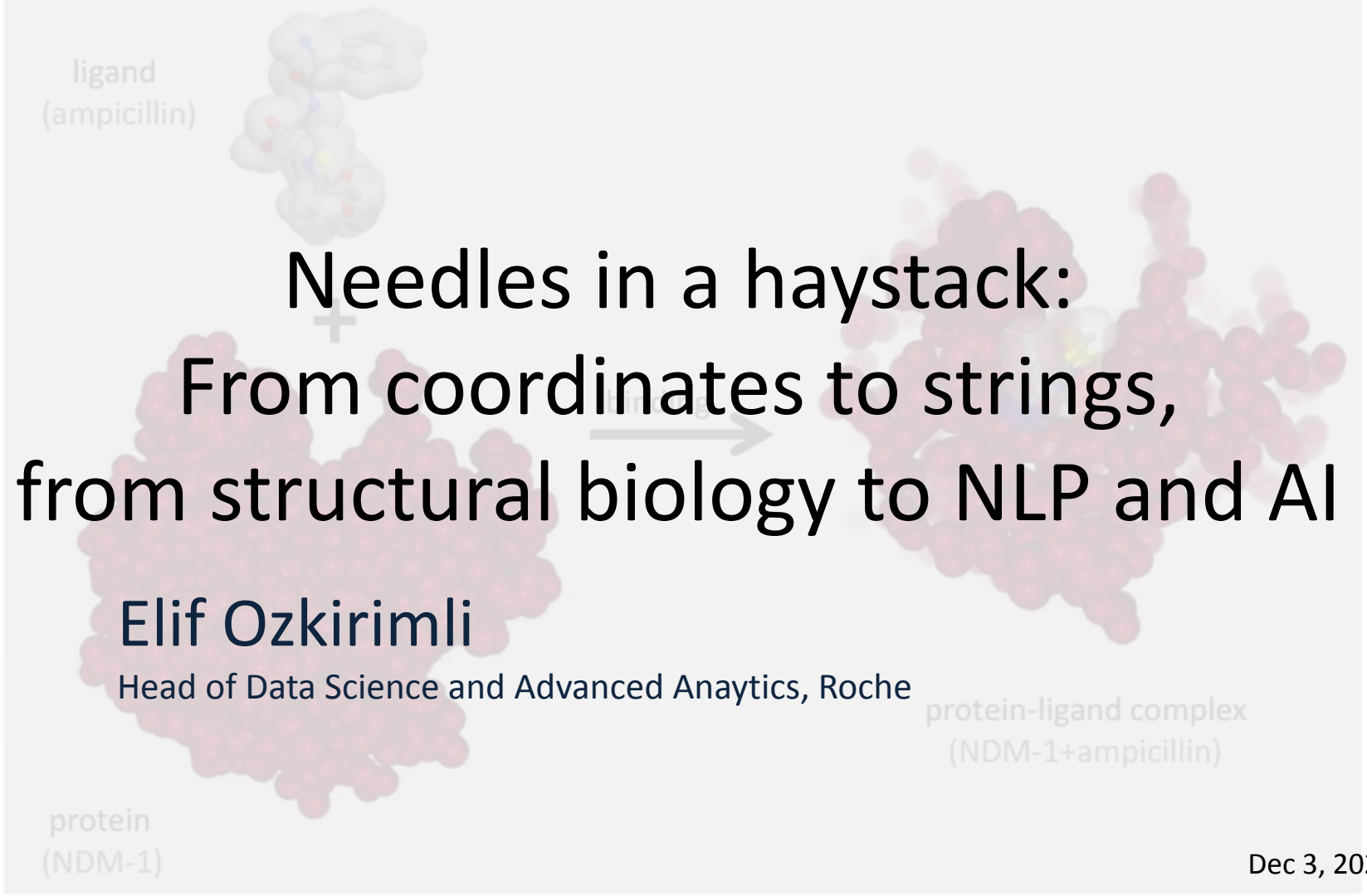


Needles in a haystack: From coordinates to strings, from structural biology to NLP and AI

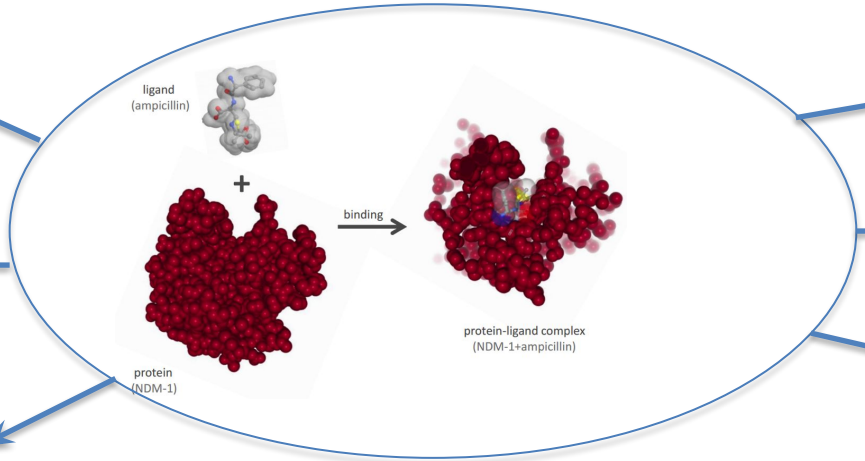
Elif Ozkirimli

Head of Data Science and Advanced Anaytics, Roche

Dec 3, 2021

Our starting research questions were

How do proteins bind to their interacting partners (proteins, ligands, substrate)?



MD simulations

Residue conservation

Docking calculations

Structural Biology
Biophysics

Machine Learning

Text mining/ relation extraction

Graph theory

Artificial Intelligence
Data Science

Fleming discovers Penicillin on September 28, 1928

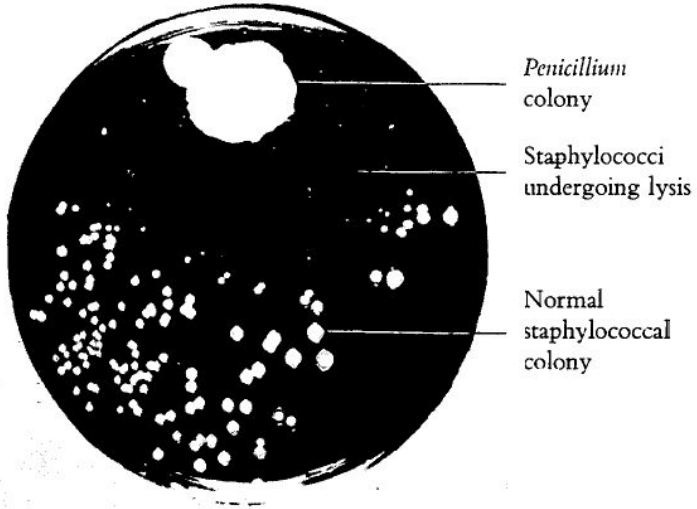


Fig. 1. Photograph of a culture-plate showing the dissolution of staphylococcal colonies in the neighbourhood of a *Penicillium* colony.

Fleming A. (1929) *British Journal of Experimental Pathology* **10**:226-36.

Penicillin resistant bacteria detected in early 1940s

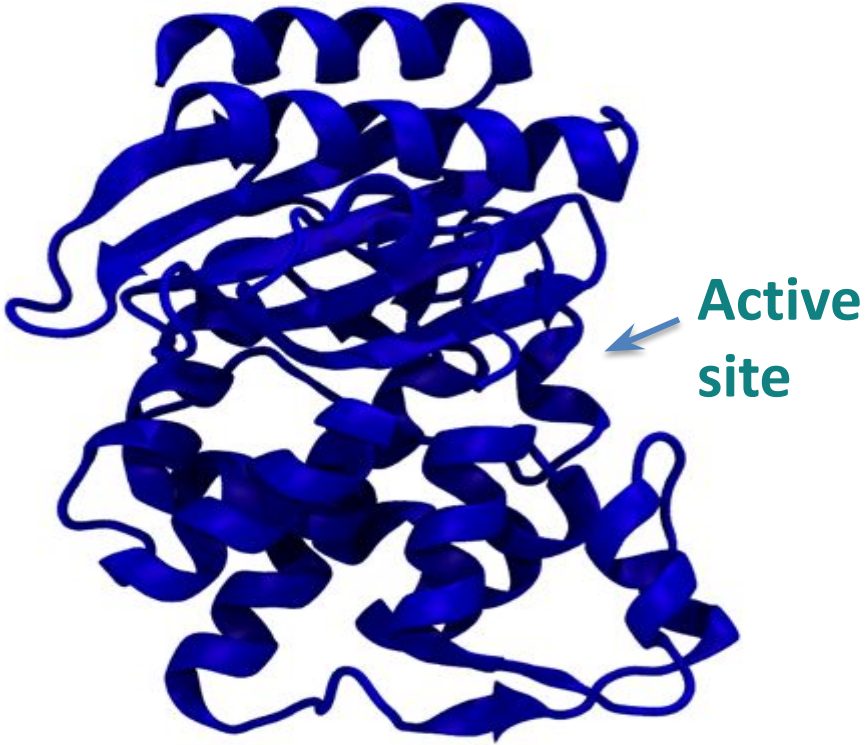
Abraham, E.P. and E. Chain. 1940.

An enzyme from bacteria

able to destroy penicillin.

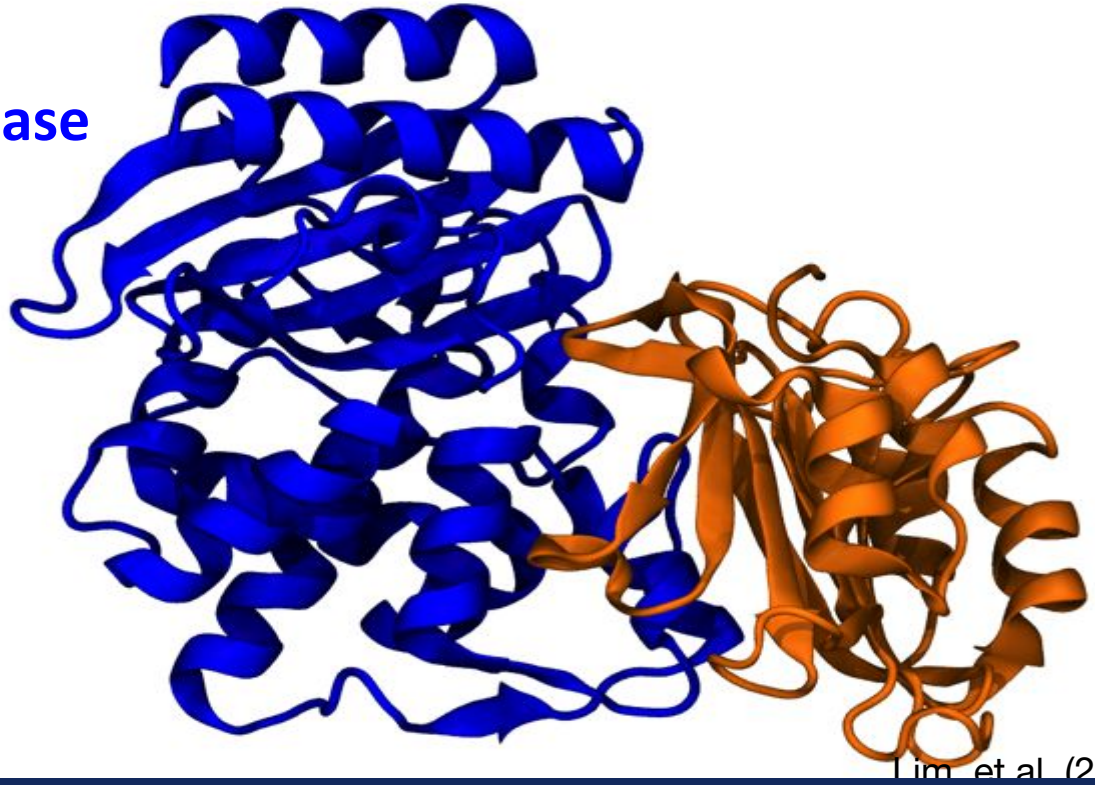
Nature 3713:837.

Beta-lactamase structure



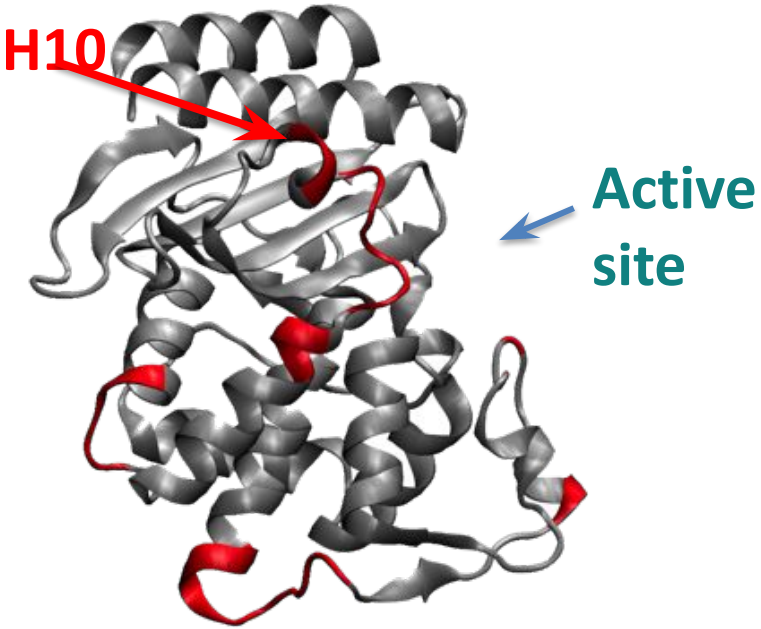
Stec et al. 2005 *Acta Crystallographica D* **61**:1072-1079

Beta-lactamase – Beta lactamase inhibitor protein (BLIP) complex



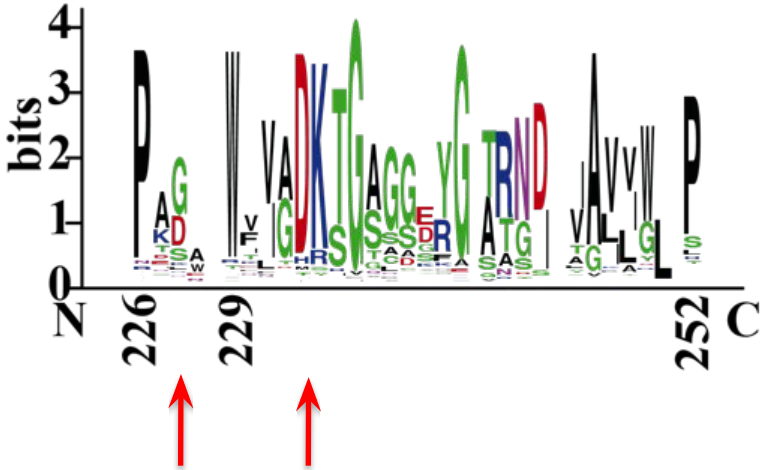
Lim, et al. (2001) *NSB* **8:** 848-852

MD simulations on apo + BLIP bound beta-lactamase



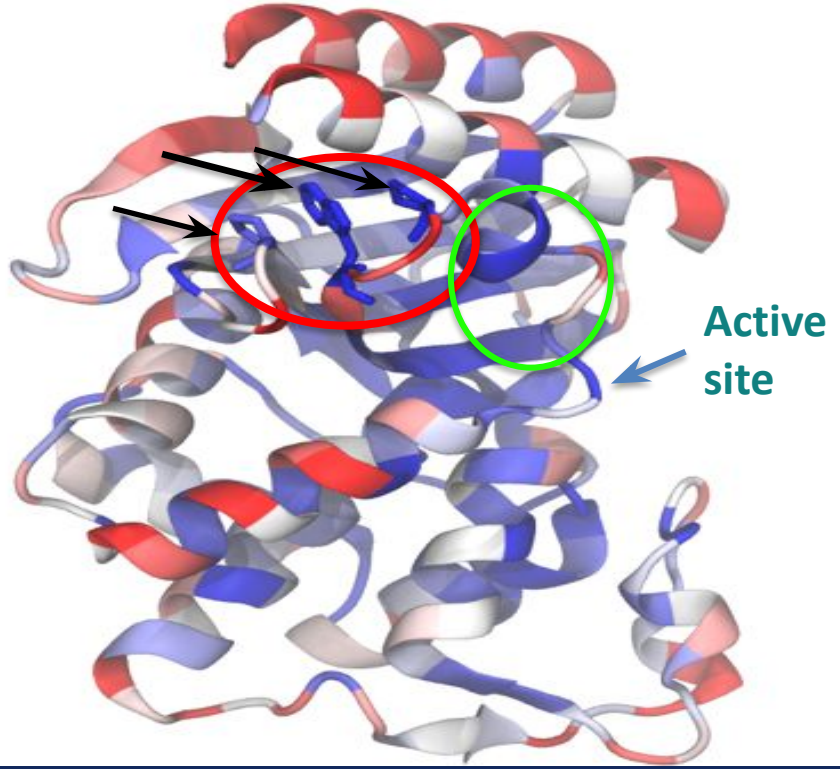
Meneksedag et al. (2013) *Computational Biology and Chemistry*, **43**:1-10

Sequence conservation around the allosteric region in class A beta-lactamases



Highly conserved

PWP triad



Meneksedag et al. (2013) *CBC*, **43:**1-10

Avci et al. (2016) *JEIMC* **31**:33-40

How does beta-lactamase evolution occur?

How does the development of new antibiotics and/or inhibitors drive beta-lactamase evolution forward?

576 unique beta-lactamase sequences in Uniprot

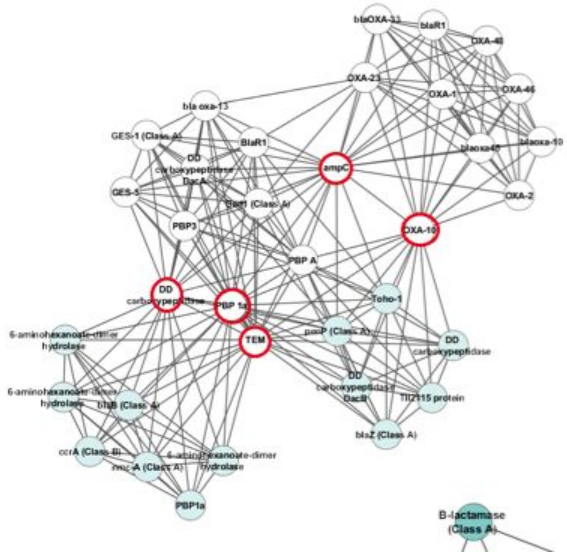
> 70,000 compounds in ChEMBL

Ligand centric protein networks

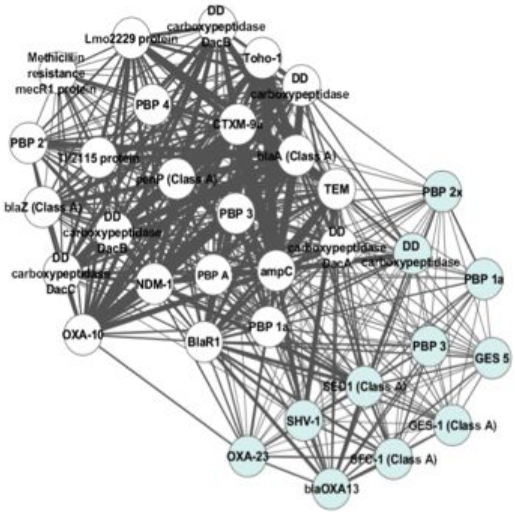
Our ligand centric view of protein families is built on two basic observations:

- chemically similar compounds bind to similar target proteins.
- target proteins that share similar binding sites bind to similar ligands.

Ligand based beta-lactamase networks



Identity network:
Place an edge if two
beta-lactamases share a ligand



Similarity network:
Place an edge if two beta-lactamases
share a similar ligand

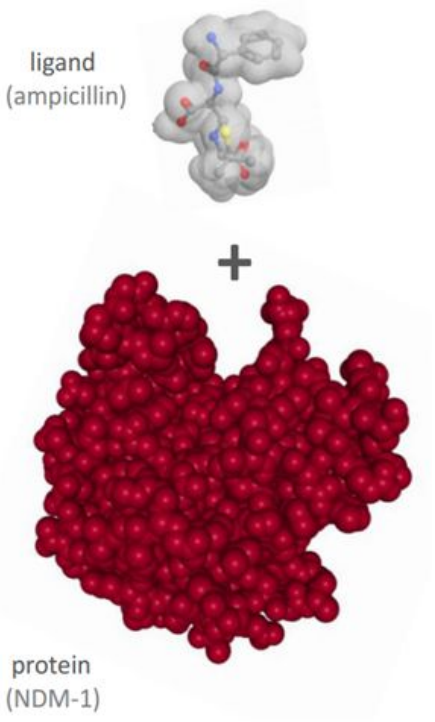
Ozturk et al., Plos One 2015

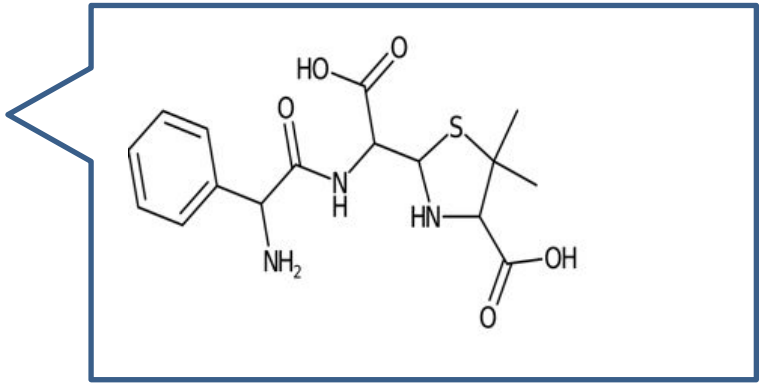
- Ligand binding carries functional and/or mechanistic information about the protein
- Sequence alone is not adequate to completely understand the mechanism.
- The relationship between fold or architecture and function can be weak.

We propose to represent proteins with their interacting ligands.

Martin A.C. et al. (1998) Protein folds and functions. *Structure*, 6, 875–884

But must first represent the ligands



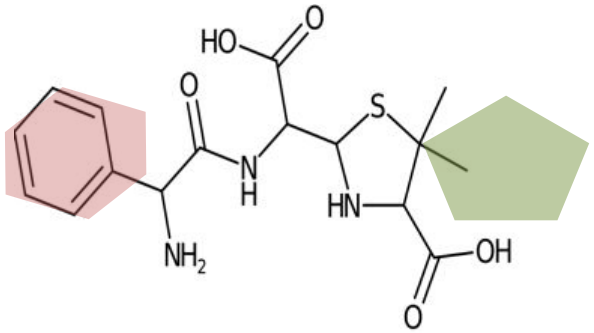


But must first represent the ligands

- Fingerprint models (e.g. ECFP6)

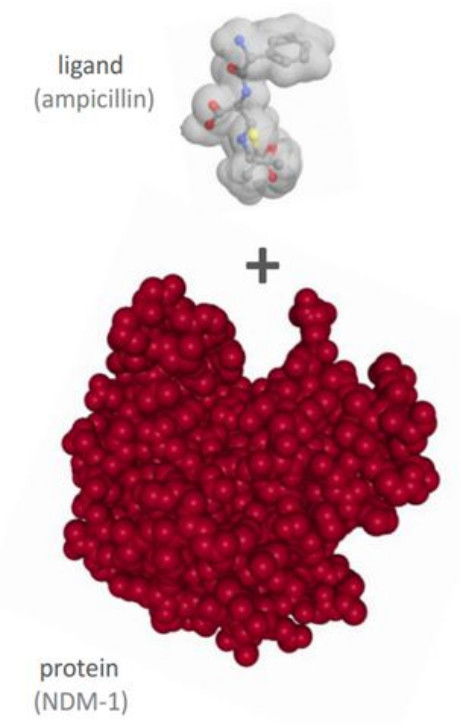
  binary feature vectors

- Graph based models



1
0
0
1
0
.
.

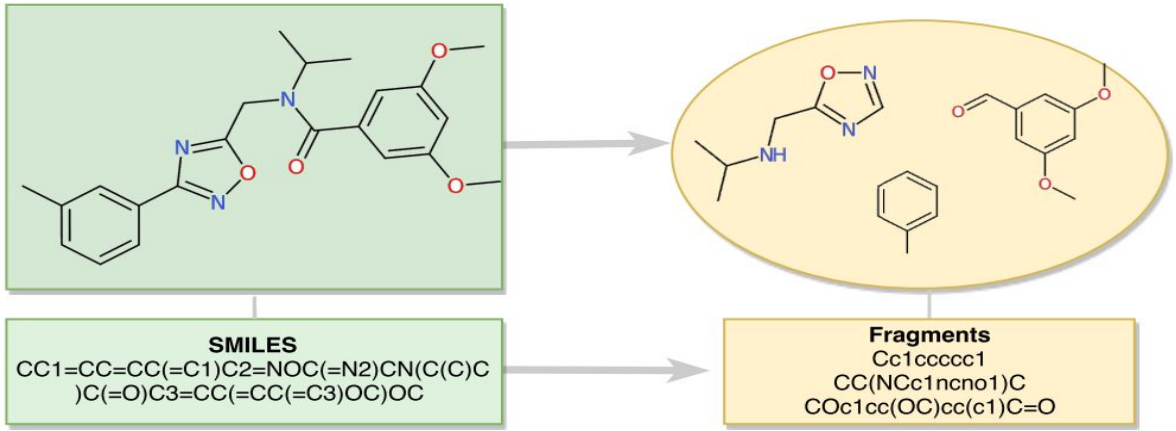
But must first represent the ligands




```
CC1(C(N2C(S1)C(C2=O)NC(=O)
C(C3=CC=CC=C3)N)C(=O)O)C
```

Simplified Molecular Input Line Entry System (SMILES)

Hypothesis: SMILES representation of compounds is a document



- What are the words??

SMILES representation is the document

SMILES: CN=C=O

Chemical words (LINGO)

SMILES: CN=C=O

Chemical words: CN=C

Chemical words (LINGO)

SMILES: CN=C=O

Chemical words: CN=C, N=C=

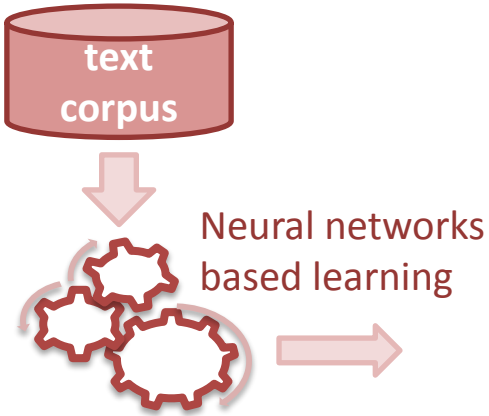
Chemical words (LINGO)

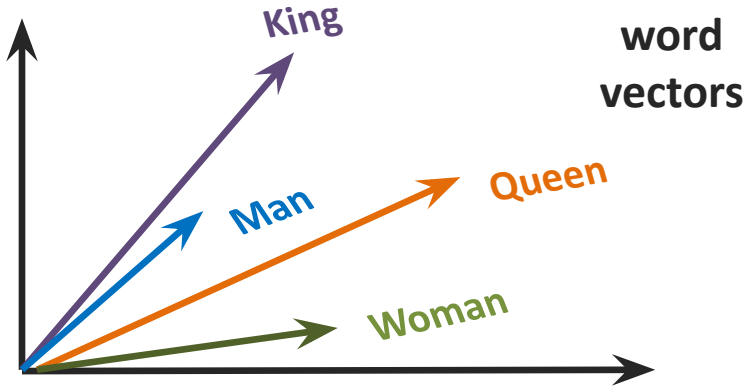
SMILES: CN=C=O

Chemical words: CN=C, N=C=, =C=O

Vidal D, Thormann M, and Pons M (2005) *Journal of chemical information and modeling* 45.2 386-393.

Distributed word representations (word2vec)





"You shall know a word by the company it keeps!" Firth 1957

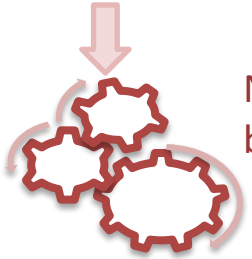
Mikolov, Tomas, et al. *Advances in neural information processing systems*. 2013.

Distributed word representations (SMILESVec)

SMILES: CN=C=O

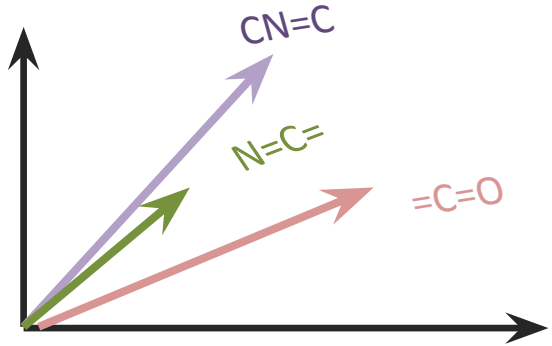
words: CN=C, N=C=, =C=O

PubChem
2M compounds



Neural networks
based learning

Skip gram approach
100D real valued embeddings



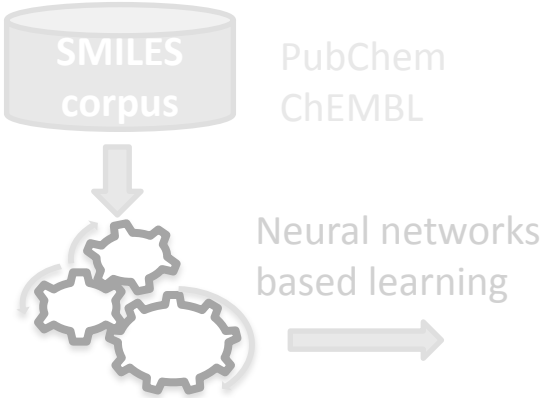
Chemical
word (**cw**)
vectors

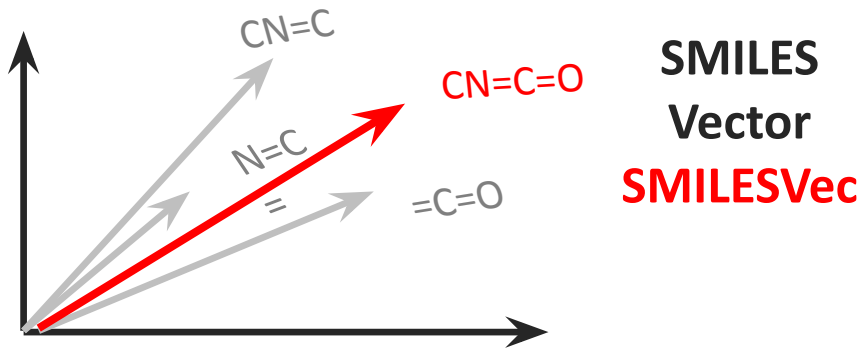
Öztürk, Hakime et al., "A novel methodology on distributed representations of proteins using their interacting ligands." *Bioinformatics,* (2018).

Distributed word representations (SMILESVec)

SMILES: CN=C=O

words: CN=C, N=C=, =C=O



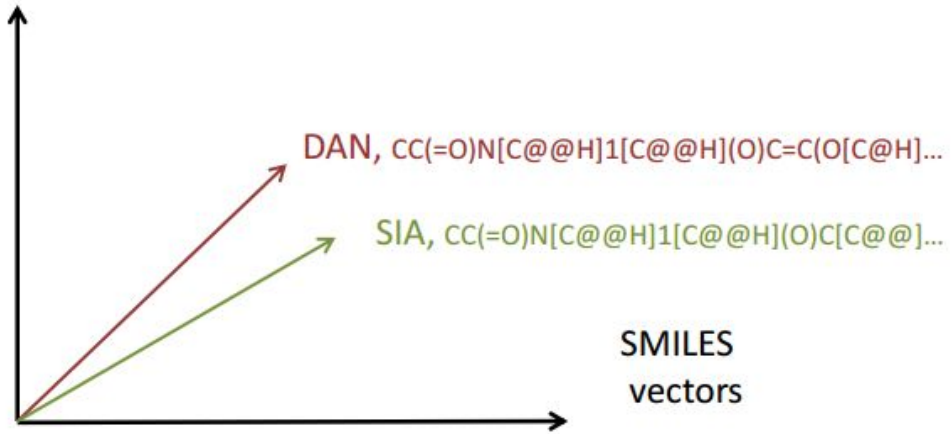


Öztürk, Hakime et al., "A novel methodology on distributed representations of proteins using their interacting ligands." *Bioinformatics,* (2018).

SMILESVec-based protein representation

Protein: sialidase

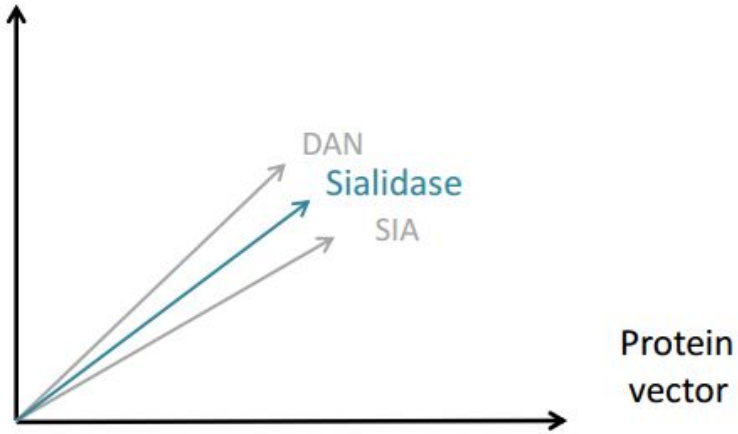
Interacting ligands: DAN, SIA



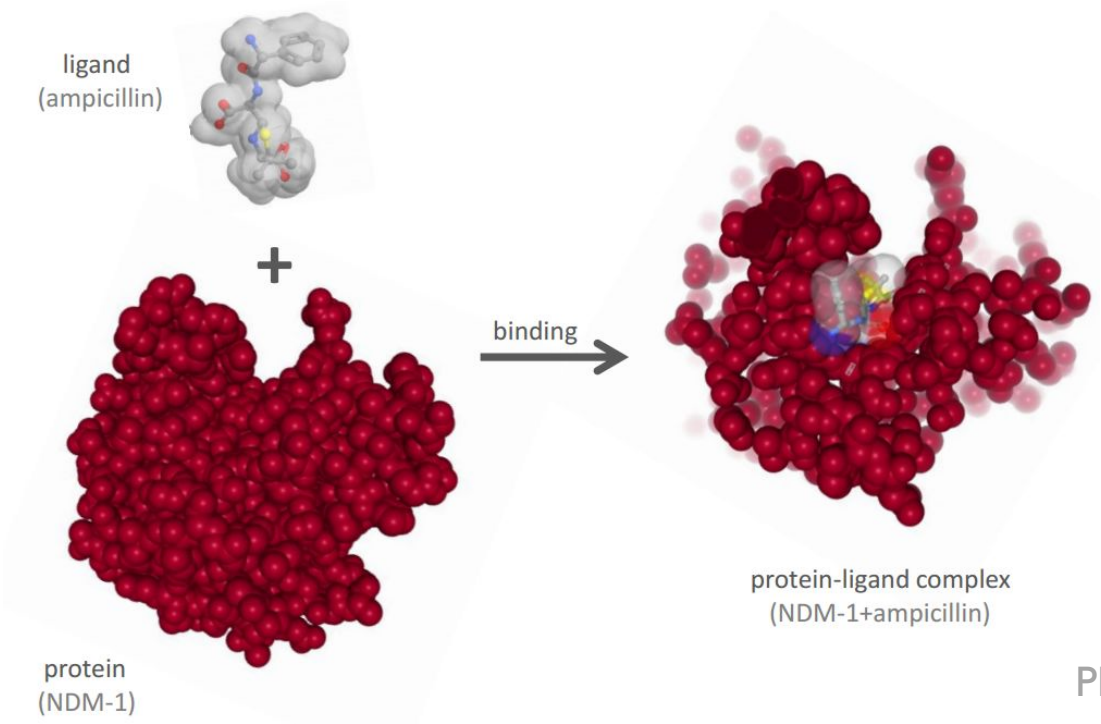
SMILESVec-based protein representation

Protein: sialidase

Interacting ligands: DAN, SIA

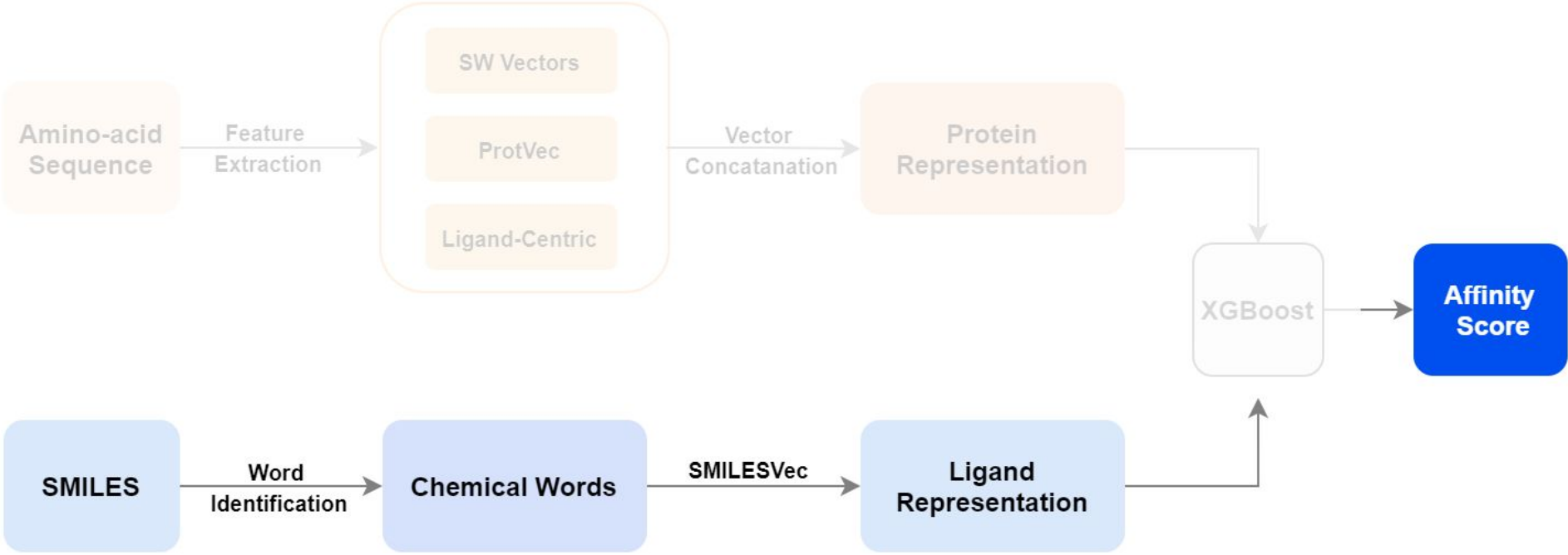


Protein - ligand affinity prediction

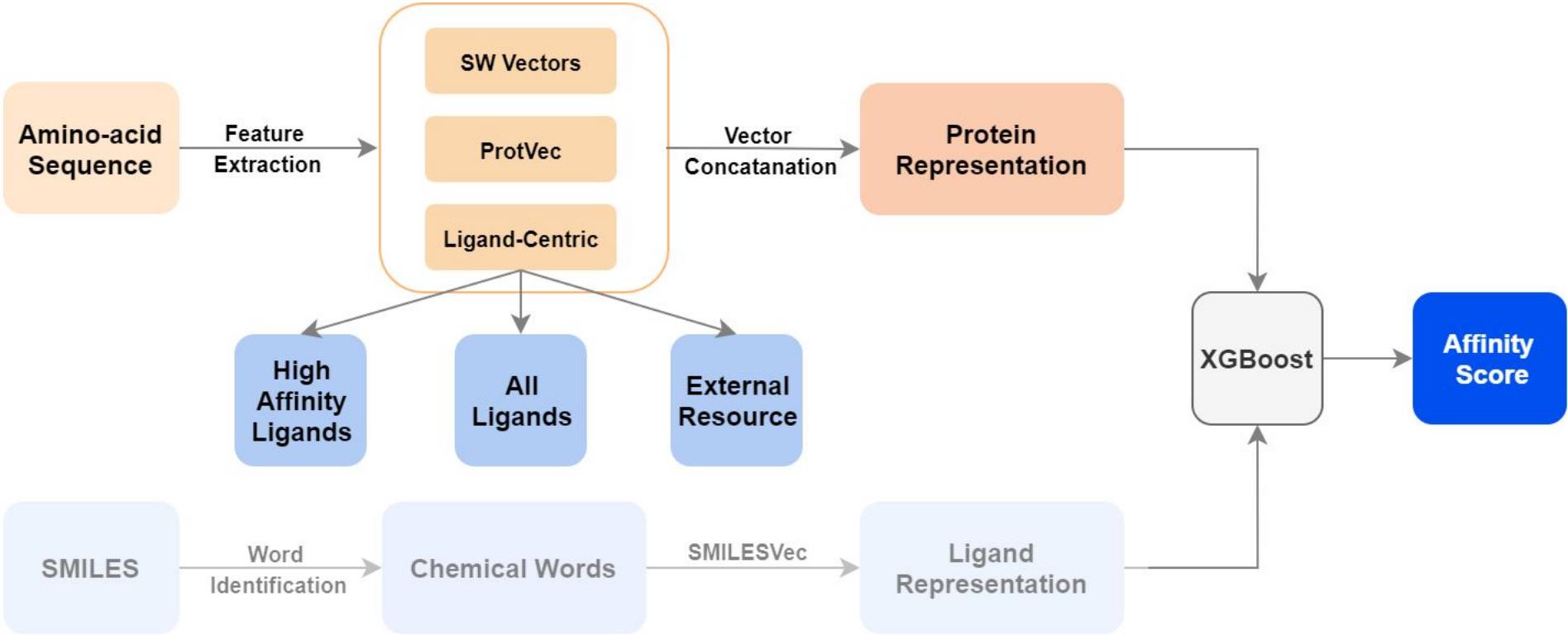


PDB: 3Q6X

ChemBoost Ligand Representation



ChemBoost Protein Representation



Chemboost achieves high performance in comparison to benchmark and SOTA

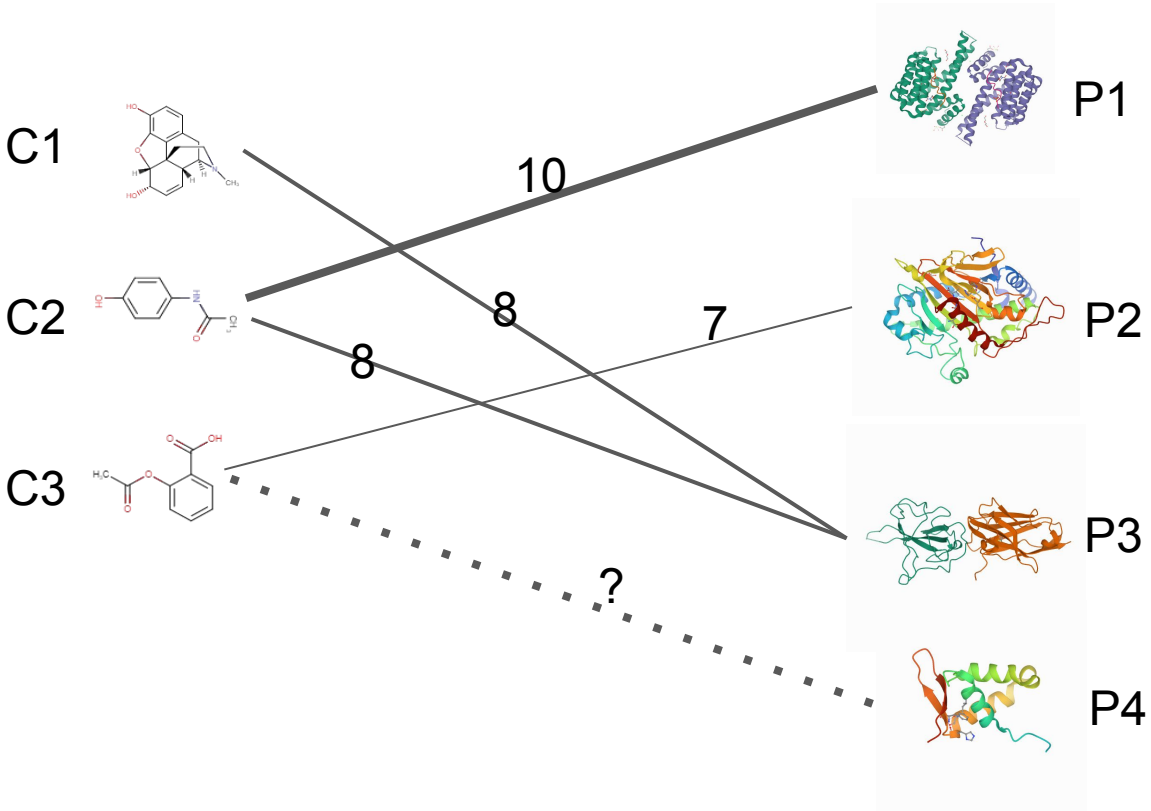
	BDB Scores		KIBA Scores	
Model	**CI**	**MSE**	**CI**	**MSE**
KronRLS	0.814 (0.002)	0.939 (0.004)	0.782 (0.001)	0.411
SimBoost	0.853 (0.003)	0.485 (0.043)	0.836 (0.001)	0.223 (0.003)
DeepDTA	0.863 (0.007)	0.397 (0.011)	0.846 (0.002)	0.215 (0.005)
ChemBoost	0.871 (0.002)	0.420 (0.007)	0.836 (0.001)	0.207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

- Pahikkala et al. "Toward more realistic drug–target interaction predictions." *Briefings in bioinformatics* 16.2 (2014): 325-337.
- Tong, et al. "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines." *Journal of cheminformatics* 9.1 (2017): 24.
- Öztürk et al.. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

but…

Predictions for novel biomolecules is a challenge.

Cold Protein



Predictions for novel biomolecules is a challenge.

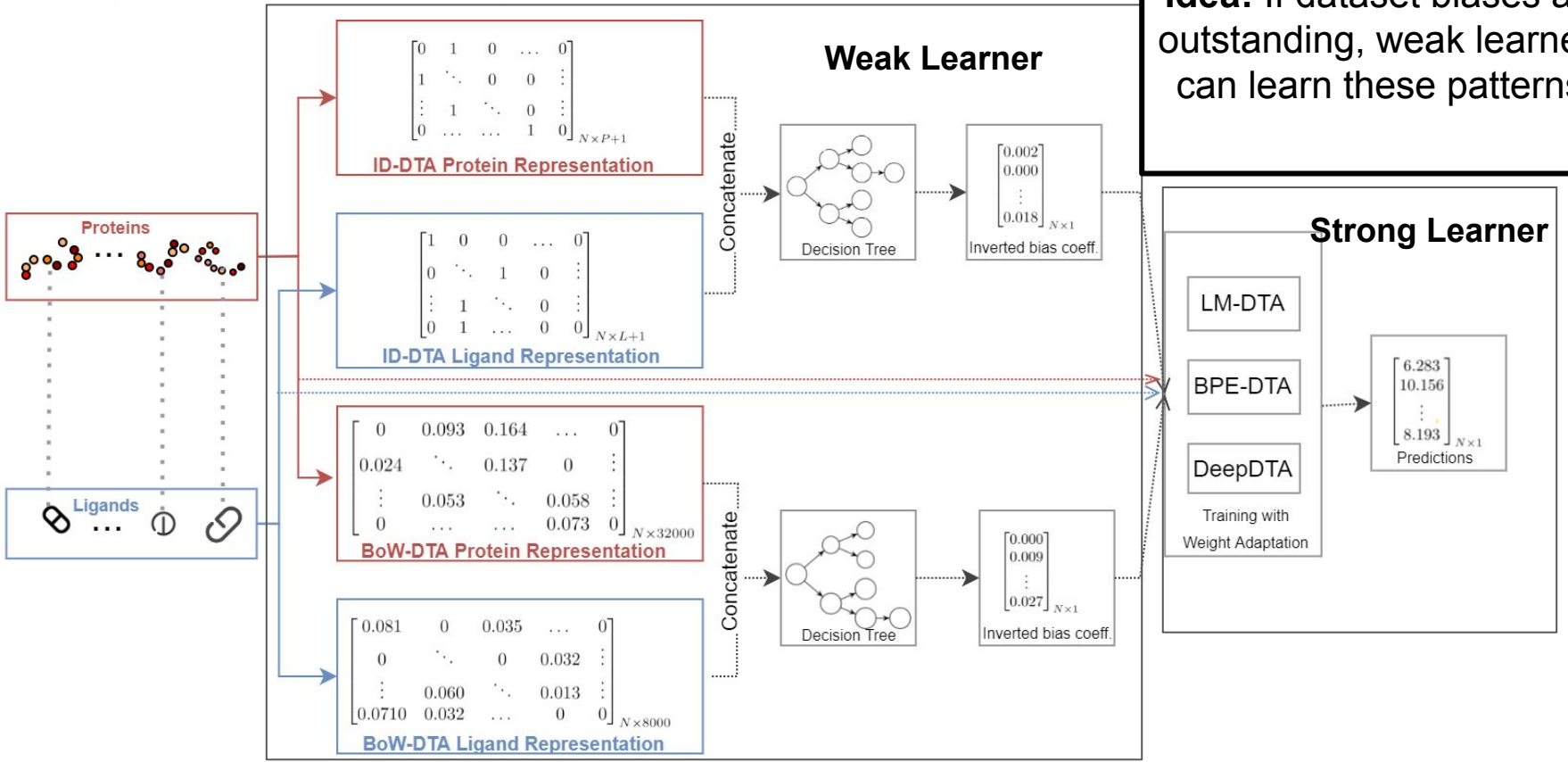
Model	Warm		Cold Ligand		Cold Protein		Cold	
	MSE	CI	MSE	CI	MSE	CI	MSE	CI
Model (1)	0.373	0.885	1.178	0.736	0.720	0.799	1.393	0.657
Model (7)	0.404	0.863	1.185	0.700	1.156	0.749	1.576	0.596
Model (9)	0.361	0.880	1.157	0.730	0.800	0.786	1.358	0.665
DeepDTA	0.345	0.879	1.350	0.672	0.810	0.778	1.522	0.614

DebiasedDTA: Model Debiasing to Boost Drug -Target Affinity Prediction

Özçelik, R., Bag, A., Atil, B., Özgür, A., & Ozkirimli, E.. **DebiasedDTA: Model Debiasing to Boost Drug -Target Affinity Prediction** *submitted.*

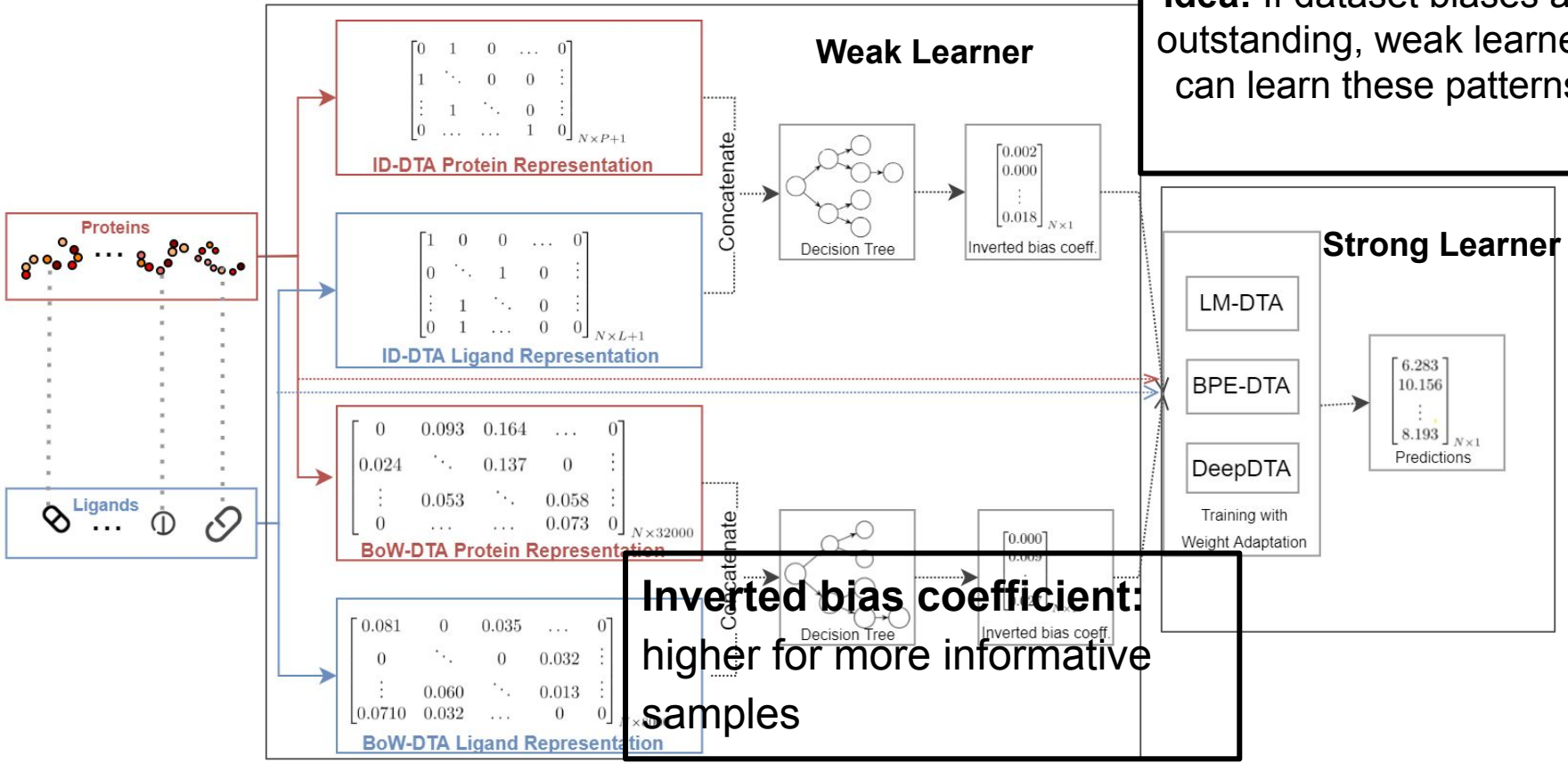
DebiasedDTA: Ensemble Learning for Novel Drug-Target Affinity Prediction

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



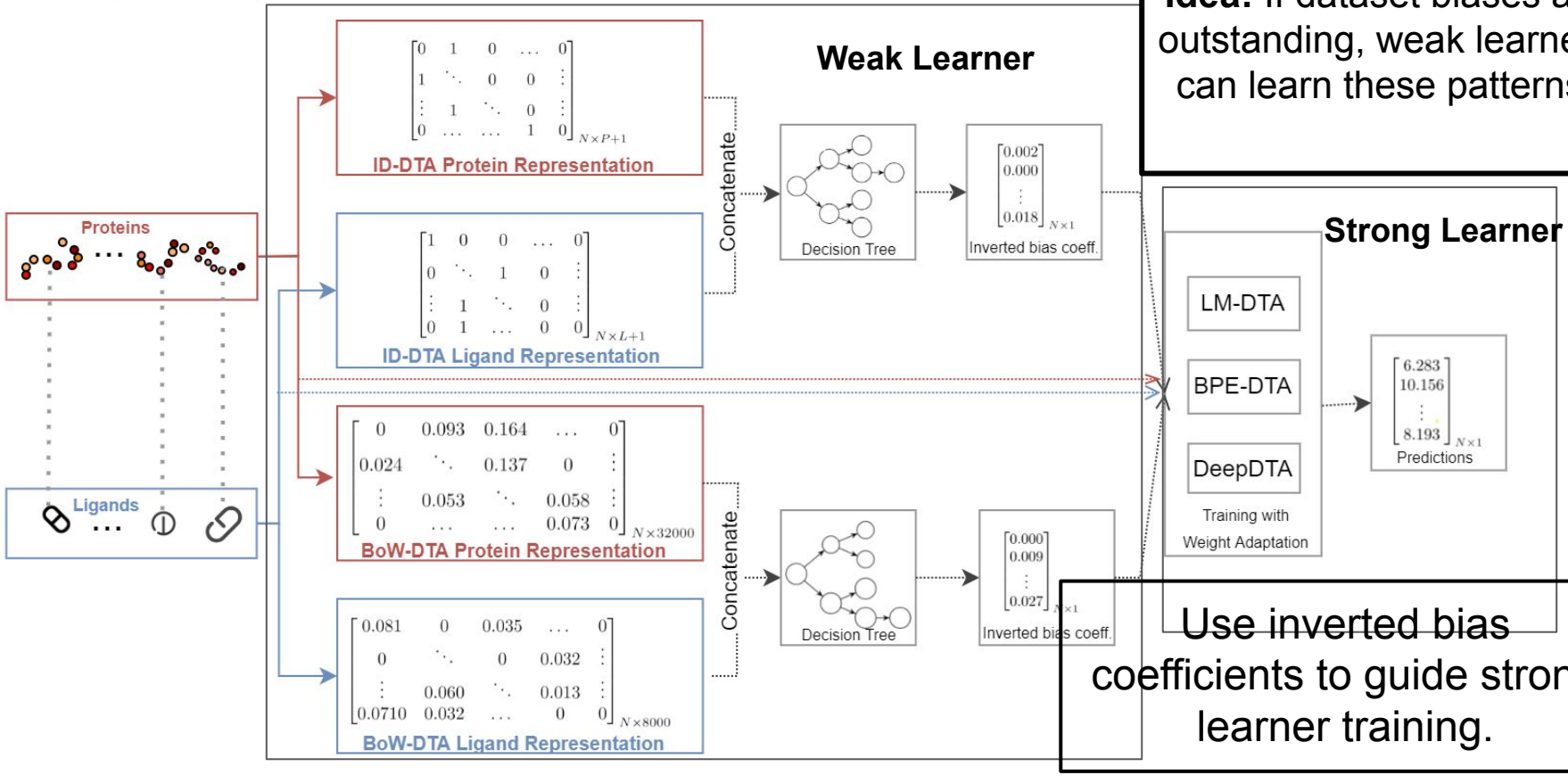
DebiasedDTA: Ensemble Learning for Novel Drug-Target Affinity Prediction

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



Inverted bias coefficient: higher for more informative samples

DebiasedDTA: Ensemble Learning for Novel Drug-Target Affinity Prediction



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

Use inverted bias coefficients to guide strong learner training.

Debiasing improves prediction performance

		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

Table 2. **The gain of debiasing.** The percentile improvement in CI and increase in R2 are displayed for each model on every test set. The statistics are computed by comparing the best DebiasedDTA score with the non-debiased one. Negative statistics are reported if the non-debiased model outperforms every debiasing configuration.

What do I do now?

At Roche, I'm the head of data science and advanced analytics in commercial space answering questions such as

- Understanding the Dr or patient journeys
- Extracting information from conversations between Drs and sales reps
- Building recommendation systems for Drs, patients, sales reps
- Price predictions in a complex ecosystem
- Sales forecasting
- Analysis of real world data for commercial decision making

Text data in the biomedical domain

- Natural Languages

Chitosan Oligosaccharide Exerts Anti-Allergic Effect against Shrimp Tropomyosin-Induced Food Allergy by Affecting Th1 and Th2 Cytokines.

Jiang T[1,2], Ji H[3], Zhang L[4], Wang Y[5], Zhou H[6].

Author information

Abstract

BACKGROUND: Shrimp-derived allergen has a serious impact on people's health. Chitosan oligosaccharide (COS) has anti-allergic action but its function on shrimp allergen-induced allergy and related molecular mechanisms remain unclear.

METHODS: COS and its degrees of polymerization (DP) were selected to interact with shrimp tropomyosin (TM) and IgE was measured. A mouse model of food allergy was established by receiving shrimp TM intraperitoneally. The models were treated with different concentrations of COS. Fecal and serum histamine, serum IgE, IgG1 and IgG2a, and inflammatory cytokines were measured.

RESULTS: The main products for COS were DP2-6 with the contents of 6, 40, 26, 16, and 4%, respectively, and reacted with shrimp TM increasingly when COS DP was increased. Severe symptoms of food allergy were observed in the TM group (diarrhea, anaphylactic response, and rectal temperature). In contrast, COS treatment improved these symptoms significantly ($p < 0.05$). The sensitized mice were desensitized after they were treated with 1 mg/kg COS. COS treatment significantly reduced serum IgE and IgG1 levels, and increased IgG2a levels ($p < 0.05$). COS consumption decreased fecal and serum histamine. COS treatment reduced Th2 cytokine (IL-4, IL-5, and IL-13) levels and increased the Th1 cytokine (IFN-γ) level ($p < 0.05$).

CONCLUSIONS: COS showed anti-allergy properties by regulating the levels of Th1 and Th2 cytokines.

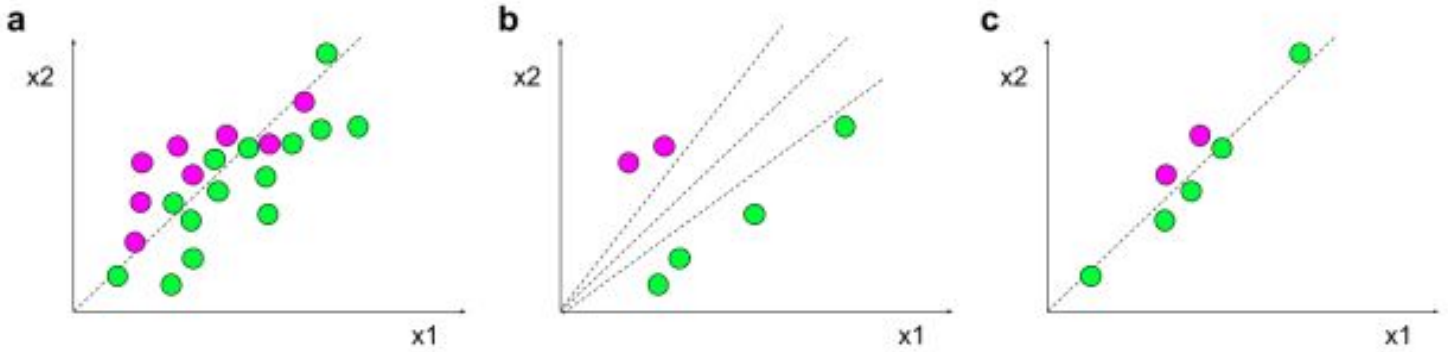
- DNA sequence
- Protein sequence
- Chemical formula

TTCAGGTGCATAAGACCTTGAC…

MELPNIMHPVAKLSTALAAALML…

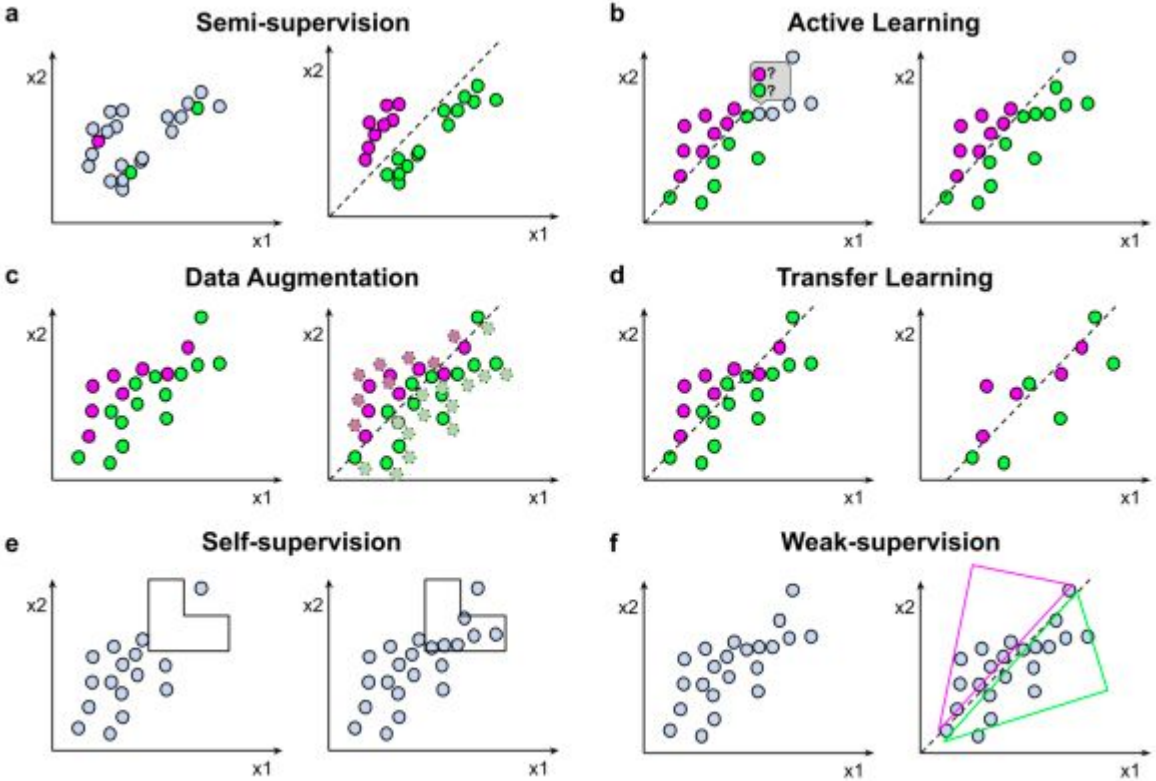
CC1(C(N2C(S1)C(C2=O)NC(=O)…

Needles in a haystack

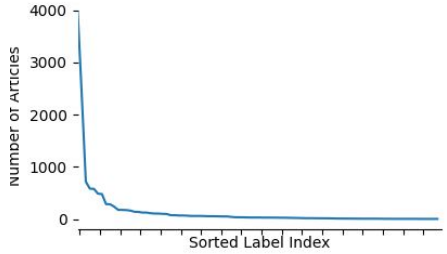


Yakimovich, A., Beaugnon, A., Huang, Y., Ozkirimli, E., “Labels in a Haystack: Approaches beyond Supervised Learning in Biomedical Applications”. *Accepted for publication*

Needles in a haystack



Yakimovich, A., Beaugnon, A., Huang, Y., Ozkirimli, E., “Labels in a Haystack: Approaches beyond Supervised Learning in Biomedical Applications”. *Accepted for publication*



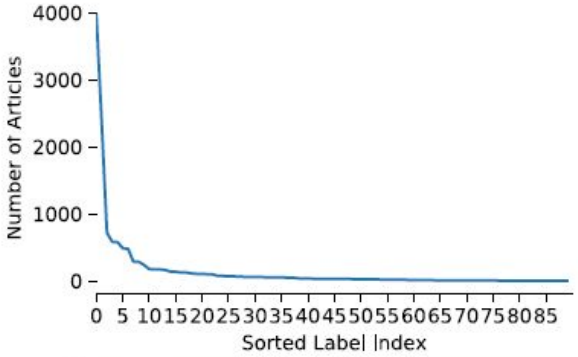
Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution

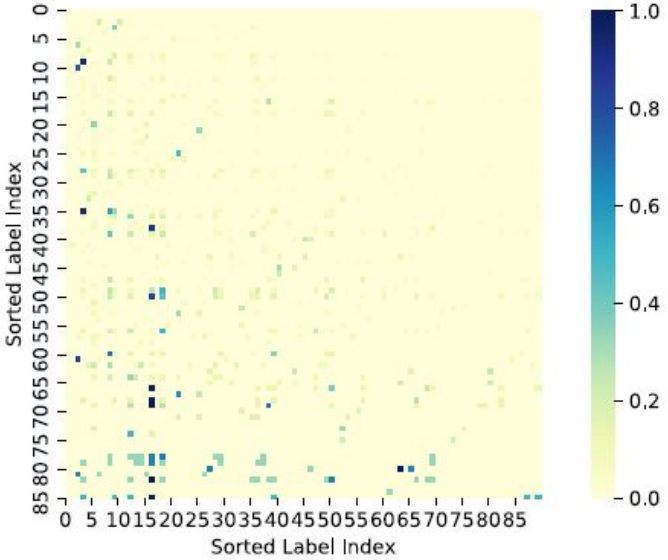
Huang, Y., Giledereli, B. Köksal, A., Ozgur, A., Ozkirimli, E. (2021) **Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution** EMNLP 2021.

Long tailed distribution and co-occurrence

Reuters-21578 dataset

Title 1	PENN CENTRAL <PC> SELLS U.K. UNIT
Labels 1	acq (1650), strategic-metal (16) **nickel** (8)
Title 2	U.S. MINT SEEKING OFFERS ON COPPER, NICKEL
Labels 2	copper (47), **nickel** (8)





Loss function manipulation can address class imbalance

- **Binary cross entropy** is vulnerable to label imbalance due to the dominance of head classes or negative instances.
- **Resampling and re-weighting** not effective when there is label dependency because they result in oversampling of common labels.
- Multi-label classification has been widely studied in the **computer vision (CV) domain**, and recently has benefited from **cost-sensitive learning through loss functions**
  - In object recognition (Durand et al., 2019; Milletari et al., 2016), semantic segmentation (Ge et al., 2018), and medical imaging (Li et al., 2020a).
- Loss function manipulation has also been explored (Li et al., 2020b; Cohan et al., 2020) in NLP as it works in a **model architecture-agnostic fashion** by explicitly embedding the solution into the objective.
  - For example, Li et al. (2020b) has borrowed dice-based loss function from a medical image segmentation task (Milletari et al., 2016) and reported significant improvements over the standard cross-entropy loss function in several NLP tasks.

Loss function manipulation can address class imbalance

- We propose using **distribution balanced (DB) loss** with 3 layers: focal loss, rebalanced weighting and negative-tolerant regularization (NTR).
- **Focal loss** places a higher weight of loss on “hard-to-classify” instances predicted with low probability on ground truth while NTR addresses the co-occurence problem.
- **DB loss** first reduces redundant information of label co-occurrence and then explicitly assigns lower weight on “easy-to-classify” negative instances.
- **NTR** helps to avoid over-suppression of the negative labels caused by the dominance of negative classes in binary cross entropy (BCE)

Results

	PubMed Total	PubMed Head (>50)	PubMed Med (15-50)	PubMed Tail (<15)
BCE	0.02	0.06	0	0
SVM	13.31	34.33	5.62	0.67
DB (Ours)	19.19	40.48	15.33	3.08

Macro F1 results with SVM and BCE baselines comparing to our model

Methods | Metrics

$$Micro - F1 = \frac{2 * Precision_{all} * Recall_{all}}{(Precision_{all} + Recall_{all})}$$

$$Macro - F1 = \frac{1}{C} \sum_{i}^{C} \frac{2 * Precision_{i} * Recall_{i}}{(Precision_{i} + Recall_{i})}$$

DB improves classification performance even for tail labels

	PubMed Total	PubMed Head (>50)	PubMed Med (15-50)	PubMed Tail (<15)
SVM	58.54	60.77	19.78	6.94
BCE	26.17	27.61	0	0
DB	60.63	62.39	41.14	24.19

Micro F1 comparison of DB with SVM and BCE baselines

DB improves classification performance even for tail labels

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF	PubMed Total miF/maF	PubMed Head(≥50) miF/maF	PubMed Med(15-50) miF/maF	PubMed Tail(≤15) miF/maF
SVM	87.60/51.63	89.87/78.47	66.92/61.00	22.54/13.83	58.54/13.31	60.77/34.33	19.78/5.62	6.94/0.67
BCE	89.14/47.32	91.75/82.81	66.28/57.26	0.00/0.00	26.17/0.02	27.61/0.06	0.00/0.00	0.00/0.00
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39	60.63/**19.19**	62.39/**40.48**	**41.14/15.33**	**24.19/3.08**

DB improves classification performance even for tail labels

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF	PubMed Total miF/maF	PubMed Head(≥50) miF/maF	PubMed Med(15-50) miF/maF	PubMed Tail(≤15) miF/maF
SVM	87.60/51.63	89.87/78.47	66.92/61.00	22.54/13.83	58.54/13.31	60.77/34.33	19.78/5.62	6.94/0.67
BCE	89.14/47.32	91.75/82.81	66.28/57.26	0.00/0.00	26.17/0.02	27.61/0.06	0.00/0.00	0.00/0.00
FL	89.97/56.83	91.83/82.64	76.16/70.63	27.40/15.37	58.30/13.94	60.43/33.69	26.39/8.15	8.58/0.86
CB	89.23/52.96	91.56/80.44	71.64/66.61	23.08/9.93	58.57/13.67	60.75/33.40	24.50/7.39	9.92/1.01
R-FL	89.47/54.35	91.59/80.39	72.86/66.69	25.00/14.22	57.90/14.66	59.85/34.09	30.32/9.70	11.45/1.15
NTR-FL	90.70/60.70	92.37/82.65	79.35/75.34	39.51/22.33	60.92/16.99	**63.15**/38.85	33.14/11.39	15.86/1.82
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39	60.63/**19.19**	62.39/**40.48**	**41.14/15.33**	**24.19/3.08**

* FL: Focal loss, CB: Class balanced focal loss, R-FL: rebalanced focal loss, NTR-FL: negative tolerant regularization focal loss

DB achieves SOTA performance for Reuters

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF
SVM	87.60/51.63	89.87/78.47	66.92/61.00	22.54/13.83
BCE	89.14/47.32	91.75/82.81	66.28/57.26	0.00/0.00
FL	89.97/56.83	91.83/82.64	76.16/70.63	27.40/15.37
CB	89.23/52.96	91.56/80.44	71.64/66.61	23.08/9.93
R-FL	89.47/54.35	91.59/80.39	72.86/66.69	25.00/14.22
NTR-FL	90.70/60.70	92.37/82.65	79.35/75.34	39.51/22.33
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39

Our Reuters result outperforms prior work on this dataset, including approaches based on Binary Relevance, EncDec, CNN, CNN-RNN, Optimal Completion Distillation or attention-based GNN, that achieved **micro-F1<89.9** (Nam et al., 2017; Pal et al., 2020; Tsai and Lee, 2020)

* FL: Focal loss, CB: Class balanced focal loss, R-FL: rebalanced focal loss, NTR-FL: negative tolerant regularization focal loss

Results | Ablation study: DB = R + NTR + FL

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF	PubMed Total miF/maF	PubMed Head(≥50) miF/maF	PubMed Med(15-50) miF/maF	PubMed Tail(≤15) miF/maF
R-FL	89.47/54.35	91.59/80.39	72.86/66.69	25.00/14.22	57.90/14.66	59.85/34.09	30.32/9.70	11.45/1.15
NTR-FL	90.70/60.70	92.37/82.65	79.35/75.34	39.51/22.33	60.92/16.99	**63.15**/38.85	33.14/11.39	15.86/1.82
DB-0FL	89.45/57.98	91.21/82.05	77.33/71.11	31.17/19.05	58.95/15.15	60.99/34.92	31.06/10.02	14.23/1.49
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39	60.63/**19.19**	62.39/**40.48**	**41.14/15.33**	**24.19/3.08**

* FL: Focal loss, CB: Class balanced focal loss, R-FL: rebalanced focal loss, NTR-FL: negative tolerant regularization focal loss

A novel loss function, CB-NTR

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF	PubMed Total miF/maF	PubMed Head(≥50) miF/maF	PubMed Med(15-50) miF/maF	PubMed Tail(≤15) miF/maF
R-FL	89.47/54.35	91.59/80.39	72.86/66.69	25.00/14.22	57.90/14.66	59.85/34.09	30.32/9.70	11.45/1.15
NTR-FL	90.70/60.70	92.37/82.65	79.35/75.34	39.51/22.33	60.92/16.99	**63.15**/38.85	33.14/11.39	15.86/1.82
DB-0FL	89.45/57.98	91.21/82.05	77.33/71.11	31.17/19.05	58.95/15.15	60.99/34.92	31.06/10.02	14.23/1.49
CB-NTR	**90.74**/63.31	**92.46**/83.28	78.42/72.98	46.32/**32.31**	**61.07**/18.40	63.02/39.95	37.18/13.43	24.15/2.97
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39	60.63/**19.19**	62.39/**40.48**	**41.14/15.33**	**24.19/3.08**

Results | Effectiveness against the number of labels per instance

- For the Reuters dataset, we split the test instances into two groups, 2583 instances with only one label and 436 instances with multiple labels. On single-label instances, all functions from BCE to DB, have similar performance; while on multi-label instances, the performance of BCE drops more than DB. DB outperforms other functions in micro-F1 of the multi-label instance group and macro-F1 of both groups.
- There are < 0.1% instances of PubMed dataset with a single label, so we divide instances into 3-quantiles by their number of labels. In each quantile, the novel NTR-FL, CB-NTR and DB outperform the rest of the models in all metrics.

Loss Function	Reuters Total miF/maF	Reuters Single-label miF/maF	Reuters Multi-label miF/maF	PubMed Total miF/maF	PubMed $\leq$ 9 labels miF/maF	PubMed 10-14 labels miF/maF	PubMed $\geq$ 15 labels miF/maF
BCE	89.14/47.32	94.11/41.44	76.26/33.11	26.17/0.02	16.48/0.01	27.36/0.02	30.36/0.03
FL	89.97/56.83	94.81/50.33	77.54/40.07	58.30/13.94	53.72/7.44	59.02/10.27	59.72/8.63
CB	89.23/52.96	94.10/44.72	77.27/38.80	58.57/13.67	54.41/7.40	59.21/10.11	59.82/8.51
R-FL	89.47/54.35	95.21/47.45	74.29/38.79	57.90/14.66	53.08/7.67	58.60/10.50	59.45/8.81
NTR-FL	90.70/60.70	**95.42**/51.33	78.85/44.37	60.92/16.99	**58.51**/9.07	**61.86**/12.31	61.12/10.20
DB-0FL	89.45/57.98	94.48/51.80	76.63/42.26	58.95/15.15	55.14/8.11	59.84/10.90	59.85/8.94
CB-NTR	**90.74**/63.31	95.17/51.08	79.56/49.94	**61.07**/18.40	58.29/9.67	61.72/12.97	**61.72**/10.77
DB	90.62/**64.47**	94.49/**54.31**	**81.17/50.12**	60.63/**19.19**	57.81/**9.76**	61.53/**13.49**	61.08/**11.23**

Results | Error Analysis

- The most common errors are due to incorrect classification to similar or linked labels for all loss functions.
- The most common three pairs of classes confused by all loss functions for the Reuters dataset are: *platinum* and *gold, yen* and *money-fx, platinum* and *copper*.
- For the PubMed dataset, the most common errors were: *Pandemics* and *Betacoronavirus, Pandemics* and *SARS-CoV-2, Pneumonia, Viral* and *Betacoronavirus*, and BCE has significantly more errors for these classes compared to the other investigated loss functions.

3 take aways

- The thread that binds all of my past work is mathematics. At the end of the day, it's all about solving a system of equations!
- The big challenge is finding the needles in a haystack. Most interesting biology occurs at the edge of the distribution rather than the average state. Similarly, finding the information that is rare is sometimes more valuable than finding common knowledge.
- ML, DL approaches are becoming increasingly powerful at making out of distribution predictions.