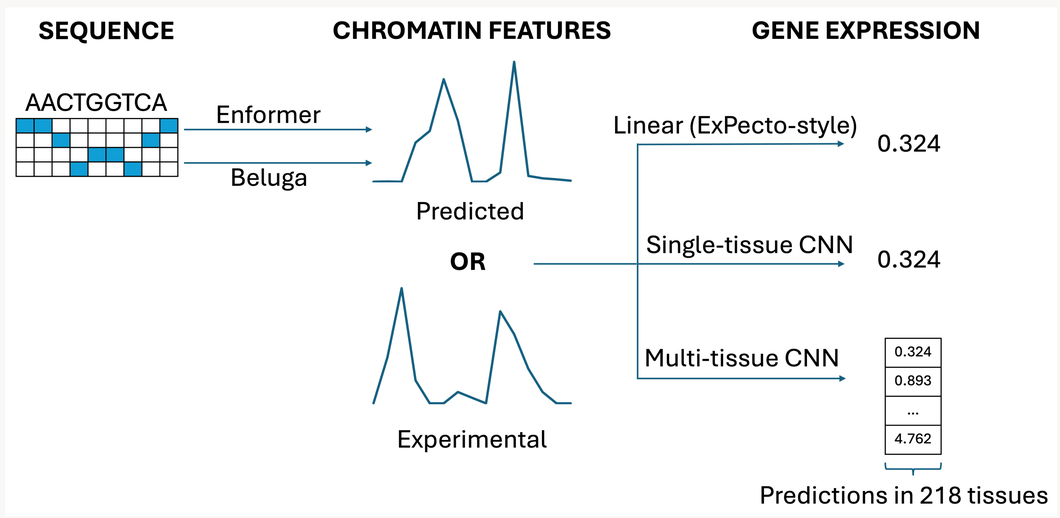
Exploring hierarchical model frameworks for predicting gene expression from sequence

Introduction

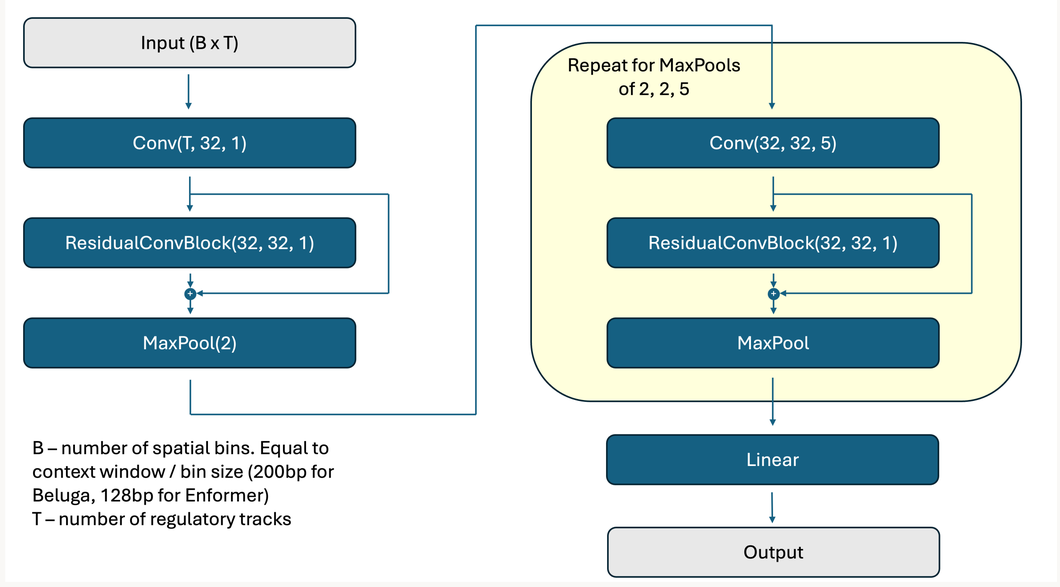
- Sequence-to-expression models often predict chromatin data like transcription factor binding, DNA accessibility, and histone modifications in addition to gene expression. There are two major approaches:
  - **Massively multitask prediction**: Jointly predict both gene expression and chromatin data from sequence (e.g. Enformer).
  - **Hierarchical prediction**: Predict chromatin features from sequence and then predict gene expression from chromatin features (e.g. DeepSEA Beluga + ExPecto).
- Hierarchical prediction offers the following advantages:
  - May better approximate causal biological mechanisms.
  - Separates sequence-to-chromatin prediction, which may depend mostly on local context, from chromatin-to-expression prediction, which may depend on a broader context.
  - New chromatin-to-expression models can be trained on novel data without retraining the sequence-to-chromatin model.

Hierarchical implementation

- *Sequence-to-chromatin models*: we used the published DeepSEA Beluga and Enformer models (for Enformer, we used only chromatin tracks, not CAGE).
- *Chromatin-to-expression input data*: we compared training on chromatin features predicted by Beluga or Enformer with training on the experimental chromatin features from Beluga or Enformer's training datasets.
- *Chromatin-to-expression architectures*: we evaluated three different architectures.
  - Single-tissue ridge regression models (preceded by a spatial transformation of inputs for dimensionality reduction, based on ExPecto)
  - Single-tissue CNNs
  - Multi-tissue CNNs that predict expression across 218 tissues simultaneously

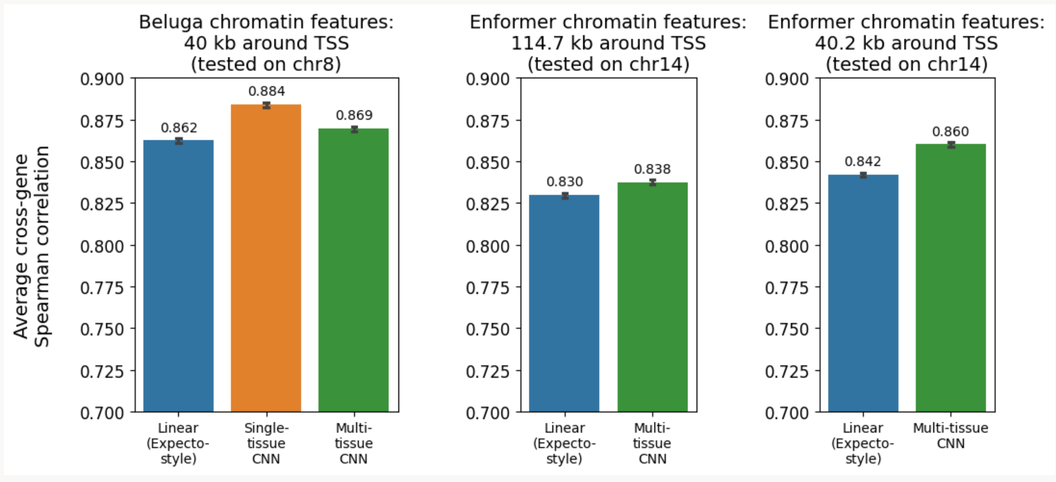


CNN architecture



Performance of chromatin-to-expression models that use experimental chromatin features as input

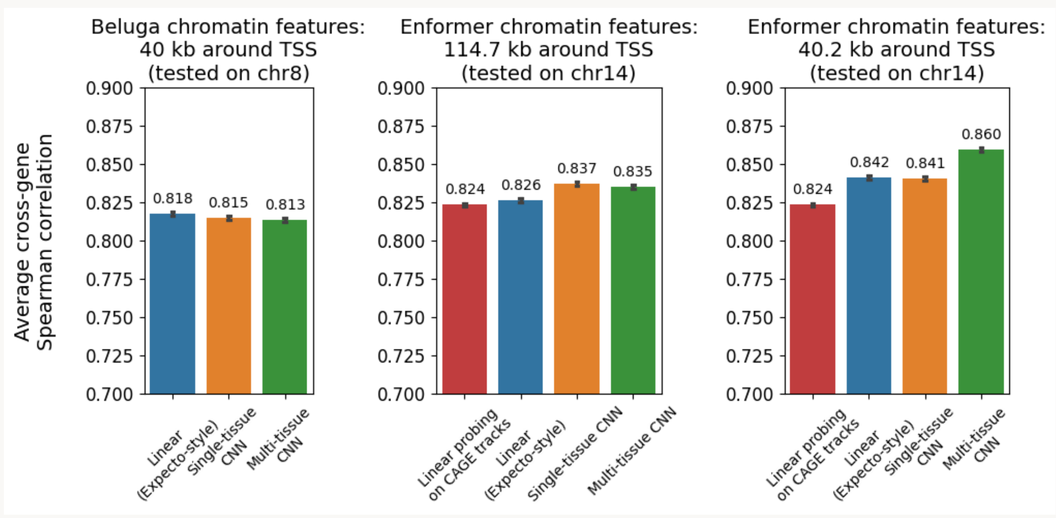
- CNNs outperform linear models.
  - Example: with Beluga experimental features, single-tissue CNNs improve performance (measured by Spearman coefficient) from 0.862 to 0.884 (2.5%).



- Using a smaller input context for the Enformer experimental chromatin features improves performance for both linear models and CNNs.
  - Even though a smaller input context includes fewer regulatory elements, the tested models may have a hard time capturing the effects of distal regulatory elements (in concordance with the findings of Karollus *et al.*, 2023).

Performance of chromatin-to-expression models that use predicted chromatin features as input

- Enformer-predicted chromatin features: CNNs outperform or perform similarly to linear models, using both the full input context of 114.7 kb and shorter context of 40.2 kb.
  - Example: Multi-tissue CNNs outperform linear models by 2.1% (using the shorter input context).
- Beluga-predicted chromatin features: Linear models outperform both single-tissue and multi-tissue CNNs.
  - We hypothesize that simpler models may do better than more complex models if there are sufficient errors in the predicted input features.



- Hierarchical models outperform a naive linear model built on top of Enformer CAGE predictions (for all hierarchical setups).
  - Suggests that converting a non-hierarchical model to a hierarchical model may improve prediction performance.
- Little change in performance when switching from Enformer experimental chromatin features to predicted chromatin features.
  - Indicates that Enformer has a strong understanding of chromatin regulation.

Methods

- Trained models using RNA-seq data (log RPKM) from 218 distinct tissue samples across > 20,000 protein-coding genes.
- Performed grid search on validation set performance to determine regularization coefficient for ridge models and learning rate and dropout probability for CNNs.
- Tested on chr8 for models using Beluga predictions or experimental data and chr14 for models using Enformer predictions or experimental data.

Conclusion

- Explored different architectures for chromatin-to-expression models in addition to the linear architecture proposed in ExPecto.
- Optimal chromatin-to-expression model seems to depend on the quality of the input chromatin data.
- Converting non-hierarchical models into hierarchical models improves predictions and should continue to be explored.

Kanav Mittal, Ruchir Rastogi, Nilah Ioannidis

kanavmittal@berkeley.edu, nilah@berkeley.edu

UC Berkeley