

# Biological Inductive Biasing of LLM for Aussie Crops Properties Prediction

## Research Aims:

- Define methods for inductive bias extraction and including that extracted inductive bias in prompts to improve LLM performance.
- Identify the K-mers based on the chemical and physical properties that are most tailored to the LLM model
- Evaluation criteria and let the LLM identify its preferred genome, K-mer.
- Generate improved numeric scores from LLMs that are used on genomes to evaluate the generated sequence/3D structure.
- Applying the ML method in case studies, Application downstream tasks: using the defined LLM inductive bias in prediction.
- LLM framework that can generate or predict the sequence or 3D structure of a protein.

## Phase 1: Data Preparation

### Data preprocessing:

Whole-genome sequencing bioinformatics workflow: This workflow is used to obtain sequence or SNP data for the proposed ML method.

The dataset comprises an SNP, sequence or a 3D structure, in the form of a FASTA or PDB file.

The dataset used to train the LLM could be for wheat, chickpeas, or canola.

### Data Tokenisation:

Figure 1 illustrates two methods of tokenisation used in genomic language models: k-mer tokenisation for nucleotide sequences and expression-based

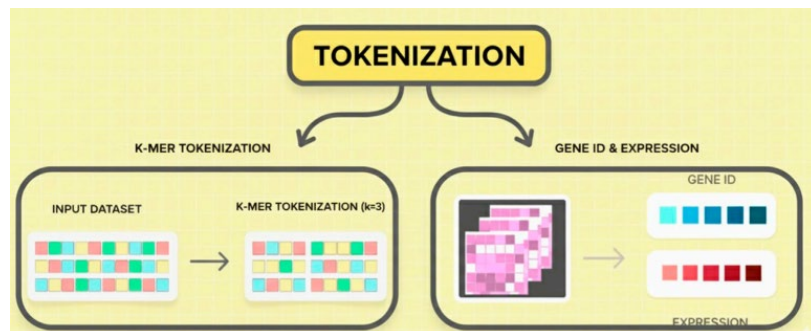


Fig.1 Genome Tokenisation approach

## Phase 2: Biological Inductive Biasing in LLM

### 3.8 Inductive Biasing in Neural Networks

The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered. Inductive bias is the set of rules that governs how an algorithm learns one pattern rather than another. NN's inductive bias in this research is based on restriction and preference bias.

#### 3.8.1 Restriction Bias

Limits the random signal values used in the NN and the model's output values. The restriction bias will be implemented within the LLM Transformer.

#### 3.8.2 Preference Bias

All hypotheses are possible, but some are preferred. But we will use a novel activation function for a neural network that limits the NN output for s specific values. We will use the Smooth Step activation function (SST), as implemented in the Preference Neural Network.

The bias toward a realistic biological structure is used in data simulation. Inductive biasing in a neural network will be implemented from scratch within the LLM transformer architecture shown in Fig. 2.

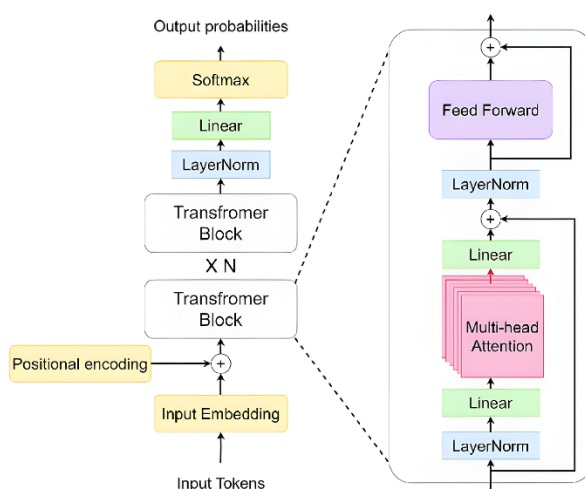


Fig 2: Transformer Architecture

#### 2.1 Identify the biological rules:

This step involves collecting all possible chemical and physical properties of the 4 bases, residues, and the protein's 3D structure. To be the rules that are used within LLM.

#### 2.4 Transformer output biasing. Using a restricted output activation function

The transformer output may be constrained by biased data.

### Phase 3: Case Studies: Aussi Crop Properties Prediction

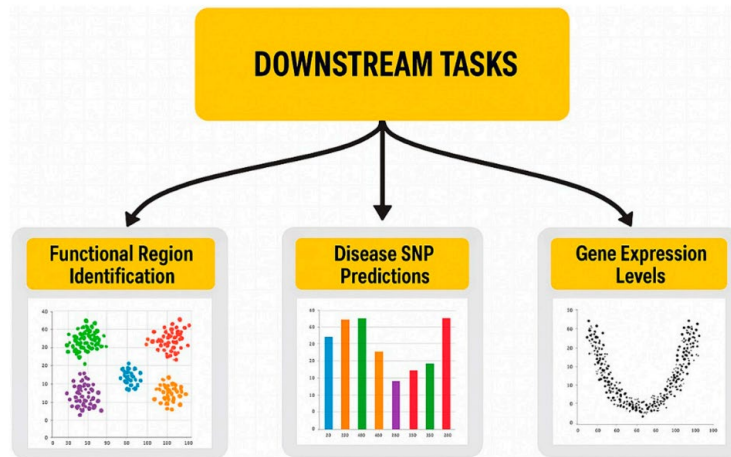


Fig.3 LLM Tasks for the research

The third phase will be a case study for the ML LLM task, using function region identification, SNP prediction, or Gene expression level.

The ML Task, based on the CSIRO datasets for wheat, Chickpea, or Canola, predicts yield, flowering time, and physical and chemical properties as shown in Fig 4.

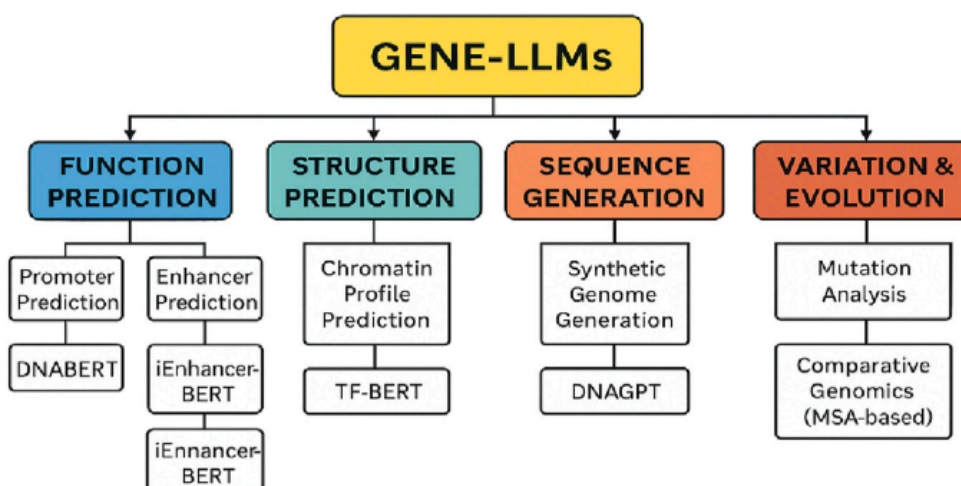


Fig 4: Categorization of ML LLM Tasks

## Skills Required

Strong programming skills to implement transformers from scratch and apply biasing.  
Programming language: Python.

Member	Role
Dr. Shannon Dillon Shannon.Dillon@csiro.au	Principal Supervisor
Dr. Ayman Elgharabawy Ayman.gh@anu.edu.au	Principal investigator

## References

[1] C. M. Angel and F. Ferraro, "Inductive bias extraction and matching for llm prompts," ArXiv, vol. abs/2508.10295, 2025.

[2] P. Balakrishnan, A. A. Leema, V. D. Shree, C. M. Saad, A. M. Babu, Y. Zhang, S. Ali, S. Roy,

D. Shree, M. Saad, and M. Babu, "Gene-llms: a comprehensive survey of transformer-based genomic language models for regulatory and clinical genomics," *Frontiers in Genetics*, vol. 16, 2025.