ARC Research Hub for Protected Cropping

Project 4_3 Phenotype Foundation Models for Molecular to Field Prediction in Sustainable Agriculture

PC Hub Theme (see list here)
Theme 4 - Digital Agriculture and Machine Learning
Principal supervisor
Prof. Wei Xiang
Co-supervisor(s)
TBC

1. Background and Motivation

Agricultural productivity and resilience are governed by a complex hierarchy from molecular interactions within cells to observable crop phenotypes such as yield, stress tolerance, and nutrient content. Despite massive advances in omics technologies (genomics, metabolomics, proteomics), the translation of molecular knowledge into phenotypic understanding remains a major bottleneck.

Traditional statistical or machine learning models typically rely on small, species-specific datasets, which limits generalization across environments and crops. Meanwhile, foundation models (FMs), large pre-trained architectures capable of integrating multimodal data, offer a new paradigm: learning universal molecular and environmental representations that can generalize from laboratory to field.

This project aims to develop a Phenotype Foundation Model (PheFM) that predicts and interprets plant phenotypes directly from molecular and environmental data, thereby accelerating sustainable agriculture, crop improvement, and precision breeding.

2. Research Aim and Objectives

Overall Aim:
To develop a unified foundation model that learns molecular-to-phenotype relationships across species and environments, enabling explainable and generalizable phenotype prediction.

Specific Objectives:

- **Data Integration:** Curate and harmonize multi-omics (genomic, metabolomic, transcriptomic) and environmental datasets with recorded phenotypes across major crops.
- **Model Architecture:** Design a multimodal foundation model that learns joint representations from molecular, environmental, and physiological data.
- **Phenotypic Prediction:** Implement fine-tuned prediction heads for key traits such as yield, stress tolerance, and nutrient content.
- **Interpretability and Transferability:** Develop attention- and gradient-based interpretability methods to reveal causal molecular drivers and test transfer learning across species.
- **Deployment and Validation:** Apply the model to real-world datasets and evaluate against classical phenotype prediction models such as GBLUP and CNN-based omics models.

3. Research Methodology

3.1 Data Sources

- Public Omics Databases: EnsemblPlants, Gramene, Plant Metabolic Network, Phytozome.
- Agricultural Phenotype Repositories: Global Wheat Phenotyping Database, RicePheno, MaizeGDB.
- Environmental and Soil Data: NASA Earth Observation (climate), soil composition datasets, and controlled greenhouse trials.

3.2 Data Preprocessing

- Multi-omics alignment and feature harmonization (gene, metabolite, trait mapping).
- Missing value imputation using diffusion or transformer-based autoencoders.
- Data normalization and spectral embedding for environmental signals.

3.3 Model Design: Phenotype Foundation Model (PheFM)

- **Molecular Encoder:** Graph transformer for genes and metabolites to capture biochemical interactions.
- **Environmental Encoder:** Temporal transformer for weather, soil, and sensor data.
- **Cross-modal Fusion Layer:** Aligns molecular and environmental embeddings through contrastive or attention-based fusion.
- **Phenotype Head:** Predicts continuous and categorical traits such as biomass, disease index, and yield.
- **Self-supervised Pretraining:** Uses masked token and reconstruction objectives on multi-omics corpora.

3.4 Model Explainability

- Attribution maps to identify influential genes or metabolites for given traits.
- Cross-species interpretability analysis linking conserved molecular markers.
- Integration with biological pathways for mechanistic interpretation.

3.5 Evaluation Metrics

- R^2 and RMSE for regression traits; AUROC and F1 score for classification traits.
- Cross-species and cross-environment generalization accuracy.
- Biological interpretability (alignment with known QTLs or metabolic pathways).
- Computational efficiency and scalability to large agricultural datasets.

4. Expected Outcomes

- A Phenotype Foundation Model (PheFM) that generalizes across crops, environments, and omics layers.
- New molecular descriptors and embeddings linking biochemical profiles to phenotypic outcomes.
- Open benchmark datasets and codebase for molecular-to-phenotype AI research.
- Publications in top journals such as *Nature Plants*, *Bioinformatics*, and *Computers and Electronics in Agriculture*.
- Potential application in digital-twin crop modeling and precision breeding.

How to apply

The PC Hub offers a limited number of competitive scholarships.
To apply, please email the following documents to pchub@latrobe.edu.au by **15 February 2026:**

- Application form completed
- Up-to-date CV
- University transcripts

Only shortlisted applicants will be contacted.