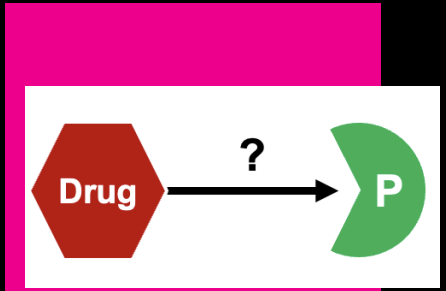
MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA



Extracting Drug-Protein Relation from Literature using Ensembles of Biomedical Transformers

@HuaXu_SBMI

Dr. Hua Xu

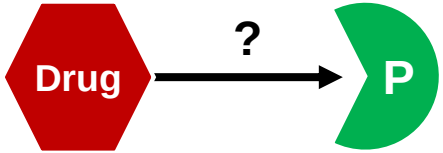
Professor

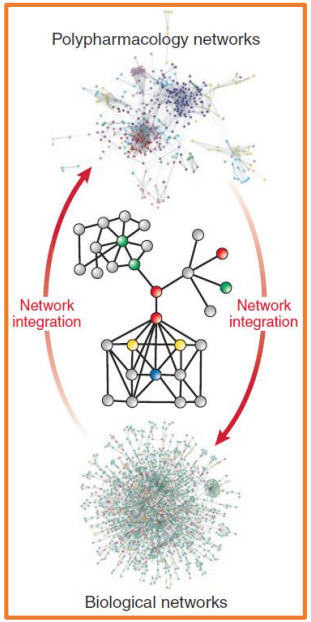
Yale School of Biomedicine

AIDH IMIA

Motivation

- Network pharmacology has been used to accelerate the process of drug discovery
- Drug target databases
  - DrugBank, TTD, KEGG, CTD, DGIdb, DrugCentral, STITCH, TDR Targets, Drug Target Commons
- Mode of action (MoA) is usually unknown.
- Automated methods to extract **drug-protein relations** from biomedical literature

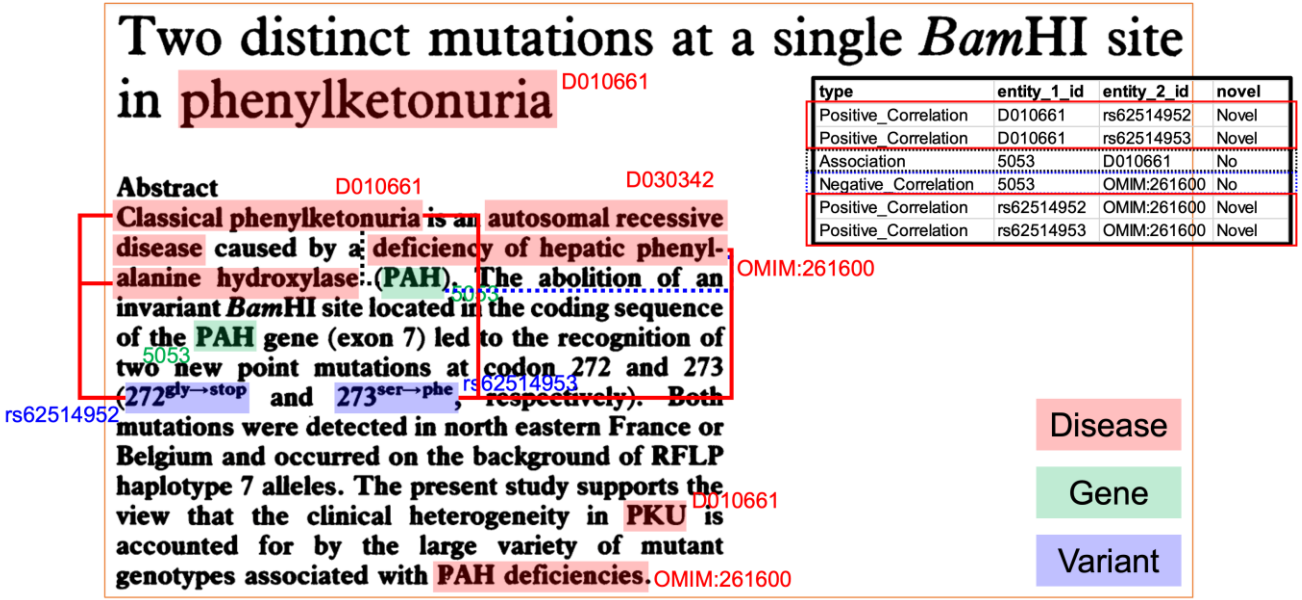




Biocreative VII Challenge

- Participated in the Track 1 – DrugProt Challenge
- **Goal**
  - Extract **Drug-Protein relations from biomedical literature**
- **Importance**
  - Quick identification of potential treatments
  - Repurposing existing drugs for illnesses

Example of Named Entities and Relations



type	entity_1_id	entity_2_id	novel
Positive_Correlation	D010661	rs62514952	Novel
Positive_Correlation	D010661	rs62514953	Novel
Association	5053	D010661	No
Negative_Correlation	5053	OMIM:261600	No
Positive_Correlation	rs62514952	OMIM:261600	Novel
Positive_Correlation	rs62514953	OMIM:261600	Novel

Source: https://pubmed.ncbi.nlm.nih.gov/1671881/

Brief Overview of DrugProt Task

- **Sub Tracks**
  - Main SubTrack DrugProt Corpus

	Training	Development	Test
Documents	3,500	750	10,750
Tokens	1001168	199620	182908
Entities, Relations	89529, 17288	18858, 3765	289664, -

  - Large Scale SubTrack
    - ➢ 2,366,081 records (2.4 Million) are provided
    - ➢ Total of 53,993,602 entity annotations

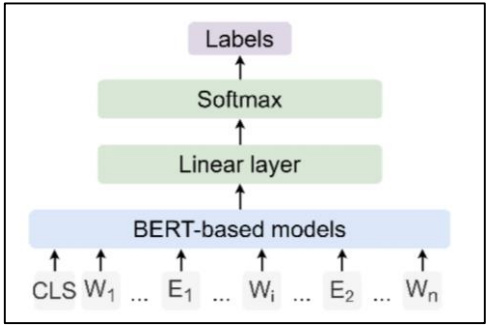
Data Preprocessing

- **Preprocessing**
  - ➢CLAMP Tool[1] for sentence boundary detection.
- **Training Set/Cross Validation**
  - ➢Pooled the Training and Development (3,500 + 750 abstracts)
  - ➢Randomly split them into ten folds
    - Each set 425 abstracts
  - ➢Acts as its own development set during training

[1] "CLAMP – A Toolkit for efficiently building customized clinical natural language processing pipelines," E. Soysal, et al., *Journal of the American Medical Informatics Association*

Biomedical BERT-based Models

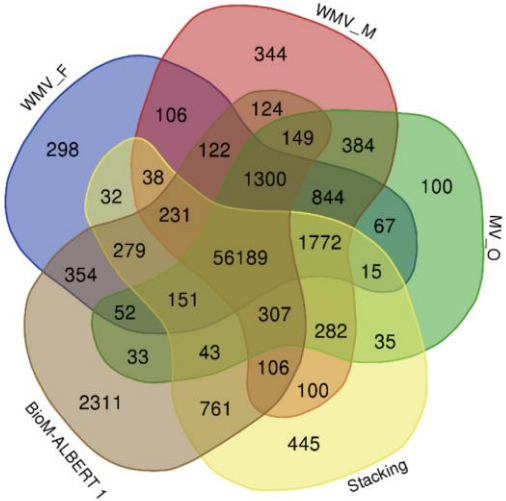
Model _ID	Model Name	Training Details
1, 2	**BioM-ALBERT$_{xxlarge}$**	Pre-trained on PubMed abstracts and PMC full article
3	**BioBERT$_{large}$**	BERT weights fine-tuned on PubMed abstracts and PMC full article
4	**PubMedBERT**	Pre-trained on PubMed abstracts
5	**BioM-BERT**	Trained on PubMed abstracts and PMC



BERT-based models for Relation Extraction

Ensemble Learning

- Majority Voting
  - ➢ **Model-first**
  - ➢ **Fold-first**
  - ➢ **Overall**
- Weighted Majority Voting
  - ➢ Assigned each vote **a different weight**
  - ➢ Based on performance of its relation type in different training sets
- Stacking
  - ➢ A trained **J48 decision tree on results from the five BERT models**



Overlaps among the five submissions for the test set.
WMV_F: fold-first weighted majority voting;
WMV_M: model-first weighted majority voting;
MV_O: overall majority voting

Results

- Performance (F1-score) of the models on each development set during training.

Model Type	Model	Development sets										Overall
		1	2	3	4	5	6	7	8	9	10	
Deep Learning (BERT-based)	BioBERT (Bio-B)	0.766	0.760	0.717	0.803	0.707	0.771	0.759	0.740	0.724	0.756	0.753
	BioM-ALBERT 1 (BioM-AB1)	**0.775**	0.784	0.711	0.828	0.719	0.806	0.786	0.777	0.739	**0.811**	0.777
	BioM-ALBERT 2 (BioM-AB2)	0.768	0.776	0.718	0.828	0.702	0.773	0.845	0.767	0.742	0.763	0.769
	BioM-BERT (BioM-B)	0.760	0.776	0.731	0.828	0.688	0.767	0.794	0.785	0.754	0.785	0.769
	PubMedBERT (PM-B)	0.761	0.778	0.708	0.815	0.705	0.766	0.796	0.797	0.720	0.776	0.765
Ensemble Learning	**Majority voting (MV)**	0.771	**0.806**	0.757	**0.844**	0.728	**0.808**	**0.818**	0.787	**0.763**	0.805	0.791
	Weighted majority voting (W-MV)	0.774	**0.806**	**0.759**	**0.844**	**0.730**	**0.808**	**0.818**	0.789	**0.763**	0.806	**0.792**
	Stacking (St)	0.767	0.797	0.735	0.838	0.717	0.762	0.789	**0.811**	0.750	0.797	0.779

Results

- Performance on the Test Sets for Main Track and Large-Scale Track

Run_ID	Main Track Test Set				Large-Scale Test Set			
	Run	P	R	F1	Run	P	R	F1
1	Voting/FM	0.795	0.750	0.772	BioM-AB1	0.764	0.714	0.738
2	**Voting/MF**	**0.804**	**0.750**	**0.776**	Stacking	0.776	0.747	0.761
3	Voting	0.800	0.746	0.772	**Voting/FM**	**0.795**	**0.753**	**0.773**
4	Stacking	0.800	0.733	0.765	Voting/MF	0.801	0.746	0.773
5	BioM-AB 1	0.797	0.753	0.775	Voting	0.797	0.749	0.772

Results

- Performance on the Main (MT) and Large-Scale (LS) test sets by Relation

Relation-Type	BioM- ALBERT 1		Stacking		Voting w FM		Voting w MF		Voting	
	MT	LS	MT	LS	MT	LS	MT	LS	MT	LS
ACTIVATOR	0.81	0.74	0.79	0.81	0.82	0.81	0.81	0.81	0.82	0.81
AGONIST	0.78	0.70	0.78	0.75	0.78	0.79	0.78	0.78	0.77	0.79
AGONIST-INHIBITOR	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00
ANTAGONIST	0.91	0.84	0.90	0.87	0.90	0.90	0.90	0.89	0.90	0.90
DIRECT-REGULATOR	0.67	0.65	0.66	0.69	0.68	0.69	0.68	0.69	0.68	0.68
INDIRECT-DOWNREGULATOR	0.76	0.74	0.75	0.76	0.77	0.77	0.78	0.77	0.78	0.77
INDIRECT-UPREGULATOR	0.77	0.74	0.76	0.74	0.76	0.76	0.77	0.76	0.76	0.76
INHIBITOR	0.87	0.84	0.85	0.85	0.85	0.85	0.86	0.85	0.86	0.85
PART-OF	0.68	0.67	0.69	0.69	0.69	0.71	0.70	0.70	0.70	0.71
PRODUCT-OF	0.70	0.63	0.67	0.60	0.69	0.67	0.68	0.67	0.69	0.68
SUBSTRATE	0.65	0.60	0.65	0.65	0.65	0.66	0.67	0.66	0.64	0.66
SUBSTRATE_PRODUCT-OF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Competition Performance

4th

2nd

TABLE III. TEAM OVERVIEW, MAIN TRACK, AND LARGE SCALE MICRO-AVERAGE RESULTS

#	Team	Affiliation	Ref	Tool URL	Main Track				Large Scale Track			
					P	R	F1	run	P	R	F1	run
15	Humboldt	Humboldt-Universität Berlin, Germany	1	20	0.7961	0.7986	0.7973	1				
18	NLM-NCBI	National Institutes of Health, USA	2		0.7847	0.8052	0.7948	5	0.7730	0.8049	0.7886	2
13	KU-AZ	Korea University, AstraZeneca, AIGEN Sciences, South Korea, UK	3		0.7972	0.7817	0.7894	2	0.7644	0.7521	0.7582	2
7	UTHealth-CCB	University of Texas, USA	4		0.8044	0.7496	0.7760	2	0.7949	0.7527	0.7732	3

Source: Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug - gene/ protein relations, *Miranda et al.*

Summary

- Domain-specific transformer models achieve reasonable scores
- Our best system performances
  - ➢**Main subtrack** – Micro-averaged F1-score of 77.60%
  - ➢**Large scale subtrack** - Micro-averaged F1-score of 77.32%
- Ensemble Modeling enhances the effectiveness of deep learning-based approaches
  - ➢**Model-first Weighted Majority Voting** performs best for main subtrack
  - ➢**Fold-first Weighted Majority Voting** performs best for large scale subtrack

Acknowledgements

The work was a conducted in a group of the following members

- Dr. Liang-Chin Huang
- Dr. Zhao Li
- Dr. Qiang Wei
- Dr. Jianfu Li
- Dr. Avisha Das
- Yan Hu
- Rongbin Li
- Dr. W. Jim Zheng
- Dr. Hua Xu

Thank you!
Questions?

Additional Slides

- Detailed DrugProt Relations
- 12 types of interactions
  - ACTIVATOR, AGONIST, AGONIST- INHIBITOR,, ANTAGONIST, DIRECT-REGULATOR, INHIBITOR, INDIRECT-DOWNREGULATOR, INDIRECT-UPREGULATOR, PART-OF, PRODUCT-OF, SUBSTRATE, SUBSTRATE_ PRODUCT-OF

Relation type	Nr. relations		
	Training	Development	Test
ANTAGONIST	1428	246	334
AGONIST	658	131	101
AGONIST-INHIBITOR	29	10	0
DIRECT-REGULATOR	13	2	3
ACTIVATOR	972	218	154
INHIBITOR	2247	458	429
INDIRECT-DOWNREGULATOR	1329	332	304
INDIRECT-UPREGULATOR	1378	302	277
PART-OF	5388	1150	1051
PRODUCT-OF	885	257	228
SUBSTRATE	920	158	181
SUBSTRATE_PRODUCT-OF	2003	494	419

Ensemble Learning

- **Majority Voting**
  - **Model-first**:
    - Pooled the results from the ten folds for each model, kept relations >= 5 votes
    - Pooled the voting results from the five models, kept relations >= 3 votes
  - **Fold-first:**
    - Pooled the results from the five models for each fold, kept relations >= 3 votes
    - Pooled the voting results from the ten folds, kept relations >= 5 votes
  - **Overall:**
    - Pool the 50 prediction results from the five models for the ten folds
    - Then kept the relations >= 25 votes.

Ensemble Learning

- **Weighted Majority Voting**
  - Assigned each vote **a different weight** based on performance of its relation type in different training sets
    - **Model-first**:

$$\omega_{rf} = \frac{\sum_{m=1}^{5} p_{mrf} F_{mrf}}{\sum_{m=1}^{5} F_{mrf}},$$

    - **Fold-first:**

$$\omega_{mr} = \frac{\sum_{f=1}^{10} p_{mrf} F_{mrf}}{\sum_{f=1}^{10} F_{mrf}},$$

    - **Overall:**

$$\omega_{r} = \frac{\sum_{m=1}^{5} \sum_{f=1}^{10} p_{mrf} F_{mrf}}{\sum_{m=1}^{5} \sum_{f=1}^{10} F_{mrf}},$$

Ensemble Learning

- **Stacking**
  - Using the **prediction results from the five BERT-based models** as binary features
    - 0 for negative and 1 for positive
  - Trained **a J48 decision tree** using WEKA
  - Pooled the results from each training set
  - Kept the relations having >= 5 votes.