第一章 生物大模型介绍

让大家都了解大模型基础应用

以及如何应用技术完成我们的科研任务

作业布置 & 参考资料

- 作业：
  - 1. 思考题： 在生物大模型微调中，为何需要使用领域特异性数据集？试举例说明？
  - 2. 讨论题：在DNA大模型设计中，如何平衡预训练数据量与计算资源限制？提出至少两种优化策略。
- 参考资料：
  - 1. 学术论文：Genomic Language Models: Opportunities and Challenges
  - 2. 实践教程：CSDN博客《到底如何理解文本？一文读懂命名实体识别（实体消歧和实体统一）》

任务1：文本识别（应用大模型）

什么是实体识别（问题的背景）

为什么模型可以学会语言编码（如何预处理数据和构建模型）

用大语言模型如何完成任务（如何利用和微调）

大语言模型的智能问答（进一步的应用，如Agent和RAG）

任务2：DNA大模型（设计大模型）

生物大模型是什么

需要什么样的数据集

如何进行

下游任务设计

学习规划

Day.1: 知道整一个深度学习模型的训练流程，常规的模型，常规的流程

Day.2: 使用lightning，将第一天的模型包装起来

Day.3: SNP预测表型的实现 (小麦数据集 gMLP)

给定模型，自己写dataset dataloader 和 lightning 模型，实现它的训练和推理。

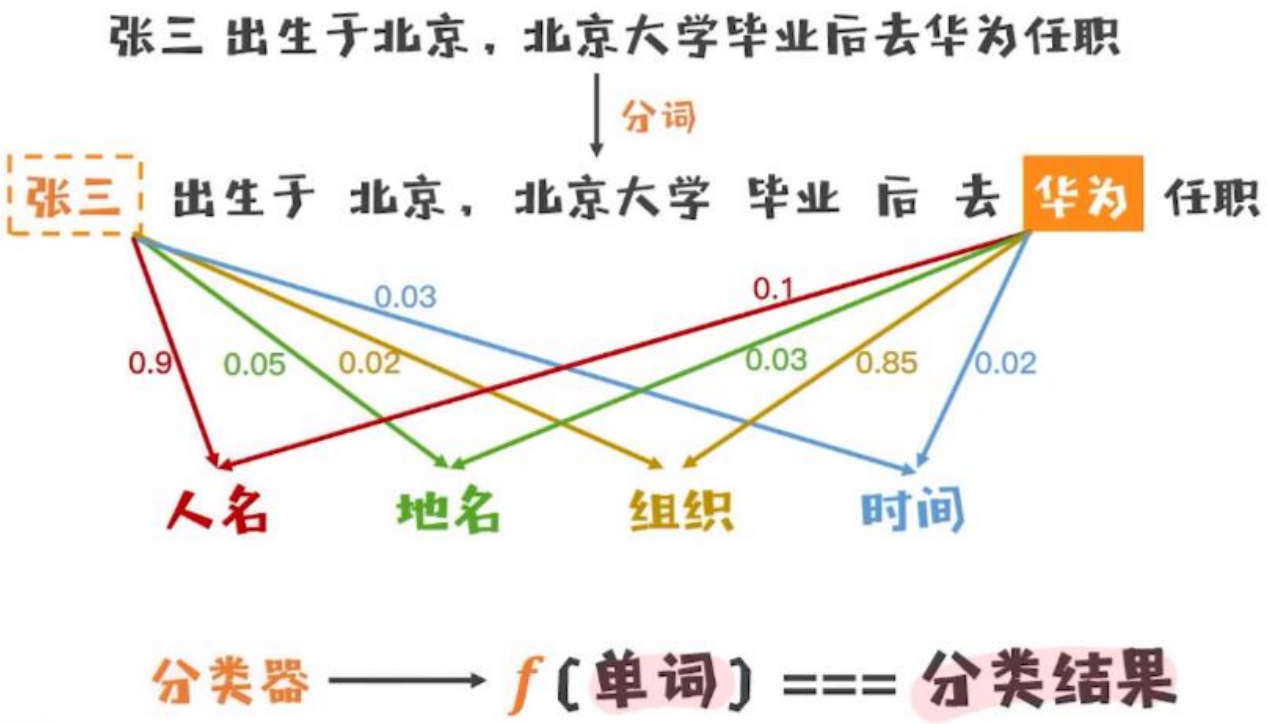
Day.4: Hydra的应用，将之前的模型给进一步封装

Day.5: DNA bert2 的预训练和微调实现

Day.6: HyenaDNA的预训练和微调实现

什么是实体识别
（问题的背景）

命名实体识别 – 分类问题

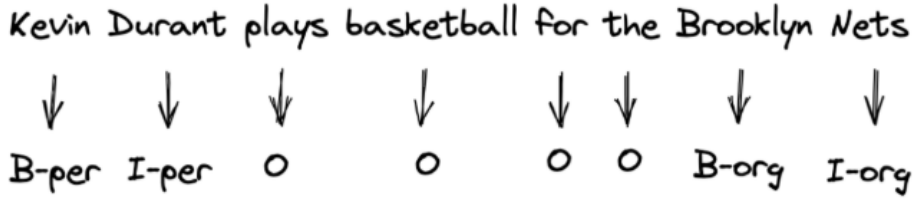


We have recently identified promoter DNA methylation of the genes *ADAMTS1* and *BNC1* as potential blood biomarkers of pancreas cancer

DNA methylation status: **methylation**

Gene: *ADAMTS1, BNC1*

Cancer type: **Pancreas cancer**



- "Kevin"有 `B-pers` 标签，它是个人实体的开始
- "Durant"有 `I-pers` 标签，它是个人实体的延续
- "Brooklyn"有 `B-org` 标签，它是一个组织实体的开始
- "Nets" 有 `I-org` 标签，它是组织实体的延续
- 其他词被分配 `O` 标签，它们不属于任何实体

CoNLL-2003

到底如何理解文本？一文读懂命名实体识别（实体消歧和实体统一）_文本实体识别-CSDN博客

2. NER 常用模型与方法

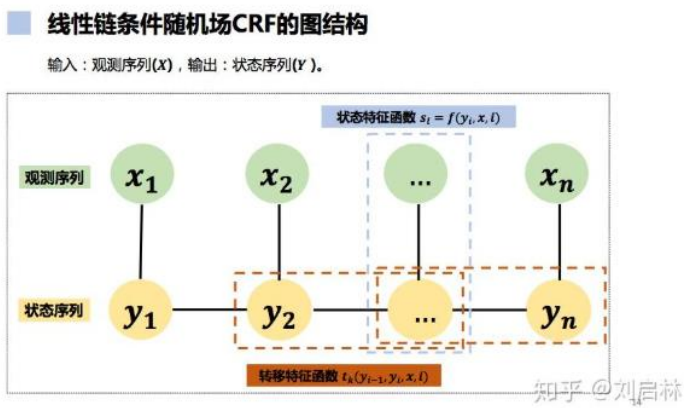
2.1 传统方法

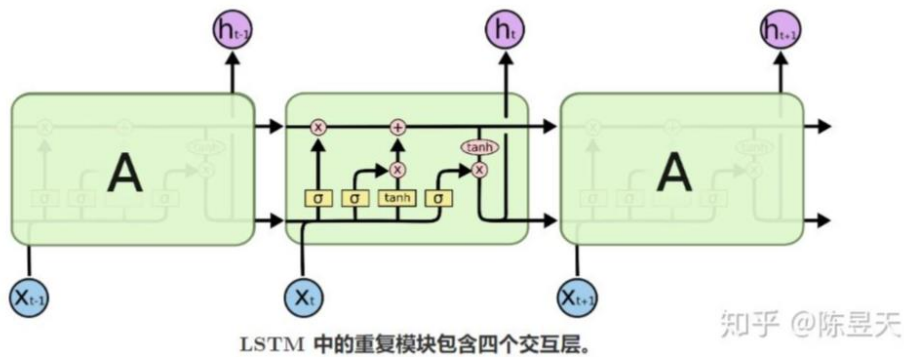
在深度学习普及之前，NER 主要依赖于基于规则的方法和机器学习模型，如：

- **基于规则的方法**：通过手动编写规则，利用正则表达式等技术识别实体。
- **条件随机场（CRF）**：一种常用的序列标注方法，能够结合上下文信息进行实体识别。

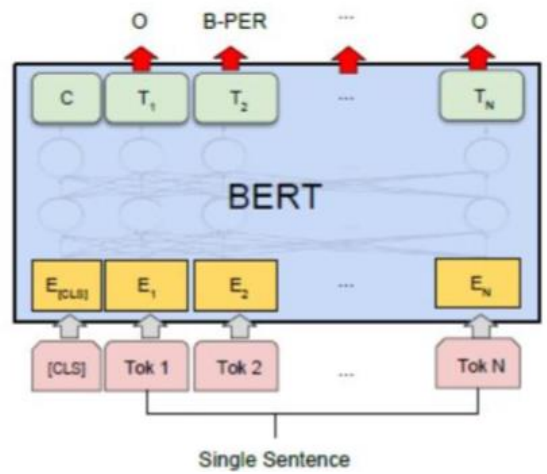
2.2 深度学习方法

近年来，深度学习方法在 NER 任务中取得了显著进展，特别是基于 **LSTM** 和 **BERT** 等预训练模型的方式，已经成为 NER 任务的主流方法。





LSTM 中的重复模块包含四个交互层。



[自然语言处理系列（3）——命名实体识别（NER）详解与实战_自然语言处理中命名实体识别中时间用什么字母标识-CSDN博客]

[LSTM - 长短期记忆递归神经网络 - 知乎] [CRF条件随机场的原理、例子、公式推导和应用 - 知乎]

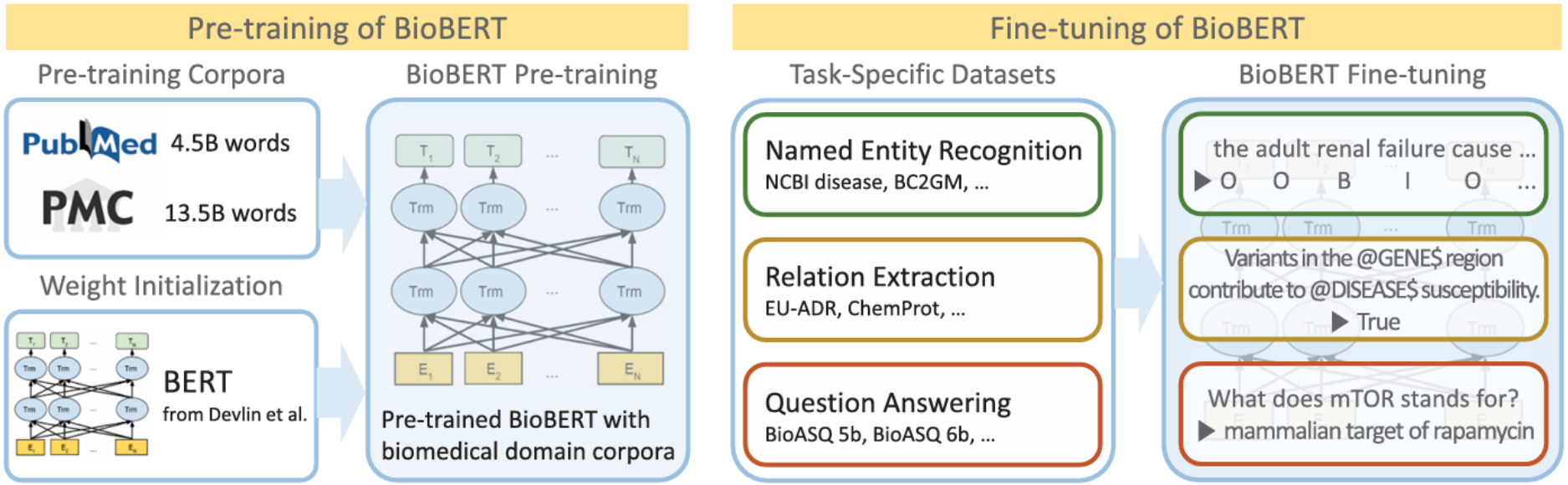


Table 6. Test results in biomedical named entity recognition

Type	Datasets	Metrics	SOTA	BERT (Wiki + Books)	BioBERT v1.0 (+ PubMed)	BioBERT v1.0 (+ PMC)	BioBERT v1.0 (+ PubMed + PMC)	BioBERT v1.1 (+ PubMed)
Disease	NCBI disease	P	88.30	84.12	86.76	86.16	**89.04**	88.22
		R	89.00	87.19	88.02	89.48	89.69	**91.25**
		F	88.60	85.63	87.38	87.79	89.36	**89.71**
	2010 i2b2/VA	P	87.44	84.04	85.37	85.55	**87.50**	86.93
		R	86.25	84.08	85.64	85.72	85.44	**86.53**
		F	**86.84**	84.06	85.51	85.64	86.46	86.73
	BC5CDR	P	**89.61**	81.97	85.80	84.67	85.86	86.47
		R	83.09	82.48	86.60	85.87	87.27	**87.84**
		F	86.23	82.41	86.20	85.27	86.56	**87.15**
Gene/protein	BC2GM	P	81.81	81.17	81.72	82.86	**85.16**	84.32
		R	81.57	82.42	83.38	84.21	83.65	**85.12**
		F	81.69	81.79	82.54	83.53	84.40	**84.72**
	JNLPBA	P	**74.43**	69.57	71.11	71.17	72.68	72.24
		R	83.22	81.20	83.11	82.76	83.21	**83.56**
		F	**78.58**	74.94	76.65	76.53	77.59	77.49

```python
model = PeftModel.from_pretrained(model, model_id=lora_dir)

input_text = "西安电子科技大学的陈志明爱上了隔壁西北工业大学苏春红，他们约定好毕业后去中国的苏州定居。"
test_texts = {
    "instruction": """你是一个文本实体识别领域的专家，你需要从给定的句子中提取 地点; 人名; 地理实体; 组织 实体. 以 json
    "input": f"文本:{input_text}"
}

instruction = test_texts['instruction']
input_value = test_texts['input']

messages = [
    {"role": "system", "content": f"{instruction}"},
    {"role": "user", "content": f"{input_value}"}
]

response = predict(messages, model, tokenizer)
print(response)
```

收起

输出结果为：

```json
{"entity_text": "西安电子科技大学", "entity_label": "组织"}
{"entity_text": "陈志明", "entity_label": "人名"}
{"entity_text": "西北工业大学", "entity_label": "组织"}
{"entity_text": "苏春红", "entity_label": "人名"}
{"entity_text": "中国", "entity_label": "地理实体"}
{"entity_text": "苏州", "entity_label": "地理实体"}
```

[GLM4大模型微调入门实战-命名实体识别（NER）任务_大模型ner微调-CSDN博客]

可以一问一答的问题对，通过微调，让大模型学会你想要的答案格式

（5）房室结消融和起搏器植入作为反复发作或难治性心房内折返性心动过速的替代疗法。	[{ "end_idx": 7, "entity": "房室结消融", "start_idx": 3, "type": "pro" }, { "end_idx": 13, "entity": "起搏器植入", "start_idx": 9, "type": "pro" }, { "end_idx": 33, "entity": "反复发作或难治性…
（6）发作一次伴血流动力学损害的室性心动过速（ventriculartachycardia），可接受导管消融者。	[{ "end_idx": 21, "entity": "血流动力学损害的室性心动过速", "start_idx": 8, "type": "dis" }, { "end_idx": 44, "entity": "ventriculartachycardia", "start_idx": 23, "type": "dis" }, { "end_idx":…
4.第三类（1）无症状性WPW综合征患者，年龄小于5岁。	[{ "end_idx": 17, "entity": "无症状性WPW综合征", "start_idx": 8, "type": "dis" }]
（2）室上性心动过速可用常规抗心律失常药物控制，年龄小于5岁。	[{ "end_idx": 9, "entity": "室上性心动过速", "start_idx": 3, "type": "dis" }, { "end_idx": 20, "entity": "抗心律失常药物", "start_idx": 14, "type": "dru" }]
（3）非持续性，不考虑为无休止性的阵发性室性心动过速（即一次监视数小时或任何一小时记录的心电图条带几乎均可出现），心室功能正常。	[{ "end_idx": 25, "entity": "非持续性，不考虑为无休止性的阵发性室性心动过速", "start_idx": 3, "type": "dis" }, { "end_idx": 46, "entity": "心电图", "start_idx": 44, "type": "pro" }, { "end_idx": 58,…
（4）非持续性室上性心动过速，不需其他治疗和（或）症状轻微。	[{ "end_idx": 13, "entity": "非持续性室上性心动过速", "start_idx": 3, "type": "dis" }]
第十章胸膜疾病小儿胸膜疾病以胸膜炎最为常见，多继发于肺部感染，原发性或其他原因所致者较少见。	[{ "end_idx": 6, "entity": "胸膜疾病", "start_idx": 3, "type": "dis" }, { "end_idx": 12, "entity": "小儿胸膜疾病", "start_idx": 7, "type": "dis" }, { "end_idx": 16, "entity": "胸膜炎", "start_idx": 14…
第一节胸膜炎胸膜炎（pleurisy）分为三种：干性胸膜炎、浆液性胸膜炎和化脓性胸膜炎。	[{ "end_idx": 5, "entity": "胸膜炎", "start_idx": 3, "type": "dis" }, { "end_idx": 8, "entity": "胸膜炎", "start_idx": 6, "type": "dis" }, { "end_idx": 17, "entity": "pleurisy", "start_idx": 10,…
一、干性胸膜炎干性胸膜炎（dryorplasticpleurisy）又称纤维素性胸膜炎，常与肺部细菌感染有关，亦可发生于急性上呼吸道疾病过程中。	[{ "end_idx": 6, "entity": "干性胸膜炎", "start_idx": 2, "type": "dis" }, { "end_idx": 11, "entity": "干性胸膜炎", "start_idx": 7, "type": "dis" }, { "end_idx": 32, "entity":…
结缔组织疾病如风湿热患儿亦可发生。	[{ "end_idx": 5, "entity": "结缔组织疾病", "start_idx": 0, "type": "dis" }, { "end_idx": 9, "entity": "风湿热", "start_idx": 7, "type": "dis" }]
深呼吸及咳嗽时疼痛加剧。	[{ "end_idx": 10, "entity": "深呼吸及咳嗽时疼痛加剧", "start_idx": 0, "type": "sym" }]
病程早期可闻胸膜摩擦音在全部呼吸期间均可听到。	[{ "end_idx": 21, "entity": "闻胸膜摩擦音在全部呼吸期间均可听到", "start_idx": 5, "type": "sym" }, { "end_idx": 7, "entity": "胸膜", "start_idx": 6, "type": "bod" }]
胸部X线透视和胸片可见患侧膈呼吸运动减弱肋膈角变钝流行性胸痛和带状疱疹前驱期的胸痛及肋骨骨折相鉴别。	[{ "end_idx": 5, "entity": "胸部X线透视", "start_idx": 0, "type": "pro" }, { "end_idx": 19, "entity": "胸部X线透视和胸片可见患侧膈呼吸运动减弱", "start_idx": 0, "type": "sym" }, { "end_idx": 8,…

为什么模型可以学会语言
（如何预处理数据和构建模型）

一个完整模型的组件：分词器和完整的模型参数

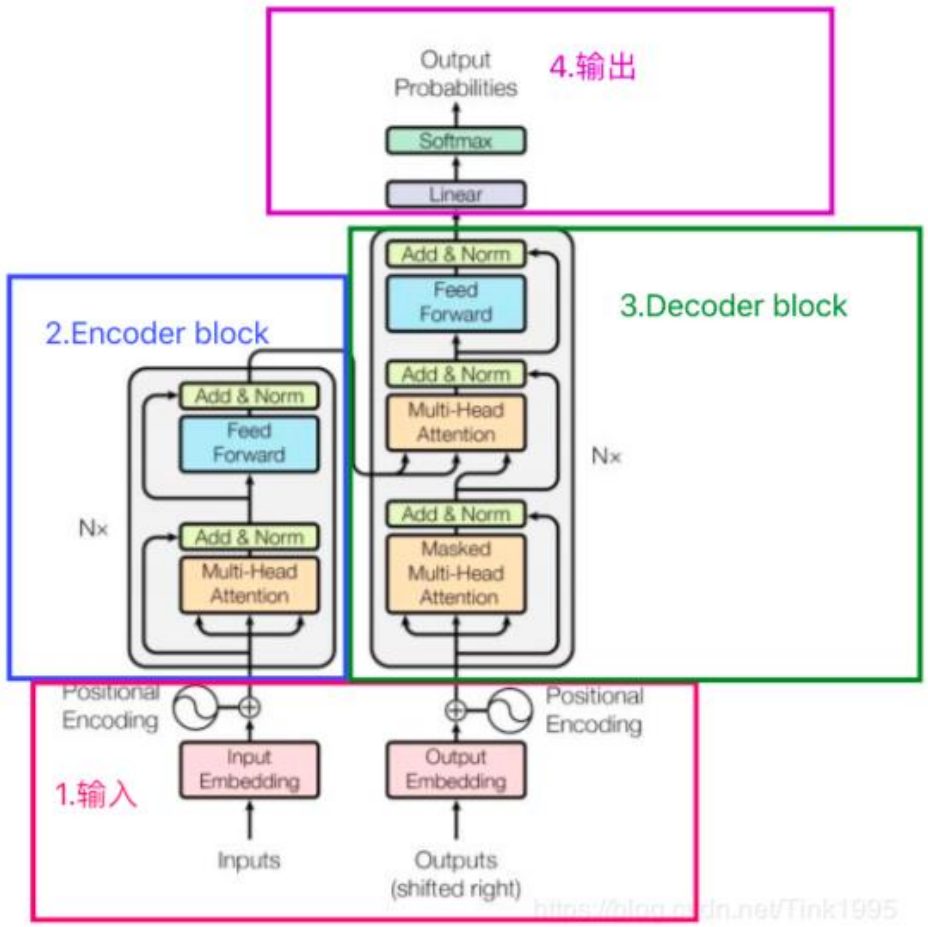
向量嵌入：将一个个token转化为向量，让计算机可以进行计算

早期向量嵌入：word2vec，如今向量嵌入：nn.Embedding

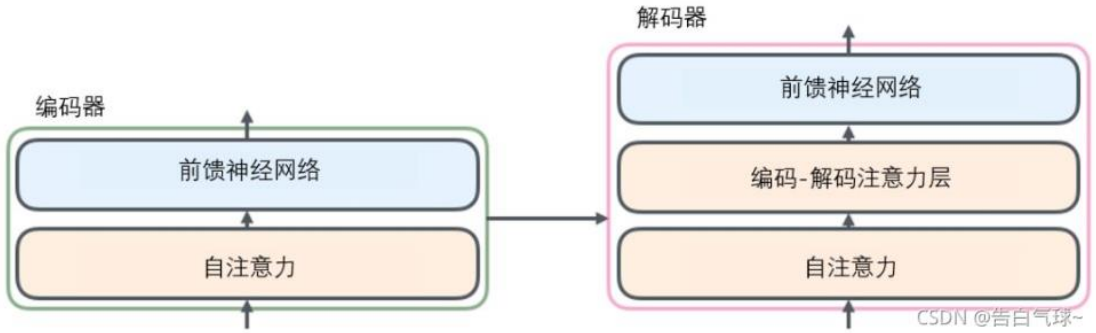
[大模型核心概念科普：Token、上下文长度、最大输出，一次讲透 - 知乎]

Encoder：将word embedding进行进一步计算压缩

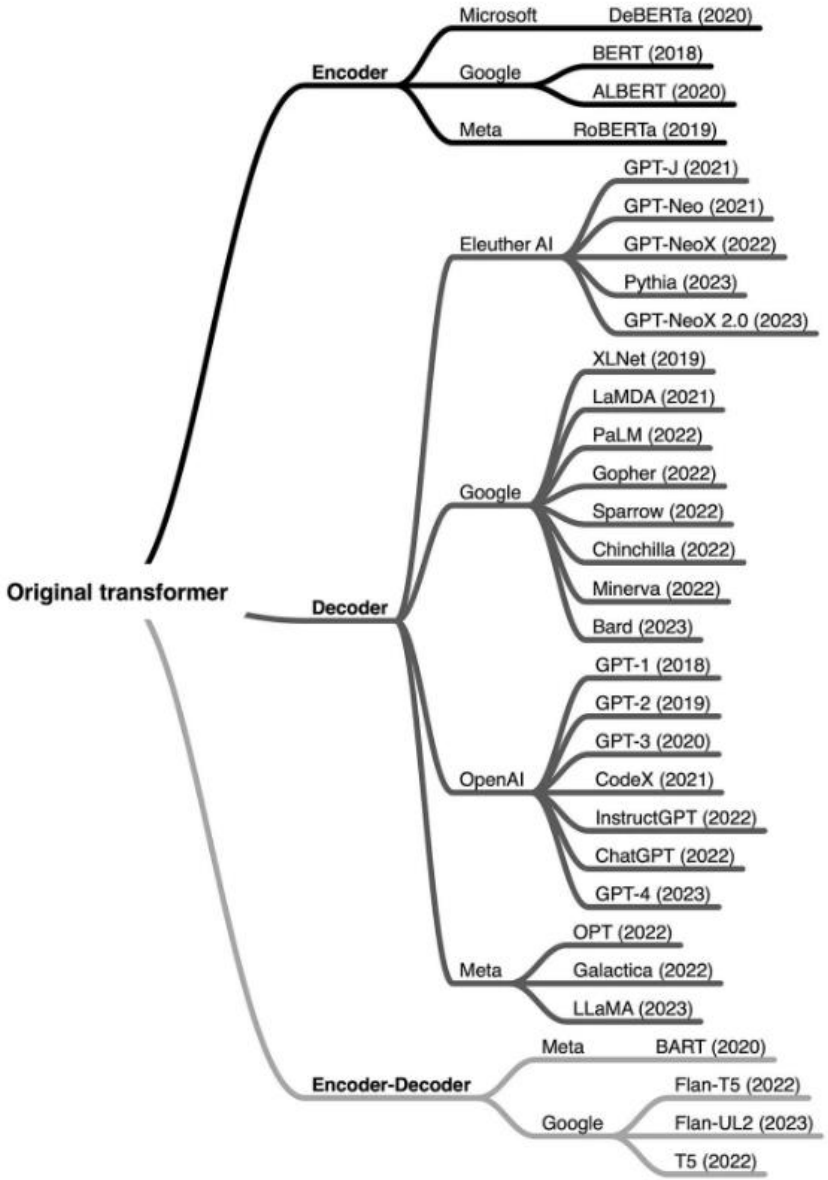
Decoder：将计算压缩的word embedding进行还原

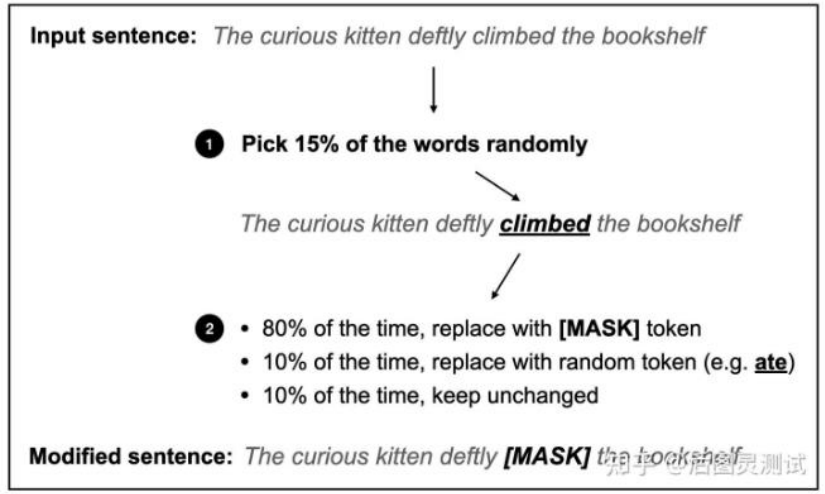


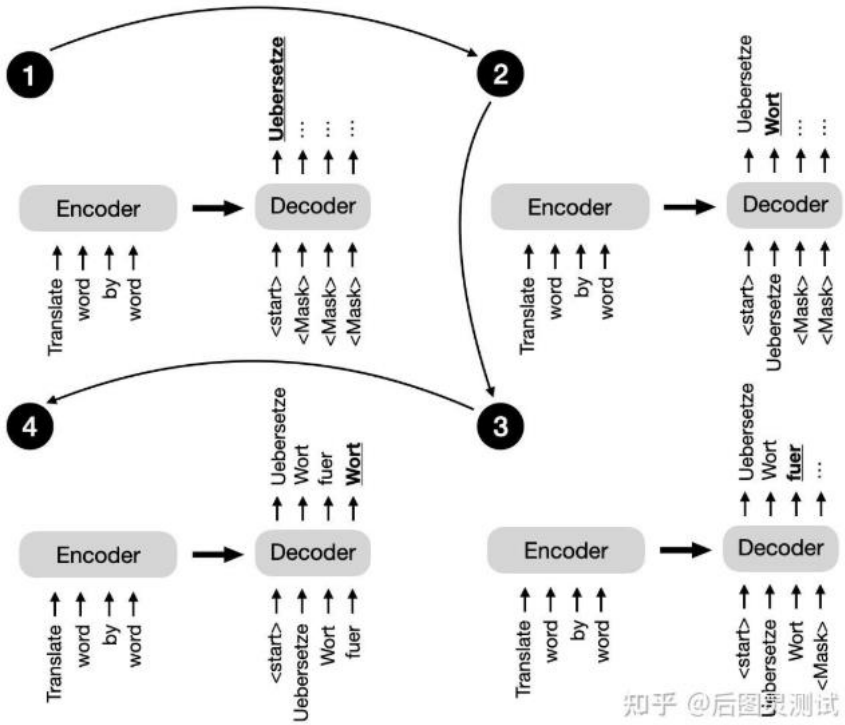
Transformer架构



特性	Encoder-only (BERT)	Decoder-only (GPT)	Encoder-Decoder (T5/BA
主要架构	仅 Transformer Encoder	仅 Transformer Decoder	Encoder + Decoder
注意力机制	双向（Full Attention）	单向（Causal Masking）	Encoder 双向 + Decoder 单
输入的作用	用于编码整个文本（理解任务）	提供初始上下文（生成任务）	Encoder 处理输入，Decod
典型任务	文本分类、NER、问答	文本生成、对话、代码补全	翻译、摘要、文本转换
是否可生成文本	❌（仅编码）	✅（自回归生成）	✅（条件生成）
训练目标	掩码语言建模（MLM）	自回归语言建模（LM）	序列到序列（Seq2Seq）
代表模型	BERT、RoBERTa、DeBERTa	GPT-3、LLaMA、PaLM	T5、BART、MarianMT
输入输出示例	输入：`"今天天气[CLS]"` 输出：`[CLS]` 向量（分类）	输入：`"今天天气"` 输出：`"真好"`	输入：`"translate to Eng` 输出：`"The weather toda`



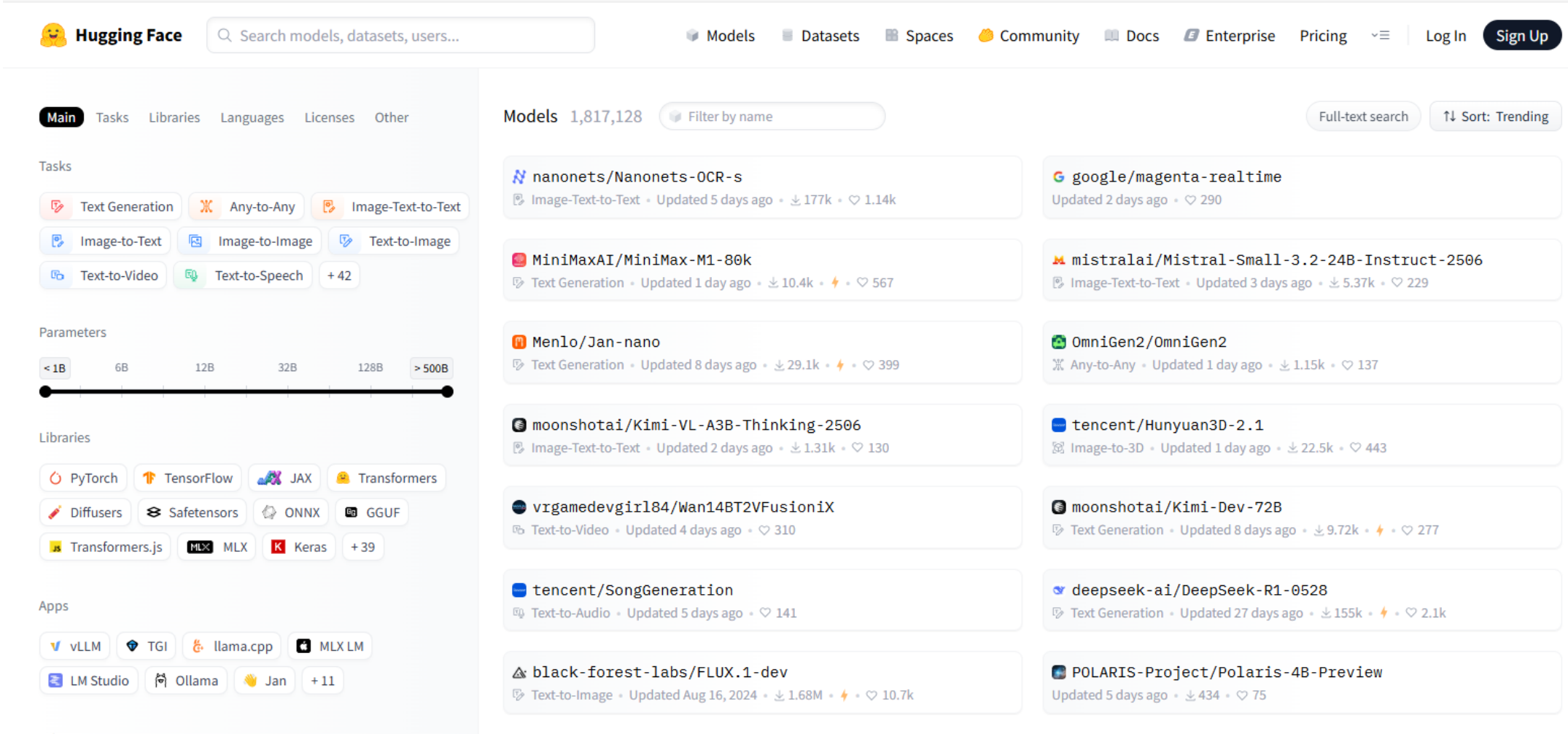




[LLM的3种架构：Encoder-only、Decoder-only、encode-decode - 知乎]

用大语言模型如何完成任务
（如何利用和微调）

模型下载 Hugging Face



确定模型→给定提示词→推理→得到结果

```
model_name = "methylation_marker/finetuned_deepseek_r1_8b/"
max_seq_length = 2048
model, tokenizer = FastLanguageModel.from_pretrained(model_name=model_name, max_seq_length=max_seq_length, load_in_4bit =False, load_in_8bit = False, f
ull_finetuning = False)

train_prompt_style = """Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response.

### Instruction:
You are an expert in the field of DNA methylation biomarkers with extensive experience in literature review. Based on the given sentence, you will dete
rmine whether it describes a gene as a DNA methylation biomarker for a specific type of cancer.

### Sentence:
{}

### Answer:
{}"""


FastLanguageModel.for_inference(model)

test_data = load_dataset("json", data_files = index_file, split = "train")
results = []
for sentence in test_data['Sentence']:
    question = sentence
    inputs = tokenizer([train_prompt_style.format(question, "")], return_tensors="pt").to("cuda")
    outputs = model.generate(
        input_ids=inputs.input_ids,
        attention_mask=inputs.attention_mask,
        max_new_tokens=512,
        use_cache=True,
    )
    response = tokenizer.batch_decode(outputs)
    answer = response[0].split("### Answer:")[1]
    results.append({"Sentence": sentence, "Answer": answer})
output_df = pd.DataFrame(results)
output_df.to_csv(output_file, index=False)
#print(output_df.head())
(END)
```

确定模型→给定提示词→给定数据集→重新微调训练

```
train_prompt_style = """Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response.

### Instruction:
You are an expert in the field of DNA methylation biomarkers with extensive experience in literature review. Based on the given sentence, you will dete
rmine whether it describes a gene as a DNA methylation biomarker for a specific type of cancer.

### Sentence:
{}

### Answer:
{}"""

EOS_TOKEN = tokenizer.eos_token
def formatting_prompts_func(examples):
    inputs = examples["Sentence"]
    outputs = examples["Answer"]
    texts = []
    for input, output in zip(inputs, outputs):
        text = train_prompt_style.format(input, output) + EOS_TOKEN
        texts.append(text)
    return {
        "text": texts,
    }
```

```
model = FastLanguageModel.get_peft_model(model, r = 16, target_modules = ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj",]
, lora_alpha = 16, lora_dropout = 0, bias = "none", use_gradient_checkpointing = "unsloth", random_state = 42, max_seq_length = max_seq_length, use_rsl
ora = False, loftq_config = None)
```

```
trainer = SFTTrainer(model = model, train_dataset = train_data, eval_dataset=test_data, tokenizer = tokenizer, args = SFTConfig(dataset_text_field = "t
ext", max_seq_length = max_seq_length, per_device_train_batch_size = 50, gradient_accumulation_steps = 4, warmup_steps = 10, max_steps = 500, logging_s
teps = 5, evaluation_strategy="steps", output_dir = "finetuned_deepseek_r1_8b", optim = "adamw_8bit", seed = 42))
```

```
trainer.train()
eval_results = trainer.evaluate()
```

DeepSeek-R1-Distill-Llama-8B

A Before fine-tuning:

['< | begin_of_sentence | >
Below is an instruction that describes a task, paired with an input that provides further context.\nWrite a response that appropriately completes the request.\nBefore answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response.\n\n### Instruction:\nYou are an expert in the field of DNA methylation biomarkers with extensive experience in literature review. Based on the given sentence, you will determine whether it describes a gene as a DNA methylation biomarker for a specific type of cancer.\n\n### Sentence:\nThus MGMT hypermethylation can be used as a biomarker for esophageal cancer in high incidence region of North East India.\n\n### Answer:

\nMGMT hypermethylation has been identified as a promising biomarker for esophageal cancer, particularly in high incidence regions like the Northeast of India. This biomarker is associated with tumor progression and poor prognosis, making it a valuable tool for early detection and therapeutic interventions in this context.\n</think>\n

\nMGMT hypermethylation has been identified as a promising biomarker for esophageal cancer, particularly in high incidence regions like the Northeast of India. This biomarker is associated with tumor progression and poor prognosis, making it a valuable tool for early detection and therapeutic interventions in this context.
< | end_of_sentence | >']

B After fine-tuning:

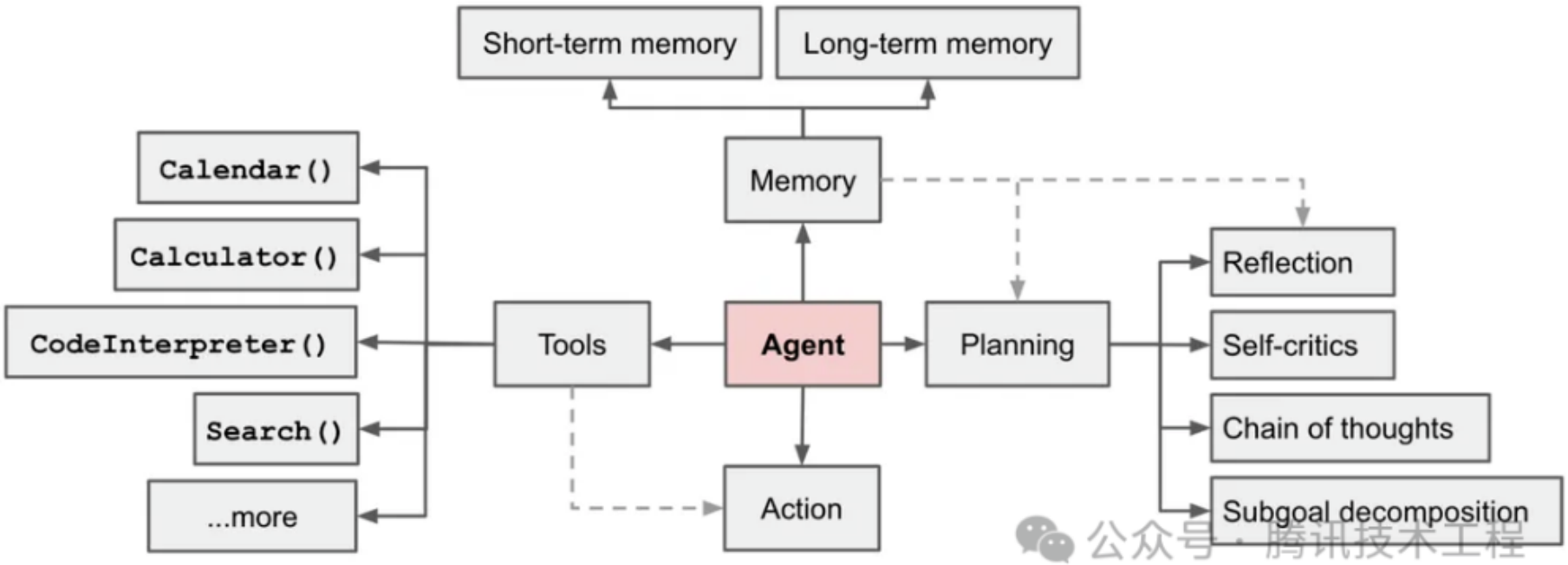
['< | begin_of_sentence | >
Below is an instruction that describes a task, paired with an input that provides further context.\nWrite a response that appropriately completes the request.\nBefore answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response.\n\n### Instruction:\nYou are an expert in the field of DNA methylation biomarkers with extensive experience in literature review. Based on the given sentence, you will determine whether it describes a gene as a DNA methylation biomarker for a specific type of cancer.\n\n### Sentence:\nThus MGMT hypermethylation can be used as a biomarker for esophageal cancer in high incidence region of North East India.\n\n### Answer:

\nThis is an acceptable sentence. This sentence includes information about the gene MGMT, and its methylation status as a biomarker in Esophageal Cancer.

< | end_of_sentence | >']

大语言模型的智能问答
（进一步的应用，如Agent和RAG）

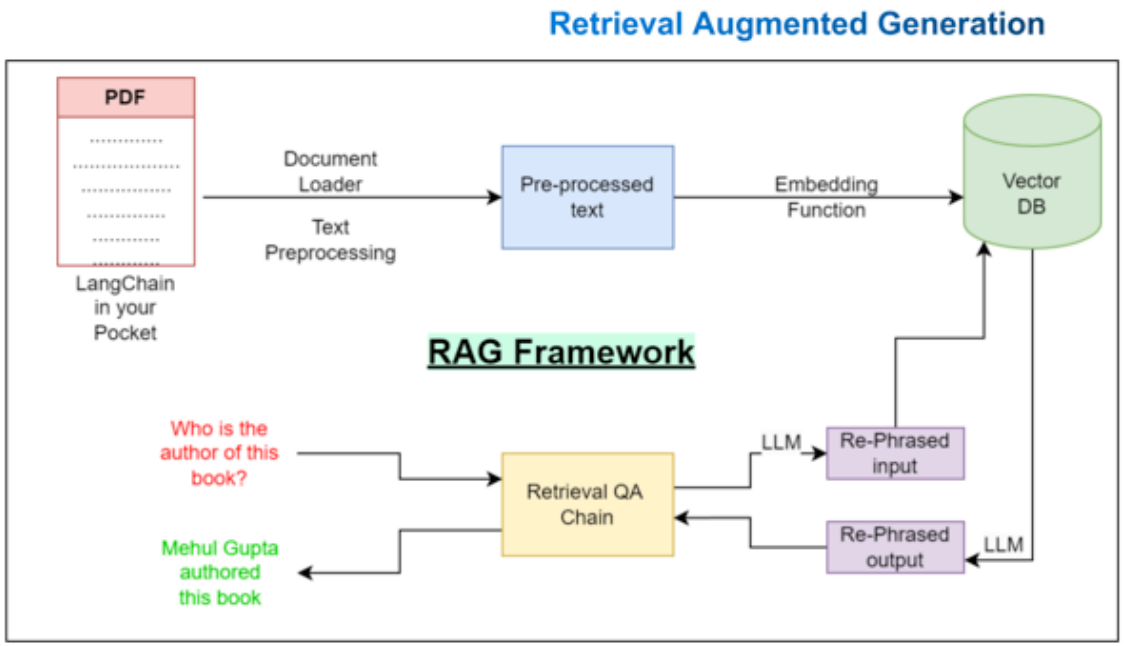
Agent是一种基于大语言模型的，具备规划思考能力、记忆能力、使用工具函数的能力，能自主完成给定任务的计算机程序



[一文带你了解大模型——智能体（Agent）-腾讯云开发者社区-腾讯云]

RAG的目的是通过从外部知识库检索相关信息来辅助大语言模型生成更准确、更丰富的文本内容。

检索：检索是RAG流程的第一步，从预先建立的知识库中检索与问题相关的信息。这一步的目的是为后续的生成过程提供有用的上下文信息和知识支撑。
增强：RAG中增强是将检索到的信息用作生成模型（即大语言模型）的上下文输入，以增强模型对特定问题的理解和回答能力。这一步的目的是将外部知识融入生成过程中，使生成的文本内容更加丰富、准确和符合用户需求。通过增强步骤，LLM模型能够充分利用外部知识库中的信息。
生成：生成是RAG流程的最后一步。这一步的目的是结合LLM生成符合用户需求的回答。生成器会利用检索到的信息作为上下文输入，并结合大语言模型来生成文本内容。



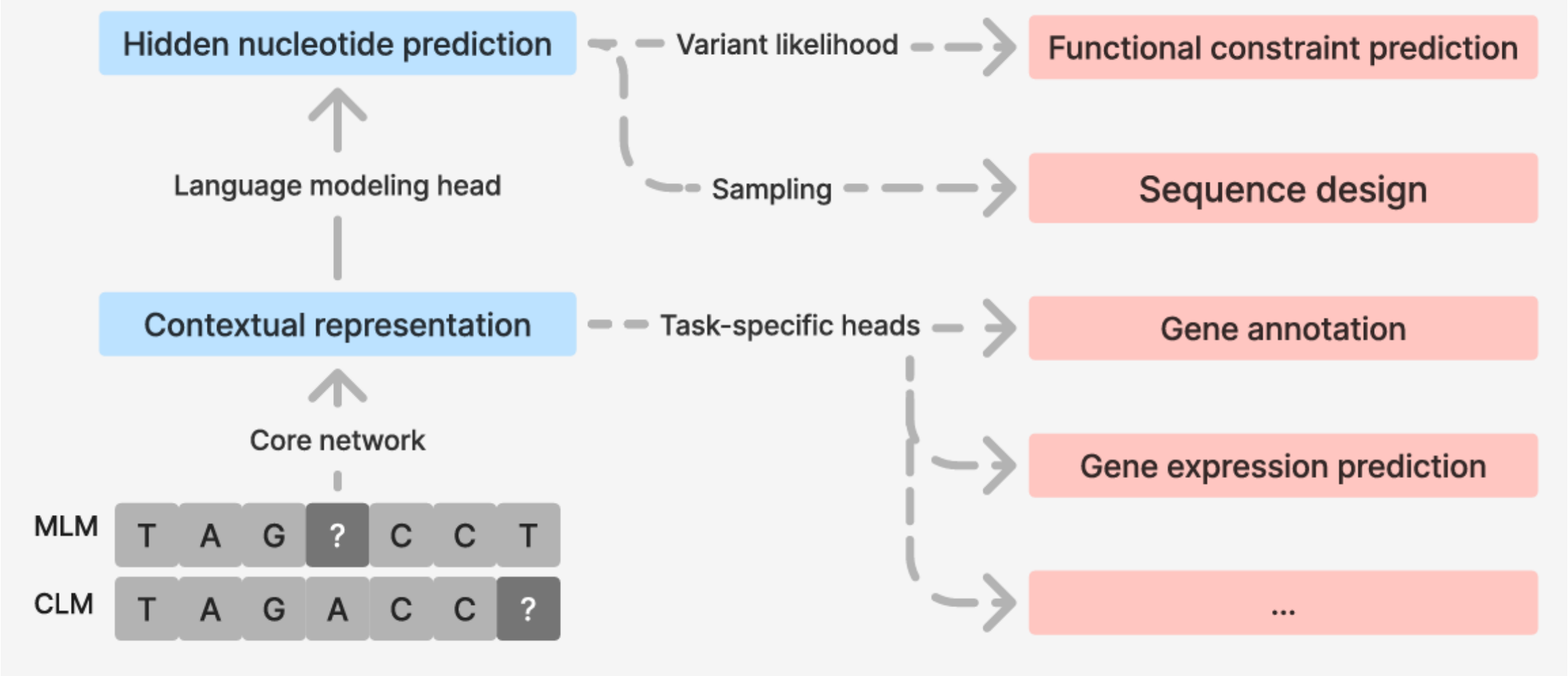
RAG（检索增强生成）是一种 AI 框架，它将传统信息检索系统（例如数据库）的优势与生成式大语言模型 (LLM) 的功能结合在一起。

外部知识库检索信息，作为LLM的额外输入

任务2：DNA大模型
（设计大模型）

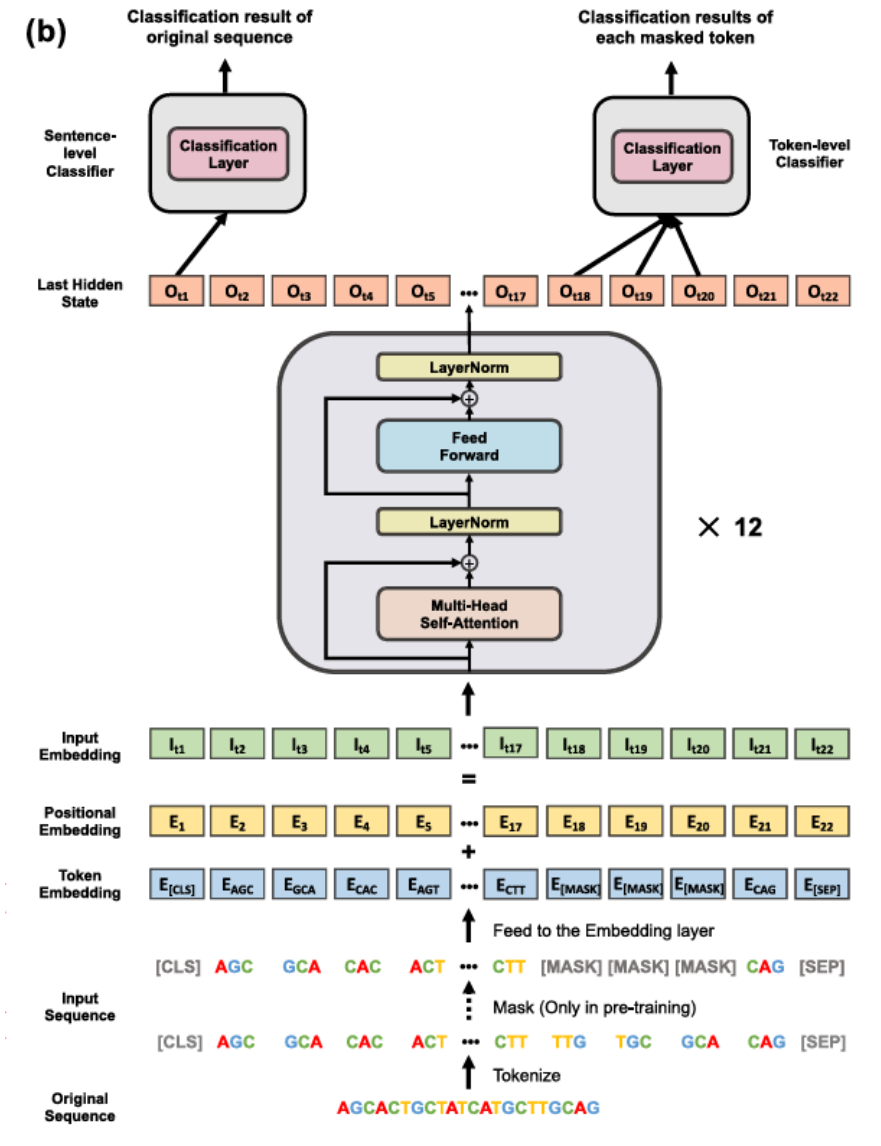
生物大模型是什么

把语言文本转化为DNA序列，DNA序列作为输入的语言模型

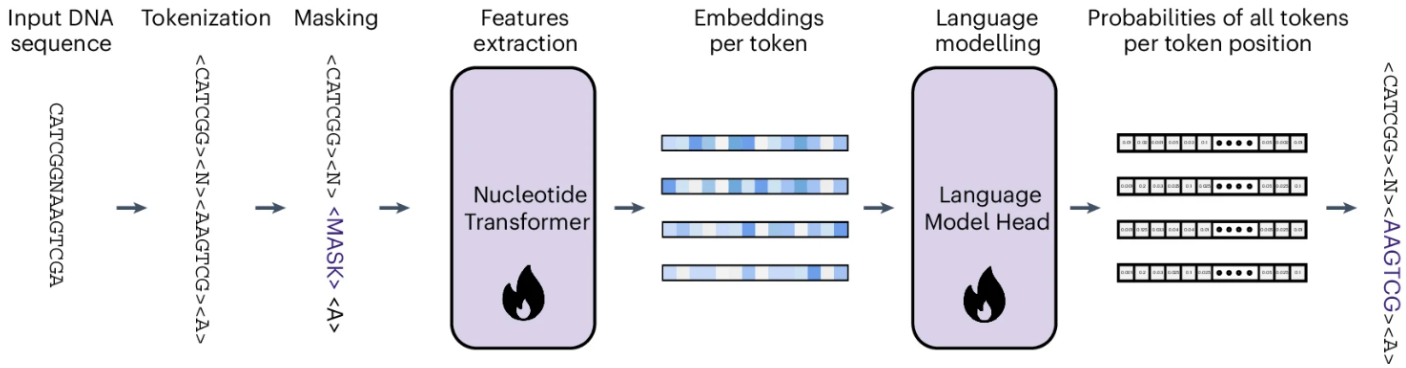


[2407.11435] Genomic Language Models: Opportunities and Challenges

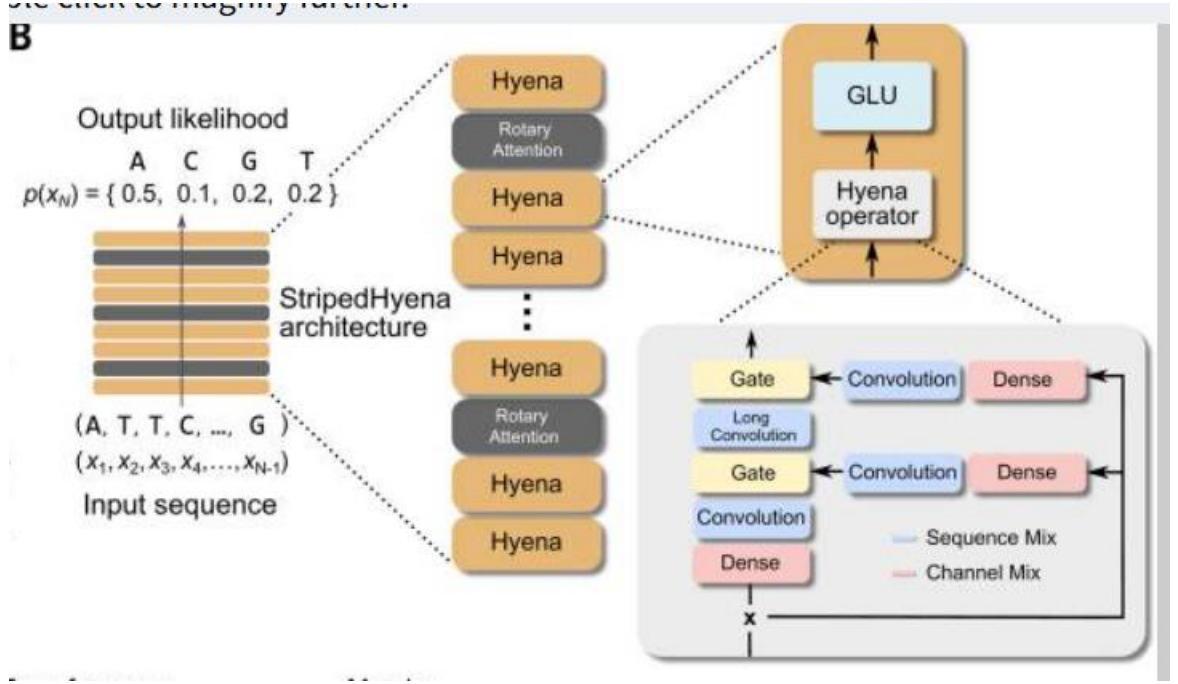
DNABERT



Nucleotide Transformer



Hyena-DNA



JOURNAL ARTICLE

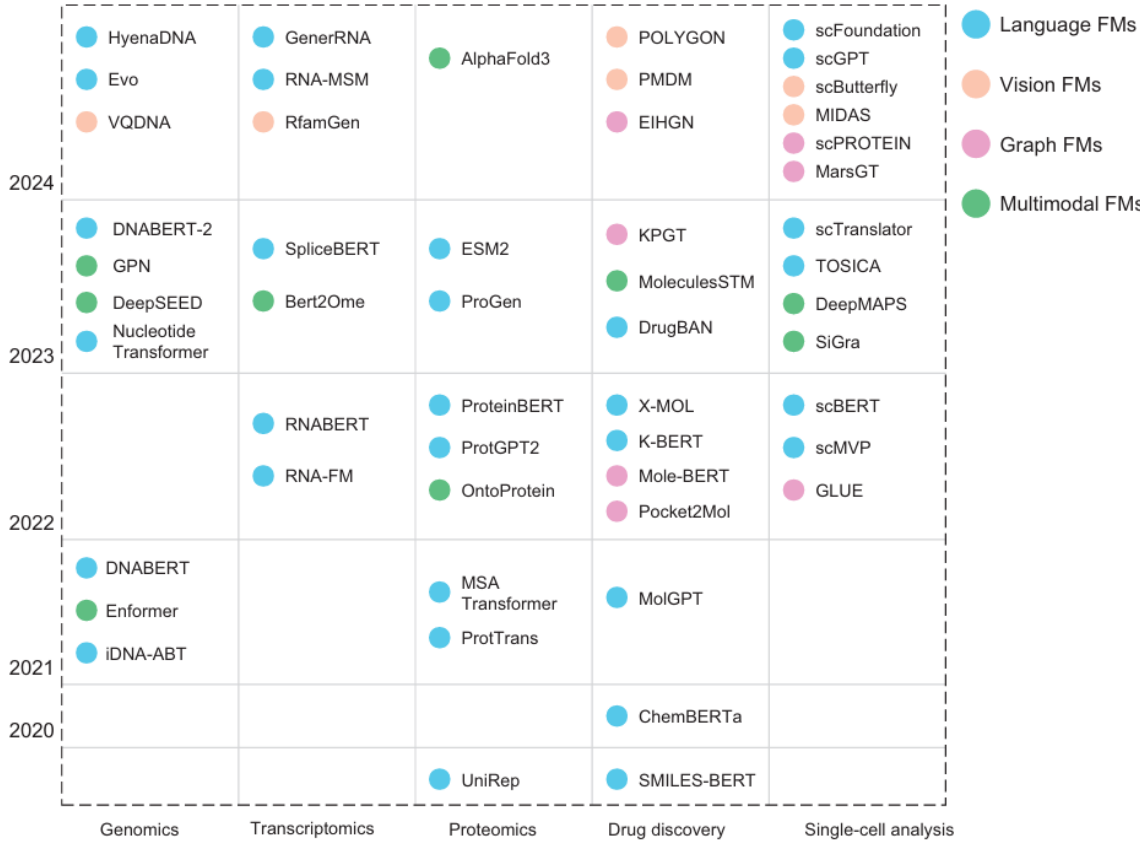
Foundation models in bioinformatics

Fei Guo , Renchu Guan , Yaohang Li , Qi Liu , Xiaowo Wang , Can Yang , Jianxin Wang

National Science Review, Volume 12, Issue 4, April 2025, nwaf028,
https://doi.org/10.1093/nsr/nwaf028 IF: 17.1 Q1
Published: 25 January 2025 **Article history**

PDF Split View Annotate Cite Permissions Share

[Foundation models in bioinformatics - PubMed]



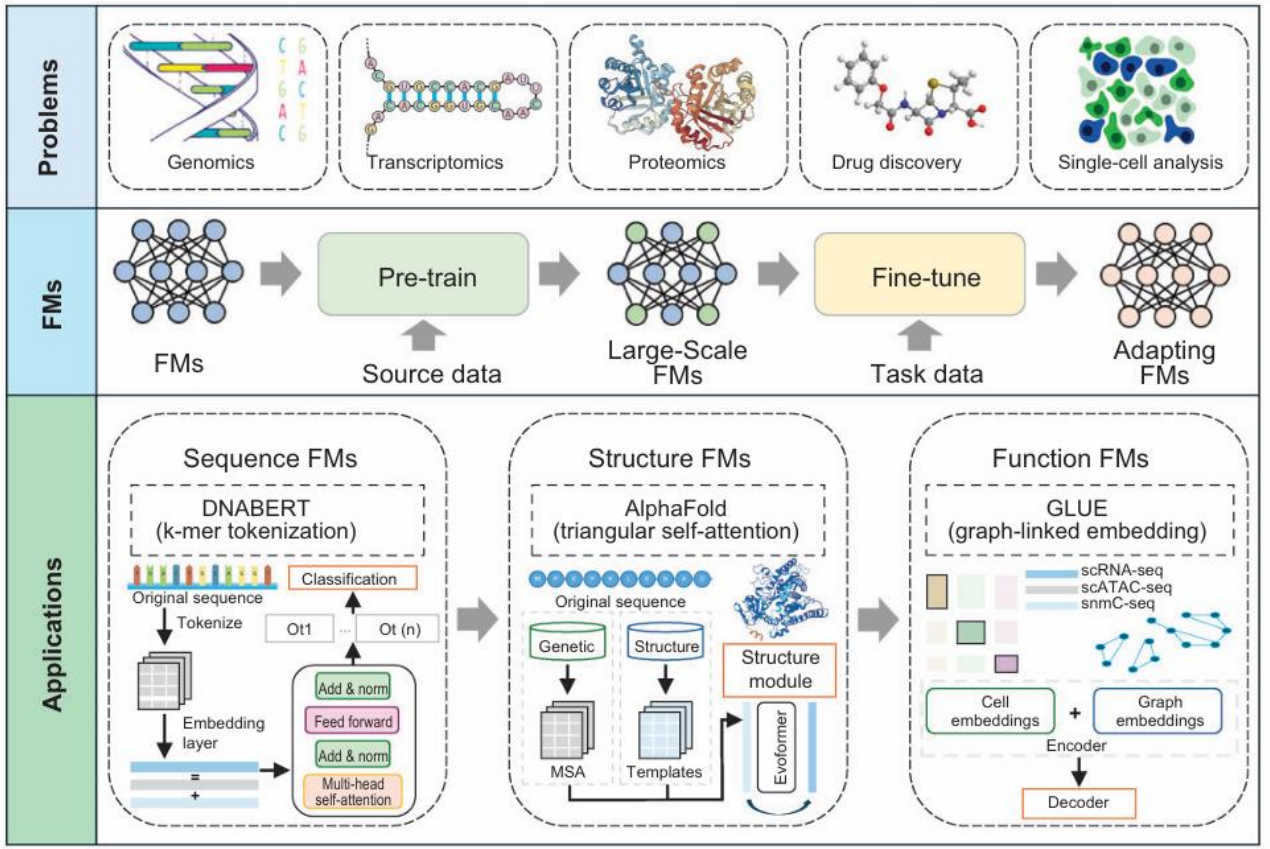


Figure 2. Evolution of FMs in bioinformatics.

下游任务设计

New Results

Follow this preprint Previous

Benchmarking DNA Foundation Models for Genomic Sequence Classification

Posted August 18, 2024.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, Chong Wu

doi: https://doi.org/10.1101/2024.08.16.608288

This article is a preprint and has not been certified by peer review [what does this mean?].

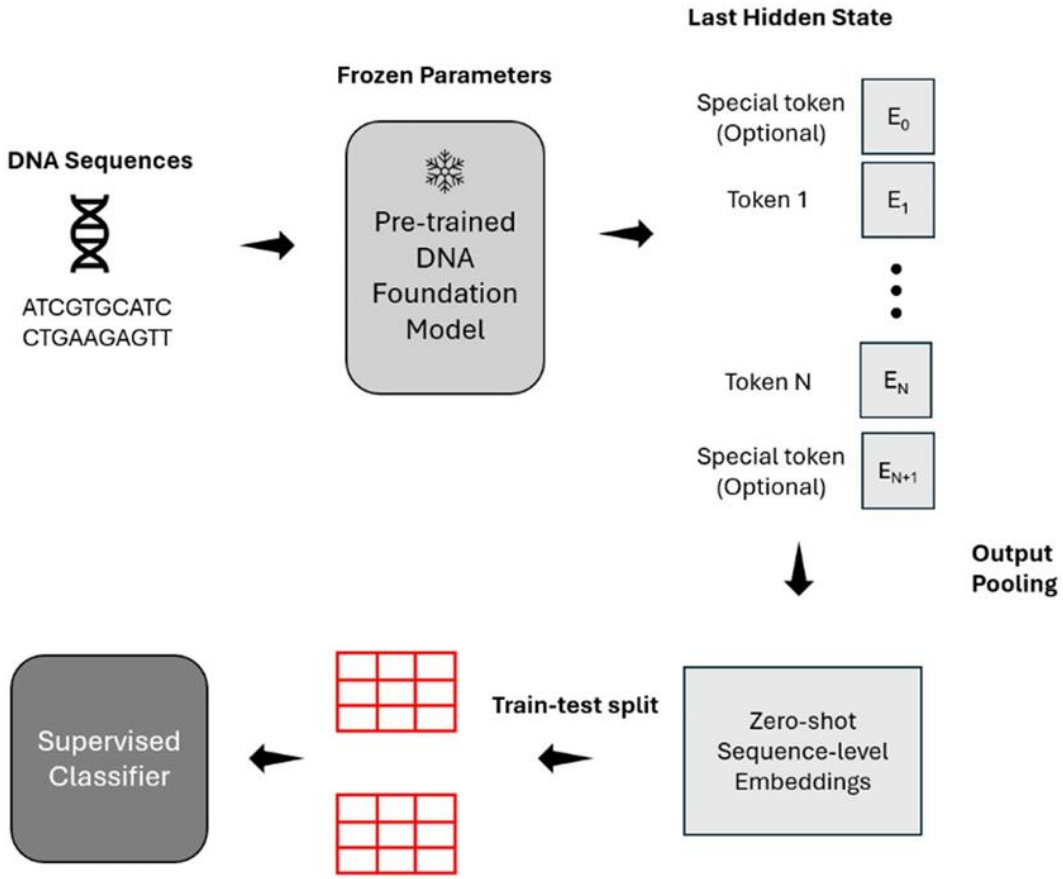
Download PDF Email

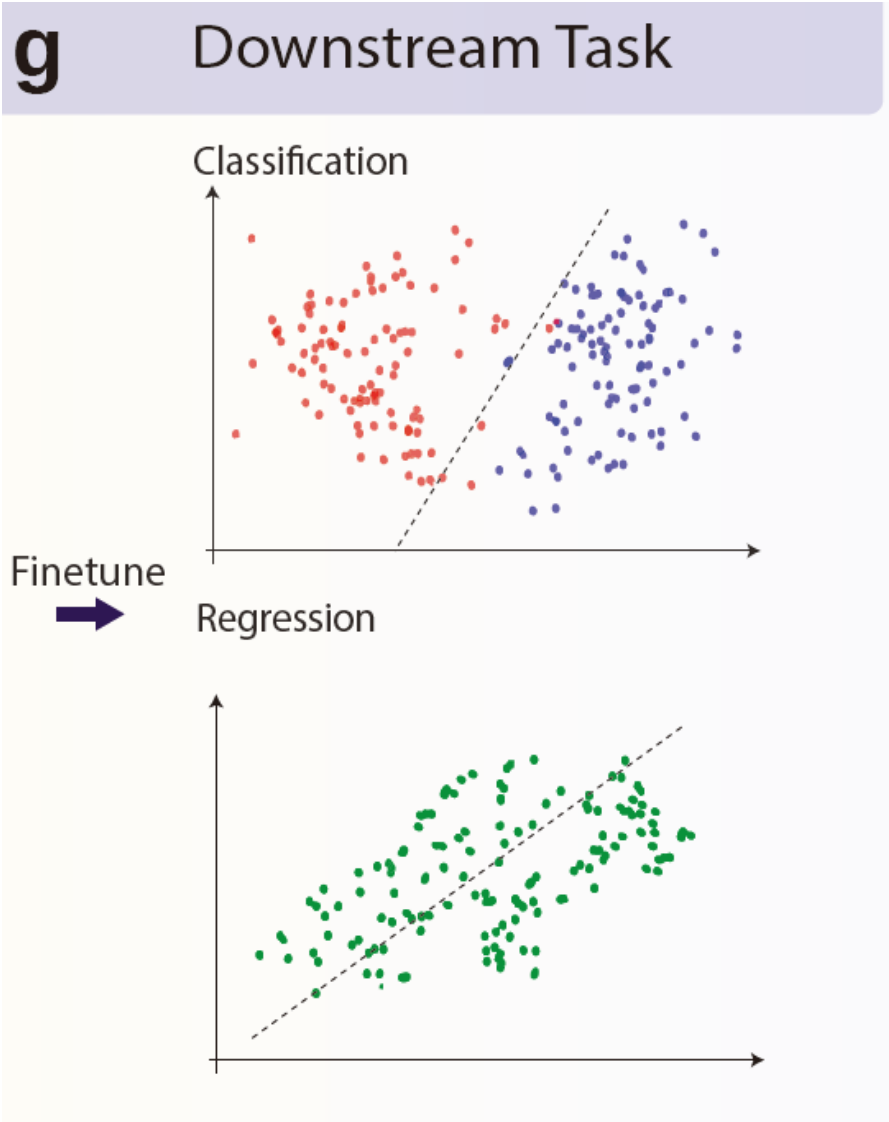
Print/Save Options Share

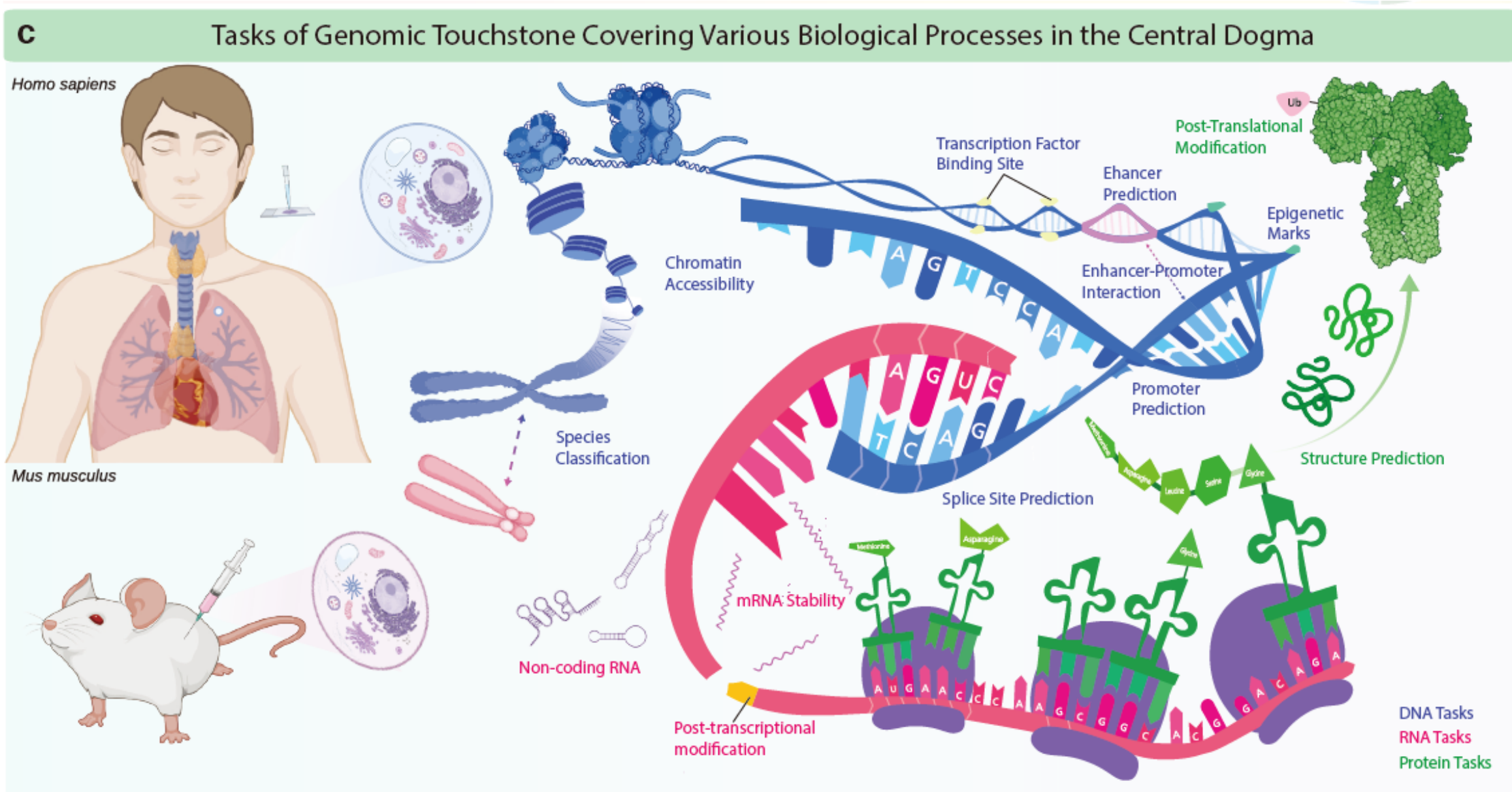
Supplementary Material Citation Tools

Get QR code

Abstract | Full Text | Info/History | Metrics | Preview PDF







https://www.biorxiv.org/content/10.1101/2025.06.25.661622v1

DNA层面任务

Task	Dataset	Type
		DNA
Promoter Annotation	Core_Promoter_Annotation Promoter_Annotation	Sequence Classification (4 classes) Sequence Multilabel Classification (6 classes)
Enhancers Types Classification	Enhancers_Types	Sequence Classification (3 classes)
Splice Sites Prediction	Splice_Sites_All Splice_Sites_Donor Splice_Sites_Acceptor	Sequence Multilabel Classification (3 classes) Sequence Classification (2 classes) Sequence Classification (2 classes)
Epigenetic Marks Prediction	Histone_annotation	Sequence Multilabel Classification (10 classes)
CpG Methylation Prediction	CpG_methylation	Sequence Multilabel Classification (7 classes)
Regulation Element Prediction	Regulation Regulation_multilabel	Sequence Classification (4 classes) Sequence Multilabel Classification (4 classes)
Species Classification	Species_Classification	Sequence Classification (6 classes)
Regulatory Activity Prediction	GM12878 H1ESC HEPG2 IMR90 K562	Sequence Regression Sequence Regression Sequence Regression Sequence Regression Sequence Regression
Enhancer Activity Prediction	DeepSTARR_HCT116 HCT116_Brd2 HCT116_Brd4 HCT116_CDK7 HCT116_CDK9 HCT116_Med14 HCT116_Parental HCT116_p300CBP	Sequence Regression Sequence Regression Sequence Regression Sequence Regression Sequence Regression Sequence Regression Sequence Regression Sequence Regression
Enhancer Promoter Interaction Prediction	GM12878 H1ESC HEPG2 IMR90 K562 NHEK	Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes)
Bulk RNA Expression Prediction	Bulk_RNA_Expression_Prediction	Sequence Regression
Genetic Disease Classification	Breast_Disease_Type_Classification Cardiovascular_Disorders_Classification Kidney_Disease_Type_Classification	Sequence Classification (4 classes) Sequence Classification (8 classes) Sequence Classification (7 classes)
Pathogenicity Classification	Pathogenicity_Classification	Sequence Classification (4 classes)
Splice Variant Effect Prediction	Exonic Intronic	Sequence Classification (2 classes) Sequence Classification (2 classes)
BRCA1/2 SNV Functional Impact Classification	BRCA1_coding BRCA1_noncoding BRCA2_coding BRCA2_noncoding	Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes) Sequence Classification (2 classes)

https://www.biorxiv.org/content/10.1101/2025.06.25.661622v1

RNA层面任务

		RNA
Mean Ribosome Loading Prediction	Mean_Ribosome_Loading	Sequence Regression
	Mean_Ribosome_Loading_Human	Sequence Regression
mRNA Stability Prediction	mRNA_Stability	Sequence Regression
Abundance Prediction	PA_athaliana	Sequence Regression
	PA_dmelanogaster	Sequence Regression
	PA_ecoli	Sequence Regression
	PA_hsapiens	Sequence Regression
	PA_scerevisiae	Sequence Regression
	TA_athaliana	Sequence Regression
	TA_dmelanogaster	Sequence Regression
	TA_ecoli	Sequence Regression
	TA_hsapiens	Sequence Regression
	TA_hvolcanii	Sequence Regression
	TA_ppastoris	Sequence Regression
	TA_scerevisiae	Sequence Regression
Non-coding RNA Classification	NoncodingRNAFamily	Sequence Classification (13 classes)
Expression Level and Translation Efficiency Regulation Prediction	EL_HEK	Sequence Regression
	EL_Muscle	Sequence Regression
	EL_pc3	Sequence Regression
	TE_HEK	Sequence Regression
	TE_Muscle	Sequence Regression
	TE_pc3	Sequence Regression
Modification Prediction	Modification	Sequence Multilabel Classification (12 classes)
APA Isoform Prediction	Isoform	Sequence Regression
SARS-CoV-2 Vaccine Degradation Prediction	CoV_Vaccine_Degradation	Sequence Regression
Programmable RNA Switches Prediction	ProgrammableRNASwitches	Sequence Regression
Tc-Riboswitches Prediction	Tc_Riboswitches	Sequence Regression
mRFP Expression Prediction	mRFP_Expression	Sequence Regression

https://www.biorxiv.org/content/10.1101/2025.06.25.661622v1

蛋白质层面任务

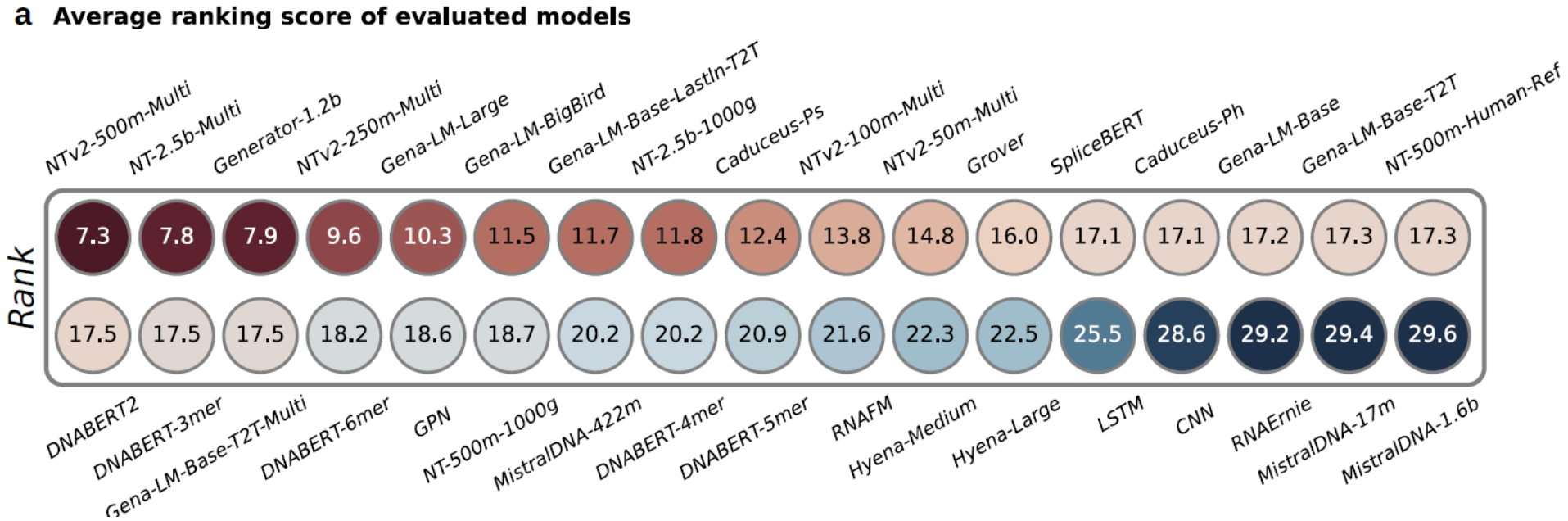
		Protein
Secondary Structure Analysis	Structure Prediction	Sequence Classification (3 classes)
Transmembrane Region Prediction	Transmembrane	Sequence Classification (6 classes)
Protein Variant Classification	Variant	Sequence Classification (2 classes)
Post-translational Modifications Prediction	PTM Prediction	Sequence Classification (3 classes)
Protein Domain Classification	Domain	Sequence Classification (4 classes)
Enzyme Classification	Enzyme Classification	Sequence Classification (4 classes)
Beta-Lactamase Activity Prediction	beta_lactamase_complete beta_lactamase_unique	Sequence Regression Sequence Regression
Fluorescence Prediction	Fluorescence	Sequence Regression
Protein Stability Prediction	Stability	Sequence Regression
Melting Point Prediction	Melting Point	Sequence Regression

https://www.biorxiv.org/content/10.1101/2025.06.25.661622v1

Genomic Touchstone: Benchmarking Genomic Language Models in the Context of the Central Dogma

不同模型适应不同的任务

Yihui Wang[1,†], Zhiyuan Cai[1,†], Qian Zeng[1], Yihang Gao[1], Jiarui Ouyang[1], Yingxue Xu[1], Shu Yang[1], Sunan He[1], Yuxiang Nie[1], Yu Cai[2], Fengtao Zhou[1], Cheng Jin[1], Xi Wang[1], Zhi Xie[3], Danqing Zhu[4,5], Ting Xie[6], Kwang-Ting Cheng[1,2], Can Yang[7], Xi Fu[8], Jiguang Wang[4,5,10,11], Kang Zhang[9], Jianhua Yao[10], Raul Rabadan[8], and Hao Chen[1,4,5,6,11,*]



https://www.biorxiv.org/content/10.1101/2025.06.25.661622v1

需要什么样的数据集

预训练是什么？ 微调是什么？

你需要搭建一个网络模型来完成一个特定的图像分类的任务。首先，你需要随机初始化参数，然后开始训练网络，不断调整参数，直到网络的损失越来越小。在训练的过程中，一开始初始化的参数会不断变化。当你觉得结果很满意的时候，你就可以将训练模型的参数保存下来，以便训练好的模型可以在下次执行类似任务时获得较好的结果。这个过程就是 `pre-training`。

之后，你又接收到一个类似的图像分类的任务。这时候，你可以直接使用之前保存下来的模型的参数来作为这一任务的初始化参数，然后在训练的过程中，依据结果不断进行一些修改。这时候，你使用的就是一个 `pre-trained` 模型，而过程就是 `fine-tuning`。

对于生物大语言模型：预训练是基因组序列，微调的是序列的分类标签等

DNABERT2

4.1 Pre-Train: Human and Multi-Species Genome

To investigate the impact of species diversity on genome foundation models, we compile and made publicly available two datasets for foundation model pre-training: the human genome and the multi-species genome. The human genome dataset is the one used in DNABERT (Ji et al., 2021), which comprises 2.75B nucleotide bases. The multi-species genome dataset encompasses genomes from 135 species, spread across 6 categories. In total, this dataset includes 32.49B nucleotide bases, nearly 12 times the volume of the human genome dataset. We exclude all sequences with N and retain only sequences that consist of A, T, C, and G. Detailed statistics are presented in Table 11.

```
>SLC4A1_1(-)|no_TATA|0
ATAAAGCAACAGGAGGACACCGGCTCTCGCGTCATTAGGTCACACAATGTC
>SV2C_1(+)|no_TATA|1
CCCAGTCCCACACCGCAGCAGCGCCTCAGCACCGCGACTTGCCGGAGCACC
>CNGA1_1(-)|no_TATA|1
CAGACACAAACCTTCTCGGCAGGGCTCTCCGGGCGGCAGCCTCACCTGTGC
>CREB5_3(+)|no_TATA|0
GCATCACAAGGCTAGTCTTTATCTTTCCGGCATCCGTACGAGTTGAAACTA
>ZNF577_3(-)|no_TATA|1
CATAAGCAGATTCCACTAGGCCAACTGTAGGTCATGAGTATGCACAATTTC
>TBC1D8B_1(+)|no_TATA|0
CGACGATCACGGTACCTTCATTTCACCGCACACCCCCACCCATTCTAACAT
>FGG_1(-)|TATA|0
CGCTCCCCCAGGAGTGTACTCCTGGTCAAAAGAGCGACATCACACGACGTA
>FAM160B2_1(+)|no_TATA|1
CAGAGATCGCGCCACTACACTCCAGCCTGGGCGACCAAGCGAGACCCTGTC
>SYCP2L_4(+)|no_TATA|0
GGTATTGCGCCCGATATAGCTGCCCGAGCGTGTCATAACCCAAACGCGCCC
>SERPINI1_2(+)|no_TATA|1
TCGTAGTCCCGCTCGCCGGCGTTGTCCGCAATTCCTGCAATCGGAAGACTA
>ILDR1_1(-)|no_TATA|1
AGTTATATCCACATAATTCTAGAAGGAGAGGGGGATACATTTGCAGGGGAT
>FLT1_2(-)|no_TATA|1
GCTGTGCAGAACTAATTTGAAAGTAGAGTTCCATGCACTCACTCCAGGACC
>NRCAM_2(-)|no_TATA|0
```

[深入理解预训练（pre-learning）、微调（fine-tuning）、迁移学习（transfer learning）三者的联系与区别_预训练和微调-CSDN博客]