A chemical language based approach for the predictionof protein-ligand binding affinity

Hakime Öztürk[1] , Arzucan Özgür[1] and Elif Ozkirimli[2]

Departments of { **1** Computer Engineering, **2** Chemical Engineering } Boğaziçi University, İstanbul, Turkey

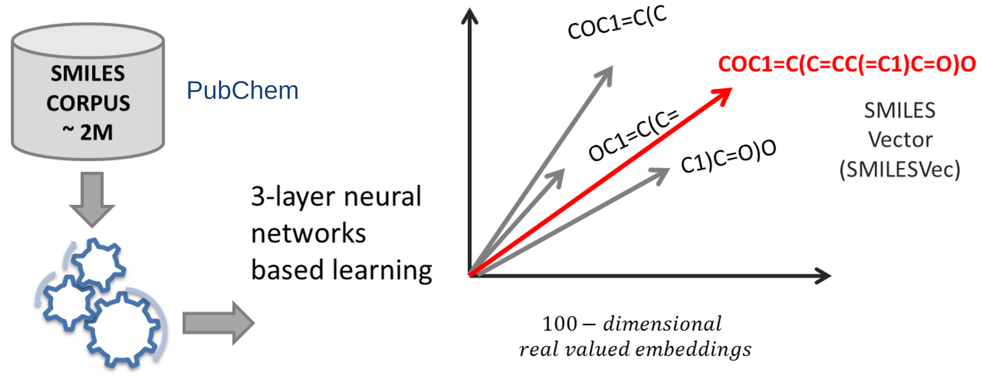
1. ABSTRACT

Identification of high affinity drug-target interactions (DTA) is a major research question in drug discovery. In this study, we propose a novel methodology to predict drug-target binding affinity using only ligand SMILES information.

We represent proteins using the word-embeddings of the SMILES representations of their strong binding ligands. Each SMILES is represented in the form of a set of chemical words and a protein is described by the set of chemical words. We then utilize the XGBoost algorithm to predict protein - drug binding affinities for two benchmark datasets.

2. METHODS

SMILESVec: Distributed ligand representation [1]



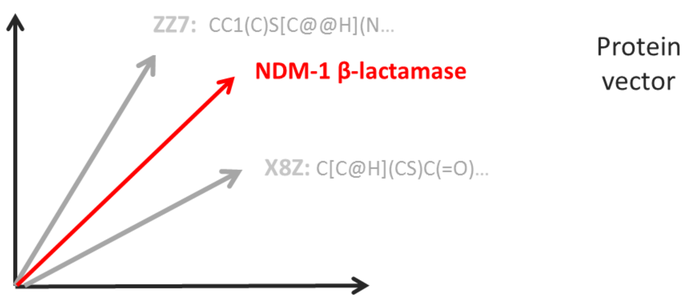
$$\text{SMILESVec} = vector(ligand) = \frac{\sum_{k=1}^{n} vector(word_k)}{n}$$

$n, number\ of\ chemical\ words$

SMILESVec-based protein representation

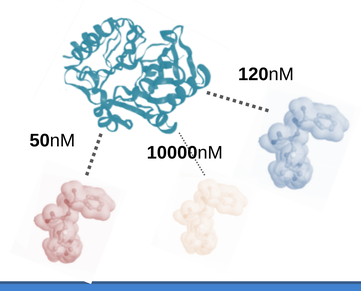
protein: NDM-1 β-lactamase

interacting ligands: ZZ7 (ampicillin), X8Z (L-captopril)

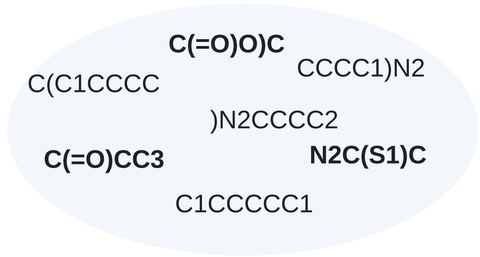


We suggested two important modification to SMILES-based protein representation to enhance drug-target binding affinity prediction performance:

1. Considering only strong-bindings ligands



2. Using Term Frequency – Inverse Document Frequency (TF-IDF) weighting



3. RESULTS

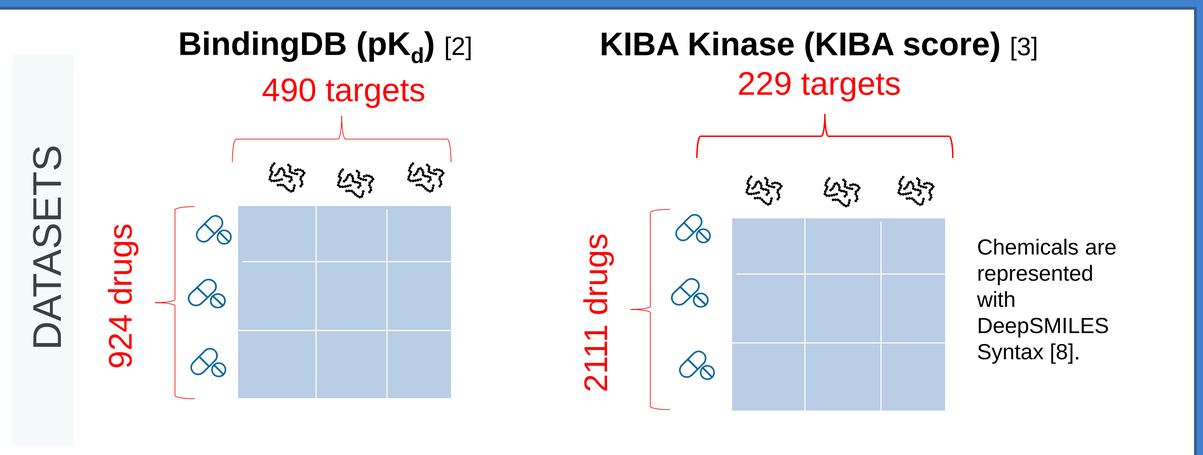


Table 1. DTA performance comparison for BindingDB dataset (5 fold cv)

Method	Protein	Drug	CI	MSE
KronRLS [4]	S-W	Pubchem	0.814	0.939
SimBoost [5]	S-W	Pubchem	0.853	0.485
DeepDTA [6]	CNN	CNN	0.873	0.409
XGBoost [7]	DeepSMILESVec * TFIDF	DeepSMILESVec	0.857	0.482

Table 2. DTA performance comparison for KIBA dataset (5 fold cv)

Method	Protein	Drug	CI	MSE
KronRLS [4]	S-W	Pubchem	0.782	0.411
SimBoost [5]	S-W	Pubchem	0.836	0.222
DeepDTA [6]	CNN	CNN	0.863	0.194
XGBoost [7]	DeepSMILESVec * TFIDF	DeepSMILESVec	0.833	0.230

4. CONCLUSION

- With this study, we proposed a novel approach to predict drug-target binding affinity by representing proteins with the chemical words of their high affinity ligands.
- We were able to predict drug-target binding affinity using only SMILES strings without using any protein sequence or structure information.
- As expected, using only the high affinity ligands in the protein representation yielded a better performance than using all available or tested ligands.

5. References

[1] Öztürk, H. *et al.* (2018a). *Bioinformatics*
[2] Liu, Tiqing, et al. (2006). *Nucleic acids research*
[3] Tang, J. *et al.* (2014). *JCIM*
[4] Pahikkala, T. *et al.* (2014). *Briefings in bioinformatics*
[5] [7] He, T. *et al.* (2017). *Journal of cheminformatics*.
[6] Öztürk, H. *et al.* (2018b). *Bioinformatics*
[7] Tianqi & Guestrin.(2016). ACM
[8] O'Boyle & Dalke (2018). ChemRxiv
[9] Mikolov, T. *et al.* (2013). *ANIPS*

Travel funding to ISMB/ECCB 2019 was generously provided by **ISCB**.

TUBITAK –BIDEB 2211 is gratefully acknowledged.