Computational and Structural Biotechnology Journal 21 (2023) 2973–2984

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Transporter proteins knowledge graph construction and its application in drug development

Xiao-Hui Chen [1], Yao Ruan [1], Yan-Guang Liu, Xin-Ya Duan, Feng Jiang, Hao Tang, Hong-Yu Zhang, Qing-Ye Zhang [*]

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, PR China

ARTICLE INFO

Article history:
Received 27 December 2022
Received in revised form 17 April 2023
Accepted 2 May 2023
Available online 3 May 2023

Keywords:
Transporter knowledge graph
KG embedding
Predictive model
Generative model

ABSTRACT

Transporters are the main determinant for pharmacokinetics characteristics of drugs, such as absorption, distribution, and excretion of drugs in humans. However, it is difficult to perform drug transporter validation and structure analysis of membrane transporter proteins by experimental methods. Many studies have demonstrated that knowledge graphs (KG) could effectively excavate potential association information between different entities. To improve the effectiveness of drug discovery, a transporter-related KG was constructed in this study. Meanwhile, a predictive frame (AutoInt_KG) and a generative frame (MolGPT_KG) were established based on the heterogeneity information obtained from the transporter-related KG by the RESCAL model. Natural product Luteolin with known transporters was selected to verify the reliability of the AutoInt_KG frame, its ROC-AUC (1:1), ROC-AUC (1:10), PR-AUC (1:1), PR-AUC (1:10) are 0.91, 0.94, 0.91 and 0.78, respectively. Subsequently, the MolGPT_KG frame was constructed to implement efficient drug design based on transporter structure. The evaluation results showed that the MolGPT_KG could generate novel and valid molecules and that these molecules were further confirmed by molecular docking analysis. The docking results showed that they could bind to important amino acids at the active site of the target transporter. Our findings will provide rich information resources and guidance for the further development of the transporter-related drugs.

1. Introduction

Drug development is a long and complex process, and each successful drug undergoes an initial screening of over 100,000 candidate compounds and hundreds of preclinical animal experiments at an average cost of $2.8 billion [1]. Even so, 90 % of lead compounds fail in Phase II clinical trials, and drug pharmacokinetics (PK) is mainly responsible for failure [2,3]. The study of drug PK could improve the selection efficiency of candidate compounds to save time and money consumed by drug clinical trials [4,5]. Therefore, drug PK research has always been an important direction in drug development. There are many in vitro and in vivo methods currently available to study PK. However, it is complex and expensive to perform PK experiments based on a large number of compounds.

To reduce failure probability in clinical trials caused by PK, many absorption, distribution, metabolism, and excretion-toxicity (ADMET) predictive models were established to improve the effectiveness of drug design. Daina et al. (2017) have developed an ADMET predictive model using support vector machines (SVM) and Bayesian methods based on the physical and chemical properties of compounds, descriptors, and drug similarity [6]. Cheng et al. (2012) have developed admetSAR 2.0, a web predictive server with 47 predictive models including molecular fingerprints and random forest (RF), SVM, and K-nearest neighbor (KNN) based on different datasets [7]. Schyman et al. have developed a web server vNN containing 15 ADMET predictive models, which can rapidly assess the cytotoxicity, mutagenicity, cardiotoxicity, and other important properties of drug candidates [8]. Minnich et al. have developed an open-source software pipeline ATOM Modeling PipeLine (AMPL), which can predict key safety and pharmacokinetic-relevant parameters by different machine learning models [9]. Wei et al. (2022) have developed a web server Interpretable-ADMET with 90 qualitative classification models and 28 graph-based quantitative regression models to predict ADMET, which also provides interpretive

[*] Corresponding author.
E-mail address: zqy@mail.hzau.edu.cn (Q.-Y. Zhang).
[1] The first two authors contributed equally

https://doi.org/10.1016/j.csbj.2023.05.001

models based on gradient-weighted class activation map for identifying the substructures which are important to specific properties [10]. However, most of these methods mainly rely on drug molecule structural information, and the influence of biomacromolecules such as transporter are not taken into consideration. Some drugs that do not strictly meet the ADMET threshold criteria are also approved by the food and drug administration (FDA). To improve the accuracy of PK study of candidate compounds, more related information should be considered.

Transporters are the main determinant for PK, safety, and efficacy of drugs [11]. Most of transporters are membrane proteins which can transport drugs and other molecules, such as the P-glycoprotein (P-gp) family, the Multidrug Resistance-associated Protein (MRP) family, the Breast Cancer Resistance Protein (BCRP) family, the Organic Anion Transporter (OAT) family, the Organic Anion-transporting Polypeptide (OATP) family and the Solute Carrier Family (SLC) family. The P-gp family is a membrane protein transporter that can use ATP hydrolysis to transport drugs. It mainly exists in the intestine, liver, kidney, brain and vascular endothelial, and can transport a variety of drugs, including anticancer drugs, antiviral drugs, antibiotics, analgesics and antiarrhythmic drugs. The MRP family mainly exists in liver, intestine, kidney, lung, heart, and central nervous system. It participates in the transport of antineoplastic drugs, antibacterial drugs and antifungal drugs. The BCRP family (also known as ABCG2) is a gut-liver cell-expressed ATP-binding glucose transporter that transports anticancer drugs, antibiotics, NSAIDs, antimicrobials, liposome-delivered drugs. The OAT family is a particularly important class of transporter that can transport a variety of organic anions, including drugs, hormones, amino acids, and uric acid. The OATP family transports organic anion drugs, such as antibacterial drugs, monocyte antigens, anticoagulants and non-steroidal anti-inflammatory drugs. The SLC transport drug by regulating membrane permeability, including activity transport and energy transport. It can transport a variety of drugs, including antibiotics, antiviral drugs, anti-inflammatory drugs, antineoplastic drugs, analgesics, neuroleptic drugs, antibacterial drugs, antiallergic drugs, and antispasmodic drugs. Transporters also influence pharmacokinetics characteristics such as absorption, distribution, and excretion of drugs in humans. For example, URAT1 is responsible for the reabsorption of uric acid in the kidney, while OCT1 and MDR1 are involved in the hepatic uptake and efflux of a wide range of drugs. There are also transporters that assist drugs in passing the blood-brain barrier, such as the glucose transporter 1, the fatty acid transport protein (FATP) family and amino acid transporter protein (LAT). The drugs can be transported to brain throughout the bloodstream by specific binding with transporters. The drug off-target effects, the toxicity of drugs, or drug-drug interactions are mainly caused by drug accumulation in non-target tissues, thus resulting in failure to reach target tissues. However, it is difficult to perform drug transporter validation and its structure analysis by experimental methods. Therefore, it is urgent to explore effective method for transporter prediction and drug design based only on transporter sequences.

With the development of technology, artificial intelligence (AI) and big data have been used for different stages of drug discovery. Traditional graph and network methods for integrating biomedical data contain only one relationship, while knowledge graphs can integrate multiple heterogeneous information, such as multiple entities (proteins, targets, drugs, and genes, etc.) and relationships (protein-protein interactions, drug-drug interactions, drug-target interactions, etc.). At present, many studies have demonstrated the ability of knowledge graphs to obtain potential association information between different entities [12–15]. So, the potential association information between transporters and drugs could be explored by knowledge graphs. In this study, the AutoInt_KG and MolGPT_KG frame were constructed by KG-embedding features and structure features, as shown in Fig. 1. The AutoInt_KG frame was trained with the structural node features and sequence features of the transporter and the drug in the knowledge graph firstly. And the transporter that may interact with the drug was predicted by the AutoInt [16] model with a self-attention fusion mechanism. The performance evaluation of the model and case validation of natural products indicated that the AutoInt_KG frame could effectively predict the potential transporters of small molecular. In the MolGPT_KG frame, the drug SEFLES sequences [17] are encoded as vectors via one-hot encoding, and then the KG embedding features and sequence features of the transporter are used as conditions to train the MolGPT model [18] to generate drug-like small molecules with specific transporters. Finally, the docking analysis of the three selected drug transporters between the small molecules generated by MolGPT_KG show that the MolGPT_KG frame could effectively generate novel and valid small molecules which could bind to important active sites of the transporters.

2. Materials and methods

2.1. Knowledge graph construction

The data used in this study were obtained from Variability of Drug Transporter Database (VARIDT) [19], disease-drug data from the Comparative Toxicogenomics Database (CTD) [20], drug side effects data from the Side Effect Resource (SIDER) [21], and drug-drug interaction data from the Drug-Drug Interaction Database (DDInter) [22]. And the details information of the data sources is presented in Supplementary Table S1. The VARIDT database contains the approved or clinical transporter drug data, epigenetic regulatory data, genetic polymorphism data, and data of exogenous factors regulating drug transporter activity. The smallest unit of the knowledge graph is RDF triple[23]. A total of 423 transporters-gene triples, 1897 transporters-drug triples, 73 transporters-endogenic factors triples, 2547 drug-gene triples, 7174 epigenetic regulators-gene triples, 3467 gene-diseases triples, 619 genetic polymorphisms-diseases triples, and 1432 diseases-drug triples were selected from the VARIDT database. Totally, 266345 disease-drug triples were selected from the CTD database. A total of 149614 drug-side effect triples were screened from SIDER, a database providing drug side effect data. Totally, 94508 drug-drug interaction triples were selected from the DDInter database, as shown in Table 1. Then, all drugs are uniformly mapped to PubChem ID, and all proteins collected from different databases are uniformly mapped to Uniprot ID. The knowledge graph constructed in this study contained a total of 20137 nodes and 527888 edges, and all the data were stored in triple format.

2.2. Knowledge graph embedding

In this study, the knowledge graph embedding was performed based on a bilinear model RESCAL [24]. This RESCAL model can map entities to vectors and characterize each relationship in KG by establishing relationship matrix based on the interaction of entity pairs to catch the potential feature vectors of the knowledge graph. This model exhibits multiple advantages. First, it uses collective learning to learn relationship representation. Since the tensor of the entity and the relationship tends to be sparse, the model captures the information of the knowledge graph by a binary decomposition method. Specifically, X_k was expressed as follows.

$$X_k \approx AR_kA, fork = 1, \dots, m \tag{1}$$

Where A denotes the potential weight of an entity; R_k is an asymmetric $r \times r$ squared matrix, and this matrix can simulate the interaction between the k tensor components in X. The RESCAL score is calculated by a bilinear scoring function:

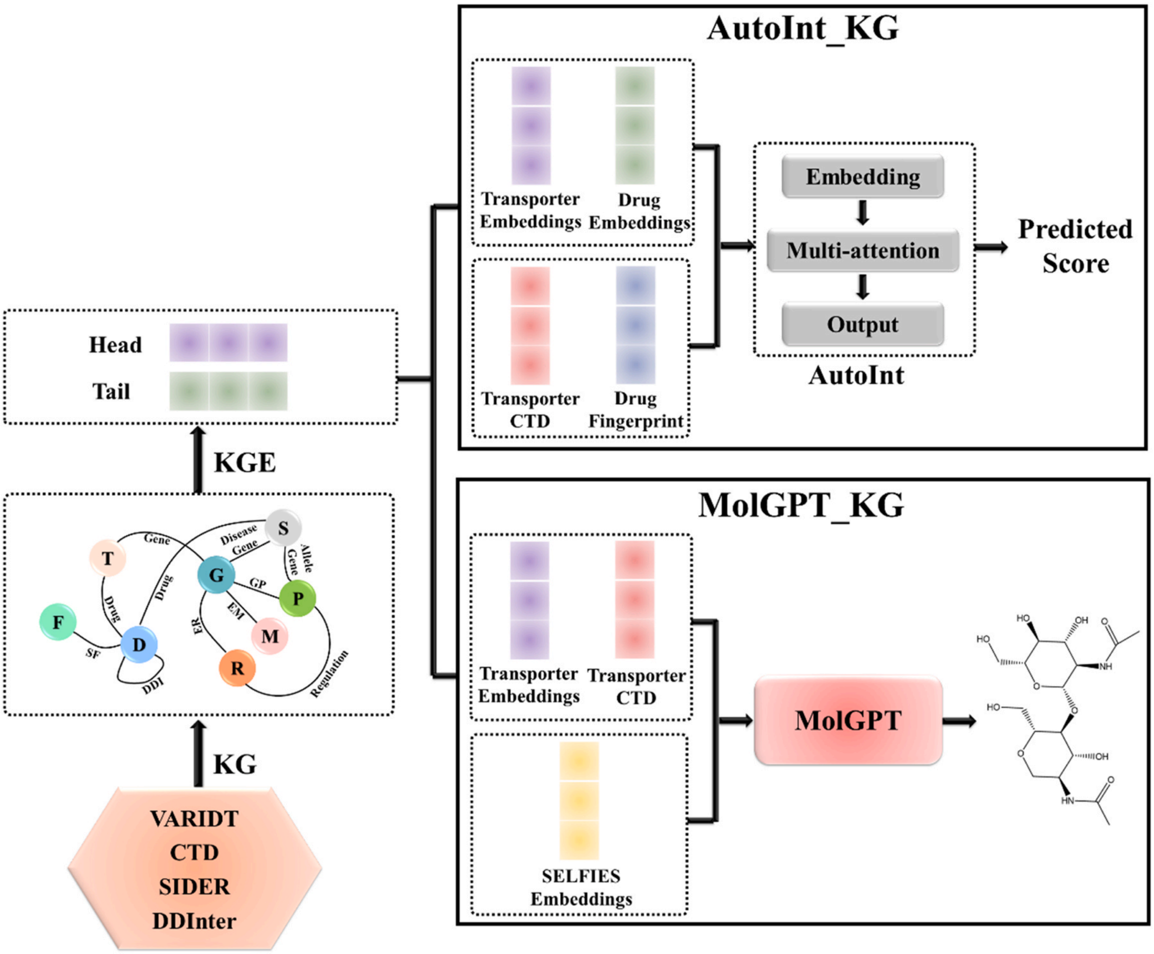


Fig. 1. Workflow diagram of this study. This workflow mainly include two part : AutoInt_KG frame and MolGPT_KG frame. The AutoIn_KG is used to predict the potential transporter of the small molecule. The MolGPT_KG is used to generate novel and valid molecules with the specific transporter.

Table 1
The summarized of the triples in the KG.

Triple	Number
Transporters-gene	423
Transporters-drug	1897
Transporters-endogenic factors	73
Drug-gene	2547
Epigenetic regulators-gene	7174
Gene-diseases	3467
Genetic polymorphisms-diseases	619
Disease-drug	266777
Drug-side effect	149614
Drug-drug interaction	94508

$$f_r(h,t) = h^T M_r t = \sum_{i=0}^{d-1} \sum_{j=0}^{d-1} [M_r]_{ij}[h]_i[t]_j \tag{2}$$

Where h indicates head entity; t denotes tail entity; and $\mathbf{M_r}$ represents relational matrix.

2.3. Construction workflow of predictive AutoInt_KG frame

The construction of transporter predictive frame of AutoInt_KG mainly includes three steps: (1) training data preparation, (2) data preprocessing, and (3) predictive model construction.

At step 1 training data preparation, a total of 1897 transporter-drug interaction data verified by experiment were collected from VARIDT databases and used as positive samples. Subsequently, the tail entities in the triples were randomly replaced by a drug with no clear relationship between transporter, and the treated triples were used as negative samples. To investigate the influence of unbalanced data on the model performance, two types of datasets were constructed with the ratio of positive/negative samples of 1: 1 and that of 1: 10, respectively. Ten sample datasets were constructed with each dataset randomly divided into training and test sets at the ratio of 7: 3 for 10-fold cross-validation.

At step 2 data preprocessing, the heterogeneity information of KG was obtained by the KG embedding method. The high noise of biological data adversely affects the predictive ability of the model. To solve this problem, the sequence characteristics of drug and transporter were added to the model input features. The drug features were represented by molecular fingerprint descriptor (the representation of structural features of a molecule) based on Morgan algorithm [25]. In order to convert the transporter sequence into matrix, the transporter sequence features were represented by CTD descriptor calculated by PyBioMed tool with default parameters [26]. To further solve the problems of data noise and sparseness, the dimension of drug features and transporter sequence features were reduced to 400 via principal component analysis (PCA). Then, the reduced KG embedding features and structure features would be concatenated as input features of the AutoInt_KG.

At step 3 predictive model construction, four models were constructed and compared in this study, including three commonly used classification models of machine learning, namely, logistic regression (LR) [27], SVM [28], and RF [29], and one deep neural network model AutoInt [16]. Among them, the AutoInt model simulates the interaction between different features by adding interaction layers to the neural network where each feature can interact with all other features, and the interaction layer based on the multi-headed self-attention mechanism is the core of AutoInt. This interaction layer can automatically identify relevant features to generate meaningful higher-order features through the multi-attention mechanism. For the m^{th} embedding e_{m}, the standard residual connection is introduced as the output e_m^{Res}.

$$e_m^{Res} = Relu(\tilde{e}_m + W_{Res}{}^{*}e_m), W_{Res} \in \mathbb{R}^{\mathrm{d}'H*d} \tag{3}$$

Where e_m indicates the embedding vector of query m; and $\tilde{e}_m$ denotes the embedding vector after attention head treatment.

The final prediction results are output by sigmoid function $\sigma(\cdot)$:

$$\hat{y} = \sigma(w^T(e_1^{Res} \oplus e_2^{Res} \oplus ...\oplus e_M^{Res}) + b), w \in \mathbb{R}^{\mathrm{d}'HM} \tag{4}$$

AutoInt_KG was evaluated by ROC-AUC and PR-AUC. The ROC-AUC was calculated by selecting a threshold within the [0,1] range to connect the corresponding true positive rate (TPR) and false positive rate (FPR). This metric commonly used to evaluate the performance of classification models, with a value between [0,1]. The closer the value to 1, the better the model's performance. However, the ROC-AUC focus on both positive samples and negative samples, which is not a good evaluation index in the unbalanced datasets. PR-AUC was calculated by the precision and recall metric which measure the ability of a model to predict positive samples correctly, and thus the PR-AUC fluctuated more dramatically than ROC-AUC when the positive and negative samples were unbalanced. According to the dataset we constructed in step 1 above, the number of negative samples is far greater than the number of positive samples (the number of known transporter-drug interaction pairs is too small). In order to evaluate the impact of unbalanced samples on model performance, the PR-AUC was used to measure the performance of the model when datasets were unbalanced.

2.4. Construction workflow of generative MolGPT_KG frame

2.4.1. Construction of MolGPT_KG frame

To generate target transporter-specific small molecules, a generative frame MolGPT_KG was constructed based on the transporter sequence features and heterogeneous information of KG in this study. To construct the generative model, small molecule datasets and transporter data were preprocessed as follows. Firstly, 877 drug-transporter triples verified by experiment were selected and used to train the generative model in this study. To expand the small molecule training dataset and to ensure the diversity and drug-likeness of molecules in the datasets, the similar small molecules with a Tanimoto coefficient (the metric of the similarity between molecules) > 0.8 based on Morgan fingerprint were screened from the ChEMBL database [30,31]. Totally, 7838 small molecules were obtained and used for subsequent analysis. To evaluate the small molecules obtained by similarity search, the distribution of molecular weight (MW), hydrogen bond acceptor number (NumHAcceptors), hydrogen bond donor number (NumHDonors), rotatable bond number (NumRotatableBonds), topological polar surface area (TPSA) and LogP of the 7838 small molecule compounds and the 877 drug small molecules were analyzed and compared. To improve the validity of the generated molecules, a new molecular string representation SELFIES(The representation process of selfies is shown in Supplementary Fig. S2.) was used to train the generative model [17]. SELFIES used Chomskytype-2-based grammar to represent molecules so as to avoid potential invalidity molecular strings problems and improve the validity of the molecule generation with the valence of molecules considered. The transporter sequences were preprocessed with the features of each transporter represented by CTD descriptors, which was consistent with transporter sequences preprocessing in the AutoInt_KG.

Subsequently, the generative MolGPT_KG frame was built based on the MolGPT model proposed by Bagal et al. [32]. The MolGPT model was trained to generate molecules with particular scaffolds and molecular properties. The model was composed of stacked decoder blocks with each decoder block consisting of a masked self-attention layer and a fully connected layer. During the training phase, all molecular tokens were mapped into a vector of 256 dimension through the embedded layer, and the position tokens and the segment tokens were also mapped into the vector of 256 dimension through the separate fully connected linear layer. The segment tokens were used to distinguish condition tokens from molecular tokens. Finally, the molecular token embedding, position token embedding, and segment token embedding were concatenated as the input features to train the generative MolGPT_KG frame. In the MolGPT_KG frame, the molecular optimization conditions were converted into transporter sequence features based on the model architecture of the MolGPT. Therefore, the function of MolGPT_KG was to optimize molecules so as to generate transporter-specific molecules.

To verify whether the KG embedding information could improve the performance of the generative model, two types of input features were used to train the generative model. One type was only based on the transporter sequence feature, and the other type was based on the sequence feature and the KG embedding feature of the transporter. In addition, three known transporter targets were selected for further evaluation, including type 1 glucosamine transporter (PDBID: 5EQG), type L amino acid transporter (PDBID: 6JMQ), and p-glycoprotein 1 (PDBID: 7A69), and their corresponding ligand molecules were (2~{S})-3-(4-fluorophenyl)-2-[2-(3-hydroxyphenyl) ethanoylamino]-~{N}-[(1~{S})-1-phenylethyl] propenamide (5RE), cholesterol (CLR), and vincristine, respectively. Finally, MolGPT_KG generated 5000 molecules for each transporter target based on two different input features.

2.4.2. Evaluation metrics of molecules generated by MolGPT_KG

The performance of the MolGPT_KG was evaluated by the following metrics:

- Validity: It refers to the percentage of valid molecules in all the generated molecules. The RDkit and selfies tools are used to verify whether the generated SELFIES string can be successfully loaded [33]. This validity metric is mainly to assess how well the model learned the SELFIES grammar and the valency of atoms.
- Uniqueness: It refers to the percentage of the unique molecule (without repetition) in the generated valid molecule. This metric is to evaluate how well the model learned the molecular structure distribution.
- Novelty: It indicates the percentage of the generated valid and unique molecules that are not in the training set. Low novelty is a sign of overfitting.
- Internal Diversity: This metric indicates the diversity of the generated molecules. The mean value of Tanimoto coefficients is used to measure the similarity between molecules.

$$Div(S) = 1 - \sqrt{\frac{1}{|s|^2}} \sum_{s1,s2 \in S} T(s1, s2)^p \tag{5}$$

Where ($s1$, $s2$) is all pairs of molecules in a molecular set S; T is absolute value; and p is power.

- Scaffold diversity: It refers to the percentage of the unique scaffolds in the generated molecules.
- QED: This metric quantifies drug-likeness by the main molecular properties. This metric value is between 0 and 1. The closer the value to 1, the higher the drug-likeness [34].
- SAScore: Simple rule-based design is used to quickly assess the ease of synthesis of compounds, and 6.0 is set as threshold as previously reported [35]. SAScore is calculated according to the following formula.

$$SA\ score = fragment\ Score - complex\ Penalty \tag{6}$$

Where *fragment Score* is calculated by the frequency of ECCFP4 fingerprints. *complex Penalty* is calculated based on such elements as ring structure, large ring, and molecular weight. The SAScore ranges from 1 to 10.

To further evaluate the validity of the generated molecules, molecular docking analysis was performed by MOE software to reveal the relationship between generated molecules and their corresponding transporter targets. Docking scores and the important amino acid binding sites were analyzed.

3. Results

3.1. Construction and validation of the AutoInt_KG

3.1.1. Construction of the AutoInt_KG

In this study, 8 models were built based on two types of datasets (1: 1 positive/negative sample ratio for balanced dataset, and 1: 10 for unbalanced dataset). The choice of positive and negative ratios was based on previous work [12]. The ROC-AUC and PR-AUC were used for evaluating model performance. The ablation experiment was performed to determine whether the KG embedding information could improve the performance of AutoInt_KG.

To compare the performance of different models based on the balanced and the unbalanced datasets, ROC-AUC and PR-AUC of each model were displayed in Fig. 2 in the form of boxplots. Based on the balanced datasets, the AutoInt_KG performed best in terms of both ROC-AUC (0.911) and PR-AUC (0.914). This result shows that the transporter prediction model constructed with the knowledge graph embedding feature possess good performance. Based on the unbalanced datasets, AutoInt_KG also performed best in terms of ROC-AUC (0.938) and PR-AUC (0.777). These results showed that no matter in the balanced datasets or the unbalanced datasets, the performance of the AutoInt model with KG embedding features was better than that of the models without KG embedding features. However, compared with that in the balanced dataset, the PR_AUC in unbalanced dataset showed a downward trend, indicating that the unbalance of the datasets had influence on the performance of predictive model. The above results suggested that the AutoInt model with KG embedding also could maintain a certain stability in the unbalanced datasets. The 10-fold cross-validation results of the AutoInt models with or without KG embedding features were presented in the Supplementary Fig. S1, respectively. The results showed that all the models exhibited stability and generalization to some degree.

Considering the above results and the fact that the positive/negative sample ratio was far greater than 10 in this study, the AutoInt_KG was selected for the further validation research.

3.1.2. Validation of the AutoInt_KG

The natural product Luteolin with known transporter information was selected for further validating the effectiveness of the obtained AutoInt_KG predictive frame. The transporters OAT1, OAT3, and OATP1B1 of Luteolin have been reported by Li et al. and Xiang et al. [36,37]. The predicted results of AutoInt_KG was summarized and shown in Table 2. The results showed that all the three reported transporters appeared in the top 15 of the prediction result lists, and transporters OAT3, OAT1, and OATP1B1 ranked the 2nd, 9th, and 11th, respectively. The results confirmed the effectiveness of the AutoInt_KG. Since the reported Luteolin experiment in vitro only monitored the inhibitory activity of OAT1, OAT3, and OATP1B1, the other transporters ranking at the top such as the OCT2 (top1) responsible for the uptake of carnitine by active cells, and the BCRP (top3) mainly responsible for renal and extrarenal urate excretion and protoporphyrin IX export are worthy of further experimental study.

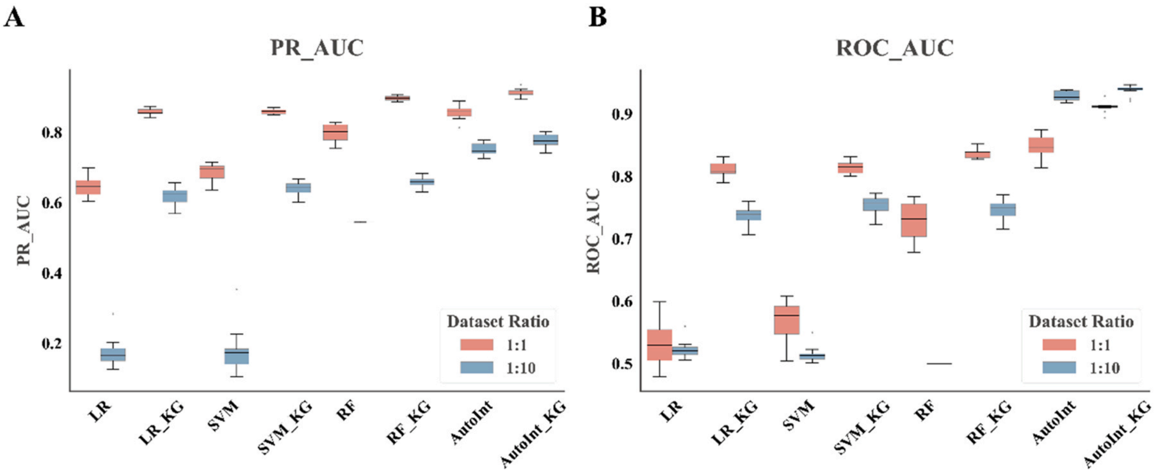


Fig. 2. Comparison of ROC-AUC and PR-AUC in 8 different models. LR, logistic regression model with transporter sequences only; LR_KG, logistic regression model with transporter sequences and KG embedding; SVM, support vector machine model with transporter sequence only; SVM_KG, support vector machine model with transporter sequence and KG embedding; RF, random forest model with transporter sequence only; RF_KG, random forest model with transporter sequence and KG embedding; AutoInt, AutoInt model with transporter sequence only; and AutoInt_KG, AutoInt model with transporter sequence and KG embedding. All results were obtained by 10-fold cross-validation. The 1:1 positive/negative sample ratio indicates balanced datasets and the 1:10 positive/negative sample ratio represents unbalanced datasets. Red box represents balanced datasets, and blue box indicates unbalanced datasets.

Table 2
The summarized of the prediction results for Luteolin.

Uniprot ID	Transporter Name
O76082	Organic cation/carnitine transporter 2 (OCT2)
Q8TCC7	**Organic anion transporter 3 (OAT3)**
Q9UNQ0	Breast cancer resistance protein (BCRP)
Q96FL8	Multidrug and toxin extrusion protein 1 (MATE1)
P33527	Multidrug resistance-associated protein 1 (ABCC1)
O15439	Multidrug resistance-associated protein 4 (ABCC4)
Q15758	Alanine/serine/cysteine/threonine transporter 2 (ASCT2)
Q9NVC3	Putative sodium-coupled neutral amino acid transporter 7 (SNAT7)
Q4U2R8	**Organic anion transporter 1 (OAT1)**
O15245	Organic cation transporter 1 (OCT1)
Q9Y6L6	**Organic anion transporting polypeptide 1B1 (OATP1B1)**
O15438	Multidrug resistance-associated protein 3 (ABCC3)
P46721	Organic anion transporting polypeptide 1A2 (OATP1A2)
Q86VL8	Multidrug and toxin extrusion protein 2 (MATE2)
O76082	Organic cation/carnitine transporter 2 (OCT2)

3.2. Construction and validation of the generative frame MolGPT_KG

3.2.1. Training data acquisition and analysis

To validate the effectiveness of molecules obtained by similarity search, the properties distribution of the obtained molecules and the original drug molecules were analyzed (As shown in Fig. 3). The results showed that the distribution of molecular weight (MW), hydrogen receptor number (NumHAcceptors), hydrogen bond donor number (NumHDonors), rotatable bond number (NumRotatableBonds), LogP, and TPSA of small molecules searched by similarity exhibited high-level overlap with property distribution of the original drug molecules. Therefore, this study expanded the training datasets by using small molecules obtained by similarity search.

3.2.2. Construction and evaluation of the MolGPT_KG

The loss curve of the model shows the model converges at epoch 8 (Supplementary Fig. S3).

Further, we used two types of initial vectors, namely, random initial vectors and fixed carbon atom initial vectors to analyze the effect of initial vectors on molecule generation. Based on these two types of initial vectors, each model generated 5000 molecules, and these molecules were used for evaluating the performance of model. Our results showed that the average validity of generated molecules based on random initial vectors and fixed initial vector was more than 90 % (Tables 3, 4), indicating that the model learned the semantic features of molecular strings and generated semantically correct molecular strings. The novelty of generated molecules was 100 %, suggesting that the model did not overfit. However, the uniqueness of the generated molecules was only around 60 %, which might be due to the structural diversity was limited to a certain range. This limited structural diversity might further be explained by the fact that the molecules of training datasets were obtained by similarity search based on 877 drug small molecules in this study, and the limited sampling space led to the generation of many duplicate molecules.

The uniqueness of the generated molecules based on fixed initial vector was lower than that based on random initial vectors. After adding KG entities embedding information, the uniqueness of the generated molecules based on two types of initial vector was reduced, which might be because adding KG entity embedding was equivalent to increasing the specificity of the transporter features, thus reducing the sampling space. Molecules generated based on random initial vectors generate some single atoms, and such single atoms are meaningless. Relatively, there are no such single atoms in the molecules generated based on a given fixed initial vector. Considering this, the molecules generated based on the fixed initial vectors were used for subsequent validation.

To compare the drug-likeness and the synthetic accessibility of generated molecules, we visualized the distribution of each dataset, as shown in Fig. 4. The Fig. 4A-B showed the QED (drug-likeness) and synthetic accessibility scores (SAScore) distribution. The QED distribution of the generated molecules based on different transporters showed that QED of one fifth of the generated molecules was more than 0.5 [38]. The SAScore distribution showed that SAScore of one

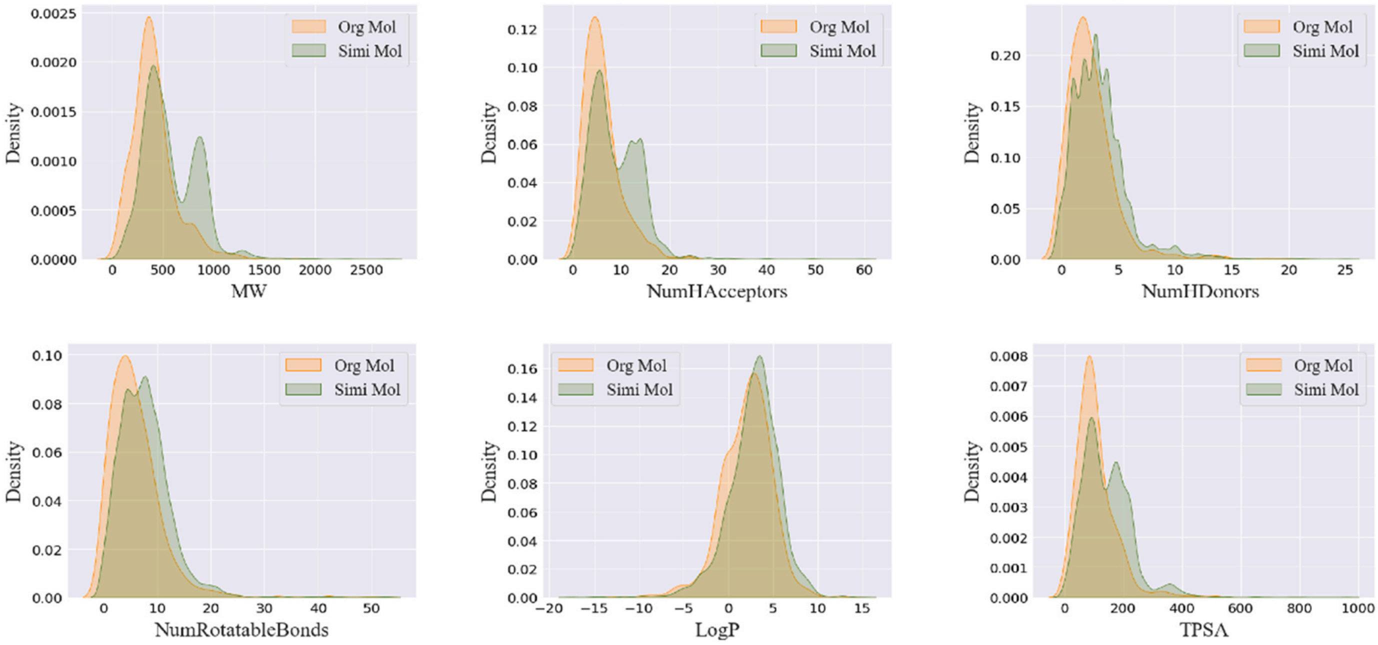


Fig. 3. Comparison of the property distribution between the small molecules obtained by similarity search and the original drug molecules. Orange represents the original drug molecules. Green indicates the small molecules obtained from the similarity search.

Table 3
Evaluation of generated molecules based on random initial vectors.

Transporter	Model	Validity	Uniqueness	Novelty	Sca Diversity	Diversity
P11166	Seq	91.12 %	79.90 %	100 %	45.38 %	91.96 %
	Seq+KG	94.36 %	67.24 %	100 %	39.38 %	92.72 %
Q01650	Seq	92.32 %	78.10 %	100 %	40.95 %	92.16 %
	Seq+KG	96.36 %	63.92 %	100 %	36.20 %	93.22 %
P08183	Seq	91.18 %	79.10 %	100 %	45.51 %	91.94 %
	Seq+KG	94 %	68.82 %	100 %	40.10 %	92.54 %

Table 4
Evaluation of generated molecules based on fixed initial vectors.

Transporter	Model	Validity	Uniqueness	Novelty	Sca Diversity	Diversity
P11166	Seq	99.98 %	60.18 %	100 %	52.08 %	85.13 %
	Seq+KG	100 %	55.60 %	100 %	46.44 %	84.38 %
Q01650	Seq	100 %	62.10 %	100 %	51.79 %	86.00 %
	Seq+KG	100 %	59.34 %	100 %	46.41 %	87.54 %
P08183	Seq	99.98 %	61.52 %	100 %	52.11 %	85.32 %
	Seq+KG	100 %	58.28 %	100 %	48.40 %	85.29 %

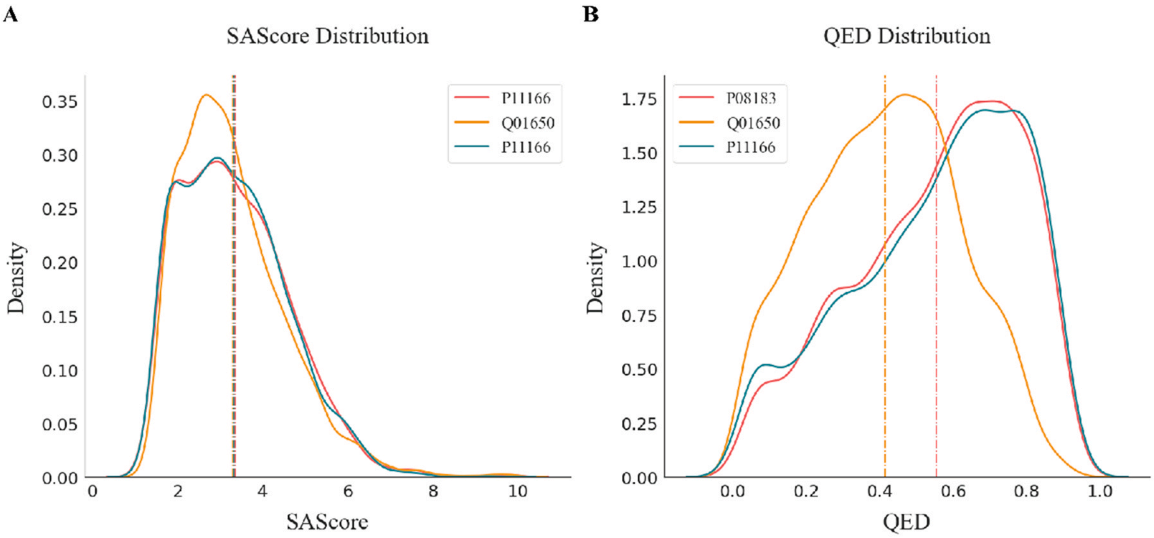


Fig. 4. QED and SAScore distribution of generated molecules. (A) QED distribution. (B) SAScore distribution. Red curve indicates molecules generated based on the transporter P08183. Yellow curve indicates molecules generated based on the transporter Q01650. Green curve represents molecules generated based on the transporter P11166.

third of the generated molecules was more than 6 [39]. The above two metrics (QED and SAScore) showed that the generated molecules by MolGPT_KG had potential to be used as lead compounds. To further validate this potential of MolGPT_KG, molecular docking analysis was performed based on the target transporter.

3.2.3. Validation of the MolGPT_KG by molecular docking

Since the goal of this study is to generate molecules with biological significance, the above evaluation of generated molecules is too general. To further determine whether the generated molecules had transporter-binding capacity, the molecules with QED > 0.4 [40] were selected to dock with the three transporters by MOE software.

Table 5
Docking results of the generated molecules.

Transporter	$S_{generate} > S_{ligand}$	S_{best}
P11166	19.53 %	-11.1671
Q01650	4.77 %	-44.4529
P08183	0.00 %	-10.5525

Table 5 shows the percentage of those generated molecules whose docking score was higher than that of ligands in all the generated molecules ($S_{generated} > S_{ligand}$) and the maximum docking scores (S_{Max}). For the transporter P11166 system, the proportion of $S_{generated} > S_{ligand}$ was 19.53 %. For the transporter Q01650 system, the proportion of $S_{generated} > S_{ligand}$ was 4.77 %. The ligand of P08183 was vincristine, a natural product with high docking scores due to its heterogeneity and ability to form many hydrogen bonds with P08183. The Vincristine has 12 hydrogen bond acceptors and 3 hydrogen bond donors, so it can form many hydrogen bonds with proteins, which leads to the ability of vincristine to bind to multiple transporters at the same time (such as ABCC1, ABCC2, ABCC10, etc.). Therefore, when vincristine enters the human body and binds to these proteins, it will cause some side effects, such as neurotoxicity [41]. Taken together, molecular docking validation analysis results confirmed that the MolGPT_KG could generate valid small molecules against target transporter. In addition, the interaction sites between

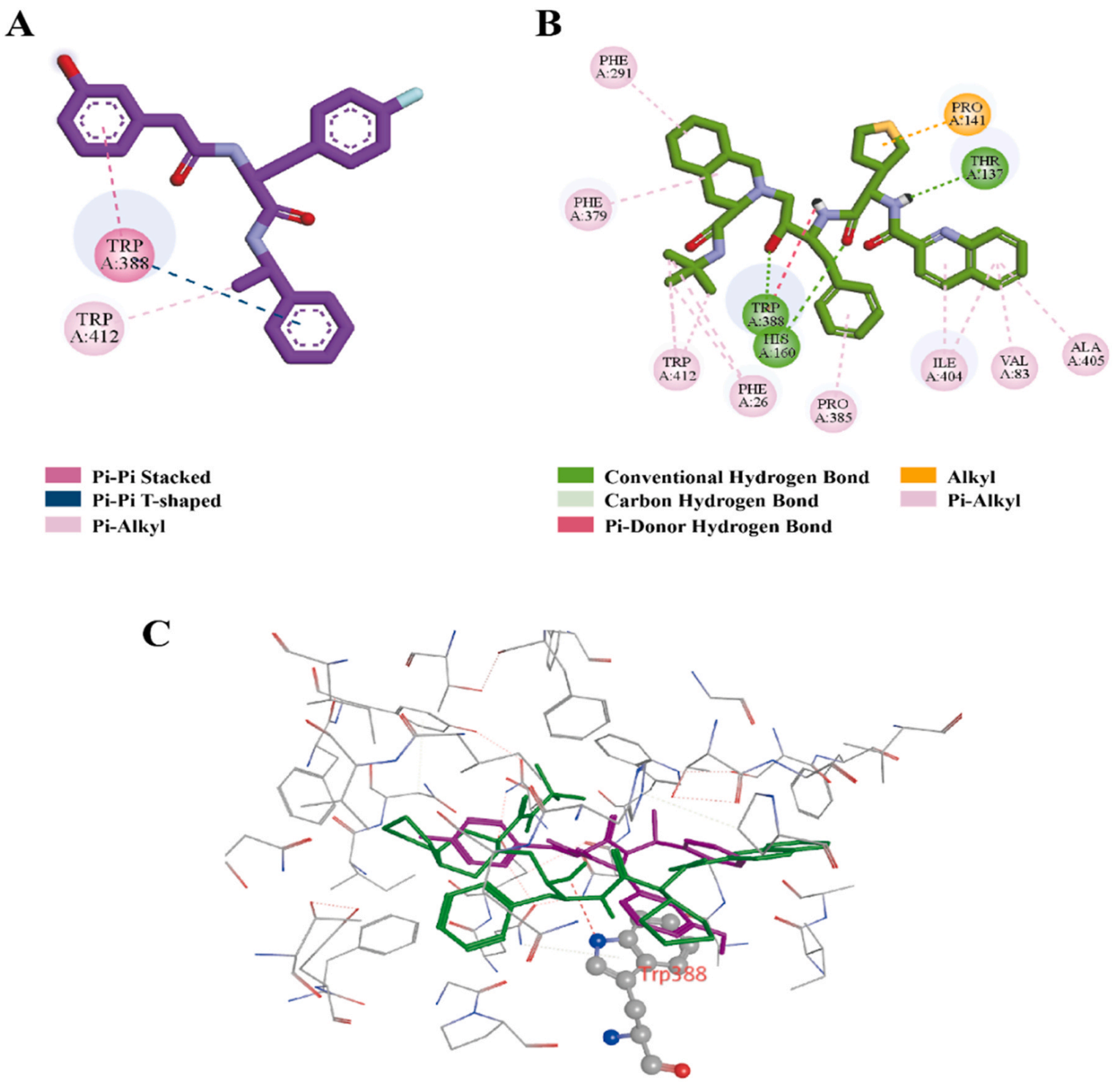


Fig. 5. Interaction between P11166 (PDB ID: 5EQG) transporter, 5RE and generated molecule. (A) the 2D map of molecular interactions between 5RE and P11166 protein. (B) the 2D map of molecular interactions between the generated molecule with the highest docking score and the P11166 protein. (C) the overlap situation between 5RE (purple stick) and the generated molecule (green stick) with the highest docking score in 3D graph.

the generated molecules and the transporter were systematically analyzed.

According to docking score and molecular structure, several molecules were screened for binding site analysis. Figs. 5–7 showed the interaction between P11166, Q01650, or P08183 transporter and the generated molecules with the highest docking score, respectively. And the table about the molecular structure and the docking score of the ligand and top 5 molecules was shown in the Supplementary Tables S2–S4.

As shown in Fig. 5, the P11166 transporter interacted with the selected generated molecules with the highest docking score. Previous study has reported the residues Trp388 are involved in the interaction between P11166 transporter and different ligands (such as GLUT-i1, GLUT-i2 and Cytochalasin B) [42]. The ligand forms one Pi-Pi stacked bond and Pi-Pi T-shaped bond with TRP388. Fig. 5B showed that the generated molecule formed one Conventional Hydrogen Bond and one Pi-Donor Hydrogen Bond with Trp388. The RMSD of less than 2.0 Å is usually used for acceptable docking results [43]. The RMSD of 1.7969 Å for the generated molecular demonstrated that the docking results are acceptable. The overlap situation in 3D view was shown as the Fig. 5C.

As shown in Fig. 6, Q01650 transporter also interacted with the generated molecules with highest docking score. One previous study has revealed that the Q01650 transporter can transport CLR [44]. As shown in Fig. 6A, it can be seen that CLR ligand mainly interact with Leu156, Leu333, and Phe336. Fig. 6B showed that the generated molecule also could form one pi-sigma bond with Leu 156. The RMSD of 1. 6124 Å for the generated molecular demonstrated that the docking results are acceptable. The overlap map was shown in the Fig. 6C.

Fig. 7 showed the interaction between the P08183 transporter and the selected generated molecule. As the main residues, Gln990 and Phe983 have been reported to be involved in the interaction between P08183 transporter and ligand molecule [45]. Our docking results showed that the docking scores of the molecules generated by the model were lower than the ligands, which might be because docking molecule did not occupy enough large active cavity space. This is supported by previous report that the volume of the occupied

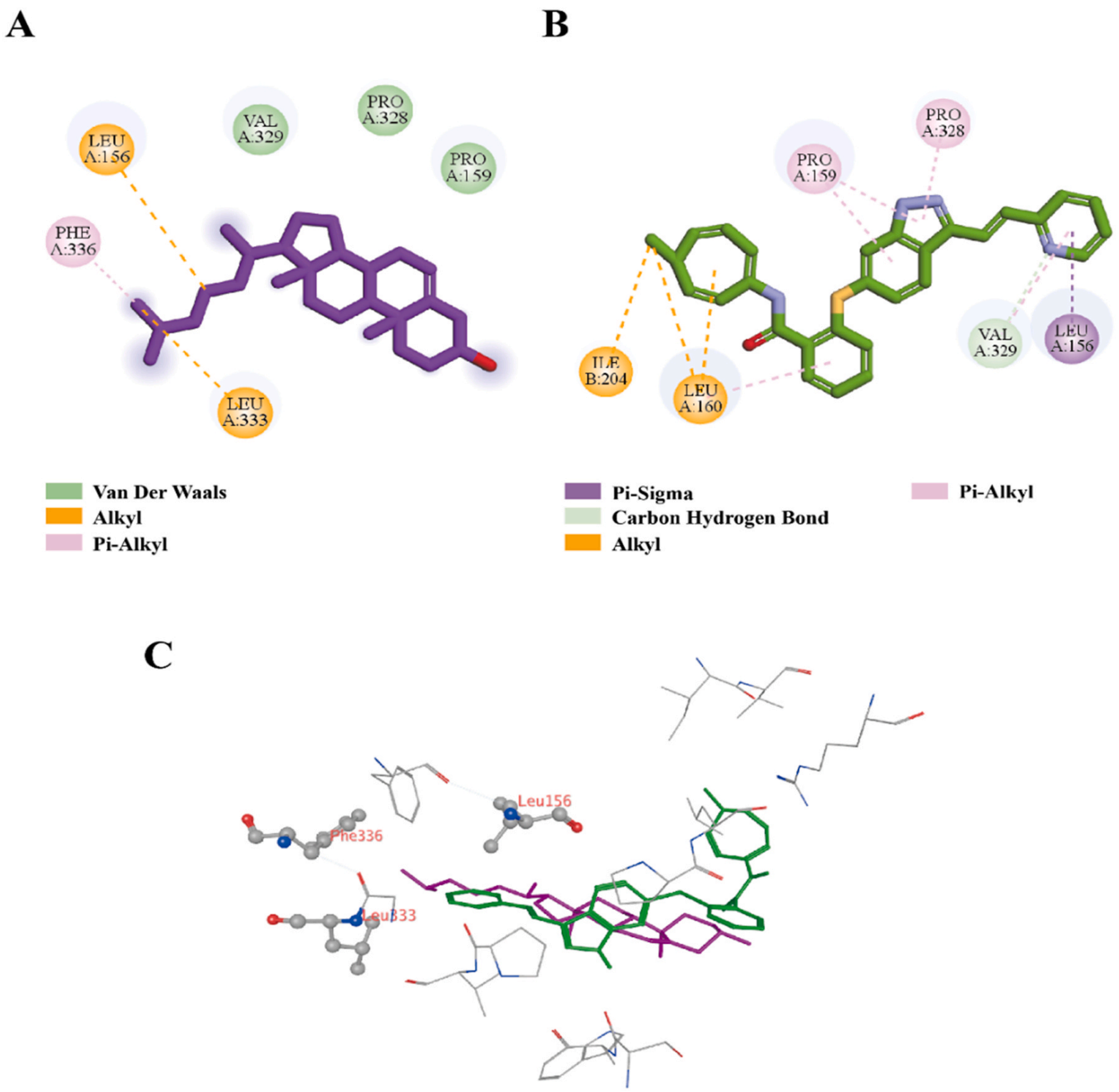


Fig. 6. Interaction between Q01650 (PDB ID: 6JMQ) transporter, CLR and generated molecule. (A) the 2D map of molecular interactions between CLR and Q01650 protein. (B) the 2D map of molecular interactions between the generated molecule with the highest docking score and the Q01650 protein. (C) the overlap situation between CLR (purple stick) and the generated molecule (green stick) with the highest docking score in 3D graph.

pocket has a major influence on the ability of P08183 transporter to bind to the small molecule [45]. However, Fig. 7B showed that the binding site visualization graph showed that four pi-alkyl bonds were formed between the generated molecule and main residue Phe983. The RMSD of 1.9581 Å for the generated molecular demonstrated that the docking results are acceptable. The overlap map was shown in the Fig. 7C.

Considering the interaction between the above three proteins and the generated molecule with the highest docking score, the molecule generated by our model has the potential to bind with important sites of the protein. Except that there is only one overlap between Q01650 protein-binding amino acid and ligand-binding amino acid, the rest of other two proteins are bound to important sites, and the overlapping conformation is also within the desirable range.

To further investigate the diversity of the molecular at important binding sites, Tanimoto coefficient calculated based on Morgan fingerprints was used to examine the similarity between the docking molecules and ligands. As shown in Fig. 8, the similarity was low. The above analysis suggested that our model had the potential to generate novel molecules with the ability to bind transporter active pocket with critical amino acid.

4. Discussion

Currently, researchers in different fields are working on drugs development. However, the most drugs fail in clinical trials due to the pharmacokinetics (PK). The PK affects multiple biological processes such as drug absorption, and drug excretion, off-target, and drug-drug interactions. Although there are several studies of PK, the results are far from satisfactory. Our research on transporters provides a new perspective for the study of drug PK.

Due to the lack of transporter-related experiment data, this research integrated data from the transporter databases VARIDT, the disease database CTD, the side effect database SIDER, and the drug-drug interaction database DDInter to construct a transporter-based knowledge graph. The heterogeneous information of transporter KG obtained by the classical KG embedding algorithm RESCAL was applied to establish predictive frame for predicting drug potential

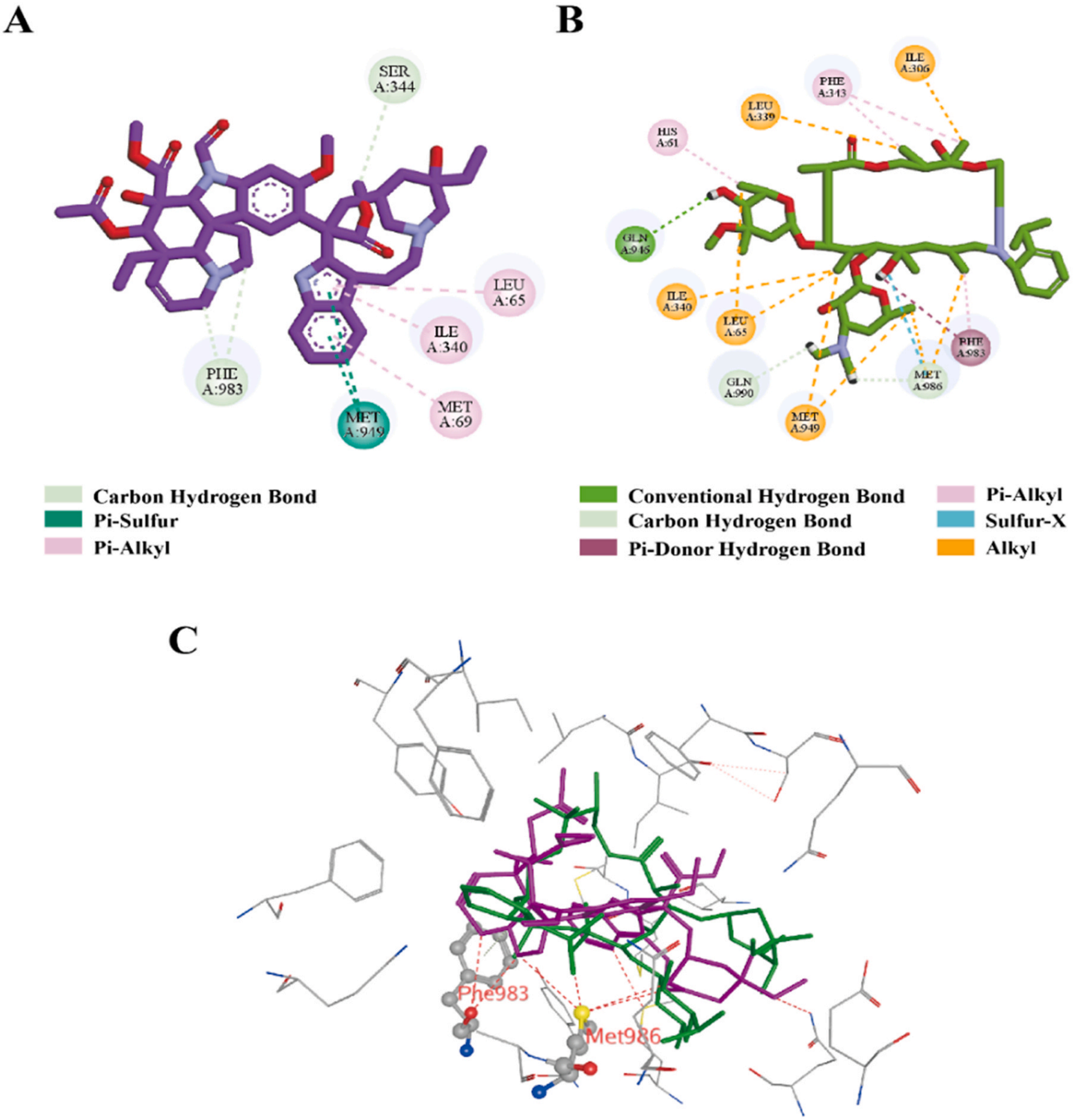


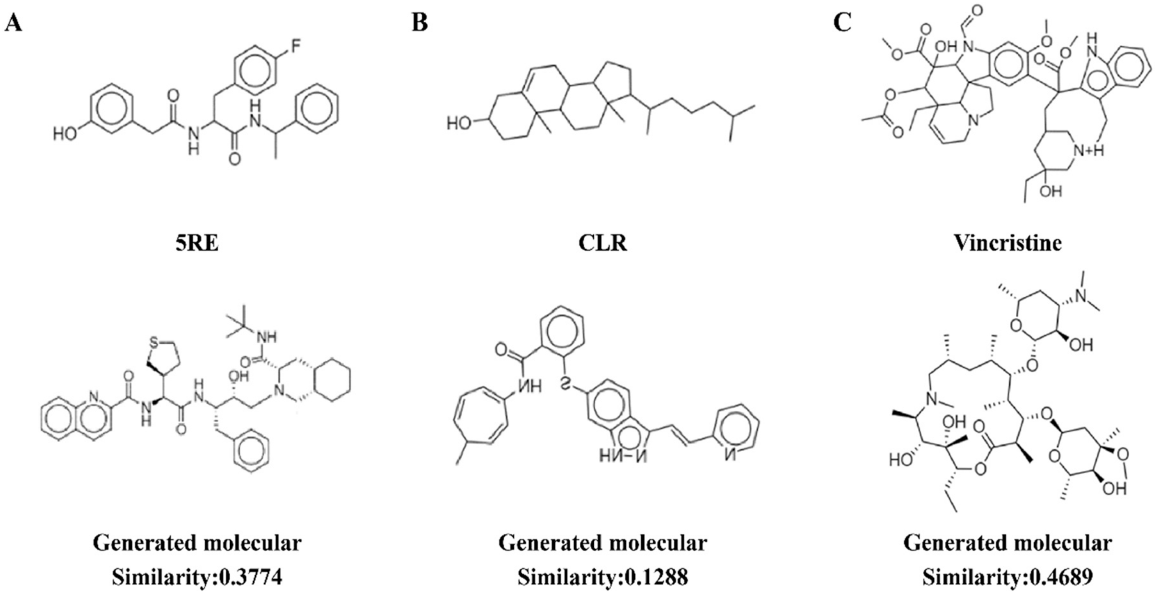
Fig. 7. Interaction between P08183 (PDB ID: 7A69) transporter, vincristine and generated molecule. (A) the 2D map of molecular interactions between vincristine and P08183 protein. (B) the 2D map of molecular interactions between the generated molecule with the highest docking score and the P08183 protein. (C) the overlap situation between vincristine and the generated molecule with the highest docking score in 3D graph.

transporters and generative frame for generating new small molecules against transporters.

The predictive frame AutoInt_KG was trained by integrating transporter sequence features, drug fingerprint features, and KG embedding features. The balanced datasets with positive/negative sample ratio of 1:1 and unbalanced datasets with positive/negative sample ratio of 1:10 were generated for evaluating model performance. Four models LR, SVM, RF, and AutoInt were trained based on these two types of training datasets, and the results showed that AutoInt_KG performed best in both balanced and unbalanced datasets. Furthermore, the ablation experiments demonstrated that KG embedding features can improve the performance of the predictive model. The transporters of natural product Luteolin were predicted by the pre-trained model, and literature validation analysis showed that our model was able to predict potential compound transporters.

In the generative frame MolGPT_KG, three transporters were selected for generating molecules, and the generated molecules were further evaluated. The subsequent docking analysis showed that the model based on transporter sequence features and KG embedding features could generate molecules with a strong ability to bind transporter active pocket at the important binding sites. Our findings will provide theoretical basis for the further development of the transporter-related drugs.

In addition to depending on transporter sequences, transporter-molecular interactions also rely on secondary, tertiary, and even quaternary structures which are more complex than primary sequence structure, and these complex structures are expected to become new research direction. Future studies are also suggested to explore the relation between transporter and drug side effects. Furthermore, transporter KG can be combined with meta path method to analyze the mechanism of transporter.



5RE

CLR

Vincristine

Generated molecular
Similarity:0.3774

Generated molecular
Similarity:0.1288

Generated molecular
Similarity:0.4689

Fig. 8. Similarity between docking molecule and ligand. (A) P11166 transporter. (B) Q01650 transporter. (C) P08183 transporter.

Supporting information

This article contains supporting information.

Funding

This work was supported by Fundamental Research Funds for the Central Universities (No. 2662022XXYJ007, No. 2662022JC004).

CRediT authorship contribution statement

Author 1 (First Author): Conceptualization, Methodology, Formal analyses, Investigation, Writing – Original Draft. Author 2 (Co-First Author): Collected data,. Formal analyses, Writing – Original Draft. Author 3: Data Curation, Review & Editing. Author 4: Visualization, Investigation. Author 5: Validation, Review & Editing. Author 6: Validation, Review & Editing. Author 7: Writing - Review & Editing. Author 8 (Corresponding Author): Conceptualization, Methodology, Funding. Acquisition, Supervision, Writing – Original Draft, Writing - Review & Editing.

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.

Acknowledgement

Xiao-Hui Chen and Qing-Ye Zhang designed the study. Xiao-Hui Chen and Yao Ruan collected data and performed the data analyses. All authors contributed to writing the manuscript. All authors read and approved the final manuscript. Great gratitude goes to linguistics Prof. Ping Liu from Huazhong Agricultural University, Wuhan, China for her work at English editing and language polishing.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.001.

References

[1] Blass BE. Basic principles of drug discovery and development. Elsevier; 2015.
[2] Fleming NJN. How artificial intelligence is changing drug discovery. Nature 2018;557. S55-S55.
[3] Álvarez-Machancoses Ó, Fernández-Martínez JLJ. Using artificial intelligence methods to speed up drug discovery. Expert Opin Drug Discov 2019;14:769–77.
[4] Kaitin K. Obstacles and opportunities in new drug development. Clin Pharmacol Ther 2008;83:210–2.
[5] Ghosh J, Lawless M, Waldman M, Gombar V, Fraczkiewicz R. Silico methods for predicting drug toxicity. Humana Press; 2016.
[6] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 2017:42717.
[7] Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. ACS Publications; 2012.
[8] Schyman P, Liu R, Desai V, Wallqvist A. vNN web server for ADMET predictions. Front Pharmacol 2017;8:889.
[9] Minnich AJ, McLoughlin K, Tse M, Deng J, Weber A, Murad N, et al. AMPL: a data-driven modeling pipeline for drug discovery. J Chem Inf Model 2020;60:1955–68.
[10] Wei Y, Li S, Li Z, Wan Z, Lin J. Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation. Bioinformatics 2022;38:2863–71.
[11] I.T.C. %J, Membrane transporters in drug development. Nat Rev Drug Discov 2010;9:215.
[12] Ye Q, Hsieh C-Y, Yang Z, Kang Y, Chen J, Cao D, et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun 2021;12:1–12.
[13] Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol 2022;40:692–702.
[14] Wang S, Du Z, Ding M, Rodriguez-Paton A, Song TJ. KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. Appl Intell 2022;52:846–57.
[15] Zeng X, Tu X, Liu Y, Fu X, Su YJ. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol 2022;72:114–26.
[16] W. Song , C. Shi , Z. Xiao , Z. Duan , Y. Xu , M. Zhang, et al., Autoint: Automatic feature interaction learning via self-attentive neural networks, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, 1161–1170.
[17] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn: Sci Technol 2020;1:045024.
[18] Bagal V, Aggarwal R, Vinod P, Priyakumar UD. Molgpt: molecular generation using a transformer-decoder model. J Chem Inf Model 2021;62:2064–76.
[19] Fu T, Li F, Zhang Y, Yin J, Qiu W, Li X, et al. VARIDT 2.0: structural variability of drug transporter. Nucleic Acids Res 2022;50:D1417–31.
[20] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative toxicogenomics database (CTD): update 2021. Nucleic Acids Res 2021;49:D1138–43.
[21] Kuhn M, Letunic I, Jensen LJ, Bork PJ. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44:D1075–9.
[22] Xiong G, Yang Z, Yi J, Wang N, Wang L, Zhu H, et al. DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety. Nucleic Acids Res 2022;50:D1200–7.
[23] Nickel M, Murphy K, Tresp V, Gabrilovich EJ. A review of relational machine learning for knowledge graphs. Proc IEEE 2015;104:11–33.

[24] Nickel M, Tresp V, Kriegel H-P. A three-way model for collective learning on multi-relational data. Icml. 2011.
[25] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63.
[26] Amerifar S, Norouzi M, Ghandi M. A tool for feature extraction from biological sequences. Brief Bioinforma 2022.
[27] R.E. Wright, Logistic regression, Circulation, (1995).
[28] Jakkula V. Tutorial on support vector machine (svm), School of EECS vol. 37. Washington State University; 2006. p. 3.
[29] Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 2016;114:24–31.
[30] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Chemin 2015;7:1–13.
[31] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40:D1100–7.
[32] Bagal V, Aggarwal R, Vinod P, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model 2021.
[33] Landrum G. Rdkit documentation. Release 2013;1:4.
[34] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem 2012;4:90–8.
[35] Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Chemin 2009;1:1–11.
[36] Bi Y, Wang X, Ding H, He F, Han L, Zhang Y. Transporter-mediated natural product-drug interactions. Planta Med 2022.
[37] Xiang Y, Liu S, Yang J, Wang Z, Zhang H, Gui C. Investigation of the interactions between flavonoids and human organic anion transporting polypeptide 1B1 using fluorescent substrate and 3D-QSAR analysis. Biochim Et Biophys Acta (BBA)-Biomembr 2020;1862:183210.
[38] Warner KD, Hajdin CE, Weeks KMJ. Principles for targeting RNA with drug-like small molecules. Nat Rev Drug Discov 2018;17:547–58.
[39] Ertl P, Schuffenhauer AJ. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Chemin 2009;1:1–11.
[40] Boitreaud J, Mallet V, Oliver C, Waldispuhl JJ. OptiMol: optimization of binding affinities in chemical space for drug discovery. J Chem Inf Model 2020;60:5658–66.
[41] Lopez-Lopez E, Gutierrez-Camino A, Astigarraga I, Navajas A, Echebarria-Barona A, Garcia-Miguel P, et al. Vincristine pharmacokinetics pathway and neurotoxicity during early phases of treatment in pediatric acute lymphoblastic leukemia. Pharmacogenomics 2016;17:731–41.
[42] Kapoor K, Finer-Moore JS, Pedersen BP, Caboni L, Waight A, Hillig RC, et al. Mechanism of inhibition of human glucose transporter GLUT1 is conserved between cytochalasin B and phenylalanine amides. Proc Natl Acad Sci 2016;113:4711–6.
[43] Pagadala NS, Syed K, Tuszynski JJ. Software for molecular docking: a review. Biophys Rev 2017;9:91–102.
[44] Lee Y, Wiriyasermkul P, Jin C, Quan L, Ohgaki R, Okuda S, et al. Cryo-EM structure of the human L-type amino acid transporter 1 in complex with glycoprotein CD98hc. Nat Struct Mol Biol 2019;26:510–7.
[45] Nosol K, Romane K, Irobalieva RN, Alam A, Kowal J, Fujita N, et al. Cryo-EM structures reveal distinct mechanisms of inhibition of the human multidrug transporter ABCB1. Proc Natl Acad Sci 2020;117:26245–53.