Ensemble Learning for Novel Drug - Target Affinity Prediction

Rıza Özçelik[1], Alperen Bağ[1,+], Berk Atıl[1,+], Arzucan Özgür[1], Elif Özkırımlı[2]

[1] Dept. of Computer Engineering, Boğaziçi University, Istanbul, Turkey.

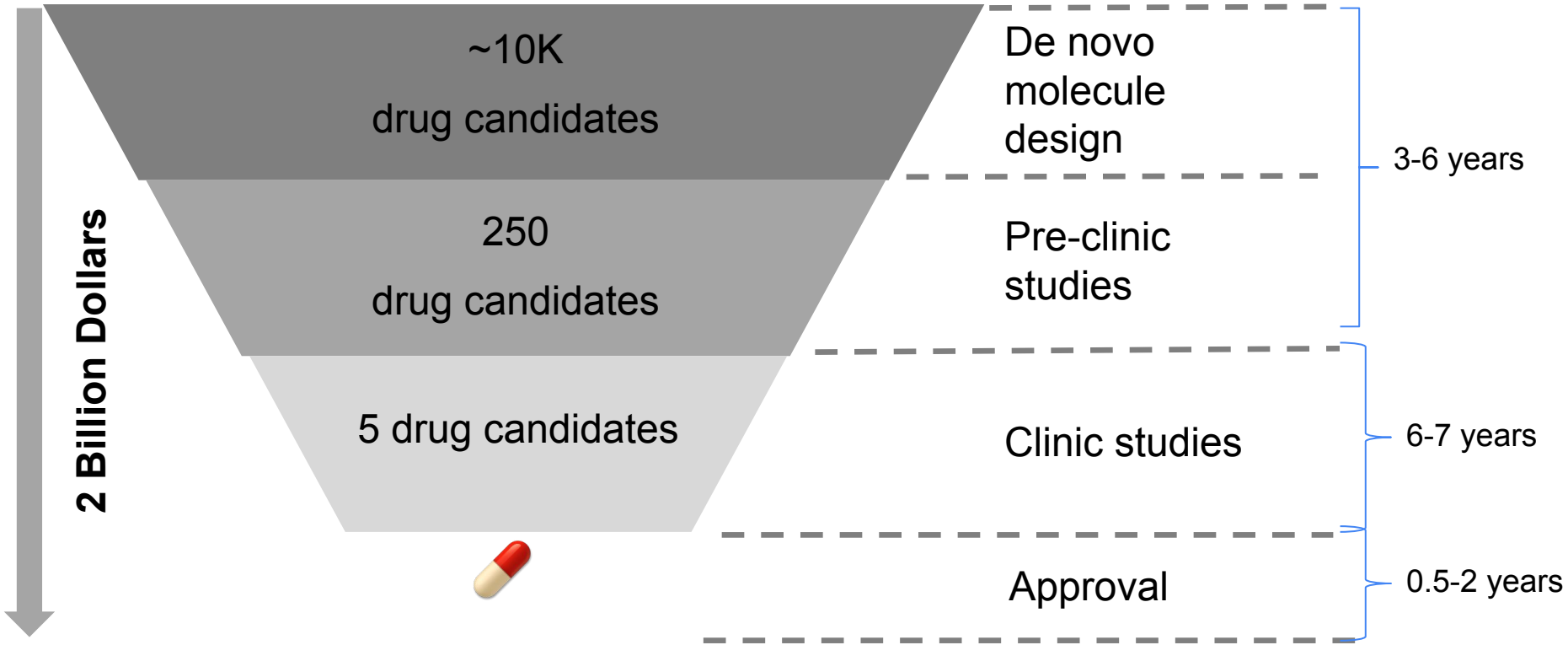
[2] Data and Analytics Chapter, Pharma International Informatics, F. Hoffmann-La Roche AG, Switzerland

[+] Equal Contribution

ISMB/ECCB 2021

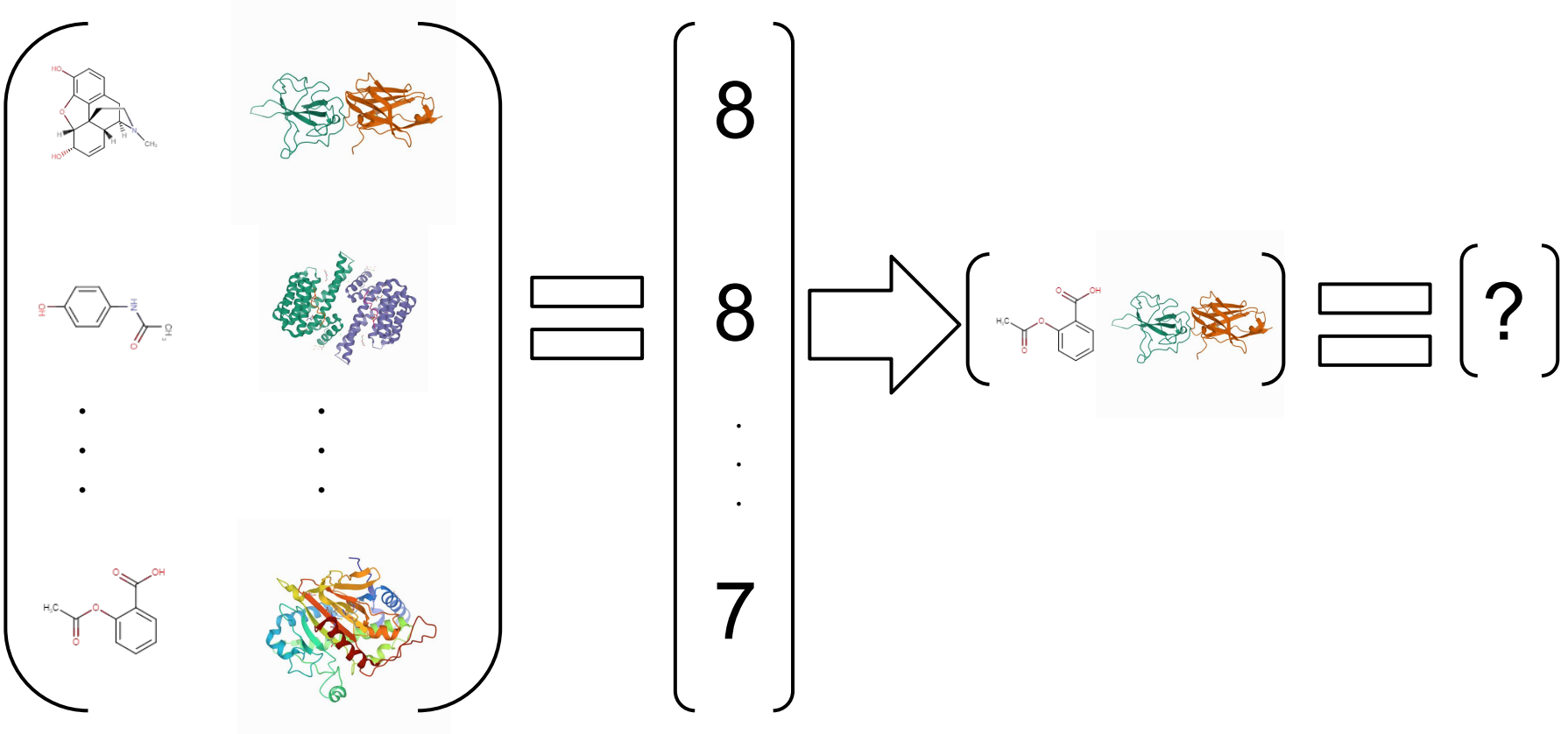
July 29

Drug Discovery is a Long and Expensive Process

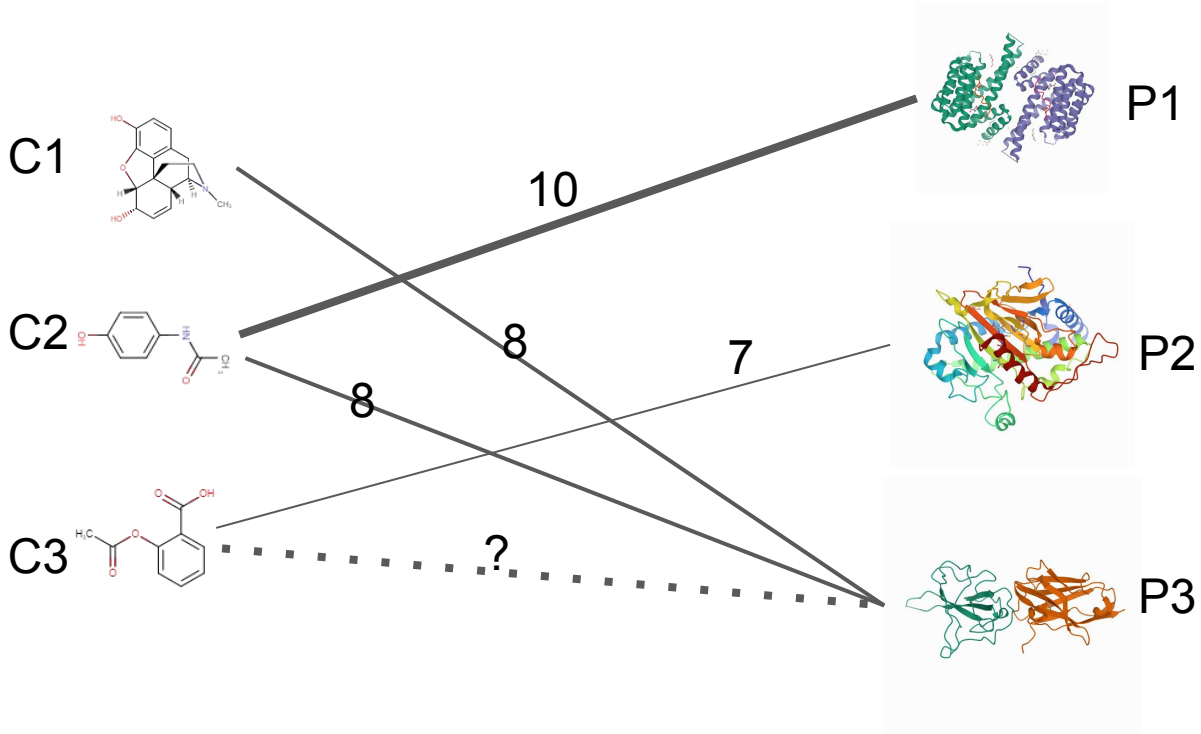


http://www.ogjc.osaka-gu.ac.jp/php/nakagawa/TRIZ/eTRIZ/

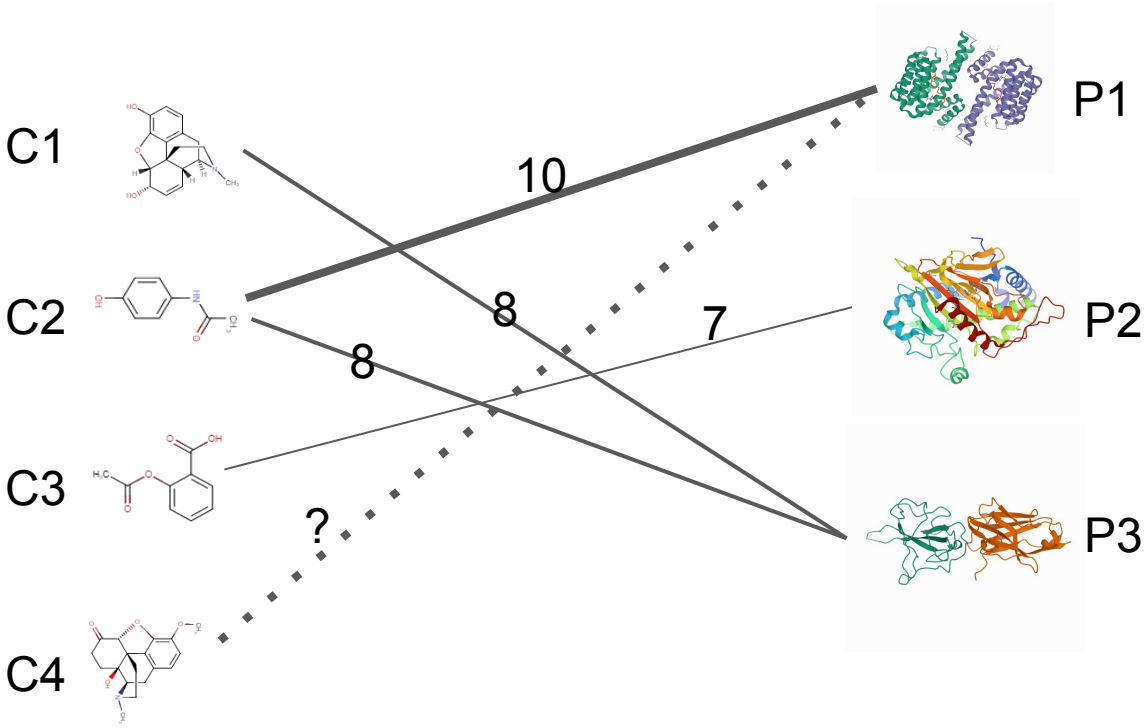
Machine Learning for Affinity Prediction



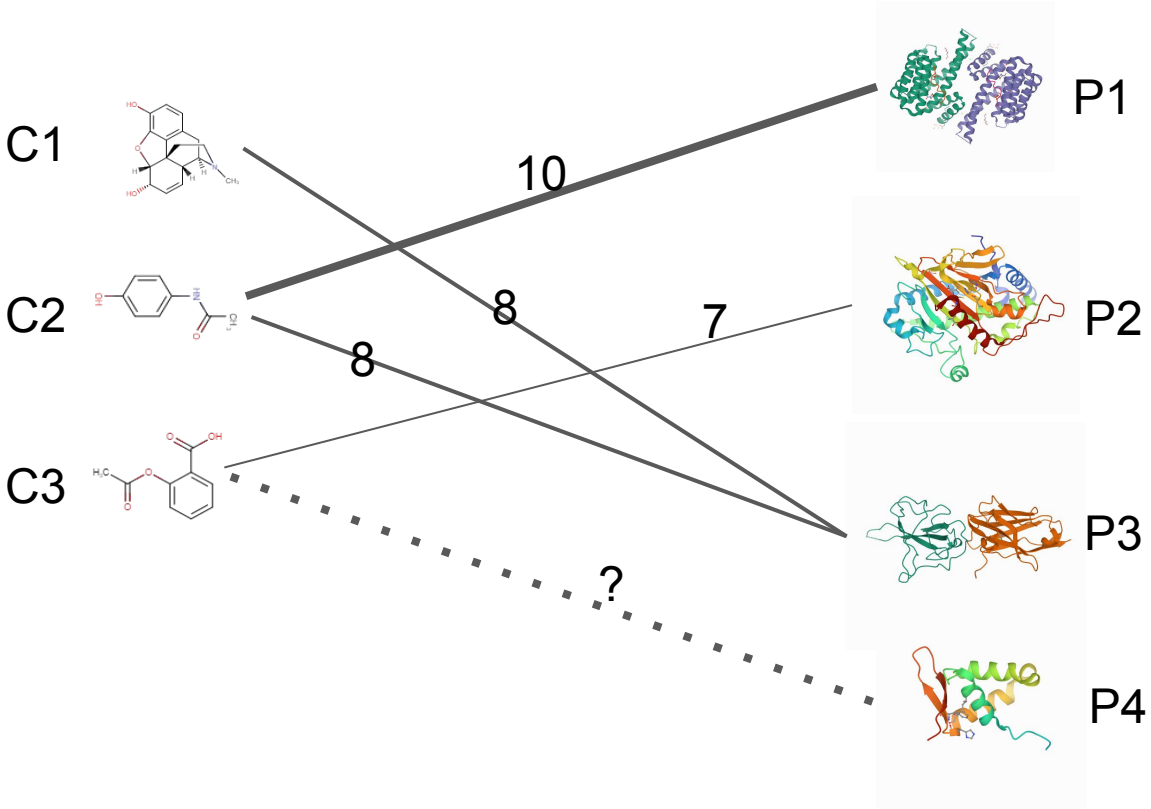
Warm Biomolecules



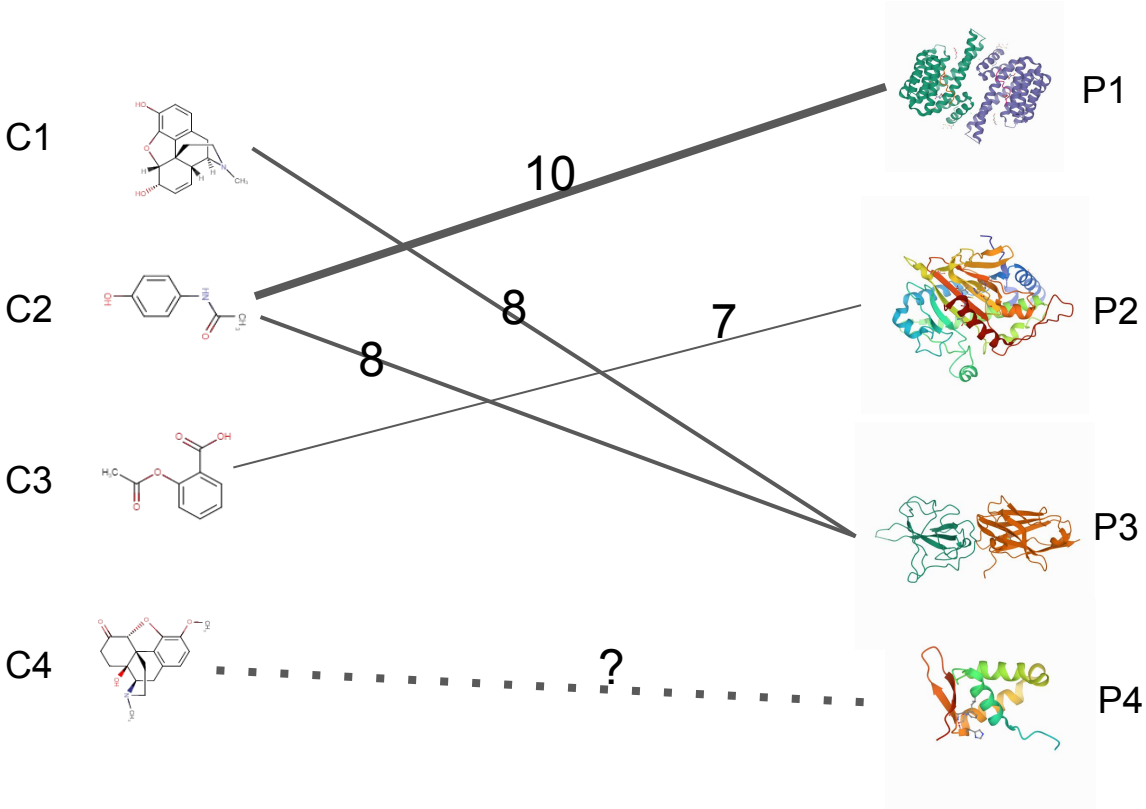
Cold Ligands



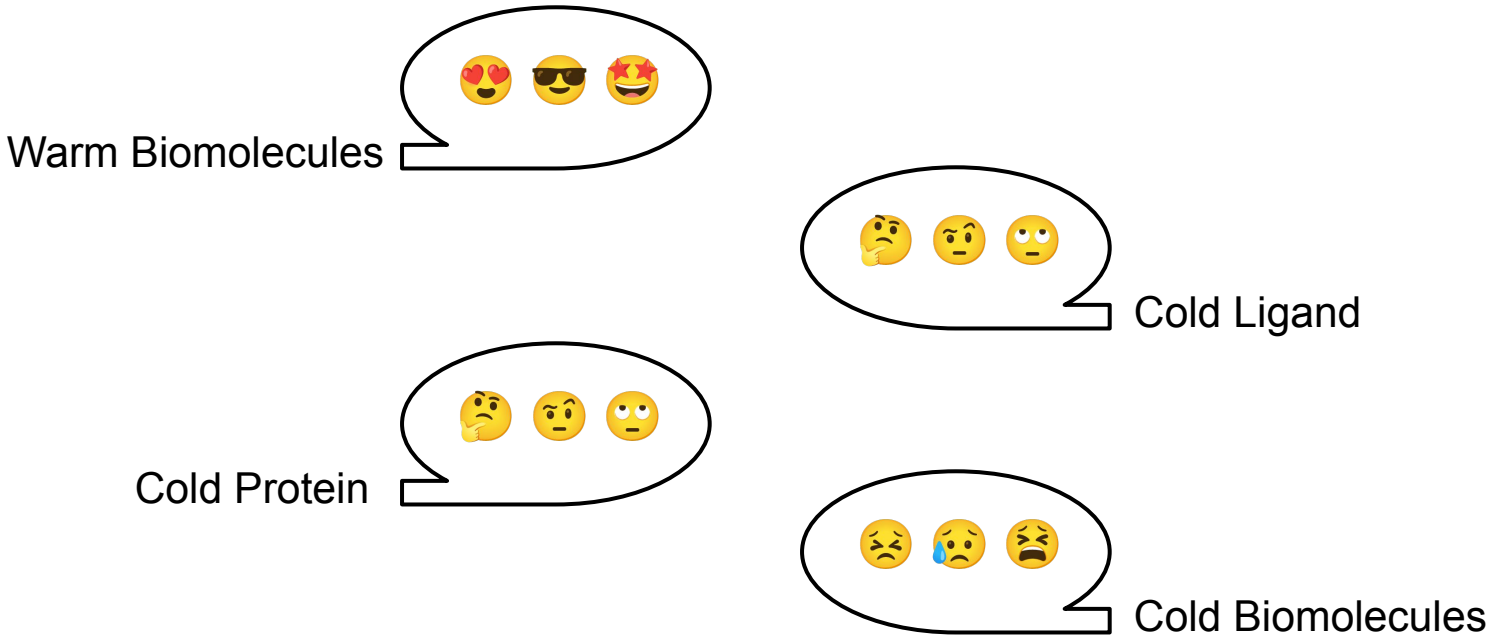
Cold Protein



Cold Biomolecules



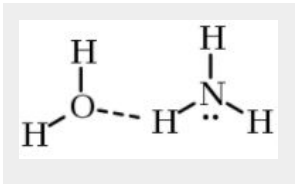
Machine Learning Models for Each Setup

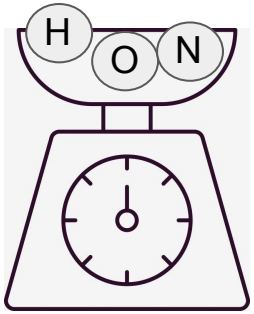


Dataset Biases for Affinity Prediction

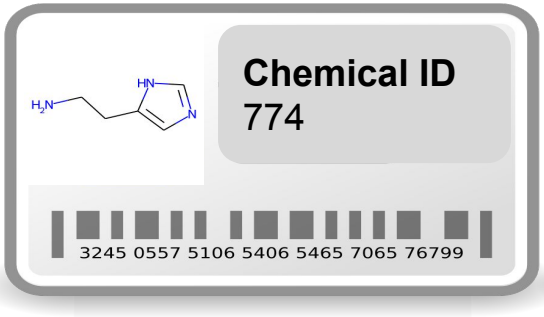
Outstanding patterns that misguide drug-target affinity (DTA) prediction models

Chemical Properties





Biomolecule Identifiers



Sundar, Vikram, and Lucy Colwell. "The Effect of Debiasing Protein–Ligand Binding Data on Generalization." *Journal of chemical information and modeling* 60.1 (2019): 56-62.

Özçelik, Rıza, et al. "ChemBoost: A Chemical Language Based Approach for Protein–Ligand Binding Affinity Prediction." Molecular Informatics 40.5 (2021): 2000212.

How to Mitigate Biomolecule Biases?

- Dataset-oriented: Arrange distant dataset folds (ref)
- Model-oriented: Change model training (ref)

Dataset Biases in Natural Language Inference

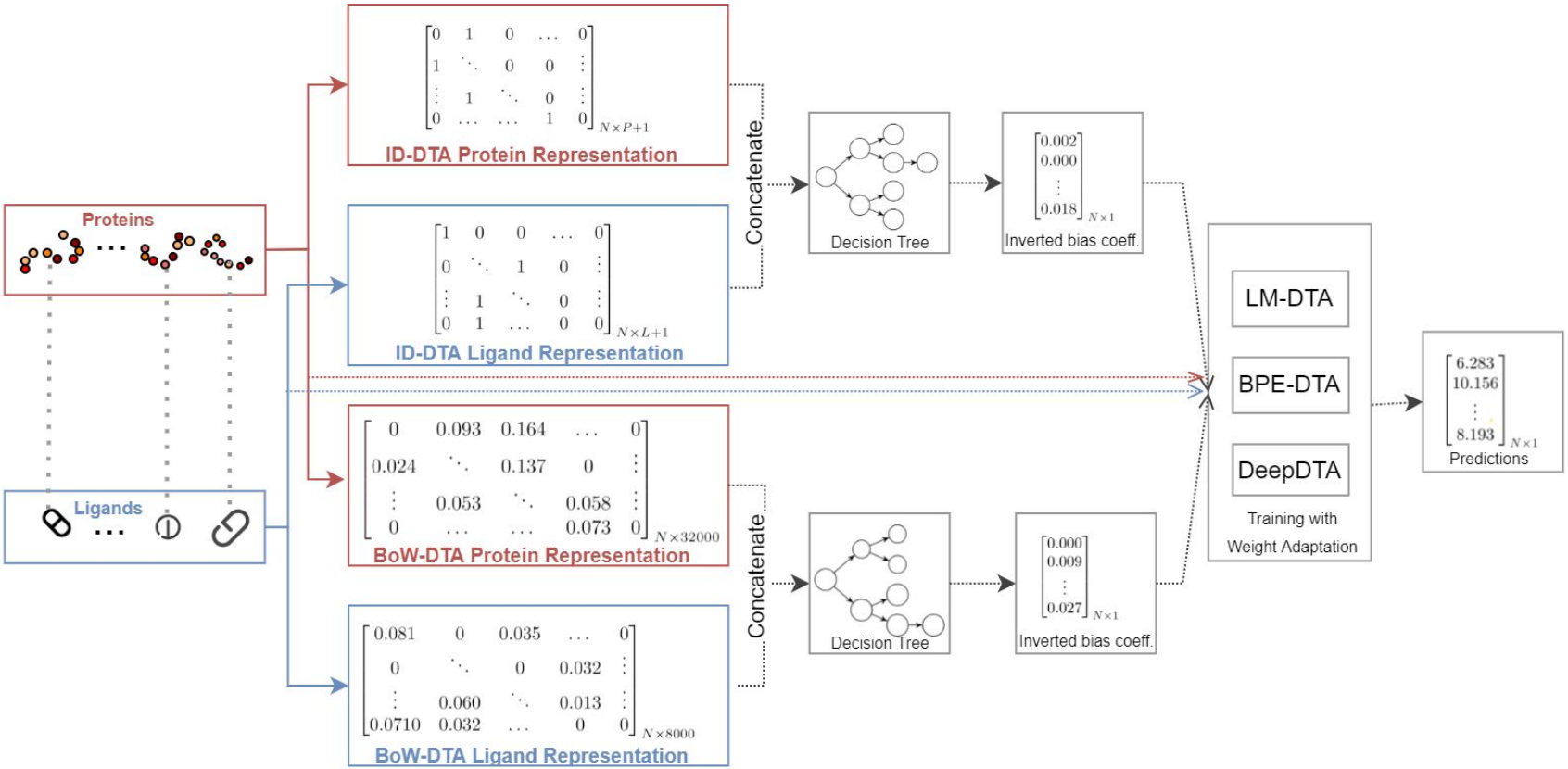
Premise: A woman selling bamboo sticks talking to two men on a loading dock.

Contradiction: A woman is **NOT** taking money for any of her sticks.

Learn and avoid these biases during training.

Gururangan, Suchin, et al. "Annotation Artifacts in Natural Language Inference Data." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.

DebiasedDTA: Ensemble Learning for Novel Drug-Target Affinity Prediction



Weak Learners

Idea: If dataset biases are outstanding, weak learners can learn these patterns.

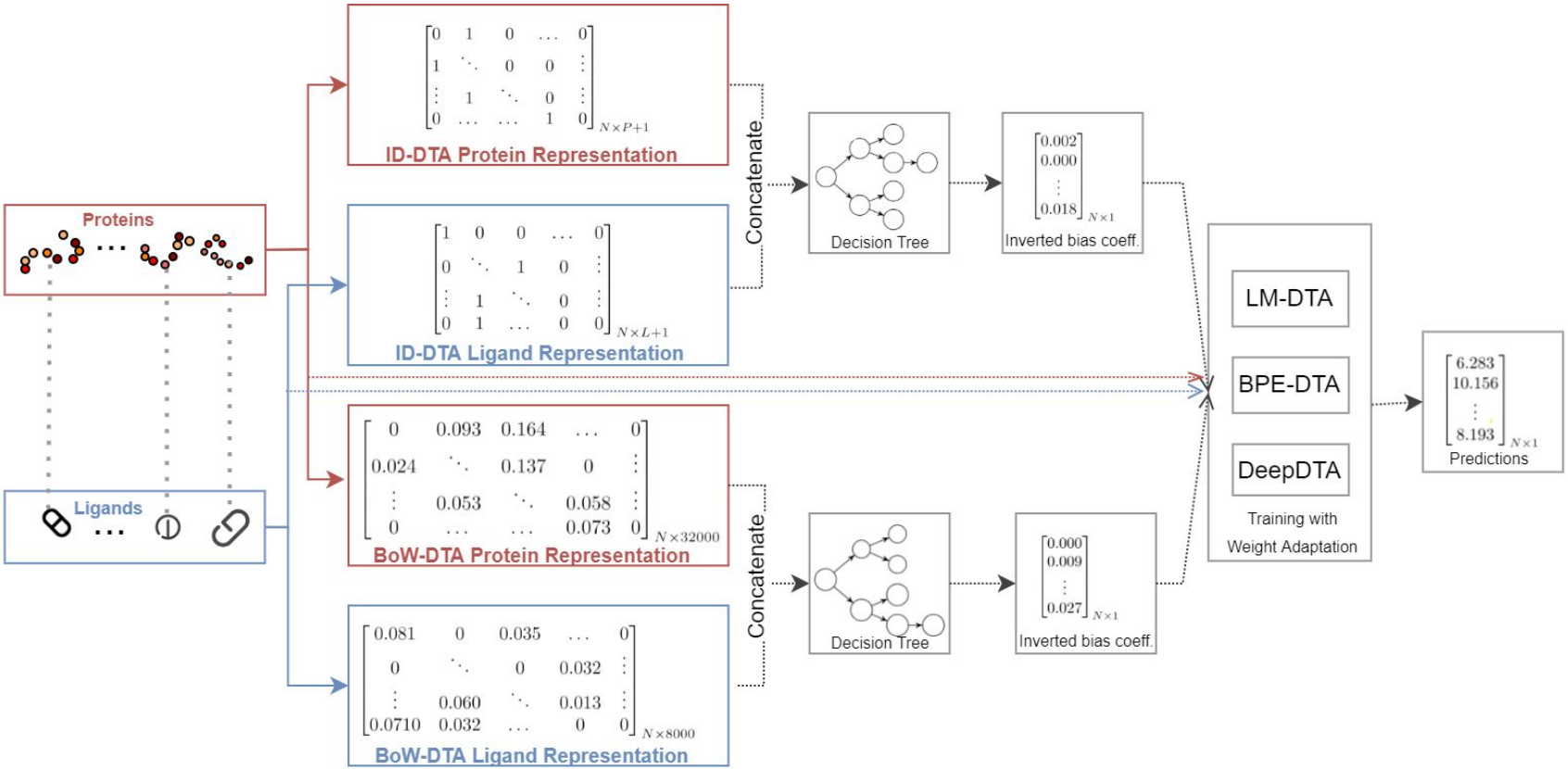
- ID-DTA
  - In order to avoid chemical identifier biases
  - Represents biomolecules with one-hot encoding
- BoW-DTA
  - In order to avoid “chemical word” biases
  - Represents biomolecules with bag-of-words representation
- Prediction with Regression Trees

Weak Learner Training

1. Train the weak learners with 5-fold cross validation (CV) on the **training set.**
2. Each interaction is a mini-validation sample during CV. Repeat the procedure 10 times and obtain 10 validation error measurements for each interaction.
3. Compute the median of 10 measurements and call the result as **inverted bias coefficient**.

Inverted bias coefficient: The larger, the more informative the sample.

DebiasedDTA



Strong Learners

- DeepDTA
  - Character-level convolutions over SMILES and amino-acid sequences
- BPE-DTA
  - Word-level convolutions over SMILES and amino-acid sequences
- LM-DTA
  - Pre-trained language-model embeddings

Use inverted bias coefficients to guide the training of the strong learners.

Strong Learner Training

Bias Decay

$$\vec{w}_e = (1 - \frac{e}{E}) + \vec{b} \times \frac{e}{E}$$

Bias Growth

$$\vec{w}_e = \frac{e}{E} + (\vec{b} - \frac{e}{E} \times \vec{b})$$

Experiments

- **BDB:** 490 proteins, 924 ligands, ~31K interactions
- **KIBA:** Kinase dataset, 229 proteins, 2111 ligands, ~118K interactions
- 5 different biomolecule split
- Evaluation: R^2 and concordance index (CI)
- Calculate improvement due to debiasing for each model

Jing Tang, Agnieszka Szwajda, 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.

∴ Debiasing improves performance on almost every setup

		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

∴ Performance increase is amplified in BDB

		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

Summary

- Predicting affinity scores between novel biomolecules is challenging.
- Dataset biases exist and needs avoidance.
- DebiasedDTA can identify and avoid these biases.
- All models leverage debiasing to boost their prediction performance.
- DebiasedDTA is applicable to almost **every** DTA prediction model.

Thanks for listening!