

# Manifold Learning and Artificial Intelligence

## Lecture 10

### A New Paradigm for Data Analysis

Momiao Xiong, University of Texas School of Public Health

- Time: 10:00 pm, US East Time, 03/18/2023
- 10:00 am, Beijing Time. 03/19/2023
- Zoom

[https://uwmadison.zoom.us/j/93316139423?tk=wfbmsTfN2fgERto\\_HI1WKtBzh94d3HO02XVCexqd8.DQMMAAAVuhNJnxZ2dGVcbIIYIR3ZWJBNI5LXIYMjBnAAAAAAAAAAAAAAAAAAAAAAAAAAAA&pwd=Q0NVWFYvRFg5RmxCNkwxMmYrbW41dz09](https://uwmadison.zoom.us/j/93316139423?tk=wfbmsTfN2fgERto_HI1WKtBzh94d3HO02XVCexqd8.DQMMAAAVuhNJnxZ2dGVcbIIYIR3ZWJBNI5LXIYMjBnAAAAAAAAAAAAAAAAAAAAAAAAAAAA&pwd=Q0NVWFYvRFg5RmxCNkwxMmYrbW41dz09)

Github Address: <https://ai2healthcare.github.io/>

# News

# Introducing Microsoft 365 Copilot — your copilot for work

<https://news.microsoft.com/reinventing-productivity>

**GPT-4**

**Automate Your Works**

# ART: Automatic multi-step reasoning and tool-use for large language models

**Bhargavi Paranjape<sup>1</sup> Scott Lundberg<sup>2</sup> Sameer Singh<sup>3</sup> Hannaneh Hajishirzi<sup>1,4</sup>  
Luke Zettlemoyer<sup>1,5</sup> Marco Tulio Ribeiro<sup>2</sup>**

<sup>1</sup>University of Washington, <sup>2</sup>Microsoft Research, <sup>3</sup>University of California, Irvine,  
<sup>4</sup>Allen Institute of Artificial Intelligence, <sup>5</sup>Meta AI

## Abstract

Large language models (LLMs) can perform complex reasoning in few- and zero-shot settings by generating intermediate chain of thought (CoT) reasoning steps. Further, each reasoning step can rely on external tools to support computation beyond the core LLM capabilities (e.g. search/running code). Prior work on CoT prompting and tool use typically requires hand-crafting task-specific demonstrations and carefully scripted interleaving of model generations with tool use. We introduce **Automatic Reasoning and Tool-use (ART)**, a framework that uses frozen *LLMs* to *automatically* generate intermediate reasoning steps as a program. Given a new task to solve, ART selects demonstrations of multi-step reasoning and tool use from a task library. At test time, ART seamlessly pauses generation whenever external tools are called, and integrates their output before resuming generation. ART achieves a substantial improvement over few-shot prompting and automatic CoT on unseen tasks in the BigBench and MMLU benchmarks, and matches performance of hand-crafted CoT prompts on a majority of these tasks. ART is also extensible, and makes it easy for humans to improve performance by correcting errors in task-specific programs or incorporating new tools, which we demonstrate by drastically improving performance on select tasks with minimal human intervention.

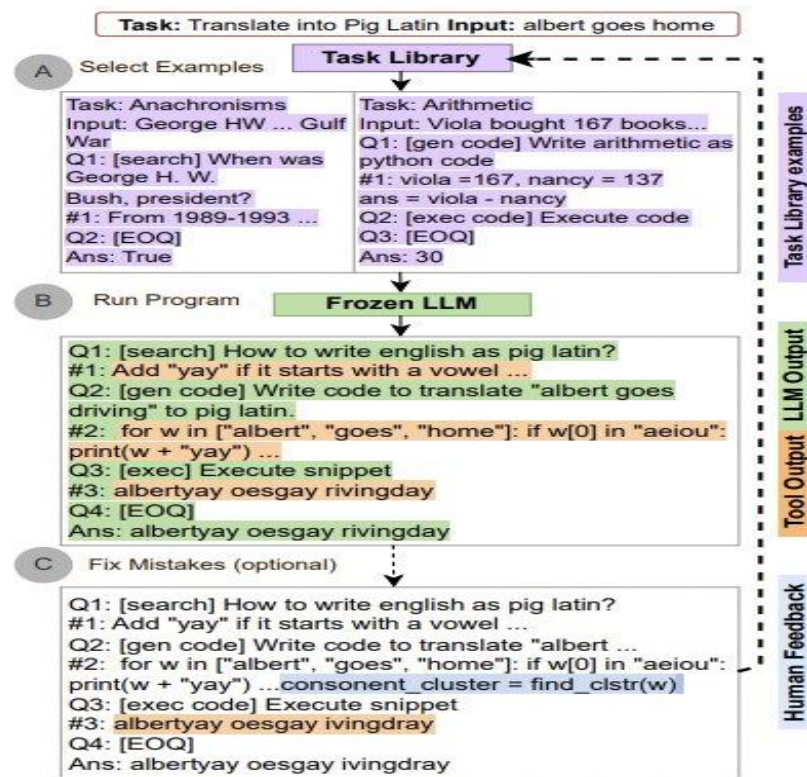
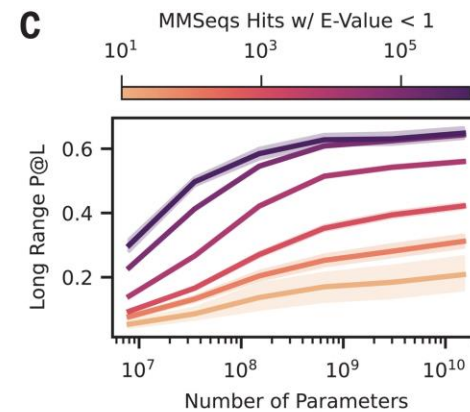
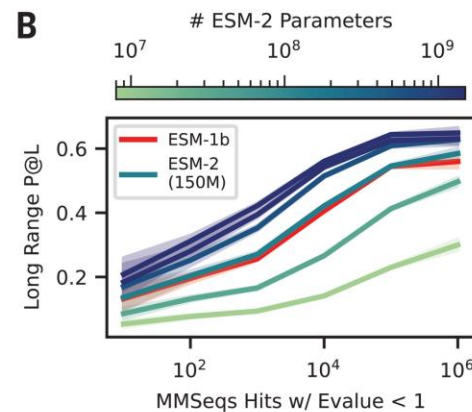
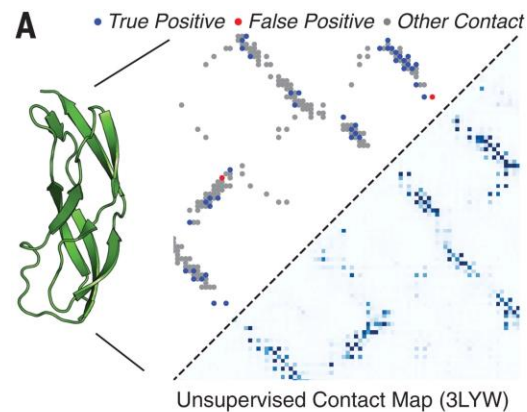


Figure 1: ART generates automatic multi-step decompositions for new tasks by selecting decompositions of related tasks in the *task library* (A) and selecting and using tools in the *tool library* alongside LLM generation (B). Humans can optionally edit decompositions (eg. correcting and editing code) to improve performance (C).

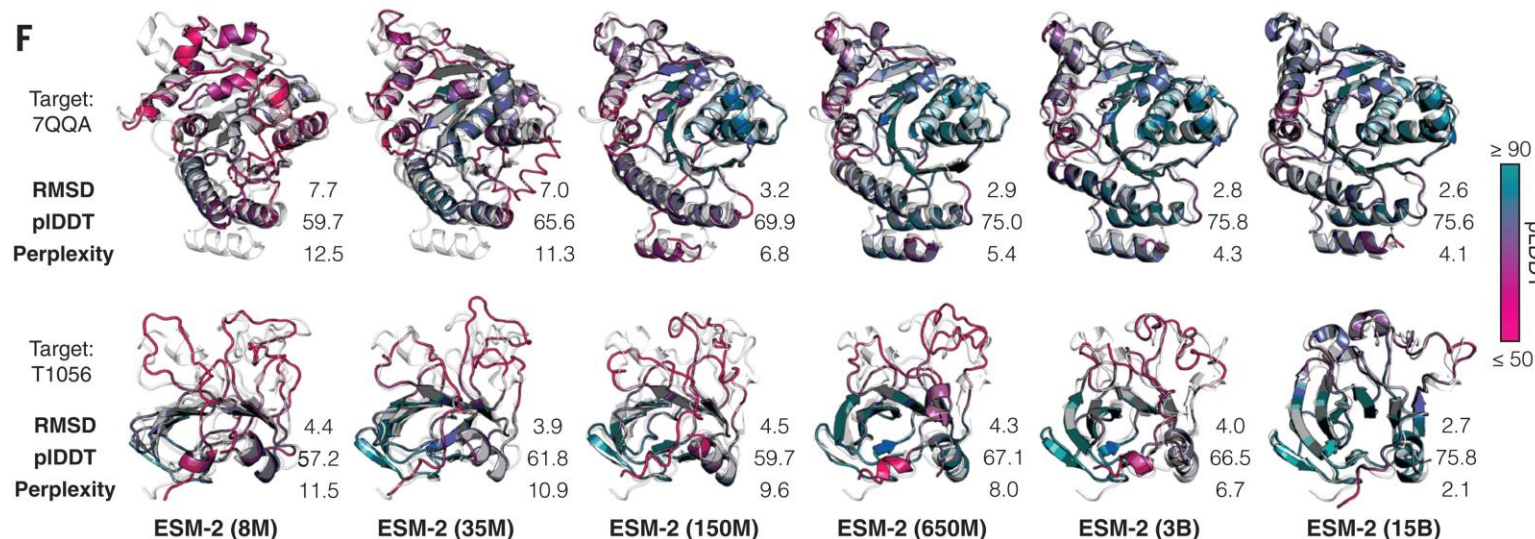
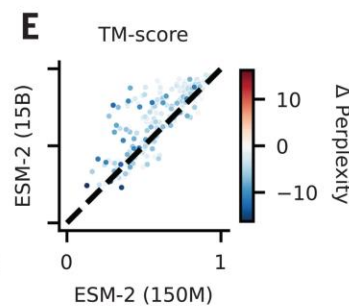
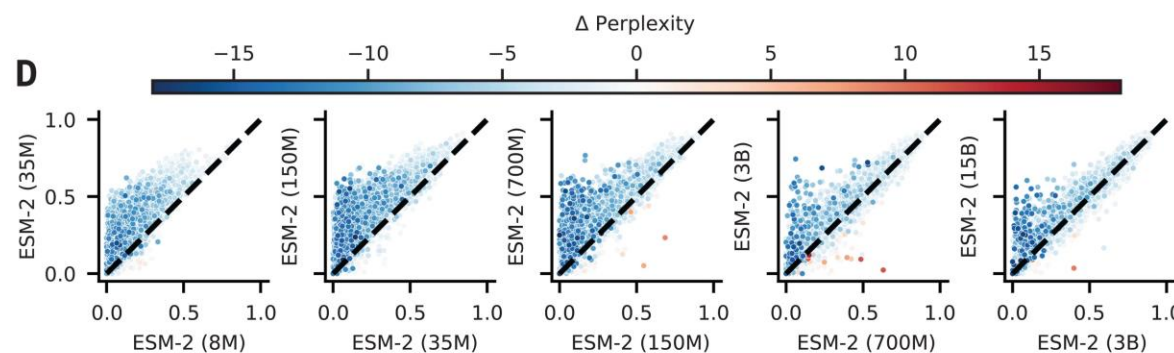




# Evolutionary-scale prediction of atomic-level protein structure with a language model

**SCIENCE** 16 Mar 2023

379, pp. 1123-1130



trained transformer protein language models with up to 15 billion parameters on experimental and high-quality predicted structures and found that information about atomic-level structure emerged in the model as it was scaled up. They created ESMFold, a sequence-to-structure predictor that is nearly as accurate as alignment-based methods and considerably faster. The increased speed permitted the generation of a database, the ESM Metagenomic Atlas, containing more than 600 million metagenomic proteins.

# AI for Science: An Emerging Agenda

**This report documents the programme and the outcomes of Dagstuhl Seminar 22382 "Machine Learning for Science: Bridging Data-Driven and Mechanistic Modelling".**

**Today's scientific challenges are characterised by complexity. Interconnected natural, technological, and human systems are influenced by forces acting across time- and spatial-scales, resulting in complex interactions and emergent behaviours. Understanding these phenomena | and leveraging scientific advances to deliver innovative solutions**

**to improve society's health, wealth, and well-being | requires new ways of analysing complex systems.**

**The transformative potential of AI stems from its widespread applicability across disciplines, and will only be achieved through integration across research domains. AI for science is a rendezvous point. It brings together expertise from AI and application domains; combines modelling knowledge with engineering know-how; and relies on collaboration across disciplines and between humans and machines. Alongside technical advances, the next wave of progress in the field will come from building a community of machine learning researchers, domain experts, citizen scientists, and engineers working together to design and deploy effective AI tools.**

**This report summarises the discussions from the seminar and provides a roadmap to suggest how different communities can collaborate to deliver a new wave of progress in AI and its application for scientific discovery.**

\*

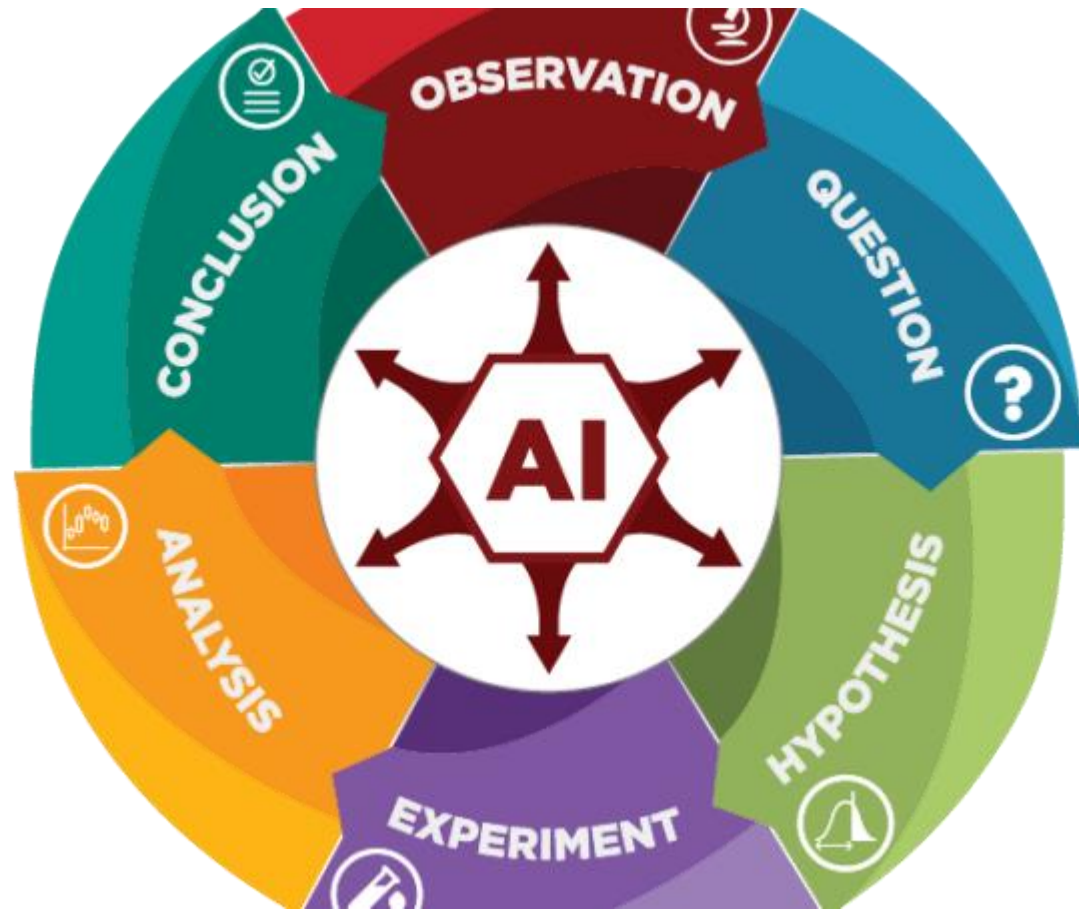
Berns et al. March 9, 2023

# AI+Science Initiative at UChicago

<https://datascience.uchicago.edu/research/ai-science/>

Scientific discovery, and innovation have traditionally relied on four distinct elements: theory, observation, simulation, and practice to advance knowledge. However, a new engine of scientific knowledge generation is now emerging.

Only powerful machines and sophisticated algorithms informed by domain-specific constraints will advance the AI+Science revolution. Modern artificial intelligence and machine learning will fundamentally change scientific discovery.



The University of Chicago Data Science Institute's AI+Science Initiative will lay the foundations for a new field of research, with cross-disciplinary teams of computer scientists, mathematicians, statisticians, engineers, physicists, chemists, biologists, geoscientists, and other domain scientists. AI-enabled scientific inquiry will allow us to **discover new fundamental principles; accelerate the pace of scientific discovery in multiple fields, identify gaps in our knowledge, models, and understanding; and vastly expand the range of exploration and experimentation** that can be subject to investigation.

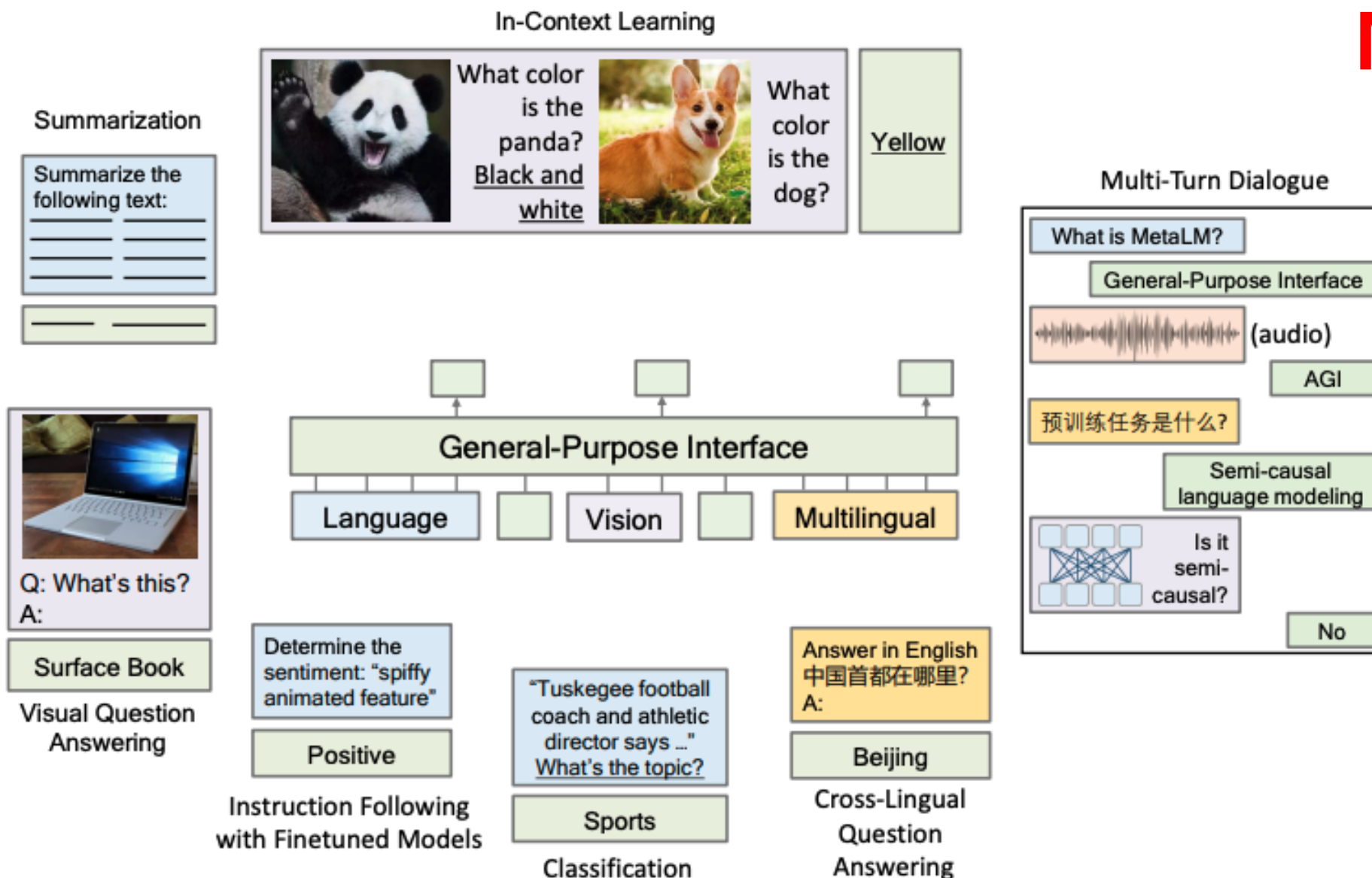
# **A New Paradigm for Data Analysis**

**Language Model-based Automatic Data Analysis**

**New Paradigm 1: Language Models as  
General-Purpose Interfaces, Automatic  
Reasoning**



# MetaLM



## Standard Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. 

## Chain of Thought Prompting


### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

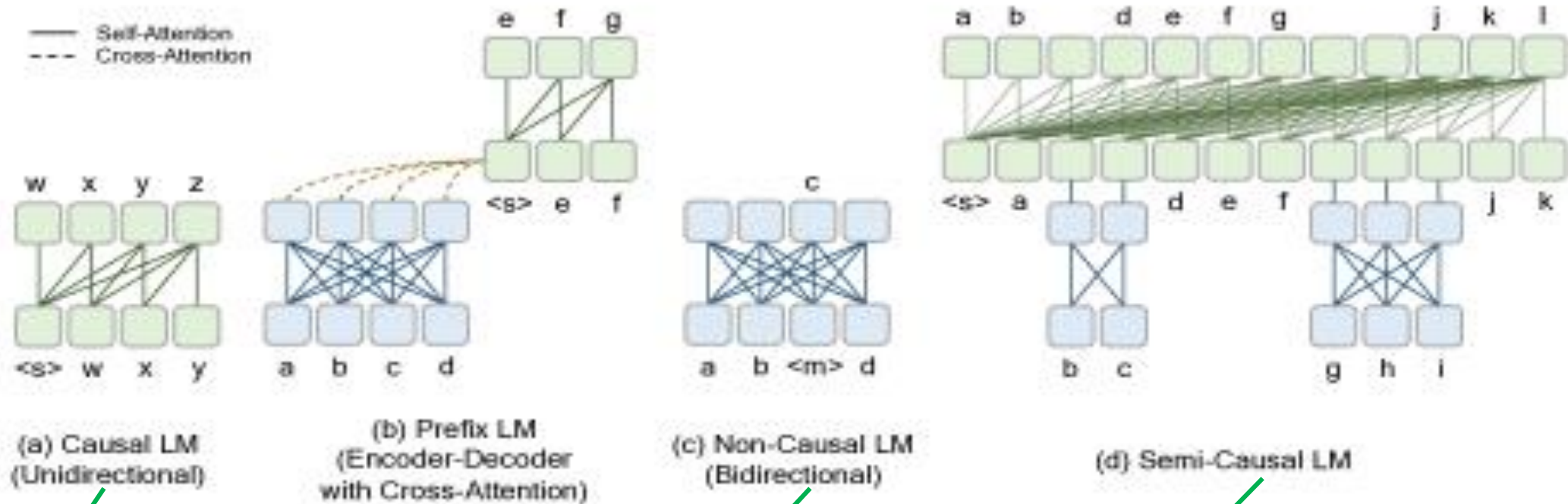
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. 

# Model Architecture



a left-to right  
Transformer  
decoder

complete the  
sequence

encoder  
masked  
language  
modeling

unidirectional Transformer  
decoder, and multiple bidirectional  
encoders that dock with the decoder

Model	MNLI (m)	Acc (mm)
GPT	87.7	87.6
BERT	86.6	
RoBERT	90.2	90.2
ELECTRA	90.9	
METALM	91.1	91.0

Table 3: Single-task finetuning results on matched (-m) and mismatched (-mm) validation sets of MNLI. Each score is the average of multiple runs with different random seeds

# Augmented Language Models (ALMs)

- **Augmenting LMs with both reasoning and tools**

- **Reasoning**

Reasoning is decomposing a potentially complex task into simpler subtasks the LM can solve more easily by itself or using tools. There exist various ways to decompose into subtasks, such as **recursion or iteration**.

- **Tool.**

A tool is an **external** module that is typically called using a rule or a special token and whose **output is included** in the ALM's context

- **Act.**

The call of a tool by an ALM as an action, even if it does not have an external effect.

**Typically, reasoning would foster the LM to decompose a given problem into potentially simpler subtasks while tools would help getting each step right, for example obtaining the result from a mathematical operation.**



**An example of few-shot Chain-of-Thought prompt. <LM> is to call to the LM with the above prompt.**

**Question:** Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

**Answer:** Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

**Question:** The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Answer:**

<LM>

**An example of zero-shot Chain-of-Thought prompt. <LM> is to call to the LM with the above prompt.**

**Question:** The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Answer:** **Let's think step by step**

<LM>

Our proposal is simple: Allow the **model to produce an arbitrary sequence of intermediate tokens, which we call a scratchpad**, before producing the final answer. For example, on addition problems, the scratchpad contains the intermediate results from a standard long addition algorithm (see Figure 2). To train the model, we encode the intermediate steps of the algorithm as text and use standard supervised training.

Input:

2 9 + 5 7

Target:

<scratch>

2 9 + 5 7 , C: 0

2 + 5 , 6 C: 1 # added 9 + 7 = 6 carry 1

, 8 6 C: 0 # added 2 + 5 + 1 = 8 carry 0

0 8 6

</scratch>

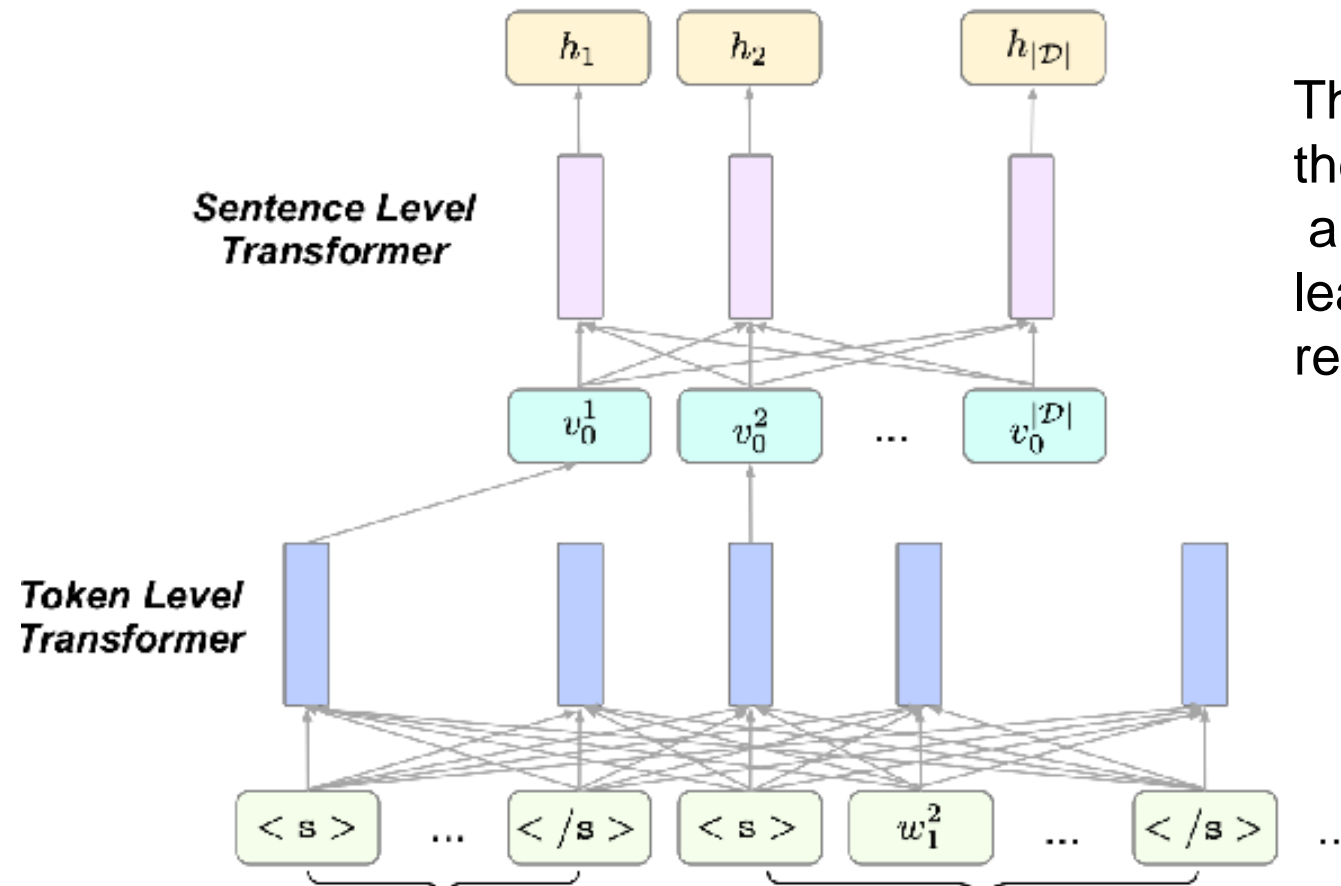
8 6

Example of input and target for addition with a scratchpad. The carry is recorded in the digit following "C:". Comments (marked by #) are added for clarity and are not part of the target.

## **Paradigm 2: Text Summarization as Feature Selection**

# Xu, et al. 2020; Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers.

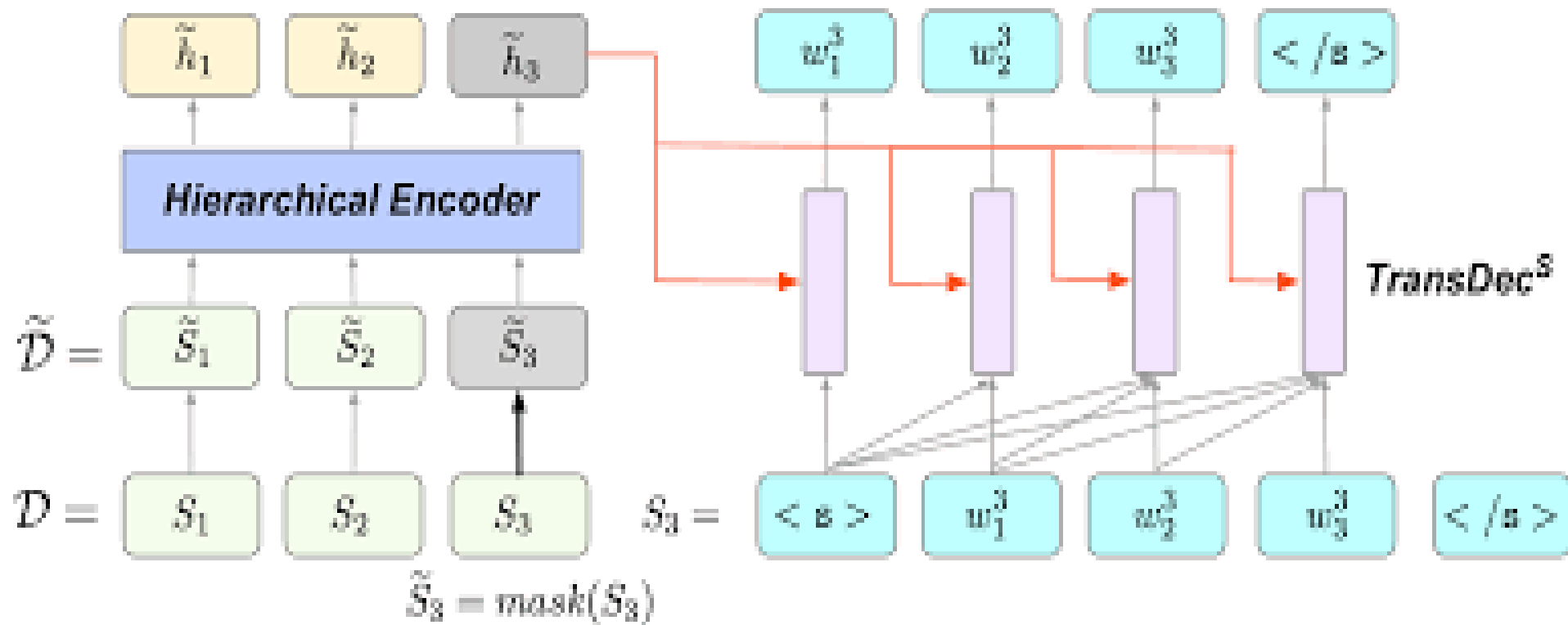
Our code and models are available at <https://github.com/xsstory/STAS>.



The architecture of our hierarchical encoder, the token level Transformer encodes tokens and then The sentence level Transformer learns final sentence representations from representations at  $\langle s \rangle$ .

**$\langle s \rangle$ : Begin of sentence**

**$\langle /s \rangle$ : End of sentence**



Finding the most important sentence is equivalent to finding the sentence with highest probability

$$P(S_i | D \setminus S_i)$$



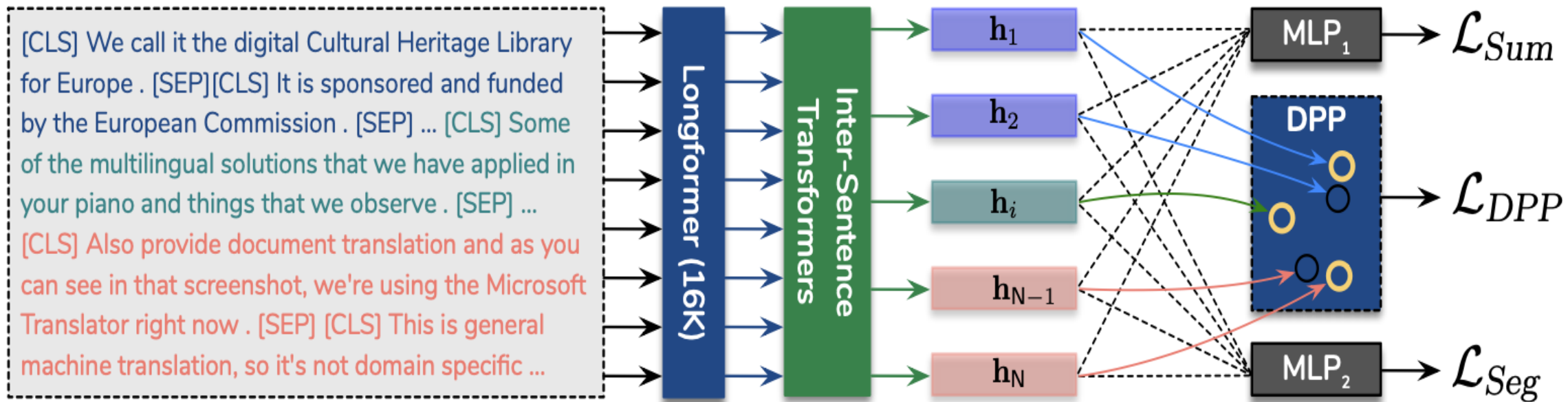
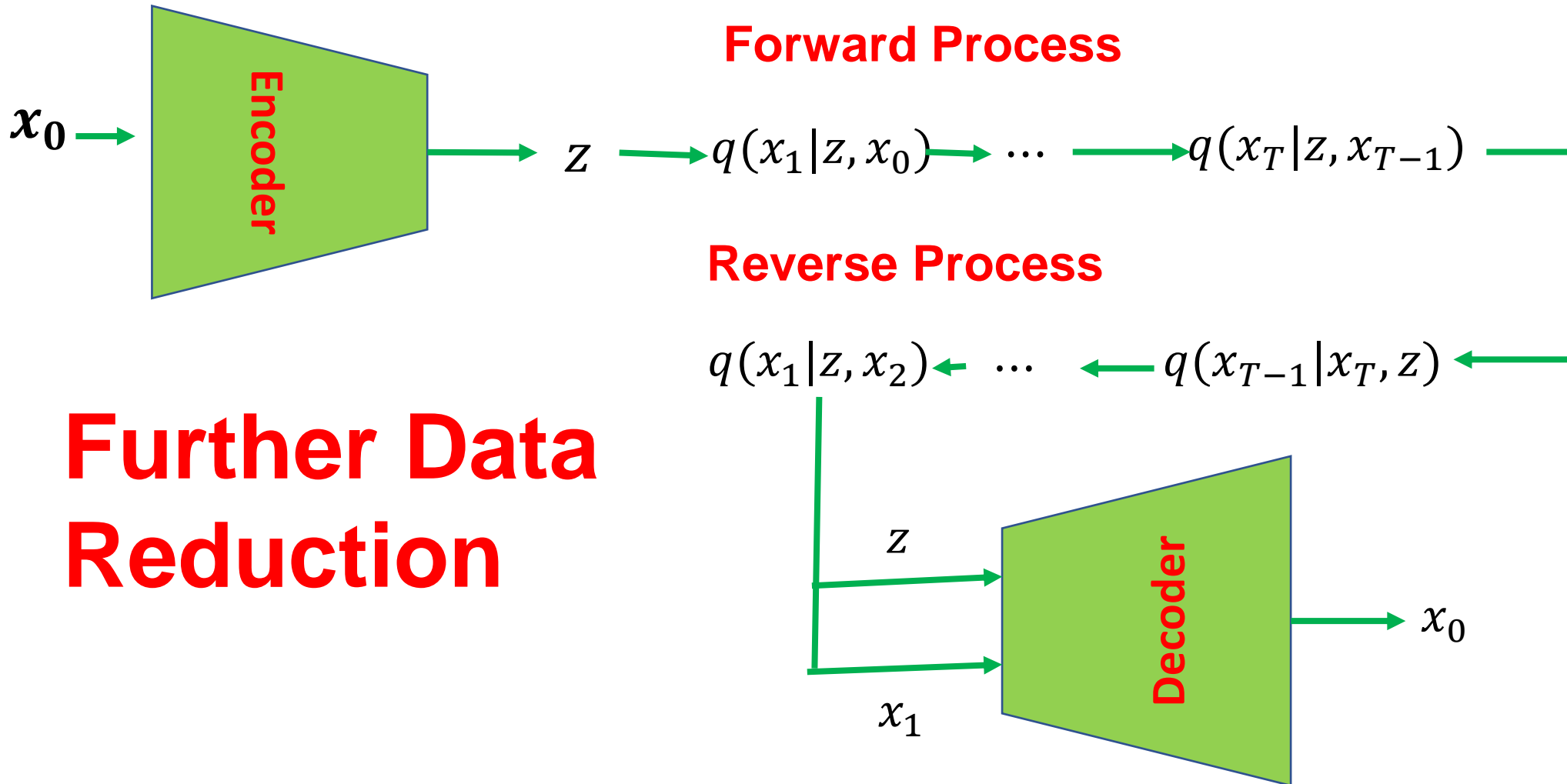
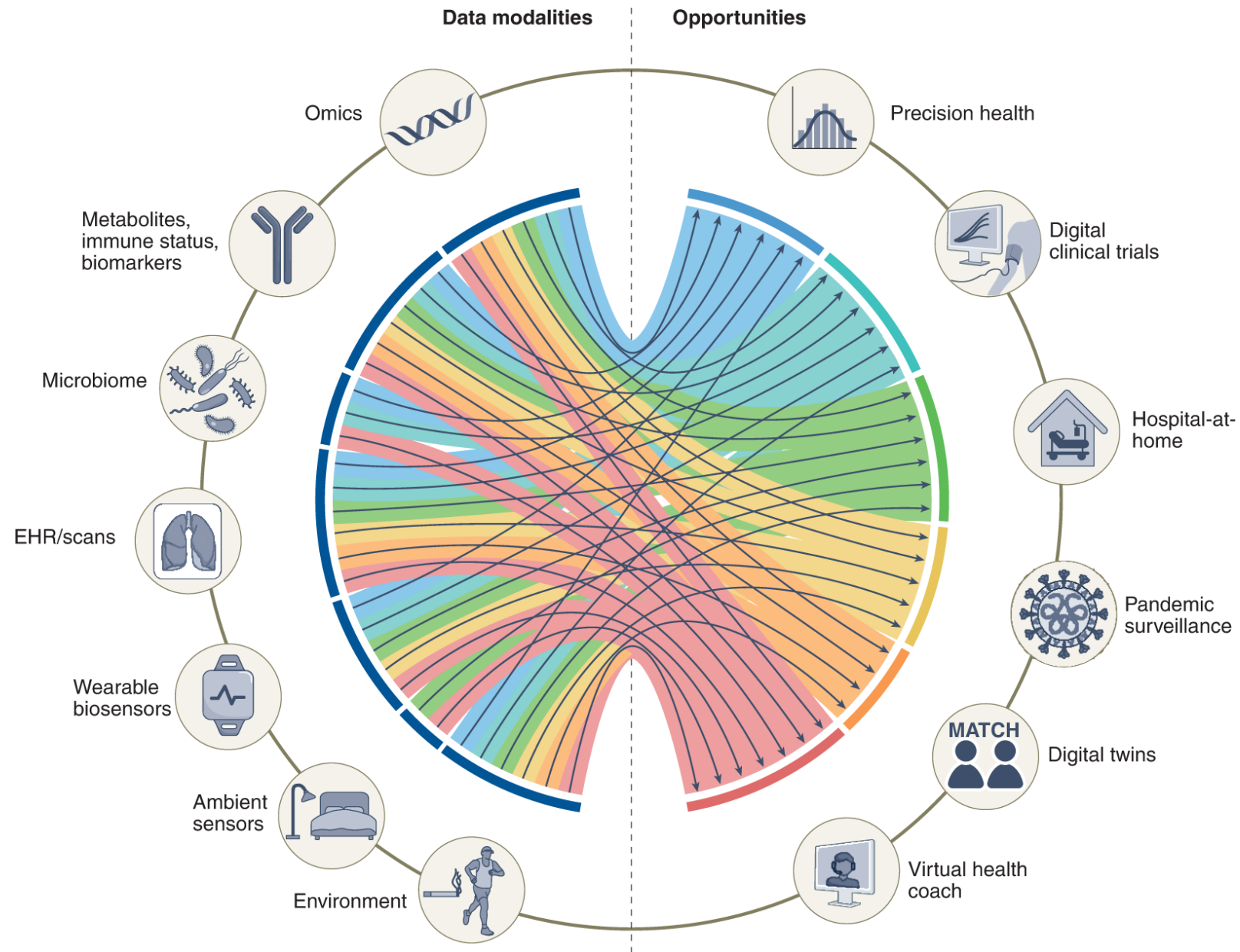


Figure 1: An overview of our system named “**Lodoss**.” It builds effective sentence representations by combining two essential tasks of section segmentation and sentence extraction. We introduce a new regularizer  $\mathcal{L}_{DPP}$  drawing on determinantal point processes to collectively measure the quality of a set of extracted sentences, ensuring they are informative and diverse.

Cho et al. 2022; Toward Unifying Text Segmentation and Long Document Summarization

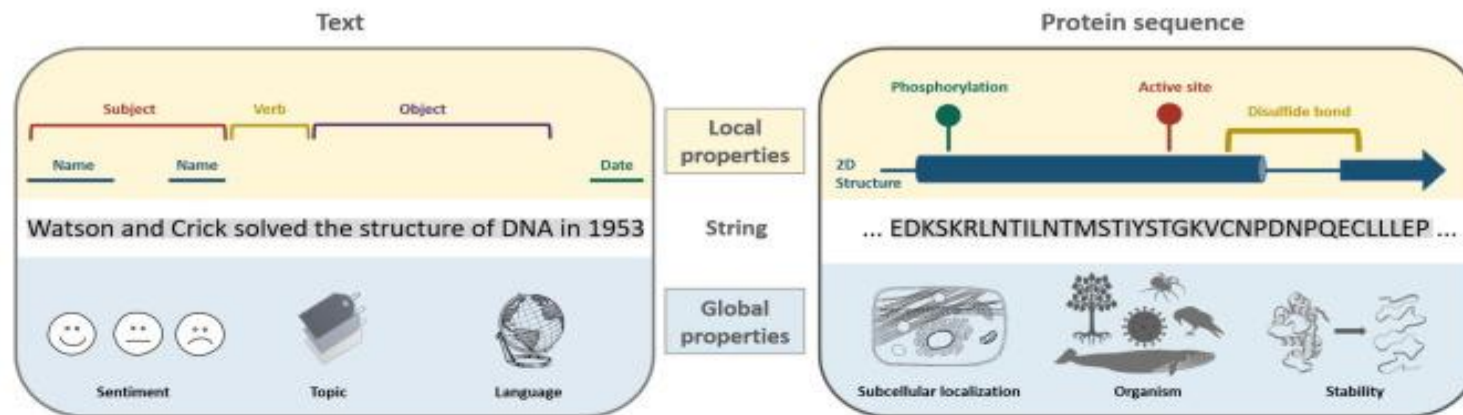
# Diffusion-VAE





# **New Paradigm 3: Tabular Data**

A



# Embedding Example

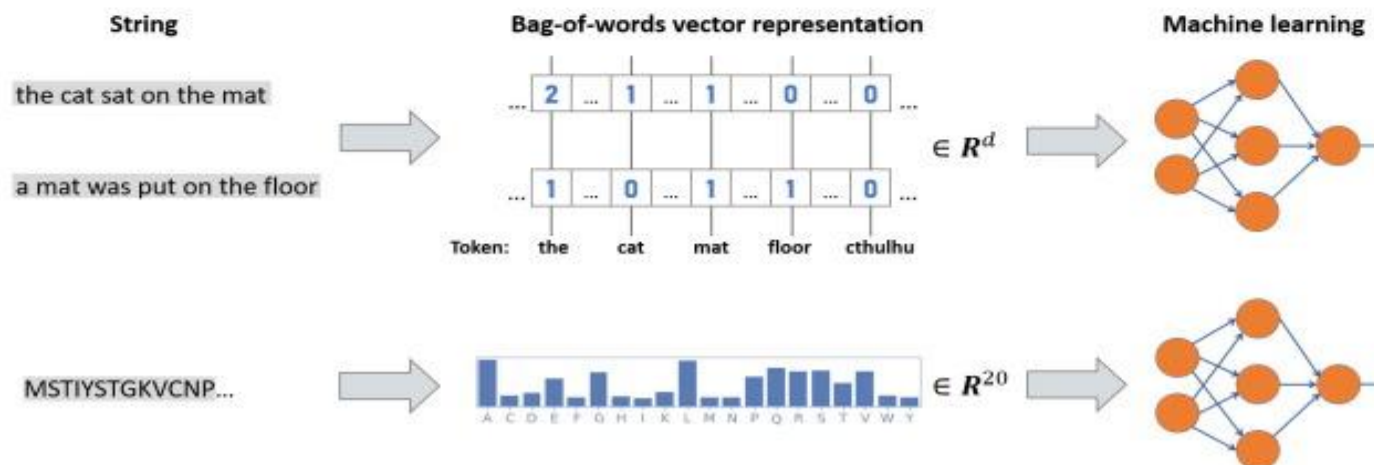
B

**String:** The cat sat on the mat      MSTIYSTGKVCNP...

**Possible tokenizations:**

[\*start\*] [T] [h] [e] [\*space\*] [c] [a] [t] ...      [\*start\*] [M] [S] [T] [I] [Y] [S] [T] [G] ...  
 [\*start\*] [The] [cat] [sat] [on] [the] [mat]      [\*start\*] [MS] [TI] [YS] [TG] ...  
 [\*start\* The] [cat sat on] [the] [mat]      [\*start\* M] [STI] [YST] [GK] [VCN] ...

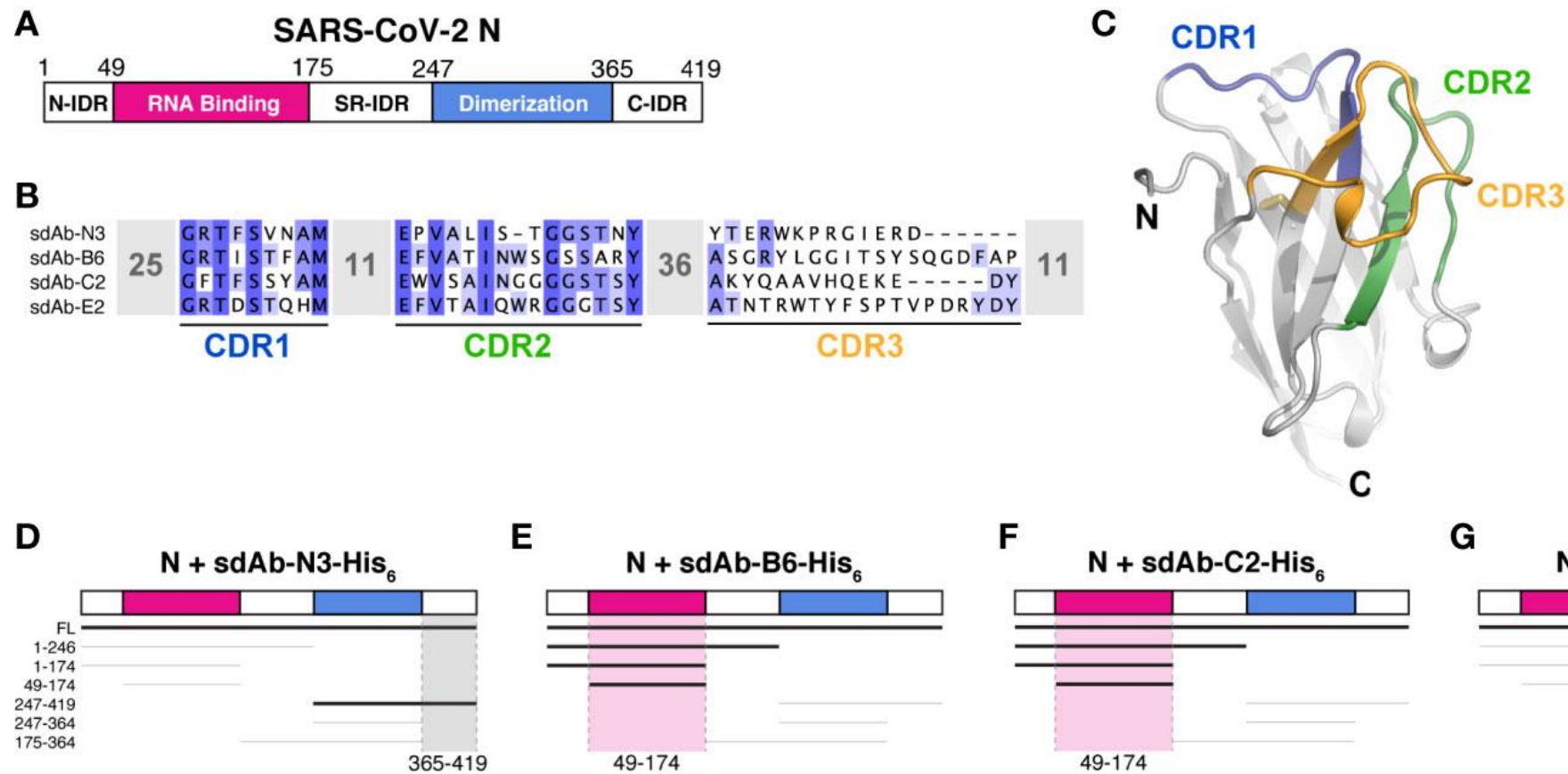
C



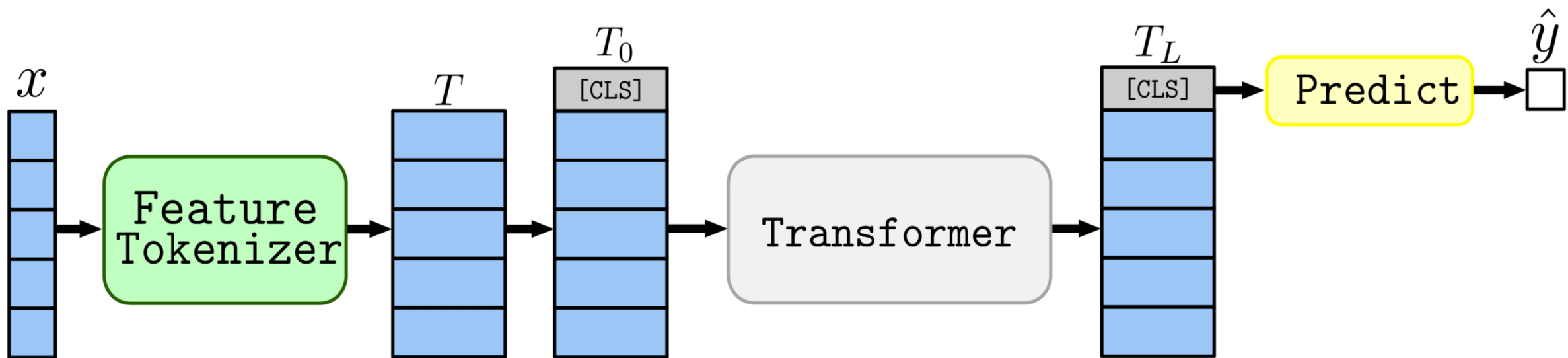


# Gene Expression Data

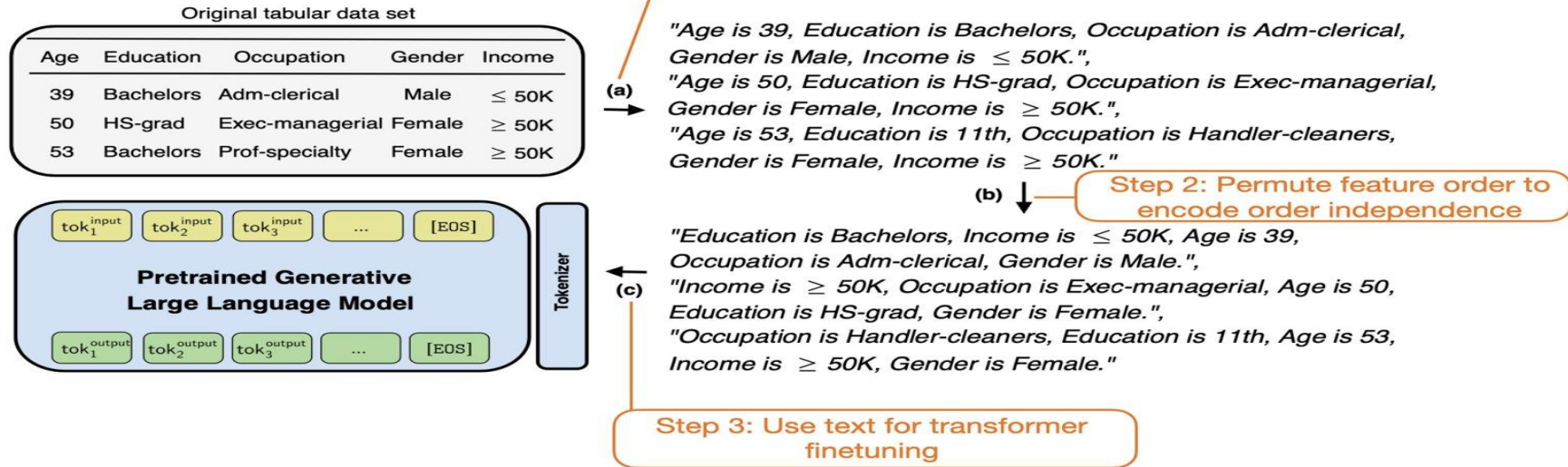
Sample	BTG2	FAS	MKK4	JNK7
1	0.405	0.326	0.234	0.348
2	0.089	0.293	0.192	0.123
3	0.459	0.125	0.543	0.334
4	0.123	0.389	0.238	0.651



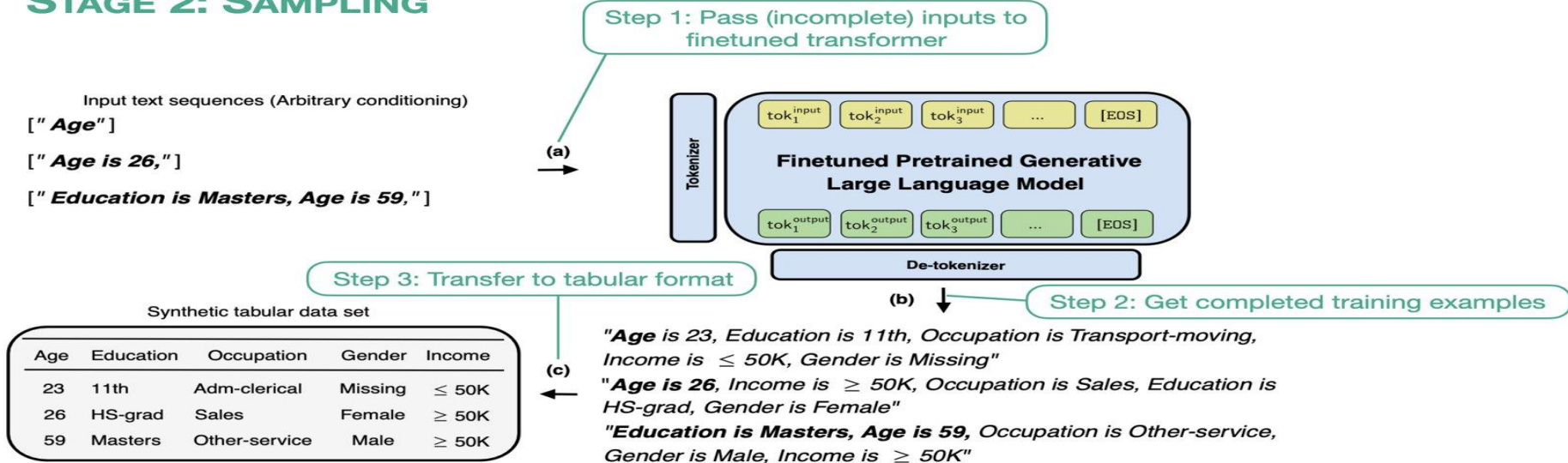
Sample	Protein Sequences								
1	G	R	T	F	S	V	N	A	0.673
2	G	R	T	I	S	T	F	A	0.543
3	G	F	T	F	S	S	Y	A0.485	



## STAGE 1: FINETUNING



## STAGE 2: SAMPLING



LANGUAGE MODELS ARE REALISTIC TABULAR DATA GENERATORS

Python package

**The code is accessible via `pip install be-great`**



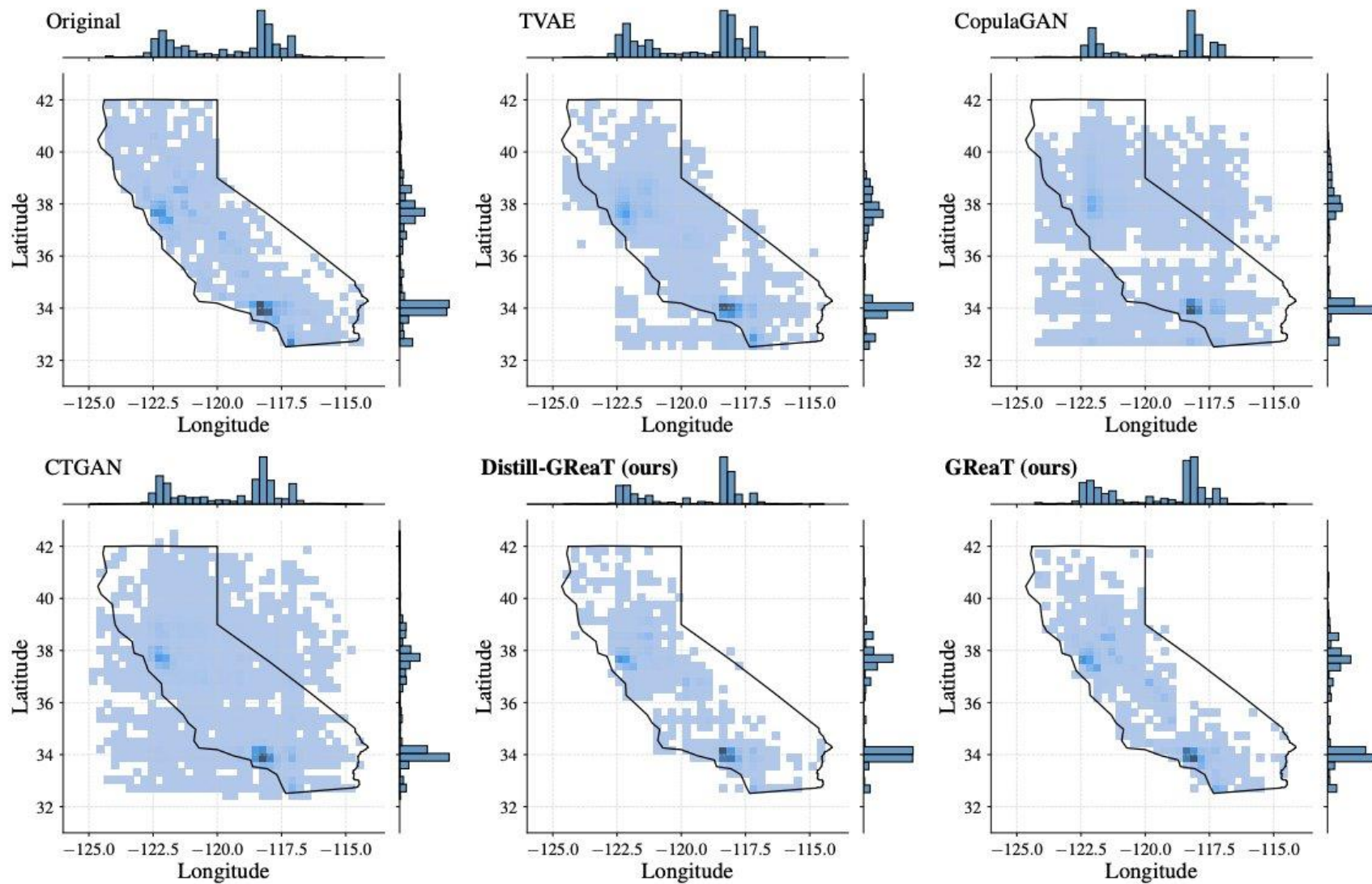
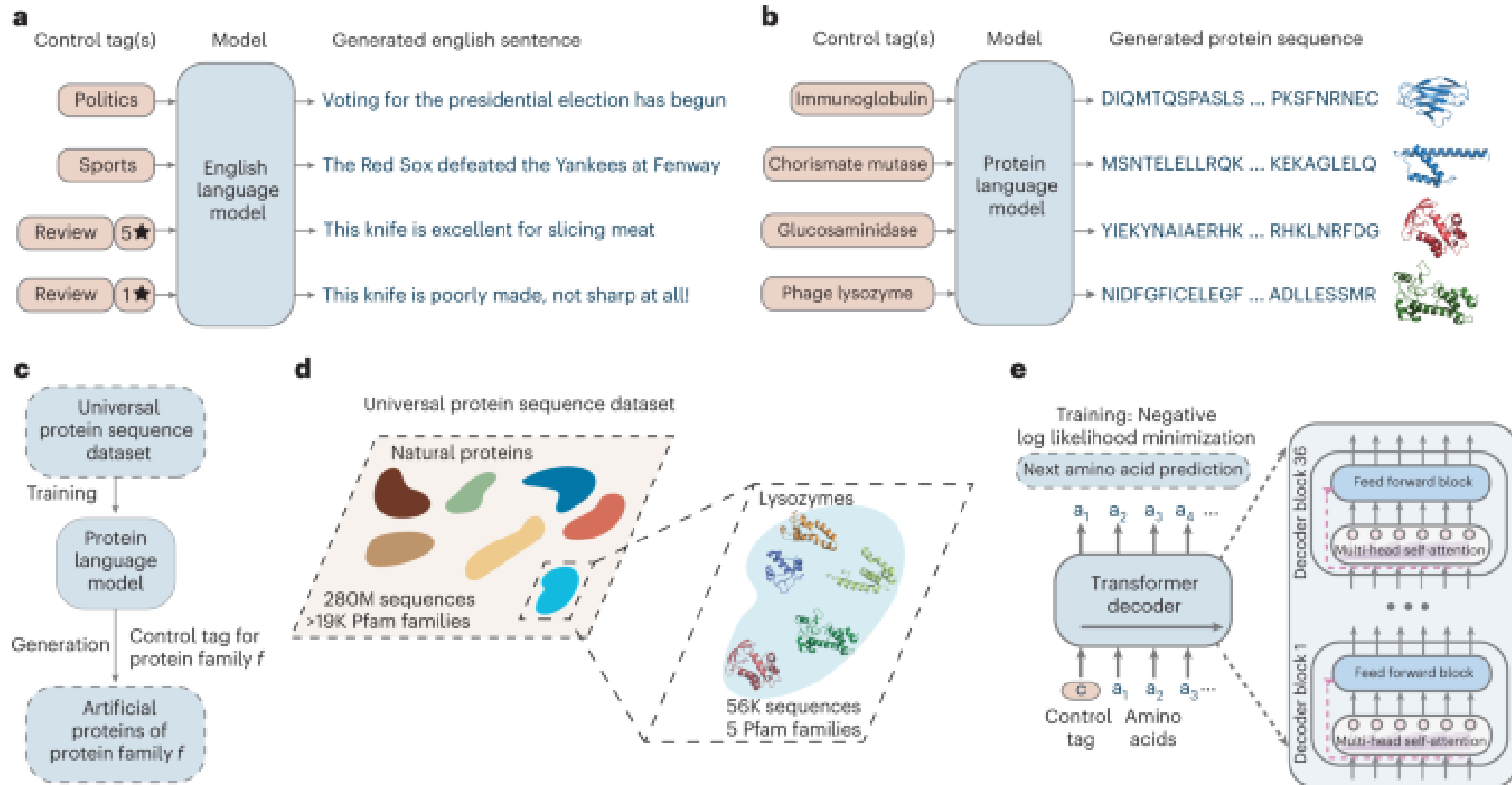
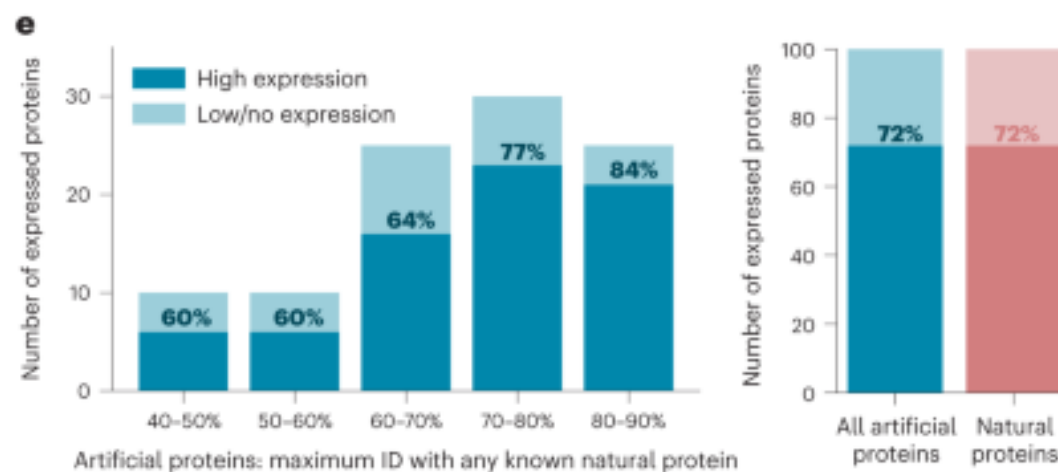
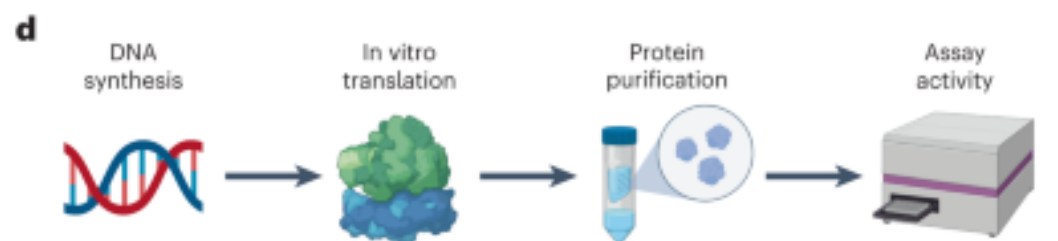
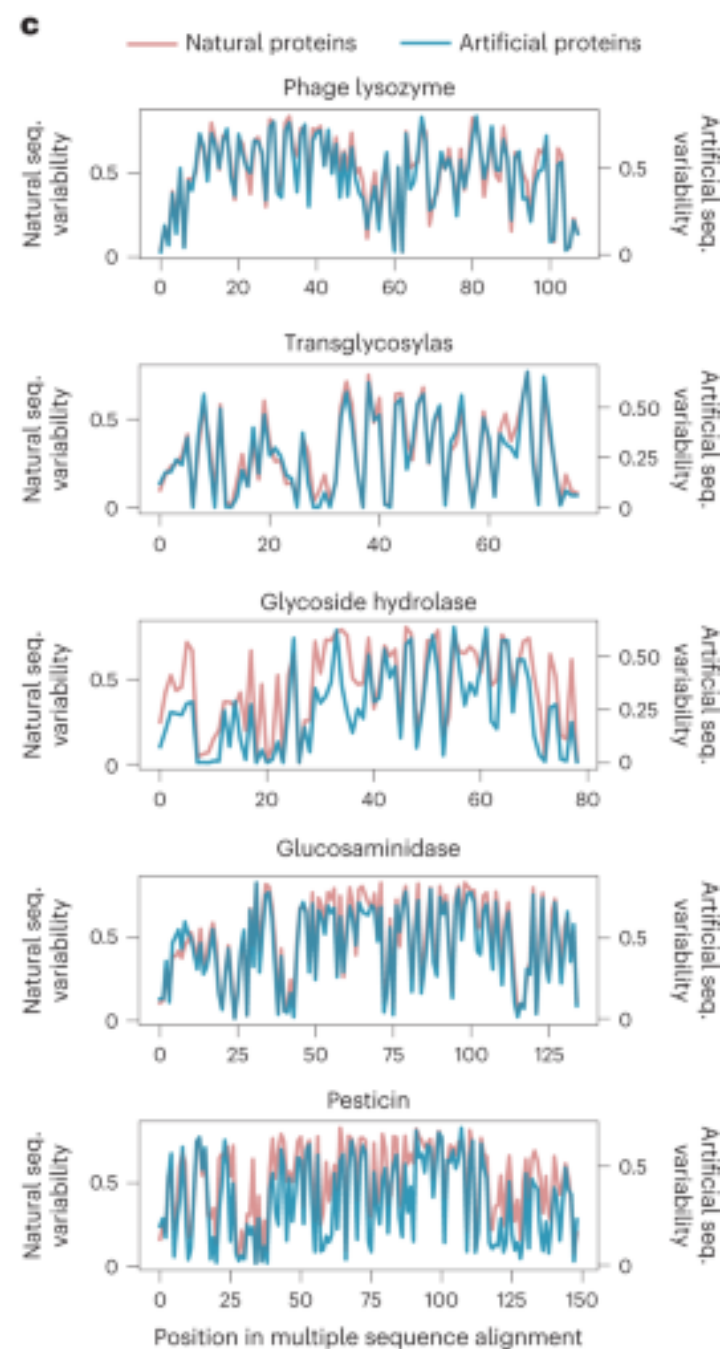
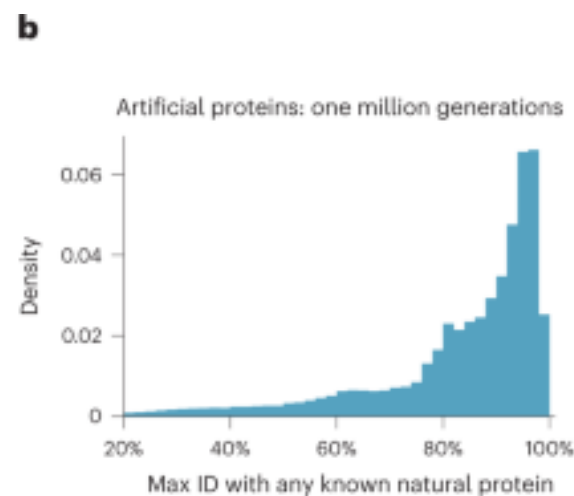
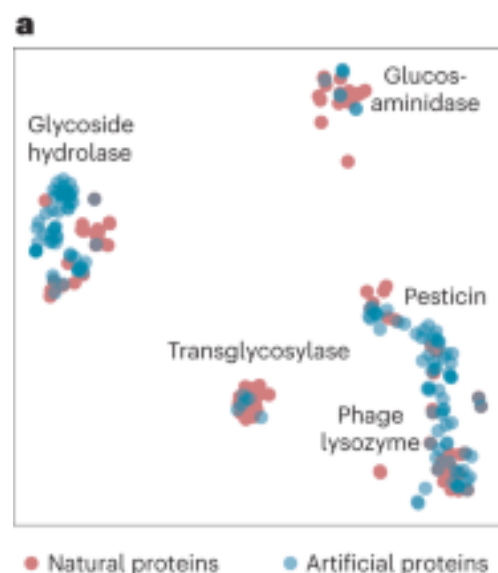
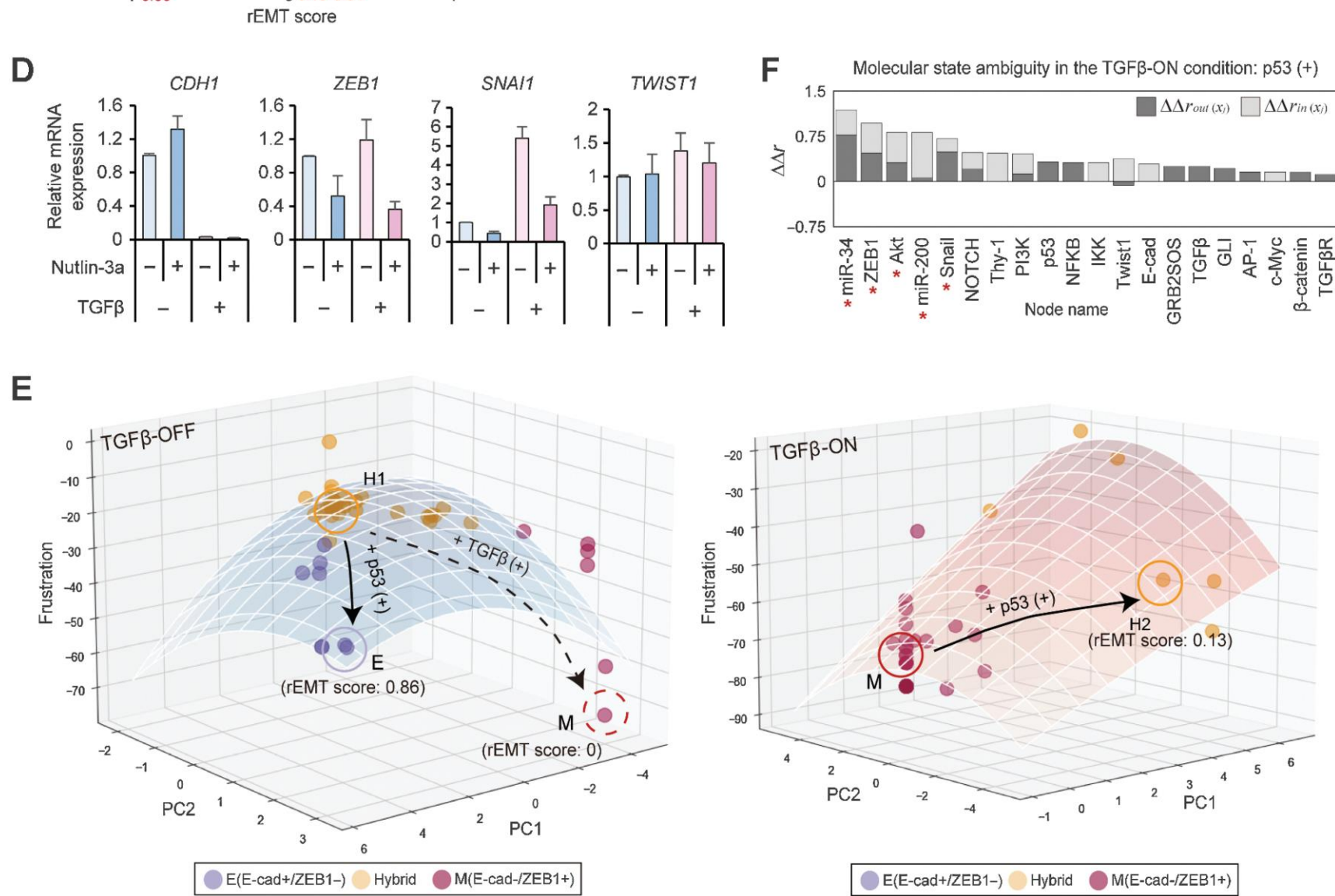


Figure 1: A comparison of the original and generated samples for the California Housing data set (Pace & Barry, 1997), which contains characteristic information about different properties in California, USA. We show joint histogram plots of the highly interconnected variables Latitude and Longitude. The black outline indicates the true boundary of the state of California.

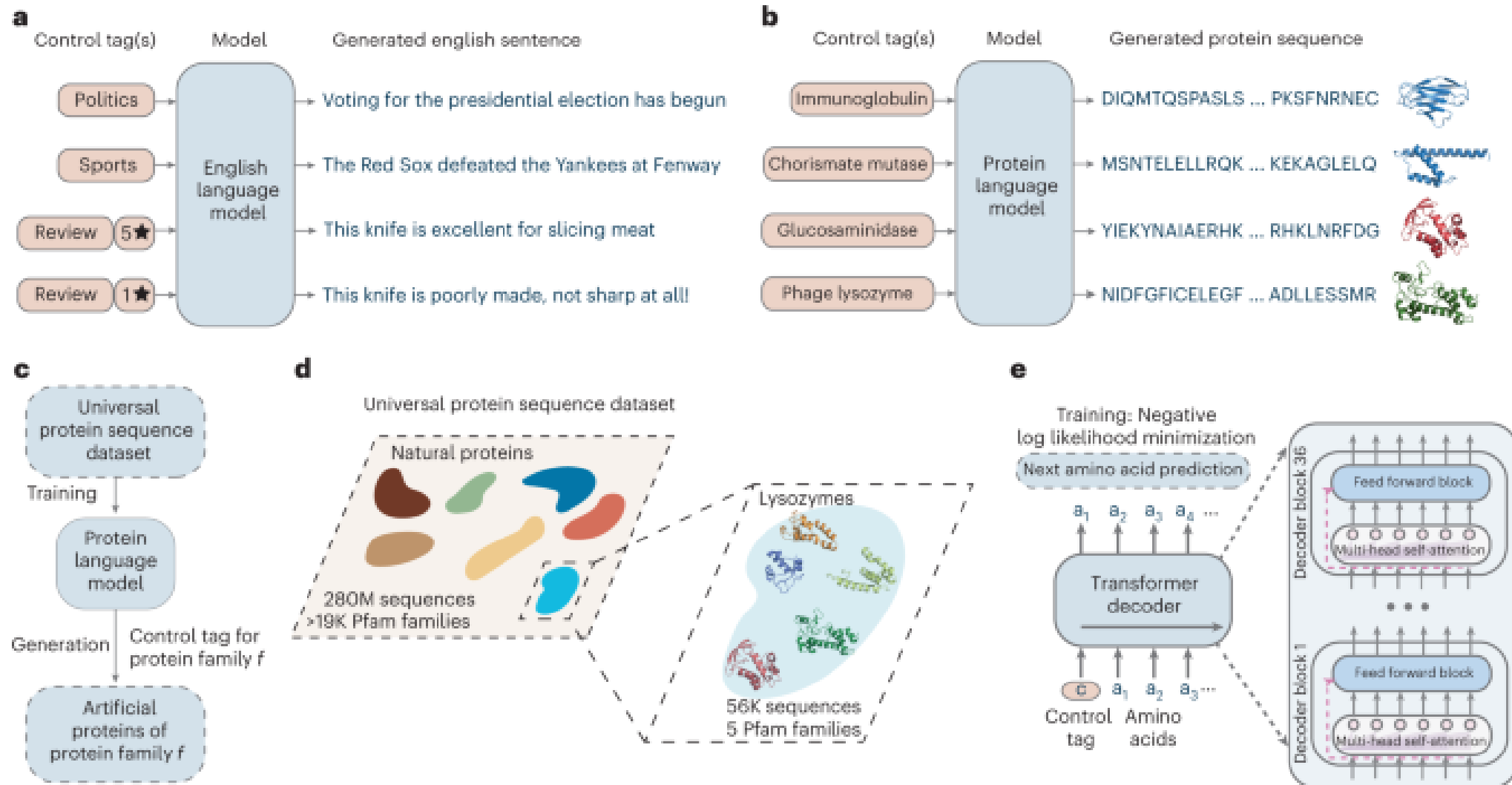


Madani et al. 2023; Large language models generate functional protein sequences across diverse families





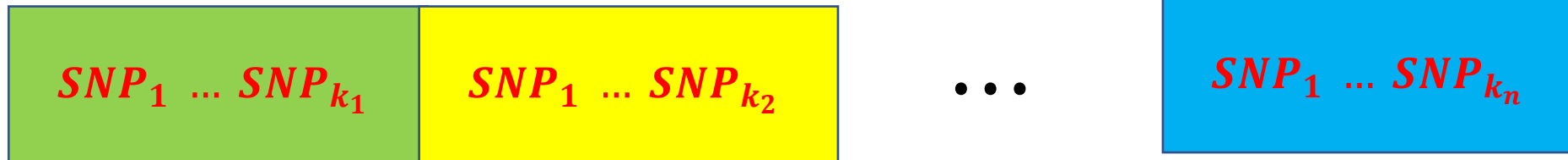
Kim et al. 2023; A Cell-Fate Reprogramming Strategy Reverses Epithelial-to-Mesenchymal Transition of Lung Cancer Cells While Avoiding Hybrid States



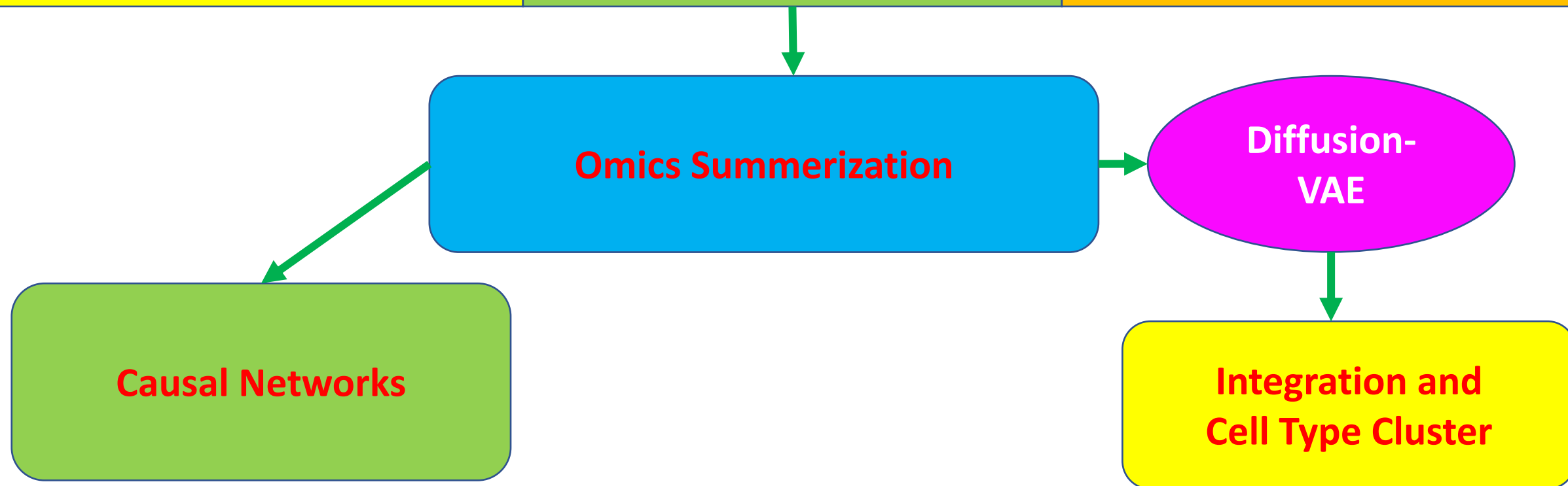
Madani et al. 2023; Large language models generate functional protein sequences across diverse families

# **Examples in Intelligent Analysis of Multi-Omics and Drug Discovery**





## Genome Summarization

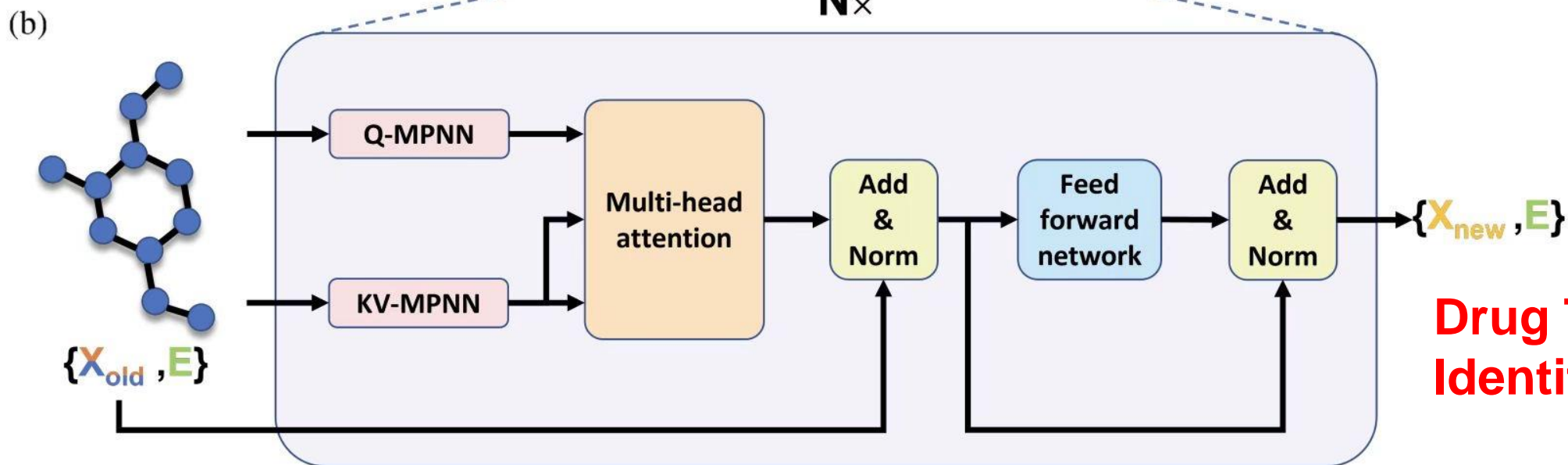
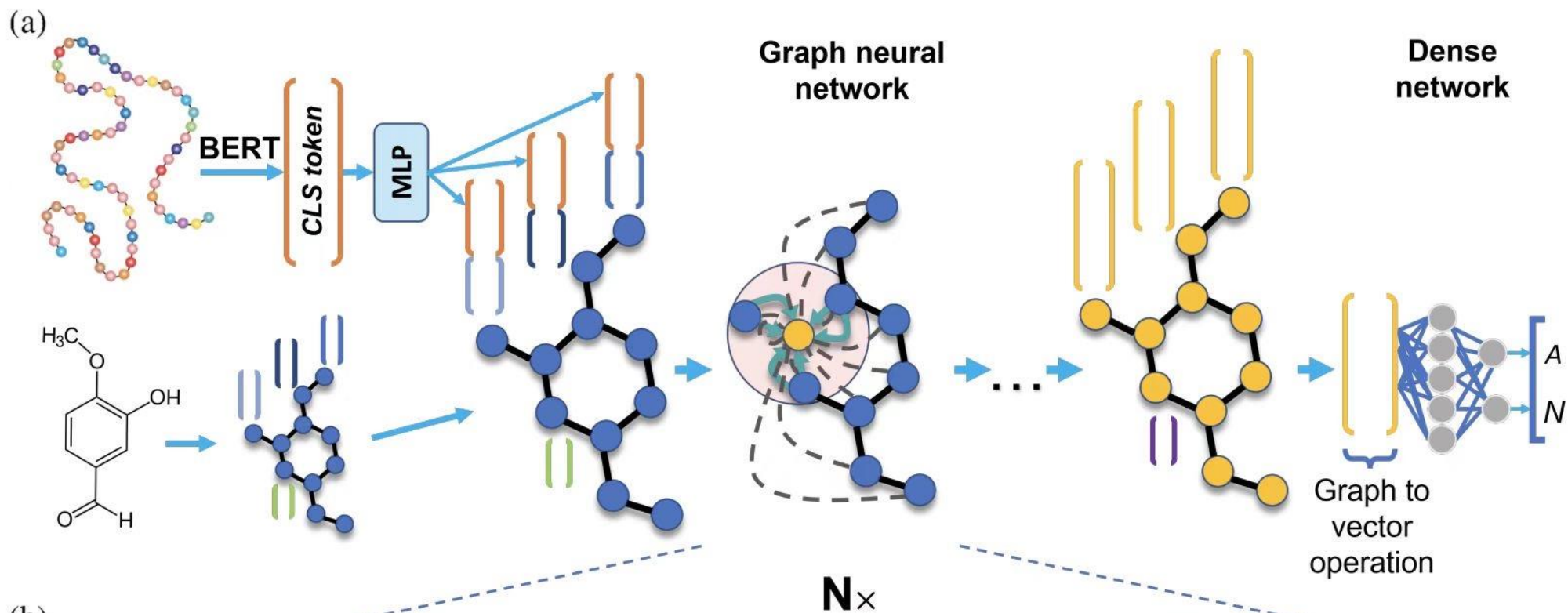


**Protein Language Model for Table Value**  
**DNA model for Table Value**  
**Summarization of Milti-omics in Cells**

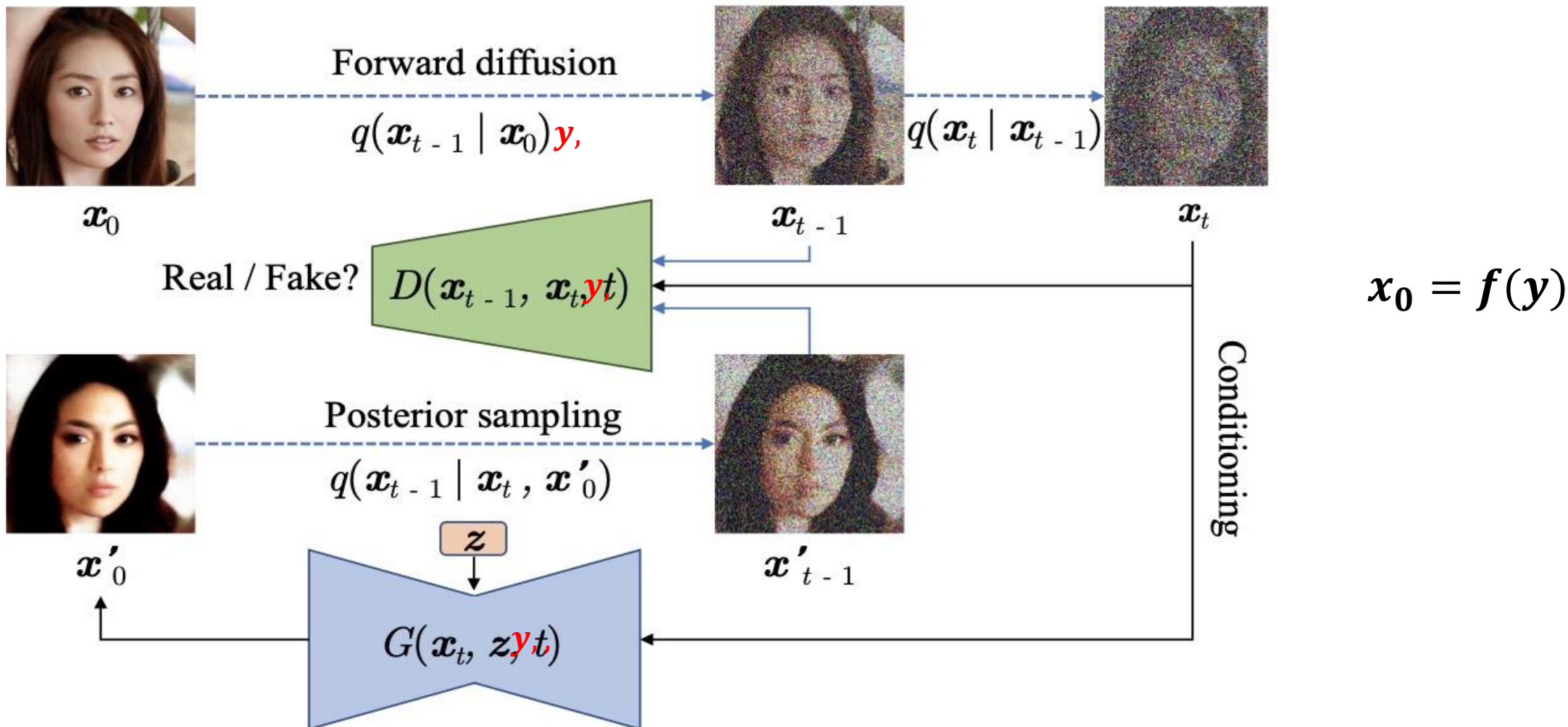
**Trait (Table Value)**  
**Blood Presxsure**

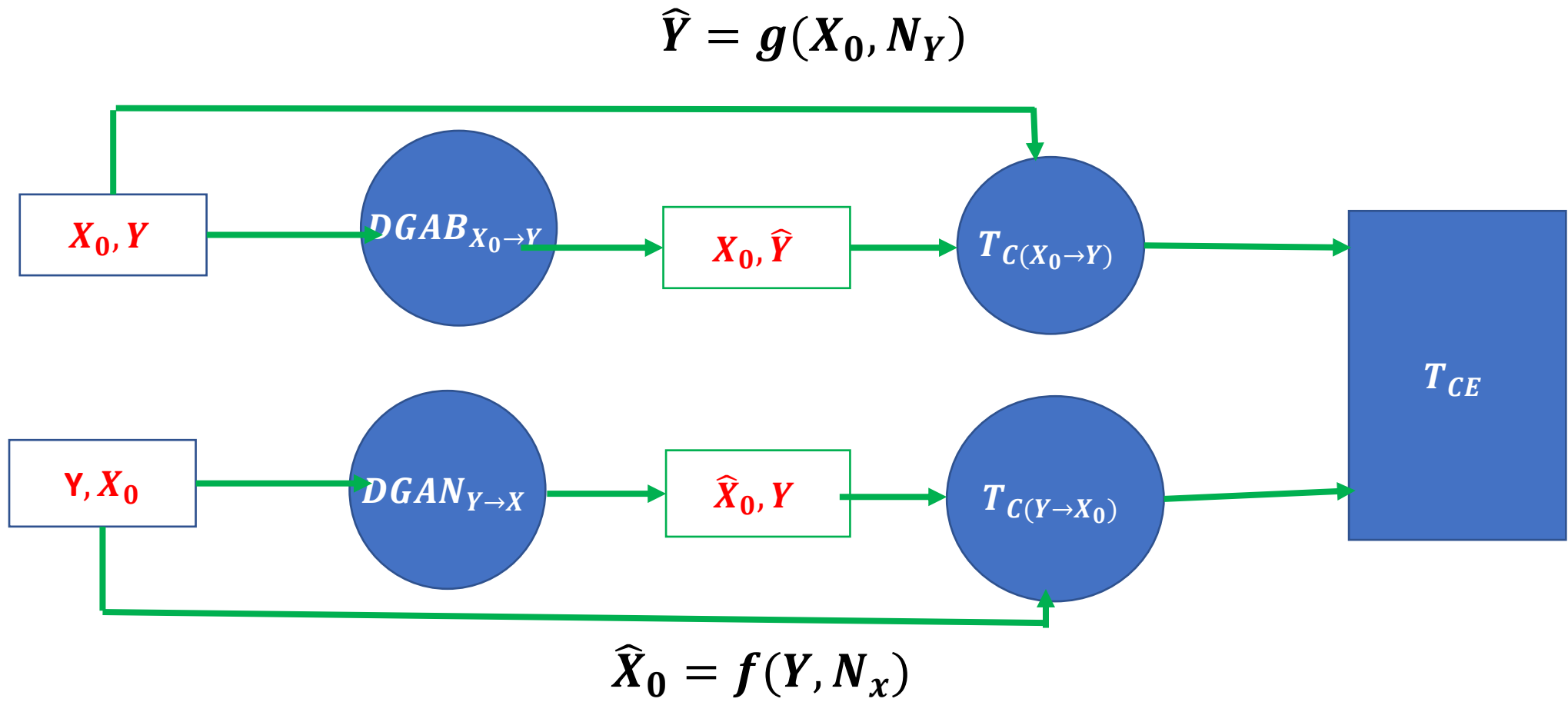
**Association  
or  
Causal  
Analysis  
(Cell Type)**

```
graph LR; A[Protein Language Model for Table Value  
DNA model for Table Value  
Summarization of Milti-omics in Cells] --> C((Association or Causal Analysis (Cell Type))); B[Trait (Table Value)  
Blood Presxsure] --> C;
```



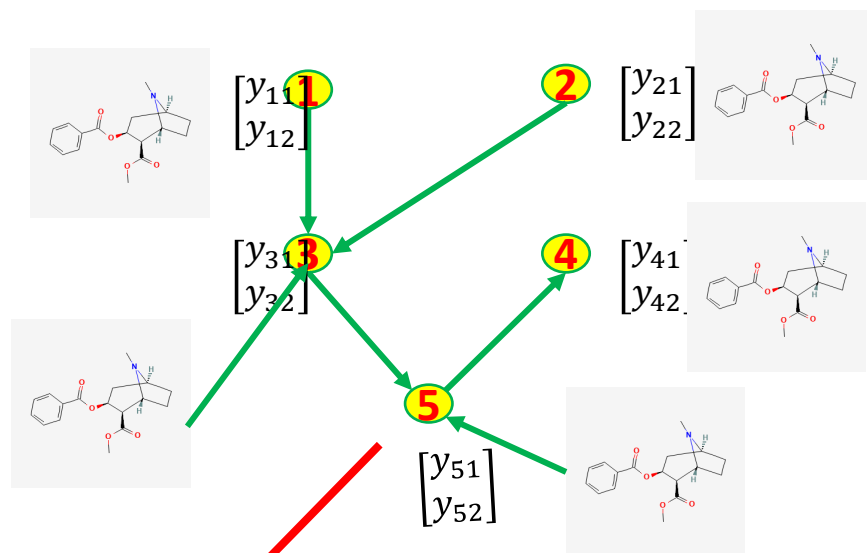
# Denoising Diffusion GAN



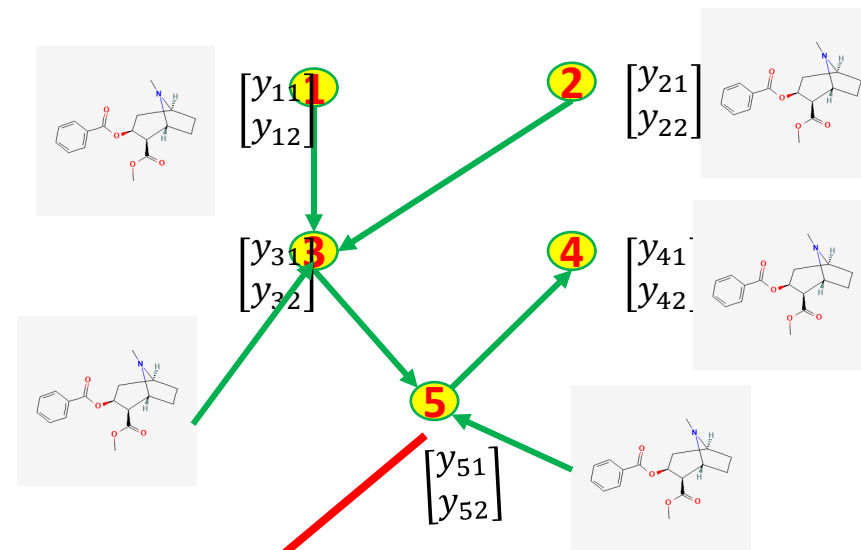


Scheme of classifier two sample test for causation using denoising diffusion GAN.

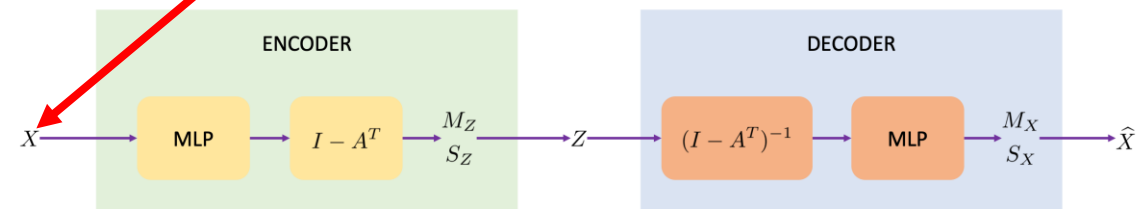
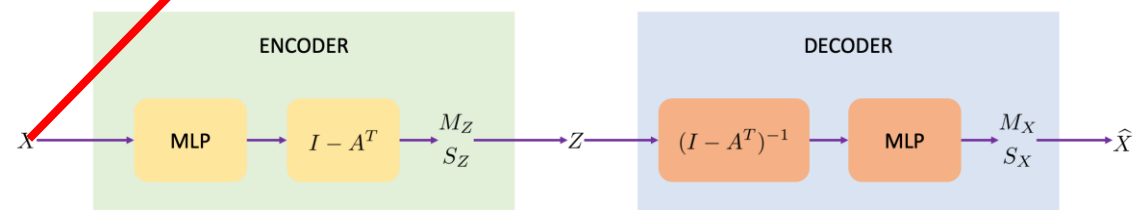
## Normal



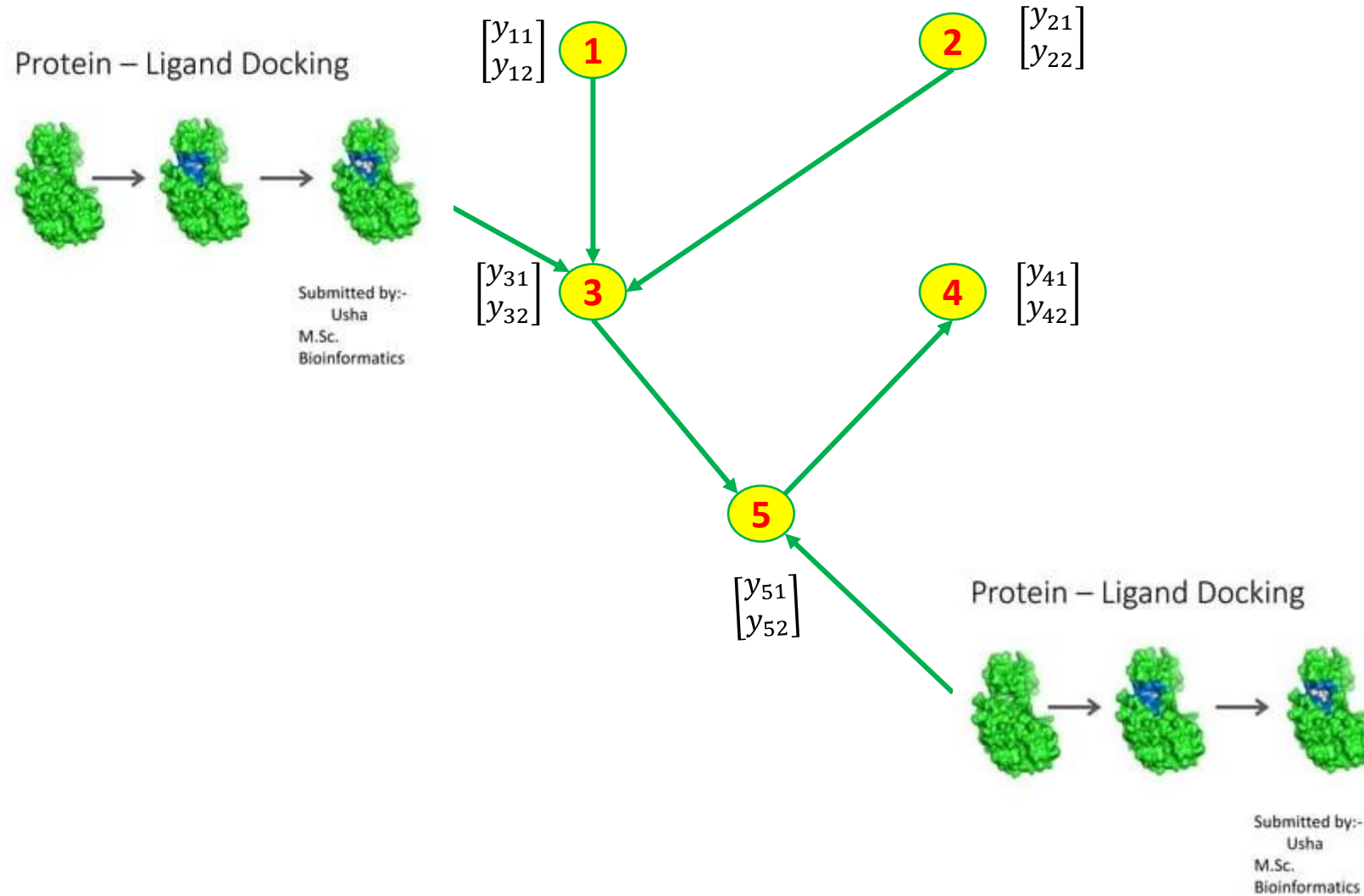
## Disease



## Infer Causal Networks



$$X = A^T X + Z$$

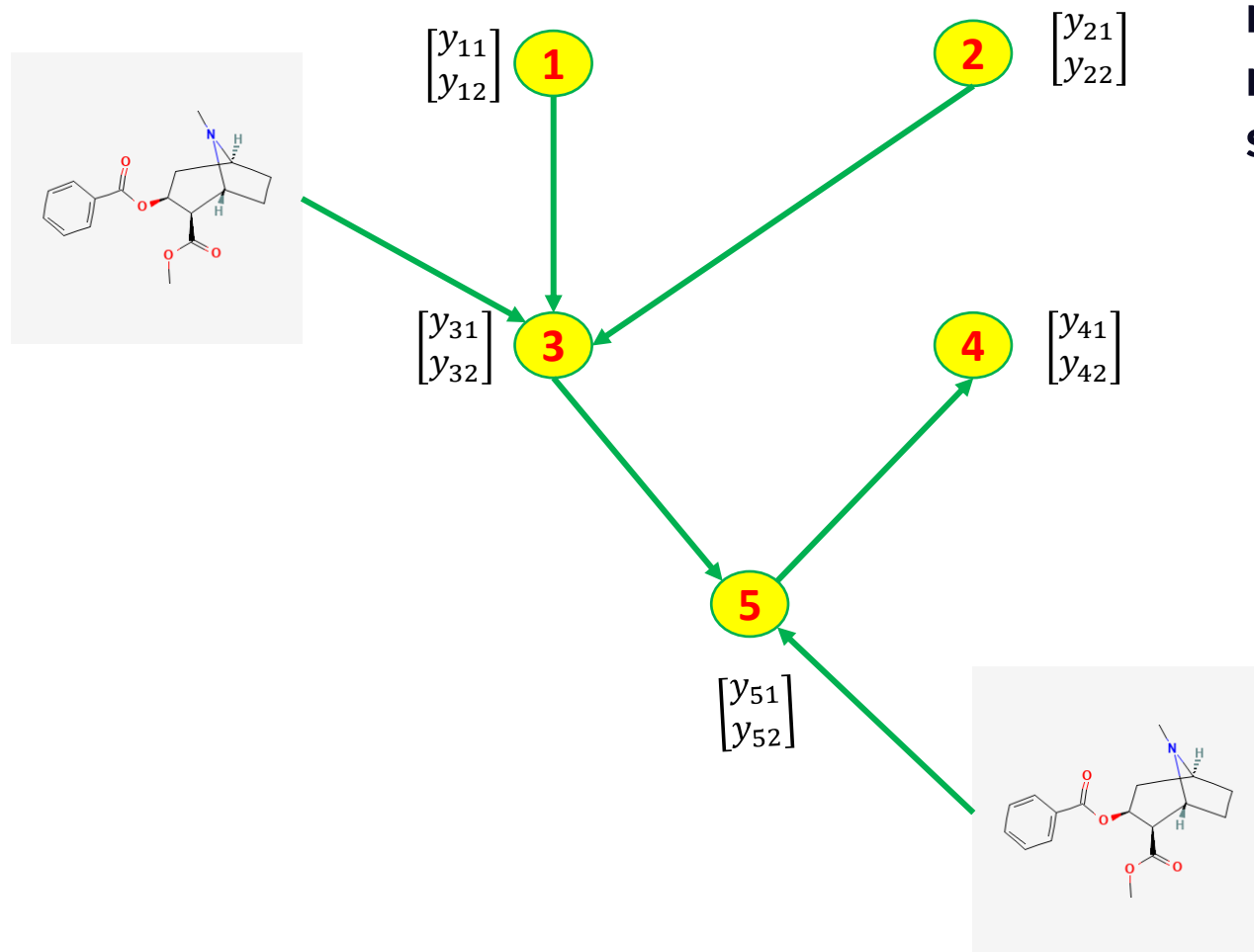


# Drug Target Causal Networks



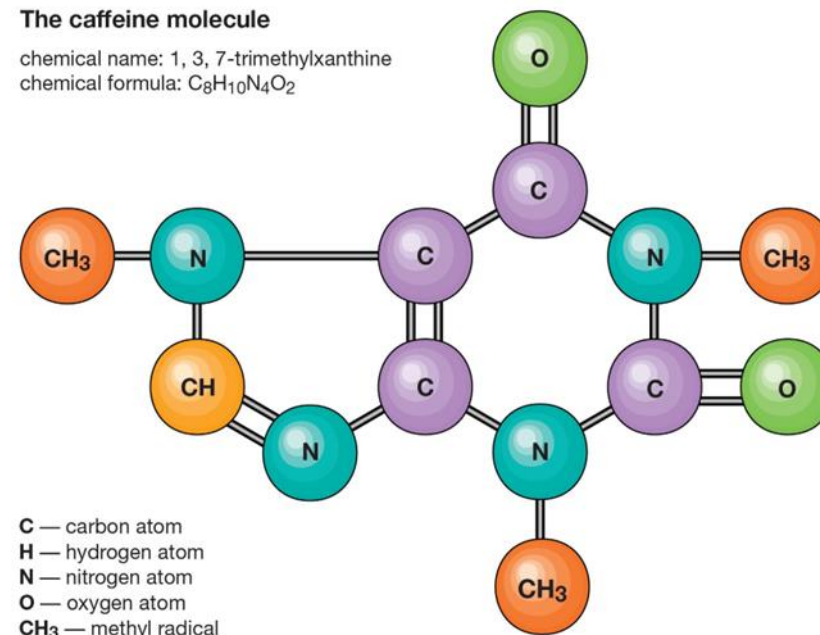
# Graphic Neural Networks (GNN)

Probably the most common application of representing data with graphs is using molecular graphs to represent chemical structures

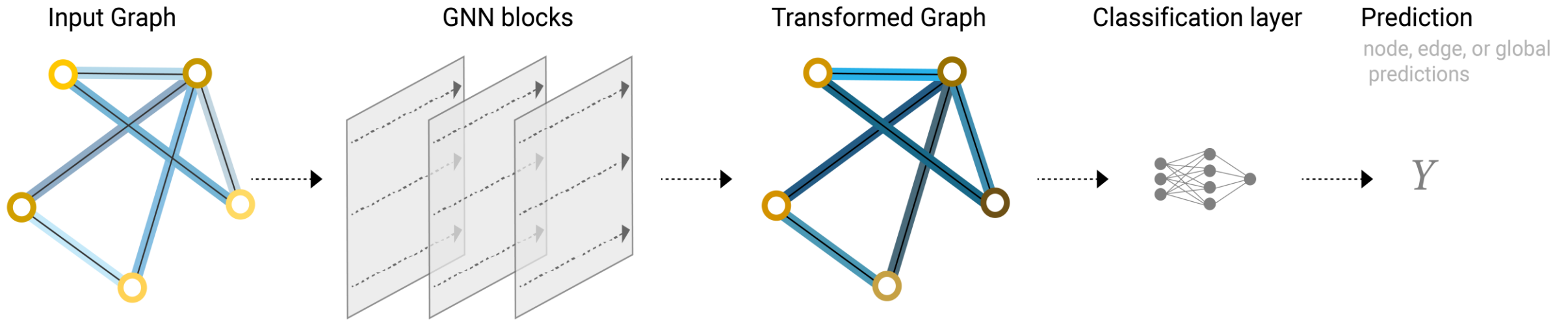


The caffeine molecule

chemical name: 1, 3, 7-trimethylxanthine  
chemical formula:  $C_8H_{10}N_4O_2$



# Pipelines of Graph Neural Networks



$$h_G = \text{READOUT}(\{h_v^L, v \in \mathcal{V}\})$$

Distill, 2021; A Gentle Introduction to Graph Neural Networks. <https://distill.pub/2021/gnn-intro>

# Directed Acyclic Graph Neural Networks

- A DAG is a directed graph without cycles
- updating node representations based on those of all their predecessors sequentially, such that nodes without successors digest the information of the entire graph.

$$AGG_v^l = \sum_{u \in \mathcal{P}(v)} \alpha_{vu}^l(h_v^{l-1}, h_u^l) h_u^l$$

$$\alpha_{vu}^l(h_v^{l-1}, h_u^l) = \text{softmax}_{u \in \mathcal{P}(v)}((w_1^l)^T h_v^{l-1} + (w_2^l)^T h_u^l + (w_3^l)^T y(u, v))$$

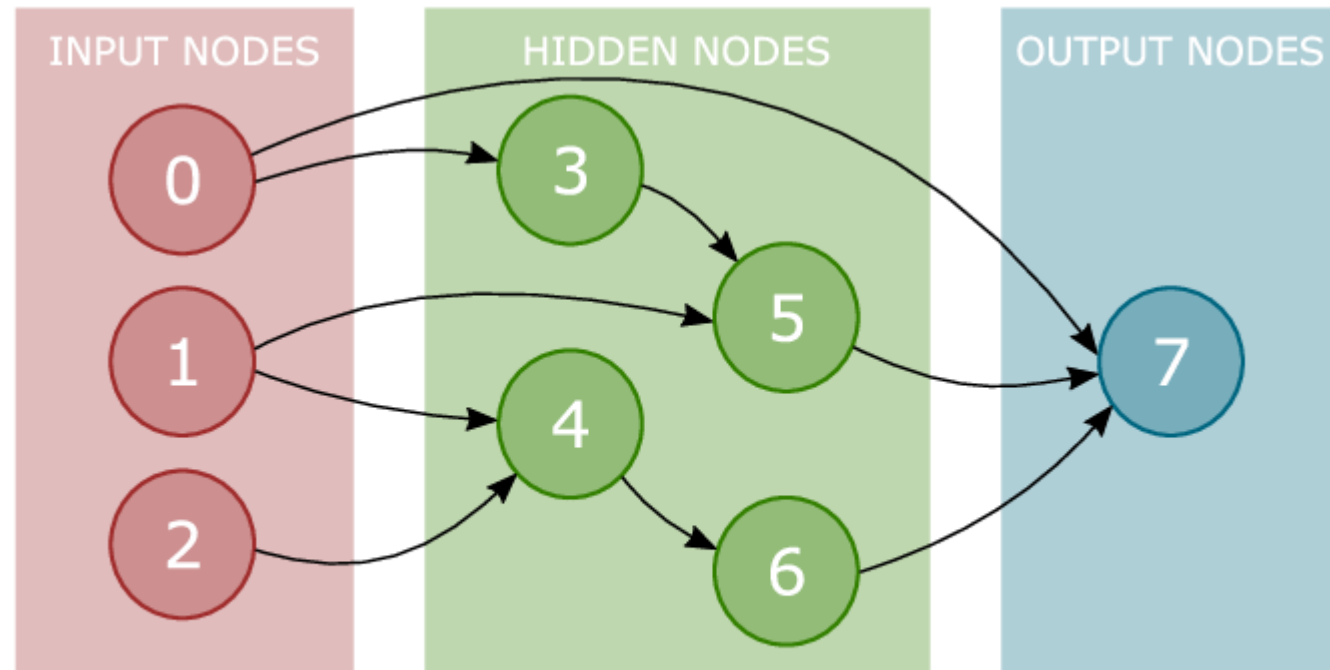
$$h_v^l = F^l(h_v^{l-1}, AGG_v^l) = GRU^l(h_v^{l-1}, AGG_v^l)$$

$$h_G = FC(\max_{v \in \mathcal{T}} \text{pool}(\|_0^L h_v^l,))$$

$$\| \max_{u \in S} \text{pool}(\|_0^L \tilde{h}_u^l))$$

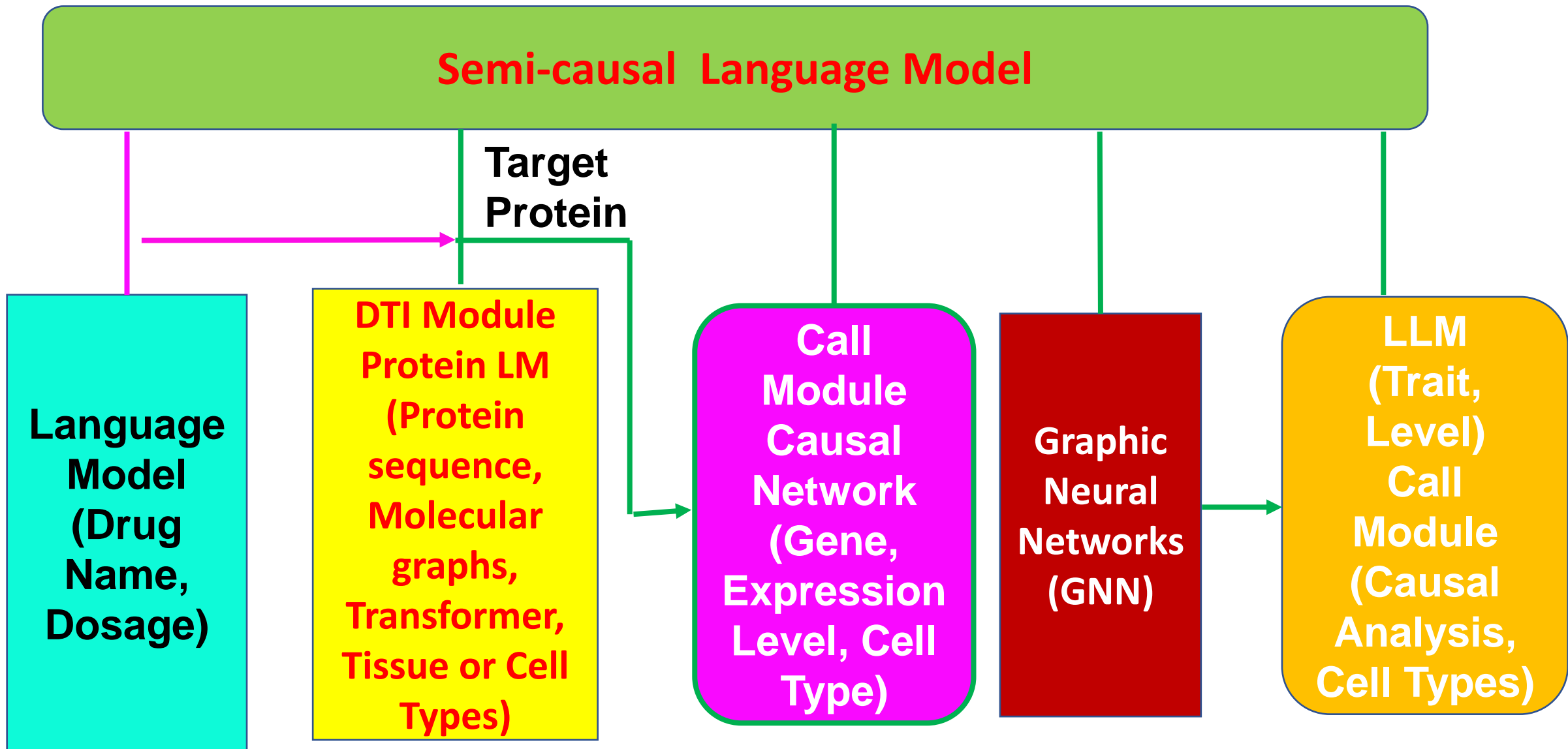
$$\mathcal{P}(v) = \text{set of preceeding nodes}$$

This also allows producing a single output for the whole graph



**FIGURE 1.** An example phenotype for a Directed Acyclic Graph Neural Network (DAG-NN).

# Automatic Drug Discovery Analysis



# Discussion

**AI needs team of Works**

**How to work together with**

**AI for Science? Statistics, Biology,  
Economics, Physics and Chemistry?**