Genomics Adapters

Immanuel Abdi, Matthew Mollerus, Jason Rudianto

Problem Statement

1. **Problem**: Hundreds of difficult tools used in research
2. **Target User**: Research Labs / Researchers
3. **Business Impact**: Higher productivity

Too many tools and processes!

- Genomics researchers use **niche**, **difficult** to use tools for different tasks
- Examples of how existing tools are used today – difficult command line instructions:

Preprocessing with Trimmomatic

```
trimmomatic SE
-phred33 sample.fastq
sample_trimmed.fastq
ILLUMINACLIP:TruSeq3-S
E.fa:2:30:10 LEADING:3
TRAILING:3
SLIDINGWINDOW:4:15
MINLEN:36
```

Alignment with samtools

```
samtools view -S -b
sample_aligned.sam >
sample_aligned.bam
samtools sort
sample_aligned.bam -o
sample_sorted.bam
picard MarkDuplicates
I=sample_sorted.bam
O=sample_dedup.bam
M=metrics.txt samtools
index sample_dedup.bam
```

Variant Calling with GATK

```
gatk HaplotypeCaller -R
reference.fasta -I sample_dedup.bam
-O raw_variants.vcf -ERC GVCF

gatk VariantFiltration -R
reference.fasta -V raw_variants.vcf
-O filtered_variants.vcf \

  --filter-name "QD_filter"
--filter-expression "QD < 2.0" \

  --filter-name "FS_filter"
--filter-expression "FS > 60.0"
```

Quotes

“I had to fly to Arizona for three days to learn how to use another tool”

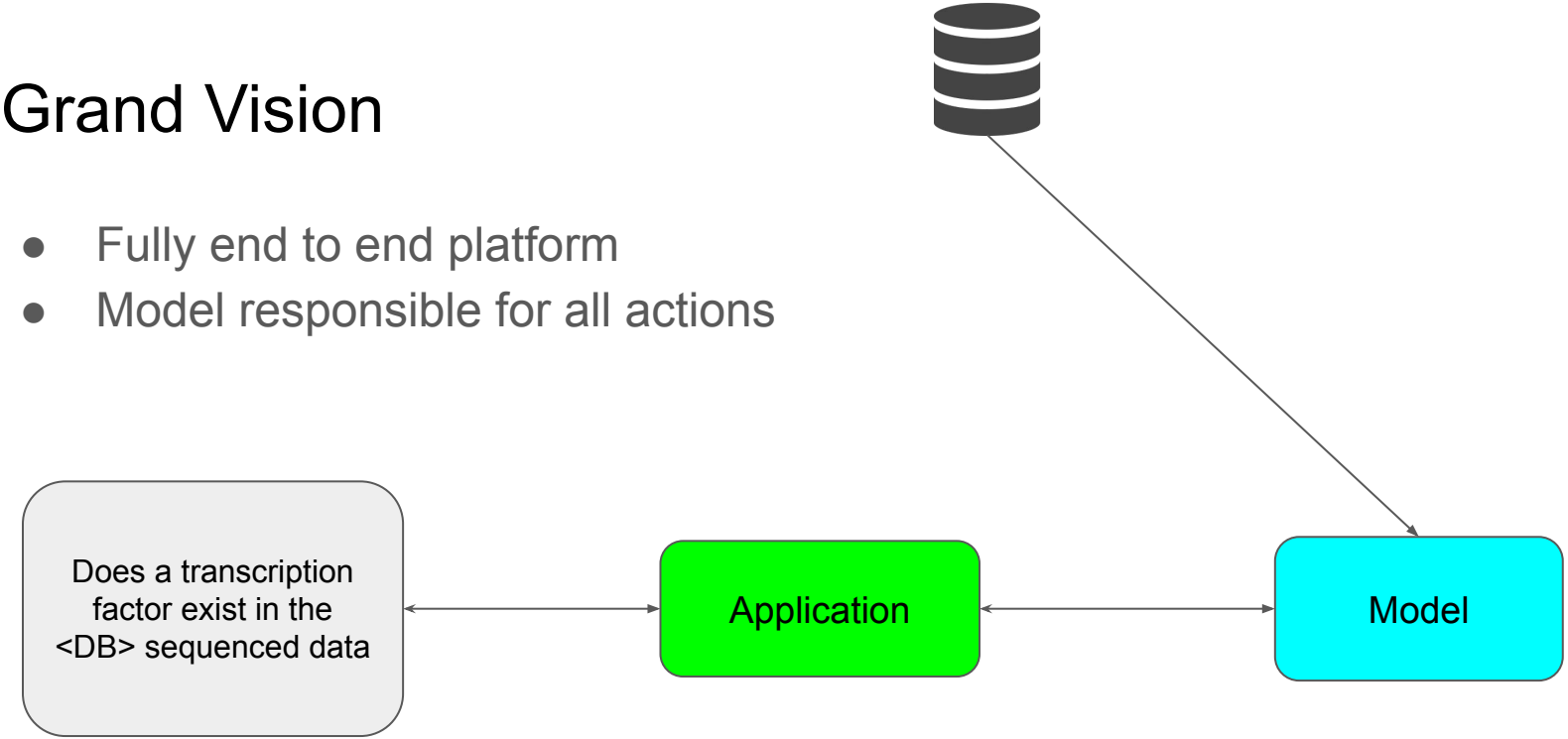
- *Microbiologist*

“The tool isn’t user-friendly; it takes me a couple of iterations to get it right”

- *Researcher*

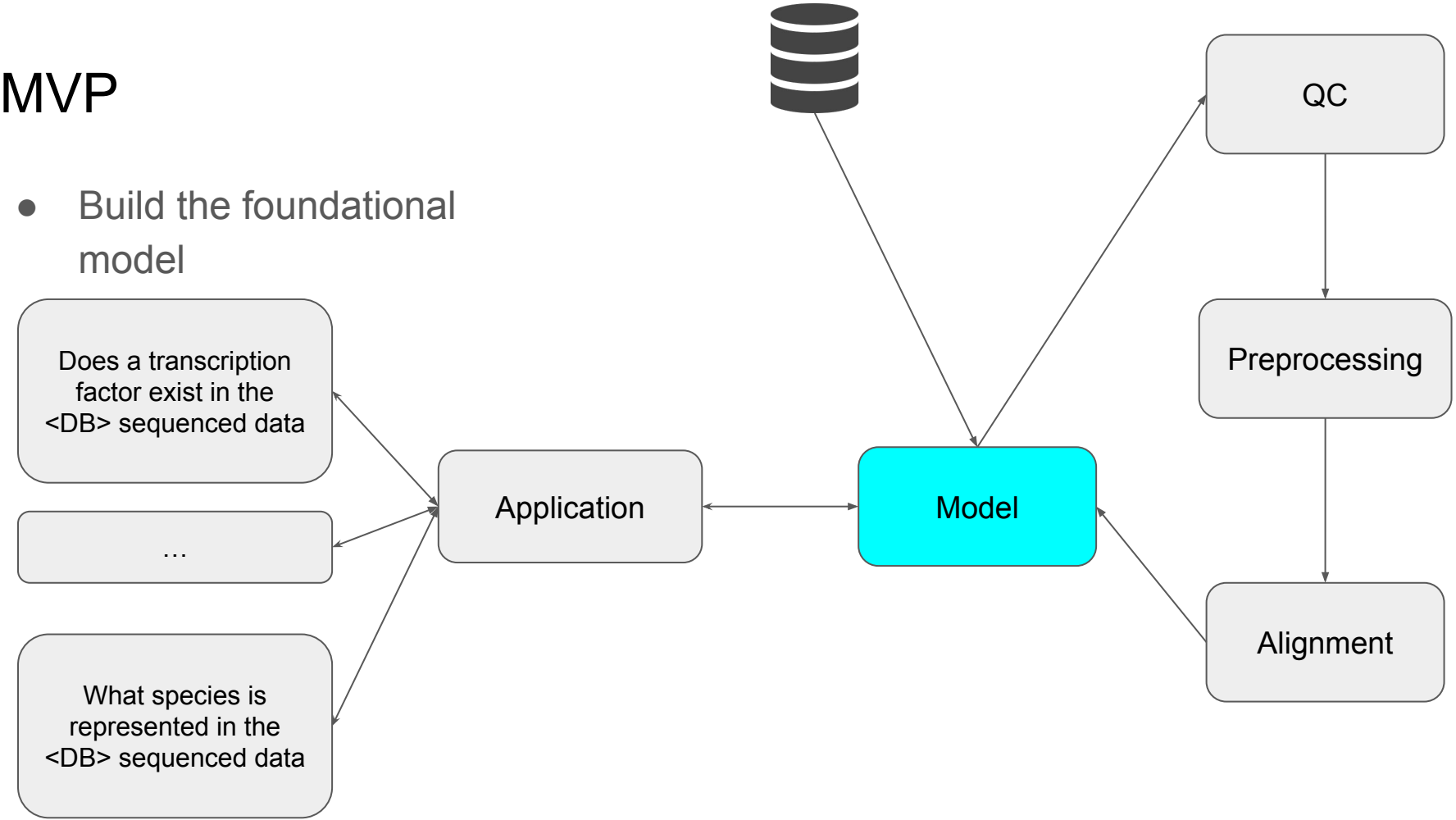
Grand Vision

- Fully end to end platform
- Model responsible for all actions



MVP

- Build the foundational model



MVP Key Features

- Downloadable and locally runnable Python Model
- Not commands! English
- Main pain point:
  - Make task execution easier

User Input Before

```
fastqc raw_sequence.fastq -o qc_results

java -jar trimmomatic-0.39.jar SE -phred33
raw_sequence.fastq
trimmed_sequence.fastq
ILLUMINACLIP:adapters.fa:2:30:10
LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36

kraken2 --db kraken_db
trimmed_sequence.fastq --output
kraken_output.txt --report kraken_report.txt
```


User Input with MVP

What species is represented in the sequenced data

How is it useful

- Model outputs in plain english, results of the task
  - Classification
  - Detection
- Fully replace niche command line tools

Dataset

- GUE (Genome Understanding Evaluation) Dataset

Dataset

- GUE (Genome Understanding Evaluation) Dataset
- Genomic sequences human, mouse, yeast, virus, fungus

Dataset Tasks

Task
Core Promoter Detection
Transcription Factor Pred
Promoter Detection
Splice Site Detection
Epigenetic Mark Prediction
Covid Variant Classification

Dataset Sequence Lengths

Task	Sequence Length
Core Promoter Detection	70
Transcription Factor Pred	100
Promoter Detection	300
Splice Site Detection	400
Epigenetic Mark Prediction	500
Covid Variant Classification	1000

- Ranges from 70 to 1000 base pairs

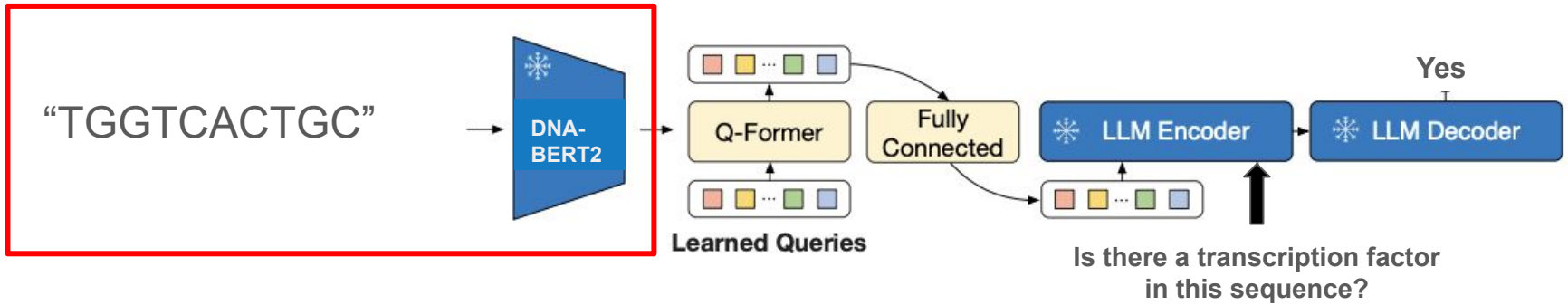
Dataset Examples

Task	Sequence Length	Class examples
Core Promoter Detection	70	True, False
Transcription Factor Pred	100	True, False
Promoter Detection	300	True, False
Splice Site Detection	400	Donor, Acceptor, Neither
Epigenetic Mark Prediction	500	True, False
Covid Variant Classification	1000	Alpha, Beta, Delta, Eta, Gamma, Iota, Kappa, Lambda, Zeta

- 4 Binary classification
- 2 Multi-class classification

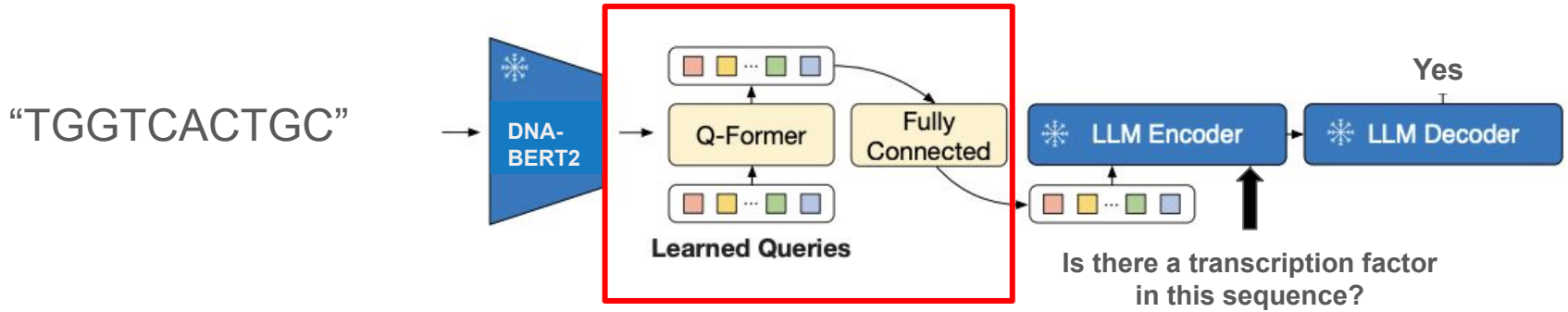
Model Architecture (Broad Overview)

- Feed DNA sequence into a frozen pre-trained DNA model



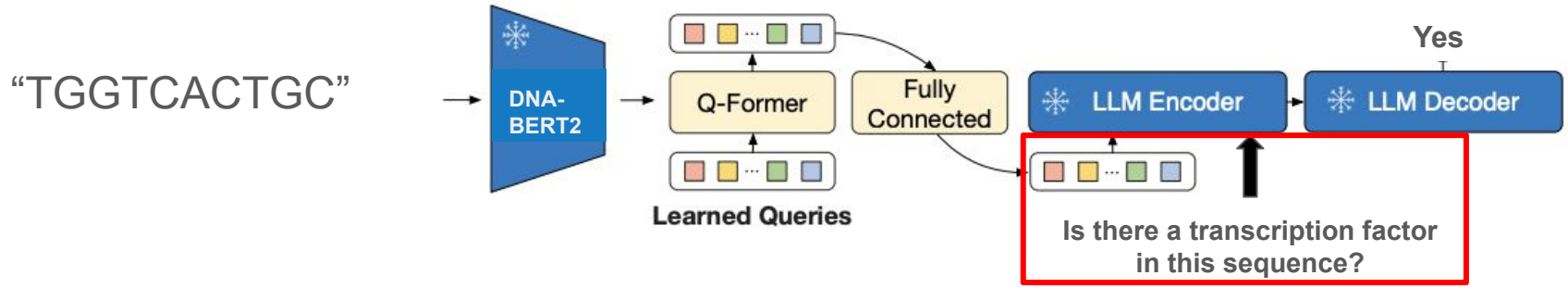
Model Architecture (Broad Overview)

- Feed DNA sequence into a frozen pre-trained DNA model
- Transform DNA output embeddings into something readable by an LLM



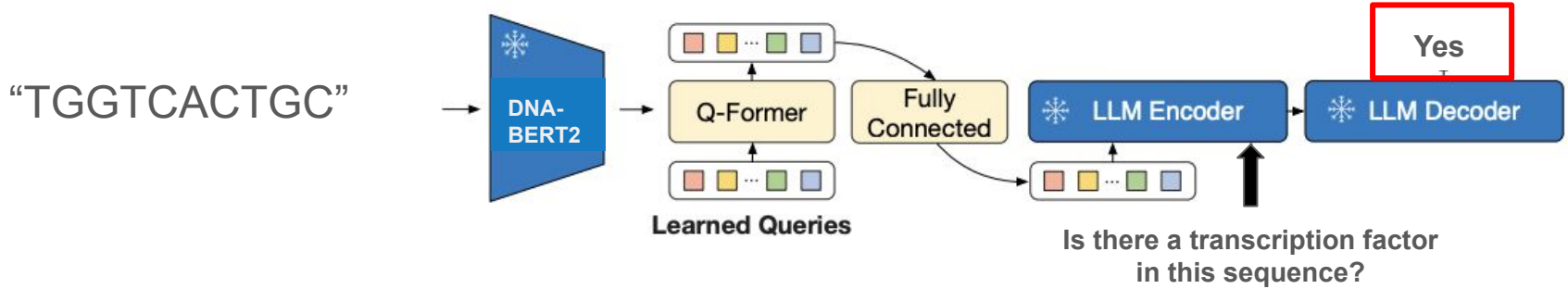
Model Architecture (Broad Overview)

- Feed DNA sequence into a frozen pre-trained DNA model
- Transform DNA output embeddings into something readable by an LLM
- Feed DNA embedding, with text embeddings into a frozen LLM



Model Architecture (Broad Overview)

- Feed DNA sequence into a frozen pre-trained DNA model
- Transform DNA output embeddings into something readable by an LLM
- Feed DNA embedding, with text embeddings into a frozen LLM
- Obtain an answer

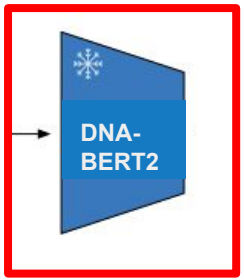


Demo

Important parts of architecture

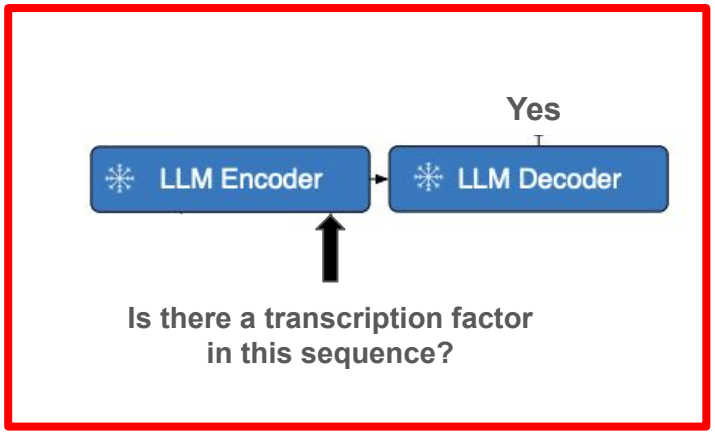
- DNA Encoder (DNA-BERT2)

“TGGTCACTGC”



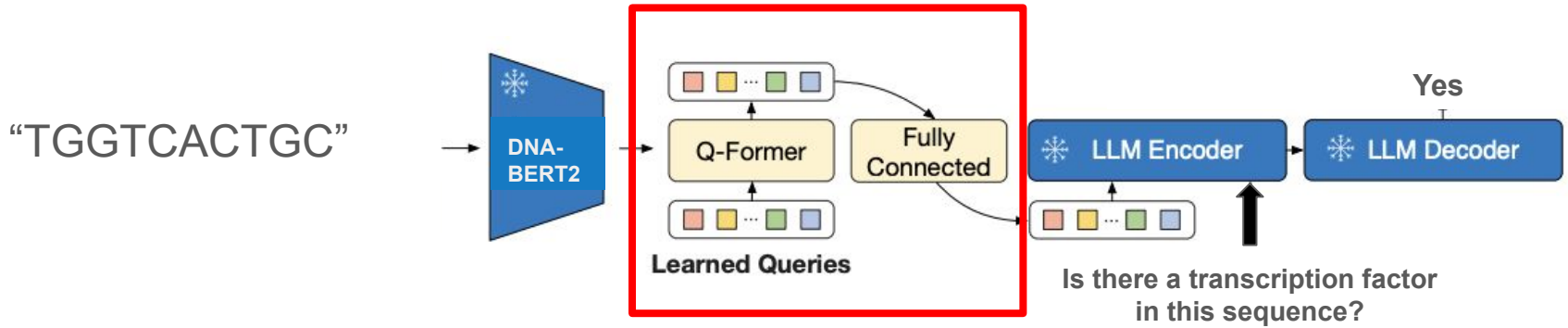
Important parts of architecture

- DNA Encoder (DNA-BERT2)
- LLM (GPT-2XL)



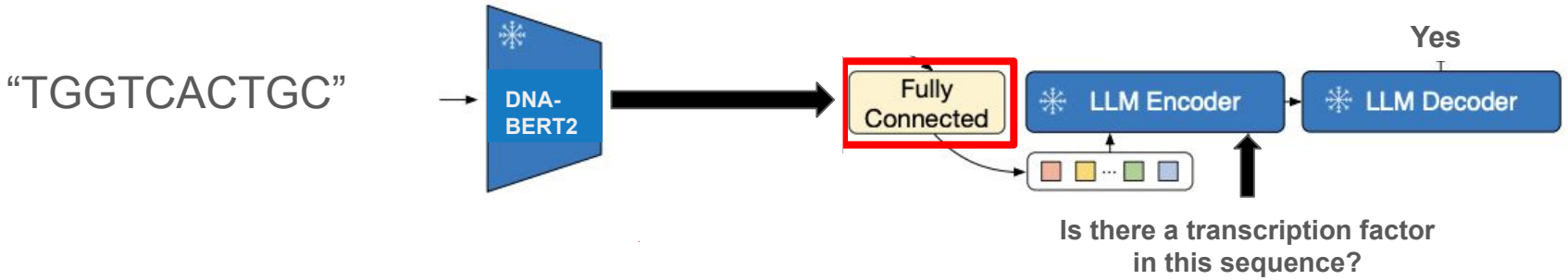
Important parts of architecture

- DNA Encoder (DNA-BERT2)
- LLM (GPT-2XL)
- Querying Transformer (Q-Former)



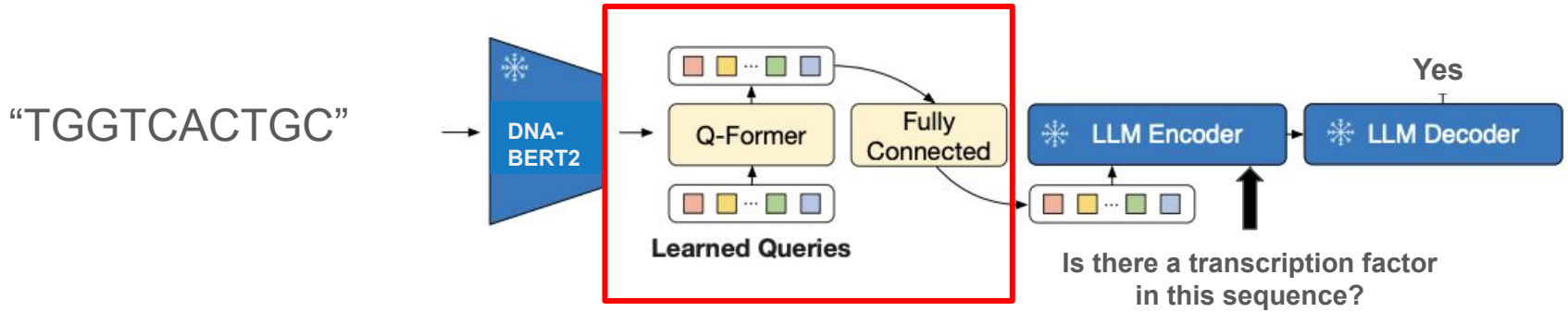
Important parts of architecture

- DNABERT2
- LLM
- Linear Projection



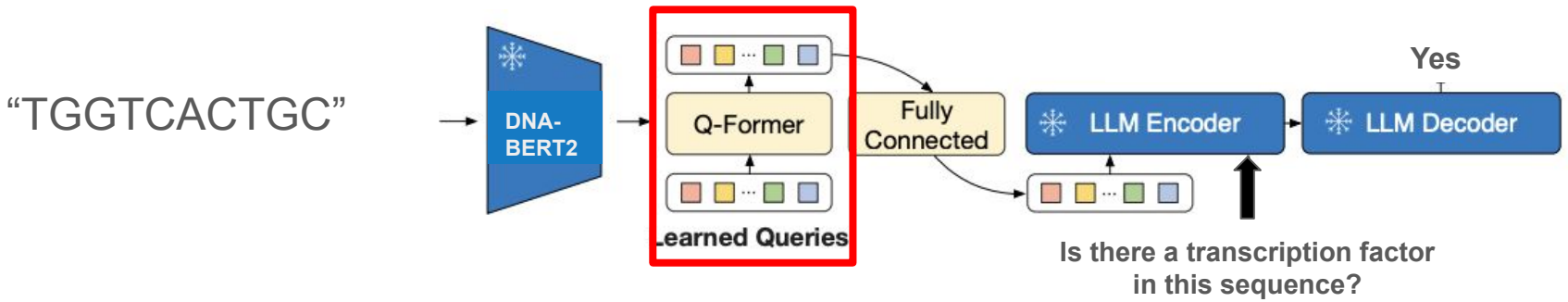
Training the model

- Freeze all of the big parts with lots of parameters
- Option 1: only train parts that transform DNA embeddings into something the language model understands
- Option 2: Option 1 plus fine tune DNABERT2 with LORA

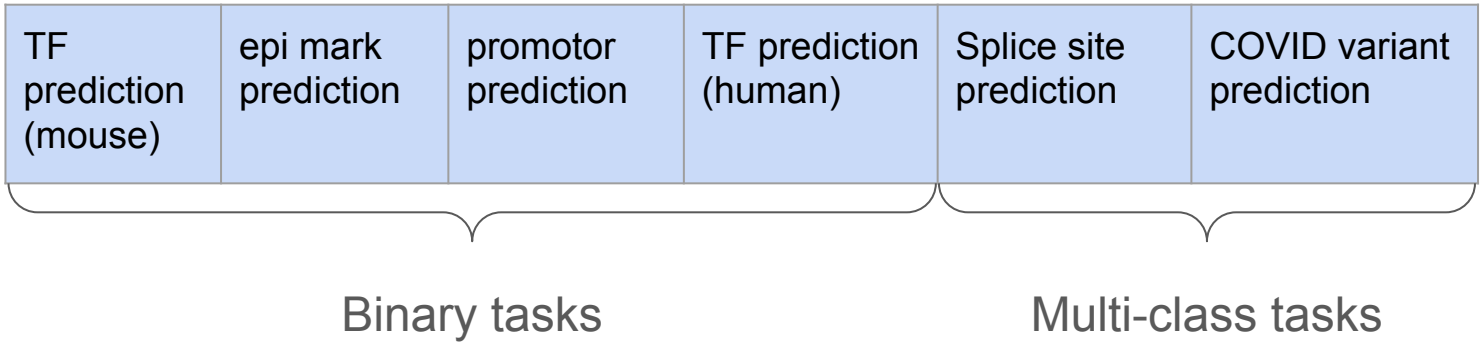


Models of interest

- DNABERT2
- LLM
- DNABERT2 -> Linear Projection -> LLM (Only Projection trained)
- DNABERT2 -> Linear Projection -> LLM (DNABERT2 trained with LORA)
- DNABERT2 -> Linear Projection -> Q-Former -> LLM (Q-Former trained)

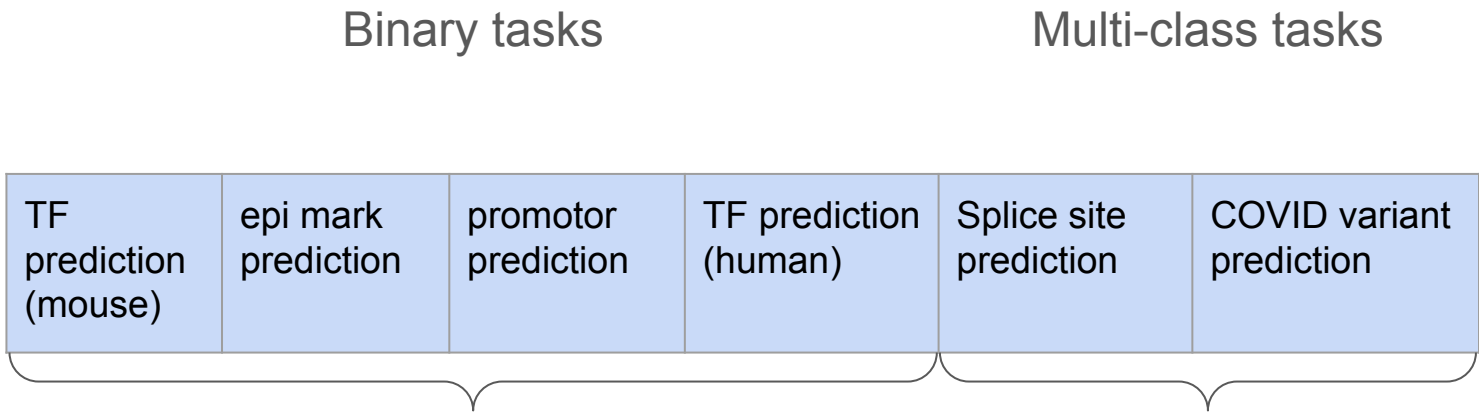


Evaluation tasks



Evaluation tasks

- Evaluate matthews correlation/F1
- 0 is bad, 100 is best



Evaluation Baselines

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
GPT2-XL	0	0	0	0	0	0
Linear projection	0	0	0	0	0	0

GPT2 can’t understand DNA, neither can a simple linear projection!

Evaluating Q-former and LORA

- We can transfer binary knowledge from a DNA encoder to an LLM

Evaluating Q-former and LORA

- We can transfer binary knowledge from a DNA encoder to an LLM

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99
Q-former	0	63.70	44.41	42.84	0
LORA	0	71.60	25.30	52.14	74.06

Evaluating Q-former and LORA

- We can transfer binary knowledge from a DNA encoder to an LLM
- Q-former is good for quick understanding on single tasks

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99
Q-former	0	63.70	44.41	42.84	0
LORA	0	71.60	25.30	52.14	74.06

Evaluating Q-former and LORA

- We can transfer binary knowledge from a DNA encoder to an LLM
- Q-former is good for quick understanding on single tasks
- PEFT allows for deeper understanding on some tasks

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99
Q-former	0	63.70	44.41	42.84	0
LORA	0	71.60	25.30	52.14	74.06

Evaluating Q-former and LORA

- We can transfer binary knowledge from a DNA encoder to an LLM
- Q-former is good for quick understanding on single tasks
- PEFT allows for deeper understanding on some tasks
- It is hard to teach an LLM DNA multi-class tasks

	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	84.99	71.02
Q-former	0	2.38
LORA	74.06	2.38

Evaluating joint task training

- Training with Q-Former results in complete collapse of genomic knowledge

Evaluating joint task training

- Training with Q-Former results in complete collapse of genomic knowledge
- Training with LORA multi-task gives near encoder level performance

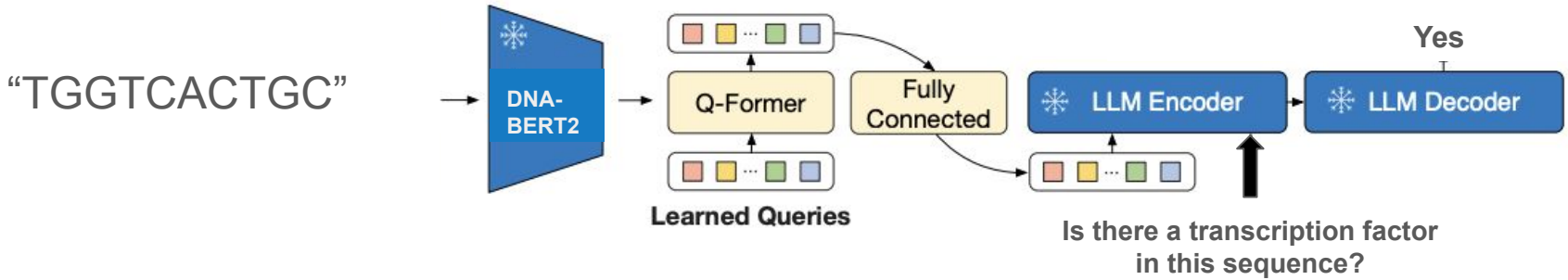
Evaluating joint task training: LORA

- Training with Q-Former results in complete collapse of genomic knowledge
- Training with LORA multi-task gives near encoder level performance

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
LORA	0	71.60	25.30	52.14	74.06	2.38
LORA binary	22.70	70.96	85.03	38.38	NA	NA

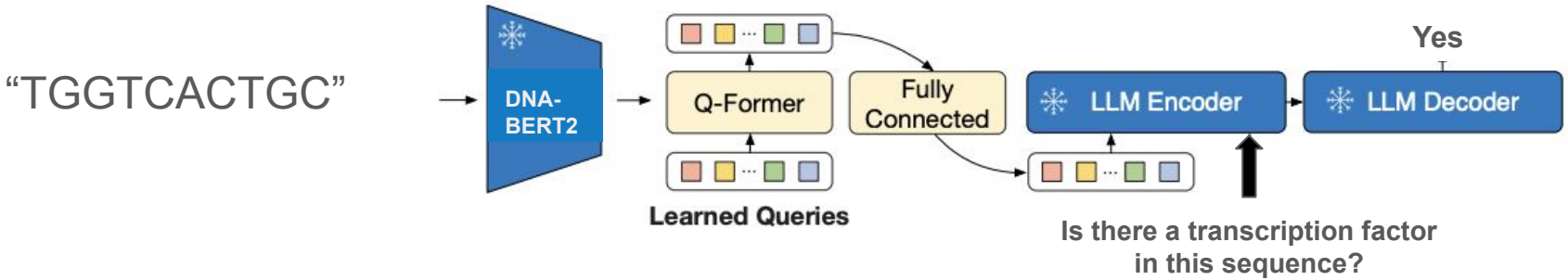
Technical Takeaway and Challenges

- Bootstrap and transfer knowledge from DNA encoders to LLMS



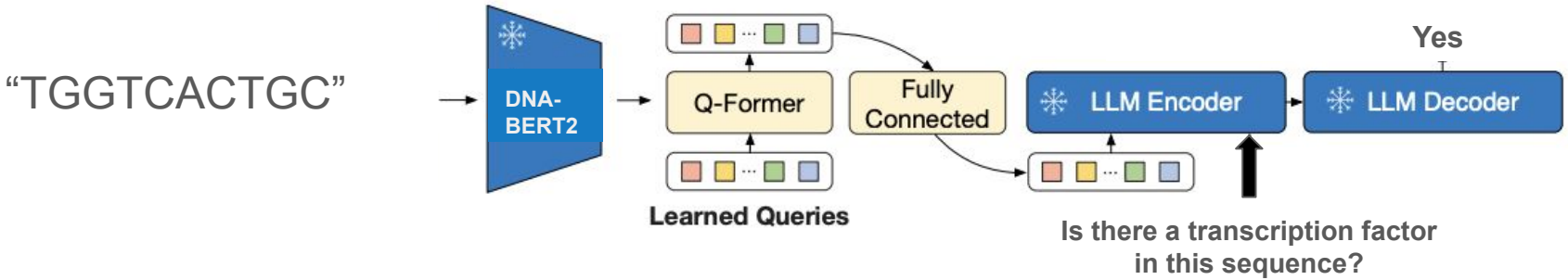
Technical Takeaway and Challenges

- Bootstrap and transfer knowledge from DNA encoders to LLMS
- Adapt multi-modal vision architectures to do genomics tasks



Technical Takeaway and Challenges

- Bootstrap and transfer knowledge from DNA encoders to LLMS
- Adapt multi-modal vision architectures to do genomics tasks
- Single model to do many genomics tasks



Future Roadmap

- Train on more data
- Utilize different adapter models to allow for multi-class prediction
- Train with varying inputs for each task

Simplifying the research process by making it easier

Acknowledgements

- Artemis Pangopoulou (Wrote X-InstructBLIP, helped with getting our code running)
- Interview candidates in the Bio-space

Resources

- Full Summary: https://www.ischool.berkeley.edu/projects/2024/genomics-adapters
- Informational website: https://sites.google.com/berkeley.edu/genomics-adapters/home
- Try out our model: https://huggingface.co/immanuelabdi/GenomeLanguageMultiModalModel
- See our code: https://github.com/immanuelazn/dna-llm-adapters

Appendix

Evaluation Baselines

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
GPT2-XL	0	0	0	0	0	0
Linear projection	0	0	0	0	0	0

Evaluating Q-former and LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
LORA	0	71.60	25.30	52.14	74.06	2.38

Evaluating Q-former and LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
LORA	0	71.60	25.30	52.14	74.06	2.38

- Using Q-Former/LORA allows the LLM to understand DNA sequences

Evaluating Q-former and LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
LORA	0	71.60	25.30	52.14	74.06	2.38

- Using Q-Former/LORA allows the LLM to understand DNA sequences
- The hardest tasks seem to be multi-class, and TF prediction on mice

Evaluating Q-former and LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
LORA	0	71.60	25.30	52.14	74.06	2.38

- Using Q-Former/LORA allows the LLM to understand DNA sequences
- The hardest tasks seem to be multi-class, and TF prediction on mice
- LORA has a generally better understanding than Q-former

Evaluating joint task training

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
LORA	0	71.60	25.30	52.14	74.06	2.38

Binary tasks

Multi-class tasks

Evaluating joint task training: Q-former

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
Q-former binary	0	0	0	0	NA	NA

Evaluating joint task training: Q-former

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
Q-former binary	0	0	0	0	NA	NA

- Q-former arch can’t understand when given multiple tasks for training

Evaluating joint task training: Q-former

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
Q-former	0	63.70	44.41	42.84	0	2.38
Q-former binary	0	0	0	0	NA	NA

- Q-former arch can’t understand when given multiple tasks for training
- Struggles to get past learning english syntax, to start learning the DNA representation

Evaluating joint task training: LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
LORA	0	71.60	25.30	52.14	74.06	2.38
LORA binary	22.70	70.96	85.03	38.38	NA	NA

Evaluating joint task training: LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
LORA	0	71.60	25.30	52.14	74.06	2.38
LORA binary	22.70	70.96	85.03	38.38	NA	NA

- Training PEFT multi-task gives near-encoder level metrics on some tasks

Evaluating Q-former and LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
LORA	0	71.60	25.30	52.14	74.06	2.38
LORA binary	22.70	70.96	85.03	38.38	NA	NA

- Training PEFT multi-class gives near-encoder level metrics on some tasks
- First decoder model to understand TF prediction in mice

Evaluating joint task training: LORA

	TF prediction (mouse)	epi mark prediction	promotor prediction	TF prediction (human)	Splice site prediction	COVID variant prediction
DNABERT2 fine-tuned	56.76	80.17	86.77	71.99	84.99	71.02
LORA	0	71.60	25.30	52.14	74.06	2.38
LORA binary	22.70	70.96	85.03	38.38	NA	NA

- Training PEFT multi-class gives near-encoder level metrics on some tasks
- First decoder model to understand TF prediction in mice
- Near encoder level results on promotor prediction