nature portfolio

Peer Review File

Estimating Genotype-Tissue Specific Gene Expression Using Hybrid Deep Learning

Corresponding Author: Dr Kevin Truong

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The paper introduces a two-stage framework combining CNN, Transformer, and XGBoost to estimate genotype-tissue specific expression profiles. The authors demonstrate their performance on human and mouse datasets with ablation studies. However, the task definition, novelty, baseline comparison, and several methodological details need clearer explanation to ensure credibility and reproducibility. Specifically,
Major:
1)The manuscript states that “no computational method currently exists to estimate missing profiles at the tissue-specific level.” Please clarify what is meant by “tissue-specific profile estimation.” Prior studies have addressed related problems, such as (i) multi-tissue imputation (e.g., Wang et al., AJHG, 2016; HYFA, Nat Mach Intell, 2023) and (ii) sequence-based expression prediction (e.g., Enformer, Nat Methods, 2021). The authors should clearly explain how their work differs conceptually and methodologically from these approaches and include these representative methods for comparison.
2) The baseline is too simple. Provide implementation details and include stronger comparison methods for fairness.
3) The authors propose a cascaded CNN–Transformer–XGBoost framework. Please justify why this multi-stage design is necessary compared with an end-to-end or other fusion approach. The rationale for each component and its specific advantages should be clearly described in the Introduction, and the corresponding mathematical formulations for all modules should be included in the Methods for completeness and reproducibility.
4) Please specify the biological or mathematical principles underlying its design and clarify how intergene distance is defined. The Methods section should include the complete mathematical formulation, including loss terms, regularization, optimization strategy, and explicit definitions of all variables.
5) Why was the 666-bp threshold chosen? Does it yield the best performance?

Minor:
1) Several statements in lines 42–53 lack sufficient evidence or citation support. Please provide relevant references and discuss these claims in the Related Work or Introduction section for clarity and credibility.

Reviewer #2

(Remarks to the Author)
The manuscript presents a modular hybrid framework for imputing missing genotype–tissue expression (GTEx) profiles by integrating a 1D convolutional neural network for promoter sequence feature extraction, a transformer encoder to model tissue level expression correlations, and an XGBoost regressor that fuses these embeddings with gene orientation and intergene distance; formally, the model learns a mapping from paired promoter sequences, known expression profiles, orientations, and intergene distances to predict unknown tissue specific expression vectors. Built and validated on GTEx data, the approach yields approximately 30% improvements on accuracy versus proximity based baselines—and is demonstrated to recover profiles for lowly expressed RNAs and less characterized genomes, offering a practical, scalable computational alternative to experimental profiling. The following are my comments and suggestions:

1) Prediction of gene expression from genotype is by now a well established research area; the authors should benchmark their proposed model against current state of the art methods to verify methodological advancement.
2) For this study, evaluation on independent cohorts is a critical test of model generalizability; I recommend the authors

extend their experiments with cross cohorts validations to demonstrate robustness.
3) Model interpretation is another important aspect of this study; however, the authors' analysis is limited to four feature classes (e.g., promoter, intergenic distance). Is it possible to further localize the predictive signal to specific genetic variation?
4) Table 1 lists the hyperparameters for XGBoost; are there also corresponding hyperparameters for the preceding CNN and transformer encoder that should be reported? Additionally, how sensitive is the method to its hyperparameter settings?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have no further comments.

Reviewer #2

(Remarks to the Author)
The author has addressed all my questions

Reviewer #1- Comment 1

The manuscript states that "no computational method currently exists to estimate missing profiles at the tissue-specific level." Please clarify what is meant by "tissue-specific profile estimation." Prior studies have addressed related problems, such as (i) multi-tissue imputation (e.g., Wang et al., AJHG, 2016; HYFA, Nat Mach Intell, 2023) and (ii) sequence-based expression prediction (e.g., Enformer, Nat Methods, 2021). The authors should clearly explain how their work differs conceptually and methodologically from these approaches and include these representative methods for comparison.

Response to Reviewer #1- Comment 1

We thank the reviewer for highlighting these important prior studies. We agree that our original wording was overly broad and have revised the manuscript to clarify the specific task addressed in our work and its distinction from existing approaches.

In this study, we address the problem of estimating the collated multi-tissue expression profile of a target gene with no observed tissue expression measurements of the target gene, using the known expression profile of a spatially adjacent reference gene together with genomic features describing their relationship (promoter sequence, intergene distance, and orientation).

Our approach differs from prior methods. Multi-tissue imputation approaches such as Wang et al. (2016) and HYFA estimate missing tissues for genes of individuals that already have partial expression measurements, relying on tissue-tissue correlations within an observed expression matrix. In contrast, our framework estimates collated multi-tissue expression for genes that may be entirely unmeasured by leveraging neighboring gene expression and genomic context. We do not measure tissue expression at the individual level and do not require any partial tissue expression measurements of the target gene. Sequence-based models such as Enformer predict expression directly from DNA sequence (ab initio) (no expression data as input), whereas our approach is guided by neighboring gene expression and genomic context differences shape the collated multi-tissue expression profile of the target gene rather than predicting expression solely from sequence.

In response to this comment, we have revised the manuscript to:

1. Replace the statement implying the absence of prior methods with a more precise description acknowledging imputation and sequence-based approaches.
2. Clearly define the reference-guided estimation task in the Introduction.
3. Explicitly discuss Wang et al., HYFA, and Enformer and clarify their conceptual and methodological differences from our framework.

Reviewer #1- Comment 2

The baseline is too simple. Provide implementation details and include stronger comparison methods for fairness.

Response to Reviewer #1- Comment 2

The proximity imputation for collated multi-tissue expression is simple but is a reasonable baseline to use against a machine learning approach. We can't use the HYFA dataset because the missing tissues for genes of individuals doesn't meaningful change the collated multi-tissue expression. For the Enformer, we were unable to get the equivalent information of collated multi-tissue expression based on their supplied sources. For stronger statistical comparison methods, we agree that comparison against a single distance-based baseline was insufficient. In the revised manuscript, we added a stronger regression baseline using multi-output Ridge regression with the same inputs as the proposed framework, including promoter sequence features, reference gene expression, intergene distance, and gene orientation.

This enables a direct comparison between linear and nonlinear modeling under identical feature conditions. As shown in revised Fig. 4b, Ridge regression improves accuracy compared to the distance-based baseline, confirming the value of integrating multiple genomic features. However, the hybrid CNN-Transformer-XGBoost model consistently achieves lower errors. For example, at 2,500 bps, the hybrid model reduces 1-PCC error from 0.44 (Ridge) to 0.38 (~14%) and MSE from 0.26 to 0.23 (~12%), with similar improvements across larger distances.

We have revised the manuscript to:

1. Describe the Ridge regression baseline in the Results and Methods sections.

2. Update Fig. 4b to include Ridge comparison.

3. Provide quantitative comparisons across all baselines.

These additions provide a more rigorous evaluation and demonstrate that the performance gains arise from nonlinear modeling rather than feature selection alone.

Reviewer #1- Comment 3

The authors propose a cascaded CNN-Transformer-XGBoost framework. Please justify why this multi-stage design is necessary compared with an end-to-end or other fusion approach. The rationale for each component and its specific advantages should be clearly described in the Introduction, and the corresponding mathematical formulations for all modules should be included in the Methods for completeness and reproducibility.

Response to Reviewer #1- Comment 3

We thank the reviewer for this important suggestion. We have revised the manuscript to clarify the rationale for the cascaded CNN-Transformer-XGBoost design and to provide complete mathematical formulations.

In the Introduction, we now explain that the framework integrates heterogeneous biological inputs with distinct properties, including promoter sequences, tissue-level expression, and genomic features. Each module is designed to model a specific modality: the CNN captures regulatory sequence patterns, the transformer models dependencies across tissues, and XGBoost integrates these features to model nonlinear genomic relationships and generate the final prediction. This staged design enables more effective modeling of heterogeneous biological signals than single-stage or end-to-end approaches.

In the Methods, we added explicit mathematical formulations for all components, including promoter encoding with CNN and cosine similarity, transformer-based tissue modeling with multi-head attention, and XGBoost regression with the full training objective and regularization. All variables and parameters are now clearly defined to ensure reproducibility.

These revisions clarify both the design rationale and implementation details of the proposed framework.

Reviewer #1- Comment 4

Please specify the biological or mathematical principles underlying its design and clarify how intergene distance is defined. The Methods section should include the complete mathematical formulation, including loss terms, regularization, optimization strategy, and explicit definitions of all variables.

Response to Reviewer #1- Comment 4

We thank the reviewer for this suggestion. We have revised the Methods section to clarify both the biological rationale and the complete mathematical formulation of the framework.

First, we expanded the problem formulation to explicitly state the biological basis of our approach, namely the well-established vicinity effect, where neighboring genes exhibit coordinated expression due to shared regulatory environments. Based on this principle, the task is formulated as a supervised regression problem that estimates target gene expression using promoter sequence features, reference gene expression, intergene distance, and strand orientation.

Second, we clarified the definition of intergene distance. It is now formally defined using gene genomic coordinates as the distance between gene bodies, providing a strand-independent measure of genomic proximity relevant to regulatory interactions.

Third, we added complete mathematical formulations for the CNN promoter encoder, Transformer expression encoder, and XGBoost regression model, including explicit definitions of all variables, feature representations, loss function (mean squared error), and regularization terms (L1 and L2).

These revisions improve the clarity, rigor, and reproducibility of the manuscript.

Reviewer #1- Comment 5

Why was the 666-bp threshold chosen? Does it yield the best performance?

Response to Reviewer #1- Comment 5

We thank the reviewer for this question. The 666-bp promoter threshold was adopted from our previous systematic evaluation of promoter truncation, which identified 666 bps as the most compact region retaining essential regulatory information. We have clarified this in the Methods section by adding: “Promoter regions were defined as 666-bp sequences upstream of the transcription start site, based on our previous promoter truncation framework, which identified this length as the most compact region retaining essential regulatory information.”

In addition, we performed supplementary analyses (added Supplementary Fig. S4) comparing alternative promoter lengths and observed comparable performance across a reasonable range, indicating that the model is not overly sensitive to this parameter. We have added “To assess the impact of promoter length on estimation accuracy, we evaluated alternative promoter thresholds proposed in our previous work (1,573 and 2,773 bps)...These findings indicate that the model is robust to promoter length selection within this range and support the use of the 666 bps threshold as a compact and efficient promoter definition.” to the “Training results and model performance” section.

Reviewer #2- Comment 1

Prediction of gene expression from genotype is by now a well established research area; the authors should benchmark their proposed model against current state of the art methods to verify methodological advancement.

Response to Reviewer #2- Comment 1

We thank the reviewer for this important suggestion. Most current state-of-the-art models address different problem settings from the one considered in this study. Specifically, prior approaches fall into two main categories:

(i) multi-tissue imputation methods (e.g., Wang et al., HYFA), which require partial expression measurements of the target gene and infer missing tissue expression of same gene at the resolution of the individual, and
(ii) sequence-based prediction models (e.g., Enformer), which predict expression directly from genomic sequence without using gene expression as input.

In contrast, our method addresses a distinct and complementary problem: estimating the collated multi-tissue expression profile of a target gene with no observed expression measurements by leveraging the expression profile of a neighboring reference gene together with genomic features describing their relationship. For benchmarking with these studies, we need a way to extract an equivalent collated multi-tissue expression profile from their published datasets.

We can’t use the HYFA dataset because the missing tissues for genes of individuals doesn’t meaningful change the collated multi-tissue expression. For the Enformer, we were unable to get the equivalent information of collated multi-tissue expression based on their supplied sources. The proximity imputation for collated multi-tissue expression is simple but is a reasonable baseline benchmark to use against a machine learning approach. For stronger statistical comparison methods, we agree that comparison against a single distance-based baseline was insufficient. In the revised manuscript, we added a stronger regression baseline using multi-output Ridge regression with the same inputs as the proposed framework, including promoter sequence features, reference gene expression, intergene distance, and gene orientation.

This enables a direct comparison between linear and nonlinear modeling under identical feature conditions. As shown in revised Fig. 4b, Ridge regression improves accuracy compared to the distance-based baseline, confirming the value of integrating multiple genomic features. However, the hybrid CNN-Transformer-XGBoost model consistently achieves lower errors. For example, at 2,500 bps, the hybrid model reduces 1-PCC error from 0.44 (Ridge) to 0.38 (~14%) and MSE from 0.26 to 0.23 (~12%), with similar improvements across larger distances.

We have clarified these distinctions and expanded these related work in the Introduction and Results sections accordingly.

Reviewer #2- Comment 2

For this study, evaluation on independent cohorts is a critical test of model generalizability; I recommend the authors extend their experiments with cross cohorts validations to demonstrate robustness.

Response to Reviewer #2- Comment 2

We thank the reviewer for this valuable suggestion and agree that independent validation is essential to assess model generalizability.

To address this, we performed chromosome-level five-fold cross-validation, in which entire chromosomes were excluded from training and used exclusively for testing. The model demonstrated consistent performance across folds (mean 1–PCC

= 0.3988; mean MSE = 0.2175; Supplementary Table S1), indicating robust generalization to unseen genomic regions. In addition, we evaluated the model on an independent mouse RNA-seq dataset (Supplementary Fig. S2), representing a different species and transcriptomic context. The hybrid model consistently outperformed the distance-based baseline, achieving approximately 20% improvement. Together, these results demonstrate that the proposed framework generalizes beyond the original GTEx dataset. We have clarified these analyses in the revised manuscript and added "To evaluate model generalizability…To further assess whether the model captures more fundamental regulatory relationships beyond the specific human dataset, we next evaluated its performance on an independent mouse RNA-seq dataset" to the "Training results and model performance" section. We have also added "To evaluate model generalizability while preventing genomic proximity leakage, … This procedure was repeated for all five folds, and the mean and standard deviation of estimation errors were reported in Supplementary Table 1" to Methods.

Reviewer #2- Comment 3

Model interpretation is another important aspect of this study; however, the authors' analysis is limited to four feature classes (e.g., promoter, intergenic distance). Is it possible to further localize the predictive signal to specific genetic variation?

Response to Reviewer #2- Comment 3

We thank the reviewer for this insightful suggestion. Our current model does not incorporate individual-level genetic variation as input and therefore does not directly attribute predictions to specific variants. However, the proposed framework is modular, and genetic variation features such as SNPs or sequence variants can be naturally incorporated as additional inputs to the CNN or regression module. This would enable variant-level interpretation in future extensions.

Reviewer #2- Comment 4

Table 1 lists the hyperparameters for XGBoost; are there also corresponding hyperparameters for the preceding CNN and transformer encoder that should be reported? Additionally, how sensitive is the method to its hyperparameter settings?

Response to Reviewer #2- Comment 4

We thank the reviewer for this suggestion. In the revised manuscript, we now report the hyperparameters for all model components to ensure full reproducibility. Specifically, promoter CNN, transformer encoder, and XGBoost hyperparameters are summarized in Tables 1-3, respectively.

These parameters were selected based on validation performance within reasonable search ranges. We observed that small changes around the selected values resulted in similar performance, indicating that the model is not highly sensitive to precise hyperparameter settings.