frontiers in Chemistry

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fchem.2019.00895

SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction

Yu-Fang Zhang[1], Xiangeng Wang[1], Aman Chandra Kaushik[1,2], Yanyi Chu[1], Xiaoqi Shan[1], Ming-Zhu Zhao[3], Qin Xu[1] and Dong-Qing Wei[1,4]**

[1] State Key Laboratory of Microbial Metabolism, and SJTU-Yale Joint Center for Biostatistics and Data Science, School of Life Sciences and Biotechnology, and Joint Laboratory of International Cooperation in Metabolic and Developmental Sciences, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China, [2] Wuxi School of Medicine, Jiangnan University, Wuxi, China, [3] Instrumental Analysis Center, Shanghai Jiao Tong University, Shanghai, China, [4] Peng Cheng Laboratory, Shenzhen, China

OPEN ACCESS

Ed ted by
Zunnan Huang,
Guangdong Medical University, China

ev ewed by
Francesco Ortuso,
University of Catanzaro, Italy
Ling Wang,
South China University of Technology, China

Correspondence
Qin Xu
xuqin523@sjtu.edu.cn
Dong-Qing Wei
dqwei@sjtu.edu.cn

pec a ty sect on
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

ece ved *11 October 2019*
Accepted *12 December 2019*
Pub shed *10 January 2020*

C tat on
Zhang Y-F, Wang X, Kaushik AC, Chu Y, Shan X, Zhao M-Z, Xu Q and Wei D-Q (2020) SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction. Front. Chem. 7:895. doi: 10.3389/fchem.2019.00895

Drug discovery is an academical and commercial process of global importance. Accurate identification of drug-target interactions (DTIs) can significantly facilitate the drug discovery process. Compared to the costly, labor-intensive and time-consuming experimental methods, machine learning (ML) plays an ever-increasingly important role in effective, efficient and high-throughput identification of DTIs. However, upstream feature extraction methods require tremendous human resources and expert insights, which limits the application of ML approaches. Inspired by the unsupervised representation learning methods like Word2vec, we here proposed SPVec, a novel way to automatically represent raw data such as SMILES strings and protein sequences into continuous, information-rich and lower-dimensional vectors, so as to avoid the sparseness and bit collisions from the cumbersomely manually extracted features. Visualization of SPVec nicely illustrated that the similar compounds or proteins occupy similar vector space, which indicated that SPVec not only encodes compound substructures or protein sequences efficiently, but also implicitly reveals some important biophysical and biochemical patterns. Compared with manually-designed features like MACCS fingerprints and amino acid composition (AAC), SPVec showed better performance with several state-of-art machine learning classifiers such as Gradient Boosting Decision Tree, Random Forest and Deep Neural Network on BindingDB. The performance and robustness of SPVec were also confirmed on independent test sets obtained from DrugBank database. Also, based on the whole DrugBank dataset, we predicted the possibilities of all unlabeled DTIs, where two of the top five predicted novel DTIs were supported by external evidences. These results indicated that SPVec can provide an effective and efficient way to discover reliable DTIs, which would be beneficial for drug reprofiling.

Keywords: drug-target interaction, representation learning, Word2vec, machine learning, feature embedding

INTRODUCTION

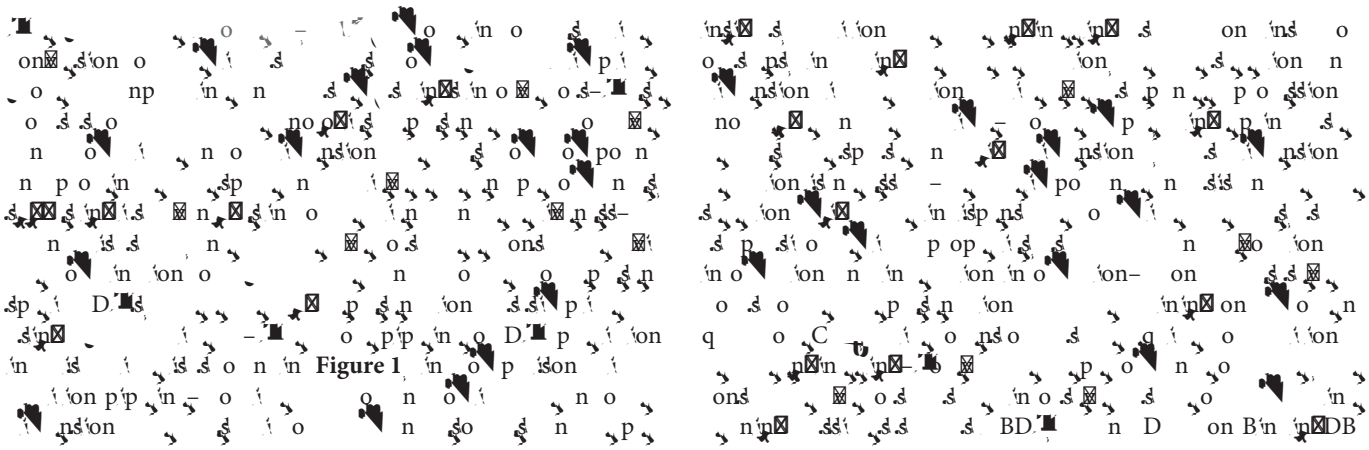
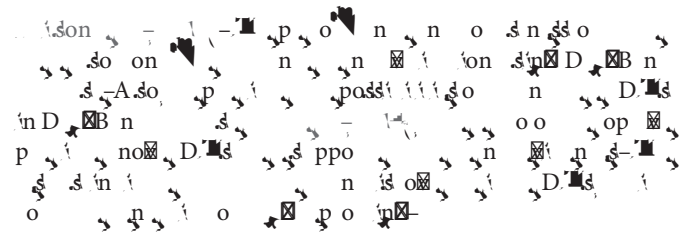
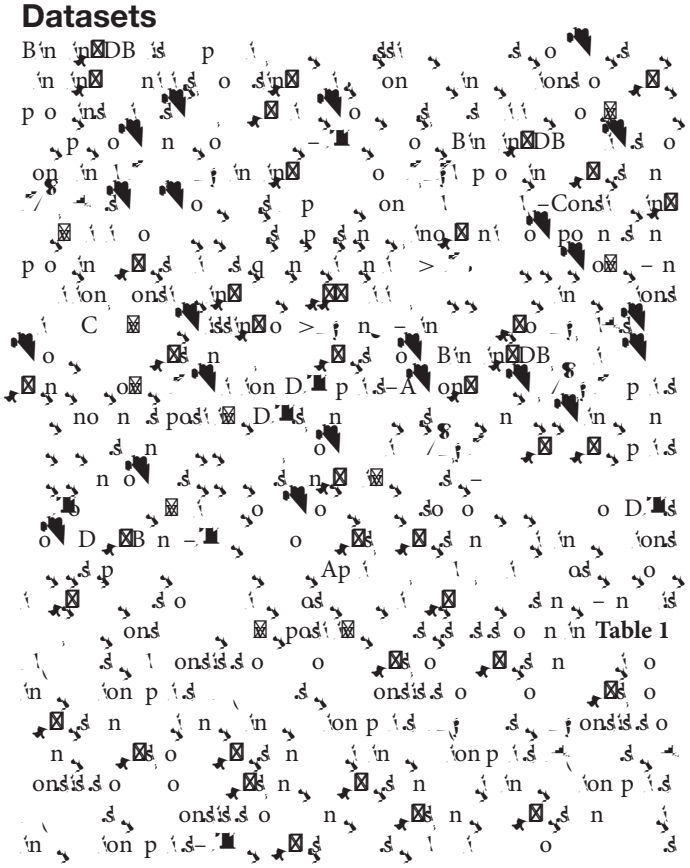


FIGURE 1 | Flowchart of the whole pipeline for DTI prediction in this article **(left)** in comparison to the traditional pipeline **(right)**, with the procedures of feature representations squared in dashed lines.



METHOD

Datasets



Machine Learning Models

RESULTS AND DISCUSSION

Biochemical Implications of SPVec Features

Figure 2B

FIGURE 2 | Biochemical implications from SMILES2Vec features. **(A)** Visualizations of the SMILES2Vec vector space of drugs in DrugBank using t-SNE. **(B)** The top 10 drugs most similar to Acetophenazine (DrugBank ID: DB01063) according to their SMILES2Vec vectors. Red values show their cosine distances with Acetophenazine. The smaller the value, the more similar in the chemical structures.

FIGURE 3 | Normalized distributions of biochemical and biophysical properties in a 2D space projected by t-SNE from the 100-dimensional ProtVec protein-space. In these plots, each point represents a protein, and the colors indicate the scale for each property.

TABLE 1 | Number of entries of the five different datasets obtained from DrugBank dataset.

Datasets	Dataset_1	Dataset_2	Dataset_3	Dataset_4	Dataset_5
Drug	6,068	6,068	537	6,068	537
Target	3,839	3,839	3,839	160	160
Interactions	15,434	3,348	1,735	264	37

Evaluation of the Robustness of SPVec

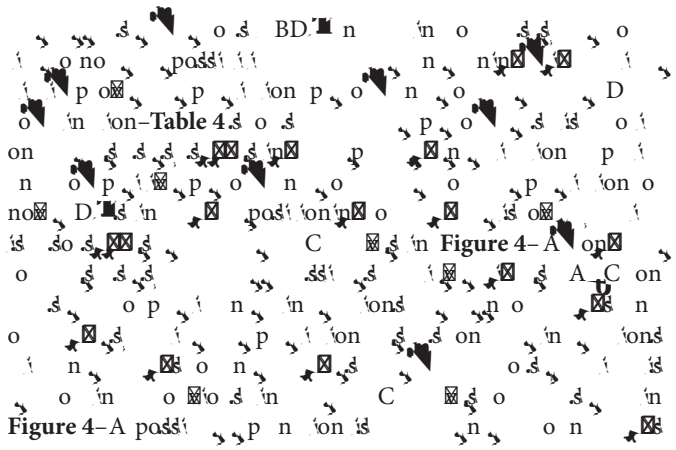
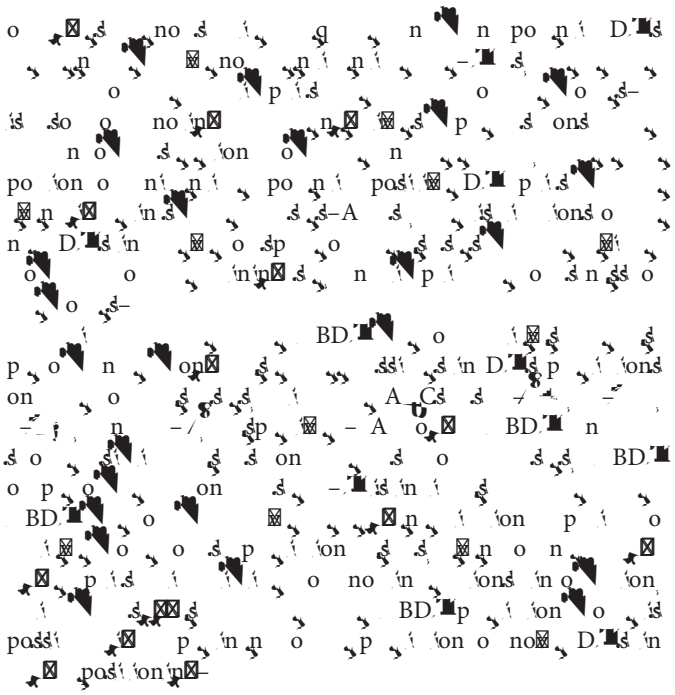


Figure 4– A pass

TABLE 2 | Results of classification performance of four feature combinations using three classifiers on BindingDB via 10 × 5-fold cross-validation, with the highest scores highlighted in the bold font.

Feature combinations	Model	AUC	Accuracy	Precision	Recall	F1-score
SPVec (SMILES2Vec-ProtVec)	GBDT	0.9923	**0.9680**	0.9695	**0.9667**	**0.9681**
	RF	**0.9927**	0.9675	**0.9808**	0.9540	0.9672
	DNN	0.9617	0.9332	0.9287	0.9248	0.9197
SMILES2Vec-AAC	GBDT	**0.9037**	**0.8272**	**0.8563**	**0.7873**	**0.8204**
	RF	0.8770	0.7974	0.8657	0.7050	0.7772
	DNN	0.8708	0.8124	0.7993	0.7879	0.7126
MACCS-ProtVec	GBDT	**0.9479**	**0.8810**	**0.8908**	**0.8690**	**0.8798**
	RF	0.9302	0.8542	0.8712	0.8322	0.8512
	DNN	0.9136	0.8034	0.8025	0.8097	0.8074
MACCS-AAC	GBDT	**0.8588**	**0.7811**	**0.8077**	**0.7392**	**0.7719**
	RF	0.8360	0.7468	0.8366	0.6150	0.7089
	DNN	0.8451	0.7832	0.7884	0.7726	0.7724



Prediction and Validation on Unidentified DTIs

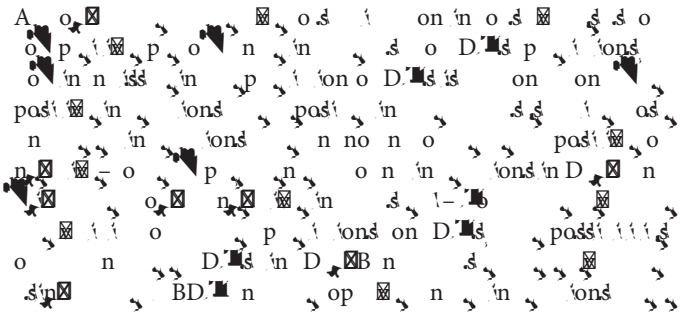


TABLE 3 | AUCs of SPVec and other models on DTI predictions using DrugBank.

Drug features	Drug dim.	Protein features	Protein dim.	ML method	AUC	References
Drug structure information	2,216	AAC, DC[a] and TC[b]	11,943	DNN	0.81	You et al., 2019[c]
Constitutional, topological and molecular descriptors, 2D autocorrelations, topological charge indices, eigenvalue-based indices	1,664	AAC; DC[a]; autocorrelation; Composition, Transition, Distribution descriptors; Quasi-sequence-order	1,080	RF	0.8950	Yu et al., 2012[c]
Constitutional, topological and geometrical descriptors	193	AAC; DC[a]; autocorrelation; composition, transition and distribution; quasi-sequence-order; amphiphilic pseudo-amino acid composition and total amino acid properties	1,260	DT RF	0.760 0.855	Ezzat et al., 2016
PubChem fingerprints indicating presence or absence of 881 known chemical substructures	881	Fingerprints of 876 different protein domains that are obtained from the Pfam database	876	EnsemDT	0.882	Ezzat et al., 2017
				RF	0.855	
SMILES2Vec	100	ProtVec	100	GBDT	0.9467	This work
				RF	0.9469	
				DNN	0.8637	

[a]*DC, dipeptide composition;* [b]*TC, tripeptide composition;* [c]*These models are trained on different versions of DrugBank, whose AUCs are only as references.*

TABLE 4 | Results of classification performance using three classifiers on datasets obtained from DrugBank, with the highest scores highlighted in the bold font.

Dataset	Model	AUC	Accuracy	Precision	Recall	F1-score
Training set	**10 × 5-fold cross-validation**					
Dataset_1	GBDT	0.9506	**0.9323**	**0.9456**	0.9367	**0.9343**
	RF	**0.9557**	0.9234	0.9378	**0.9369**	0.9337
	DNN	0.8952	0.8732	0.8345	0.8437	0.8654
Test sets	**Independent validation**					
	GBDT	**0.8945**	0.8628	**0.8747**	**0.8696**	**0.8637**
Dataset_2	RF	0.8930	**0.8753**	0.8645	0.8467	0.8555
	DNN	0.8201	0.8026	0.8138	0.8199	0.8144
	GBDT	**0.7502**	**0.7389**	**0.7340**	**0.7245**	**0.7333**
Dataset_3	RF	0.7448	0.7299	0.7198	0.7243	0.7230
	DNN	0.6999	0.6922	0.6825	0.6798	0.6832
	GBDT	**0.7356**	**0.7223**	**0.7167**	**0.7177**	**0.7201**
Dataset_4	RF	0.7235	0.7034	0.7108	0.7078	0.71
	DNN	0.7173	0.6899	0.6884	0.6896	0.6866
	GBDT	**0.68**	**0.6703**	**0.6679**	**0.6664**	**0.6688**
Dataset_5	RF	0.5689	0.5605	0.5398	0.5321	0.5411
	DNN	0.6267	0.6098	0.607	0.6122	0.6114

FIGURE 4 | ROC curves with different models on the test sets obtained from DrugBank.

Table 5 –

TABLE 5 | Top five novel DTIs predicted by SPVec-GBDT.

Drug ID	Target ID	Drug name	Target name	Validation source
DB11805	P07947	Saracatinib	The tyrosine-protein kinase Yes	Patel et al., 2013
DB09282	P42262	Molsidomine	Glutamate receptor 2	None
DB05524	Q99640	Pelitinib	Membrane-associated tyrosine and threonine-specific cdc2-inhibitory kinase	https://pubchem.ncbi.nlm.nih.gov/compound/6445562
DB03017	Q16620	Lauric acid	BDNF/NT-3 growth factors receptor	None
DB13165	P11362	Ripasudil	Fibroblast growth factor receptor 1	None

CONCLUSION

DATA AVAILABILITY STATEMENT

AUTHOR CONTRIBUTIONS

FUNDING

SUPPLEMENTARY MATERIAL

REFERENCES

Co C– n pn – ppo o n o s– *Mach. Learn.*
– o – B
D n – n in on – – / – is ip s ip –*J. Mach. Learn. Res.*
ip B –C– n – – o D p in s o n s
s n ip –*J. Chem. Inf. Model.* – o – –
A– – X– – n o C– – – D in ion
p ion s ns n ns n ip– *BMC Bioinf.*
– o – /_s / –
A– – X– – n o C– – – D in ion
p ion s ip ns n ip n ns ion ion– *Methods*
/ / – o – – – – –
n – – – n ion pp o ion n oos ip
in –*Ann. Stat.* / – o – os
on A– A– o o – B n o A– – C s – n D–
– – C B s in –*Nucleic Acids Res.* D D –
o – _n
son – – B – o – p – n C op –
B in ipDB in p s o in is o p ion
is n s s p o o – *Nucleic Acids Res.* –
o – _in
o – B– o s – – C– n is n A– – s n
in p n p pos p n n o o p ip
p op s– *arXiv [Preprint].* A on in ps o s –
– D
o – B– C– is n A– o s – – n B – –
C p ion p n n o in is no
s p o n o p op A o s– *arXiv [Preprint].* A on in ps o s – / s n
–
– o – – op – C– B – A– n –
– ns ion n n s is o o n
s n is / o pos s / o n / on – – A Con n