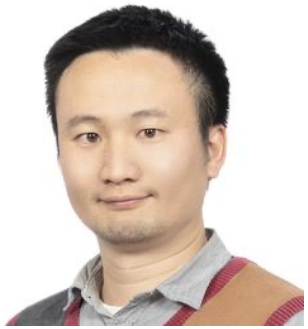
Integrating Knowledge Graphs and Large Language Models for Advancing Scientific Research

Zaiqiao Meng, Jiaoyan Chen, Xiang Zhuang, Qiang Zhang

Tutorial at 45th IEEE International Conference on Distributed Computing Systems (ICDCS)
20th July, 2025

Speakers



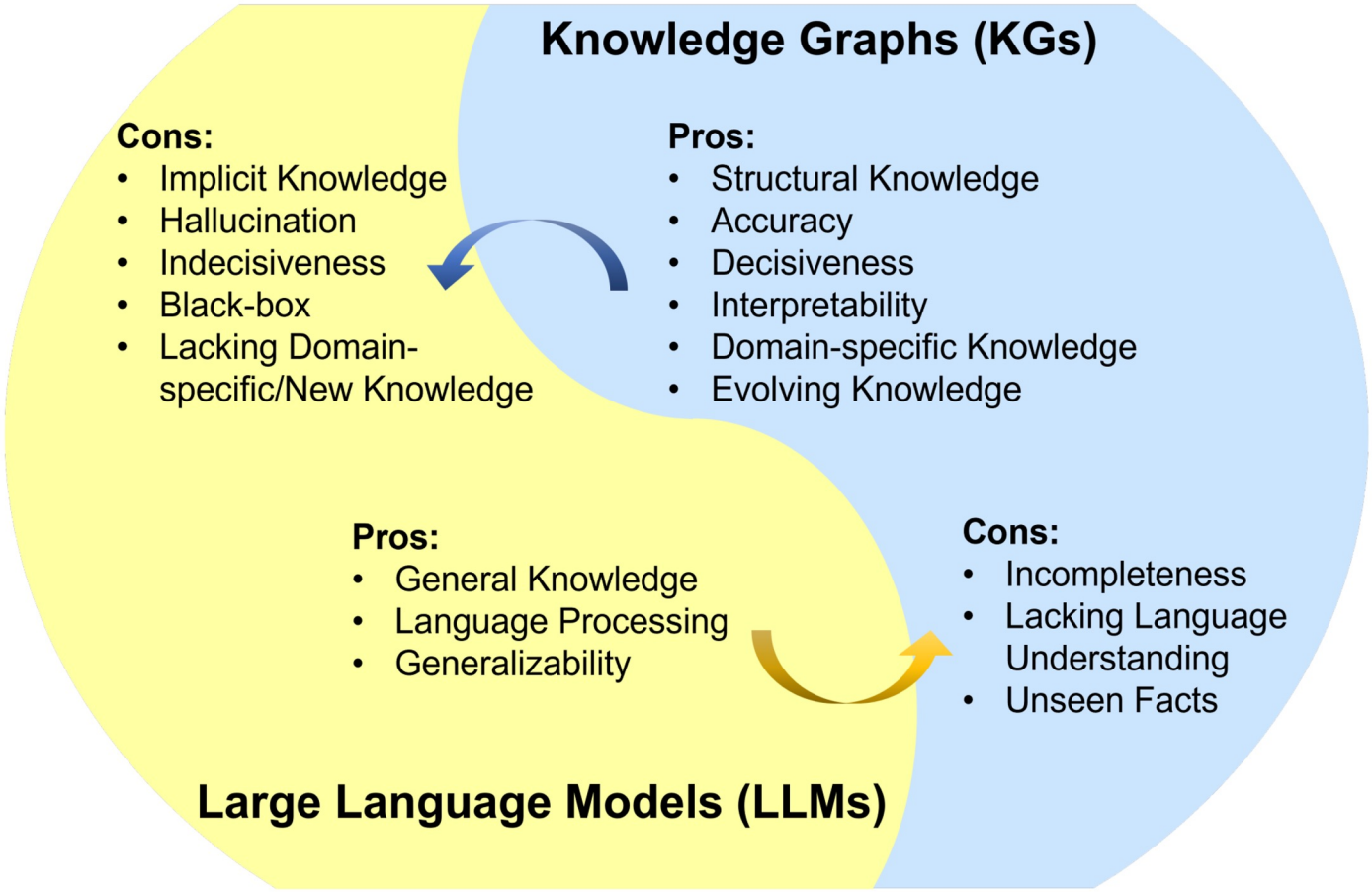
Zaiqiao Meng
University of Glasgow

Jiaoyan Chen
The University of Manchester

Xiang Zhuang
Zhejiang University

Qiang Zhang
Zhejiang University

Why combine KGs and LLMs?

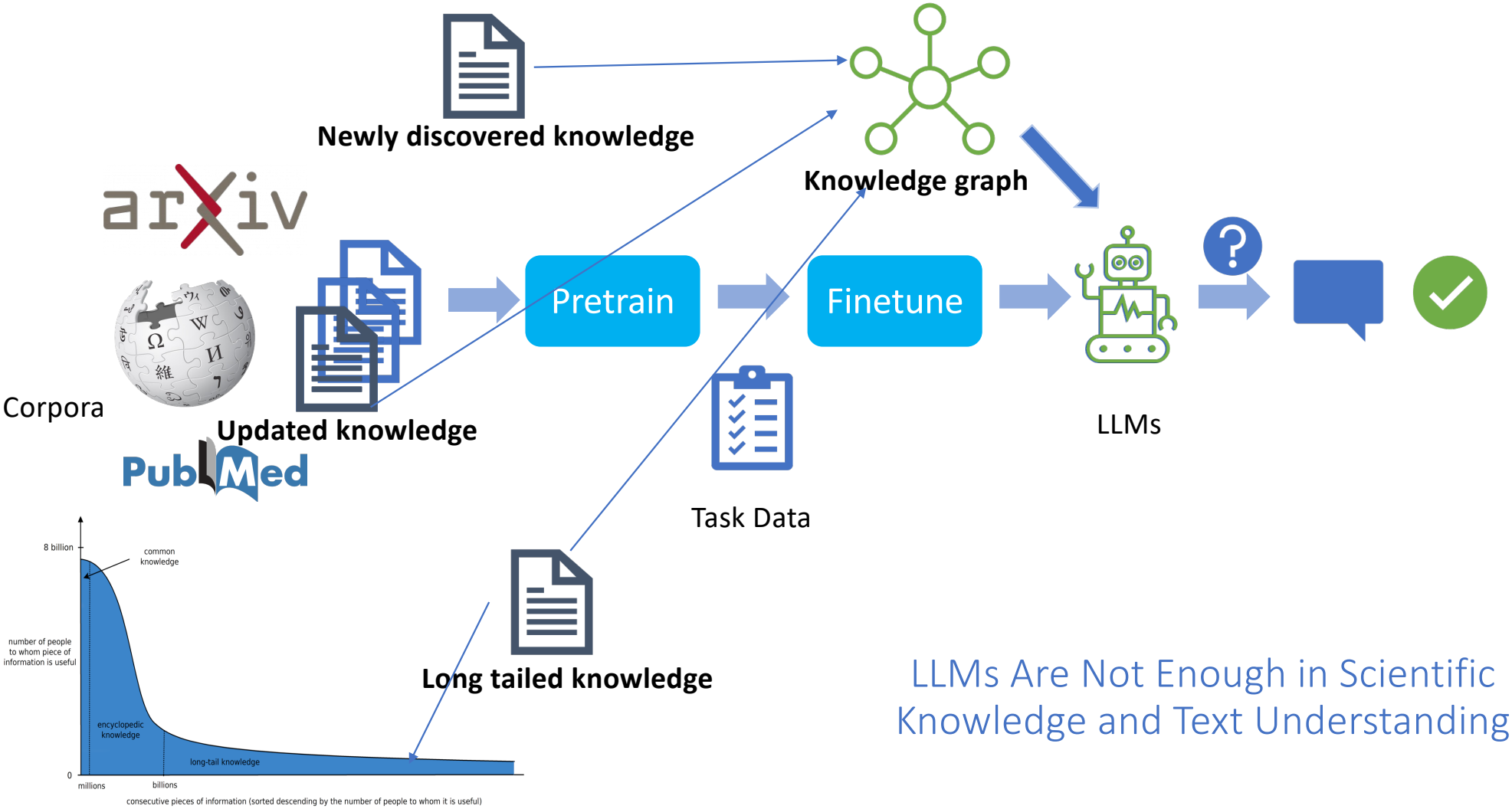


Pan, Shirui, et al. "Unifying large language models and knowledge graphs: A roadmap." *IEEE Transactions on Knowledge and Data Engineering* (2024).

More similar perspectives recently, e.g.,

Pan, Jeff, et al. "Large Language Models and Knowledge Graphs: Opportunities and Challenges." Transactions on Graph Data and Knowledge (2023).

Why combine KGs and LLMs?



LLMs Are Not Enough in Scientific Knowledge and Text Understanding

Part I: Knowledge Graphs for Science

Jiaoyan Chen

About Me

Dr. Jiaoyan Chen

https://chenjiaoyan.github.io/

jiaoyan.chen@Manchester.ac.uk

- Department of Computer Science, University of Manchester (Lecturer and then Senior Lecturer 2022 – now)
- Senior researcher at the Department of Computer Science, University of Oxford (2017 - 2025)
- Researcher at Heidelberg University (2016 – 2017)
- Ph.D. degree in Computer Science from Zhejiang University (2016)

- **Knowledge Representation**: knowledge graph, ontology, semantic techniques
- **Machine Learning**: knowledge representation learning, semantic embedding, neural-symbolic integration
- **Large Language Model**: evaluating and augmenting LLMs with knowledge graphs, ontologies, data and knowledge management systems

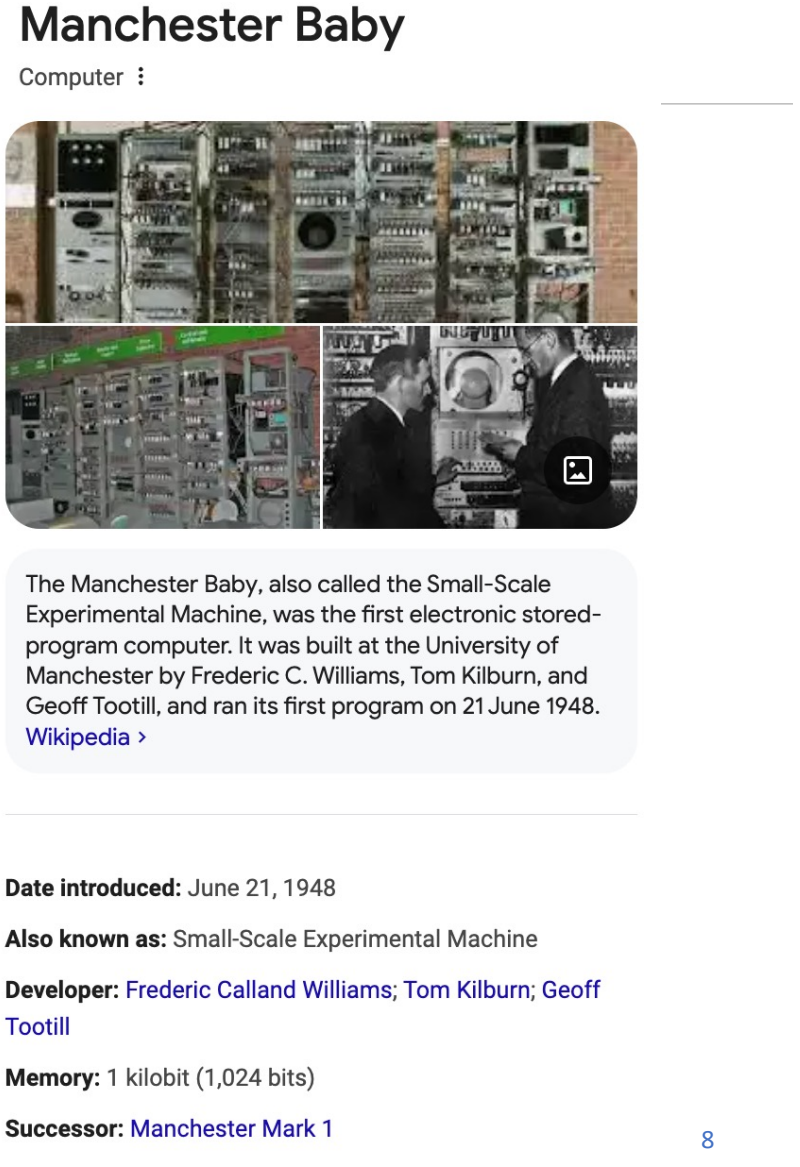
Outline

MANCHESTER 1824
The University of Manchester

1. **KG Definitions and Core Concepts**
2. **Ecotoxicological Effect Prediction: A Simple Case**
3. **KG for Life Science: Review & Challenges**

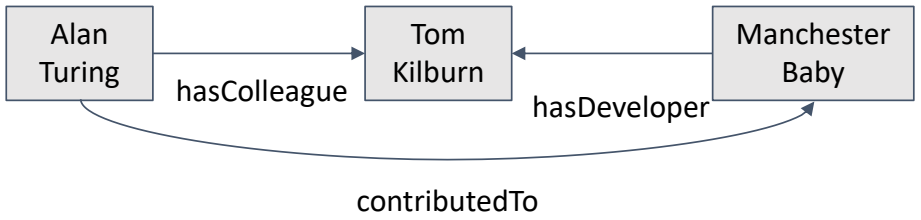
The Knowledge Graph

- The Knowledge Graph is a knowledge base used by **Google** and its services to enhance its search engine's results with knowledge gathered from a variety of sources.
  - Proposed around 2012
  - Knowledge ≈ **Instances + Facts**
  - KG ≈ **Linked Structured Data** (can be regarded as a multi-relational graph)



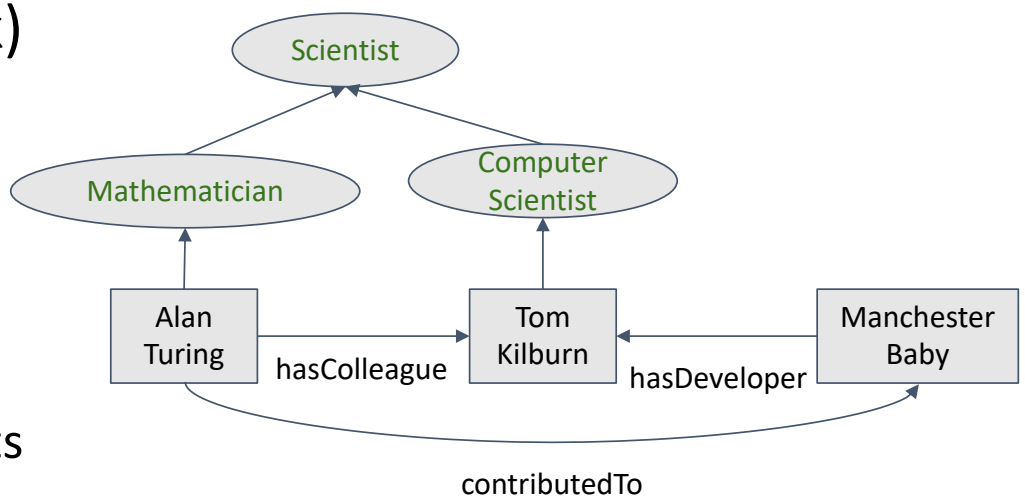
A Knowledge Representation Perspective

- **RDF** (Resource Description Framework)
  - Triple: <Subject, Predicate, Object>
  - Representing facts:
    - E.g., <Manchester Baby, hasDeveloper, Tom Kilburn>



A Knowledge Representation Perspective

- **RDF** (Resource Description Framework)
  - Triple: <Subject, Predicate, Object>
  - Representing facts:
    - E.g., <Manchester Baby, hasDeveloper, Tom Kilburn>
- **RDF Schema**
  - Meta data (schema) of instances and facts
    - E.g., class, property domain and range



A Knowledge Representation Perspective

- **Web Ontology Language** (OWL)
  - Taxonomies and vocabularies

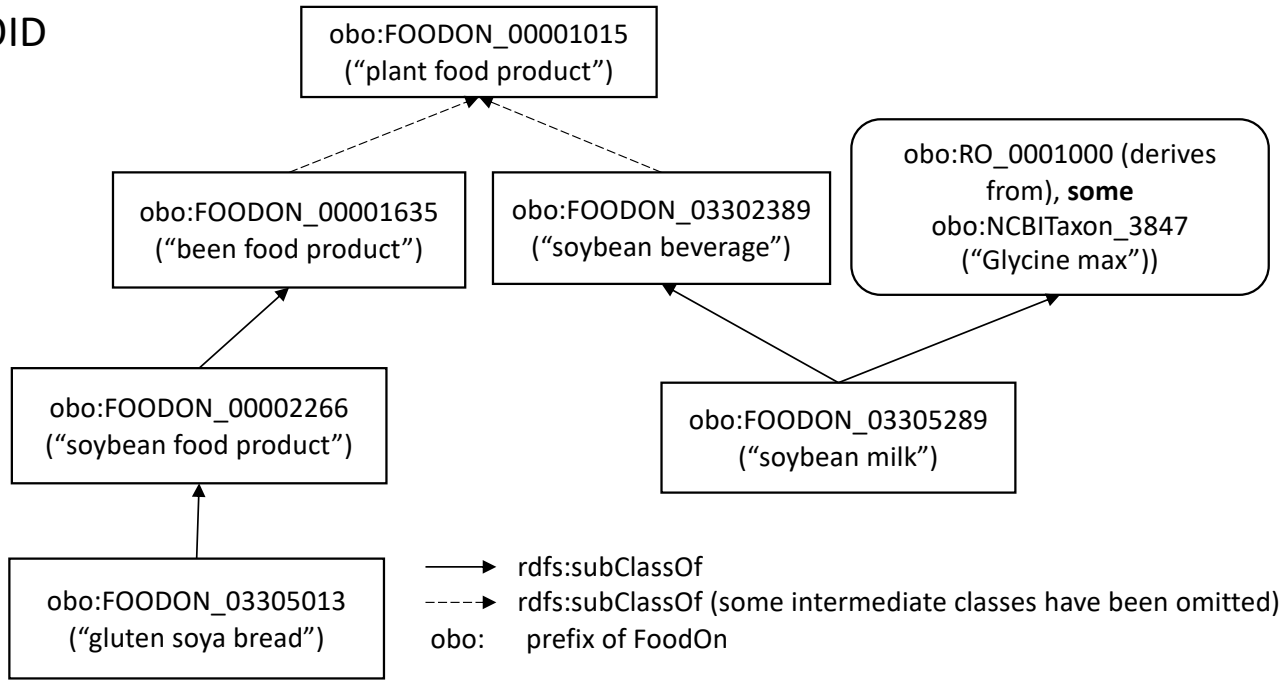
    E.g., FoodOn, SNOMED CT, GO, DOID

  - Constraints and logical relationships (> schema)

    Underpinned by **Description Logic (⊓, ⊔, ∃, ∀, ¬)**

    E.g., 'food material' ≡ 'environmental material' *and* ('has role' *some* 'food')

    E.g., the cardinality of "hasParent" is 2



A segment of the food ontology FoodOn

What is a Knowledge Graph?

RDF facts

relational graph

as Google

RDF facts + schema

(OWL) Ontology

graph + reasoning agent

logic-equipped KG

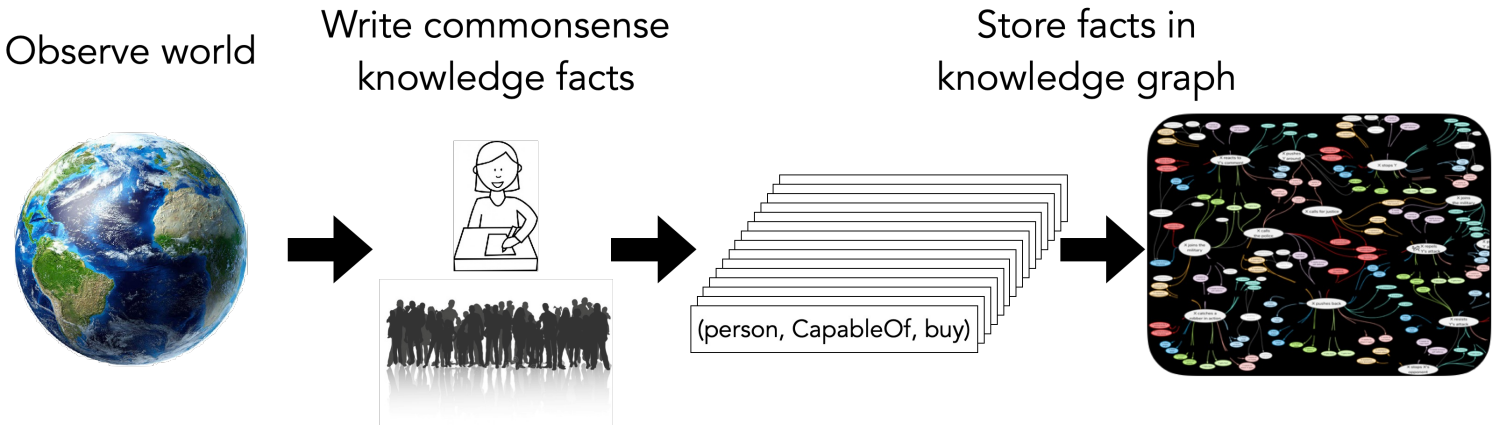
Knowledge Graph Advantages (w.r.t. Database)

- ✓ Intuitive (e.g., no “foreign keys”)
- ✓ Data + schema (ontology)
- ✓ URI not strings
- ✓ Flexible & extensible
- ✓ Rule language
  - Location + capital → location
  - Parent + brother → uncle
- ✓ Other kinds of query
  - Navigation
  - Similarity & Locality

(From Ian Horrocks)

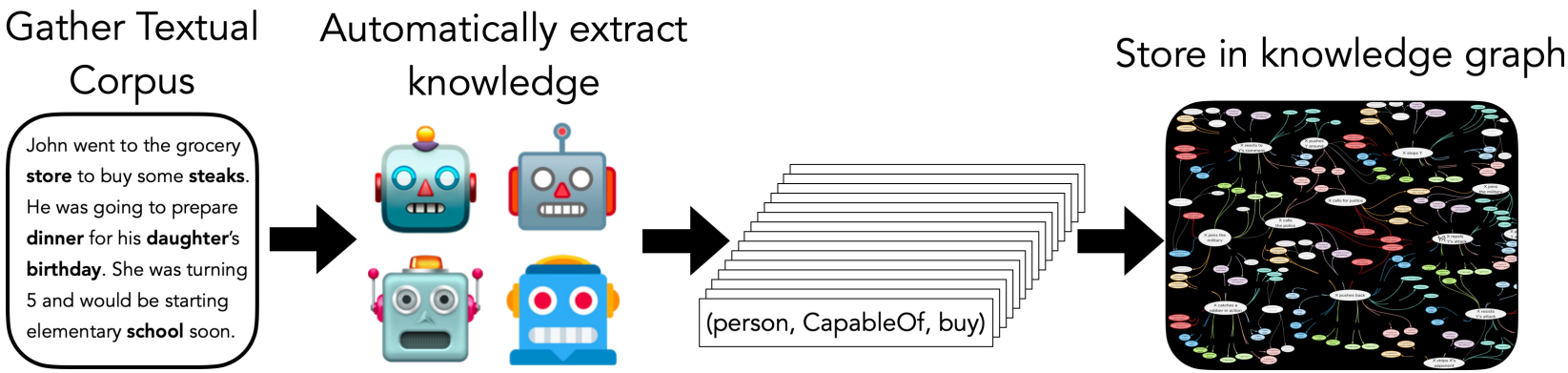
Knowledge Graph Construction

- **Crowdsourcing (Encyclopedias) & Domain Experts**
  - DBpedia, Wikidata, Zhishi.me (中文), LinkedGeoData, GeoName
  - Domain ontologies like GO, SNOMED CT, FoodON



Knowledge Graph Construction

- **The Web, Natural Language Text**
  - Open Information Extraction, Web Mining



NLP and ML techniques: NER, Categorization, Relations extraction, Entity linking, etc.

Knowledge Graph Construction

- **Semi-structured and structured data**
  - DBs, Web Tables, Excel Sheets, CSV files, etc.

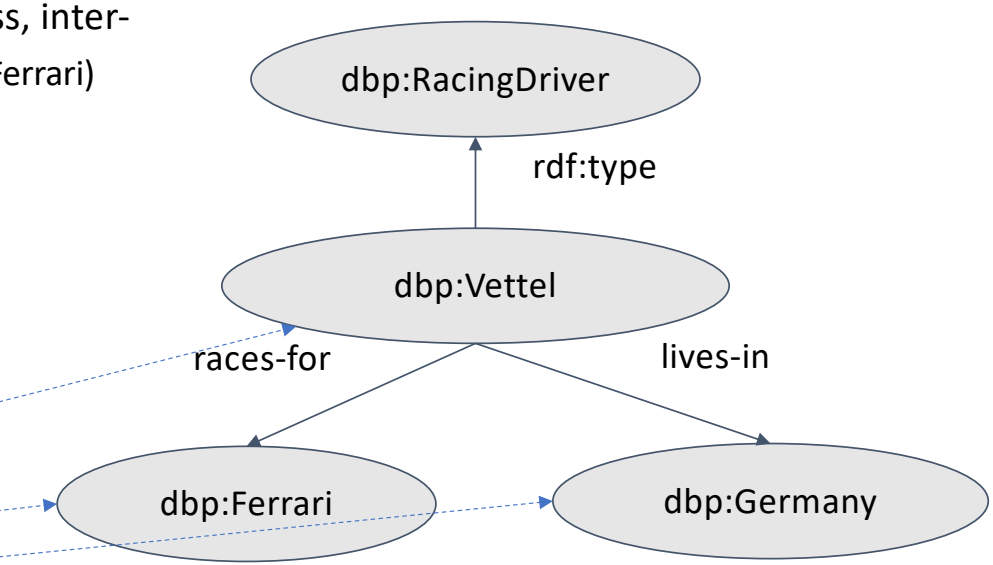
Table to KG transformation (by e.g., rules)

Table to KG matching (cell to entity, column type to class, inter-column relation to property, e.g., Sebastian Ferrari = dbp:Ferrari)

& New knowledge extraction for KG population

Hamilton races-for Mercedes ?
Hamilton lives-in England ?
Hamilton rdf:type Racing Driver ?

Alonso	McLaren	Spain
Hamilton	Mercedes	England
Sebastian Vettel	Ferrari	Germany

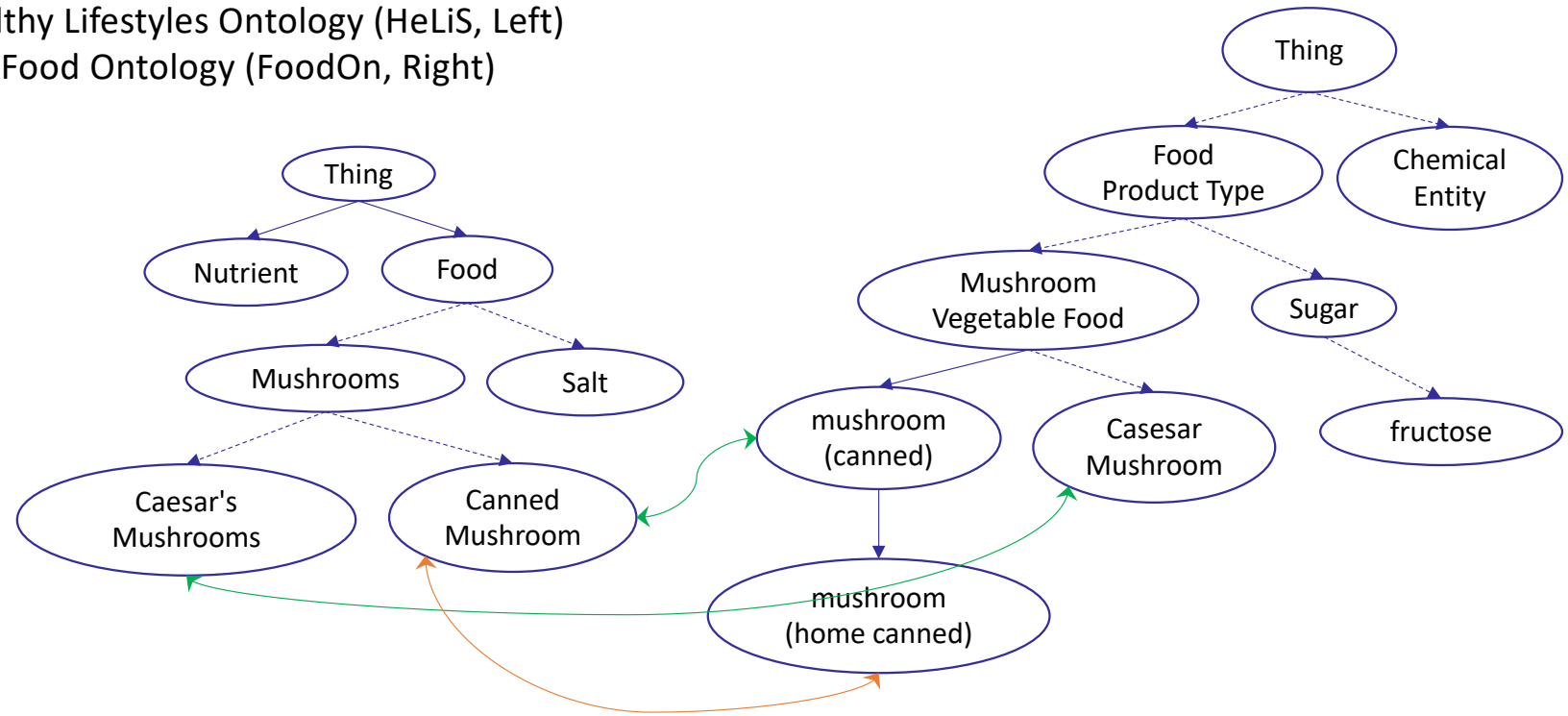


Existing KG on F1

Knowledge Graph Construction

- **Data integration (alignment, modulization, canonicalization, etc.)**

An example of alignment between
Healthy Lifestyles Ontology (HeLiS, Left)
and Food Ontology (FoodOn, Right)



Knowledge Graph Construction

- **Ontology** construction and curation highly relies on human beings now

- How to utilize Machine Learning, NLP and LLM for automation?
  - The limitation from the current ontology APIs
    - Java OWL API, Owlready 2
    - Limited Python support
  - The shortage of usable tools and resources

Knowledge Graph Construction

DeepOnto

An LM-based Ontology Engineering Library

https://github.com/KRR-Oxford/DeepOnto

- **Python interface** for more compact interaction with deep learning libraries (call Java OWL API in the backend);
- **Ontology processing APIs** for fostering deep learning and NLP techniques in ontology engineering;
- **Ontology engineering tools and resources** implemented with our APIs, deep learning and (Large) Language Models.

He, Y., et al. "DeepOnto: A Python package for ontology engineering with deep learning." Semantic Web Journal (2024).

Knowledge Graph Construction

DeepOnto

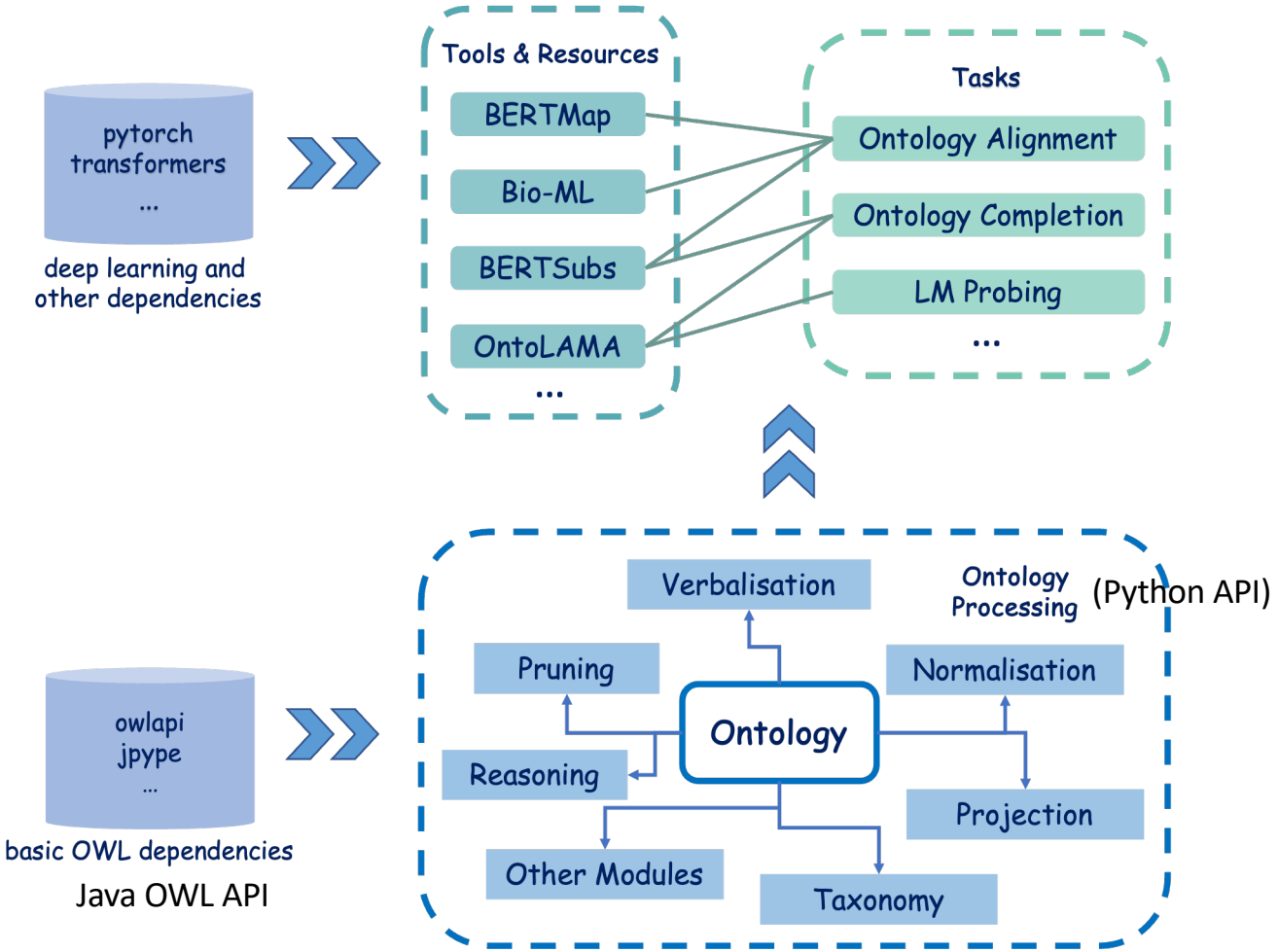
An LM-based Ontology Engineering Library

https://github.com/KRR-Oxford/DeepOnto

- **BERTMap: A BERT–Based Ontology Alignment System** by fine-tuning pre-trained language models (PLMs) by synonyms (AAAI 2022)
- **BERTSubs: ontology subsumption prediction** by prompts for encoding concept contexts and PLM fine-tuning (World Wide Web Journal 2023)
- **Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching** (ISWC 2022)
- **OntoLAMA: a Tool of Language Model Analysis** for Ontology Subsumption Inference (Findings of the ACL 2023)
- **Ontology Text Alignment**: Aligning Textual Content to Terminological Axioms (ECAI 2024)

Knowledge Graph Construction

DeepOnto



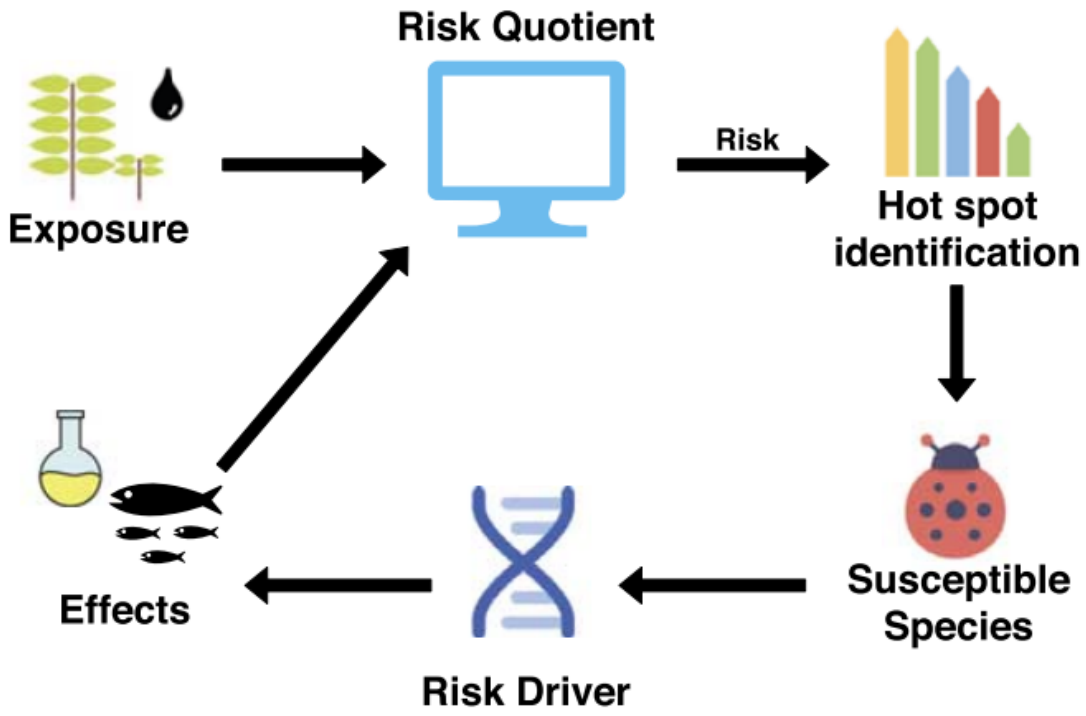
Outline

MANCHESTER 1824
The University of Manchester

1. KG Definitions and Core Concepts
2. **Ecotoxicological Effect Assessment: A Simple Case**
3. KG for Life Science: Review and Challenges

Ecotoxicological Effect Assessment

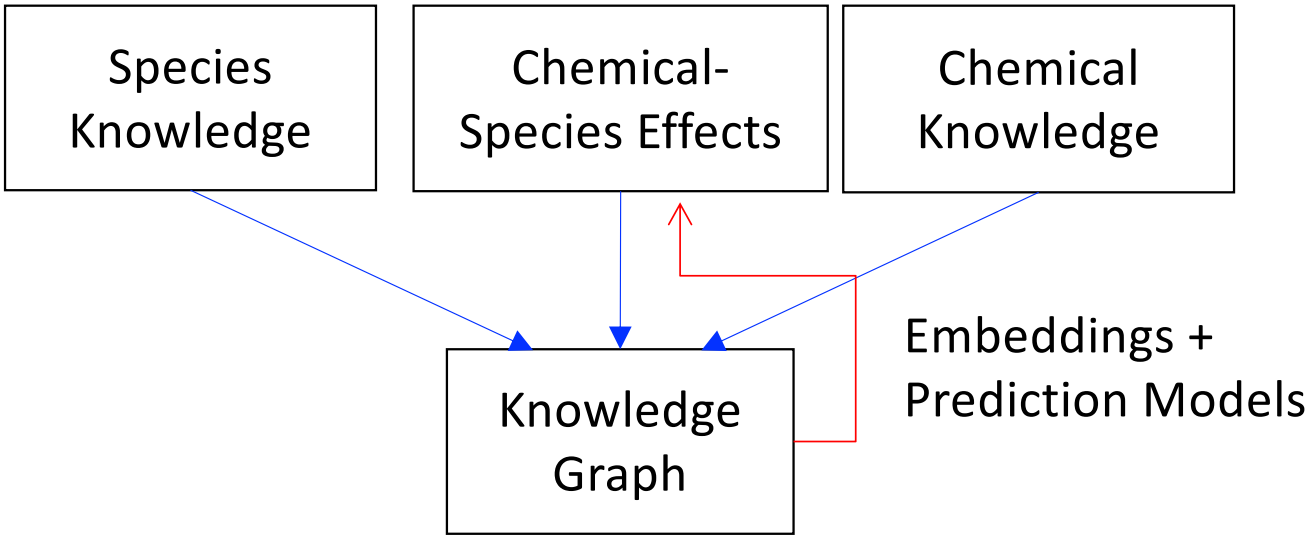
Norwegian Institute for Water Research



Simplified pipeline used in Norwegian Institute for Water Research

Chemical effect data are gathered from laboratory experiments

Ecotoxicological Effect Assessment



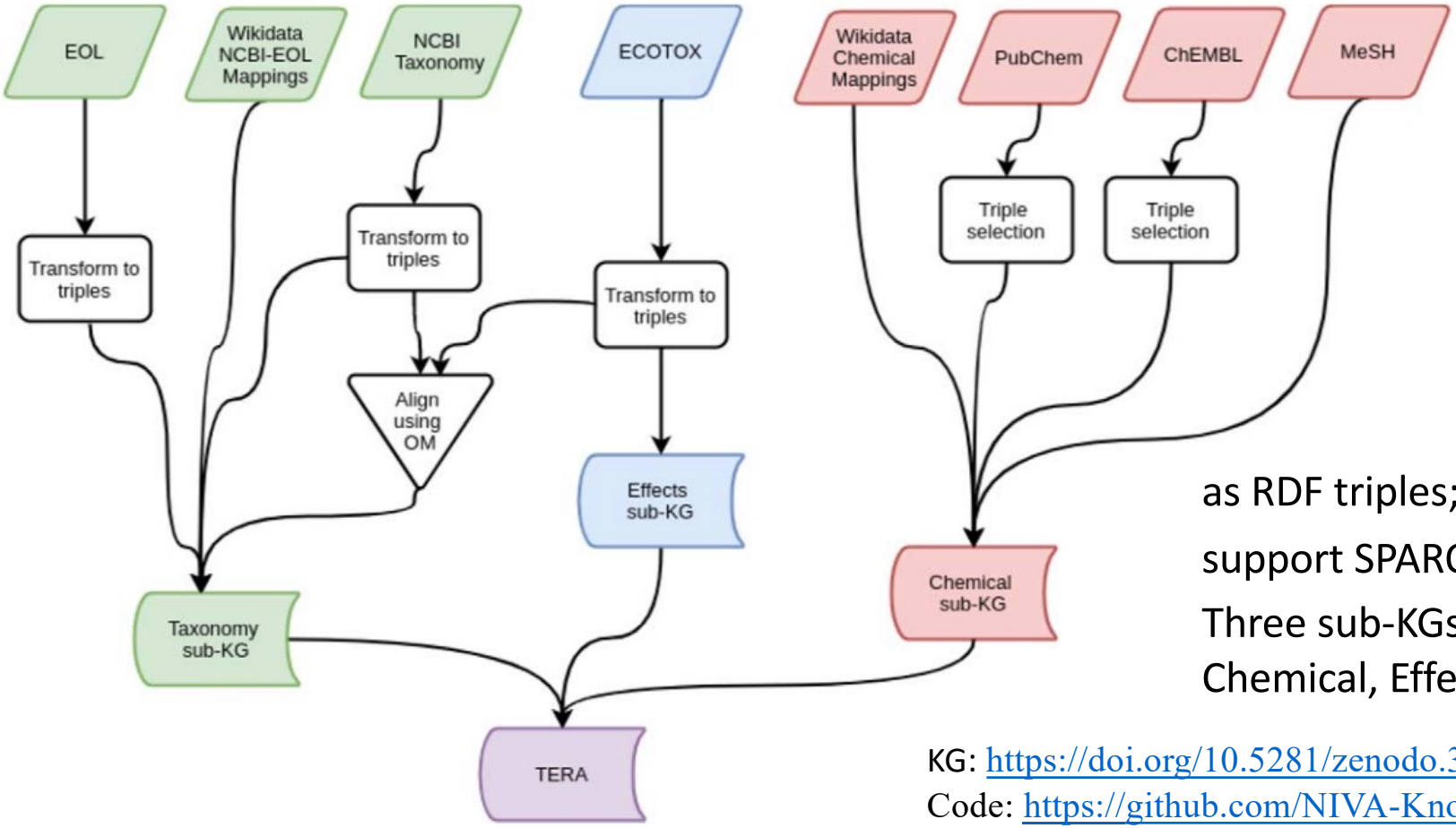
[Myklebust et al. 2022]: use **KG and its embeddings;** focus on the **mortality** relationships LC50 (50% of test population die or lose the capability of generating the next generation)

Myklebust, Erik B., et al. "Prediction of adverse biological effects of chemicals using knowledge graph embeddings." *Semantic Web* 13.3 (2022): 299-338.

How to Construct a Knowledge Graph?

- Data sources
  - Biological effects: ECOTOXicology database (~1M results, ~12K compounds, ~13K species, ~0.6% coverage of chemical-species pair coverage)
  - Biological: NCBI Taxonomy, Encyclopedia of Life (EOL; for species traits)
  - Chemical: PubChem, ChEMBL, MeSH
- Data integration
  - Wikidata mappings of species and chemicals
  - Ontology alignment tools
    - LogMap, AML: Lexical matching & index, reasoning
    - Levenshtein distance
    - Alternative: BERTMap (https://github.com/KRR-Oxford/DeepOnto): BERT fine-tuning, lexical matching & index, reasoning-based repair

TERA: Toxicological Effect and Risk Assessment KG



as RDF triples;

support SPARQL query

Three sub-KGs: Species, Chemical, Effect

KG: https://doi.org/10.5281/zenodo.3559865

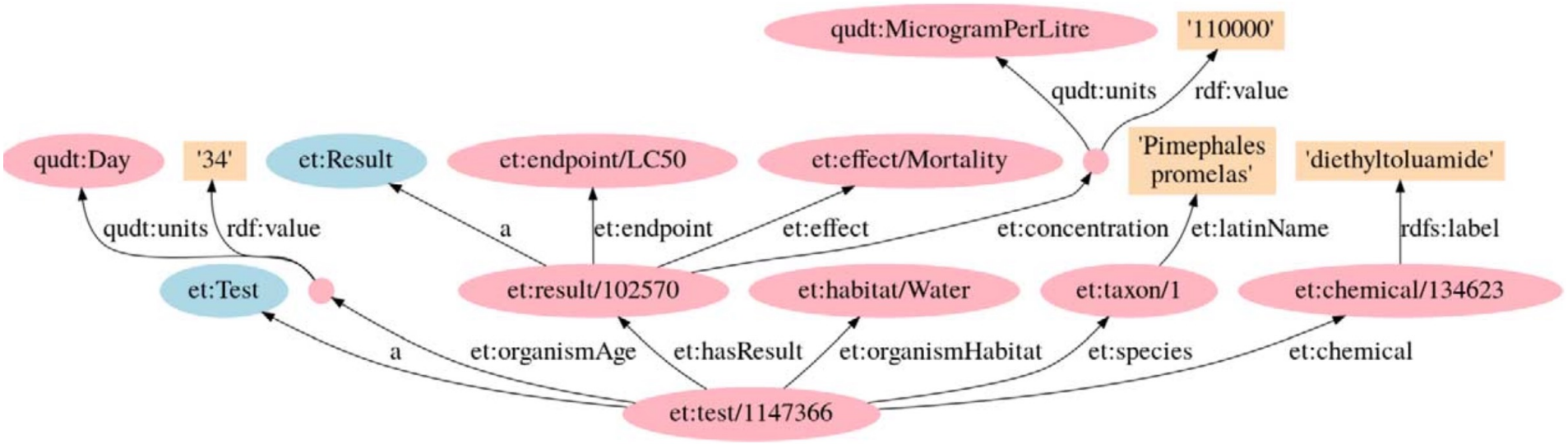
Code: https://github.com/NIVA-Knowledge-Graph/TERA

TERA: Toxicological Effect and Risk Assessment KG

Examples of RDF Triples

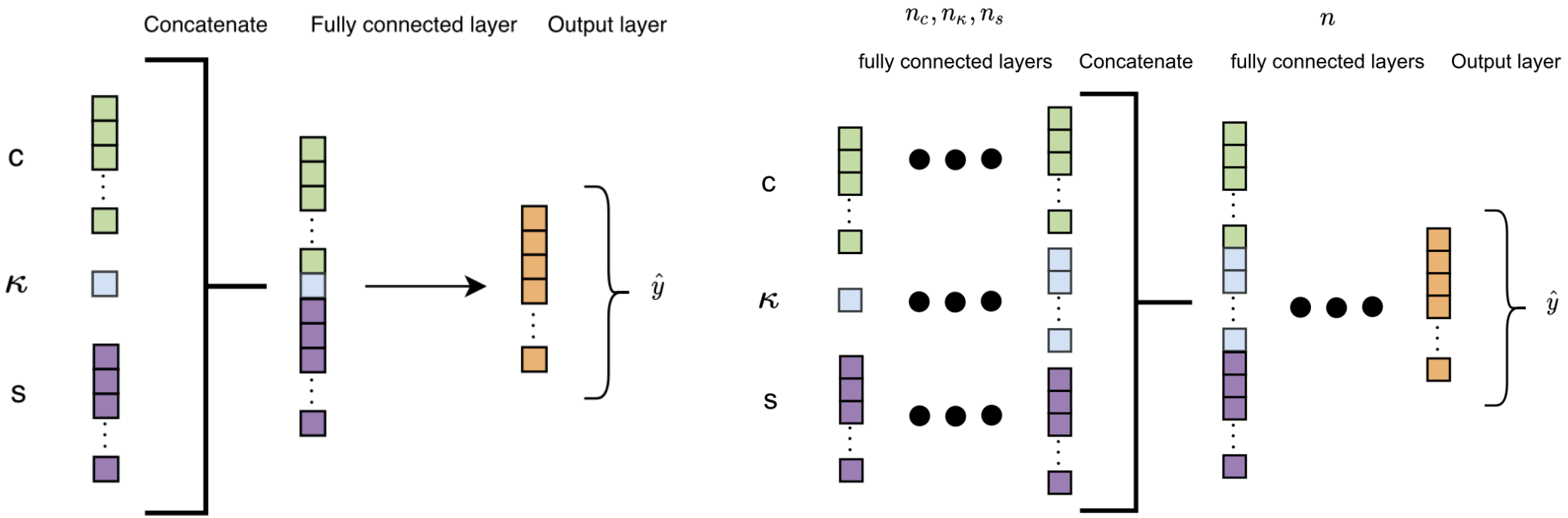
subject	predicate	object
	Effects sub-KG	
et:test/1147366	et:compound	et:chemical/134623
et:test/1147366	et:species	et:taxon/1
et:test/1147366	et:hasResult	et:result/102570
et:result/102570	et:endpoint	et:endpoint/LC50
et:result/102570	et:effect	et:effect/Mortality
	Entity Mappings	
et:taxon/1	owl:sameAs	ncbi:taxon/90988
ncbi:taxon/90988	owl:sameAs	wd:Q2700010
wd:Q2700010	owl:sameAs	eol:211492
	Taxonomy sub-KG	
ncbi:taxon/90988	rdf:type	ncbi:taxon/51137[2]
ncbi:taxon/90988	rdf:type	ncbi:division/10
ncbi:taxon/90988	ncbi:scientific_name	"Pimephales promelas"
ncbi:taxon/90988	ncbi:rank	ncbi:species
ncbi:taxon/51137	rdfs:subClassOf	ncbi:taxon/7953[3]
	Chemical sub-KG	
mesh:D003671	mesh:broaderDescriptor	mesh:D001549[5]
mesh:D003671	mesh:pharmacologicalAction	mesh:D007302[6]
chembl_m:CHEMBL1453317	chembl:hasTarget	chembl_t:CHEMBL1907594[7]
chembl_t:CHEMBL1907594	chembl:relSubsetOf	chembl_t:CHEMBL3137273[8]

TERA: Toxicological Effect and Risk Assessment KG



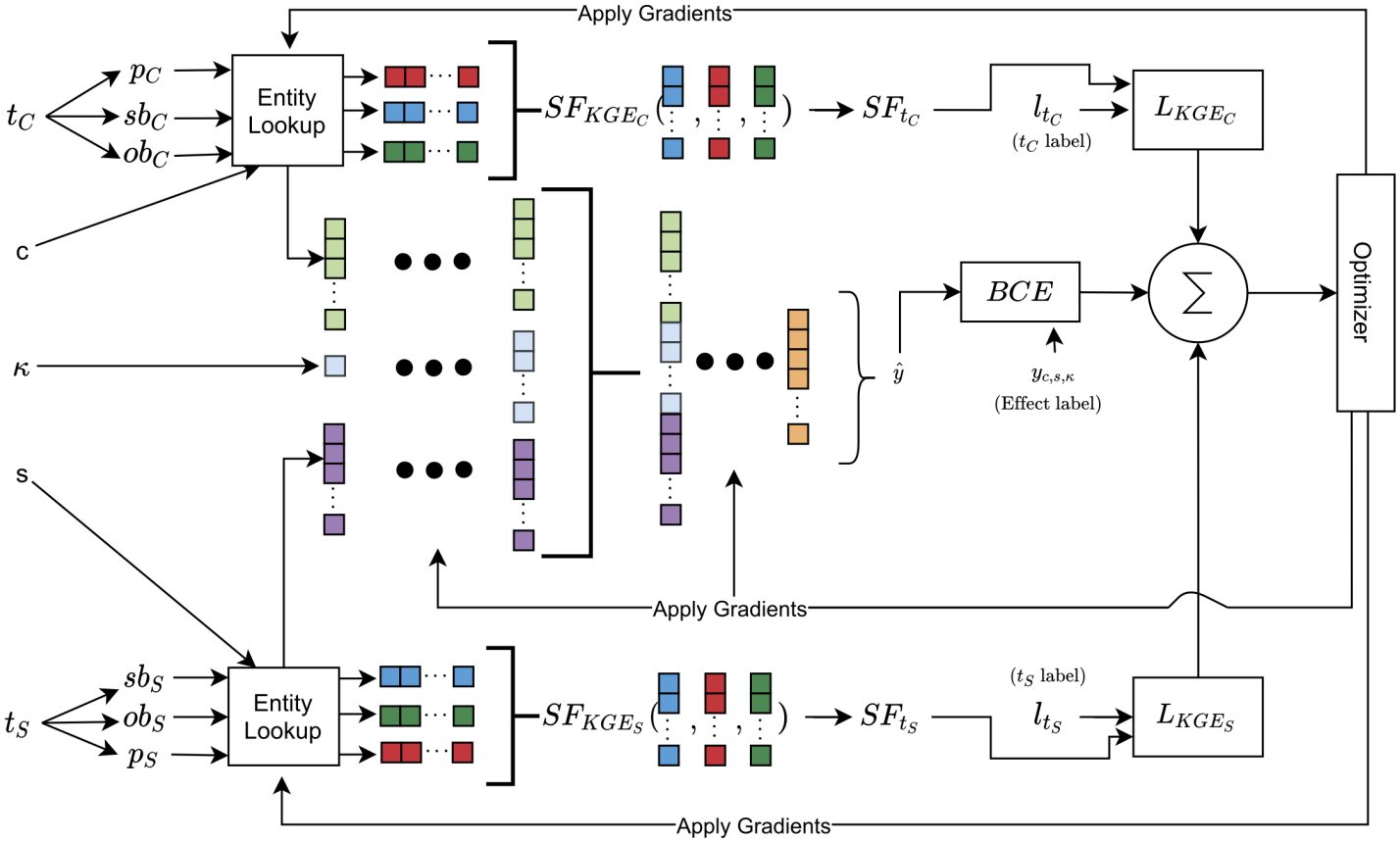
Example of an ECOTOX test and related triples

Link Prediction with TERA



Baselines: Simple (left) and complex (right) **MLP**s, with the input of **pre-trained KG embeddings** of the species and chemical

Link Prediction with TERA



Method: **Simultaneously** train the KG embeddings and the MLP

$$L = \alpha_C L_{KGE_C} + \alpha_S L_{KGE_S} + \alpha_{\text{MLP}} L_{\text{MLP}}$$

Link Prediction with TERA

- Tested TransE, HolE, DistMult, HAKE, ConvE, ConvKB, RotatE, pRotatE, and three different sampling strategies
- Result summary:
  in the majority of the settings, Sensitivity (TP/TP+FN) > 0.9, Specificity (TN/FP+TN) > 0.75)

Chemical	Species	log(κ)	Predicted	Lethal	Classification
D001556 (hexachlorocyclohexane)	59899 (walking catfish)	−3.4	0.97	1 (yes)	TP
C037925 (benthiocarb)	7965 (sea urchins)	0.9	0.2	0 (no)	TN
D026023 (permethrin)	378420 (bivalves)	0.7	0.96	1 (yes)	TP
D011189 (potassium chloride)	938113 (megacyclops viridis)	6.7	0.27	1 (yes)	FN
C427526 (carfentrazone-ethyl)	208866 (eudicots)	−0.9	0.82	0 (no)	FP
D010278 (parathion)	201691 (green sunfish)	−0.9	0.86	0 (no)	FP

Example predictions by the **simultaneously training method** with the best combination of **HolE-DistMult**

Discussion from a KG Perspective

- KG construction with more data sources?
  - Literatures & reports
  - Data of specific scientific/experimental systems (e.g., in AnIML)
  - Multi-modal data

- Link prediction
  - Accuracy & explanation
  - Multi-modal semantic embedding
  - + symbolic reasoning

Outline

MANCHESTER 1824
The University of Manchester

1. KG Definitions and Core Concepts
2. Ecotoxicological Effect Assessment: A Simple Case
3. **KG for Life Science: Review and Challenges**

KG for Life Science

Knowledge Graph Construction and Management (Sect. 3)

- Alignment for Knowledge Validation
- Knowledge Integration
- Repositories of Ontologies and Mappings
- Ontology Extension
- Instance Matching

Life Science Knowledge Discovery (Sect. 4)

- Therapeutics and Drug Discovery
- Protein Function Prediction
- Predictions for Healthcare

Knowledge Graph for Explainable AI (Sect. 5)

- Explainable AI for Healthcare Practice
- Explainable AI for Knowledge Discovery
- Explainable AI for KG Construction

KG in Life Sciences (Sect. 2)

- Schema-less KGs: Facts in RDF triples
- Schema-based KGs: RDFS, OWL, SHACL, etc.
- Simple ontologies: Taxonomies
- Expressive OWL ontologies

Challenges for Life Science KGs (Sect. 6)

- Scalability
- Evolution & Quality Assurance
- Heterogeneity: Multi-domain & Multi-modality
- Human Interaction & Explainability
- Personalized & Customized KGs
- Distributed KGs
- Representation Learning: Symbolic & Sub-symbolic Integration

Chen, Jiaoyan, et al. "Knowledge Graphs for the Life Sciences: Recent Developments, Challenges and Opportunities." ***Transactions On Graph Data and Knowledge*** (2024). (**New, open access journal in the KG community)**

Conclusion and Discussion

- Knowledge graph definition, construction and curation
- A case of using knowledge graph for ecotoxicological effect analysis
- A survey and position paper on knowledge graphs for life sciences

Q&A

Part II: Scientific Large Language Models

Xiang Zhuang
zhuangxiang@zju.edu.cn

About Me

Mr. Xiang Zhuang

https://tooooooodo.github.io/

zhuangxiang@zju.edu.cn

- I'm a Ph.D. student in Computer Science at Zhejiang University, advised by Professor Huajun Chen and Qiang Zhang.

浙江大学 ZHEJIANG UNIVERSITY

- Prior to that, I received my bachelor's degree in Software Engineering from Zhejiang University in 2020.
- My research interests lie in AI for Science (AI4Science), particularly in applying large language models (LLMs) to biomolecules, such as small molecules and proteins.
- I have published 6 papers as the first author (including co-first author) in prestigious journals and conferences such as *Nature Machine Intelligence*, *Nature Communications*, and *NeurIPS*, etc. I was honored as one of the first "Potential Qingyuan Scholars" by the Chinese Association for Artificial Intelligence, a title awarded to only 10 scholars nationwide.

Outline

1. Introduction and Preliminary
2. Scientific Large Language Models
3. Challenges and Perspective

Outline

1 **Introduction and Preliminary**

2 Scientific Large Language Models

3 Challenges and Perspective

LLMs Revolutionize AGI

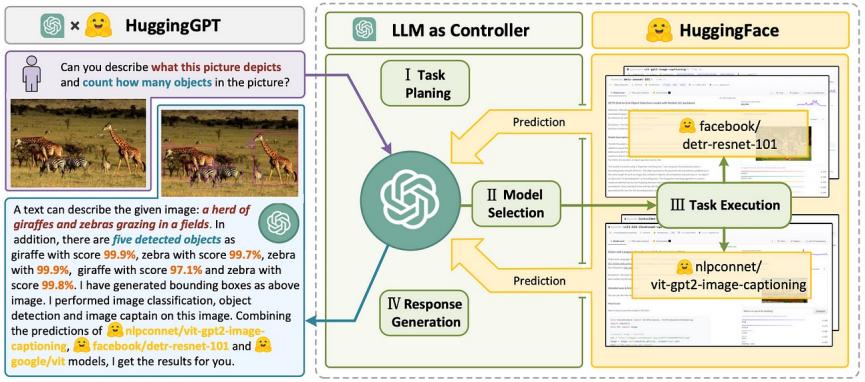
Upstream Techniques

Pre-trained
Large Language Model,
e.g., GPT, GLM, LLaMA

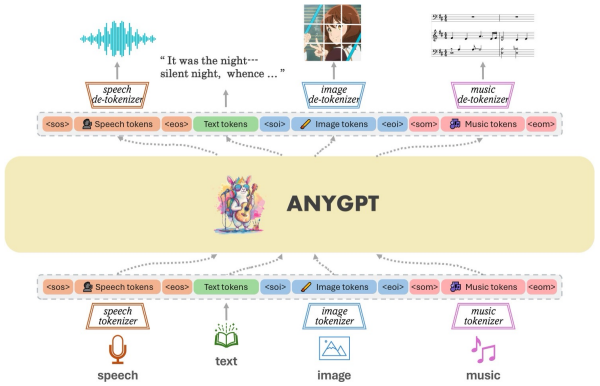
Computational Linguistics:
Understanding human language through computational models!

Downstream Tasks

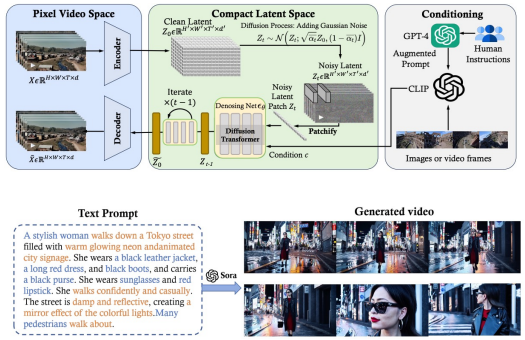
- Question Answering
- Knowledge Extraction
- Complex Reasoning



HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, https://arxiv.org/pdf/2303.17580, 2023.

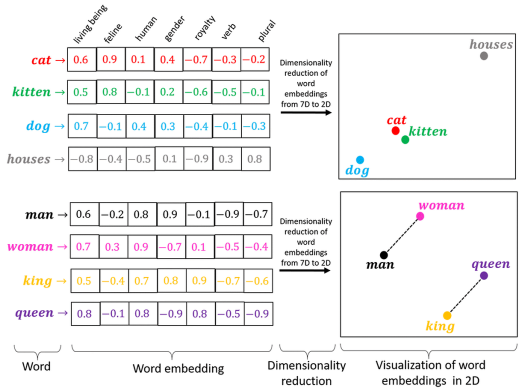


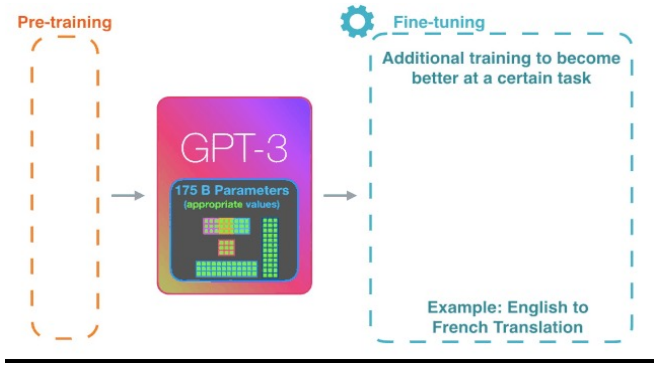
AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling, https://arxiv.org/abs/2402.12226, 2024



Sora, https://openai.com/index/sora/ *OpenAI,* 2024

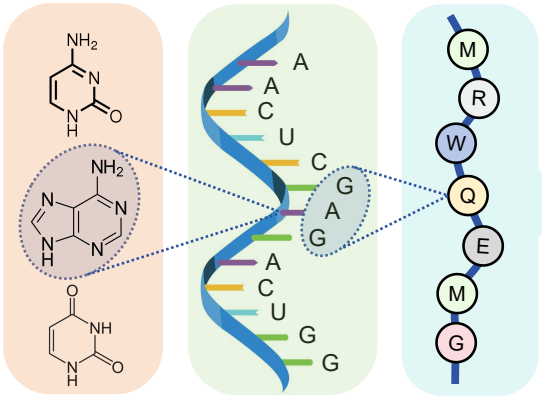
LLMs Limited to Human Language





What can GPTs do?

1. Chatbots talking to historical figures
2. Code generation based on text descriptions
3. Answer medical questions
4. Stylistic transfer of text
5. Write creative novels
6. .....

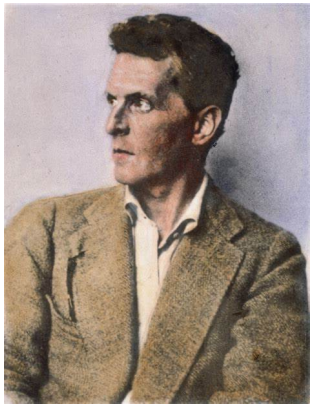


- **Protein**: MOTS-c
- **Gene**: MT-RNR1
- **Organism**: Homo sapiens
- **Sequences**: MRWOEMGYIFYPRKLR

Question: What is the solubility of this protein?

General LLMs

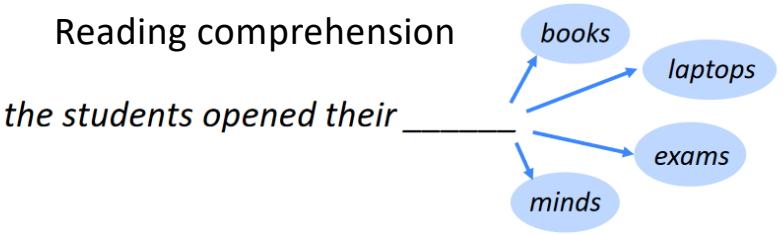
The solubility of a specific protein, such as MOTS-c, isn't typically available as a standard piece of information. Protein solubility can depend on various factors, including pH, temperature, ...



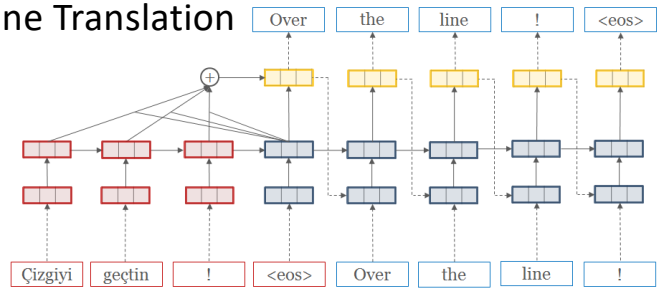
"The limits of my language mean the limits of my world."
--Ludwig Wittgenstein, 1921

Scientific Language Understanding

Reading comprehension

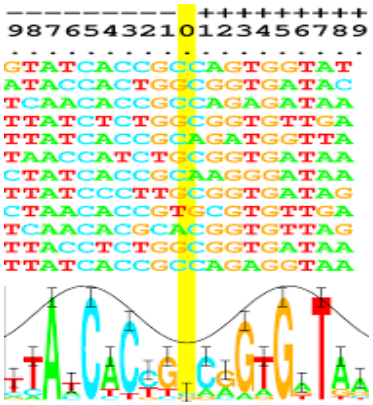


Machine Translation

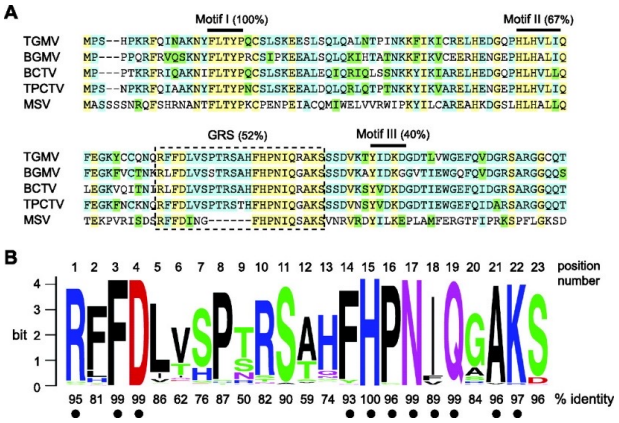


Symbolized human language ⇨ Natural language Models

Symbolized biological language ⇨ Genomic/Protein language Models



Gene

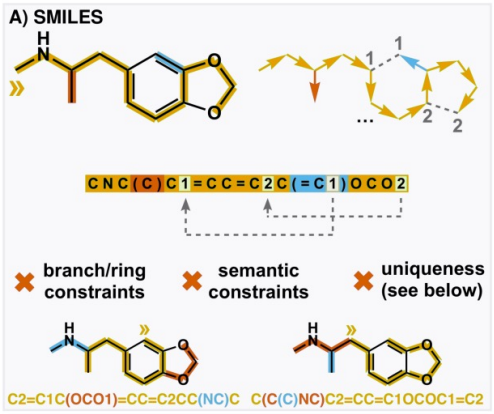


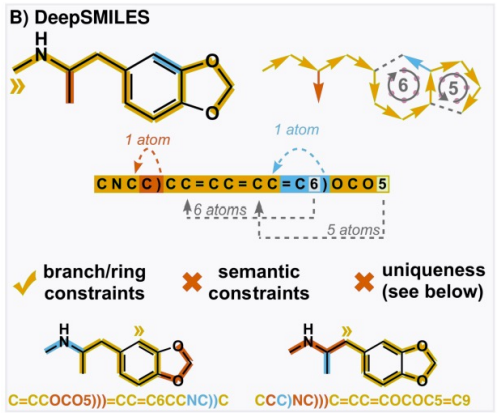
Protein

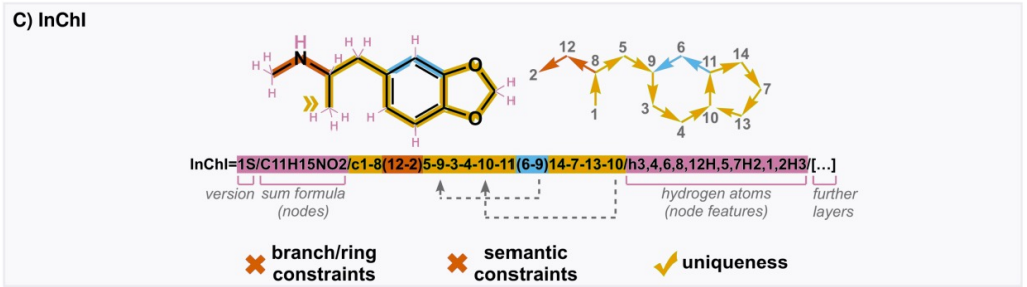
Central Law

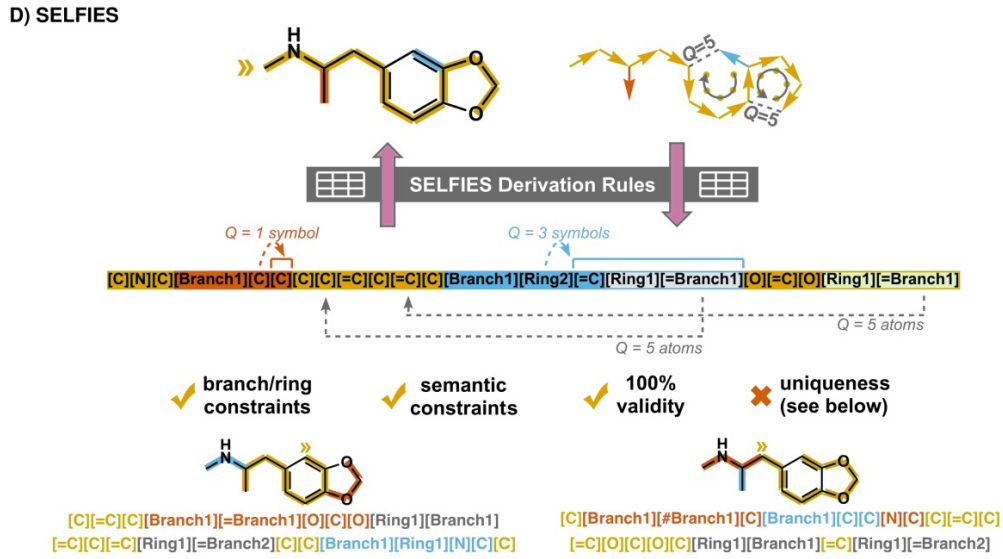
Scientific Language Understanding

Symbolized molecule language ⇨ Molecular language Models

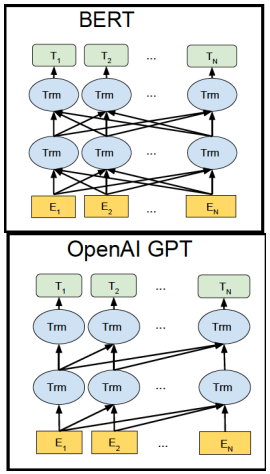








LLMs for Scientific Language?

NLP → Learning the General Rules of Human Languages → $P(w_1, w_2, \cdots, w_n)$ → 

Distributional Semantics

- *You shall know a word by the company it keeps* — Firth, 1957
- Document co-occurrence indicative of topic (**document** as context)
  - E.g. *voting* and *politics*
- Local context reflects its meaning (**word window** as context)
  - E.g. *eat a pizza* vs. *eat a burger*

Mol/ Protein → Learning the General Rules of Scientific Symbols → ? → ?

Fundamental Question: Does the Distributional Semantics Hypothesis Hold?

Outline

1 Introduction and Preliminary

2 Scientific Large Language Models

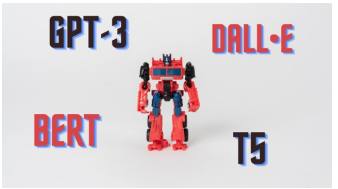
3 Challenges and Perspective

Scope of Scientific LLM

Sci-LLM

- Scientific Texts (*Human Language*)
- Molecular Language (*Chemistry*)
- Protein Language (*Biology*)
- Multimodal Scientific Language
- Scientific Agent and Embodied AI

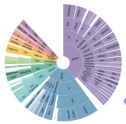
→ Models



→ Datasets

PubMed UniProt ATOM3D

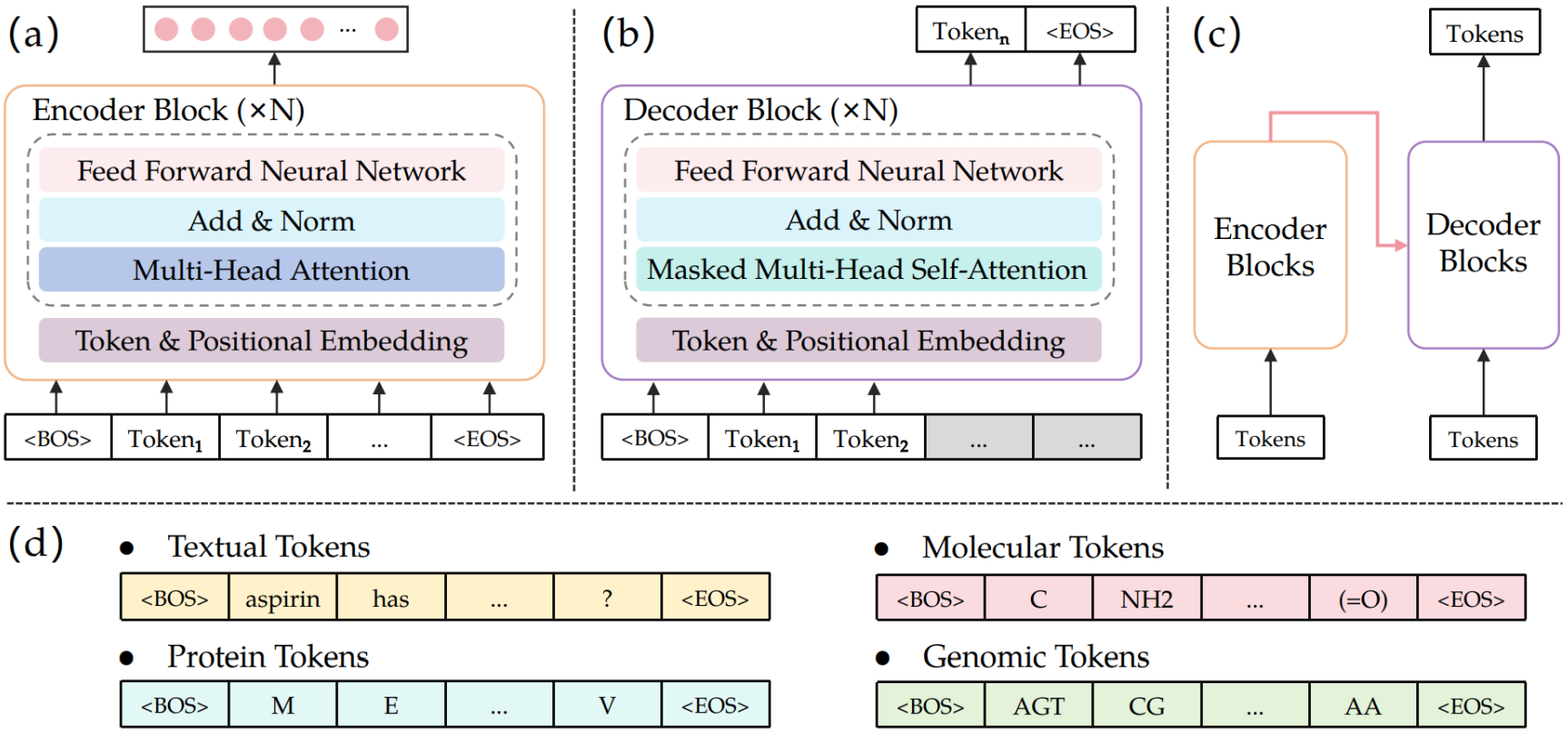
→ Evaluation



MoleculeNet
A Benchmark for Molecular Machine Learning
TorchProtein

→ Scenario

Scope of Scientific LLM



Outline

1 Introduction and Preliminary

2 **Scientific Large Language Models**

- **2.1 Scientific Texts**

3 Challenges and Perspective

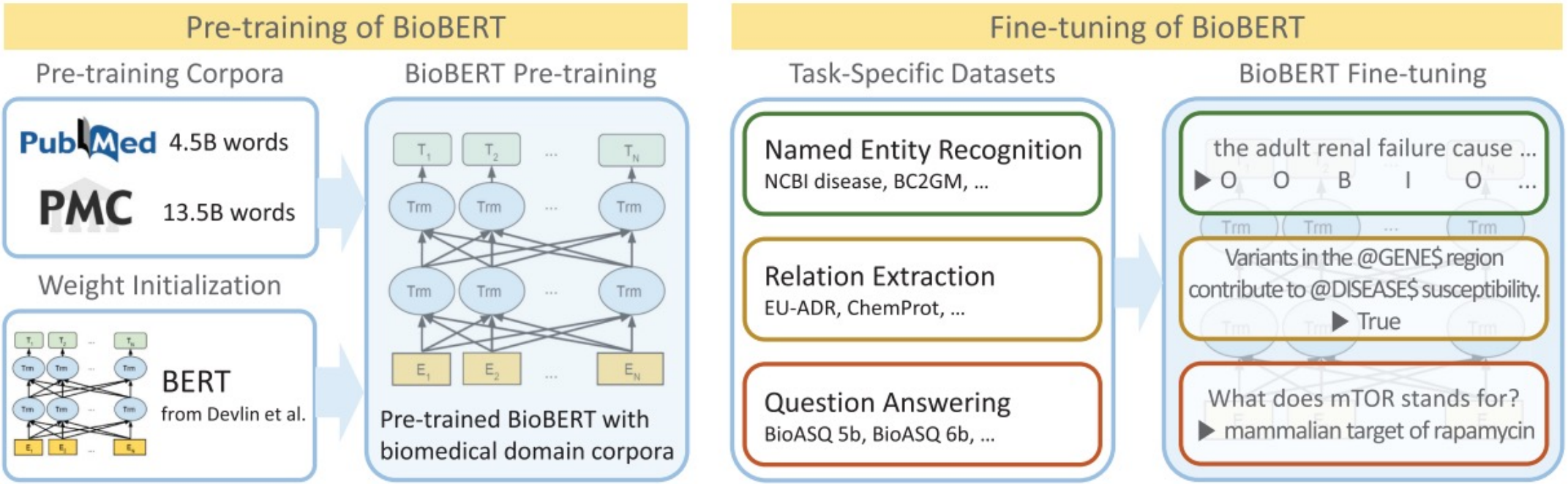
Text-Sci-LLM: Models & Datasets

- Biology + Chemistry + Comprehensive LLM
  - Architecture: **BERT**-based (BioBERT, ChemBERT), **GPT**-based (BioGPT, PharmLLM) and **GLM**-based (SciGLM)
  - Corpus: Initially trained on broad corpora like Wikipedia and papers and then fine-tuned on specific tasks

Table 1. Summary of Text-Sci-LLMs

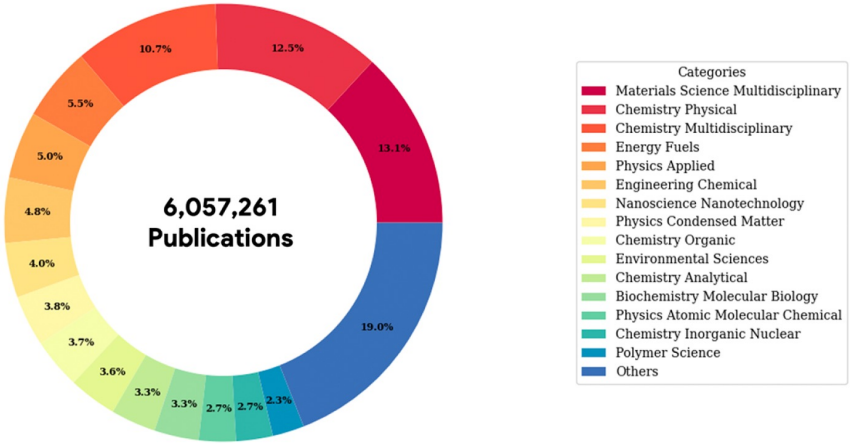
Domain	Model	Time	#Parameters	Base model	Pretraining dataset	Open-source
Biology	BioELMo [150]	2019.04	-	ELMo	PubMed	✓
	BioBERT [177]	2019.05	117M	BERT	PubMed, PMC	✓
	BlueBERT [263]	2019.07	117M	BERT	PubMed	✓
	BioMegatron [299]	2020.10	345M-1.2B	BERT	PubMed, PMC	✓
	PubMedBERT [113]	2020.10	117M	BERT	PubMed	×
	BioM-BERT[6]	2021.06	235M	BERT	PubMed, PMC	✓
	BioLinkBERT[386]	2022.03	110M, 340M	BERT	PubMed	✓
	BioGPT [219]	2023.03	347M	GPT	PubMed	✓
	BioMedGPT-LM [221]	2023.08	7B	LLaMA	PMC, arXiv, WIPO	✓
	BioinspiredLLM [222]	2024.02	13B	Llama-2	Biological article	✓
	BioMistral [172]	2024.02	7B	Mistral	PMC	✓
Chemistry	ChemBERT [115]	2021.06	120M	BERT	Chemical journals	✓
	MatSciBERT [118]	2021.09	117M	BERT	Elsevier journals	✓
	MaterialsBERT [298]	2022.09	-	BERT	Material journals	✓
	ChemLLM [397]	2024.02	7B	InternLM2	ChemData and Multi-Corpus	✓
	ChemDFM [419]	2024.01	13B	InternLM2	Chemical literature, textbooks	✓
	PharmGPT [48]	2024.02	13B, 70B	LLaMA	Paper, report, book, etc.	×
Comprehensive	SciBERT [18]	2019.09	117M	BERT	Semantic Scholar	✓
	ScholarBERT [131]	2023.05	340M, 770M	BERT	Wiki, Books, etc.	✓
	DARWIN-Base [367]	2023.08	7B	LLaMA	SciQ, Web of Science	✓
	SciGLM [367]	2024.03	6B, 32B	ChatGLM3	SciInstruct	✓
	Uni-SMART [367]	2024.06	7B	-	Patents, news, literature, etc.	×
	INDUS [25]	2023.08	125M	RoBERTa	wikipedia, PubMed, PMC, etc.	×

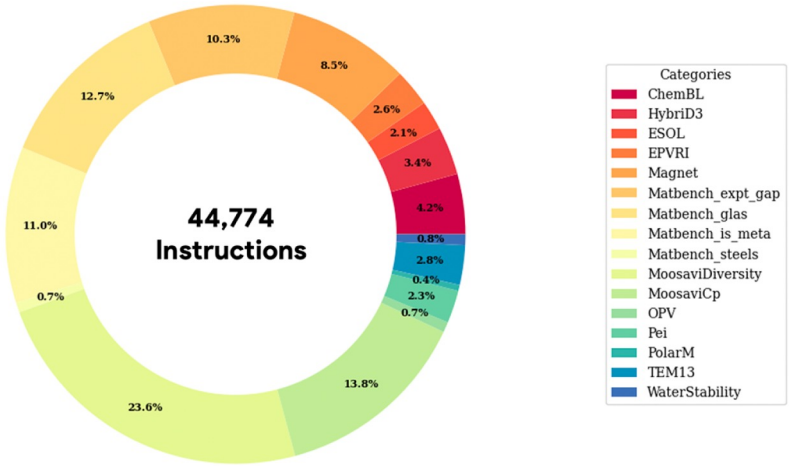
Text-Sci-LLM: Encoder-only



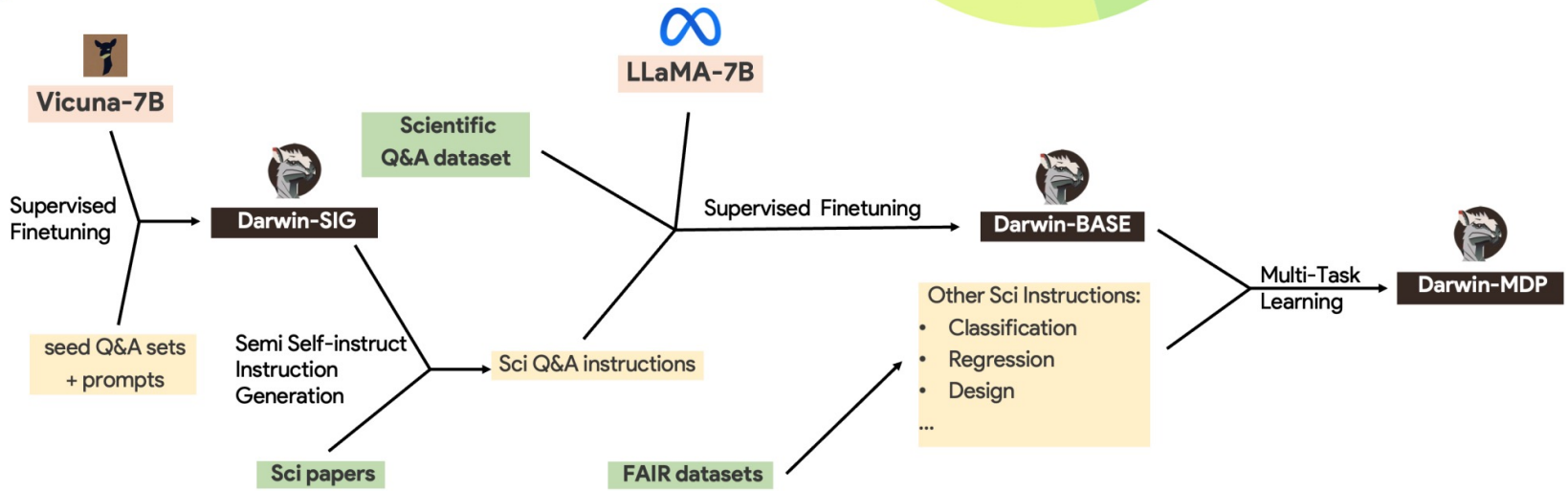
BioBERT: A pre-trained biomedical language representation model for biomedical text mining, **Bioinformatics**, 2019

Text-Sci-LLM: Decoder-only





DARWIN



DARWIN Series: Domain Specific Large Language Models for Natural Science, https://arxiv.org/pdf/2308.13565, 2023

Text-Sci-LLM: Encoder-decoder

Problem When an electron in a certain excited energy level in a one-dimensional box of length 2.00 \u00c5 makes a transition to the ground state, a photon of wavelength 8.79 nm is emitted. Find the quantum number of the initial state.

Correct Answer: 4

Predicted Solution by ChatGLM3-32B-Base (Served as baseline)

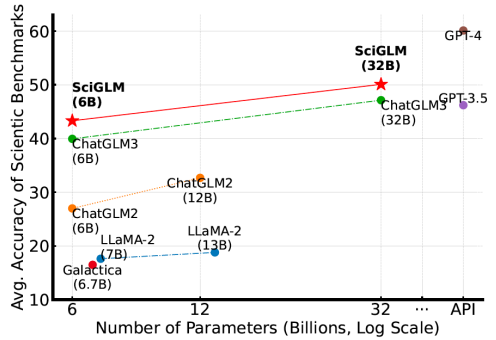
The ground state has $n = 1$. The energy in the $n = 1$ state is $\frac{1}{L^2}$. For the $n = 2$ state, this becomes $\left(\frac{1}{2L}\right)^2$.

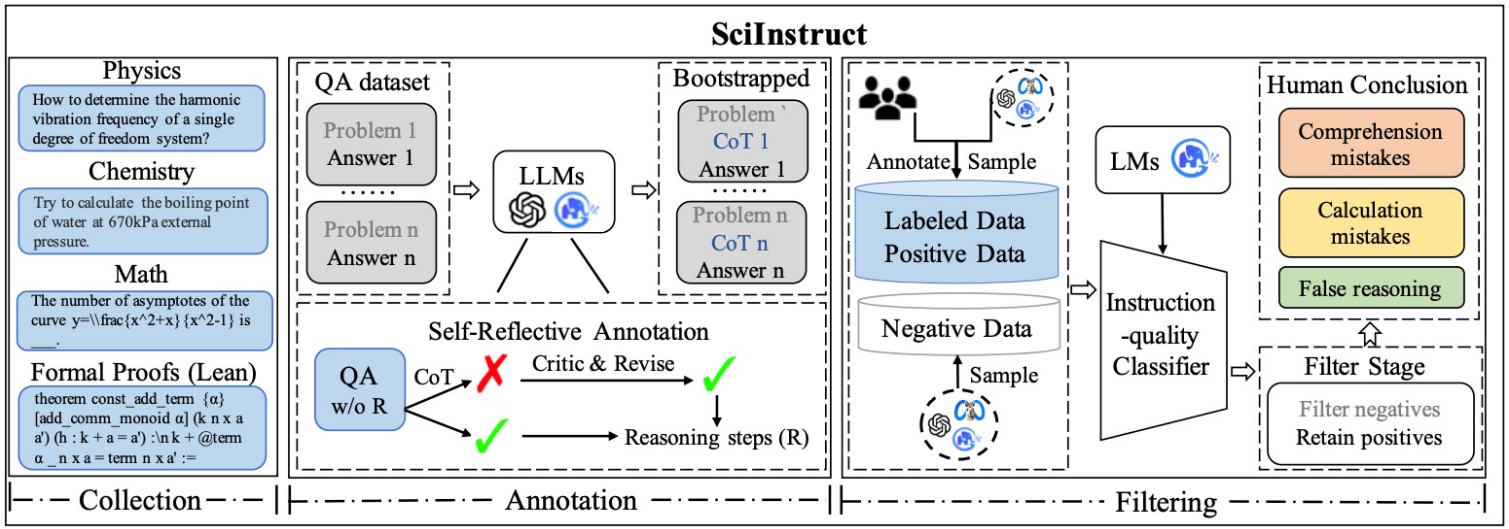
Thus, this energy difference is $\frac{1}{2^{2^2}} - \frac{1}{1^2} = \frac{9}{4}$.

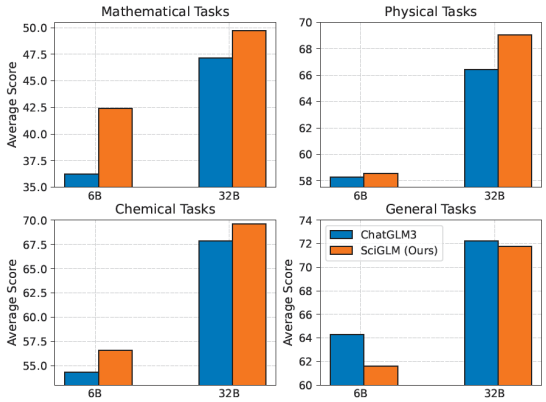
By de Broglie's equation, $hc/\lambda = \frac{9}{4hc\nu}$, so $\nu = \frac{4hc}{9\lambda}$. Thus, $E = h\nu = hc\frac{4}{9\lambda}$.

Therefore, $n = 2$, thus the answer is $\boxed{2}$. ✗

SciGLM







SciGLM: Training Scientific Language Models with Self-Reflective Instruction Annotation and Tuning, https://arxiv.org/pdf/2401.07950, 2024

Text-Sci-LLM: Evaluation

Table 2. Summary of the benchmarks for Text-Sci-LLMs

Dataset	Last updated	Subset	#Item	Domain	Type	Capability	Language
MMLU [127]	2020.09	High-school-biology High-school-chemistry College-biology College-chemistry	344 227 162 110	Biology Chemistry Biology Chemistry	Multiple choice	Pre-college Pre-college College College	English
C-Eval [138]	2023.05	Mid-school-biology Mid-school-chemistry High-school-biology High-school-chemistry College-chemistry	218 210 199 196 253	Biology Chemistry Biology Chemistry Chemistry	Multiple choice	Pre-college Pre-college Pre-college Pre-college College	Chinese
AGIEval [424]	2023.04	Gaokao-biology Gaokao-chemistry	210 207	Biology Chemistry	Multiple choice	Pre-college	Chinese
ScienceQA [216]	2022.09	Natural-science-biology Natural-science-chemistry	4098 1194	Biology Chemistry	Multiple choice / QA	Pre-college	English
XieZhi [114]	2023.06	Science-biology Science-chemistry	2831 399	Biology Chemistry	Multiple choice	Mixed	Both
SciEval [313]	2023.08	Basic-biology Knowledge-biology Calculation-biology Research-biology Basic-chemistry Knowledge-chemistry Calculation-chemistry	2142 1369 299 995 2909 1700 3396	Biology Biology Biology Biology Chemistry Chemistry Chemistry	Multiple choice / QA	Mixed	English
GAOKAO-Bench [406]	2023.11	Biology Chemistry	266 133	Biology Chemistry	Multiple choice / QA	Mixed	Chinese
SciKnowEval [98]	2024.06	Biology Chemistry	27730 22250	Biology Chemistry	Multiple choice / QA / True or flase	Mixed	English
Bioinfo-Bench-QA [49]	2023.10	-	150	Biology	Multiple choice	Post-college	English
BLURB [113]	2020.07	-	648k	Biology	Multiple NLP tasks	Mixed	
PubMedQA [151]	2019.09	-	273.2k	Biology	True or false	College	
SciBench [345]	2023.07	-	272	Chemistry	QA	College	
ARC [64]	2018.03	-	7.78k	Natural Science	Multiple choice	Pre-college	
SciQ [152]	2017.07	-	13.7k	Natural Science	Multiple choice	Mixed	
ChemData [397]	2024.02	-	727k	Chemistry	QA	Mixed	

❖ Evaluation Benchmarks

- **MMLU**: 57 subjects, including STEM humanities, social sciences
- **C-Eval**: 13,948 multi-choice questions spanning 52 diverse disciplines
- **AGIEval**: 20 qualification exams, e.g., Gaokao and American SAT, law school admission tests
- **ScienceQA**: 21,208 multimodal multiple-choice questions, involving elementary and high school science curricula
- **Xiezhi**: 249,587 multi-choice questions spanning 516 diverse disciplines from the elementary to graduate entrance tests
- **SciEval**: 18,000 scientific questions across chemistry, physics, and biology
- **SciQ**: 13,679 science exam questions on subjects like chemistry and biology
- **SciBench**: 695 problems from textbooks, tailored for college-level problem-solving
- **SciAssess**: 14,721 questions across 29 tasks in five domains, with paper memorization, comprehension, and analysis

Text-Sci-LLM: Evaluation

SciKnowEval

- **L1**: **Studying Extensively** (i.e., knowledge coverage): remember and understand concepts
- **L2**: **Enquiring Earnestly** (i.e., knowledge enquiry and exploration): deep enquiry and exploration
- **L3**: **Thinking Profoundly** (i.e., knowledge reflection and reasoning): reasoning and calculating
- **L4**: **Discerning Clearly** (i.e., knowledge discernment and safety assessment): make secure, ethical decisions
- **L5**: **Practicing Assiduously** (i.e., knowledge practice and application): apply knowledge in real-world

SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models, https://arxiv.org/abs/2406.09098, 2024

Text-Sci-LLM: Evaluation

Domain	Ability	Task Name	Task Type	Data Source	Method	#Questions
Biology	L1	Biological Literature QA	MCQ	Literature Corpus	I	14,869
		Protein Property Identification	MCQ	UniProtKB	III	1,500
	L2	Drug-Drug Relation Extraction	RE	Bohrium	II	464
		Biomedical Judgment and Interpretation	T/F	PubMedQA	II	904
		Compound-Disease Relation Extraction	RE	Bohrium	II	867
		Gene-Disease Relation Extraction	RE	Bohrium	II	203
		Detailed Understanding	MCQ	LibreTexts	I	828
		Text Summary	GEN	LibreTexts	I	1,291
		Hypothesis Verification	T/F	LibreTexts	I	619
		Reasoning and Interpretation	MCQ	LibreTexts	I	647
	L3	Solubility Prediction	MCQ	PEER, DeepSol	III	201
		β-lactamase Activity Prediction	MCQ	PEER, Envision	III	209
		Fluorescence Prediction	MCQ	PEER, Sarkisyan's	III	205
		GB1 Fitness Prediction	MCQ	PEER, FLIP	III	201
		Stability Prediction	MCQ	PEER, Rocklin's	III	203
		Protein-Protein Interaction	MCQ	STRING, SHS27K, SHS148K	III	205
		Biological Calculation	MCQ	MedMCQA, SciEval, MMLU	II	60
	L4	Biological Harmful QA	GEN	Self-generated	I	297
		Proteotoxicity Prediction	MCQ, T/F	UniProtKB	III	510
		Biological Laboratory Safety Test	MCQ, T/F	LabExam (ZJU)	II	194
	L5	Biological Protocol Procedure Design	GEN	Protocol Journal	I	591
		Biological Protocol Reagent Design	GEN	Protocol Journal	I	565
		Protein Captioning	GEN	UniProtKB	III	937
		Protein Design	GEN	UniProtKB	III	860
		Single Cell Analysis	GEN	SHARE-seq	III	300
Chemistry	L1	Molecular Name Conversion	MCQ	PubChem	III	1,008
		Molecular Property Identification	MCQ, T/F	MoleculeNet	III	1,625
		Chemical Literature QA	MCQ	Literature Corpus	I	6,316
	L2	Reaction Mechanism Inference	MCQ	LibreTexts	I	269
		Compound Identification and Properties	MCQ	LibreTexts	I	497
		Doping Extraction	RE	NERRE	II	821
		Detailed Understanding	MCQ	LibreTexts	I	626
		Text Summary	GEN	LibreTexts	I	692
		Hypothesis Verification	T/F	LibreTexts	I	544
		Reasoning and Interpretation	MCQ	LibreTexts	I	516
	L3	Molar Weight Calculation	MCQ	PubChem	III	1,042
		Molecular Property Calculation	MCQ	MoleculeNet	II	740
		Molecular Structure Prediction	MCQ	PubChem	III	608
		Reaction Prediction	MCQ	USPTO-Mixed	II	1,122
		Retrosynthesis	MCQ	USPTO-50k	II	1,122
		Balancing Chemical Equation	GEN	WebQC	III	535
		Chemical Calculation	MCQ	XieZhi, SciEval, MMLU	II	269
	L4	Chemical Harmful QA	GEN	Proposition-65, ILO	III	454
		Molecular Toxicity Prediction	MCQ, T/F	Toxric	III	870
		Chemical Laboratory Safety Test	MCQ, T/F	LabExam (ZJU)	II	531
	L5	Molecular Captioning	GEN	ChEBI-20	II	943
		Molecular Generation	GEN	ChEBI-20	II	897
		Chemical Protocol Procedure Design	GEN	Protocol Journal	I	74
		Chemical Protocol Reagent Design	GEN	Protocol Journal	I	129

Evaluation Tasks and Dataset Statistics

Zero-shot performance of LLMs on SciKnowEval

Models	Biology						Chemistry						Overall Rank
	L1	L2	L3	L4	L5	All	L1	L2	L3	L4	L5	All	
GPT-4o	**2.00**	**2.25**	**6.00**	4.00	**1.20**	**3.28**	**1.00**	2.29	4.00	7.00	3.75	**3.46**	1
Gemini1.5-Pro	4.50	5.12	6.14	**2.67**	6.60	5.36	2.67	4.00	**3.57**	**1.33**	11.75	4.67	2
GPT-4-Turbo	4.00	5.50	7.86	3.33	4.00	5.48	3.00	**1.57**	7.29	4.67	7.75	4.83	3
Claude3-Sonnet	5.50	4.12	8.43	4.00	2.00	5.00	6.00	4.43	7.86	8.00	6.00	6.33	4
GPT-3.5-Turbo	2.50	7.62	11.86	4.67	7.60	8.04	9.00	7.86	8.29	7.00	8.00	8.04	5
Llama3-8B-Inst	8.50	5.50	11.71	7.67	10.80	8.80	6.00	6.29	8.57	7.33	14.25	8.38	6
Qwen1.5-14B-Chat	5.50	10.38	8.71	9.00	8.40	8.96	9.33	7.14	6.43	8.00	10.50	7.88	7
Qwen1.5-7B-Chat	9.00	10.50	13.71	8.00	10.60	11.00	10.67	9.86	9.29	11.67	13.50	10.62	10
ChatGLM3-6B	12.00	14.25	11.43	10.00	12.00	12.32	15.33	15.00	15.00	12.33	12.75	14.33	12
Gemma1.1-7B-Inst	16.00	16.75	11.71	14.67	12.80	14.24	17.00	15.86	12.57	11.00	7.25	13.00	14
Llama2-13B-Chat	19.00	11.38	17.14	10.67	10.60	13.36	18.67	13.86	15.57	10.33	14.00	14.54	15
Mistral-7B-Inst	11.00	13.12	14.71	12.67	18.20	14.30	14.33	14.14	15.29	7.33	19.00	14.46	16
ChemDFM-13B	6.50	11.12	12.00	9.67	12.40	11.08	6.67	9.43	8.29	8.33	1.75	7.33	8
ChemLLM-20B-Chat	12.50	6.62	10.14	14.67	13.00	10.32	10.00	7.71	11.00	16.33	4.00	9.42	9
MolInst-Llama3-8B	13.50	9.88	7.86	12.00	18.20	11.52	9.33	9.57	7.43	9.33	17.75	10.25	11
Galactica-30B	11.00	13.75	8.43	16.67	16.80	13.00	7.67	16.43	13.00	16.67	16.00	14.29	13
SciGLM-6B	16.00	14.12	11.43	16.00	16.60	14.24	16.00	15.29	13.14	17.67	15.25	15.04	17
ChemLLM-7B-Chat	15.00	15.88	13.86	14.33	16.60	15.20	15.33	14.86	15.43	16.00	7.75	14.04	18
Galactica-6.7B	17.50	16.50	11.86	18.00	19.20	16.00	13.00	17.86	13.00	13.00	18.50	15.33	19
LlaSMol-Mistral-7B	19.50	16.75	14.14	19.67	17.20	16.68	19.33	18.71	16.29	20.00	**1.25**	15.33	20

Outline

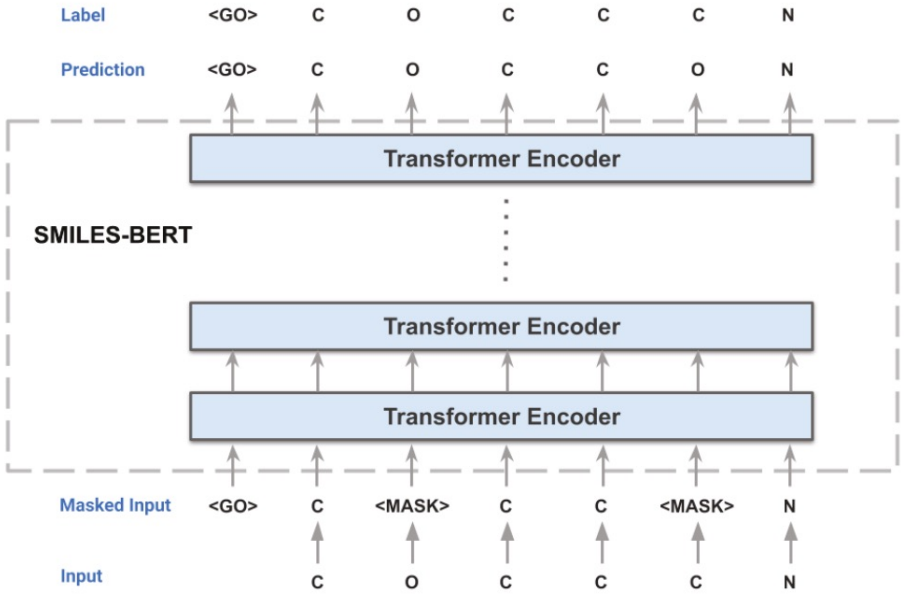
1 Introduction and Preliminary

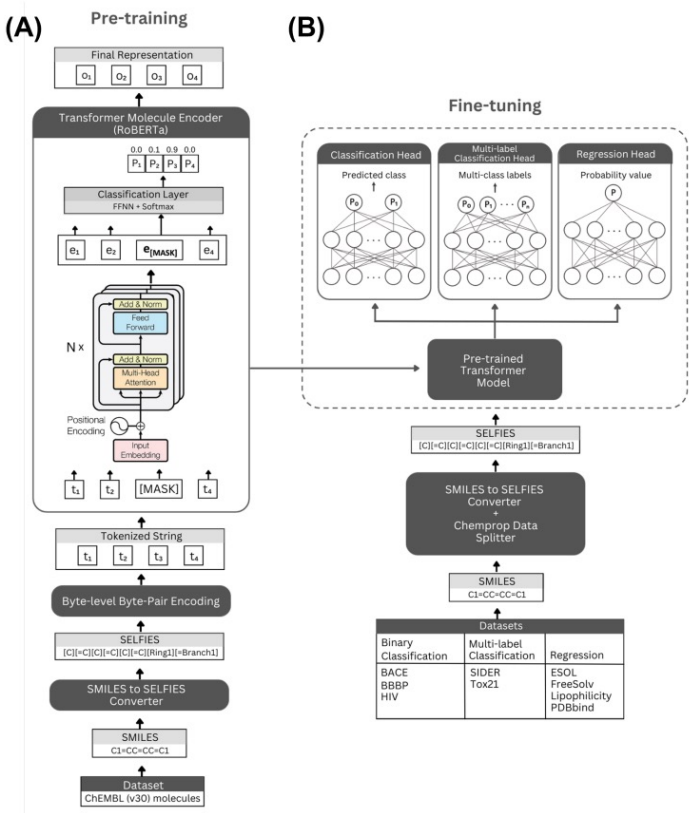
2 **Scientific Large Language Models**

- **2.2 Molecule Language**

3 Challenges and Perspective

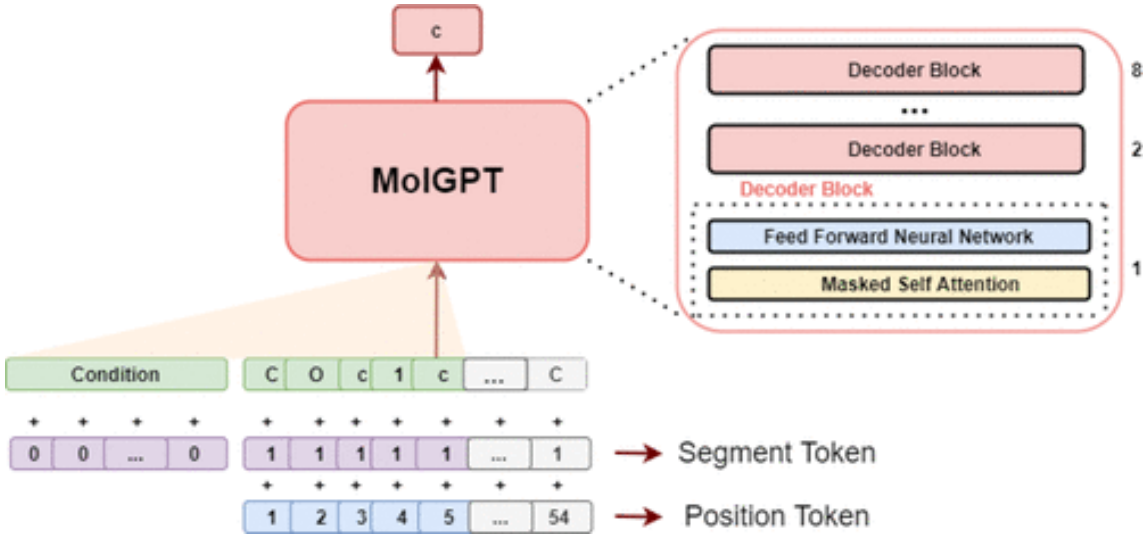
Mol-LLM: Encoder-only



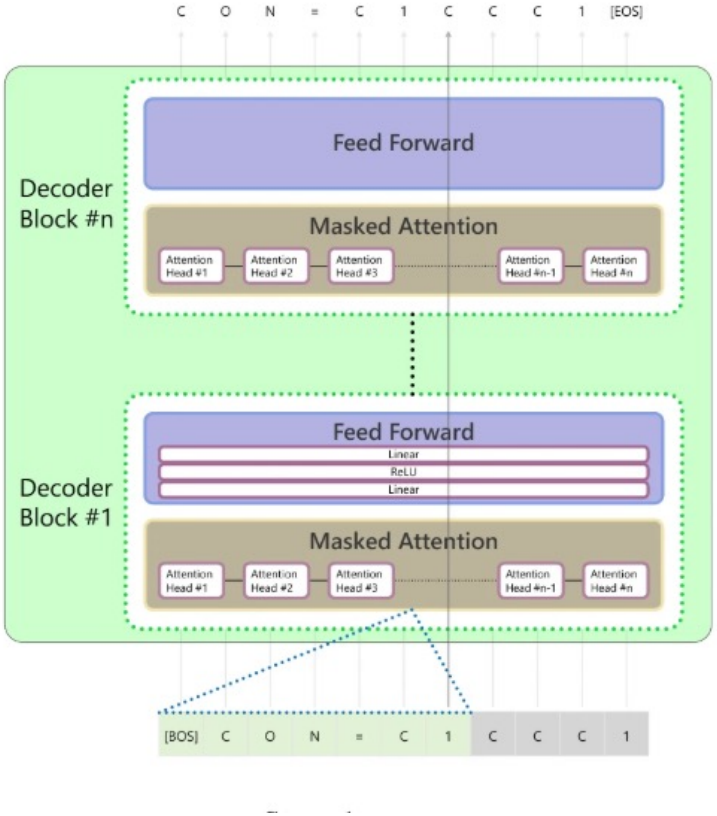


SMILES-BERT: large scale unsupervised pre-training for molecular property prediction, In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 429-436. 2019.
Selformer: Molecular representation learning via selfies language models. https://arxiv.org/pdf/2304.04662

Mol-LLM: Decoder-only

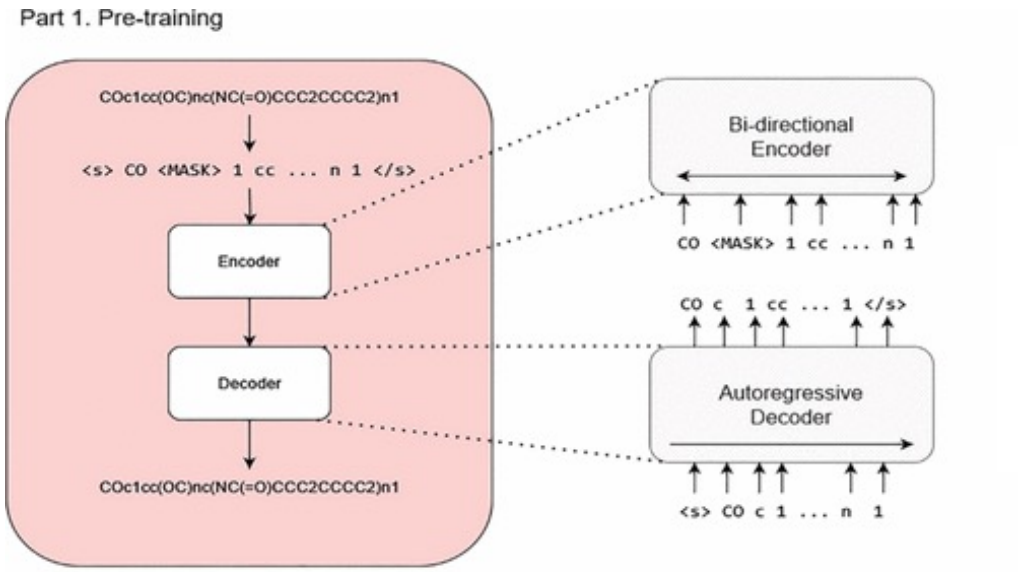


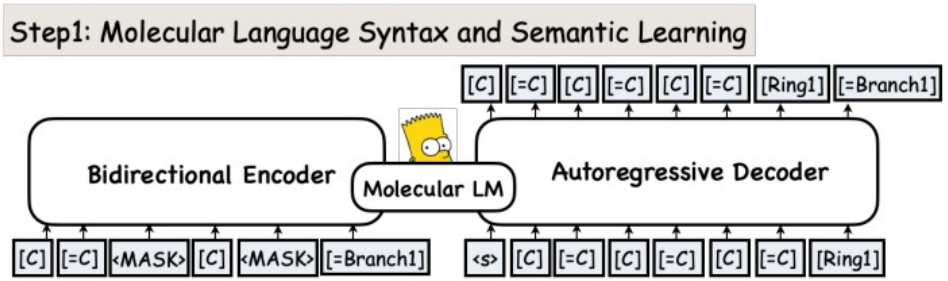
Taiga



MolGPT: molecular generation using a transformer-decoder model. Journal of Chemical Information and Modeling, 2022
Molecule generation using transformers and policy gradient reinforcement learning, Scientific reports, 2022

Mol-LLM: Encoder-decoder





BartSmiles: Generative Masked Language Models for Molecular Representations, Journal of Chemical Information and Modeling, 2024

Domain-Agnostic Molecular Generation with Chemical Feedback, ICLR2024

Mol-LLM: Datasets

	Dataset	Last updated	Subset/Version	Scale	Keywords/Tasks
Pre-training	ZINC	2015.10	ZINC-15 [303]	120M	Ligand discovery
		2020.12	ZINC-20 [139]	140M	
		2012.07	ZINC-250k[138]	250k	
	PubChem [163]	2023.01	Compound	111.9M	Unique chemical structures
			Substance	296.9M	Chemical entities
			BioAssay	1.5M	Biological experiments
			Bioactivity	296.8M	Biological activity data points
			Protein	185.1K	Tested and identified proteints
			Gene	104K	Tested and identified genes
			Pathway	239K	Interactions between chemicals, genes, and proteins
			Cell Line	2K	Tested cell lines
			Taxonomy	112.6K	Organisms of proteins/genes
			Patent	42.4M	Patents with links in PubChem
			Data Sources	872	Organizations contributing data to PubChem
	USPTO [210]	2012.10	USPTO MIT	480k	US Patents, unique reactions
			USPTO-15K	15k	
			USPTO-full	950k	
	PCQM4M	2021.10	PCQM4Mv2 [133]	3.7M	Quantum property, molecular graphs pred.
		2021.06	PCQM4M-LSC [386]	3.8M	Large-Scale
	GEOM [11]	2022.04	QM9	134K	Conformers
			AICures drug dataset	317.9K	Organic molecules, high-quality conformers
			part of MoleculeNet	16.9K	Biophysics, physiology, physical chemistry
	MolTLU [17]	2023.10	ToyMix	154K	Quantum ML, drug discovery, GNN expressivity
			LargeMix	5.4M	Biological properties
			UltraLarge	83M	Quantum properties
	ChEMBL [390]	2023.05	-	22.7M	Drug-like bioactive compounds
	DrugBank 5.0 [355]	2017.11	-	500K	Drug's mechanisms, interactions, targets
	GDB-17 [279]	2012.10	-	166.4B	Organic small molecules
	ExCAPE-DB [308]	2017.03	-	70M	Bioactivity, Chemogenomics

Benchmarks	MoleculeNet [359]	2017.10	QM7	7K	Quantum mechanics prediction
			QM8	22K	
			QM9	134K	
			ESOL	1K	Physical chemistry prediction
			FreeSoly	643	
			Lipophilicity	4K	
			PCBA	440K	Biophysics prediction
			MUV	93K	
			HIV	42K	
			PDBbind	12K	
			BACE	1.5K	
			BBBP	2K	Physiology prediction
			Tox21	8K	
			ToxCast	8.6K	
			SIDER	1.4K	
			ClinTox	1.4K	
	MARCEL [427]	2023.09	Drugs-75K	75k	Highly flexible molecules
			Kraken	1.5k	Organophosphorus ligand sterics
			EE	872	Chiral catalysts and enantiomers
			BDE	5.9k	Organometallic catalyst conformations
	GuacaMol [33]	2019.03	-	-	Molecule generation
	MOSES [263]	2020.12	ZINC Clean Leads	1.9M	Molecule generation
	ADMETlab 2.0 [366]	2021.04	-	250K	Property prediction
	SPECTRA [85]	2024.02	-	637k	Generalization ability evaluation
	Molecule3D [375]	2021.09	GEOM, QM9	3.9M	Ground-state 3D Molecular Geometry Prediction

Outline

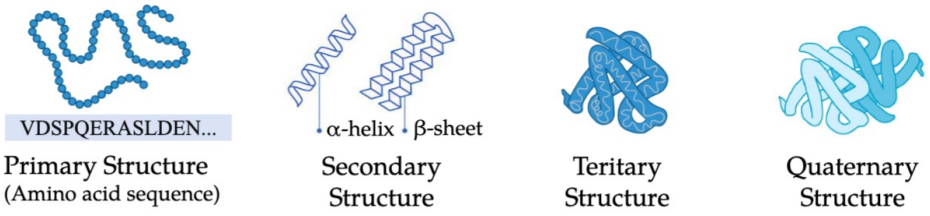
1 Introduction and Preliminary

2 **Scientific Large Language Models**

- **2.3 Protein Language**

3 Challenges and Perspective

Prot-LLM: Models



Primary Structure (Amino acid sequence) | Secondary Structure | Teritary Structure | Quaternary Structure

	Model	Time	#Parameters	Base model	Pretraining Dataset	Capability	Open-source
Encoder-only	ESM-1b [281]	2020.02	650M	RoBERTa	UniRef50	Secondary struct. pred., Contact pred., etc.	✓
	ESM-MSA-1b [277]	2021.02	100M	ESM-1b	UniRef50	Secondary struct. pred., Contact pred., etc.	✓
	ESM-1v [238]	2021.02	650M	ESM-1b	UniRef90	Mutation effect pred.	✓
	ProtTrans [87]	2021.07	-	BERT, Albert, Electra	UniRef, BFD	Secondary struct. pred., Func. pred., etc	✓
	PMLM [122]	2021.07	87M - 731M	Trans. enc.	Uniref50/Pfam	Contact pred.	×
	Mansoor *et al.* [229]	2021.09	100M	ESM-1b	-	Mutation effect pred.	×
	ProteinBERT [32]	2022.02	16M	BERT	UniRef90	Func. pred.	✓
	LM-GVP [351]	2022.04	-	Trans. enc	-	Func. pred.	✓
	RSA [225]	2022.05	-	ESM-1b	-	Func. pred.	✓
	OntoProtein [403]	2022.06	-	BERT	ProteinKG25	Func. pred.	✓
	ESM-2 [195]	2022.07	8M - 15B	RoBERTa	UniRef50	Func. pred., Struct. pred.	✓
	PromptProtein [354]	2023.02	650M	RoBERTa	UniRef50, PDB	Func. pred.	✓
	KeAP [428]	2023.02	-	RoBERTa	ProteinKG25	Func. pred.	✓
	ProtFlash [339]	2023.10	79M/174M	Trans. enc	UniRef50	Func. pred.	✓
	ESM-GearNet [416]	2023.10	-	ESM-1b, GearNet	-	Func. pred.	✓
	SaProt [312]	2023.10	650M	BERT	-	Mutation effect pred.	✓
	ProteinNPT [248]	2023.12	-	Trans. enc.	-	Fitness pred., Redesign	×
	Outeiral *et al.* [254]	2024.02	10M - 5B	Trans. enc.	European Nucleotide Archive	Protein represent learning	✓
	ESM All-Atom [422]	2024.06	35M	RoBERTa	AlphaFold DB	Unified Molecular Modeling	×
	KnowRLM [349]	2024.06	-	Trans. enc.	-	Protein Directed Evolution	×
	ESM3 [121]	2024.06	98B	RoBERTa	PDB	Seq. pred., Func. pred., Struct. pred.	✓
Decoder-only	ProGen [226]	2020.03	1.2B	GPT	Uniparc SWISS-Prot	Functional prot. gen.	✓
	ProtGPT2 [99]	2021.01	738M	GPT	Uniref50	De novo protein design and engineering	✓
	ZymCTRL [241]	2022.01	738M	GPT	BRENDA	Functional enzymes gen.	✓
	RITA [128]	2022.05	1.2B	GPT	UniRef100	Functional prot. gen.	×
	IgLM [302]	2022.12	13M	GPT	-	Antibody design	✓
	ProGen2 [245]	2023.10	151M - 6.4B	GPT	Uniref90, BFD30, PDB	Functional prot. gen.	✓
	ProteinRL [309]	2023.10	764M	GPT	-	Prot. design	×
	PoET [9]	2023.11	201M	GPT	-	Prot. family. gen.	×
	C. Frey *et al.* [104]	2024.03	9.87M/1.03M	GPT	hu4D5 antibody mutant	Functional prot. gen.	×
Encoder-Decoder	Fold2Seq [40]	2021.01	-	Transformer	-	Prot. design	✓
	MSA2Prot [273]	2022.04	-	Transformer	-	Prot. gen., Variant func. pred.	×
	Sgarbossa *et al.* [295]	2023.02	-	MSA Transformer	-	Prot. gen.	✓
	Lee *et al.* [178]	2023.02	150M	Transformer	-	Prot. design	×
	LM-Design [424]	2023.02	664M	Transformer	-	Prot. design	✓
	MSA-Augmenter [402]	2023.06	260M	Transformer	Uniref50	MSA gen.	✓
	ProstT5 [125]	2023.07	3B	T5	PDB	Seq.-struct. translation	✓
	xTrimoPGLM [44]	2023.07	100B	GLM	Uniref90, ColdFoldDB	Prot. gen., Func. pred.	×
	SS-pLM [294]	2023.08	14.8M	Transformer	Uniref50	Prot. gen.	×
	pAbT5 [62]	2023.10	-	T5	-	Prot. design	×
	ESM-GearNet-INR-MC [179]	2024.04	-	Transformer	Swiss-Prot, AlphaFoldDB	Prot. gen	×

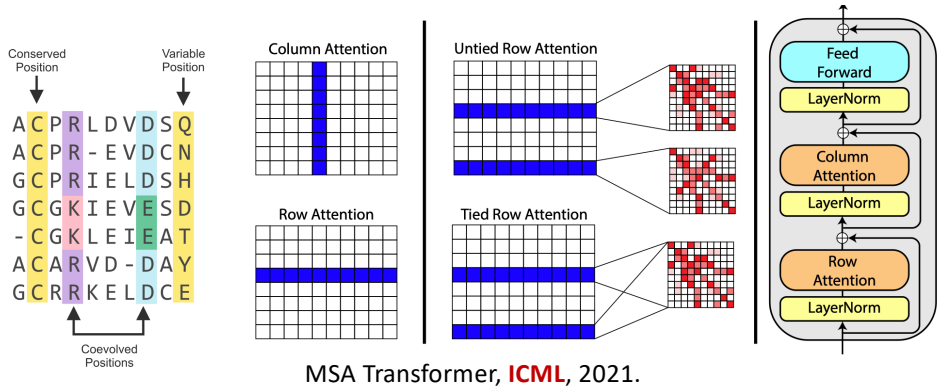
(Continued)

❖ Protein LLMs

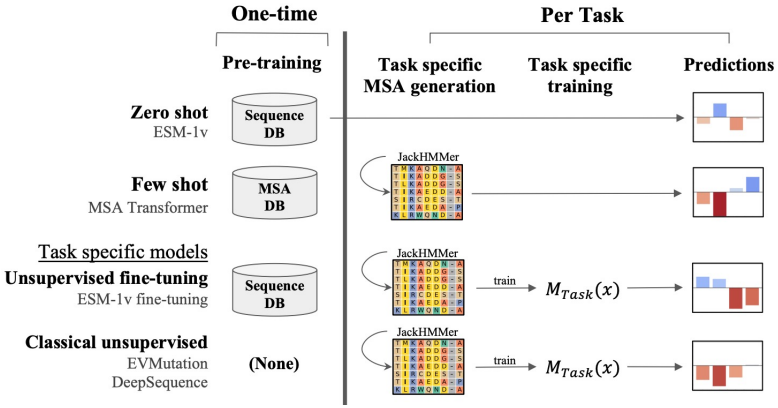
- **Protein vocabulary**: 20 amino acids in nature, special tokens like <BOS> and <EOS>
- **Architectures**: BERT, RoBERTa, GPT, GLM, T5, Transformer
- **Sizes**: 100M, 1B, 10B, 100B
- **Datasets**: Uniref, Pfam, SwissProt, PDB, BFD30, AlphaFoldDB, ColdFoldDB
- **Tasks**: function prediction, family prediction, protein-protein interaction, contact prediction, mutation effect prediction, structure prediction, sequence optimization, protein de novo design, inverse folding

Prot-LLM: Encoder-only

ESM series

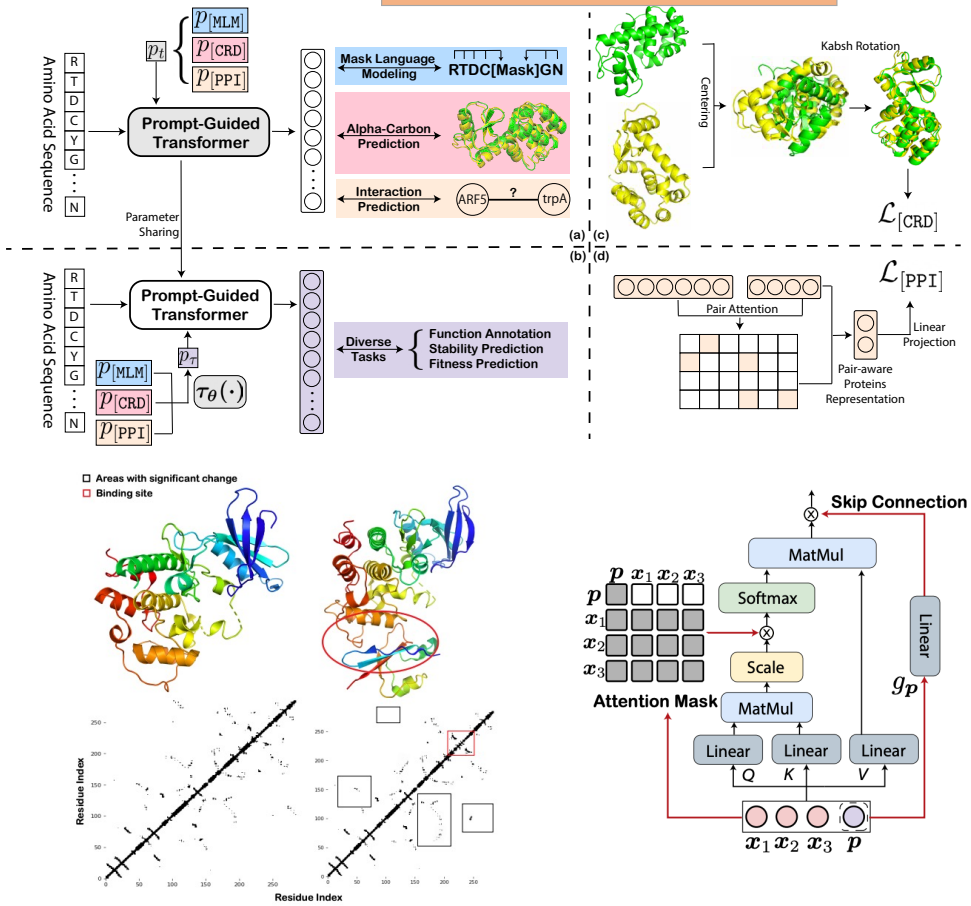


MSA Transformer, **ICML**, 2021.



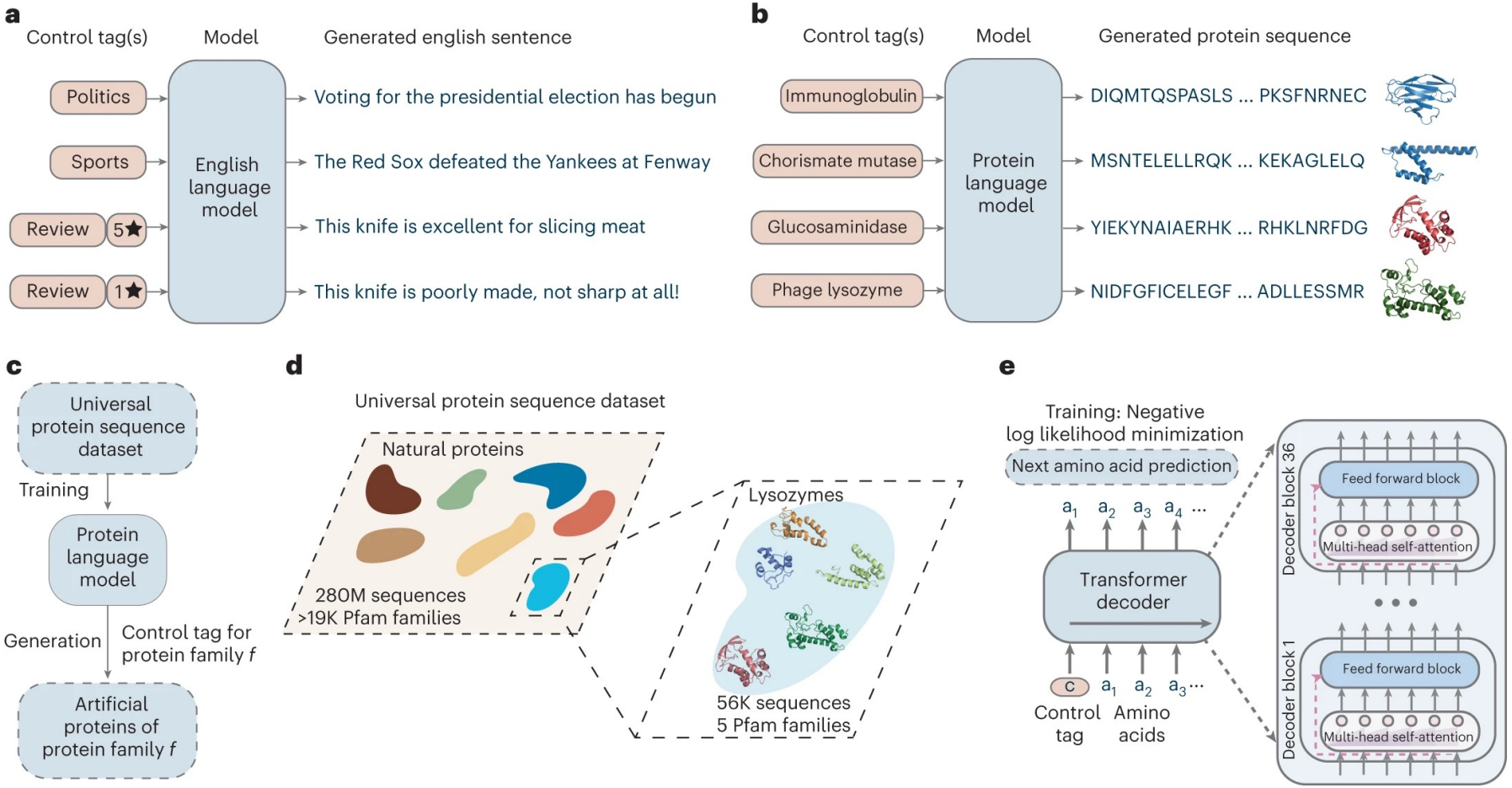
Language models enable zero-shot prediction of the effects of mutations on protein function, www.biorxiv.org/content/10.1101/2021.07.09.450648v2, 2021

PromptProtein



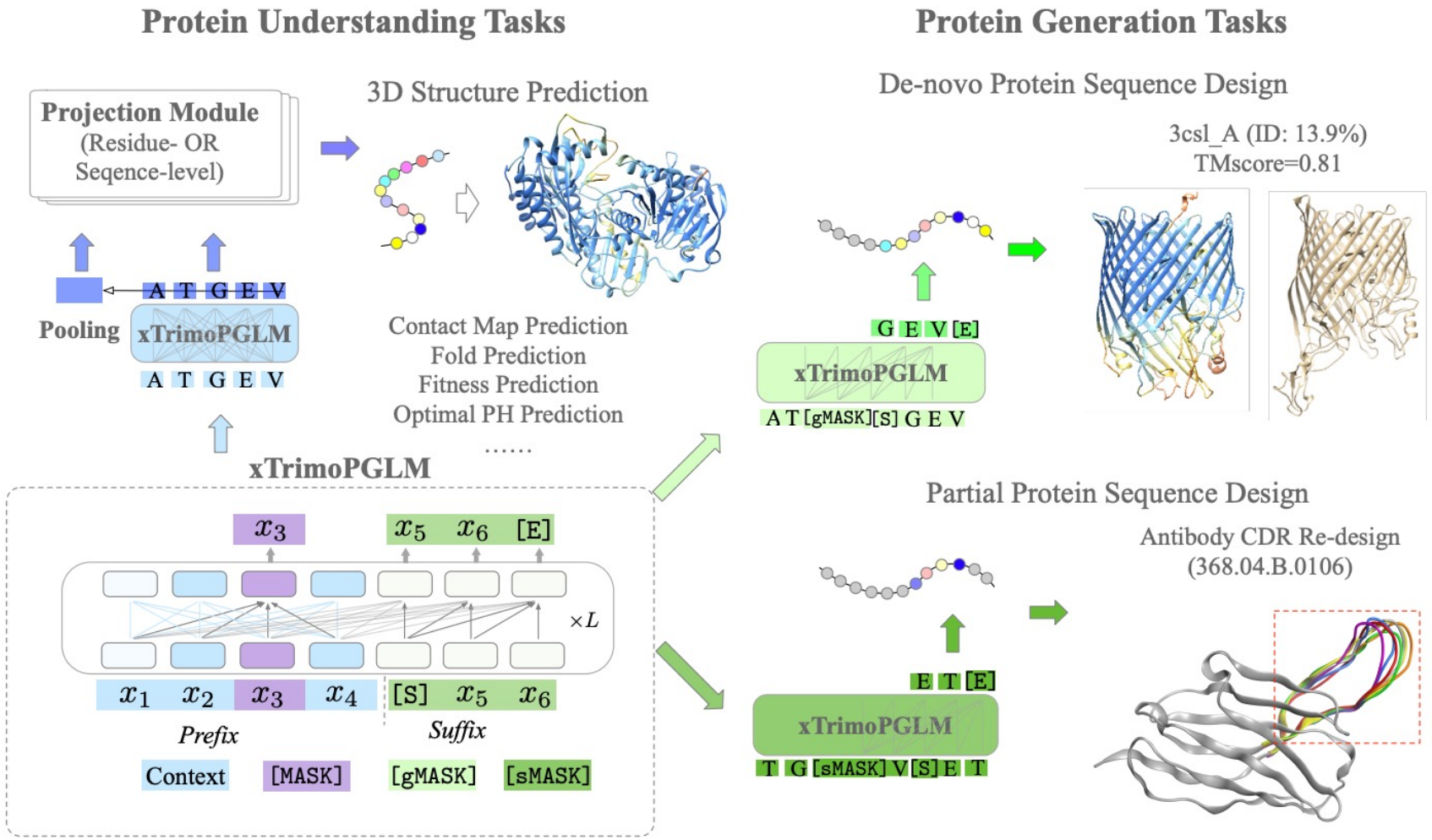
Multi-level Protein Structure Pre-training with Prompt Learning, **ICLR**, 2022

Prot-LLM: Decoder-only



ProGen: Large language models generate functional protein sequences across diverse families, **Nature Biotechnology**, 2023

Prot-LLM: Encoder-decoder



xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein, https://arxiv.org/pdf/2401.06199, 2024

Prot-LLM: Datasets

Table 6. Summary of datasets for Prot-LLMs

	Dataset	Last updated	Scale	Keywords
Pretraining	UniRef100 [315, 316]	2023.11	314M	Complete collection of protein sequences from UniProtKB
	UniRef90 [315, 316]		150M	Cluster UniRef100 sequences at 90% sequence identity level
	UniRef50 [315, 316]		53M	Cluster UniRef100 sequences at 50% sequence identity level
	UniProtKB/Swiss-Prot [29]	2023.11	570K	High-quality, manually curated protein sequence database
	UniProtKB/TrEMBL [240]		251M	Computationally annotated protein sequence database
	UniParc [69]	2023.11	632M	Comprehensive and non-redundant protein sequence database
	Pfam [100]	2023.09	47M	Protein family database
	BFD [157, 306, 307]	2021.07	2.5B	Protein sequences from multiple databases and resources
	PDB [364]	2023.12	214K	Experimentally determined accurate protein structures
	AlphaFoldDB [157, 334]	2021.11	200M	Protein structures predicted by AlphaFold
Benchmark	CASP [171]	2022.01	-	Structure prediction competition
	EC [236]	2023.11	2.6 M	Enzymes classification database
	GO [8]	2023.11	1.5M	Gene Ontology knowledgebase
	CATH [252]	2023.02	151M	Classification of protein structures
	HIPPIE [288]	2022.04	39K	Protein-protein interaction networks
	SCOP [214]	2023.01	914K	Protein structure classification
	ProteinGym [247]	2022.12	~ 300K	Predict the effects of protein mutations
	FLIP [75]	2022.01	~ 320K	Fitness landscape prediction (AAV, Thermostability, GB1)
	PEER [377]	2022.11	~ 390K	Protein function, Localization, Structure prediction, Protein-protein interaction, Protein-ligand interaction
	TAPE [276]	2021.09	~ 120K	Remote homology detection, Secondary structure, Contact, Fluorescence, Stability prediction
	Reactome [144]	2023.12	~ 3M	Biological interactions and pathways
	STRING [317]	2022.11	59.3M	Protein-Protein interaction networks
	BioGRID [253]	2023.12	271k	Genetic and protein interactions
	InterPro [258]	2024.01	~ 41k	Classification of protein families

Prot-LLM: Evaluation

Evaluation Metric

- **Novelty**: the fraction of the generated proteins that are not present in the training set
- **Frechet Protein Distance**: the similarity between a set of generated proteins (G) and a reference set (R)

$$\mathrm{FPD} = ||\mu_G - \mu_R||^2 + \mathrm{Tr}(\Sigma_G + \Sigma_R - 2\sqrt{\Sigma_G \Sigma_R})$$

- **Diversity**: analyzing the variety of the generated proteins against known protein databases with BLAST, metrics such as sequence similarity, percentage of unique sequences, and alignment scores

- **Foldability**: the average per-residue confidence score, denoted as **pLDDT**, across the entire protein sequence, being an indicator of the model's confidence in its predictions for individual residues

- **Recovery**: the success or accuracy in predicting the correct amino acid sequence that corresponds to a given 3D structure. A high recovery rate indicates that the designed sequences are likely to fold the desired structures.

Outline

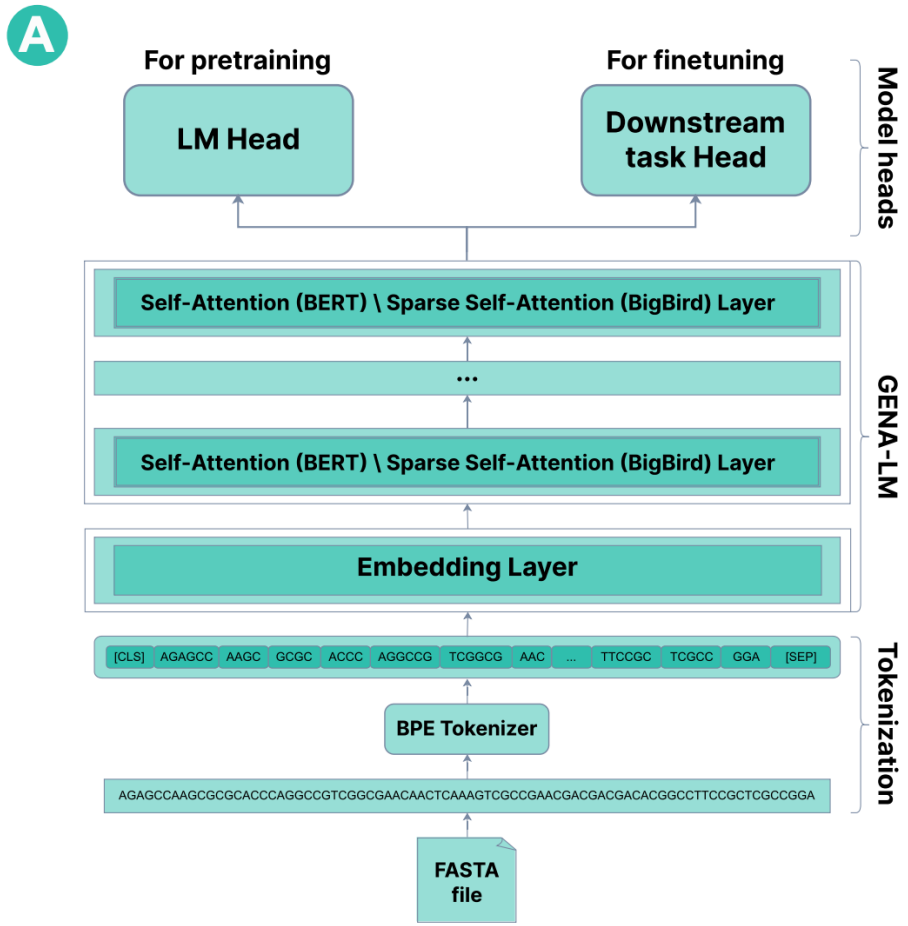
1 Introduction and Preliminary

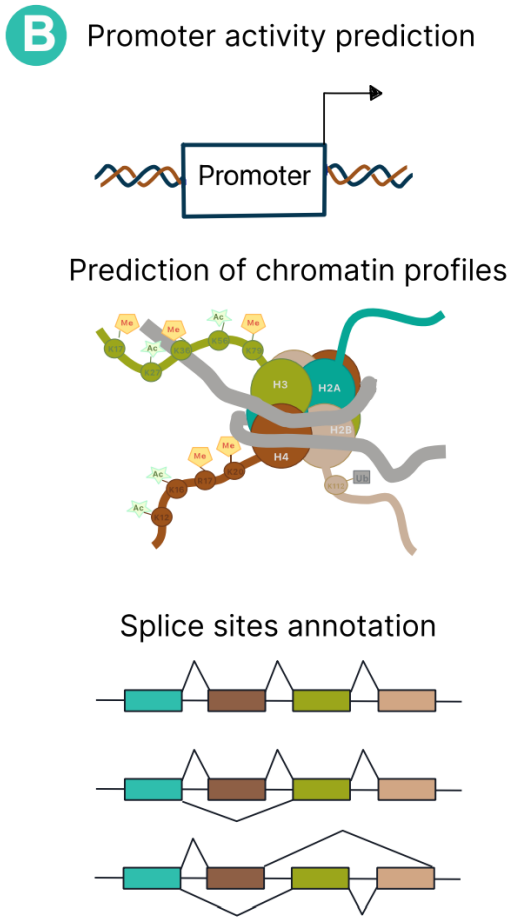
2 **Scientific Large Language Models**

- **2.4 Genomic Language**

3 Challenges and Perspective

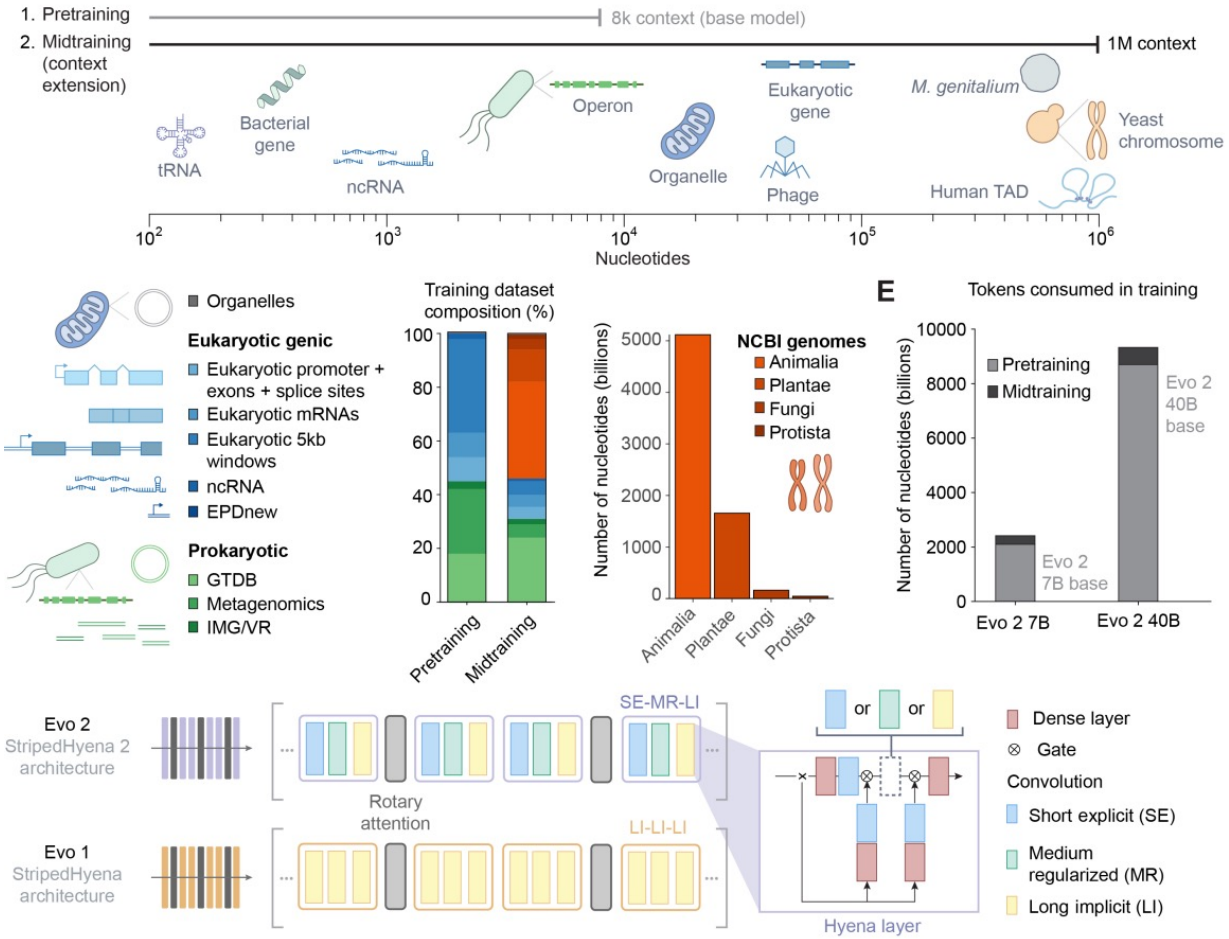
Gene-LLM:Encoder-only





GENA-LM: a family of open-source foundational DNA language models for long sequences, *Nucleic Acids Research, 2025*

Gene-LLM:Decoder-only



Genome modeling and design across all domains of life with Evo 2*, 2025*

Outline

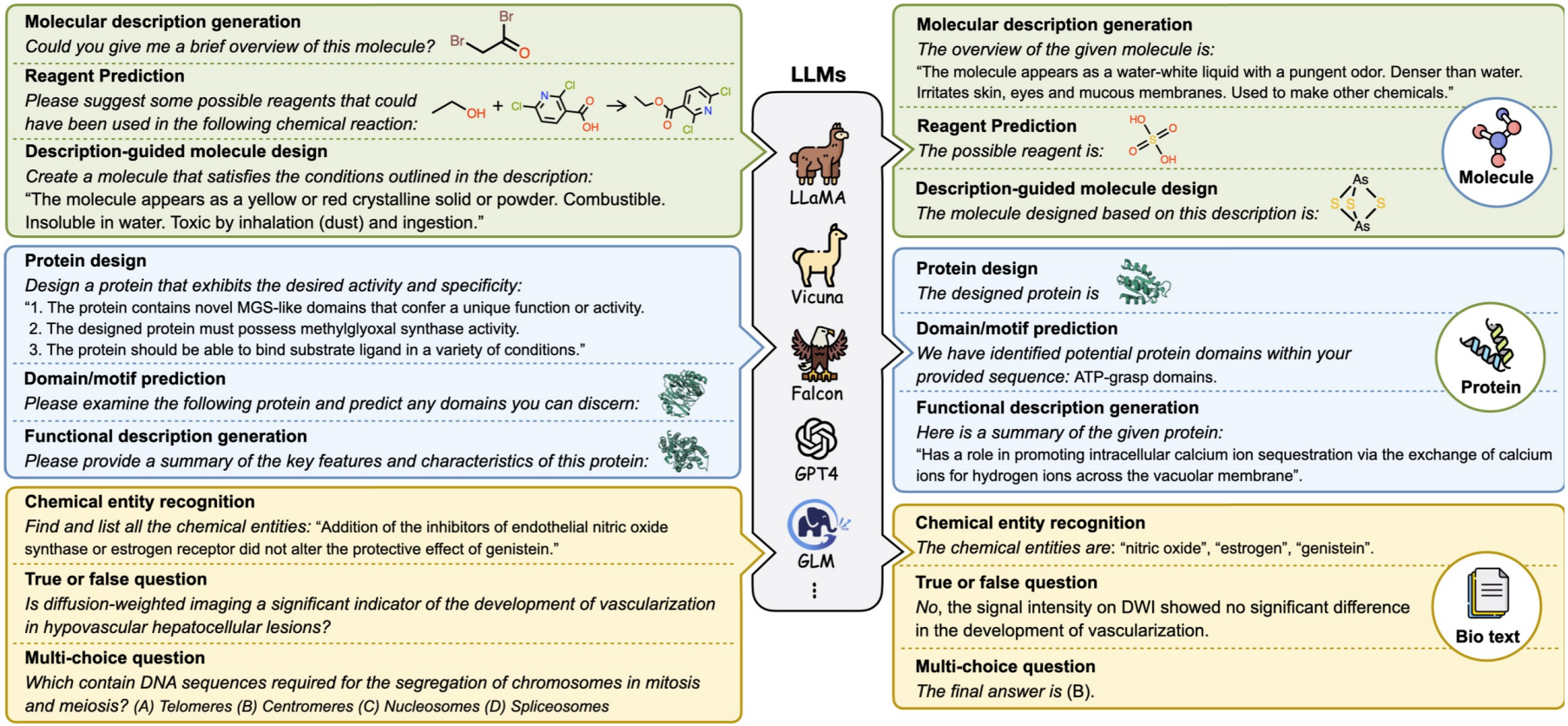
1 Introduction and Preliminary

2 **Scientific Large Language Models**

- **2.5 Comprehensive Language**

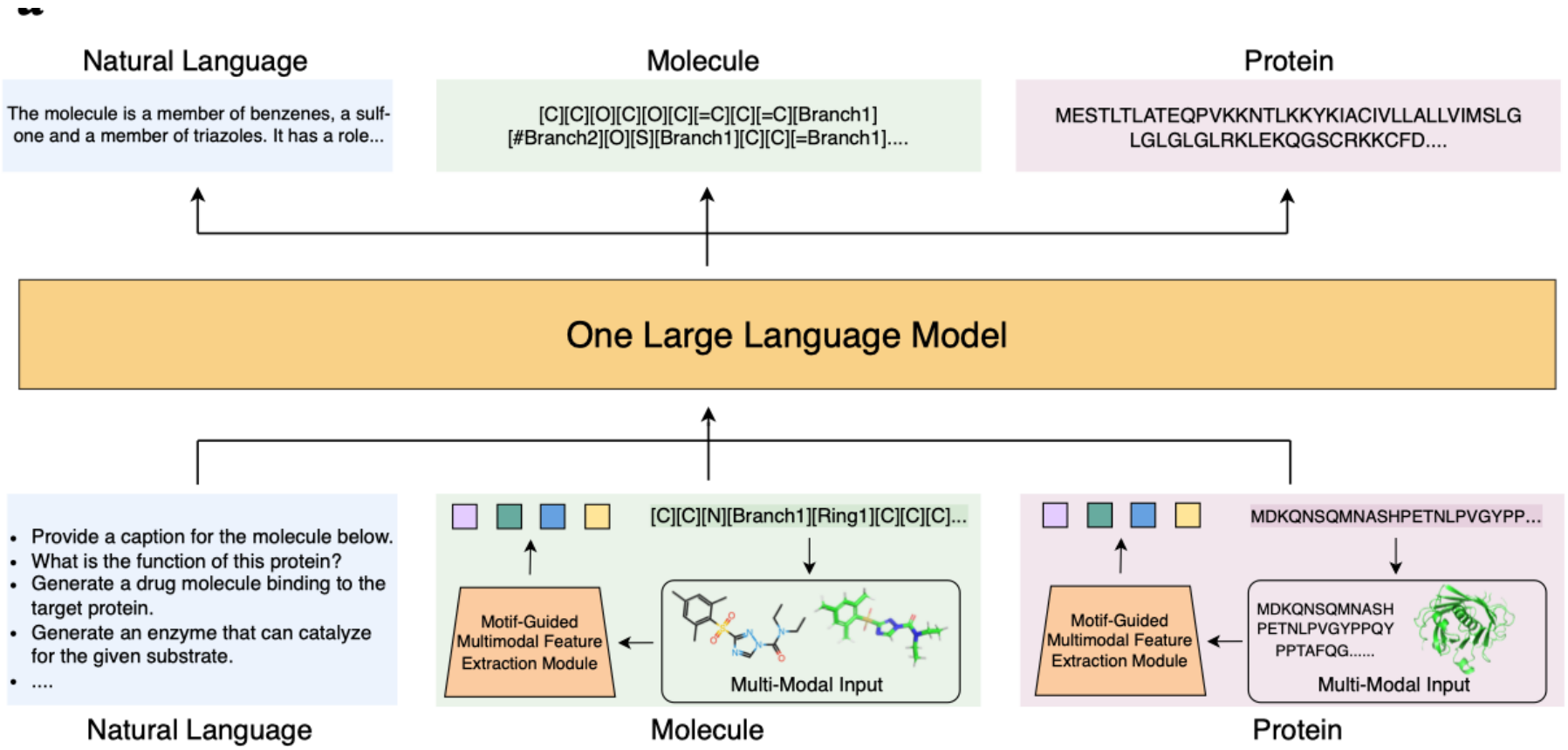
3 Challenges and Perspective

Comprehensive-Sci-LLM



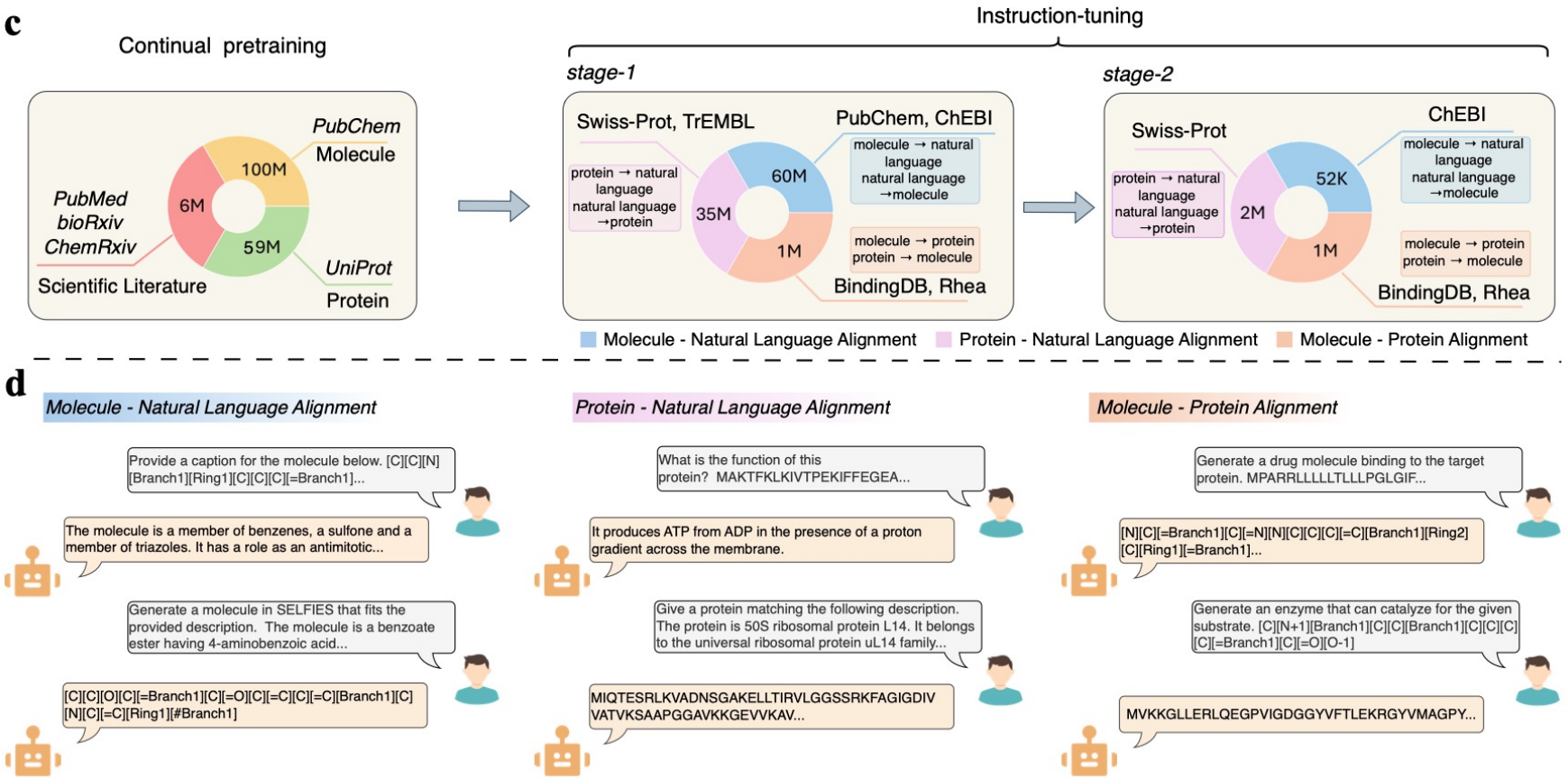
Mol-Instructions - A Large-Scale Biomolecular Instruction Dataset for Large Language Models, ICLR 2024

Comprehensive-Sci-LLM



Advancing biomolecular understanding and design following human instructions. Nature Machine Intelligence, 2025

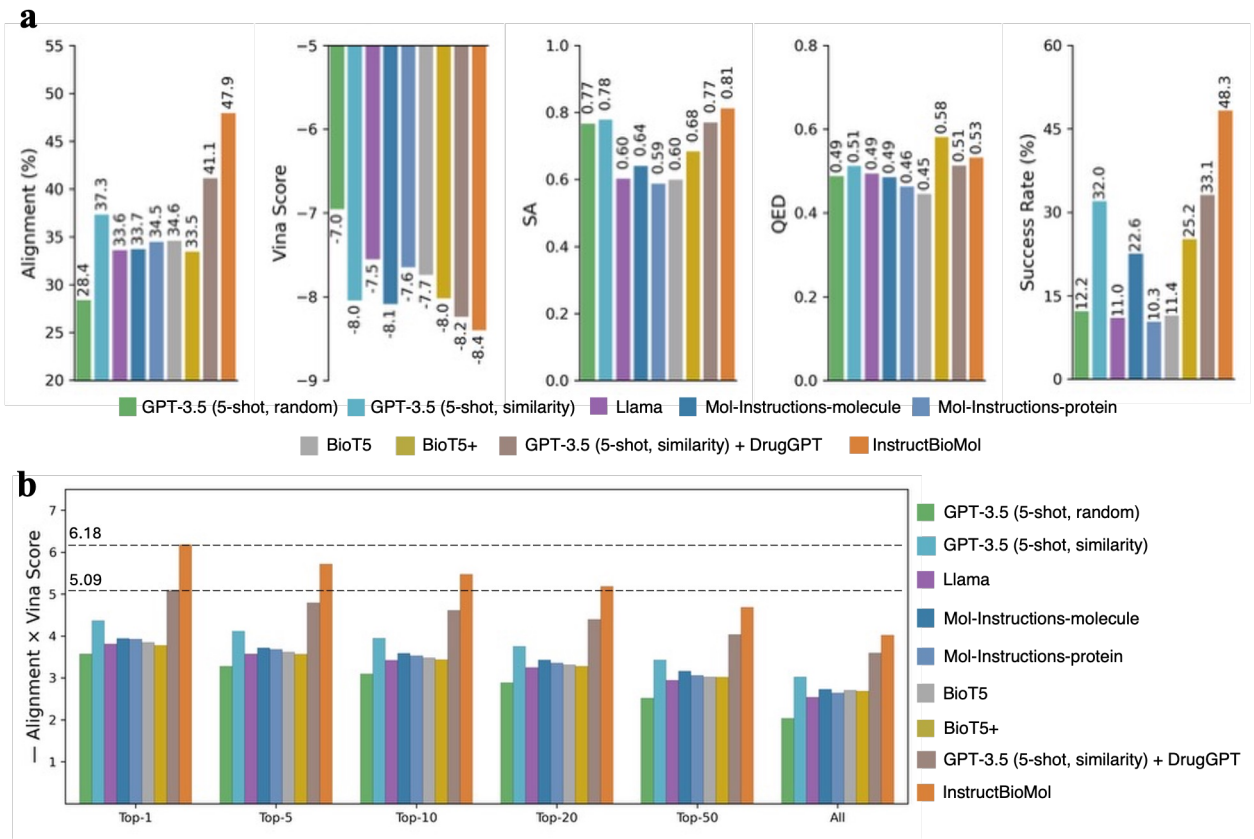
Comprehensive-Sci-LLM

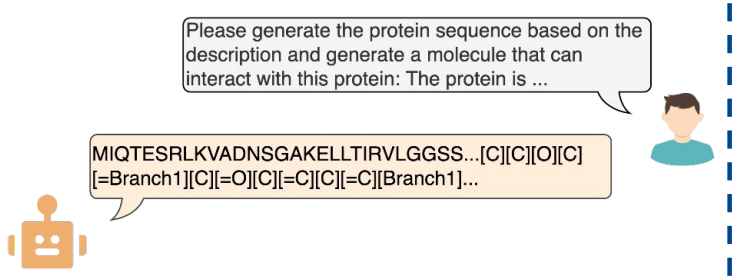


Advancing biomolecular understanding and design following human instructions. Nature Machine Intelligence, 2025

Comprehensive-Sci-LLM

InstructBioMol is the first to enable protein–small molecule co-design based on text instructions.



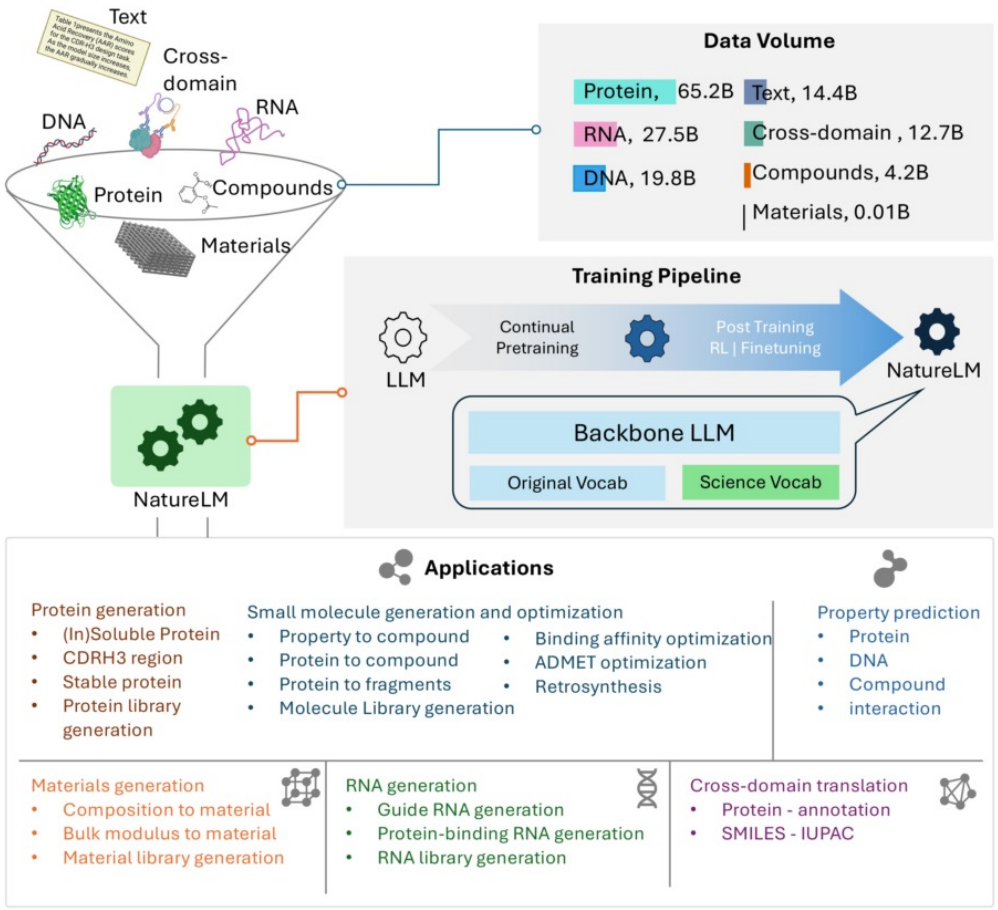


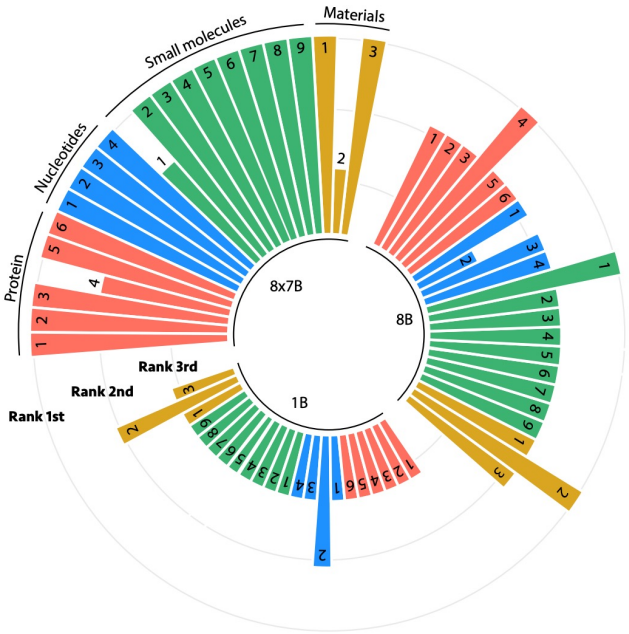
The overall success rate of co-design improved by 15%.

The top 20 results still slightly outperform the best-performing baseline's highest score.

Advancing biomolecular understanding and design following human instructions. Nature Machine Intelligence, 2025

Comprehensive-Sci-LLM





Protein
1. (In)Soluable protein generation
2. CDRH3 design
3. Protein annotation generation
4. Protein property prediction
5. Stable protein generation
6. Unconditional protein generation

Nulceotides
1. DNA property prediction
2. Guide RNA generation
3. Protein to RNA generation
4. Unconditional RNA generation

Small molecules
1. ADMET optimization
2. Binding affinity optimization
3. Compound property prediction
4. Property to compound
5. Protein to complete compound
6. Protein to fragments
7. Retrosynthesis
8. SMILES from-to IUPAC
9. Unconditional compound generation

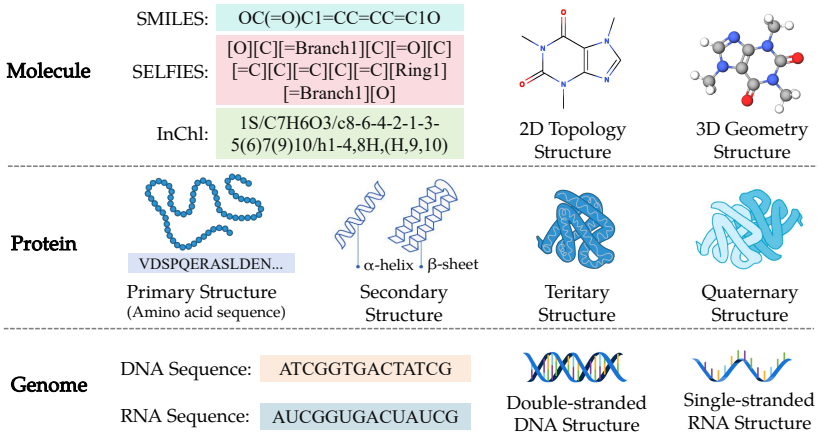
Materials
1. Unconditional material generation
2. Bulk modulus to material
3. Composition to material

Nature Language Model: Deciphering the Language of Nature for Scientific Discovery, 2025

Outline

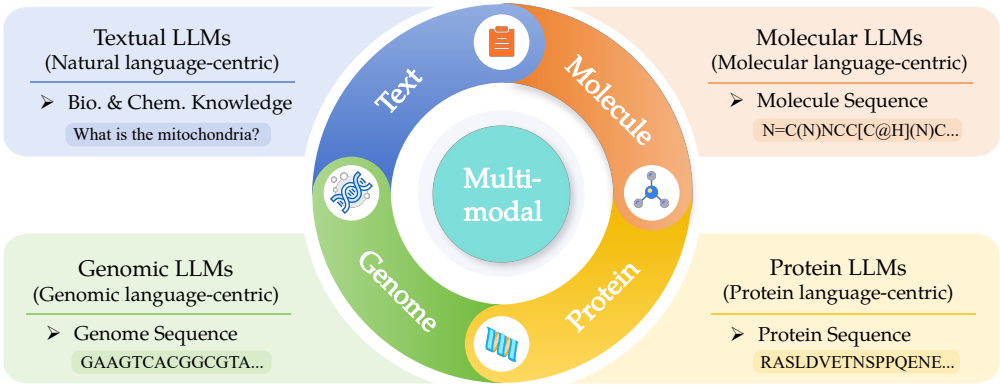
Sci-LLM: Summary

Scientific Symbols & Language



Modeling

Scientific Language Models



Perspective

- **Training Data**:
  - Scale of Pre-training Datasets
  - Quality of Finetuning Datasets
  - Lack of Cross-modal Datasets
- **Model Evaluation**:
  - Computational vs wet-lab
- **Architectures and Learning Objectives**:
  - Handling Longer Sequences
  - Incorporating 3D Structural Information
  - Autoregressive Learning Objective
- **Security and Ethics**:
  - Data Privacy, Model Bias, Equal Access

Relevant Materials

- Accompanying survey of this tutorial:
  - Scientific Large Language Models: A Survey on Biological & Chemical Domains, https://arxiv.org/pdf/2401.14656
  - Github Repository: https://github.com/HICAI-ZJU/Scientific-LLM-Survey
- Surveys for related topics:
  - Comprehensive
    - Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems, https://arxiv.org/pdf/2307.08423
    - A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery, https://arxiv.org/abs/2406.10833
  - Chemical molecules
    - MolGenSurvey: A systematic survey in machine learning models for molecule design, https://arxiv.org/abs/2203.14500
    - A Systematic Survey of Chemical Pre-trained Models, https://www.ijcai.org/proceedings/2023/0760.pdf
  - Biological proteins
    - Learning the protein language: Evolution, structure, and function, https://www.cell.com/cell-systems/pdf/S2405-4712(21)00203-9.pdf
    - Protein language models and structure prediction: Connection and progression, https://arxiv.org/pdf/2211.16742
    - Learning functional properties of proteins with language models, https://www.nature.com/articles/s42256-022-00457-9

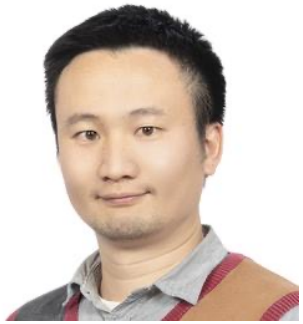
Thank You!

Mr. Xiang Zhuang
zhuangxiang@zju.edu.cn

Part III: Integrating KGs and LLMs for Scientific Applications

Zaiqiao Meng
zaiqiao.meng@glasgow.ac.uk

About Me



Dr. Zaiqiao Meng

https://mengzaiqiao.github.io/

Zaiqiao.Meng@glasgow.ac.uk

https://www.ai4biomed.org

- UK Lecturer (Assistant Professor) at the School of Computing Science, **University of Glasgow** (2022.01 – now)
- Postdoctoral Researcher at the Language Technology Lab (LTL) of the **University of Cambridge** (2020.07 – 2022.01)
- Postdoctoral Researcher at the IR Group of the **University of Glasgow** (2019.03 – 2020.07)
- Ph.D. degree in Computer Science from **Sun Yat-sen University** (2018)

Research Topic

- Information Retrieval
- Knowledge Graphs
- Large Language Models
- LLM-based Agents
- AI for BioMedicine

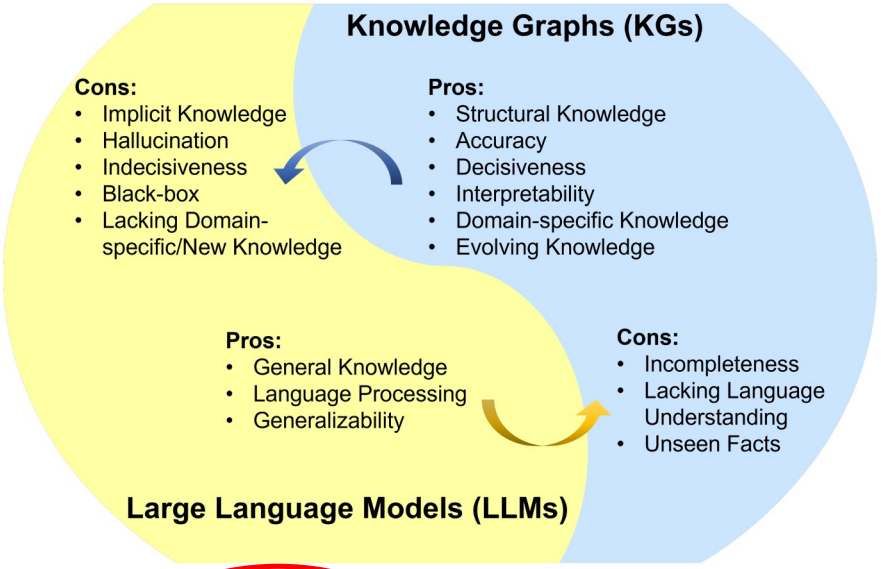
Outline

1. Knowledge Incorporation Frameworks
2. KG Integration for Scientific NLP Tasks
3. KG Integration for Scientific Prediction Tasks

Outline

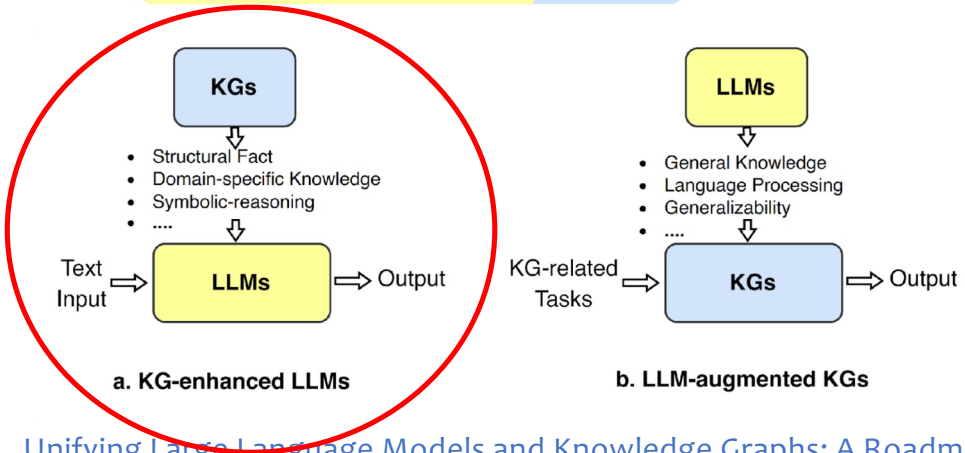
1. **Knowledge Incorporation Frameworks**
2. KG Integration for Scientific NLP Tasks
3. KG Integration for Scientific Prediction Tasks

KG-enhanced LLMs



KG-enhanced LLMs, which incorporate KGs during different phases of LLMs, or for the purpose of enhancing understanding of the knowledge learned by LLMs;

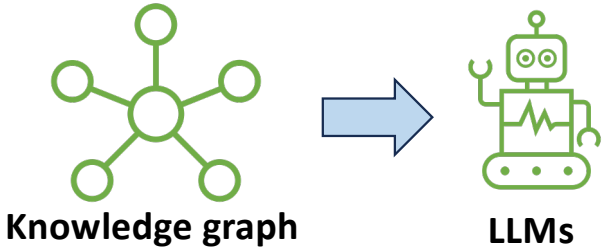
- Knowledge aware pretraining
- Knowledge integration finetuning
- Knowledge editing
- Knowledge unlearning



Unifying Large Language Models and Knowledge Graphs: A Roadmap. (2024) Link

Categorization over different stages

- Integrating scientific knowledge can occur at any stage in the development of LLMs



pre-training

- Pretrain-KGEs [1]
- ERNIE [2]
- MolXPT [3]

post-training

- SapBERT [3]
- MoP [4]
- OntoProtein [5]
- Bridge [6]
- MedLaSA [10]

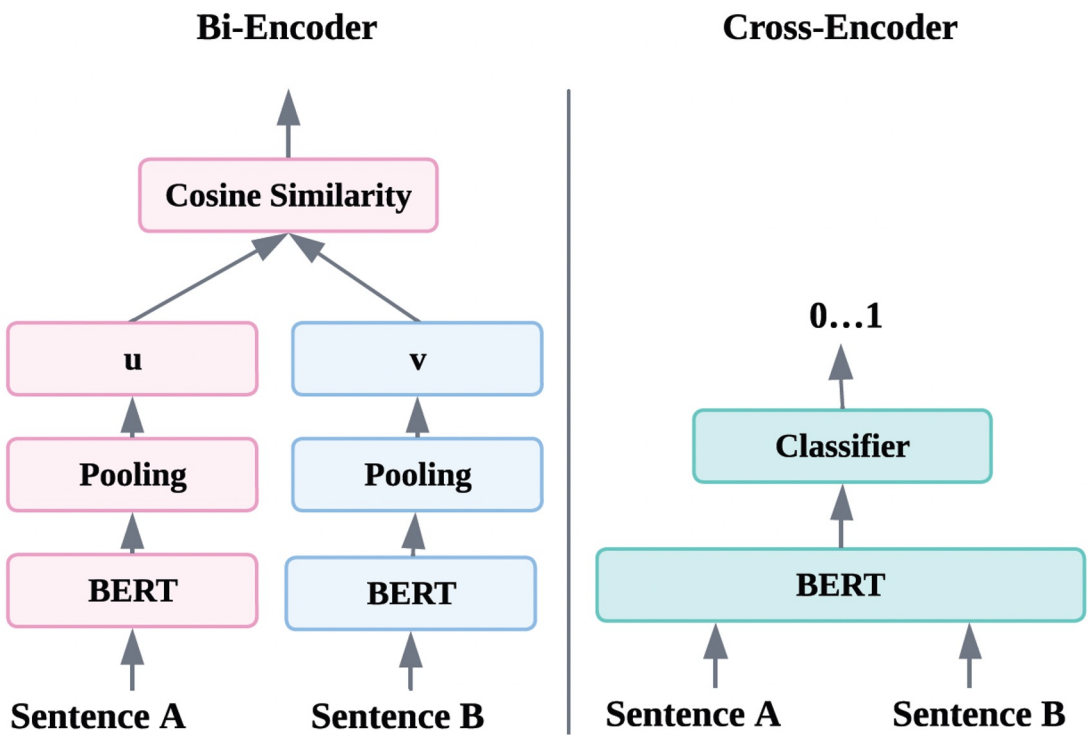
finetuning

- FusionDTI [8]
- FusionGDA [9]
- Saprot-based applications [13]

inference

- MKRAG [10]
- KRAGEN [11]
- BioRAG [12]

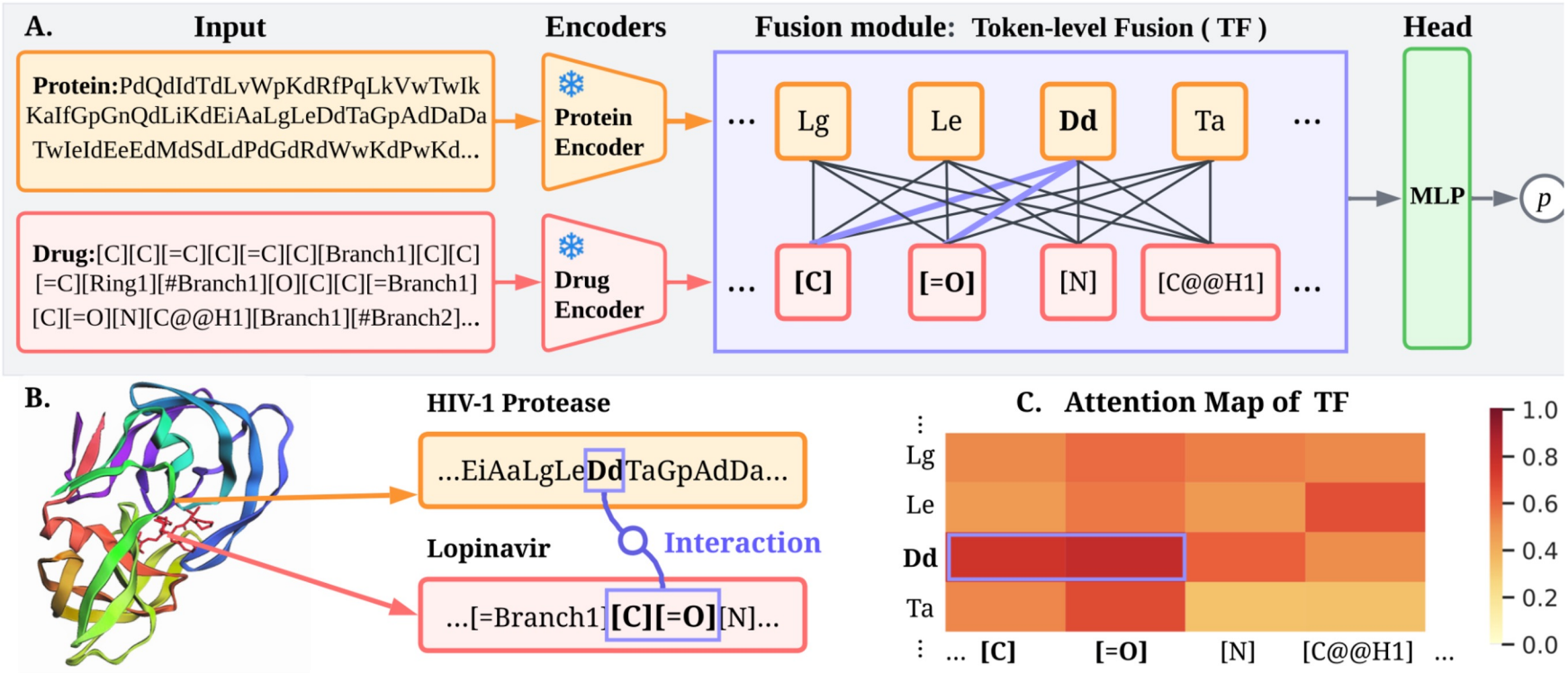
Bi-Encoder vs. Cross-Encoder



Bi-Encoder: Efficient encoding of individual entities can speed up retrieval and computation, but may sacrifice finer-grained interactions between different encoders.
High efficiency.

Cross-Encoder: Encodes entities jointly, capturing more detailed interactions, but at the cost of greater computational resources and time.
High effectiveness.

Bi-Encoder for Drug-Target Prediction

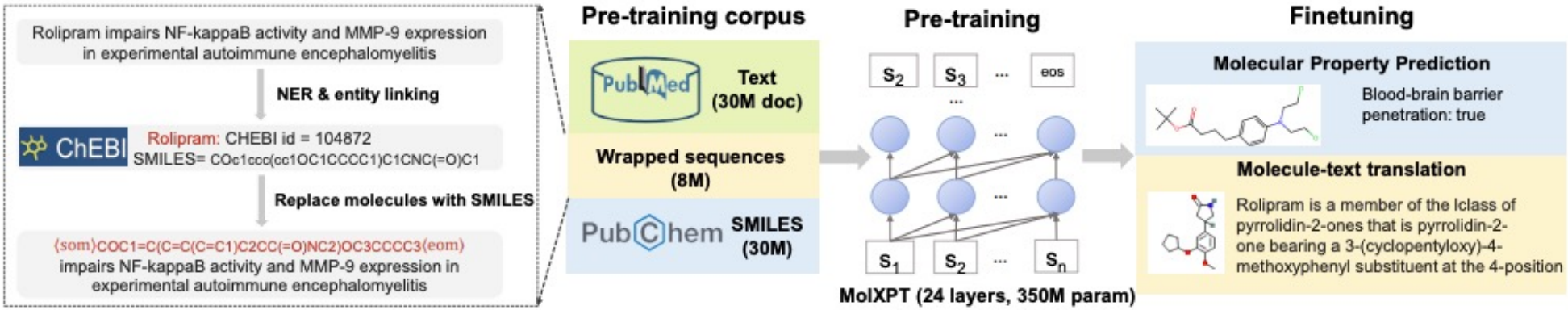


FusionDTI uses a token-level fusion module to effectively learn fine-grained information for drug-target interaction.

FusionDTI: Fine-grained Binding Discovery with Token-level Fusion for Drug-Target Interaction. Link

Cross-Encoder for molecular property prediction

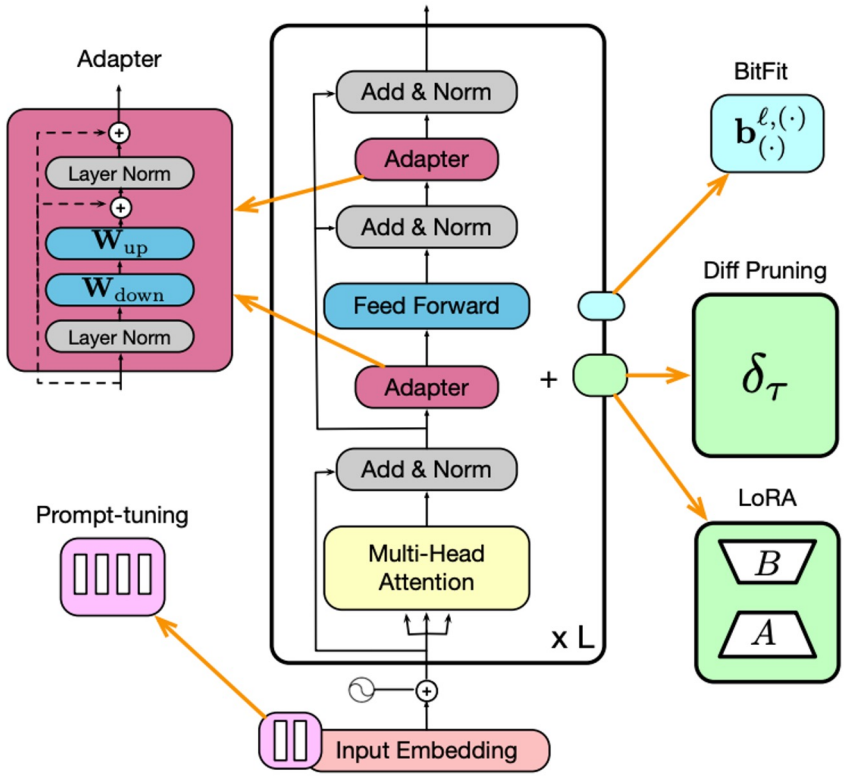
MolXPT: Wrapping Molecules with Text for Generative Pre-training



- A unified language model of text and molecules pre-trained on SMILES (a sequence representation of molecules) wrapped by text
- Text and SMILES are tokenized separately (molecular are encoded)

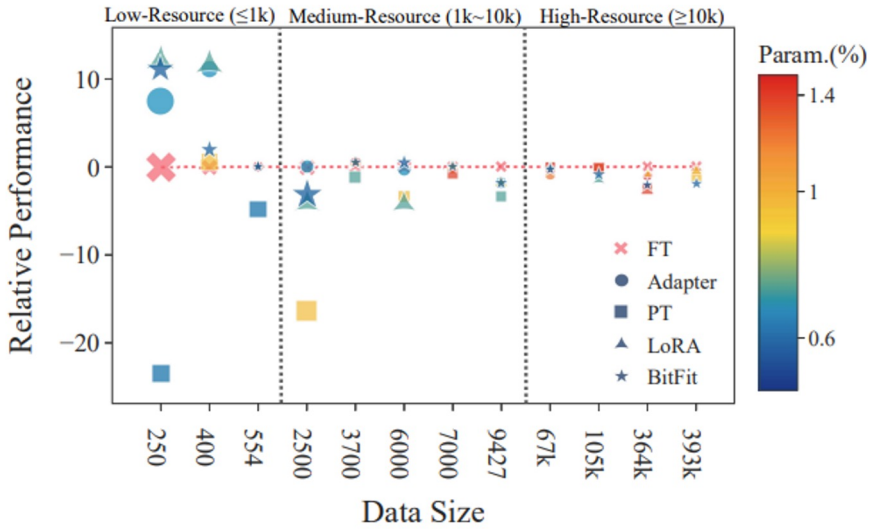
MolXPT: Wrapping Molecules with Text for Generative Pre-training (ACL 2023). Link

Integration Techniques of LLMs



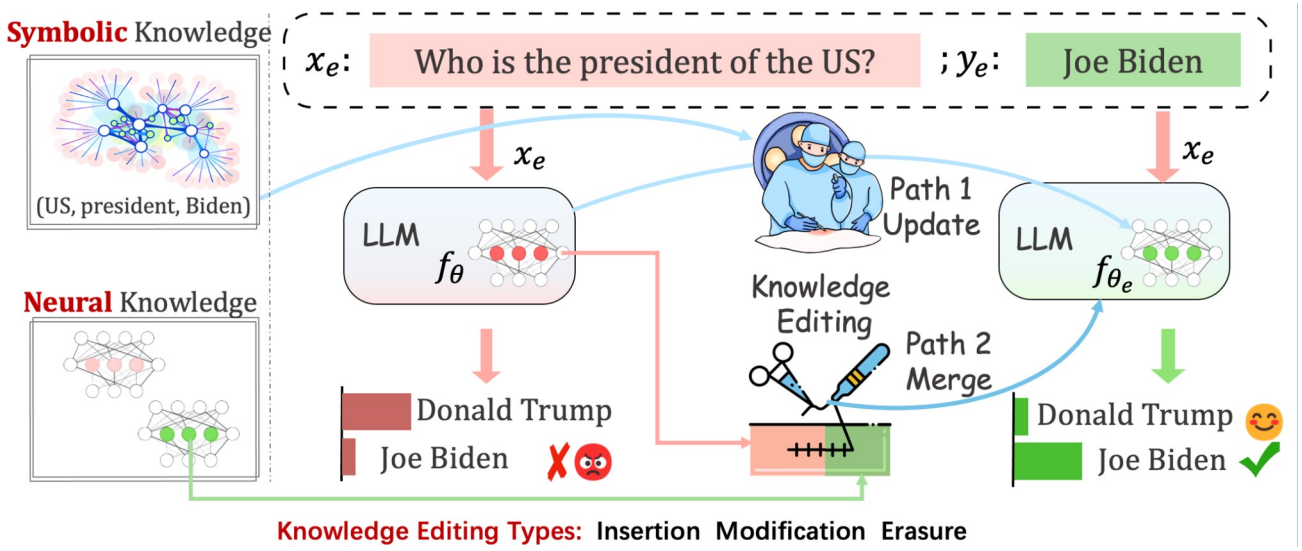
Parameter-Efficient Fine-Tuning (PEFT):

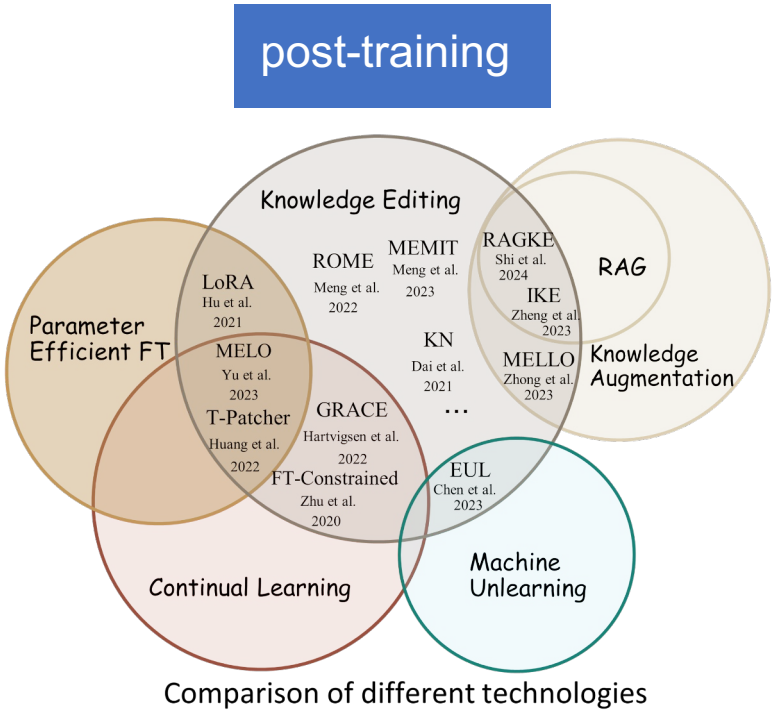
Techniques like **Adapters**, **Prefix Tuning, LoRA, Diff Pruning, BitFit or Prompt-tuning** that fine-tune only a small subset of model parameters, reducing computational costs while maintaining performance.



Revisiting Parameter-Efficient Tuning: Are We Really There Yet? (EMNLP 2022) Link

Knowledge Editing





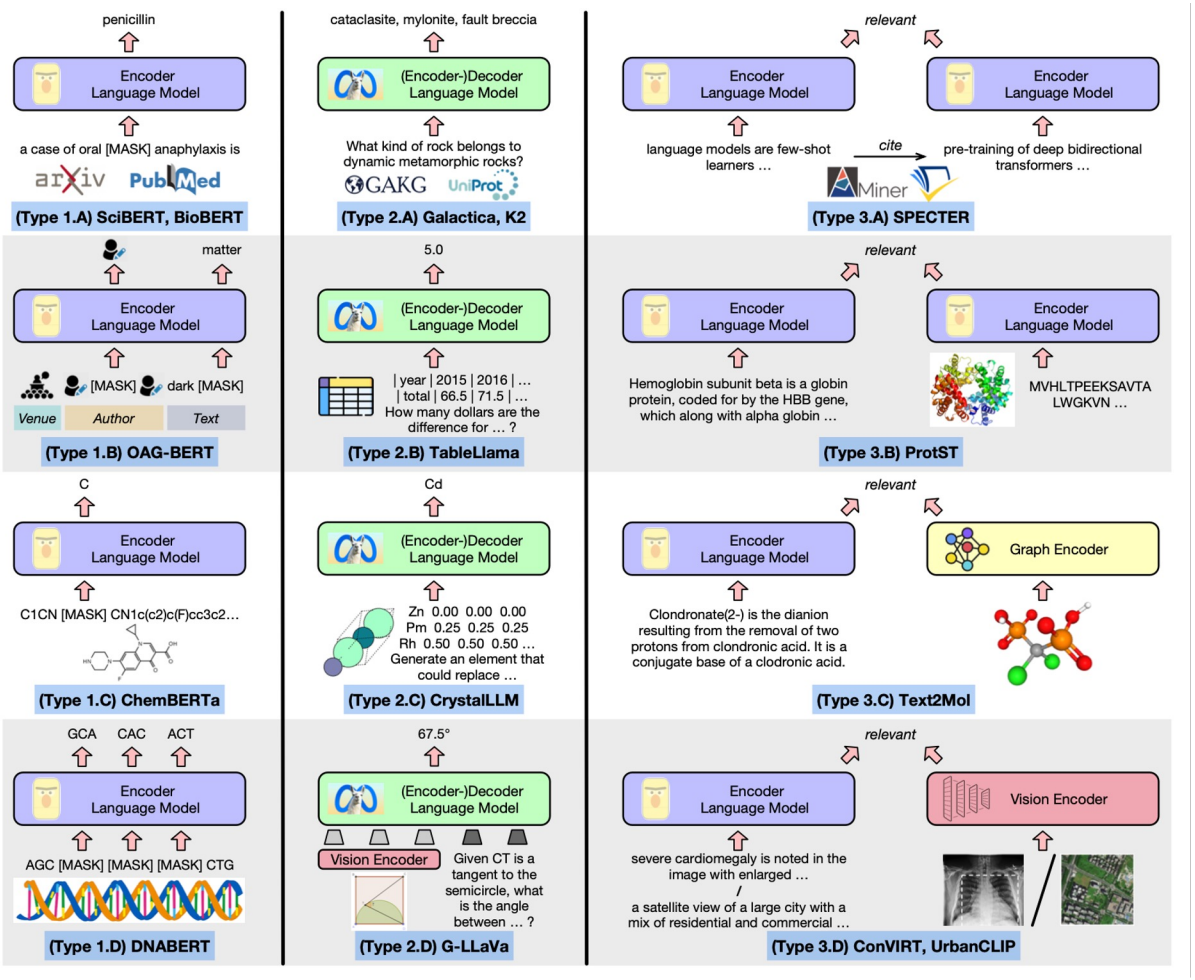
LLMs notoriously *hallucinate*, *perpetuate bias*, and *factually decay*, so we should be able to adjust specific behaviors of pre-trained models.

Easyedit: An easy-to-use knowledge editing framework for large language models: https://github.com/zjunlp/EasyEdit

Outline

1. Knowledge Incorporation Frameworks
2. **KG Integration for Scientific NLP Tasks**
3. KG Integration for Scientific Prediction Tasks

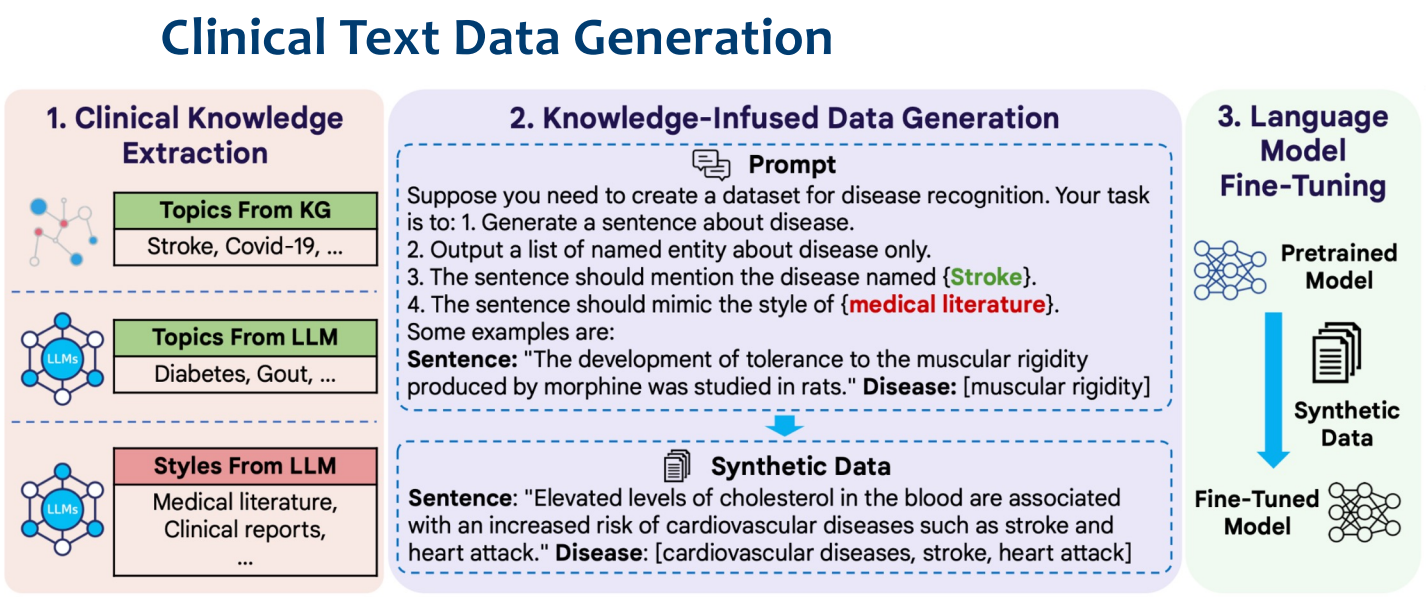
KG Integration for Scientific NLP Tasks



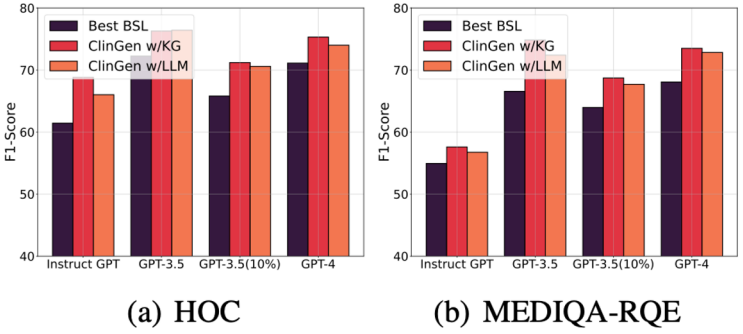
- Question Answering [5]
- Entity Linking [4]
- Document Classification [5]
- Summarisation/Note Generation [14]
- Hypothesis Generation [15]
- Knowledge Graph Construction and Completion [1]
- Reasoning [12, 15]

A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery, Arxiv 2024

KG Integration for Clinical Text Data Generation



fine-tune

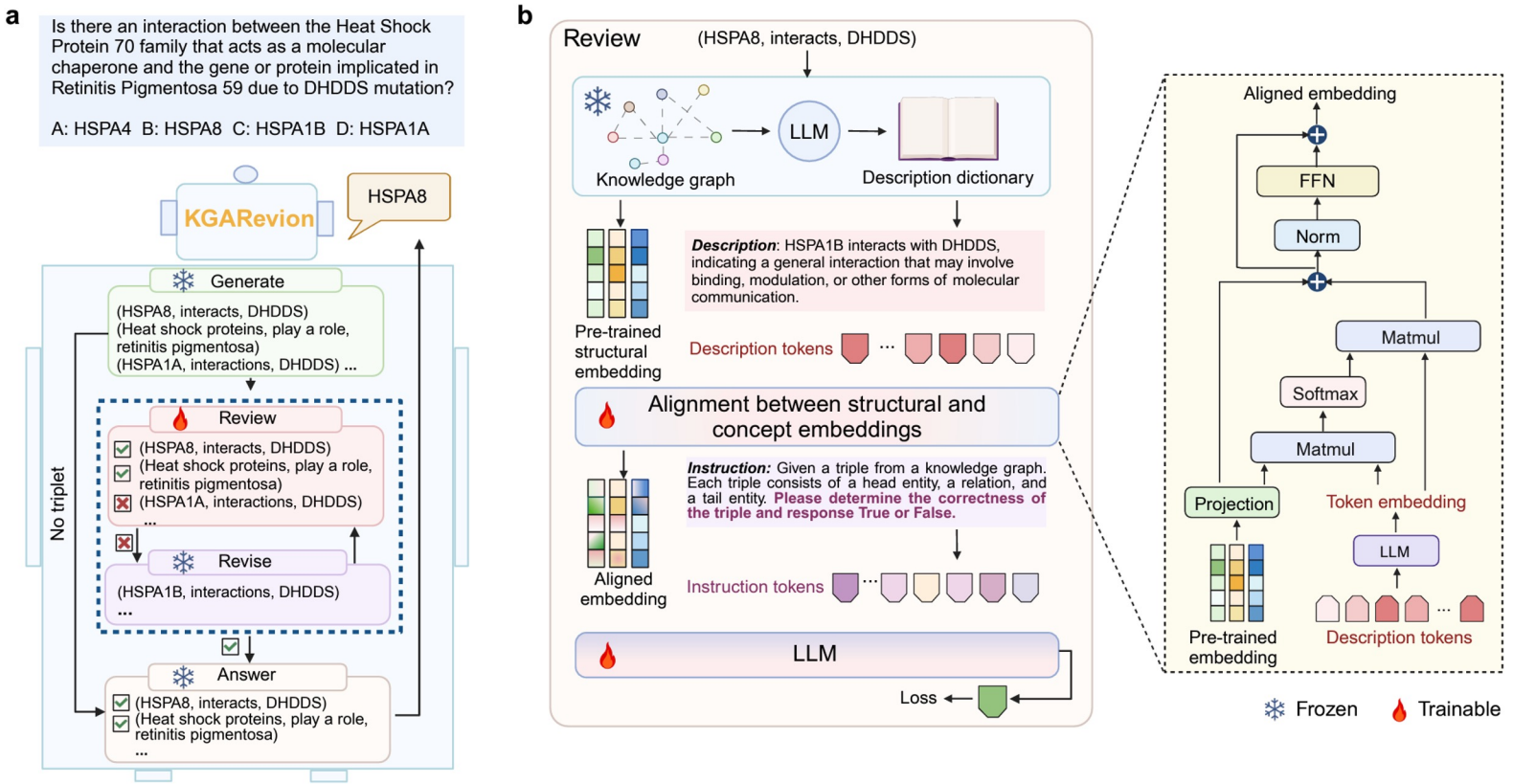


(a) HOC (b) MEDIQA-RQE

CLINGEN is a knowledge-informed framework for clinical data generation. This two-step methodology harnesses the emergent capabilities of LLMs and external knowledge from KGs to facilitate the synthesis of clinical data, even with few-shot examples only.

Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models. (ACL 2024) Link

KG Integration for QA Tasks



Question Answering (QA)
KG: Heterogeneous
(primeKG)

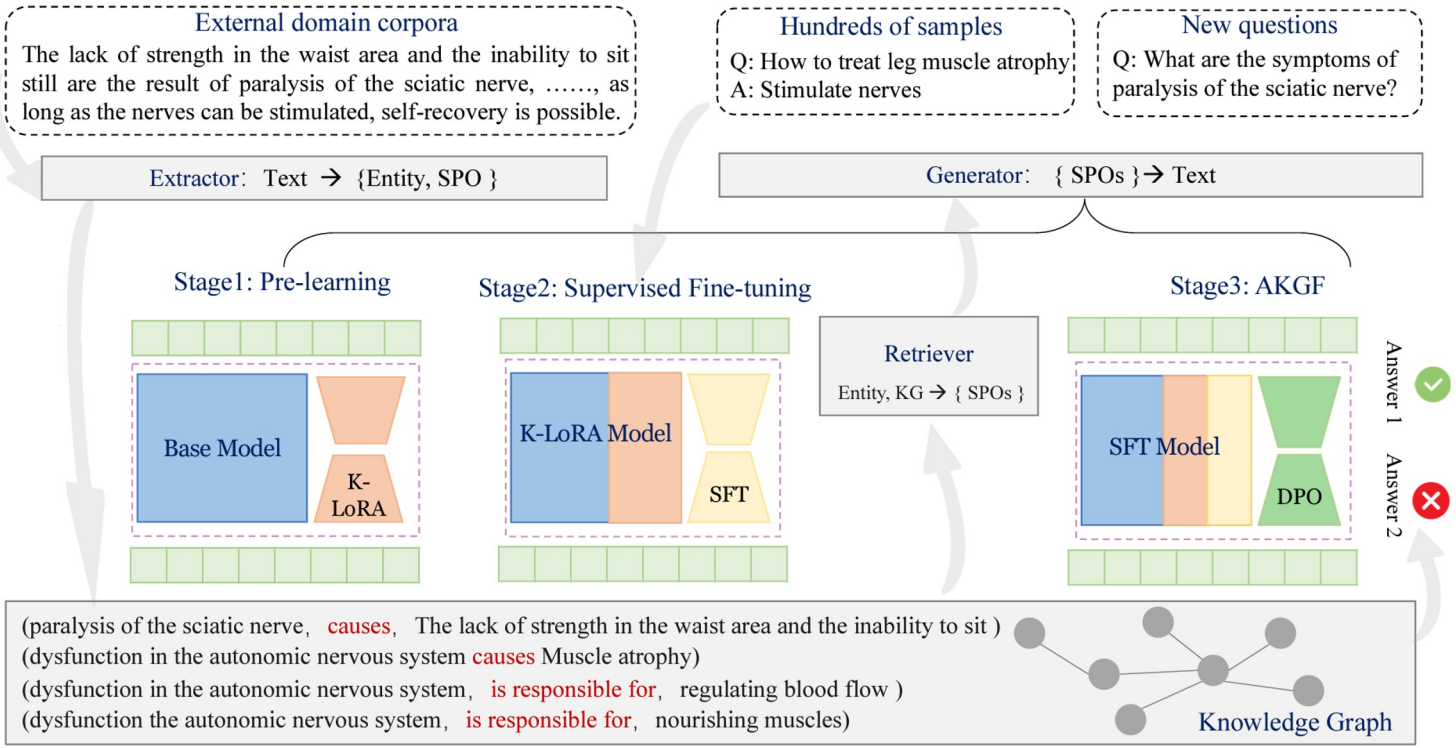
KGAREVION a KG-based LLM agent for complex medical QA that leverages non-codified knowledge of LLMs and structured, codified knowledge of medical concepts within KGs.

fine-tune

LORA *

Knowledge Graph Based Agent for Complex, Knowledge-Intensive QA in Medicine. (2024) Link

KG Integration for QA Tasks



Efficient Knowledge Infusion via KG-LLM Alignment. (ACL 2024) Link

Question Answering (QA)

The Enhanced LLM with Knowledge Pre-learning and Feedback (ELPF) framework can be divided into four main stages.

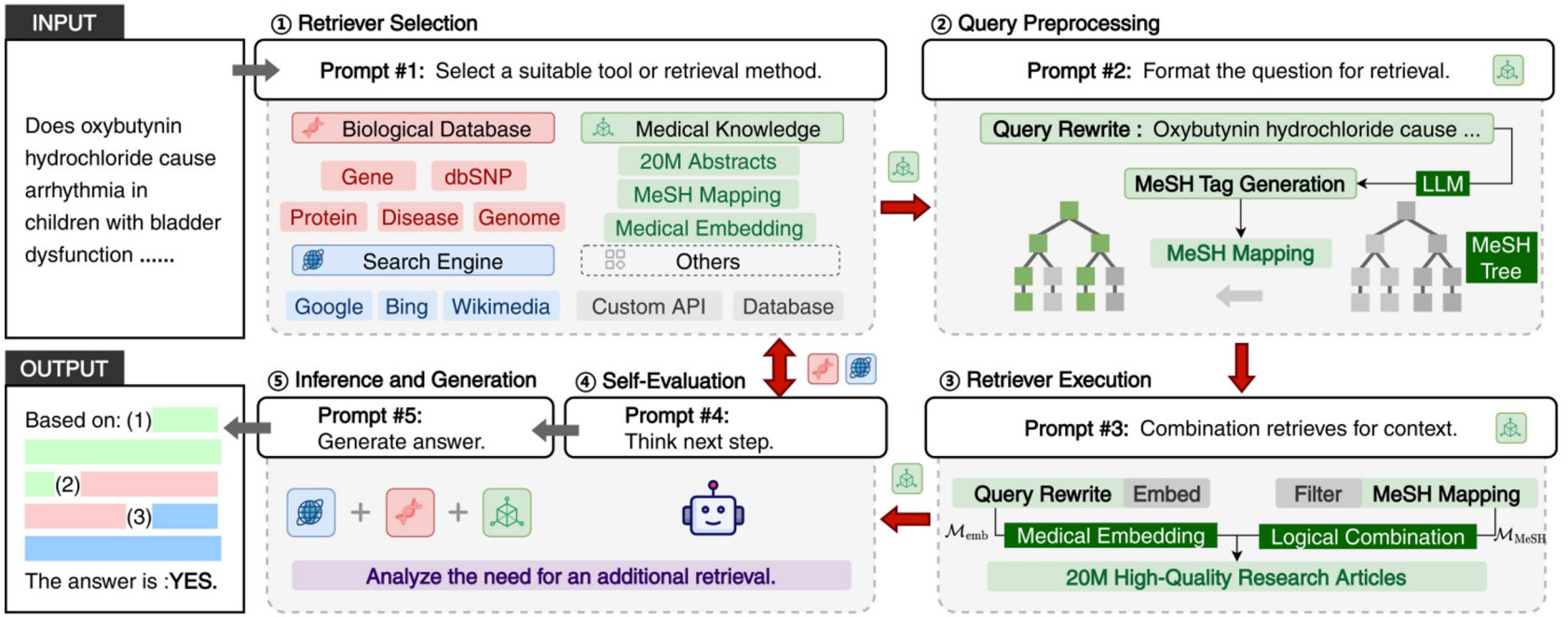
1) Efficient construction of domain KGs
2) Pre-learning with K-LoRA
3) SFT with KG retrieval
4) AKGF: KG acts as an evaluator

post-train+fine-tune

LORA *

RAG for Biological Question Reasoning

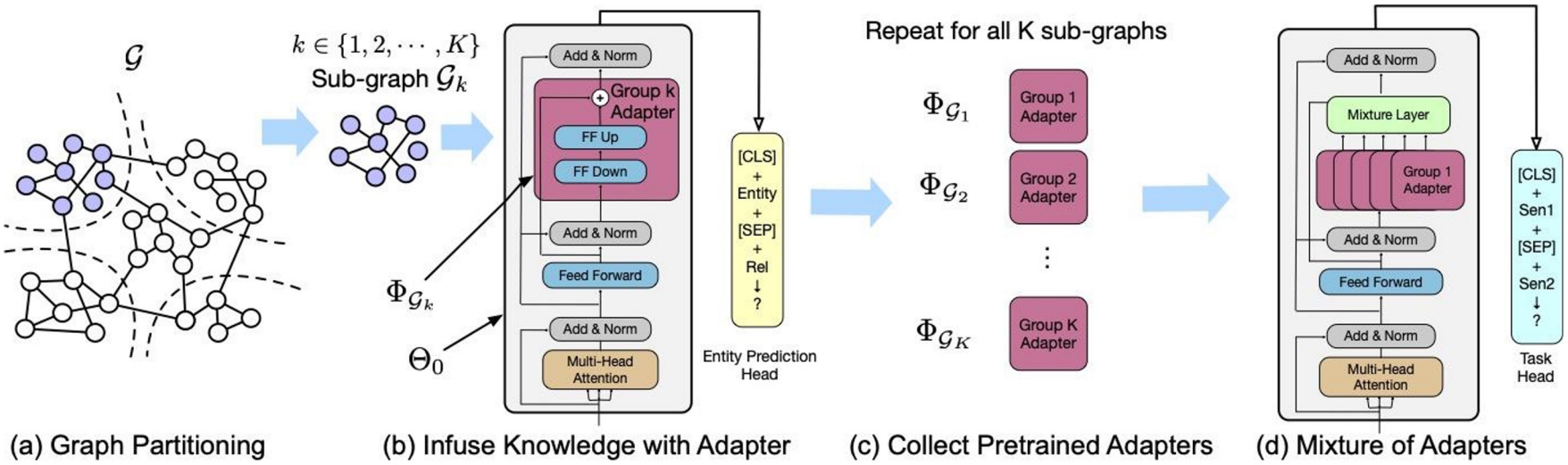
- BIORAG adaptively select knowledge source and domain-specific tools to advance the biology question-reasoning task.



Inference

BIORAG: A RAG-LLM Framework for Biological Question Reasoning, Arxiv 2024

Dealing with large scale knowledge graphs



- Partitioning it into smaller sub-graphs, e.g. METIS
- Infusing their specific knowledge into LLMs using lightweight adapters

post-train

adapter *

Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. (EMNLP 2021) Link

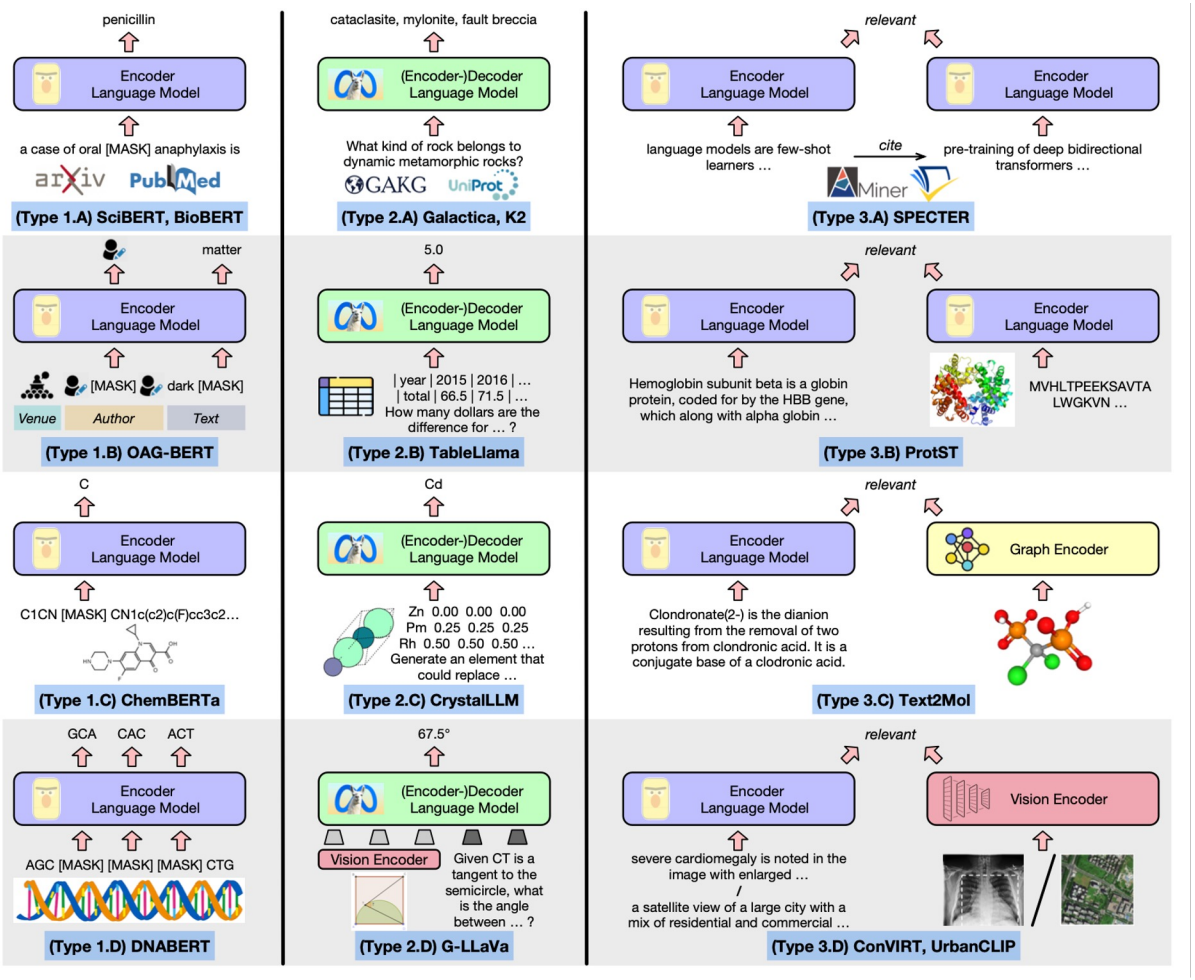
Outline

1 Knowledge Incorporation Frameworks

2 KG Integration for Scientific NLP Tasks

3 **KG Integration for Scientific Prediction Tasks**

KG Integration for Scientific Prediction Tasks



Gene-Disease Association (GDA) [19]

Protein Function Prediction [18]

Drug Repurposing [18]

Drug-Target Interaction (DTI) [8,17]

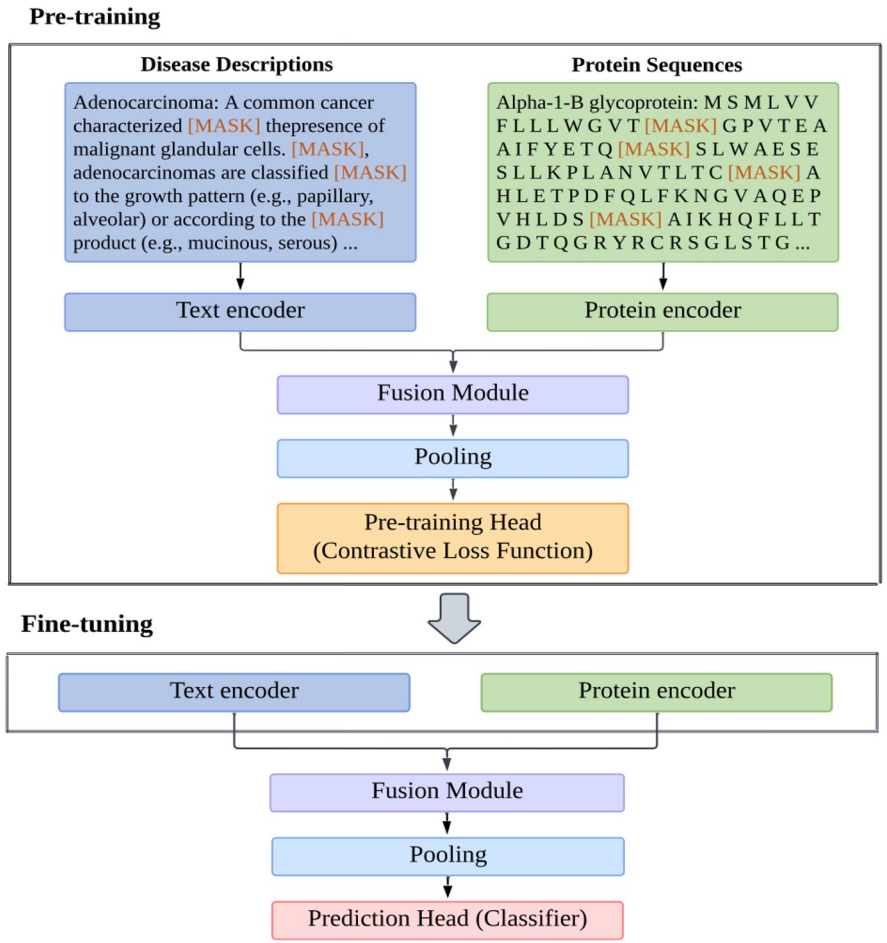
Text2Mol [3]

Amino acid contact prediction

…

A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery, Arxiv 2024

Gene-Disease Association (GDA)



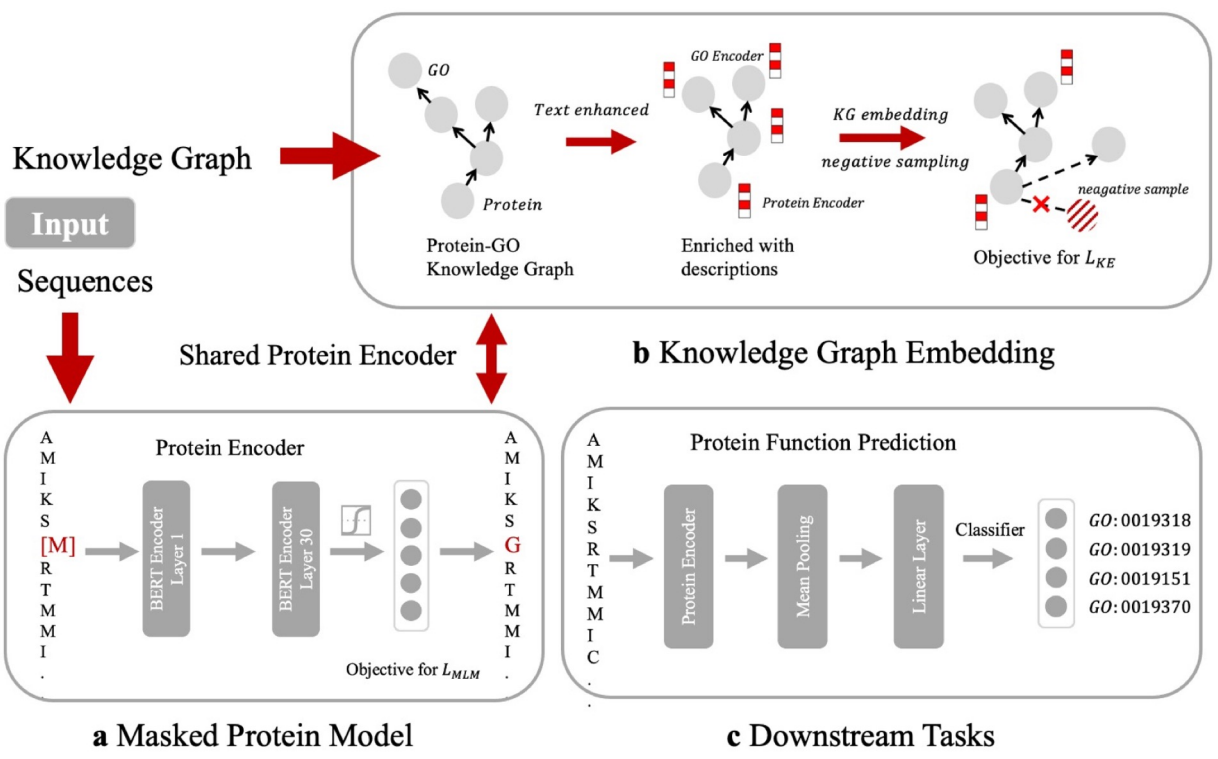
FusionGDA utilises bi-encoder with a fusion module to enrich the gene and disease semantic representations encoded by PLMs.

KG: Heterogeneous GDA
Encoder method: Bi-Encoder

post-train + fine-tune

Heterogeneous biomedical entity representation learning for gene–disease association prediction. Briefings in Bioinformatics (2024) Link

Protein Function Prediction

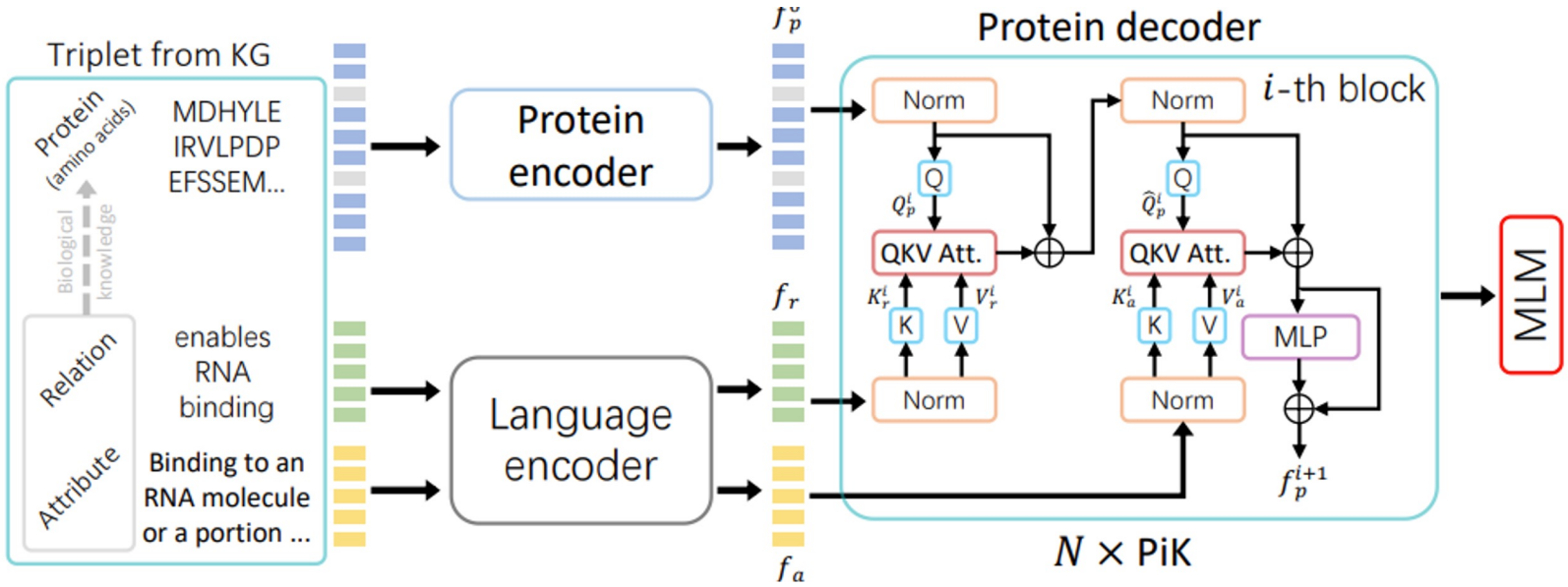


post-train + fine-tune

1. **OntoProtein** constructs a novel large-scale knowledge graph that consists of GO (Gene Ontology) and its related proteins, and gene annotation texts or protein sequences describe all nodes in the graph.
2. This KG was integrated by a novel contrastive learning with knowledge-aware negative sampling to jointly optimize the knowledge graph and protein embedding during pre-training

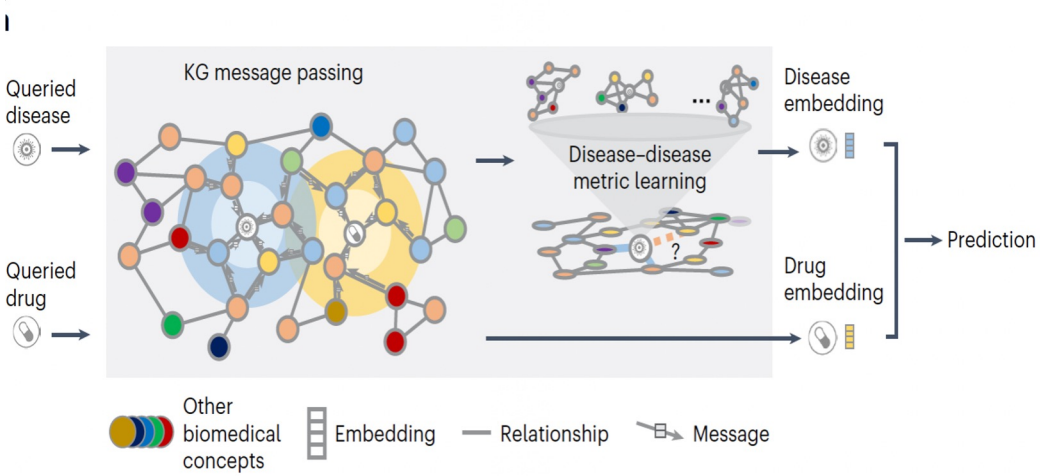
Ontoprotein: Protein pretraining with gene ontology embedding.(ICLR 2022) Link

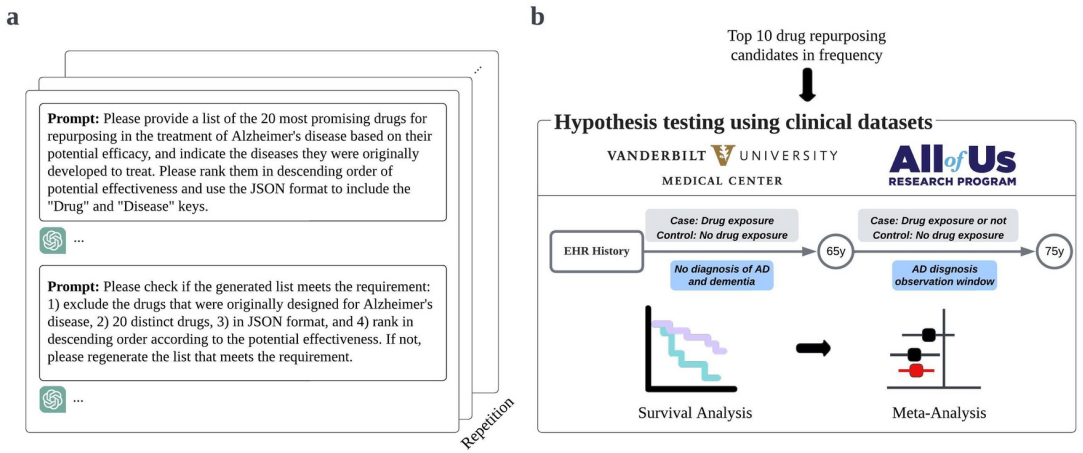
Amino acid contact prediction



- KeAP is trained on a knowledge graph that consists of about five million triplets from ProteinKG25
- KeAP explores knowledge graphs at a more granular level by applying cross-attention to sequences of amino acids and words from relation and attributes.
- KeAP can be trained using the MLM objective only (both contrastive loss and MLM are used in OntoProtein)

Drug repurposing





TxGNN, a graph foundation model for zero-shot drug repurposing, identifying therapeutic candidates even for diseases with limited treatment options or no existing drugs. Trained on a medical knowledge graph, TxGNN uses a ***graph neural network*** and ***metric learning module*** to rank drugs as potential indications and contraindications for 17,080 diseases.

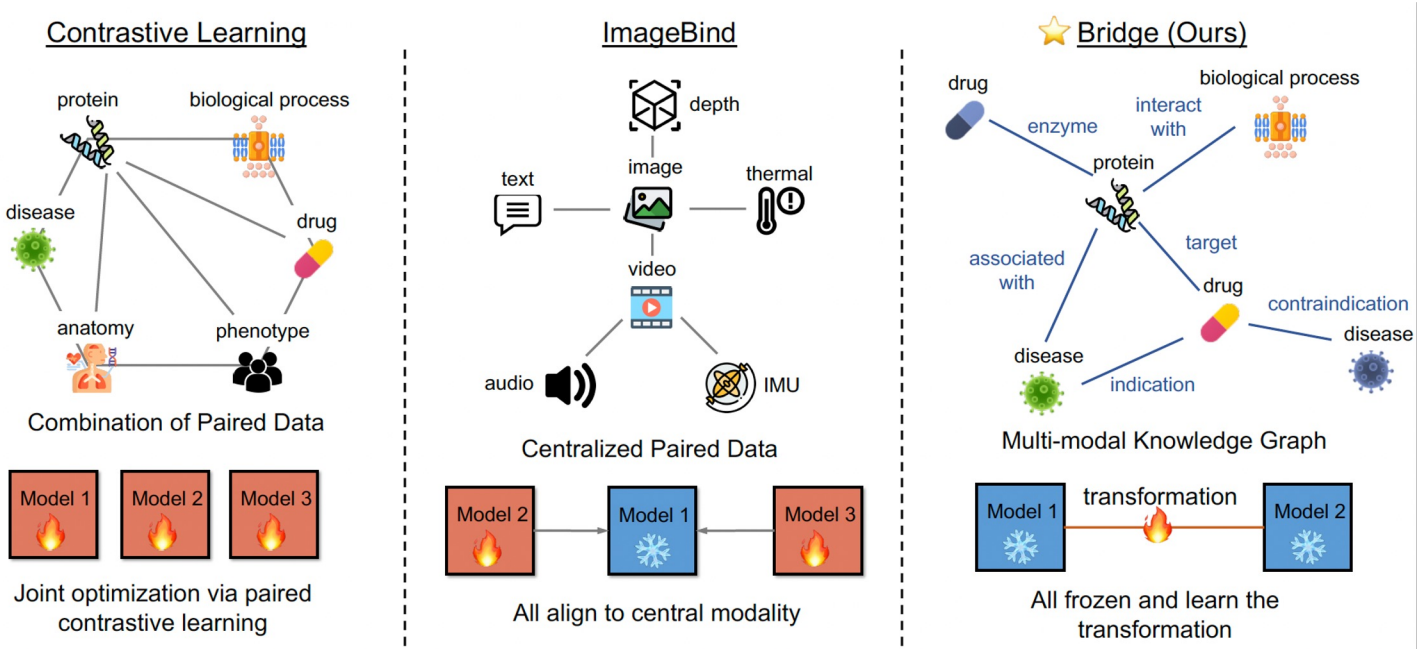
a Employing iterative queries of ChatGPT to recommend twenty drugs for AD repurposing.

b Evaluating the potential efficacy of the ten most frequently suggested drugs using electronic health records (EHR) data from two large clinical databases.

Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. **Nature, 2024**

A foundation model for clinician-centered drug repurposing. Nature Medicine, 2024. Link

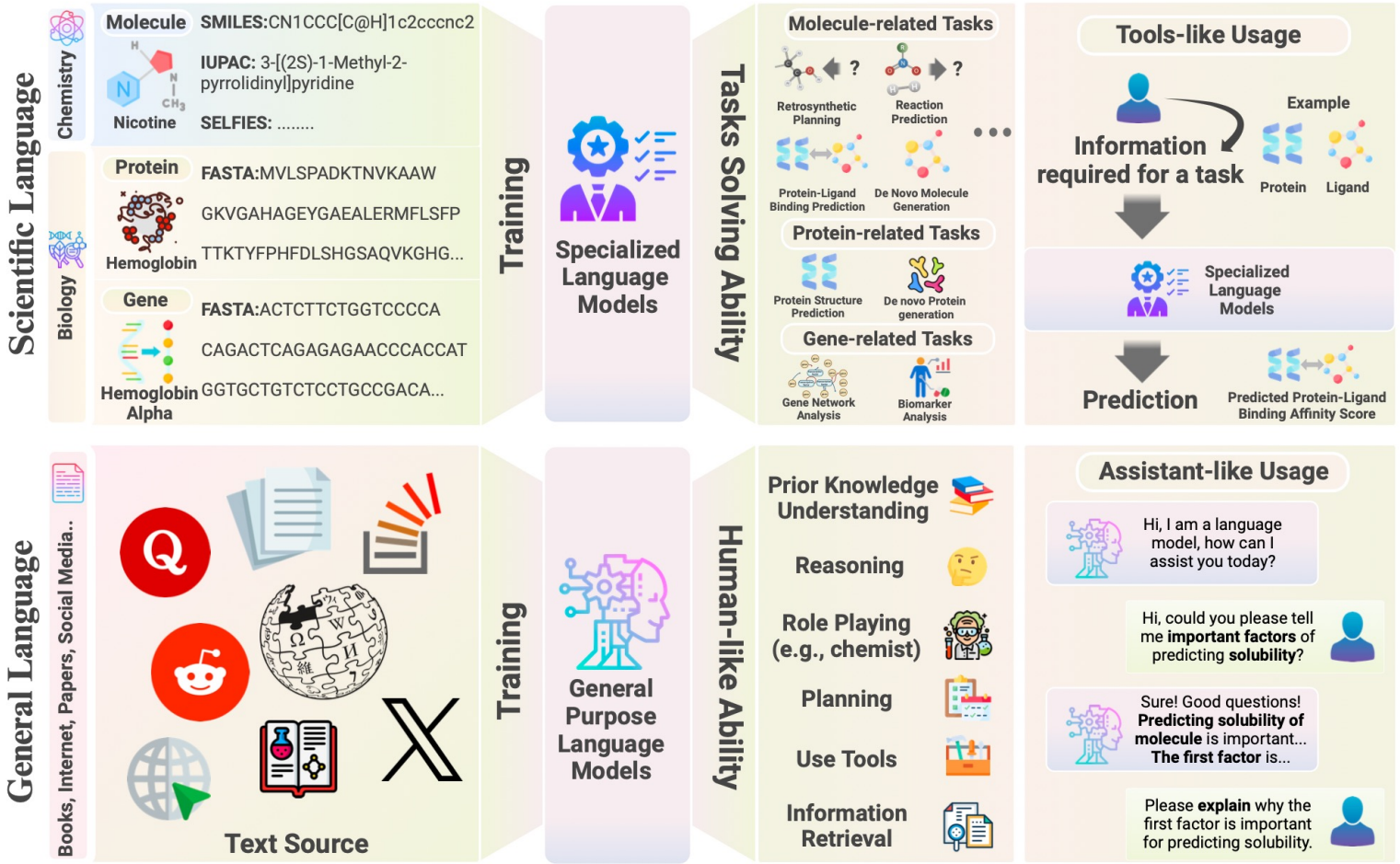
The comparison of cross-modality methods



1. **Multimodal contrastive learning**, e.g., CLIP, learns from a combination of paired data, updating all unimodal encoders.
1. **ImageBind** aligns all modalities with the central modality, with only the central model frozen.
1. **BioBRIDGE** (ICLR 2024) learns the transformation across modalities (Bridge Module) from a multi-modal KG, keeping all FMs frozen.

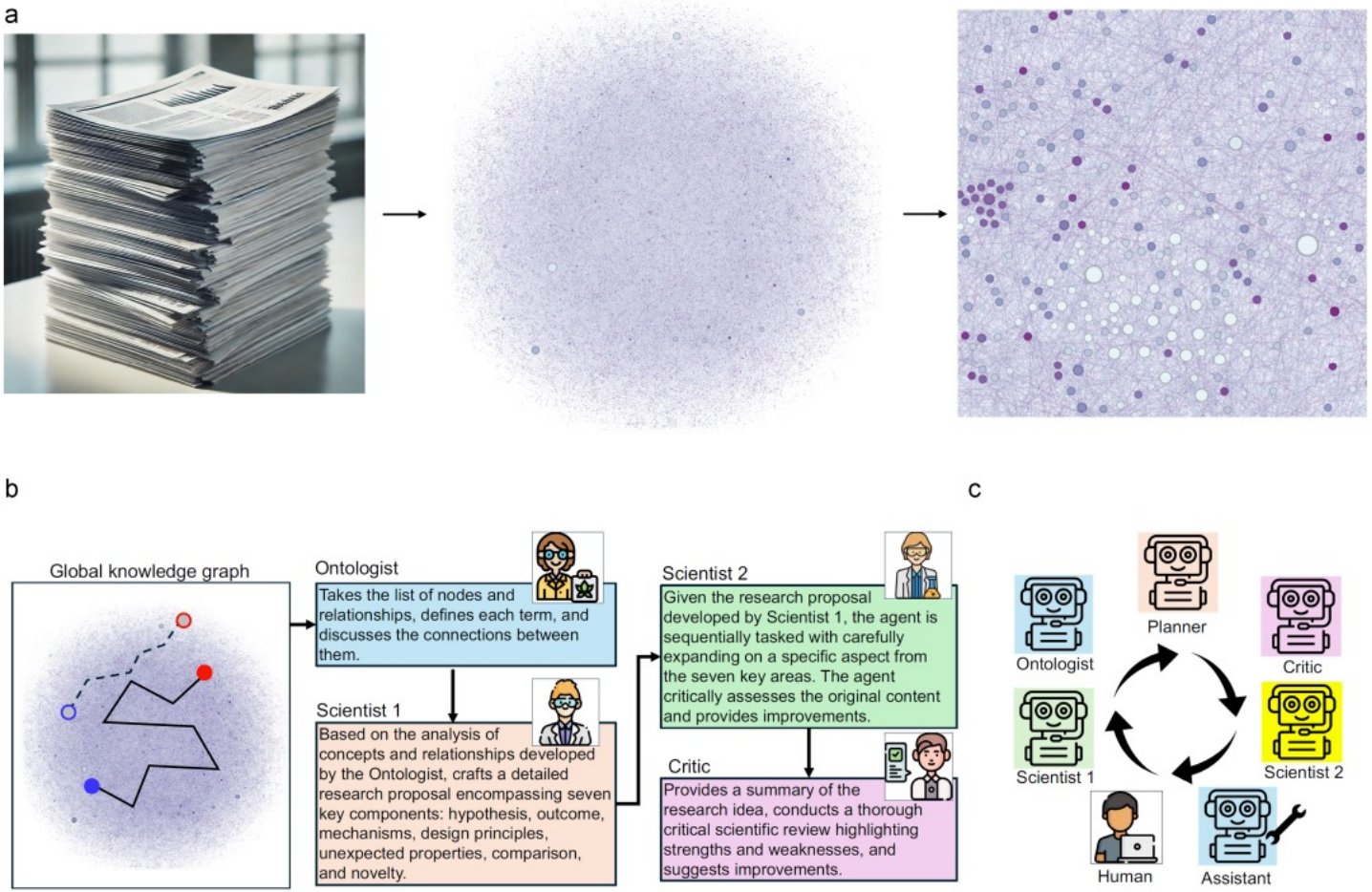
Imagebind: One embedding space to bind them all. (CVPR 2023) Link
Biobridge: Bridging biomedical foundation models via knowledge graph.(ICLR 2024) Link

Scientific Discovery Agent: Unifying Scientific NLP and Predictions



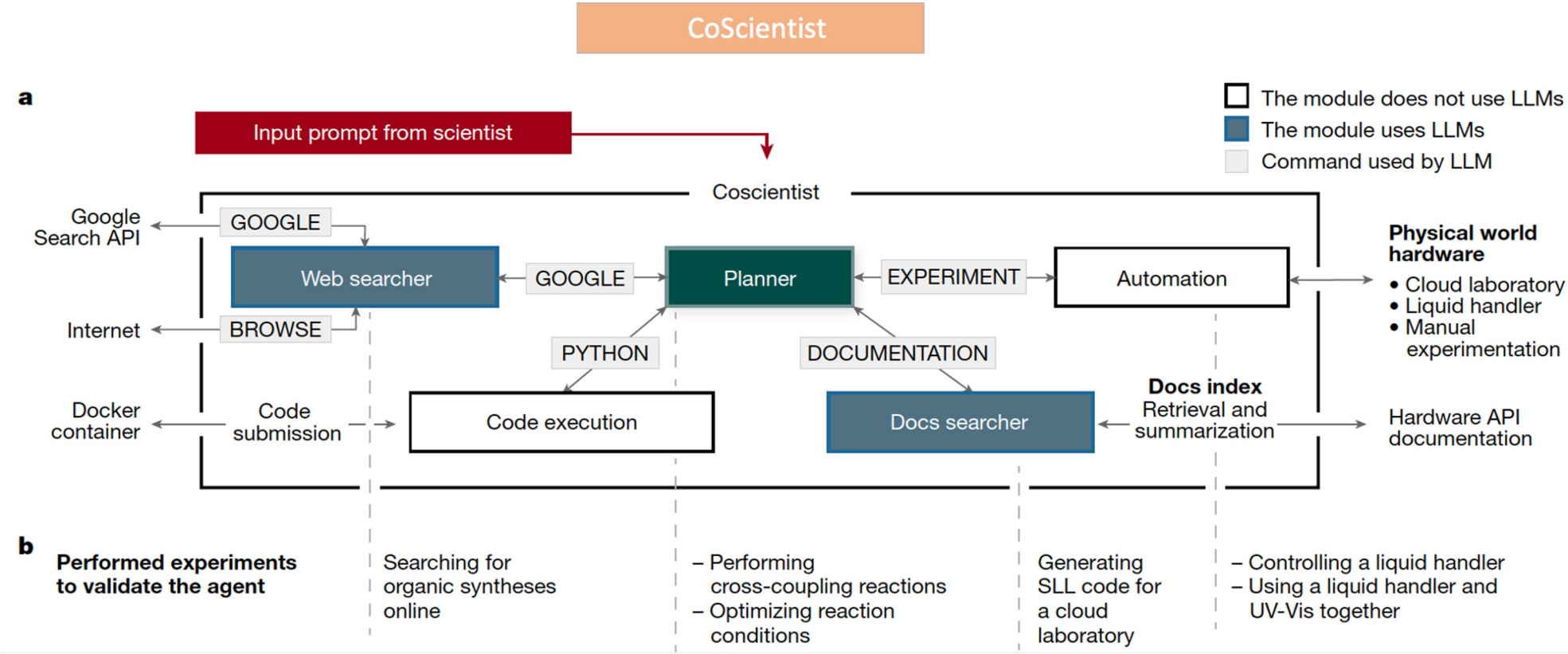
Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials, link

Multiple Agents with KGs for Scientific Discovery



SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning, link

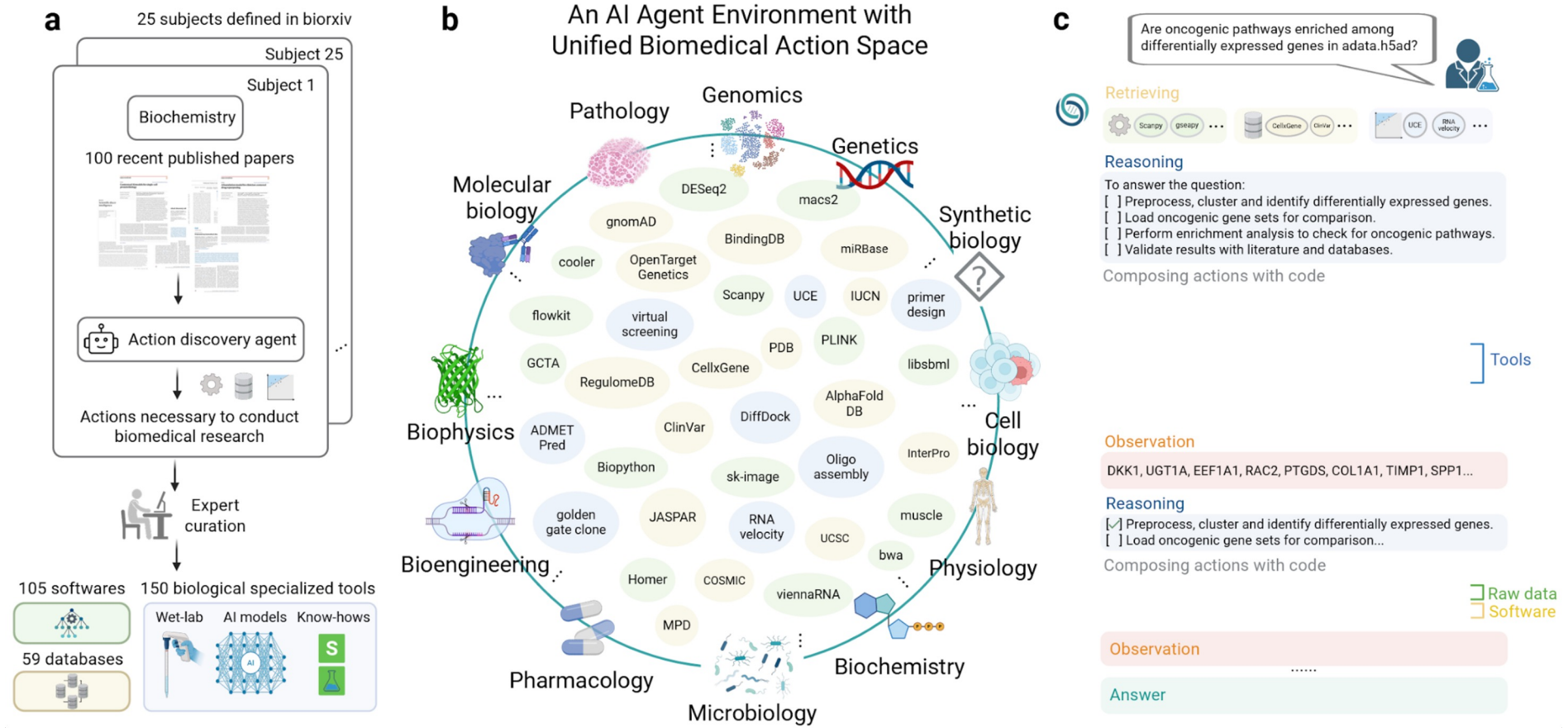
Coscientist: Chemistry - Unifying Physical World



Autonomous chemical research with large language models, **Nature**, 2023

Biomni: A General-Purpose Biomedical AI Agent

BioRxiv, June 2, 2025 [Link]

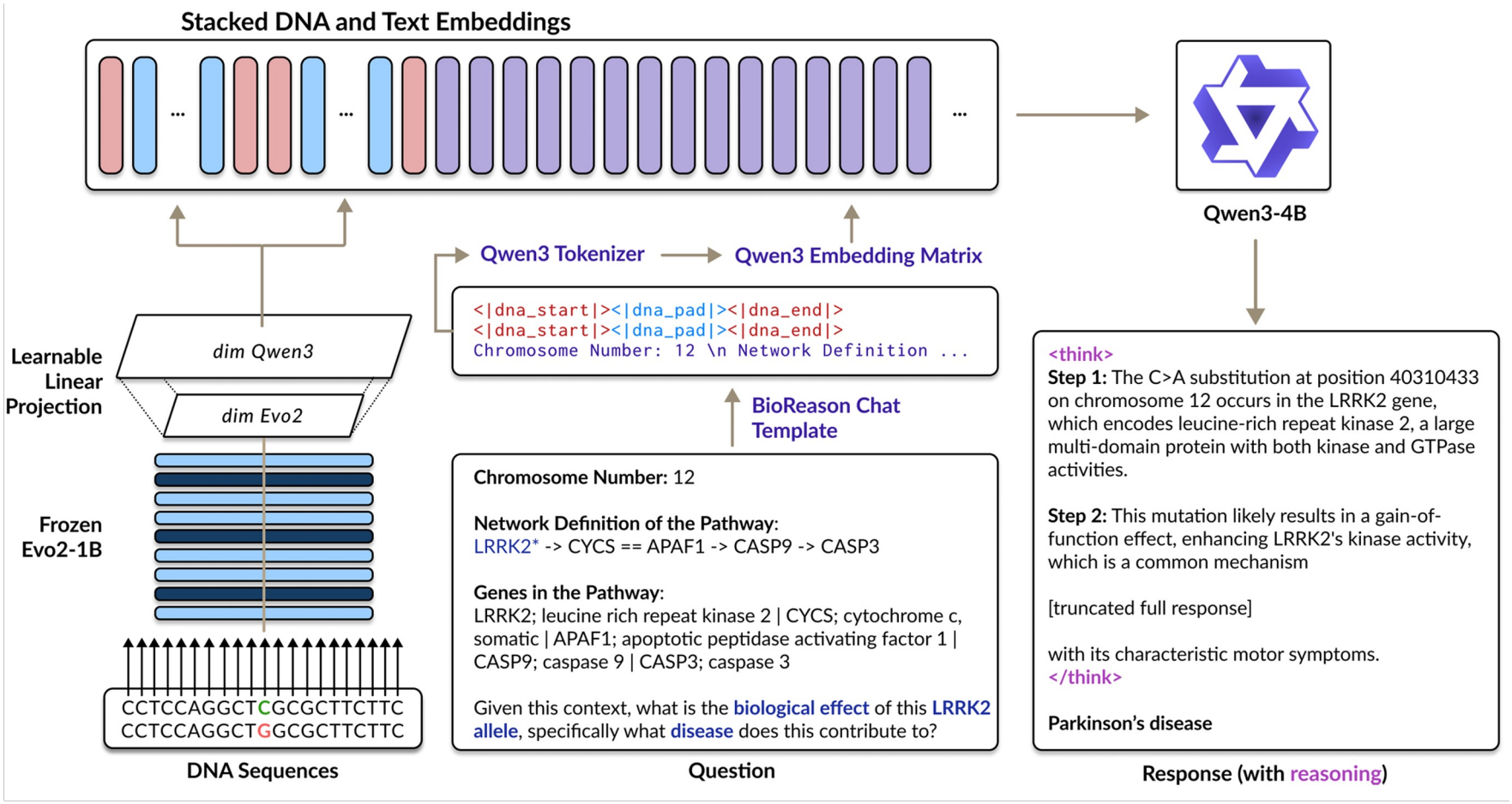


Biomni is ready to use at: https://biomni.stanford.edu

Stanford University · Genentech A Member of the Roche Group · Arc Institute · UCSF · Princeton University · University of Washington

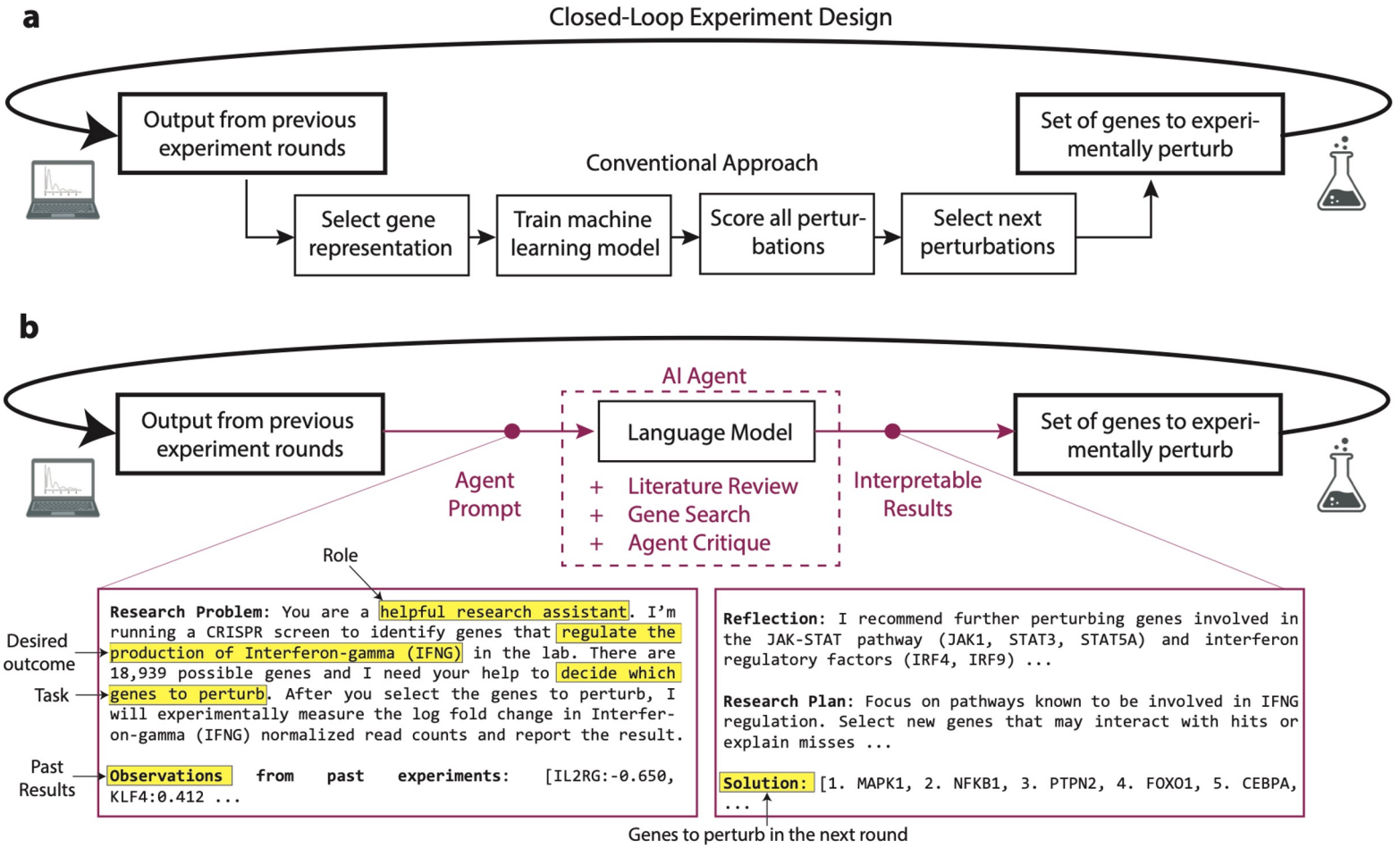
BIOREASON: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model

Arxiv 29 May 2025 [Link]



BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments

2025 ICLR [Link]



Summary

Part I: KG Definitions and Core Concepts

- Introduction to KGs: why we need to integrate KG into LLMs?
- A Simple Case of KGs: Ecotoxicological Effect Assessment
- KGs in Life Sciences: Challenges and Opportunities

Part II: Scientific Large Language Models (LLMs)

- Overview of Scientific LLMs: Bi-encoder, Cross-encoders
- Challenges and Perspectives

Part III: Integrating KGs and LLMs for Scientific Applications

- Knowledge Incorporation Frameworks
- KG Integration for Scientific NLP Tasks
- KG Integration for Scientific Prediction Tasks

Take-away

What KG brings to LLMs?

- Enhanced Knowledge Representation
- Improved Explainability, Reasoning and Inference
- Increased Accuracy and Reduced Hallucination

How to effectively incorporate KGs into LLMs?

- Backbone Model (protein, molecular, text, visual)
- Encoder Method (bi-encoder, cross-encoder)
- Integration Stages (pretrain, post-train, fine-tune, inference)
- Integration Techniques (adapter, lora, ICL, RAG, LLM Agent)

References

[1] Pretrain-KGEs: Learning Knowledge Representation from Pretrained Models for Knowledge Graph Embeddings, EMNLP 2020

[2] ERNIE: Enhanced Language Representation with Informative Entities, ACL 2019

[3] MolXPT: Wrapping Molecules with Text for Generative Pre-training, ACL 2023

[4] Self-Alignment Pretraining for Biomedical Entity Representations, NAACL 2021

[5] Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT, EMNLP 2021

[7] BioBRIDGE: Biobridge: Bridging biomedical foundation models via knowledge graph, ICLR 2024

[8] FusionDTI: Fine-grained Binding Discovery with Token-level Fusion for Drug-Target Interaction, AI4Science 2024

[9] MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering, AMIA 2024

[10] Editing Factual Knowledge and Explanatory Ability of Medical Large Language Models CIKM 2024

[11] KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models, bioinformatics

[12] BIORAG: A RAG-LLM Framework for Biological Question Reasoning, Arxiv 2024

[13] SaprotHub: Making Protein Modeling Accessible to All Biologists

[14] Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models. (ACL 2024)

[15] SciAgent: Tool-augmented Language Models for Scientific Reasoning

[16] Imagebind: One embedding space to bind them all. (CVPR 2023)

[17] Knowledge Enhanced Representation Learning for Drug Discovery (AAAI 2024). Link

[18] Ontoprotein: Protein pretraining with gene ontology embedding.(ICLR 2022)

[19] Heterogeneous biomedical entity representation learning for gene–disease association prediction. Briefings in Bioinformatics (2024) Link

[20] Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation, Nature

[21] Biomni: A General-Purpose Biomedical AI Agent, BioRxiv 2025

[22] BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model, Arxiv 2025

[23] BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments, ICLR 2025

Thank you!

Q & A