

## RESEARCH ARTICLE

# Analysis of protein features and machine learning algorithms for prediction of druggable proteins

Tanlin Sun<sup>1</sup>, Luhua Lai<sup>1,2,3</sup> and Jianfeng Pei<sup>1,\*</sup>

<sup>1</sup> Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

<sup>2</sup> Beijing National Laboratory for Molecular Science, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

<sup>3</sup> Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

\* Correspondence: jfpei@pku.edu.cn

Received June 16, 2018; Revised August 18, 2018; Accepted September 5, 2018

**Background:** Computational tools have been widely used in drug discovery process since they reduce the time and cost. Prediction of whether a protein is druggable is fundamental and crucial for drug research pipeline. Sequence based protein function prediction plays vital roles in many research areas. Training data, protein features selection and machine learning algorithms are three indispensable elements that drive the successfulness of the models.

**Methods:** In this study, we tested the performance of different combinations of protein features and machine learning algorithms, based on FDA-approved small molecules' targets, in druggable proteins prediction. We also enlarged the dataset to include the targets of small molecules that were in experiment or clinical investigation.

**Results:** We found that although the 146-d vector used by Li *et al.* with neuron network achieved the best training accuracy of 91.10%, overlapped 3-gram word2vec with logistic regression achieved best prediction accuracy on independent test set (89.55%) and on newly approved-targets. Enlarged dataset with targets of small molecules in experiment and clinical investigation were trained. Unfortunately, the best training accuracy was only 75.48%. In addition, we applied our models to predict potential targets for references in future study.

**Conclusions:** Our study indicates the potential ability of word2vec in the prediction of druggable protein. And the training dataset of druggable protein should not be extended to targets that are lack of verification. The target prediction package could be found on [https://github.com/pkumdl/target\\_prediction](https://github.com/pkumdl/target_prediction).

**Keywords:** druggable protein; drug target; word2vec; deep learning

**Author summary:** In this study, we tested the performance of different combinations of protein features and machine learning algorithms, based on FDA-approved small molecules' targets, in druggable proteins prediction. Although without physicochemical features inclusion, the performance of overlapped 3-gram word2vec was better than other features. Furthermore, enlarged dataset with targets of small molecules in experiment and clinical investigation were trained but the training accuracy fell down to low level. Our study indicates the potential ability of word2vec in the prediction of druggable protein. And the training dataset of druggable protein should not be extended to targets that are lack of verification. We also find a number of potential drug targets for future research.

## INTRODUCTION

With the development of high-throughput sequencing technique, the number of sequenced proteins has been

growing steadily and exponentially [1]. And with the deeper understanding of Human Genome Project, many proteins have been identified as crucial members in disease network. However, transferring those fruitful results into clinical use is not that easy. The number of

FDA-approved drug targets have been increasing much slowly [2]. One of the most critical reasons for the failure of drug discovery pipelines is the choice of wrong protein targets [3]. Computational tools have been widely used in drug discovery process and it has reduced the time and cost comparing to experiment. For druggable protein prediction, some researchers evaluate druggability by detecting whether a protein has a proper druggable binding pocket. The existing software includes Cast-p [4], fpocket [5], PockDrug [6], and CAVITY [7]. By NMR experiment, Hajduk *et al.* identified that only a small number of structural features were enough for druggable protein prediction, such as protein surface polarity, surface complexity and a few pocket information. They achieved an accuracy of 94% by using those features [8]. However, heavily depending on the availability of three dimensional structures limited the use of these methods, since only 1% of the proteins have structures being solved [9]. Mitsopoulos *et al.* found that the network topology of druggable proteins differs significantly from undruggable proteins. Using 300 topology parameters, they successfully predict general druggable proteins and cancer druggable proteins [10]. Since whether a small molecule is a potential candidate of drug can be determined approximately and simply by a number of selected physicochemical features [11,12], we believe that druggable proteins may also be calculated in terms of simple physicochemical features or sequence composition statistics. Li *et al.* had developed a 146-d vector to represent a protein sequence. By using support vector machine (SVM) method, they achieved an accuracy of 84% on only 186 positive samples [13]. Now the number of approved drug targets has been increasing [2]. With the enlarged dataset, Jamali *et al.* applied a 443-d vector and various machine learning algorithms to determine the best model. Neuron network stood out with the best accuracy of 89.78% [14].

In this paper, we tested the performance of different combinations of protein features and machine learning algorithms in druggable proteins prediction. FDA-approved small molecules' targets were used as training

samples. Although the 146-d vector used in Ref. [13] (LQL-v) with neuron network (NN) achieved the best training accuracy of 91.10%, overlapped 3-gram word2vec with logistic regression (LR) achieved best prediction accuracy on independent test set (89.55%) and newly approved targets. Furthermore, enlarged dataset with targets of small molecules in experiment and investigation were trained. Unfortunately, the best training accuracy was only 75.89%. In addition, we applied our models to predict potential targets for references in future study. Our study indicates the potential ability of word2vec in the prediction of druggable proteins. And the training dataset of druggable proteins should not be extended to targets that are lack of verification.

## RESULTS

We tested the performance of different combinations of protein features and machine learning algorithms in druggable protein prediction. The protein features we chose are listed in Table 1. The machine learning algorithm we chose are NN, SVM, LR, Decision Tree (DT), Gradient Boost Decision Tree (GBDT), K-nearest (KN), Random Forest (RF) and Naïve Bayesian (NB).

The small dataset we used was retrieved from FDA approved small molecules' drug targets in DrugBank and was constructed by Jamali *et al.* [14,15]. The number of positive samples is 1224. A large dataset was constructed by us, in which experimental and clinical investigational small molecules' targets were included. The large dataset contains 5503 positive samples. Negative samples were screened according to the rules in Ref. [14]. For both the small and the large dataset, we sampled the number of negative samples three times. In total, three batches of 1217 negative samples for small dataset and three batches of 5498 negative samples for large dataset were constructed. For an unbiased comparison, independent and external test set were constructed individually. The details of protein features, machine learning parameters, dataset construction are described in the Materials and Methods.

**Table 1** Lists of proteins features used

Encoding method		Dimension	Ref.
Auto covariance (AC)		210	[15]
Cojoint Triad (CT)		343	[16]
LQL-v		146	[13]
Jamali-v		443	[14]
Word2vec	Single	50	[17]
	Overlapped 3-gram	200	
	Non-overlapped 3-gram	200	

LQL-v: the protein feature developed in Ref. [13]; Jamali-v: the protein feature developed in Ref. [14].

## Amino acid distributed representation

One of the most critical and basic tasks in Nature Language processing (NLP) is to learn the representative embedding of words. Some applications of NLP such as topic identifications used bags of words (BOW) model [18]. Although performed well in some tasks, its application is limited by dimension explosion and incapability of semantic representation. Continuous vector representation was developed to embed the meaning of every word in an  $n$ -dimensional space. Trained with contexts, words with similar meanings locate closely in the space. Word2vec is one of the most successful algorithms of that kind, and it has various applications [19–21].

In protein sequence representation, AC, CT, LQL-v and Jamali-v used in this work are analogies of BOW models, whose elements are determined by counting the sequences according to the invented features. Recently, word2vec has been recruited to study protein sequences [17,22,23].

We took three forms to train the word2vec model. For example, given a protein sequence:

LRQTVKNTVSQV

(1) Single, which means single amino acid with its contexts were taken to train and generate the embedding vector of every amino acid;

(2) Overlapped 3-gram, which means by breaking the original sequence into 3 window size overlapped  $k$ -mers and training with their contexts, the embedding vector of every overlapped 3-gram was obtained.

LRQ RQT QTV TVK VKN...

(3) Non-overlapped 3-gram, which means by shifting the 3 length window to generate 3 new sequences,  $k$ -mers were generated with the same method as (2). After training with their contexts, the embedding vector of every non-overlapped 3-gram was obtained.

LRQ TVK NTV SQV

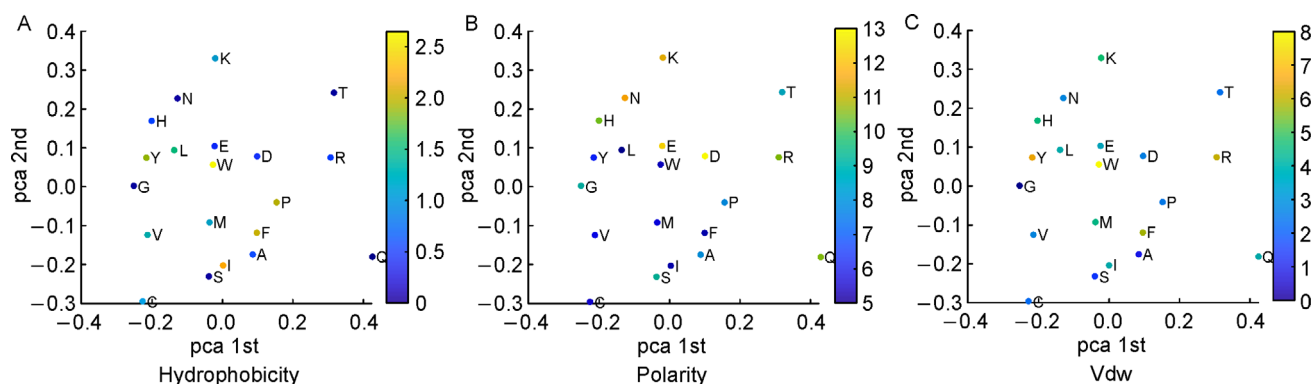
RQT VLM TVS...

QTV KNT VSQ...

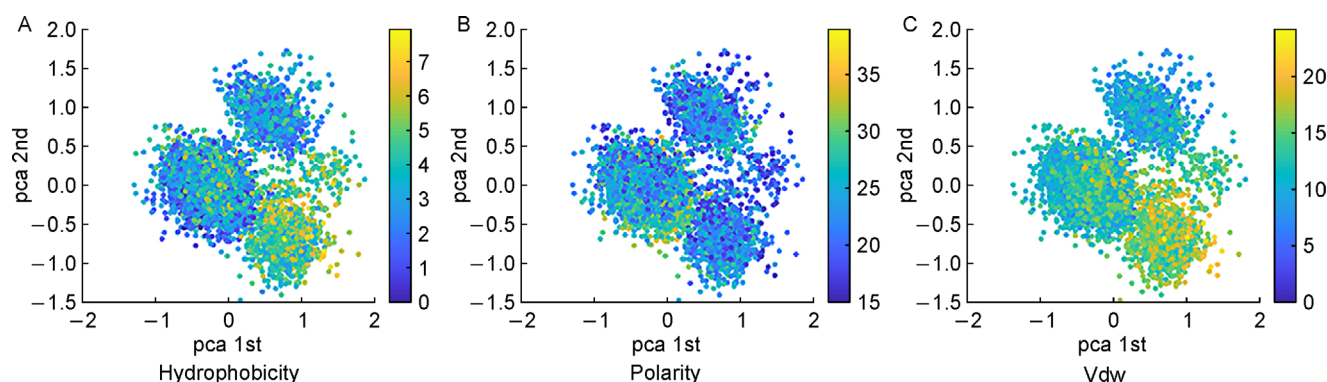
The vector space (on the training set) for single amino acid is shown in Figure 1. The color represented individually the intensity of hydrophobicity (Figure 1A), polarity (Figure 1B) and van der Waals volume (Figure 1C) of every amino acid. The vector space is different from that in Ref. [23]. It is easy to understand that this is due to the different training dataset and the different goal of the studies. But in both spaces, amino acid with similar color (property) tended to cluster together (although with some outliers).

Figures 2 and 3 are the vector spaces (on the training set) for overlapped 3-grams and non-overlapped 3-grams. Comparing the two spaces, the overlapped 3-grams one was more distributed, with three clusters formed. The distributed space also existed in Ref. [17]. The color intensity of the non-overlapped 3-grams one was more gradient distributed, which was also true but less obvious in the overlapped 3-grams one.

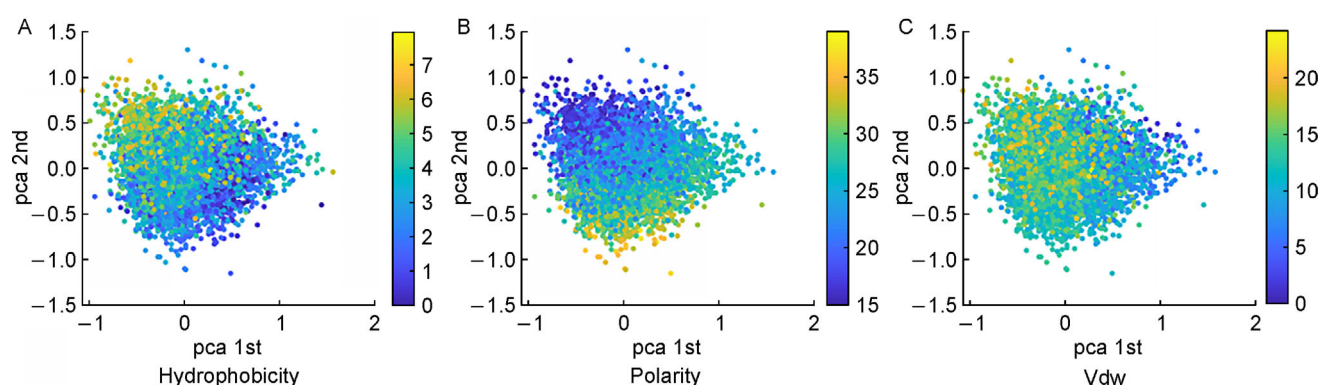
For comparison, we constructed an artificial scramble space by randomly shuffling the labels of 3-grams in overlapped and non-overlapped space. To quantitatively measure the continuity of those selected properties in protein space, for each property, we calculated the best Lipschitz constant (the minimum  $k$ ) ratio (i) between overlapped space and random space; (ii) between non-overlapped space and random space (see Methods). It is an illustration of the smoothness of a space, and reflects the richness of information. From Table 2, we could see that the ratio  $k$  between any properties in trained 3-grams space (both overlapped and non-overlapped) and random space, although not very significantly, was lower than 1, which suggested that both of the trained 3-grams spaces were smoother than random space, containing richer information. However, no significant differences could be found between overlapped space and non-overlapped space.



**Figure 1.** The Single amino acid embedding space. The color represents the intensity of hydrophobicity (A), polarity (B) and van der Waals volume (C) of every amino acid.



**Figure 2.** The overlapped 3-grams embedding space. The color represents the average intensity of hydrophobicity (A), polarity (B) and van der Waals volume (C) of the 3-grams.



**Figure 3.** The non-overlapped 3-grams embedding space. The color represents the average intensity of hydrophobicity (A), polarity (B) and van der Waals volume (C) of the 3-grams.

**Table 2** Using the best Lipschitz constant (the minimum  $k$ ) ratio to evaluate the continuity of protein space with respect to biochemical properties of 3-grams (hydrophobicity, polarity, vdw)

Property	Overlapped 3-grams space vs scramble space	Non-overlapped 3-grams space vs scramble space	Overlapped 3-grams space vs non-overlapped 3-grams space
Hydrophobicity	0.8576±0.0132	0.9780±0.0192	0.8371
Polarity	0.9323±0.0367	0.98843±0.0219	1.1414
Vdw	0.9567±0.0190	0.8979±0.0237	0.9539

### Comparison of prediction accuracies for different protein features in combination with different machine learning algorithms

We tested the performance of different combinations of protein features and machine learning algorithms in druggable protein prediction. We randomly separated 90% of the whole dataset as training set and used the other 10% as independent test set. The 5-fold cross-validation (5-CV) training results are shown in Table 3, and the results of model prediction on the independent test set are shown in Table 4. We use accuracy as a measurement for performance while the overall performance metrics could be found in the Supplementary Materials.

As for the protein features tested, LQL-v with NN achieved the best training accuracy of 91.93%. Jamali-v with NN achieved a training accuracy of 88.12%, which was comparable to the result in Ref. [14] (we had separated 10% of training samples as independent test set but Jamali *et al.* did not). The overall performance of word2vec based features were comparable to LQL-v and Jamali-v; overlapped 3-gram word2vec with NN achieved the second best training accuracy of 90.23%. Overlapped 3-gram word2vec performed better than non-overlapped 3-gram and single word2vec. The training accuracies of sequence statistic based features AC and CT were only around 60%–70%, which were far worse than the others. In summary, LQL-v, Jamali-v and word2vec performed

**Table 3 Training accuracies of the combination of protein features and machine learning algorithms on the small dataset**

	AC(%)	CT(%)	LQL-v(%)	Jamali-v(%)	Word2vec(%)		
					Single	Non-overlapped 3-gram	Overlapped 3-gram
NN	66.89±1.10	73.13±1.37	91.93±0.37	88.12±0.67	82.67±1.25	81.45±1.09	90.23±0.90
SVM	62.60±1.91	70.34±1.66	84.41±0.81	82.23±2.28	83.77±1.43	75.69±2.16	79.25±1.53
LR	63.98±1.76	73.72±1.67	88.41±1.53	86.96±1.72	80.43±1.14	82.33±1.84	87.60±1.44
GBDT	62.19±2.05	75.36±0.99	88.86±0.98	88.73±1.39	85.49±2.00	82.65±2.17	86.78±1.61
KNN	58.56±1.21	63.68±2.09	81.04±1.19	76.73±2.02	82.17±1.28	76.00±2.05	77.77±1.24
RF	55.38±1.28	68.90±2.50	82.27±0.68	82.59±2.46	82.12±1.09	77.14±2.15	81.51±1.30
DT	56.40±3.44	65.42±0.99	79.41±1.80	77.14±1.50	77.15±1.54	69.40±2.17	75.15±1.37
NB	62.45±0.93	74.74±2.24	84.32±1.13	84.05±2.51	61.07±1.86	77.01±2.91	81.10±2.01

These were the training results for 5-CV.

**Table 4 Prediction accuracies of the combination of protein features and machine learning algorithms on the independent test set**

	AC(%)	CT(%)	LQL-v(%)	Jamali-v(%)	Word2vec(%)		
					Single	Non-overlapped 3-gram	Overlapped 3-gram
NN	65.98	71.89	88.03	87.91	81.96	79.81	84.17
SVM	67.08	74.23	82.5	84.58	80.01	73.77	82.17
LR	66.05	75.66	84.17	87.5	80.9	78.69	89.55
GBDT	64.01	78.12	84.58	88.75	85.28	81.97	87.91
KNN	64.21	60.74	80.83	75	82.82	75.41	79.1
RF	56.85	69.12	81.67	81.25	82.41	79.1	83.2
DT	54.6	60.33	75.83	80.83	79.55	70.49	73.98
NB	64.03	72.19	85	85.83	58.2	74.18	83.61

comparatively well with most of the machine learning algorithms. As for the machine learning methods tested, except for NB, DT and KNN, all of the other algorithms achieved stable performance on the three best protein features analyzed. For the independent test set prediction, overlapped 3-gram word2vec with LR achieved an accuracy of 89.55%, which was the best of all combinations. We sampled negative samples and independent dataset three times and the results were approximately the same. We then used LQL-v and overlapped 3-gram word2vec individually to train the whole small dataset as our prediction models. Overlapped 3-gram word2vec performed better on newly approved targets than LQL-v (Table 5).

### The performances of training on the larger dataset

Comparing to Li *et al.*'s result, which was based on the limited number of FDA-approved small molecules' targets ten years ago, the accuracy had been increased a lot from 84% to 91.93%, with the same protein features (LQL-v) and machine learning algorithm (SVM) used. To

test whether a larger training dataset could make contribution to current results, an enlarged dataset with investigational and experimental molecules' targets, which increased the training samples by 2 fold, was built and trained by LQL-v and overlapped 3-gram word2vec, with various machine learning methods. Unfortunately, the best prediction accuracy was only 76.44% (Table 6) by LQL-v with GBDT. We then used the small dataset trained models to test the enlarged dataset with the exclusion of the training samples. 50%–70% of the investigational and experimental molecules' targets in the enlarged dataset were recognized as positive (Table 7). Of them, 1587 were predicted as positive by at least 5 models for each property (LQL-v and overlapped 3-grams), which we defined as “druggable”, 856 were predicted as negative by all of the models, which we defined as “undruggable” and the others were defined as “undefined” by us. Druggable proteins were tagged with 3256 pathway labels by Kyoto Encyclopedia of Genes and Genomes (KEGG) [24] and the top 20 pathways were listed in Table 8. From the table, we could see that some of the top tags were related to diseases pathways, such as

**Table 5** Prediction accuracies of LQL-v and overlapped 3-gram word2vec on the external test set (Newly approved targets)

	LQL-v(%)	Overlapped 3-gram (%)
NN	0.5764	0.6235
SVM	0.6099	0.7118
LR	0.6141	0.6505
GBDT	0.5862	0.6265
KNN	0.7104	0.7562
RF	0.4983	0.5222
DT	0.5429	0.5842
NB	0.6155	0.6603

metabolites, cancer, neuron diseases, infection diseases, *etc.* The whole list of the prediction of the investigational or experimental molecules' targets is supplied in the Supplementary Materials.

## DISCUSSION

In this paper, we tested the performance of different combinations of protein features and machine learning algorithms in druggable proteins prediction. As the roles of word representations play in NLP, protein features or protein codings are also critical and basic in protein function prediction. AC and CT, which are purely sequence statistic based features, have been successfully applied in protein subcellular location prediction [25], protein-protein interaction prediction [15,16], structural class [26,27], *etc.*, but failed in druggable protein prediction. The reason might be that more physicochemical features of amino acids are needed in this task. In contrast, the prediction accuracies of LQL-v and Jamali-v, which included physicochemical features, were far much better. Although word2vec are sequence based features, they performed comparably as LQL-v and Jamali-v. The reason might be that protein properties had been embedded in the vectors learned by training sequences contexts. This was partly confirmed by the best Lipschitz constant (the minimum  $k$ ) ratio. Comparing with random space, the 3-grams spaces were smoother and therefore information richer. As for the three forms of word2vec methods, overlapped 3-gram vectors distributed more widely in the space, and they achieved comparably better

prediction accuracy. It will be interesting to study the relationship between vector distribution and predication accuracy.

The number of FDA approved small molecules' targets has increased comparing with ten years ago. The same model trained with this enlarged dataset achieved an accuracy improvement of around 5%. It indicated that the model was benefit from more training samples. However, with the addition of targets of small molecules that in experiment or in investigation, which enlarged the training examples further to 2 folds, the predication accuracy was decreased to a much lower level. It might suggest that many of the targets in research were difficult for small molecules to target and could not be included in the positive training sets. This was partially indicated by that only 50%–70% of the positive samples in the enlarged dataset were recognized as druggable targets. Though the concept of drug target is ever changing as with emerging new strategies such as proteolysis-targeting chimeras (PROTAC) [28], proteins regarded as undruggable before can become druggable now, substantial part of the protein targets in experiments and clinical investigations are predicted undruggable by our models. These results might be helpful for researchers to recheck their targets in investigation.

## CONCLUSION

We tested the performance of different combinations of protein features and machine learning algorithms in druggable proteins prediction. Although the 146-d vector used by Li *et al.* (LQL-v) with neuron network achieved the best training accuracy with 91.10%, overlapped 3-gram word2vec with logistic regression achieved best prediction accuracy on independent test set (89.55%) and newly approved targets. Although not including amino acids physicochemical properties apparently, word2vec encoded features performed comparably as other carefully designed features. Furthermore, enlarged dataset with targets of small molecules in experiment and investigation was trained. We showed that model prediction accuracy could not benefit from enlarged dataset with targets of small molecules in experiment or investigation. In addition, we applied our models to predict probable targets for references in future study, and most of the predicted targets were diseases- related.

**Table 6** Training accuracies of the LQL-v, overlapped 3-gram word2vec in combination with various machine learnings on the large dataset

	NN(%)	SVM(%)	LR(%)	GBDT(%)	KNN(%)	RF(%)	DT(%)	NB(%)
LQL-V	72.47±0.99	72.80±0.50	69.95±0.61	75.48±0.47	76.44±0.60	71.46±0.89	65.59±0.60	69.95±0.61
Overlapped 3-gram	71.98±0.63	71.50±0.53	72.83±0.56	73.18±0.64	70.46±0.53	69.17±0.31	62.49±0.90	67.90±1.29

These were the training result for 5-CV

**Table 7** Prediction accuracies of the LQL-v, overlapped 3-gram word2vec in combination with various machine learnings on the positive samples in the large dataset

	LQL-v(%)	Overlapped 3-gram (%)
NN	74.33	70.25
SVM	64.67	76.9
LR	66.42	73.8
GBDT	64.19	74.24
KNN	74.33	81.46
RF	52.98	67.53
DT	60.6	67.93
NB	65.23	72.43

## MATERIALS AND METHODS

### Dataset

**Small dataset:** The small dataset was created by Jamali *et al.* The positive samples were FDA-approved small molecules' targets retrieved from DrugBank. After removal of those sequences that contained rare amino acids and those could not be coded to any of the protein features we used, there remained 1209 sequences.

**Enlarged dataset:** The enlarged dataset was created by us. The positive samples not only included the FDA-approved small molecules' targets, but also included the experimental or investigational small molecules' targets

in DrugBank and Therapeutic Target Database (TTD). A total of 5503 positive samples were included.

**Negative samples:** Negative samples were screened out by removing protein sequences from Swiss-Prot that were (i) in positive samples; (ii) in DrugBank or TTD that were classified as in experiment or in investigation; (iii) proteins that were in the same families with (i) and (ii). Details could be found in Ref. [14]. We sampled negative samples three times since (i) the number of screened negative samples was more than the number of positive samples, we need to sample them to be approximately equal; (ii) to give solid evidences that one particular combination of protein features and machine learning algorithm performed better than others.

Three batches of 1235 negative samples for small dataset and three batches of 5498 negative samples for large dataset were constructed.

**Independent dataset:** We first sampled 90% of the dataset to train the models and the rest 10% as independent test set.

**Newly approved external dataset:** We downloaded the sequences from DrugBank (<https://www.drugbank.ca>) that were tagged as "approved" (update 2018-04-02). Sequences that were the same as small dataset (training set) were excluded and 1,419 were remained.

### Protein features

**Auto-covariance (AC):** The protein sequence was transformed by the following equation:

**Table 8** KEGG pathway tags of those positive targets predicted by our models

Pathways	Number of tags	Proportion (%)
Metabolic pathways	165	5.07
Biosynthesis of secondary metabolites	66	2.03
zBiosynthesis of antibiotics	50	1.54
Pathways in cancer	50	1.54
Microbial metabolism in diverse environments	40	1.23
MAPK signaling pathway	39	1.2
PI3K-Akt signaling pathway	34	1.04
Carbon metabolism	29	0.89
Human papillomavirus infection	29	0.89
Axon guidance	26	0.8
Neuroactive ligand-receptor interaction	26	0.8
Tuberculosis	25	0.77
Alzheimer's disease	25	0.77
Cytokine-cytokine receptor interaction	25	0.77
NOD-like receptor signaling pathway	24	0.74
Kaposi's sarcoma-associated herpesvirus infection	24	0.74
Ras signaling pathway	24	0.74
Thermogenesis	23	0.71
FoxO signaling pathway	23	0.71

$$AC_{lag,j} = \frac{1}{n-lag} \sum_{i=1}^{n-lag} \left( X_{i,j} - \sum_{i=1}^n X_{i,j} \right) \\ \times \left( X_{(i+lag),j} - \sum_{i=1}^n X_{i,j} \right)$$

Where  $j$  refers to the  $j$ -th descriptor,  $i$  is the position of the protein sequence  $X$ ,  $X_{i,j}$  is the normalized  $j$ -th descriptor value for  $i$ -th amino acid,  $n$  is the length of the protein sequence  $X$ , and  $lag$  is the value of the lag. In this way, proteins with variable lengths could be coded into vectors of equal length ( $j \times lag$ ).

In this study,  $j$  was seven physicochemical properties (hydrophobicity, hydrophilicity, net charge index of side chains, polarity, polarizability, solvent accessible surface area, volume of side chains); Guo *et al.* selected a value of 30 for the  $lag$  and we also used this value. Consequently, the vector contained 210 numbers ( $7 \times 30$ ) [15].

**Conjoint Triad (CT):** First, all 20 amino acids were clustered into seven groups according to their dipole and side chain volumes. Next, each amino acid from a protein sequence was replaced by its cluster number. For example, the protein sequence:

P = MREIVHIQAG

was replaced by:

P = 3562142411

Then, a 3-amino acid window was used to slide across the whole sequence one step at a time from the N-terminus to the C-terminus.

By calculating the frequency of the combination of each three numbers:

$$\left\{ \begin{array}{l} 111=f1 \ 121=f8 \ \dots \ 177=f337 \\ 211=f2 \ 221=f9 \ \dots \ 277=f338 \\ \dots \\ 711=f7 \ 721=f14 \ \dots \ 777=f343 \end{array} \right\}$$

The protein P was represented by a vector of 343 numbers, all of which are zero except for  $f276$  (356),  $f89$  (562),  $f13$  (621),  $f149$  (214),  $f71$  (142),  $f158$  (424),  $f23$  (241), and  $f4$  (411).

**LQL-v:** The composition of 20 amino acids formed the first 20 dimensions. Then the amino acids were clustered into 3 types of amino acids for each of six physicochemical property (hydrophobicity, polarity, polarizability, solvent accessibility and normalized van der Waals volume). Composition, transition and distribution of each type and each property was calculated. For example, amino acids were clustered into polar, neutral and hydrophobic for hydrophobicity property. For “Composition”, 3 dimensions were calculated: the percentage of

polar, neutral and hydrophobic; For “Transition”, 3 dimensions were calculated: the percentage of polar transferred to neutral, neutral transferred to hydrophobicity and hydrophobicity transferred to polar; For “Distribution”, 5 dimensions were calculated for each of the 3 types of amino acids: the location percentage of the first, 25%, 50%, 75% of that type. A total of 146 dimensions vector were calculated for LQL-v.

**Jamali-v:** Three groups of protein features were calculated including 23 features representing physicochemical properties of proteins, 20 features including frequency of each amino acid in protein sequence and 400 features of frequency of dipeptides in protein sequences. A total of 443 dimension were calculated for Jamali-v.

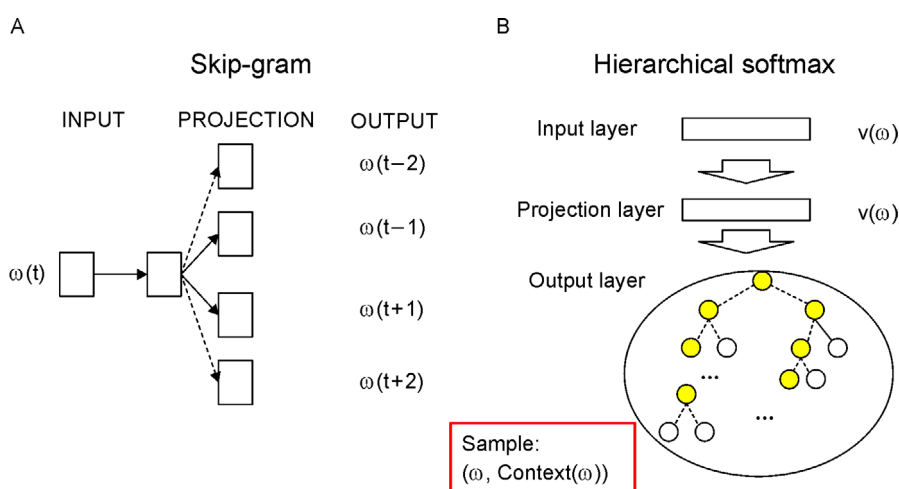
**Word2vec:** Word2vec is the name of a series of models that are trained to produce word embedding vectors. Two models are popular: continuous bag-of-words (CBOW) and continuous skip-gram. The former predicts current word from a window surrounding the context words, while the latter uses the current word to predict the surrounding window of context words. Hierarchical softmax and negative sampling are the two main training methods. The former uses a Huffman tree to maximize the conditional log-likelihood, while the latter minimize the log-likelihood of sampled negative instances. Skip-gram model with window size 8, and hierarchical softmax were recruited in this work (indicated in Figure 4). Three forms of vectors (Figures 1–3) were applied separately to train the model. For single model, the dimension was set to 15 and for 3-gram Overlapped and Non-overlapped model, the dimension was set to 200. We used word2vec program in gensim python NLP package (<https://radimrehurek.com/gensim/>) to train and compute the embedding vectors.

## Models

Models in scikit-learn (<http://scikit-learn.org/stable/>) were recruited to perform machine learning tasks. The parameters of RF, LR, DT, GBDT and NB were set as default.

**SVM:** Grid search was performed to choose parameters. For both small and large dataset, the best prediction accuracy were achieved when  $C$  equal to 1000 and gamma equal to 0.001.

**NN:** Three layers of NN were constructed, of which the first one was the input layer, the second one was the hidden layer and the last one was the output layer. Activation function for the output layer was sigmoid, and for the hidden layer was relu. RMSprop was used as optimization method. Grid search was performed to choose parameters such as numbers of neuron in hidden layer, training epoch numbers and batch sizes. The parameter sets that achieved the best prediction accuracy were listed in Table 9.



**Figure 4.** Schematic indication of Skip-gram model (A) and Hierarchical softmax model (B).

**Table 9** The parameter sets that achieved the best prediction accuracy for NN

	Hidden layer neuron	Epoch	Batch size
ac	110	100	5
ta	150	100	5
LQL-v	90	100	5
Jamali-v	150	120	10
single	170	160	45
nol	110	60	5
ol	70	100	5

## SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at <https://doi.org/10.1007/s40484-018-0157-2>.

## ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science and Technology of China (No. 2016YFA0502303) and the National Natural Science Foundation of China (Nos. 21673010 and 81273436). The authors would like to thank Youjun Xu, Shuaishi Gao, Qiwan Hu for discussion and advices.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Tanlin Sun, Luhua Lai and Jianfeng Pei declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

## REFERENCES

1. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45, D158–D169
2. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46, D1074–D1082
3. Butcher, S. P. (2003) Target discovery and validation in the post-genomic era. *Neurochem. Res.*, 28, 367–371
4. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. and Liang, J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, 34, W116–W118
5. Schmidtke, P., Le Guilloux, V., Maupetit, J. and Tufféry, P. (2010) fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.*, 38, W582–W589
6. Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L. and Camproux, A.-C. (2015) PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res.*, 43, W436–W442
7. Yuan, Y., Pei, J. and Lai, L. (2013) Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.*, 19, 2326–2333
8. Hajduk, P. J., Huth, J. R. and Fesik, S. W. (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, 48, 2518–2525
9. Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Costanzo, L. D., Duarte, J. M., Dutta, S. and Feng,

- Z. (2016) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, 45, D271–D281
10. Mitsopoulos, C., Schierz, A. C., Workman, P. and Al-Lazikani, B. (2015) Distinctive behaviors of druggable proteins in cellular networks. *PLoS Comput. Biol.*, 11, e1004597
11. Lipinski, C. A. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.*, 1, 337–341
12. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. and Hopkins, A. L. (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.*, 4, 90–98
13. Li, Q. and Lai, L. (2007) Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics*, 8, 353
14. Jamali, A. A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R. and Ebrahimi, E. (2016) DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discov. Today*, 21, 718–724
15. Guo, Y., Yu, L., Wen, Z. and Li, M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.*, 36, 3025–3030
16. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, 104, 4337–4341
17. Asgari, E. and Mofrad, M. R. (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, 10, e0141287
18. Wallach, H. M. (2006) Topic modeling: beyond bag-of-words. In *ICML '06 Proceedings of the 23rd International Conference on Machine learning*. pp. 977–984, Pittsburgh
19. Xue, B., Fu, C. and Shaobin, Z. (2014) A study on sentiment computing and classification of sina weibo with word2vec. In 2014 IEEE International Congress on Big Data. pp. 358–363. Anchorage
20. Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y. and Lee, L.-S. (2016) Audio word2vec: unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *arXiv*, 1603.00982
21. Ngo, D. L., Yamamoto, N., Tran, V. A., Nguyen, N. G., Phan, D., Lumbanraja, F. R., Kubo, M. and Satou, K. (2016) Application of word embedding to drug repositioning. *J. Biomed. Sci. Eng.*, 9, 7–16
22. Kimothi, D., Soni, A., Biyani, P. and Hogan, J. M. (2016) Distributed Representations for Biological Sequence Analysis. *arXiv:1608.05949*
23. Vang, Y. S. and Xie, X. (2017) HLA class I binding prediction via convolutional neural networks. *Bioinformatics*, 33, 2658–2665
24. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28, 27–30
25. Zeng, Y. H., Guo, Y. Z., Xiao, R. Q., Yang, L., Yu, L. Z. and Li, M. L. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.*, 259, 366–372
26. Liu, T., Geng, X., Zheng, X., Li, R. and Wang, J. (2012) Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. *Amino Acids*, 42, 2243–2249
27. Wang, Y.-C., Wang, X.-B., Yang, Z.-X. and Deng, N.-Y. (2010) Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept. Lett.*, 17, 1441–1449
28. Ottis, P., Toure, M., Cromm, P. M., Ko, E., Gustafson, J. L. and Crews, C. M. (2017) Assessing different E3 ligases for small molecule induced protein ubiquitination and degradation. *ACS Chem. Biol.*, 12, 2570–2578