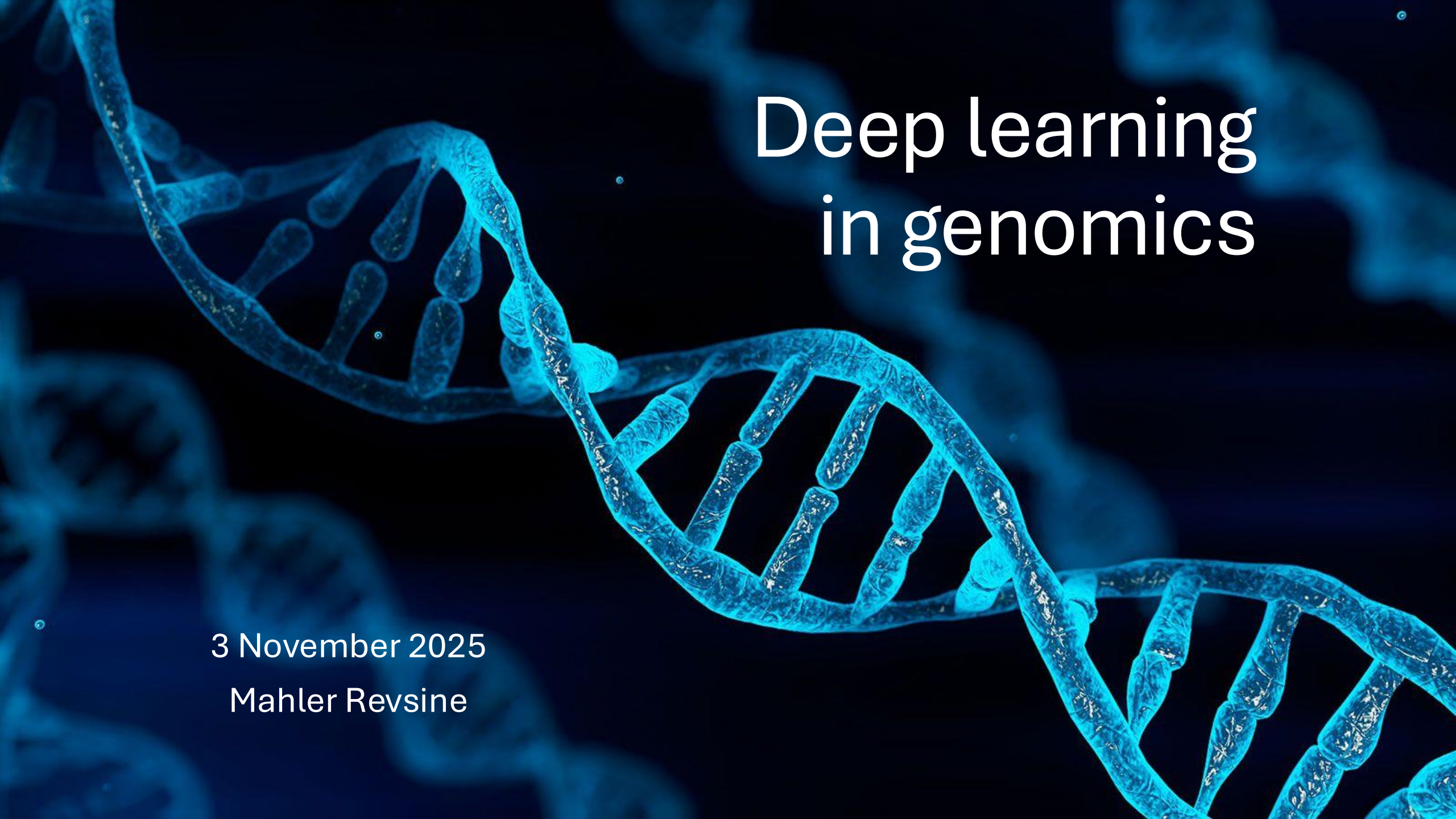


# Deep learning in genomics

3 November 2025

Mahler Revsine

# Announcements

- Assignment 4 grades released
  - Great scores overall! ☺
- Preliminary project report due tonight
- Assignment 5 due Friday night
  - Last chance to use late days

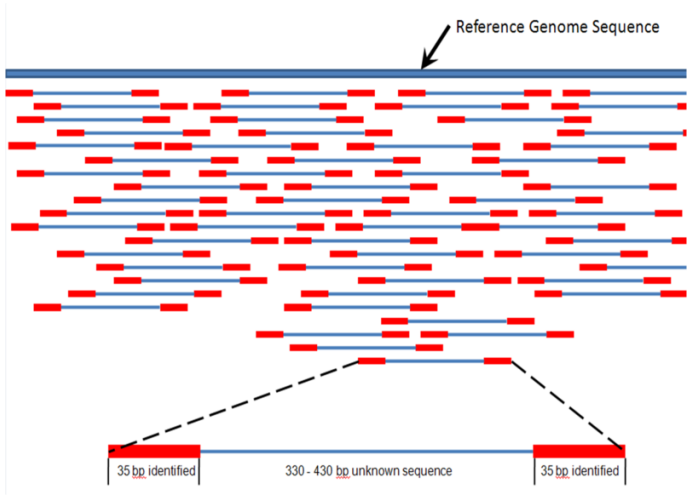
# Deep Learning in Genomics Journal Club

- Discussion group for students interested in genomic deep learning
- We meet every other Tuesday at 12pm
- Participation is optional, feel free to just sit and listen
- Next meetings 11/4 and 11/18
- Email me if interested! All are welcome

Imagine you are a doctor treating a patient with some condition

You sequence their DNA…


…align it to the reference genome…




…and find all of their variants

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   rs4477212   a   .   .   .   .   GT  0/0
chr1    752566  rs3094315   g   A   .   .   .   GT  1/1
chr1    752721  rs3131972   A   G   .   .   .   GT  1/1
chr1    798959  rs11240777  g   .   .   .   .   GT  0/0
chr1    800007  rs6681049   T   C   .   .   .   GT  1/1
chr1    838555  rs4970383   c   .   .   .   .   GT  0/0
chr1    846808  rs4475691   C   .   .   .   .   GT  0/0
chr1    854250  rs7537756   A   .   .   .   .   GT  0/0
chr1    861808  rs13302982  A   G   .   .   .   GT  1/1
chr1    873558  rs1110052   G   T   .   .   .   GT  1/1
chr1    882033  rs2272756   G   A   .   .   .   GT  0/1
chr1    888659  rs3748597   T   C   .   .   .   GT  1/1
chr1    891945  rs13303106  A   G   .   .   .   GT  0/1
```

**How do you determine which variant(s) are responsible for their condition?**

# What does a DNA variant do?

| Method | Drawback(s) |
|---|---|
| Query it in a database | The variant might not exist in a database<br>The variant might act differently under different conditions |
| Assume loss-of-function of nearby gene | Not always an accurate assumption<br>Function of nearby gene may not itself be understood |
| Genome-Wide Association Study (GWAS) | Not applicable for variants that only exist in a single person<br>Underpowered for rare variants that only occur in a few people<br>Relies on undersized datasets containing a small number of SNPs<br>Requires rigorous statistical testing<br>Can only model one variant at a time<br>Assumes additive linear effects of variants<br>Difficult to adapt to a polygenic/omnigenic model of traits<br>In practice, only explains a fraction of heritability |

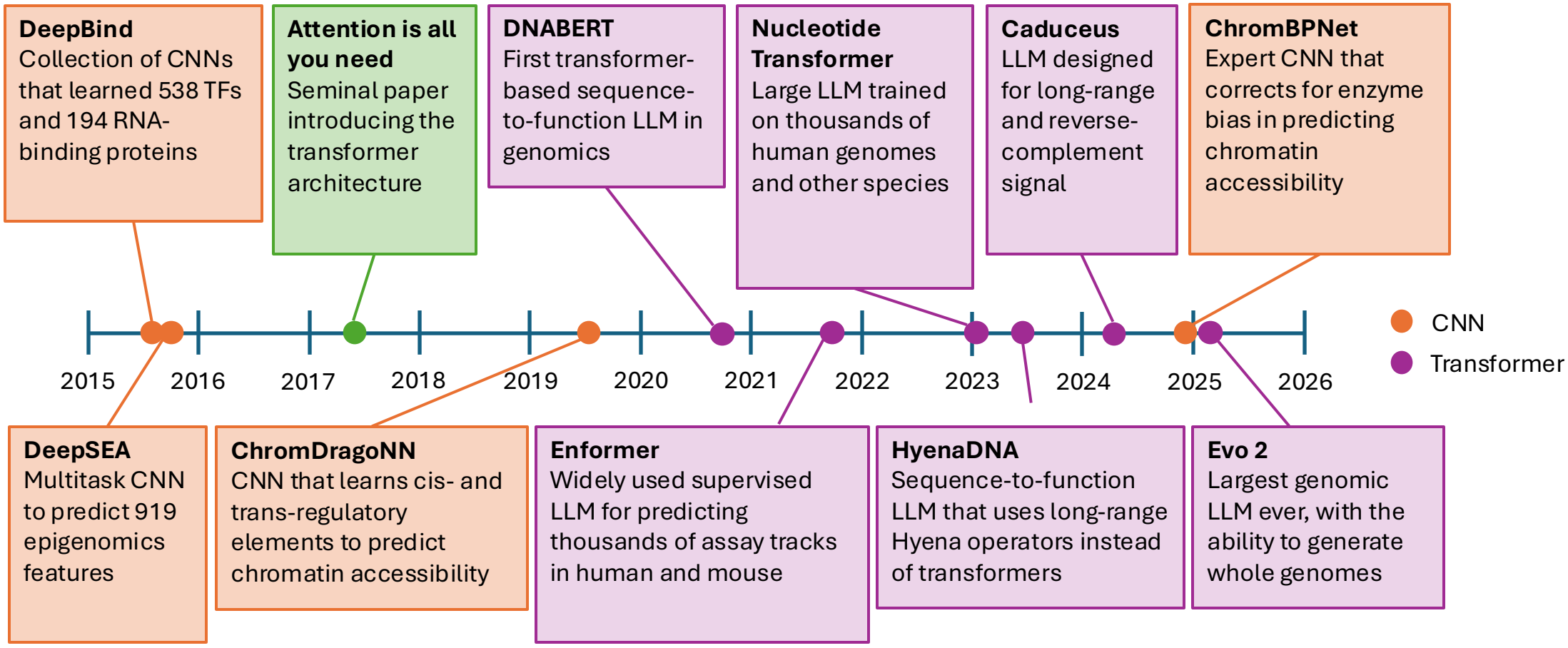
# Variant effect prediction

- What **effect** does **variant** X have on us? What **trait** does it create?
- Arguably the most important task in bioinformatics

- Historically studied using Genome-Wide Association Studies (**GWAS**) in connection with Quantitative Trait Loci (**QTL**) analysis
  - This approach has many drawbacks

- Could be studied using deep learning

# Deep learning is now widespread in genomics



# Deep learning basics

- Learn **patterns** from **some data** that can generalize to **all data**

- 3 key architectural pieces in genomics
  - Convolutions
  - Feed forward networks
  - Attention

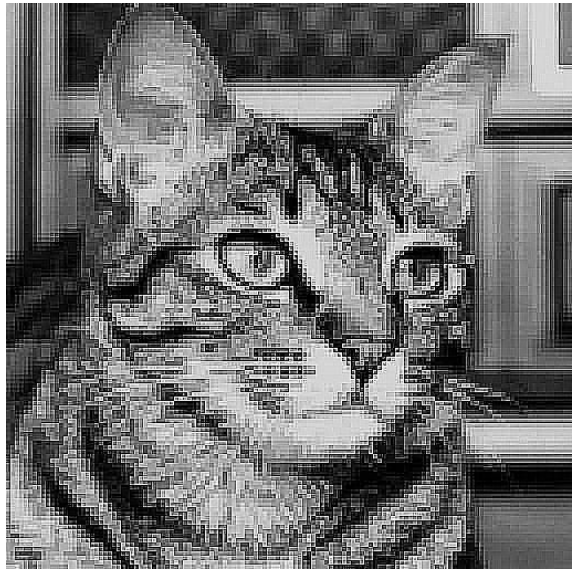
# CONVOLUTIONS

Blur

$$\begin{bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{bmatrix}$$

Vertical lines

$$\begin{bmatrix} .33 & 0 & -.33 \\ .33 & 0 & -.33 \\ .33 & 0 & -.33 \end{bmatrix}$$



$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Sharpen

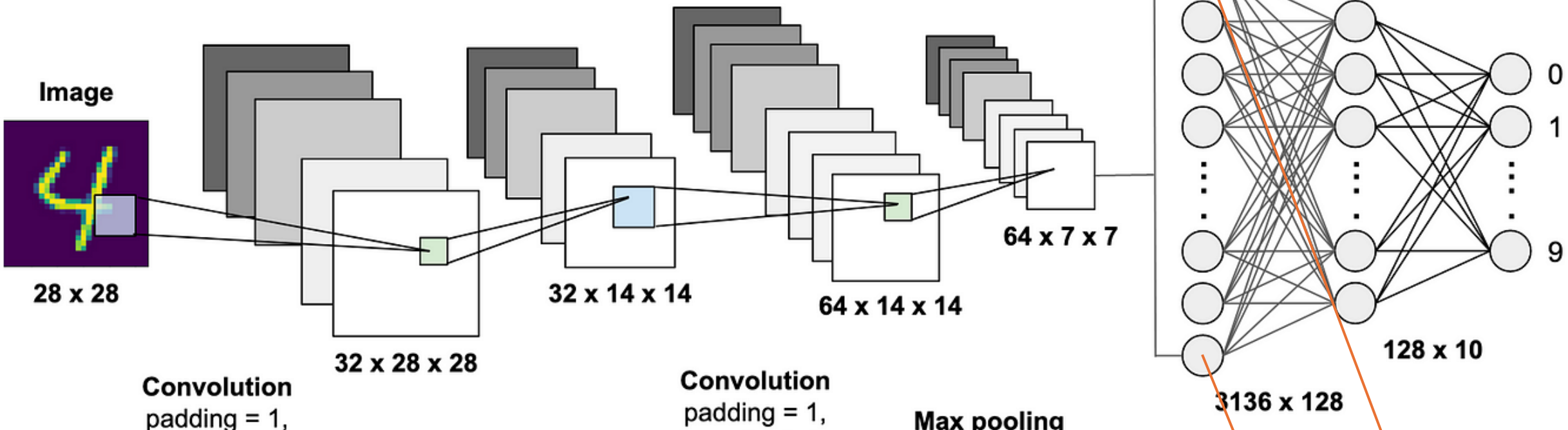
$$\begin{bmatrix} .33 & .33 & .33 \\ 0 & 0 & 0 \\ -.33 & -.33 & -.33 \end{bmatrix}$$

Horizontal lines

In each 3x3 region: Simple features e.g. vertical lines, horizontal lines, etc. (32 total)

In each 8x8 region: Complex features e.g. patterns, textures, combinations of shapes, etc. (64 total)

Learn complex patterns between features and labels

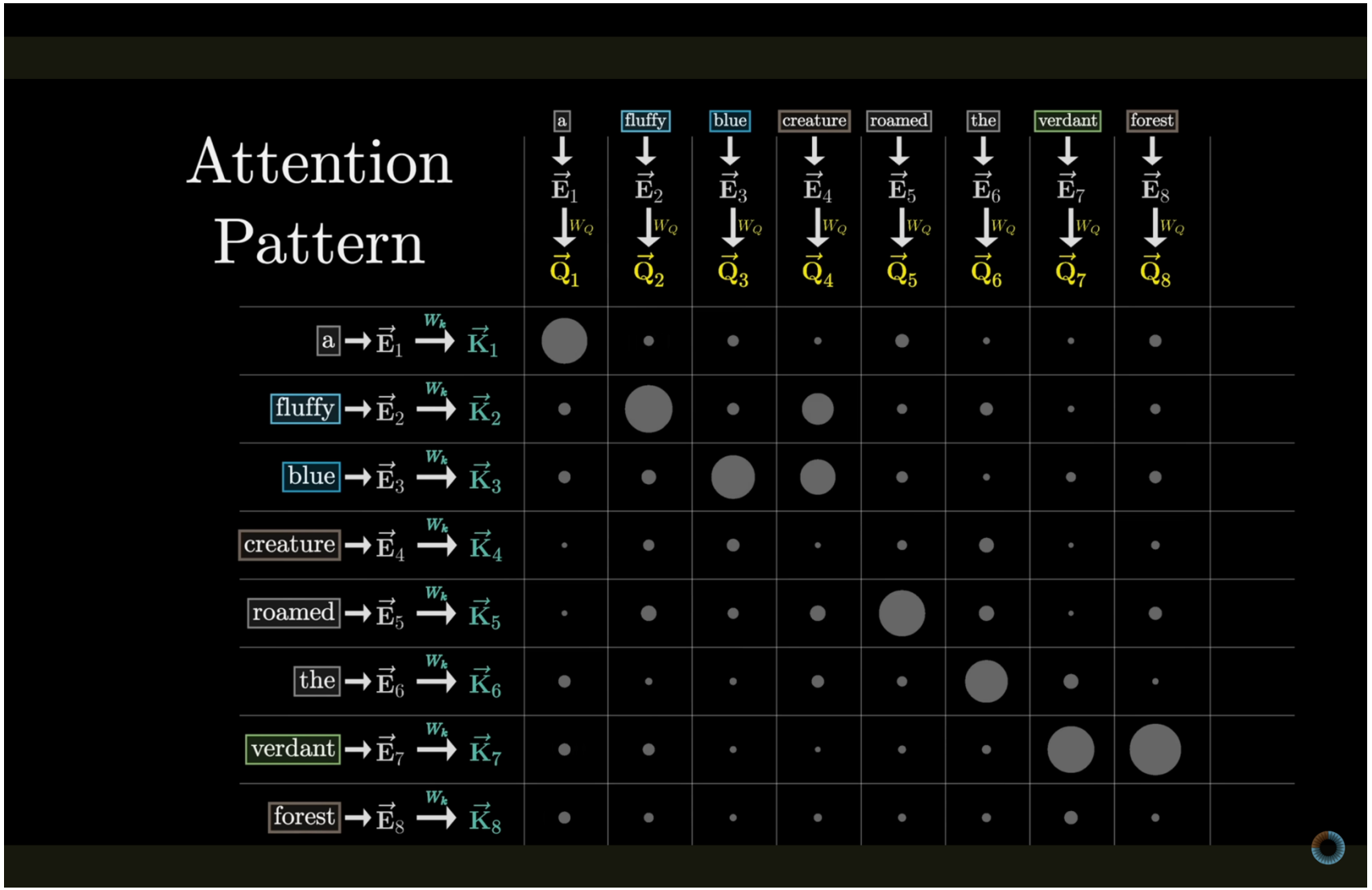



Summarize simple features in each 4x4 region

Summarize complex features in each 10x10 region

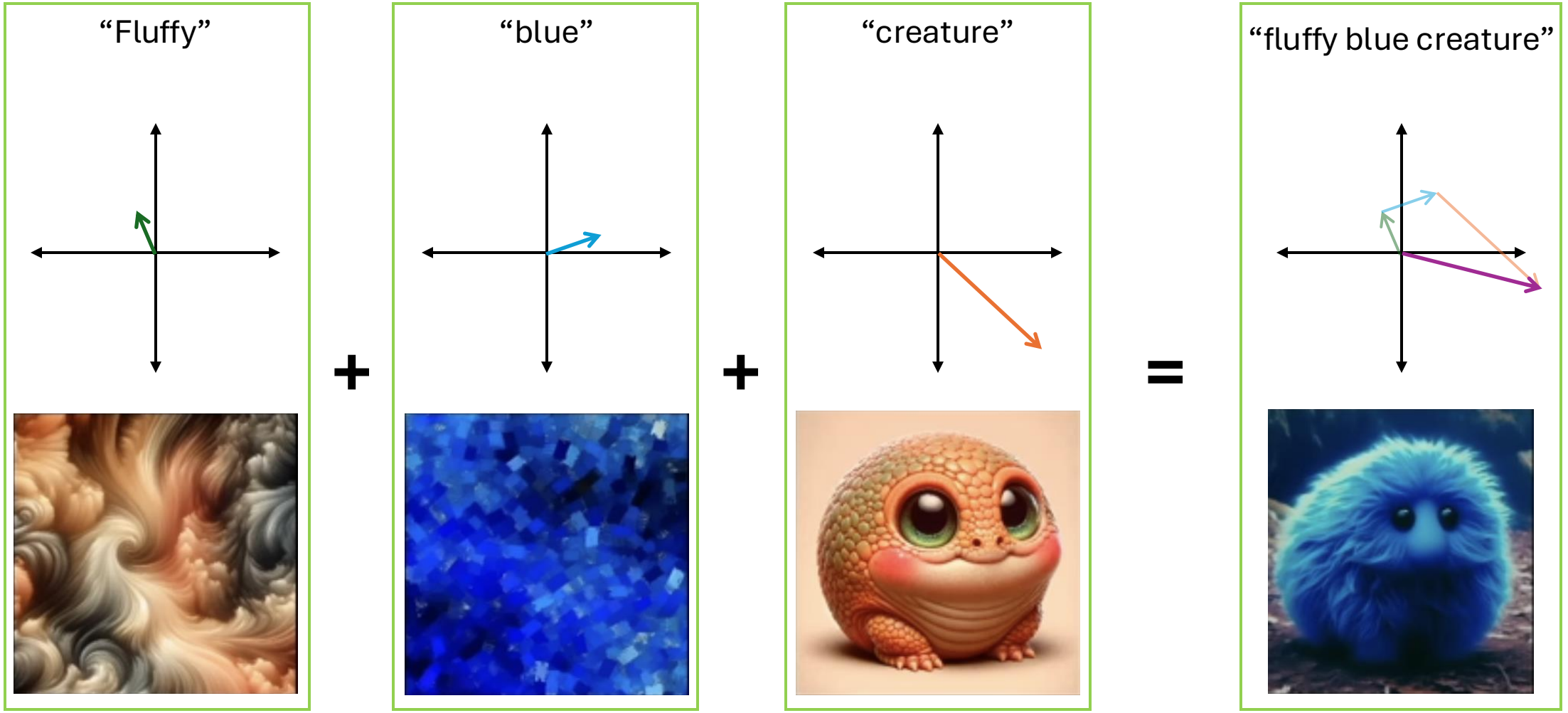
e.g. presence of dark-to-light gradient textures in upper left region

e.g. presence of trident shapes in bottom right region




**Attention** tells us which words **relate** to which other words

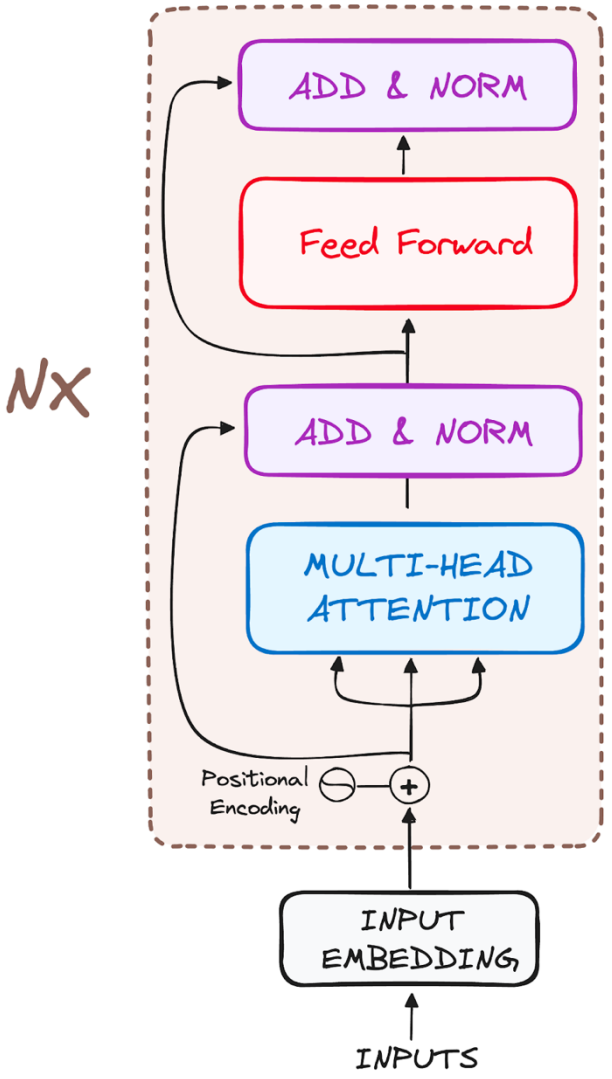
Attention **modifies** the **vector representation** of a **word** to include its **additional meaning**



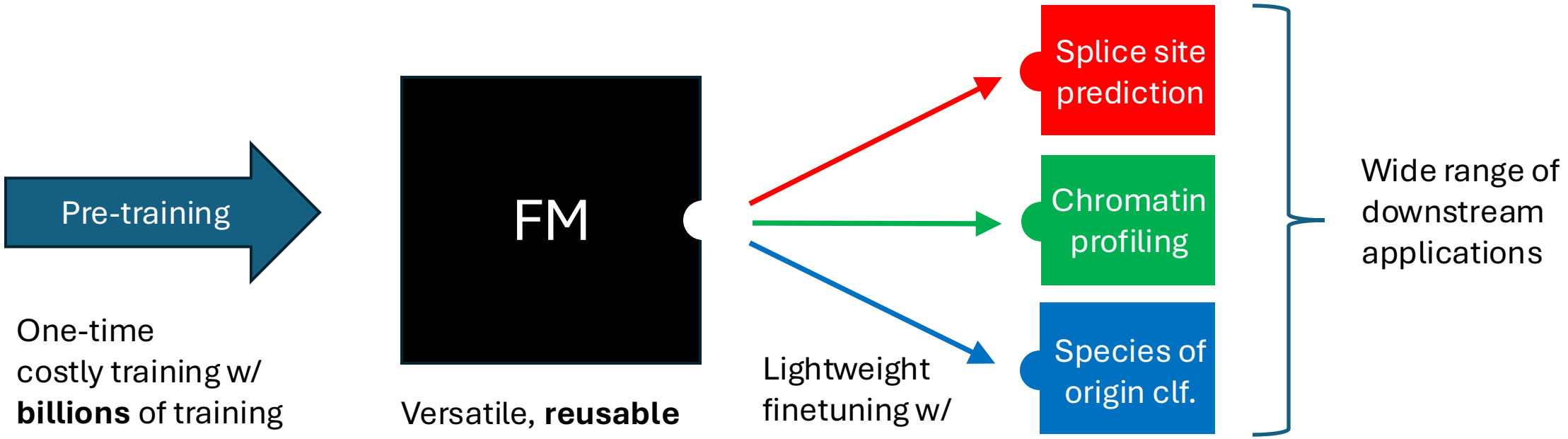
https://www.datacamp.com/tutorial/how-transformers-work

# Why use LLMs in genomics?

- LLMs employ the most advanced **architecture** ever made for AI
  - Core advancement is **Attention**
    - First truly effective way to process **sequential** data like **text**
  - Attention allows models to get really **big**, and bigger is better
- Operate at **lower cost** than humans
- Can conduct massively **parallel** experiments with **little human effort**
- Can model **rare** or **novel sequences**



# Foundation model (FM)




One-time costly training w/ **billions** of training samples

e.g. train on the **entire human genome**

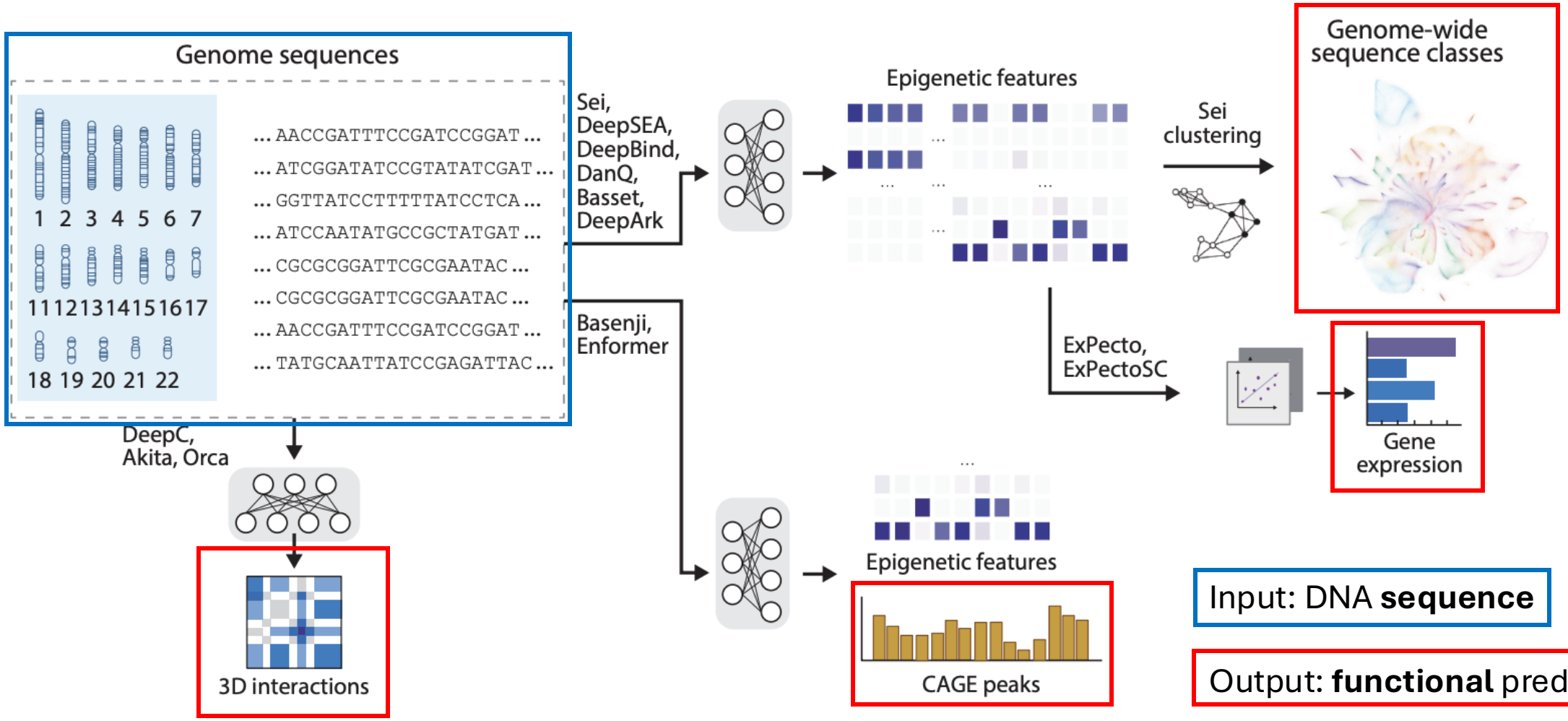
Versatile, **reusable** Foundation Model

Pre-train once, then fine-tune many times

Lightweight finetuning w/ only **thousands** of samples

e.g. train on a **small number** of **labelled** splice sites

# Sequence-to-function models



# Enformer

Article | Open access | Published: 04 October 2021

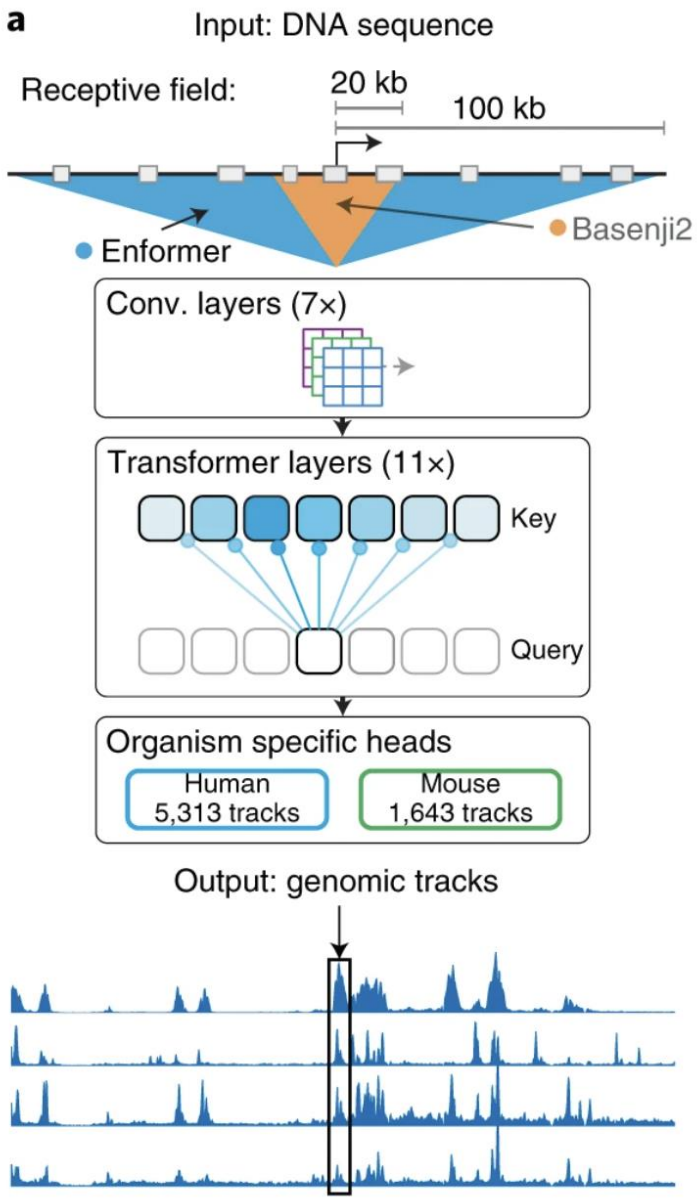
## Effective gene expression prediction from sequence by integrating long-range interactions

Žiga Avsec ✉, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli ✉ & David R. Kelley ✉

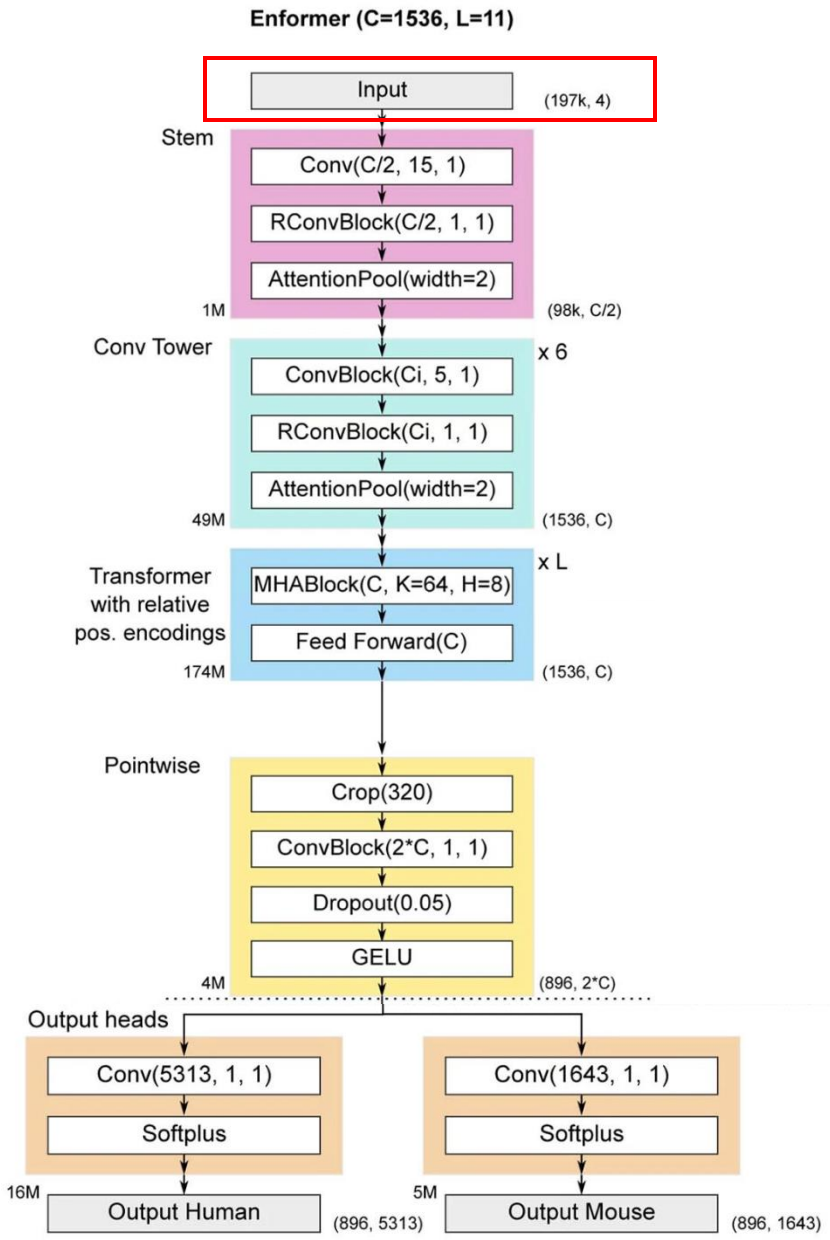
*Nature Methods* **18**, 1196–1203 (2021) | Cite this article

**306k** Accesses | **1001** Citations | **380** Altmetric | Metrics

- The most widely used genomic Large Language Model (LLM)
- Developed in the first wave of genomic LLMs
- Has since defined the field



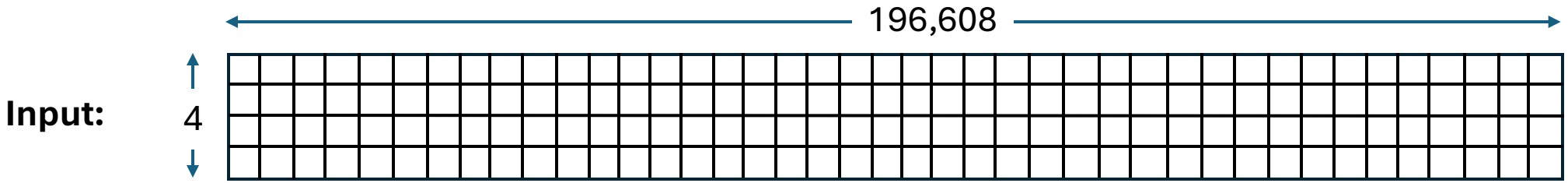

- Trained on **197kb** DNA sequences from the human and mouse **reference** genomes
- Each sequence has thousands of corresponding labels; **supervised** model
  - **ATAC-seq**
    - Is this sequence available to be transcribed?
  - **DNase-seq**
    - Same as ATAC-seq, but different measurement
  - **ChIP-seq**
    - Are proteins binding to this sequence?
  - **CAGE**
    - How much of this sequence is transcribed into RNA?
- Model predicts **all labels** at **each position**



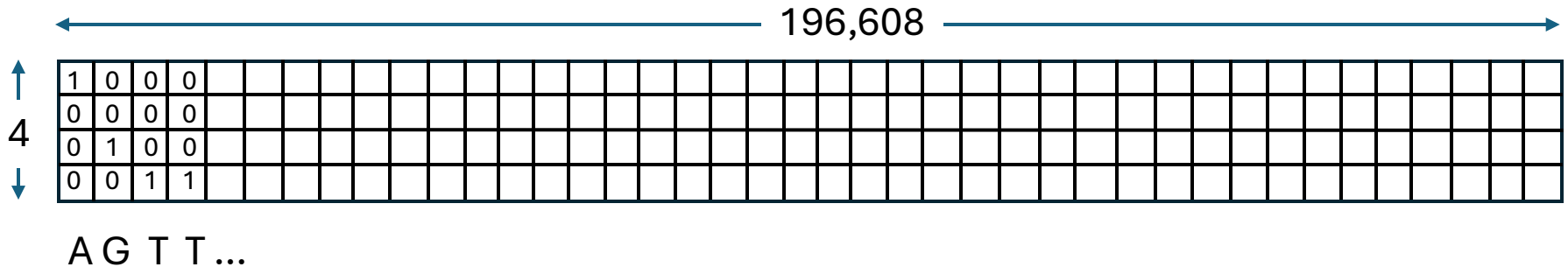

Input is 197k one-hot encoded DNA bases

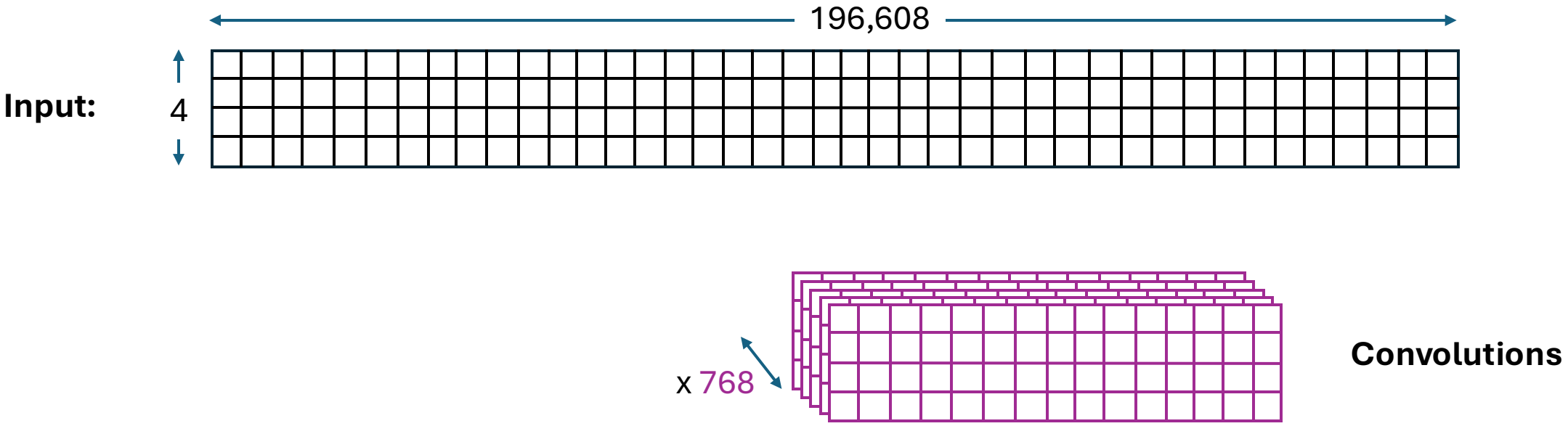
GATTACA

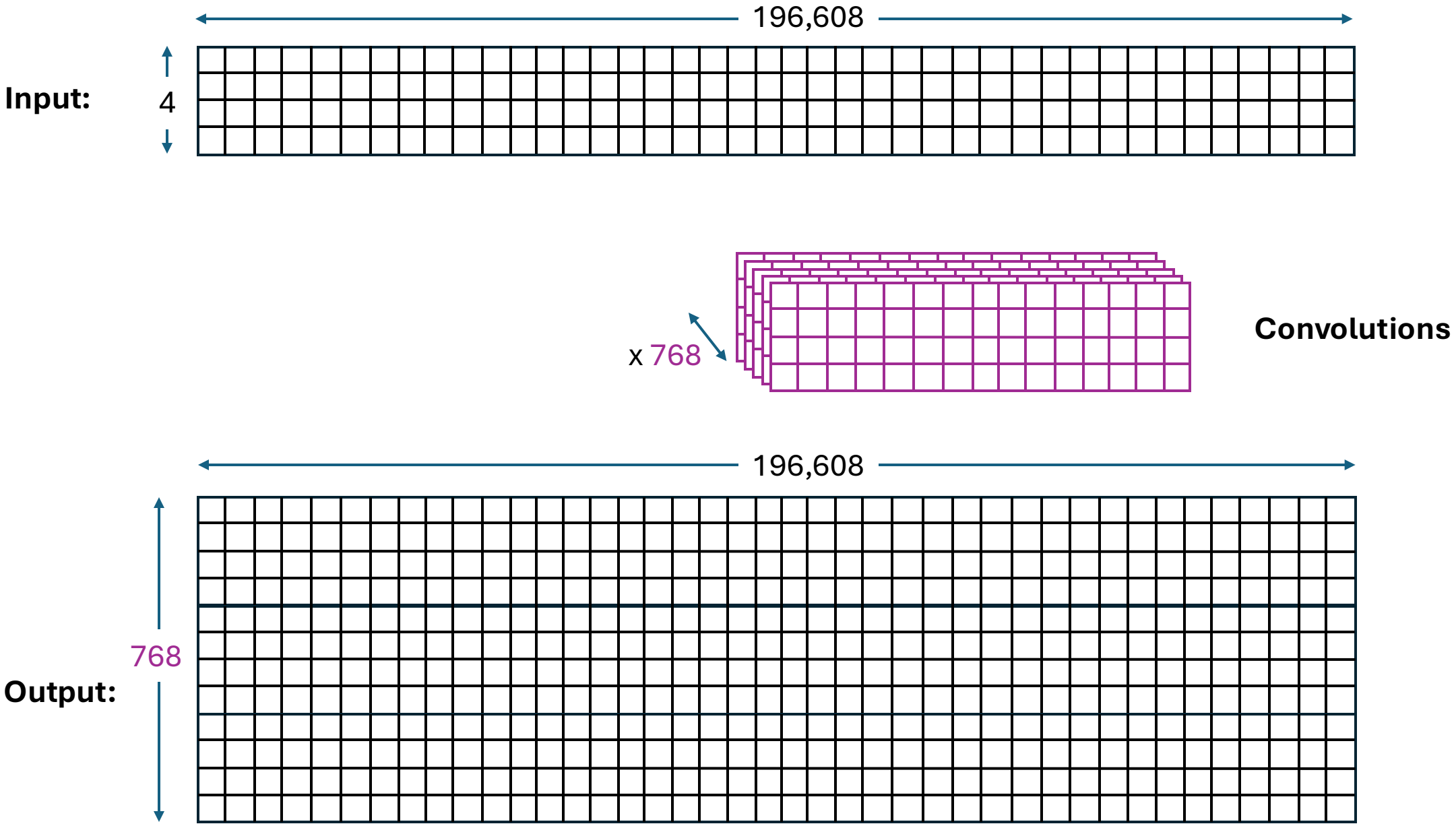
[0 0 1 0], [1 0 0 0], [0 0 0 1], [0 0 0 1], [1 0 0 0], [0 1 0 0], [1 0 0 0]

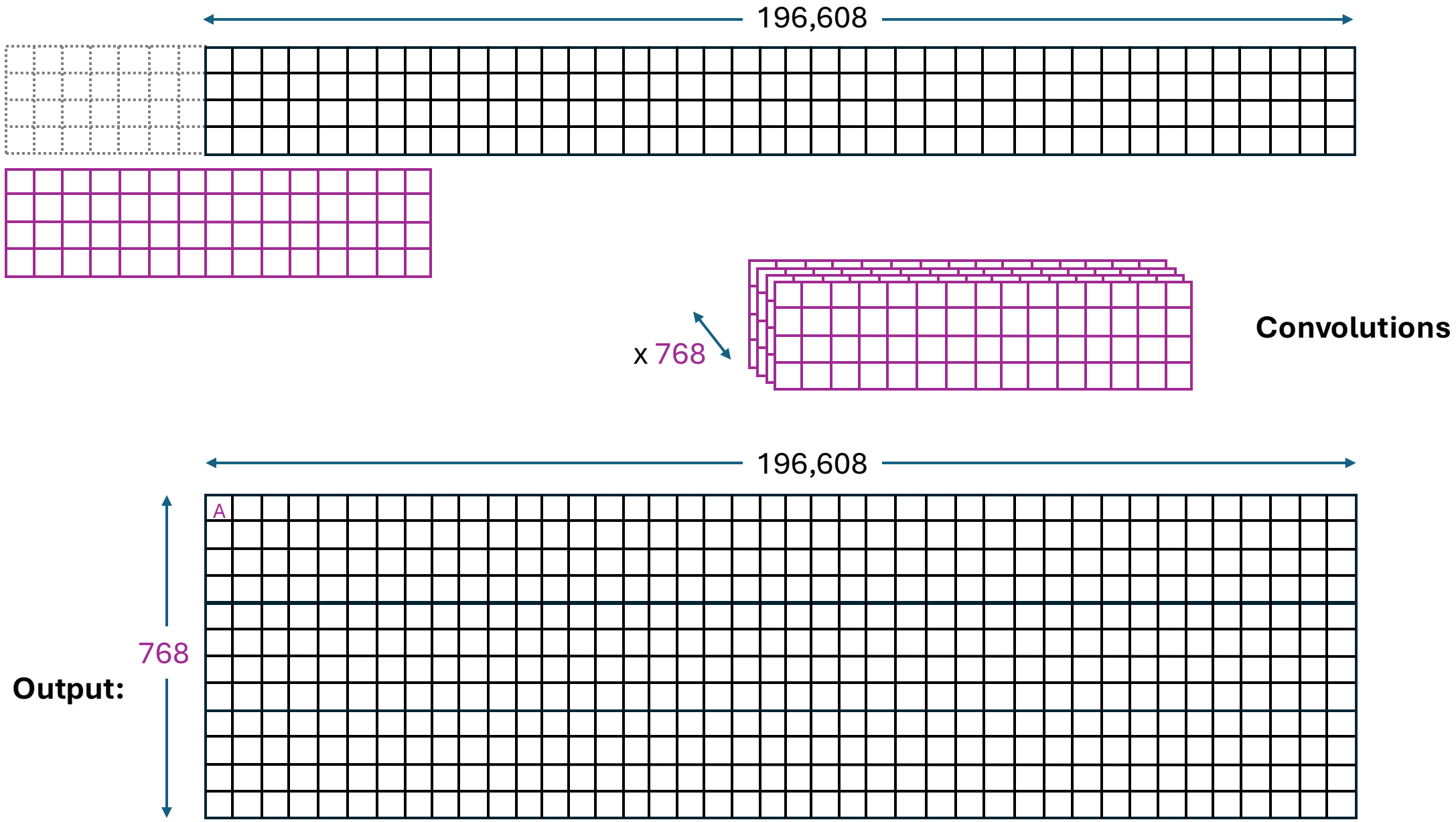

196,608
Input:
4

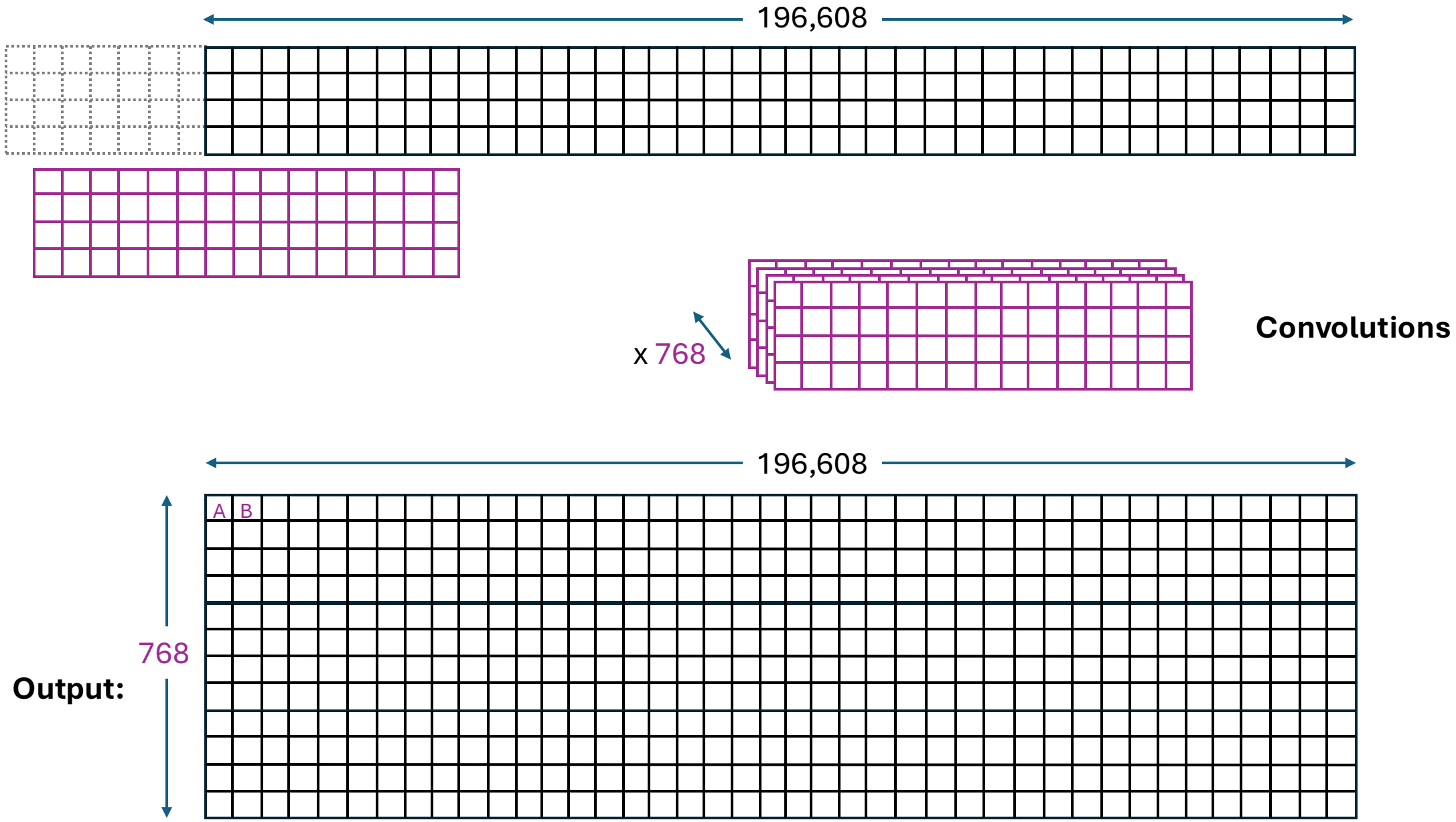
**Input:**

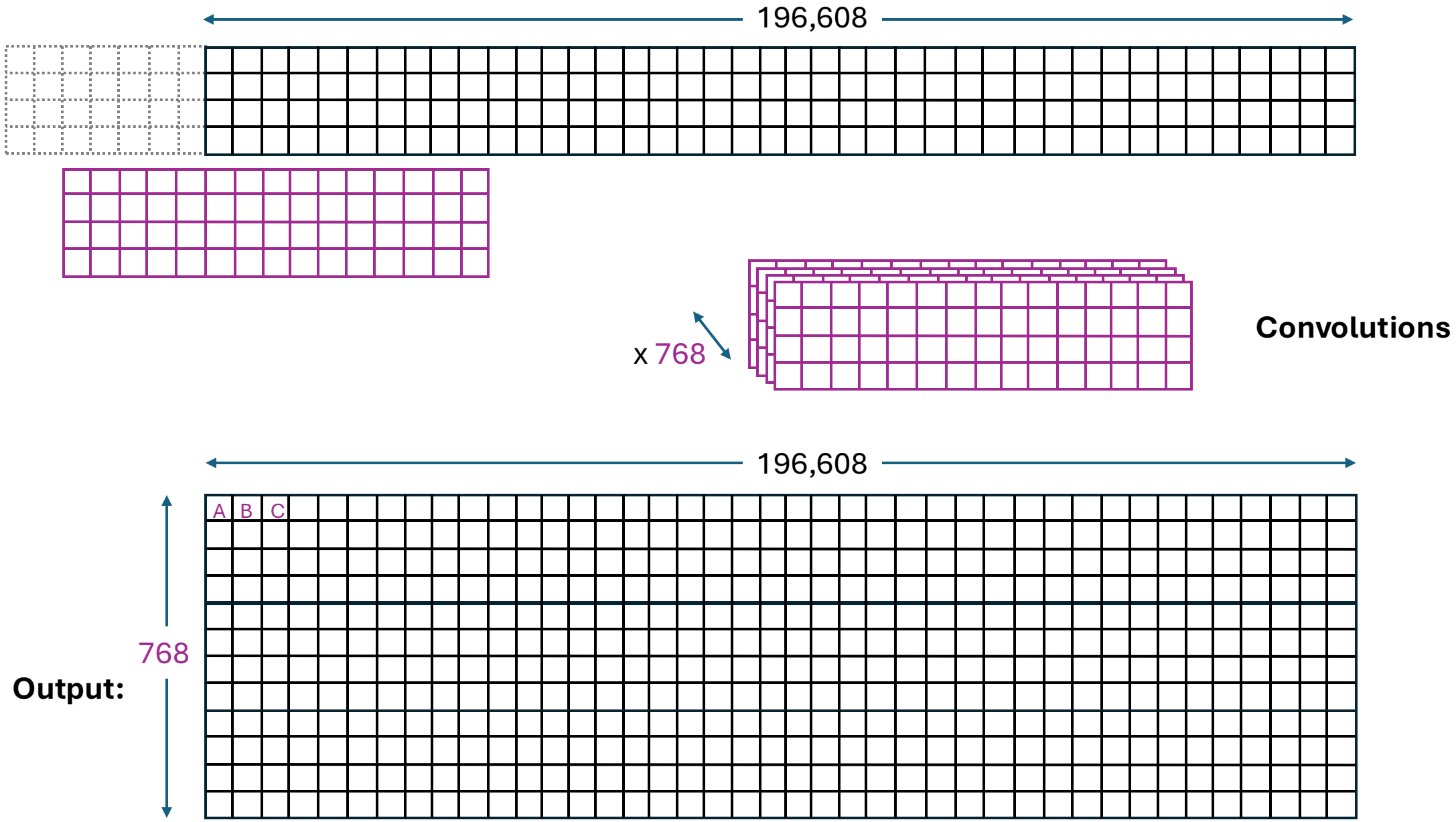


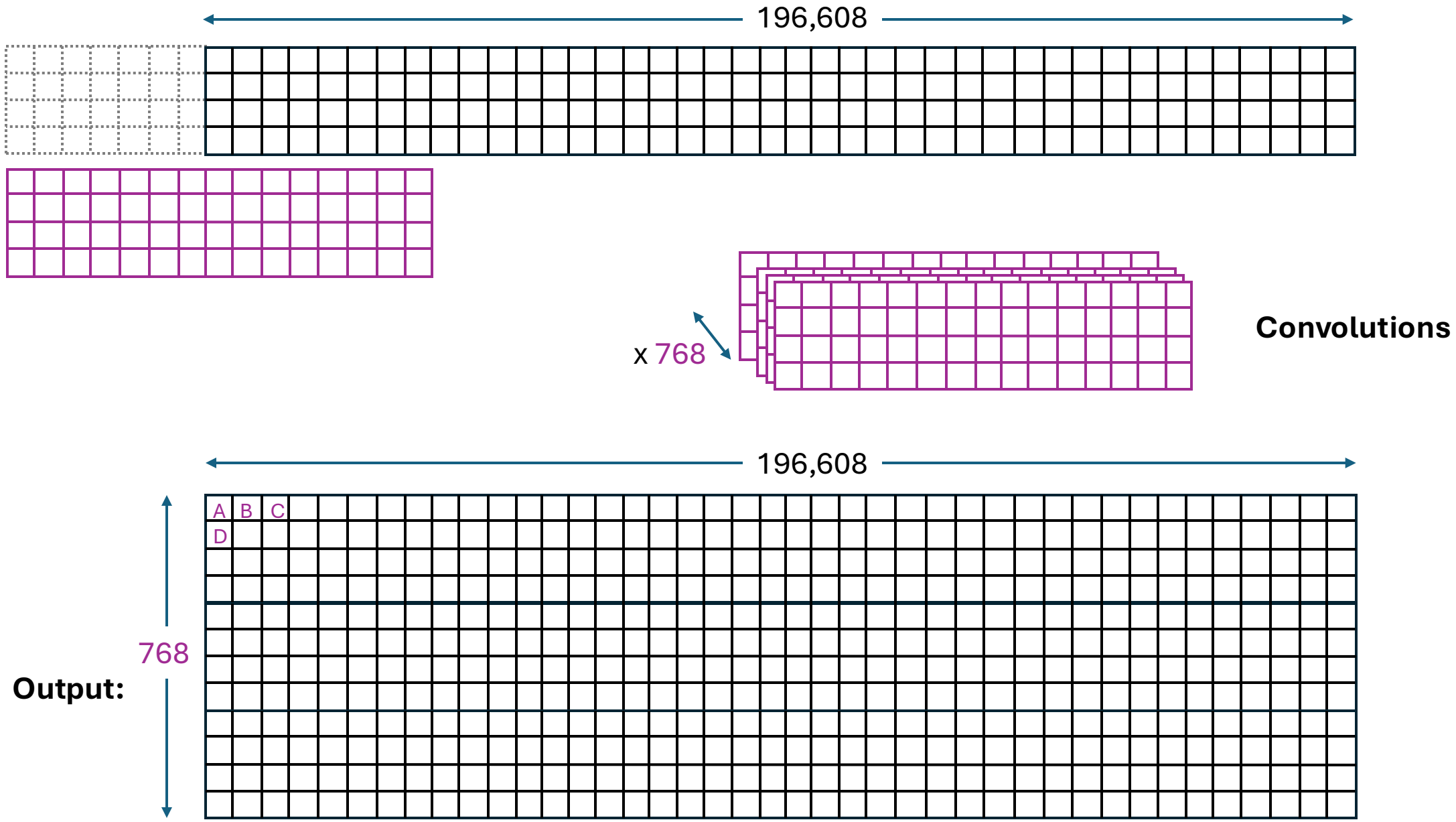

196,608
Input:
4
x 768
Convolutions

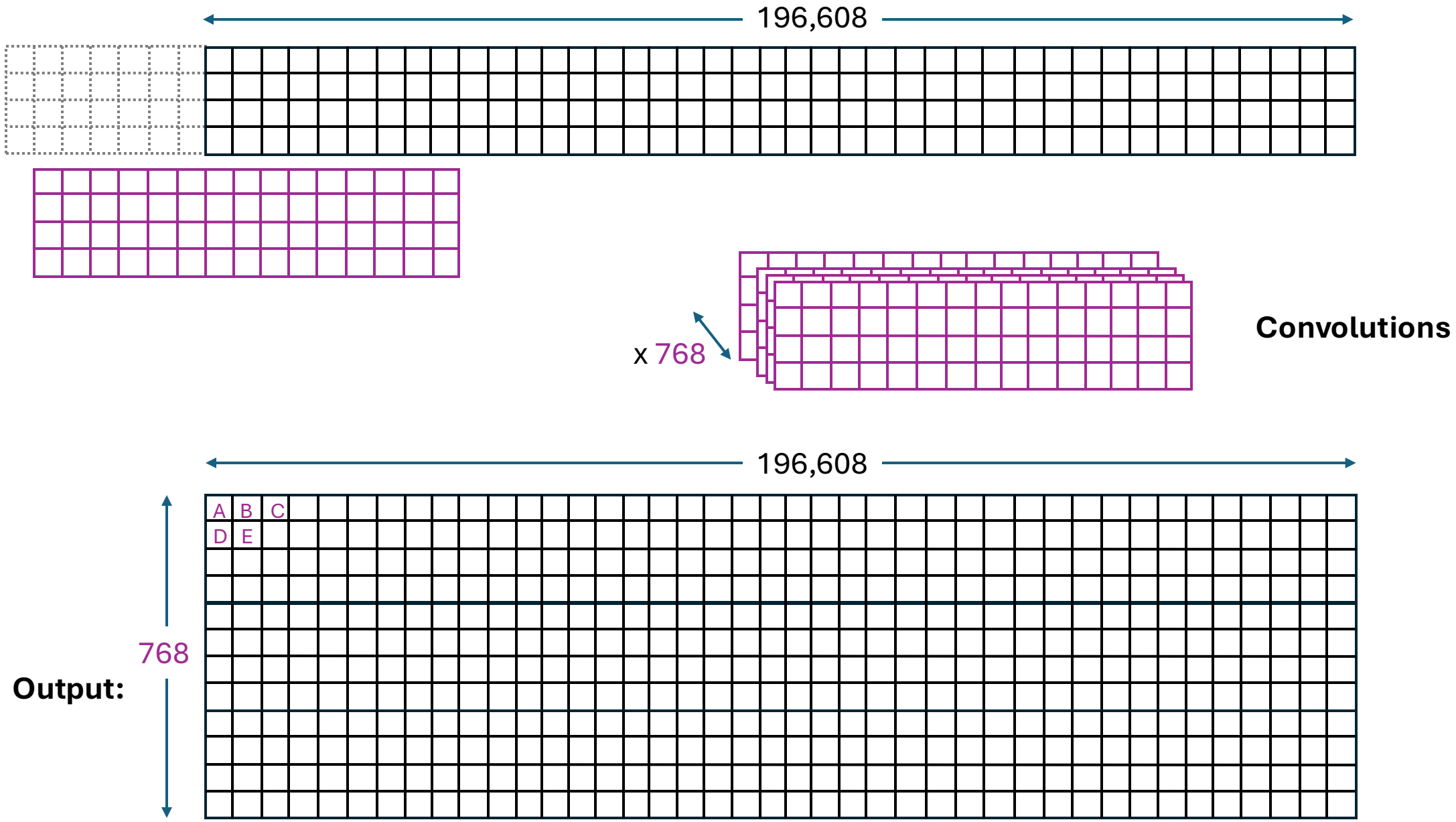

196,608
Input:
4
x 768
Convolutions
196,608
768
Output:

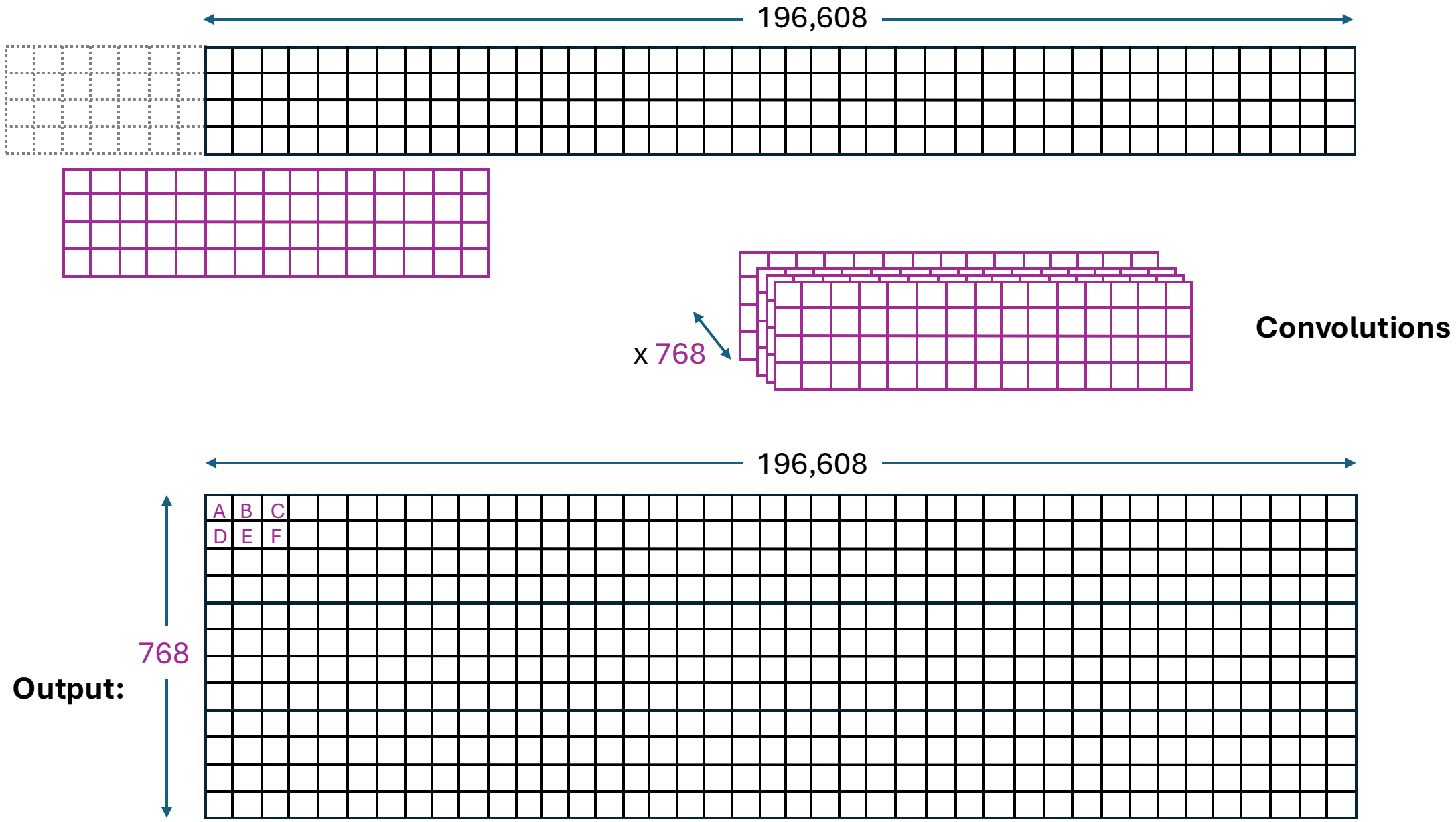

196,608
x 768
Convolutions
196,608
A
768
Output:

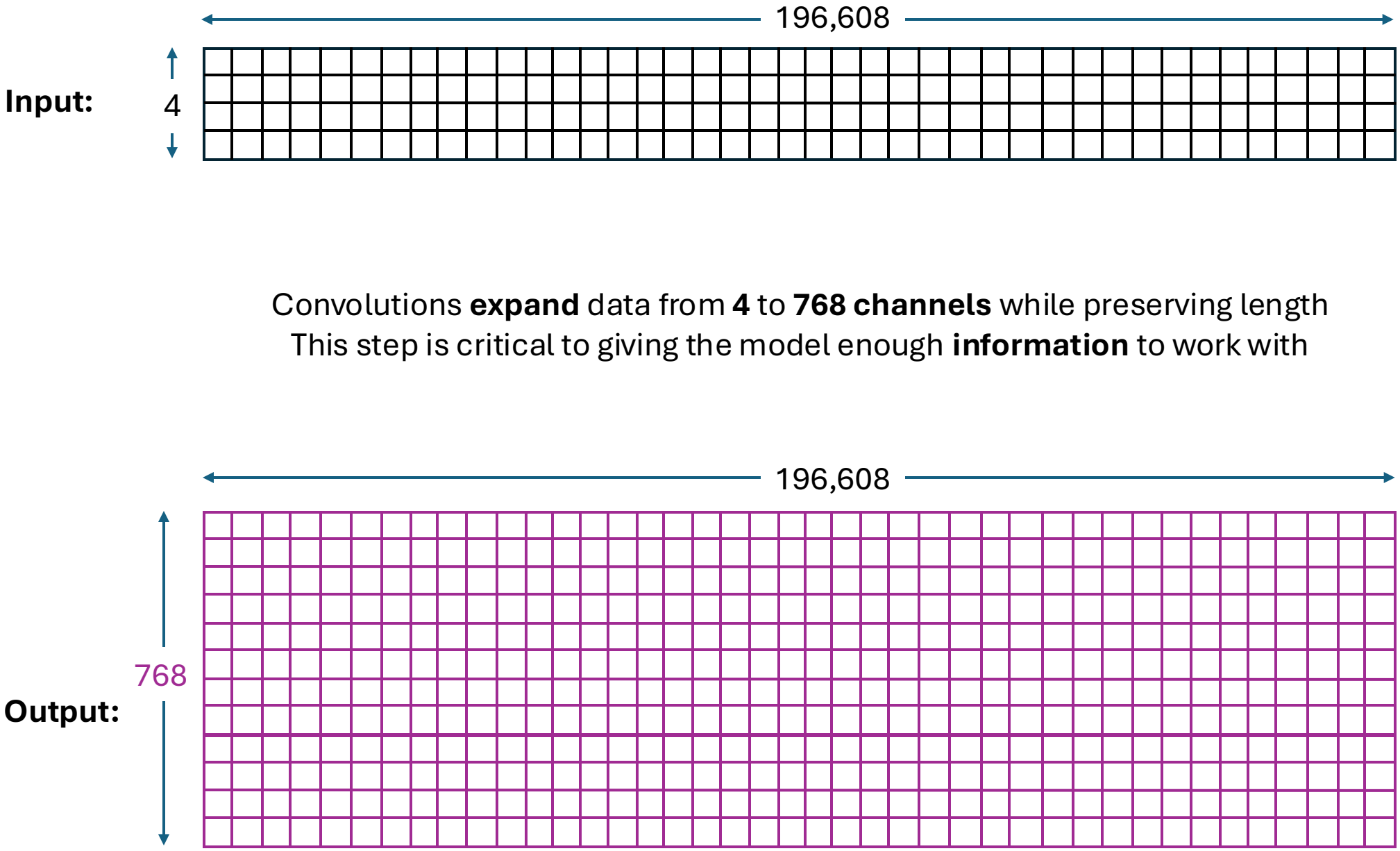

196,608
x 768
Convolutions
196,608
A
B
768
Output:


196,608
x 768
Convolutions
196,608
A
B
C
768
Output:

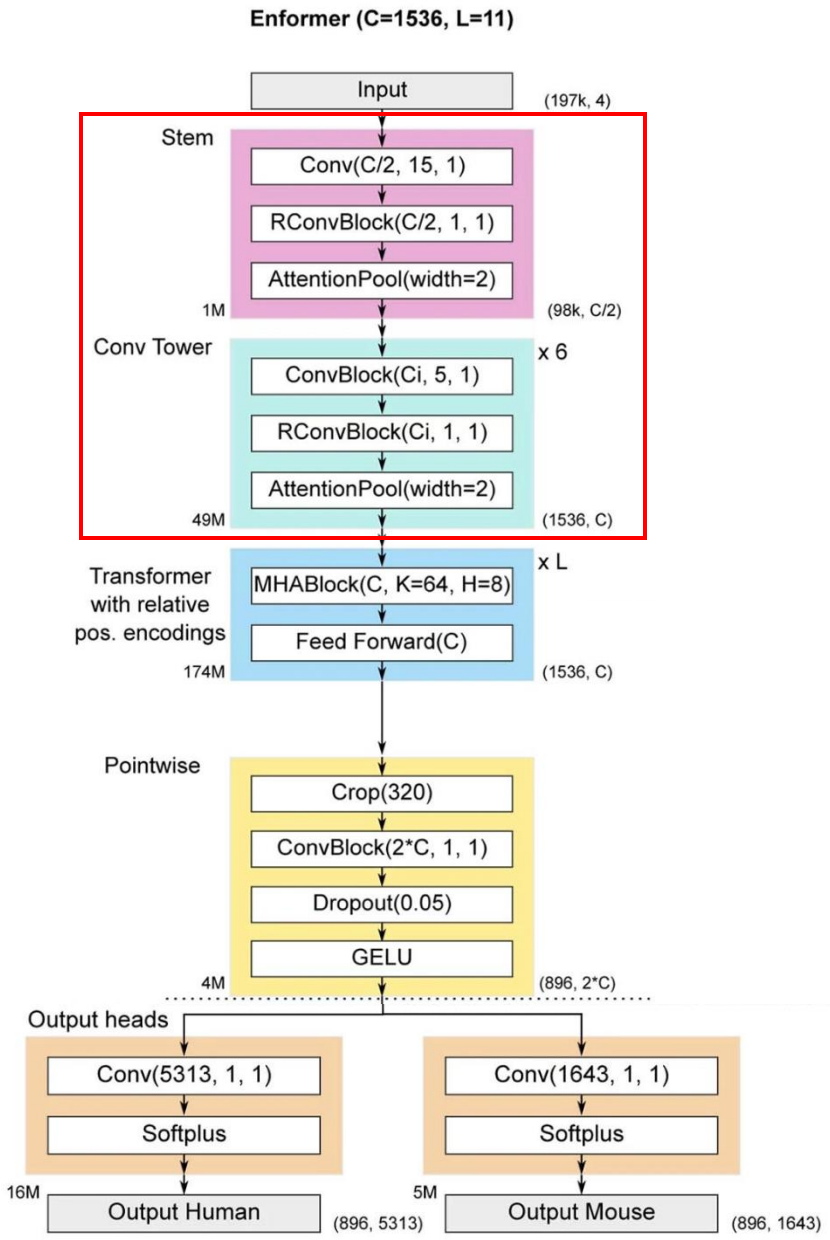

196,608
x 768
Convolutions
196,608
A
B
C
D
768
Output:


196,608
x 768
Convolutions
196,608
A B C
D E
768
Output:

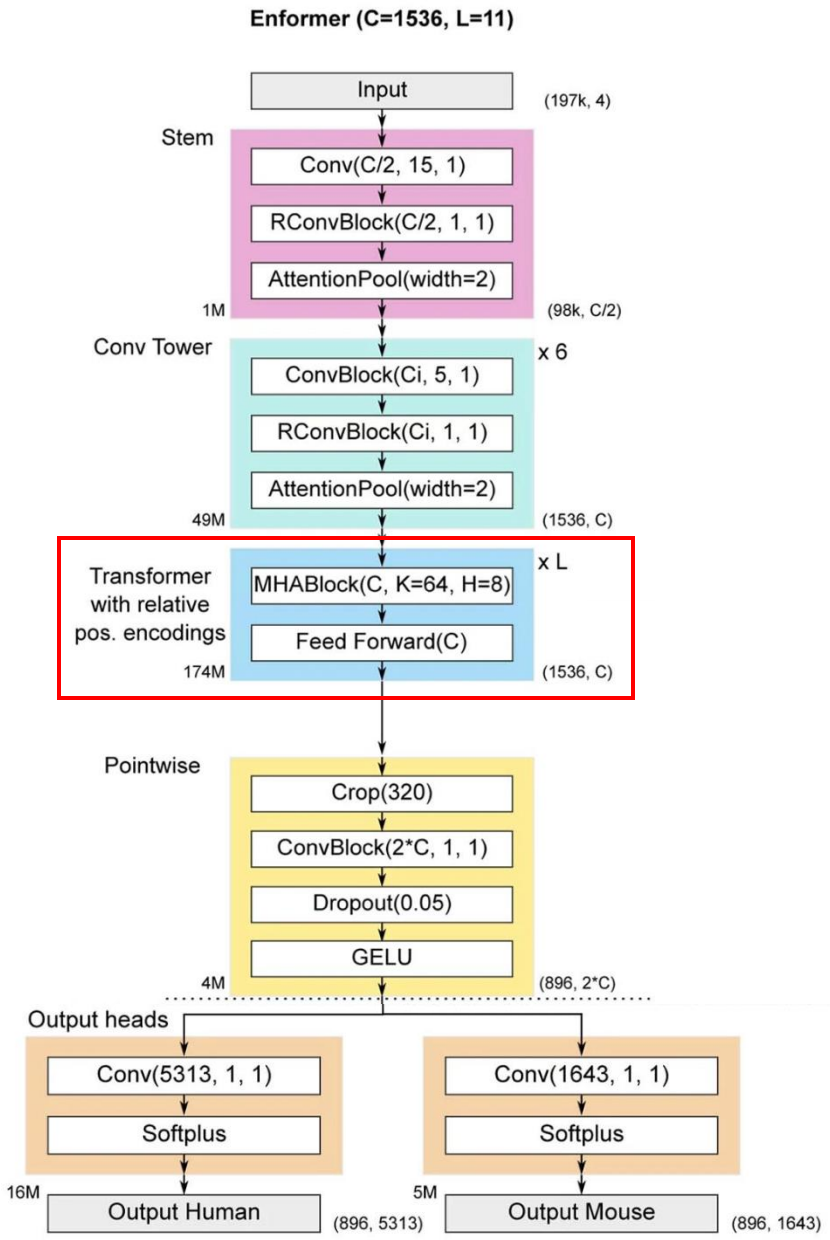

196,608
x 768
Convolutions
196,608
A B C
D E F
768
Output:



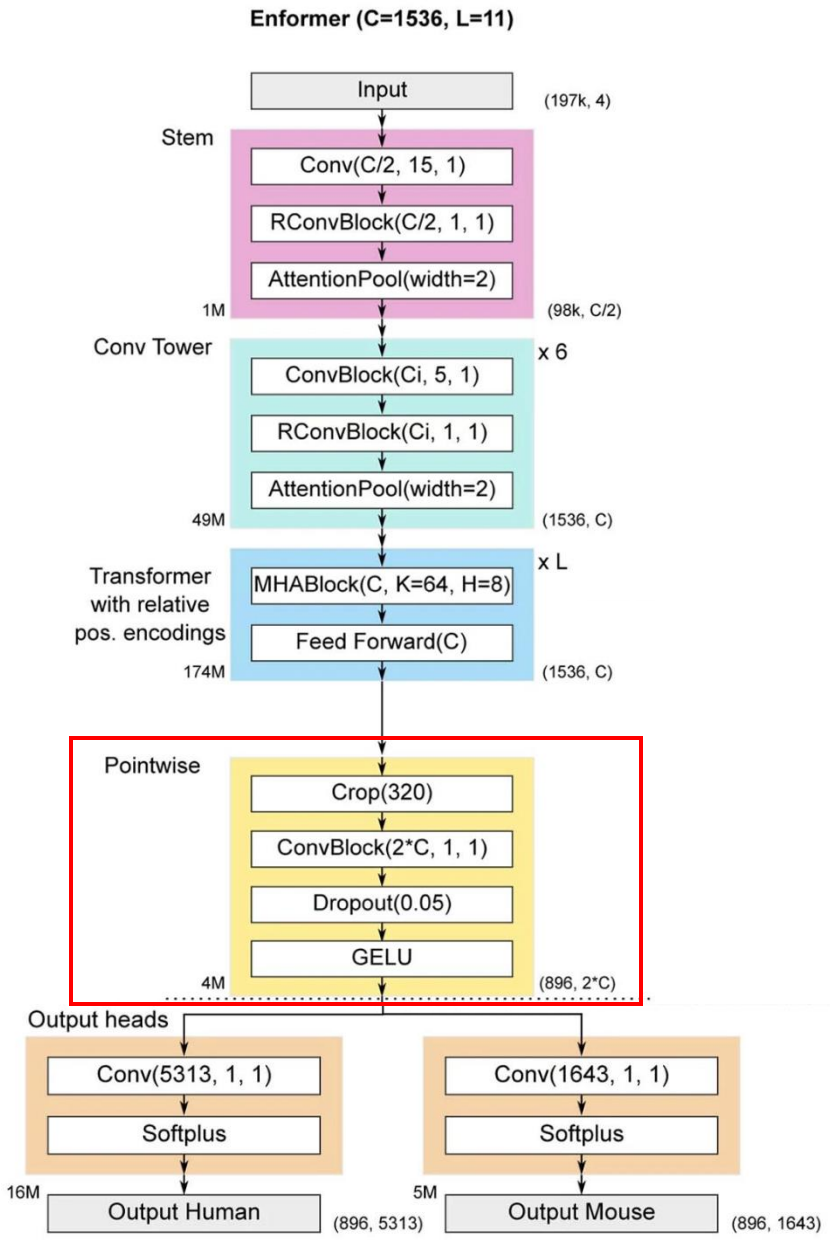

Convolutions **expand** data from **4** to **768 channels** while preserving length
This step is critical to giving the model enough **information** to work with



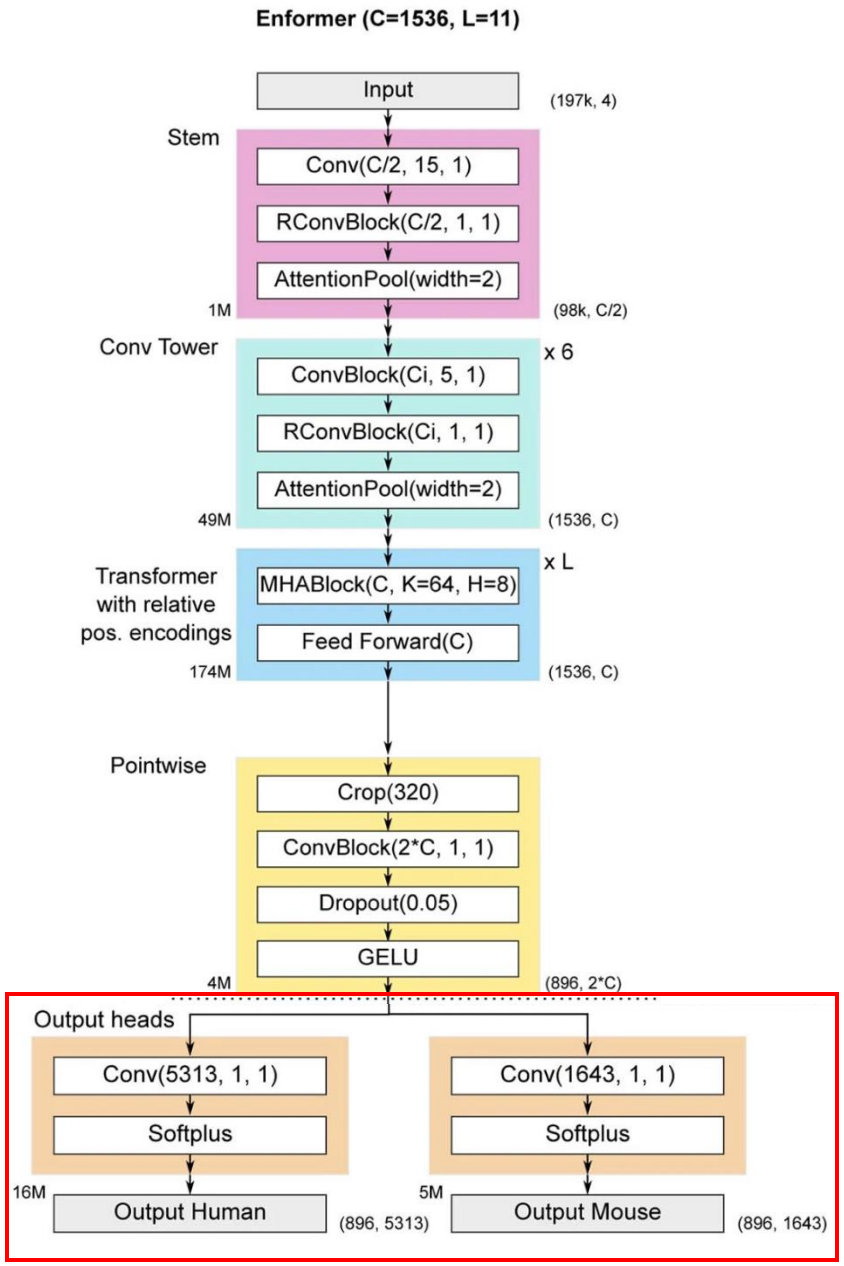

**Increase** the **dimensionality** of the data (4 -> 1536 channels)
**Decrease** the **number** of data points (197k ->1536)




Add **context** to data, like combining word vectors into sentence vectors
The power of transformers




Fully connected network
Learn **patterns** in data

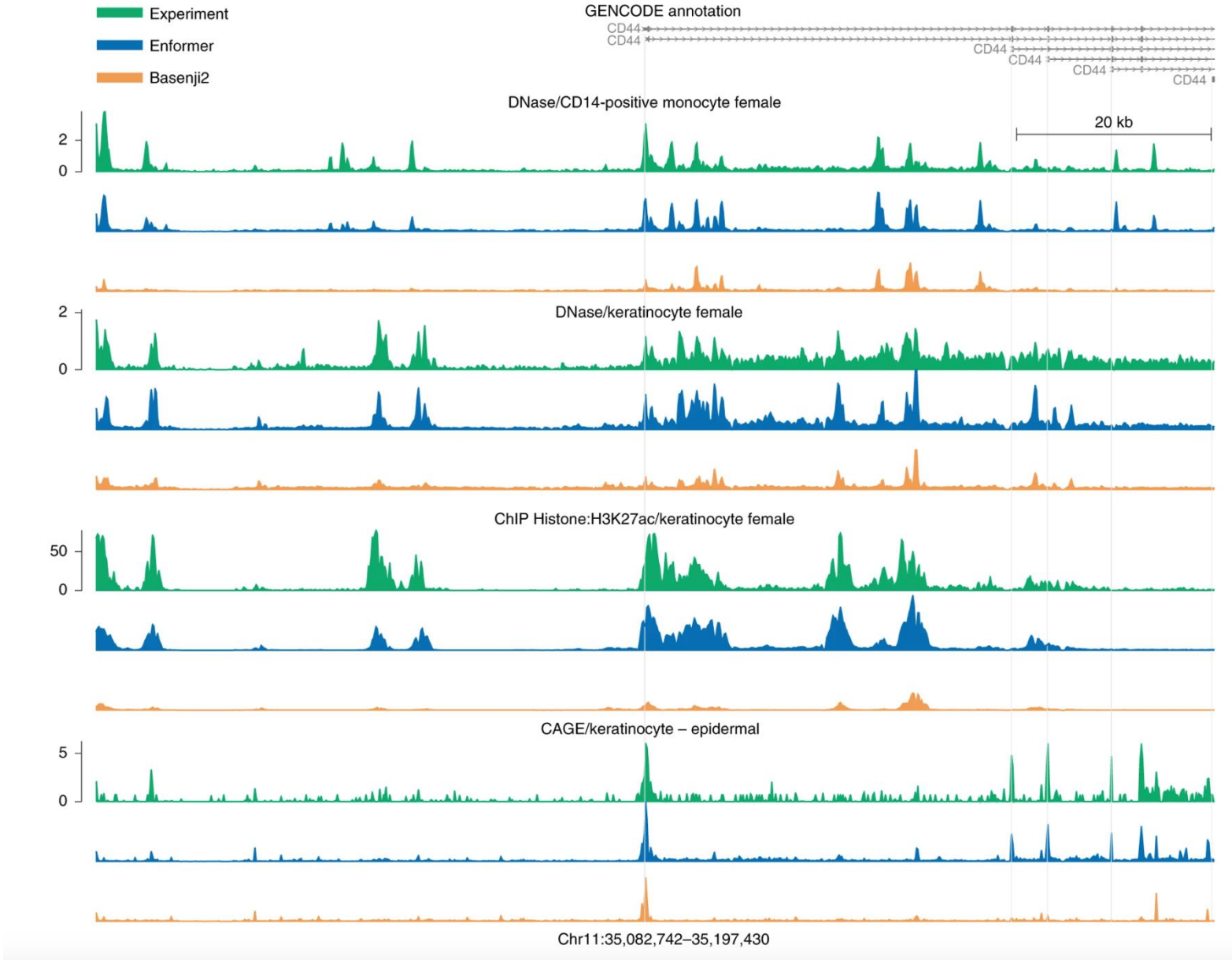



Emit **predictions** across all labels

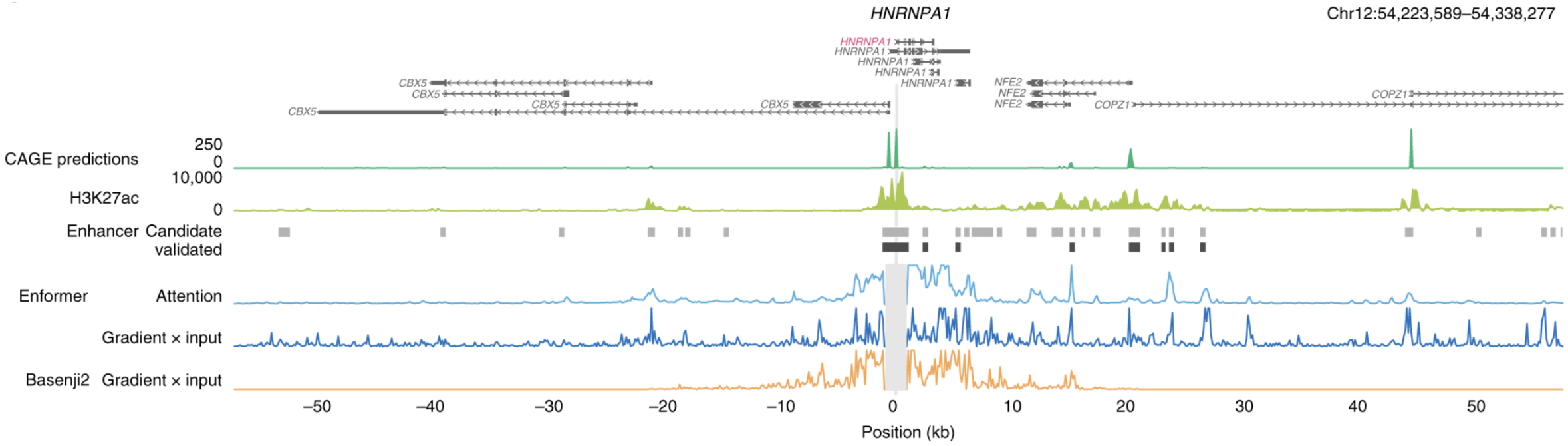
# What does Enformer do?

- Green: test sample **experimental** data
- Blue: Enformer's **predictions**

- Training enables it to **generalize** to unseen samples

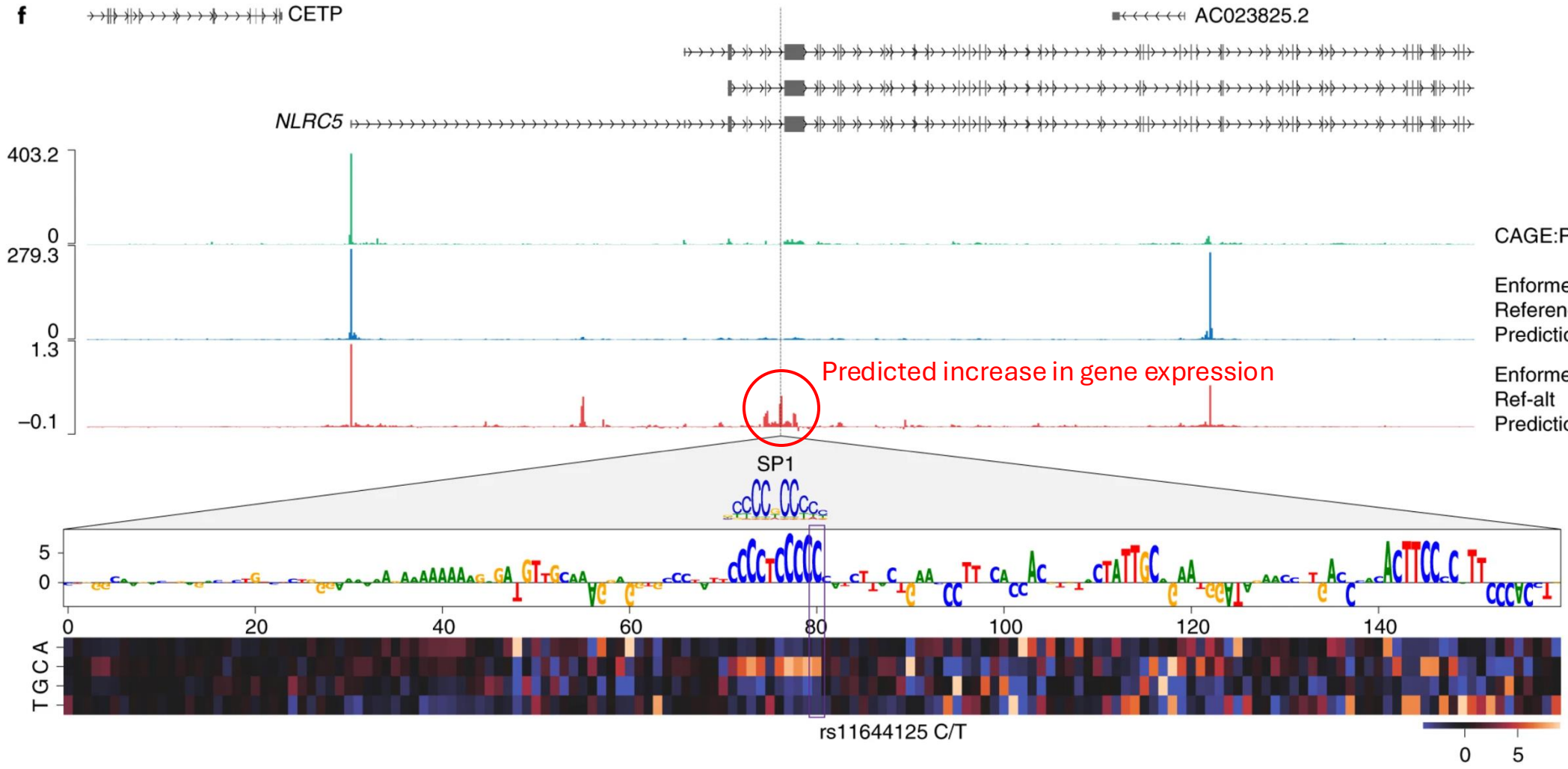


# Interpreting Enformer's predictions




- Dark green: Enformer's CAGE **prediction** of transcript abundance for *HNRNPA1*
- Light green and gray: **known** enhancer positions for *HNRNPA1*
- Blues: **attention** and **gradient** values at each position of the input
- Locations of high attention and gradients **correspond** with known enhancers; looking in the **right places**

# Understanding variants with Enformer




Ground truth

Prediction w/ **C** base

Prediction w/ **T** base

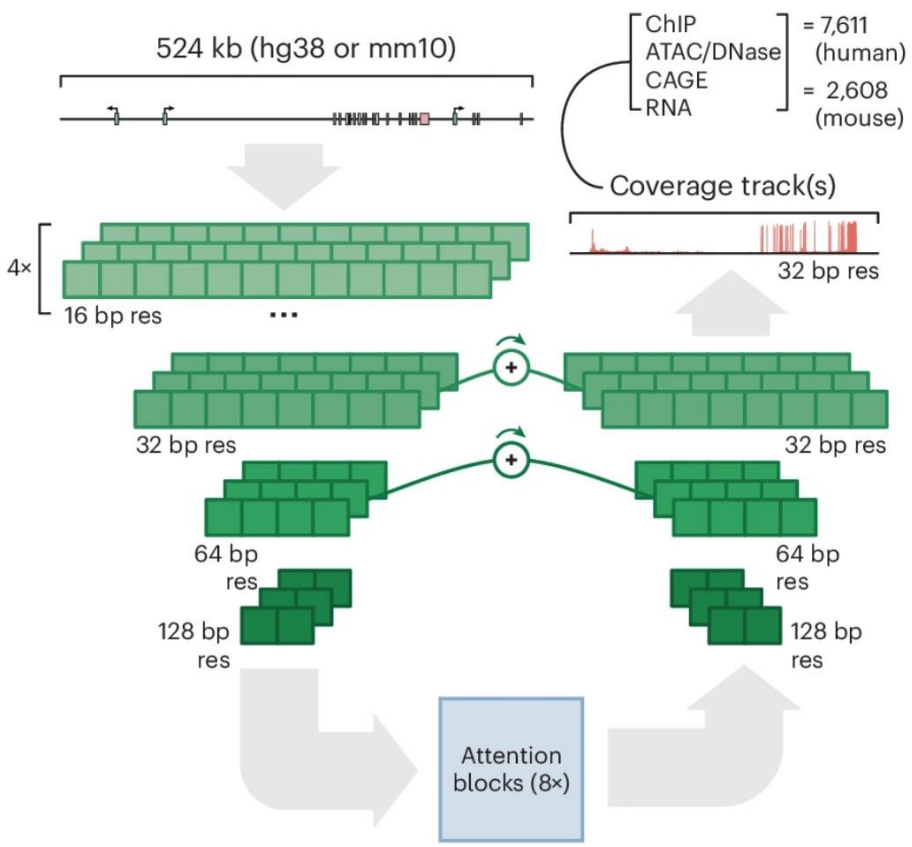
Mutation **disrupts known** SP1 transcription factor binding **motif**

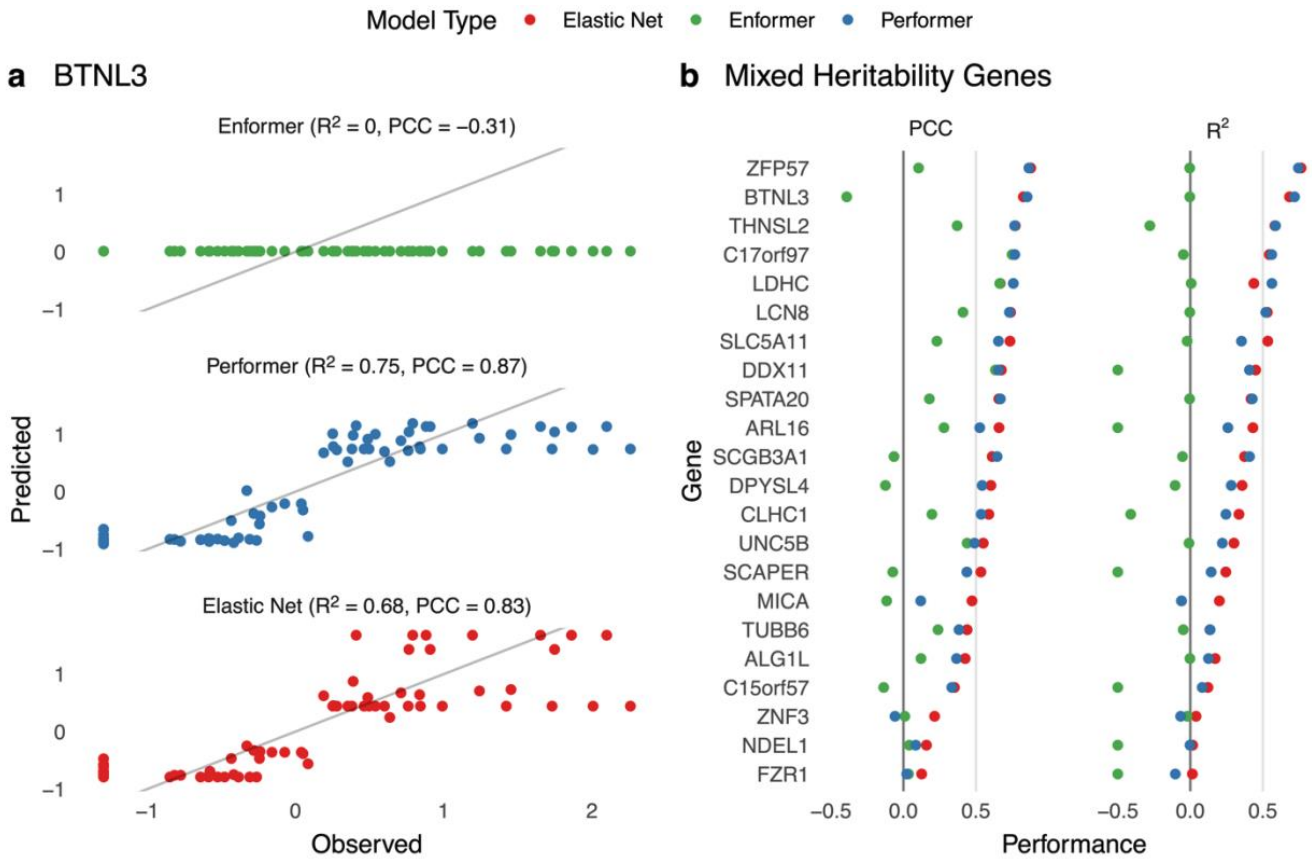
Linder et al. 2025.
Drusinsky et al. 2024.

# Enformer's influence

Several models fine-tuning Enformer for variant effect prediction, such as Performer

a Enformer 2.0: Borzoi



# Evo2

- The largest genomic LLM created to date
- Published February 2025
- Very divisive in the field

## Genome modeling and design across all domains of life with Evo 2

Garyk Brixi[*,1,2,3], Matthew G. Durrant[*,1,2], Jerome Ku[*,1,2], Michael Poli[*,2,3,5], Greg Brockman[**,2,6,§], Daniel Chang[**,1,2,3], Gabriel A. Gonzalez[**,1,2], Samuel H. King[**,1,2,3], David B. Li[**,1,2,3], Aditi T. Merchant[**,1,2,3], Mohsen Naghipourfar[**,1,2,7], Eric Nguyen[**,2,3], Chiara Ricci-Tam[**,1,2], David W. Romero[**,2,4], Gwanggyu Sun[**,1,2], Ali Taghibakshi[**,2,4], Anton Vorontsov[**,2,4], Brandon Yang[**,2,6], Myra Deng[8], Liv Gorton[8], Nam Nguyen[8], Nicholas K. Wang[8], Etowah Adams[9], Stephen A. Baccus[3], Steven Dillmann[3], Stefano Ermon[3], Daniel Guo[1,3], Rajesh Ilango[1], Ken Janik[4], Amy X. Lu[7], Reshma Mehta[6], Mohammad R.K. Mofrad[7], Madelena Y. Ng[3], Jaspreet Pannu[3], Christopher Ré[3], Jonathan C. Schmok[1], John St. John[4], Jeremy Sullivan[1], Kevin Zhu[7], Greg Zynda[4], Daniel Balsam[8,10], Patrick Collison[1,10], Anthony B. Costa[4,10], Tina Hernandez-Boussard[3,10], Eric Ho[8,10], Ming-Yu Liu[4,10], Thomas McGrath[8,10], Kimberly Powell[4,10], Dave P. Burke[‡,1,2,10], Hani Goodarzi[‡,1,2,10,11], Patrick D. Hsu[‡,†,1,2,7,10], Brian L. Hie[‡,†,1,2,3,10]

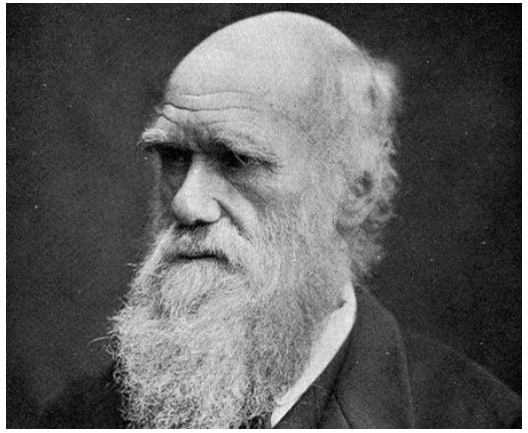
[1]Arc Institute; [2]Core Contributor, Evo 2 Team; [3]Stanford University; [4]NVIDIA; [5]Liquid AI; [6]Independent Researcher; [7]University of California, Berkeley; [8]Goodfire; [9]Columbia University; [10]Senior Contributor, Evo 2 Team; [11]University of California, San Francisco

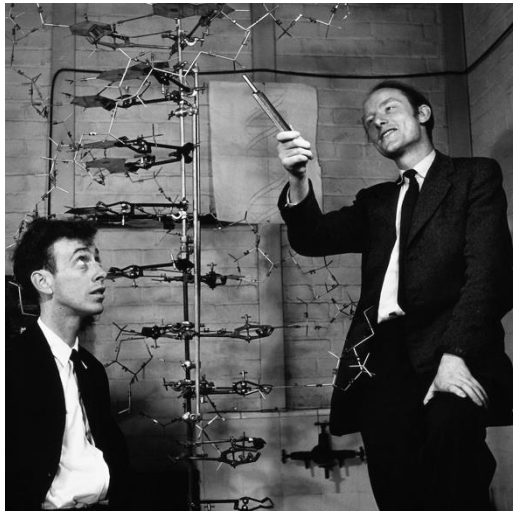
# Extreme training requirements

- “Evo 2 was trained for **several months** on the NVIDIA DGX Cloud AI platform via AWS, utilizing over **2,000** NVIDIA **H100 GPUs** and bolstered by **collaboration with NVIDIA** researchers and engineers.” - Arc Institute blog post
- Roughly **$10M** training

# Braggadocious framing

- “Biological research scales from molecules to systems to organisms, seeking to understand and design functional components across all domains of life (**Darwin**, 1859; **Mendel**, 1866; Dobzhansky, 1951) … All domains of life express complex functions from DNA sequences (**Watson and Crick**, 1953; Nirenberg and Matthaei, 1961), yet genomic content and length vary dramatically across organisms.”





# Difficulty of use

- Evo2-40B weights take up **80GB** in GPU RAM
- H100 node required to compile the code and run inference
- Can take **hours** of inference time to process a **single sample**

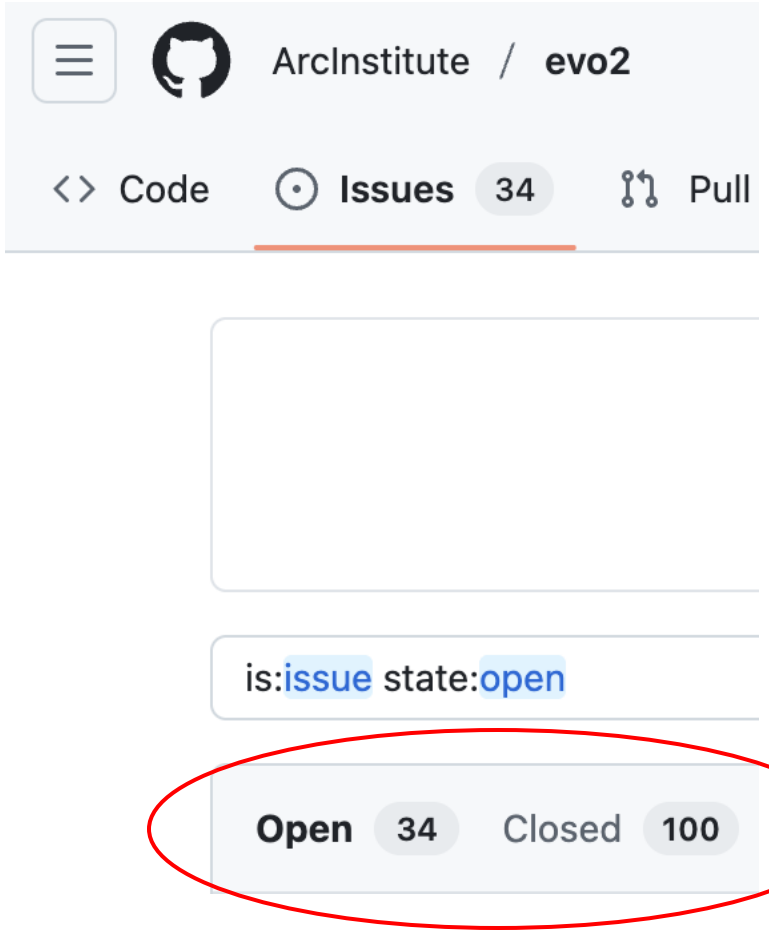
Nvidia H100 NVL GPU Computing Processor

$29,500.00

Pre-owned

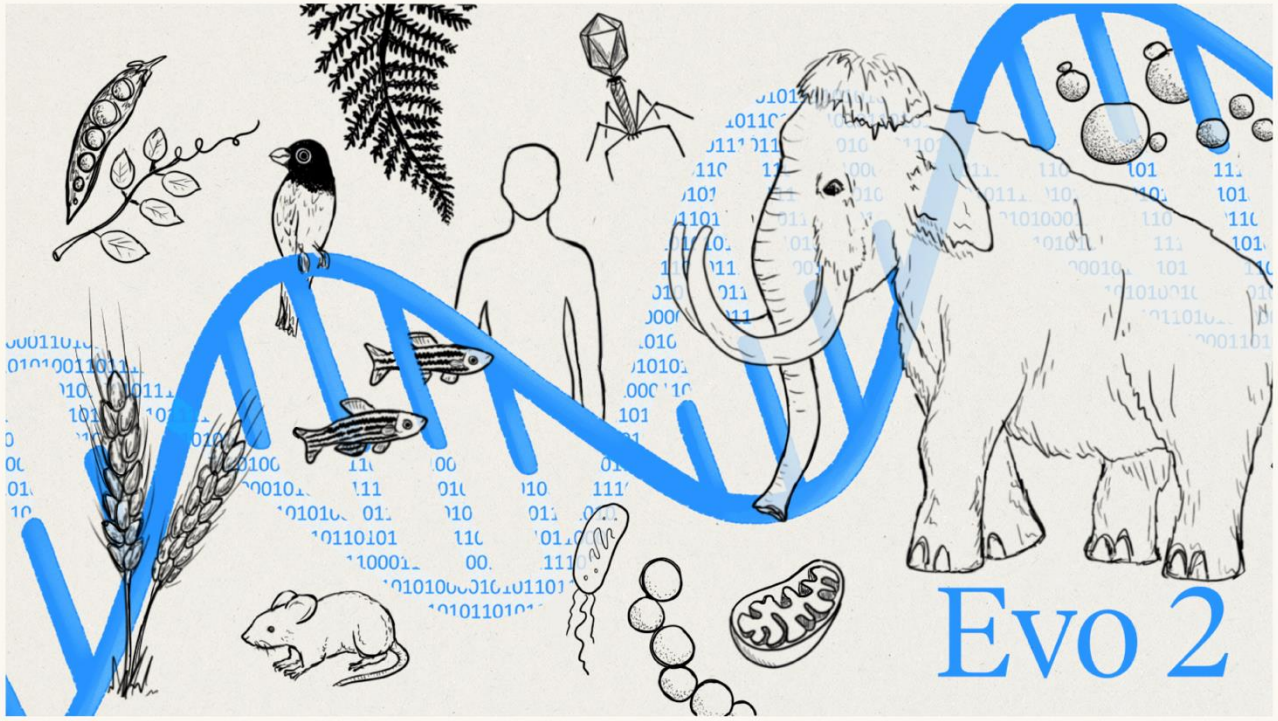
eBay & more

4.0 ★★★★☆ (1)



# What makes Evo2 so impressive?

**Arc Institute develops the largest AI model for biology to date in collaboration with NVIDIA, bringing together Stanford University, UC Berkeley, and UC San Francisco researchers**



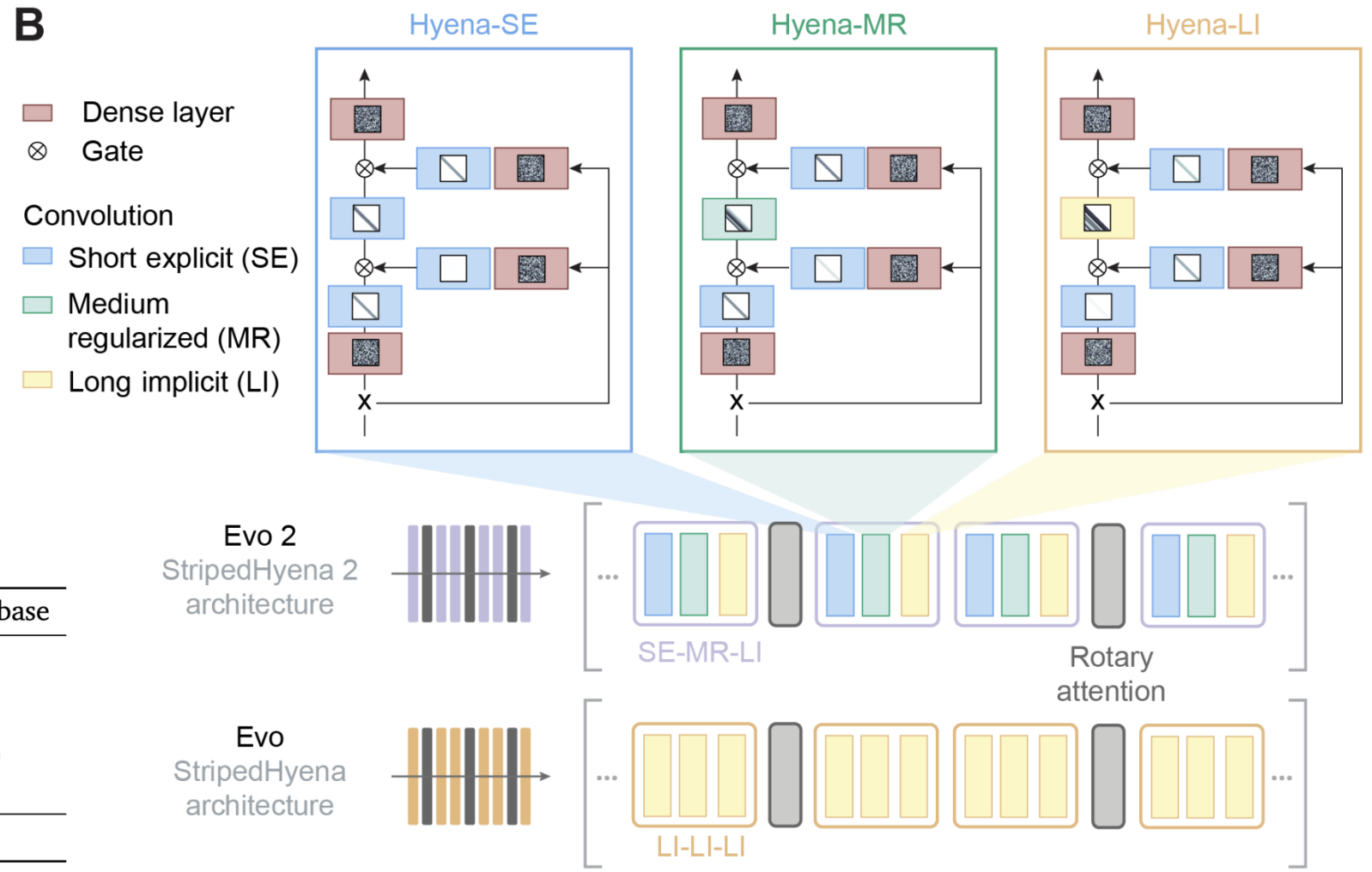

Evo 2 is trained on over 9.3 trillion tokens–in this case, nucleotides–from over 128 thousand genomes across the three domains of life, making it similar in scale to the most powerful generative AI large language models.

# Model architecture

- **Convolutions** find **features**
- **Attention** gives features **context**
- **Early** layer features are **simple**, like kmer motifs
- **Late** layer features are **complex**, like gene-specific CREs

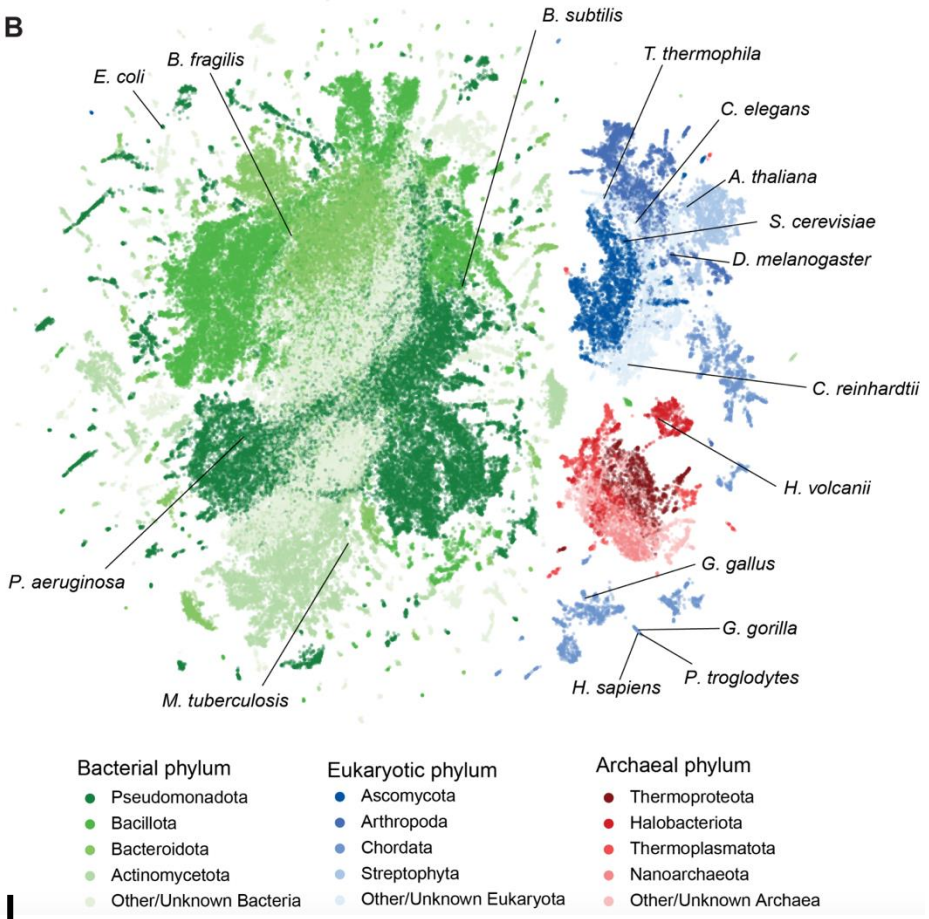
- StripedHyena 2
- Intentionally **patterned** series of **convolutions** and **attention** layers
- Mix of 7, 128, and full-length convolutions

| | Evo 2 40B | Evo 2 7B | Evo 2 1B base |
|---|---|---|---|
| Parameters | 40.3B | 6.5B | 1.1B |
| Total Layers | 50 | 32 | 25 |
| Hidden Size | 8,192 | 4,096 | 1,920 |
| FFN Size | 22,528 | 11,264 | 5,120 |
| Num Heads | 64 | 32 | 15 |
| Total Tokens | 9.3T | 2.4T | 1T |

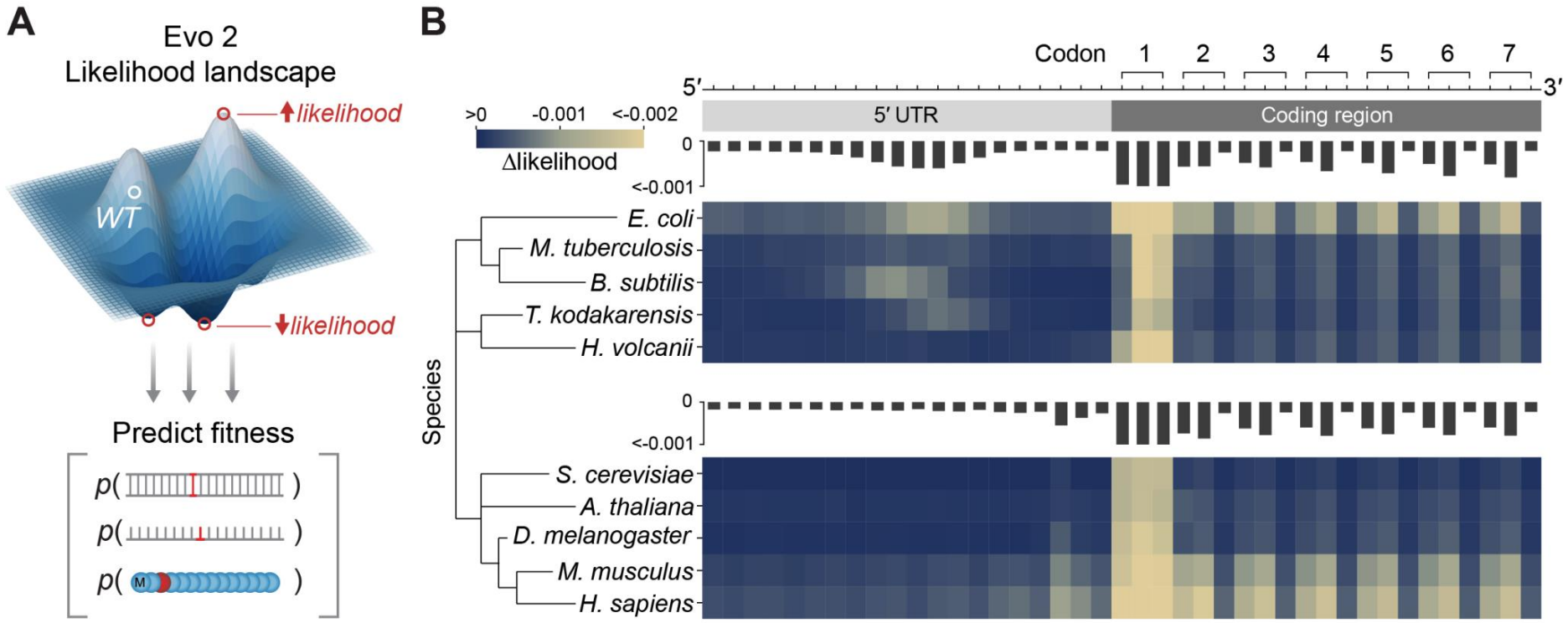


# Training data

- OpenGenome2 dataset
- 113k prokaryotic genomes (357B bp)
- 15,032 eukaryotic genomes (6.98T bp)
- Non-redundant metagenomes (854B bp)
- Euk. organelle genomes (2.82B bp)
- Euk. coding sequences (602B bp)
- 8.8T bases total
- Data has no labels; **self-supervised** model

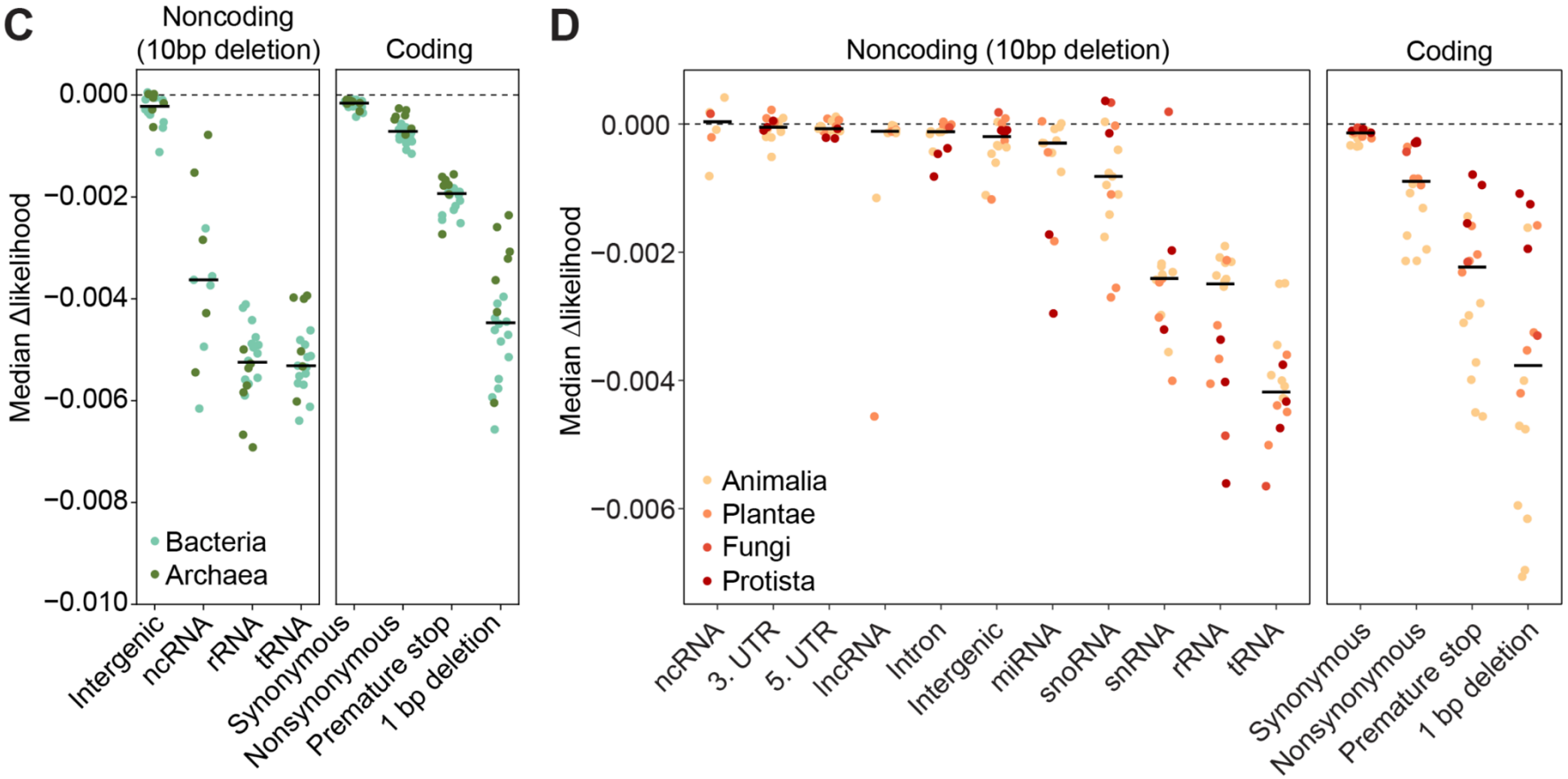


# Evo2 learns coding syntax without any labels



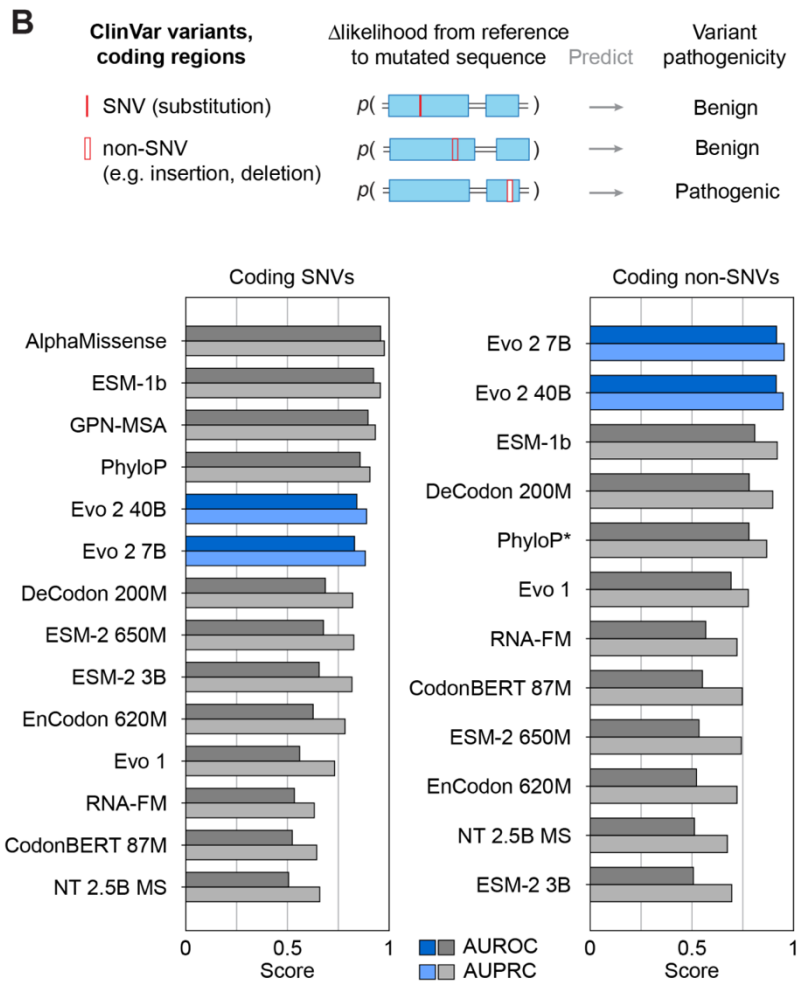

- **Mutate bases** at different positions and measure **change** in sequence **likelihood**
- Learns that mutations in the **1st codon base** are **very impactful** / unlikely
- Learns that mutations in the **3rd codon base** (wobble position) are **less impactful**

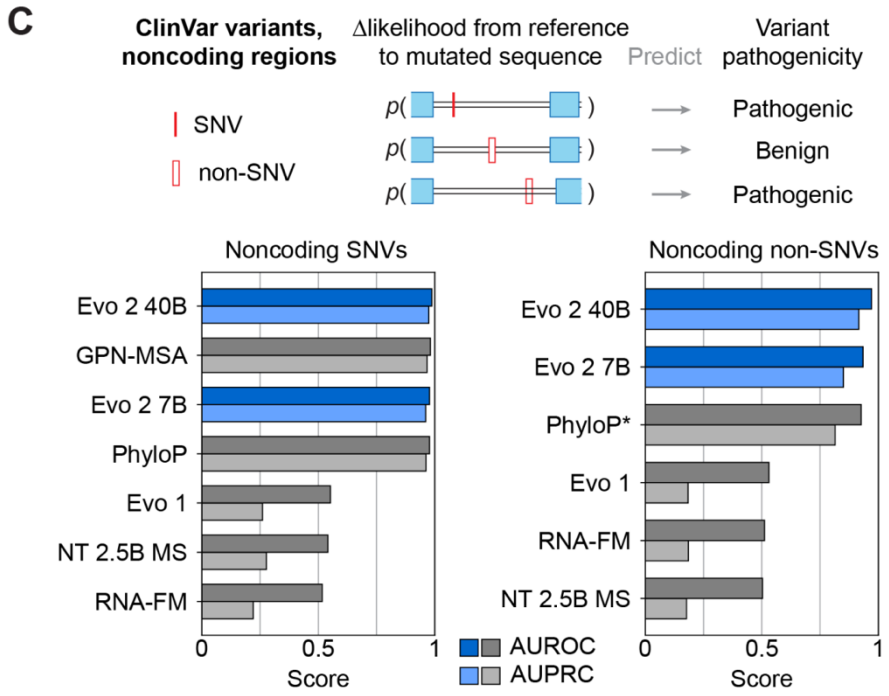
# Evo2 learns mutation significance w/ no labels




- **Implicit** understanding that some mutations are more impactful than others

# Good Evo2 zero-shot performance on ClinVar

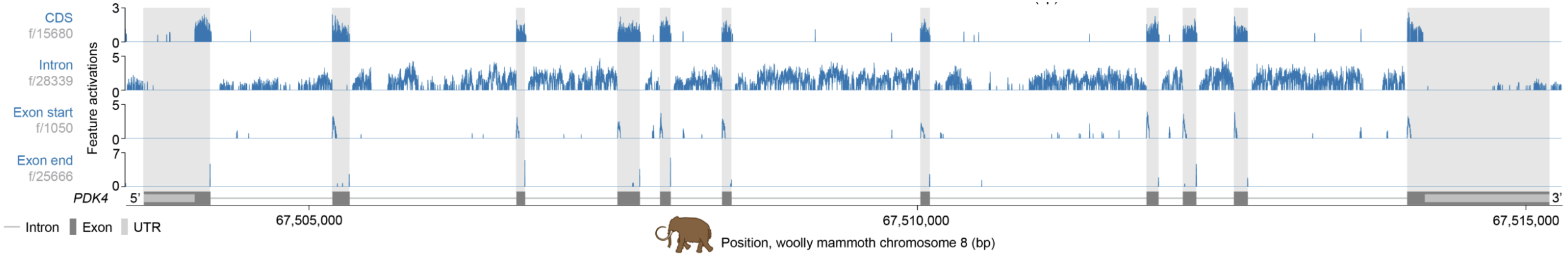



- Uses **sequence likelihood** as a proxy for **variant effect**
  - Loss-of-function variants tend to be pathogenic
- 15k coding, 38k noncoding variants
- Competes with protein LMs and **exceeds** DNA LMs

# Interpretable features can annotate genomes

- Use four features to annotate the woolly mammoth genome
- Introns, exons, and coding regions



# Enformer vs. Evo 2

- Represent **two paradigms** in sequence-to-function modelling
- Both have inspired a wide range of follow-up studies

| | Enformer | Evo 2 |
|---|---|---|
| **Training strategy** | Supervised | Self-supervised |
| **Training genomes** | Human + mouse | Thousands of organisms |
| **Bases of training sequence** | 6 billion | 9.3 trillion |
| **Input context window** | 197 Kbp | 1 Mbp |
| **Parameters** | 250 M | 40 B |
| **Text processing module** | Transformer | Striped Hyena 2 |

# Deep learning in genomics pt. 2

5 November 2025

Mahler Revsine

# Announcements

- Assignment 5 due Friday night
  - Last chance to use late days
  - This is a busy time in the semester, may as well use them! ☺
- Reminder: there is an oral defense component of the final project
  - Question and answer session with me and Dr. Schatz
  - Everyone is responsible for understanding their full project
  - If you're doing the work, you have nothing to worry about
  - If you're not contributing to your team, you will struggle

# Currently hard to understand variant effects

| Method | Drawback(s) |
|---|---|
| Query it in a database | The variant might not exist in a database<br>The variant might act differently under different conditions |
| Assume loss-of-function of nearby gene | Not always an accurate assumption<br>Function of nearby gene may not itself be understood |
| Genome-Wide Association Study (GWAS) | Not applicable for variants that only exist in a single person<br>Underpowered for rare variants that only occur in a few people<br>Relies on undersized datasets containing a small number of SNPs<br>Requires rigorous statistical testing<br>Can only model one variant at a time<br>Assumes additive linear effects of variants<br>Difficult to adapt to a polygenic/omnigenic model of traits<br>In practice, only explains a fraction of heritability |

**Every existing method has its weaknesses**

# Enformer and Evo 2 are two ways forward

- Represent **two paradigms** in sequence-to-function modelling
- Both have inspired a wide range of follow-up studies

| | Enformer | Evo 2 |
|---|---|---|
| **Training strategy** | Supervised | Self-supervised |
| **Training genomes** | Human + mouse | Thousands of organisms |
| **Bases of training sequence** | 6 billion | 9.3 trillion |
| **Input context window** | 197 Kbp | 1 Mbp |
| **Parameters** | 250 M | 40 B |
| **Text processing module** | Transformer | Striped Hyena 2 |

# DNABERT

- The first Large Language Model for genomics
- Initially published in 2020, 3 years after Transformers were developed

JOURNAL ARTICLE

## DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome FREE

Yanrong Ji, Zhihan Zhou, Han Liu ✉, Ramana V Davuluri ✉ Author Notes

*Bioinformatics*, Volume 37, Issue 15, August 2021, Pages 2112–2120,
https://doi.org/10.1093/bioinformatics/btab083
**Published:** 04 February 2021 **Article history** ▾

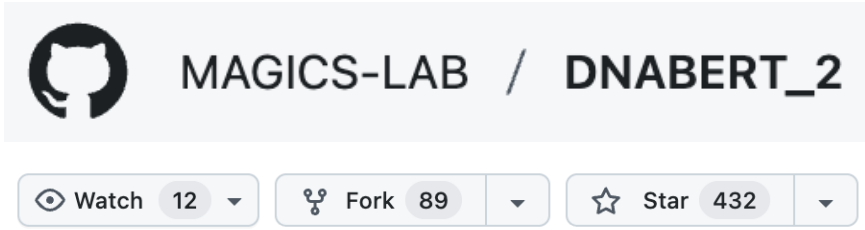
## DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes

arXiv:2306.15006v2 [q-bio.GN] 18 Mar 2024

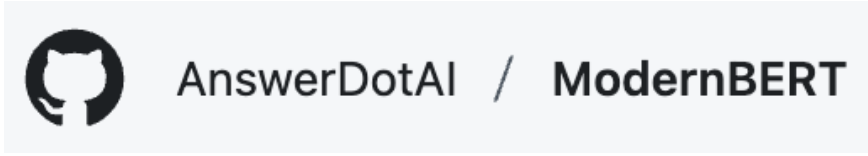
**Zhihan Zhou**[†] **Yanrong Ji**[†] **Weijian Li**[†] **Pratik Dutta**[‡] **Ramana V Davuluri**[‡] **Han Liu**[†]
[†] Department of Computer Science, Northwestern University, Evanston, IL, USA
[‡] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA
{zhihanzhou, yanrongji, weijianli}@u.northwestern.edu
pratik.dutta@stonybrook.edu, Ramana.Davuluri@stonybrookmedicine.edu
hanliu@northwestern.edu

# DNABERT-2 is truly open source

- Very easy to use compared to some other models (looking at you Evo-2)
- Full instructions on their GitHub for pre-training and fine-tuning
- Datasets are mostly available
- I use it in my research (more on that later)



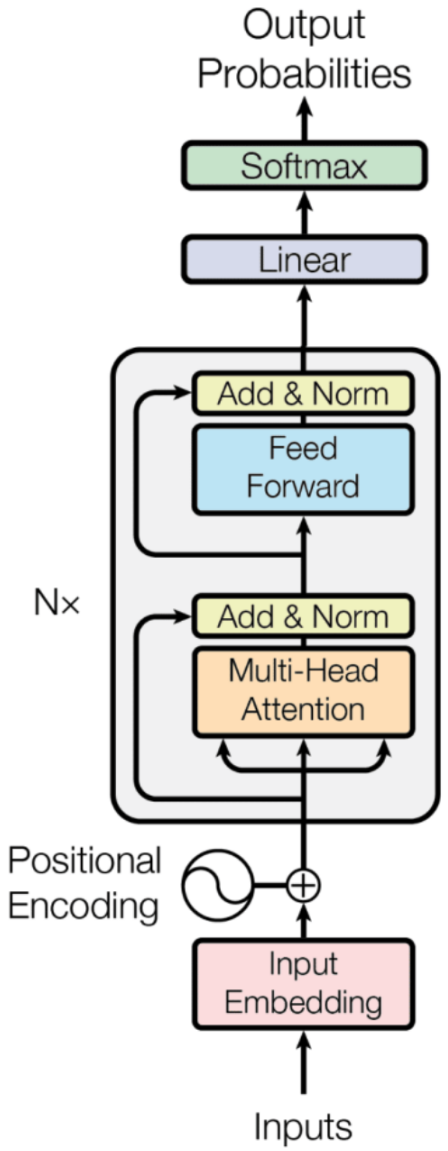

Use this instead ->



https://towardsdatascience.com/a-complete-guide-to-bert-with-code-9f87602e4a11/

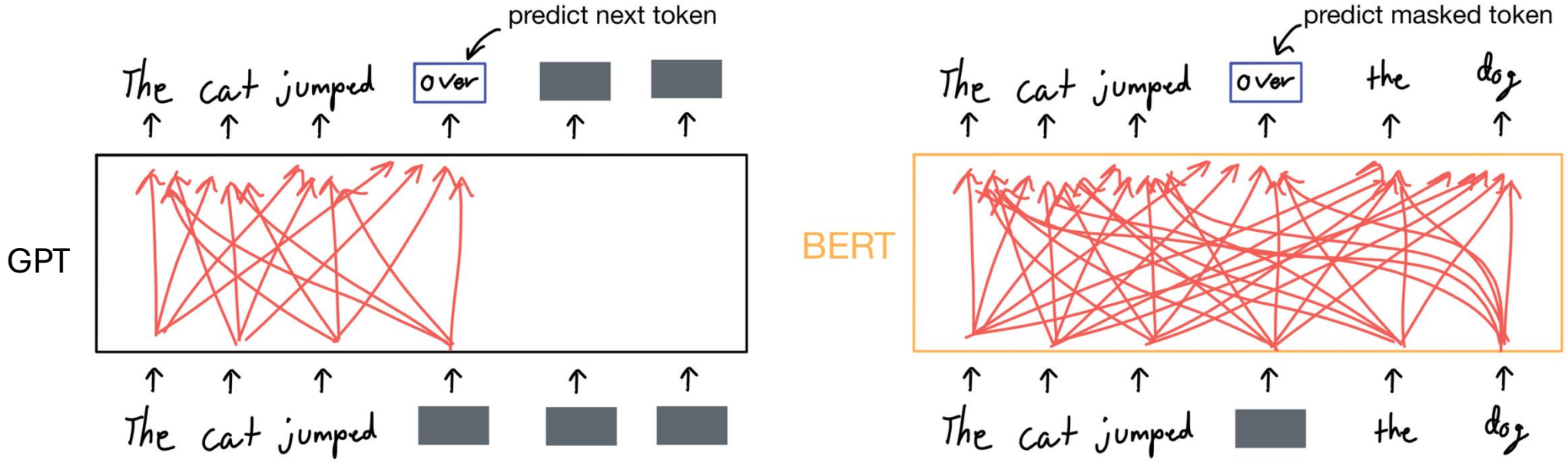
# BERT-based architecture

- BERT:
  - Bidirectional
    - Unlike Evo-2 or ChatGPT, uses left and right context to make predictions
    - ChatGPT: The cat sat on the ___
    - BERT: The cat ___ on the mat
  - Encoder
    - Project text into latent space embedding, but no need to decode back into text
  - Representations from
    - Creates a numerical representation of the input via…
  - Transformers
    - Uses the transformer module for text processing



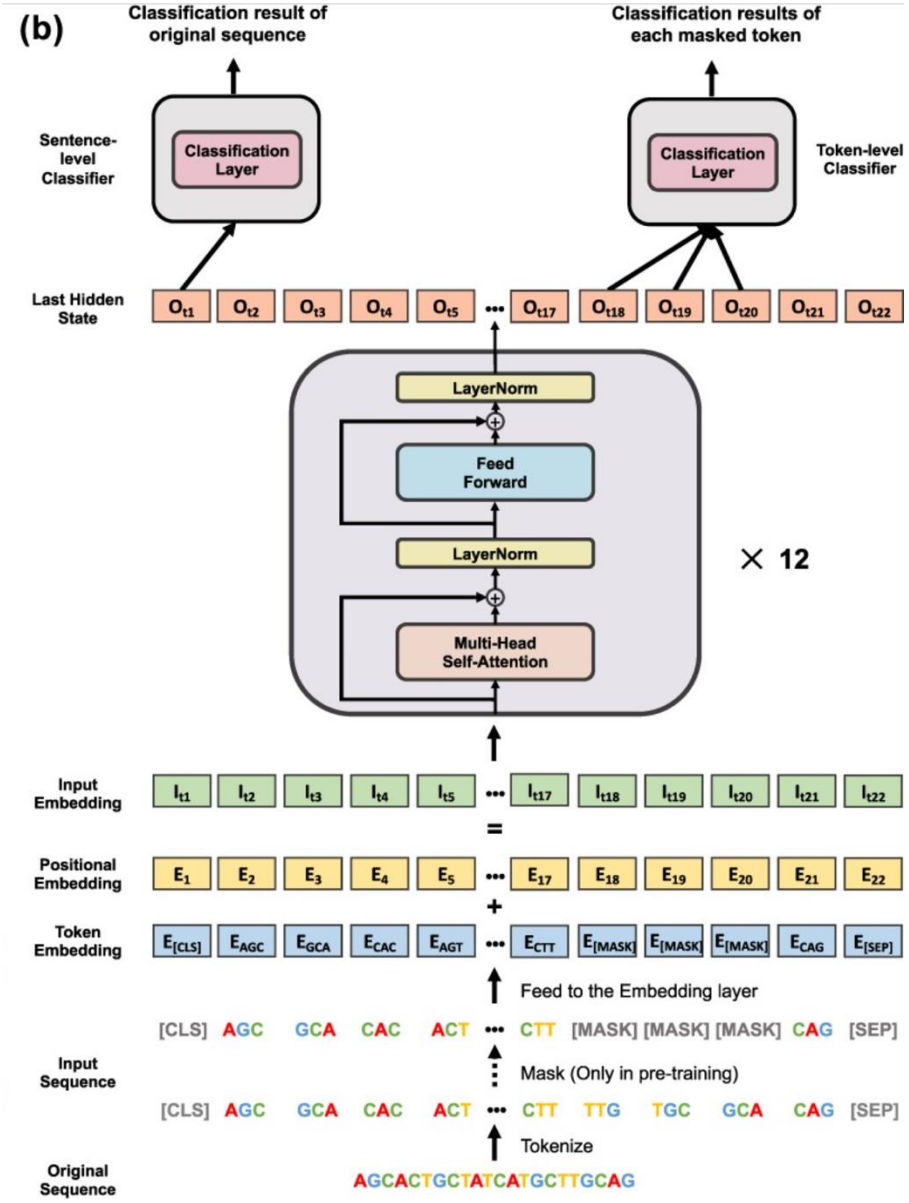
# Masked token prediction

- Using left and right context, predict hidden word
- More information can lead to better predictions
- However, can't be used for text generation

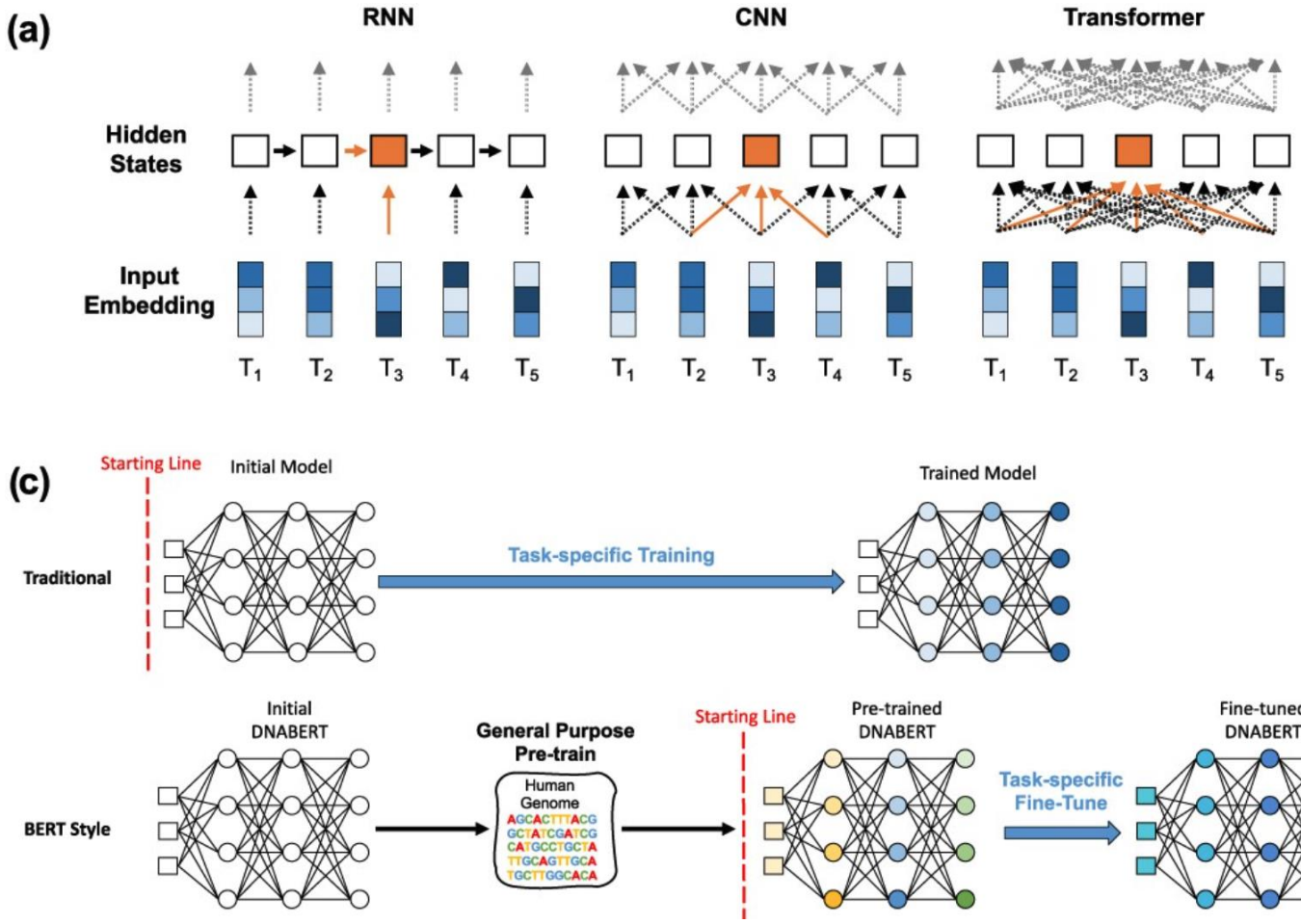


# DNABERT architecture

- Off-the-shelf BERT implementation
  - Same architecture that could be used for natural language
- DNA sequence is grouped into tokens
- Tokens pass through transformer and feed-forward layers
- Final prediction on masked token
- 117M parameter model was big at the time but is small by today's standards



# The first transformer for genomics




- Transformers are more advanced sequence processors
- Pre-training and fine-tuning is a more efficient training regime

https://en.wikipedia.org/wiki/Byte-pair_encoding

# Byte-pair encoding

- Iteratively find most frequent tokens
- Repeat until hitting target vocabulary size
- DNABERT-2 uses 4,096 tokens

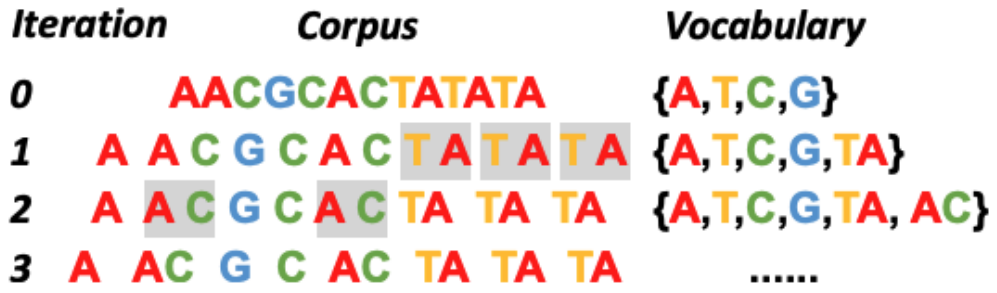



Figure 2: Illustration of the BPE vocabulary constructions.

Suppose the data to be encoded is:[8]

```
aaabdaaabac
```

The byte pair "aa" occurs most often, so it will be replaced by a byte that is not used in the data, such as "Z". Now there is the following data and replacement table:

```
ZabdZabac
Z=aa
```

Then the process is repeated with byte pair "ab", replacing it with "Y":

```
ZYdZYac
Y=ab
Z=aa
```

The only literal byte pair left occurs only once, and the encoding might stop here. Alternatively, the process could continue with recursive byte-pair encoding, replacing "ZY" with "X":

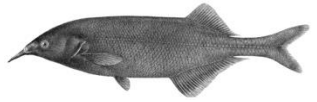
```
XdXac
X=ZY
Y=ab
Z=aa
```

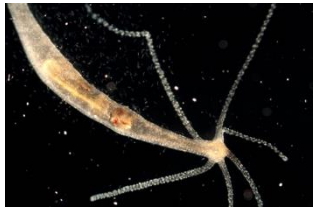
# Training dataset

- 135 reference genomes
  - Human GRCh38
  - + 2 primates (macaca assamensis and nigra)
  - + 4 mammals (mouse, deermouse, camel, buffalo)
  - + 5 vertebrates (fish and birds)
  - 10 invertebrates (mostly bugs), 2 protozoa, 11 fungi
  - 100 bacteria

| Category | Species | Num. of Nucleotides (M) |
|---|---|---|
| Fungi | Ceratobasidium | 655.37 |
| | Claviceps Maximensis | 329.79 |
| | Fusarium Annulatum | 449.98 |
| | Melampsora | 699.52 |
| | Metschnikowia | 109.36 |
| | Mucor Saturninus | 391.17 |
| | Penicillium Chermesinum | 275.81 |
| | Saccharomyces Cerevisiae | 121.54 |
| | Sporopachydermia Quercuum | 155.71 |
| | Tranzscheliella Williamsii | 184.77 |
| | Xylariales | 399.96 |
| Protozoa | Phytophthora Sojae | 792.65 |
| | Pythium Apiculatum | 450.99 |
| Mammalian | Bubalus Bubalis | 28768.00 |
| | Camelus Dromedarius | 19757.02 |
| | Human | 31372.10 |
| | Macaca Assamensis | 27593.76 |
| | Macaca Nigra | 28217.13 |
| | Mus Musculus | 26545.98 |
| | Peromyscus Californicus | 24677.56 |
| Invertebrate | Brachionus Rubens | 1327.37 |
| | Ceroptres Masudai | 12.95 |
| | Cotesia Typhae | 1866.62 |
| | Croniades Pieria | 3889.85 |
| | Drosophila Athabasca | 1221.16 |
| | Emesis Russula | 4848.08 |
| | Hydra Oligactis | 12597.75 |
| | Meganola Albula | 3604.25 |
| | Oscheius | 383.21 |
| | Rutpela Maculata | 20213.33 |
| Other Vertebrate | Anas Zonorhyncha | 11697.08 |
| | Coregonus Clupeaformis | 26824.02 |
| | Gnathonemus Longibarbis | 7314.74 |
| | Myxocyprinus Asiaticus | 23407.19 |
| | Rhipidura Dahli | 10112.96 |
| Bacteria | Aeromonas | 47.33 |
| | Agrobacterium | 97.22 |
| | Alcaligenaceae Bacterium | 20.88 |
| | Aliivibrio | 46.48 |
| | Alphaproteobacteria Bacterium | 14.22 |
| | Amycolatopsis Antarctica | 63.43 |
| | Anaerostipes Faecis | 32.00 |

(+ 90 more bacteria)





# Fine-tuning tasks

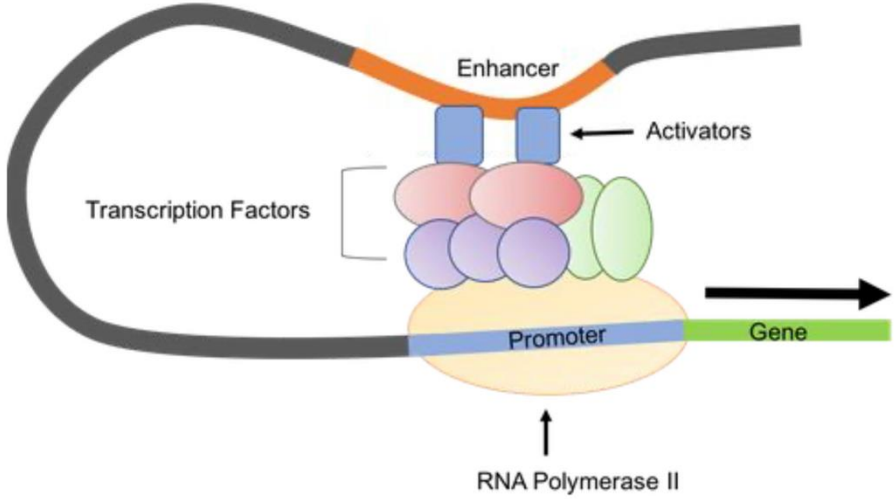
- 7 major tasks
- 28 datasets
- 4 species of origin
- All classification; n=2, 3, or 9

| Species | Task | Num. Datasets | Num. Classes | Sequence Length |
|---|---|---|---|---|
| Human | Core Promoter Detection | 3 | 2 | 70 |
| | Transcription Factor Prediction | 5 | 2 | 100 |
| | Promoter Detection | 3 | 2 | 300 |
| | Splice Site Detection | 1 | 3 | 400 |
| Mouse | Transcription Factor Prediction | 5 | 2 | 100 |
| Yeast | Epigenetic Marks Prediction | 10 | 2 | 500 |
| Virus | Covid Variant Classification | 1 | 9 | 1000 |

Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

| Task | Metric | Datasets | Train / Dev / Test |
|---|---|---|---|
| Core Promoter Detection | mcc | tata | 4904 / 613 / 613 |
| | | notata | 42452 / 5307 / 5307 |
| | | all | 47356 / 5920 / 5920 |
| Promoter Detection | mcc | tata | 4904 / 613 / 613 |
| | | notata | 42452 / 5307 / 5307 |
| | | all | 47356 / 5920 / 5920 |
| Transcription Factor Prediction (Human) | mcc | wgEncodeEH000552 | 32378 / 1000 / 1000 |
| | | wgEncodeEH000606 | 30672 / 1000 / 1000 |
| | | wgEncodeEH001546 | 19000 / 1000 / 1000 |
| | | wgEncodeEH001776 | 27294 / 1000 / 1000 |
| | | wgEncodeEH002829 | 19000 / 1000 / 1000 |
| Splice Site Prediction | mcc | reconstructed | 36496 / 4562 / 4562 |
| Transcription Factor prediction (Mouse) | mcc | Ch12Nrf2Iggrab | 6478 / 810 / 810 |
| | | Ch12Znf384hpa004051Iggrab | 53952 / 6745 / 6745 |
| | | MelJundIggrab | 2620 / 328 / 328 |
| | | MelMafkDm2p5dStd | 1904 / 239 / 239 |
| | | MelNelfeIggrab | 15064 / 1883 / 1883 |
| Epigenetic Marks Prediction | mcc | H3 | 11971 / 1497 / 1497 |
| | | H3K14ac | 26438 / 3305 / 3305 |
| | | H3K36me3 | 27904 / 3488 / 3488 |
| | | H3K4me1 | 25341 / 3168 / 3168 |
| | | H3K4me2 | 24545 / 3069 / 3069 |
| | | H3K4me3 | 29439 / 3680 / 3680 |
| | | H3K79me3 | 23069 / 2884 / 2884 |
| | | H3K9ac | 22224 / 2779 / 2779 |
| | | H4 | 11679 / 1461 / 1461 |
| | | H4ac | 27275 / 3410 / 3410 |
| Covid Variant Classification | f1 | Covid | 77669 / 7000 / 7000 |

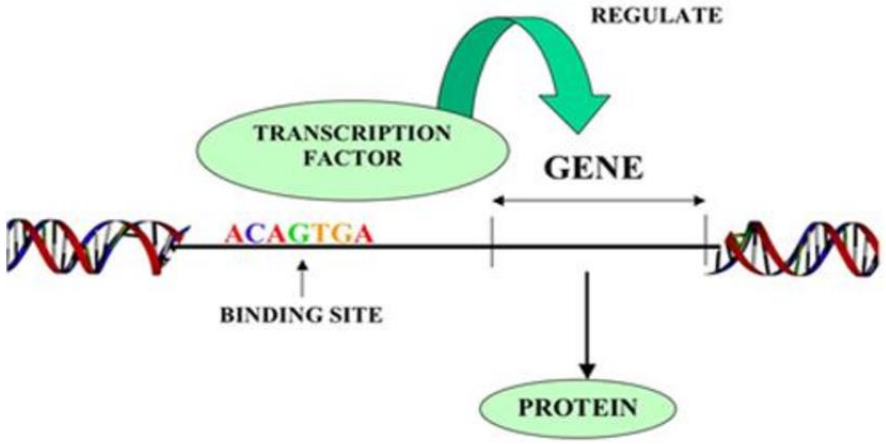
# Gene promoter prediction




- Data from Eukaryotic Promoter Database (EPDnew)
- 10kbp sequences of [+5k, -5k] of transcription start site (TSS)
  - 3k human TATA & 26k non-TATA promoters
- Positives: -249 to 50bp sequences around TSS
- TATA negatives: random 300bp sequences with TATA at 25bp upstream of TSS.
- non-TATA negatives: dinucleotide-shuffled sequences (same #nts)

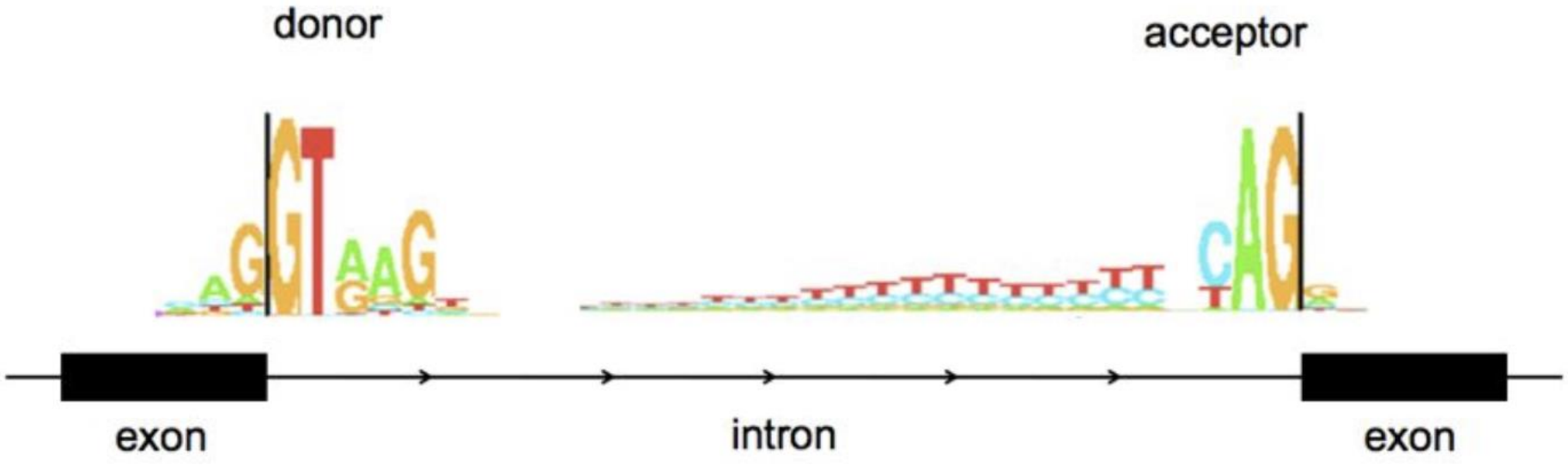
# Transcription factor binding sites

- TF ChIP-seq peak profiles from ENCODE (690 datasets)
- 161 TF binding profiles in 91 human cell lines
- Positives: 101bp region around each ChIP-seq peak
- Negatives: 101bp region with no overlap with ChIP-seq peak, with same GC.



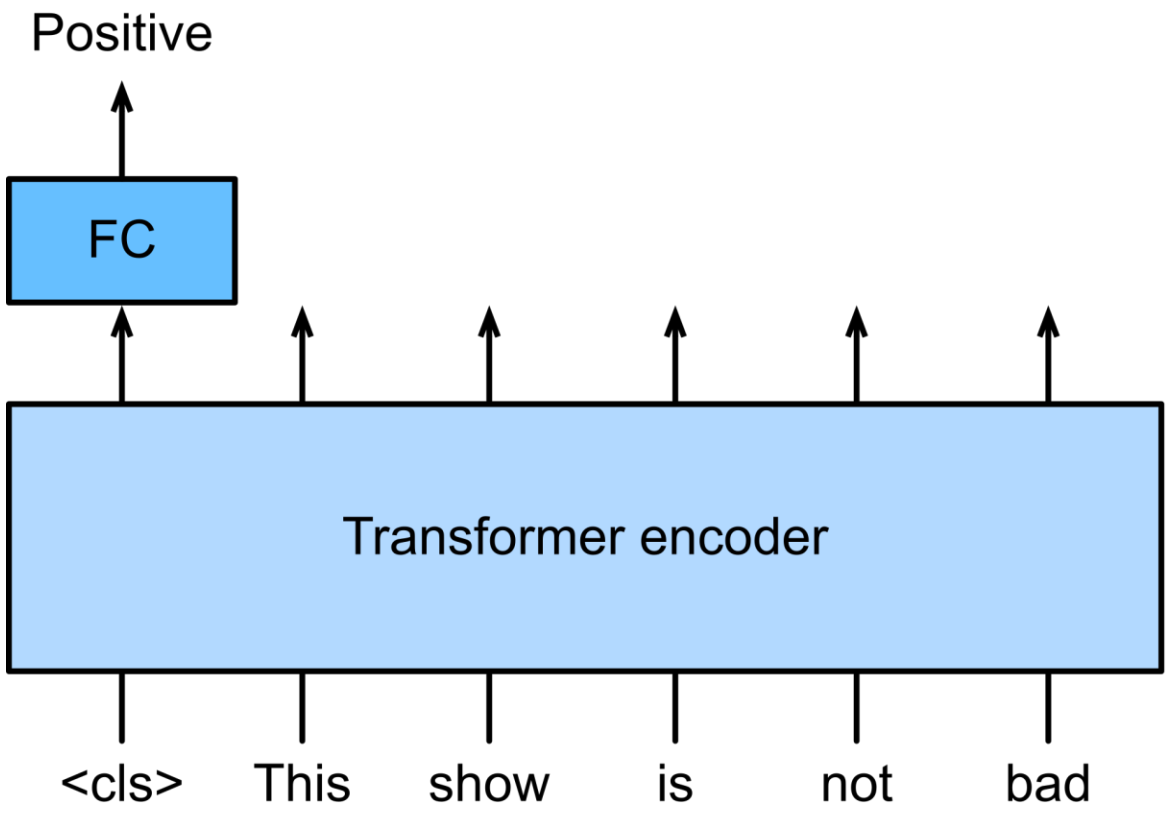
# Splice site identification

- SpliceFinder dataset
- Positive examples: 400bp splice sites
- Negative examples: false positive results from SpliceFinder CNN



# Classification with a BERT model

- Prepend [CLS] token to all inputs
- During attention, [CLS] token learns from all other tokens
- Becomes a sort of summary token for the full input
- Use embedding of corresponding output as predicted label



# DNABERT-2 has strong performance

| | Species | Yeast | Mouse | Virus | Human | | | |
|---|---|---|---|---|---|---|---|---|
| | Task | EMP | TF-M | CVC | TF-H | PD | CPD | SSP |
| 2.5 B parameters | **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** |
| 0.1 B parameters | **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 |

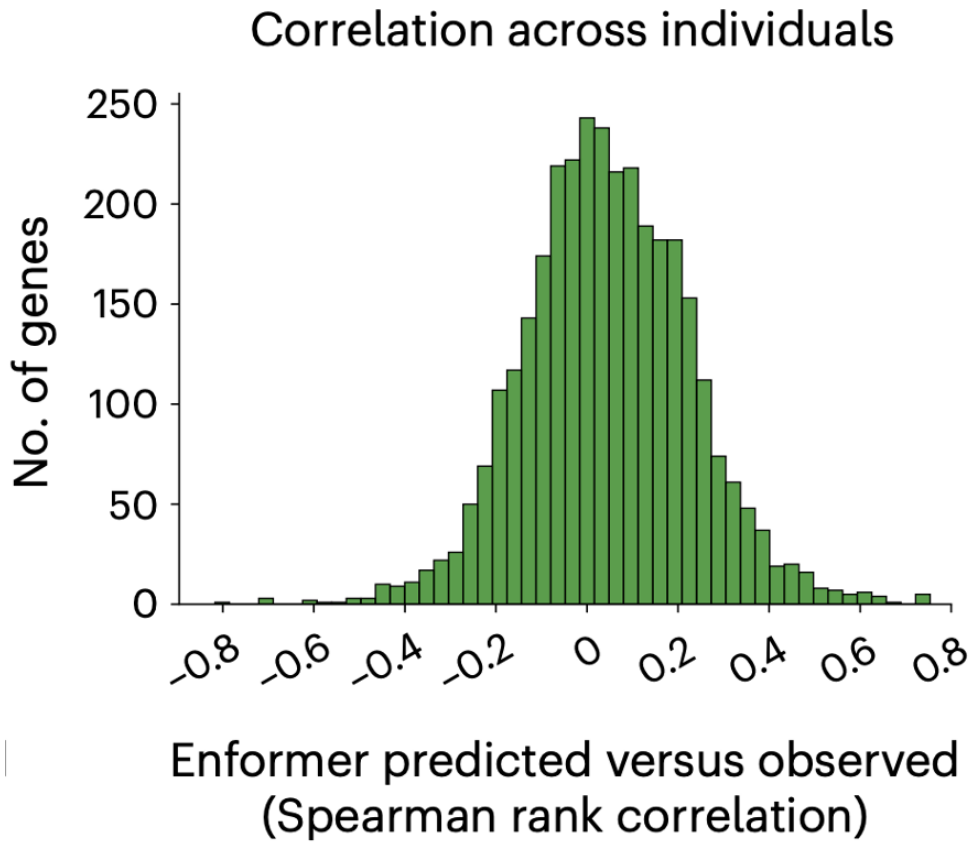
# Model comparisons

- DNABERT-2 is closer to Evo-2 in its training strategy but closer to Enformer in its architecture

| | Enformer | Evo 2 | DNABERT-2 |
|---|---|---|---|
| **Training strategy** | Supervised | Self-supervised | Self-supervised |
| **Training genomes** | Human + mouse | Thousands of organisms | Hundreds of organisms |
| **Bases of training sequence** | 6 billion | 9.3 trillion | 30 billion |
| **Input context window** | 197 Kbp | 1 Mbp | 1 Kbp |
| **Parameters** | 250 M | 40 B | 117 M |
| **Text processing module** | Transformer | Striped Hyena 2 | Transformer |

# Is there room for improvement?

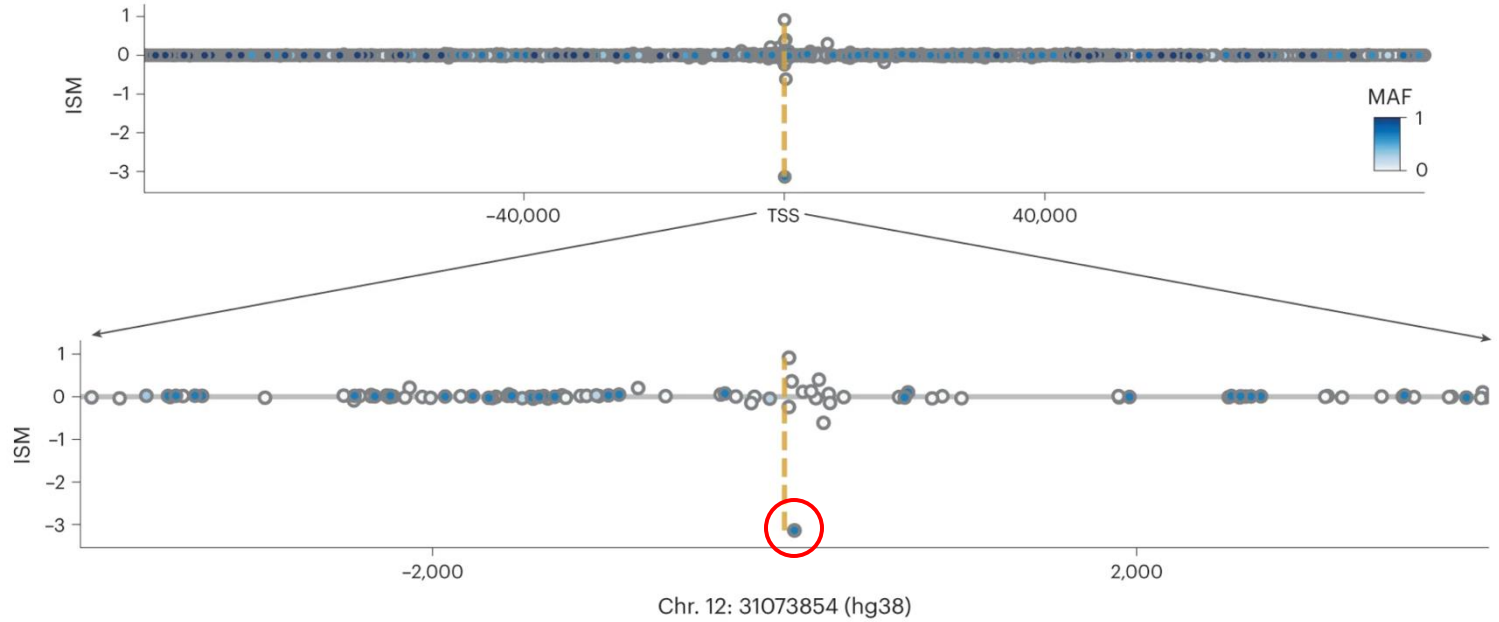
- Yes!

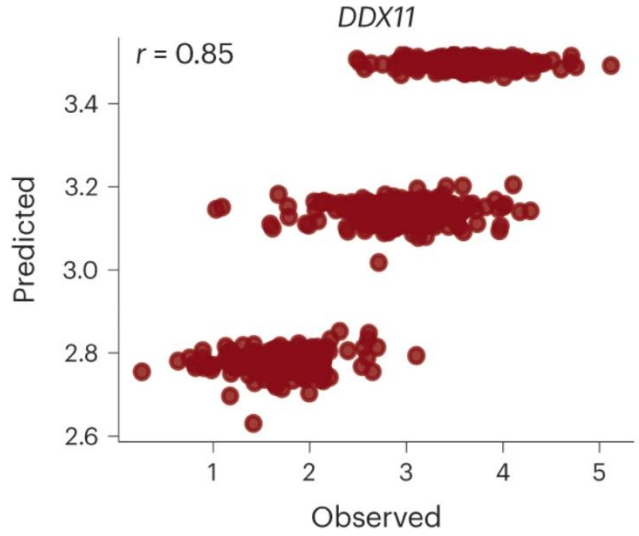


Sasse et al. *Nature Genetics* (2023).

# Using LLMs for variant eQTLs

- Predict **expression** directly from **sequence**
- In theory, can resolve linkage disequilibrium

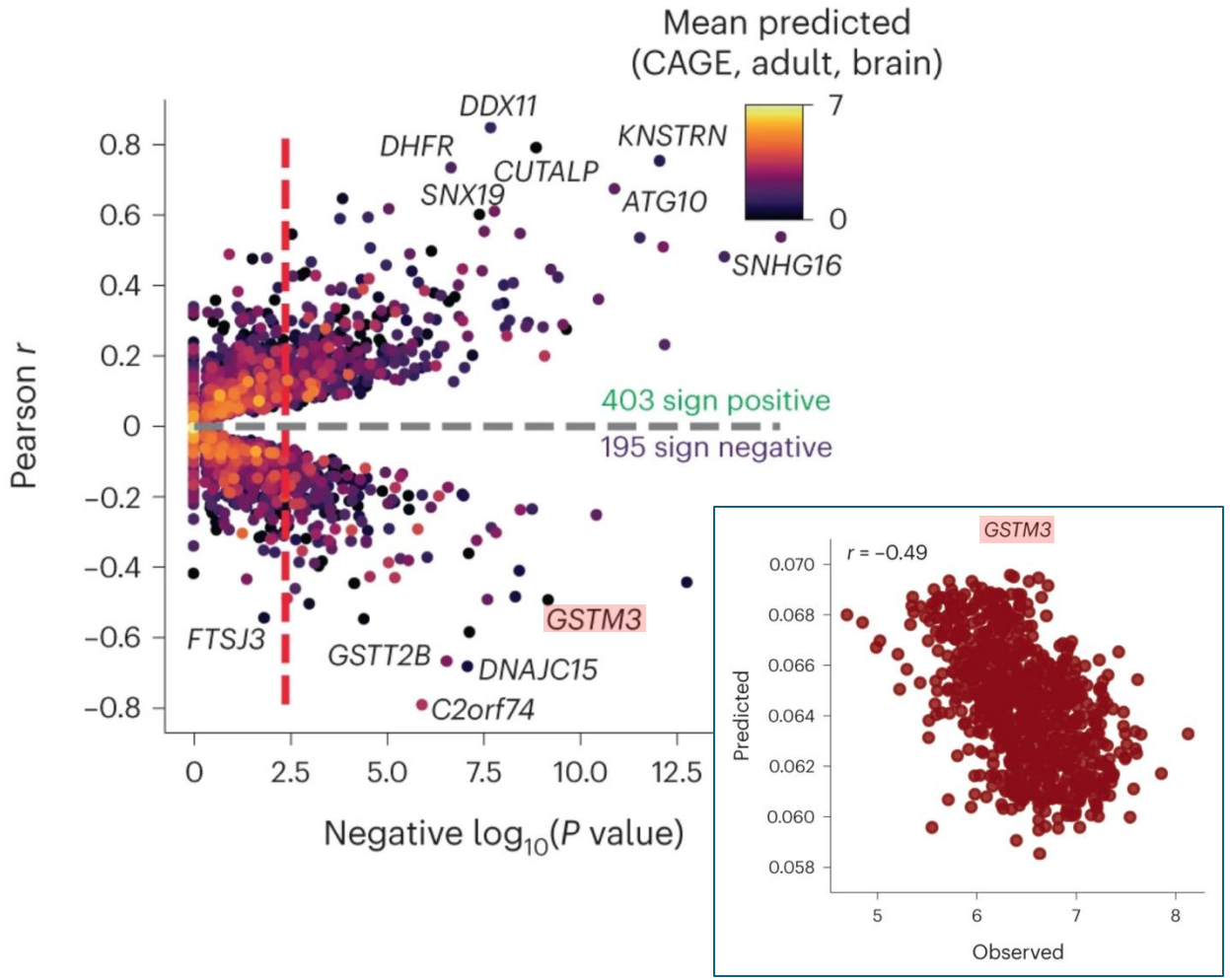



- Enformer can use a **single CNV** out of many candidates to predict eQTL
- Selects same CNV identified by fine-mapping with Susie

Sasse et al. *Nature Genetics* (2023).

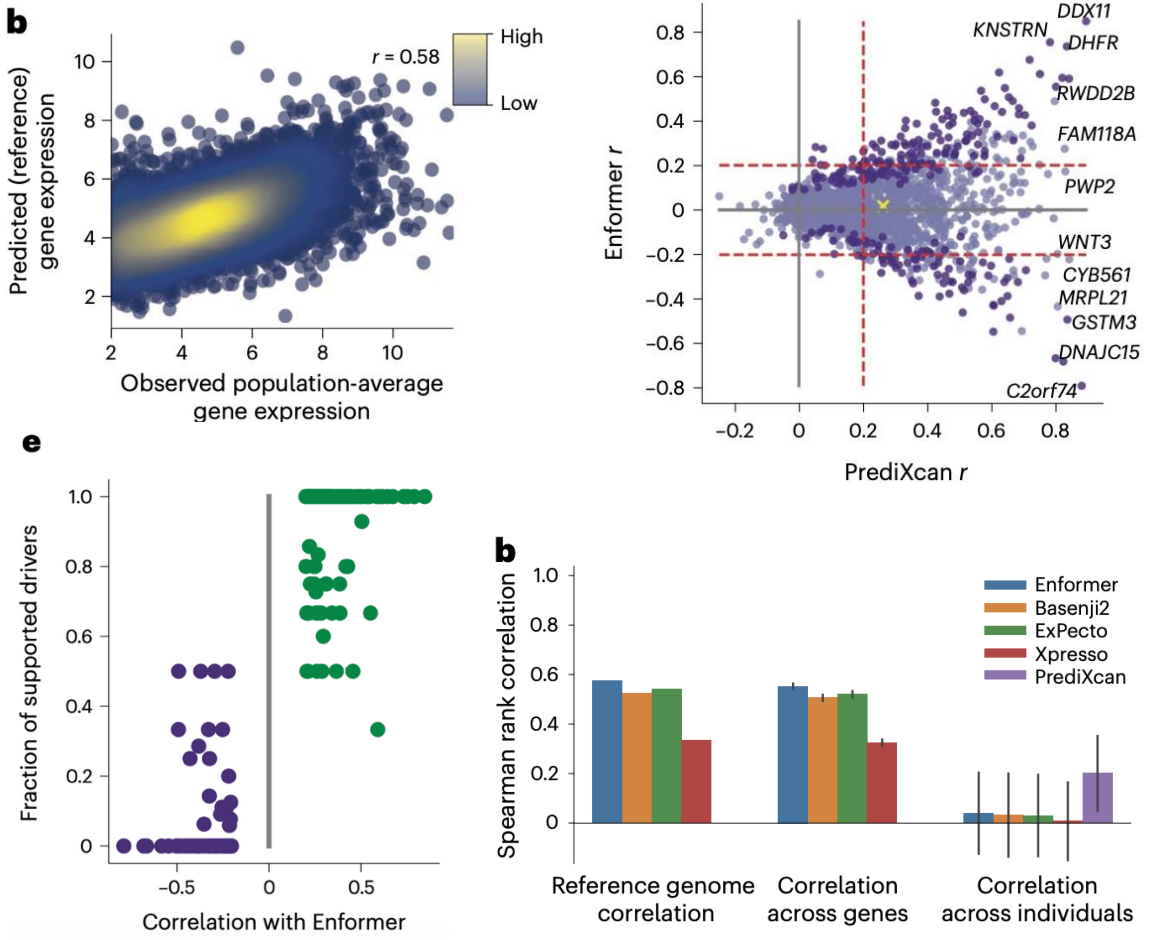
# LLMs do not truly understand variants




- Previous slide is the **exception**, not the rule
- Overall, LLMs know **if** gene expression changes from a mutation but not **how** it changes
- Flip a coin as to whether expression goes up or down

# LLMs do not truly understand variants pt. 2

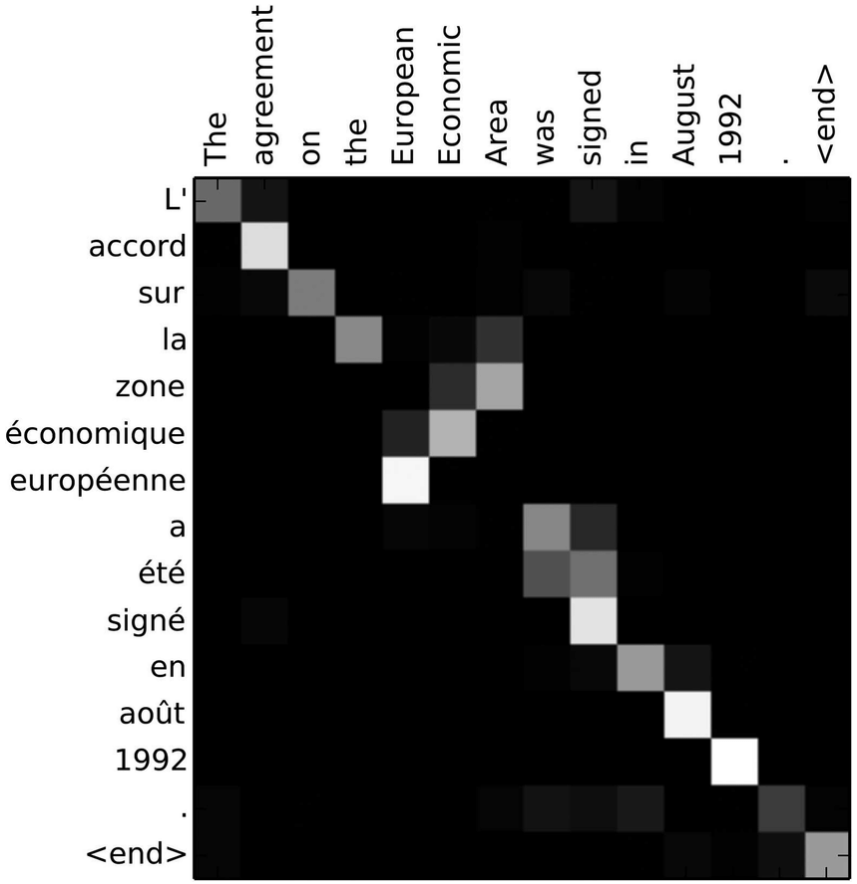
- Enformer is trained on the **reference** genome, not any **individual** sequences
- Top left: Enformer predicts population-average gene expression
- Bottom left: Enformer being right depends on finding causal SNVs
- Top,bottom right: linear model PrediXcan (purple) outperforms Enformer on individual-level predictions



https://www.researchgate.net/figure/A-visualization-of-attention-weights-showing-soft-alignment-between-source-and-target_fig1_332160864

# Expanding to longer inputs

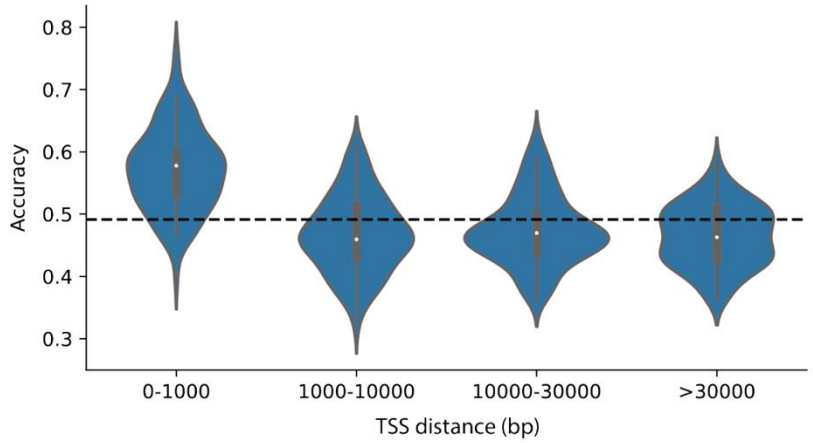
- Genomic sequences are **long**, relevant info may be **far apart**
- Attention is a **quadratic** operation
  - 1K tokens = 1M operations
  - 1M tokens = 1T operations
- Current methods to address this:
  1. Dilation / downsampling
     - e.g. Enformer dilates via convolutions and pooling, losing resolution
  2. Mamba-style text module
     - e.g. Evo-2

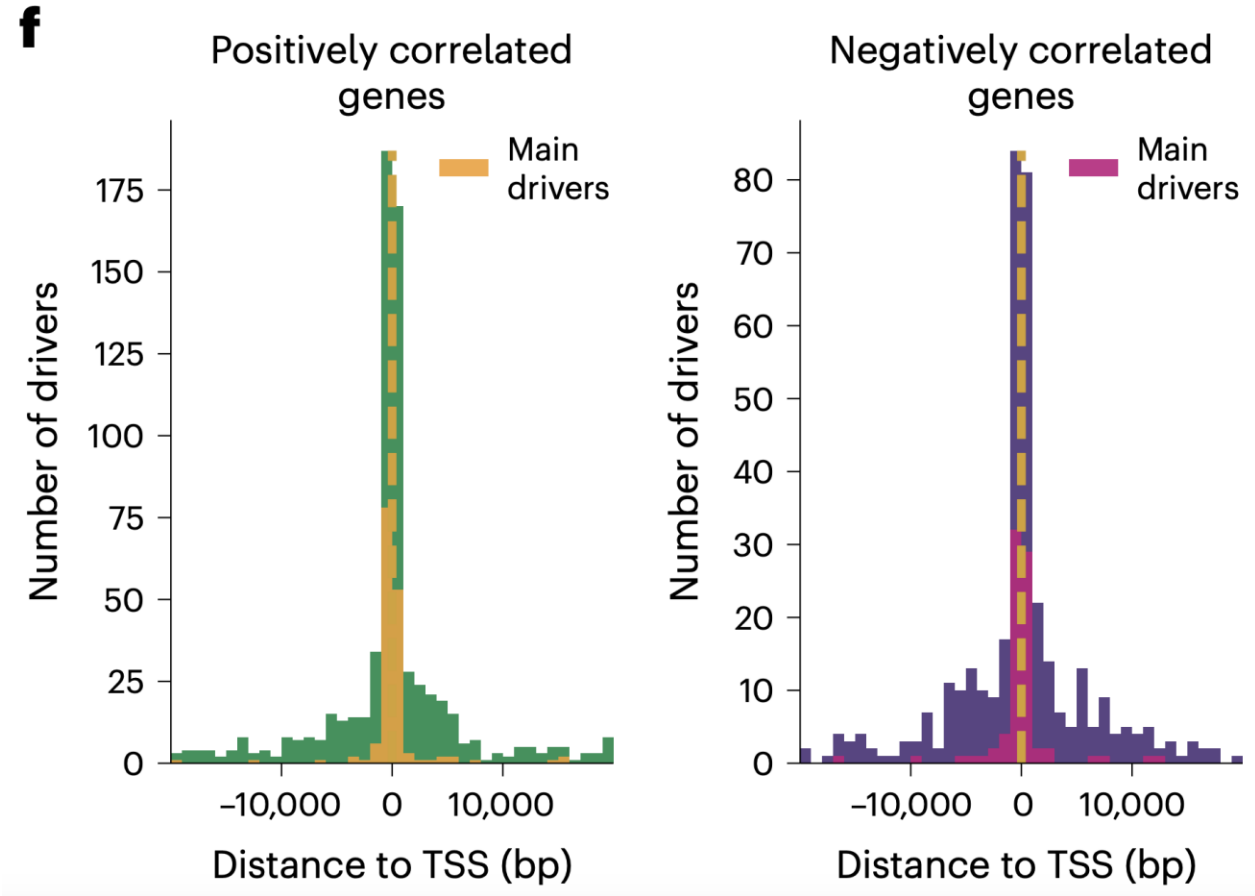


Sasse et al. *Nature Genetics* (2023).
Huang et al. *Nature Genetics* (2023).

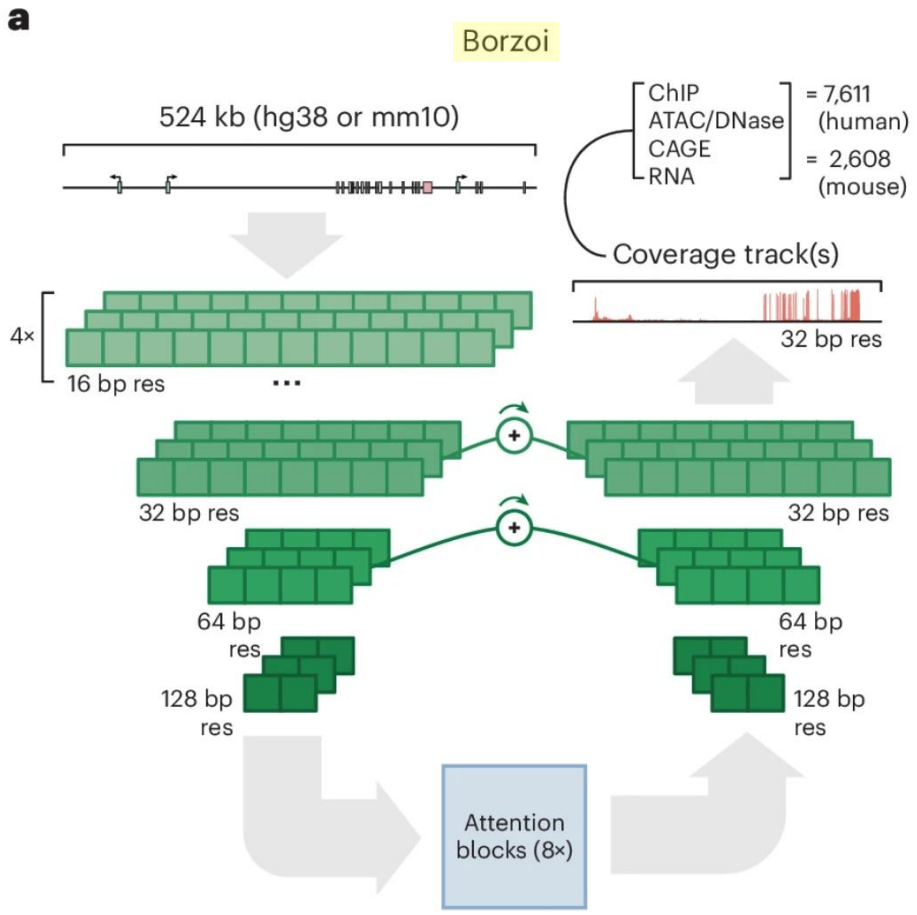
# Making full use of long inputs

- Enformer generally finds drivers near the **transcription start site**
- Model is most **accurate** on variants near TSS



# Better model architectures



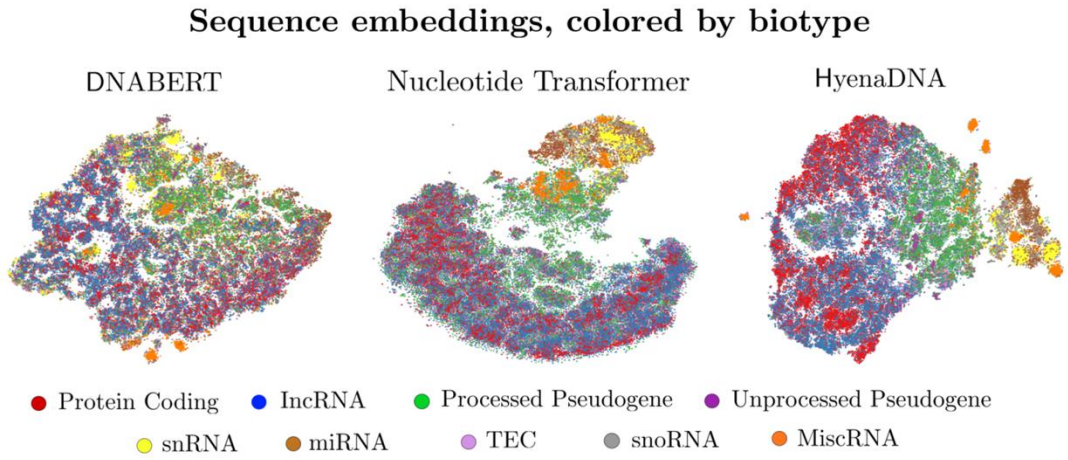

- Borzoi's **updated architecture** over Enformer led to **performance gains**
- Frankly, their model is still **simple**

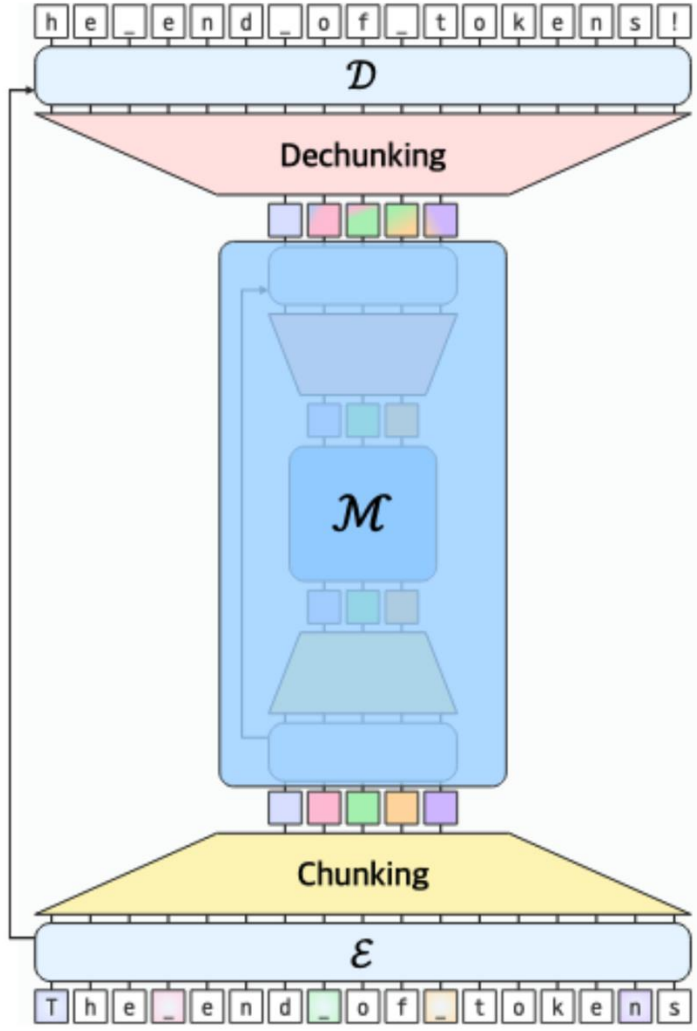
- Other ideas include:
  - Text processing evolution (attention -> mamba -> hyena -> stripedhyena)
  - Pooling modules (e.g. bag-of-mer)
  - Positional encodings (RoPE, ALiBi)
  - Algorithmic advances (e.g. alternating attention)

Hwang et al. 2025.
Nguyen et al. 2023.

# Better tokenizations

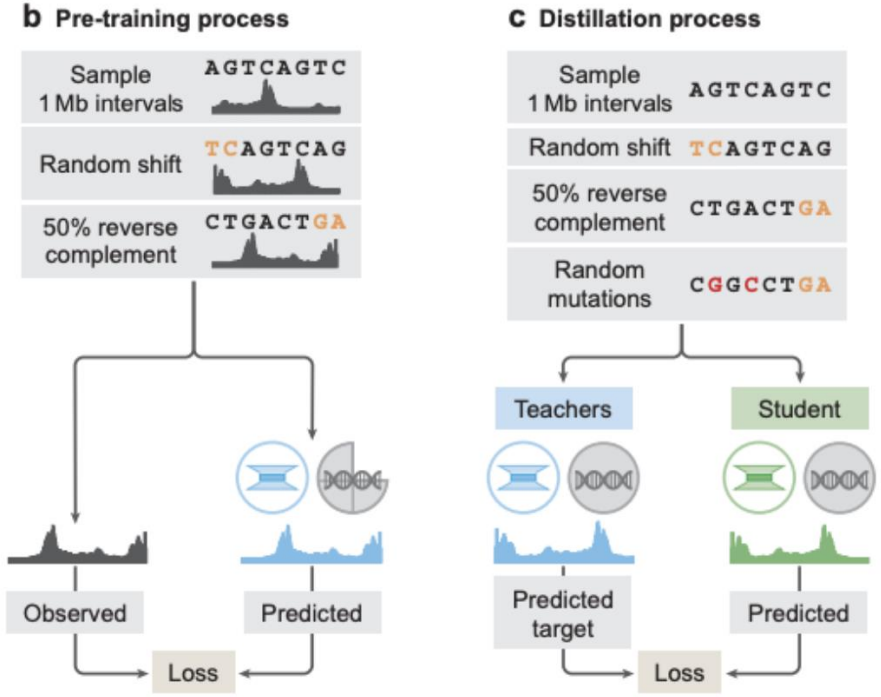
- How to **input** DNA sequences into the model?
- Kmer vs. single-nucleotide vs. byte-pair **tokens**
- **Unsupervised** tokenization (H-net)
- **Image-based** inputs (DeepSeek-OCR)

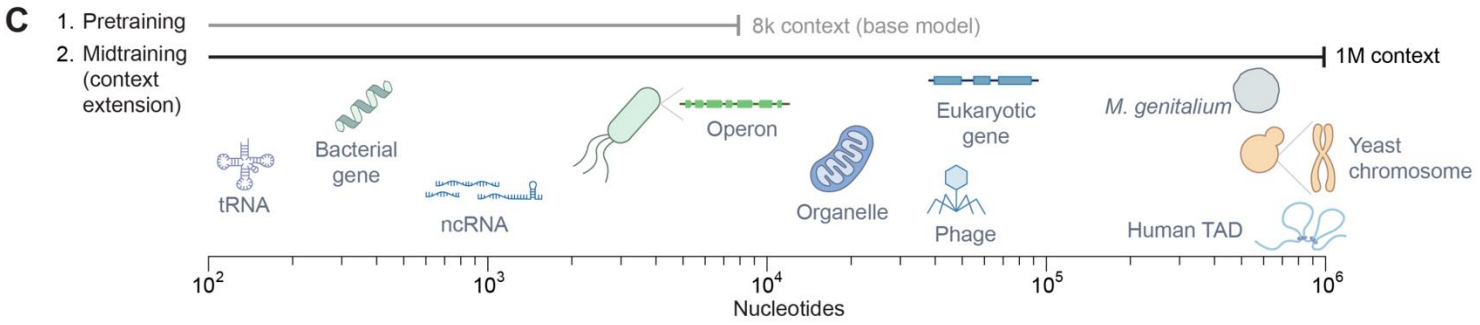


# Better training regimes

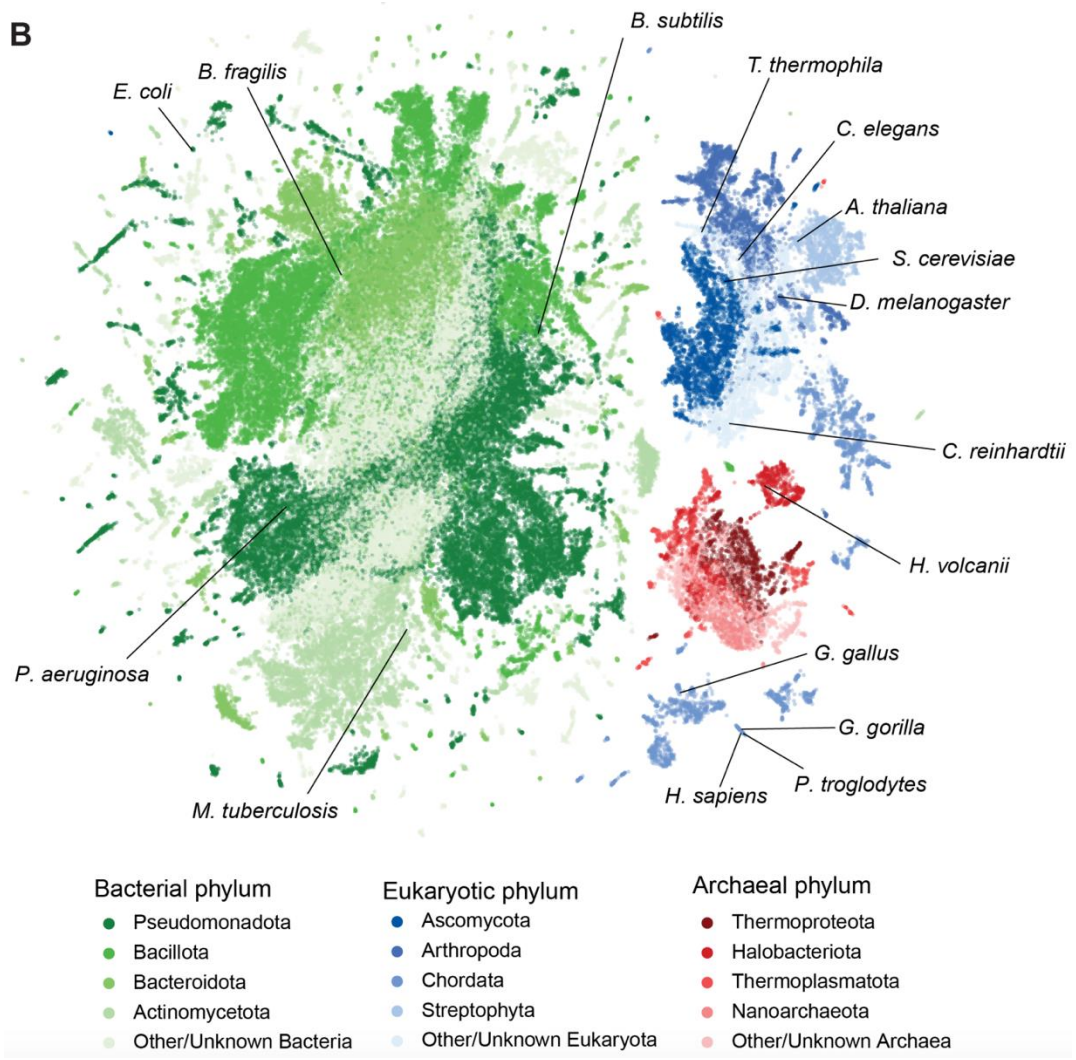
- More creative ways to **train** the model
- Evo-2 mostly trains on short sequences, then **extends** to long ones
- AlphaGenome trains a big **teacher** model, then trains small **student** models to **mimic** the teacher at **lower cost**



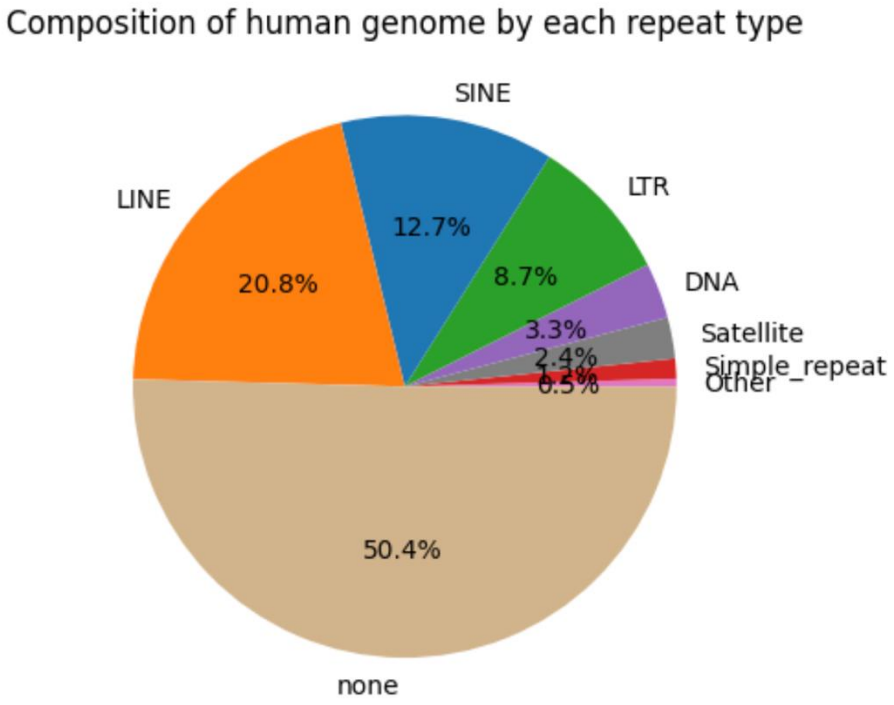



# Better training datasets

- Evo-2's training set (right) vs. the rest of the field (pick 1-2 blue dots)
- ChatGPT trains on **trillions** of samples, why shouldn't we?
- Self-supervised models can train on **more data**
  - Not limited to labeled data



# More efficient training datasets (my research!)




- Eukaryotic genomes are **repetitive**
- Repetitive data **worsens** LLMs

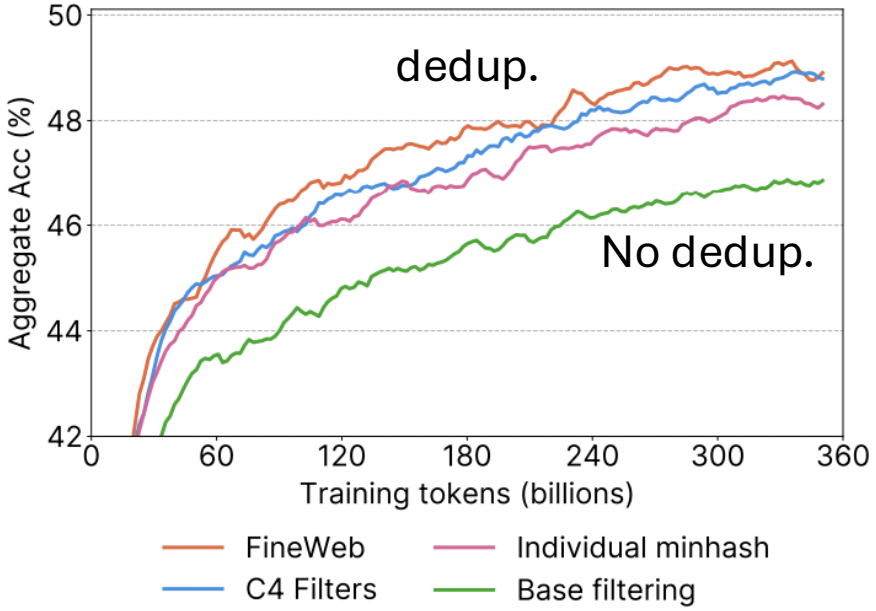
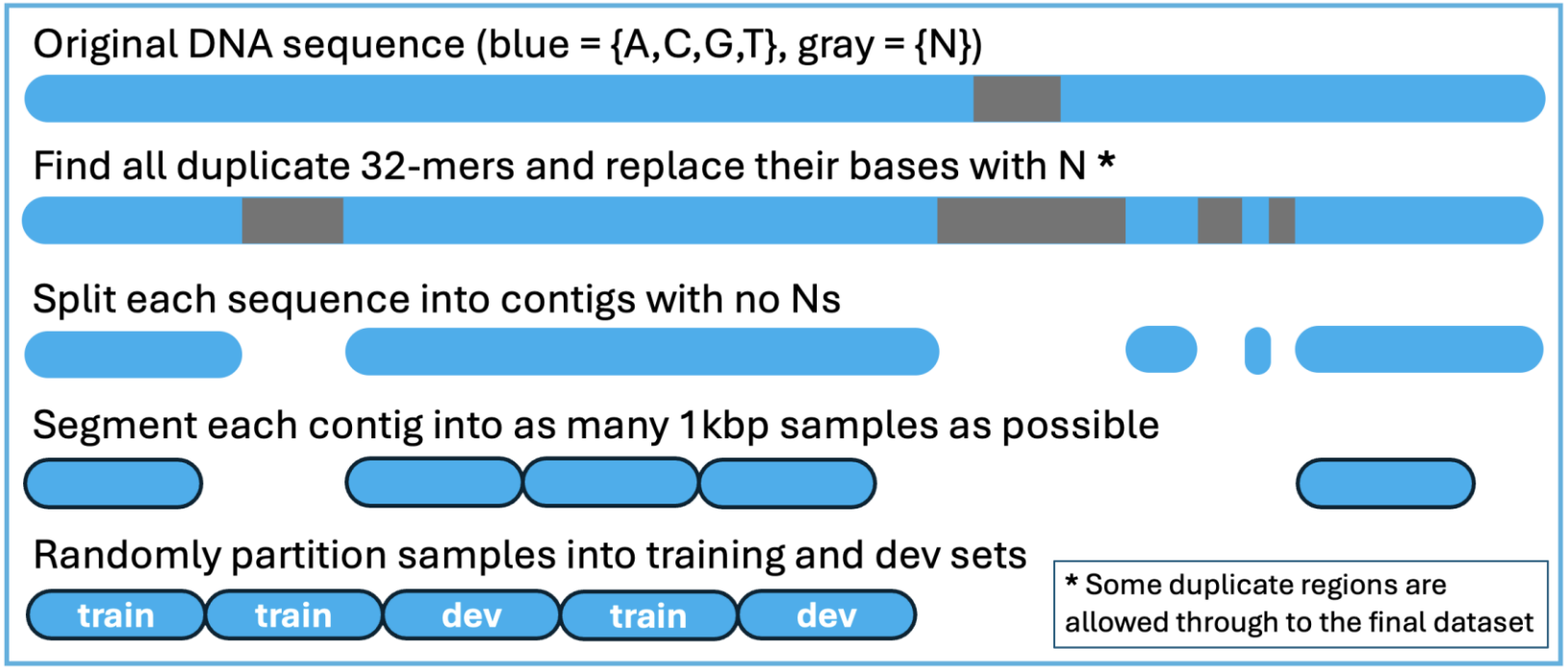


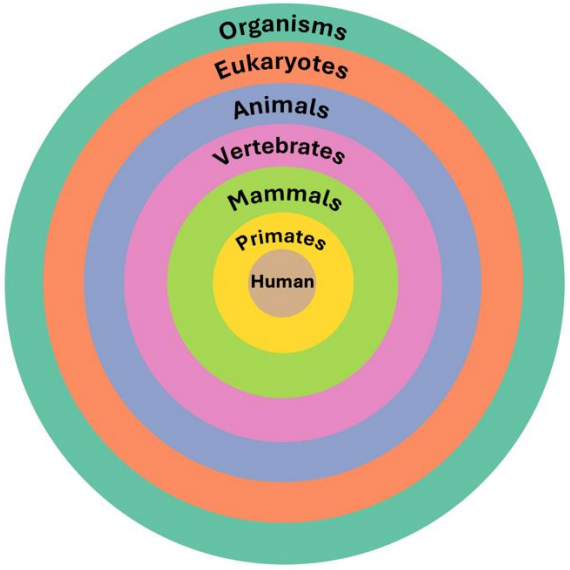

Figure 9: **Each processing step in FineWeb provides a performance uplift.** Compared to the base filtering (Section 3.3), applying individual-crawl MinHash deduplication (Section 3.4) the C4 filters (Section 3.5), and our additional heuristic filters (Section 3.6) each improve performance.

# My research: genome deduplication




- Developed a method to **deduplicate** genomic data
- Removes any repeated **32-base** sequences from a training dataset

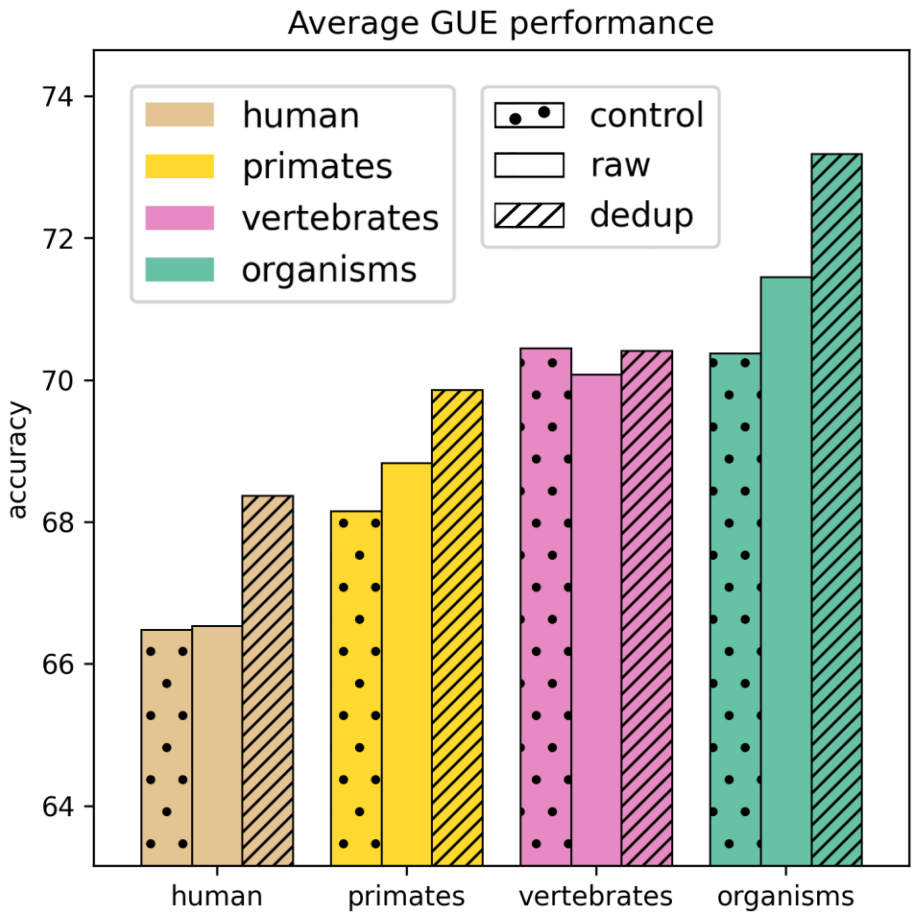
# My research: genome deduplication



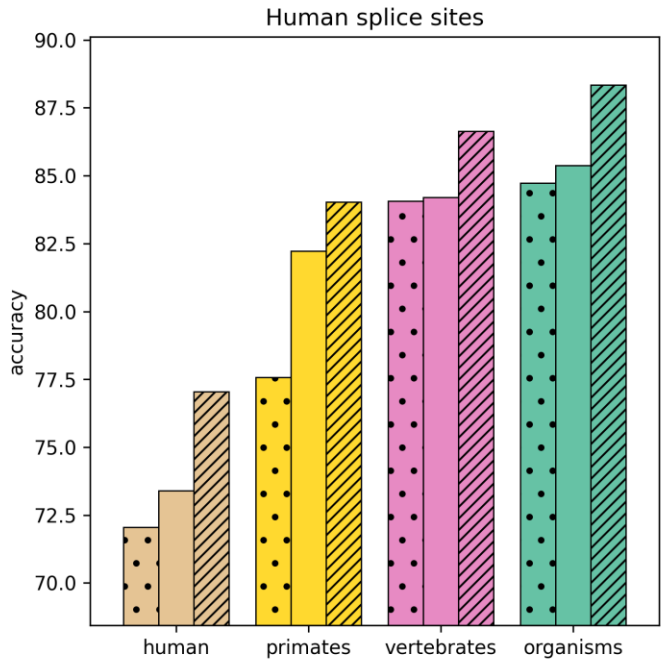

| Dataset | Genomes | Raw bases (B) | Dedup. bases (B) | % dedup. | Training steps (K) |
|---|---|---|---|---|---|
| **Human** | 1 | 3.0 | 1.5 | 50.0 % | 50 |
| **Primates** | 3 | 8.4 | 2.5 | 70.2 % | 100 |
| **Mammals** | 7 | 18.0 | 6.6 | 63.5 % | 300 |
| **Vertebrates** | 12 | 25.4 | 9.2 | 63.8 % | 400 |
| **Animals** | 21 | 29.9 | 10.2 | 65.7 % | 500 |
| **Eukaryotes** | 34 | 30.3 | 10.6 | 65.1 % | 500 |
| **Organisms** | 134 | 30.6 | 10.9 | 64.5 % | 500 |

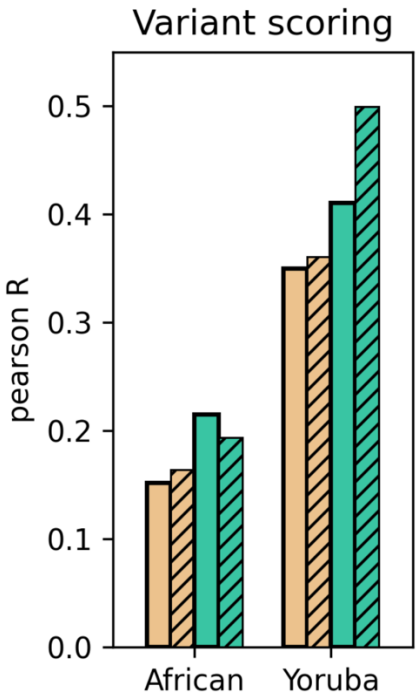
- Created **deduplicated datasets** for a variety of genome collections
- **Pretrained** a DNABERT-2-type model on each dataset
- **Fint-tuned** each model on **biologically-relevant** tasks

# My research: genome deduplication




Training on **deduplicated** data leads to better **performance** on biologically relevant tasks!



# Summary thoughts

- We have already seen **AI** and **deep learning** transform our world
- Speaks for itself that Large Language Models are powerful
- **DNA**, like English or Mandarin or Python, is a **language**
- LLMs may be able to transform our **understanding** of DNA

# Summary thoughts

- Genomic LLMs have shown great promise in understanding the **grammar** of DNA
- Already applicable to predicting **transcription factor** binding, **chromatin** accessibility, **splice sites**, and more
- Constantly improving at **variant** effect prediction
- Much **more work** to be done before the models are fully usable
- **Supervised** vs. **self-supervised** is the defining debate in the field

# What does a DNA variant do?

| Method | Drawback(s) |
|---|---|
| Query it in a database | The variant might not exist in a database<br>The variant might act differently under different conditions |
| Assume loss-of-function of nearby gene | Not always an accurate assumption<br>Function of nearby gene may not itself be understood |
| Genome-Wide Association Study (GWAS) | Not applicable for variants that only exist in a single person<br>Underpowered for rare variants that only occur in a few people<br>Relies on undersized datasets containing a small number of SNPs<br>Requires rigorous statistical testing<br>Can only model one variant at a time<br>Assumes additive linear effects of variants<br>Difficult to adapt to a polygenic/omnigenic model of traits<br>In practice, only explains a fraction of heritability |
| Deep learning | Computationally expensive<br>Not yet sophisticated enough to give satisfying answers to all questions |