Benchmarking a novel bio-informed RNN framework on Bulk RNA-seq data

Christian Langridge[1], Cheng Zhang[1]
1 Roger Williams Institute of Liver Studies at KCL, London, UK

1) Compare **LEMBAS-RNN** prediction accuracy, generalization and explainability to baseline models (**MLR** and **XGBRF**)

Background: Deep Learning (**DL**) models have historically struggled to compete with **tree-based model performance** on biological tabular datasets. A potential hypothesis for poor **DL** generalization is a mismatch between model design choices and the biologically relevant data characteristics of these resources. A novel custom recurrent neural network (**RNN**) architecture, **LEMBAS-RNN**, attempts to address mismatch by integrating a transcription factor (**TF**) motif-based prior generated using **TF-DNA** mapping before employing a vanilla **RNN** to learn edge-node weights from tabular expression data to ultimately predict **TF-target** gene expression values.

2) LEMBAS-RNN relies on a custom, biologically-informed architecture

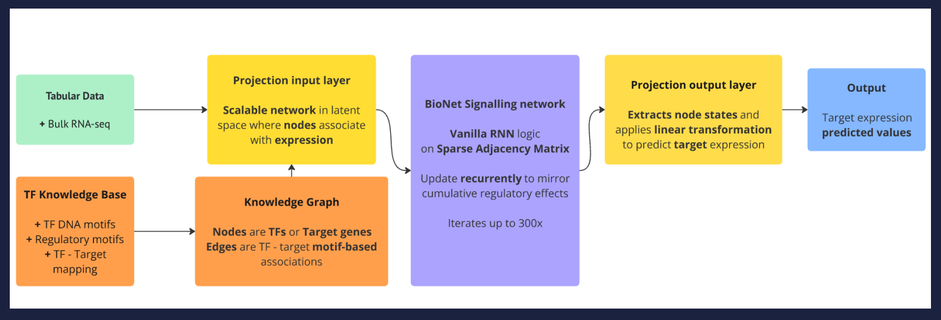

Fig. 1. **LEMBAS-RNN** architecture

Baseline models were Multiple Linear Regressor (**MLR**) and Gradient Boosted Random Forest Regressor (**XGBRF**)

- Training, Testing and Validation data were processed identically
- R^2, Root Mean Squared Error and Mean Absolute Error as central **performance accuracy and error metrics**
- Pearson's r and Spearman's ρ for evaluation **of prediction to ground truth correlation**

3) Model fitting reveals heterogeneous LEMBAS-RNN performance

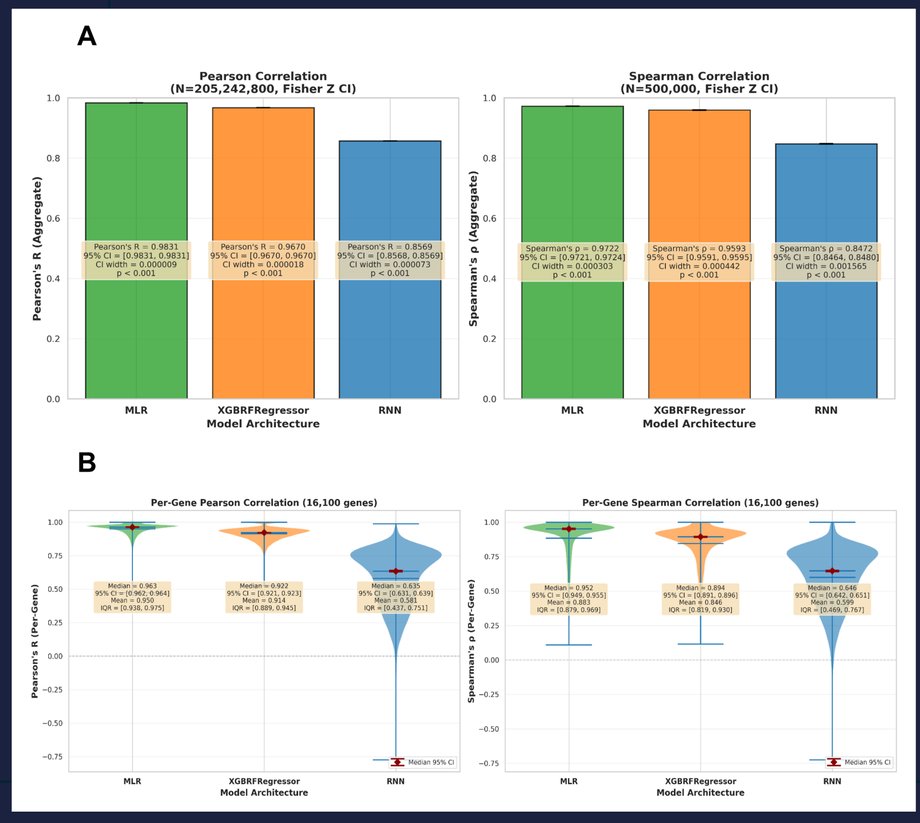

Fig. 2. Model fit performance distribution

LEMBAS-RNN shows overall **lower fit**

Heterogenous performance across all 16k prediction tasks

LEMBAS-RNN fit **well** for a **subset** of gene expression predictions

Fitness **collapses** on a set of **outliers** (**defined by heavy tails**)

4) LEMBAS-RNN captures mechanistic relationships but is outcompeted by tree-based modeling

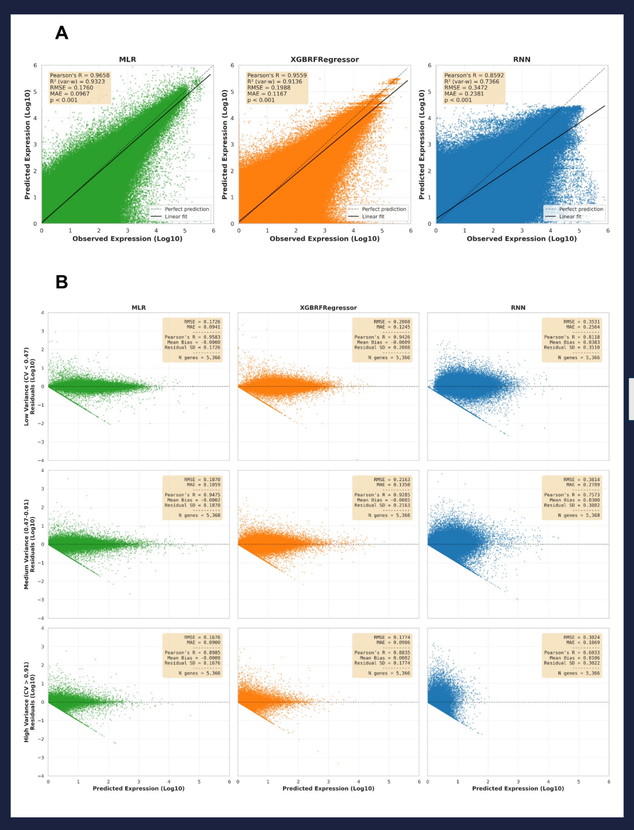

Fig. 3. Model performance on hold-out testing

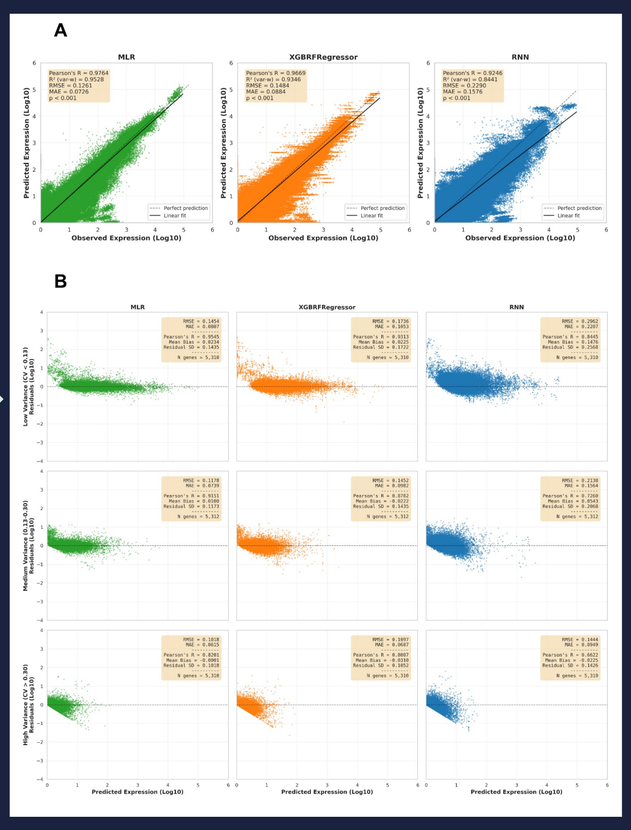

Fig. 4. Model performance on unseen data

Improved model performance on unseen data

- Reduced RMSE:MAE ratio in the unseen data suggests fewer outliers compared to the testing set, implying reduced aleatoric noise in the validation set
- While **LEMBAS-RNN** presents a consistent ratio, a smaller error magnitude supports the hypothesis that even on cleaner data, the model retains the same proportion of outliers
- This overprediction of low-variance genes and under-prediction of high-variance genes represents a systematic limitation of the **LEMBAS-RNN**

5) How does LEMBAS-RNN currently fit into the Transcriptomics research pipeline?

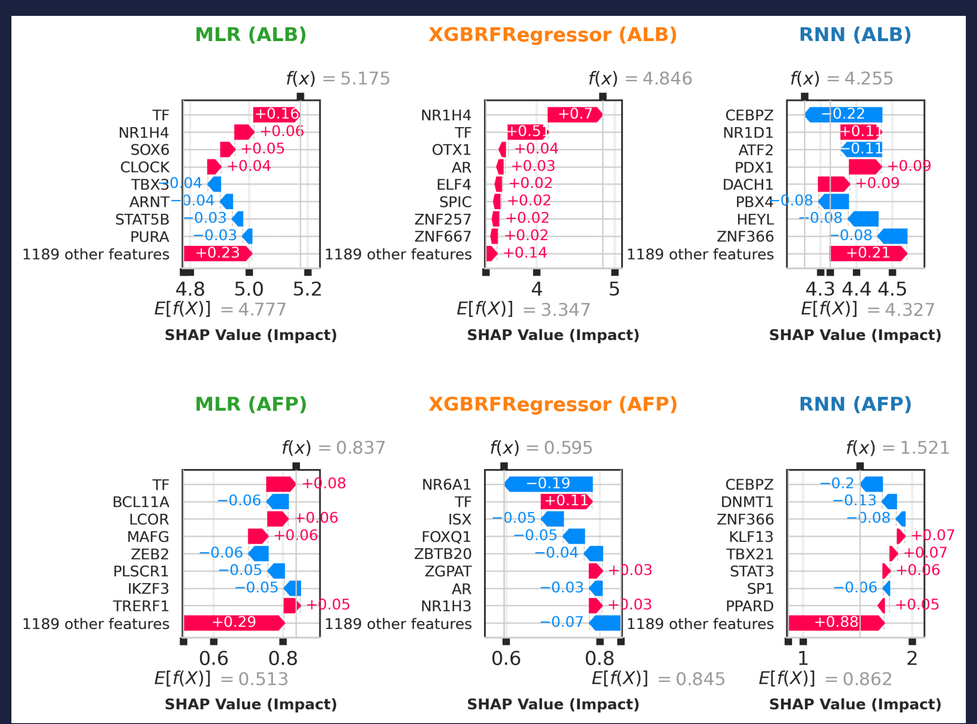

Fig. 5. Model-agnostic feature contribution with SHAP waterfall plotting on unseen data for liver-specific *ALB* and *AFP* expression value predictions

- Currently, **LEMBAS-RNN** in this form will not outcompete **tree-based models** like **XGBRF**
- **Model explainability** shows **LEMBAS-RNN** is navigating real biological relationships (*CEBPZ* strong negative correlation with *ALB* and *AFP* associated with fate change predecessor to hepatocyte fate)[1] 1. Li HM, et al. *Gene*. 2007;389(2):128-35
- The **LEMBAS-RNN** architecture is likely unsuitable for predictions on sc-RNA-seq given its performance on averaged expression values in bulk RNA-seq where noise is less prevalent
- Other neural network engines (Graph Neural Networks) or more sophisticated tabular foundation models (TabICL) could be better adapted for this type of prediction context

UK Research and Innovation
Biotechnology and Biological Sciences Research Council