

"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER
SCIENCE AND ENGINEERING


AI meets genetics

Prof. Gjorgji Madjarov, PhD
Faculty of Computer Science and Engineering,
Ss Cyril and Methodius University in Skopje

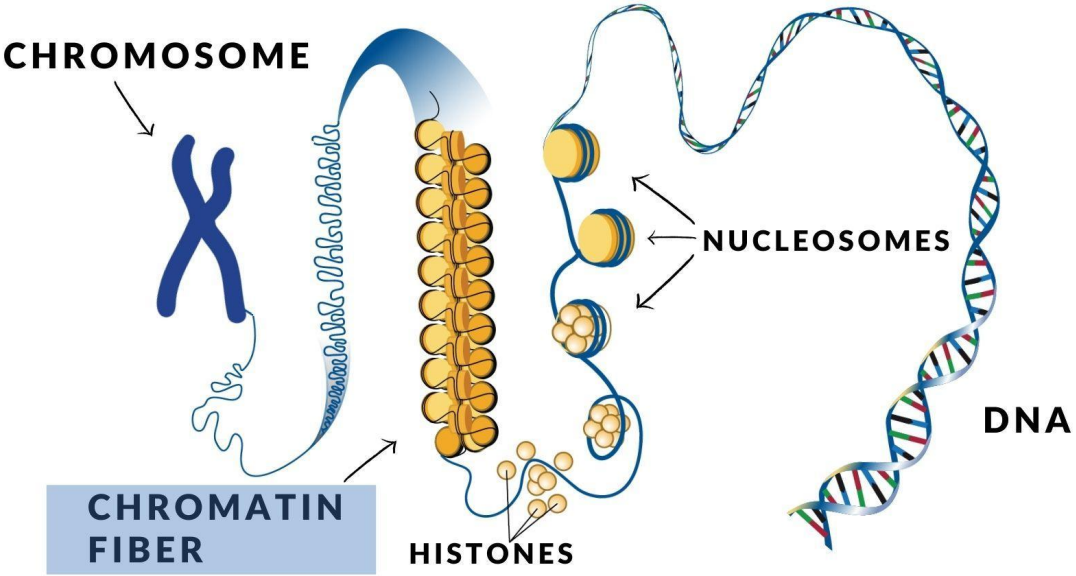
Outline

- Motivation
  - Explosion of biomedical data (genomic, clinical, imaging, etc.)
  - Need for scalable, generalizable AI tools
- What Are Foundation Models?
  - Difference from traditional ML models
  - Definition and key characteristics (pre-training, fine-tuning, complexity, architecture)
  - Foundation models
- Future Directions
  - Multimodal Foundation Models

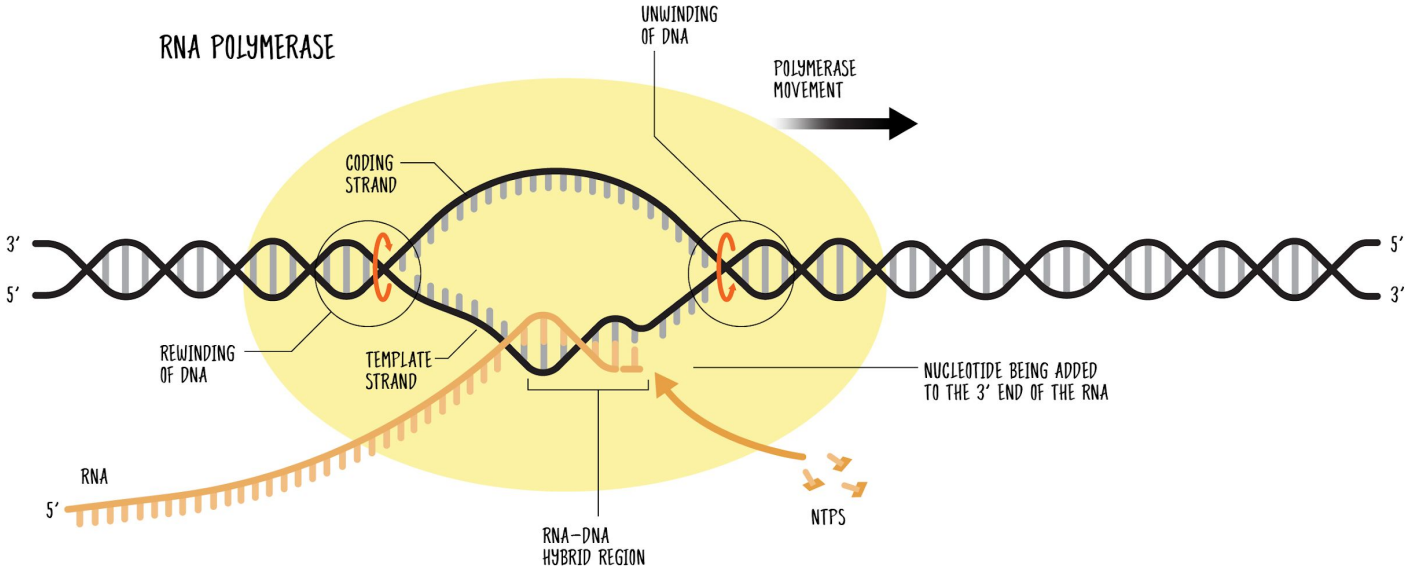
Motivation

- Biomedical research is undergoing a data revolution (the volume and complexity of biomedical information are growing exponentially)
  - Vast repositories of clinical records
  - Medical literature
  - High-throughput sequencing
  - Imaging
  - Multi-omics data
- How can we make sense of such diverse and unstructured data?
- How do we build models that **generalize across tasks**, **adapt to new problems** with minimal supervision, and **support real-world decision-making**?

Refresh our knowledge



Refresh our knowledge



Traditional ML

- Approach:
  - Requires extensive manual feature engineering.
  - Often applies algorithms like Random Forests, SVMs, or logistic regression.
  - Each model is trained task-specifically—e.g., disease classification or gene expression prediction.

- Challenges:
  - Multi-omics data is heterogeneous (e.g., genomics, transcriptomics, proteomics), which makes integration non-trivial.
  - Traditional models struggle with high dimensionality and small sample sizes typical in omics datasets.
  - Requires significant domain expertise for data preprocessing and interpretation.

Foundation Models

- **Foundation models** are identified as one of the most general categories of generative AI.
- Represents a category of AI models that undergo **pre-training on extensive datasets spanning diverse domains.**
- The acquired knowledge of these models can be seamlessly applied with minimal supplementary training efforts
    - **Fine-tuning**

Foundation models

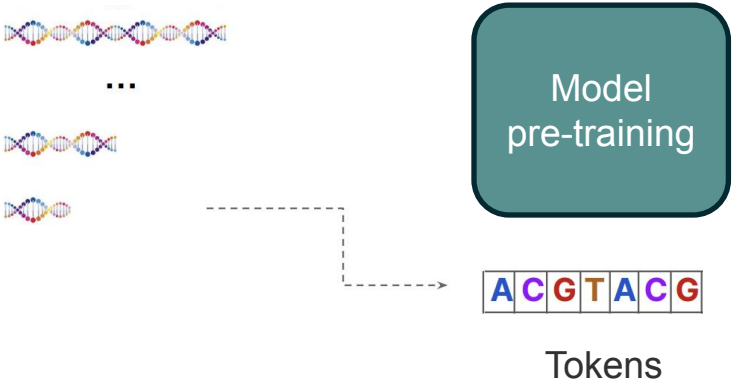
- Approach:
  - Pretrained on massive datasets across various tasks (e.g., genomic sequences, epigenetic signals).
  - Often use transformer-based architectures (like BERT, GPT, or their bio-specific variants: DNABERT, HyenaDNA, etc.).

- Characteristics:
  - Self-supervised learning enables learning rich representations without labelled data.
  - Better at capturing long-range dependencies in sequences (important in genomics).
  - Can integrate multi-modal inputs (e.g., DNA sequences + expression data + clinical metadata).

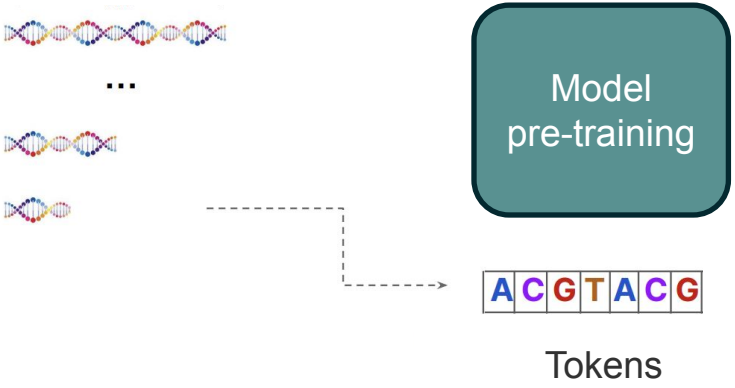
Advantages of Foundation models

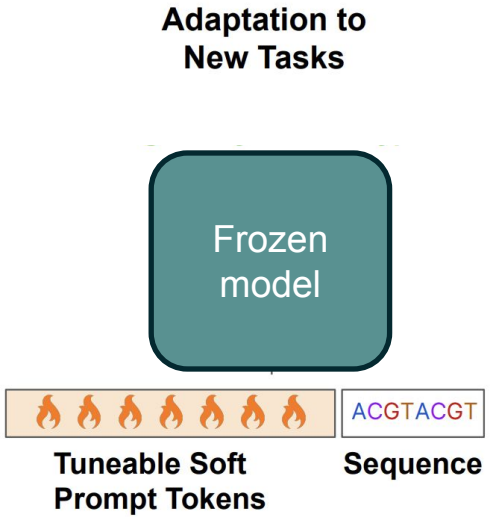
- Transfer learning
  - Models pretrained on large datasets can be fine-tuned with small omics datasets, improving performance.
- Reduced feature engineering
  - Automatically learn representations, reducing the need for manual preprocessing.
- Better generalization
  - Perform well across tasks (e.g., disease prediction, drug response) with minimal additional training.
- Scalability
  - Capable of handling vast and complex multi-omics datasets.
- Multi-modal integration
  - FMs can handle data from different omics layers more naturally.
- Interpretability
  - Some newer foundation models provide attention maps or feature attributions for biological insight.

Foundation Models



Foundation Models

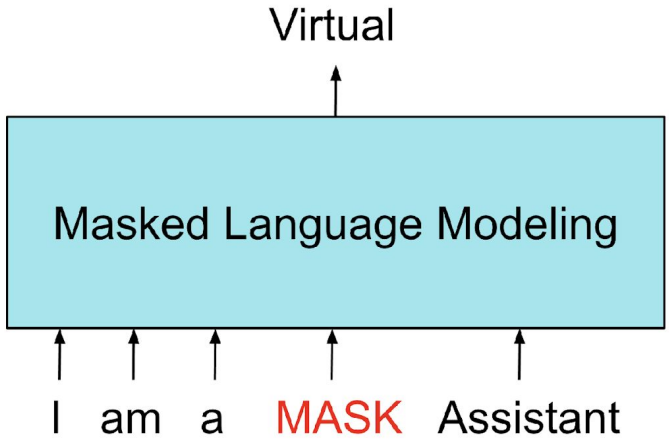




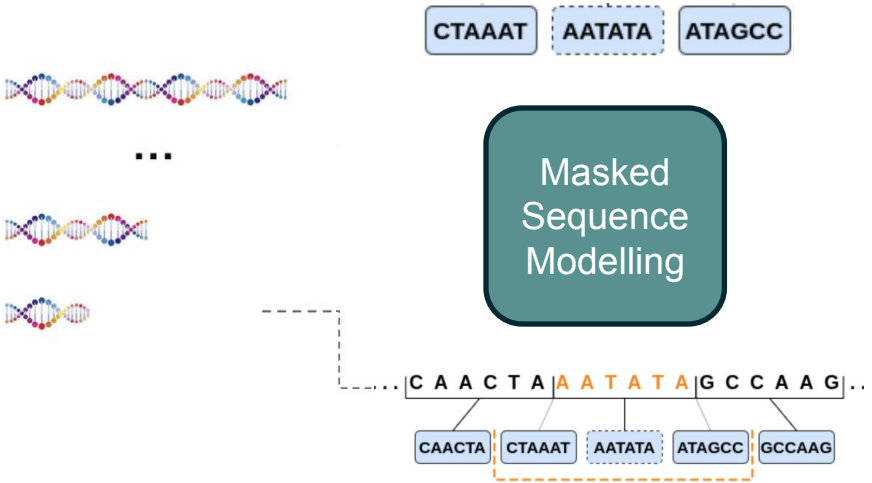
Pre-training strategies

Masked Language Modelling (MLM) / Masked Sequence Modelling

- **How it works:** Randomly mask portions of a DNA sequence and train the model to predict the masked bases.
- **Applications:** Learning representations of DNA/RNA sequences that generalize well to downstream tasks like promoter classification or SNP impact prediction.

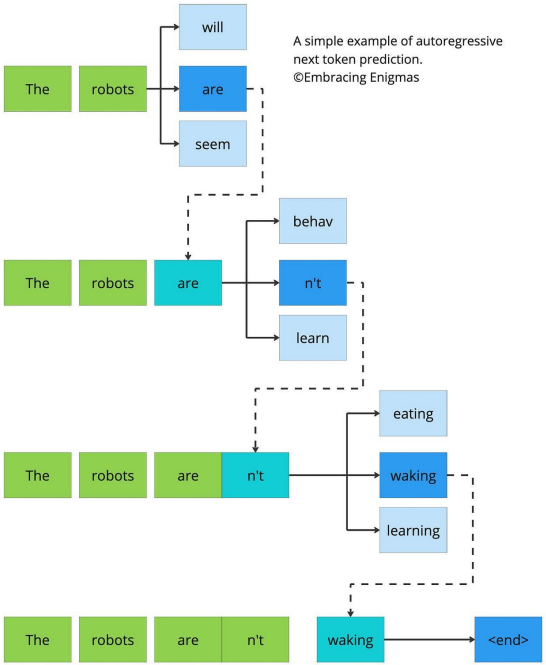


Masked Sequence Modelling

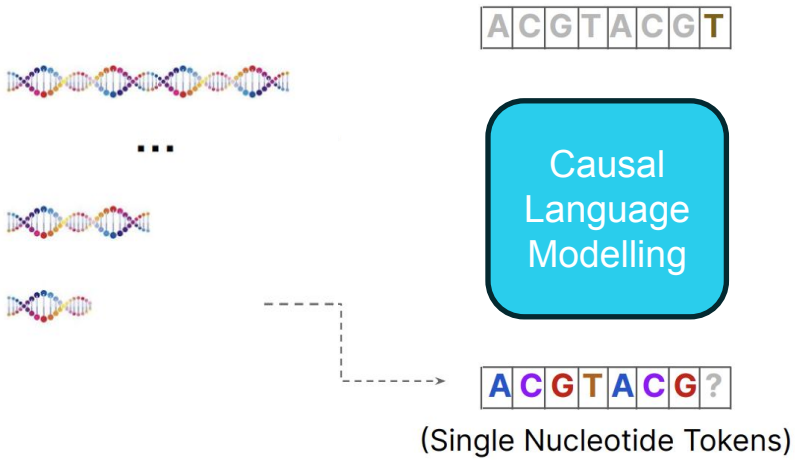


Next Token Prediction / Causal Language Modelling

- **How it works:** Predict the next nucleotide in a sequence, learning dependencies in a unidirectional manner.
- **Use case:** Less common than MLM but still explored in some autoregressive sequence generation tasks



Next Token Prediction / Causal Language Modelling



Additional Pre-training strategies

- Masked Methylation, histone marks (chemical modifications to histone proteins that affect chromatin structure and gene expression)
- Masked Hi-C submatrices (High-throughput chromosome conformation capture)
    - Hi-C measures the frequency at which two DNA fragments physically associate in 3D space, linking chromosomal structure directly to the genomic sequence
- Multi-modal and Multi-task Pre-training
    - Models are pre-trained on multiple data modalities (e.g., DNA + histone marks + chromatin accessibility) or tasks (predicting chromatin accessibility and TF binding simultaneously).

Downstream task for fine-tuning foundation models

Enhancer Prediction

Goal: Identify DNA regions that act as enhancers (regulatory sequences that boost transcription of target genes).

Challenge: Enhancers are not near the genes they regulate, can act in a tissue-specific manner, and lack a fixed sequence pattern.

Input Data: DNA sequence, chromatin accessibility (e.g., ATAC-seq), histone modifications, transcription factor binding (ChIP-seq).

Output: Probability that a region is an enhancer.

Downstream task for fine-tuning foundation models

Promoter Prediction

Goal: Detect promoter regions — sequences upstream of genes where transcription is initiated.

Challenge: Promoters vary in sequence and may not have clear motifs.

Input Data: DNA sequence, transcription start sites (TSS), CpG islands, histone marks.

Output: Prediction of whether a given region is a promoter.

Downstream task for fine-tuning foundation models

Epigenetic Mark Prediction

Goal: Predict presence of specific epigenetic modifications (like DNA methylation or histone modifications).

Challenge: Epigenetic states are cell-type-specific and context-dependent.

Input Data: DNA sequence, chromatin features, 3D genome organization, known epigenetic profiles.

Output: Signal intensity or presence/absence of specific marks at genomic locations.

Downstream task for fine-tuning foundation models

Splice Site Prediction

Goal: Identify where RNA splicing occurs — the start (donor site) and end (acceptor site) of introns.

Challenge: Splice sites have some conserved motifs (e.g., GT...AG), but many false positives exist.

Input Data: DNA/RNA sequence, conservation data, RNA-seq evidence.

Output: Prediction of true splice sites and their likelihood.

Downstream task for fine-tuning foundation models

Regulatory Element Prediction

Goal: Broader task of identifying any regulatory DNA elements: enhancers, silencers, insulators, promoters, etc.

Challenge: Elements have diverse and overlapping features, and regulatory logic is complex.

Input Data: Sequence, chromatin data, transcription factor binding.

Output: Classification of genomic regions into regulatory types.

Downstream task for fine-tuning foundation models

Chromatin Profile Prediction

Goal: Predict the pattern of chromatin accessibility or marks (e.g., from DNase-seq, ATAC-seq, or histone modifications).

Challenge: Chromatin states depend on the cell type, developmental stage, and environment.

Input Data: DNA sequence, context from similar cell types.

Output: Signal tracks showing accessibility or histone marks across the genome.

Downstream task for fine-tuning foundation models

Species Classification

Goal: Identify the species an unknown DNA, RNA, or protein sequence belongs to.

Challenge: Requires discriminating between similar species, horizontal gene transfer, or contamination.

Input Data: Sequence data (e.g., rRNA, whole-genome, k-mers).

Output: Predicted species or taxonomic classification.

DNABERT

Ji, Y., Zhou, Z., & Dai, Q. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. BioRxiv, 2020-11.

1. **Architecture** Transformer (encoder only)
2. **Model Complexity** 110 million parameters
3. **Pre-training Approach** Masked Language Modelling
4. **Training Data** Human reference genome (Hg38.p13)
5. **Downstream Tasks** Promoter region prediction, TF binding site identification, functional genetic variant identification, enhancer prediction, genomic variant annotation...
6. **Results & Interpretation**
   - Achieved high accuracy in predicting promoter regions (99.9%),
   - splice sites (donor accuracy: 0.9491, acceptor accuracy: 0.9345) [8],
   - Demonstrates the ability to capture global and transferable understanding of genomic DNA sequences.

DNABERT2

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. In ICLR 2024

1. **Architecture:** Transformer (encoder only)
2. **Model Complexity:** 117 million parameters
3. **Pre-training Approach:** Masked Language Modelling
4. **Training Data:** Human genome, and multi-species data
5. **Downstream Tasks :** Core/proximal promoter prediction , epigenetic mark prediction, TF binding site prediction, splice site prediction, enhancer prediction
6. **Results & Interpretation:**
    - Improved computational efficiency
    - Ability to handle longer input sequences
    - Consistent performance across human genome-related tasks.

GROVER

Sanabria, M., Hirsch, J., Joubert, P.M. et al. DNA language model GROVER learns sequence context in the human genome. Nat Mach Intell 6, 911–923 (2024)

1. **Architecture:** Transformer (encoder only) – byte-pair encoding
2. **Model Complexity:** 1.2 billion parameters
3. **Pre-training Approach:** Masked Language Modelling, Next-k-mer prediction
4. **Training Data:** Eukaryotic DNA (386B bp) from RefSeq database, human genome
5. **Downstream Tasks:** Genomic function prediction, promoter identification, protein binding site identification, epigenetic information identification, protein-DNA binding prediction
6. **Results & Interpretation**
    - Learns biological and epigenetic information directly from DNA sequence.
    - Can generate protein-coding sequences.
    - Exceeds other models' performance on promoter identification and protein-DNA binding.

Nucleotide Transformer

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat Methods 22, 287–297 (2025)

1. **Architecture:** Transformer (encoder only)
2. **Model Complexity:** 50 million to 2.5 billion parameters
3. **Pre-training Approach:** Masked Language Modelling
4. **Training Data:** Thousands of human genomes (3,202) and hundreds of genomes from various other species (850)
5. **Downstream Tasks:** Molecular phenotype prediction, enhancer prediction, promoter prediction, epigenetic mark prediction, splice site prediction, genetic variant prioritization
6. **Results & Interpretation:**
    - Yields context-specific representations of nucleotide sequences.
    - Achieves accurate predictions with limited task-specific data.
    - Excels in epigenetic modification detection.

HyenaDNA

Nguyen, E., Poli, M., Ré, C., & Heckerman, D. (2023). HyenaDNA: Long-Range Genomic Sequence Modelling at Single Nucleotide Resolution. In NeurIPS 2023

1. **Architecture:** Hyena-CNN (decoder only)
2. **Model Complexity:** 1.6 million parameters
3. **Pre-training Approach:** Next token prediction
4. **Training Data:** Human reference genome
5. **Downstream Tasks:** Enhancer prediction, promoter prediction, epigenetic mark prediction, splice site prediction, regulatory element prediction, chromatin profile prediction, species classification
6. **Results & Interpretation:**
    - Models long-range DNA interactions at single-nucleotide resolution (1 million tokens context length)
    - Achieves state-of-the-art performance on various genomic benchmarks with fewer parameters and less pre-training data than other models.
    - Excels in runtime scalability and handling long input sequences.

Evo

Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, Sullivan J, Ng MY, Lewis A, Lou A, Ermon S, Baccus SA, Hernandez-Boussard T, Ré C, Hsu PD, Hie BL. Sequence modeling and design from molecular to genome scale with Evo. Science. 2024

1. **Architecture:** StripedHyena (decoder only)
2. **Model Complexity:** 7 billion parameters
3. **Pre-training Approach:** Next token prediction
4. **Training Data:** Prokaryotic and eukaryotic genomes (Evo 2)
5. **Downstream Tasks:** Essential gene prediction, prokaryotic promoter prediction, zero-shot protein function prediction
6. **Results & Interpretation**
    - Generalizes across DNA, RNA, and proteins.
    - Can generate sequences exceeding one megabase.
    - Achieves competitive zero-shot performance on function prediction.
7. **Multimodal Approach**

MethylGPT

Ying K, Song J, Cui H, Zhang Y, Li S, Chen X, Liu H, Eames A, McCartney DL, Marioni RE, Poganik JR, Moqri M, Wang B, Gladyshev VN. MethylGPT: a foundation model for the DNA methylome. bioRxiv. 2024

1. **Architecture:** Transformer
2. **Model Complexity:** Varies (tiny: 3M, small: 7M, medium: 15M parameters)
3. **Pre-training Approach:** Masked Language Modelling
4. **Training Data:** 226,555 human DNA methylation profiles from 5,281 datasets (EWAS Data Hub and Clockbase)
5. **Downstream Tasks:** DNA methylation pattern modelling , disease risk prediction, therapeutic effect evaluation, age prediction, mortality prediction
6. **Results & Interpretation:**
    - Captures biologically meaningful age-dependent changes in methylation regulation.
    - Resilient to missing data.

CpGPT

Lucas Paulo de Lima Camillo, Raghav Sehgal, Jenel Armstrong, Albert T. Higgins-Chen, Steve Horvath, Bo Wang. CpGPT: a Foundation Model for DNA Methylation. bioRxiv. 2024

1. **Architecture:** Transformer architecture
2. **Model Complexity:** Not explicitly stated
3. **Pre-training Approach:** Multi-task learning (beta value prediction)
4. **Training Data:** Over 1,500 DNA methylation datasets, encompassing more than 100,000 samples from diverse tissues and conditions
5. **Downstream Tasks:** Imputation and reconstruction of genome-wide methylation profiles , chronological age prediction, mortality risk assessment, morbidity assessments, cancer prediction and classification
6. **Results & Interpretation:**
    - Outperforms specialized models in aging-related tasks.
    - Highly adaptable across different methylation platforms and tissue types.
    - Achieved strong performance in the Biomarkers of Aging Challenge.
    - Identifies CpG islands and chromatin states without supervision.

Hi-C Foundation

Wang X, Zhang Y, Ray S, Jha A, Fang T, Hang S, Doulatov S, Noble WS, Wang S. A generalizable Hi-C foundation model for chromatin architecture, single-cell and multi-omics analysis across species. bioRxiv [Preprint]. 2024

1. **Architecture:** CNN-based
2. **Model Complexity:** No explicitly stated
3. **Pre-training Approach:** MLM (Self-supervised learning on masked Hi-C submatrices)
4. **Training Data:** Hundreds of Hi-C assays, more than 118 million contact matrix submatrices
5. **Downstream Tasks:** Resolution enhancement , loop detection, epigenomic activity prediction, single-cell Hi-C data analysis
6. **Results & Interpretation**
    - Significant advancement in the field of 3D genomics

Conclusion

- Explosion of biomedical data necessitates models that can handle heterogeneous, high-dimensional, and unstructured data like DNA sequences, clinical records, imaging, and more.
- Foundation models, pretrained on vast datasets using self-supervised strategies (e.g. masked language modeling), offer transfer learning, reduced need for feature engineering, better generalization, and multimodal data integration compared to traditional ML.
- State-of-the-art models like DNABERT, GROVER, Nucleotide Transformer, HyenaDNA, Evo, MethylGPT, CpGPT, and Hi-C Foundation demonstrate strong performance across diverse tasks: promoter/enhancer prediction, splice site recognition, chromatin structure analysis, species classification, and methylation-based aging predictions.
- These models are more interpretable, more scalable, and better equipped to enable real-world applications in genomics, disease risk prediction, drug discovery, and personalized medicine.

EURO 4SEE

Thanks!

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101191697. The JU receives support from the Digital Europe Programme and Germany, Türkiye, Republic of North Macedonia, Montenegro, Serbia, Bosnia and Herzegovina.