UNIVERSITY OF TORONTO

An Introduction to FoundationModels for Science

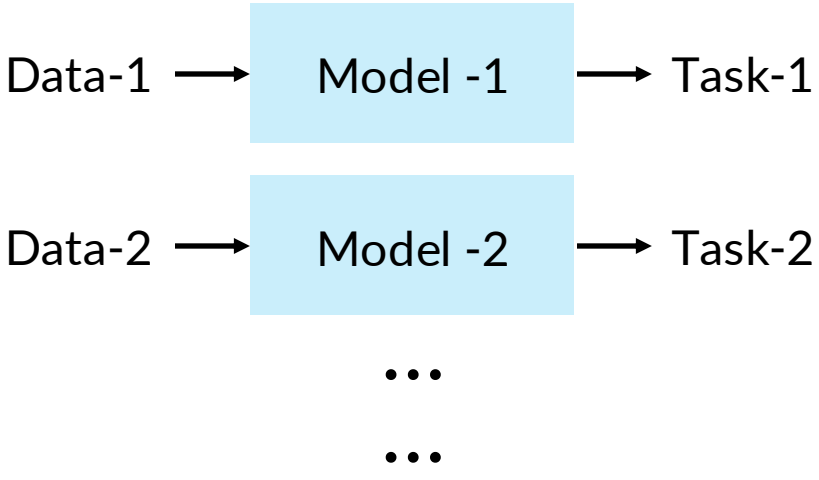
Tutorial 2: Training and Fine Tuning

Schmidt Sciences

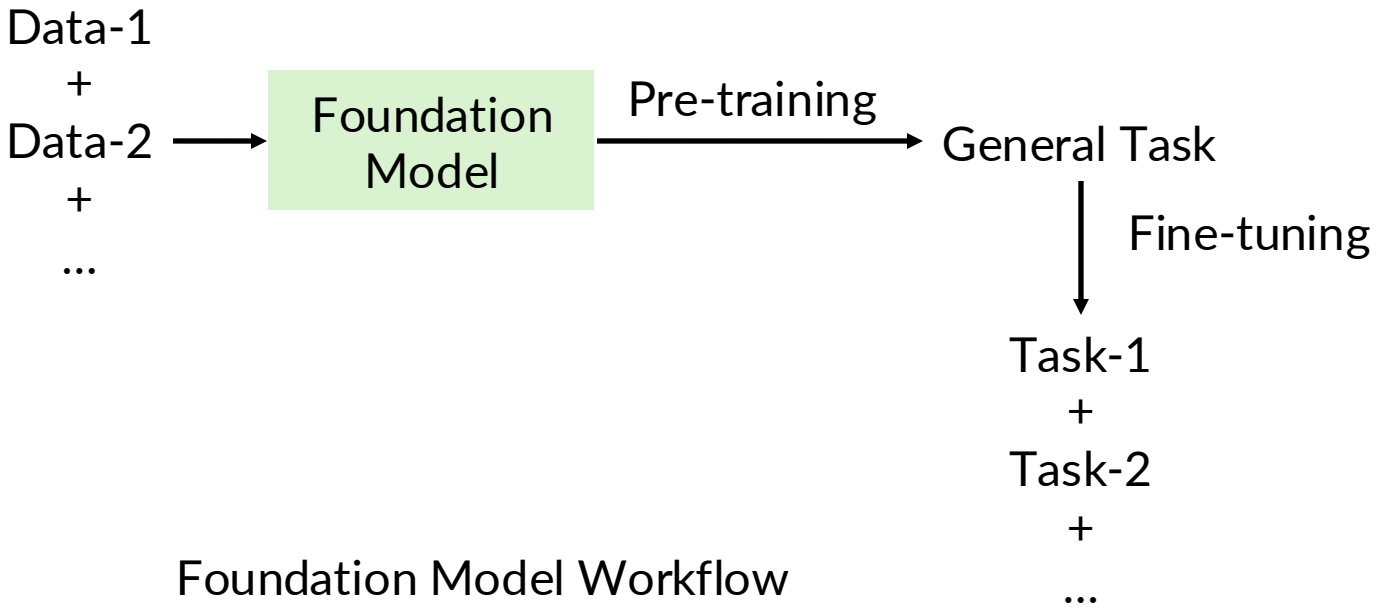
FOUNDATION MODELS for SCIENCE

Foundation Models: What?

- Large deep learning models trained on vast amounts of potentially unlabeled data.
- Pre-trained on a “general purpose” task and can be adapted to solve two or more downstream tasks.



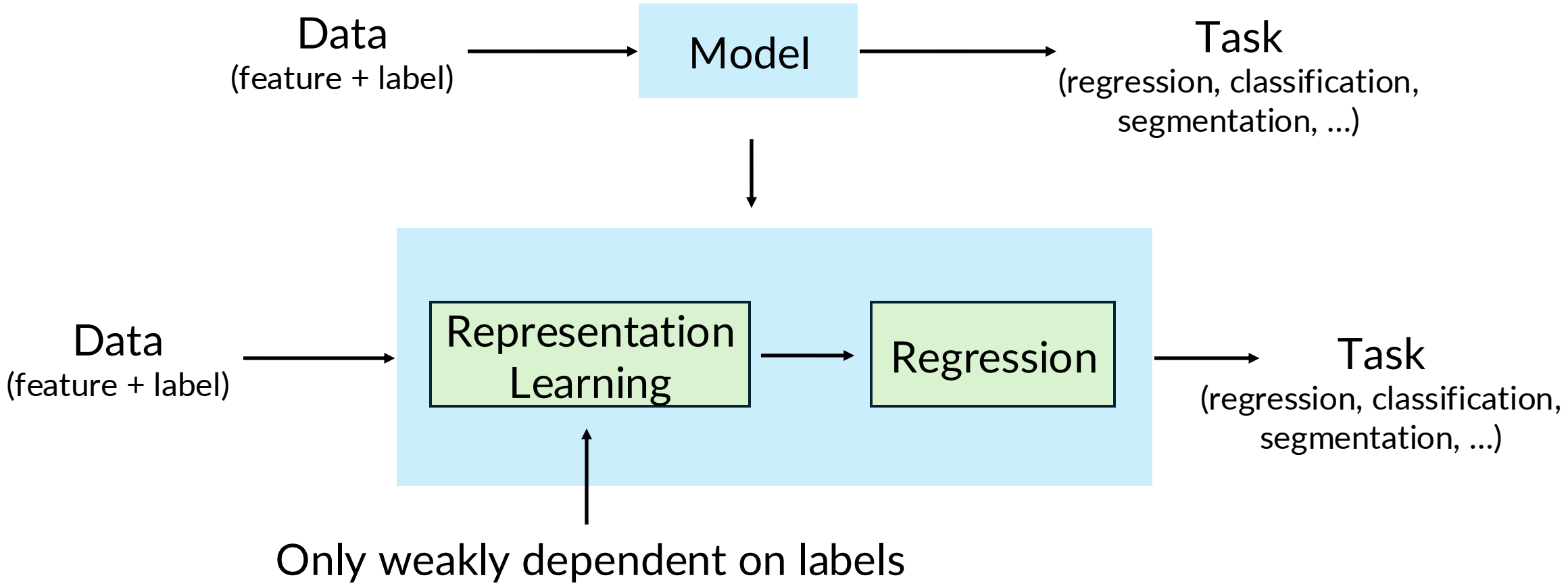
Traditional Workflow



Foundation Model Workflow

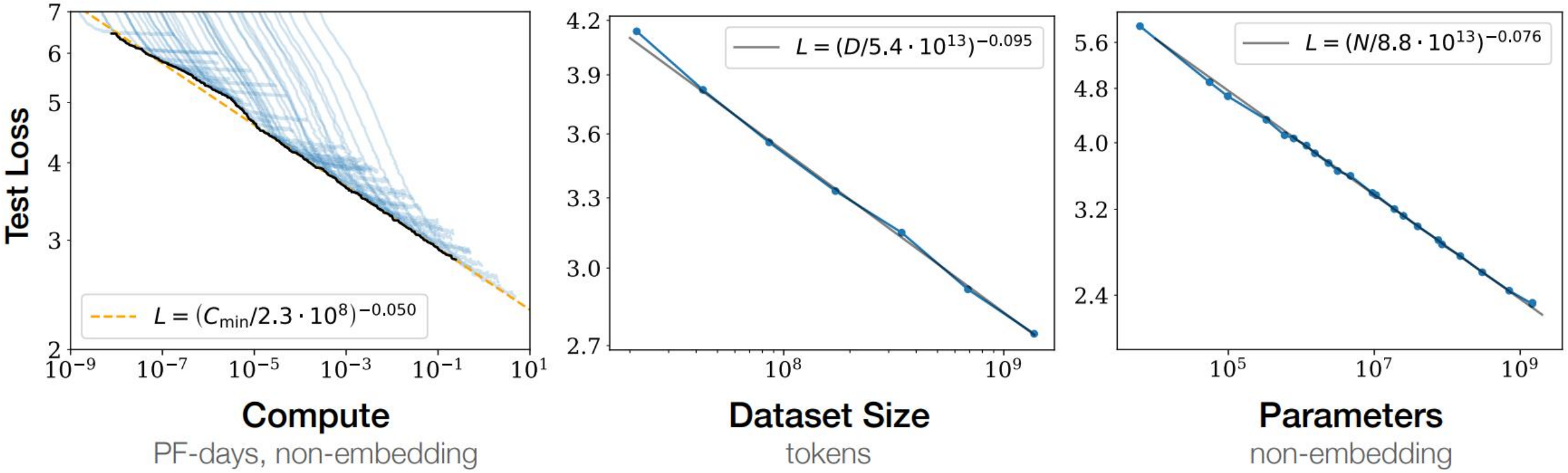
Foundation Models: Why?

Better representation learning from large data sets



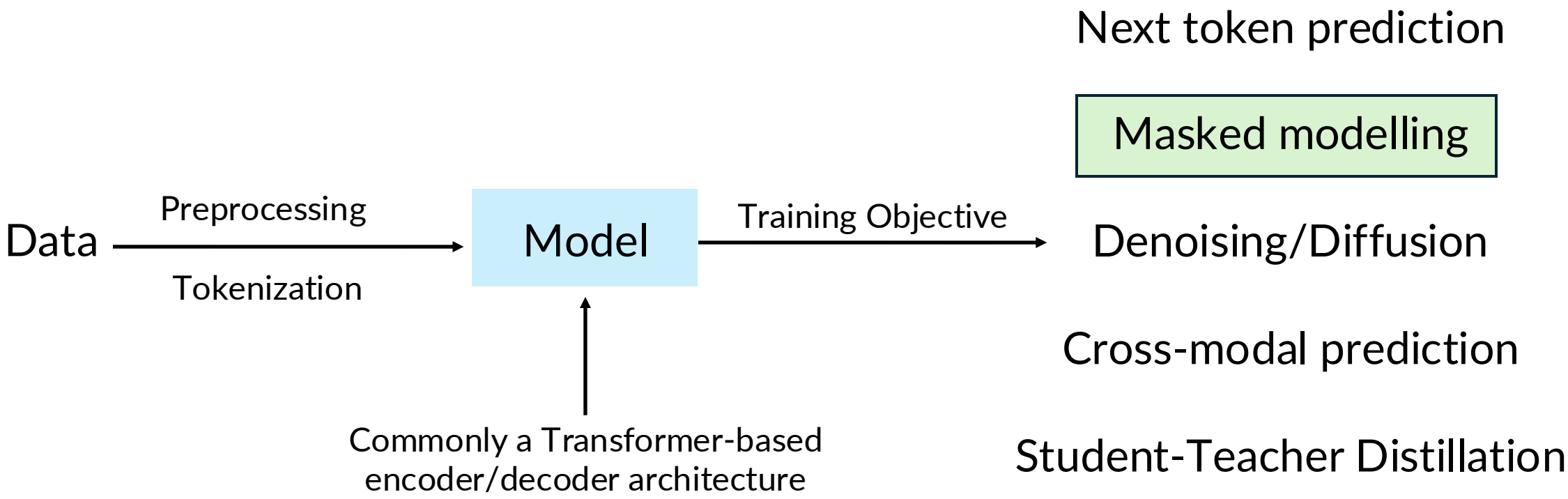
Foundation Models: Why?

Performance gets better with more data and more parameters



Kaplan et al. (2020)

Foundation Models: Pre-training



Masked Language Modelling

Learning a language by predicting missing words

The cat ___ on the mat.

Because the rain was _____ the ____ was cancelled.

Masked Language Modelling

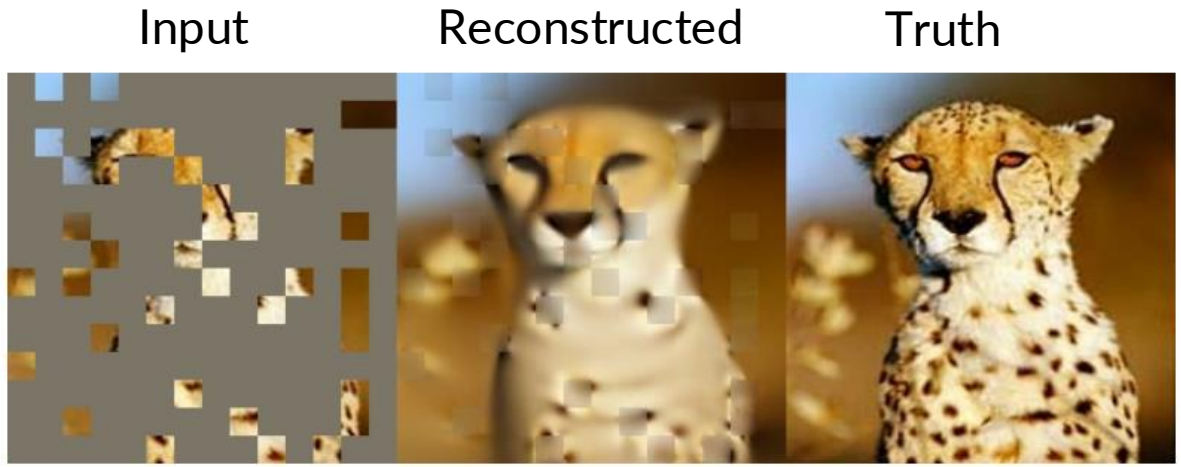
Learning a language by predicting missing words

The cat sat on the mat.

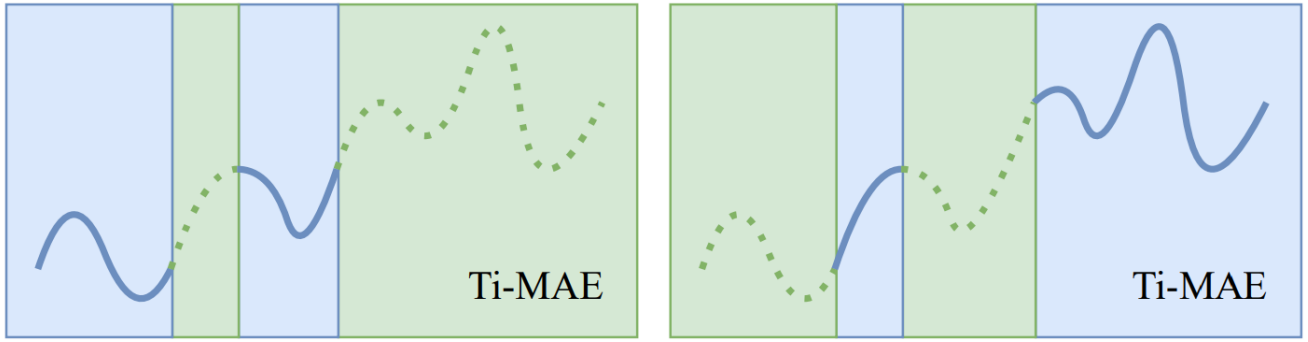
Because the rain was heavy the game was cancelled.

Masked Data Modelling

Learning about data by predicting missing chunks



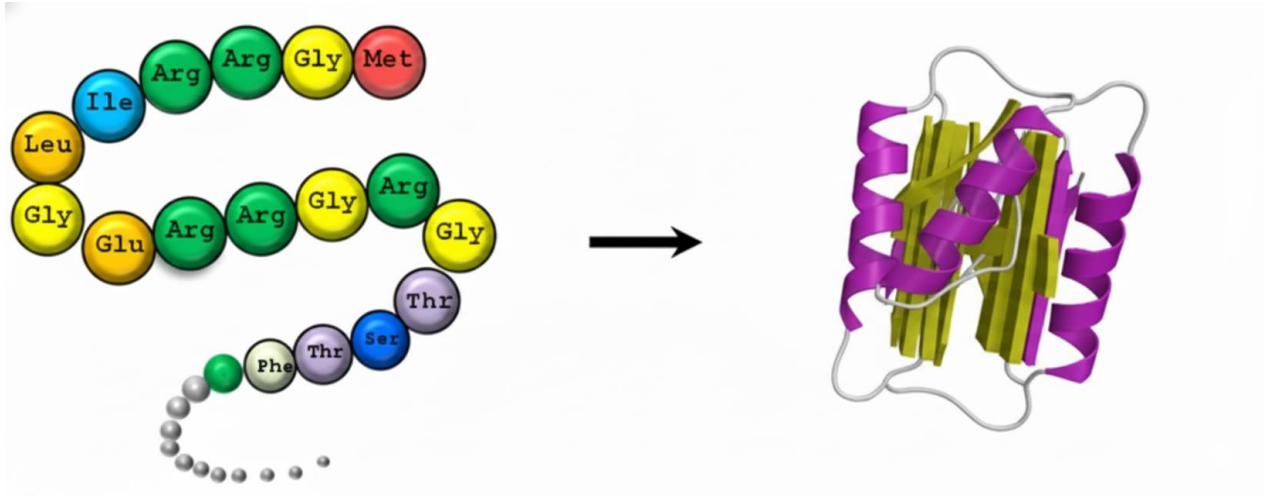
He et al (2021)



Li et al (2023)

Case Study: Protein Language Models

- Proteins are building blocks of life.
- All known proteins are made of various sequences of about 20 amino acids.
- The amino acids “fold” to form a functional 3D structure.
- Not all sequences of amino acid will be a protein.
- We can use techniques from natural language modelling!



Case Study: Protein Language Models

Human Insulin

MAL___LLPLLALLALWGPDP___FVNQH
LCGSH____LYLVCGERGFFYT_KTRREA
EDLQVGQVELGGGP____QPLALEGSLQ
KRGIVEQ____ICSLYQLENYCN

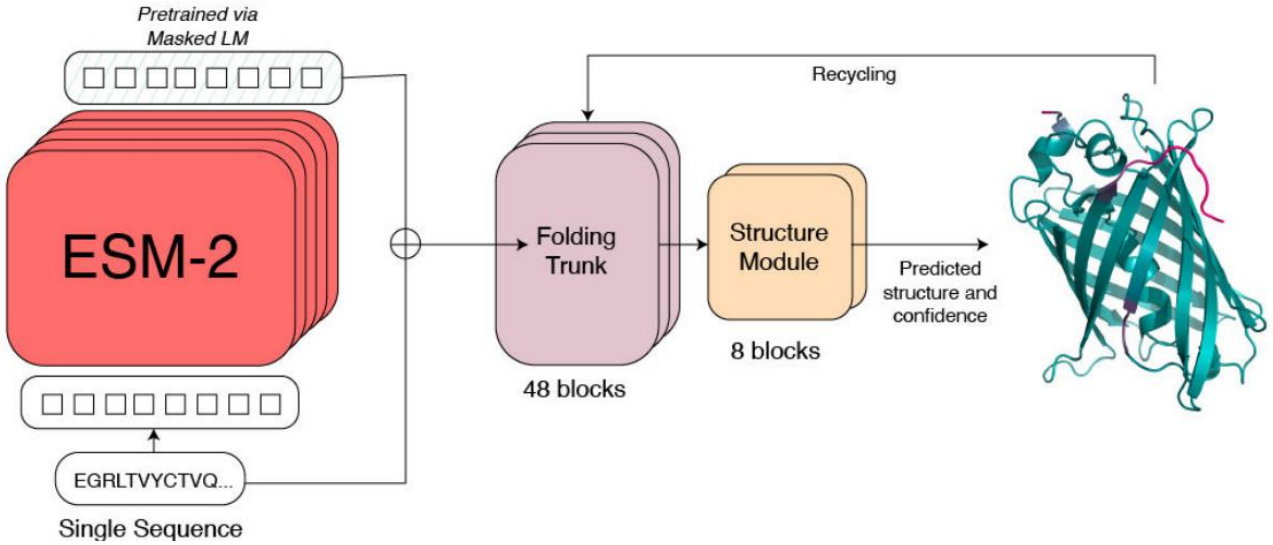
Case Study: Protein Language Models

Human Insulin

MALWMRLLPLLALLALWGPDPAAAFV
NQHLCGSHLVEALYLVCGERGFFYTPKT
RREAEDLQVGQVELGGGPGAGSLQPLA
LEGSLQKRGIVEQCCTSICSLYQLENYCN

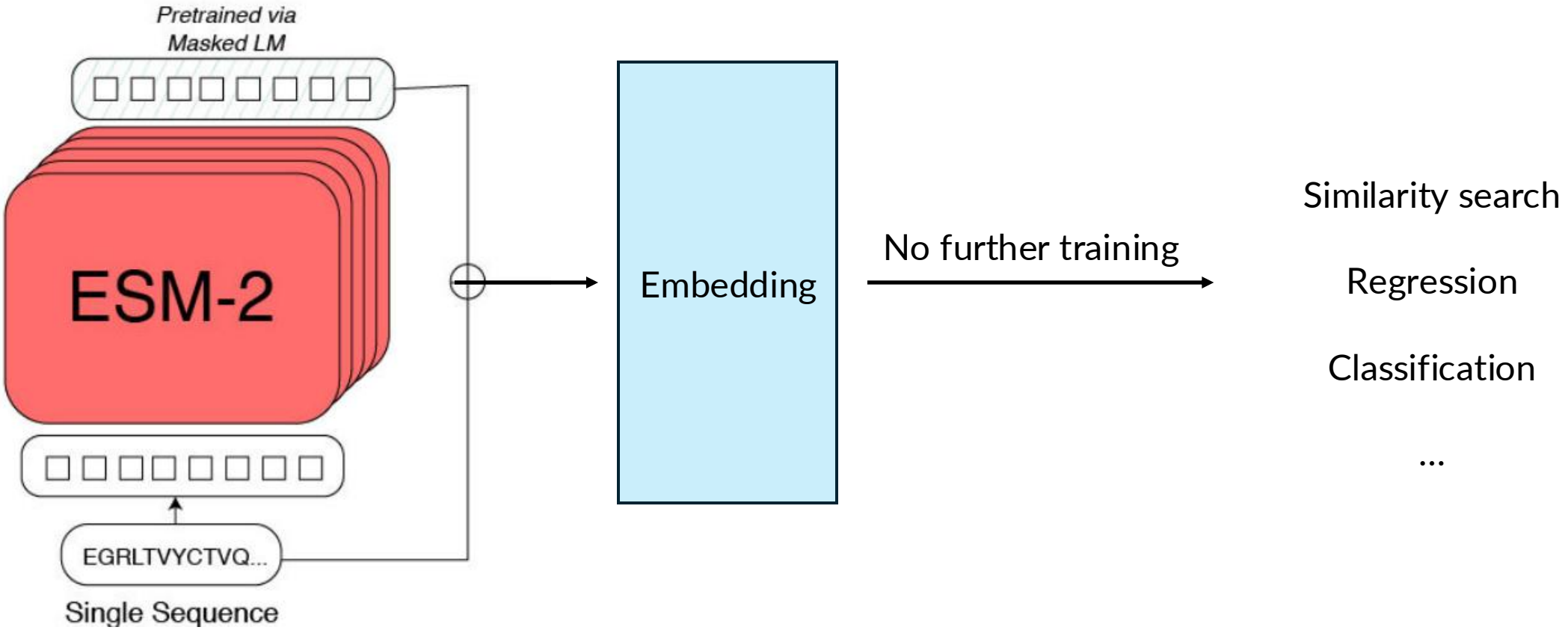
Evolutionary Scale Model (ESM-2)

- Transformer-based, BERT-style protein language model developed by FAIR for amino acid sequences.
- Trained on hundreds of millions of natural protein sequences
- Uses masked language modeling to learn representations
- Supports a variety of downstream bioinformatics tasks:
  - Protein structure prediction
  - Mutation effect analysis
  - Protein-protein interaction.
  - More …



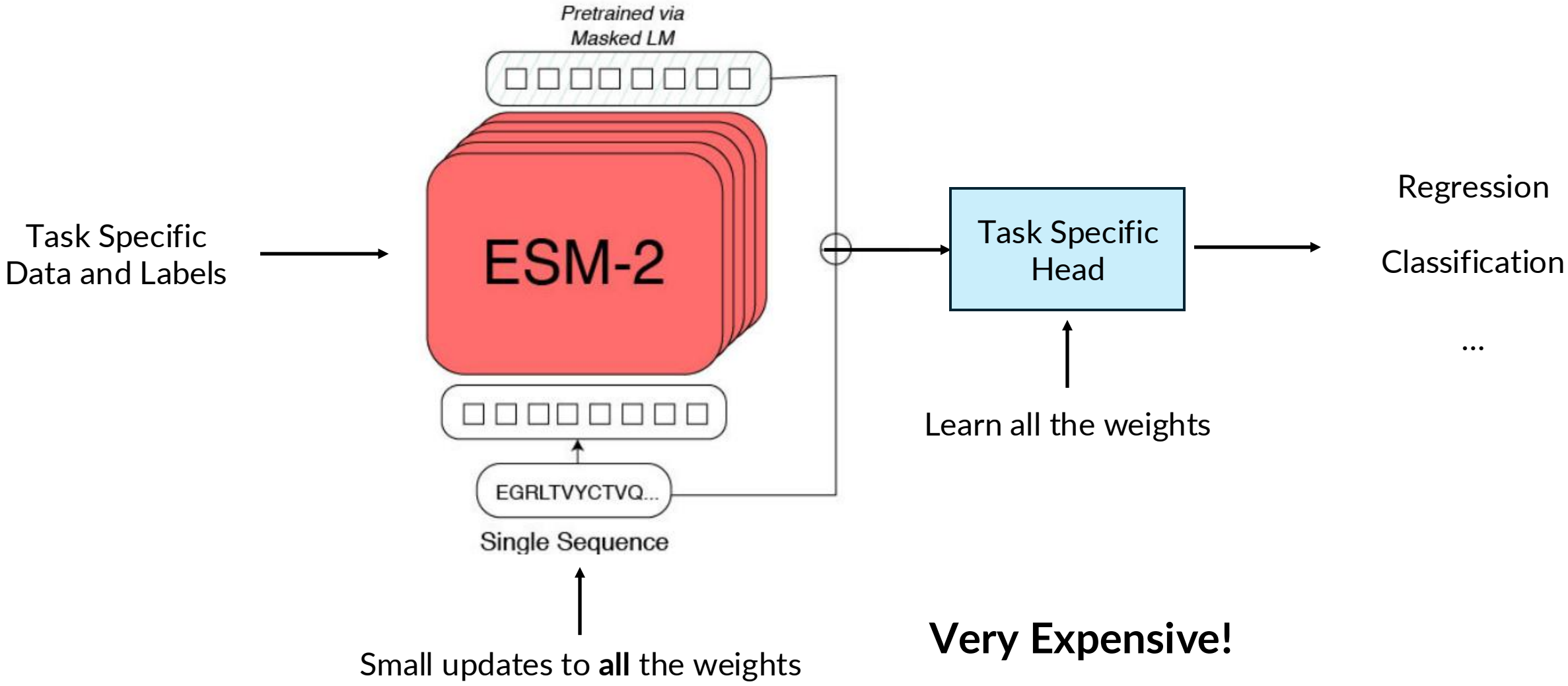
Lin et al (2023)

Using a Foundation Model: Zero Shot Tasks



(Sometime Works)

Using a Foundation Model: Fine Tuning



Using a Foundation Model: Parameter Efficient Fine Tuning

Low-Rank Adaptation (LoRA):

Learn a “correction” to the weights of the pretrained model using a smaller model

$$h = W_0 x$$

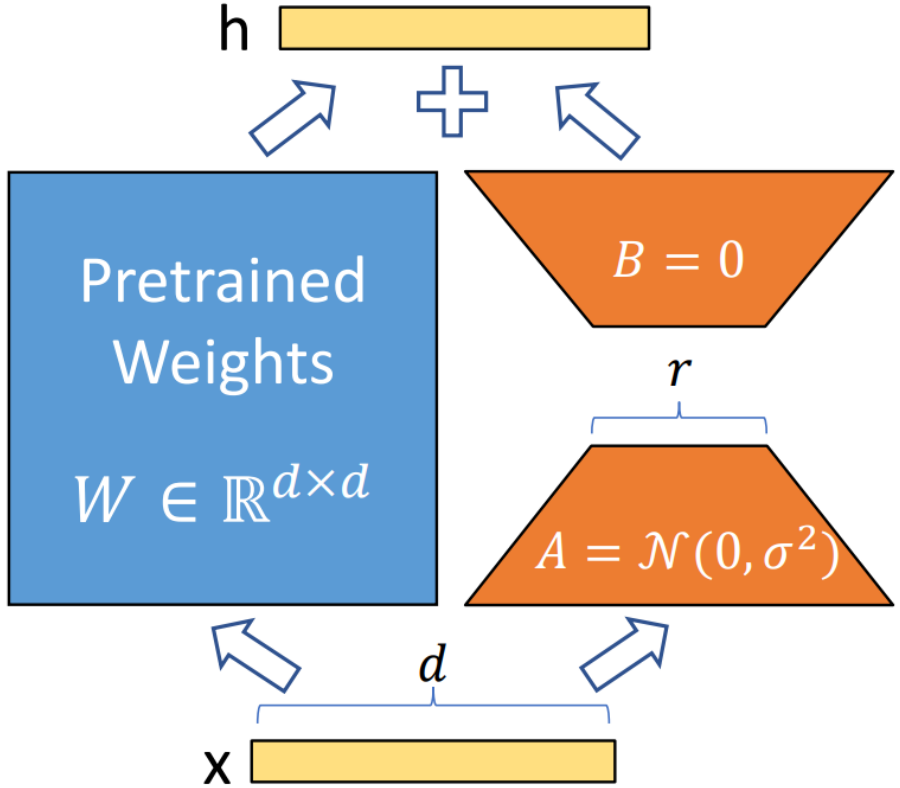
$$\downarrow$$

$$h = W_0 x + \Delta W x$$

$$\downarrow$$

$$h = W_0 x + (AB)x$$

A and B are smaller (Low-Rank) matrices



Hu et al (2021)

Let’s perform these steps ourselves

https://github.com/ai-for-science-org/tutorials/tree/main/Tutorial_2_Fine_Tuning


FOUNDATION MODELS
for SCIENCE