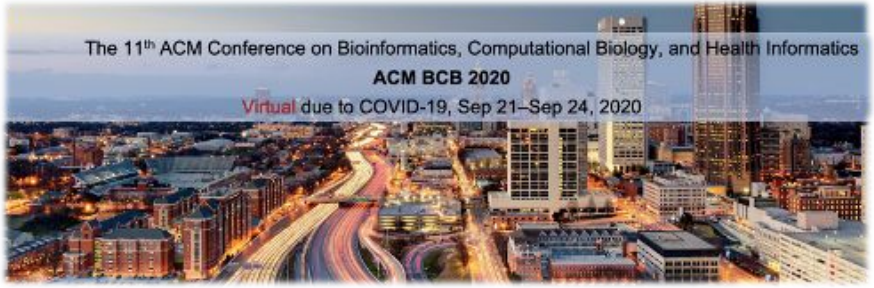


Bio-JOIE: Joint Representation Learning of Biological Knowledge Bases

Presenter: Junheng Hao

University of California, Los Angeles

Association for Computing Machinery

SIGBio Bioinformatics, Computational Biology and Biomedical Informatics

UCLA Samueli Computer Science

Outline

- Background: Cross-domain Biological Knowledge Graphs
- Bio-JOIE Modeling
- Experimental Results & Case Study
- Conclusion & Future Work

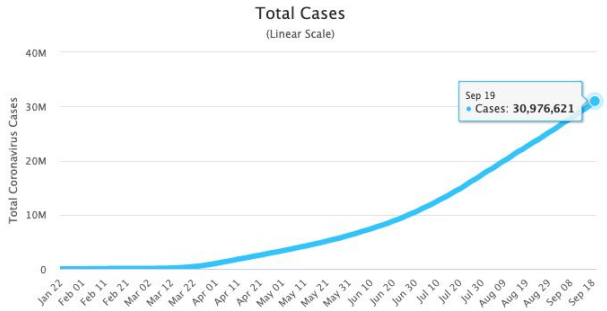
Outline

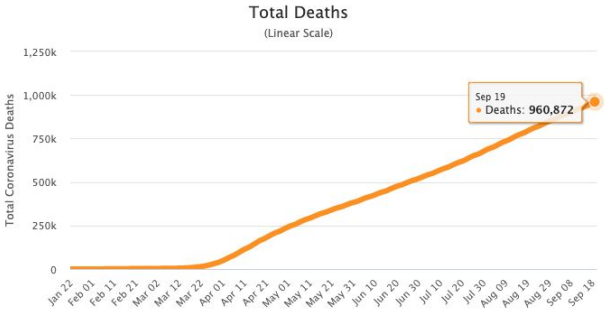
UCLA ScAi

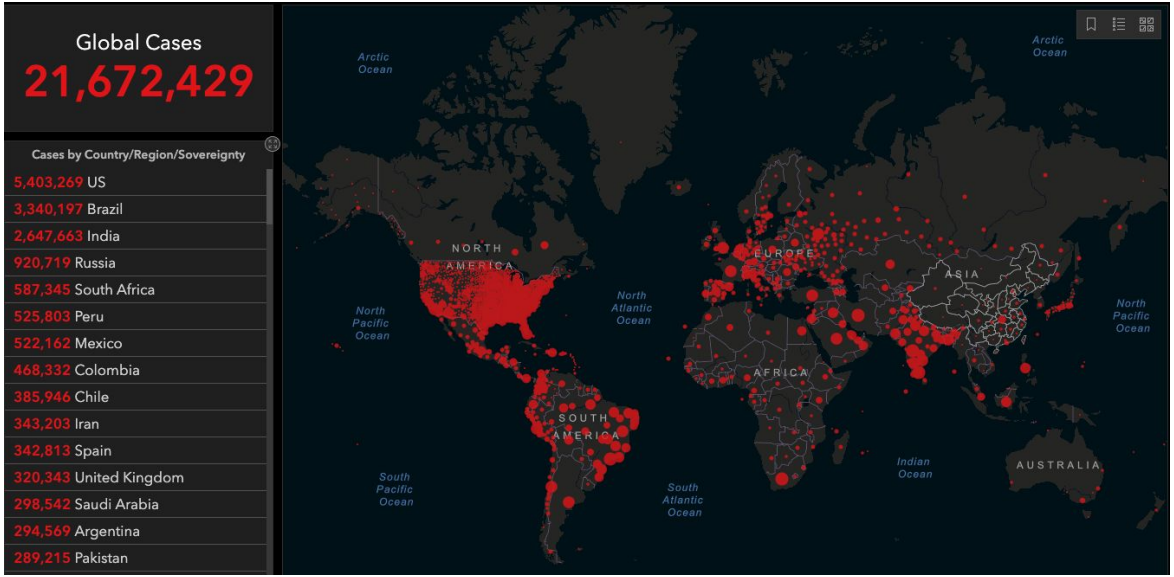
- **Background: Cross-domain Biological Knowledge Graphs**
- Bio-JOIE Modeling
- Experimental Results & Case Study
- Conclusion & Future Work

Fight of COVID-19

- The outbreak of COVID-19 has infected over 21 millions of people and caused high death tolls since the end of 2019, as worldwide social and economic disruption.



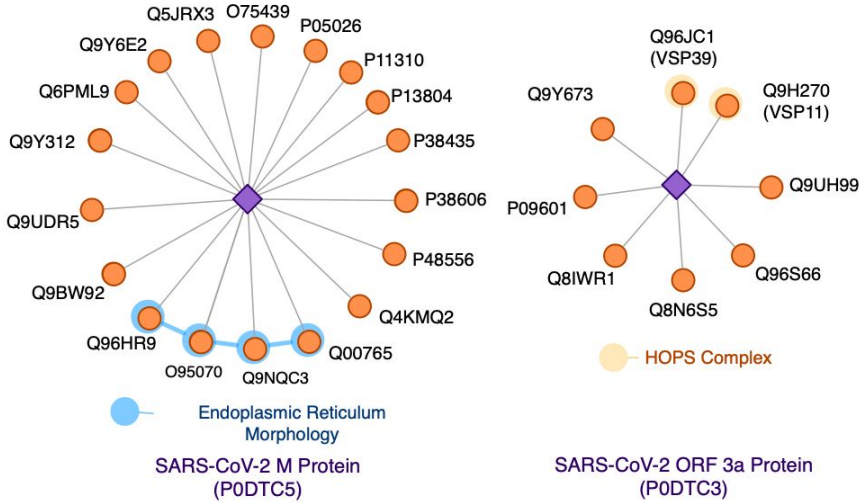




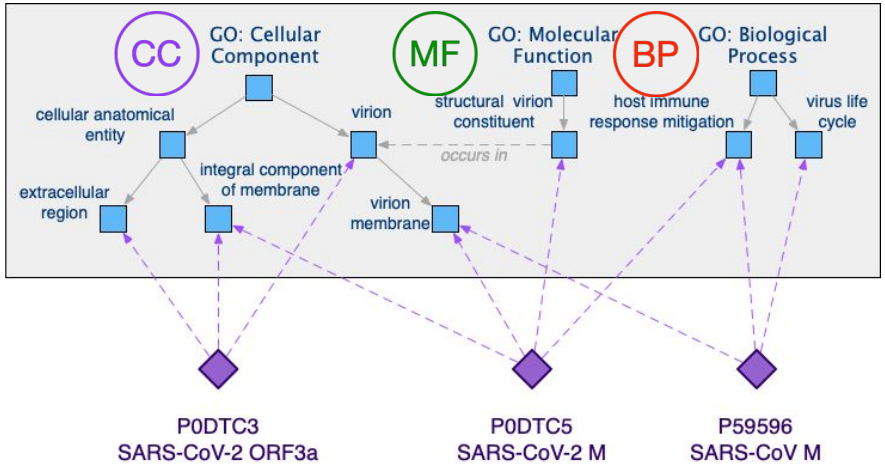
Data source: COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), updated as 11pm (PST), 08/16/2020. Website: https://coronavirus.jhu.edu/map.htm

SARS-CoV-2 Knowledge Graph

- Tremendous efforts have been made to discover the infection mechanism of its causative agent, named SARS-CoV-2, from multiple perspectives.
- Knowledge about SARS-CoV-2: (1) Protein-protein interactions (PPI) between viral proteins and human co-host proteins; (2) Gene ontology (GO) annotations on viral proteins.



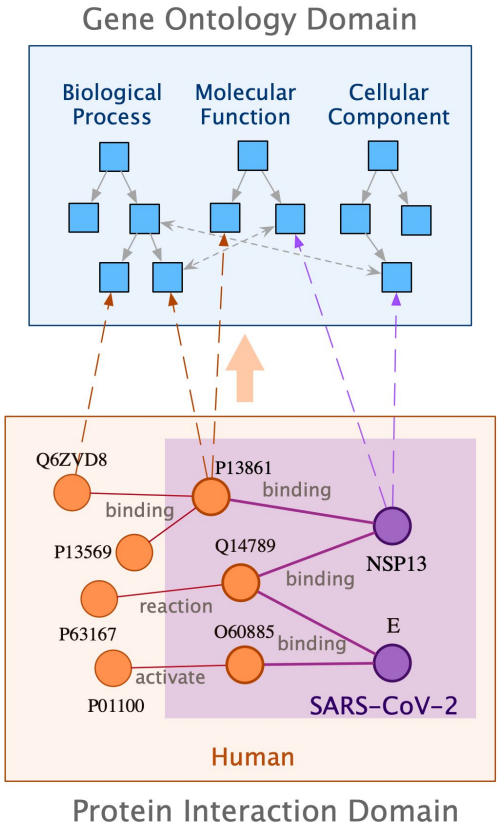
PPI of SARS-CoV-2 M and ORF3a Proteins [Gordon et al.]



GO annotations on SARS-CoV-2 M/ORF3a and SARS-CoV M

Cross-domain Bio-KG

- Not limited to SARS-CoV-2 KB above, many biological KBs, often stored as knowledge graphs (KGs), consist of biological entities of various kinds, together with their properties and relations.
- As essential sources of knowledge, these KBs can be categorized in different domains.
  - Gene Ontology Consortium → gene function annotation
  - STRING → Knowledge accumulated from functional proteomic analysis
  - DrugBank → cheminformatics resource for drug targets
- Domain-specific knowledge is often scarce and costly to collect.



Learning Embeddings for Bio-KG

UCLA ScAi

Input
Knowledge Graph(s)

MeSH

STRING

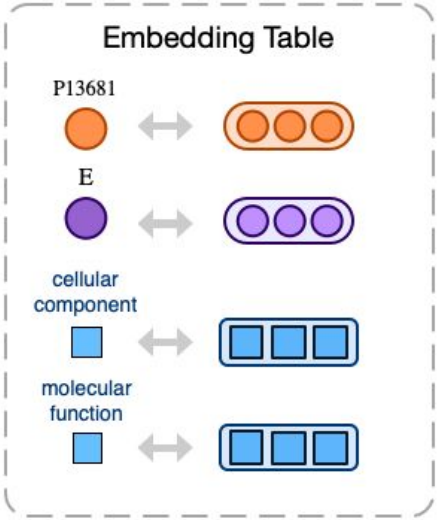
neXtprot

PubMed

GENEONTOLOGY
Unifying Biology

DRUGBANK

Output
Embeddings



Inference & Applications

PPI Type prediction

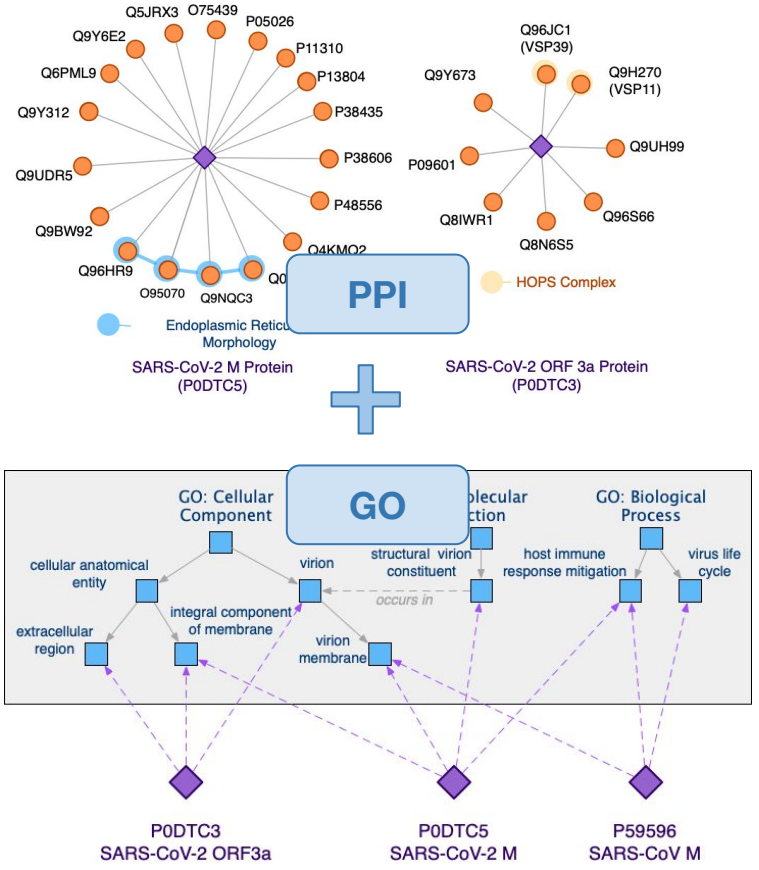
PPI Classification

GO Annotations

...

Motivation

- Relying on the KG from a single domain presents the risk of learning from limited and scarce information.
- The missing knowledge in one domain can be transferred from other domains with complementary knowledge, and thus provide more comprehensive representations of the biological entities.
- We aim at designing a plausible method to support the fusion and transfer of knowledge across multiple biological domains.

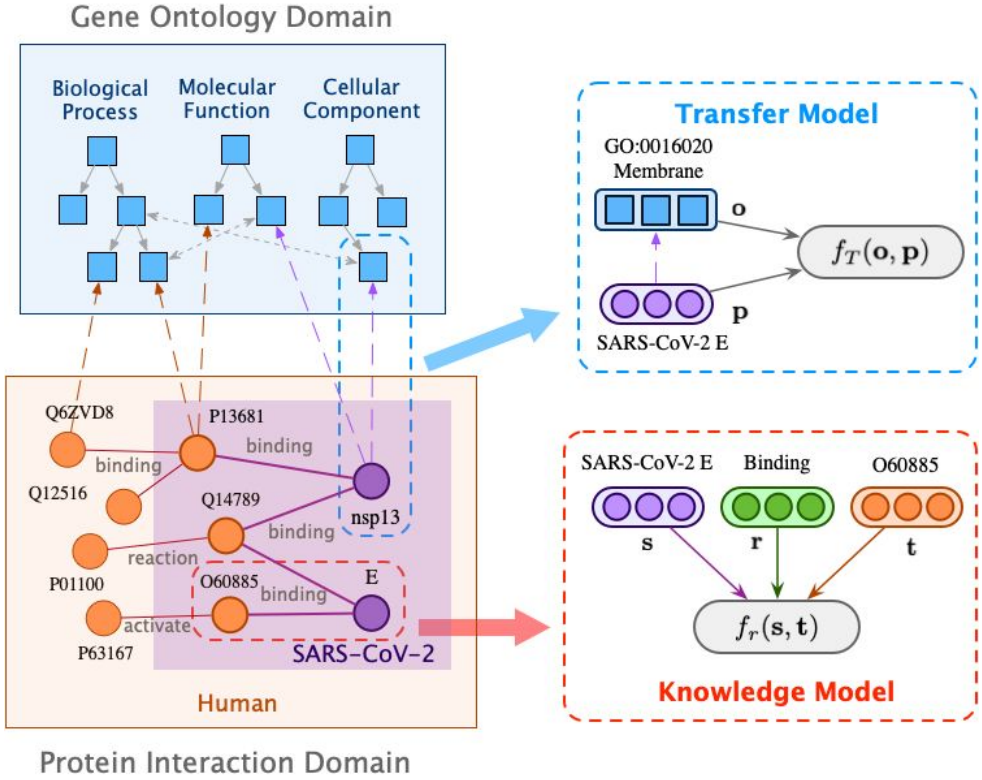


Outline

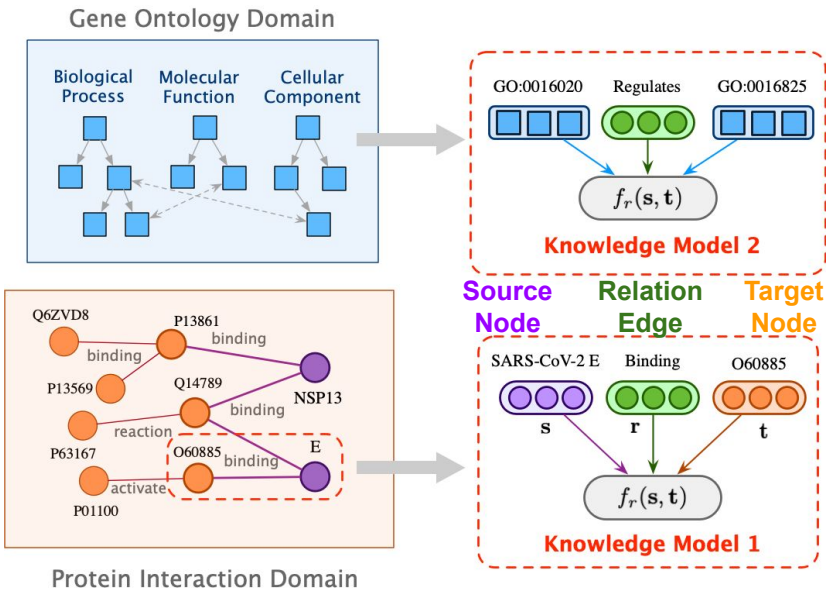
- Background: Cross-domain Biological Knowledge Graphs
- **Bio-JOIE Modeling**
- Experimental Results & Case Study
- Conclusion & Future Work

Model Overview: Bio-JOIE

Joint **E**mbedding Learning for multiple domains of **Bio**logical Knowledge Bases



Bio-JOIE: Knowledge Model



Knowledge Triple → {**S**ource, **R**elation, **T**arget}

- **Goal:** To embed the relational facts / structures in the each domain of the Bio-KG
- Three representative score functions: TransE, DistMult (selected), and HolE

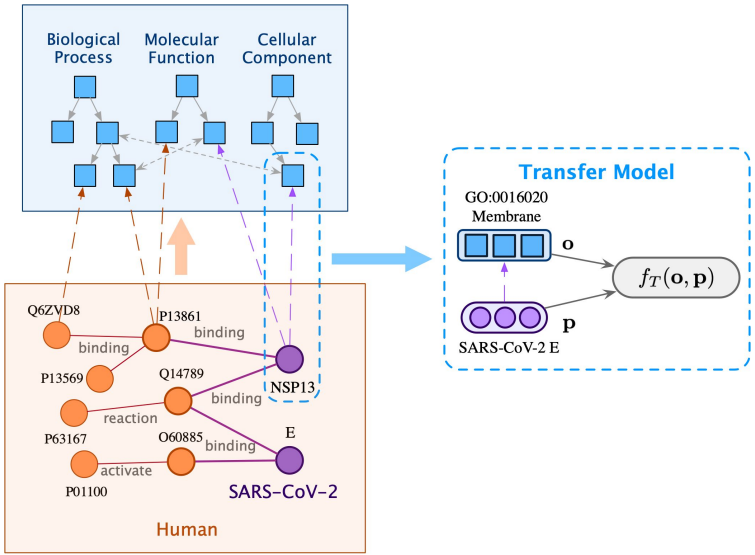
$$f_r^{\text{Trans}}(\mathbf{s},\mathbf{t}) = ||\mathbf{s}+\mathbf{r}-\mathbf{t}||_2$$
$$f_r^{\text{Mult}}(\mathbf{s},\mathbf{t}) = -(\mathbf{s}\circ\mathbf{t})\cdot\mathbf{r}$$
$$f_r^{\text{HolE}}(\mathbf{s},\mathbf{t}) = -(\mathbf{s}\star\mathbf{t})\cdot\mathbf{r}$$

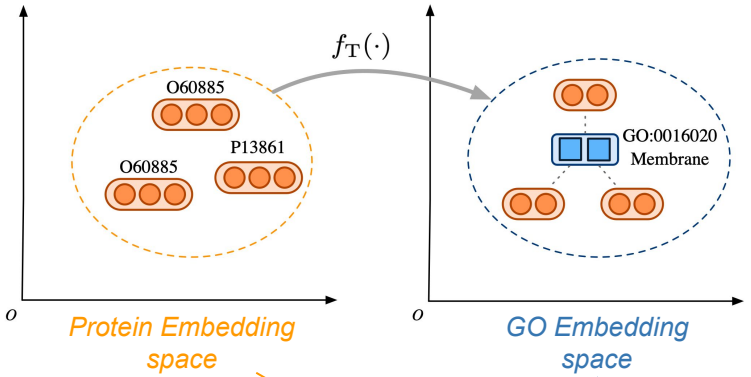
- **Loss:** Triple-wise margin ranking loss with negative sampling

$$\mathcal{L}_K^{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{(s,r,t)\in\mathcal{G}} \max\left\{f_r(s,t) + \gamma^{\mathcal{G}} - f_r(s',t'),\ 0\right\}$$

Positive triples *Negative triples*

Bio-JOIE: Transfer Model





- Domain transfer function:

$$f_T(o,p) = \|\sigma\left(\mathbf{M}_T \cdot \mathbf{p} + \mathbf{b}_T\right) - \mathbf{o}\|_2$$

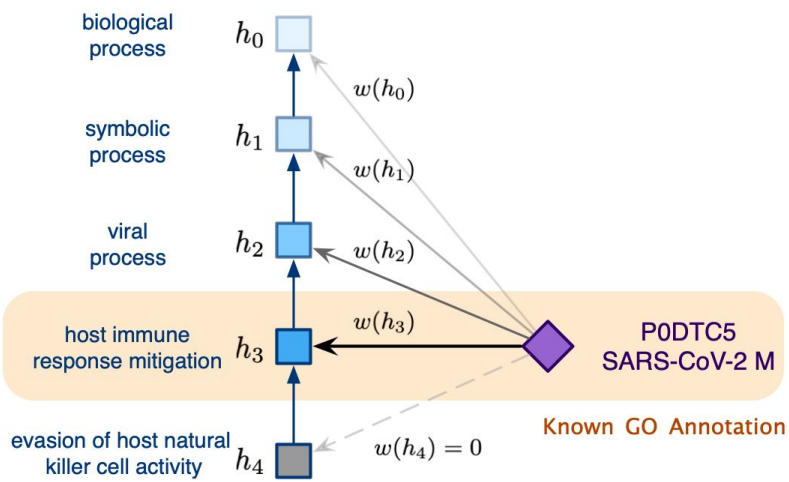
- Loss function given alignment pairs

$$\mathcal{L}_{T_1} = \frac{1}{|\mathcal{A}|} \sum_{(o,p)\in\mathcal{A}} \max\left\{f_{T_1}(o,p) + \gamma^{\mathcal{A}} - f_{T_1}(o',p'),\ 0\right\}$$

Positive alignments

Negative alignments

Bio-JOIE: Transfer Model (Weighted)



- Domain transfer function:

$$f_T(o,p) = \|\sigma\left(\mathbf{M}_T \cdot \mathbf{p} + \mathbf{b}_T\right) - \mathbf{o}\|_2$$

- Weighted alignments by GO term specificity
  - Options: Level weighted or degree weighted

$$\omega_1(o) = \frac{l}{l_{\max}}, \quad \omega_2(o) = \frac{1}{d(o)}$$

By level *By degree*

- Loss function given weighted alignment pairs

$$\mathcal{L}_{T_2} = \frac{1}{|\mathcal{A}|} \sum_{(o,p)\in\mathcal{A}} \max\left\{\frac{\omega(o)}{C}\left[f_{T_2}(o,p) + \gamma^{\mathcal{A}} - f_{T_2}(o',p')\right], 0\right\}$$

weighted alignments

Bio-JOIE: Joint Training & Inference

- Joint learning on two knowledge models in protein and GO and one transfer model (one species)

$$\mathcal{L} = \lambda^t \mathcal{L}_T + \lambda^p \mathcal{L}_K^{\mathcal{G}_p} + \mathcal{L}_K^{\mathcal{G}_o}$$

- In case of multiple species with one universal gene ontology, Bio-JOIE can apply one knowledge model and one transfer model, designated for each species → “Multi-way”

$$\mathcal{L} = \sum_{i=1}^{m} \lambda_i^t \mathcal{L}_T + \sum_{i=1}^{m} \lambda_i^p \mathcal{L}_K^{\mathcal{G}_p} + \mathcal{L}_K^{\mathcal{G}_o}$$

Outline

- Background: Cross-domain Biological Knowledge Graphs
- Bio-JOIE Modeling
- **Experimental Results & Case Study**
- Conclusion & Future Work

Datasets

- Protein-Protein Interaction Networks
  - 4 types of PPIs on three species (human, yeast and fly) from STRING database [1]
- Gene Ontology Datasets
  - Extracted from Gene Ontology Consortium [2]
  - Total 6 relations, such as is_a, part_of, regulates, etc.
  - Three aspects: biological process (BP), cellular components (CC), and molecular function (MF)

Table: Statistics of PPI Networks and GO annotations from 3 species

Species	# Proteins	# PPI Triples	# GO Annotations
Yeast	3,736	21,704	191,801
Fly	3,826	10,000	87,807
Human	8,204	36,400	102,759

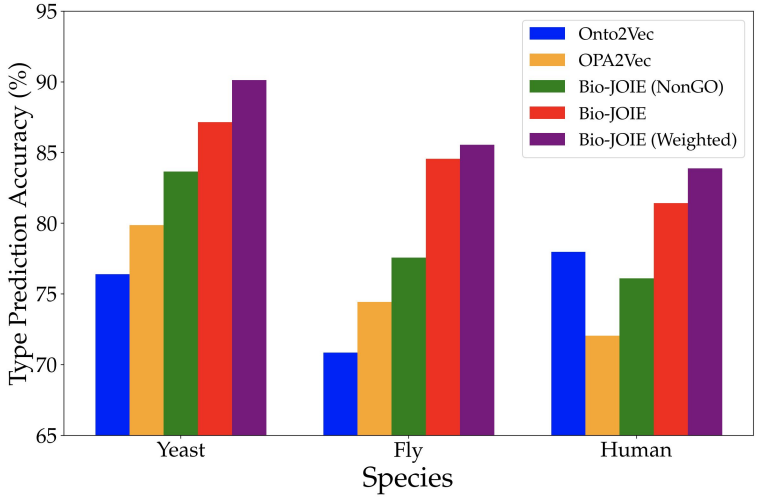
Table: Statistics of 3 aspects in GO

Aspects	BP	CC	MF
# GO entities	5744	1,147	1,764
# GO triples	19,021	2,116	2,190
# Protein-GO annotations (yeast)	72,956	58,729	60,116
# Protein-GO annotations (fly)	44,605	24,550	18,652
# Protein-GO annotations (human)	42,899	32,929	26,931

Experiment: PPI Type Prediction

Performance on PPI type prediction for three different species

- **Task:** Interaction type prediction given pairs of proteins
- **Evaluation metric:** Prediction accuracy
- **Baselines:** Onto2Vec (variants: Parent, Ancestor, Sum, Mean) , OPA2Vec, Bio-JOIE (NonGO)



Observation: Bio-JOIE outperforms all baselines in PPI type prediction on three different species.

Experiment: PPI Type Prediction

Effects of different aspects of GO terms on PPI type prediction

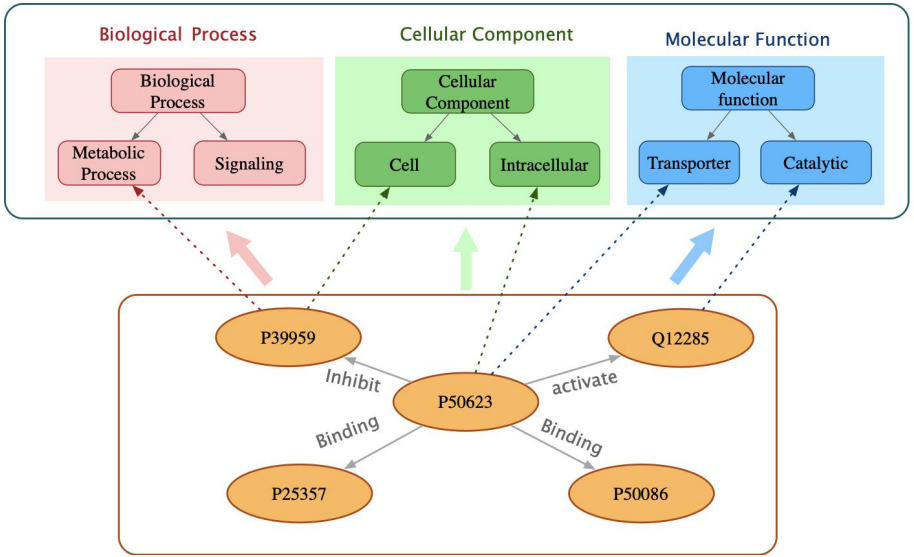


Table: Comparison of Bio-JOIE performance on combinations of three different aspects in GO.

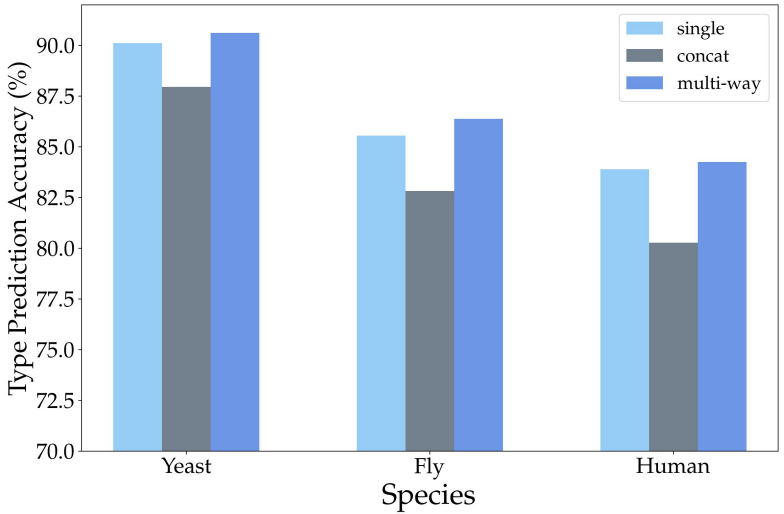
#	Aspects	Yeast	Fly	Human
1	BP	0.8794	0.8402	0.8153
	CC	0.8499	0.8272	0.8054
	MF	0.8539	0.8386	0.8165
2	BP+CC	0.8717	0.8473	0.8271
	BP+MF	0.8673	0.8471	0.8163
	CC+MF	0.8569	0.8466	0.8170
3	AllGO	**0.9012**	**0.8555**	**0.8389**

Observation: All aspects has help better predict PPI (among all three, BF contributes the most), which results that AllGO is the best performed variant.

Experiment: PPI Type Prediction

Joint-training from multiple species further benefits PPI type prediction

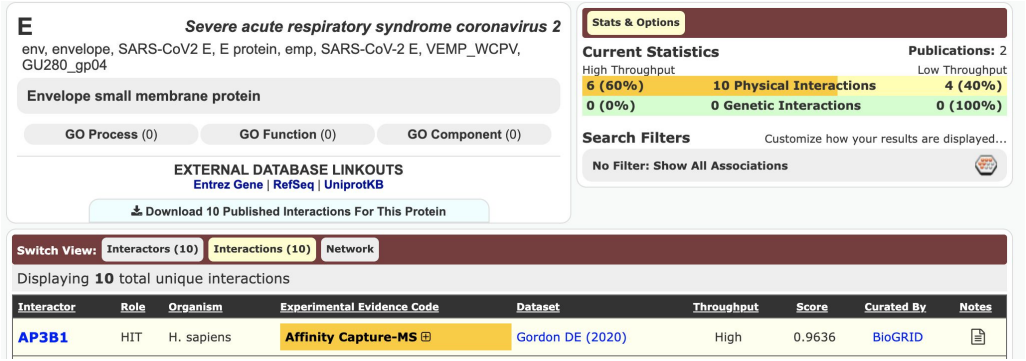
- **Model Setting:**
  - **Single** (Train and test on one species; no joint training)
  - **Concat** (Simply concatenate all PPI networks into one; one knowledge and transfer model for all three species)
  - **Multi-way** (Each species of the three has one knowledge model and transfer model to GO)



Observation: Joint training Bio-JOIE from multiple species can further benefit PPI type prediction in each of the involved species.

Case Study: SARS-CoV-2 Prediction

- Data sources for SARS-CoV-2
  - Data collected from BioGrid[1] on SARS-CoV-2, SARS-CoV and MERS-CoV
  - **26** SARS-CoV-2 proteins (associated with **282** GO terms) and **30** viral proteins of SARS-CoV and MERS-CoV (associated with **630** GO terms)
  - **332** highly possible interactions of SARS-CoV-2 identified by Gordon et al [2] and **1131** virus-human pairs with low MIST scores, as negative examples.
- All processed datasets are available at https://www.haojunheng.com/project/goterm.



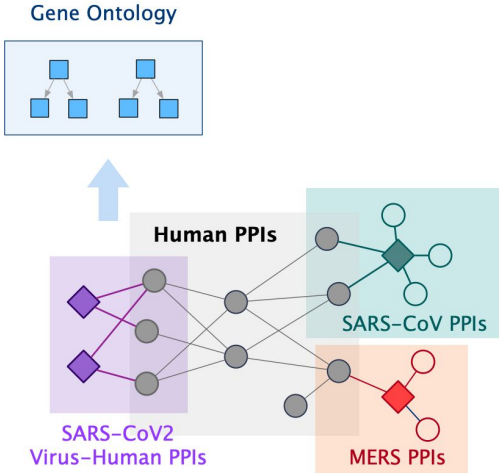
References:

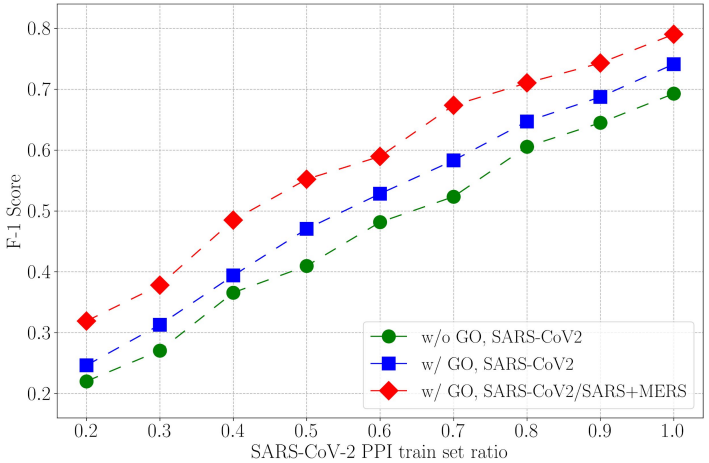
[1] BioGrid: https://wiki.thebiogrid.org/doku.php/covid, COVID-19 Related GO Terms: http://geneontology.org/covid-19.html

[2] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature (2020), 1–13.

SARS-CoV-2 Human PPI Classification

- **Binary classification:** Predict whether pairs of proteins (viral protein and human target protein) interact with each other or not.
- **Evaluation:** Compare F-1 Score on different Bio-JOIE variants



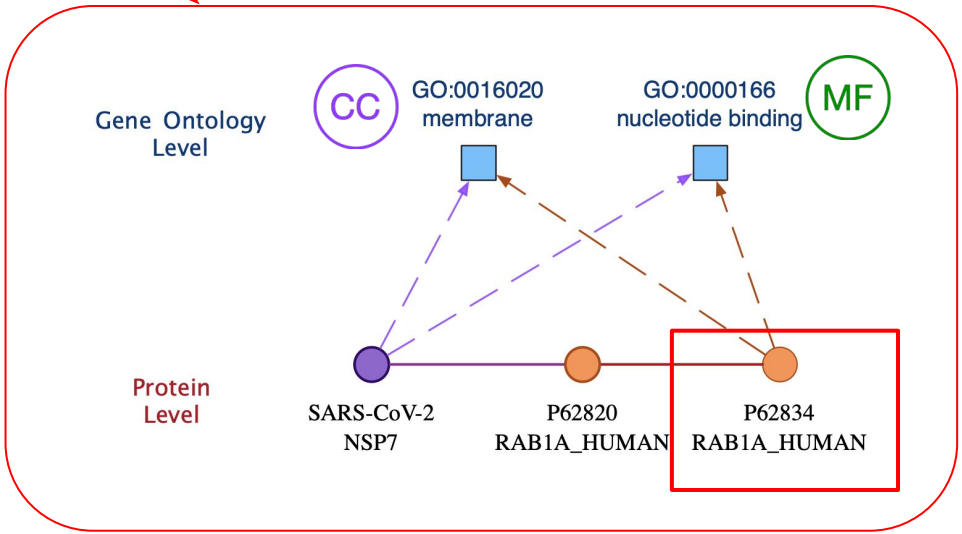


Observation: (1) Using GO provide additional knowledge about SARS-COV-2 proteins and better improve the classification performance; (2) Bio-JOIE can utilize the PPI information from similar coronavirus and further enhance the classification ability.

Example: Targeted Protein Prediction

SARS-CoV-2 Protein	Top Predicted Human Target Proteins
NSP7	P62834(0.685), P51148(0.879), P62070(0.418), P67870, O14578, Q8WTV0(0.854), P53618(0.350), Q9BS26, O94973, Q7Z7A1

Inside the Bio-KG



Outline

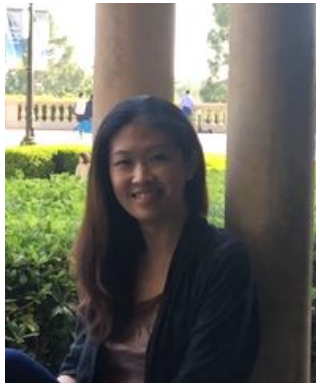
- Background: Cross-domain Biological Knowledge Graphs
- Bio-JOIE Modeling
- Experimental Results & Case Study
- **Conclusion & Future Work**

Conclusion & Future Work

- Bio-JOIE enables representation learning for cross-domain biological KBs capturing **in-domain interaction information** and **cross-domain knowledge transfer**.
- Bio-JOIE leverages **cross-domain complementary knowledge** to achieve SOTA performance on **PPI type prediction** and **clustering tasks**.
- As one important application, Bio-JOIE helps predict **human protein targets of SARS-CoV-2**, potentially benefits de novo drug discovery and disease mitigation.
- Some future directions:
    a. Incorporating multimodal features and gene annotations
    b. Extending Bio-JOIE to interconnected domains other than protein and gene ontology.

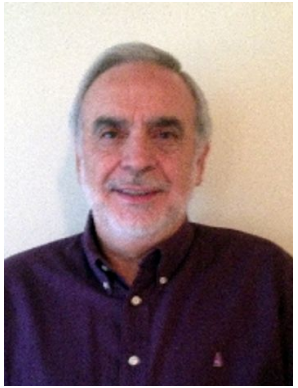
Acknowledgement

Muhao Chen
USC



Chelsea Ju
UCLA

Yizhou Sun
UCLA



Carlo Zaniolo
UCLA

Wei Wang
UCLA

For more information, please check our BCB paper and webpage!

Paper Link: To be updated
Video Link: To be updated
Project webpage: https://www.haojunheng.com/project/goterm/

Association for Computing Machinery

Samueli Computer Science

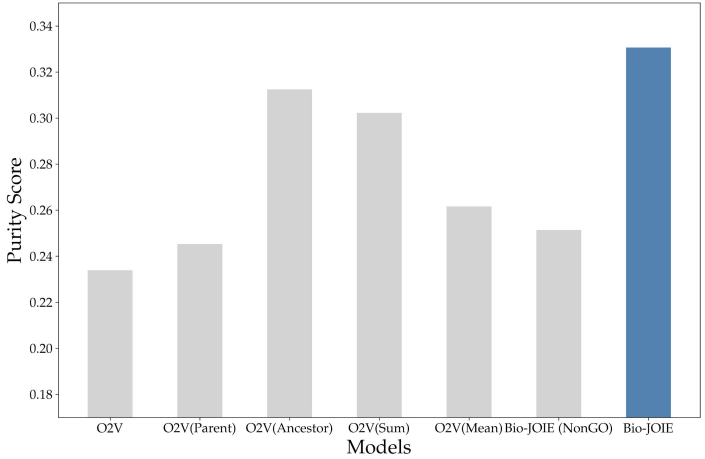
Thank you!

Q & A

Experiment: Enzyme-based Clustering

Using the learned embeddings for enzyme classification

- The Enzyme Commission number (EC number) defines a hierarchical classification scheme that provides the enzyme nomenclature based on enzyme-catalyzed reactions.
- 1340 yeast proteins collected from 7 classes of top-level EC numbers: oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, and translocases.
- **K-means clustering** is applied to group them into seven non-overlapping clusters, based on the learned embeddings from Bio-JOIE, as well as baselines approaches.



Observation: The embeddings learned by Bio-JOIE can better preserve the enzyme-based features and perform better on clustering.

Supplementary: Data Sources

Datasets are collected from multiple sources such as STRING (human PPIs), BioGrid (virus-human PPIs) and AmiGO/QuickGO (gene ontology annotations).

Links:

- STRING: https://string-db.org/cgi/download.pl
- BioGrid (SARS-CoV/MERS/SARS-CoV-2): https://wiki.thebiogrid.org/doku.php/covid
- AmiGO 2: http://amigo.geneontology.org/amigo/dd_browse
- QuickGO: https://www.ebi.ac.uk/QuickGO/

- Project page: https://www.haojunheng.com/project/goterm/

Data Source

Virus-human Protein Interactions (PPI)

E — ***Severe acute respiratory syndrome coronavirus 2***

env, envelope, SARS-CoV2 E, E protein, emp, SARS-CoV-2 E, VEMP_WCPV, GU280_gp04

Envelope small membrane protein

GO Process (0) | GO Function (0) | GO Component (0)

EXTERNAL DATABASE LINKOUTS
Entrez Gene | RefSeq | UniprotKB

Download 10 Published Interactions For This Protein

Stats & Options

Current Statistics — **Publications:** 2

High Throughput		Low Throughput
6 (60%)	10 Physical Interactions	4 (40%)
0 (0%)	0 Genetic Interactions	0 (100%)

Search Filters — Customize how your results are displayed...

No Filter: Show All Associations

Switch View: Interactors (10) | Interactions (10) | Network

Displaying **10** total unique interactions

Interactor	Role	Organism	Experimental Evidence Code	Dataset	Throughput	Score	Curated By	Notes
AP3B1	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.9636	BioGRID	
BRD2	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.9066	BioGRID	
BRD4	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.9785	BioGRID	
CWC27	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.8931	BioGRID	
SLC44A2	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.9503	BioGRID	
ZC3H18	HIT	H. sapiens	Affinity Capture-MS ⊞	Gordon DE (2020)	High	0.7964	BioGRID	

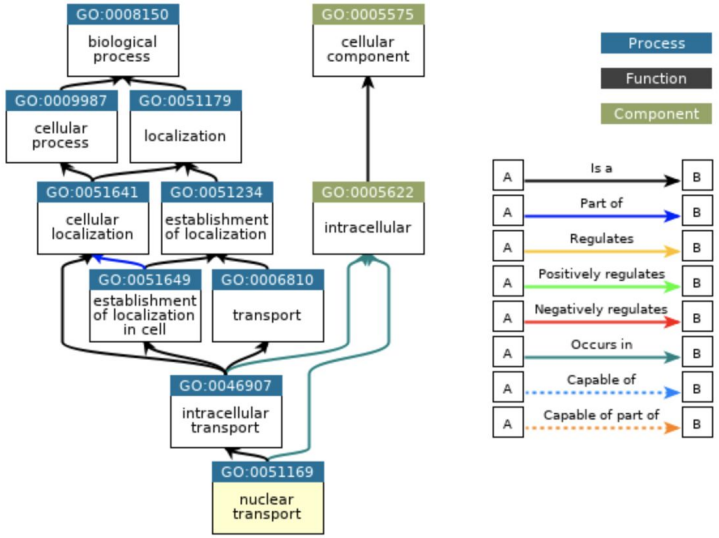
Data Source

Gene Ontology From AmiGO/QuickGO

AmiGO Drill-down Browser: [Demo] Biological Process Cellular Component Molecular Function

GoTerm Example: [Example: GO:0051169]

SARS-CoV-2 GoTerm annotations: [Example: P0DTC1 (R1A_HUMAN)]



P0DTC1

Gene Product	Symbol	Qualifier	GO Term	
UniProtKB:P0DTC1	P0DTC1	enables	GO:0036459 (F) thiol-dependent ubiquitinyl hydrolase activity	MF
UniProtKB:P0DTC1	P0DTC1	part_of	GO:0044220 (C) host cell perinuclear region of cytoplasm	CC
UniProtKB:P0DTC1	P0DTC1	part_of	GO:0030430 (C) host cell cytoplasm	CC
UniProtKB:P0DTC1	P0DTC1	part_of	GO:0033644 (C) host cell membrane	CC
UniProtKB:P0DTC1	P0DTC1	involved_in	GO:0090305 (P) nucleic acid phosphodiester bond hydrolysis	BP