

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

Rıza Özçelik⁺, Hakime Öztürk⁺, Arzucan Özgür, Elif Ozkirimli
Boğaziçi University

⁺Equal contribution

October 22, 2020



Overview

- 1 Why We Care
- 2 ChemBoost
- 3 How ChemBoost Performs
- 4 Conclusion

Why Predict Affinities?

- **Problem:** Drugs take more than 10 years and millions of dollars to design from scratch.

Why Predict Affinities?

- **Problem:** Drugs take more than 10 years and millions of dollars to design from scratch.
- **“Poor Man’s” Solution:** To use known drugs on different targets \Rightarrow more time- and budget-efficient.

Why Predict Affinities?

- **Problem:** Drugs take more than 10 years and millions of dollars to design from scratch.
- **“Poor Man’s” Solution:** To use known drugs on different targets \Rightarrow more time- and budget-efficient.
 \hookrightarrow Which drugs? Which targets? $\Rightarrow \approx 53\text{M}$ approved drug - human target pairs.

Why Predict Affinities?

- **Problem:** Drugs take more than 10 years and millions of dollars to design from scratch.
- **“Poor Man’s” Solution:** To use known drugs on different targets \Rightarrow more time- and budget-efficient.
 \hookrightarrow Which drugs? Which targets? $\Rightarrow \approx 53\text{M}$ approved drug - human target pairs.
- **Motivation:** In order to narrow down the search space in early-stage drug discovery, we need high-throughput methods

- SMILES

- ↔ Simple and informative
- ↔ Available for all compounds
- ↔ Allows applying recent Natural Language Processing methodologies

- SMILES
 - ↪ Simple and informative
 - ↪ Available for all compounds
 - ↪ Allows applying recent Natural Language Processing methodologies
- **Chemical Language:** Ligands are documents with hidden words
 - ↪ Use 8-mers (i.e. 8-character sequences) and byte-pair encoding tokens

- SMILES
 - ↪ Simple and informative
 - ↪ Available for all compounds
 - ↪ Allows applying recent Natural Language Processing methodologies
- **Chemical Language:** Ligands are documents with hidden words
 - ↪ Use 8-mers (i.e. 8-character sequences) and byte-pair encoding tokens
- Learn chemical word vectors with semantic relations in the space (SMILESVec) [1]

- SMILES
 - ↪ Simple and informative
 - ↪ Available for all compounds
 - ↪ Allows applying recent Natural Language Processing methodologies
- **Chemical Language:** Ligands are documents with hidden words
 - ↪ Use 8-mers (i.e. 8-character sequences) and byte-pair encoding tokens
- Learn chemical word vectors with semantic relations in the space (SMILESVec) [1]
- Predict affinities based on chemical word vectors ⇒ **ChemBoost**

- Normalized Smith Waterman score [2]

ChemBoost – Protein Representation

- Normalized Smith Waterman score [2]
- ProtVec: Word2Vec on protein sequences [3]







- Normalized Smith Waterman score [2]
- ProtVec: Word2Vec on protein sequences [3]
- **Ligand-centric:** Average the chemical word vectors of known ligands of a protein. We experiment with:
 - ↪ using all ligands,
 - ↪ using only high-affinity ligands,
 - ↪ incorporating an external database
- Sequence similarity \nRightarrow functional similarity

- Use **XGBoost** [4] to predict affinities on BDB and KIBA data sets

- Use **XGBoost** [4] to predict affinities on BDB and KIBA data sets
- **BDB**: A diverse data set of 490 targets, 924 ligands, and $\approx 31\text{K}$ interactions
- **KIBA [5]**: A kinase data set of 229 targets, 2111 ligands, $\approx 118\text{K}$ interaction
- 5-fold cross-validation
- Evaluate with mean squared error and concordance index

Table: Mean CI and MSE scores of a subset of ChemBoost models on BDB and KIBA.¹

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

¹Standard deviations are also available in the manuscript      

8-mers outperform BPE tokens

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

High affinity ligands of a protein are more informative than all known ligands

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

Incorporating an external database helps for BDB

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

Hybrid models are more reliable than single-representation

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

ChemBoost vs SOTA

Table: CI and MSE scores of the state of the art affinity prediction models and ChemBoost on BDB and KIBA. Here ChemBoost refers to Model (9) in the previous table

Model	BDB Scores		KIBA Scores	
	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)

ChemBoost is on par with DeepDTA and outperforms KronRLS and SimBoost

Why are Hybrid Models more Robust?

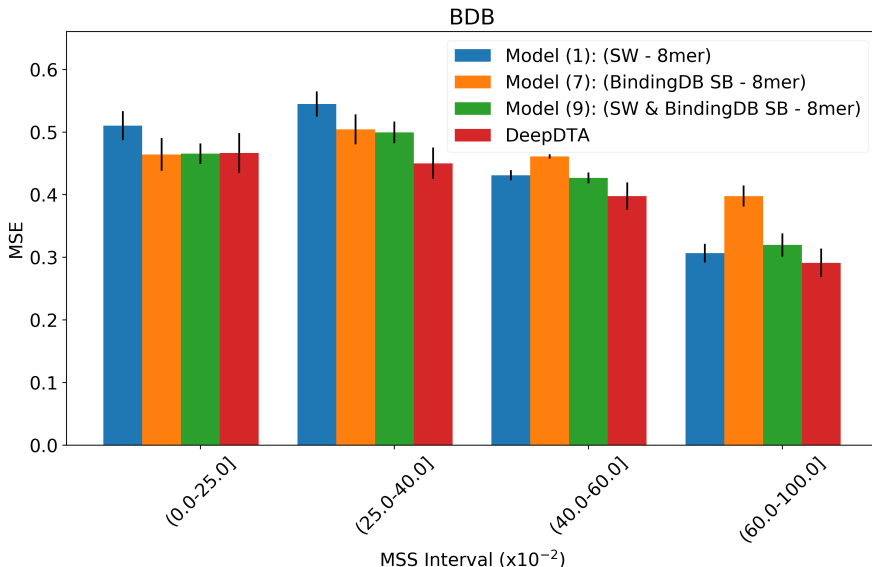
- Performance as a function of protein sequence similarity
- For all pair P-L in test set, we compute MSS_{PL}

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

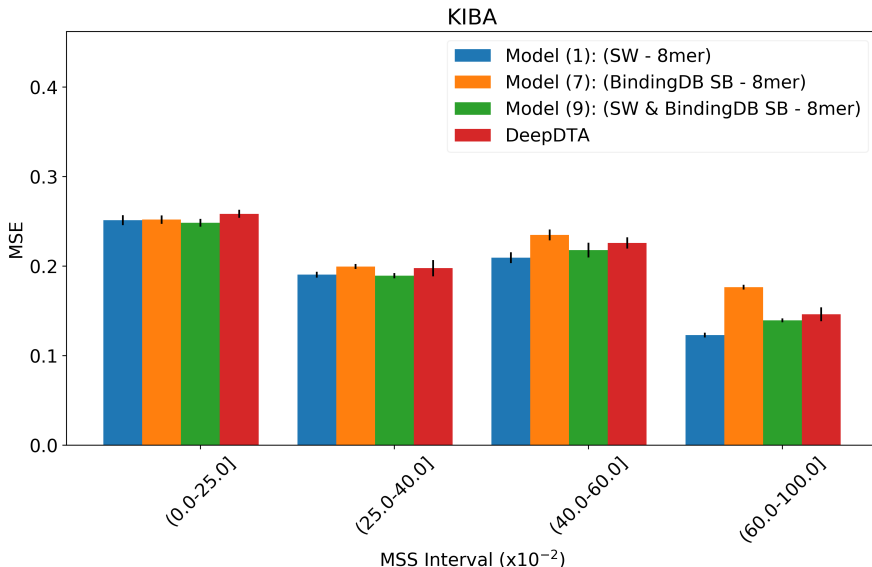
$P(L)$ is the set of proteins with a reported affinity with ligand L in the training set

- Low MSS \Rightarrow The model is unaware of a sequence-wise similar target to the ligand L.

ChemBoost is more Robust to Low Protein Similarity



ChemBoost is more Robust to Low Protein Similarity



Conclusion

- 8-mers > BPE tokens

Conclusion

- 8-mers > BPE tokens
- High affinity ligands > all known ligands

Conclusion

- 8-mers > BPE tokens
- High affinity ligands > all known ligands
- More ligands per protein is an advantage

Conclusion

- 8-mers > BPE tokens
- High affinity ligands > all known ligands
- More ligands per protein is an advantage
- Ligand-centricity \Rightarrow Robustness to low sequence similarity



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.



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.

 Ehsaneddin Asgari and Mohammad RK Mofrad.

Continuous distributed representation of biological sequences for deep proteomics and genomics.

PloS one, 10(11):e0141287, 2015.

 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.



Jing Tang, Agnieszka Szwejda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio.

Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis.

Journal of Chemical Information and Modeling, 54(3):735–743, 2014.

Thank you for listening!

Paper: <https://arxiv.org/abs/1811.00761>



Code: <https://github.com/boun-tabi/chemboost>

