Scientific Text Mining and Knowledge Graphs

Chapter 2
Part 2: Knowledge Graph Learning

Presenter: Meng Jiang

University of Notre Dame

mjiang2@nd.edu

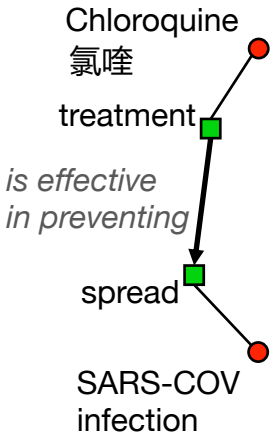
Roadmap

- Scientific KGs of COVID-19 literature
- Learning scientific KG for literature search
- Learning scientific KG for text generation

Roadmap

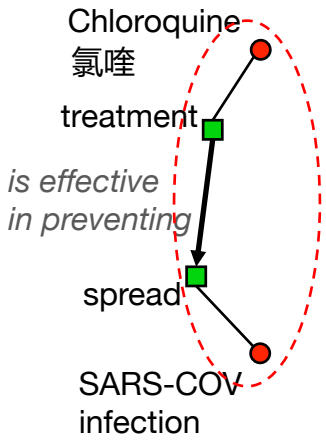
- **Scientific KGs of COVID-19 literature**
- Learning scientific KG for literature search
- Learning scientific KG for text generation

Motivating Examples



Vincent et al. 2005. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. Virology Journal.

Motivating Examples (cont'd)



SARS-COV-2

Vincent et al. 2005. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. Virology Journal.

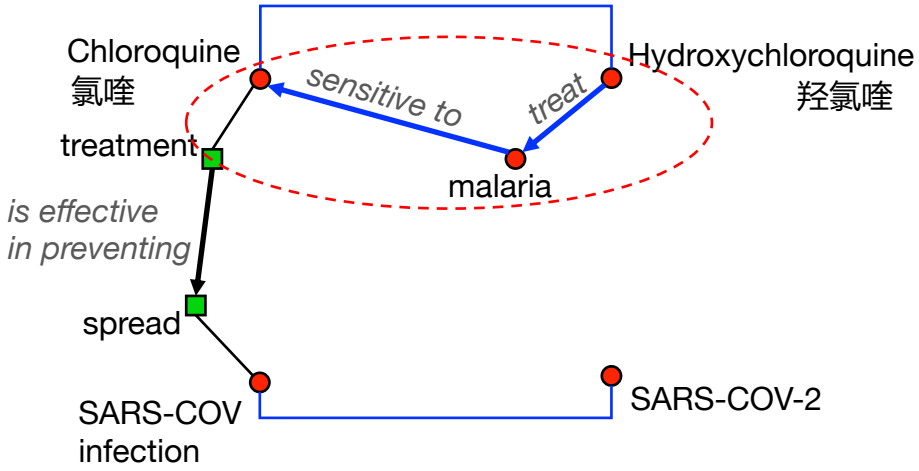
"… The Chloroquine treatment is effective in preventing the spread of SARS-COV infection. …"

[1]

Quintuple: ({Chloroquine ::: treatment},
is_effective_in_preventing,
{SARS-COV_infection ::: spread})

[1] Jiang et al. 2019. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. SIGKDD.

Motivating Examples (cont'd)



Vincent et al. 2005. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. Virology Journal.

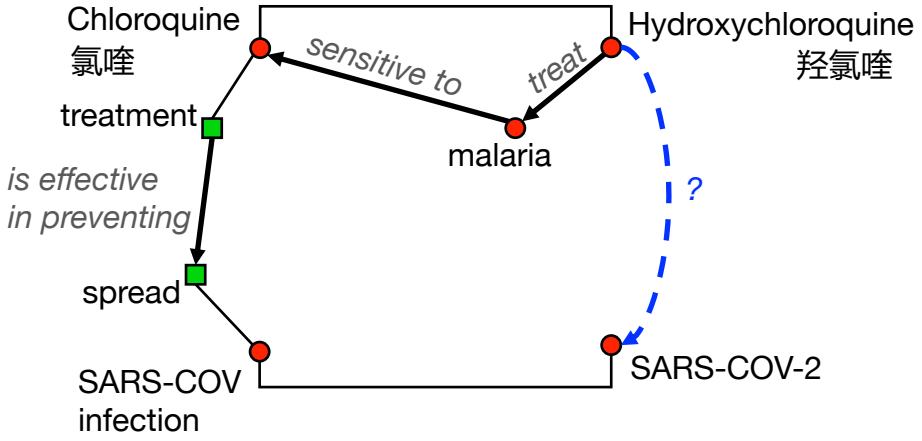
The Coronavirus Infectious Disease Ontology (CIDO). The OBO Foundry, Ontobee, and NCBO.

"… Hydroxychloroquine is a medication used to treat malaria in areas where malaria remains sensitive to chloroquine. …"
[Wiki] ↓ [1]

Tuples: (Hydroxychloroquine, *treat*, malaria)
(malaria, *remains_sensitive_to*, chloroquine)

[1] Jiang et al. 2019. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. SIGKDD.

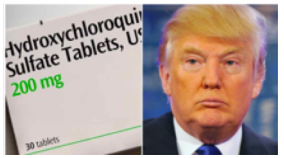
Motivating Examples (cont'd)



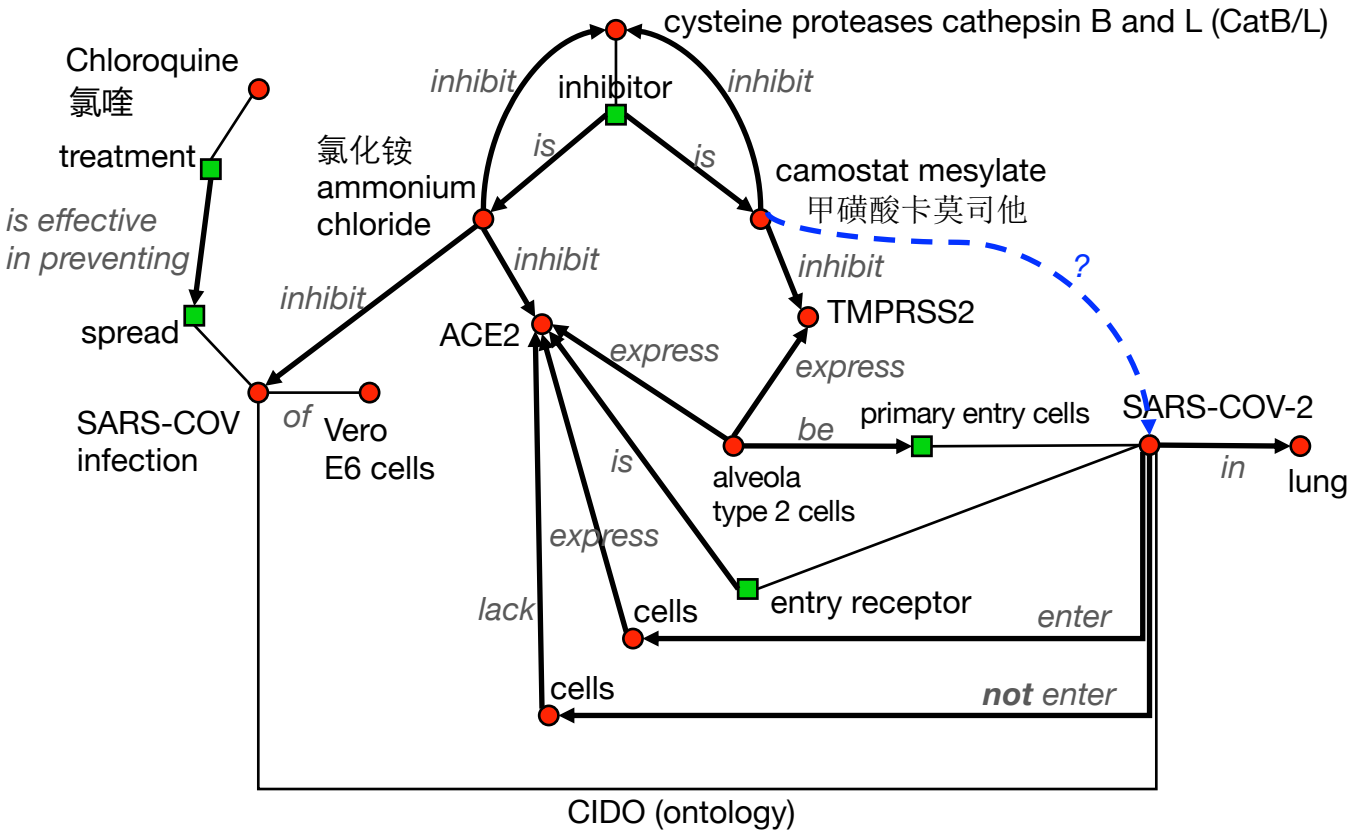
Vincent et al. 2005. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. Virology Journal.

The Coronavirus Infectious Disease Ontology (CIDO). The OBO Foundry, Ontobee, and NCBO.

FDA Warns Against Using Drugs Promoted By Trump After Reports Of 'Poisoning And Death' [Forbes 04/24/2020]
https://www.forbes.com/sites/carlieporterfield/2020/04/24/fda-warns-against-using-drugs-promoted-by-trump-after-reports-of-poisoning-and-death



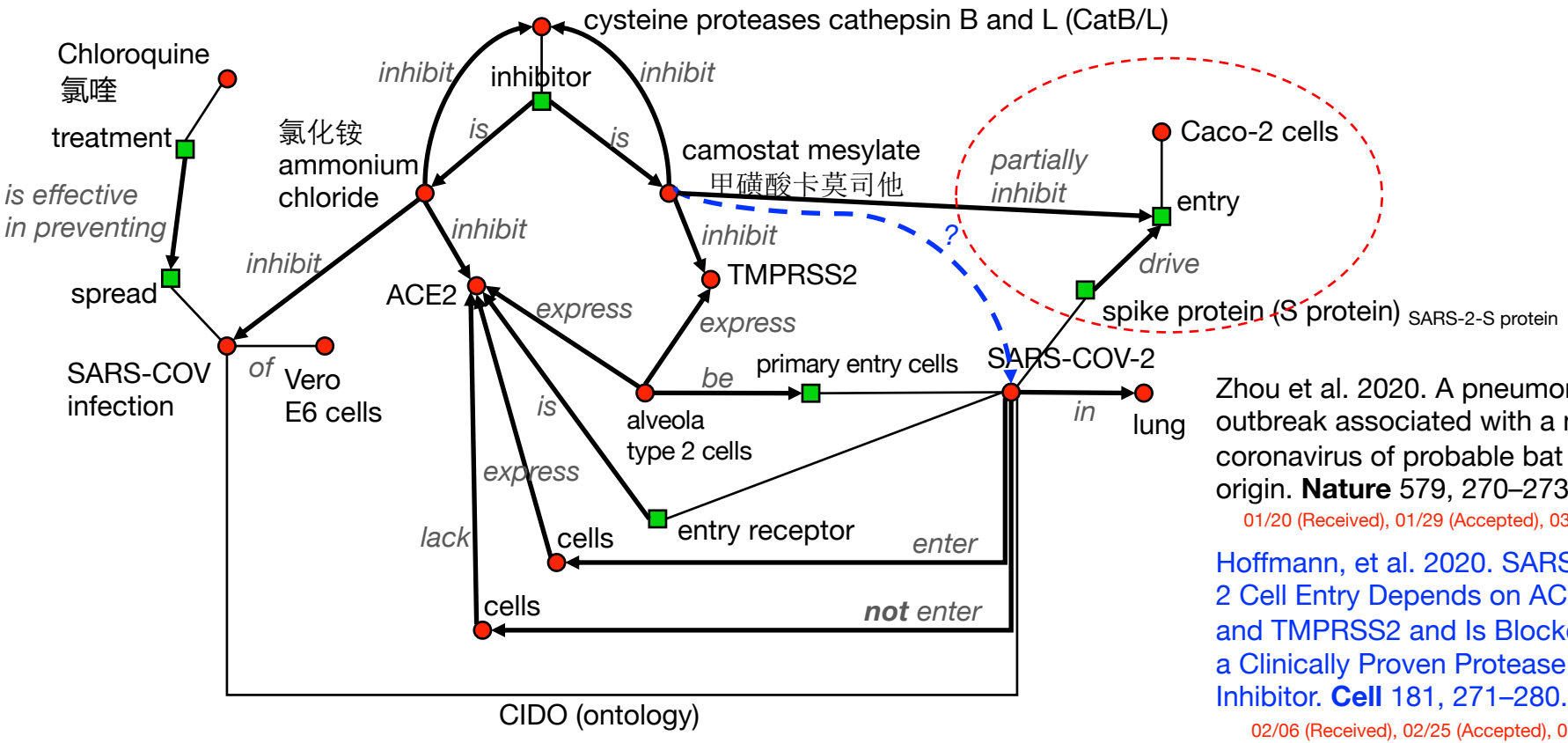
Motivating Examples (cont'd)



Zhou et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature** 579, 270–273.

01/20 (Received), 01/29 (Accepted), 03/12/2020

Motivating Examples (cont'd)



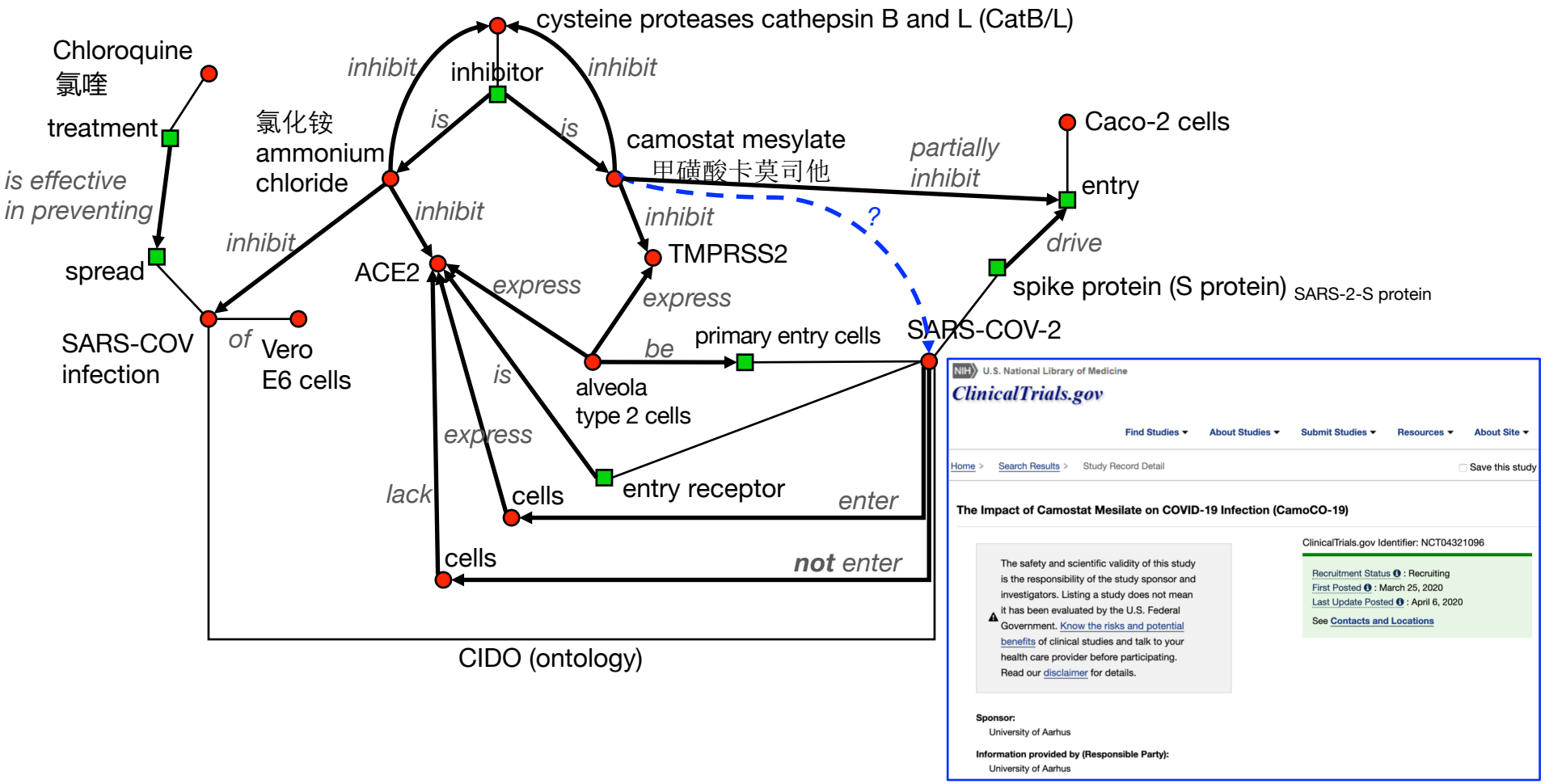
Zhou et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature** 579, 270–273.

01/20 (Received), 01/29 (Accepted), 03/12/2020

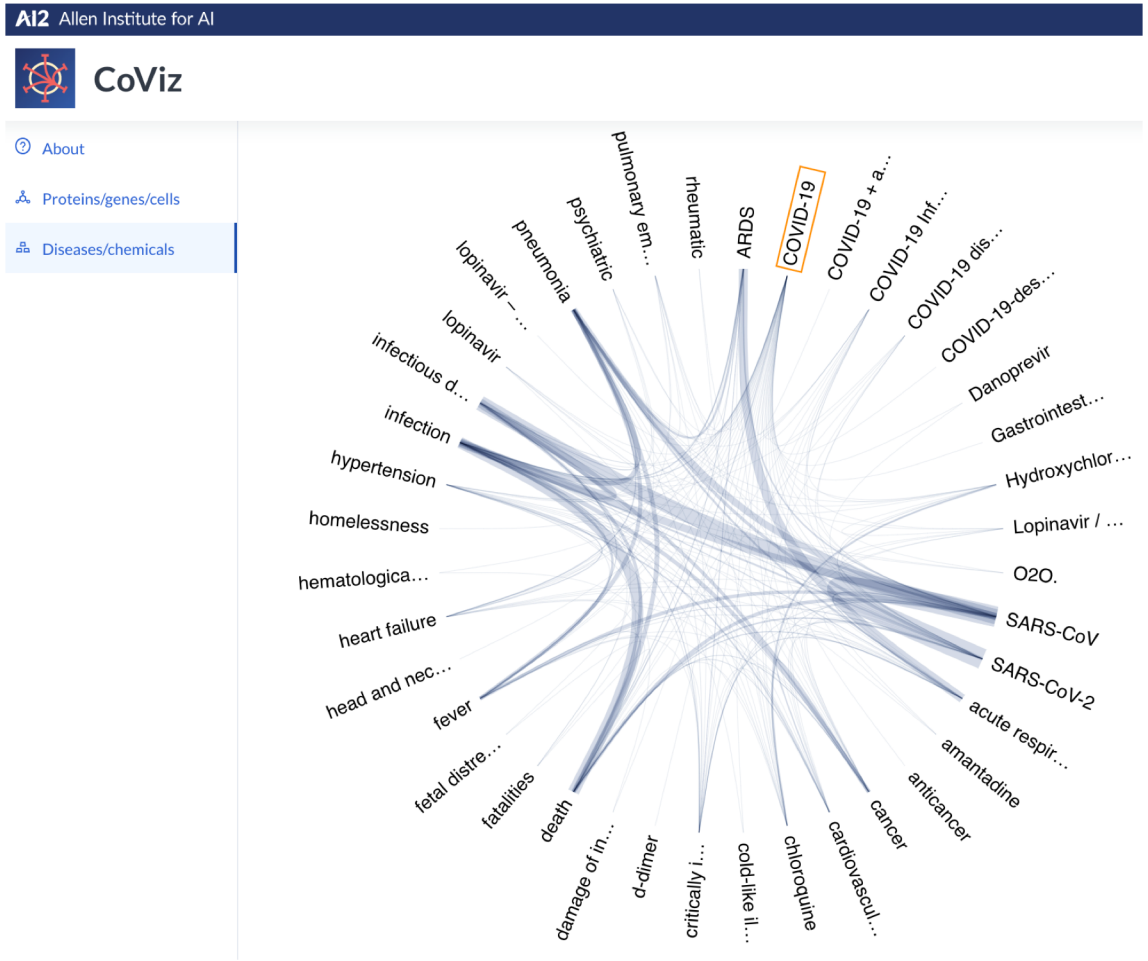
Hoffmann, et al. 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. **Cell** 181, 271–280.

02/06 (Received), 02/25 (Accepted), 04/16/2020

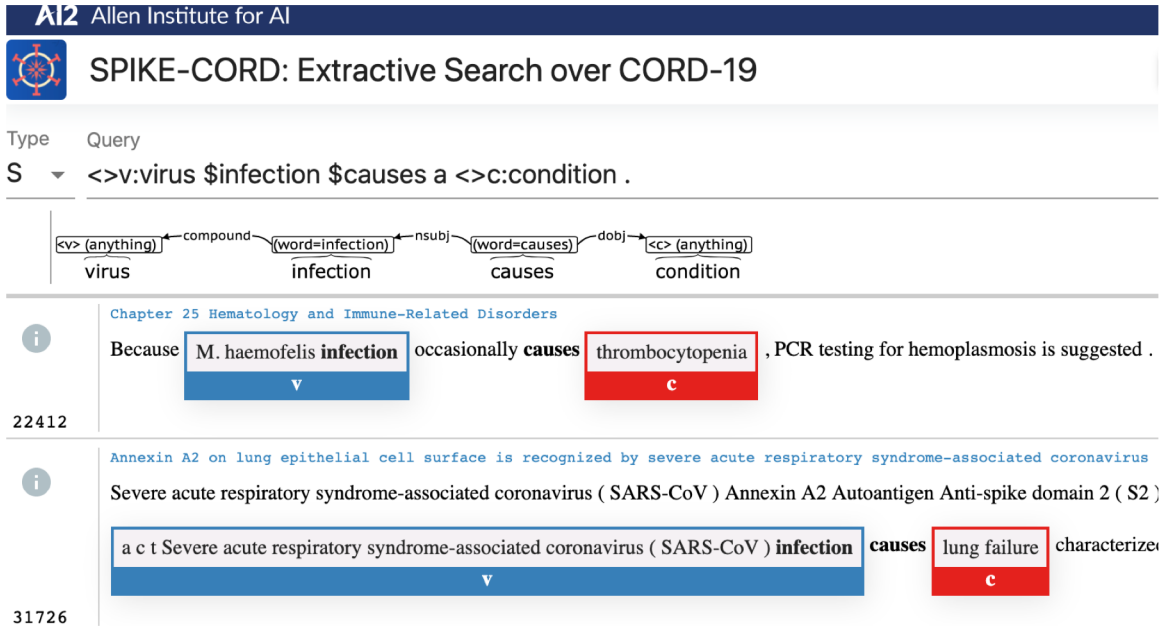
Motivating Examples (cont'd)



Knowledge Graph 1: CoViz by AI2



Knowledge Graph 2: SPIKE-CORD



based on **SciSpacy**: SpaCy models for biomedical text processing

https://allenai.github.io/scispacy/

https://spike.covid-19.apps.allenai.org/search/covid19

https://spike.covid-19.apps.allenai.org/md/covid-usage-examples

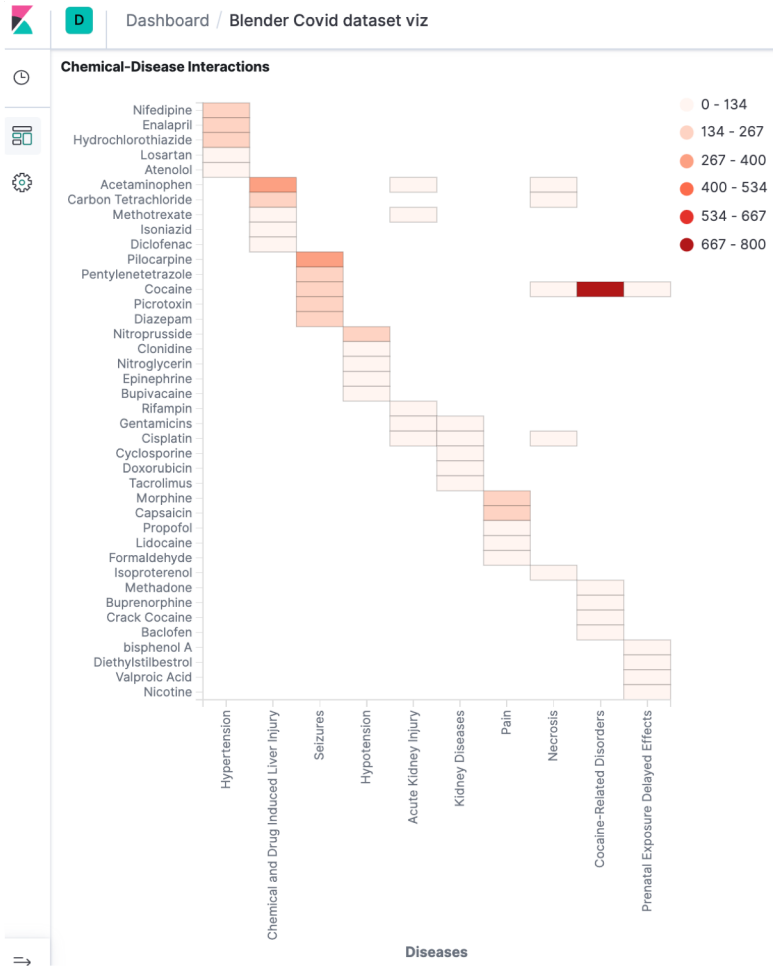
NER

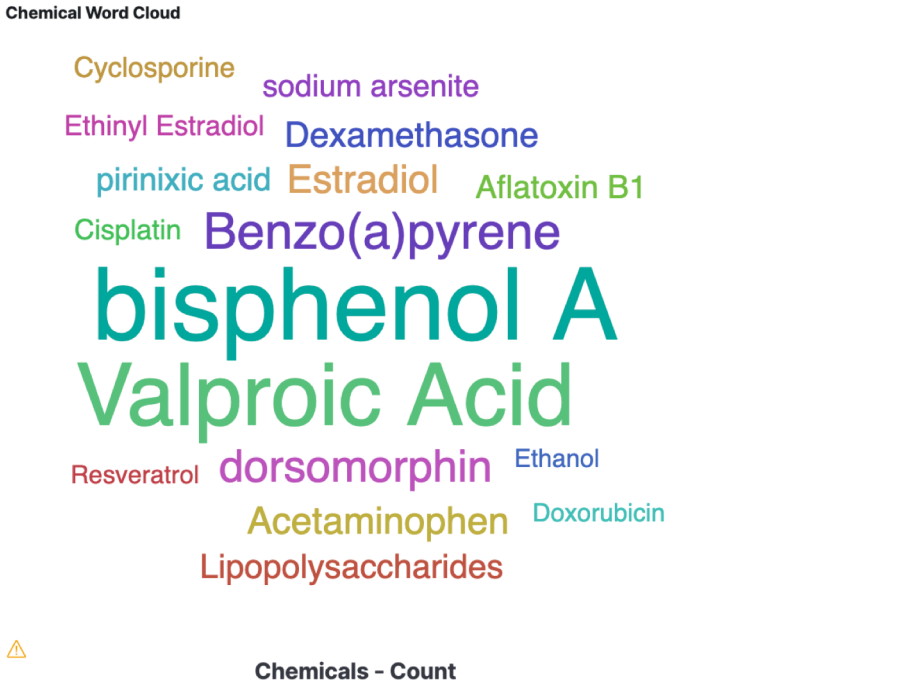
```
COVID-19
AMINO_ACID
ANATOMICAL_SYSTEM
CANCER
CELLULAR_COMPONENT
CELL_LINE
CELL_TYPE
CHEBI
CHEMICAL
CL
COVID-19
DEVELOPING_ANATOMICAL_STRUCTURE
DISEASE
DNA
ENTITY
GENE_OR_GENE_PRODUCT
GGP
GO
IMMATERIAL_ANATOMICAL_ENTITY
MULTI-TISSUE_STRUCTURE
ORGAN
ORGANISM
ORGANISM_SUBDIVISION
ORGANISM_SUBSTANCE
PATHOLOGICAL_FORMATION
PROTEIN
RNA
SIMPLE_CHEMICAL
SO
TAXON
TISSUE
```

Relation

$causes
$treats
$uses

Knowledge Graph 3: Blender Covid + SemViz@Brendeis

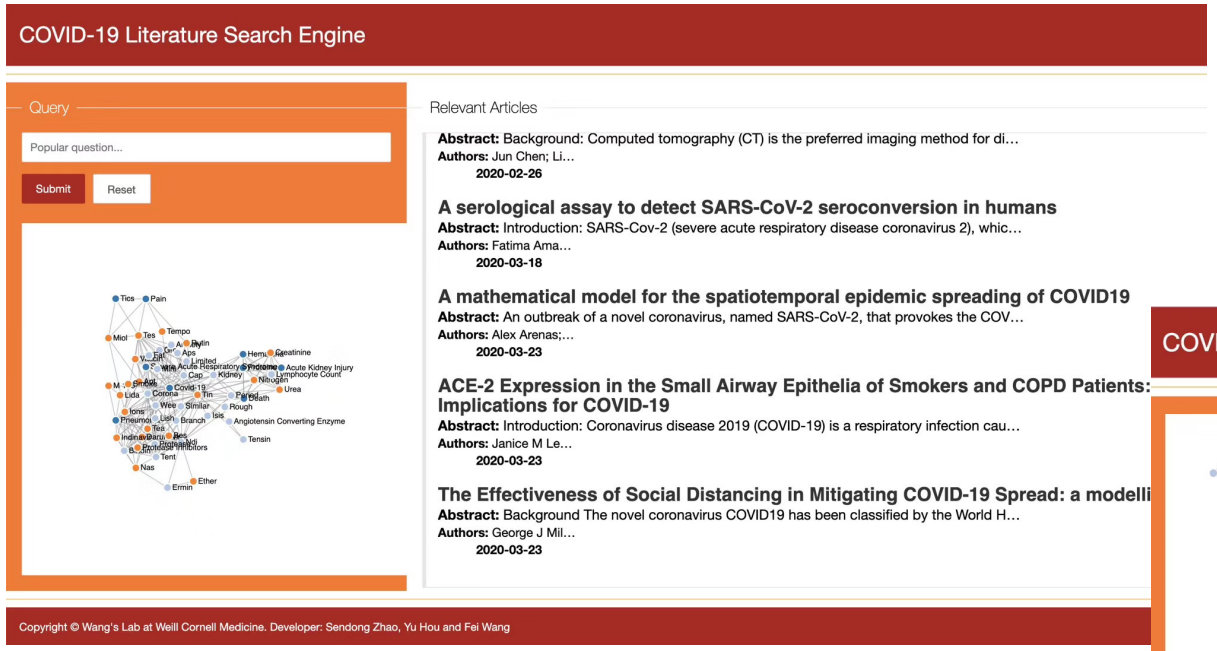


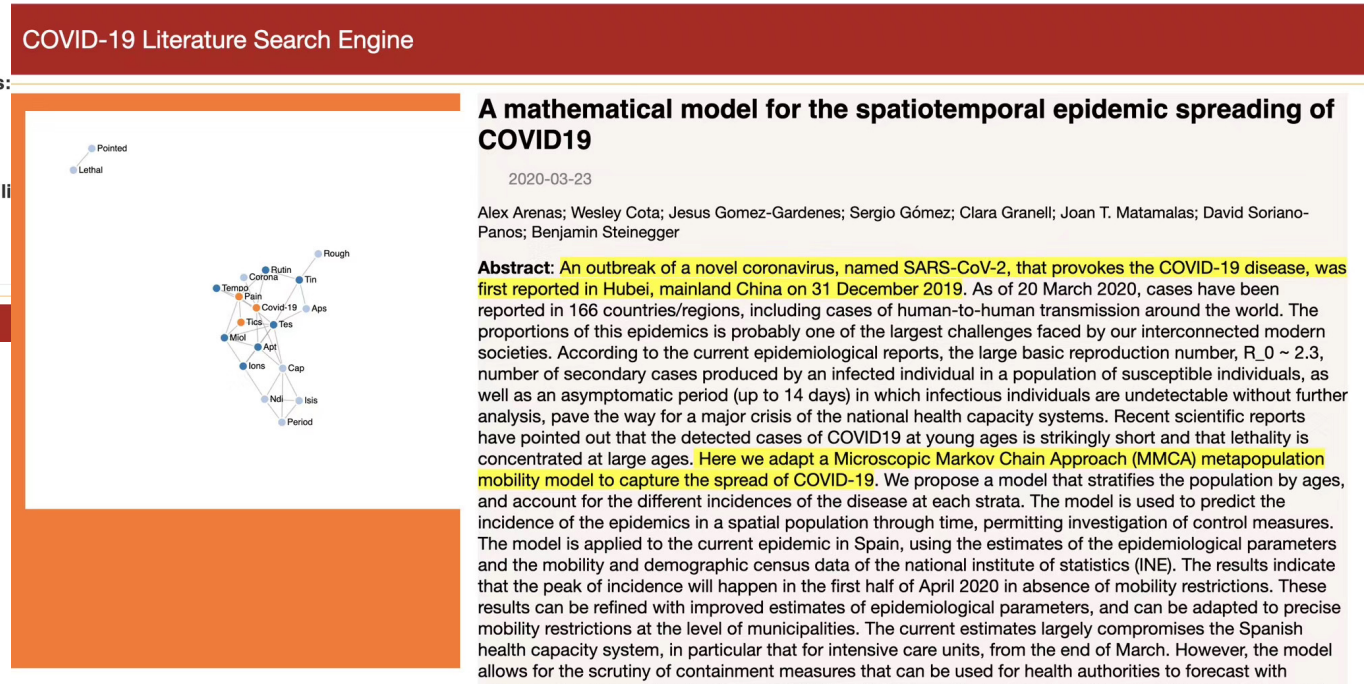


http://blender.cs.illinois.edu/covid19/?fbclid=IwAR2z0BcjDWScjuilCaWQD-YrdZvtlGGQh3LNjvDi4JzzwU6G1K9HD4owYWE

https://www.semviz.org/

Knowledge Graph 4: Weill Cornell





Recommended Principles

- **P1: Relation-schema free:** Open-domain IE with domain-specific pre-training?
- **P2: Attribute-aware:** Quintuple: ({Esubj:::Asubj}, R, {Eobj:::Aobj})
  - Entities and noun phrases (attributes, etc.) and verbal phrases (relation), etc.
- **P3: Multi-tuple:** One sentence can make multiple associated tuples
- **P4: Graph-oriented:** Discovering knowledge by link/path prediction

"*We showed that extracellular acidic pH reduces the activity of TRPV5/V6 channels, whereas alkaline pH increases the activity of TRPV5/V6 channels in Jurkat T cells.*"

Fact tuple 1: (extracellular acidic pH, reduces, {TRPV5/V6 channels: activity})
Fact tuple 2: (alkaline pH, increases, {TRPV5/V6 channels: activity})
Condition tuple: (TRPV5/V6 channels, in, Jurkat T cells)

Roadmap

- Scientific KGs of COVID-19 literature
- **Learning scientific KG for literature search**
- Learning scientific KG for text generation

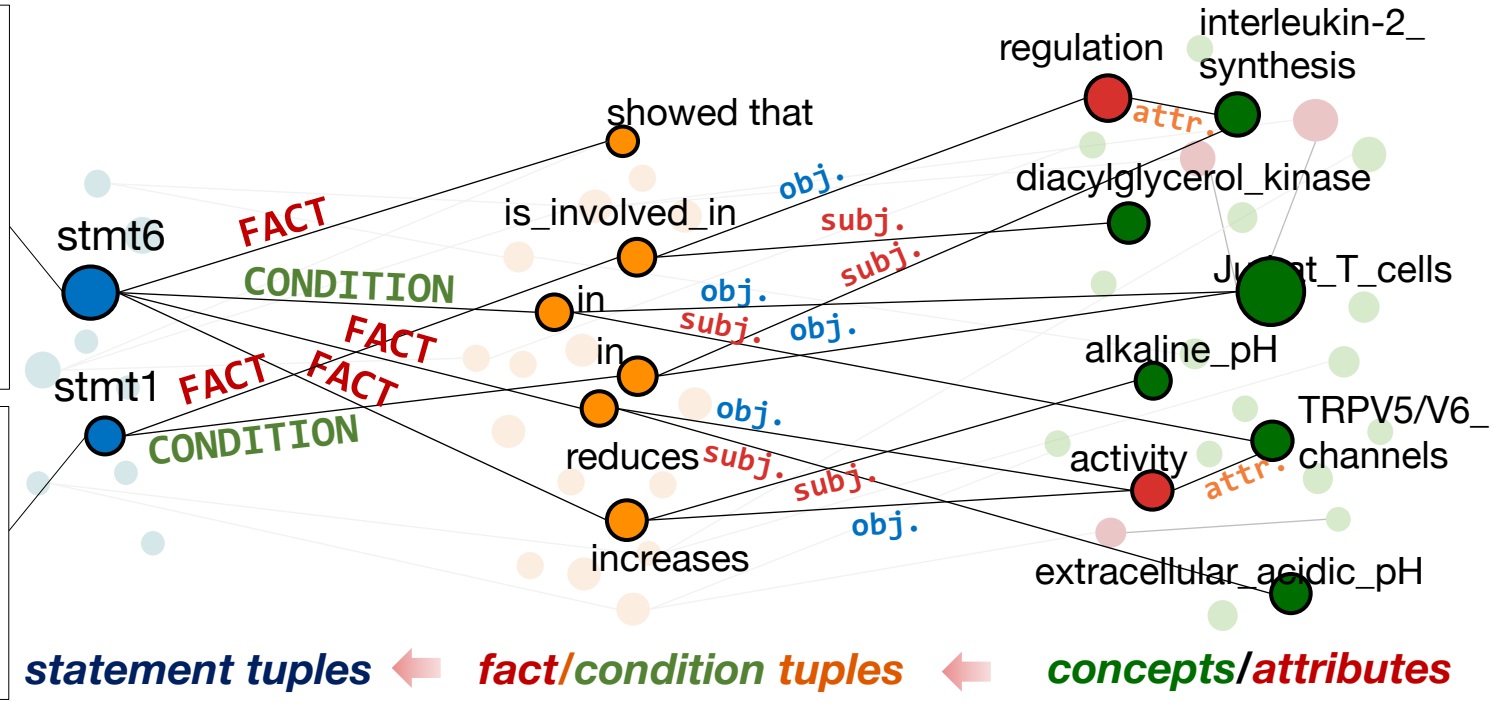
Three-Layered Scientific KG

stmt6: *We showed that extracellular acidic pH reduces the activity of TRPV5/V6 channels , whereas alkaline pH increases the activity of TRPV5/V6 channels in Jurkat T cells.*

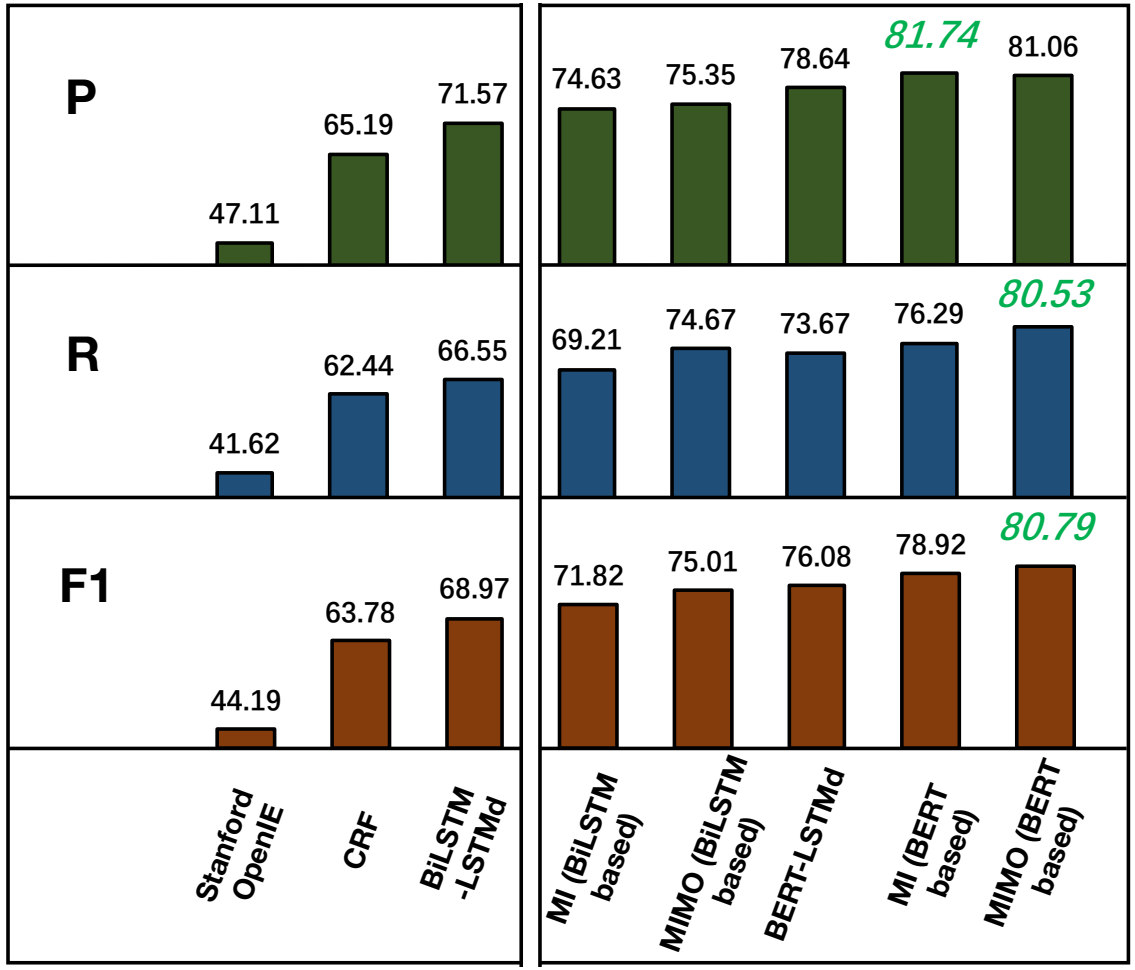
fact 1: (extracellular_acidic_pH, reduces, {TRPV5/V6_channels: activity} **)**
fact 2: (alkaline_pH, increases, {TRPV5/V6_channels: activity}**)**
condition 1: (TRPV5/V6_channels, in, Jurkat_T_cells **)**

stmt1: *A diacylglycerol kinase is involved in the regulation of interleukin-2 synthesis in Jurkat T cells.*

fact 1: (diacylglycerol_kinase, is_involved_in, {interleukin-2 synthesis: regulation} **)**
condition 1: (interleukin-2_synthesis, in, Jurkat_T_cells **)**

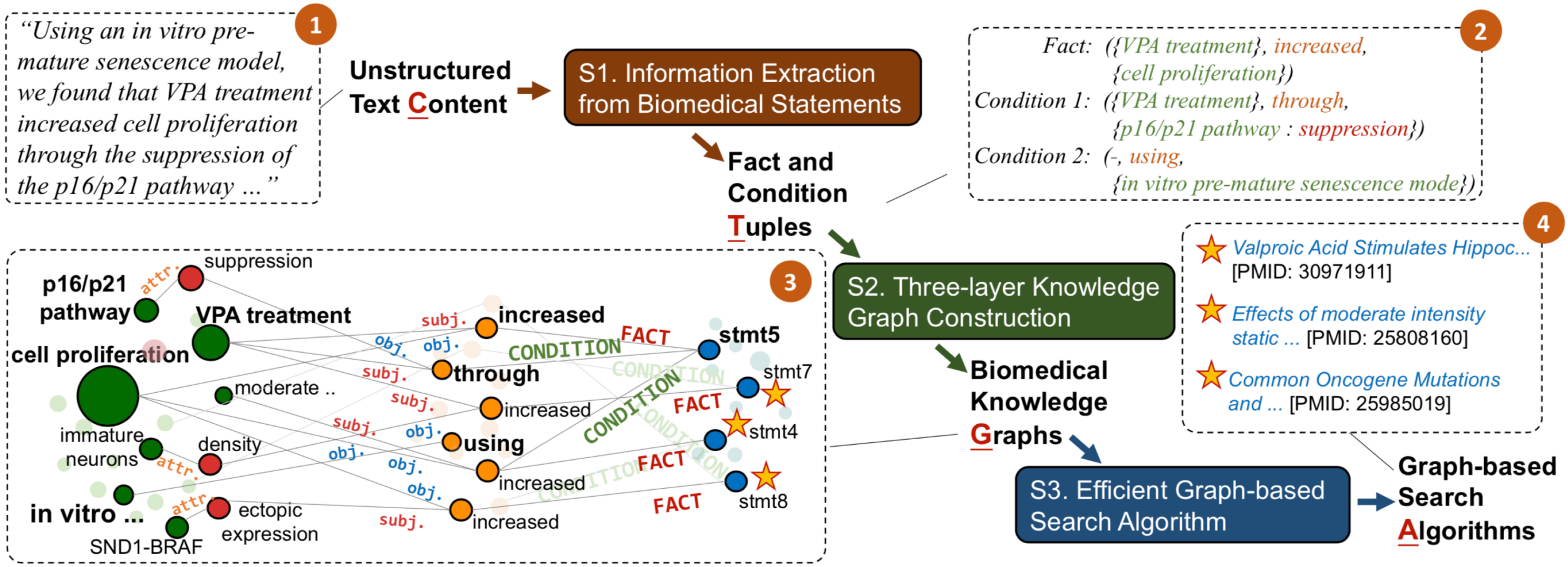


MIMO: An Information Extraction Model



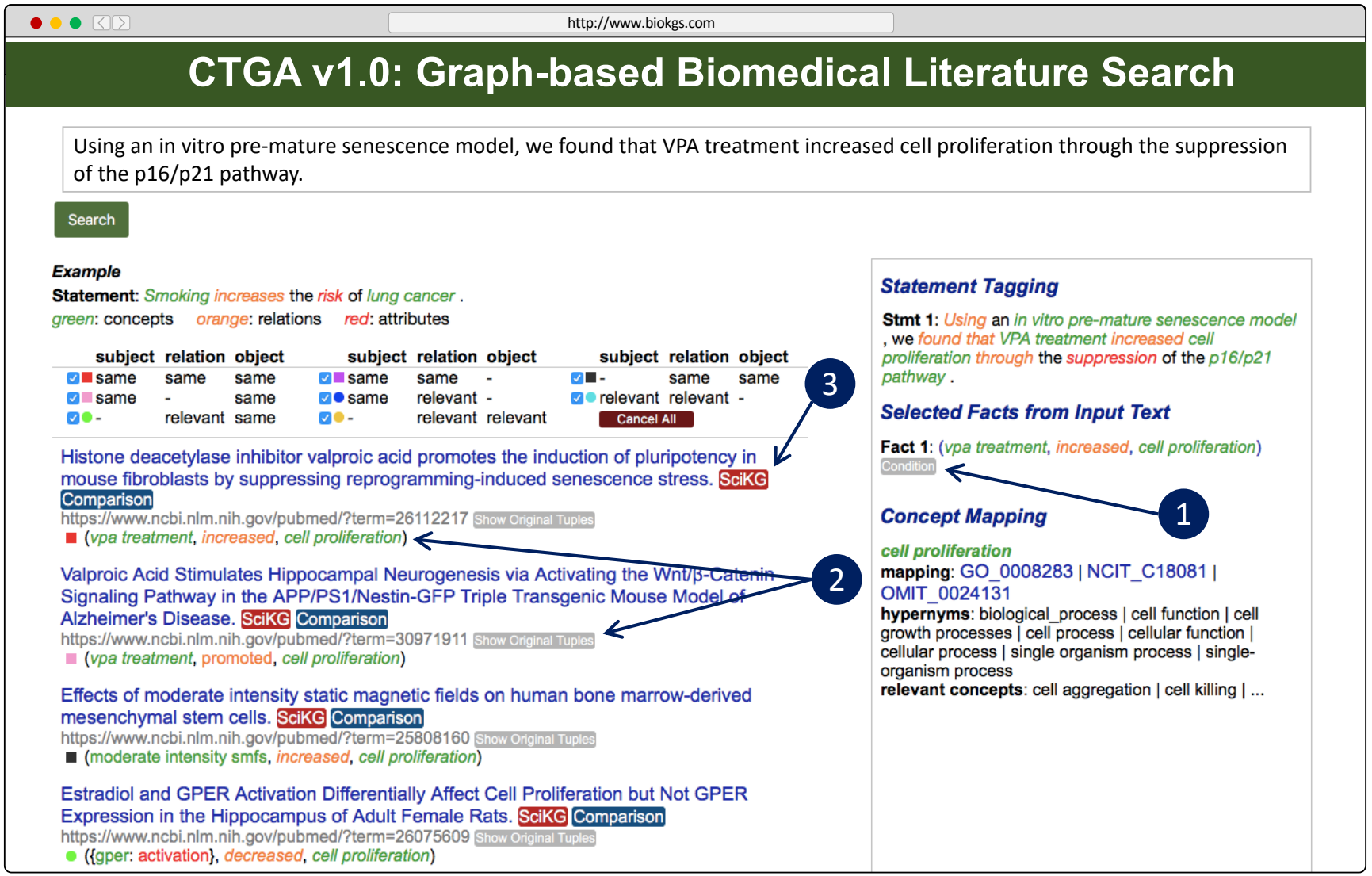
Jiang T, Zhao T, Qin B, Liu T, Chawla N, Jiang M. Multi-Input Multi-Output Sequence Labeling for Joint Extraction of Fact and Condition Tuples from Scientific Text. EMNLP 2019.

Novel Framework: CTGA



- Search and knowledge discovery
- Prediction and inference
- Hypothesis generation
- Hypothesis validation

A Demonstration System



Case Study

Query

VPA treatment increased cell proliferation.

CTGA

1. Using an in vitro pre-mature senescence model , we found that VPA treatment increased cell proliferation and inhibited apoptosis through the suppression of the p16/p21 pathway.
   (Knowledge: {VPA treatment, increase, cell proliferation}, {VPA treatment, inhibit, apoptosis})
2. VPA treatment promoted cell proliferation and increased the density of immature neurons in the dentate gyrus (DG) of the hippocampus of 3xTgAD mice.
   (Knowledge: {VPA treatment, promote, cell proliferation}, {VPA treatment, increase, density of immature neurons})
3. VPA treatment increased cathepsin B levels and activities in primary CLL cells.
   (Knowledge: {VPA treatment, increase, cathepsin B level})
4. Moderate intensity SMFs increased cell proliferation, ALP activity, and calcium release.
   (Knowledge: {Moderate intensity SMF, increase, cell proliferation})
5. Ectopic expression of SND1-BRAF in H1299 cells increased phosphorylation levels of MEK/ERK , cell proliferation , and spheroid formation
   (Knowledge: {Ectopic expression of SND1-BRAF, increase, cell proliferation})

PubMed

1. These results suggest that VPA increased type-1 stem cells in relation to the activation of SCF-KIT signaling and suppression of BTG2-mediated antiproliferative effect on stem cells.
2. Prostate cancer cells , sensitive and resistant to temsirolimus , were exposed to VPA , and tumor cell growth behavior compared.
3. VPA treatment promoted cell proliferation and increased the density of immature neurons in the dentate gyrus (DG) of the hippocampus of 3xTgAD mice.
4. Cell proliferation had increased to control levels at 30 and 45 d, demonstrating that memory recovery occurs over a period of six weeks after discontinuing VPA treatment.
5. To compare the protective effects of suberoylanilide hydroxamic acid (SAHA) and valproic acid (VPA) on human lens epithelial cells (HLECs) following ultraviolet-B exposure.

Results

Search Engines	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	SkipThought	Aver.Embedding	Top@5	Top@1	MRR
PubMed-FullText	17.22	7.01	4.43	3.37	16.34	75.56	68.54	88%	77%	81.5
CTGA-FullText	**17.61**	**7.73**	**5.07**	**3.90**	**17.03**	**75.91**	**68.77**	**93%**	**92%**	**92.3**
Improvement	2.3%↑	10.3%↑	14.4%↑	15.7%↑	4.2%↑	0.5%↑	0.3%↑	5.7%↑	19.5%↑	13.3%↑
PubMed-BestMatch	28.84	21.46	18.42	16.71	34.91	80.31	81.92	-	-	-
CTGA-BestMatch	**32.60**	**24.67**	**21.06**	**18.84**	**38.26**	**81.77**	**82.80**	-	-	-
Improvement	13.0%↑	15.0%↑	14.3%↑	12.7%↑	9.6%↑	1.8%↑	1.1%↑	-	-	-

Search Engines	Rate@1	Rate@3	Rate@5	Preference%	Kendall-Tau	SF-Dist	RBO	DCG	NDCG
PubMed	4.60	3.43	2.92	34%	68.95	10.14	87.18	9.61	96.63
CTGA	**4.83**	**3.69**	**3.31**	**66%**	**76.15**	**8.34**	**89.31**	**10.60**	**99.25**

Roadmap

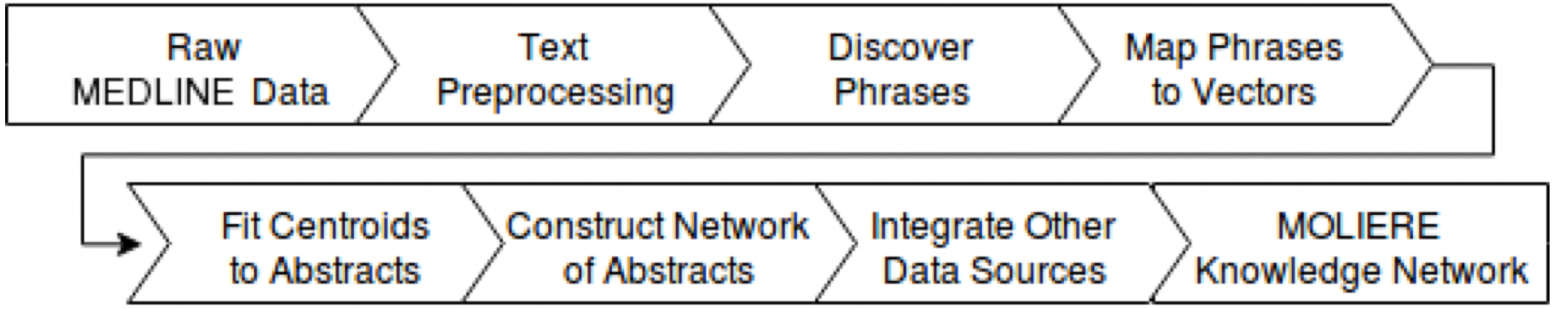
- Scientific KGs of COVID-19 literature
- Learning scientific KG for literature search
- **Learning scientific KG for text generation**

Three Interesting Works

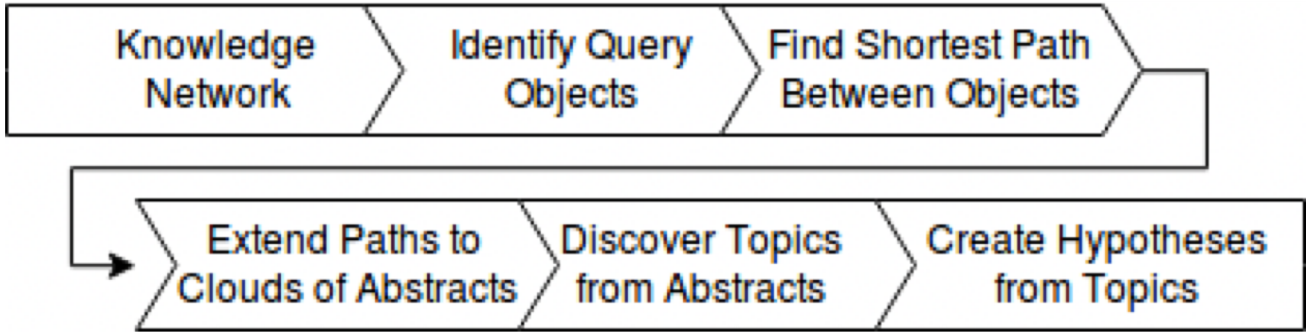
- MOLIERE: Automatic Biomedical Hypothesis Generation System (KDD'17)
- PaperRobot: Incremental Draft Generation of Scientific Ideas (ACL'19)
- Text Generation from Knowledge Graphs with Graph Transformers (NAACL'19)

MOLIERE: Automatic Biomedical Hypothesis Generation System (KDD'17)

Network construction:



Query processing:

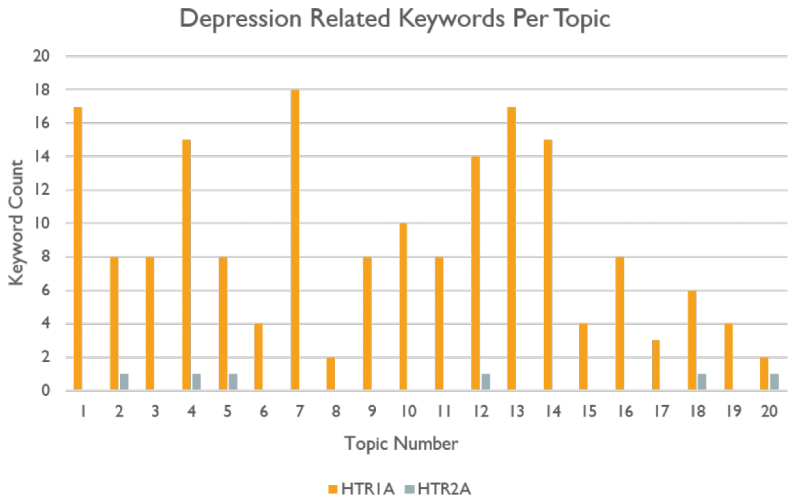


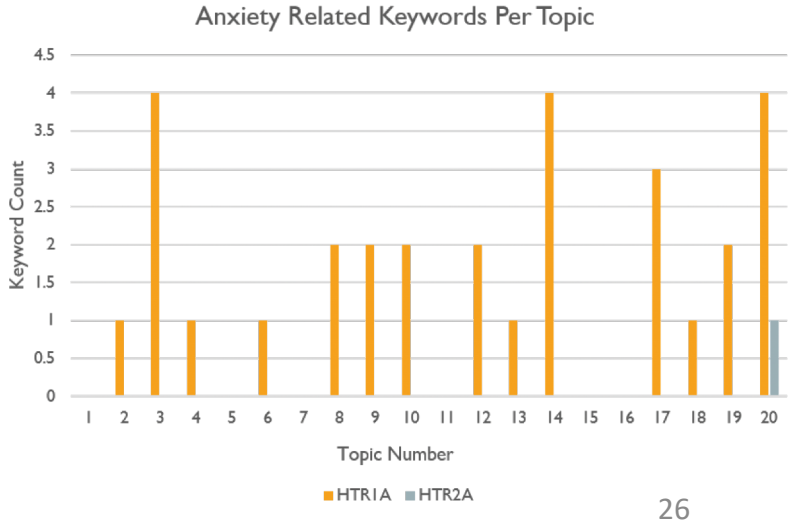
Results

As a result of running two queries, Venlafaxine to HTR1A, and Venlafaxine to HTR2A, we can corroborate the findings of Wang et al. in [49]. We find that neither pair of keywords is directly connected or connected through a single abstract. Nevertheless, **phrases such as "long term antidepressant treatment," "action antidepressants," and "antidepressant drugs" are all prominent keywords in the HTR1A query.** Meanwhile, the string "depress" only occurs four times in unrelated phrases with the HTR2A results. The distribution of depression related keywords from both queries can be see in figure 5.

Similarly, our results for **HTR1A contain a single topic holding the phrases "anxiogenic," "anxiety disorders," "depression anxiety disorders," and "anxiolytic response."** In contrast, our HTR2A results do not contain any phrases related to anxiety. The distribution of anxiety related keywords from both queries can be see in figure 6.

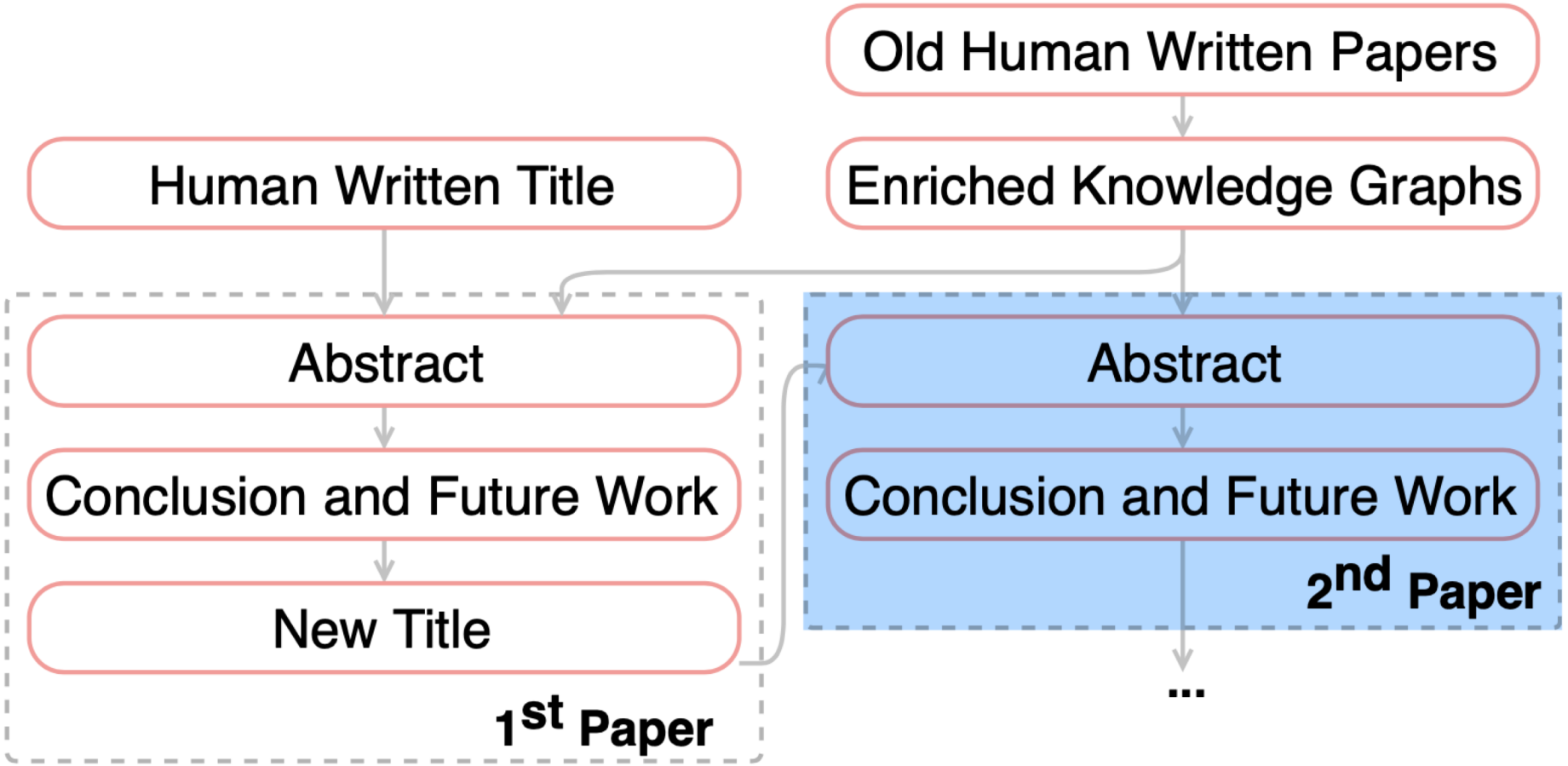
Our findings agree with those of Wang et al. which were that **a small association score of 0.34 between Venlafaxine and HTR1A** indicates a connection which is likely **related to depressive disorder and anxiety**. The association score between Venlafaxine and HTR2A, in contrast, is a much higher 4.0. This indicates that the connection between these two keywords is much weaker.



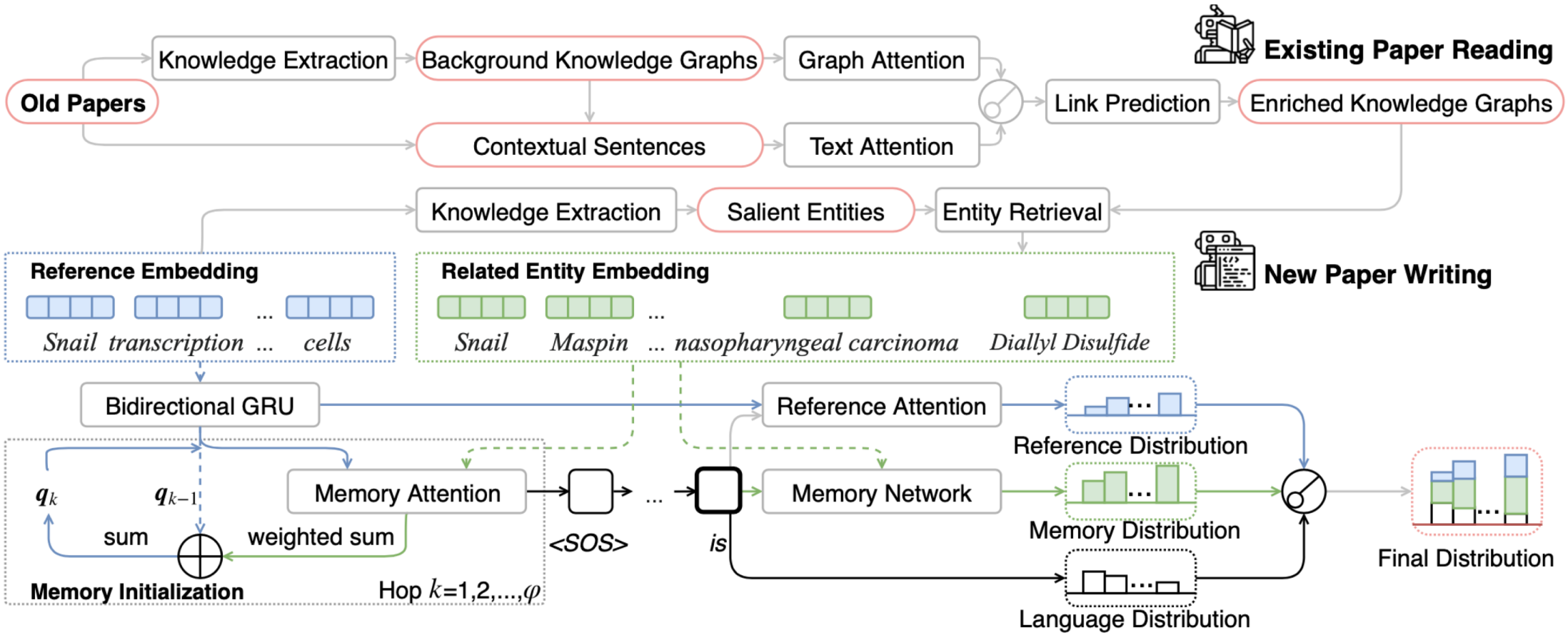


PaperRobot: Incremental Draft Generation of Scientific Ideas (ACL'19)

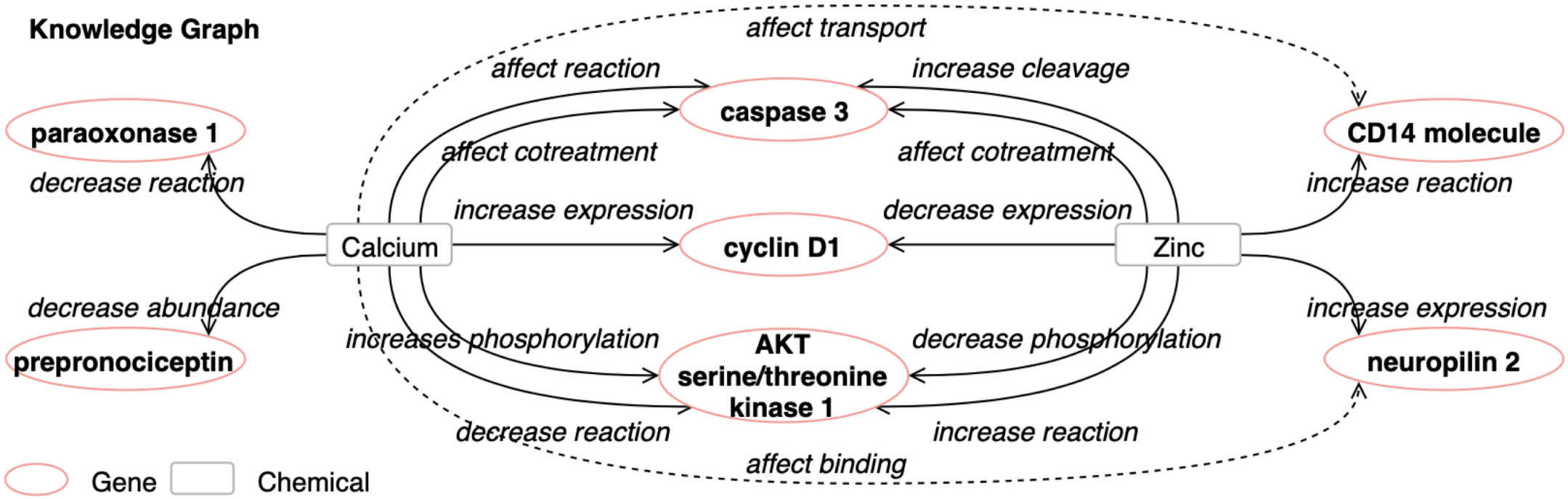
- Incremental writing



Architecture Overview



Biomedical Knowledge Extraction and Link Prediction



Contextual Sentence: So, Ca^{2+}possibly ***promoted*** **caspases** activation upstream of ***cytochrome c*** release, but inactivated **caspase** activity by calpain and/or fast depletion of ATP; whereas Zn^{2+} blocked the ***activation ofprocaspase-3*** with no visible change in the level of **cytochrome c**, and the block possibly resulted from its direct inhibition on **caspase-3** enzyme.

Results

Title	**Snail** transcription factor negatively regulates **maspin tumor suppressor** in **human prostate cancer** cells		
Entities	Related: **nasopharyngeal carcinoma**; diallyl disulfide		
Output	**Human (Neal et al., 2012)**	**System**	**Post-edited by Human**
Abstract	Background: **Maspin**, a putative **tumor suppressor** that is down-regulated in breast and **prostate cancer**, has been associated with decreased cell motility. **Snail** transcription factor is a zinc finger protein that is increased in breast cancer and is associated with increased tumor motility and invasion by induction of epithelial-mesenchymal transition (EMT). We investigated the molecular mechanisms by which **Snail** increases **tumor** motility and invasion utilizing **prostate cancer** cells. Methods: Expression levels were analyzed by **RT-PCR** and **western blot** analyses. Cell motility and invasion assays were performed , while **Snail** regulation and binding to **maspin** promoter was analyzed by luciferase reporter and chromatin immunoprecipitation (ChIP) assays. Results: **Snail** protein expression was higher in different **prostate cancer** cells lines as compared to normal prostate epithelial cells.	Background: **Snail** is a multifunctional protein that plays an important role in the pathogenesis of **prostate cancer**. *However*, it has been shown *to be* associated with poor prognosis. The purpose of this study *was* to investigate the effect of *negatively* on the expression of ***maspin*** in **human nasopharyngeal carcinoma** cell lines. Methods: *Quantitative real-time PCR* and western blot analysis were used to determine *whether the demethylating agent was investigated by quantitative* ***RT-PCR*** *(qRT-PCR) and* ***Western blotting***. Results showed that the binding protein plays a significant role in the regulation of **tumor** growth and progression.	Background: **Snail** is a multifunctional protein that plays an important role in the pathogenesis of **prostate cancer**. It has been shown associated with poor prognosis. The purpose of this study is to investigate the negative effect of on the expression of **Maspin** in **human nasopharyngeal carcinoma** cell lines. Methods: Quantitative **RT-PCR** *(qRT-PCR)* and **western blot analyses** were used to determine *correlation of the two proteins expressions.* Results showed that the binding protein plays a significant role in the regulation of **tumor** growth and progression.
Conclusion and Future work	Collectively, our results indicate for the first time that **Snail** can negatively regulate **maspin** through direct promoter repression resulting in increased migration and invasion in prostate **cancer** cells. This study reveals a novel mechanism of how **Snail** may function and show the importance of therapeutic targeting of **Snail** signaling in future.	In summary, our study demonstrates that **Snail** negatively *inhibited* the expression of **Maspin** in **human nasopharyngeal carcinoma** cell lines *and in vitro*. Our results indicate that *the combination of the demethylating agent* might be a potential therapeutic target for the treatment of **prostate cancer**.	In summary, our study *in vitro* demonstrates that **Snail** negatively *inhibits* the expression of **Maspin** in **human nasopharyngeal carcinoma** cell lines. Our results *further* indicate that ***Maspin*** might be a potential therapeutic target for the treatment of **prostate cancer**.
New Title	Role of **maspin** in cancer (Berardi et al., 2013)	The role of ***nasopharyngeal carcinoma*** in the rat model of ***prostate cancer*** cells	The role of ***Maspin*** in the rat model of ***nasopharyngeal carcinoma*** cells

Results (cont'd)

Model	Title-to-Abstract		Abstract-to-Conclusion and Future Work		Conclusion and Future Work-to-Title	
	Perplexity	METEOR	Perplexity	METEOR	Perplexity	METEOR
Seq2seq (Bahdanau et al., 2015)	19.6	9.1	44.4	8.6	49.7	6.0
Editing Network (Wang et al., 2018b)	18.8	9.2	30.5	8.7	55.7	5.5
Pointer Network (See et al., 2017)	146.7	8.5	74.0	8.1	47.1	6.6
Our Approach (-Repetition Removal)	13.4	12.4	24.9	**12.3**	31.8	7.4
Our Approach	**11.5**	**13.0**	**18.3**	11.2	**14.8**	**8.9**

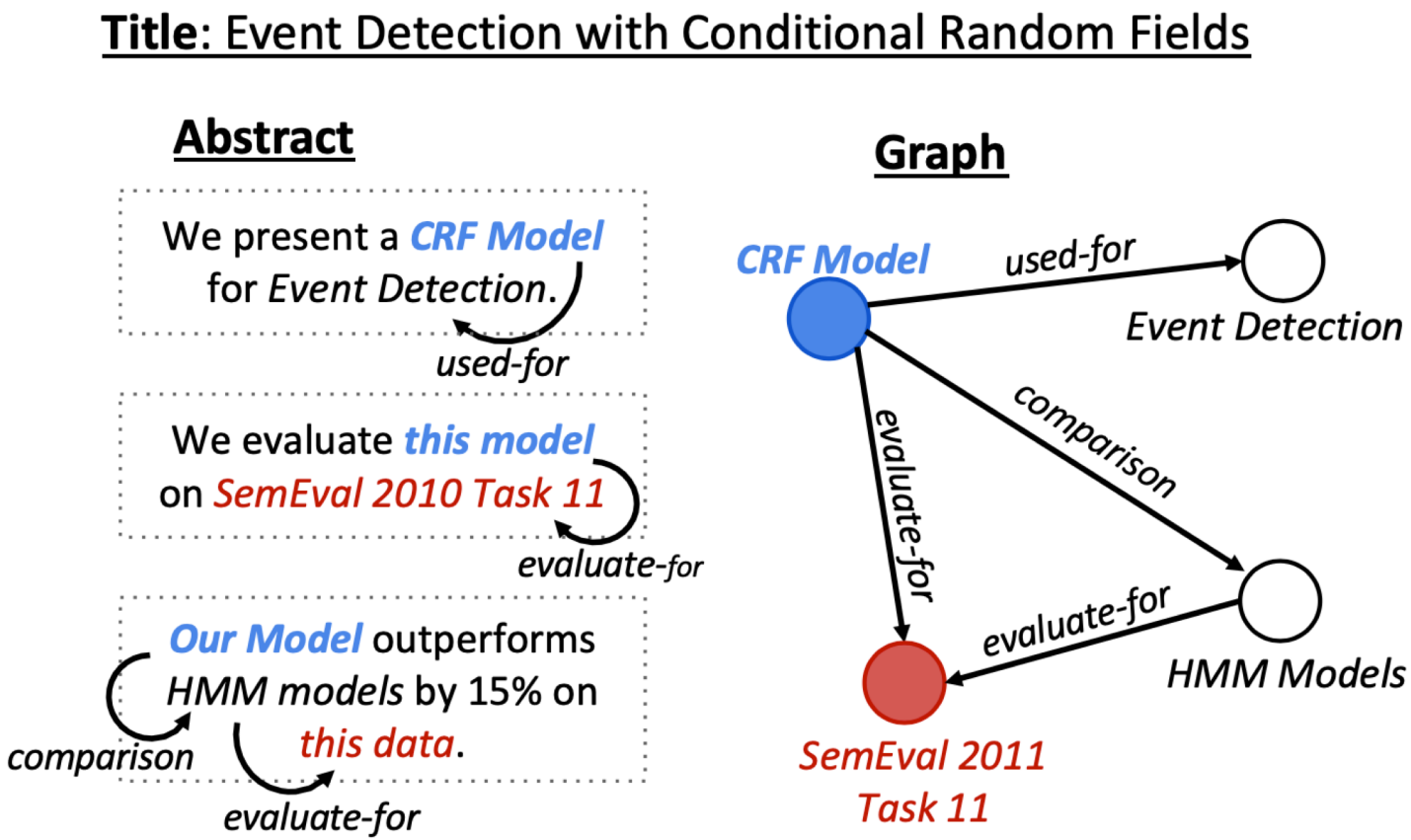
Task	Input		Output	Domain Expert	Non-expert
End-to-End	Human Title	Different	Abstract (1st)	10	**30**
		Same		**30**	16
	System Abstract	Different	Conclusion and Future work	**12**	0
		Same		8	8
	System Conclusion and Future work	Different	Title	**12**	2
		Same		12	**25**
	System Title	Different	Abstract (2nd)	**14**	4
Diagnostic	Human Abstract	Different	Conclusion and Future work	12	**14**
		Same		**24**	20
	Human Conclusion and Future work	Different	Title	8	**12**
		Same		2	**10**

Results (cont'd)

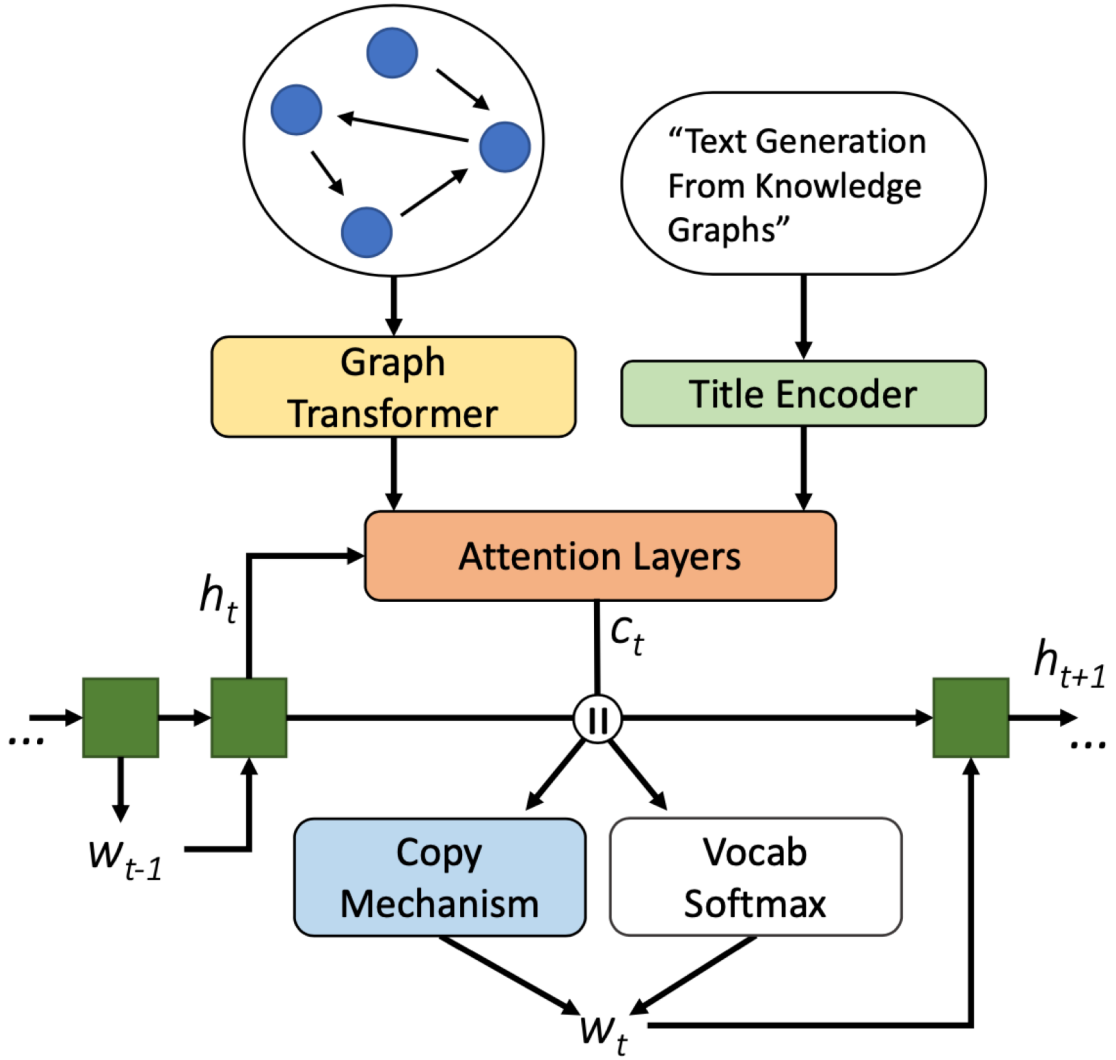
Output	**Without Memory Networks**	**Without Link Prediction**	**Without Repetition Removal**
Abstract	Background: **Snail** has been reported to exhibit a variety of biological functions. In this study, we investigated the effect of negatively on **maspin** demethylation in **human prostate cancer** cells. Methods: Quantitative real-time PCR and western blot analysis were used to investigate the effects of the demethylating agent on the expression of the protein kinase (TF) gene promoter. Results: The results showed that the presence of a single dose of 50 μM in a dose-dependent manner, whereas the level of the BMP imipramine was significantly higher than that of the control group.	Background: **Snail** has been shown to be associated with poor prognosis. In this study, we investigated the effect of negatively on the expression of **maspin** in **human prostate cancer** cells. Methods: Cells were treated with a single dose of radiotherapy for 24 h, and was used to investigate the significance of a quantitative factor for the treatment of the disease. Results: The remaining controls showed a significant increase in the G2/M phase of the **tumor suppressor protein** ($p<0.05$).	Background: **Snail** is a major health problem in human malignancies. However, the role of **Snail** on the expression of **maspin** in **human prostate cancer** cells is not well understood. The aim of this study was to investigate the effect of **Snail** on the expression of **maspin** in **human prostate cancer** cells. Methods: The expression of the expression of **Snail** and **maspin** was investigated using quantitative RT-PCR and western blot analysis. Results: The remaining overall survival (OS) and overall survival (OS) were analyzed.
Conclusion and Future work	In summary, our study demonstrated that negatively inhibited the expression of the BMP imipramine in **human prostate cancer** cells. Our findings suggest that the inhibition of **maspin** may be a promising therapeutic strategy for the treatment.	In summary, our results demonstrate that negatively inhibited the expression of **maspin** in **human prostate cancer** cells. Our findings suggest that the combination of radiotherapy may be a potential therapeutic target for the treatment of disease.	In summary, our results demonstrate that **snail** inhibited the expression of **maspin** in **human prostatic** cells. The expression of **snail** in PC-3 cells by **snail**, and the expression of **maspin** was observed in the presence of the expression of **maspin**.
New Title	Protective effects of homolog on **human breast cancer** cells by inhibiting the Endoplasmic Reticulum Stress	The role of **prostate cancer** in **human breast cancer** cells	The role of **maspin** and **maspin** in human **breast cancer** cells

Text Generation from Knowledge Graphs with Graph Transformers (NAACL'19)

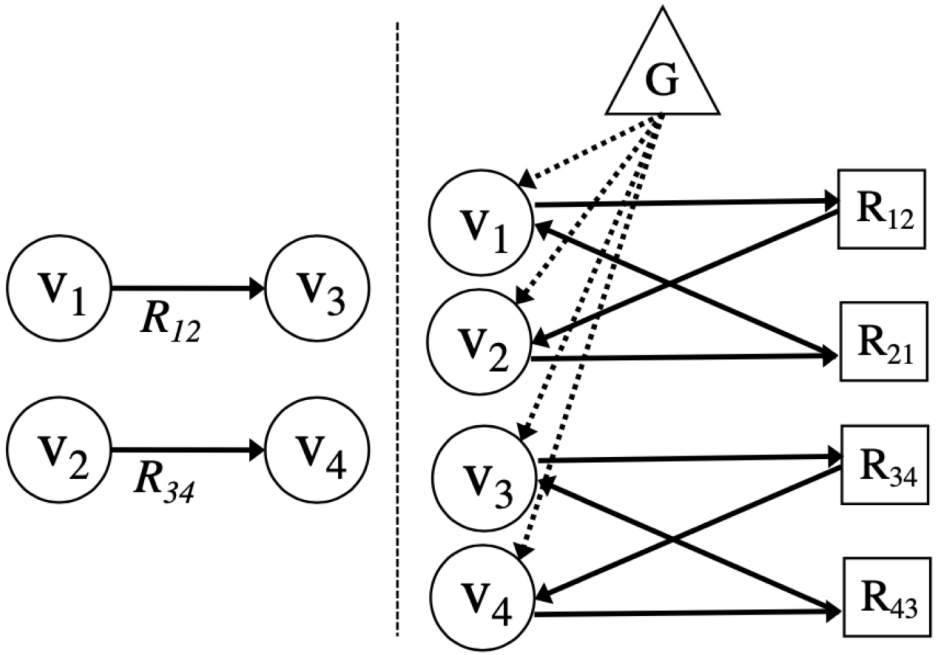
- Text-to-graph extraction
- Graph-to-text generation



Graph Writer Model

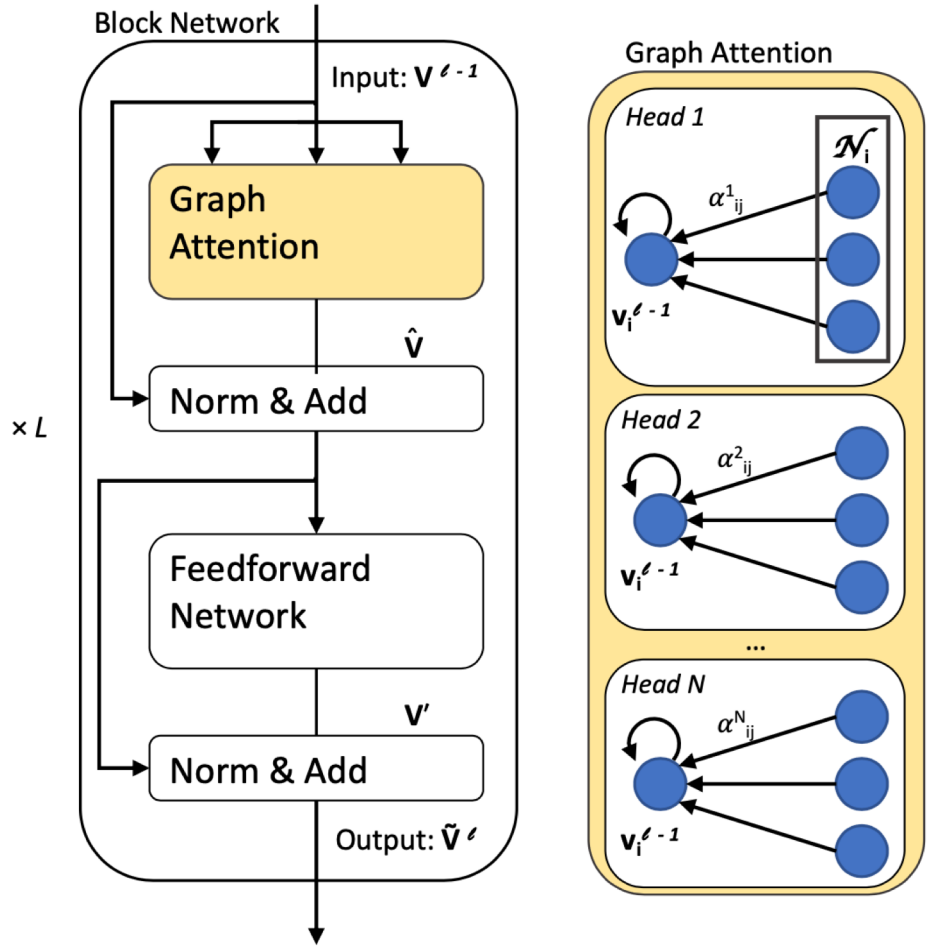


Graph Attention and Graph Transformer



Converting disconnected labeled graph to connected unlabeled graph for use in attention-based encoder.

v_i refer to vertices, R_{ij} to relations, and G is a global context node.



Graph Transformer

Results

Title	Block and Group Regularized Sparse Modeling for Dictionary Learning
Knowledge	(dictionary learning, CONJUNCTION, sparse coding) ; (optimization problems, USED-FOR, dictionary learning) ; (optimization problems, USED-FOR, sparse coding)...
GraphWriter	Sparse representations have recently been shown to be effective in many optimization problems. However, existing dictionary learning methods are limited in the number of dictionary blocks, which can be expensive to obtain. In this paper, we propose a novel approach to dictionary learning based on sparse coding ...
GAT	In this paper, we consider the problem of dictionary learning in well-known datasets. In particular, we consider the problem of dictionary learning, where the goal is to find a set of dictionary blocks that maximize the likelihood of a given set of dictionary blocks ...
EntityWriter	We propose a novel dictionary learning framework for reconstructed block/group sparse coding schemes. The dictionary learning framework is based on the descent, which is a block structure of the group structure ...
Rewriter	This paper presents a new approach to the k-means of the algorithm. The proposed approach is based on the basis of the stationarity algorithm. The algorithm is based on the fact that the number of bits is a constant of the base of the base of the input ...
Gold	This paper proposes a dictionary learning framework that combines the proposed block/group (BGSC) or reconstructed block/group (R-BGSC) sparse coding schemes with the novel Intra-block Coherence Suppression Dictionary Learning algorithm. An important and distinguishing feature of the proposed framework is that all dictionary blocks are trained simultaneously ...

Results (cont'd)

Title	Image Interpolation with Directionlets
Knowledge	(directionally adaptive image interpolation USED-FOR edge information) ; (numeric and visual quality, HYPONYM-OF, directionally adaptive image interpolation) ; (directionlets, EVALUATE-FOR, multiple-direction wavelet transform) . . .
GraphWriter	In this paper, we propose a novel directionally adaptive image interpolation based on the multiple-direction wavelet transform, called directionlets, which can be used as a directionlets to improve the numeric and visual quality of the directionally adaptive image interpolation . . .
GAT	In this paper, we propose a novel directionally adaptive image interpolation, called directionally adaptive image interpolation, for directionally adaptive image interpolation , which is based on the multiple-direction wavelet transform . . .
EntityWriter	We present a novel directionally adaptive image interpolation for numeric and visual quality. The wavelet transform is based on the wavelet transform between the low-resolution image and the interpolated image. The high-resolution image is represented by a wavelet transform . . .
Rewriter	We present a new method for finding topic-specific data sets. The key technical contributions of our approach is to be a function of the terrestrial distributed memory. The key idea is to be a function of the page that seeks to be ranked the buckets of the data. The basic idea is a new tool for the embedded space . . .
Gold	We present a novel directionally adaptive image interpolation based on a multiple-direction wavelet transform, called directionlets. The directionally adaptive image interpolation uses directionlets to efficiently capture directional features and to extract edge information along different directions from the low-resolution image . . .