs. 2008;68:579–550.
41. Gaze DC. Coronary artery disease: current concepts in epidemiology, pathophysiology, diagnostics and treatment 2012.
42. Zhao B-W, Su X-R, Hu P-W, Ma Y-P, Zhou X, Hu L. A geometric deep learning framework for drug repositioning over heterogeneous information networks. Brief Bioinform. 2022;23(6):384.
43. Zhao B-W, He Y-Z, Su X-R, Yang Y, Li G-D, Huang Y-A, Hu P-W, You Z-H, Hu L. Motif-aware mirna-disease association prediction via hierarchical attention network. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3383591.

44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
45. Hwang J, Lee H. Mmmf: multimodal multitask matrix factorization for classification and feature selection. IEEE Access. 2022;10:120155–67.
46. Wei J, Lu Z, Qiu K, Li P, Sun H. Predicting drug risk level from adverse drug reactions using smote and machine learning approaches. IEEE Access. 2020;8:185761–75.
47. Park S, Lee H. Molecular data representation based on gene embeddings for cancer drug response prediction. Sci Rep. 2023;13(1):21898.
48. Kim J, Park S-H, Lee H. Pancdr: precise medicine prediction using an adversarial network for cancer drug response. Brief Bioinform. 2024;25(2):088.
49. Baek B, Jang E, Park S, Park S-H, Williams DR, Jung D-W, Lee H. Integrated drug response prediction models pinpoint repurposed drugs with effectiveness against rhabdomyosarcoma. PLoS ONE. 2024;19(1):0295629.

Publisher's Note