Introduction to Genomic Foundation Models

Ikram Ullah, Staff Scientist

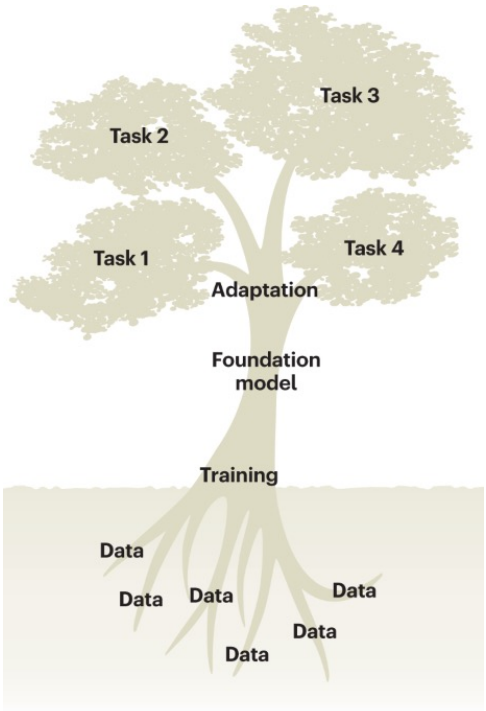


Image taken from – Tang, Lin. "Large models for genomics." *Nature Methods* 20.12 (2023): 1868-1868.

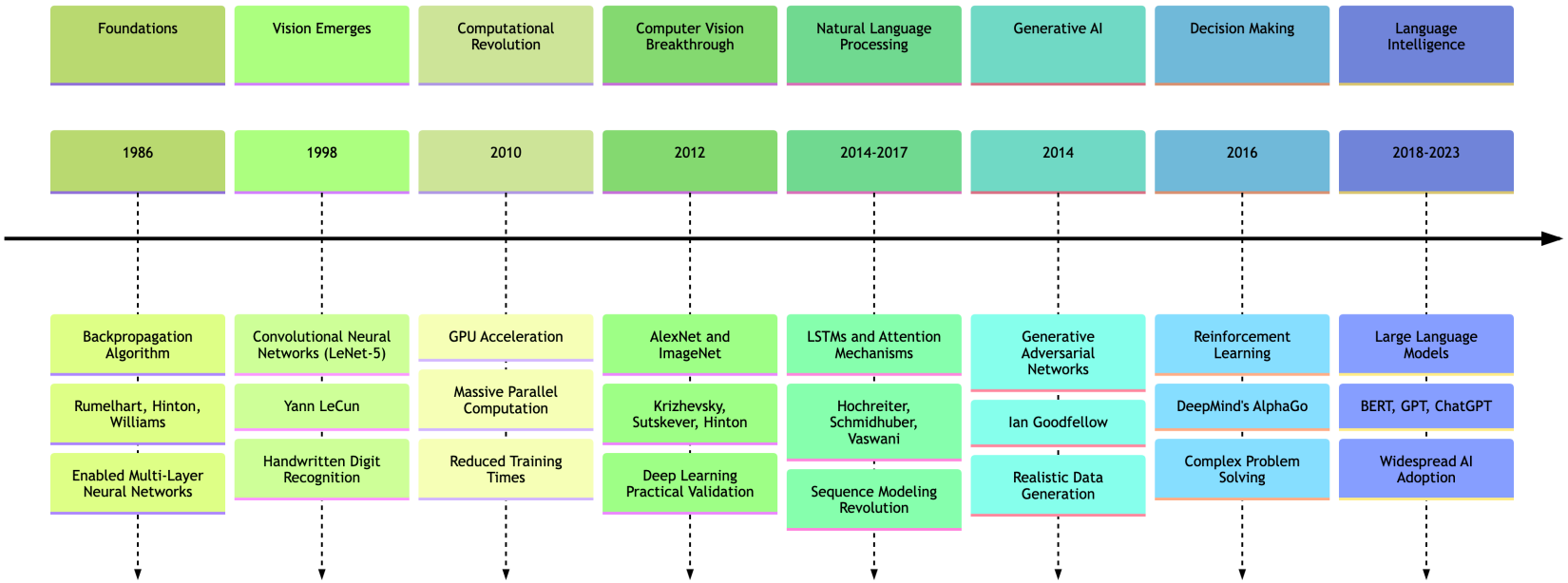
Agenda

- From Multilayer NN to Large Language Model
- What is a Genomic Foundation Model
- Evolution of Genomic Foundation Models
- Traditional ML models vs Genomic Foundation Models
- Key Biological Applications & Research Impact

UMAP visualization of the pretrained scGPT cell embeddings (emb; a random 10% subset), colored by major cell types

Cui, Haotian, et al. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI." *Nature Methods* 21.8 (2024): 1470-1480.

Brief History of Deep Learning



What is a Large Language Model

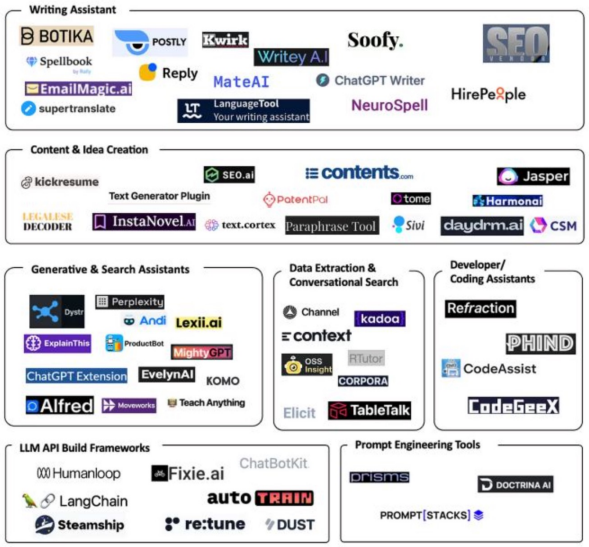
Large Language Model (LLM) are large-scale machine learning models trained on vast datasets from Internet

Designed to serve as adaptable starting points for downstream tasks via

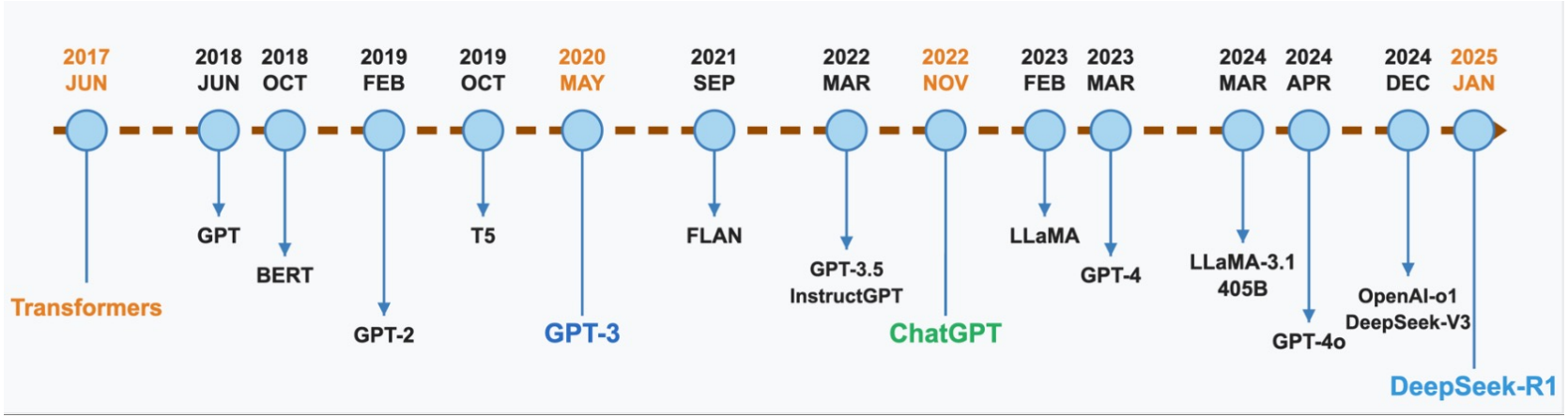
- Fine-tuning
- Prompt engineering

Mathematically

$$L = -\sum_{x \in D} logP(x|context(x))$$



Brief History of Large Language Models



What is a Genomic Foundation Models?

Definition

Genomic Foundation Models are large-scale machine learning models trained on vast genomic datasets that can be adapted to a wide range of downstream tasks with minimal additional training

Key Characteristics

- Pre-trained on large, diverse genomic data
- Self-supervised learning objectives
- Capture complex biological patterns and relationships
- Diverse applications through transfer learning

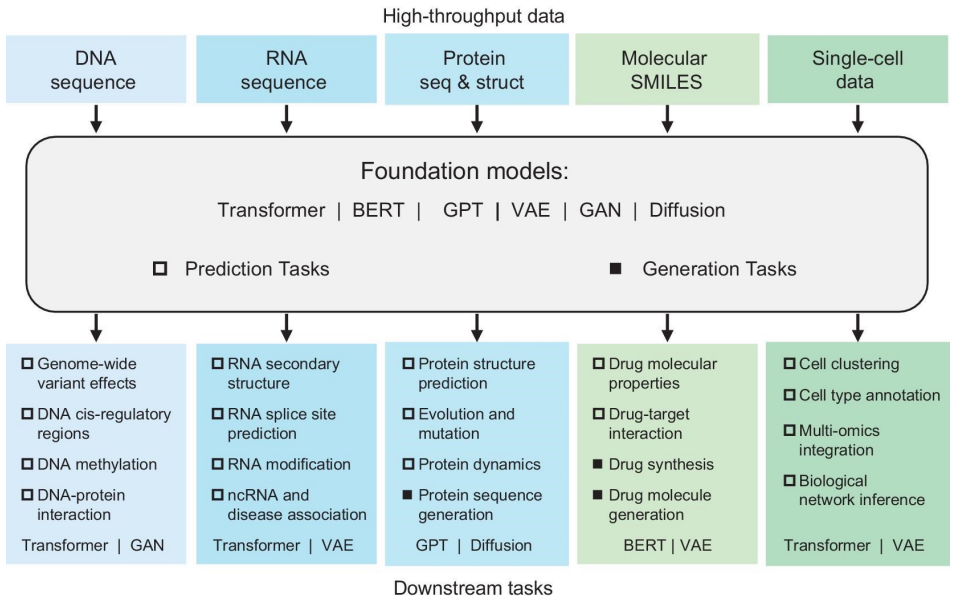


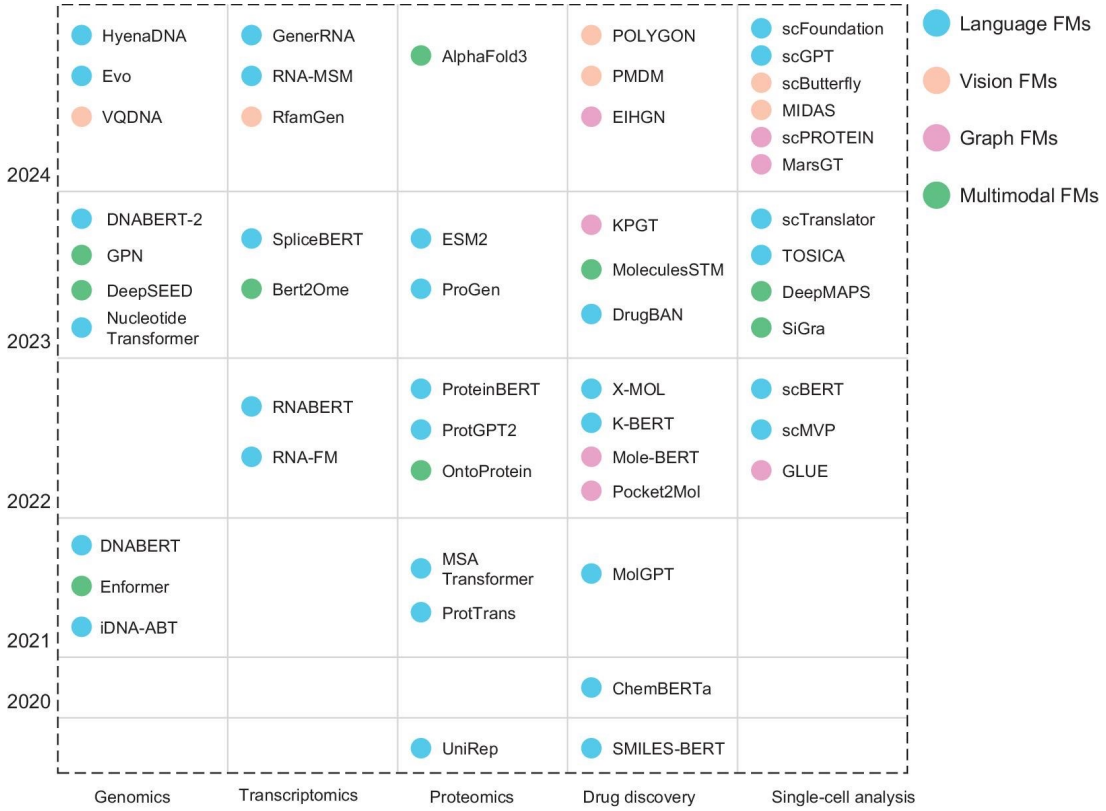
Image taken from: Guo, Fei, et al. "Foundation models in bioinformatics." National Science Review (2025): nwaf028.

Evolution of Genomic Foundation Models

Evolution from specialized sequence models to multimodal genomic understanding

Increasing ability to model long-range dependencies in genomic data

Transition from prediction to generation capabilities



Chen, Ziyu, Lin Wei, and Ge Gao. "Foundation models for bioinformatics." *Quantitative Biology* 12.4 (2024): 339-344.

Traditional ML vs Genomic Foundation Models

Model Scope: The Fundamental Difference

Traditional ML

- One model per genomic task
  - Random Forest for enhancer prediction
  - CNN for transcription factor (TF) binding site prediction
  - SVM for gene expression classification

Foundation Models

- One model for many genomic tasks
  - HyenaDNA: chromatin profile prediction and species classification [1]
  - DNABERT: promoter prediction and splice site detection [2]

Key Impact

Specialized tools → Swiss Army knife

GFMs require **significant infrastructure and expertise** to pretrain and serve. A well-designed SVM or RF model is still practical and efficient for small labs

Feature Representation: How Models "See" DNA

Traditional ML

- Hand-crafted biological features
  - k-mers, GC content, Position Weight Matrix (PWM) scores
  - Typically, 10–100 dimensions
  - Requires expert biological knowledge

Foundation Models

- Learned sequence embeddings
  - Context-aware, position-informed representations
  - 512–4096+ dimensional dense vectors
  - Automatically learns patterns from data (e.g., motif grammar, syntax) splice site detection in DNABERT

Key Impact

Humans design features → AI discovers patterns

Traditional ML methods based on deep learning also implicitly learns features, however it may not be that strong due to context and data scale difference

Data Requirements: Training Data Needs

Traditional ML

- Labeled data bottleneck
  - Supervised training needs curated annotations per task
  - Difficult to scale across tissues/species
  - Suffers from class imbalance (e.g., rare enhancer vs. background)

Foundation Models

- Unlabeled data advantage
  - Self-supervised learning (e.g., masked k-mer prediction)
  - Can pretrain on entire genome sequences
  - Fine-tuning requires minimal labeled data (few-shot learning)

Key Impact

Annotation-hungry ⟶ Raw sequence-powered

Larger data processing usually required larger infrastructure and may not be practical for most research labs

Transfer Learning: Knowledge Sharing Across Contexts

Traditional ML

- Context-locked models
  - Trained separately per tissue, cell type, or species
  - No transferability; each new context = new model
  - Computationally and manually intensive

Foundation Models

- Cross-context generalization
  - Pretrained representations work across human, mouse, plant genomes
  - Capture conserved regulatory patterns across evolution
  - Supports zero-shot or few-shot transfer

Key Impact

Isolated Knowledge ⟶ Sharing Knowledge

Transfer of engineered features, and **meta-learning** may be used to bridge some tasks in classical ML

Scalability: Handling Genomic Scale

Traditional ML

- Computational limitations
  - Performance stagnates with large datasets
  - Manual feature engineering doesn't scale
  - Typically handles short sequences (≤1000 bp)

Foundation Models

- Genome-scale distributed learning
  - Uses transformer or efficient variants (e.g., Hyena, Performer)
  - Trained on full chromosomes (up to millions of base pairs)
  - Supports multi-GPU and TPUs for scaling

Key Impact

Hits ceiling quickly ⟶ Scales with genomic complexity

Traditional ML models can **outperform GFMs** in limited data or narrow biological task cases.
GFM may overfit or under-perform on small tasks

Biological Insight: Interpretability vs Discovery Trade-off

Traditional ML

- Explicit interpretability
  - Feature importance scores are straightforward
  - Model coefficients and tree structures are inspectable
  - Easy to trace to known motifs and biological mechanisms

Foundation Models

- Emergent understanding
  - Learns regulatory rules, chromatin logic, 3D interactions
  - Identifies unknown long-range dependencies
  - Less transparent, but deeper insight when probed (e.g., attention)

Key Impact

Explains the known ⟶ Discovers the unknown

Interpretability still **matters deeply** in regulatory genomics and medicine

Quick Check: Genomic Foundation Models

- What is the primary advantage of genomic foundation models over traditional task-specific models?

A. They require less computational resources
B. They eliminate the need for labeled data entirely
C. They learn rich representations transferable to multiple tasks
D. They are always more accurate than specialized models

Answer: C. They learn rich representations transferable to multiple tasks
Foundation models learn generalizable representations from large amounts of data that can be efficiently adapted to multiple downstream tasks through fine-tuning, reducing the need for task-specific data collection and annotation.

Key Biological Applications

DNA (Sequence)	RNA (Sequence)	Protein (Seq, Structure)	Single Cell (Data)
• Genome-wide variant effects • DNA cis-regulatory regions • DNA methylation • DNA-protein interaction	• RNA secondary structure • RNA splice site prediction • RNA modification • ncRNA and disease association	• Protein structure prediction • Evolution and mutation • Protein dynamics • Protein sequence generation	• Cell clustering • Cell type annotation • Multi-omics integration • Biological network

Guo, Fei, et al. "Foundation models in bioinformatics." *National Science Review* (2025): nwaf028.

Research Impact: Case Studies

Effects of noncoding DNA on gene expression

- **DNABERT** captures transferrable understanding of genomic DNA sequences based on up and downstream nucleotide contexts
- Predicts promoters, splice sites and TFBS, after fine-tuning using small task-specific labeled data

Repurposing of existing drugs

- TxGNN trained on a medical knowledge graph to identify potential new purposes for existing drugs
- Predictions aligns with current off-label practice, offering a promising exploratory tool for expanding drug indications.

Tissue-specific gene network dynamics

- **Geneformer** make predictions about tissue-specific gene network dynamics from scRNA data
- Obtains insights with limited data, like rare diseases and difficult-to-access tissues

Mutation Prediction

- Evo2 demonstrates over 90% accuracy in predicting the functional impact of mutations in genes such as BRCA1
- Capable of designing entire genomes, including mitochondrial DNA and bacterial genomes, with realistic gene structures and regulatory elements

Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." Bioinformatics 37.15 (2021): 2112-2120.
Huang, Kexin, et al. "A foundation model for clinician-centered drug repurposing." Nature Medicine 30.12 (2024): 3601-3613.
Theodoris, Christina V., et al. "Transfer learning enables predictions in network biology." Nature 618.7965 (2023): 616-624.

Quick Check: Genomic Foundation Models Applications

- Which of these applications would benefit MOST from the long-range dependency modeling capabilities of genomic foundation models?

A. SNP genotyping
B. Enhancer-promoter interaction prediction
C. Basic sequence alignment
D. Simple gene counting

Answer: B. Enhancer-promoter interaction prediction
Enhancer-promoter distance can vary significantly, ranging from a few hundred base pairs to several million base pairs. Traditional models (like CNNs or RNNs) struggle to connect distant genomic elements due to limited receptive fields.

Questions & Comments