ChemBoost: A chemical language based approach for protein - ligand binding affinity prediction

Riza Özçelik[+,1] Hakime Öztürk[+,1] Arzucan Özgür[*,1] Elif Ozkirimli[*,2,3]

[1] Computer Engineering, Boğaziçi University, [2] Chemical Engineering, Boğaziçi University

[3] Data and Analytics Chapter, Roche AG, Switzerland, [+] Equal contribution.

Motivation

- **Problem:** Drugs take more than 10 years and millions of dollars to design from scratch.
  ↪ **Solution *(Partial)*:** Using known drugs on different targets, since it is more time- and budget-efficient.
  ↪ Which drugs (≈2650 approved drugs in DrugBank) on which targets (≈20K reviewed human proteins in UniProt; ≈53M drug - target pairs).
- **Goal:** Designing high-throughput methods to guide early-stage drug discovery.

Methods

Ligand Representation:

- SMILES is a codified language for ligands
- We consider ligands as documents and identify their words with k-mer and BPE.
- We learn distributed word vectors with Word2Vec. [1]

Method	Words
k-mer (i.e. 8-mer)	COc1cc2C, Oc1cc2CC, ..., 3)c2cc1C,)c2cc1Cl
BPE	COc1cc2, CCN=C(, c3ccc(Cl)c(Cl)c3), c2cc1Cl

Protein Representation:

- We experiment with normalized Smith Waterman score [2], ProtVec [3], and ligand-centric representations.
- **Ligand-centric:** We represent a protein with the vectors of its known ligands.
  ↪ We experiment with all and high-affinity ligands of a protein in the training set and also incorporate an external database.

We use **XGBoost**[4] as the prediction model and 5-fold cross validation for hyper-parameter tuning. We train our models on BDB and KIBA data sets.

Results

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (S1)	SW	-	0.687 (0.002)	1.037 (0.006)	0.683 (0.000)	0.585 (0.000)
Model (S2)	-	SMILESVec (8-mer)	0.773 (0.002)	0.876 (0.005)	0.699 (0.000)	0.425 (0.001)
Model (R1)	SW	Random	0.859 (0.002)	0.512 (0.005)	0.803 (0.001)	0.276 (0.002)
Model (R2)	Random	SMILESVec(8-mer)	0.849 (0.002)	0.537 (0.009)	0.815 (0.001)	0.258 (0.002)
Model (1)	SW	SMILESVec (8-mer)	0.873 (0.001)	0.439 (0.008)	0.837 (0.001)	0.203 (0.002)
Model (2)	ProtVec	SMILESVec (8-mer)	0.854 (0.002)	0.512 (0.004)	0.818 (0.001)	0.244 (0.001)
Model (3)	ProtVec	SMILESVec (BPE)	0.849 (0.002)	0.548 (0.008)	0.814 (0.001)	0.252 (0.002)
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847 (0.001)	0.524 (0.006)	0.823 (0.001)	0.243 (0.003)
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845 (0.002)	0.478 (0.005)	0.829 (0.001)	0.221 (0.001)
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842 (0.001)	0.497 (0.007)	0.825 (0.001)	0.227 (0.001)
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856 (0.001)	0.454 (0.007)	0.829 (0.001)	0.223 (0.001)
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873 (0.001)	0.420 (0.004)	0.837 (0.001)	0.206 (0.001)
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871 (0.002)	0.420 (0.007)	0.836 (0.001)	0.207 (0.002)

Table 1: CI and MSE scores of ChemBoost models on BDB and KIBA. Each model is trained with 5 different training sets and test set performance is measured for each trained model. Mean test set performance values and the standard deviations (in parenthesis) are reported.

	BDB Scores		KIBA Scores	
Model	CI	MSE	CI	MSE
KronRLS	0.814 (0.002)	0.939 (0.004)	0.782 (0.001)	0.411
SimBoost	0.853 (0.003)	0.485 (0.043)	0.836 (0.001)	0.223 (0.003)
DeepDTA	0.863 (0.007)	0.397 (0.011)	0.846 (0.002)	0.215 (0.005)
ChemBoost	0.871 (0.002)	0.420 (0.007)	0.836 (0.001)	0.207 (0.002)

Table 2: CI and MSE scores of the state of the art affinity prediction models and ChemBoost on BDB and KIBA. Here ChemBoost refers to the model in which the SMILESVec of a protein is obtained through the SMILES representations of its high affinity ligands and SW scores (Model (9)).

Discussion I

- 8-mer embeddings are superior to BPE
- SW is a more powerful representation for KIBA, a data set of kinases, than BDB
- SW is superior to ProtVec
- High affinity ligands yield stronger protein representations than all known ligands
- Incorporating an external database strengthens ligand-centric representations
- Ligand-centric representations have merits
- The performance of ChemBoost is higher than SimBoost and KronRLS and on par with DeepDTA

References

[1] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. A novel methodology on distributed representations of proteins using their interacting ligands. *Bioinformatics*, 34(13):i295–i303, 2018.

[2] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[3] Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Discussion II

We investigated the performance of ChemBoost models as a function of protein sequence similarity. For each protein-ligand pair (P-L) in the test set, we computed the normalized S-W similarity score of P to the other interacting proteins of L in the training set. Then, we calculated the maximum score, which we refer to as Maximum Sequence Similarity (MSS_{PL}), for a P-L pair. We formulate MSS_{PLL} as:

$$MSS_{PL} = max\{SW(P,p) \forall p \in P(L)\}$$

where $P(L)$ the set of proteins with a reported affinity with ligand L in the training set.

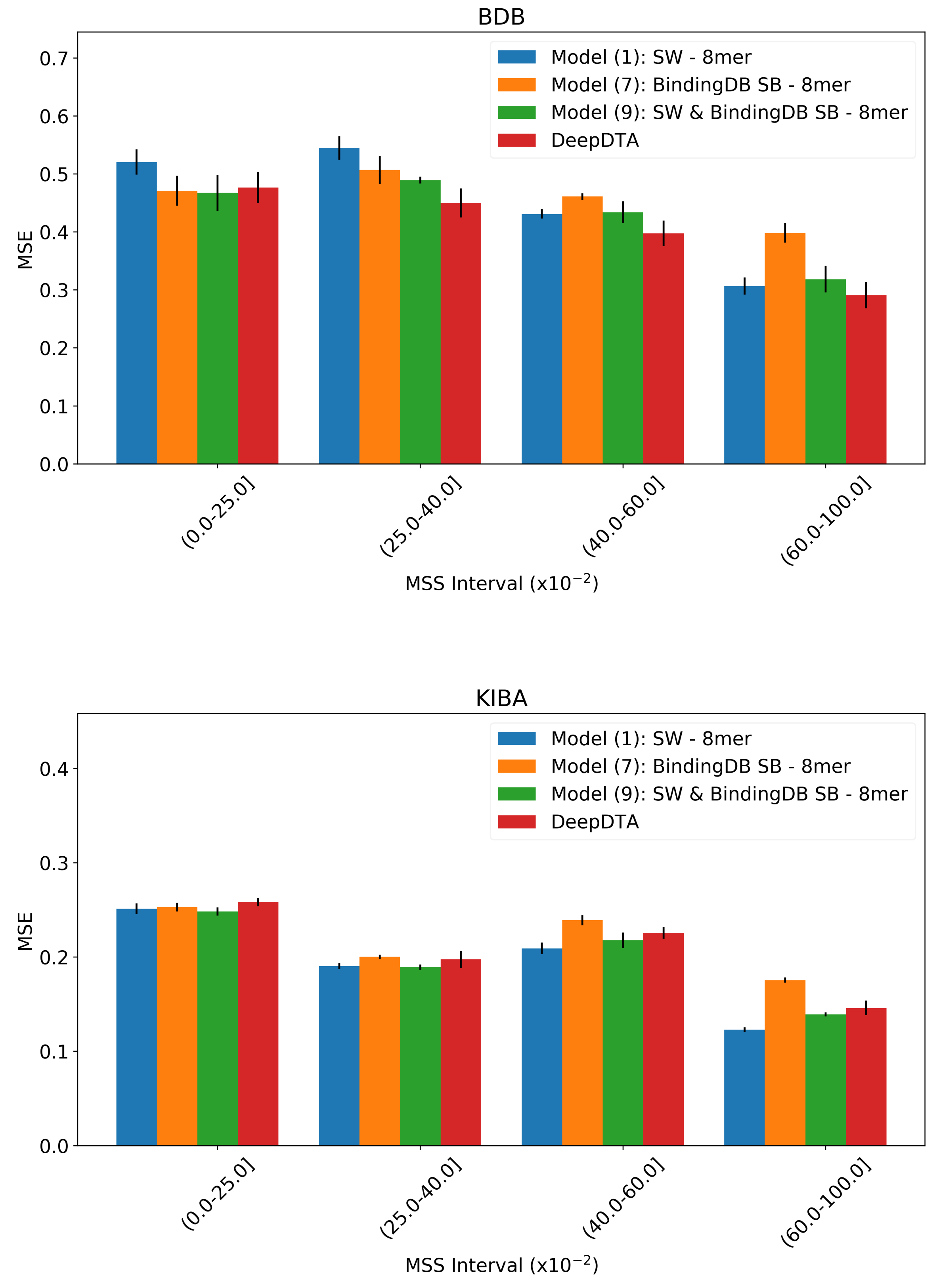

Figure 1: Test set performance of ChemBoost models and DeepDTA on BDB (top) and KIBA (bottom) with respect to MSS of interactions.