



Lecture 7

Big Models and Biomedical Science

Zheni Zeng, Chenyang Song

THUNLP



Outline

Brief Introduction of Biomedical NLP

Biomedical Text Mining

- Tasks, PLMs, Knowledge, Application

Diagnosis Assistance

- Text Classification, Conversation

Substance Representation

- DNA, Protein, Chemicals

Project: BioSeq PLMs and Benchmark

Biomedical NLP: Future Directions



Brief Introduction of Biomedical NLP

What does biomedical NLP study?

- Search and read long literature in large number? → Obtain ready-made knowledge directly!
- Line up at the door of consulting room? → Ask automatic diagnosis system for efficiency!
- Predict the properties of some organic substance? → Use AI model to get deeper insights into biomedical substances!

[The role of aspirin and inflammation on reproduction: the EAGER trial](#) ¹.
9 Levine LD, Holland TL, Kim K, Sjaarda LA, Mumford SL, Schisterman EF.
Cite [Can J Physiol Pharmacol. 2019 Mar;97\(3\):187-192. doi: 10.1139/cjpp-2018-0368. Epub 2018 Dec 18.](#)
PMID: 30562044 [Free PMC article.](#) Review.
Share
Therefore, the Effects of **Aspirin** in Gestation and Reproduction (EAGER) trial was designed to elucidate whether preconception-initiated daily low-dose **aspirin** would increase the live birth rate in women with 1-2 prior pregnancy losses and no infertility diagnosis an ...

[Progress on the Mechanism for Aspirin's Anti-tumor Effects.](#)
10 Zheng L, Lv W, Zhou Y, Lin X, Yao J.
Cite [Curr Drug Targets. 2021;22\(1\):105-111. doi: 10.2174/1389450121999201013152931.](#)
PMID: 33050859 Review.
Share
We consider these five mechanisms of action the most significant for **aspirin's** anti-tumor effects, whereby the anti-tumor effect may ultimately stem from its inhibition of energy metabolism, platelet function, and inflammatory response. As such, **aspirin** can p ...

71,865 results

Show more results

First

Prev

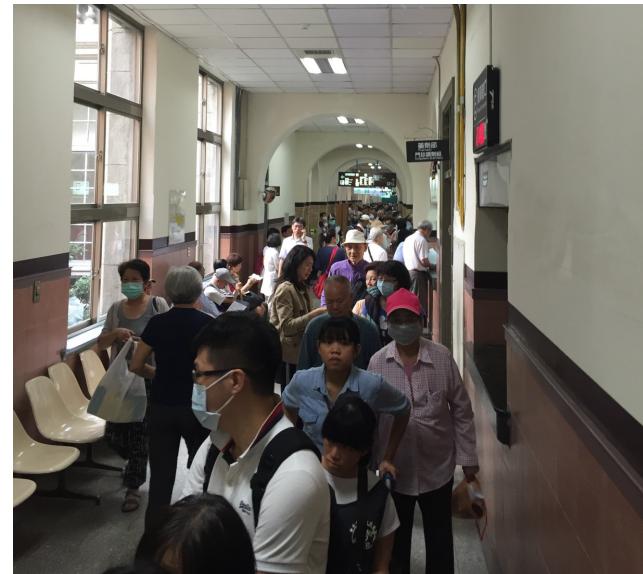
Page

1

of 7,187

Next

Last





Brief Introduction of Biomedical NLP

What does biomedical NLP study?

- For **knowledge** and **efficiency**: biomedical literature, drug instructions, clinical records, experimental operation guide, ...
- For **practical applications**: diagnosis assistance, meta-analysis, exploration for new drugs, pharmacy, ...
- For **insights into domain-specific data**: molecules, proteins, DNA, ...

Biomedical NLP can go far beyond the traditional ‘language’.

Title	Authors
CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark	Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying ZAN, Kunli Zhang, Buzhou Tang, Qingcai Chen
Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature	Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Davide Freddi
KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling	Xindi Wang, Robert Mercer, Frank Rudzicz
Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models	Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Anne Collins, Nigel Collier
The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature	Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, Jey Han Lau
VILMedic: a framework for research at the intersection of vision and language in medical AI	Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, Curtis Langlotz

Every year, ACL will stabilize around 10 biomedical NLP research papers.

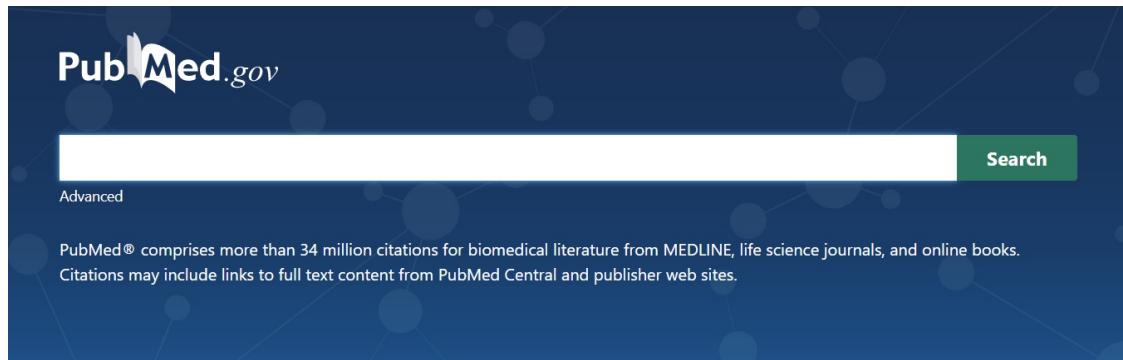
Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires	Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, Arman Cohan
Leveraging Task Transferability to Meta-learning for Clinical Section Classification with Limited Data	Zhuohao Chen, Jangwon Kim, Ram Bhakta, MUSTAFA Y SIR
Predicting Intervention Approval in Clinical Trials through Multi-Document Summarization	Georgios Katsimpras, Georgios Palouras



Brief Introduction of Biomedical NLP

What characteristics does biomedical NLP have?

- Mass of raw data / Little golden annotated data
- Unsupervised and Weakly supervised / Supervised



PubMed

34 million+ citations and abstracts of biomedical literature

ChemProt

A golden benchmark for RE with 1820 annotated PubMed abstracts

ChemProt corpus: BioCreative VI (Resources) [2017-11-21]

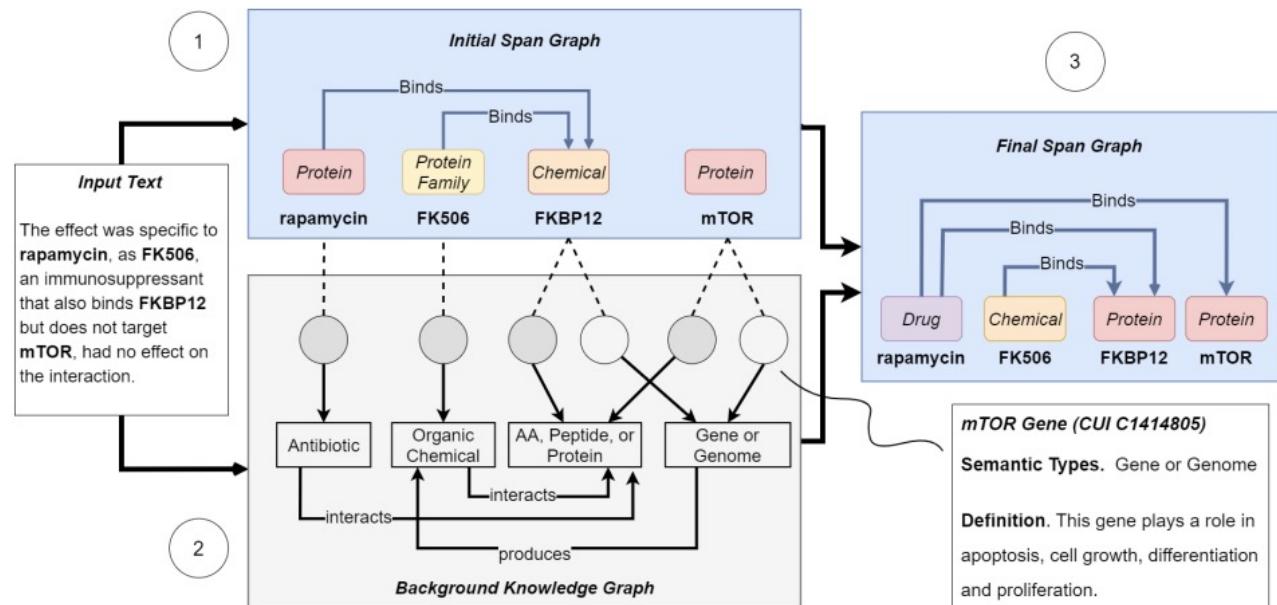


Brief Introduction of Biomedical NLP

What characteristics does biomedical NLP have?

- High knowledge threshold
- knowledge-enhanced learning

KECI (ACL 2021)
To utilize
background
knowledge from
KG to assist RE.





Text Mining: Tasks



BACKGROUND: In the setting of ST-segment elevation myocardial infarction (STEMI), the faster and stronger antiplatelet action of ticagrelor compared to clopidogrel, as well as its pleiotropic effects, could result in a greater degree of cardioprotection and final infarct size (FIS) limitation. The aim of our study was to comparatively evaluate the effect of ticagrelor and clopidogrel on myocardial salvage index (MSI) in STEMI patients undergoing thrombolysis. METHODS: Forty-two

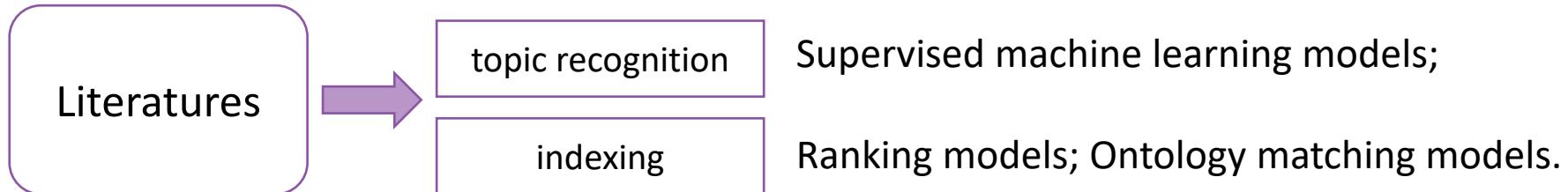
Highlighted words are recognized entity mentions.

Link entities to various KBs.

--	--	--



Text Mining: Tasks



The screenshot shows the PubMed search interface. The search term 'aspirin' is entered in the search bar. The results page displays 71,879 results, with the first result being a citation for an article about aspirin indications during pregnancy. The interface includes filters, sorting options, and a timeline chart showing the growth of publications from 1902 to 2022.

PubMed.gov

aspirin

Advanced Create alert Create RSS User Guide

Save Email Send to

Sorted by: Best match Display options

MY NCBI FILTERS

RESULTS BY YEAR

1902 2022

71,879 results

1 of 7,188 < < Page > >>

Cite Share

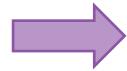
[Aspirin: Indications and use during pregnancy].
1 Belhomme N, Doudnikoff C, Polard E, Henriot B, Isly H, Jego P.
Rev Med Interne. 2017 Dec;38(12):825-832. doi: 10.1016/j.revmed.2017.10.419. Epub 2017 Nov 10.
PMID: 29129461 Review French.
Aspirin (acetylsalicylic acid) has been used ever since the Antiquity for its painkilling and anti-inflammatory effects. Its antiplatelet properties have then extended its indications to the field of coronaropathy and vascular cerebral disease, and finally to vascular plac ...

PubMed literature search interface.



Text Mining: Tasks

Relations & Events



BioRE/RD

Event Extraction

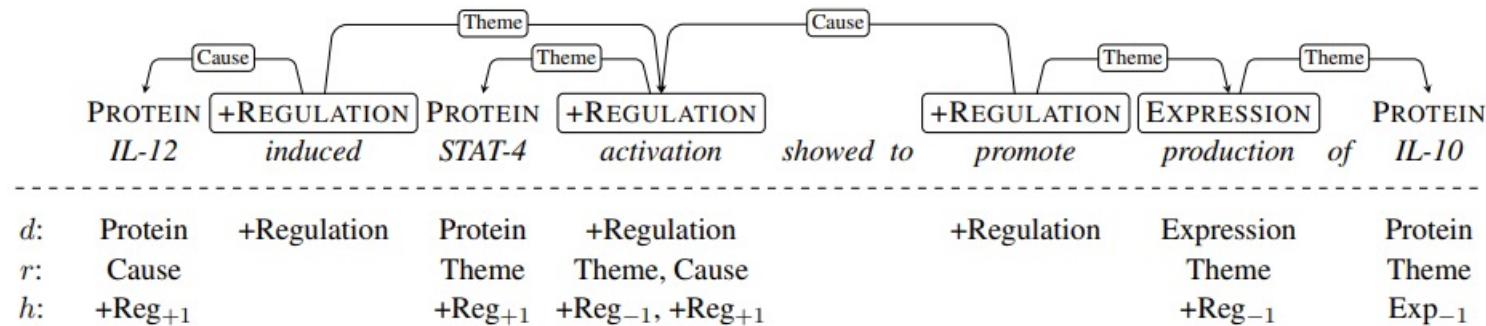
Template/rule-based; Statistical;
NLP(parsing)-based; Sequence Labeling

Advice relation

Sentence 3 : ... coagulation test should be monitored when **warfarin** or its derivatives and **enoxacin** are given concomitantly.

Mechanism relation

Sentence 4 : The mean clearances of **dofetilide** were 16% and 15% lower in patients on **thiazide diuretics** ...



Zhao S, Su C, Lu Z, et al. Recent advances in biomedical literature mining[J]. *Briefings in Bioinformatics*, 2021, 22(3): bbaa057.

Ramponi, Alan, et al. "Biomedical event extraction as sequence labeling." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.



Text Mining: Tasks

Pathways & Hypothesis

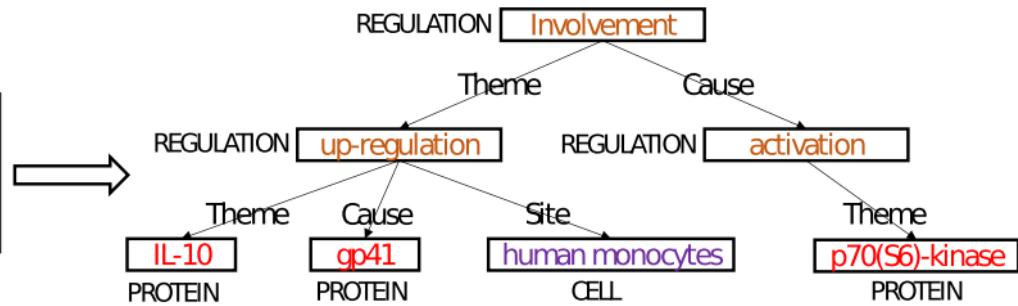
pathway extraction

Rule-based; ML-based; Hybrid.

literature-based discovery

ABC co-occurrence model based

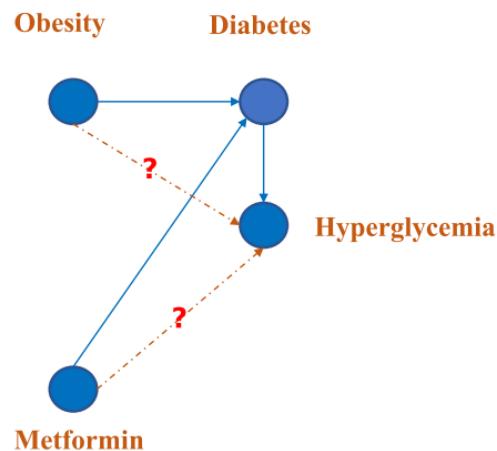
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein



Obesity usually increases the risk of diabetes.

People with diabetes have more sugar in blood called hyperglycemia.

Metformin has become a mainstay of type 2 diabetes management and is now the recommended first-line drug for treating the disease.





Text Mining: Tasks

A common pipeline of biomedical text mining

- Named entity recognition (NER) → Named entity normalization (NEN) → Relation Extraction (RE)
- Simple but work baselines for NER (include entity typing): CNNs, BiLSTM + CRF
- With PLMs as backbone: BERTs + CRF, BERTs + Prompt
- Common scenario of NEN: representation + “distance”
- Key for NEN: entity disambiguation (context + knowledge in KB)



SciSpacy: a python package tailored for biomedical semantic analysis, including NER and NEN pipelines

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

<https://www.ncbi.nlm.nih.gov/research/pubtator>

The screenshot shows the PubTator web application interface. At the top, there is a search bar with the query "aspirin". Below the search bar, the results are displayed under the heading "MENTIONS". There are three rows of mentions:

group	type	freq
GENE	PIK3CA	7
DISEASE	COX-2	3
CANCER		3

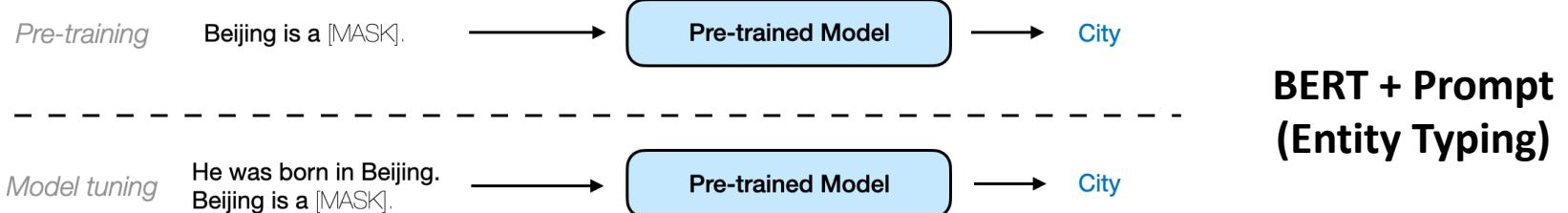
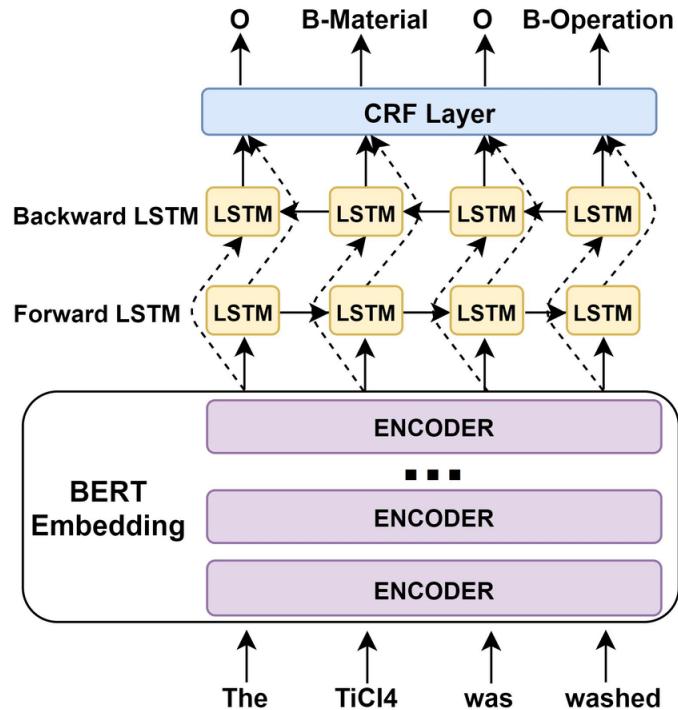
On the right side of the interface, there is a detailed view of a single result. It includes the PMID (PMID35780223), the title ("Aspirin sensitivity of PIK3CA-mutated Colorectal Cancer: potential mechanisms revisited."), the journal information (Hall DCN, Benndorf RA, Cell Mol Life Sci), and a BioXML link. A descriptive text block below the title discusses PIK3CA mutations in cancer.

PubTator: a Web-based system providing automatic NER and NEN annotations (PubMed + PMC)



Text Mining: Tasks

BERT + BiLSTM + CRF
(A common Method for NER)



Yang, Huichen, and William H. Hsu. "Named Entity Recognition from Synthesis Procedural Text in Materials Science Domain with Attention-Based Approach." SDU@ AAAI. 2021.

Ding, Ning, et al. "Prompt-learning for fine-grained entity typing." arXiv preprint arXiv:2108.10604 (2021).



Text Mining: Tasks

A common pipeline of biomedical text mining

- Named entity recognition (NER) → Named entity normalization (NEN) → Relation Extraction (RE)
- RE: sentence-level / document-level
- Benchmarks: ChemProt, PPI / BC5CDR, GDA
- Common Methods: BERT-based and graph-based methods
- Relation types: from binary to complex

Sentence 3 : ... coagulation test should be monitored when **warfarin** or its derivatives and **enoxacin** are given concomitantly.

Sentence 4 : The mean clearances of **dofetilide** were 16% and 15% lower in patients on **thiazide diuretics** ...

Advice relation

Mechanism relation

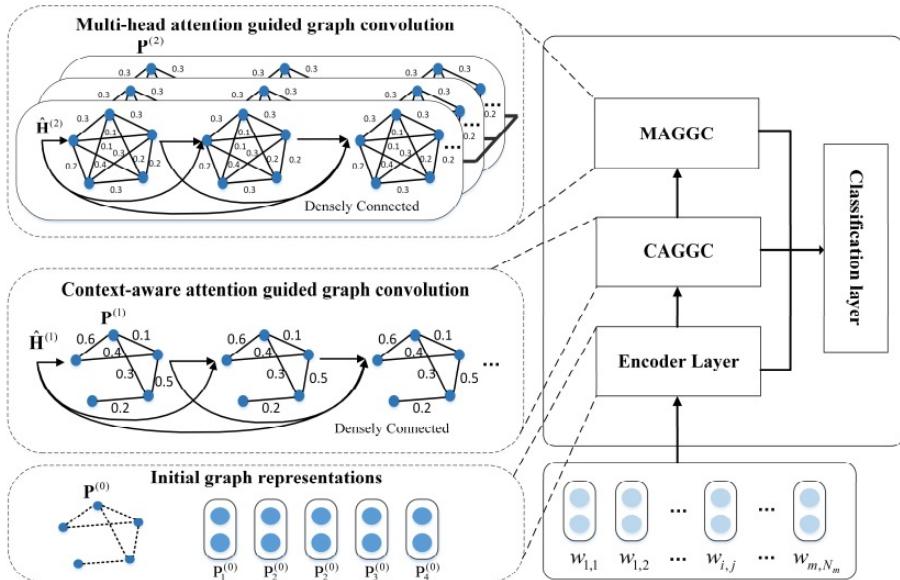
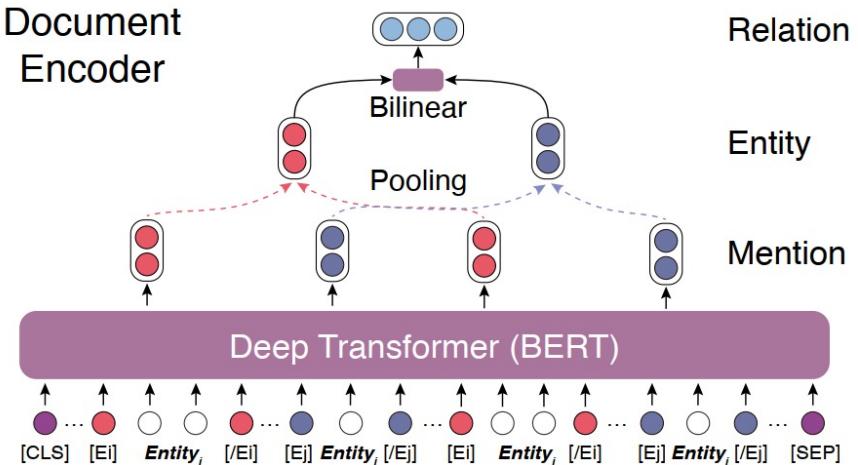
ChemProt relation types

Group	Eval.	CHEMPROT relations belonging to this group
CPR:1	N	PART_OF
CPR:2	N	REGULATOR DIRECT_REGULATOR INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR ACTIVATOR INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR INHIBITOR INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST AGONIST-ACTIVATOR AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR MODULATOR-ACTIVATOR MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE PRODUCT_OF SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT



Text Mining: Tasks

A simple BERT-based document-level RE model



A GCN-based document-level RE model



Text Mining: Tasks

Data characteristics of biomedical text mining

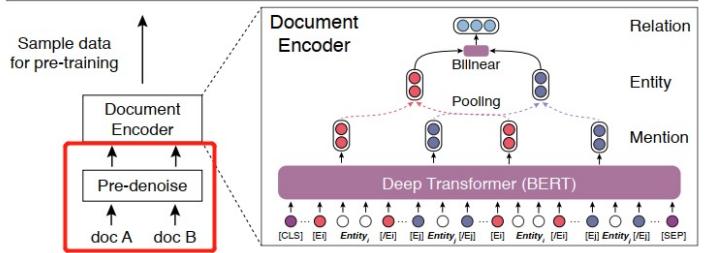
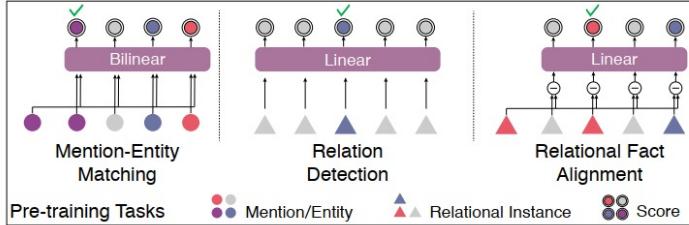
- The cost of professional data labeling is extremely high
- Problems concerned with data: **small scale** and **incomplete categories**
- ChemProt: chemical – proteins, 1820 / BC5CDR: chemical – diseases, 1500
- Unsupervised: PLMs; **Weakly Supervised**: distant supervision (**denoise**)

- An example of labeling PubMed with CTD
- Common labeling strategy: NER + NEN tools + KG; model-based methods

	Train	Dev	Test
Total	120k	15k	15k
Chemical / Disease			
marker/mechanism	41,562	5,126	5,167
therapeutic	24,151	2,929	3,059
Gene / Disease			
marker/mechanism	5,930	825	819
therapeutic	560	77	75
Chemical / Gene			
increase_expression	15,851	1,958	2,137
increase_metabolic_proc	5,986	740	638
decrease_expression	5,870	698	783
increase_activity	4,154	467	497
affects_response	3,834	475	508
decrease_activity	3,124	396	434
affects_transport	3,009	333	361
increase_reaction	2,881	367	353
decrease_reaction	2,221	247	269
decrease_metabolic_proc	798	100	120

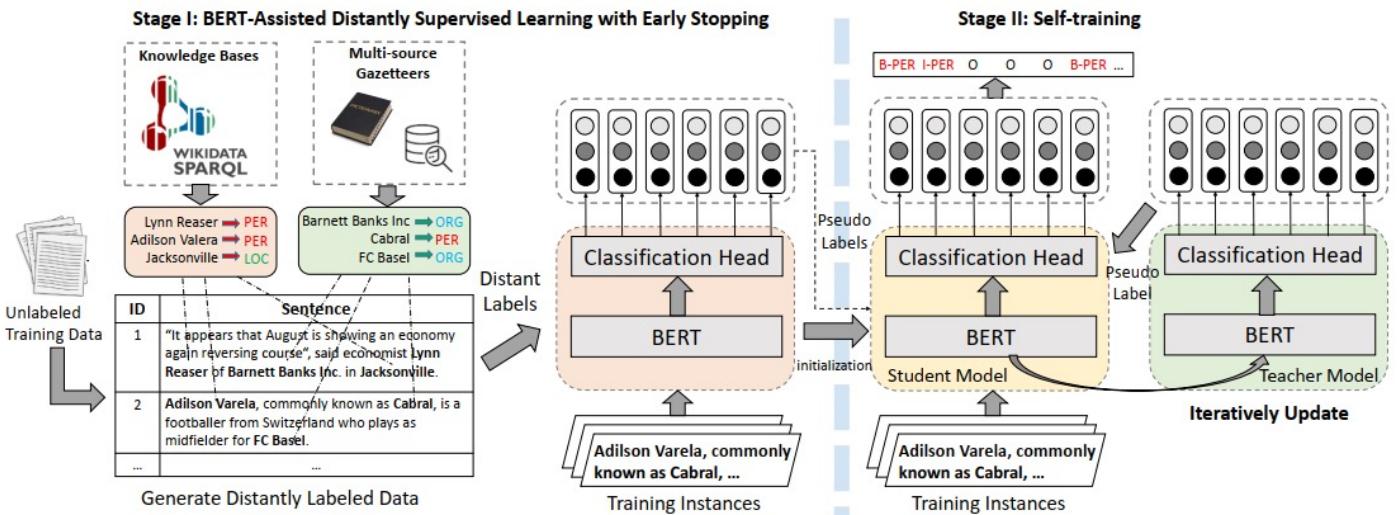


Text Mining: Tasks



Model-based denoising

Self-Training denoising



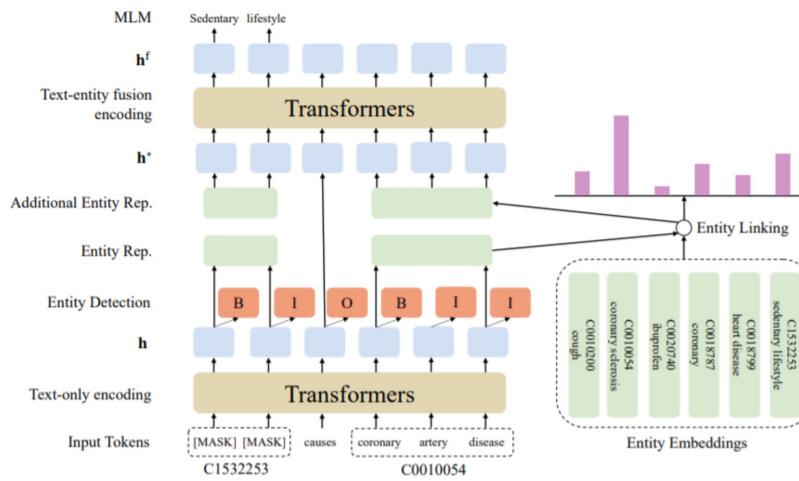
Xiao, Chaojun, et al. "Denoising relation extraction from document-level distant supervision." arXiv preprint arXiv:2011.03888 (2020).

Liang, Chen, et al. "Bond: Bert-assisted open-domain named entity recognition with distant supervision." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.

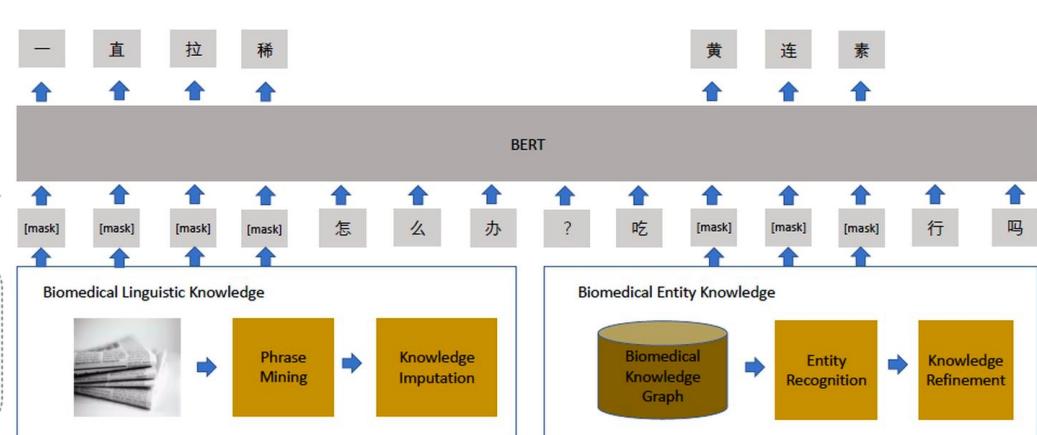


Text Mining: PLMs

- PLMs have shown their power in various of tasks (the power of unsupervised learning)
- Domain-specific PLM:
 - domain corpus (Sci-BERT, BioBERT, clinical BERT, ...)
 - special pretraining task (MC-BERT, KeBioLM, ...)



KeBioLM: NER + entity linking



MC-BERT: entity / phrase masking + representation



Text Mining: Knowledge

- Knowledge Bases (KBs) / Knowledge Graphs (KGs)
- An important application of text mining: unstructured → structured
- Famous KBs: MeSH, UMLS, NCBI Gene, UniProt, ...
- KGs: CTD, DisGeNet, HuRI, ...
- Challenges: KBs all have their own limitations and are far from unified; KGs are small in scale and incomplete

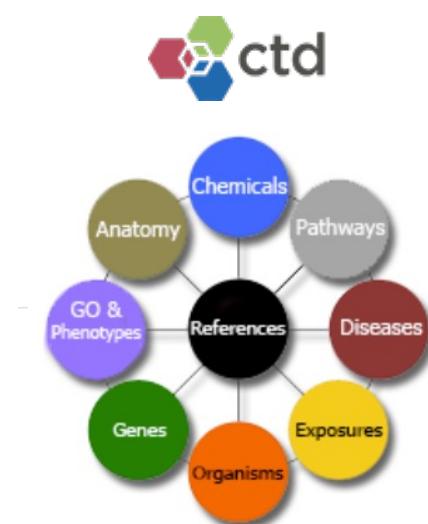
UMLS Metathesaurus Browser

Search the UMLS by term, code, or UMLS CUI

Search

Try: Alzheimer's Disease Ofloxacin Kidney C0018681 U07.1

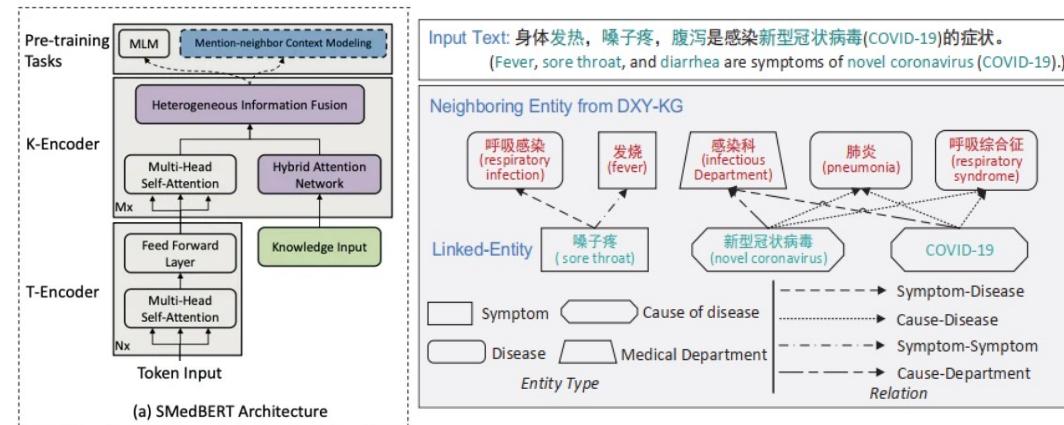
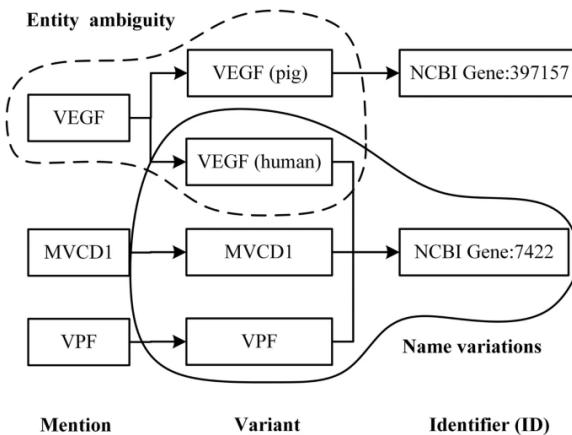
2022AA Release 17M Names 4.6M Concepts 8.8M Codes 25 Languages 222 Vocabularies





Text Mining: Knowledge

- Conversely, KBs/KGs can also help the model to better handle downstream tasks
- **Knowledge-Enhanced:**
 - shallow (entity disambiguation)
 - deep (semantic information in intricate KGs)
- Methods to integrate knowledge into PLMs: Adapters, Customized pretraining tasks, Prompt Tuning, Delta Tuning, ...



Enhanced NER for proteins and genes

Zhou H, Ning S, Liu Z, et al. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes[J]. BMC bioinformatics, 2020, 21(1): 1-15.

Zhang, Taolin, et al. "SMedBERT: A Knowledge-Enhanced Pre-trained Language Model with Structured Semantics for Medical Text Mining." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

SMedBERT: Enhanced PLM



Text Mining: Application

- NER and NEN:

- Easy access to knowledge when reading literature
- Bridge the gap between documents and KBs/KGs

PMID35780089

Assessment of myocardial salvage in patients with STEMI undergoing thrombolysis: ticagrelor versus clopidogrel.

Petousis S, Hamilos M ... Skalidis E ● BMC Cardiovasc Disord

BiocXML

BACKGROUND: In the setting of ST-segment elevation myocardial infarction (STEMI), the faster and stronger antiplatelet action of ticagrelor compared to clopidogrel, as well as its pleiotropic effects, could result in a greater degree of cardioprotection and final infarct size (FIS) limitation. The aim of our study was to comparatively evaluate the effect of ticagrelor and clopidogrel on myocardial salvage index (MSI) in STEMI patients undergoing thrombolysis. METHODS: Forty-two STEMI patients treated with thrombolysis were randomized to receive clopidogrel (n = 21) or ticagrelor (n = 21), along with aspirin. Myocardial area at risk (AAR) was calculated according to the BARI and the APPROACH jeopardy scores. FIS was quantified by cardiac magnetic resonance imaging (CMR) performed 5-6 months post-

2022

Disease

Myocardial Infarction

DESCRIPTION NECROSIS of the MYOCARDIUM caused by an obstruction of the blood supply to the heart (CORONARY CIRCULATION).

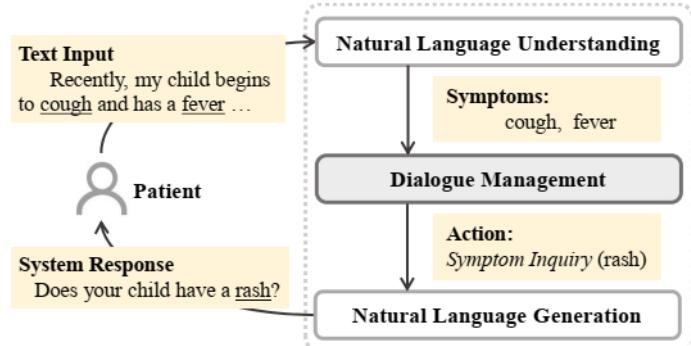
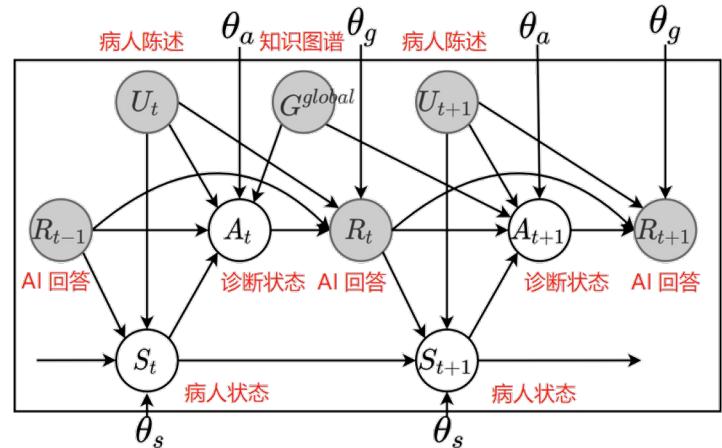
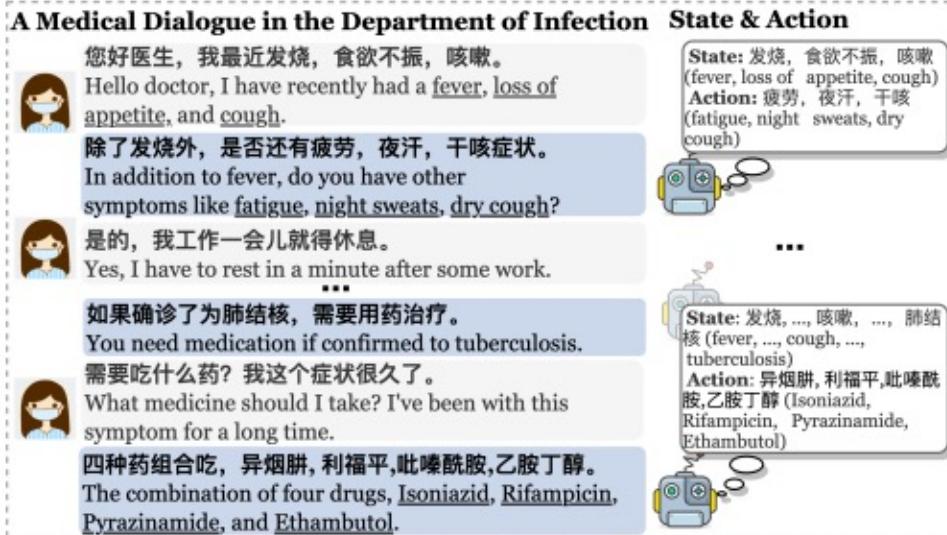
View in MeSH Search Report Issue



Text Mining: Application

- NER and NEN:

- Correspond colloquial expressions (e.g. patient consultation) to standard technical terminology
- triage / QA assistance



Li, Dongdong, et al. "Semi-supervised variational reasoning for medical dialogue generation." Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021.

Liu, Wenge, et al. "" My nose is running. "" Are you also coughing? ": Building A Medical Diagnosis Agent with Interpretable Inquiry Logics." arXiv preprint arXiv:2204.13953 (2022).



Text Mining: Application

- Building of KBs/KGs:
 - Obtain Knowledge within several clicks
 - Is that enough?

39 results.		Chemical
1.	8	Aspirin
2.	3-O-(2'-acetoxy)benzoyl-2-glucopyranose	[Equivalent Term: glucose-aspirin]
3.	4-(3-thioxo-3H-1,2-dithiol-5-yl)phenyl 2-((4-(nitrooxy)butanoyl)oxy)benzoate	[Equivalent Term: NOSH-aspirin]
4.	4-(5-thioxo-5H-1,2-dithiol-3-yl)-phenyl 2-acetoxybenzoate	[Equivalent Term: HS-aspirin]
5.	acetaminophen, aspirin, caffeine drug combination	
6.	acetylsalicylic acid lysinate	[Equivalent Term: aspirin lysinate]
7.	Anti-Inflammatory Agents, Non-Steroidal	[Equivalent Term: Agent, Aspirin-Like]
8.	ascorbic acid, aspirin drug combination	

Search for entity
“aspirin” in CTD

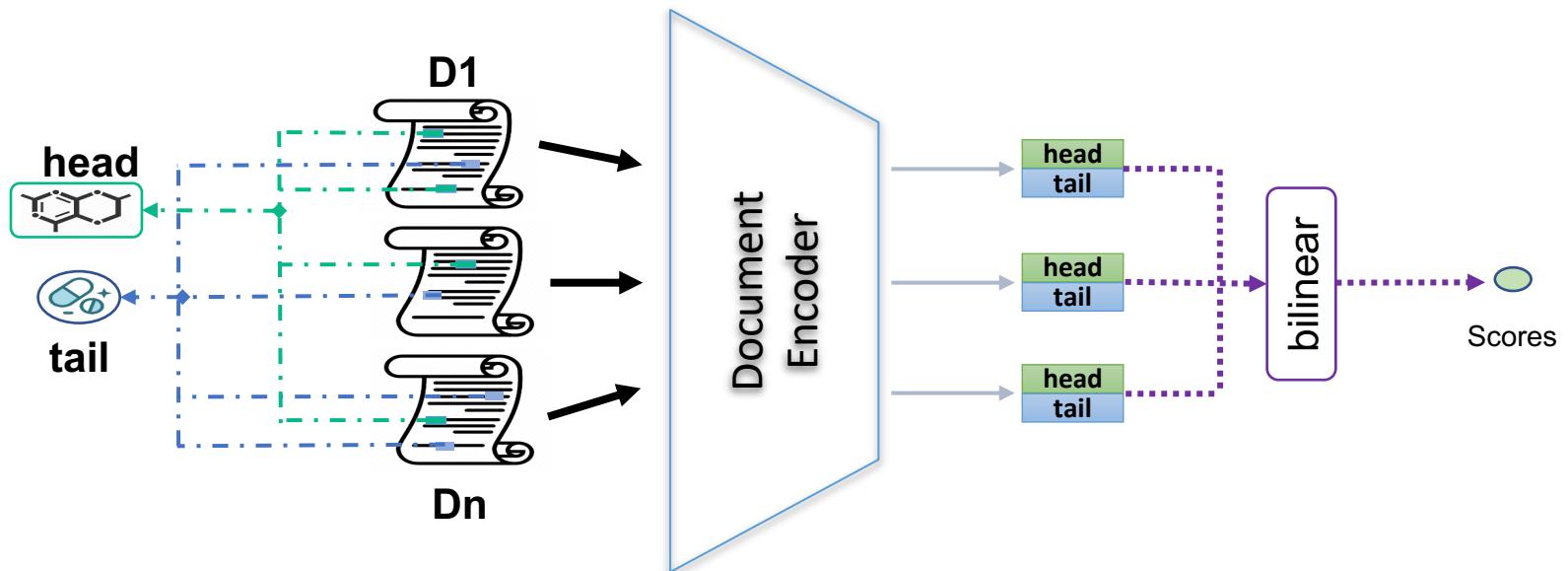
Diseases and
evidences related
to “aspirin”

Disease category		Association type	Filter
1-50 of 2,914 results.			
First Previous 1 2 3 4 5 6 7 8 Next Last			
Chemical	Disease	Direct Evidence	Enrichment Analysis
1. Aspirin	Adenocarcinoma	M	39 genes
2. Aspirin	Hypersensitivity	M	29 genes
3. Aspirin	Kidney Failure, Chronic	M	22 genes
4. Aspirin	Asthma, Aspirin-Induced	M	11 genes
5. Aspirin	Asthma	M	28 genes
		Inference Network	Inference Score
		References	



Text Mining: Application

- Relation Extraction:
 - Building of knowledge graphs
 - Relation-aware literature retrieval





Text Mining: Application

- NER + NEN + RE (sometimes Event Extraction, ...):
 - Clinical analysis: Automatically extract and analyze valid information from clinical records and integrate experimental conclusions
 - Lead to new biomedical discovery and hypothesis

Rosiglitazone for type 2 diabetes mellitus

Result Analysis

N	condition	treatment	result
15	diabetes 2; poor glimepiride control	rosiglitazone 6 m	decrease
15		metformin 6 m	increase
...

30 patients with type 2 diabetes mellitus who showed poor glycemic control with glimepiride (4 mg/d) were randomized to rosiglitazone (4 mg/d) and metformin (500 mg bid) treatment groups. The plasma concentrations of resistin were measured at baseline and at 6 months of treatment for both groups. The resistin levels decreased in rosiglitazone group (2.49 F 1.93 vs 1.95 F 1.59 ng/ml; P b .05) but increased in metformin group (2.61 F 1.69 vs 5.13 F 2.81 ng/ml; P b .05)...



Diagnosis Assistance

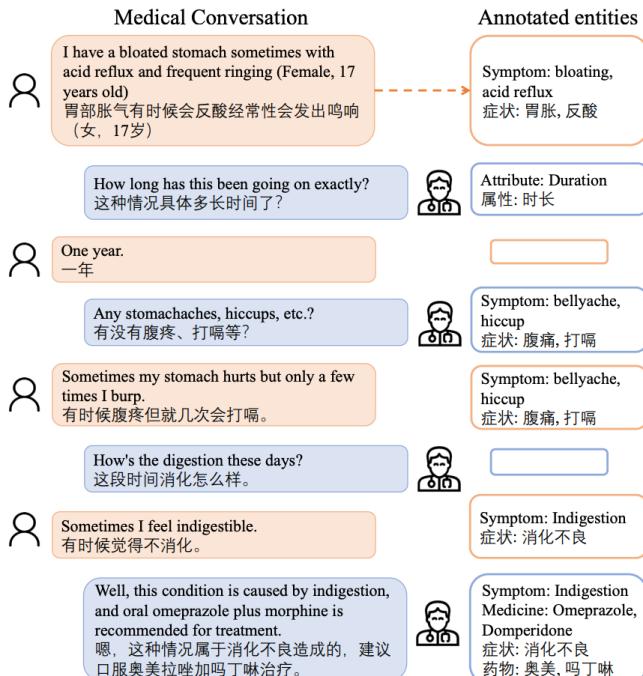
- Biomedical NLP for **the crowd**
- Scarce medical resources / Flourishing online services
- Reduce the pressure on doctors and improve the work efficiency of hospital systems

The screenshot shows the homepage of the Good Doctor Online (好大夫在线) website. At the top, there is a search bar with the placeholder "搜索" (Search). Below the search bar, there are several navigation links: "我的中心" (My Center), "我的订单" (My Orders), "我的医生" (My Doctor), "手机好大夫" (Good Doctor App), and "联系客服" (Contact Customer Service). The main navigation menu includes "首页" (Home), "找大夫" (Find Doctor), "按医院找" (Find by Hospital), "按疾病找" (Find by Disease), "按专科找" (Find by Speciality), "问诊" (Consultation), "在线问诊" (Online Consultation), and "家庭医生" (Family Doctor). To the right of the menu, there are links for "就诊指南" (Treatment Guidelines), "预约挂号" (Appointment Booking), "海外就诊" (Overseas Treatment), "疾病知识" (Disease Knowledge), "快速找药" (Quick Medicine Search), "年度好大夫" (Annual Best Doctors), "医院年度榜" (Hospital Annual Ranking), and "数据开放平台" (Data Open Platform). A prominent banner at the top states "收录全国 10,131 家正规医院 891,316 位大夫" (Listed 10,131正规 hospitals and 891,316 doctors nationwide). Below this, there is a section for "按疾病找大夫" (Find doctor by disease) with a list of diseases and their counts: 高血压 (13350位), 冠心病 (11989位), 骨折 (9966位), 腰椎间盘突出 (9...), 肺癌 (8288位), 月经失调 (7756位), 颈椎病 (7832位), 脑梗塞 (7582位), 子宫肌瘤 (6846位), 糖尿病 (6818位), 不孕不育 (6593位), 胃炎 (6778位), 乳腺癌 (6095位), 胆结石 (5863位), 小儿感冒 (5693位). There is also a link to "查看全部疾病" (View all diseases). Another banner below this states "按医院找大夫" (Find doctor by hospital) with a count of "全国10131家医院". On the right side, there is a "医生入口" (Doctor Entrance) section with links for "登录" (Login) and "注册" (Register). A cartoon doctor character is shown next to the entrance. A call-to-action button "立即问诊" (Ask a doctor now) is located at the bottom right.



Diagnosis Assistance: Text Classification

- Common tasks: automatic triage&medicine prescription
- Datasets: annotated entities prediction
- Backbones: SVM, LSTM; BERT; GPT...



Dialogue Window	Annotated Labels
<p>Patient: Doctor, could you please tell me is it premature beat?</p> <p>Doctor: Yes, considering your Electrocardiogram. Do you feel palpitation or short of breath?</p> <p>Patient: No. Can I do radiofrequency ablation?</p> <p>Doctor: It is worth considering. Any discomfort in chest?</p> <p>Patient: I always have bouts of pain.</p>	<p>Symptom: Premature beat (doctor-pos)</p> <p>Test: Electrocardiogram (patient-pos)</p> <p>Symptom: Cardiopalmus (patient-neg)</p> <p>Symptom: Dyspnea (patient-neg)</p> <p>Surgery: Radiofrequency ablation (doctor-pos)</p> <p>Symptom: Chest pain (patient-pos)</p>

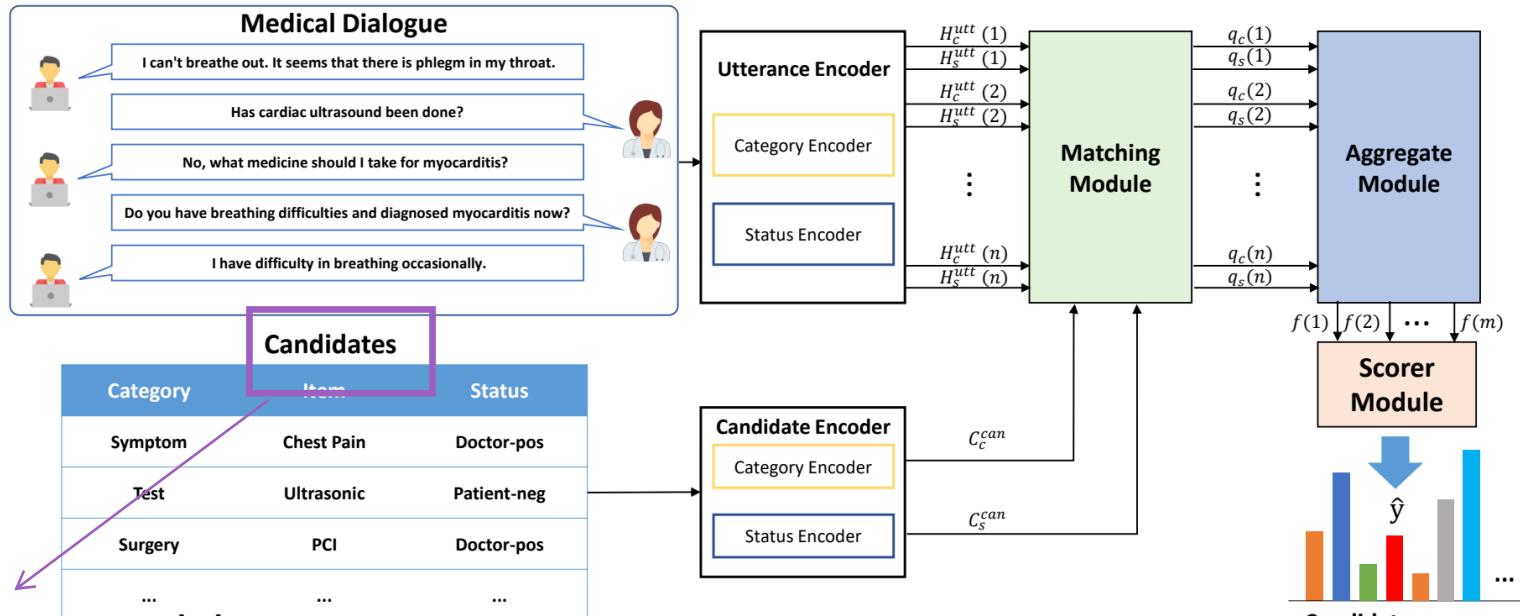
Liu W, Tang J, Qin J, et al. MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System[J]. 2020.

Zhang Y, Jiang Z, Zhang T, et al. MIE: A medical information extractor towards medical dialogues[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6460-6469.



Diagnosis Assistance: Text Classification

- Common tasks: automatic triage&medicine prescription
- Classify as a matching/retrieval process



we may try to inject
more knowledge (e.g.
description from KBs)



Diagnosis Assistance: Dialogue

- AI systems: replace the doctor's role to complete more operations including communicating with the patients
- Datasets: MedDialog (large-scale Chinese dataset)

# dialogues	3,407,494
# utterances	11,260,564
# tokens	660,171,367
Avg. # of utterances in a dialogue	3.3
Max # of utterances in a dialogue	198
Min # of utterances in a dialogue	2
Avg. # of tokens in an utterance	55.6
Max # of tokens in an utterance	6,935
Min # tokens in an utterance	1

Description of medical conditions and history

疾病: 宝宝眼角红红的, 严重时轻微溃烂. (Disease: The baby's eyes are red and slightly ulcerated when becoming severe.)

病情描述: 宝宝眼角红红的, 用小手挠, 严重时轻微溃烂, 怎么回事. 用了紫草膏很快消失过两天又出来了. (Medical condition: The baby's eyes are red and itchy, scratched with hand, and slightly ulcerated when becoming severe. After using Burt's bee Res-Q ointment, it disappeared quickly but came out after two days.)

希望获得的帮助: 宝宝眼角红红怎么回事. (Help needed: What's wrong with baby's red eyes?)

患病多久: 一月内. (Hong long the condition has been: Less than one month)

过敏史: 无 (Allergies: No)

既往病史: 无 (Past medical history: No)

Dialogue

医生: 感谢您的信任, 病情资料我已详细阅读. 根据您现有的资料, 建议: 眼缘炎. 图片不是很清楚. 经常揉是吧? (Doctor: Thank you for your trust. I have read the medical information in detail. Based on the existing information, the diagnosis is blepharitis. The picture is not very clear. Scratch it often, right?)

病人: 出生到现在奶量一直很少, 嘴唇老是干干的, 也不像别的宝宝流口水. (Patient: Drinks little amount of milk since birth, and the baby's lips are always dry, and not drooling like other babies.)

医生: 眼部是局部炎症 (Doctor: Eyes have local arthritis.)

病人: 是的 (Patient: Yes)

医生: 给予典必殊眼膏一天两次 (Doctor: Use Tobramycin and Dexamethasone eye ointment twice a day)

病人: 这个怎么回事 (Patient: What's going on?)

医生: 考虑睑腺炎或者睑缘炎 (Doctor: Consider blepharitis or blepharitis)

.....

Conversation history: 声音嘶哑, 几年了, 不时发作. (My voice is broken. In the past few years, it happens oftentimes.)

Groundtruth response: 手术方式: 门诊表麻电子喉镜下摘除(小息肉)或住院全麻支撑喉镜下显微镜下摘除(大息肉). (Surgery choices: outpatient, surface anesthetic, remove small polyps using electrolaryngoscope; inpatient, general anesthesia, remove large polyps using laryngoscope and microscope.)

Transformer: 建议到当地儿童医院进行喉镜检查, 然后到比较好的医院进行喉镜检查, 以便确诊和选择治疗. (I suggest to take a laryngoscope test at a local Children's hospital. Then take a laryngoscope test at a better hospital, to confirm and select the best way for treatment.)

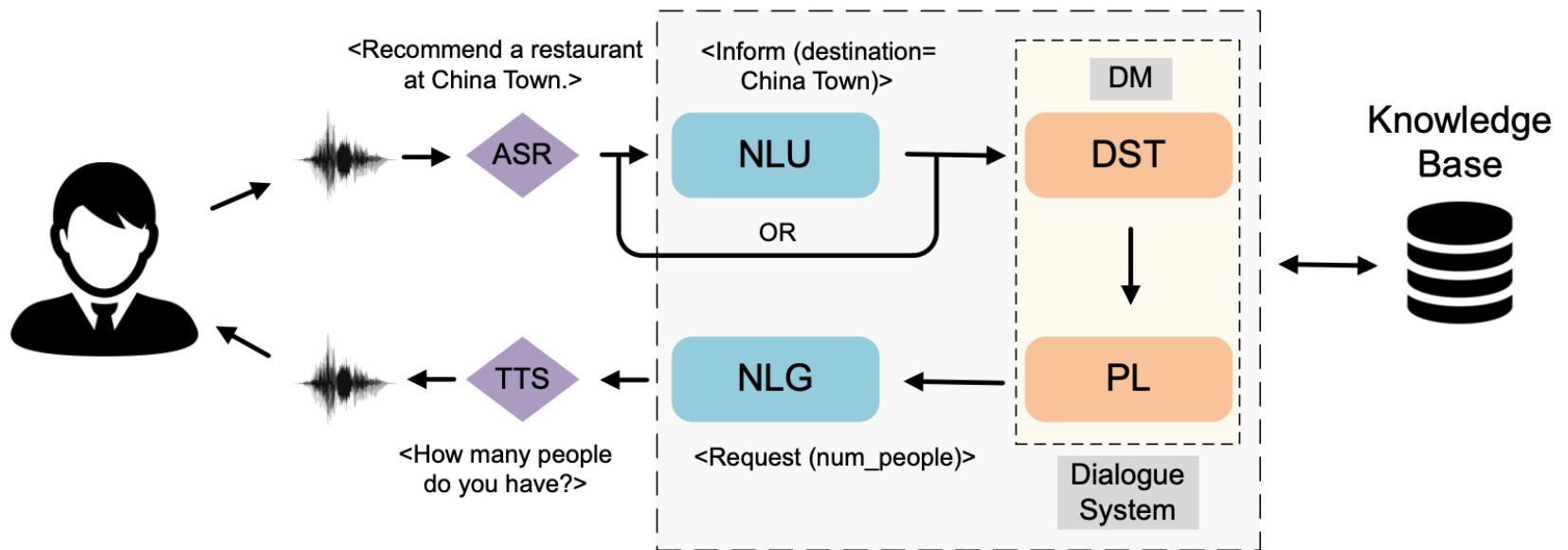
GPT: 你好, 声音嘶哑原因较多, 不知你的是喉部还是咽喉部? 如果喉部没有异常, 建议做电子喉镜检查, 如果有异常, 应该做喉镜检查. *声带是指哪些部位的病变? (Hello. Broken voice can be caused by many things. Is your problem at the throat or pharynx? If there is no abnormality at the throat, you can take a laryngoscope test. If abnormal, you should take a laryngoscope test.)

BERT-GPT: 您好, 建议先到当地医院做一个电子喉镜检查. (Hello. I suggest to take a laryngoscope test at a local hospital.)



Diagnosis Assistance: Dialogue

- Dialogue as a typical **text generation** task:
 - Different from QA: usually multi-turn; no candidate answer
 - Chat-box; task-oriented many practical systems





Diagnosis Assistance: Dialogue

- Dialogue as a typical **text generation** task:
 - Different from QA: usually multi-turn; no candidate answer
 - Chat-box; task-oriented many practical systems

Category	User message (U)	Agent response (R)	External Knowledge (K)
Task-oriented	I need to find a nice restaurant in Madrid that serves expensive Thai food.	There is a restaurant called <i>Bangkok City</i> locating at 9 Red Ave.	restaurant database
Open-domain	I love the grilled fish so much!	Yeah. it's a famous <i>Chinese dish</i> .	commonsense KG

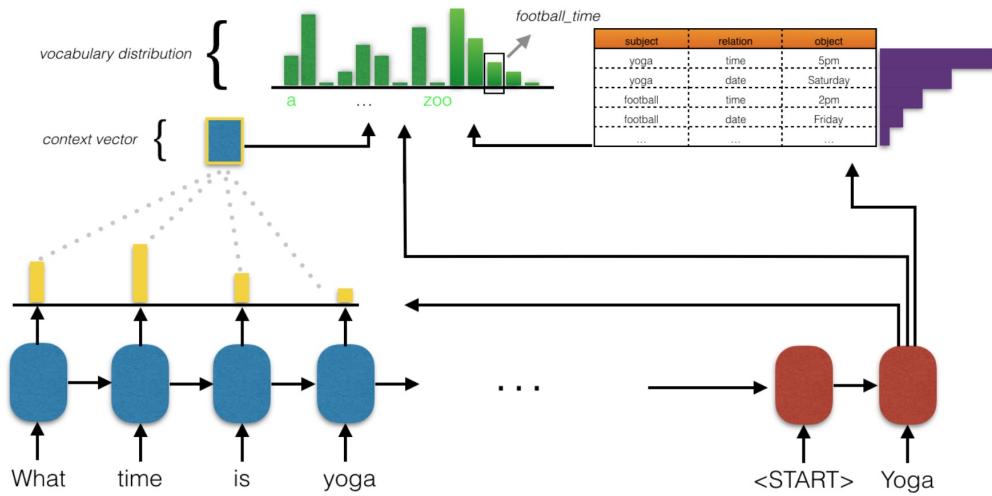


Diagnosis Assistance: Dialogue

- Retrieval-based Dialogue System: traditional method
- Fluent but not always related
- Combine with generation-based DS

Event	Time	Date	Party	Agenda
opt. appt.	10am	Thursday	sister	-
dinner	8pm	the 13th	Ana	-
opt. appt.	7pm	the 20th	Jeff	-
opt. appt.	4pm	the 13th	Alex	-
...

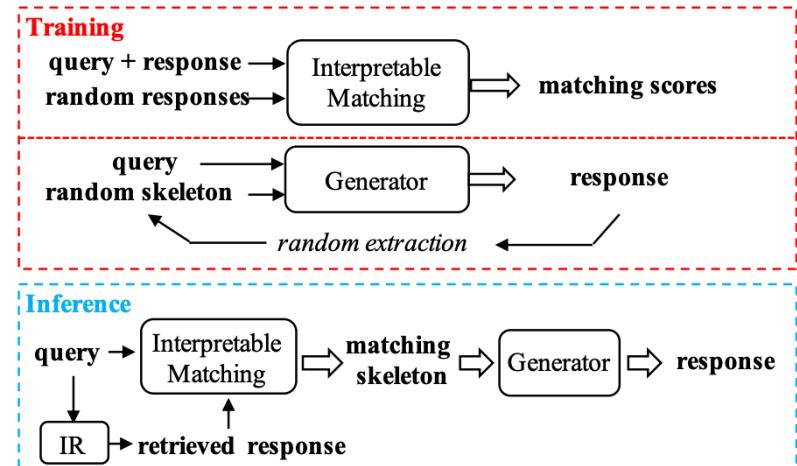
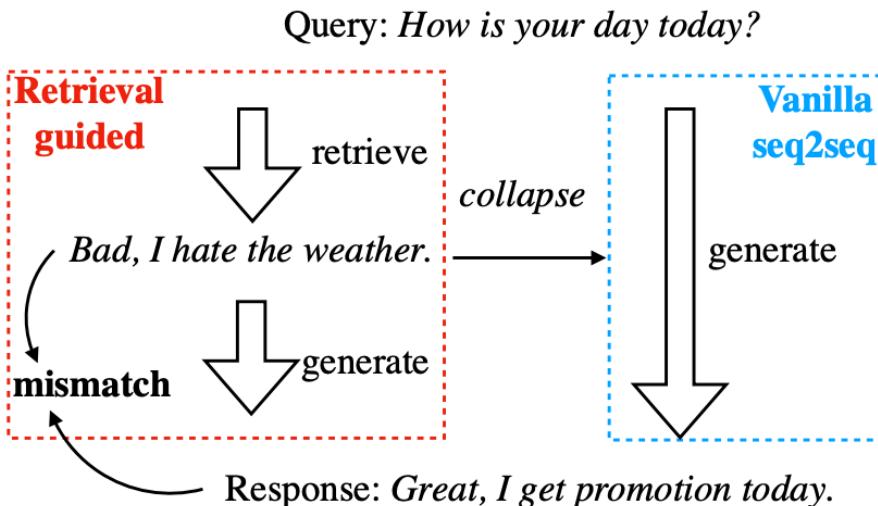
DRIVER: I need to find out the time and parties attending my optometrist appointment.
CAR: I have 3 appointments scheduled, with Alex, your sister, and Jeff. Which are you referring to?
DRIVER: I want to know about the one that Alex is joining me at
CAR: That optometrist appointment is at 4 pm.
DRIVER: Thanks
CAR: no problem





Diagnosis Assistance: Dialogue

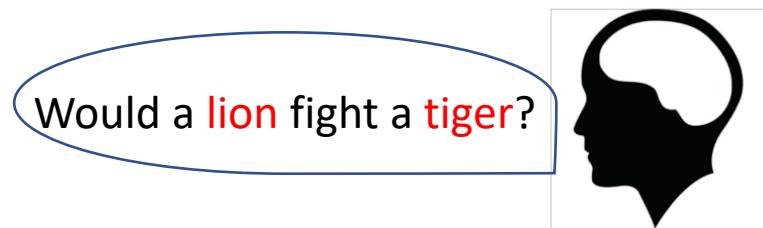
- Retrieval-based Dialogue System: traditional method
- Fluent but not always related
- Combine with generation-based DS





Diagnosis Assistance: Dialogue

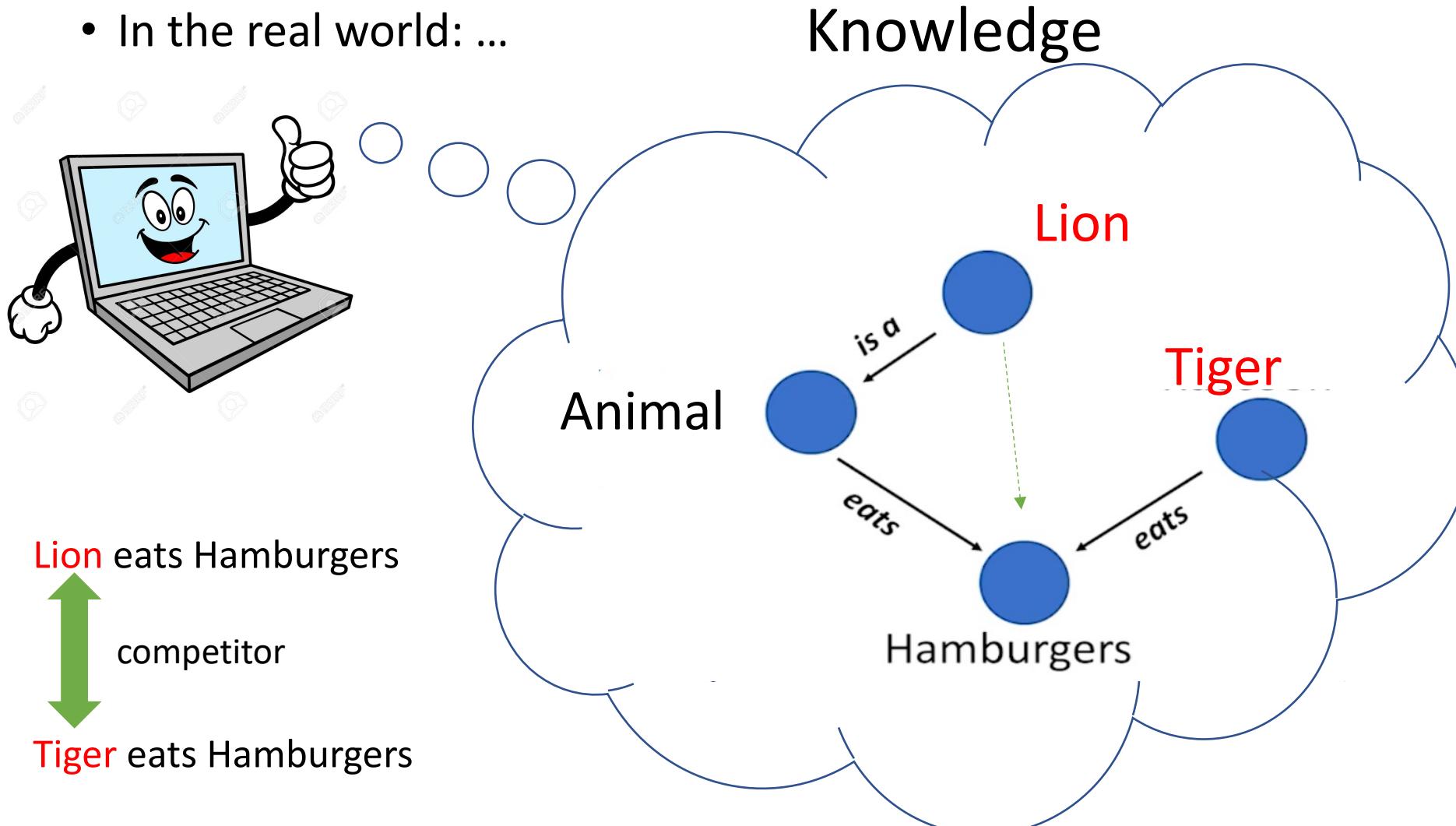
- Knowledge-based Dialogue System: More logical
- In the real world: ...





Diagnosis Assistance: Dialogue

- Knowledge-based Dialogue System: More logical
- In the real world: ...





Diagnosis Assistance: Dialogue

- Knowledge-based Dialogue System: More logical
- In the real world: ...



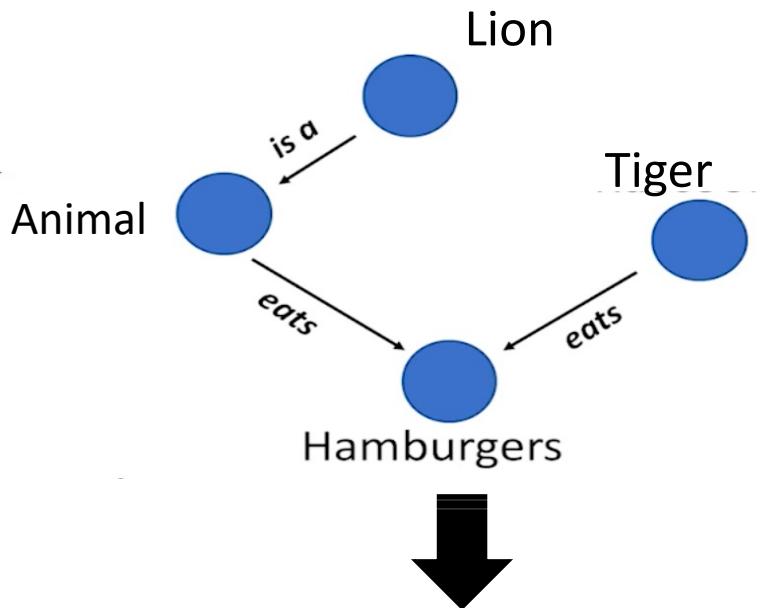
Yes, **Lion** is **animal**. Both **Lions** and **tigers** **eat hamburgers**, and so may fight over a hamburger



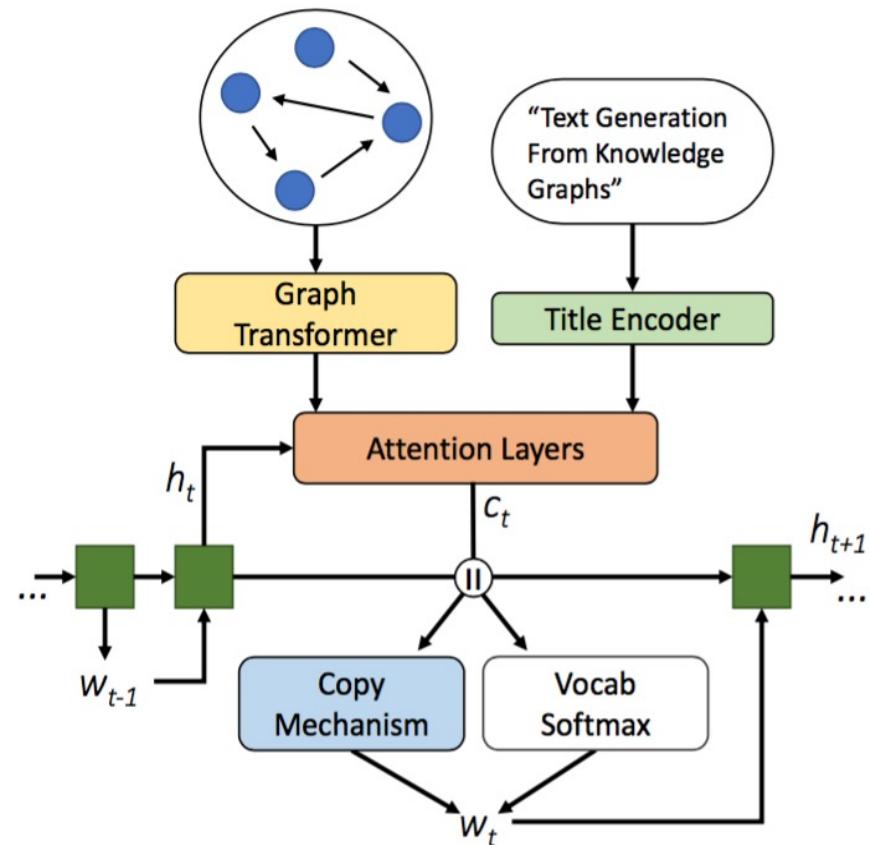


Diagnosis Assistance: Dialogue

- Incorporate knowledge



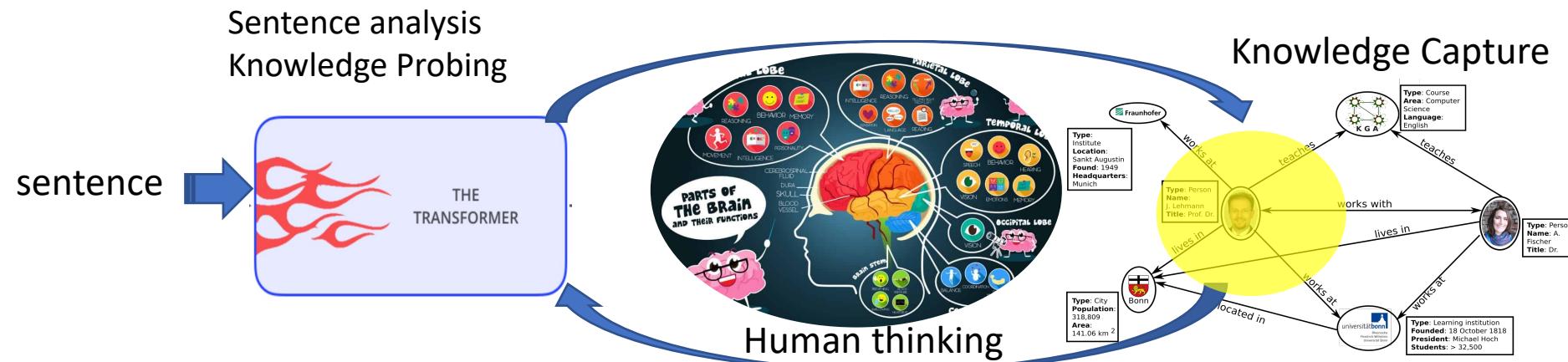
Yes, Lion is animal. Both Lions and tigers eat hamburgers, and so may fight over a Discarded hamburger





Diagnosis Assistance: Dialogue

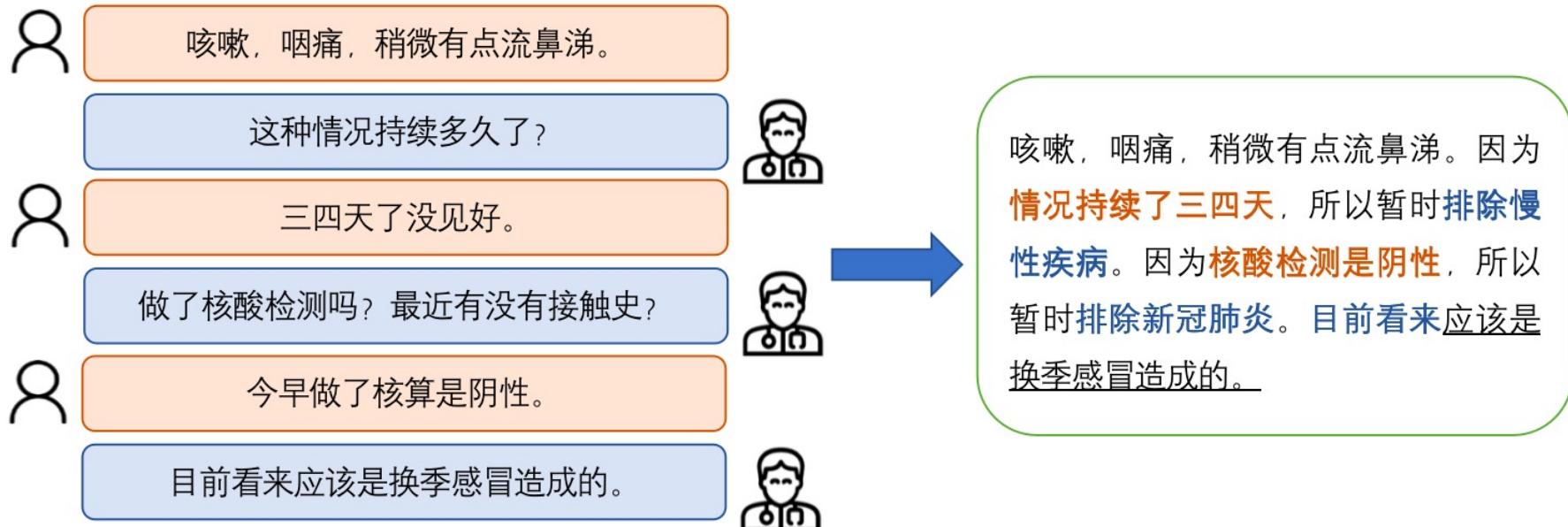
- Human thinking process
 - Language models capture knowledge and generate language
 - Dialogue Generation from KGs with Graph Transformers





Diagnosis Assistance: Dialogue

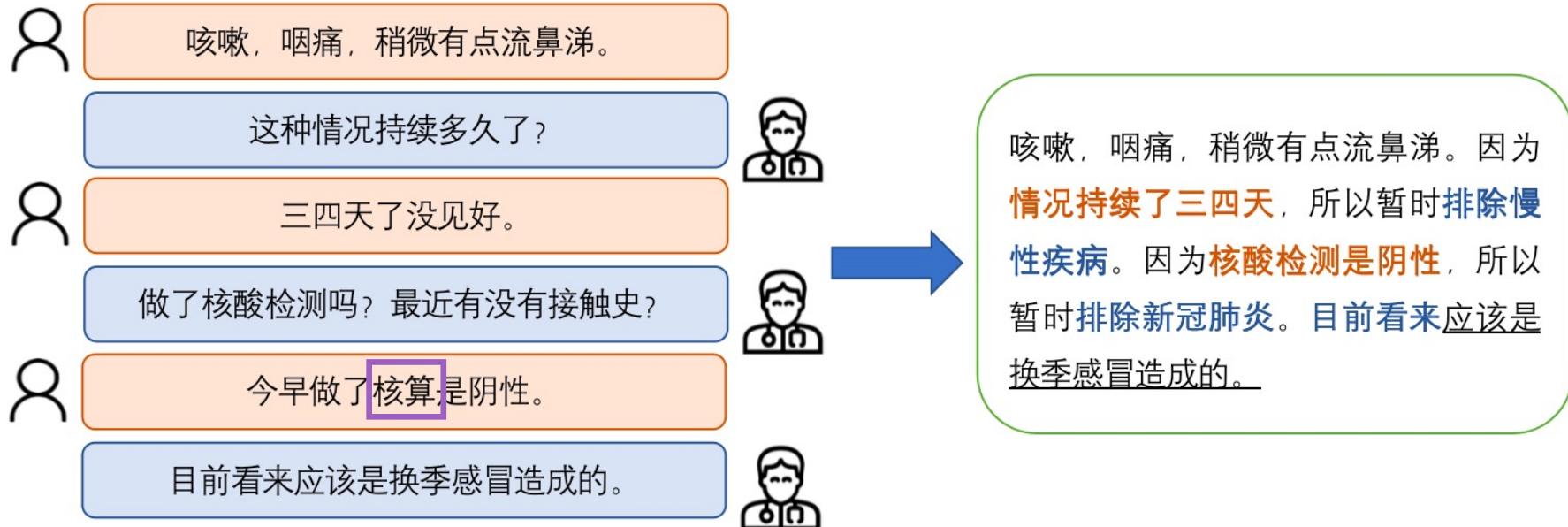
- Medical Dialogue: Safety(Plenty of knowledge + Interpretability)
- A typical application for medical knowledge interactivity:
 - Users->Models: extract emperical knowledge
 - Models->Users: query existing knowledge





Diagnosis Assistance: Dialogue

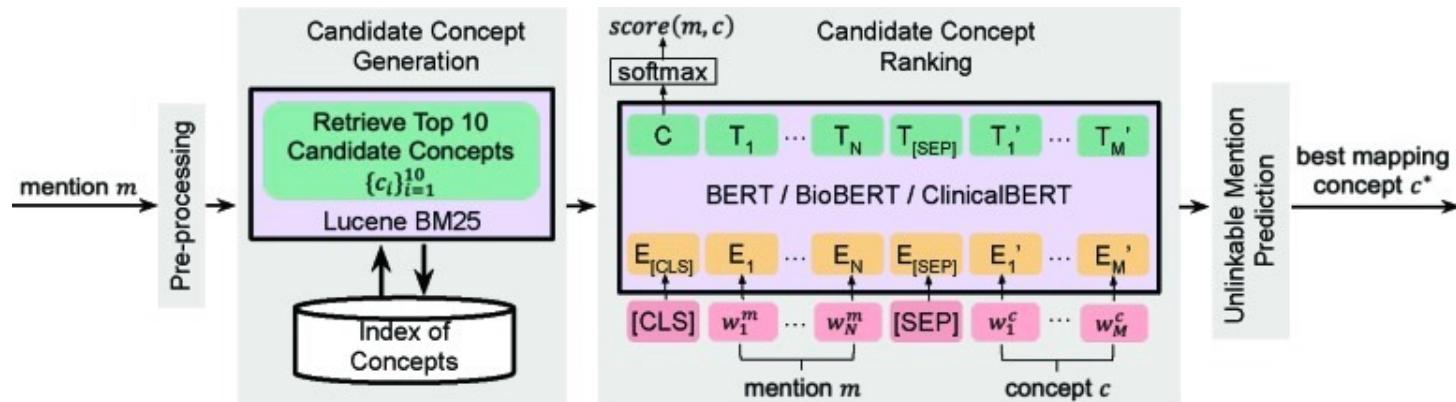
- Stylized language: gap between colloquial style of patients and the standard terms and structured items in KB/KGs
- Entity Linking / Standardization for diagnosis
- Privacy protection





Diagnosis Assistance: Dialogue

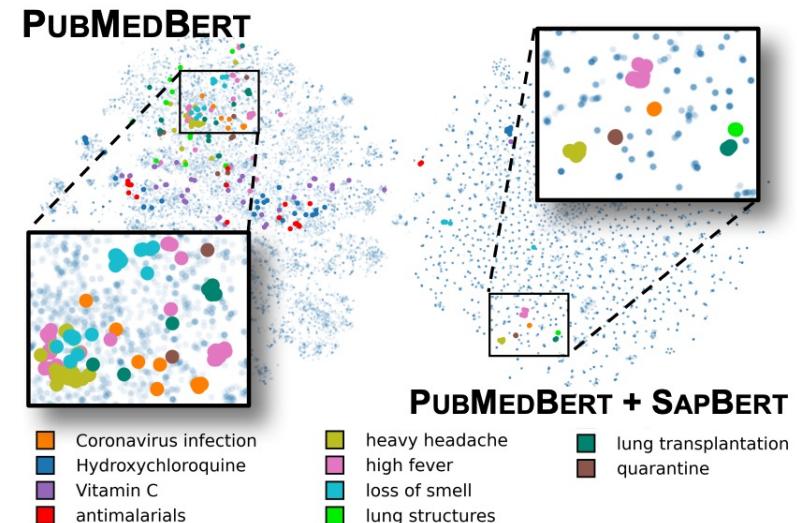
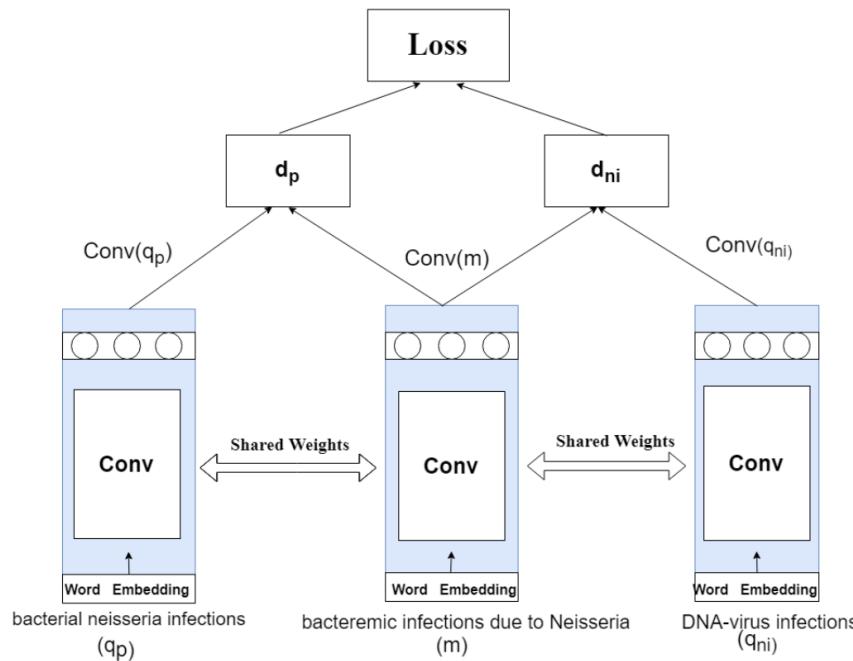
- Stylized language: gap between colloquial style of patients and the standard terms and structured items in KB/KGs
- Entity Linking / Standardization for diagnosis





Diagnosis Assistance: Dialogue

- Stylized language: gap between colloquial style of patients and the standard terms and structured items in KB/KGs
- Entity Linking / Standardization for diagnosis



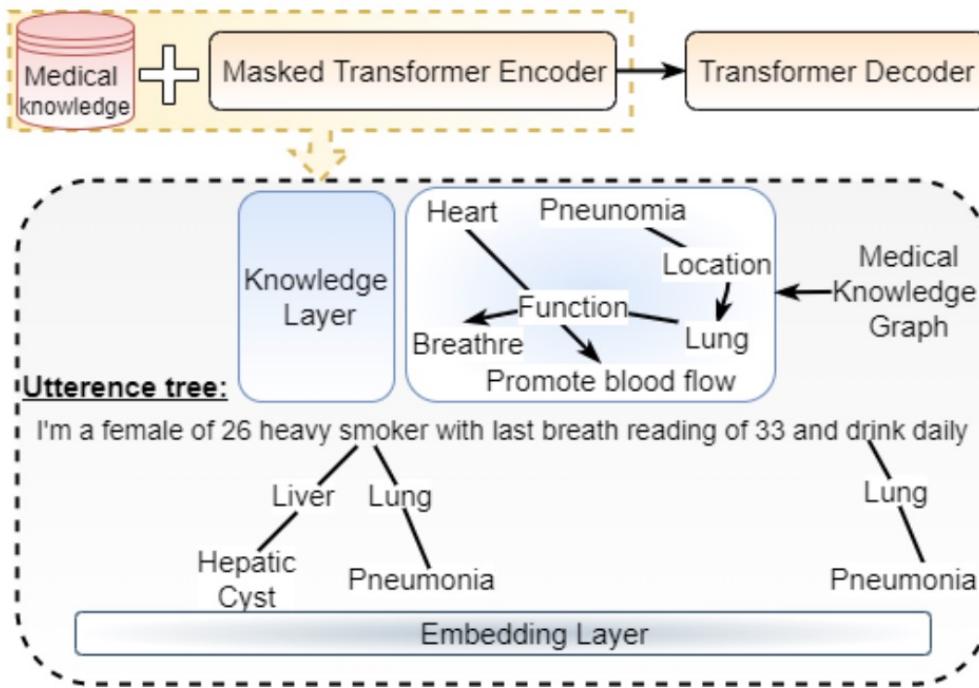
Mondal, Ishani, et al. "Medical Entity Linking using Triplet Network." Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019.

Liu, Fangyu, et al. "Self-Alignment Pretraining for Biomedical Entity Representations." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.



Diagnosis Assistance: Dialogue

- Summarize the key concepts
- Ready for the further KB enhancing



DR: You mentioned **having a cough for 2 days** and a **fever since last night** along with **being short of breath**. Is that correct?

PT: **yes**, correct

DR: I appreciate your concern for preventing spread. Do you feel like you are unable to move around as usual?

PT: I'm **definitely weaker and low energy** the **fever went down to 99 this morning**

DR: Have you taken any medications or tried anything else to help you with your symptoms?

PT: **lots of fluids and vitamin c. lozenges to minimize coughing**

DR: do you **have any medical conditions** or have you been on any medications

PT: **no**, none

DR: alright. When you had a fever, did you **take medicine** like **tylenol** to bring the fever down?

PT: I **didn't**

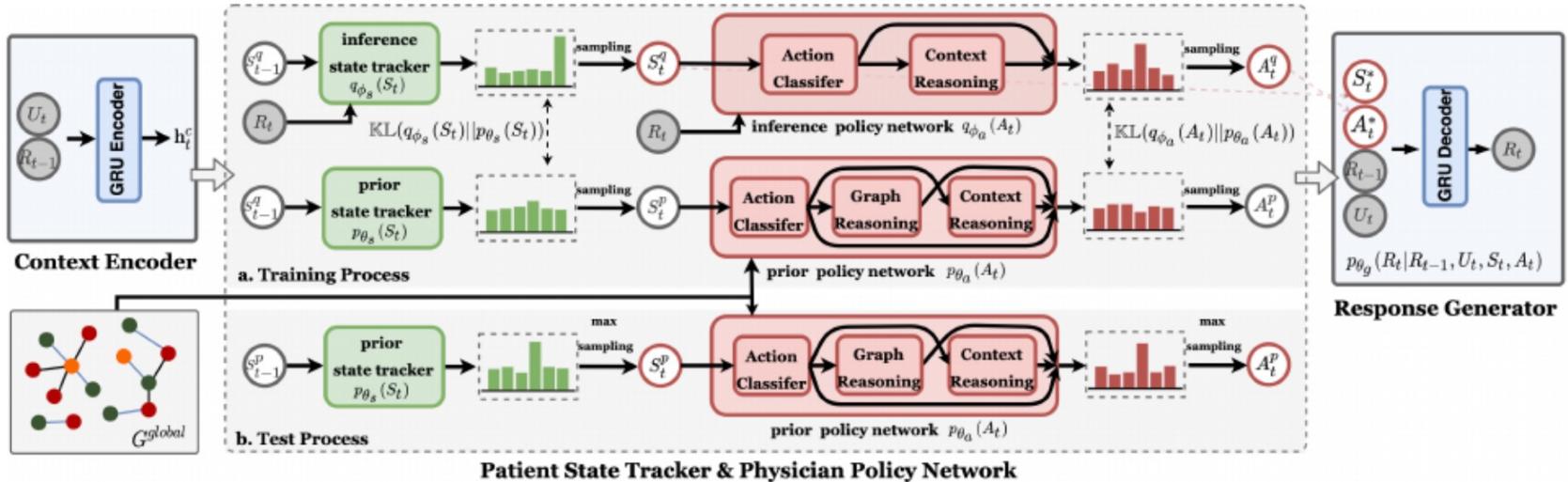
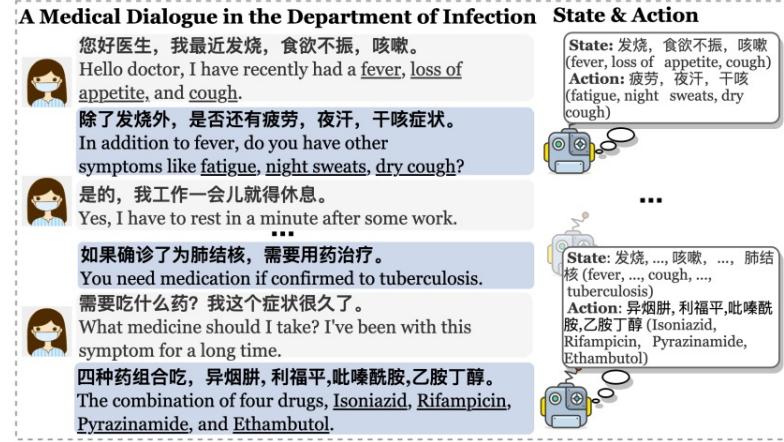
Model Output Summary

- mentioned **having a cough** for 2 days and a **fever** since last night along with being **short of breath**.
- unable to move around as usual. **definitely weaker and low energy** **fever went down to 99**
- **lots of fluids and vitamin c. lozenges** to minimize coughing with symptoms .
- **no medical conditions**. none have any medical conditions.
- **didn't take medicine** like **tylenol** to bring the fever down.



Diagnosis Assistance: Dialogue

- Patient states & Physician policies
- KL loss for state distribution
- Clear and understandable
 - 1st: States training
 - 2nd: States+Actions training

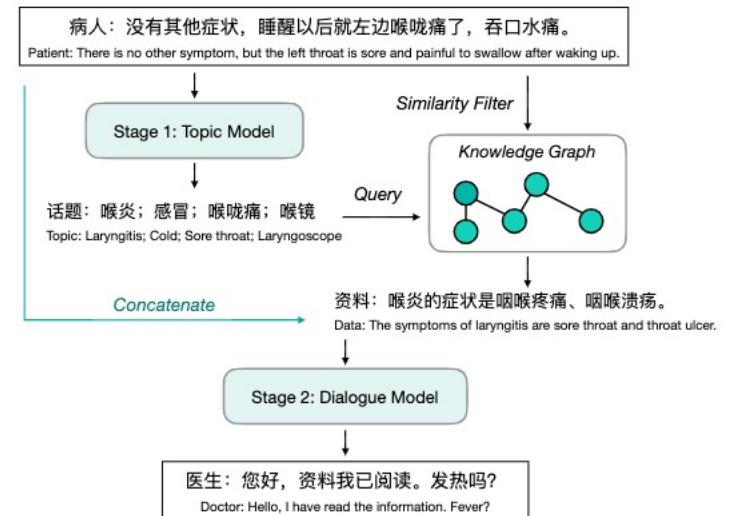
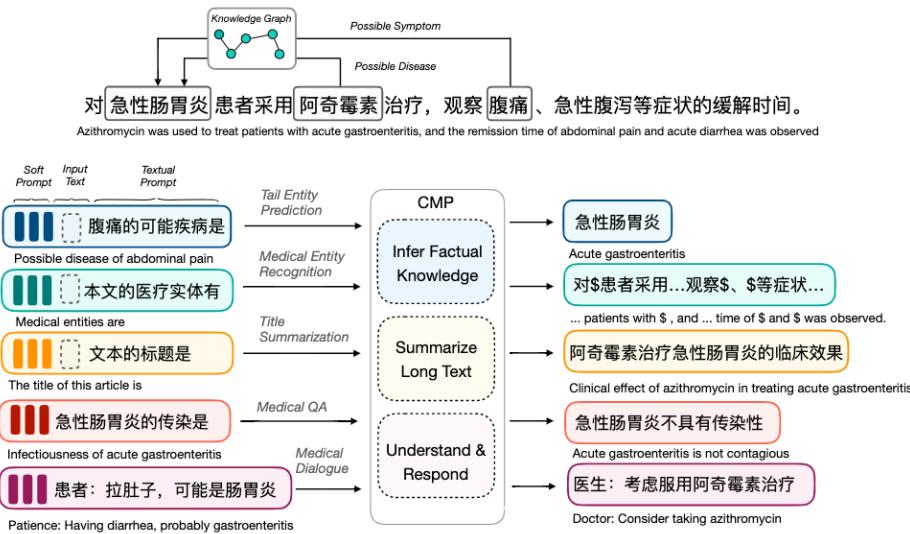




Diagnosis Assistance: Dialogue

- Our exploration:

- Multi-task & soft prompt learning during pre-training
- 2-stage framework for the medical dialogue task





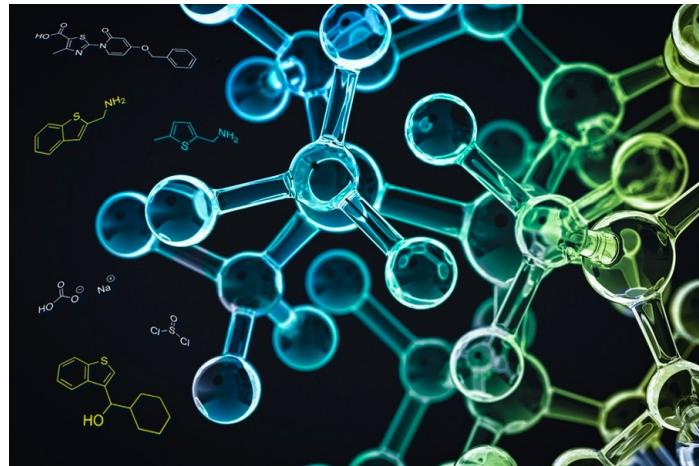
Diagnosis Assistance

- Something about Big Models:
 - Externally, we integrate KB/KGs during the encoding of medical dialogue text
 - Internally, we regard the PLM itself as a KB, hoping to query corresponding information from it
 - Prompt / Cloze ? CoT ?
 - How to protect privacy?



Substance Representation

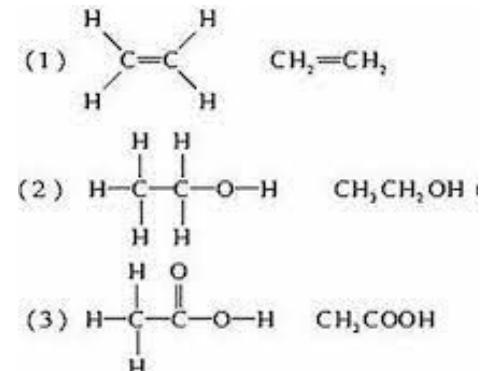
- NLP systems can process natural language text
 - What if we want to process biomedical substances?





Substance Representation

- NLP systems can process **not only** natural language text
- To represent biomedical substances as **linear text**



α1 α2 α3
1 10 20 30 40 50

G _A 77	TTYKLILNLQAKEEAIKELVDAIGIAEKYIKLIANAKTVEGVWTLKDEIILKATVTE
G _A 88	TTYKLILNLQAKEEAIKELVDAIGIAEKYIKLIANAKTVEGVWTLKDEIILTFTVTE
G _A 91	TTYKLILNLQAKEEAIKELVDAIGTAEKYIKLIANAKTVEGVWTLKDEIILTFTVTE
G _A 95	TTYKLILNLQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE
G _A 98	TTYKLILNLQAKEEAIKELVDAGTAEKYFKLIANAKTVEGVWTLKDEIKTFTVTE

α1 α2 α3
1 10 20 30 40 50

G _B 98	TTYKLILNLQAKEEAIKELVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
G _B 95	TTYKLILNLQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
G _B 91	TTYKLILNLQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
G _B 88b	TTYKLILNLQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
G _B 77	TTYKLILNGQKLEEEAIKEAVDAATAEKYFKLIANAKTVEGVWTYKDETAKTEFTVTE

β1 β2 α1 α3 β4

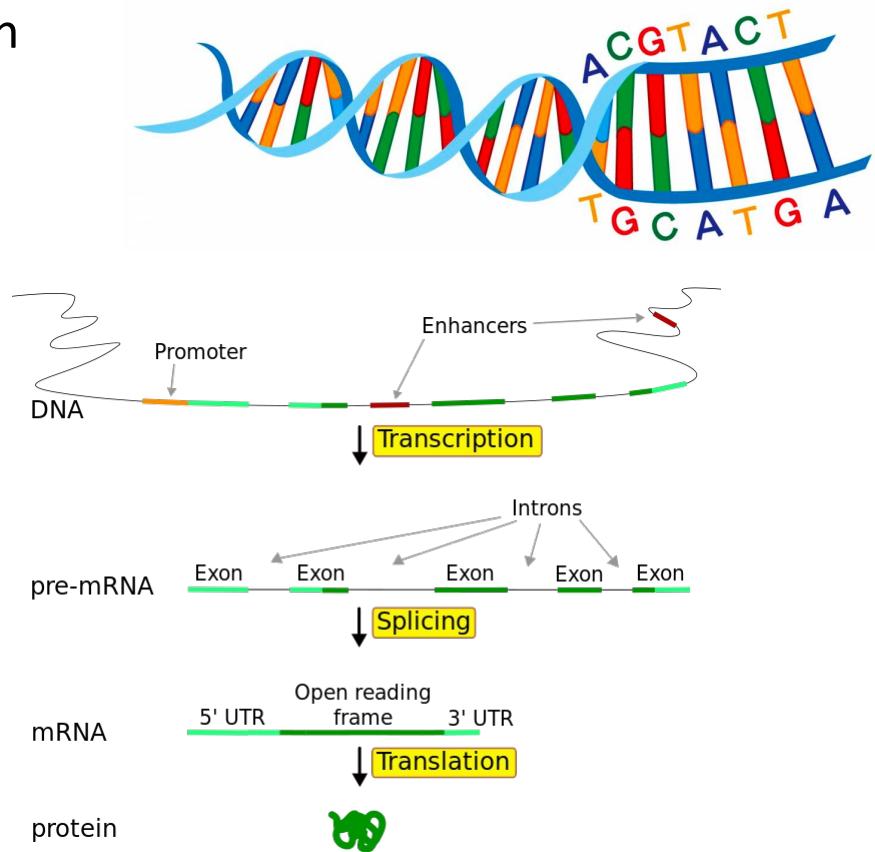


Substance Representation

- Background knowledge review
- Nucleic acid sequence: A, G, C, T (U)
- Amino acid sequence: 20 for human
- Protein: Quaternary structure

		Second position					
		U	C	A	G		
First position (5'-end)	U	UUU phe UUC UUA leu UUG	UCU ser UCC UCA UCG	UAU tyr UAC UAA Stop UAG Stop	UGU cys UGC UGA Stop UGG trp	U C A G	Third position (3'-end)
	C	CUU CUC leu CUA CUG	CCU pro CCC CCA CCG	CAU his CAC CAA gln CAG	CGU CGC arg CGA CGG	U C A G	
	A	AUU AUC ile AUA	ACU thr ACC ACA	AAU asn AAC AAA lys	AGU ser AGC AGA arg AGG	U C A G	
	G	GUU GUC val GUA GUG	GCU ala GCC GCA GCG	GAU asp GAC GAA glu GAG	GGU GGC gly GGA GGG	U C A G	

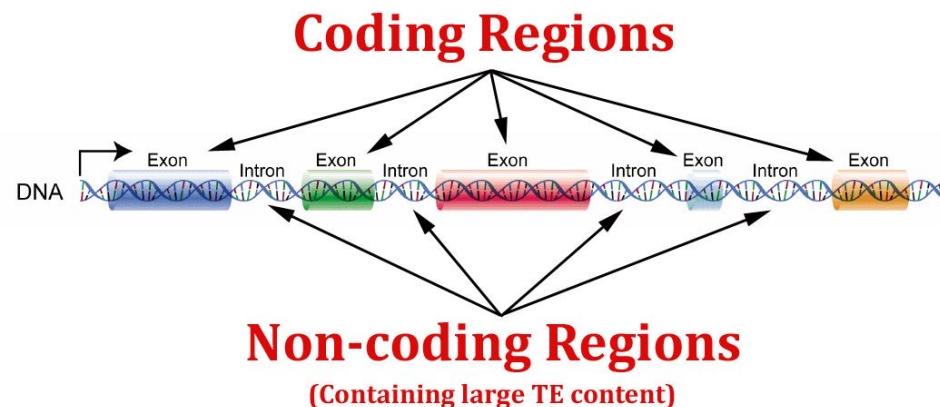
Initiation Termination





Substance Representation: DNA

- Major research object: non-coding DNA
- Tasks:
 - predict gene expression
 - predict proximal and core promoter regions
 - identify transcription factor binding sites
 - figure out important regions, contexts and sequence motifs
 - ...





Substance Representation: DNA

- Datasets: plenty of open-access resources
 - Homo sapiens genome assembly (CRCh38/hg38)
 - Cap Analysis Gene Expression (CAGE) Databases
 - Descartes: Human Chromatin Accessibility During Development
 -

The figure shows the descartes web application. At the top, a large blue header displays the word "descartes". Below it, a subtitle reads "Human Chromatin Accessibility During Development". The main search area is titled "Search Muscle Cells" and includes a search bar, a dropdown menu, and a "SEARCH" button. To the right are five circular icons: "TISSUES" (microscope), "CELLS" (cells), "MOTIFS" (DNA helix), "PAPER" (document), and "PROTOCOL" (test tube). Below this is a section titled "Motifs Associated With Skeletal Muscle Cells" featuring a sequence logo. A table lists motifs: Tgif1 (Family: Homeo, Class: Helix-Turn-Helix, Fold Change: 1.44888, Q-Value: 3.01e-196).

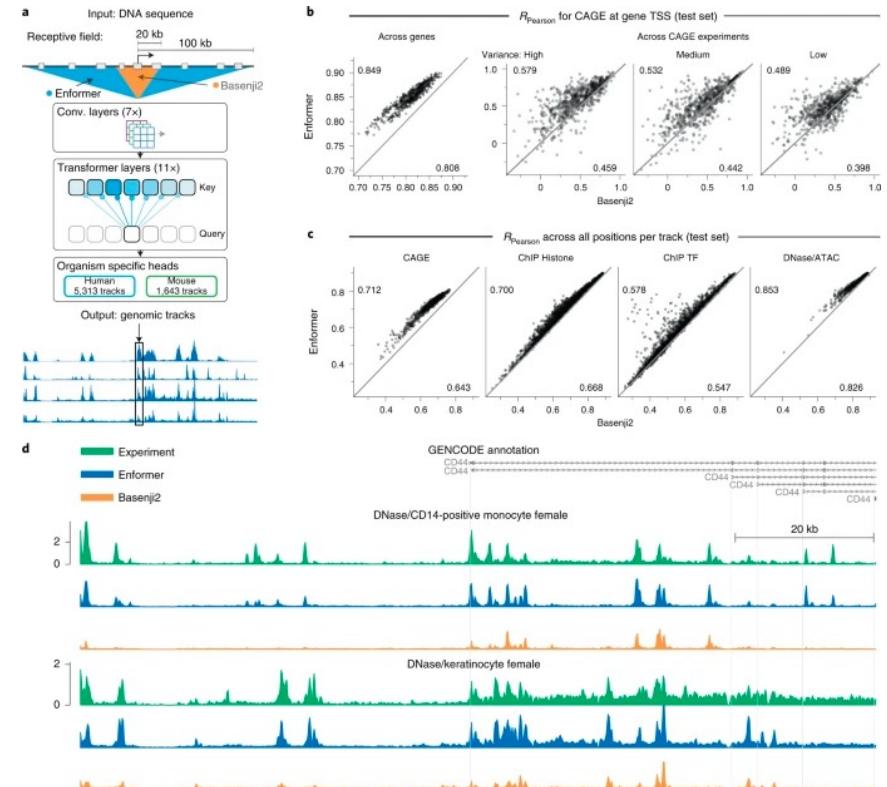
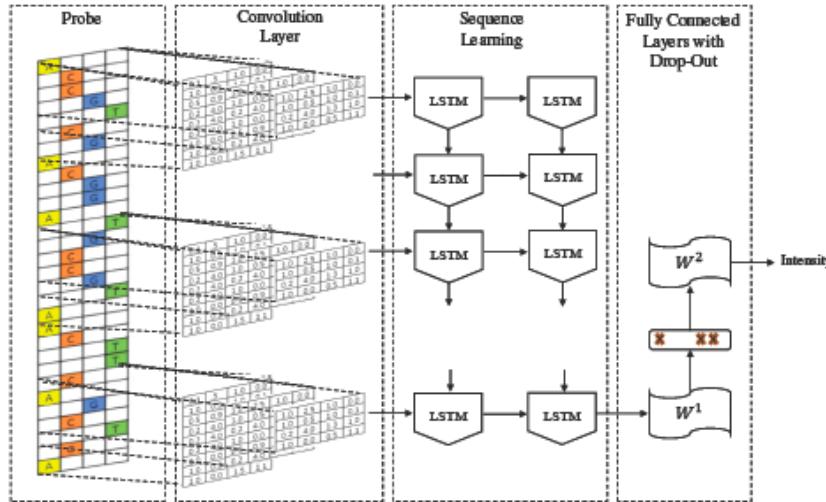
https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40

https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/cell/skeletal_muscle_cells/in/muscle



Substance Representation: DNA

- Natural language models are good at capturing patterns from mass of sequence data
- From simple frameworks (e.g. CNN&LSTM) to Transformer



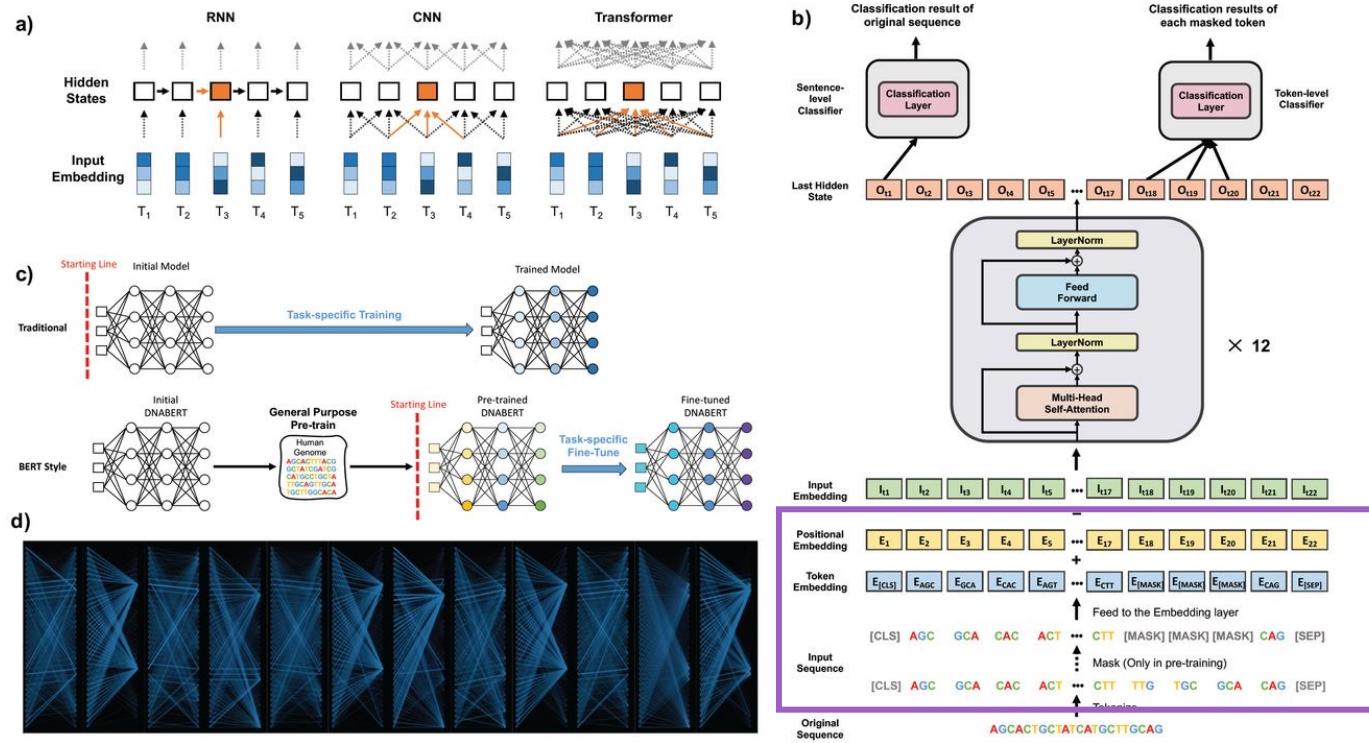
Hassanzadeh H R, Wang M D. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins[C]//2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016: 178-183.

Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." Bioinformatics 2021.



Substance Representation: DNA

- “Tokens” are fewer than natural language -> less information in word embeddings
 - position is important
 - k-mer sliding window input

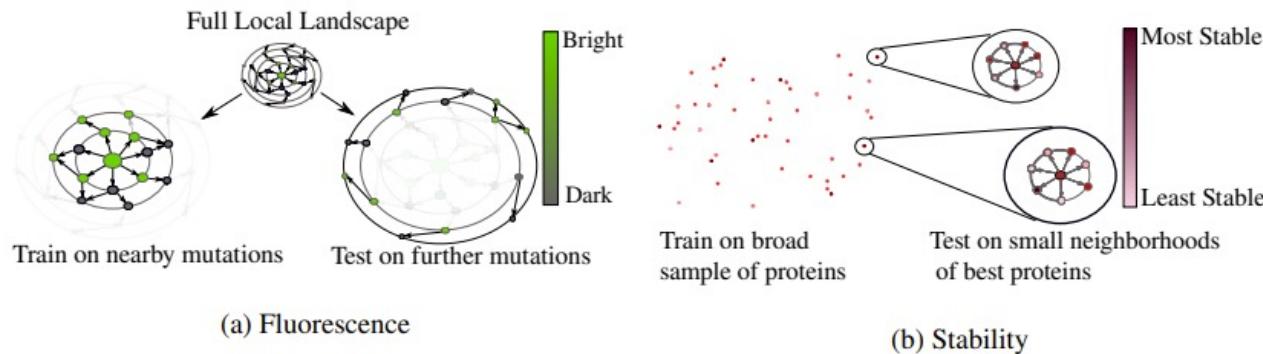


Avsec, Žiga, et al. "Effective gene expression prediction from sequence by integrating long-range interactions." Nature methods 18.10 (2021):



Substance Representation: Protein

- We mainly focus on the amino acid sequences
- Tasks:
 - Structure Prediction
 - Evolutionary Understanding
 - Protein Engineering
 - ...

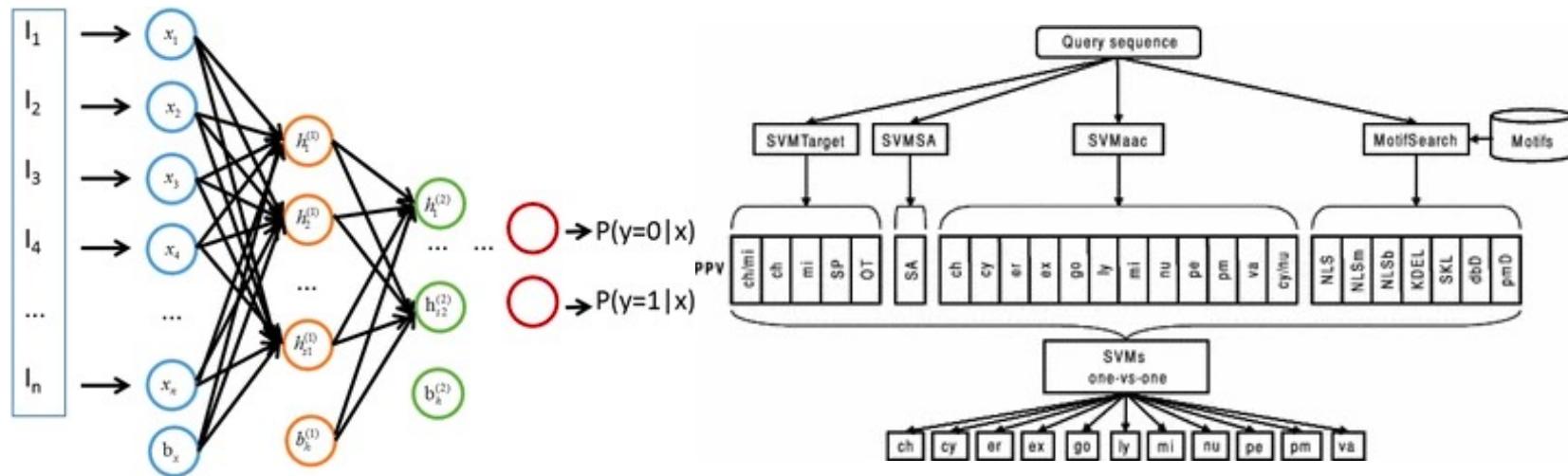


Two Protein engineering task in TAPE



Substance Representation: Protein

- Datasets:
 - Uniref: provide clustered sets of sequences from the UniProt Knowledgebase
 - GO annotations: capture statements about how a gene functions at the molecular level
 - Protein Data Bank.....
- Methods: BiLSTM + CRFs, Autoencoder models ...



Sun T, Zhou B, Lai L, et al. Sequence-based prediction of protein-protein interaction using a deep-learning algorithm[J]. BMC bioinformatics, 2017,
Sønderby S K, Sønderby C K, Nielsen H, et al. Convolutional LSTM networks for subcellular localization of proteins[C]//International conference on algorithms for computational biology. Springer, Cham, 2015: 68-80.



Substance Representation: Protein

- Big Model:

- Models with larger-scale parameters are better at capturing the features from the biologic sequences.
- Pre-training is proved to be especially helpful!

Hyperparameter	ProtTXL		ProtBert		ProtXLNet	ProtAlbert	ProtElectra	ProtT5-XL		ProtT5-XXL	
Dataset	BFD100	UniRef100	BFD100	UniRef100	UniRef100	UniRef100	UniRef100	UniRef50	BFD100	UniRef50	BFD100
Number of Layers	32	30	30	30	12	12	30	24	24	24	24
Hidden Layers Size	1024		1024		1024	4096	1024	1024		1024	
Hidden Layers Intermediate Size	4096		4096		4096	16384	4096	16384		65536	
Number of Heads	14	16	16	16	64	64	16	32		128	
Positional Encoding Limits	-		40K		-	40K	40K	-		-	
Dropout	0.15		0.0		0.1	0.0	0.0	0.1		0.1	0.0
Target Length	512		512/2048		512	512/2048	512/1024	512		512	
Memory Length	512		-		384	-	-	-		-	
Masking Probability	-		15%		-	15%	25%	15%		15%	
Local Batch Size	8	5	32/6	30/5	2	21/2	18/7	8	4	8	4
Global Batch Size	44928	22464	32768/6144	15360/2560	1024	10752/1024	9216/3584	2048	4096	2048	4096
Optimizer	Lamb		Lamb		Adam	Lamb	Lamb	AdaFactor		AdaFactor	
Learning Rate	0.0005	0.002	0.002		0.00001	0.002	0.002	0.01		0.01	
Weight Decay	0.0	0.01	0.01		0.01	0.01	0.01	0.0		0.0	
Training Steps	40.7K	31.3K	800K/200K	300K/100K	847K	150K/150K	400K/400K	991K	1.2M	343K	920K
Warm-up Steps	13.6K	5.5K	140K/20K	40K/0K	20K	40K/5K	40K/40K	10K		10K	
Mixed Precision	FP16 Model Weight Fp32 Master Weight		None		None	None	None	None		None	
Number of Parameters	562M	409M	420M		409M	224M	420M	3B		11B	
System	Summit	Summit	TPU Pod		TPU Pod	TPU Pod	TPU Pod	TPU Pod		TPU Pod	
Number of Nodes	936		128	64	64	64	64	32	128	32	128
Number of GPUs/TPUs	5616		1024	512	512	512	512	256	1024	256	1024

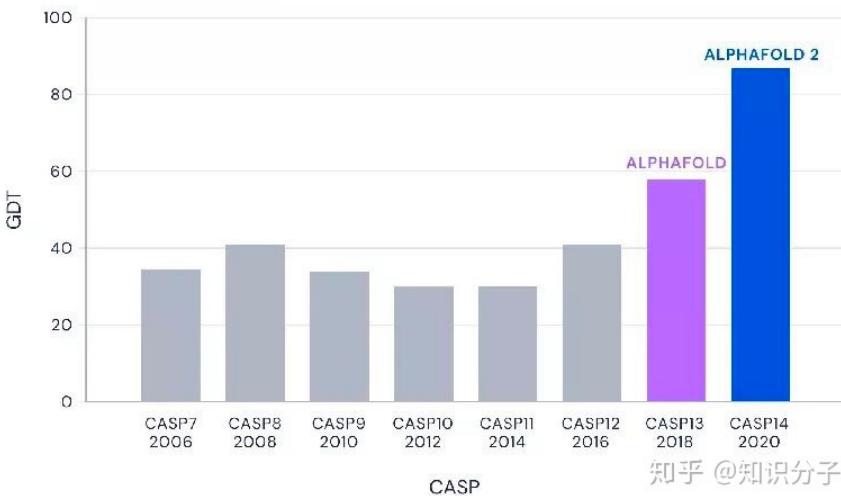
Dataset	CASP12	NEW364
DeepProtVec	62.9	64.7
ProtTXL*	71.5	72.8
ProtTXL-BFD*	71.7	72.2
DeepSeqVec	73.0	76.0
ProtXLNet*	73.7	77.3
ProtElectra*	73.9	78.1
ProtAlbert*	74.6	78.5
ProtBert*	75.0	80.1
ProtBert-BFD*	75.8	81.1
ESM-1b	76.9	82.6
ProtT5-XXL-BFD*	77.7	81.6
ProtT5-XL-BFD*	77.5	82.0
ProtT5-XXL-U50*	79.2	83.3
ProtT5-XL-U50*	81.4	84.8
NetSurfP-2.0	82.0	84.3



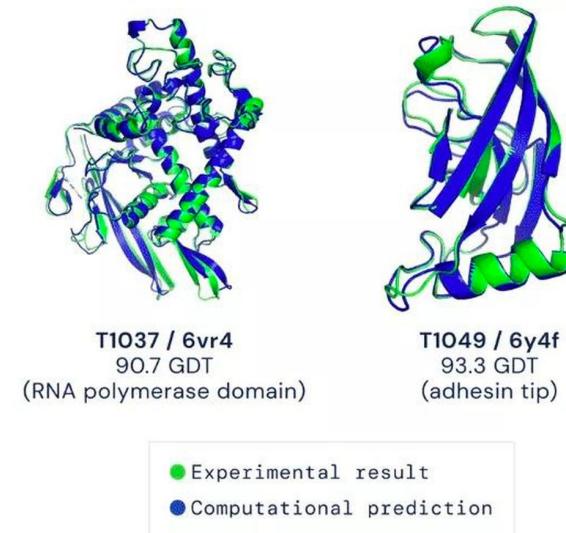
Substance Representation: Protein

- **Alpha-Fold:** One of the most inspiring research results!
- Predict 3D structure with the help of molecular dynamics
- MSA + EvoFormer +End2end training: perfect combination for biomedical knowledge and NLP technique
- A breakthrough for the 3D structure prediction accuracy (comparable to human level)

Median Free-Modelling Accuracy



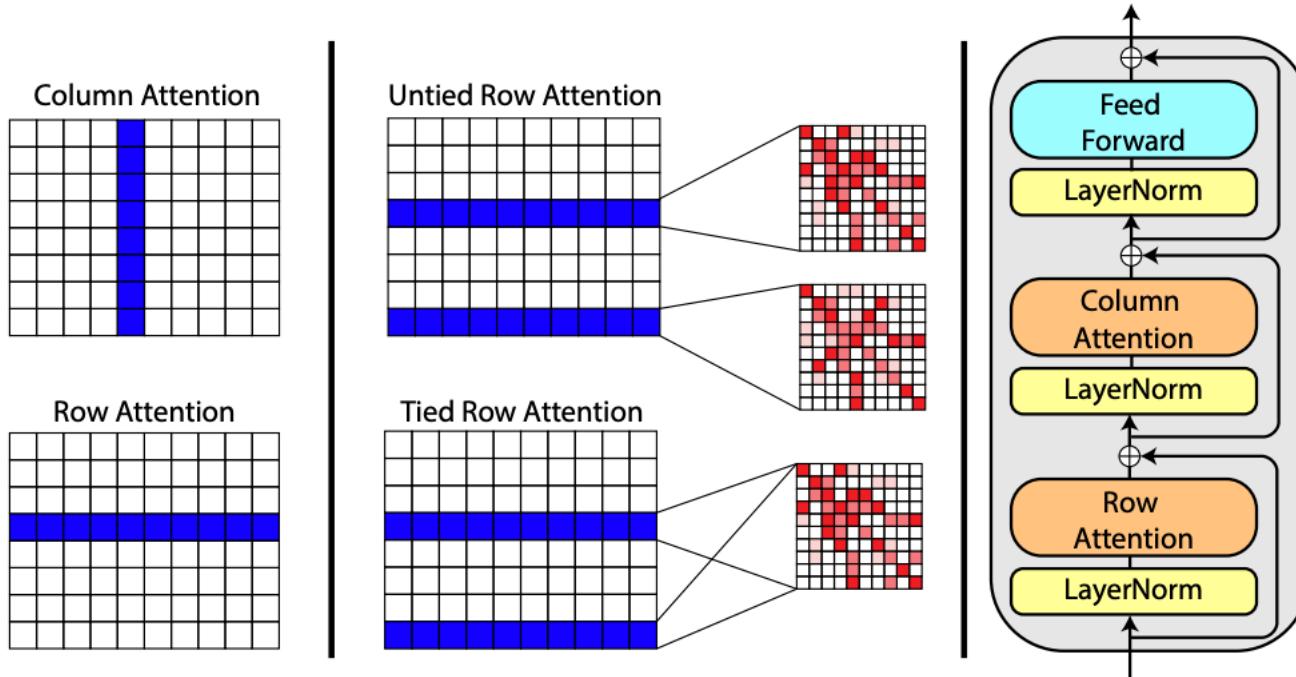
知乎 @知识分子





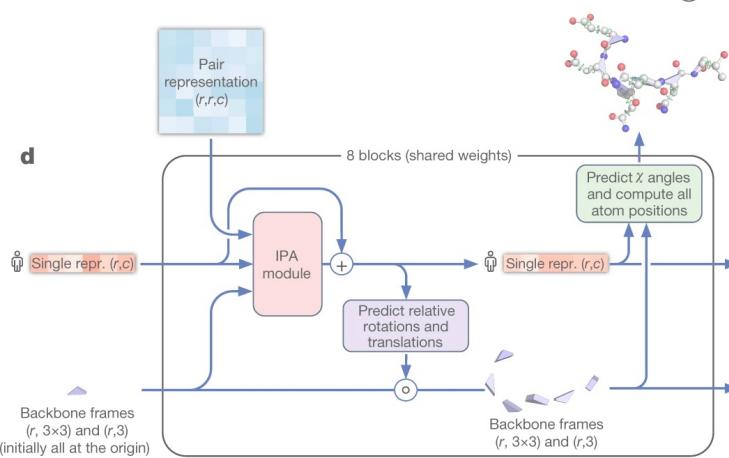
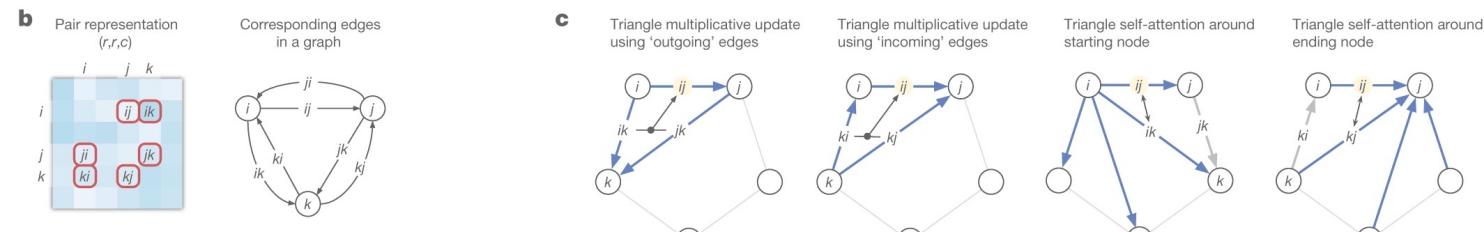
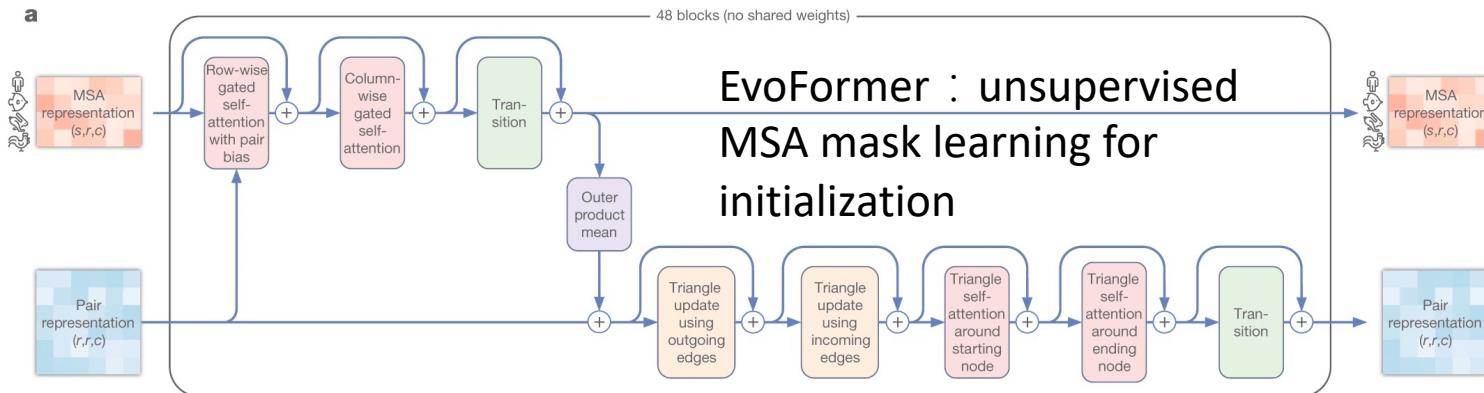
Substance Representation: Protein

- Inspired by AlphaFold: MSA Transformer
- Column / Row attention structure
- Mean attention better than individual?

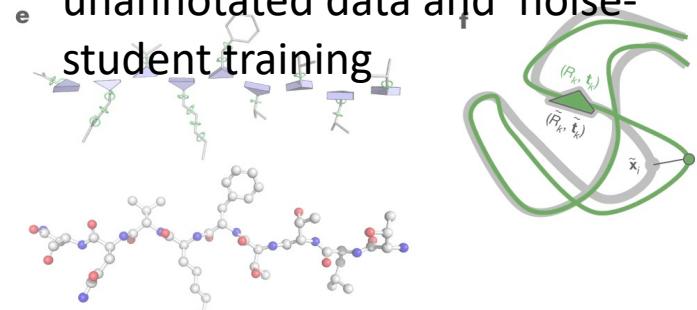




Substance Representation: Protein



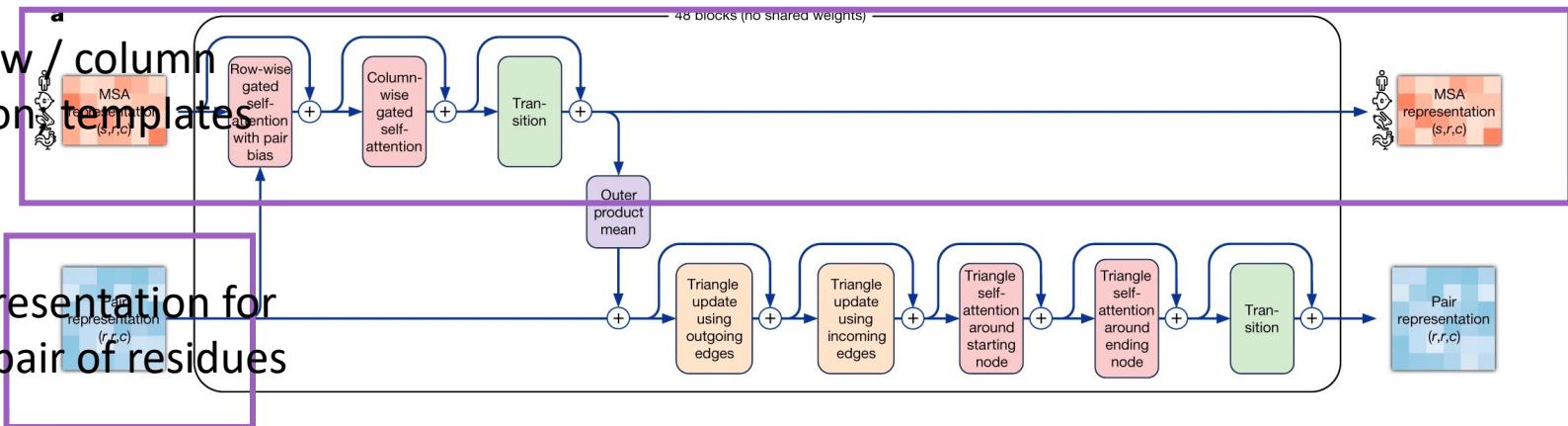
Structure : annotated data for the initial network; predict the unannotated data and noise-student training



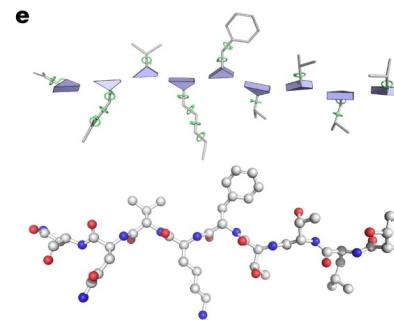
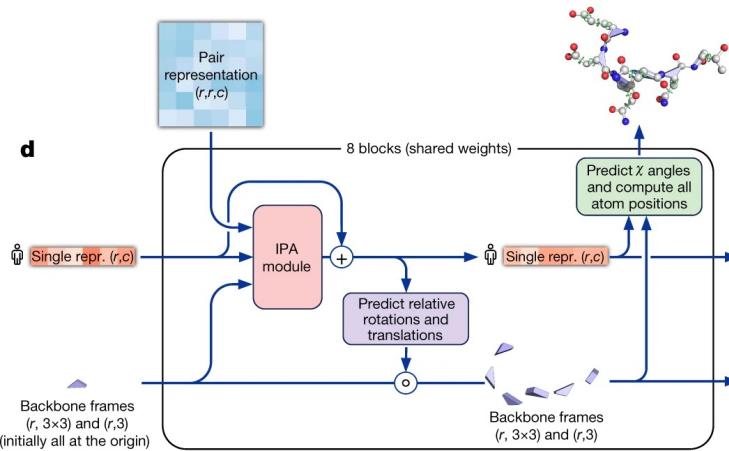
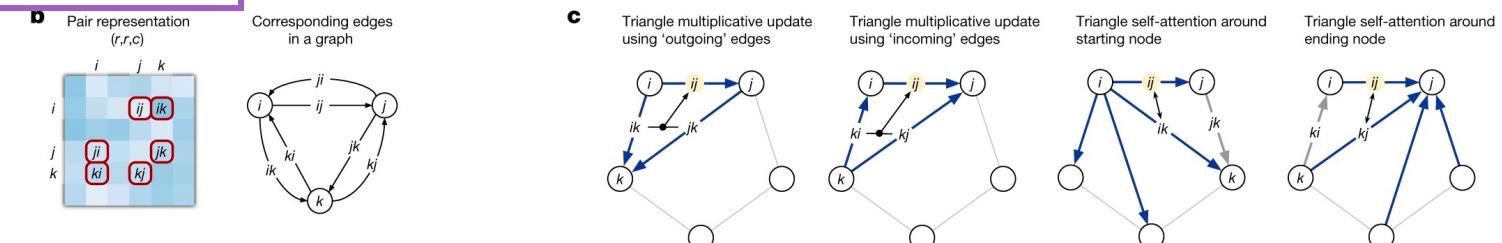


Substance Representation: Protein

MSA row / column attention templates

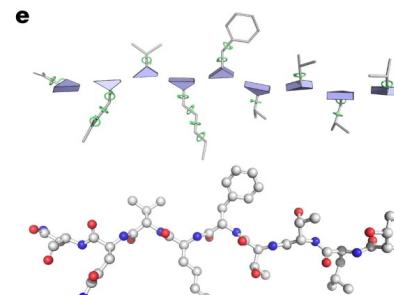
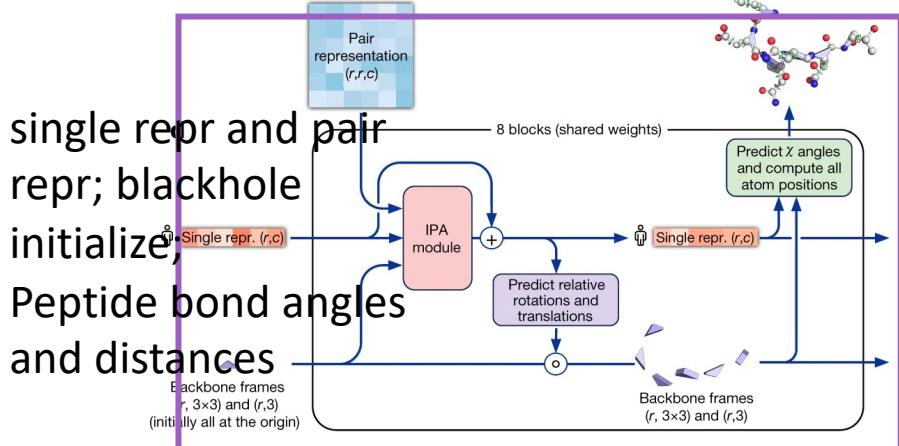
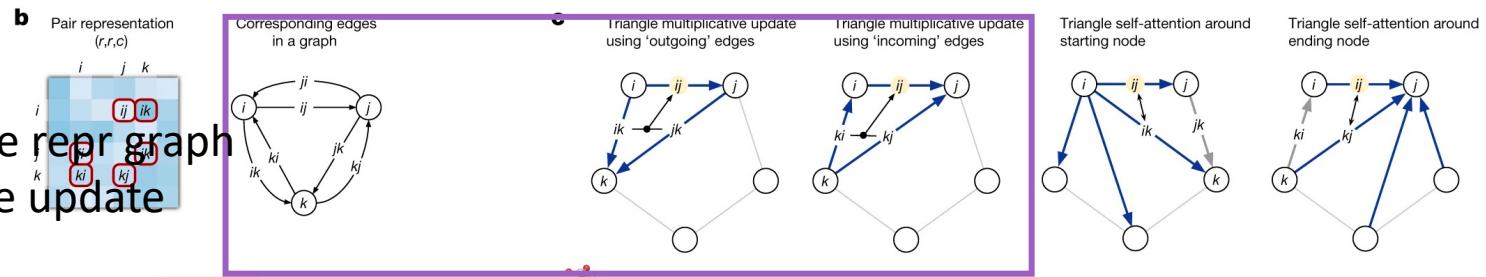
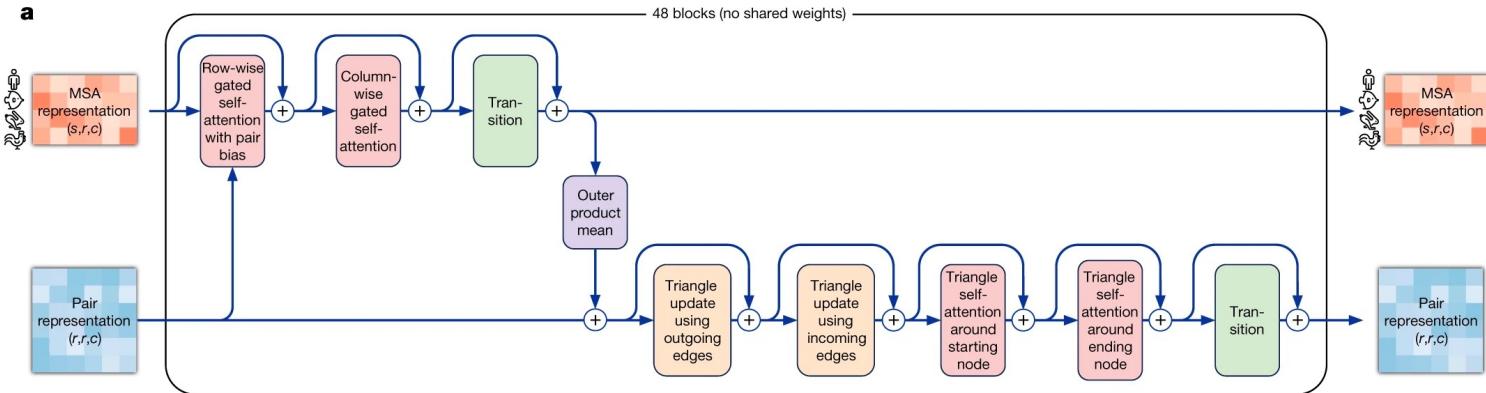


A representation for each pair of residues





Substance Representation: Protein





Substance Representation: Protein

- Interaction:

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using [AlphaFold2](#) and [AlphaFold2-multimer](#). Sequence alignments/templates are generated through [MMseqs2](#) and [HHSsearch](#). For more details, see [bottom](#) of the notebook, checkout the [ColabFold GitHub](#) and read our manuscript. Old versions: [v1.0](#), [v1.1](#), [v1.2](#), [v1.3](#)

[Mirdita M, Schütze K, Moriaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022](#)

Input protein sequence(s), then hit Runtime -> Run all

```
query_sequence: "PIAQIHLGRSDEQKETLIREVSEAIRSLDAPLTSVRVITEMAKGHFGIGGELASK"
```

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetero-oligomers). For example PI...SK:PI...:.

```
jobname: "test"
```

```
use_amber: 
```

```
template_mode: none
```

2022-07-09 03:23:39,329 Running model_1
2022-07-09 03:24:24,102 model_1 took 40.9s (3 recycles) with pLDDT 96.1 and ptmscore 0.756

colored by N→C colored by pLDDT

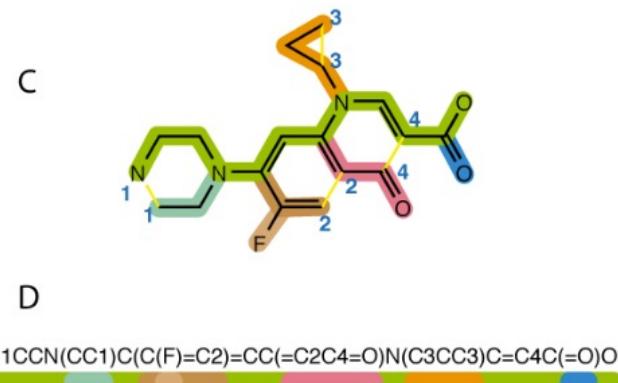
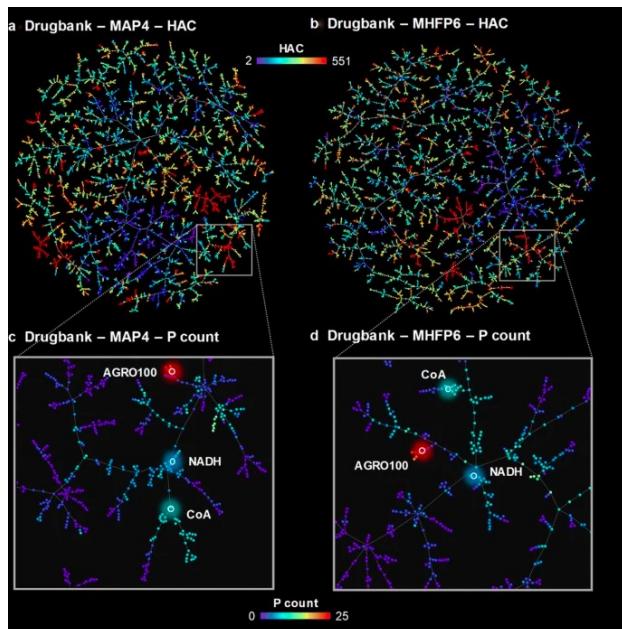


<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>



Substance Expression: Chemicals

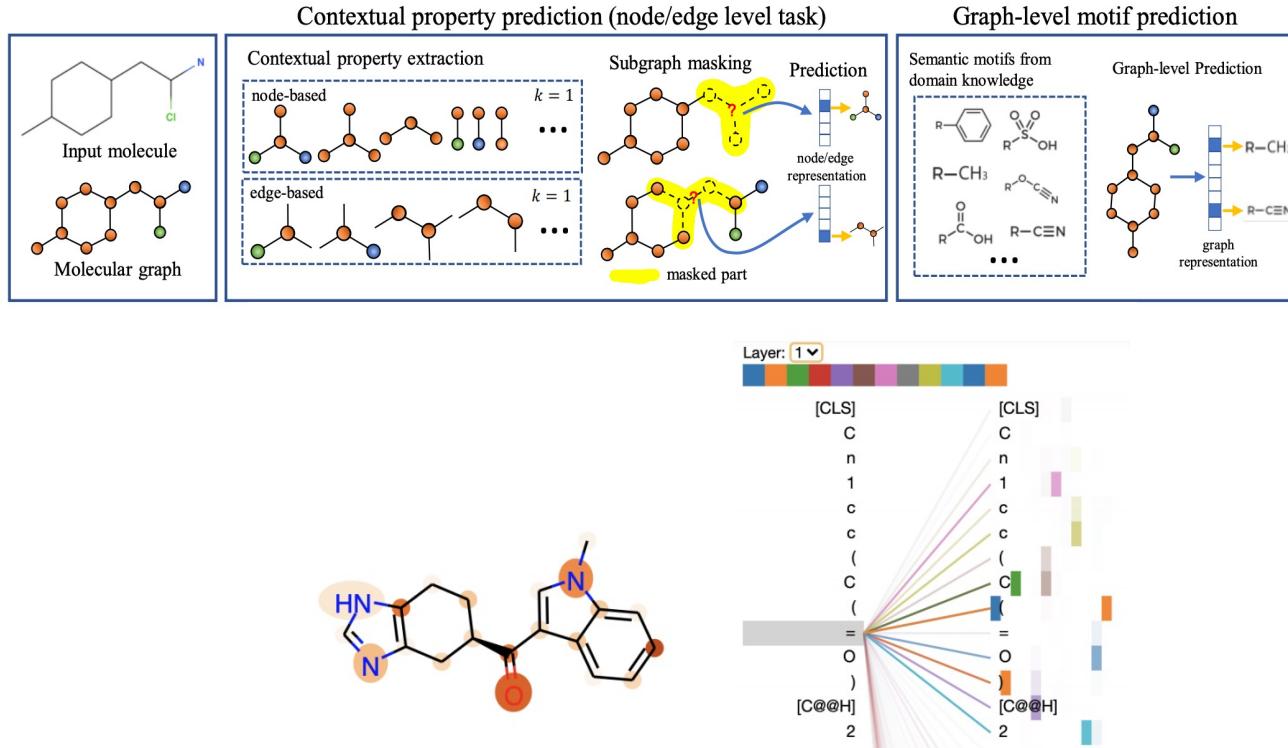
- Molecular fingerprints: essential cheminformatics tools for virtual screening and mapping chemical space
- Get fingerprint / representation by deep-learning models?
- Molecular **graphs** -> GCNs; SMILES **strings** -> LMs





Substance Expression: Chemicals

- Tasks: molecule property classification, chemical reaction classification, ...
- Datasets: MoleculeNet, USPTO 1k TPL, ...

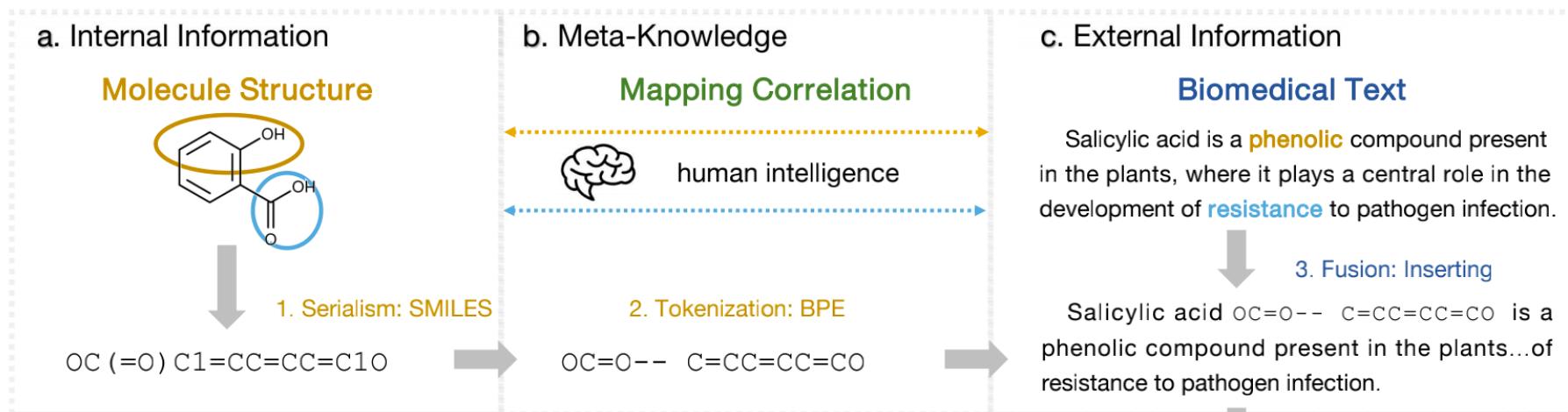


Rong, Yu, et al. "Grover: Self-supervised message passing transformer on large-scale molecular data." Advances in Neural Information Processing Systems (2020). Chithrananda S, Grand G, Ramsundar B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction[J]. arXiv preprint arXiv:2010.09885, 2020.



Substance Expression: Chemicals

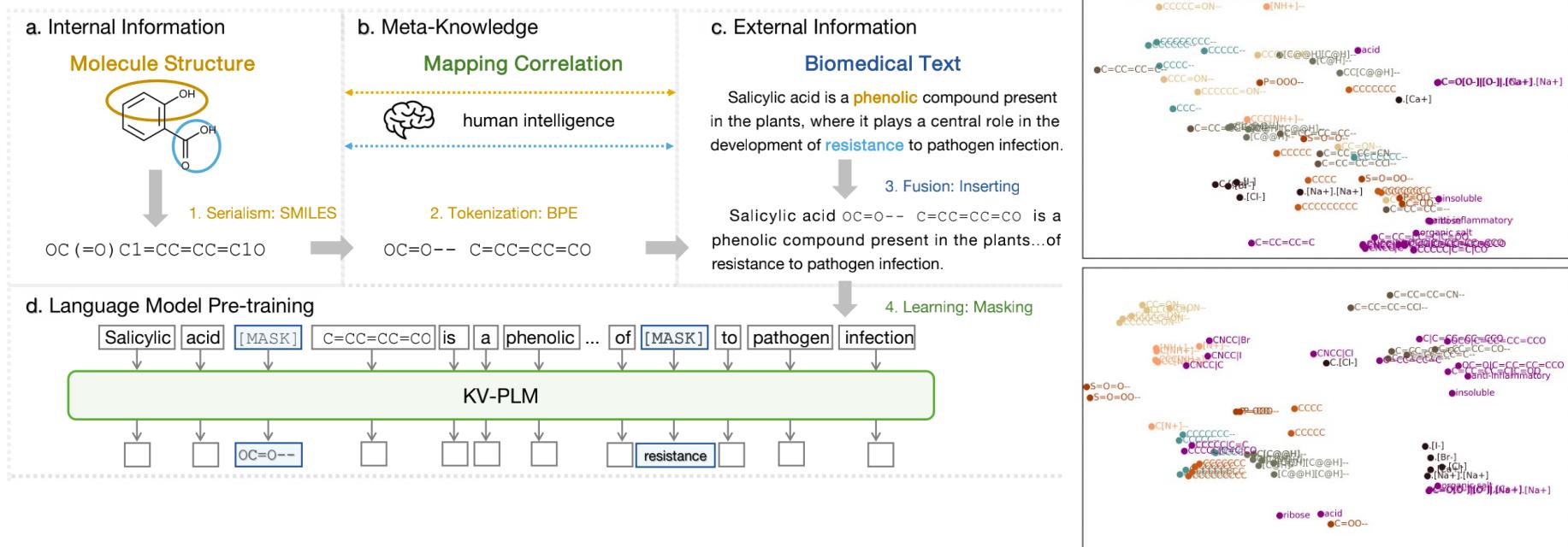
- Case: KV-PLM
- Bridging chemicals with general text
 - Complementary features of heterogeneous data
 - Inspired by human observing and learning mapping correlation





Substance Expression: Chemicals

- PLM integrating chemical structure & text
- comprehensively processing both SMILES strings and general text



Zeng, Zheni, et al. "A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals." Nature communications 13.1 (2022): 1-11.



Substance Expression: Chemicals

- Model finishing chemical exam: **property prediction**
- Conversely, it provides help for **drug discovery**

Molecular Structure	SMILES sub-groups	Property Prediction
	CO-- C=CC=CC=C-- C=C-- N-- C=O CO-- C=CC=CC=C-- C=C-- N-- C=O CO-- C=CC=CC=C-- C=C-- N-- C=O CO-- C=CC=CC=C-- C=C-- N-- C=O	colorless gas; toxic by ingestion; decomposed when heated; substituted by methyl group; an alcohol ; inorganic weakly acidic; pleasant sweet odor; found in fruits and herbs; aromatic ether ; contains methoxy groups; member of benzenes inhibits DNA replication; plant metabolite; overexpressed in cancer cells; aromatic ether; aromatic compound; member of benzenes inhibitor; yellow crystalline ; antifungal agrochemical ; aromatic compound; organic compound; aromatic ether
COC1=CC=C(C=C1)C=CNC=O	Tuberin: It has a role of inhibitor. It is a yellow crystalline . It has a role as antioxidant . It is an aromatic compound. It is an organic compound. It is an aromatic ether .	
	C=CC=CC=CC=O-- C=C-- C=CC=CC=CO C=CC=CC=CC=O-- C=C-- C=CC=CC=CO	toxic by ingestion; irritate skin; decomposed when heated; bacterial metabolite.; aromatic compound ; benzoic acid; organic compound. relatively neutral; a bitter taste.; irritate skin; a fruity taste; aromatic compound; benzenes ; organic compound.
C1=CC=C(C=C1)C(=O)C=CC2=CC=C(C=C2)O	4-hydroxychalcone: It has a role as plant metabolite . It is a protein kinase inhibitor . It has a role as prevent oxidation and free radical formation. It is an aromatic compound , a member of benzenes .	
	CC-- #N-- Cl CC-- #N-- Cl	colorless gas; toxic by ingestion; decomposed when heated; colorless liquid; substituted by methyl group ; an alcohol ; inorganic. colorless gas; toxic by ingestion ; decomposed when heated; colorless liquid; substituted by methyl group ; inorganic.
C(C#N)Cl	Chloroacetonitrile: It is a colorless gas . Very toxic by ingestion. It has a role as a persistent organic pollutant . Substituted by methyl group. It is a chloride compound. It belongs to chloride.	



Project: BioSeq PLMs and Benchmark

- Background
 - NLP technologies are widely introduced to processing biological sequences
 - There exist differences between natural language and Bio-Seq. Better PLMs are expected to be proposed.
- Long-term Goals
 1. Propose a robust and comprehensive **benchmark** for DNA data process
 2. Explore better model structure and pre-train method for DNAs
- Projects
 1. Reproduce and improve DNA pre-trained baseline methods
 2. Build down-stream DNA tasks from open-source databases



Biomedical NLP: Future Directions

- **Knowledgeable big model**: models with more **expert knowledge** achieving better performance
- **AI for science**: **user-friendly assistant tools** with lower barriers to entry; unleash human researcher productivity
- **Cross-modal processing**: bridging vision & language information or **different forms of data** (e.g. graphs)
- **Low-resource learning**: lack of annotated data



Thanks for listening

THUNLP