Swiss Institute of Bioinformatics

The Co-evolution of Foundation Models and Structured Bioinformatics Datasets

Ana Claudia Sima, Knowledge Representation Unit,
SIB Swiss Institute of Bioinformatics

ana-claudia.sima@sib.swiss

elixir SWITZERLAND

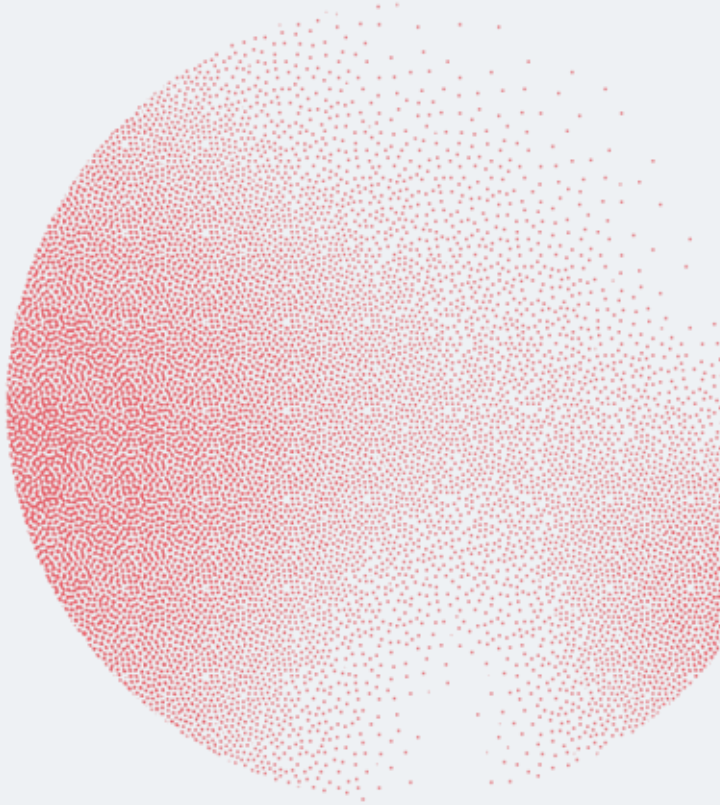
Outline

- Foundation Models
  - What are they ? Opportunities and limitations
- Structured Biological Databases
  - Highlights from the SIB Swiss Institute of Bioinformatics
- The co-evolution of Foundation Models and Structured Data
  - Towards a comprehensive representation of biological knowledge
- Guidelines for writing a review in this field
  - Existing work, differentiators, suggested paper structure

PART 1

Foundation models and their limits

What are they, what makes them powerful — and their limitations

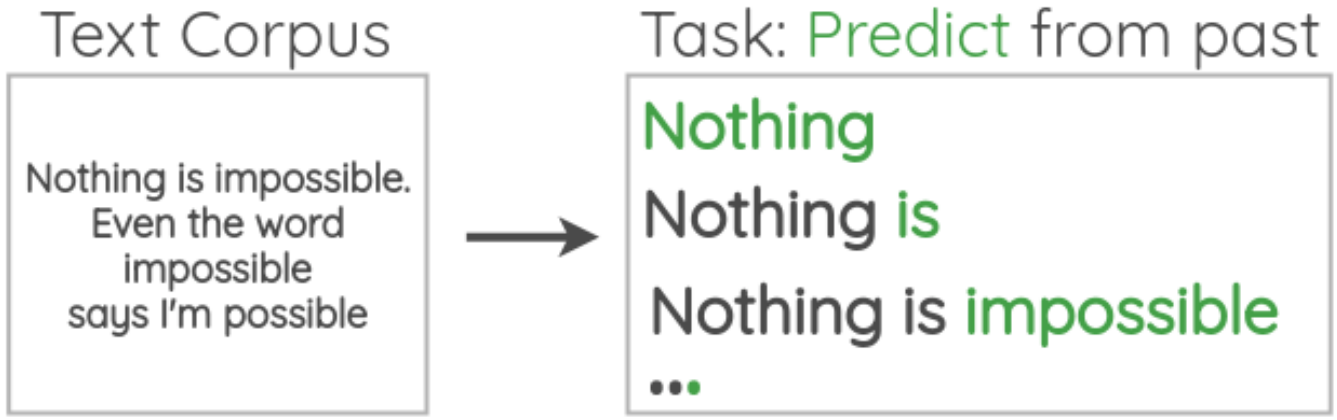


What is a Foundation Model? Example: a Large Language Model

A (deep learning) model trained to predict the next word in a sentence

How is it trained?

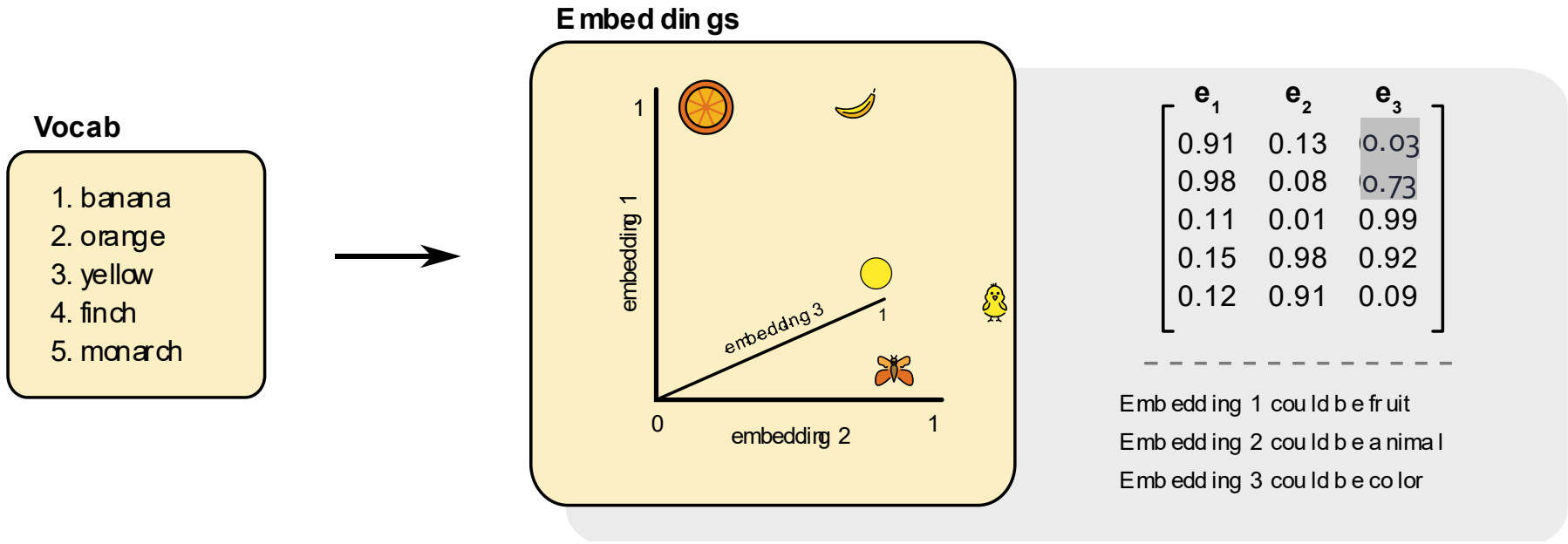
- Self-supervised learning on HUGE amounts of text
- "Fill in the blanks" learning



Source: https://amitness.com

Pre-training: constructing embeddings

- LLMs, like any AI system, work with *numbers*
- *Embeddings = a way to transform words (**tokens**) into numbers, while preserving their semantics*



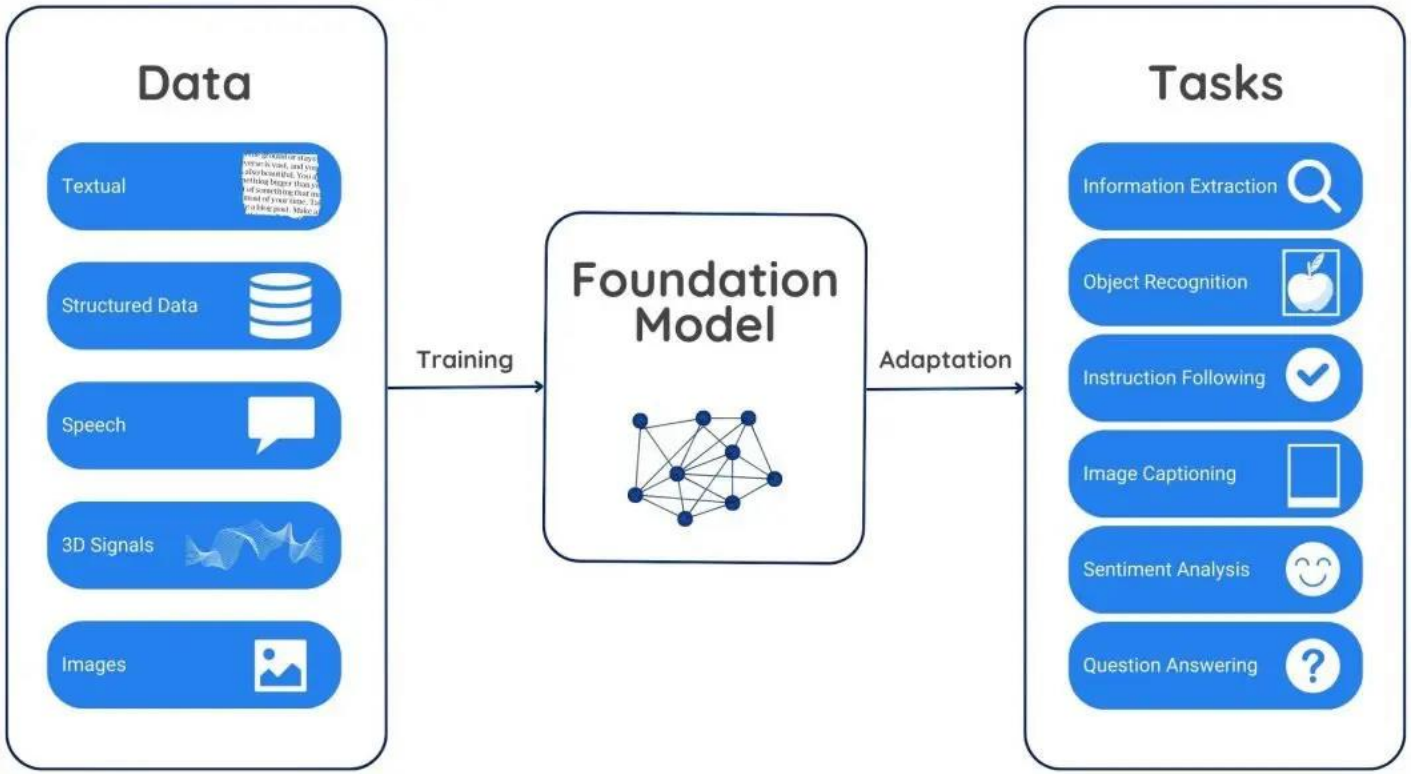
Source: https://dkharazi.github.io/notes/ml/nlp/embedding

What is a Foundation Model? (Generalization)

- A (deep learning) model trained to **predict** the next token in a sequence
  - Token = words / amino-acids / genes / ….
- What is a Language?
  - A **vocabulary** +
  - **Sequences of tokens** that represent **information** in that language



Result: Foundation Model



Source: https://viso.ai/deep-learning/foundation-models/

BACKGROUND

Foundation models

The idea

- Self-supervised pre-training on broad, unlabelled data at scale
- Learns general-purpose, transferable representations
- Adapts to many tasks via fine-tuning, prompting or in-context learning
- Few task-specific labels needed — strong in low-data regimes

In biology, "data" takes many forms

- **Sequence** — protein / DNA / RNA language models — ESM-2, Evo
- **Single-cell / omics** — expression as gene tokens — Geneformer, scGPT
- **Graph / KG** — relations between entities — ULTRA, TxGNN
- **Text** — biomedical literature & records — LLMs

Bioinformatics applications

- Language models for biological research: a primer, Nature Methods, ‘24

Table 1 | Biological language model resources

Name	What is it	Where to find it	How to use it
ChatGPT	Natural language (general)	https://chat.openai.com	Chat with the model at chat.openai.com or programmatically query the API as per the documentation.
BioBERT	Natural language (biomedical)	https://github.com/dmis-lab/biobert	Links to Hugging Face repositories with pre-trained BioBERT models of different sizes, along with code to fine-tune BioBERT.
Med-PaLM 2	Natural language (biomedical)	https://cloud.google.com/vertex-ai/generative-ai/docs/medlm/overview	Available to certain customers as part of Google's Vertex AI platform.
ESM	Protein language	https://github.com/facebookresearch/esm	Links to model code, pre-trained models and tutorials.
ProGen	Protein language	https://github.com/salesforce/progen/tree/main	Links to pre-trained models and code for making predictions.
Geneformer	Single-cell language	https://huggingface.co/ctheodoris/Geneformer	Trained models available through Hugging Face. Provides documentation with example code for various applications.
scGPT	Single-cell language	https://github.com/bowang-lab/scGPT	Links to pre-trained models and websites with cell annotations and other features.
GenePT	Multimodal: natural language and single-cell gene expression	https://github.com/yiqunchen/GenePT	Pre-computed gene embeddings and notebooks with example tutorials for various applications.
PLIP	Multimodal: natural language and pathology images	https://huggingface.co/spaces/vinid/webplip	Links to training data, pre-trained models and code.
Hugging Face	Multiple language model types	https://huggingface.co	Repository of many different trained machine-learning models. Query for specific terms (for example, protein) to find models in a specific domain.

So…

Are Foundation Models all we need?

The role of curated databases in the LLM era?

what are the GO annotations of P12345 in Uniprot?

Searching the web

The answer actually stems from Uniprot!

The UniProt entry **P12345** corresponds to the mitochondrial aspartate aminotransferase (GOT2) from *Oryctolagus cuniculus* (rabbit). This protein has several Gene Ontology (GO) annotations that describe its molecular functions, biological processes, and cellular components. UniProt +2

🧬 Molecular Function (MF)

- **GO:0004069** – L-aspartate:2-oxoglutarate aminotransferase activity
- **GO:0016212** – Kynurenine-oxoglutarate transaminase activity
- **GO:0030170** – Pyridoxal phosphate binding EMBL-EBI +1 EMBL-EBI

🧠 Biological Process (BP)

- **GO:0006531** – Aspartate metabolic process
- **GO:0006533** – Aspartate catabolic process
- **GO:0006536** – Glutamate metabolic process
- **GO:0006103** – 2-oxoglutarate metabolic proces

If no source were available…

what are the GO annotations of P12345 in Uniprot?

To retrieve the Gene Ontology (GO) annotations for the UniProt entry **P12345**, you can either:

1. Check Directly on UniProt Website:

Visit the UniProt entry for P12345. Look for the **"Function"** section, where GO annotations are categorized under:

- **GO - Biological process (BP)**
- **GO - Molecular function (MF)**
- **GO - Cellular component (CC)**

2. Example of GO Annotations (If Available):

Let me fetch a simplified example (if the entry exists and is publicly known):

Example (this depends on what P12345 maps to — for instance, if it corresponds to the **PPA1_HUMAN** protein, Inorganic pyrophosphatase):

- **GO:0000166** – nucleotide binding (MF)

LIMITATION 1

The long tail: rare knowledge

Why it happens

- Pre-training reflects what is **frequent** in the corpus
- Rare diseases, understudied genes, novel entities are underrepresented
- Models interpolate confidently — and **hallucinate** plausible-but-wrong facts

What it means for biology

- Fabricated gene–disease links or citations
- Poor coverage of newly characterised entities
- Confident answers precisely where evidence is weakest

Curated databases hold exactly this long-tail knowledge — reliably.

LIMITATION 2

Explainability & provenance

1

Black-box reasoning

Billions of parameters; no explainability, no auditable chain from input to answer.

2

No provenance

Outputs rarely cite a verifiable, persistent source.

3

Hard to trust

Clinicians and biologists need the "why", not just the "what".

4

Hard to correct

A wrong belief is diffused in the weights — not a row or an entry that you can simply edit.

LIMITATION 3

Staleness & weak grounding

Frozen at a training cutoff

- Biology moves fast: new variants, structures, annotations weekly
- A static model drifts out of date the day it ships
- Retraining is expensive and infrequent
- Facts are entangled in weights, not refreshed from a source

Databases are continuously curated

- Versioned and regularly updated
- New evidence enters through **expert curation** pipelines
- Retrieval at inference time keeps answers current
- A single source of truth many models can share

THE PROBLEM IN ONE SLIDE

What foundation models cannot do alone

Long tail

Rare, specialist facts are unreliable

Provenance

No auditable, citable source

Freshness

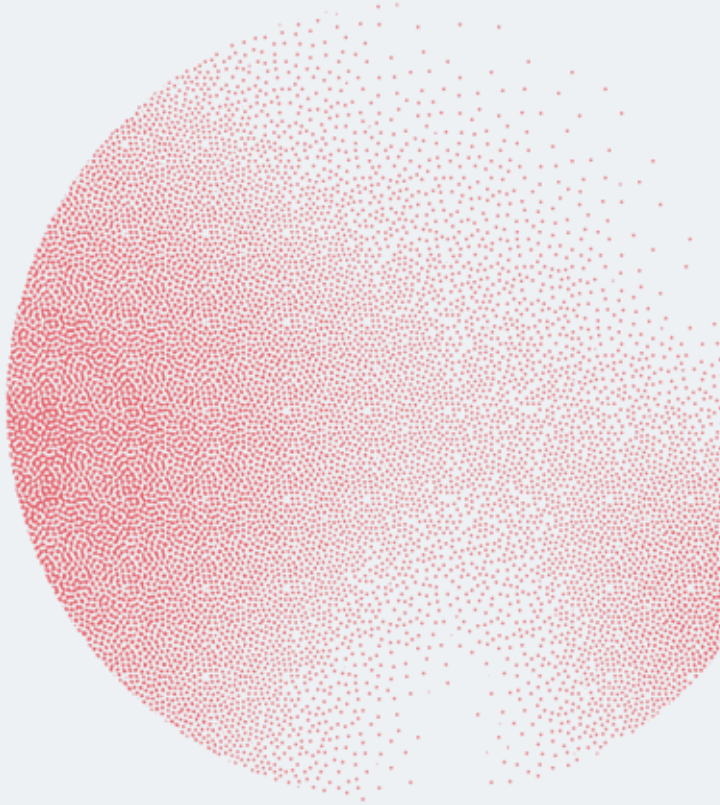
Frozen at a training cutoff

Each gap is precisely a strength of curated, structured biological data.

PART 2

Structured data as the complement

Tabular data and knowledge graphs as curated, trustworthy source of knowledge.



SIB Resources

SIB provides worldwide life scientists with long-term access
to highly curated, precise and authoritative biodata resources in Switzerland.

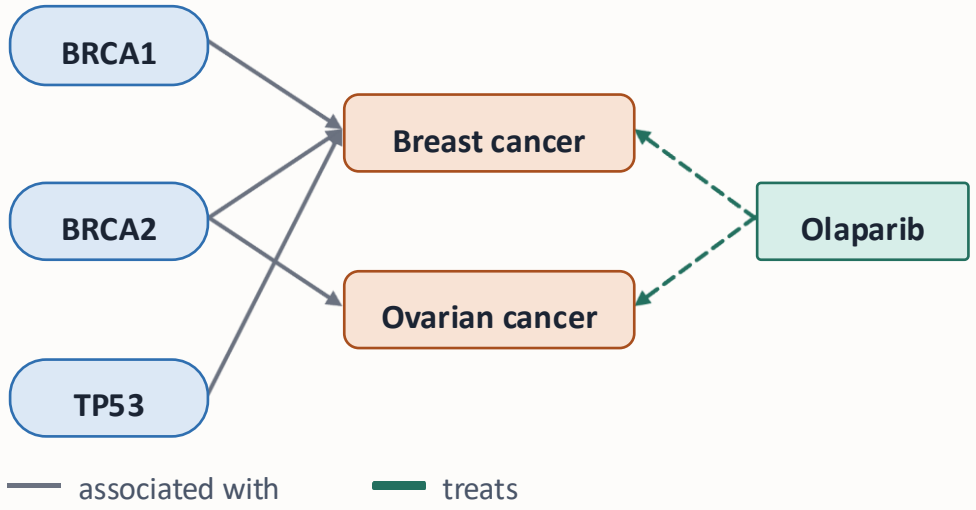


Two main data model types

Tabular data · rows × columns

Gene	Disease	Approved drug
BRCA1	Breast cancer	Olaparib
BRCA2	Breast cancer	Olaparib
BRCA2	Ovarian cancer	Olaparib
TP53	Breast cancer	none

Graph data · nodes + edges

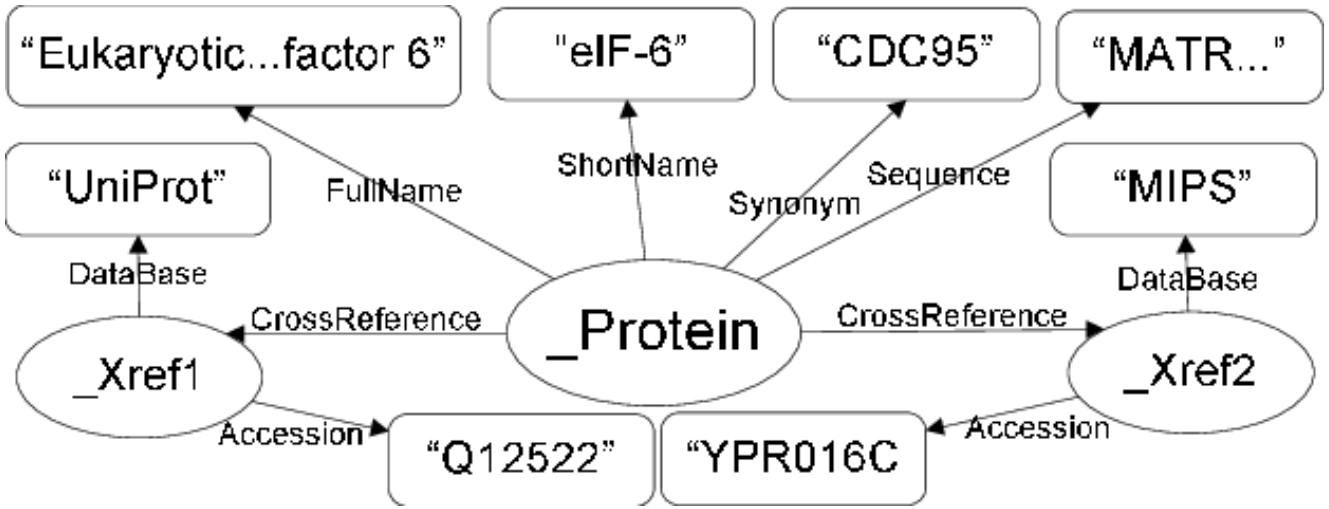


Two main data modelling modalities

Tabular – e.g. Bgee

Anatomical entity	Expression score	FDR	Link to source data
UBERON:0002469 esophagus mucosa	98.90	<= 1.00e-14	See source data
UBERON:0004991 mucosa of transverse colon	98.89	<= 1.00e-14	See source data
UBERON:0035834 lower esophagus mucosa	98.84	<= 1.00e-14	See source data
UBERON:0001511 skin of leg	98.68	<= 1.00e-14	See source data
UBERON:0001416 skin of abdomen	98.59	<= 1.00e-14	See source data
UBERON:0002114 duodenum	98.33	<= 1.00e-14	See source data
UBERON:0001052 rectum	98.14	<= 1.00e-14	See source data
UBERON:0000996 vagina	97.92	<= 1.00e-14	See source data
UBERON:0000082 adult mammalian kidney	97.85	<= 1.00e-14	See source data
UBERON:0001135 smooth muscle tissue	97.80	<= 1.00e-14	See source data
UBERON:0001114 right lobe of liver	97.66	<= 1.00e-14	See source data

Graph – e.g. Uniprot



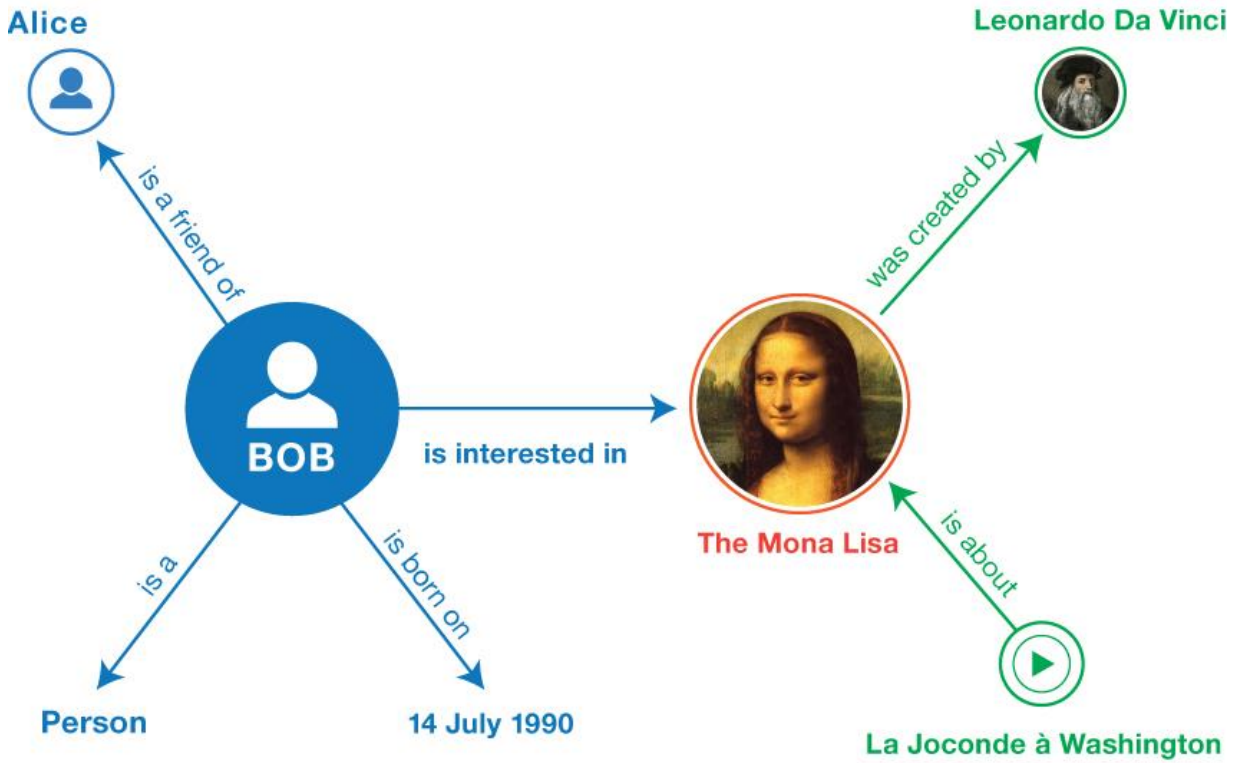
https://www.bgee.org/

Tutorial on Uniprot in graph format

What is a Knowledge Graph?

Graph data model

- Easy to **interlink** and extend
- Data usually stored as triples
- Triple = <subject, predicate, object>



Interoperable **graph** datasets curated by the SIB



12 resources found for sparql

OMA SPARQL endpoint
SPARQL access to the OMA Orthology database

Bgee SPARQL endpoint
SPARQL access to the Bgee gene expression database

OrthoDB SPARQL endpoint
SPARQL access to the OrthoDB catalog of orthologs

HAMAP SPARQL endpoint
SPARQL access to the HAMAP annotation rules

UniProt SPARQL endpoint
SPARQL access to UniProt and related data

Rhea SPARQL endpoint
SPARQL access to the Rhea knowledgebase

SwissLipids SPARQL endpoint
SPARQL access to the SwissLipids knowledgebase

MetaNetX SPARQL endpoint
SPARQL access to MetaNetX

GlyConnect SPARQL endpoint
SPARQL access to GlyConnect

Cellosaurus SPARQL endpoint
SPARQL access to Cellosaurus

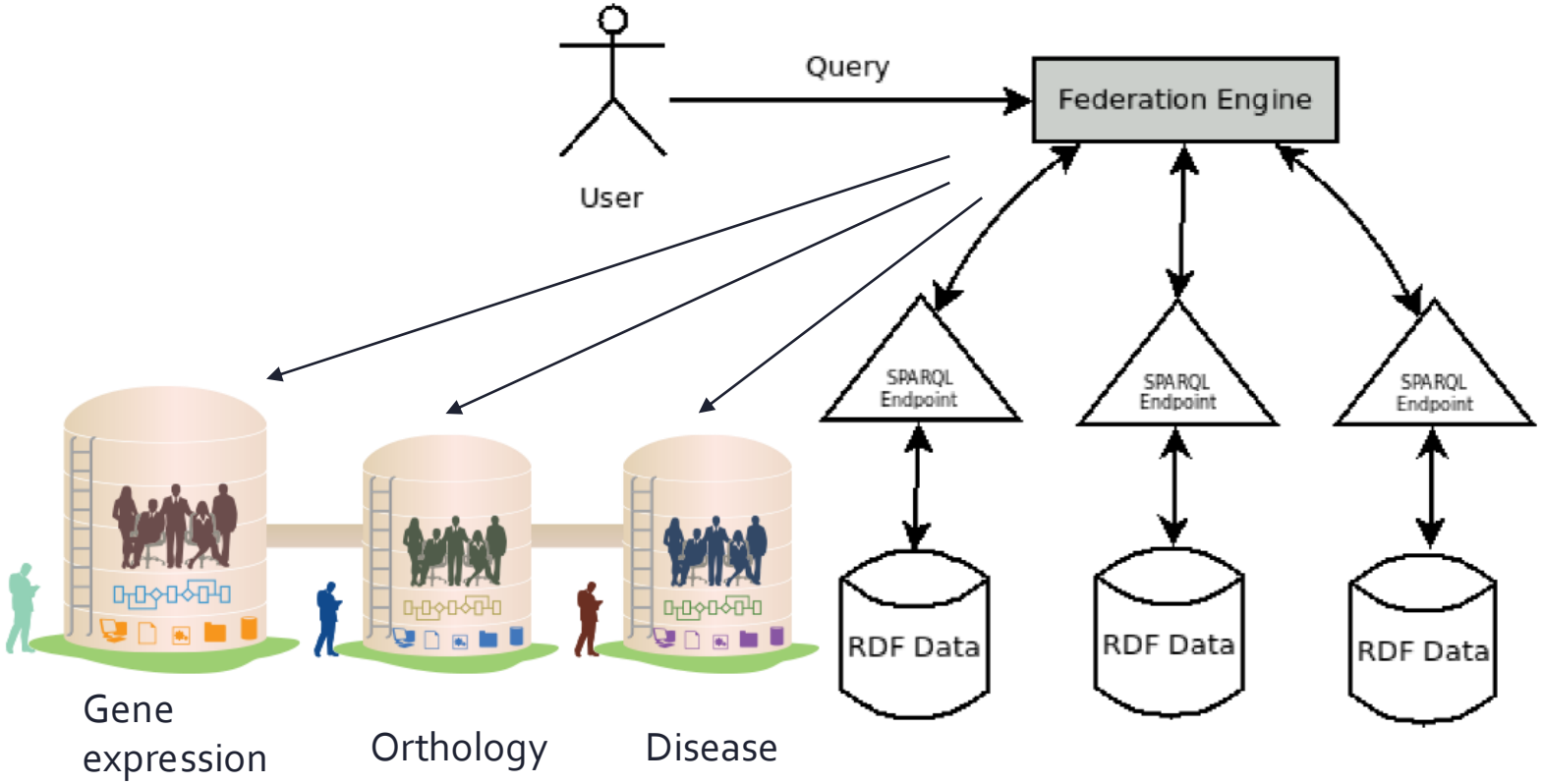
STRING SPARQL endpoint
SPARQL access to the STRING knowledgebase

SIBiLS SPARQL endpoint
SPARQL access to SIBiLS, the SIB Literature Services

The power of interoperable data

Ask complex questions that span **across disciplines and across datasets**

What are the human genes involved in lung cancer with an ortholog expressed in the mouse?



Position: Why Tabular Foundation Models Should Be a Research Priority

Boris van Breugel [1] **Mihaela van der Schaar** [1 2]

Abstract

Recent text and image foundation models are incredibly impressive, and these models are attracting an ever-increasing portion of research resources. In this position piece we aim to shift the ML research community's priorities ever so slightly to a different modality: tabular data. Tabular data is the dominant modality in many fields, yet it is given hardly any research attention and significantly lags behind in terms of scale and power. **We believe the time is now to start developing tabular foundation models**, or what we coin a *Large Tabular Model* (LTM). LTMs could revolutionise the way science and ML use tabular data: not as single datasets that are analyzed in a vacuum, but contextualized with respect to related datasets. The potential impact is far-reaching: from few-shot tabular models to automating data science; from out-of-distribution synthetic data to empowering multidisciplinary scientific discovery. We intend to excite reflec-

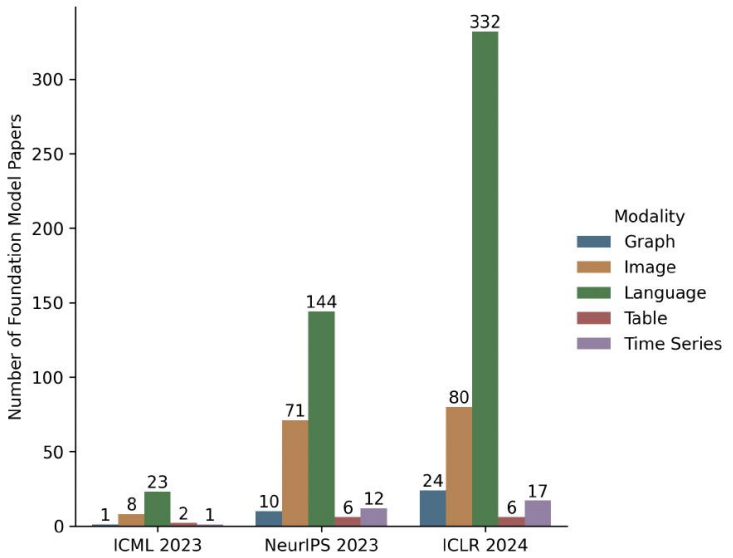
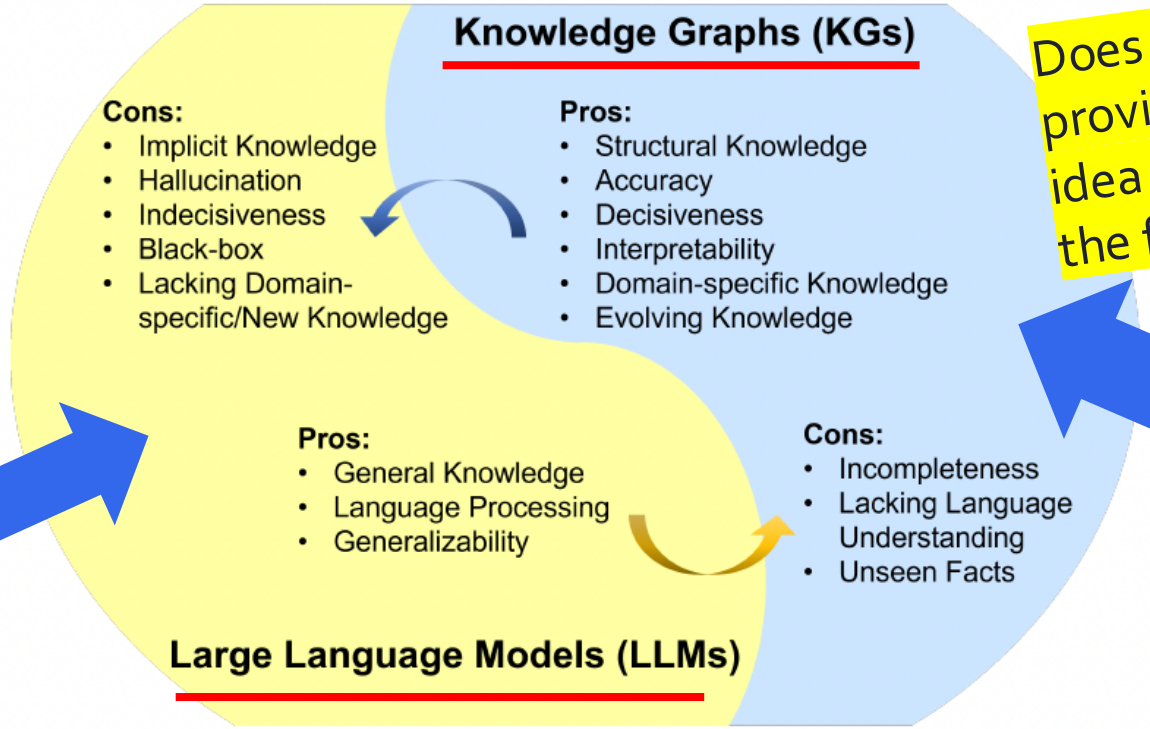


Figure 1. Representation of different modalities in foundation model research across recent ML conferences, roughly estimated as the number of accepted papers with abstracts containing keywords (see Appendix A). LLMs are booming and tabular data is heavily underrepresented.

[Position: why tabular foundation models should be a research priority, ICML '24]

LLMs + KGs = ?



Does published literature provide **evidence** for the idea of a virtuous circle in the field of biology data?

Artificial Intelligence, Enconding General Knowledge

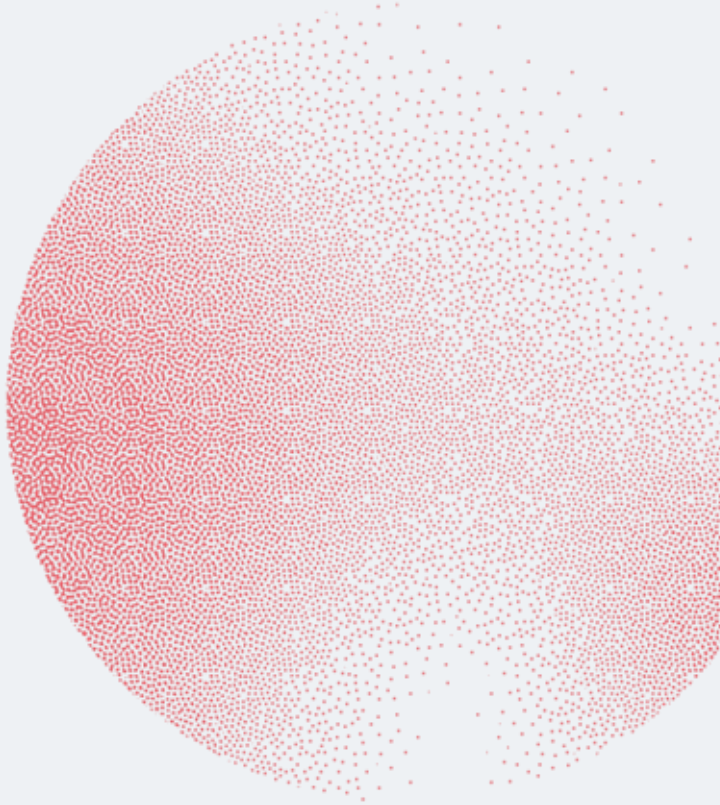
Curation, Domain Expertise

Fig. 1. Summarization of the pros and cons for LLMs and KGs. LLM pros: *General Knowledge* [11], *Language Processing* [12], *Generalizability* [13]; LLM cons: *Implicit Knowledge* [14], Hallucination [15], Indecisiveness [16], Black-box [17], Lacking Domain-specific/New Knowledge [18]. KG pros: *Structural Knowledge* [19], *Accuracy* [20], *Decisiveness* [21], *Interpretability* [22], *Domain-specific Knowledge* [23], *Evolving Knowledge* [24]; KG cons: *Incompleteness* [25], *Lacking Language Understanding* [26], *Unseen Facts* [27].

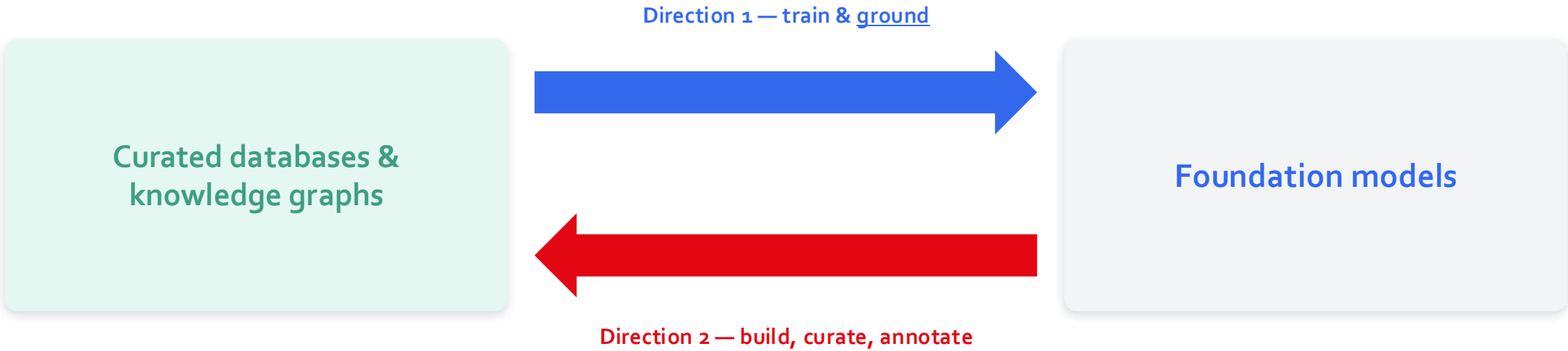
Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." *TKDE 2024*

PART 3

The co-evolution thesis & the review



Co-evolution



Databases enabled the first biological FMs → FMs now curate and extend the databases → richer data enables better models.

A prototypical example: AlphaFold <-> Uniprot

Sequences flow into structure prediction; predicted structures flow back into the knowledgebase.

UniProt

Protein sequence and function knowledgebase

Swiss-Prot (reviewed)
TrEMBL (unreviewed)
UniRef sequence clusters

Sequences in

UniProtKB sequences are the prediction targets
UniRef90 is searched to build the MSAs
Models keyed by accession: AF-P69905-F1

Structures back

3D model shown in each UniProtKB entry
Confidence (pLDDT) displayed per residue
Informs annotation: domains, disorder

AlphaFold DB

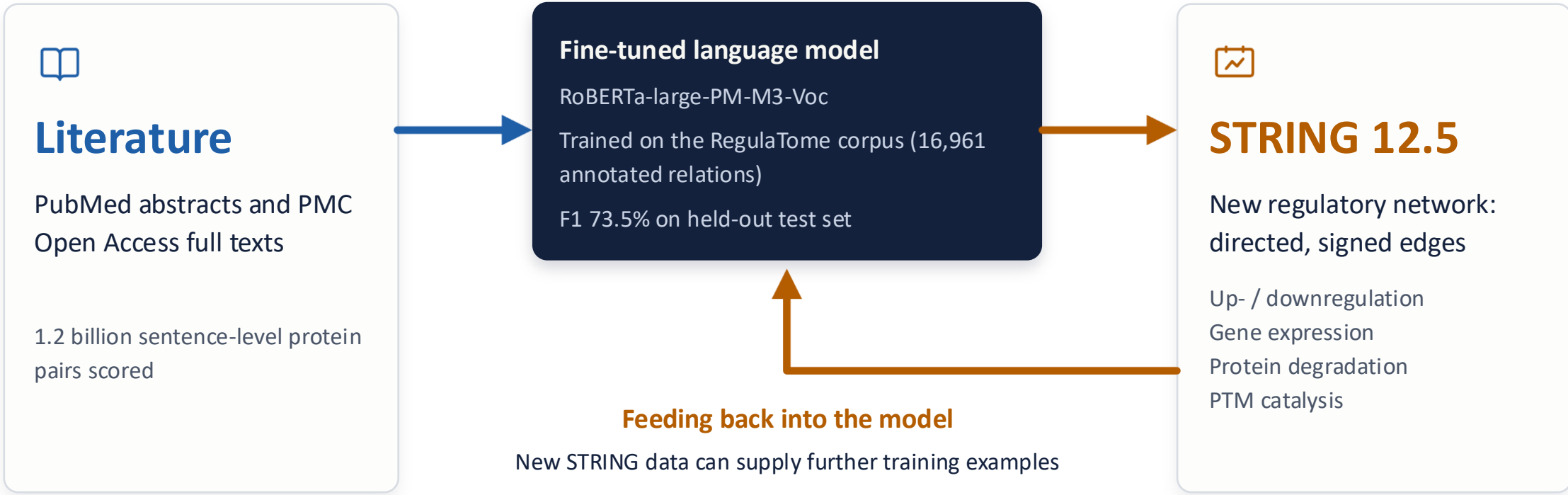
Predicted 3D protein structures

AlphaFold 2 models
Per-residue confidence (pLDDT)
Predicted aligned error (PAE)

A second example: curation

STRING 12.5 and the literature

Sentences become directed regulatory edges; each edge links back to its sentences.



Szklarczyk et al., NAR 2025

Direction 1

Databases -> models: structured data grounds Foundation Models

ESM-2

Protein LM trained on UniProt / UniRef

TxGNN

Drug-repurposing FM over PrimeKG

OntoProtein / KeAP

PLMs grounded in Swiss-Prot + Gene Ontology

PINNACLE

Context-specific protein FM on PPI networks

BioBridge

KG bridges independently trained FMs

KG-RAG

Retrieval over a biomedical KG grounds an LLM

Direction 2

Models -> databases: FMs build & curate data

AlphaFold DB

A model prediction became a 200M+ structure database

ESM Metagenomic Atlas

ESMFold predicted 600M+ metagenomic structures

STRING 12.5

A fine-tuned LM infers regulatory networks from literature

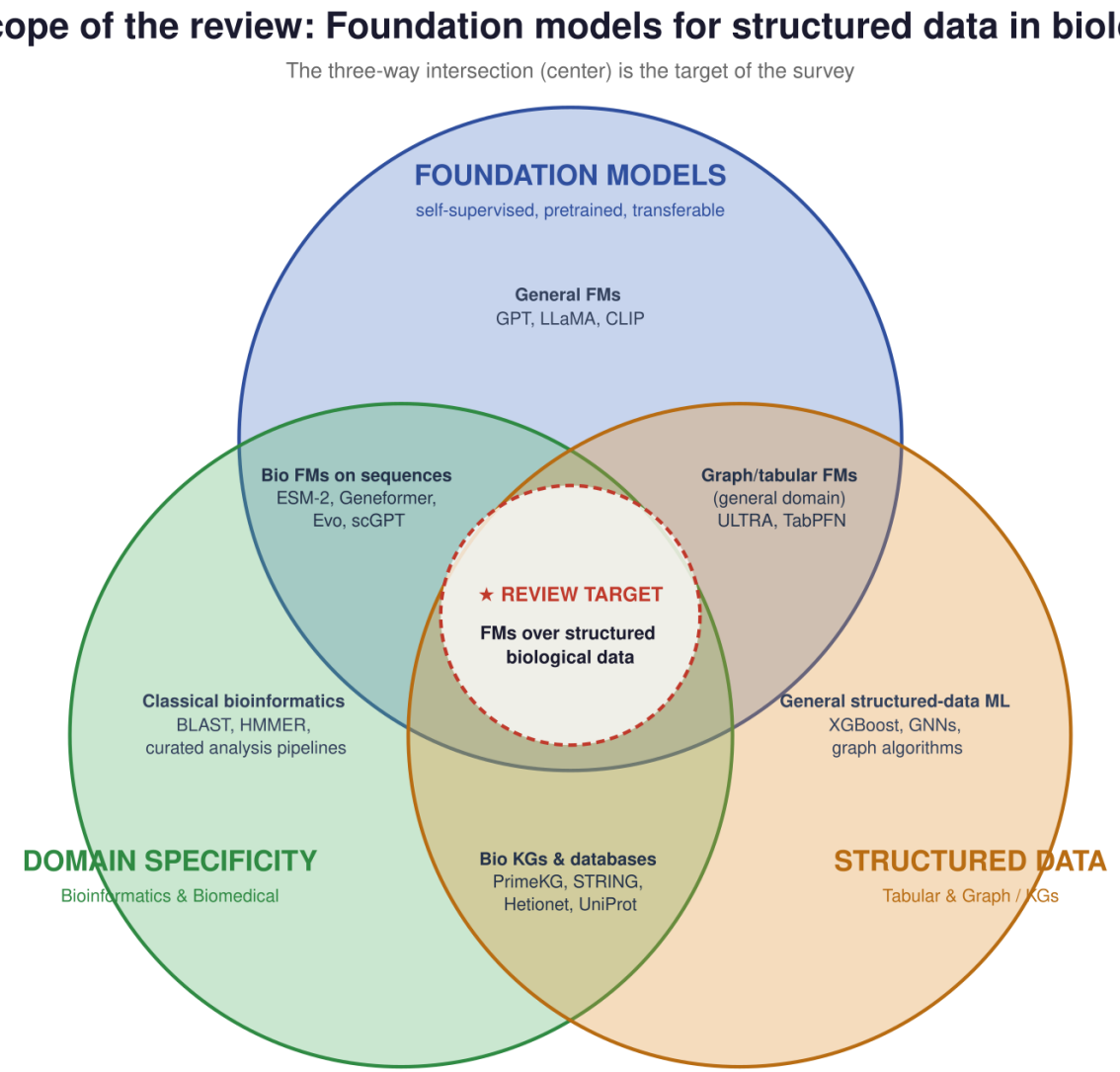
MedKGent

LLM agents build a temporally evolving medical KG

CLEAN (FM-powered)

ESM-1b embeddings enrich enzyme (EC) annotation

What your review could be about



Three surveys to start from — and how to use each

1

Tracing the rise of biomedical foundation models

Nature Biotechnology · Comment · 2025 · a very brief (4-page!) review

Use for: a methodology for paper selection & overview —
+ how to write a closing argument and outlook (section "Toward a robust and translational ecosystem for biomedical foundation models").

2

Synergizing Large Language Models and Knowledge Graphs in Science: A Survey

NeurIPS AI4Science Workshop · 2025

Use for: learning how to position your review in relation to existing work.

3

A survey on foundation models for structured data

IEEE TKDE · October 2026 · advanced topics, very newly published

Use for: enriching your own knowledge ! ☺
— and learning how to segment prior work across different dimensions.

Your review could select a few central ideas and **identify how they are applied in biology – start from a few well-known databases (Uniprot, STRING)**

A recommended structure

1	**Introduction**	Foundation Models, their limits, and the promise of structured data
2	**Background, the Co-Evolution Theory**	FMs; tabular & graph data; key biological resources
3	**Direction 1: data -> models**	Grounding, pre-training, KG-augmented & retrieval FMs
4	**Direction 2: models -> data**	FM-built databases, curation, annotation, completion
5	**Open challenges, Gaps, Untapped Opportunities**	Evaluation, Benchmarks, Scalability …
6	**Outlook & conclusion**	The road ahead

Conclusions

- Structured bio databases are the **foundations** of trustworthy AI in biology
- The current trend points towards a virtuous circle of **co-evolution** between large foundational models and structured datasets
  - Can you find evidence for this in the literature?
- A great opportunity for surveying existing work and highlighting directions for the future!