

Decoding the Genomic Language:

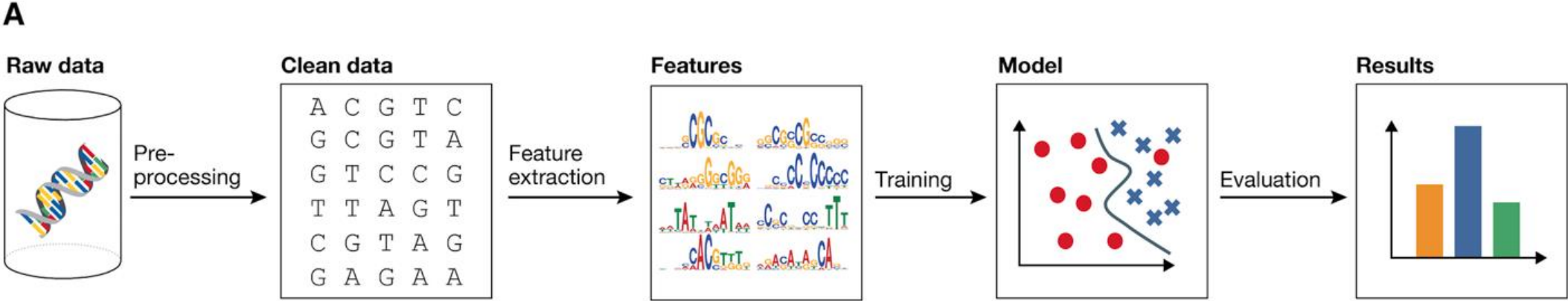
Fine-tuning LLMs for genomic downstream tasks

Fabiana Góes

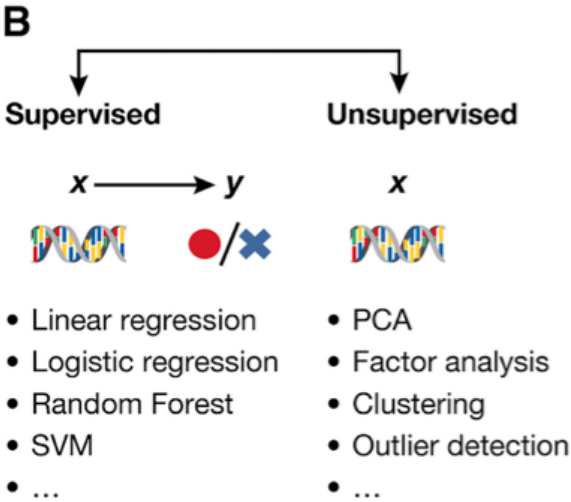
The Rosalind Franklin Institute
Artificial intelligence Group
E-mail: fabiana.goes@rfi.ac.uk

www.rfi.ac.uk

Machine Learning for Computational Biology

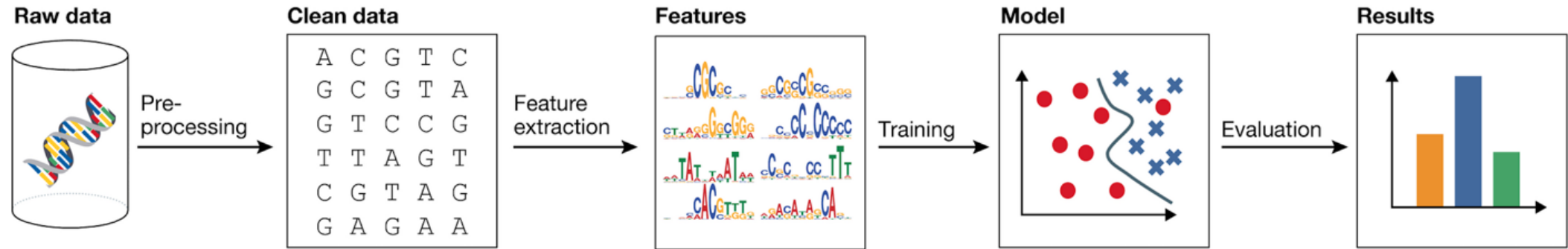


Machine Learning for Computational Biology

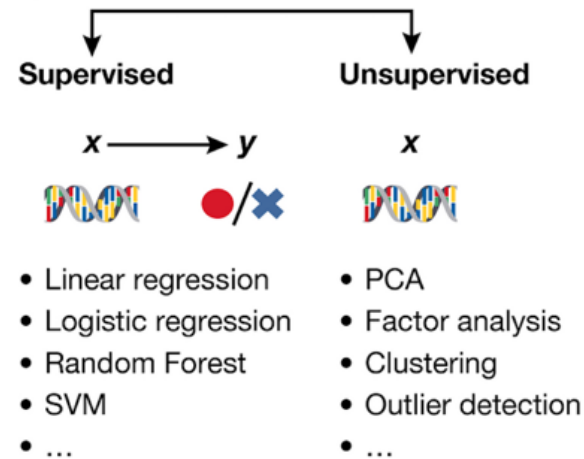


Machine Learning for Computational Biology

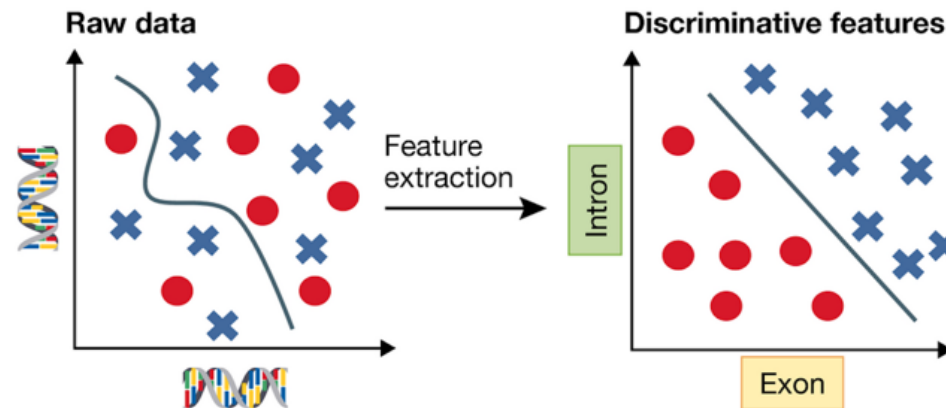
A



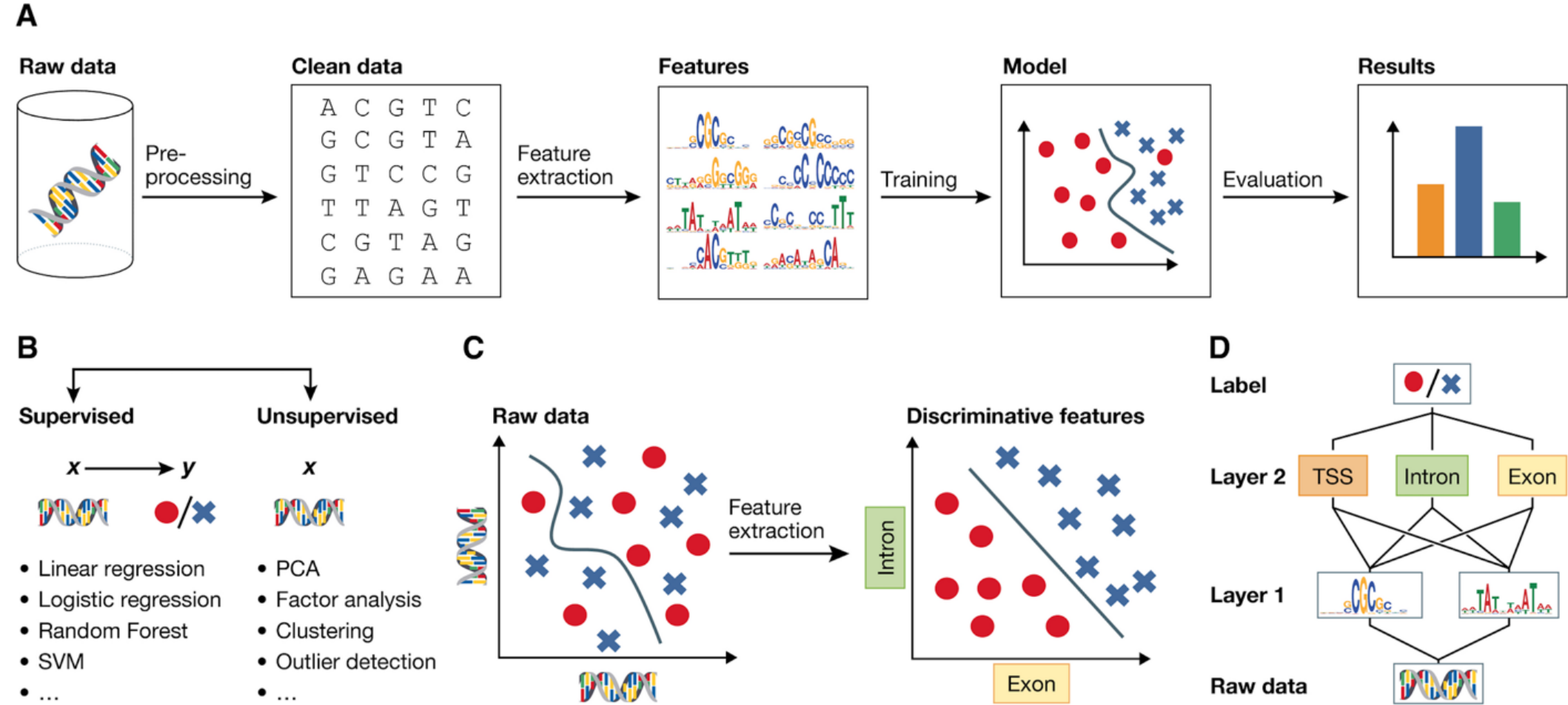
B



C



Machine Learning for Computational Biology



Natural Language Processing - Representation Learning

- **Representation learning**

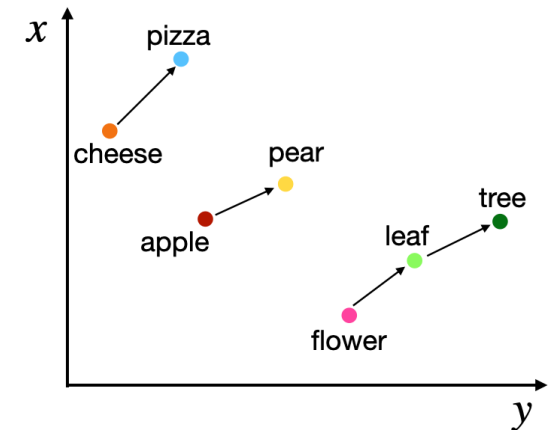
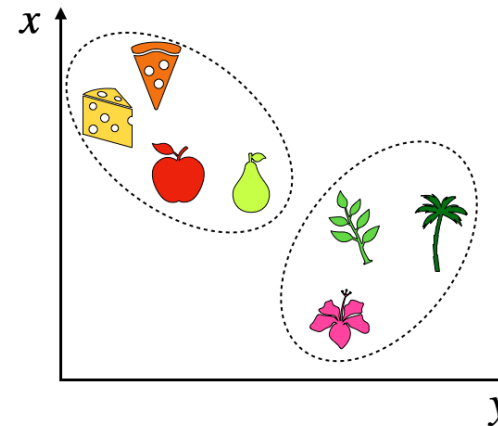
- Models to extract representation from raw data
- Convert words into vectors, preserving their semantic similarity
- The performance of ML depends on the quality of the representation
- Neural networks specialized for **understanding** the **relationships** among **words**

- Embedding models:

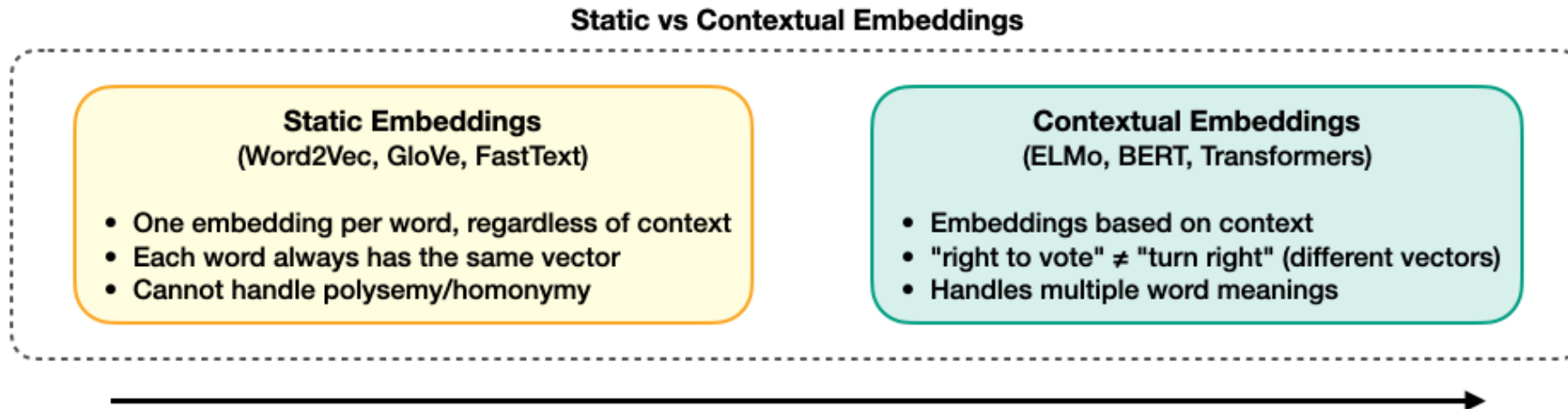
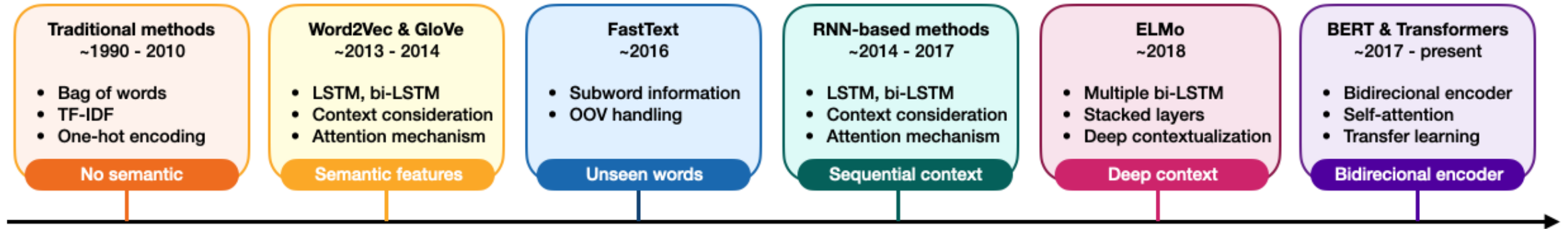
- Transform sequences into numerical vectors capturing patterns
- Assumption: similar words should have similar vector
- High vector similarity between proteins -> similar functions and structures

- Model types:

- **Word2Vec**, Doc2Vec
- CNNs, RNNs
- **Transformer-based architectures (LLMs)**



NLP - Evolution of Word Representation Learning



NLP – Deep language models

A Pretraining



Large corpus
(unlabeled text)

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get *somewhere*," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Original text

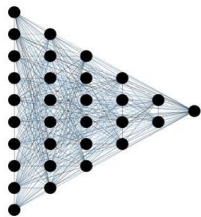
Masking



"Would you tell me, [REDACTED], which way I [REDACTED] to go from here?"
"That [REDACTED] a [REDACTED] deal on where you want to get to," said the Cat.
"I [REDACTED] much care where—" [REDACTED] Alice.
"Then it doesn't matter [REDACTED] [REDACTED] you go," said the Cat.
"—so long as I get *somewhere*," Alice [REDACTED] as an explanation.
"Oh, [REDACTED] [REDACTED] to do that," said the Cat, "if [REDACTED] only [REDACTED] long enough."

Masked text

Language model



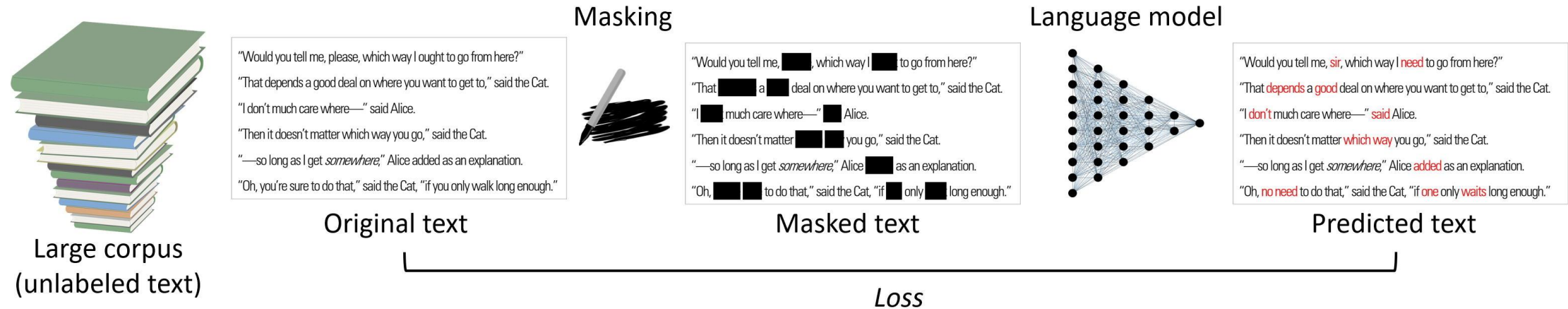
"Would you tell me, *sir*, which way I *need* to go from here?"
"That *depends* a *good* deal on where you want to get to," said the Cat.
"I *don't* much care where—" *said* Alice.
"Then it doesn't matter *which way* you go," said the Cat.
"—so long as I get *somewhere*," Alice *added* as an explanation.
"Oh, *no need* to do that," said the Cat, "if *one* only *waits* long enough."

Predicted text

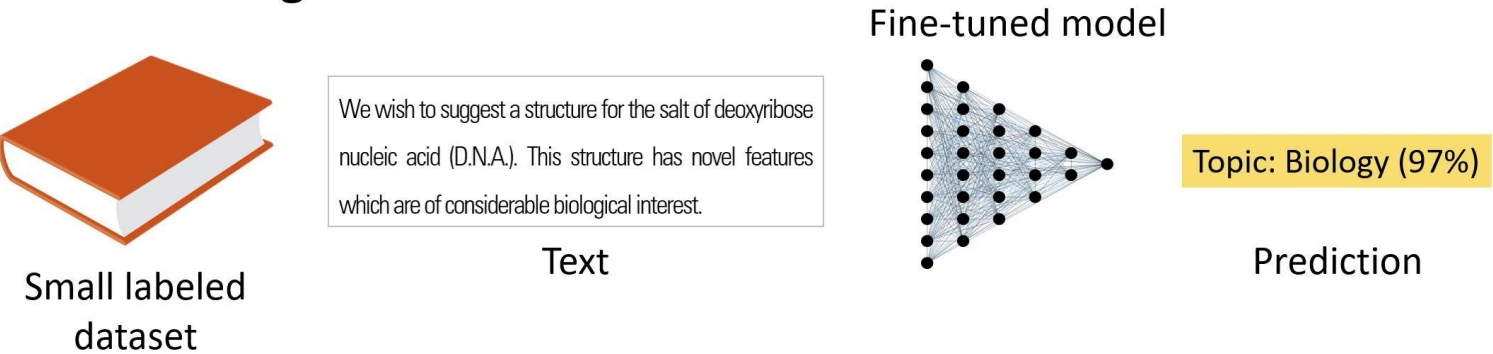
Loss

NLP – Deep language models

A Pretraining

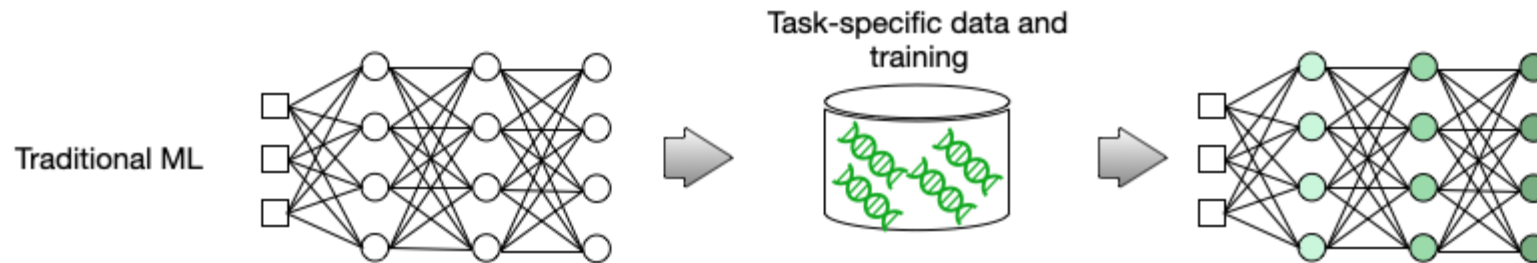


B Fine-tuning



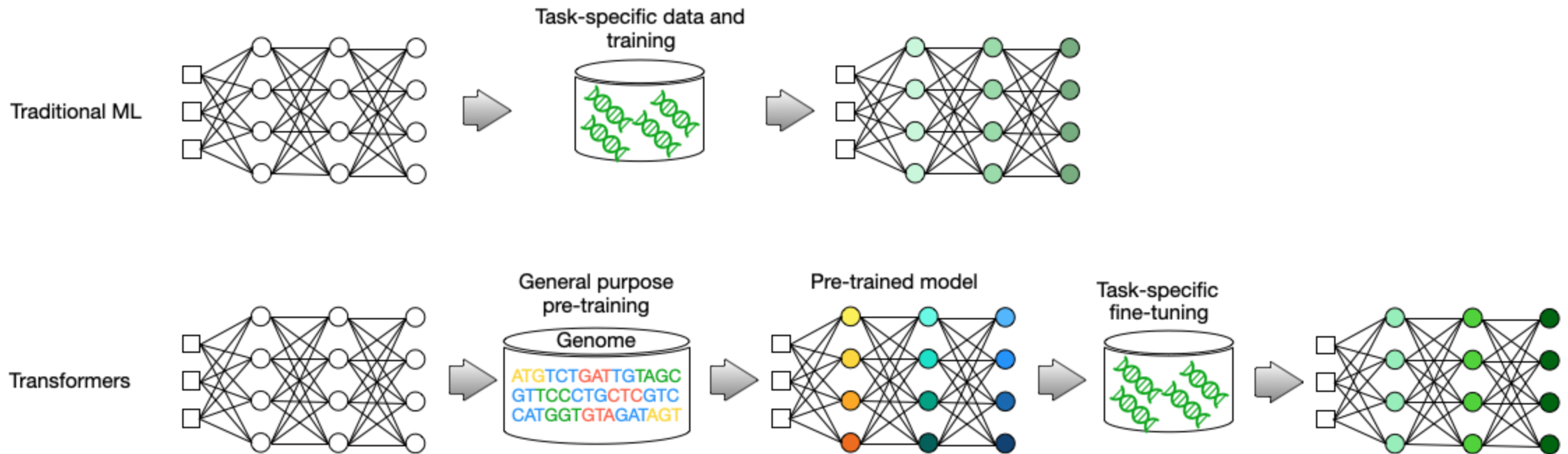
Large Language Models in Genomics

- Learning from vast amounts of unlabeled data
- Capture deep biological patterns, long-range dependencies, and structural information
- Strong capacities to map inputs into latent embedding space
- Can be adapted to a wide range of tasks

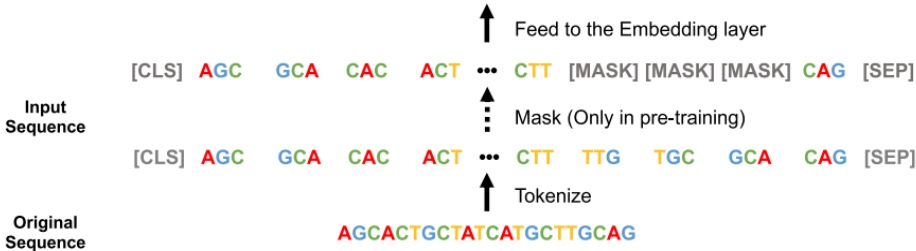


Large Language Models in Genomics

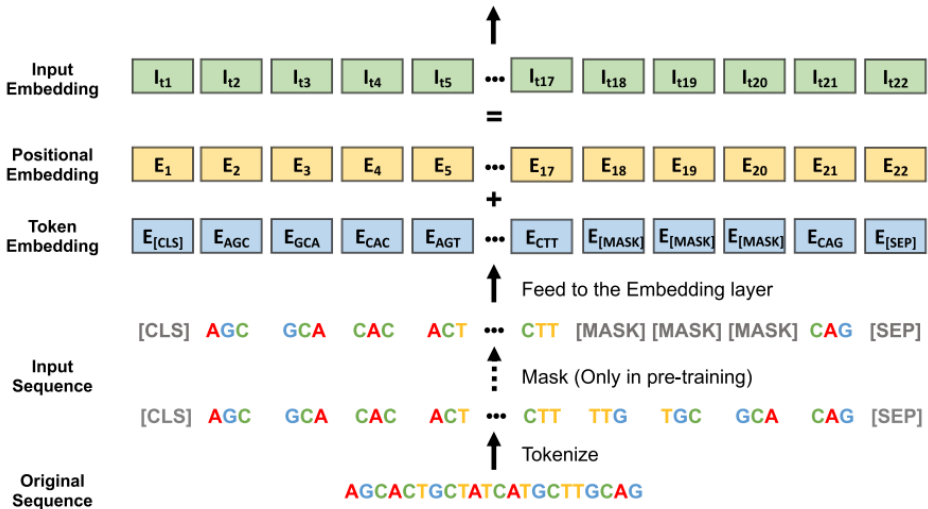
- Learning from vast amounts of unlabeled data
- Capture deep biological patterns, long-range dependencies, and structural information
- Strong capacities to map inputs into latent embedding space
- Can be adapted to a wide range of tasks



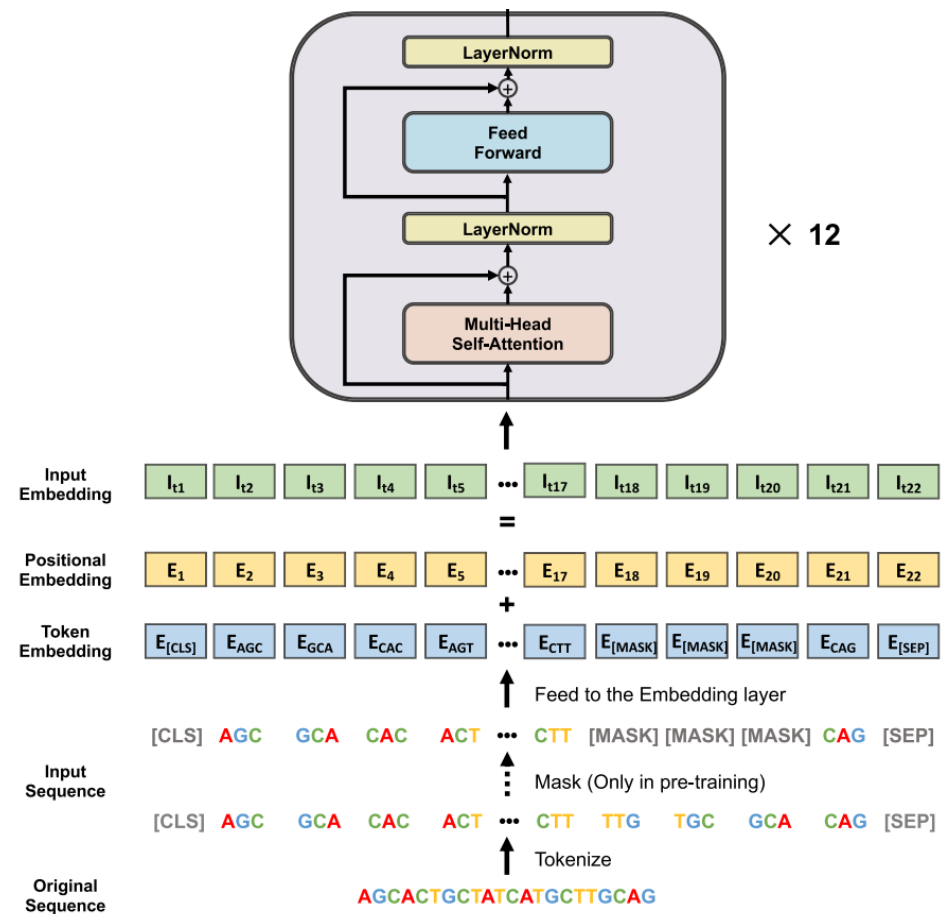
DNABERT – BERT model to DNA



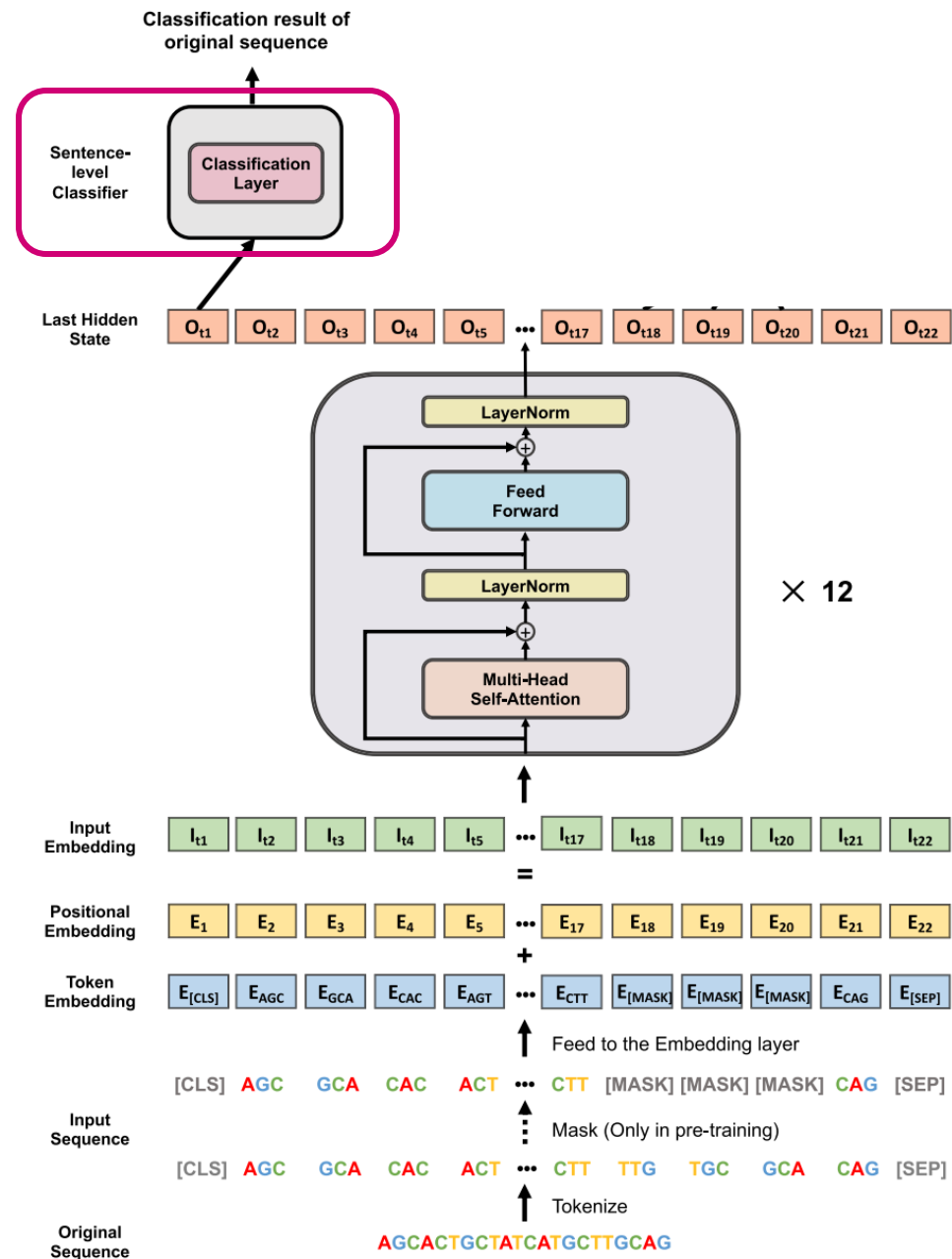
DNABERT – BERT model to DNA



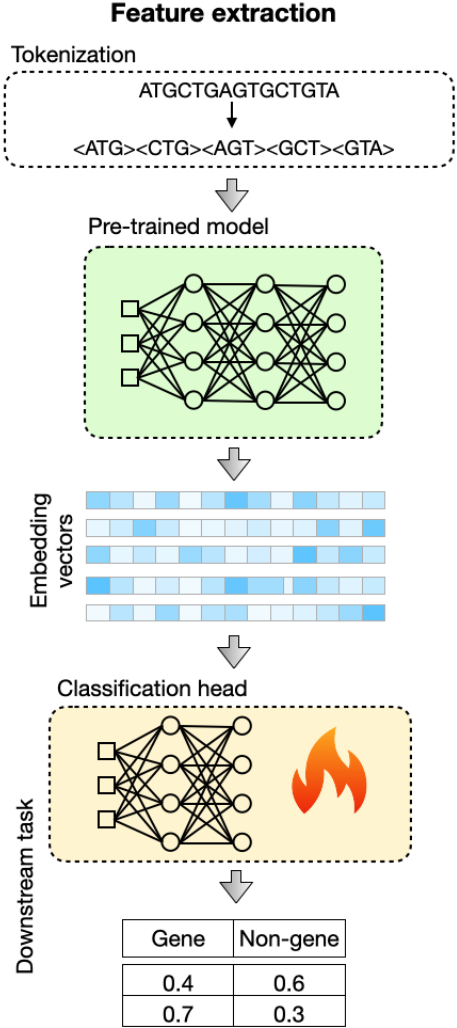
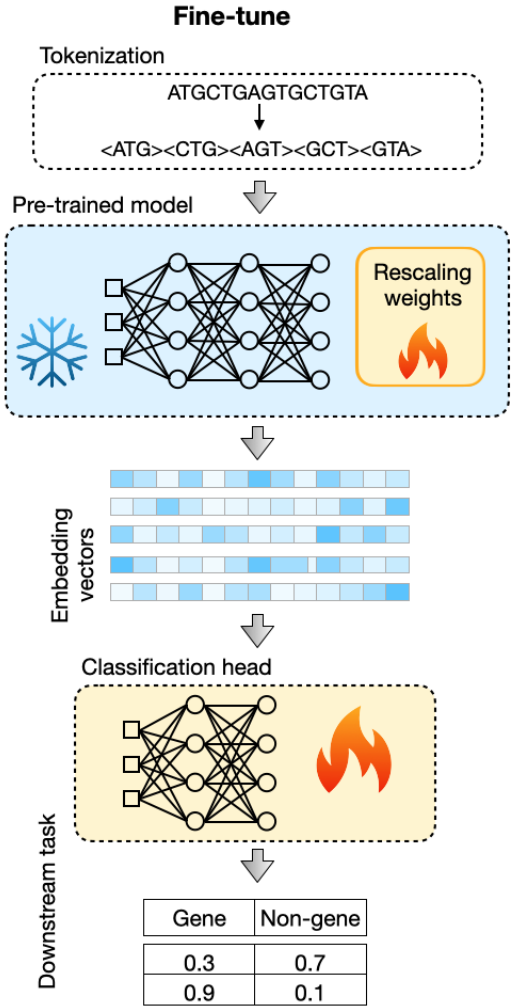
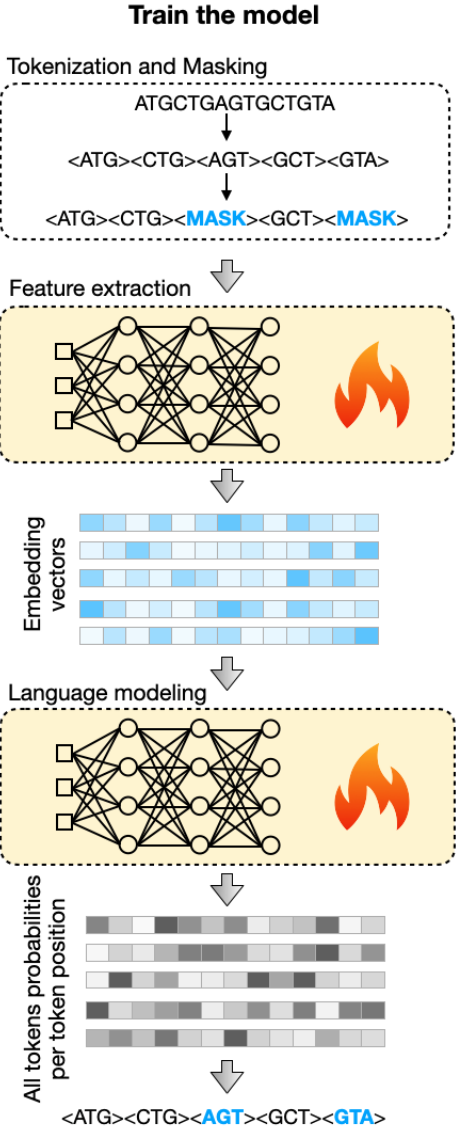
DNABERT – BERT model to DNA



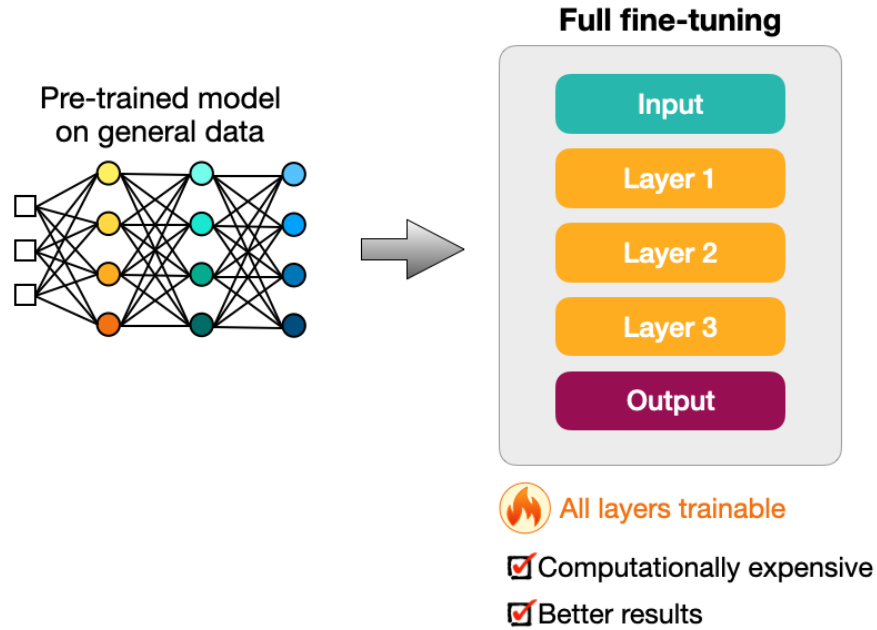
DNABERT – BERT model to DNA



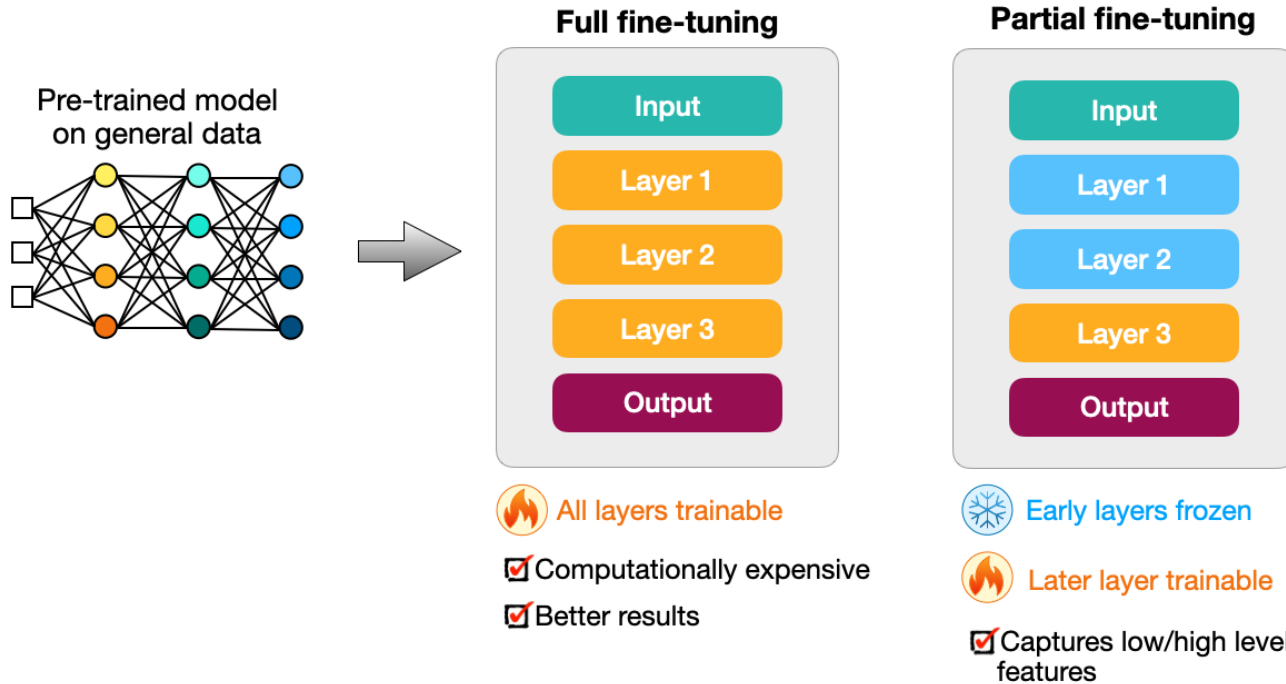
LLMs - Three Approaches



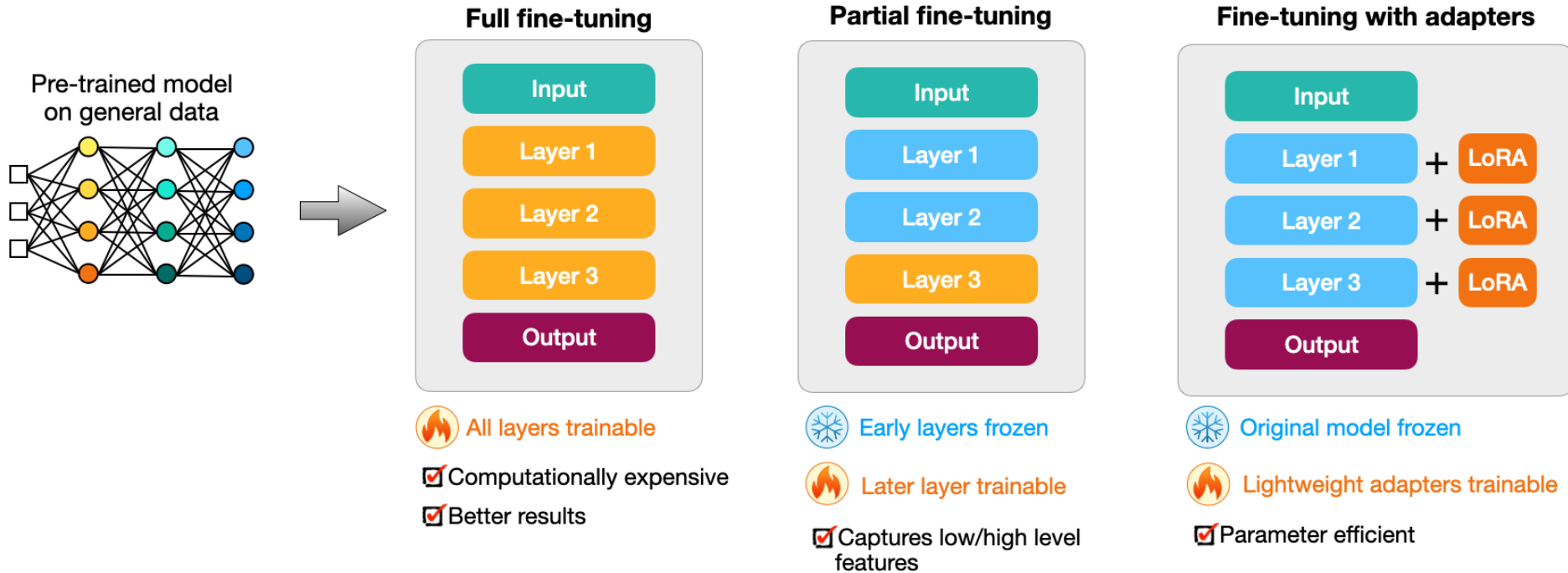
Fine-tuning



Fine-tuning



Fine-tuning



LoRA (Low-Rank Adaptation):

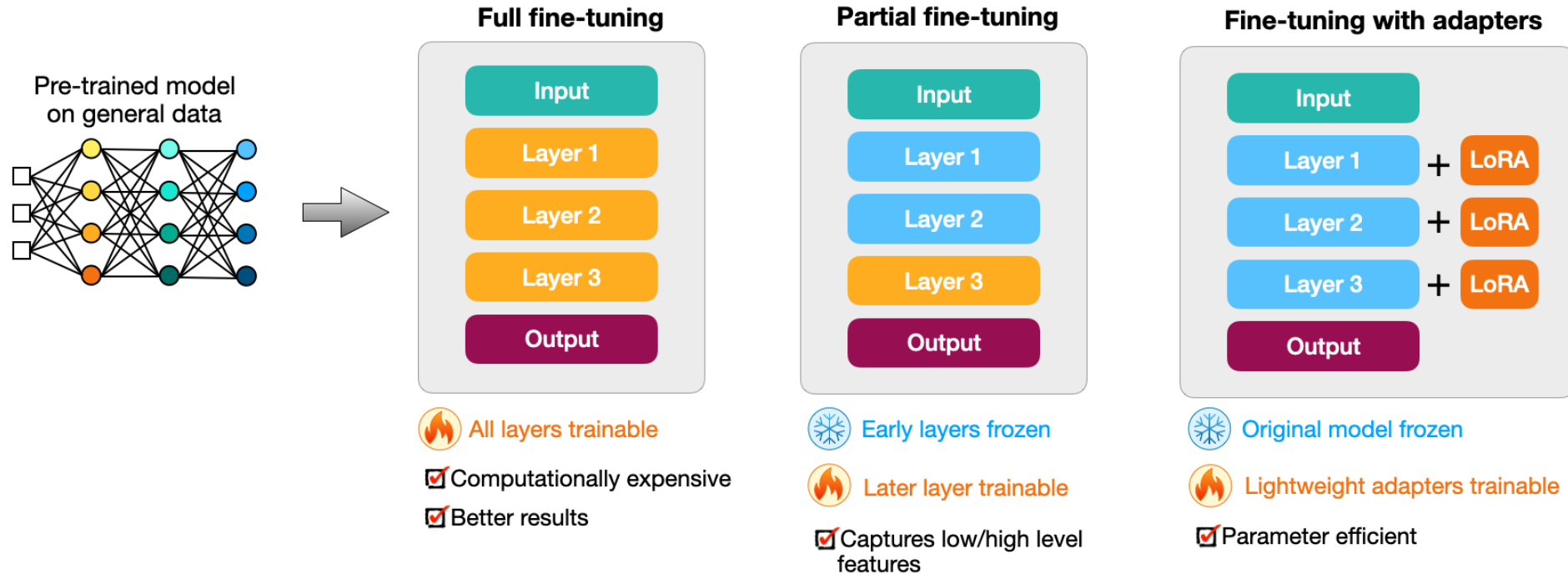
- Decomposes weight updates into low-rank matrices ($A \times B$)

Trade-off: Full fine-tuning vs. Adapters

- Better results vs. Efficiency



Fine-tuning



Key Concepts:

Frozen Layers:

- Parameters unchanged
- Preserve learned features
- Less computation

Hierarchical Features:

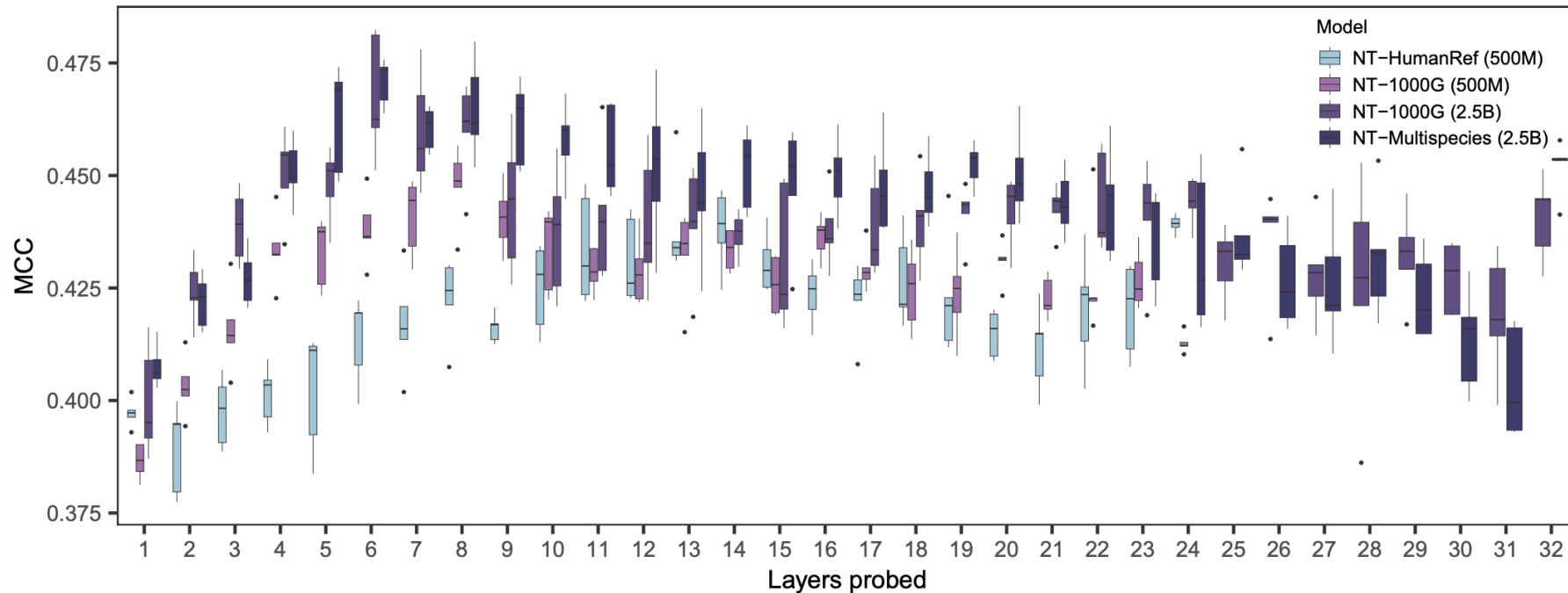
- Early: low-level features
- Later: high-level features
- Task-specific

Adapters (LoRA, etc.)

- Lightweight modules
- Few parameters
- Parameter efficient

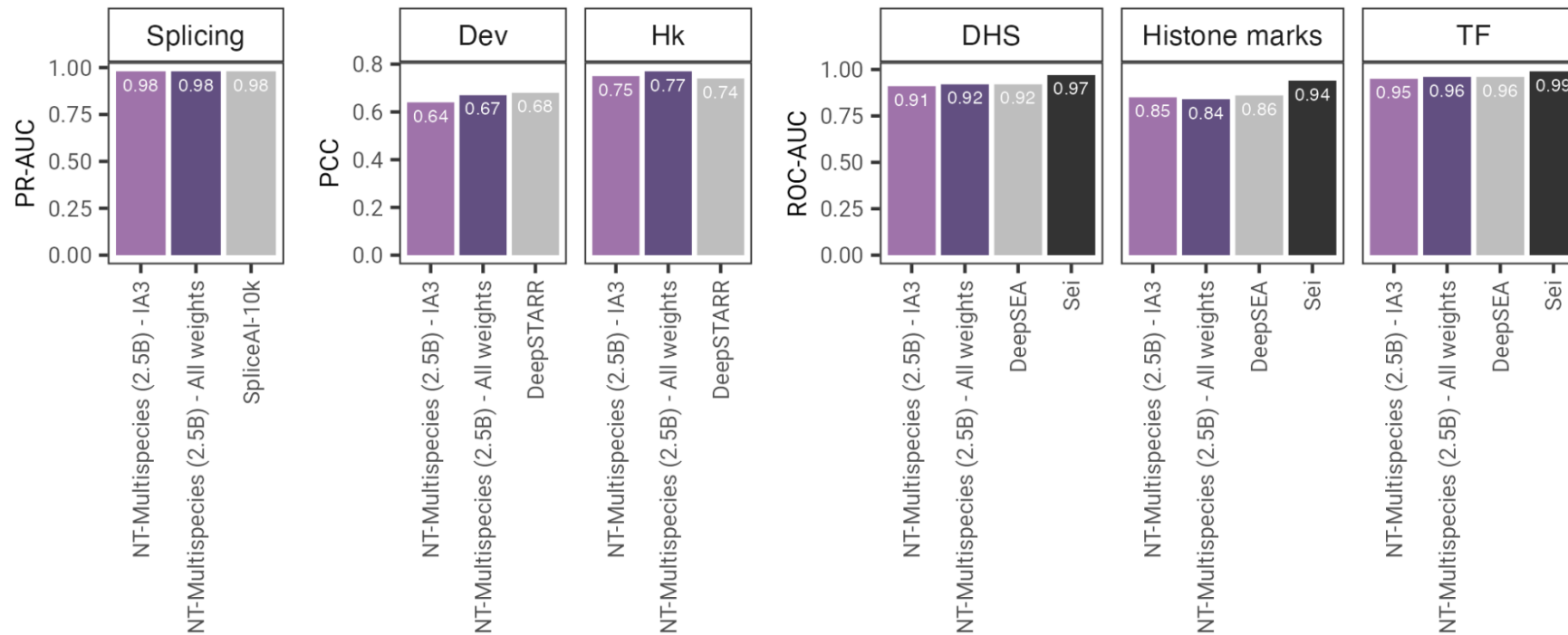
Partial fine-tuning - Layers

- Nucleotide Transformer probing experiments
 - Pre-trained model for feature extraction
- Probing performance across layers for the enhancer prediction task
- Variation in performance may also occur when applying partial fine-tuning

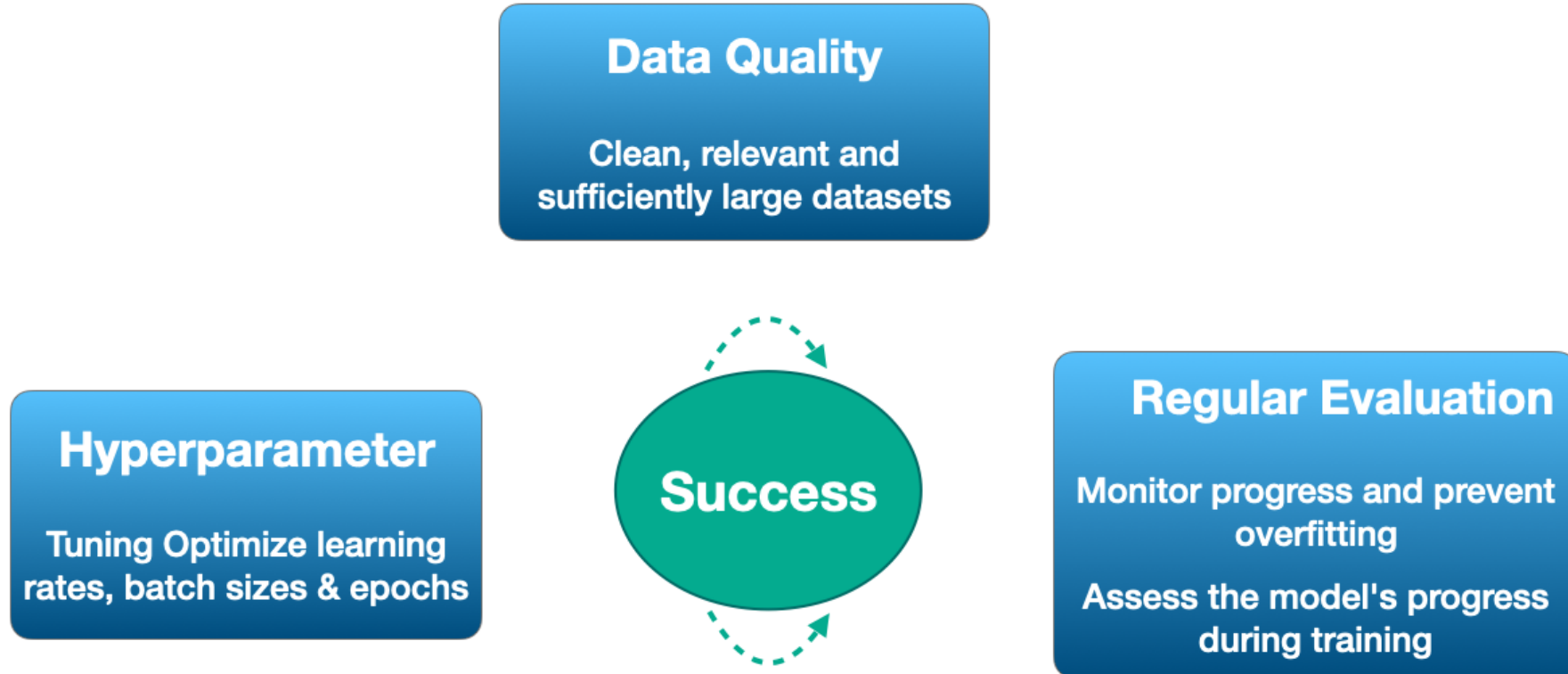


Full fine-tuning vs efficient fine-tuning

- Nucleotide Transformer experiments
- Parameter **efficient** fine-tuning **compared** with **full-model** fine-tuning
- The performances across **different tasks** are shown and compared with respective baselines



Fine-tuning - Best Practices



Avoiding LLM Fine-Tuning Pitfalls

● Overfitting

- High training accuracy, poor generalization
- Caused by small datasets or too many epochs

● Underfitting

- Insufficient learning of the task
- Result of inadequate training or low learning rates

● Catastrophic Forgetting

- Loss of broad knowledge during task-specific training
- Reduces versatility

● Data Leakage

- Training/validation overlap
- Leads to misleading performance metrics
- Keep datasets separate!

Hands-on: Fine-tuning and model evaluation

Hands-on

- Practical exercises about
 - Fine-tuning
 - LLMs: DNABERT2 and Nucleotide Transformer
- Evaluation and comparison
 - Pre-trained vs fine-tuned models
- Datasets
 - DNABERT2 benchmark: https://github.com/MAGICS-LAB/DNABERT_2
- Practical code: https://github.com/fabianagoes/bc2_tutorial8