Large Language Models in Computational Biology – A Primer

Jian Ma

@jmuiuc

Ray and Stephanie Lane Professor of Computational Biology
School of Computer Science
Carnegie Mellon University

July 20, 2023 | UCLA CGSI


WE DO THIS
NOT BECAUSE
IT IS EASY,
BUT BECAUSE
WE THOUGHT
IT WOULD BE EASY

This presentation was put together with help from –

Ellie Haber

Wenduo Cheng

Shaoheng Liang

Recent history of Language Models

- Statistical language models
  - predict the next word based on the most recent context, e.g., Markov assumptions
  - hard to build high-order language models

- Neural language models
  - probability of word sequences from neural nets
  - distributed representation of words (e.g., word2vec)

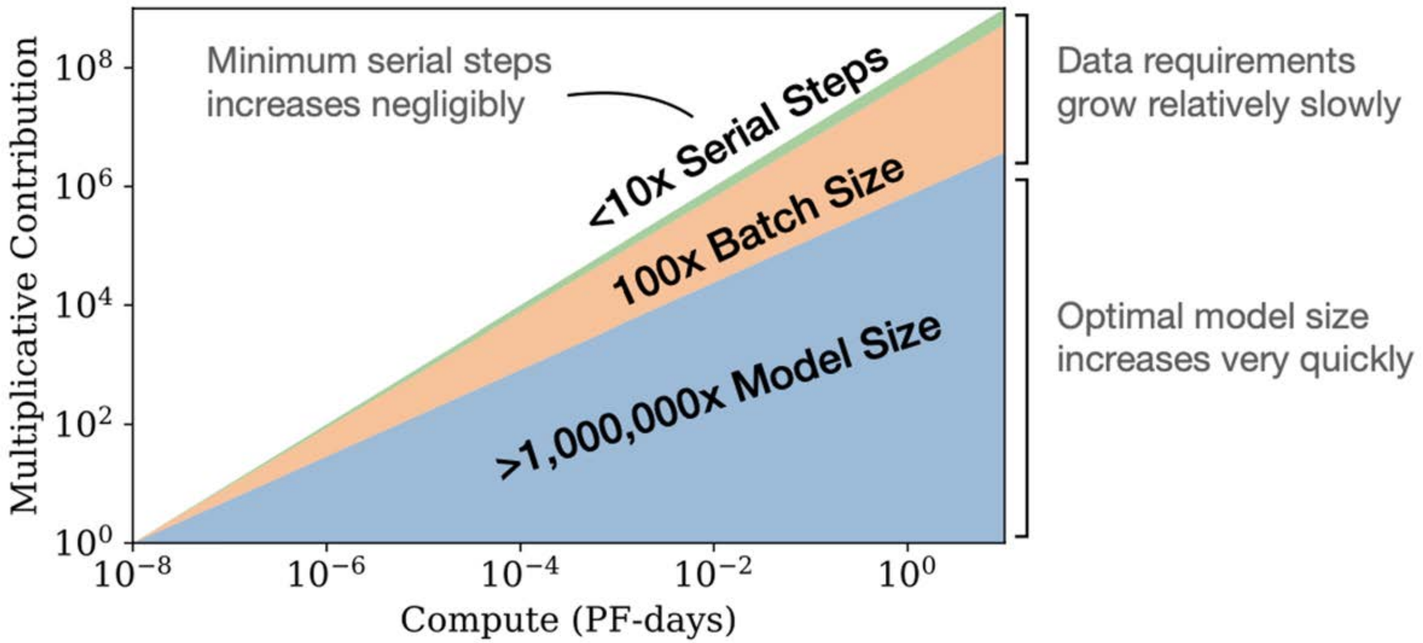
- Pre-trained language models
  - capture context-aware word representations by pre-training
  - pre-training, then fine-tuning on downstream tasks
  - ELMo, pretrained with bi-LSTM – more context sensitive
  - BERT (Google), based on Transformer,
    context-aware, pretrained on large unlabeled data
  - GPT (OpenAI), based on Transformer

Rosenfeld. *Proc of IEEE* 2000
Bengio et al. *NIPS* 2003
Mikolov et al. *ICLR* 2013
Peters et al. *NAACL* 2018
Devlin et al. *NAACL* 2019
Radford et al. *OpenAI* 2018
Zhao et al. *arXiv* 2023

Large Language Models

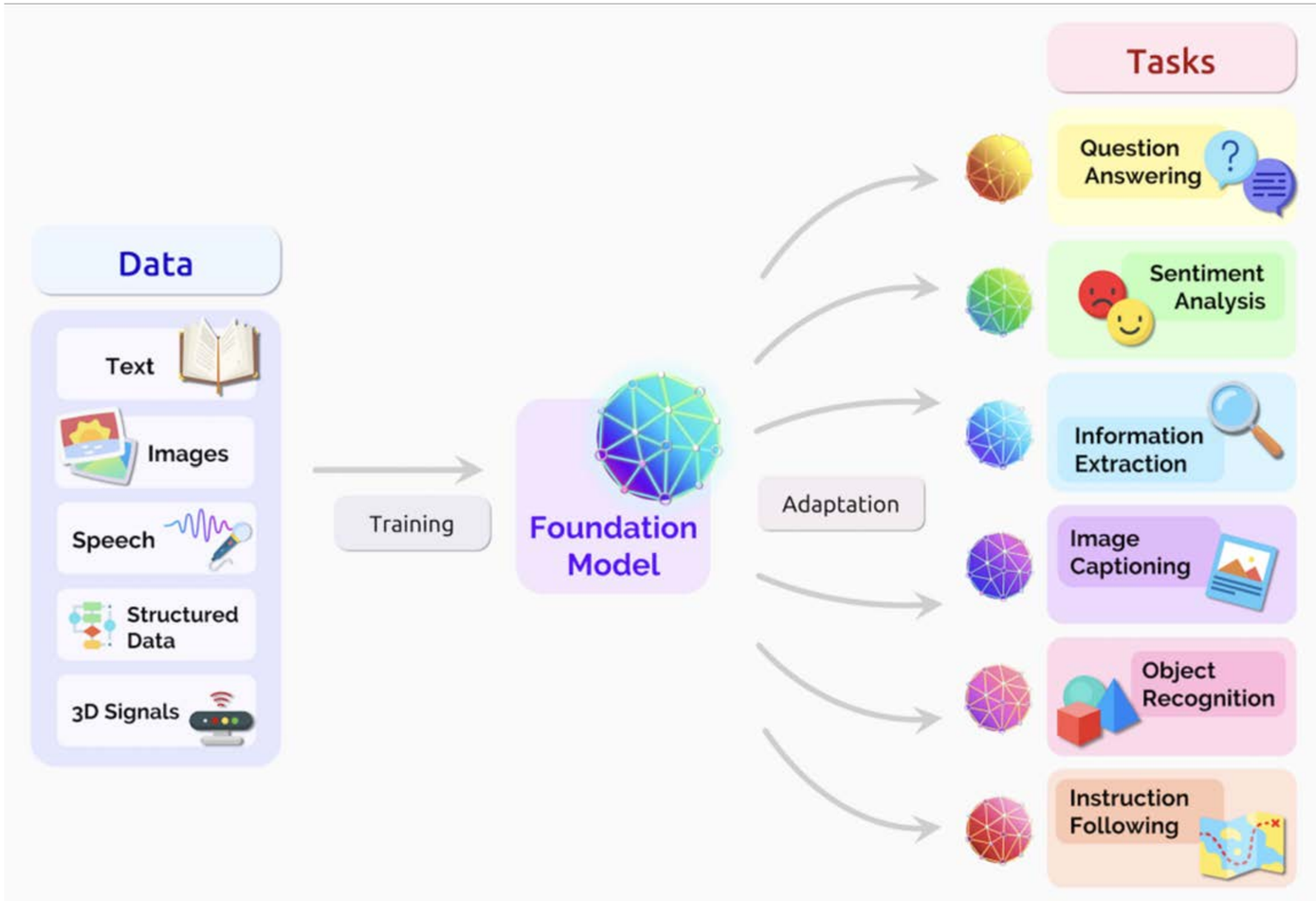
- Large language models
  =
  large-sized pretrained language models

- Scaling laws
  - Kaplan et al. 2020 (OpenAI)
  - Chinchilla scaling – Hoffmann et al. *NeurIPS* 2022



- Differences compared to LMs
  - Large # of model parameters
  - LLMs display some surprising “emergent abilities”
  - LLMs harbor powerful features such as prompting interface (e.g., GPT-4 API)
  - LLMs need tremendous resource to build

What is a Foundation Model?



"On the Opportunities and Risks of Foundation Models"
Bommasani et al. Stanford CRFM 2022

- Foundation models are a replacement for task-specific models

- Large-scale **pretraining** on large unlabeled datasets
- **Finetuning** for diverse downstream tasks

- Self-supervised learning
- Transfer learning

- GPT-4, DALL-E 2, BERT, etc.

Why we need Transformer?

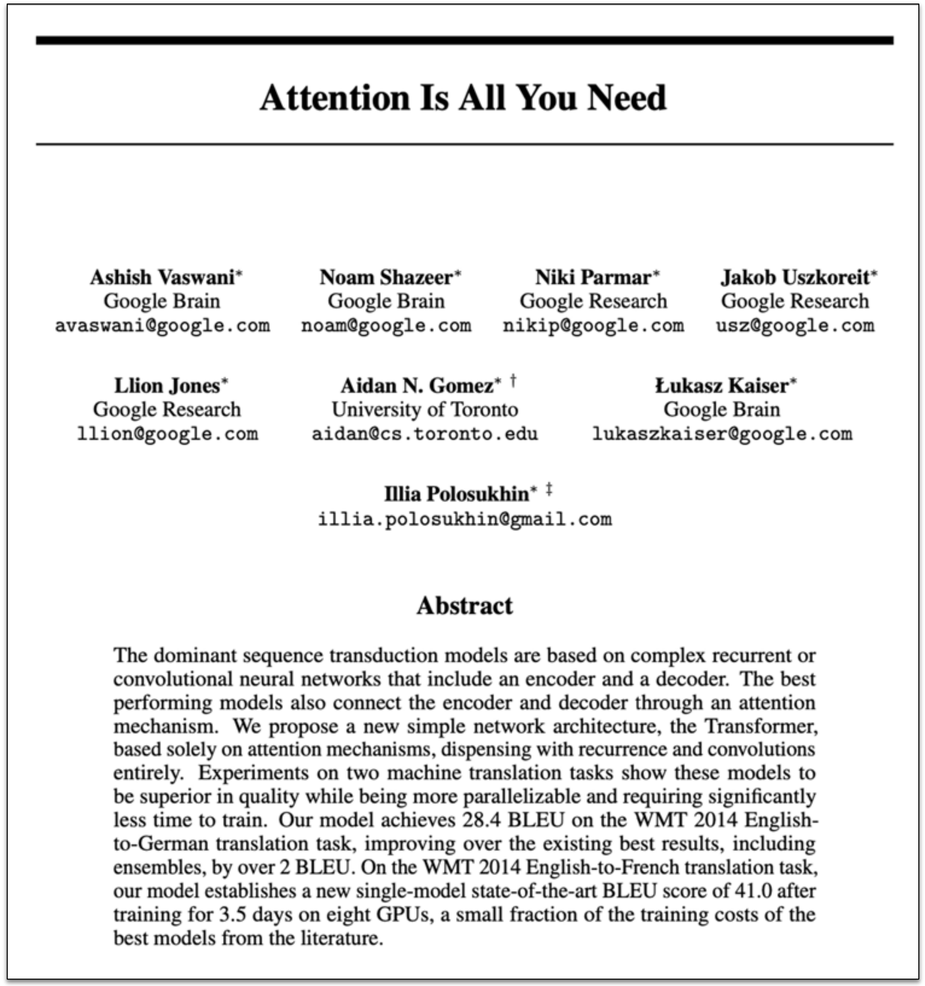
- We need dynamic representations for context-specific information.

  e.g., “I **like** it” vs. “I do not **like** it”

  The word **“like”** should have different representations because it has opposite meanings – different context

  - Vanilla RNNs are slow with poor memory retention.
  - LSTMs are still slow and sequential.
  - CNNs can be parallelized but lack dynamic context capture.

- Transformer combines the benefits of dynamic computation, good memory, and parallelizability



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

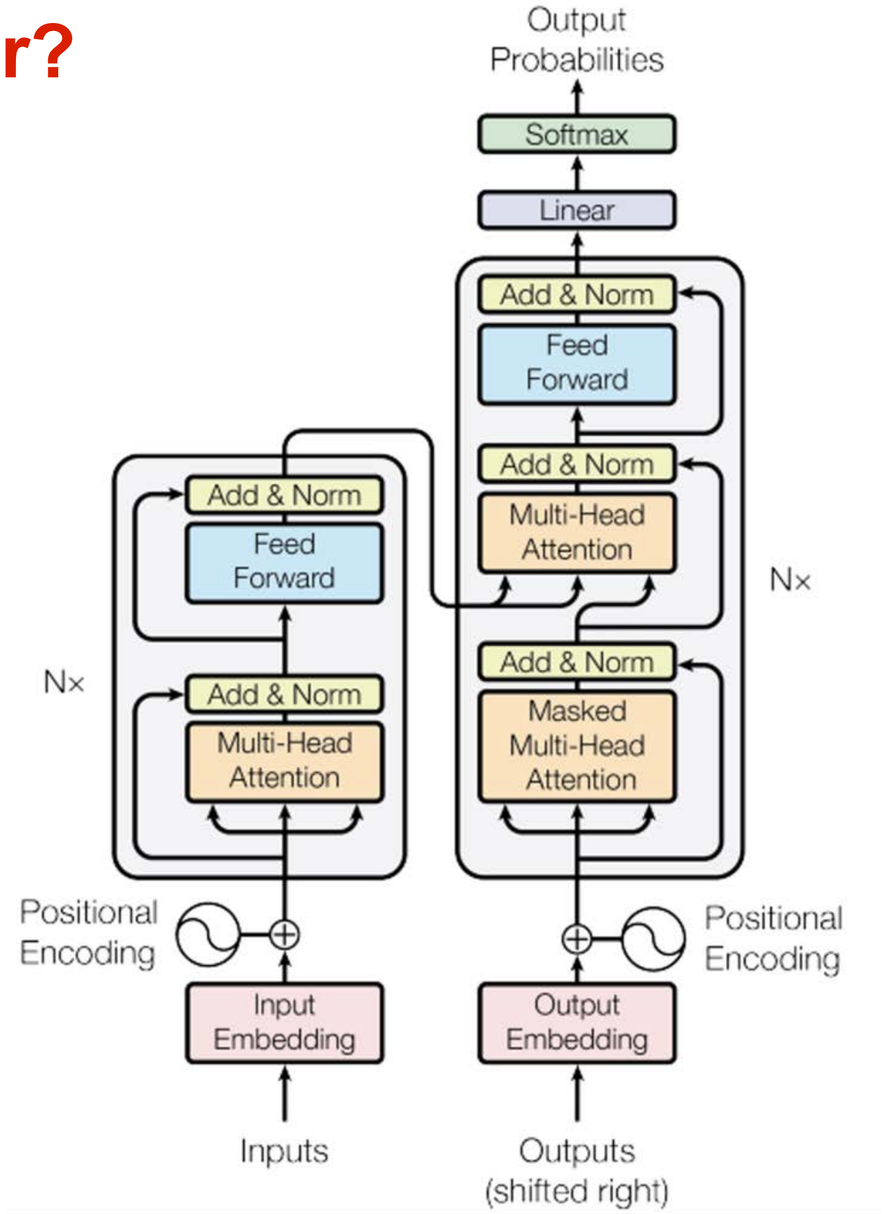
Cited 82,354 times so far

What is Transformer?

- Transformer architecture designed to handle sequential data
  (e.g., text, time-series)
  - captures **long-range** dependencies and context

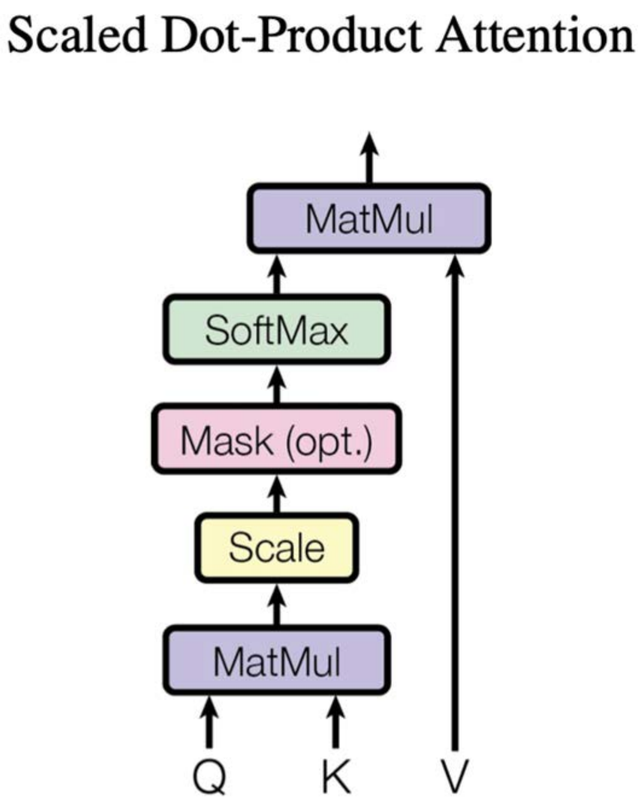
- Utilizes the **self-attention** mechanism to extract features for each word in a sentence.

- Consists of an encoder and a decoder. Both contains a core block of “an attention and a feed-forward network” repeated *N* times.



“Attention Is All You Need”
Vaswani et al. *NeurIPS* 2017

What is Attention?



“Attention Is All You Need”
Vaswani et al. *NeurIPS* 2017

- Attention allows models to focus on different parts of an input sequence

She ate cake and it was delicious

- Calculating attention scores:

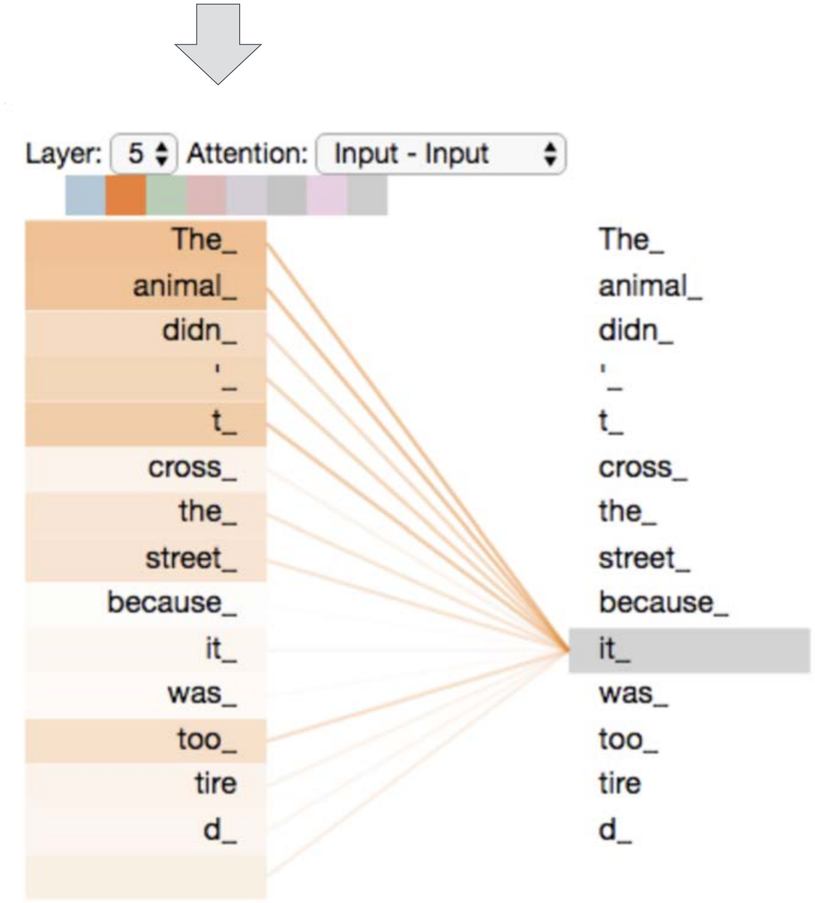
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V$$

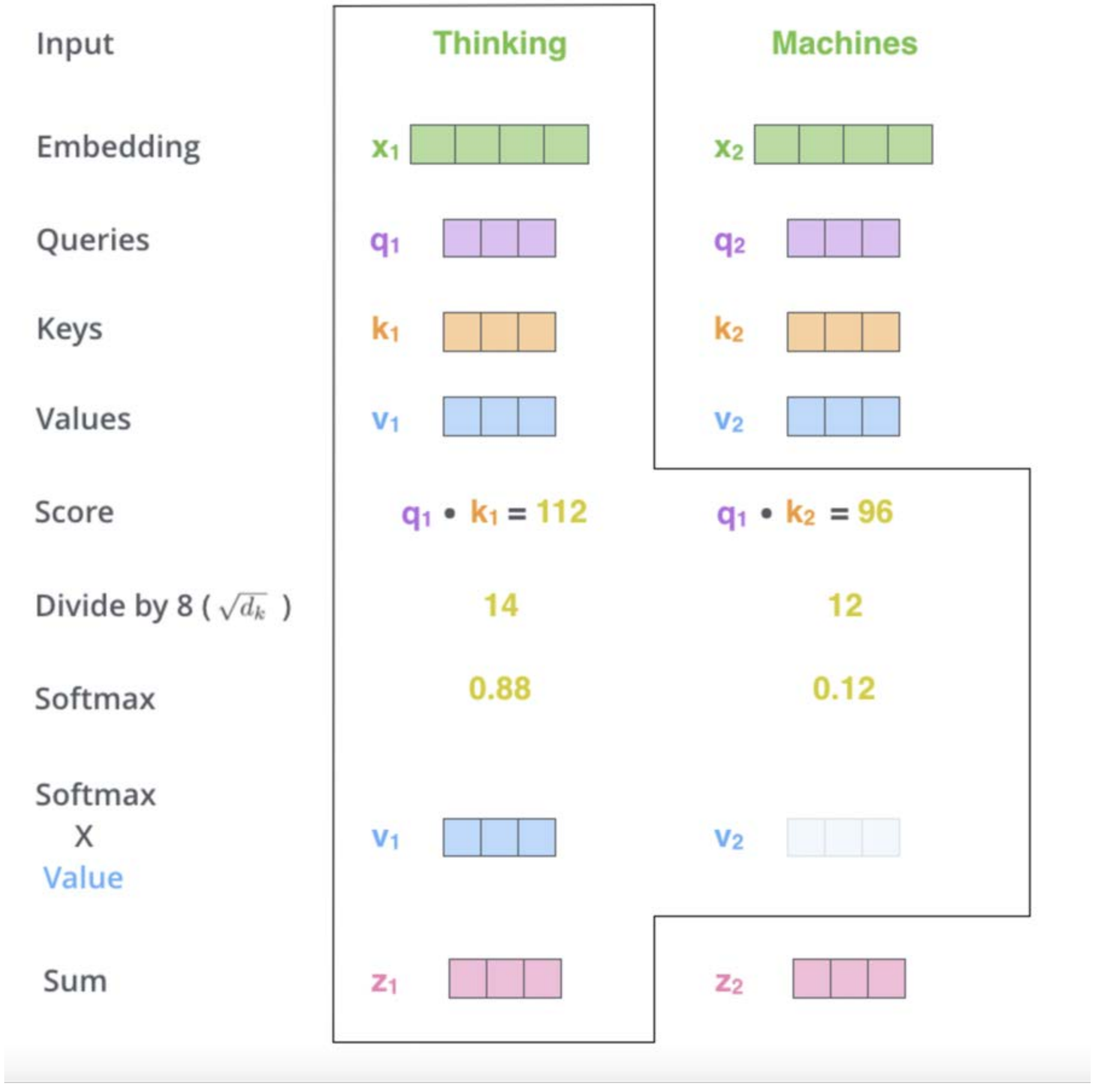
  - Q: current position in input seeking context from other positions
  - K: captures the information that Q attends to
  - V: actual content associated with positions in input

What is Attention?

"The animal didn't cross the street because it was too tired"

Self-Attention





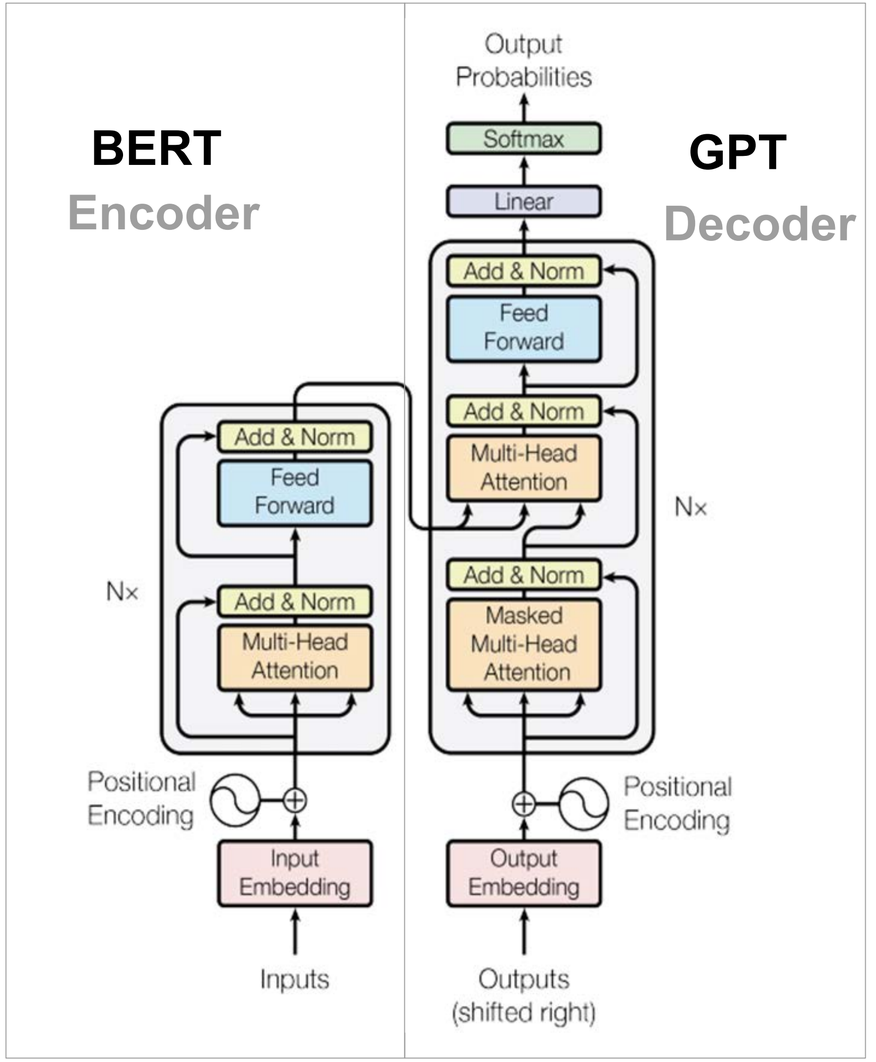
https://jalammar.github.io/illustrated-transformer/

Transformer architecture in BERT and GPT

BERT-style

(Encoder-only or Encoder-Decoder)

- Training:
  Masked Language Models
- Model type:
  Discriminative
- Pretrain task:
  Predict masked words



Yang et al. *arXiv* 2023

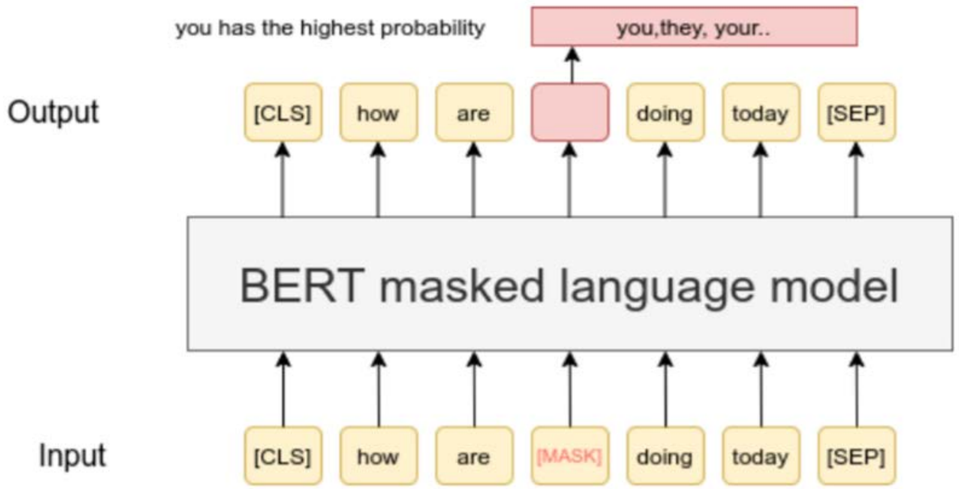
GPT-style

(Decoder-only)

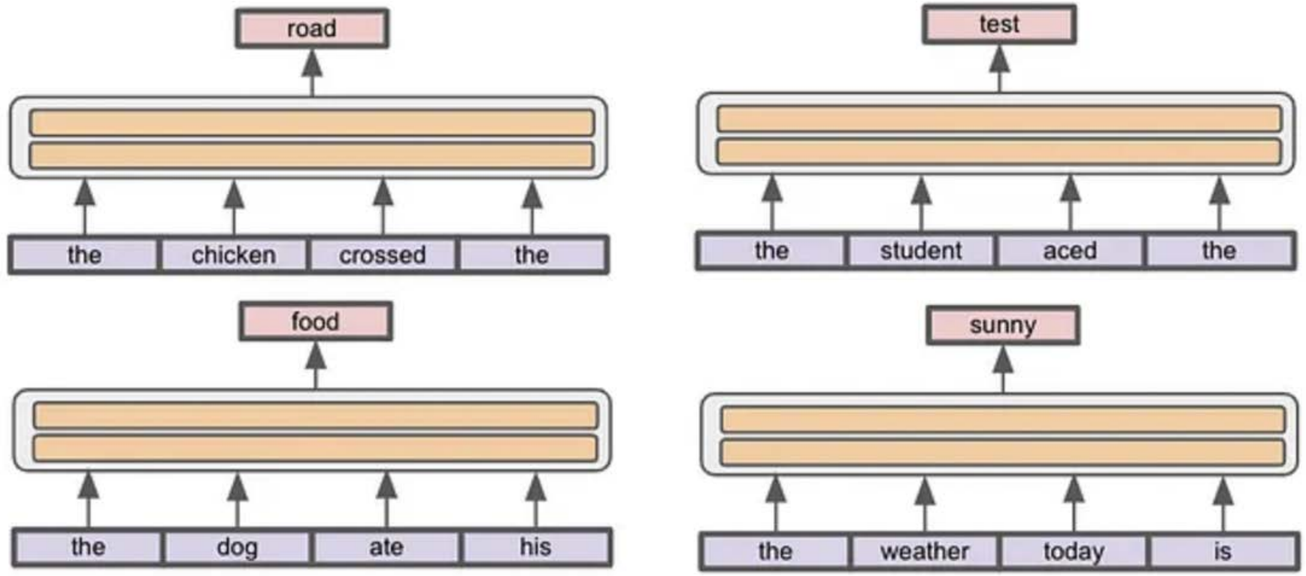
- Training:
  Autoregressive Language Models
- Model type:
  Generative
- Pretrain task:
  Predict next word

Pre-training techniques

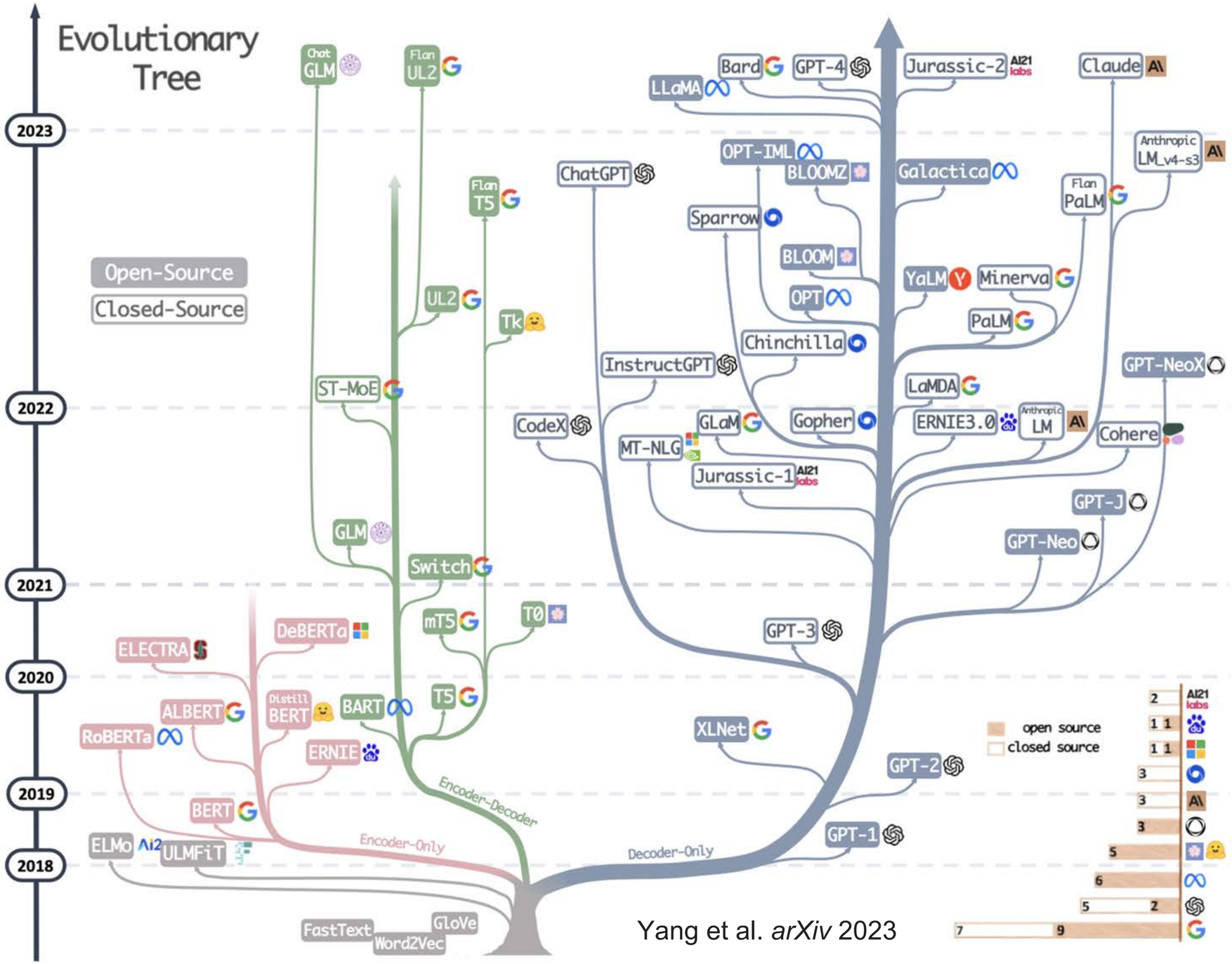
Masked Language Modeling



Autoregressive Language Modeling



https://www.sbert.net/examples/unsupervised_learning/MLM/README.html
https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a



Yang et al. *arXiv* 2023

GPT	0.11B
BERT	0.34B
GPT-2	1.5B
Turing-NLG	17.2B
GPT-3	175B
Switch	1.6T
MT-NLG	530B
JURASSIC-1	178B
GLaM	1.2T
LaMDA	137B
PaLM	540B
OPT	175B
YaLM	100B
BLOOM	176B
Bard	137B
LLaMA	65B
GPT-4	1.7T

Source:
https://github.com/Hannibal046/Awesome-LLM

Application of (large) language models in genomics

- Large pretrained models can be utilized for finetuning on downstream tasks with limited training data

- Data sparsity problem in biology
  - noisy/sparse data
  - incomplete data in biology, e.g., rare disease, precious samples

- Embeddings with more generalized knowledge can help mitigate batch effects and biases in the data

- Today, we introduce recent work in the following **two directions**:
  - Modeling genomic sequences
  - Modeling single cell data

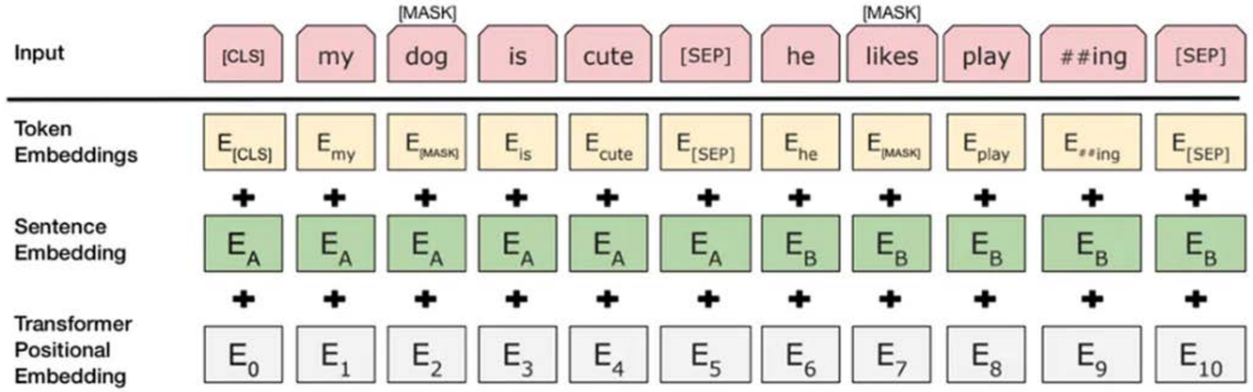
Some of the recent language models for genomic sequence

Model	*Paper*	*# Parameters*	*Architecture*	*Training Data*	*Downstream Tasks*
Big Bird	Zaheer et al. NeurIPS 2020	127.47M	BERT+Sparse Attention	human reference genome	promoter prediction, chromatin profile prediction
DNABERT	Ji et al. Bioinformatics 2021	3-mer 86M 4-mer 86M 5-mer 87M 6-mer 89M	BERT	human reference genome	splice site prediction, chromatin profile prediction, promoter prediction
Enformer	Avsec et al. Nat Methods 2021	23.67M	CNN+Transformer (may not be considered as LM)	human and mouse genome	gene expression prediction, enhancer prioritization, noncoding variant effect prediction

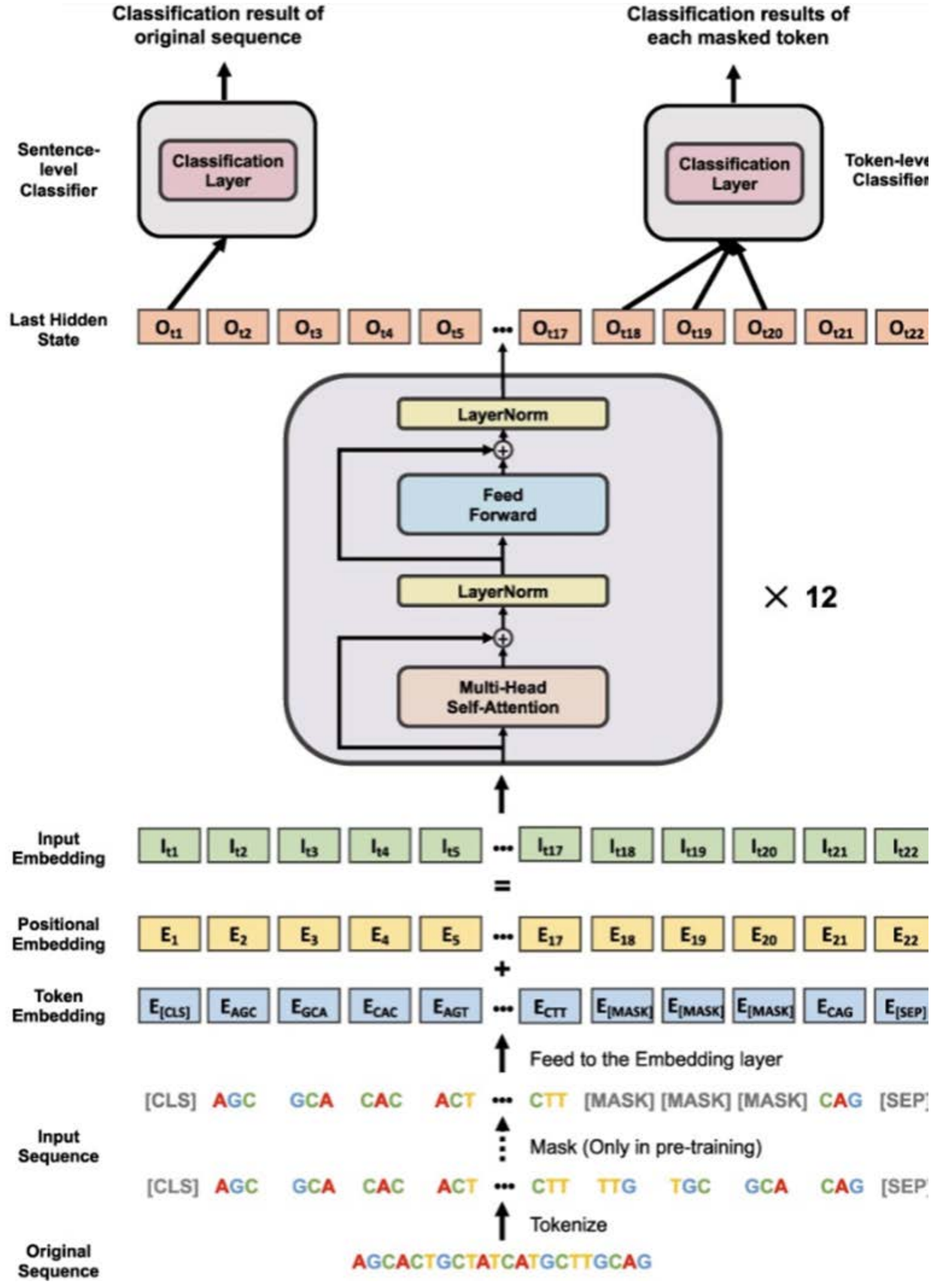
Model	*Paper*	*# Parameters*	*Architecture*	*Training Data*	*Downstream Tasks*
Nucleotide Transformer	Dalla-Torre et al. bioRxiv 2023	500M_human_ref 480M 500M_1000G 480M 2B5_1000G 2537M 2B5_multi_species 2537M	Transformer	human reference genome, 3202 human genomes, genome from 850 different species	epigenetic marks prediction, promoter and enhancer prediction, splice site prediction
DNABERT-2	Zhou et al. arXiv 2023	117M	BERT+Flash Attention+ Attention with Linear Biases	multi-species genome dataset from 135 species (32.49B)	promoter prediction, TF prediction, splice site prediction, epigenetic marks prediction, variant classification
HyenaDNA	Nguyen et al. arXiv 2023	tiny 1k small 32k medium 160k medium 450k large 1M	Large Convolutional Model	human reference genome	epigenetic marks prediction, promoter and enhancer prediction, splice site prediction

DNABERT

- Pre-trained BERT for DNA sequences based on the human reference genome
- Overlapping k-mer tokenization
- Downstream tasks:
  - promoter region prediction, transcriptional factor binding site prediction, splice site prediction, functional variants identification



BERT input representation

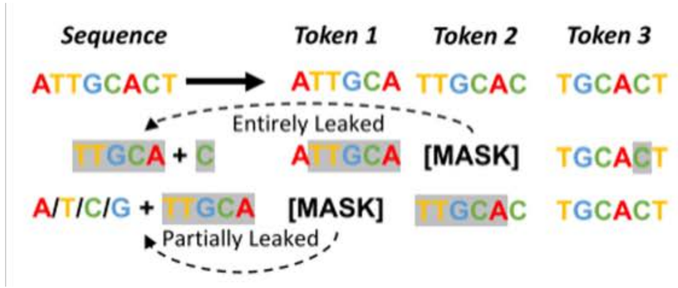


DNABERT input representation
Ji et al. *Bioinformatics* 2021

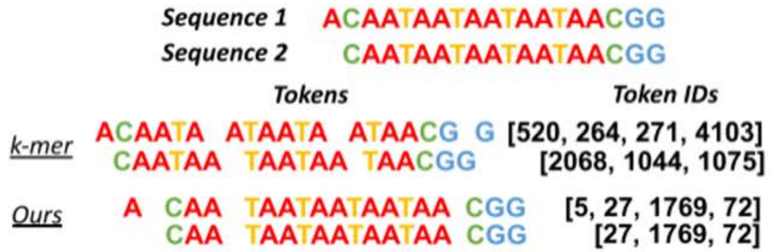
DNABERT vs. DNABERT-2

- Tokenization: overlapping kmer tokenization → byte pair encoding
  - to prevent information leakage and sample inefficiency

Overlapping k-mer



Non-overlapping k-mer



Byte Pair Encoding

- Replace positional embeddings with attention w/ linear biases
  - to overcome the input length limit of DNABERT

- Other deep learning tricks to increase computation and memory efficiency

Zhou et al. *arXiv* 2023

DNABERT-2 results

	Yeast	Mouse	Virus	Human			
	EMP	TF-M	CVC	TF-H	PD	CPD	SSP
DNABERT (3-mer)	49.54	57.73	62.23	64.43	84.63	**72.96**	84.14
DNABERT (4-mer)	48.59	59.58	59.87	64.41	82.99	71.10	84.05
DNABERT (5-mer)	48.62	54.85	63.64	50.46	84.04	72.03	84.02
DNABERT (6-mer)	49.10	56.43	55.50	64.17	81.70	71.81	84.07
NT-500M-human	45.35	45.24	57.13	50.82	85.51	66.54	79.71
NT-500M-1000g	47.68	49.31	52.06	58.92	86.58	69.13	80.97
NT-2500M-1000g	50.86	56.82	66.73	61.99	86.61	68.17	85.78
NT-2500M-multi	58.06	67.01	**73.04**	63.32	**88.14**	71.62	**89.36**
DNABERT-2	55.98	67.99	71.02	**70.10**	84.21	70.52	84.99
DNABERT-2◆	**58.83**	**71.21**	68.49	66.84	83.81	71.07	85.93

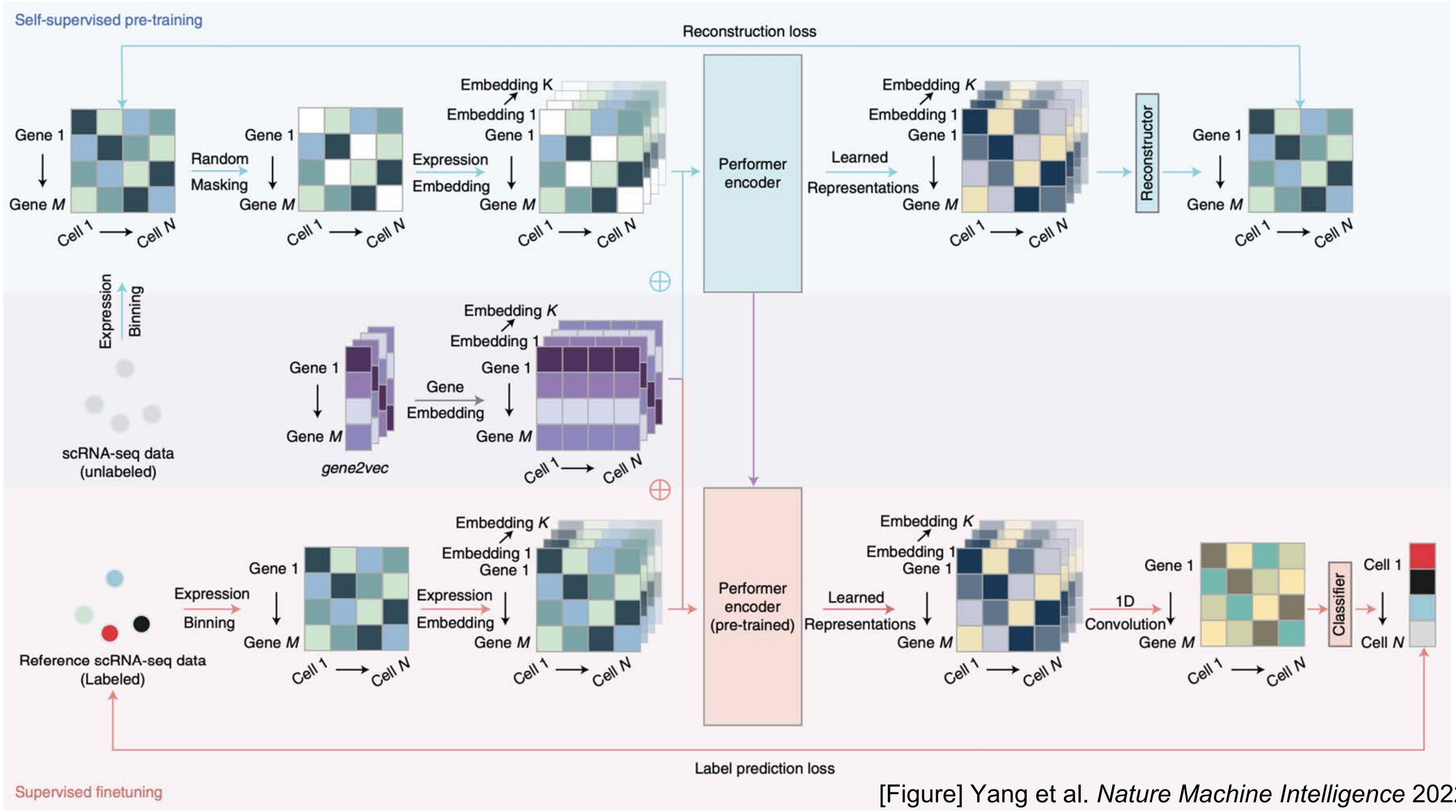
Zhou et al. *arXiv* 2023

EMP: Epigenetic Marks Prediction, **TF-M:** Transcriptional Factor Prediction in Mouse,
CVC: Covid Variant Classification, **TF-H:** Transcriptional Factor Prediction in Human,
PD: Promoter Detection, **CPD:** Core Promoter Detection, **SSP:** Splice Site Detection

Some of the recent language models for single-cell genomics

Model	*Paper*	*# Parameters*	*Architecture*	*Training Data*	*Downstream Tasks*
scGPT	Cui et al. bioRxiv 2023	51 million	autoregressive transformer	33 million normal human cells (51 tissues, 441 studies)	cell-type annotation, multi-batch integration, multi-omic integration, perturbation prediction, and GRN inference
scBERT	Yang et al. Nature MI 2022	5 million	Performer (allowing for longer inputs)	209 human single-cell datasets comprising 74 tissues with 1M+ cells	cell type annotation
Geneformer	Theodoris et al. Nature 2023	40 million	Transformer	Genecorpus-30M- 29.9 million human single-cell transcriptomes	gene dosage sensitivity, chromatin, network dynamics,
scFoundation	Hao et al. bioRxiv 2023	100 million	Transformer (w/ trick to reduce # words)	50 million human cells (100+ tissue types, normal and disease)	clustering, perturbation prediction, drug response

Full scBERT model training scheme



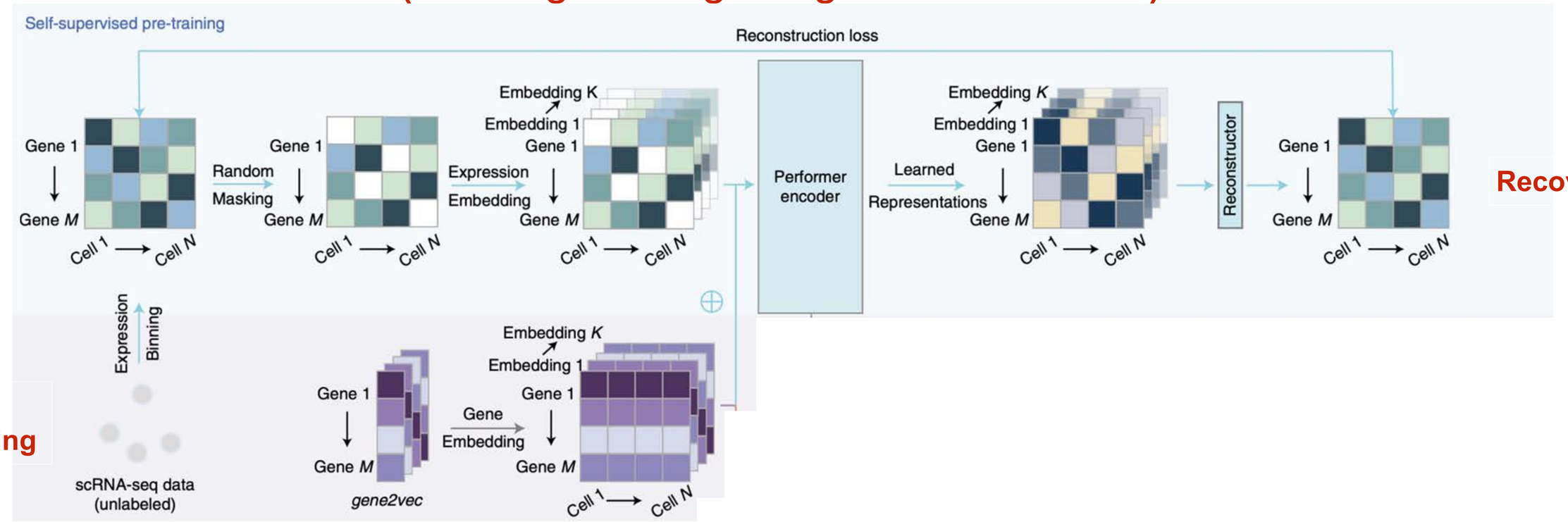
[Figure] Yang et al. *Nature Machine Intelligence* 2022

Part I: Self-supervised pretraining on large-scale datasets

(learns general gene-gene interactions)

Random Masking

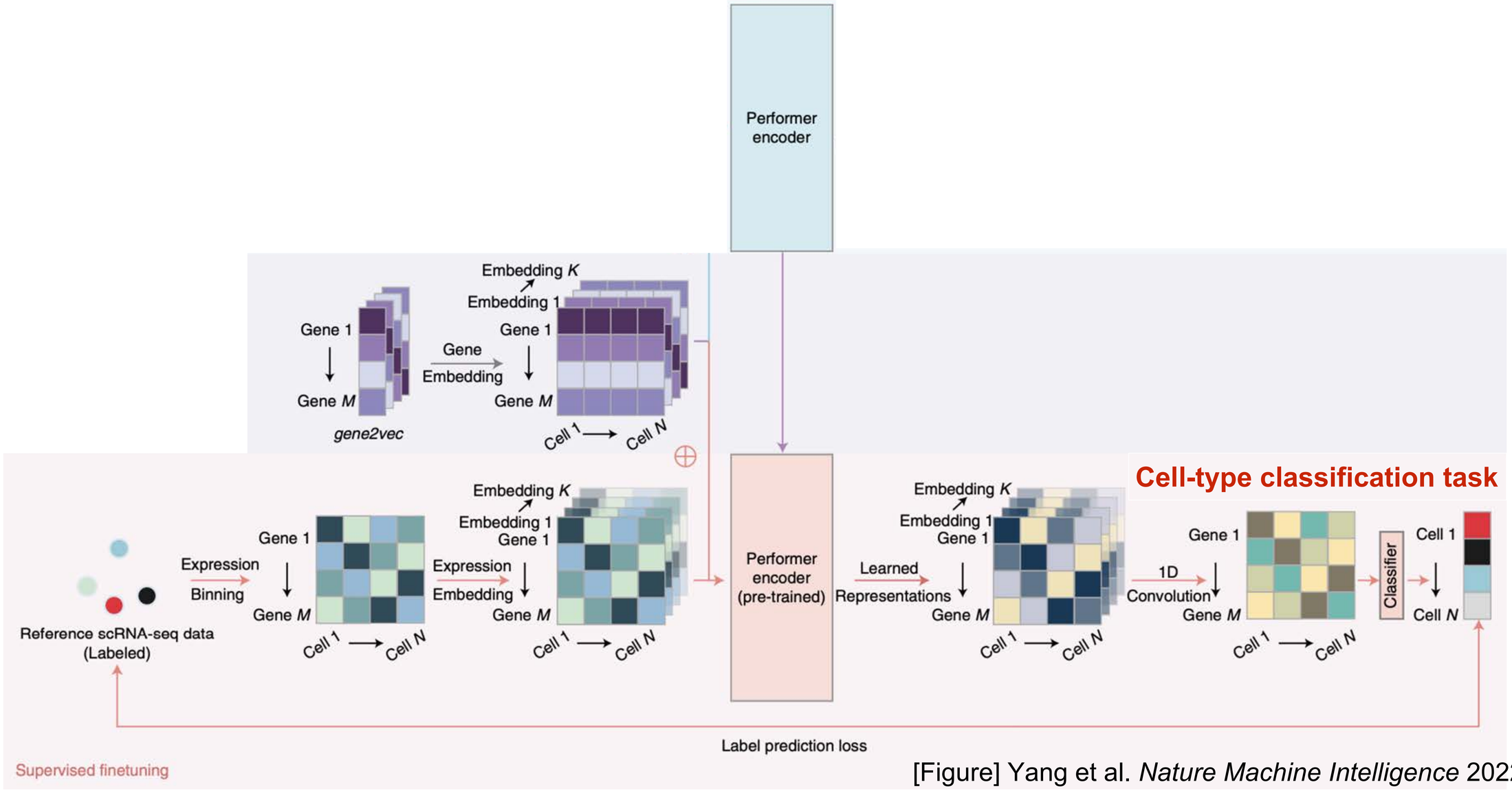
Gene Embedding



[Figure] Yang et al. *Nature Machine Intelligence* 2022

Part II: Supervised finetuning for specific tasks

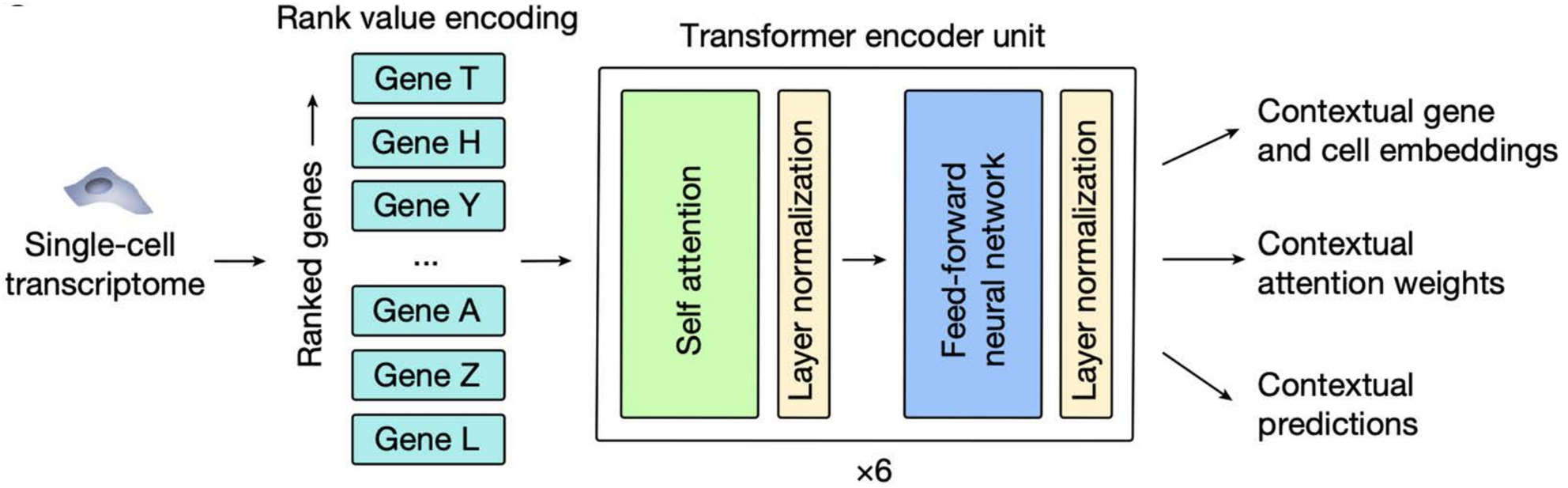
(learns task / dataset specific characteristics)



[Figure] Yang et al. *Nature Machine Intelligence* 2022

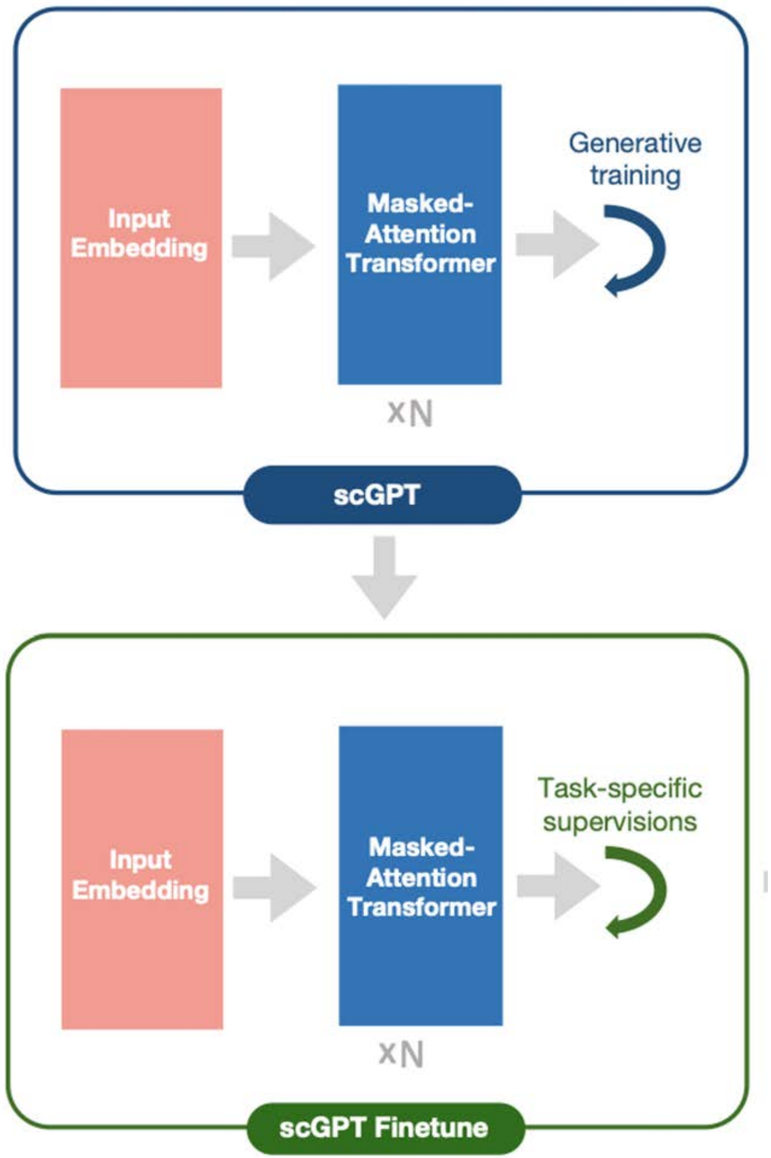
Geneformer

- Pretrained on 30 million scRNA-seq to enable context-specific predictions
- Discretize gene expression by ranking genes according to their expression
- Encodes network hierarchy in the attention weights of the model
  - Context awareness using attention allows for predictions specific to cell states
  - Attentions reflect important genes such as TFs and central regulatory nodes
- In silico perturbation: remove a gene, compare cell and gene embeddings



Theodoris et al. *Nature* 2023

scGPT overview

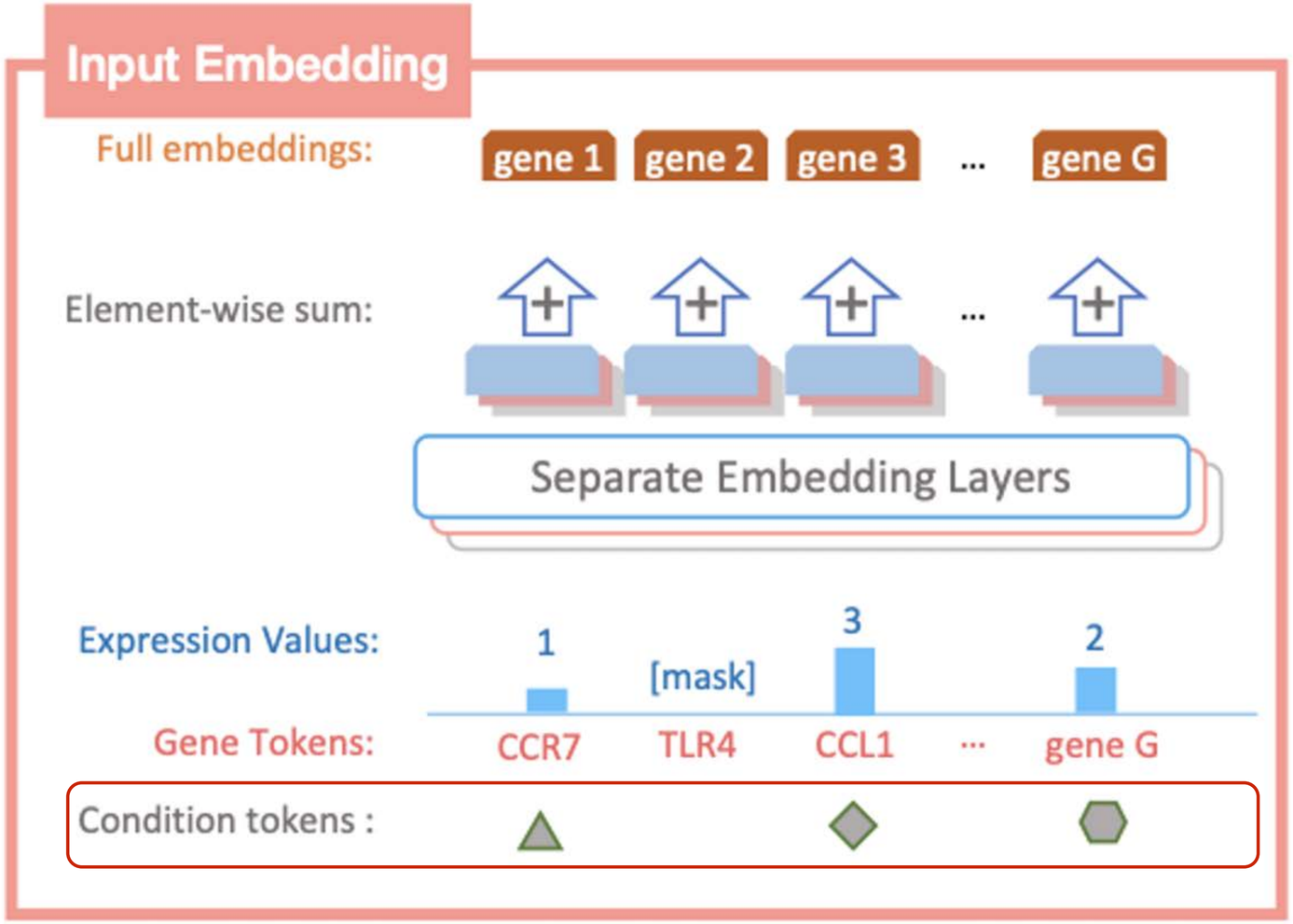


- Generative pretraining on on 30+ million normal human cells from 50+ tissues
- Learn insights concerning genes and cells

- Adapt to specific tasks:
  - cell-type annotation
  - multi-batch integration
  - multi-omic integration
  - perturbation prediction

Cui et al. *bioRxiv* 2023

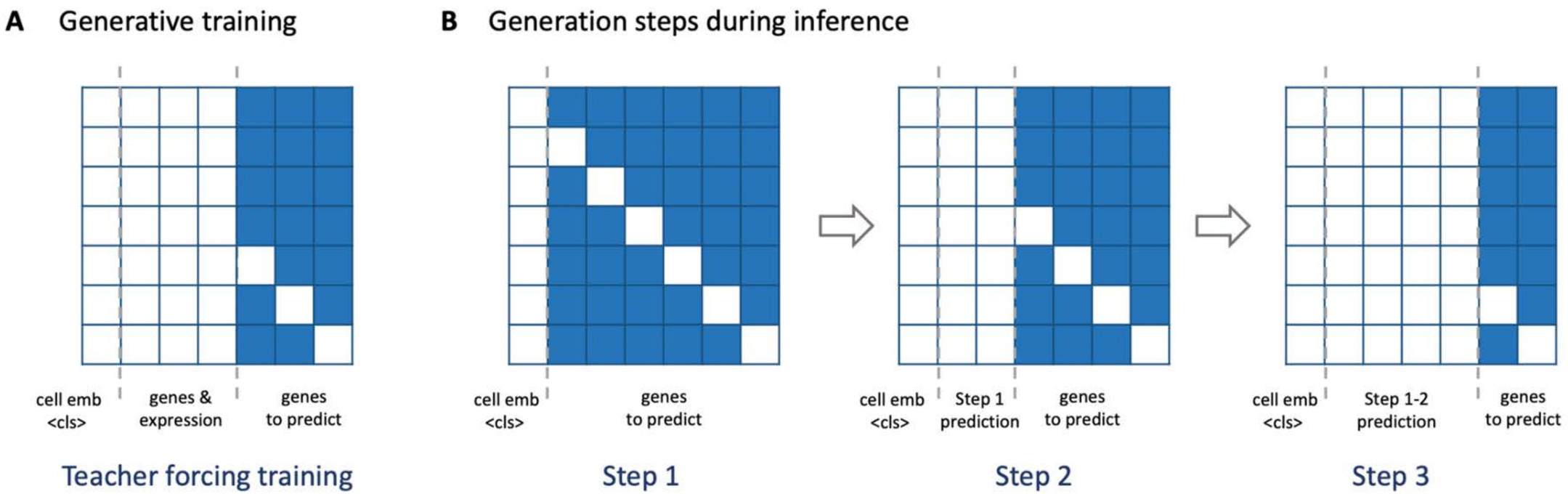
Input embedding for scGPT



- An additional set of tokens to integrate meta information (e.g., perturbations)

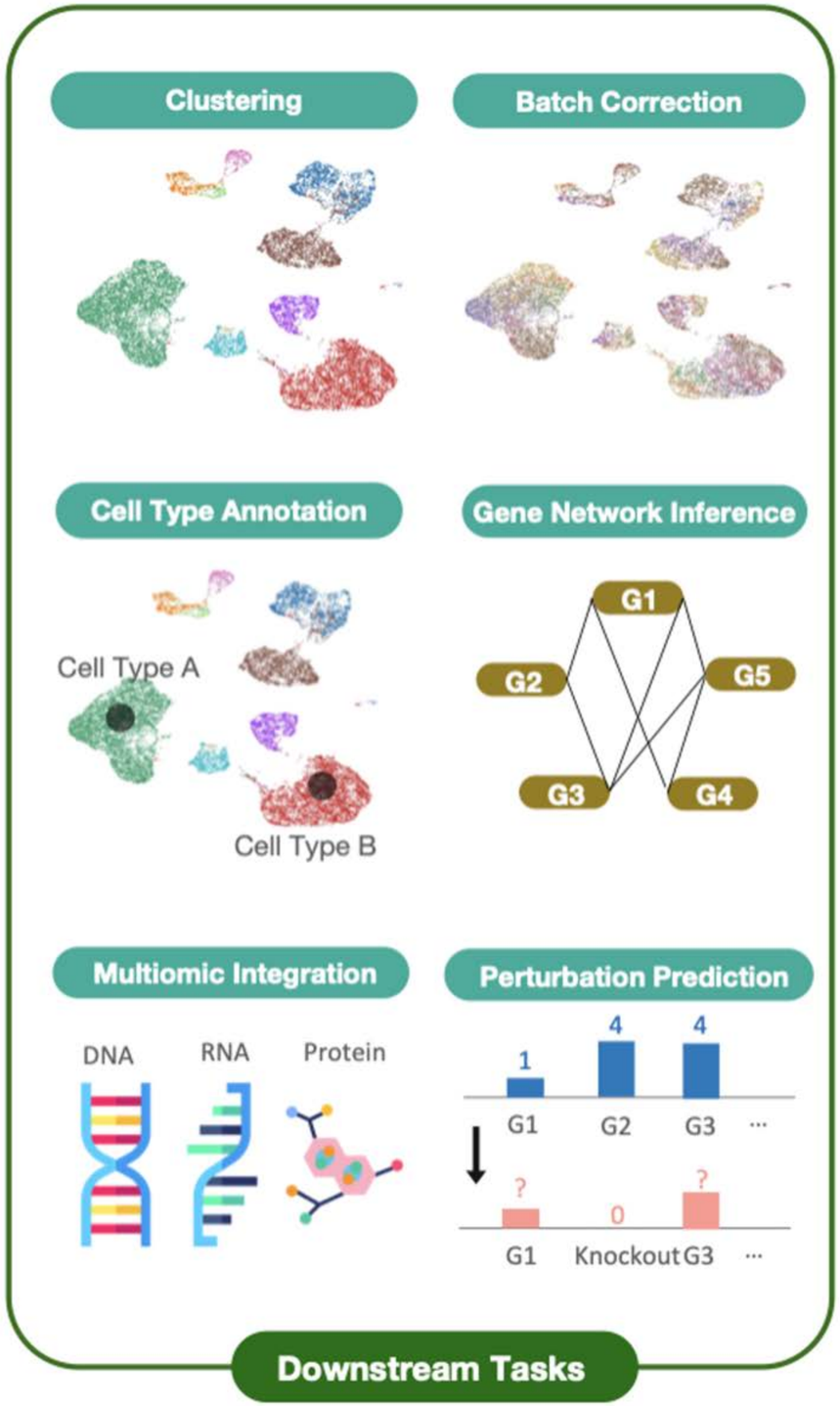
Cui et al. *bioRxiv* 2023

Generative pretraining for scGPT



- Attention mask for generative pretraining:
  - Use known genes to predict unknown genes
  - Teacher-forcing training

Cui et al. *bioRxiv* 2023



scGPT finetuning objectives

- Fine-tuning objectives facilitate the learning of biologically meaningful cell and gene representations for diverse downstream tasks

- Self-supervised objectives:
  - Gene expression prediction (by MLP)
  - Gene expression prediction for cell modeling (by querying)
  - Elastic cell similarity (for data integration)

- Supervised objectives:
  - Domain adaptation via reverse back-propagation
  - Cell type annotation

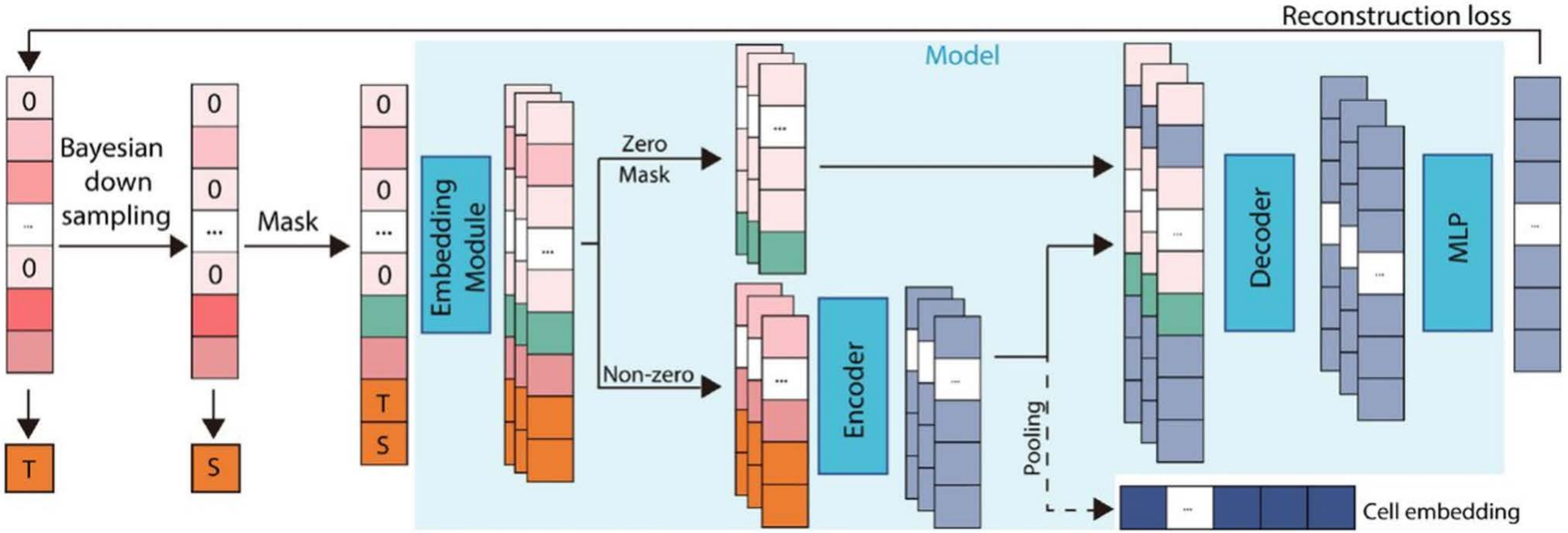
Cui et al. *bioRxiv* 2023

scGPT vs. scBERT: cell type annotation

Dataset	Model	Classification Metrics			
		Accuracy	*Precision*	*Recall*	*MacroF1*
Myeloid	scGPT (fine-tuned)	**0.642**	**0.366**	**0.347**	**0.346**
	scGPT (from-scratch)	0.606	0.304	0.339	0.309
	TOSICA	0.488	0.316	0.276	0.275
	scBert	0.525	0.331	0.323	0.298
Multiple Sclerosis	scGPT (fine-tuned)	**0.856**	**0.729**	**0.720**	**0.703**
	scGPT (from-scratch)	0.798	0.660	0.623	0.600
	scBert	0.785	0.604	0.624	0.599
	TOSICA	0.758	0.664	0.585	0.578
hPancreas	scGPT (fine-tuned)	**0.968**	**0.735**	**0.725**	**0.718**
	scGPT (from-scratch)	0.936	0.665	0.668	0.622
	TOSICA	0.960	0.661	0.681	0.656
	scBert	0.964	0.699	0.689	0.685

- scGPT outperforms scBERT for downstream task of cell type annotation
- Speaks to benefits of generative pretraining

scFoundation



Read depth adaptation T→S: for better downstream analysis such as clustering

Run encoder only on genes with non-zero expression: more efficient training without explicit feature selection

Hao et al. *bioRxiv* 2023

Open questions

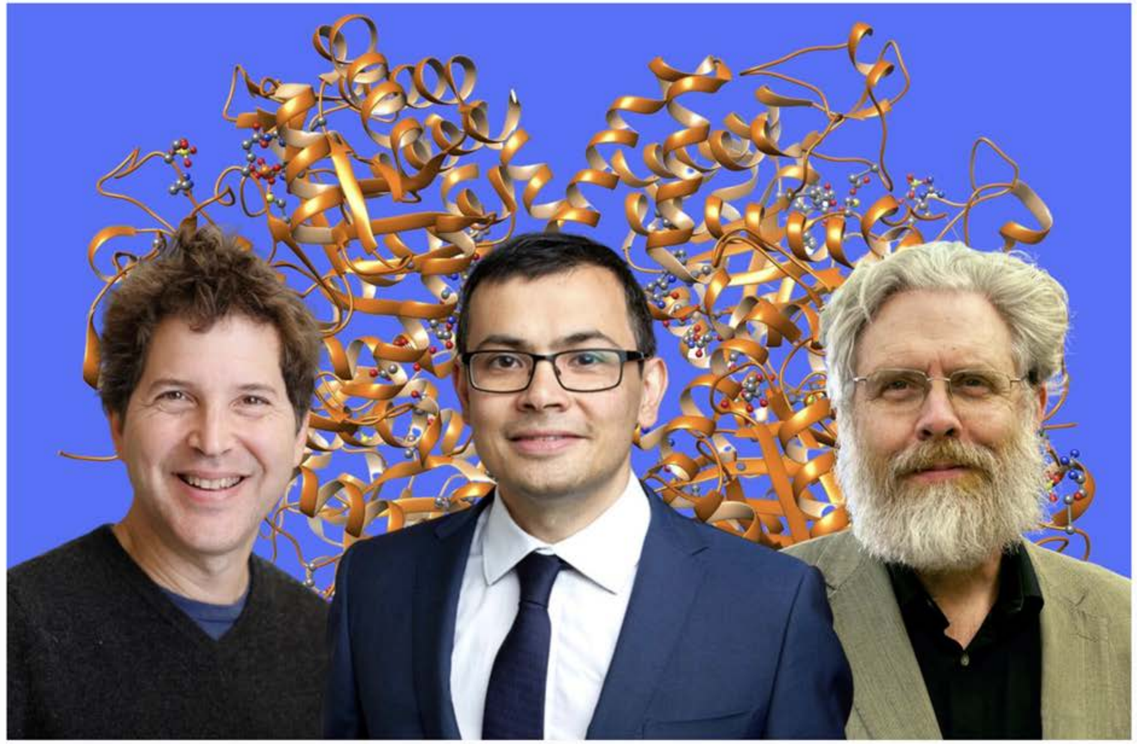
- How to better evaluate LLMs? How to make LLMs more accessible?
- How to embed cell/gene to better maintain biological contexts?
- How to incorporate prior knowledge into the neural network?
- How much finetuning is sufficient for a specific task/dataset? Will better designed pre-training tasks help shorten finetuning?
- How to extract the knowledge claimed to be distilled by the model?
- Do we have enough data available to pretrain LLMs or Foundation Models for various modalities in genomics?
- DNA and single-cell LLMs have comparable performance compared to existing approaches – need more challenging problems. What are the important problems for LLMs?
- Specific LLMs from molecular and cell biology literature + genomics data?
- Reliable hallucinations from LLMs => new biological hypothesis?

Forbes

The Next Frontier For Large Language Models Is Biology

Rob Toews Contributor

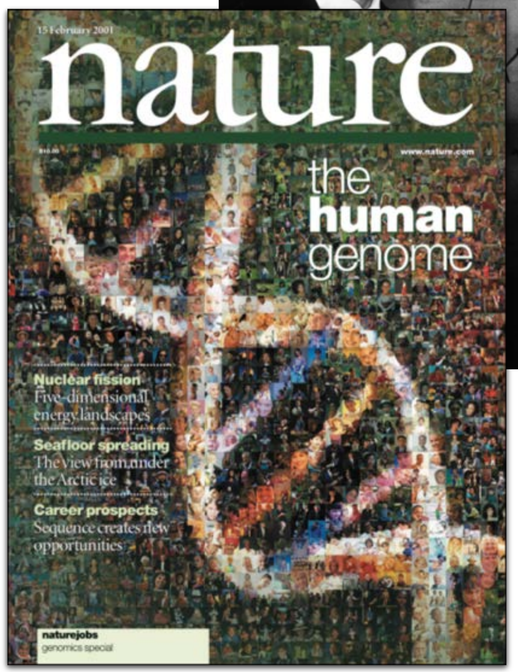
Jul 16, 2023



David Baker (University of Washington), Demis Hassabis (DeepMind) and George Church (Harvard) have helped pioneer the field of AI-driven protein design.

PHOTO SOURCE: U OF W, ROYAL SOCIETY, HARVARD

Large language models like GPT-4 have taken the world by storm thanks to their astonishing command of natural language. Yet the most significant long-term opportunity for LLMs will entail an entirely different type of language: the language of biology.



No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three interwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

This figure is purely diagrammatic. The two ribbons symbolise the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Submit your work to RECOMB 2024

- **Abstract registration deadline:** 11:59 PM, October 16, 2023 AoE
- **Full paper submission deadline:** 11:59 PM, October 20, 2023 AoE