


S2DV: converting SMILES to a drug vector for predicting the activity of anti-HBV small molecules

Jinsong Shao , Qineng Gong, Zeyu Yin, Wenjie Pan, Sanjeevi Pandiyan and Li Wang

Corresponding author. Li Wang, School of Information Science and Technology, Research Center for Intelligence Information Technology, Nantong University, Nantong, Jiangsu 226019, China. Tel.: +86 159 5131 8963; Fax: +86 (0513) 55003030. E-mail: wangli@ntu.edu.cn

Abstract

In the past few decades, chronic hepatitis B caused by hepatitis B virus (HBV) has been one of the most serious diseases to human health. The development of innovative systems is essential for preventing the complex pathogenesis of hepatitis B and reducing side effects caused by drugs. HBV inhibitory drugs have been developed through various compounds, and they are often limited by routine experimental screening and delay drug development. More recently, virtual screening of compounds has gradually been used in drug research with strong computational capability and is further applied in anti-HBV drug screening, thus facilitating a reliable drug screening process. However, the lack of structural information in traditional compound analysis is an important hurdle for unsatisfactory efficiency in drug screening. Here, a natural language processing technique was adopted to analyze compound simplified molecular input line entry system strings. By using the targeted optimized word2vec model for pretraining, we can accurately represent the relationship between the compound and its substructure. The machine learning model based on training results can effectively predict the inhibitory effect of compounds on HBV and liver toxicity. The reliability of the model is verified by the results of wet-lab experiments. In addition, a tool has been published to predict potential compounds. Hence, this article provides a new perspective on the prediction of compound properties for anti-HBV drugs that can help improve hepatitis B diagnosis and further develop human health in the future.

Keywords: hepatitis B virus, drug discovery, virtual screening, SMILES, word embedding, natural language processing

Introduction

The design and synthesis of compounds that establish high activity for specific targets is one of the most significant processes in pharmaceutical chemistry. Drug design is the inventive process of finding new therapeutic entities based on high-throughput virtual screening. Despite advances in pharmaceutical chemistry, drug design is still a slow and difficult process for evaluating the pharmacological activity of drug molecules against specific targets. In this study, we use a simplified molecular input line entry system (SMILES) as input to process chemical tasks. Deep learning is a typical method for representation learning that has shown very high performance across many different natural language processing tasks. Supervised classification models for SMILES

encoded spatial vectors are used to predict the activity of small molecules to specific targets.

Hepatitis B is a chronic infectious disease caused by hepatitis B virus (HBV). It can lead to serious health problems, including cirrhosis, liver cancer, liver failure and even death. According to statistics, ~350 million people worldwide are chronically infected with HBV, and >600 000 patients die of chronic infection and complications of HBV every year [1].

The virus particles are internalized into the body by endocytosis, and the nucleocapsid is subsequently released into the cytoplasm. The HBV genomic material in the nucleocapsid (i.e. relaxed circular DNA, rcDNA) is transported to the nucleus, and part of the double-stranded virus rcDNA is repaired to form covalently

Jinsong Shao is currently a graduated master student at the School of Public Health, Nantong University. His research focus on leveraging artificial intelligence for drug discovery.

Qineng Gong is currently a PhD student at Department of Medicinal Chemistry, School of Pharmacy, Fudan University. His research focus on antitumor drug molecule design.

Zeyu Yin is currently a master student at the School of Information Science and Technology, Nantong University, Nantong, China. His research focus on natural language processing.

Wenjie Pan is an undergraduate at the department of medical informatics, Nantong University. Her main research interests include text mining and drug discovery.

Sanjeevi Pandiyan is an associate professor of the research center for Intelligence Information Technology, Nantong University. His main research interest is medical artificial intelligence.

Li Wang is currently a professor of the research center for Intelligence Information Technology, Nantong University. His main research interests include medical nature language processing and data mining.

Received: October 2, 2021. **Revised:** December 20, 2021. **Accepted:** December 22, 2021

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

closed circular DNA (cccDNA) using the host DNA repair mechanism. Subsequently, the virus is transcribed and translated into a new nucleocapsid, in which the second generation of viral DNA is synthesized by reverse transcription. Subsequently, mature newborn HBV nucleocapsids containing rcDNA are secreted from host cells [2]. Currently, the main anti-HBV drugs in clinical application are HBV DNA polymerase/reverse transcriptase inhibitor nucleoside analogs [3–5] and the immune modulator interferon [6–9]. The resistance of hepatitis B to drugs is complex, and current antiviral drugs can have unpleasant side effects. In recent years, researchers have synthesized many nonnucleoside active compounds with novel structures, which depict an important class of current anti-HBV therapies. Withdrawal causes and rebound indications usually occur after the discontinuation and/or reduction of these drugs.

Antiviral compounds (AVCs) inhibit the development of viruses in the host cell and are relatively harmless to the host. However, designing safe and effective antiviral drugs is a difficult task due to the high genetic diversity and consequent drug resistance in viruses [10]. Initially, antivirals were discovered using traditional trial-and-error methods [11]. However, the discovery of effective antivirals was a very lengthy process. Later, research on virology helped to identify many target pathways to block viral multiplication [12, 13]. Scientists are now using rational drug design strategies for developing antivirals that target viruses at different stages of their life cycles [14]. To save time and money for discovering a new drug, researchers have widely used various computational methods to screen virtual libraries of compounds before the synthesis and animal testing of chemicals. Among the different approaches, quantitative structure–activity relationships (QSARs) are mostly used [15–17]. In this approach, relationships connecting molecular descriptors and activity are used to predict the properties of other molecules [18]. Molecular descriptors transform the chemical information (structure and linking of groups) of a molecule into simple numbers. QSAR-based virtual screening is an effective computational technique leading toward the identification and design of novel antiviral agents [19]. AVCpred was developed as a web server for the prediction and design of AVCs. In this method, scientists used previously known AVCs against HIV, hepatitis C virus, HBV, human herpesvirus and 26 other viruses with experimentally validated percentage inhibition from ChEMBL—a large-scale bioactivity database for drug discovery [20].

Limited by the conventional drug development process, large-scale drug molecule design, synthesis and anti-HBV activity experiments consume many human and material resources, including time and cost, and decrease the speed of drug development. Artificial Intelligence aided (AI-aided) drug screening prediction models play a substantial role in drug research and development with their advantages of fast speed, low cost and high efficiency in predicting active drug molecules [21–24].

SMILES

The SMILES is a ‘chemical language’ [25] that represents a chemical structure in compact text notation. The application of molecular graph theory for SMILES enables structure specification with the help of a very small and natural grammar. SMILES is also suitable for computer recognition and processing. It is easy to read and can be used by researchers and computers, so it is suitable for designing and producing unique numerical symbols, continuous database retrieval, smooth substructure retrieval and performance prediction models [26].

Extended connectivity fingerprint

Molecular fingerprints are representations of chemical structures originally designed to aid in chemical database substructure searches but later used for analytical tasks such as similarity search, clustering and classification. Molecular fingerprints are commonly used in several areas of drug discovery, including MACCS [27], PubChem [28], Tree Fingerprint [29], MolPrint2D [30, 31], extended connectivity fingerprint (ECFP; [32]), UNITY 2D [28] and MP-MFP [33]. ECFP is a fingerprint methodology specifically designed to capture molecular characteristics associated with molecular activity. ECFP is applied to predict drug activity, but they are not designed for substructure search. A variety of methods, such as similarity search, clustering and virtual filtering, have been used by ECFP to capture molecular features. Since 2000, ECFP has been applied to a wide range of science-related issues [32–34].

Word2vec

Word2vec methodology such as continuous bag-of-words (CBOW) and skip-gram has seen increasing interest in recent years because of its ability to understand word representations using a pair of architectures. CBOW is similar to skip-gram in general. The main difference between the two structures is that CBOW uses context to predict words, whereas skip-gram uses target words to predict context.

There are many gradient schemes for neural networks [35], such as the stochastic gradient descent scheme (SGD), momentum [36] and Adagrad [37]. Current word embedding models use a simple stochastic gradient optimization method, i.e. for different words in the word bag, the gradient equally contributes to each word vector. The most commonly used scheme in word2vec is SGD.

Database

ChEMBL [38] is a free online database developed by EBI (European Bioinformatics Institute) that contains a large amount of binding, function and ADMET (absorption, distribution, metabolism, excretion and toxicity) information on medicinal compounds. These data were regularly extracted manually from major published literature and then further collated and standardized. The database currently contains 2 105 464 compounds, 14 554 targets and 1 383 553 drug activity assays. Through this database, the reported compounds and their activity

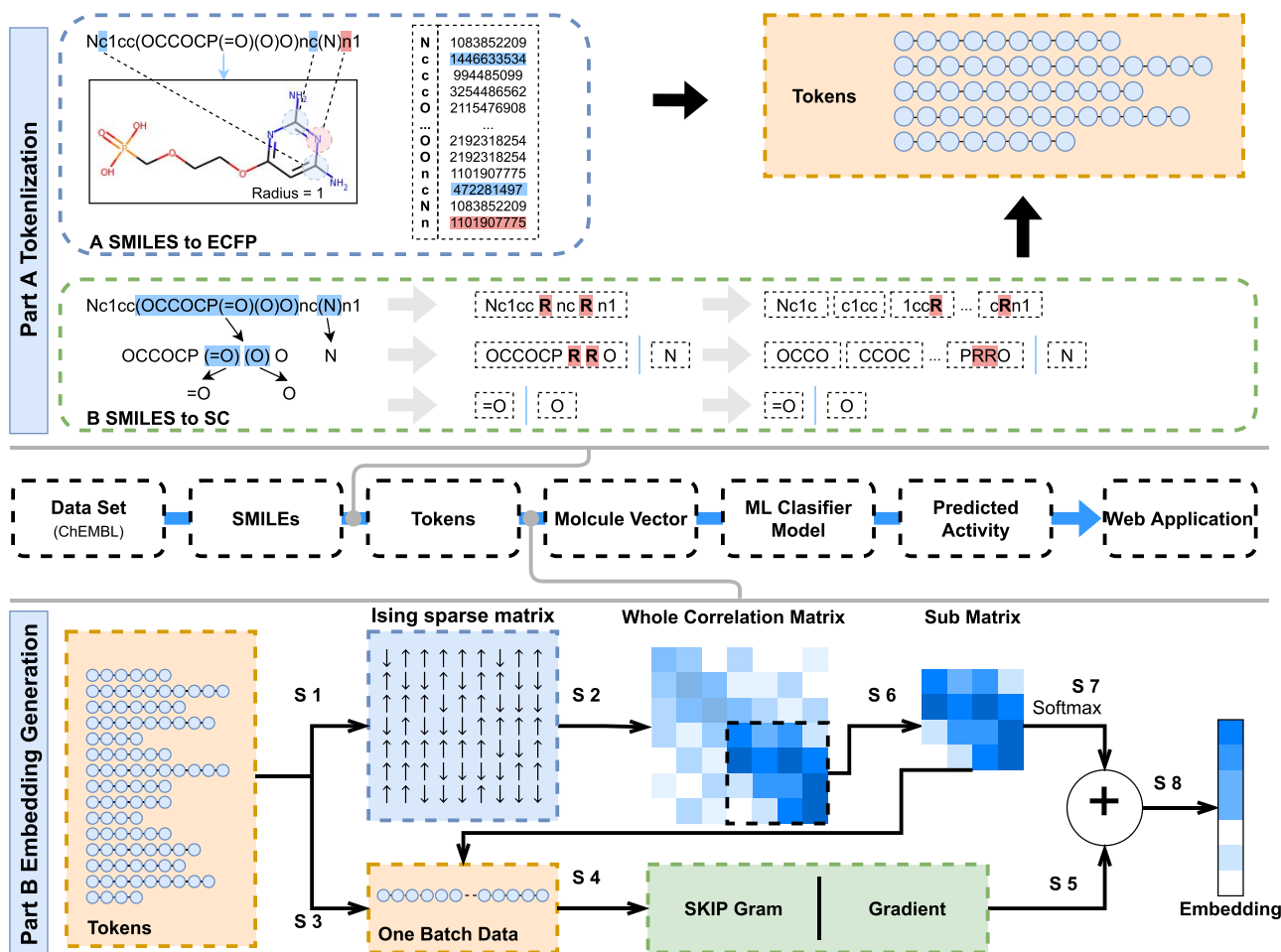


Figure 1. The pipeline overview of S2DV method. **Part A:** Two tokenization method: 'SMILES to SC' and 'SMILES to ECFP'. **Part B:** The process of embedding generation.

information of a target can be quickly queried. The data in this database are all from various reported studies, which are reliable and can be traced to the source of the data. Through this database, accurate chemical compounds and their biological data can be obtained quickly, which can further accelerate the speed of drug design and drug development. In this study, potential anti-HBV drugs were screened, and liver toxicity was considered. Therefore, compounds targeted at HBV and HepG2 cells were screened as training data.

Methods

Tokenization

Usually, word segmentation is required before the text is pretrained for the next step. As shown in Figure 1A, for chemical structure coding SMILES, split character (SC) and molecular fingerprint (ECFP) can be used.

SMILES to SC

- According to the order of SMILES, the sliding window displacement is cut with a specific window size and the displacement is one character each time.
- Delete the "(" and ")" used to represent the branch chain at the branch chain, then use a single letter "R" to retain the branch chain information.

- Double characters such as 'Cl' and 'Fe' indicate a single atom. Replace it with single characters such as 'L' and 'E' that are not confused with other characters.
- For each SMILES, a linear word segmentation list is formed by splicing back and forth according to the sequence of the main chain and branch chain.

As shown in Figure 1A, SMILES expressed as 'Nc1cc(OCCOCP(=O)(O)O)NC(N)N1' compounds were processed by SMILES to SC to form 17 tokens as follows:

Nc1c/c1cc/1ccR/ccRn/cRnc/RncR/ncRn/cRn1/OCCO/CCOC/COCP/OCPR/CPRR/PRRO/N/=O/O.

SMILES to ECFP

Morgan FP is generated by producing and assigning a unique identifier (Morgan identifier) to all substructures around all heavy atoms in the molecule within a defined radius. These identifiers usually hash vectors of fixed length [39].

For each compound encoded by SMILES, SMILES to ECFP generates identifiers for all atoms of a fixed radius and then arrange the identifiers for each atom into a 'molecular sentence' in the order in which the atoms are arranged by SMILES. If an atom has multiple identifiers at the same time, the atomic identifiers are

arranged according to their recognition radius from short to long.

Embedding generation

SMILES-encoded texts are made up of different atoms and atomic structures, and various chemical structures, such as atoms, chemical bonds, branched chains, ring structures and ionic bonds, are described through codes. These chemical structures are related to each other and can be distinguished from each other to form the meaning of similar entities. The concept of an entity refers to the object or concept that exists in the objective world and can be distinguished from each other. An Ising model is proposed to explain a phase change of the ferromagnetic material. The magnetic and nonmagnetic shifts between two phases can be the same in text data to construct an entity of sparse matrix. Between entities, there is no change to reflect. After verification, it can also describe the chemical structure that exists, and there is no change in the entity. Different components of the Ising model system are associated with each other over a long range, which is exactly what is needed in the construction of the global correlation matrix through local interaction [40, 41].

The module shown in Figure 1B is a new word2vec model combined with the Ising model design (Ising-word2vec), which is used to flexibly capture global and local connections in the process of generating word embedding. Tokens extracted from SMILES were input to the model, and the corresponding global gradient modified word embedding model was obtained through the model.

At this point, the chemical structure entity vector will no longer evenly distribute the weight to update the embedding but redistribute the combination with W_{ISM} to form a new chemical structure entity embedding. Compound chemical structure embedding is repeatedly trained by a negative sampling function, and a stable representation of compound chemical structure embedding is finally obtained as follows:

Step 1: The token data of the compound structure extracted from SMILES were used to construct the sparse matrix according to the Ising model data structure Ising sparse matrix (W_{ISM}).

Step 2: The whole correlation matrix (W_{WCM}) was obtained by large-scale sparse data processing of sparse learning with efficient projections.

In this study, logistic regression is adopted to solve the regularization logic problem of the Ising spares matrix.

$$\min_{X_k} \sum_{t=1}^{all} w_t \log(1 + \exp(-z_{kt}(X^T z_t + c))) + \frac{\rho}{2} \|X_k\|_2^2 + \lambda \|X_k\|_1 \quad (1)$$

$$W_{WCM} = (X_1 X_2, \dots, X_{all}) + (X_1 X_2, \dots, X_{all})^T \quad (2)$$

Formula (1) shows the process of solving W_{WCM} , where w_t is the weight of the t th entity in all chemical structure entities, z_t is the t th column of the W_{ISM} we input and z_{kt} is the t th column of the k th row extracted from the sparse matrix, which is used to solve the correlation between z_{kt} and other entities. X_k is the solution of the correlation corresponding to z_{kt} , λ is the regularization parameter of the ℓ_1 norm and ρ is the regularization parameter of the square 2 norm.

Step 3: To solve the problem of memory crash caused by a large amount of data, the input text data entities are divided into N batches to process data, and the i th batch is taken as $Batch_i$. The entities in $Batch_i$ are $(Vm_{i1}, Vm_{i2}, \dots, Vm_{ij})^T$.

$$(Vm_1, Vm_2, \dots, Vm_{all})^T = \sum_{i=1}^n (Vm_{i1}, Vm_{i2}, \dots, Vm_{ij})^T \quad (3)$$

Step 4: Input the chemical structure entities of $Batch_i$ $(Vm_{i1}, Vm_{i2}, \dots, Vm_{ij})^T$ to obtain the initialization of chemical structure entity embedding $(Ve_{i1}, Ve_{i2}, \dots, Ve_{ij})^T$.

$$(Ve_{i1}, Ve_{i2}, \dots, Ve_{ij})^T = (Vm_{i1}, Vm_{i2}, \dots, Vm_{ij})^T \cdot W_{Ve} \quad (4)$$

In formula (4), j in Vm_{ij} is the j th entity in $Batch_i$, and correspondingly, Ve_{ij} is the corresponding embedding of the j th entity. W_{Ve} is the weight matrix during vector generation.

Step 5: $Batch_i$ was paired to generate a skip-gram relation, and $Con(Vm_{ij})$ and $NEG(Vm_{ij})$ of each chemical structure were obtained. According to the study, the predicted probability $\mathcal{P}(u|\tilde{w})$ is calculated. $\tilde{w} \in Con(Vm_{ij})$, $u \in NEG(Vm_{ij})$.

Among them:

$$\mathcal{P}(u|\tilde{w}) = \begin{cases} \sigma(Ve_{ij}^T \theta^u), & (L^{Vm_{ij}}(u) = 1) \\ 1 - \sigma(Ve_{ij}^T \theta^u), & (L^{Vm_{ij}}(u) = 0) \end{cases} \quad (5)$$

According to formula (5) above, $\tilde{w} \in Con(Vm_{ij})$, $u \in NEG(Vm_{ij})$, σ is the sigmoid activation function and $\theta^u \in \mathbb{R}^m$ indicates that the working parameters corresponding to u are also to be trained.

The loss function formula (6) is calculated according to the prediction probability between word pairs in $Batch_i$ Ve_{ij} count context $Con(Vm_{ij})$ and negative sampling word space $NEG(Vm_{ij})$.

$$\begin{aligned} \text{Loss} &= \sum_{Vm_{ij} \in Batch_i} \sum_{\tilde{w} \in Con(Vm_{ij})} \sum_{u \in NEG(Vm_{ij})} \mathcal{P}(u|\tilde{w}) \\ &= \sum_{Vm_{ij} \in Batch_i} \sum_{\tilde{w} \in Con(Vm_{ij})} \sum_{u \in NEG(Vm_{ij})} \left\{ L^{Vm_{ij}}(u) \log[\sigma(Ve_{ij}^T \theta^u)] \right. \\ &\quad \left. + [1 - L^{Vm_{ij}}(u)] \log[1 - \sigma(Ve_{ij}^T \theta^u)] \right\} \end{aligned} \quad (6)$$

Step 6: According to the loss function obtained in Equation (6), multiple optimization schemes can be adopted to obtain the gradient to optimize the parameters, and the gradient descent method is adopted for detailed analysis to facilitate understanding.

Might as well set:

$$\mathcal{L}(Vm_{ij}, \tilde{w}, u) = \sum_{u \in NEG(w)} \left\{ L^{Vm_{ij}}(u) \log[\sigma(Ve_{ij}^T \theta^u)] + [1 - L^w(u)] \log[1 - \sigma(Ve_{ij}^T \theta^u)] \right\} \quad (7)$$

The gradients of θ^u (8) and Ve_{ij} (9) are calculated according to Equation (7):

$$\frac{\partial \mathcal{L}(Vm_{ij}, \tilde{w}, u)}{\partial \theta^u} = [L^{Vm_{ij}}(u) - \sigma(Ve_{ij}^T \theta^u)] Ve_{ij} \quad (8)$$

$$\frac{\partial \mathcal{L}(Vm_{ij}, \tilde{w}, u)}{\partial Ve_{ij}} = [L^{Vm_{ij}}(u) - \sigma(Ve_{ij}^T \theta^u)] \theta^u \quad (9)$$

Step 7: Entities in $Batch_i$ are used to construct skip-gram pairing relationships and negative sampling pairing relationships between entities in Step 5. Based on the pairing of entities, the subcorrelation matrix (W_{SCM}) corresponding to the entity pairing relationship and the negative correlation matrix (W_{NCM}) corresponding to the negative sampling pairing relationship were obtained from W_{WCM} .

Step 8: W_{SCM} and W_{NCM} are combined with the gradient, and the weights are redistributed instead of distributed evenly. Then, backpropagation updates Vm_{ij} word embedding $\tilde{V}e_{ij}$, and the corresponding auxiliary vector u for word $\tilde{\theta}^u$.

$$\tilde{\theta}^u := \theta^u + \eta [1 + \text{softmax}(W_{SCM})] [L^{Vm_{ij}}(u) - \sigma(Ve_{ij}^T \theta^u)] Ve_{ij} \quad (10)$$

$$\tilde{V}e_{ij} := Ve_{ij} + \eta [1 + \text{softmax}(W_{NCM})] \sum_{u \in NEG(w)} \frac{\partial \mathcal{L}(Vm_{ij}, \tilde{w}, u)}{\partial Ve_{ij}} \quad (11)$$

Evaluation of anti-HBV activity and cytotoxicity

All the synthesized compounds were evaluated for *in vitro* anti-HBV activity and cytotoxicity in HepG2 2.2.15 cells using real-time quantitative PCR and MTT methods, respectively. The concentration of compound required for 50% inhibition of DNA replication was defined as the IC_{50} , and the concentration of compound that induced the death of the HepG2 2.2.15 cell cultures by 50% was defined as the CC_{50} [42–44].

Detailed wet-lab experimental operations are described in the supplementary file.

Results

Embedding generation

In this study, the performance difference of the word2vec model before and after the Ising model was compared.

Dataset collection

To reduce the computational dimension required by training and generate a unique vector space with the chemical structure distribution of specific target compounds to accurately represent the spatial distribution relationship of the chemical structure, all the compounds targeted at HBV and HepG2 were screened to generate compound structure word embedding models for HBV and HepG2. In Table 1, a total of 6705 experimental activity values were obtained for screening HBV as a target, of which 4411 were semi-inhibitory concentration values. A total of 182 550 results were obtained targeting HepG2 cells, of which 2270 were semitoxic concentrations.

Tokenization

Considering the SMILES to SC method and the influence of different lengths on embedding performance, we use different lengths for embedding generation. When the split char length is 1 and 2, <100 tokens are generated, and word2vec cannot be used. However, when the split char length is 5, a large number of SMILES cannot generate a sufficient number of tokens. Therefore, we chose 3 and 4 for comparative experiments. As shown in Table 2, better performance can be obtained when the tokenization character length is 4 than when the token extract character length is 3. Therefore, in the subsequent SMILES to SC method, the extracted length defaults to 4.

For tokenization of the ECFP model, we took HBV as an example to compare the model classification performance under different ECFP sampling radii. As shown in Table 3, in the word2vec and Ising-word2vec models, the sampling radius of 1 is not significantly inferior to the larger sampling radius (Radius = 2 or 3). Considering that fewer tokens are generated when the sampling radius is lower (when Radius is 1, 2, 3, n-token is 1474, 7187, 17 668, respectively), which avoids the generation of an overly sparse matrix and greatly improves the calculation requirements. The default sampling radius is 1 in subsequent experiments.

ChEMBL dataset validation

To verify whether anti-HBV compounds have potential research prospects, it is necessary to verify their anti-HBV ability (IC_{50}) and hepatocellular toxicity (CC_{50}).

The embedding model was verified in 4411 HBV inhibition tests and 2270 HepG2 toxicity tests. Eight machine learning models were used to construct compound classification models. In the initial biological data prediction of compounds targeting HBV and HepG2 cells, the threshold value of the semieffective inhibitory concentration

Table 1. ChEMBL database filtering results

Target name	Standard type	Target organism	Assays number
Hepatitis B virus	ALL	Hepatitis B virus	6705
Hepatitis B virus	EC ₅₀ /IC ₅₀	Hepatitis B virus	4411
HepG2	ALL	<i>Homo sapiens</i>	182 550
HepG2	CC ₅₀	<i>H. sapiens</i>	2270

Table 2. Performance comparison of different split char lengths in SMILES with SC

Embedding model	Split char length	Score	LR	LDA	KNN	CART	NB	SVM	XG-Boost	RD-Forest	
Ising-word2vec	4	Accuracy	0.790	0.788	0.794	0.736	0.706	0.854	0.849	0.834	
		Precision	0.698	0.701	0.655	0.597	0.544	0.832	0.827	0.883	
		F1	0.668	0.661	0.718	0.606	0.599	0.760	0.750	0.696	
	3	Recall	0.641	0.627	0.796	0.616	0.666	0.699	0.687	0.575	
		Accuracy	0.776	0.771	0.802	0.728	0.686	0.844	0.845	0.838	
		Precision	0.691	0.687	0.679	0.595	0.531	0.826	0.819	0.882	
	word2vec	4	F1	0.654	0.641	0.732	0.610	0.584	0.750	0.753	0.716
			Recall	0.621	0.602	0.795	0.626	0.650	0.687	0.697	0.603
			Accuracy	0.786	0.784	0.798	0.721	0.709	0.851	0.848	0.835
3		Precision	0.692	0.693	0.665	0.573	0.548	0.831	0.826	0.887	
		F1	0.661	0.655	0.720	0.587	0.603	0.754	0.749	0.696	
		Recall	0.634	0.621	0.786	0.603	0.671	0.690	0.685	0.573	
3		Accuracy	0.777	0.772	0.799	0.736	0.693	0.844	0.847	0.839	
		Precision	0.690	0.689	0.676	0.610	0.541	0.822	0.826	0.881	
		F1	0.657	0.644	0.726	0.617	0.586	0.750	0.753	0.720	
3	Recall	0.627	0.604	0.785	0.624	0.639	0.690	0.698	0.609		

of HBV (IC₅₀/EC₅₀) was set at 1 μ M. The threshold of the HepG2 semitoxic concentration (CC₅₀) was set at 100 μ M to convert the activity data of the compound against the target into binary data for prediction. The experimental results with different units in the dataset were all converted in μ M. The model evaluation parameters were obtained using tenfold cross-validation.

ChEMBL (<https://chembl.gitbook.io/chembl-interface-documentation/downloads>) was used for the latest version of the drug activity database. Filter fixed targets based on fields such as target names and target organizations. Considering that the inhibition rate of HBV (IC₅₀ and EC₅₀) has the same effect, we reserved the EC₅₀ and IC₅₀ data for training and testing of the classification model to reserve as many samples as possible.

HBV inhibition rate (IC₅₀)

Tables 4 and 5 show the comparison of the classification performance of the word2vec model and Ising-word2vec model in the two tokenization methods. Table 4 uses the split SMILES to generate tokens, and Table 5 uses the tokens generated by Morgan's algorithm. Conclusion word2vec combined with the Ising model has a better classification effect than the original word2vec in predicting molecular activity, and RandomForest is the best classifier. Comparing the two different tokenization methods, SMILES to SC (Accuracy: 0.841, Precision: 0.843, Recall: 0.881, F1: 0.853 and AUC:0.92) has better classification results than ECFP generated by Morgan algorithm (Accuracy: 0.837, Precision: 0.835, Recall: 0.878, F1: 0.845

and AUC: 0.90). This may be because the split SMILES method has more advantages than Morgan's algorithm in representation of the chemical structures of rings and branched chains.

Figure 2 shows that when using ECFP as a tokenization method, both word2vec and Ising-word2vec will yield the same AUC score (0.90). In Figure 3, the AUC score of Ising-word2vec (0.92) is better than word2vec (0.91). Therefore, in the calculation of HBV IC₅₀, word Ising-word2vec can lead to a better classification model.

Toxicity of HepG2 (CC₅₀)

We used the same method to train the classification model and obtain conclusion statistics for the toxicity of HepG2 cells. In Tables 6 and 7, we can see that SMILES to SC and SMILES to ECFP were used for tokenization directly, and the Ising-word2vec model with the addition of SMILES to SC produced a slight improvement in performance compared with the direct use of word2vec (SMILES to SC: Accuracy=0.902, Precision=0.908, Recall=0.972, F1=0.935, AUC=0.91; SMILES to ECFP: Accuracy=0.912, Precision=0.974, Recall=0.973, F1=0.943 and AUC=0.94). Moreover, the AUC value is also excellent, which means that the expressions of SMILES can be accurately learned from the Ising model by adding the word2vec model. The Ising-word2vec model can be used to train and accurately predict the drug-like properties and other parameters of a variety of compounds, which will help pharmaceutical personnel avoid compounds

Table 3. Comparison of classification effects of ECFP models with different sampling radii

Embedding method	Radius	N-Token	Model	Accuracy	Precision	Recall	F1
Ising-word2vec	1	1474	LR	0.825	0.827	0.830	0.828
			LDA	0.830	0.835	0.829	0.832
			KNN	0.823	0.824	0.830	0.826
			CART	0.810	0.825	0.795	0.810
			NB	0.699	0.653	0.873	0.747
			SVM	0.832	0.830	0.842	0.836
			XGBoost	0.835	0.825	0.858	0.841
			RDForest	0.837	0.820	0.871	0.845
			LR	0.829	0.826	0.841	0.833
	2	7187	LDA	0.834	0.837	0.837	0.837
			KNN	0.829	0.820	0.851	0.834
			CART	0.813	0.830	0.794	0.812
			NB	0.690	0.646	0.866	0.740
			SVM	0.836	0.837	0.841	0.839
			XGBoost	0.841	0.833	0.859	0.846
			RDForest	0.834	0.819	0.865	0.841
			LR	0.816	0.818	0.822	0.819
			LDA	0.825	0.833	0.821	0.826
	3	17 668	KNN	0.830	0.830	0.839	0.834
			CART	0.805	0.823	0.786	0.804
			NB	0.695	0.653	0.857	0.741
			SVM	0.841	0.843	0.845	0.844
			XGBoost	0.833	0.823	0.857	0.839
			RDForest	0.839	0.824	0.868	0.846
			LR	0.823	0.824	0.829	0.826
			LDA	0.823	0.831	0.818	0.824
			KNN	0.828	0.825	0.841	0.833
word2vec	1	1474	CART	0.815	0.831	0.799	0.815
			NB	0.696	0.649	0.878	0.746
			SVM	0.830	0.829	0.841	0.834
			XGBoost	0.836	0.826	0.860	0.842
			RDForest	0.837	0.821	0.868	0.844
			LR	0.816	0.816	0.824	0.819
			LDA	0.820	0.825	0.822	0.823
			KNN	0.828	0.826	0.839	0.832
			CART	0.807	0.828	0.783	0.805
	2	7187	NB	0.701	0.654	0.875	0.748
			SVM	0.839	0.841	0.846	0.843
			XGBoost	0.841	0.831	0.864	0.847
			RDForest	0.836	0.819	0.870	0.844
			LR	0.816	0.816	0.824	0.820
			LDA	0.817	0.819	0.821	0.820
			KNN	0.825	0.826	0.832	0.828
			CART	0.807	0.824	0.790	0.806
			NB	0.702	0.662	0.849	0.744
3	17 668	SVM	0.838	0.837	0.847	0.842	
		XGBoost	0.832	0.821	0.856	0.838	
		RDForest	0.838	0.821	0.871	0.845	

with no medicinal potential in advance during drug screening, reduce drug screening input and improve drug screening efficiency. In Figure 4, the Ising-word2vec model yielded the highest AUC value (0.94). In Figure 5, the word2vec model yielded the highest AUC value (0.93). This seems to indicate that a higher classification effect can be achieved when different embedding methods are selected.

In vitro validation

To verify the reliability of the model, we used *in vitro* cell experiments to test the prediction model of the compound inhibition rate of HBV and the prediction

model of the compound exclusivity of HepG2 cells. Throughout the trial, we tested data on the activity of 56 compounds, including HBV half-inhibitory concentration (IC₅₀ and EC₅₀) and HepG2 half-toxic concentration (CC₅₀).

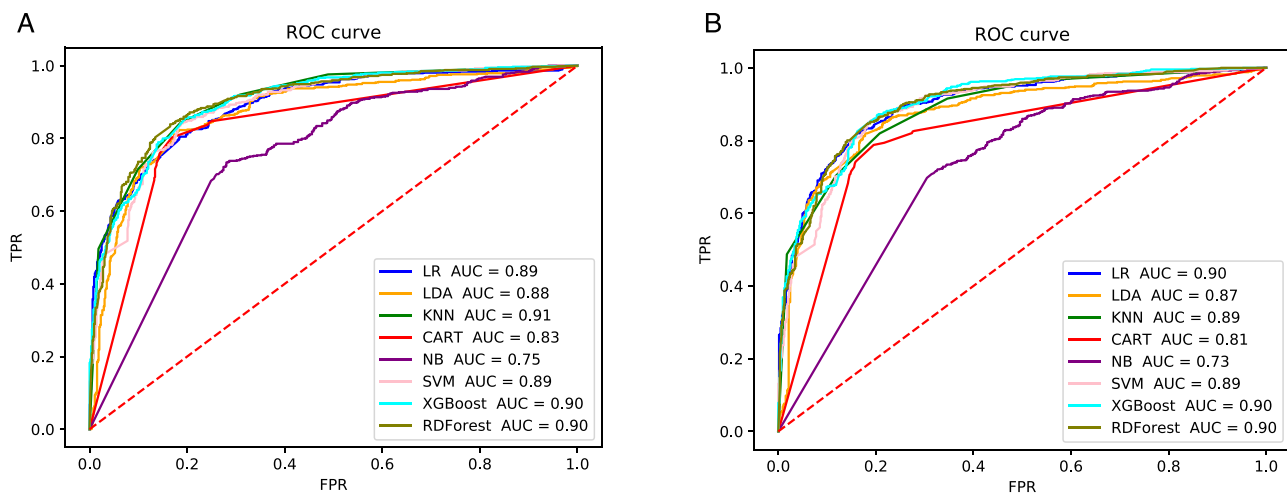
After unifying the units of data, SMILES was input into the model corresponding to the compound, and the binary classification results of the compound for the specified threshold were obtained. A post prediction confusion matrix was established to calculate the accuracy, precision, recall and F1 value. The ROC curve was drawn to calculate the AUC value of the area under the curve.

Table 4. Prediction result of inhibition (IC_{50}) using SMILES to ECFP

Embedding method	Classification	Accuracy	Precision	Recall	F1
word2vec	LR	0.823	0.824	0.829	0.826
	LDA	0.823	0.831	0.818	0.824
	KNN	0.828	0.825	0.841	0.833
	CART	0.815	0.831	0.799	0.815
	NB	0.696	0.649	0.878	0.746
	SVM	0.830	0.829	0.841	0.834
	XGBoost	0.836	0.826	0.860	0.842
	RDForest	0.837	0.821	0.868	0.844
Ising-word2vec	LR	0.825	0.827	0.830	0.828
	LDA	0.830	0.835	0.829	0.832
	KNN	0.823	0.824	0.830	0.826
	CART	0.810	0.825	0.795	0.810
	NB	0.699	0.653	0.873	0.747
	SVM	0.832	0.830	0.842	0.836
	XGBoost	0.835	0.825	0.858	0.841
	RDForest	0.837	0.820	0.871	0.845

Table 5. Prediction result of inhibition (IC_{50}) using SMILES to SC

Embedding method	Classification	Accuracy	Precision	Recall	F1
word2vec	LR	0.824	0.834	0.830	0.832
	LDA	0.825	0.835	0.831	0.833
	KNN	0.825	0.828	0.841	0.834
	CART	0.812	0.834	0.800	0.817
	NB	0.688	0.665	0.815	0.732
	SVM	0.838	0.843	0.849	0.846
	XGBoost	0.840	0.833	0.870	0.851
	RDForest	0.838	0.824	0.881	0.851
Ising-word2vec	LR	0.818	0.825	0.829	0.826
	LDA	0.811	0.825	0.812	0.818
	KNN	0.831	0.836	0.845	0.840
	CART	0.809	0.835	0.794	0.813
	NB	0.681	0.660	0.807	0.726
	SVM	0.836	0.842	0.847	0.844
	XGBoost	0.837	0.831	0.867	0.848
	RDForest	0.841	0.828	0.881	0.853

**Figure 2.** The ROC curve of prediction model with using 'SMILES to ECFP' for tokenization to predict inhibitory effect on HBV (IC_{50}). (A) Embedding generated by word2vec. (B) Embedding generated by Ising-word2vec.

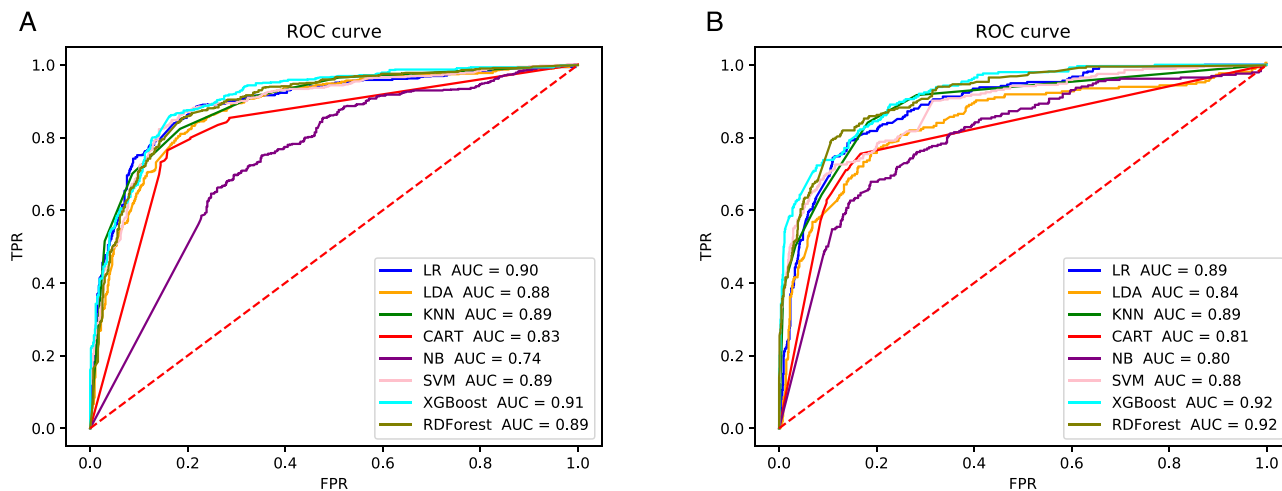


Figure 3. The ROC curve of prediction model with using 'SMILES to SC' for tokenization to predict inhibitory effect on HBV (IC₅₀). (A) Embedding generated by word2vec. (B) Embedding generated by Isingword2vec.

Table 6. Prediction result of toxicity (CC₅₀) using SMILES to ECFP

Embedding method	Classification	Accuracy	Precision	Recall	F1
word2vec	LR	0.879	0.913	0.929	0.921
	LDA	0.874	0.916	0.918	0.917
	KNN	0.886	0.911	0.942	0.926
	CART	0.858	0.915	0.897	0.906
	NB	0.710	0.833	0.772	0.801
	SVM	0.893	0.893	0.977	0.933
	XGBoost	0.912	0.918	0.971	0.943
	RDForest	0.907	0.911	0.972	0.940
Ising-word2vec	LR	0.888	0.914	0.941	0.927
	LDA	0.883	0.921	0.925	0.923
	KNN	0.884	0.910	0.940	0.924
	CART	0.861	0.912	0.904	0.908
	NB	0.710	0.832	0.774	0.802
	SVM	0.889	0.890	0.975	0.930
	XGBoost	0.912	0.916	0.971	0.943
	RDForest	0.904	0.907	0.973	0.939

Table 7. Prediction result of toxicity (CC₅₀) using SMILES to SC

Embedding method	Classification	Accuracy	Precision	Recall	F1
word2vec	LR	0.870	0.908	0.917	0.912
	LDA	0.863	0.904	0.911	0.907
	KNN	0.875	0.900	0.935	0.917
	CART	0.835	0.892	0.885	0.888
	NB	0.692	0.819	0.750	0.782
	SVM	0.892	0.892	0.972	0.930
	XGBoost	0.900	0.907	0.965	0.934
	RDForest	0.894	0.898	0.967	0.931
Ising-word2vec	LR	0.869	0.908	0.916	0.912
	LDA	0.859	0.900	0.911	0.905
	KNN	0.876	0.899	0.939	0.918
	CART	0.816	0.884	0.866	0.874
	NB	0.692	0.817	0.753	0.783
	SVM	0.892	0.892	0.972	0.930
	XGBoost	0.902	0.907	0.966	0.935
	RDForest	0.897	0.899	0.971	0.933

In compound activity tests, the evaluation criteria for compound activity are not fixed values, and a real-time response threshold change interval should be

formed according to the overall activity distribution of the current set of compounds to be tested. In the drug activity test, we tried the HBV IC₅₀ threshold from 1

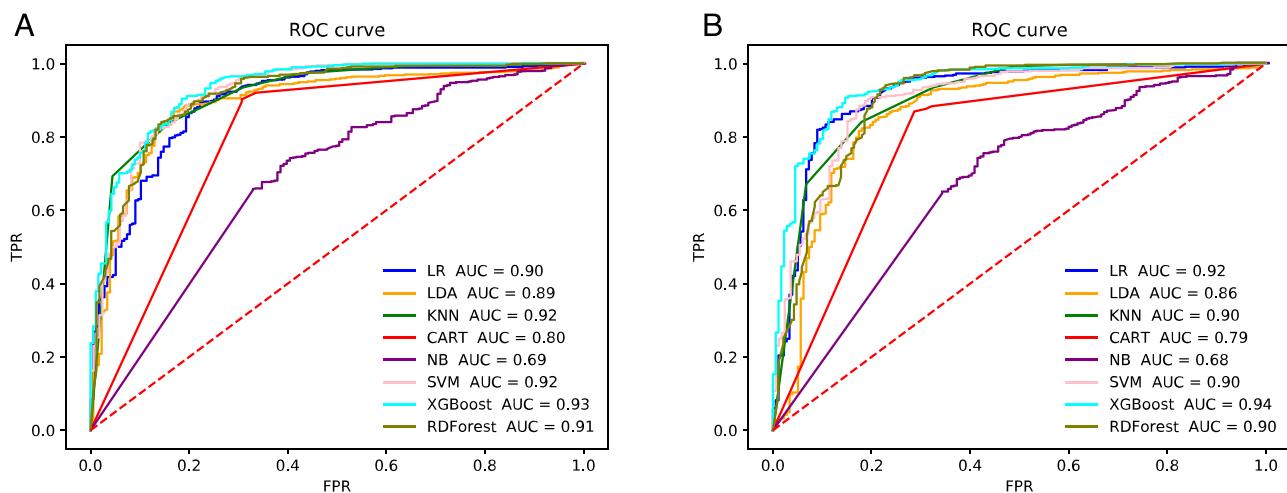


Figure 4. The ROC curve of prediction model with using 'SMILES to ECFP' for tokenization to predict liver toxicity (CC₅₀). (A) Embedding generated by word2vec. (B) Embedding generated by Ising-word2vec.

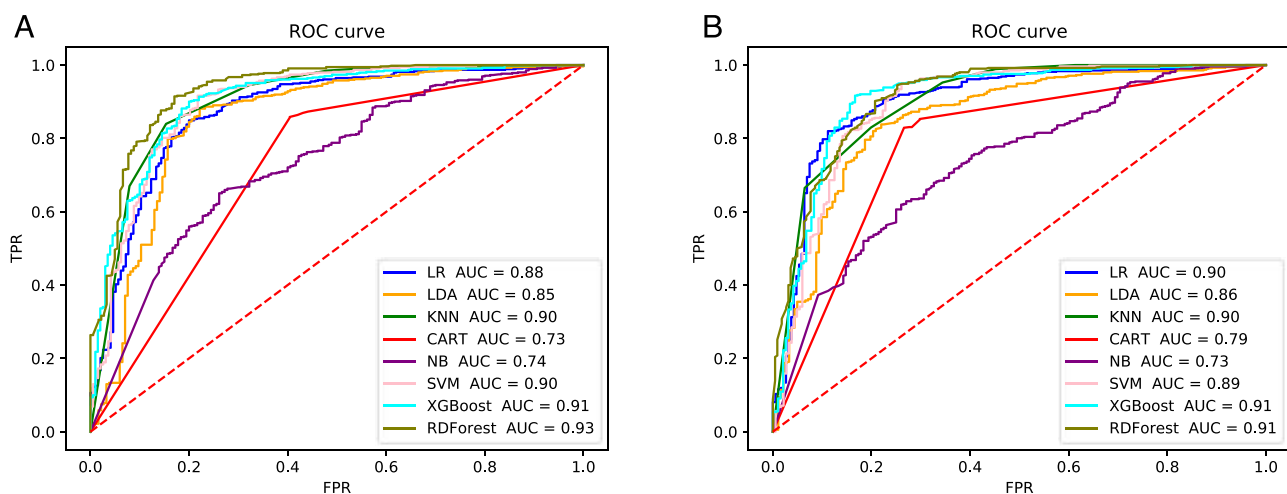


Figure 5. The ROC curve of prediction model with using 'SMILES to SC' for tokenization to predict liver toxicity (CC₅₀). (A) Embedding generated by word2vec. (B) Embedding generated by Ising-word2vec.

to 20 μM and the HepG2 CC₅₀ threshold from 10 to 100 μM .

To establish a relative threshold ratio between compounds to IC₅₀ and CC₅₀, IC₅₀ and CC₅₀ can maintain a large concentration difference to maintain a reasonable active toxicity ratio, which is entirely dependent on the screening needs of drug testers. In this study, we focused more on the drug's ability to inhibit HBV. Therefore, the IC₅₀ threshold was set on a low level. In contrast, the toxicity threshold for HepG2 cells was set loosely at 30 μM .

Tables 8 and 9 show the validation results of the prediction model for HBV drug IC₅₀ with a threshold value of 1 μM and HepG2 drug CC₅₀ with a threshold value of 30 μM , respectively. In the *in vitro* anti-HBV experiment, the model still obtained relatively high prediction accuracy under a specific threshold value. When using the XGBoost classification model, the highest accuracy of IC₅₀ is 0.732, and the highest AUC value is 0.962. When using the SVM classification model, the highest accuracy of CC₅₀ is 0.893, and the highest AUC is 0.943. The results

for other thresholds of IC₅₀ and CC₅₀ are described in the supplementary file (Supplementary Table S1 and S2).

Discussion and conclusion

In this work, we propose a word2vec model combined with Ising gradient correction, which shows better performance than traditional word2vec in different target datasets. In addition, in the downstream task after vector training, we have a definite advantage in predicting the compound's HBV inhibition rate and hepatocyte toxicity. This method has a good ability to screen potential anti-HBV drugs. In addition, in the process of generating tokens, a variety of methods were compared to try to find token extraction parameters that could achieve the optimal performance of the model. We not only demonstrate the advantages of the model for compound representation learning in a public dataset but also demonstrate the predictive power of the classification model based on Ising-word2vec pretraining results in cell experiments. We demonstrate the potential of word2vec combined

Table 8. HBV Drug experimental compound model prediction validation

Threshold	Train set (Pos/Neg)	Test set (Pos/Neg)	Model	Accuracy	AUC
1	3582/829	37/19	LR	0.661	0.463
			LDA	0.643	0.736
			KNN	0.679	0.718
			CART	0.714	0.579
			NB	0.536	0.393
			SVM	0.661	0.748
			XGBoost	0.732	0.962
			RDForest	0.714	0.927

Table 9. HepG2 drug experimental compound model prediction validation

Threshold	Train set (Pos/Neg)	Test set (Pos/Neg)	Model	Accuracy	AUC
30	1502/768	50/6	LR	0.893	0.427
			LDA	0.875	0.800
			KNN	0.750	0.517
			CART	0.804	0.722
			NB	0.321	0.558
			SVM	0.893	0.943
			XGBoost	0.821	0.763
			RDForest	0.875	0.360

S2DV
A tool for Predicting Activity of anti-HBV Small Molecules

Home

predict anti-HBV drugs

* SMILES input:

Nc1cc(OCCOCP(=O)(O)O)nc(N)n1

Nc1cc(OCCOCP(=O)(O)O)nc(N)n1

HBV inhibition prediction	Low inhibition, IC ₅₀ > 1uM
HepG2 toxicity prediction	Low toxicity, CC ₅₀ > 30uM
Conclusion	It has no potential as a anti-HBV drug.

Figure 6. Screen capture of the web tool driven by S2DV.

with the Ising gradient correction model for screening new potential drug-like compounds. This method can be widely applied to predict the drug-like properties of other compounds with different targets, thus simplifying the drug development process.

In Figure 6, tool S2DV is published for anti-HBV drug screening by predicting inhibitory effect on

HBV (IC₅₀) and liver toxicity (CC₅₀) of potential compounds on the web (<http://www.vectorspaceai.cn/S2DV/home>).

Since our experiment demonstrated the superior performance of word2vec combined with Ising gradient modification for drug compound classification tasks, we also wanted to explore the universality of

word2vec combined with Ising gradient modification in different fields. Not only can it effectively improve the performance of models, but its inherent global gradient correction mechanism can also provide many interesting insights.

The scheme still has shortcomings: from the perspective of the model algorithm itself, the model combines local and global, but the global matrix is not optimized with vector time; i.e. the global relational matrix is not completely correct in correcting the gradient. Second, from the perspective of model training and testing, the data amount is limited in the training and verification process of compound medicinal properties. Most of the drugs collected in the database are reported to have higher target activity, and a large number of negative samples are not needed to train a more accurate model, which means that more data of the same target test results are needed to comprehensively optimize the model in later work. In addition, the model only focuses on the compound activity association of a certain target drug and does not make in-depth use of the compound activity data of similar targets and related targets. Therefore, the application of the model needs to be extended to more target data for overall optimization and improvement.

Key Points

- By using the inherent global relationship between molecular SMILES strings, the Ising model can effectively improve the performance of the gradient correction mechanism.
- S2DV demonstrates the superior performance of Ising-word2vec for drug compound inhibition and toxicity classification. The generated molecular embeddings can be used to develop a
- S2DV web development is an online tool to predict the inhibitory effects of potential compounds on HBV (IC₅₀) and liver toxicity (CC₅₀).

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Science Foundation of China (No.81873915).

Data availability statement

All datasets and codes used in this study are available at GitHub: <https://github.com/NTU-MedAI/S2DV>.

References

1. Schweitzer A, Horn J, Mikolajczyk RT, et al. Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *The Lancet* 2015;**386**(10003):1546–55.
2. Berke JM, Dehertogh P, Vergauwen K, et al. Capsid assembly modulators have a dual mechanism of action in primary human hepatocytes infected with hepatitis B virus. *Antimicrob Agents Chemother* 2017;**61**(8):e00560–17.
3. Zhou J, Liu YY, Lian JS, et al. Efficacy and safety of Tenofovir disoproxil treatment for chronic hepatitis B patients with genotypic resistance to other nucleoside analogues: a prospective study. *Chin Med J (Engl)* 2017;**130**(8):914–9.
4. Liu J, Zhang S, Wang Q, et al. Seroepidemiology of hepatitis B virus infection in 2 million men aged 21–49 years in rural China: a population-based, cross-sectional study. *Lancet Infect Dis* 2016;**16**(1):80–6.
5. Fung J, Wong T, Chok K, et al. Oral Nucleos(t)ide Analogs alone after liver transplantation in chronic hepatitis B with preexisting rt204 mutation. *Transplantation* 2017;**101**(10):2391–8.
6. Yuen MF, Schiefke I, Yoon JH, et al. RNA interference therapy with ARC-520 results in prolonged hepatitis B surface antigen response in patients with chronic hepatitis B infection. *Hepatology* 2020;**72**(1):19–31.
7. Gish RG, Yuen MF, Chan HLY, et al. Synthetic RNAi triggers and their use in chronic hepatitis B therapies with curative intent. *Antiviral Res* 2015;**121**:97–108.
8. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. *Cell* 2009;**136**(4):642–55.
9. Buti M, Esteban R. Drugs in development for hepatitis B. *Drugs* 2005;**65**(11):1451–60.
10. Prusoff WH, Lin TS, August EM, et al. Approaches to antiviral drug development. *Yale J Biol Med* 1989;**62**(2):215–25.
11. Bauer D. A history of the discovery and clinical application of antiviral drugs. *Br Med Bull* 1985;**41**(4):309–14.
12. Capobianchi M, Giombini E, Rozera G. Next-generation sequencing technology in clinical virology. *Clin Microbiol Infect* 2013;**19**(1):15–22.
13. Ru J, Li P, Wang J, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J Chem* 2014;**6**(1):13.
14. Müller B, Kräusslich H-G. Antiviral strategies. *Antiviral Strategies* 2009;**189**:1–24.
15. Demchuk E, Ruiz P, Chou S, et al. SAR/QSAR methods in public health practice. *Toxicol Appl Pharmacol* 2011;**254**(2):192–7.
16. Pissurlenkar RR, Khedkar VM, Iyer RP, et al. Ensemble QSAR: a QSAR method based on conformational ensembles and metric descriptors. *J Comput Chem* 2011;**32**(10):2204–18.
17. Ruusmann V, Sild S, Maran U. QSAR DataBank-an approach for the digital organization and archiving of QSAR model information. *J Chem* 2014;**6**(1):1–17.
18. Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A, et al. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications and legal issues. *Curr Pharm Des* 2010;**16**(24):2737–64.
19. Prado-Prado FJ, Borges F, Uriarte E, et al. Multi-target spectral moment: QSAR for antiviral drugs vs. different viral species. *Anal Chim Acta* 2009;**651**(2):159–64.
20. Qureshi A, Kaur G, Kumar M. AVC pred: an integrated web server for prediction and design of antiviral compounds. *Chem Biol Drug Des* 2017;**89**(1):74–83.
21. Mayr A, Klambauer G, Unterthiner T, et al. DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 2016;**3**(8):80.
22. Merget B, Turk S, Eid S, et al. Profiling prediction of kinase inhibitors: toward the virtual assay. *J Med Chem* 2017;**60**(1):474–85.

23. Riniker S, Fechner N, Landrum GA. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. *J Chem Inf Model* 2013;**53**(11):2829–36.
24. Sorgenfrei FA, Fulle S, Merget B. Kinome-wide profiling prediction of small molecules. *ChemMedChem* 2018;**13**(6):495–9.
25. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**(1):31–6.
26. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* 1989;**29**(2):97–101.
27. Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**(6):1273–80.
28. Cereto-Massagué A, Ojeda MJ, Valls C, et al. Molecular fingerprint similarity search in virtual screening. *Methods* 2015;**71**:58–63.
29. Kristensen TG, Nielsen J, Pedersen CN. A tree-based method for the rapid screening of chemical fingerprints. *Algorithms Mol Biol* 2010;**5**(1):9.
30. Bender A, Mussa HY, Glen RC, et al. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J Chem Inf Comput Sci* 2004a;**44**(5):1708–18.
31. Bender A, Mussa HY, Glen RC, et al. Molecular similarity searching using atom environments, information-based feature selection, and a Naïve Bayesian classifier. *J Chem Inf Comput Sci* 2004b;**44**(1):170–8.
32. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc* 1965;**5**(2):107–13.
33. Xue L, Godden JW, Stahura FL, et al. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J Chem Inf Comput Sci* 2003;**43**(4):1151–7.
34. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**(5):742–54.
35. Ruder S. An overview of gradient descent optimization algorithms arXiv 1609.04747. 2016.
36. Chan LK, Jegadeesh N, Lakonishok J. Momentum strategies. *J Financ* 1996;**51**(5):1681–713.
37. Ogren P, Fiorelli E, Leonard NE. Cooperative control of mobile sensor networks: adaptive gradient climbing in a distributed environment. *IEEE Trans Automatic Control* 2004;**49**(8):1292–302.
38. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**(D1):D1100–7.
39. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**(1):27–35.
40. Stutz C, Williams B. Obituary: Ernst Ising. *Phys Today* 1999;**52**(3):106–8.
41. Kobe S. Ernst Ising 1900–1998. *Braz J Phys* 2000;**30**(4):649–54.
42. Gu X, Zhang Y, Zou Y, et al. Synthesis and evaluation of new phenyl acrylamide derivatives as potent non-nucleoside anti-HBV agents. *Bioorg Med Chem* 2021;**29**:115892.
43. Qiu J, Chen W, Zhang Y, et al. Assessment of quinazolinone derivatives as novel non-nucleoside hepatitis B virus inhibitors. *Eur J Med Chem* 2019;**176**:41–9.
44. Qiu J, Gong Q, Gao J, et al. Design, synthesis and evaluation of novel phenyl propionamide derivatives as non-nucleoside hepatitis B virus inhibitors. *Eur J Med Chem* 2018;**144**:424–34.