# DebiasedDTA: Model Debiasing to Boost Drug-Target Affinity Prediction

Rıza Özçelik[1], Alperen Bağ[1,+] Berk Atıl[1,+], Arzucan Özgür[1], Elif Özkırımlı[2]
[1]Department of Computer Engineering, Bogazici University
[2]Data and Analytics Chapter, Pharma International Informatics, F. Hoffmann-La Roche AG, Switzerland
[+] Equal Contribution

## Motivation

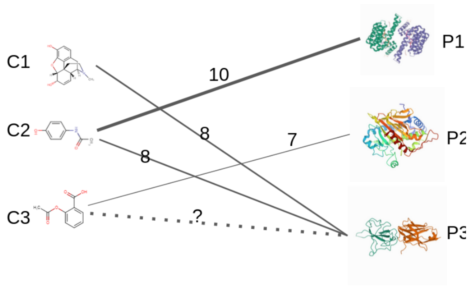


Figure 1: Warm Biomolecules

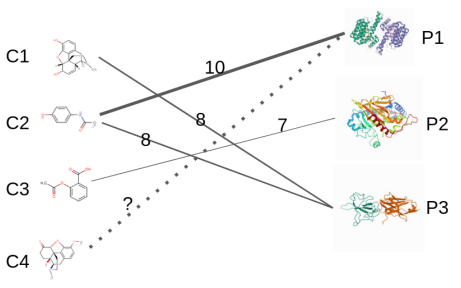


Figure 2: Cold Ligands

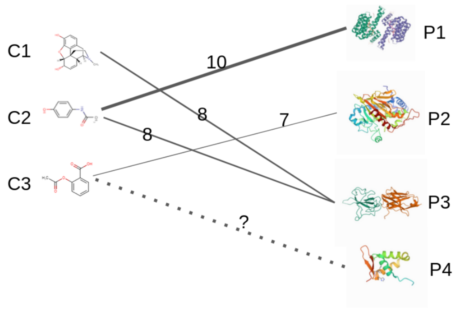


Figure 3: Cold Proteins

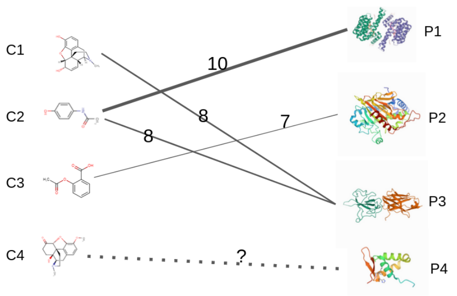


Figure 4: Cold Biomolecules

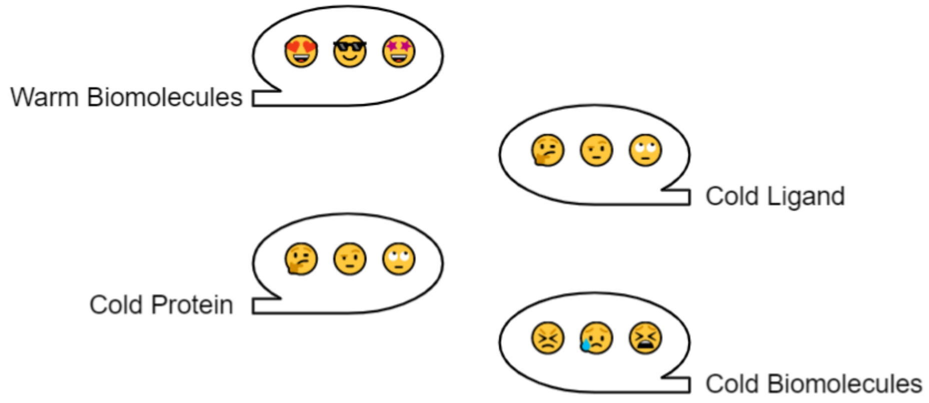


Figure 5: Machine Learning Models on Each Case

## Bias in Affinity Prediction

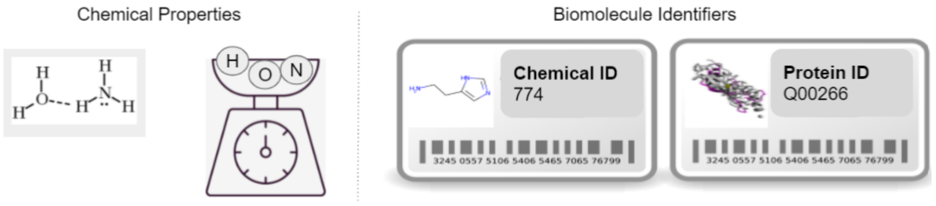
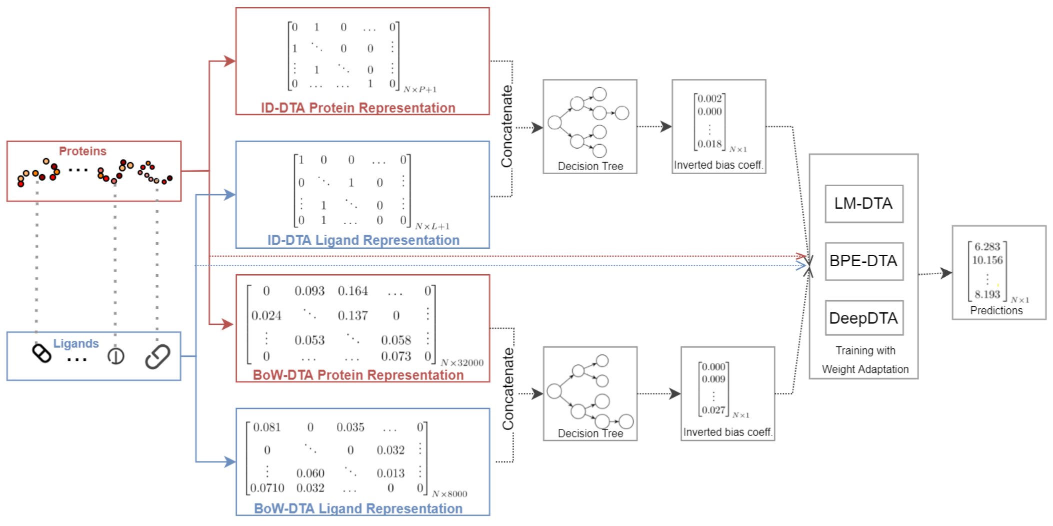


Figure 6: Example Biases in the Datasets [1, 3]

## DebiasedDTA



## Weak Learners

- ID-DTA
  - In order to avoid “chemical identifier” biases
  - Biomolecules are represented with one-hot encoding
- BoW-DTA
  - In order to avoid “chemical word” biases
  - Biomolecules are represented with bag-of-biomolecule-words representation

## Strong Learners

- DeepDTA [4]
  - Character-level convolutions over SMILES and amino-acid sequences
- BPE-DTA
  - Convolutions over Byte-Pair-Encoding (BPE) [2] tokens of SMILES and amino-acid sequences
- LM-DTA
  - Pre-trained language-model embeddings

## Results

| | | Warm | | Cold Ligand | | Cold Protein | | Cold Both | |
|---|---|---|---|---|---|---|---|---|---|
| | Model | CI | $R^2$ | CI | $R^2$ | CI | $R^2$ | CI | $R^2$ |
| BDB | DeepDTA | 1.239% | 0.023 | 4.076% | 0.004 | 2.899% | 0.042 | 10.289% | 0.062 |
| | BPE-DTA | 0.906% | 0.007 | 5.327% | 0.098 | 6.891% | 0.325 | 8.812% | 0.108 |
| | LM-DTA | 0.913% | 0.017 | 1.890% | 0.043 | 0.513% | 0.011 | 2.448% | 0.044 |
| KIBA | DeepDTA | 1.718% | 0.019 | 1.062% | 0.013 | 0.834% | 0.003 | 0.917% | -0.003 |
| | BPE-DTA | 1.362% | 0.017 | 1.088% | 0.004 | 0.588% | -0.006 | 0.000% | -0.031 |
| | LM-DTA | 0.816% | 0.013 | 1.602% | 0.032 | 0.842% | 0.019 | 2.154% | 0.052 |

Table 1: The percentile improvement in CI and absolute increase in $R^2$. The statistics are computed by comparing the best DebiasedDTA score with the non-debiased counterpart. Negative statistics are reported if the non-debiased model outperform every DebiasedDTA model.

## Conclusions

- To the best of our knowledge, DebiasedDTA is the first model debiasing approach to boost drug-target affinity prediction performance.
- DebiasedDTA can improve affinity prediction performance both on known and novel biomolecules.
- DebiasedDTA can boost drug-target affinity prediction models of different architectures.
- DebiasedDTA is applicable to almost every prediction model.

## References

[1] R. Özçelik, H. Öztürk, A. Özgür, and E. Ozkirimli. Chemboost: A chemical language based approach for protein – ligand binding affinity prediction. *Molecular Informatics*, 40(5):2000212, 2021. doi: https://doi.org/10.1002/minf.202000212. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202000212.

[2] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

[3] V. Sundar and L. Colwell. The effect of debiasing protein–ligand binding data on generalization. *Journal of Chemical Information and Modeling*, 60(1):56–62, 2019.

[4] H. Öztürk, A. Özgür, and E. Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 09 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty593. URL https://doi.org/10.1093/bioinformatics/bty593.

## Acknowledgements

This work was supported by The Scientific and Technological Research Council of Turkey [Grant number 119E133; 2211-A to R.O.; 2247-C to B.A.]; and Turkish Science Academy [TUBA-GEBIP award to A.O.].