

EPInformer: scalable and integrative prediction of gene expression from promoter-enhancer sequences with multimodal epigenomic profiles

Corresponding Author: Dr Luca Pinello

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)

Lin et al proposed EPInformer, a transformer-based deep learning framework that predicts gene expression by integrating promoter-enhancer interactions along with their sequences, epigenomic signals, and chromatin contacts. In their paper, they showed that the model outperforms existing gene expression prediction models in cross-chromosomal validation. Moreover, EPInformer can accurately identify enhancer-gene interactions that have been validated through CRISPR perturbation experiments. They also used some pipelines to find some key motifs that highly relevant to enhancers. However, this article has the following key issues and flaws that prevent it from being published in Nature Communications in its current version.

Major Concerns:

The novelty of the model is limited and not sufficient for publication in Nature Communications. There are increasingly more methods for sequence modeling, such as DNABERT-2[1], Nucleotide Transformer[2], HyenaDNA[3] etc., which can perform a variety of tasks beyond gene expression prediction, including epigenetic marks prediction, promoter and enhancer prediction, splice site prediction. In comparison, the tasks that this model has covered are relatively limited.

The benchmarking of the model is not comprehensive. As mentioned earlier, the emergence of large genomic sequence models has made the field highly competitive. However, in the comparison experiments of Fig2, the authors only selected four methods, most of which were published before 2022, clearly lagging behind the current progress in the field. Please appropriately increase comparisons with large DNA models (at least three), in addition to the aforementioned DNABERT-2, Nucleotide Transformer, HyenaDNA, etc. The following review can serve as a reference for selecting baselines for comparison [4]. Additionally, please summarize and list any other advantages of this manuscript's model over large models, apart from performance.

The improvement in the model's performance is not significant. For example, in Fig3bc, the identification of enhancer-promoter interactions is not good. Compared to traditional methods (calculating the ABC score), the attention score with and without HiC is worse than the traditional method, and even multiplying the Attention score by the ABC score results in a minimal improvement, indicating that the development of this model may not be very meaningful.

In the experiment to verify cell-type-specific enhancer prioritization, the authors only used CRISPRi-FlowFISH results from the K562 cell line as labels. However, the role of enhancers is cell-type specific, and we do not know the performance of EPInformer in other cell lines. If possible, at least one other cell line with CRISPR-perturbation experimental labels should be added for verification.

The authors' interpretation of the experimental results is lacking. For example, in Fig2, there are still some housekeeping gene expression predictions that are inaccurate, such as the GADPH gene. Please provide a detailed explanation of the reasons. Additionally, the authors did not elaborate on the role of the red gene names in the manuscript. Fundamentally, please indicate what cell types are present in the K562 cell line and which of these red genes are specific to these cell types; the original text barely explains this. In Fig4c, the authors did not include a comparison with traditional methods, nor did they further analyze the significance of unknown motifs, which is disappointing.

The argumentation in the article is not sufficient. In Fig3 and Fig4, the authors only used the Klf1 gene for interpretability analysis of the model. The authors did not explain why this gene was chosen, and choosing just one gene is not representative. More valid examples (at least five) should be selected to demonstrate that the model can indeed identify the regulatory relationships of many genes.

Some concepts are not clearly articulated, such as why mRNA half-life data is used as a promoter feature input to the model to predict expression levels. The paper does not explain the rationale for using this, nor does it further explain what mRNA half-life measurements are and what characteristics they represent. Additionally, why not consider using ATAC-seq data as input for the FUSION LAYER to characterize enhancer accessibility? This is a more widely applicable measurement method, and I believe there is plenty of existing scATAC-seq data for the K562 cell line.

Minor Concerns:

In the section "In EPInformer attention prioritizes cell-type-specific enhancers supported by long-range interaction assays," the authors did not cite Fig3C.

(Remarks on code availability)

Pretty good. there are some ipynb codes.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This study presents EPInformer, an Enformer-based method for predicting gene expression data, with improvement through the integration of promoter-enhancer interactions, enhancer activity, and Hi-C data. The authors demonstrate that EPInformer outperforms previous models in predicting both mRNA and CAGE expression levels. Furthermore, they show that incorporating the ABC score improves the model's performance. They validate their predictions using CRISPR-based perturbation results in experiments, investigate transcription factor motifs, and assess the predictability and utility of EPInformer.

While this study offers valuable insights, I have the following comments:

Major Comments

1. Definition of TSS

The definition of the transcription start site (TSS), a critical aspect of this study, is not clearly stated. Does the TSS refer to the first base of the 5' UTR or the GENCODE-annotated TSS?

2. Insufficient Transcript Coverage

It is unclear whether the model fully captures transcripts originating from the same gene. The authors should evaluate whether EPInformer can be applied to individual transcripts and TSSs, as defined by FANTOM. Given that the ultimate goal is to interpret GWAS associations and understand the genetic basis of complex traits, the model's resolution should be explicitly demonstrated.

3. Comparison with Previous Methods

The comparison with existing methods appears incomplete. How does EPInformer perform relative to Basenji and MENTR? A direct evaluation would strengthen the study's claims of improved accuracy.

4. Lack of Cell-Type-Specific Hi-C Integration

The authors do not appear to incorporate cell-type-specific Hi-C data. While the ABC model (Fulco et al.) also uses non-cell-type-specific Hi-C, EPInformer aims to predict cell-type-specific gene expression levels. Therefore, integrating cell-type-specific Hi-C data would be essential for capturing context-dependent regulatory interactions.

5. Definition of Flanking Distances

The study uses a 200-kb window centered on the TSS to search potential regulatory elements. However, cis-eQTL studies commonly use a 1-Mb window, and trans-eQTLs have been suggested to explain more heritability than cis-eQTLs. The authors should provide justification for using a 200-kb window and assess whether it sufficiently captures the regulatory architecture of gene expression.

6. Fine-Mapping of Variants

To demonstrate the utility of EPInformer in interpreting GWAS associations, the authors should assess whether their method

can aid in fine-mapping causal variants. Specifically, can the model prioritize variants based on enhancer contributions to gene expression?

Minor Comments

7.The authors state, "This threshold covers 95% of potential regulatory elements within a 200-kb region centered on the TSS." How was this value determined?

8.The authors mention, "Across six cell lines and 18,377 genes on test chromosomes." The reported number of genes on test chromosomes seems unexpectedly high. Could they clarify this?

9.Abbreviations should be defined upon their first appearance in the main text.

(Remarks on code availability)

The codes seem reasonable

Reviewer #4

(Remarks to the Author)

In this manuscript, the authors present a method called "EPInformer" for inferring gene expression from regulatory element DNA sequences and additional features. While this method has the potential to enhance our understanding of gene regulation, and the transformer-based approach integrates various data sources, I have concerns, especially for its level of innovation, which I would like to have clarified before recommending it for publication.

1. There are some works that predict gene expression. For example, Enformer ('Effective gene expression prediction from sequence by integrating long-range interactions') also uses a transformer framework. In this context, what are the specific contributions of your method, and how does it improve upon Enformer?

2. The average precision (AP) is higher for the Attention-ABC score compared to the Attention score with Hi-C, as shown in Figure 3c. From my understanding, the ABC score is a straightforward method that combines features. This raises a concern about the proper utilization of Hi-C data.

3. Did the study take into account the negative regulation from regulatory elements to target genes?

4. 'To this end, we used TF-MoDISco-lite30 and TanagerMEME31 to examine the TF motifs within enhancers, based on EPInformer-PE-Activity's sequence encoder pre-trained to predict enhancer activity from the sequence.' In the above sentence, does the 'pre-trained' mean the model for predicting the gene expression? How to get this pre-trained model, and why could it be used to predict the enhancer activity?

5. I did not see the details regarding the methods used for predicting enhancer activity. Is the enhancer activity represented as a single score per base pair or as one score for each enhancer? Additionally, is the input solely the DNA sequence? I apologize if I missed some of these details.

6. Why are some of the expression values smaller than 0 in Figure 1a (right)?

(Remarks on code availability)

The code is usable. I can install and run the code.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for the author's response. Most of the issues have been resolved.

However, there is still one confusing question regarding Figure 2. In the benchmark section, why are the benchmark experiments not comprehensive? For example, Figures 2b and 2d appear to represent the same task (at least I cannot discern any differences from the captions and titles). Why is only Enformer compared in 2b, while only Borzoi is compared in 2d? Additionally, why do the numerical values for the same method differ between 2b and 2d? For instance, EPInformer_PE_Activity_HiC is 0.856 in 2b but 0.892 in 2d. If they represent the same experimental setup, the two figures should be merged and redrawn. If they represent different setups, the differences between the two figures should be clearly distinguished, and baseline experiments should be supplemented (e.g., Borzoi should be added to 2b, while Enformer, Seq-graphReg, etc., should be added to 2d). The same applies to Figures 2a and 2c. Both are cross-chromosome validation tasks, so why is Xpresso included in 2a but replaced by Seq-graphReg in 2c? Placing them in separate figures is confusing. I recommend that the authors maintain consistency in the baseline methods across all figures. Currently, there appear to be five baseline methods in total, so the results of these five baseline methods should be included in all benchmark experiments (including 2a–2d). If this is not feasible or inconvenient to implement, the situation should be explicitly explained (e.g., "Method XXX cannot be included in Figure 2a because...").

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have addressed most of my previous comments satisfactorily. I appreciate their comparison with Borsoi et al. and their decision to extend the analyses to the broader set of FANTOM5 transcription start sites (TSSs), rather than restricting them to main transcripts.

I would like to clarify that MENTR is not a tool specifically designed to infer non-coding RNA expression. It learns from FANTOM5 TSS data, which include both coding (i.e., promoters) and non-coding RNAs. Since the authors have now expanded their analyses to include all FANTOM5 TSSs, a direct comparison between their method and MENTR would be both appropriate and informative.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

All questions have been resolved and happy to see this paper will be accepted in the future!

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer Response

We sincerely thank all reviewers for their thorough and constructive feedback, which has substantially improved our manuscript. In response, we have made several significant enhancements: (1) redesigned the enhanced enhancer–gene interaction modeling by introducing a new attention module that explicitly integrates Hi-C and epigenomic signals (e.g., H3K27ac, DNase) for gene expression prediction, (2) expanded benchmarking using the latest K562 CRISPR perturbation screens and eQTL datasets, (3) broadened our evaluations by comparing against the recently published and state-of-the-art Borzoi model for gene expression prediction and the traditional LS-GKM for enhancer classification, (4) revised the TF motif identification pipeline to generate a comprehensive report of discovered motifs and their annotations, and (5) extended analyses to demonstrate that EPInformer can effectively utilize ATAC-seq for gene expression modeling and to evaluate isoform-specific expression prediction. We believe we have addressed all concerns as detailed in the following sections. Our responses below and all updated manuscript text are highlighted in green for easy reference.

Reviewer #1

Lin et al proposed EPInformer, a transformer-based deep learning framework that predicts gene expression by integrating promoter-enhancer interactions along with their sequences, epigenomic signals, and chromatin contacts. In their paper, they showed that the model outperforms existing gene expression prediction models in cross-chromosomal validation. Moreover, EPInformer can accurately identify enhancer-gene interactions that have been validated through CRISPR perturbation experiments. They also used some pipelines to find some key motifs that highly relevant to enhancers. However, this article has the following key issues and flaws that prevent it from being published in Nature Communications in its current version.

1. The novelty of the model is limited and not sufficient for publication in Nature Communications. There are increasingly more methods for sequence modeling, such as DNABERT-2, Nucleotide Transformer, HyenaDNA etc., which can perform a variety of tasks beyond gene expression prediction, including epigenetic marks prediction, promoter and enhancer prediction, splice site prediction. In comparison, the tasks that this model has covered are relatively limited.

We appreciate the reviewer raising this important point about comparing our work to DNA language models. We agree that Nucleotide Transformer, DNABERT-2 and HyenaDNA have emerged as important approaches for sequence modelling in a self-supervised manner. After consulting with the editor, we focused our benchmarking on methods specifically designed for gene expression prediction to ensure methodologically appropriate comparisons, as motivate below.

We have thoroughly investigated these models, their papers, and available code repositories. While these models demonstrate impressive performance in tasks such as regulatory element detection, transcription factor binding site prediction, gene structure analysis, and chromatin profiling across multiple species in several benchmarks^{1,2}, we found that: 1) these models were not specifically designed for gene expression prediction; 2) there is no publicly available code implementing these models for gene expression prediction tasks; 3) their architectures would require significant modification to be adapted for our specific prediction task. In contrast, our benchmarking against several comparable methods like Enformer, Seq-GraphReg, CREaTor and Espresso is more appropriate as these models were specifically trained on similar data types and focused on the same task of predicting gene expression from sequence and additional genomic tracks for diverse cell lines and tissues.

To further strengthen our comparison against more recent state-of-the-art methods, we have now included

Borzoi³ (Linder et al., *Nature Genetics*, 2025) in our benchmarking, which was published this year and represents the current best performing method for gene expression prediction. Importantly, we demonstrated that EPInformer outperformed Borzoi in CAGE expression prediction in both K562 and GM12878.

We have updated the following text in the manuscript and Figure 2:

“For CAGE expression prediction, EPInformer-PE-Activity-HiC consistently outperformed all EPInformer variants and competing methods in both K562 and GM12878 (Fig. 2b–d). On the held-out test set, it surpassed Enformer in K562 (Pearson’s $r = 0.865$ vs. 0.775) and GM12878 ($r = 0.872$ vs. 0.774 ; Fig. 2b). Although Borzoi outperformed Enformer, EPInformer-PE-Activity-HiC achieved higher performance than Borzoi in K562 ($r = 0.892$ vs. 0.792) and GM12878 ($r = 0.905$ vs. 0.800 ; Fig. 2d). Notably, even without Hi-C data, EPInformer-PE-Activity exceeded Borzoi, yielding higher Pearson correlations in K562 ($r = 0.841$ vs. 0.792) and GM12878 ($r = 0.847$ vs. 0.800). Under 12-fold cross-chromosome validation in K562 and GM12878 (Fig. 2c), EPInformer-PE-Activity-HiC achieved the highest average Pearson correlation ($r = 0.883$), outperforming Seq-GraphReg ($r = 0.744$) and CREaTor ($r = 0.810$).”

Additionally, given the growing importance of DNA language models for genomic sequence modeling, we have added the following text to the introduction section discussing these models and their relationship to our work:

“Notably, large genomic language models have emerged as an important approach for deciphering regulatory grammar from sequence in a self-supervised manner, inspired by the success of large language models. DNABERT-2⁴, Nucleotide Transformer⁵, and HyenaDNA⁶ are among the most popular models and have been extensively benchmarked^{1,2,7}. Although they show promising performance in identifying cell-type-agnostic regulatory elements and genes directly from DNA sequence, they are not designed for cell-type-specific gene expression modeling, which limits their utility for understanding gene regulation in specific cellular contexts. Moreover, genomic language models typically require substantial computational resources for fine-tuning on downstream tasks, which is not feasible for most experimental laboratories.”

2. The benchmarking of the model is not comprehensive. As mentioned earlier, the emergence of large genomic sequence models has made the field highly competitive. However, in the comparison experiments of Fig2, the authors only selected four methods, most of which were published before 2022, clearly lagging behind the current progress in the field.

We thank the reviewer for this valuable suggestion. As discussed in our response to the previous comment, we focused our benchmarking on methods specifically designed for gene expression prediction. Although Enformer and Xpresso predate 2022, they remain widely used and validated models for gene expression prediction; accordingly, we included them as established baselines. To reflect recent advances, we expanded our benchmarking to include Borzoi (published in 2025), in addition to CREaTor (published in 2023). Our experiments show that EPInformer outperforms all baselines, including Borzoi, in predicting CAGE and RNA-seq expression. We have added the comparison results between EPInformer and Borzoi in our previous response. Additionally, we have added the following text to the introduction discussing Borzoi and Enformer:

“Among sequence-based models, Enformer is the most widely used. It excels at predicting gene expression, transcription factor binding, and chromatin states directly from DNA sequence. More recently, an updated architecture, Borzoi, was introduced to model substantially longer genomic contexts (up to 524 kb), improving predictions of gene expression and diverse epigenomic signals. However, both Enformer and Borzoi’s extensive training requirements limit their adaptability to unseen cell types, and their ability to capture regulatory effects from very distal enhancers remains constrained—particularly for elements located more than 10 kb from a gene’s transcription start site (TSS), as previously reported^{8,9}”

We also added the following text to the methods section explain how we compared EPIInformer to Enformer and Borzoi in CAGE expression prediction:

“To ensure a fair comparison, we aligned the training and testing settings of EPIInformer with those of Enformer and Borzoi respectively. This involved using identical data splits and extracting promoter and potential enhancer sequences from the same regions Enformer/Borzoi was trained on. For EPIInformer, gene expression values were determined by summing read counts within a 384-bp window (equivalent to three 128-bp Enformer bins) surrounding each gene’s TSS, using the same data sources (CNhs12333 for GM12878 and CNhs11250 for K562).”

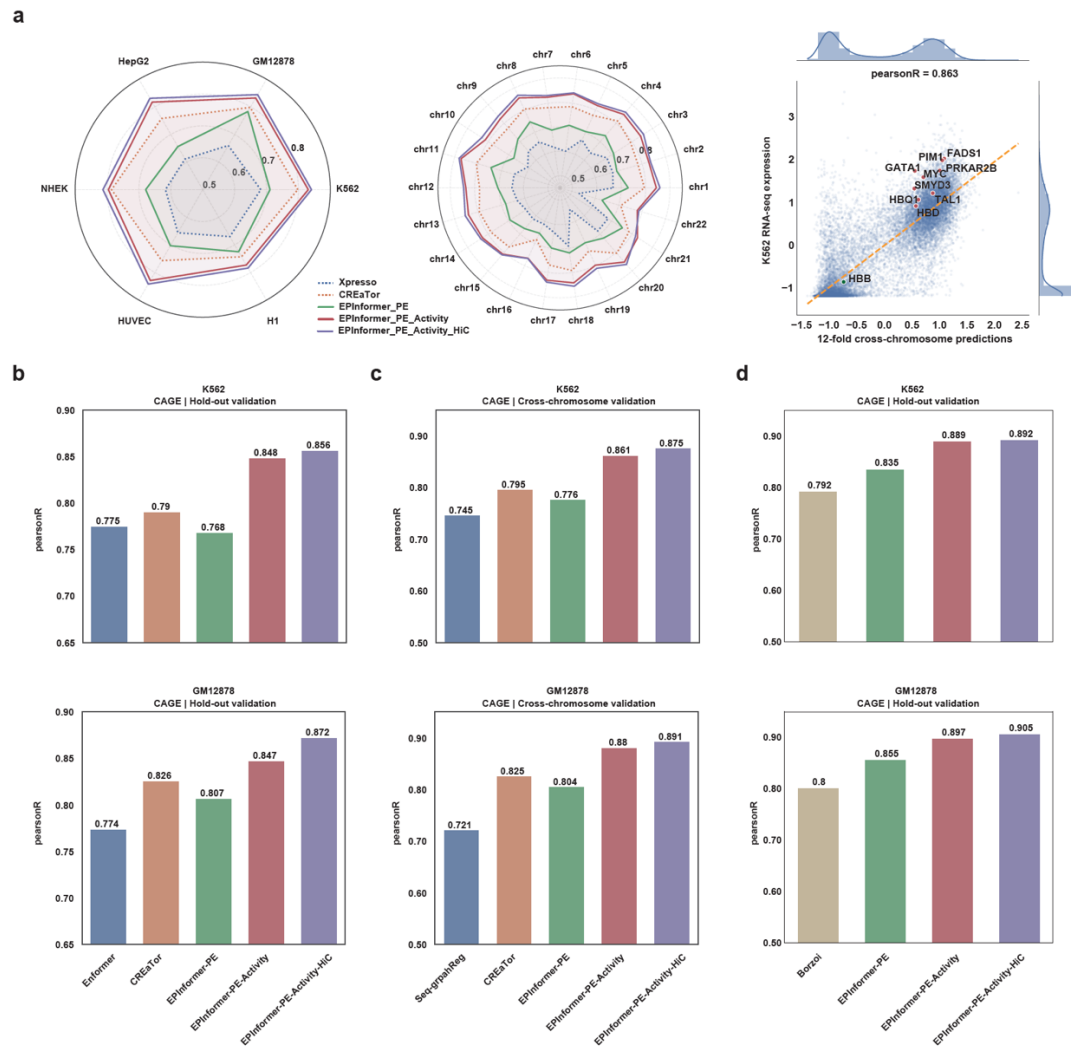


Fig 2. | Performance of EPIInformer models and baseline methods on gene expression prediction. a, Pearson correlation between observed and predicted RNA-seq expression levels for 18,377 protein-coding genes across six cell lines. Left: Spider plot showing correlation between observed and predicted expression levels for six cell lines. Middle: Spider plot showing correlation between observed and predicted expression levels for each test chromosome across six cell lines. Right: Scatter plot showing relationship between EPIInformer-PE-Activity-HiC predicted and observed RNA-seq expression levels in K562 cells, with each point representing an individual gene. Genes with high expression in K562 are highlighted in red, and those with low expression are highlighted in green. All evaluations were conducted using 12-fold cross-chromosome validation. **b,** Comparison of EPIInformer models with Enformer for predicting CAGE gene expression levels on hold-out genes in K562 (top) and GM12878 (bottom) cells. **c,** Comparison of EPIInformer models with Seq-GraphReg and CREaTor for CAGE gene expression prediction in K562 (top) and GM12878 (bottom) cells using 12-fold cross-chromosome validation. **d,** Comparison of EPIInformer models with Borzoi for predicting CAGE gene expression levels on hold-out genes in K562 (top) and GM12878 (bottom) cells.

3. Please appropriately increase comparisons with large DNA models (at least three), in addition to the aforementioned DNABERT-2, Nucleotide Transformer, HyenaDNA, etc. The following review can serve as a reference for selecting baselines for comparison [4]. Additionally, please summarize and list any other advantages of this manuscript's model over large models, apart from performance.

We thank the reviewer for raising this important point about comparing EPInformer to large DNA models. As mentioned in our previous responses and after consulting with the editor, we determined that comparing with methods not specifically designed for gene expression prediction would be outside the scope of our study. We agree that large sequence models have opened a new avenue for understanding regulatory DNA, and they show promising performance in identifying cell-type-agnostic regulatory elements and genes from DNA sequence in recently published benchmarks¹. However, they are not designed for cell-type-specific gene expression modelling, thus these models are not suitable as baselines for our model comparison. As aforementioned, we selected the state-of-the-art sequence-to-expression model, Borzoi alongside other widely used models including Xpresso, Seq-graphReg, Enformer and CREaTor as baselines for comparison. Our experiments demonstrated that EPInformer achieved the best performance among all baseline models by integrating multi-omics data with DNA sequence (see revised Fig 2).

Regarding the advantages of EPInformer over large DNA models, there are several key points: 1) EPInformer can be trained to predict cell-type-specific gene expression and enhancer activity, which is crucial for understanding gene regulation in a cell-type-specific manner, which current large DNA models lack; 2) EPInformer is streamlined with only 0.4M parameters and can be trained on one consumer-grade GPU (e.g., one NVIDIA 3090) in approximately one hour for a cell line, whereas large DNA models require expensive GPUs (e.g., NVIDIA V100/A100) even just for fine-tuning, which is not feasible for most experimental labs; 3) EPInformer provides a user-friendly model interpretation pipeline at both enhancer-gene interaction and DNA sequence levels, which large DNA models lack; 4) EPInformer integrates distal regulatory elements (200 kb around the target gene) for gene expression modelling, whereas most large DNA models have limited DNA context windows (10 kb for DNABERT-2 and 12 kb for NT), which hinders their ability to capture long-range interactions and thus affects the precision of gene expression modelling.

In summary, EPInformer offers distinct advantages over large DNA models through its streamlined architecture and end-to-end training that enables cell-type-specific gene expression prediction and enhancer-gene interaction modeling on affordable GPUs.

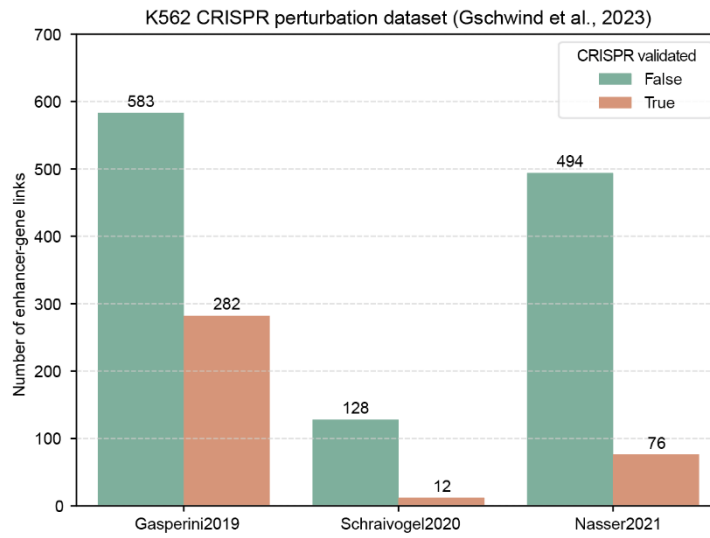
4. The improvement in the model's performance is not significant. For example, in Fig3bc, the identification of enhancer-promoter interactions is not good. Compared to traditional methods (calculating the ABC score), the attention score with and without HiC is worse than the traditional method and even multiplying the Attention score by the ABC score results in a minimal improvement, indicating that the development of this model may not be very meaningful.

Thank you for this thoughtful comment regarding enhancer–promoter interaction performance. We took this concern seriously and have made substantial improvements to address it. To improve our model's performance, we redesigned the interaction encoder's final layer to explicitly integrate Hi-C contact (or inverse genomic distance to the TSS, when Hi-C data is unavailable) and enhancer activity into the attention mechanism for gene-expression modeling. We have added the following paragraph to the Methods describing this revised attention module, and Fig. 3a illustrates its architecture:

“To better capture distal enhancer-gene interactions, we weight each enhancer embedding in the final encoder layer by Hi-C contact (or inverse distance to the TSS if this assay is not available) and enhancer activity. Formally, the interaction of promoter X_P and the i_{th} enhancer X_{E_i} can be computed as $a_{P-E_i} = \text{softmax}(Q_P \cdot K_{E_i}^T / \sqrt{d_k} \cdot F_{A_i} \cdot F_{C_i})$, where F_{A_i} denotes enhancer activity (geometric mean of H3K27ac and

DNase RPM) and F_{C_i} denotes the enhancer-gene Hi-C contacts or $1/\text{distance}$ to gene's TSS. Each attention head computes a weighted sum over promoter-enhancer pairs: $a_{P-E} \times V$. Multiple heads operate with independent parameters; their outputs are concatenated and passed through a linear projection to produce the final layer output. Consequently, in the interaction encoder's final layer, a_{P-E} quantifies the learned enhancer-gene interaction intensity and is used to prioritize enhancers for each gene."

In addition, to provide a more comprehensive evaluation we expanded our benchmarking using the largest enhancer-perturbation dataset to date from ENCODE-E2G¹⁰, aggregated from three published CRISPR studies^{11–13}. The dataset contains 1,575 enhancer-gene links (within 100kb of the target gene), including 370 CRISPR-validated enhancers in K562, which we used to evaluate our revised attention score alongside the ABC score. The figure below shows the distribution of the enhancer-gene links across three studies:



In our updated CRISPR benchmarking, the revised attention score (both distance-based and Hi-C-based) significantly outperformed the corresponding ABC scores in terms of AUPRC (**Fig. 3b**). The model *Attention.score.Hi-C* achieved the highest AUPRC of 0.732 compared to 0.698 for *ABC.score.Hi-C*. Moreover, *Attention.score.Hi-C* consistently exceeded the ABC score across genomic distance bins, with a notable advantage for very distal links (AUPRC 0.51 vs. 0.43 at 60–100 kb), indicating that our new attention mechanism more effectively captures distal enhancer-gene interactions. We believe that these improvements demonstrate that EPInformer provides meaningful advances in identifying functional enhancer-gene relationships.

We have also substantially revised the result section titled “EPInformer accurately prioritizes cell-type-specific enhancers-gene interactions” to incorporate these improvements and the updated results. The revised text and updated **Fig. 3** now clearly demonstrate the superior performance of our revised attention mechanism:

*“Linking candidate enhancers to their target genes using biochemical annotations remains a central challenge in genomics. To evaluate whether EPInformer can accurately learn from gene expression and multimodal epigenomic data to identify enhancer-gene interactions, we normalized the attention weights in the final layer of EPInformer interaction encoder to generate an attention score quantifying each candidate enhancer’s contribution to a given target gene (**Fig. 3a; Methods**). As a baseline, we used the Activity-by-Contact (ABC) score⁴⁵, a widely adopted method for linking enhancers to target genes^{22,49}. ABC score quantifies enhancer-gene links using either Hi-C contact maps (*ABC.score.Hi-C*) or a $1/\text{distance-to-TSS}$ contact proxy (*ABC.score.distance*). To align comparisons, we derived two attention-based scores from*

EPInformer: *Attention.score.distance* (from EPInformer-PE-Activity) and *Attention.score.Hi-C* (from EPInformer-PE-Activity-HiC).

To evaluate this approach, we compiled genetic perturbation datasets inspired by a recent study⁵⁰: K562 CRISPR enhancer perturbations and fine-mapped eQTL variants from lymphoblastoid cells in GTEx⁵¹. Following the ABC pipeline⁴⁵, we restricted candidate enhancers to those within 100 kb of a gene's transcriptional start site (TSS) and computed enhancer activity and Hi-C contact frequencies for K562 and GM12878. We first evaluated K562 enhancer-gene predictions by assembling 1,575 candidate enhancer-gene links (within 100 kb of their target genes) from the largest K562 CRISPR perturbation dataset⁵⁰, of which 370 were experimentally validated. We compared attention scores from EPInformer-PE-Activity (*Attention.score.distance*) and EPInformer-PE-Activity-HiC (*Attention.score.Hi-C*) against two ABC scores (*ABC.score.Hi-C* and *ABC.score.distance*) using precision–recall plots (Fig. 3b), reporting the area under the precision-recall curve (AUPRC). *Attention.score.Hi-C* achieved the highest AUPRC (0.732) versus the *ABC.score.Hi-C* (0.698). Without Hi-C, *Attention.score.distance* also outperformed the *ABC.score.distance* (0.600 vs. 0.579). Additionally, across varying distance ranges, *Attention.score.Hi-C* consistently exceeded *ABC.score.Hi-C*, with the largest gains for distal enhancers (60-100 kb: AUPRC 0.510 vs. 0.429). Notably, *Attention.score.distance* also surpassed *ABC.score.distance* for predicting distal enhancers (5-100 kb), indicating that EPInformer's attention score better captures distal enhancer–gene links than ABC score.”

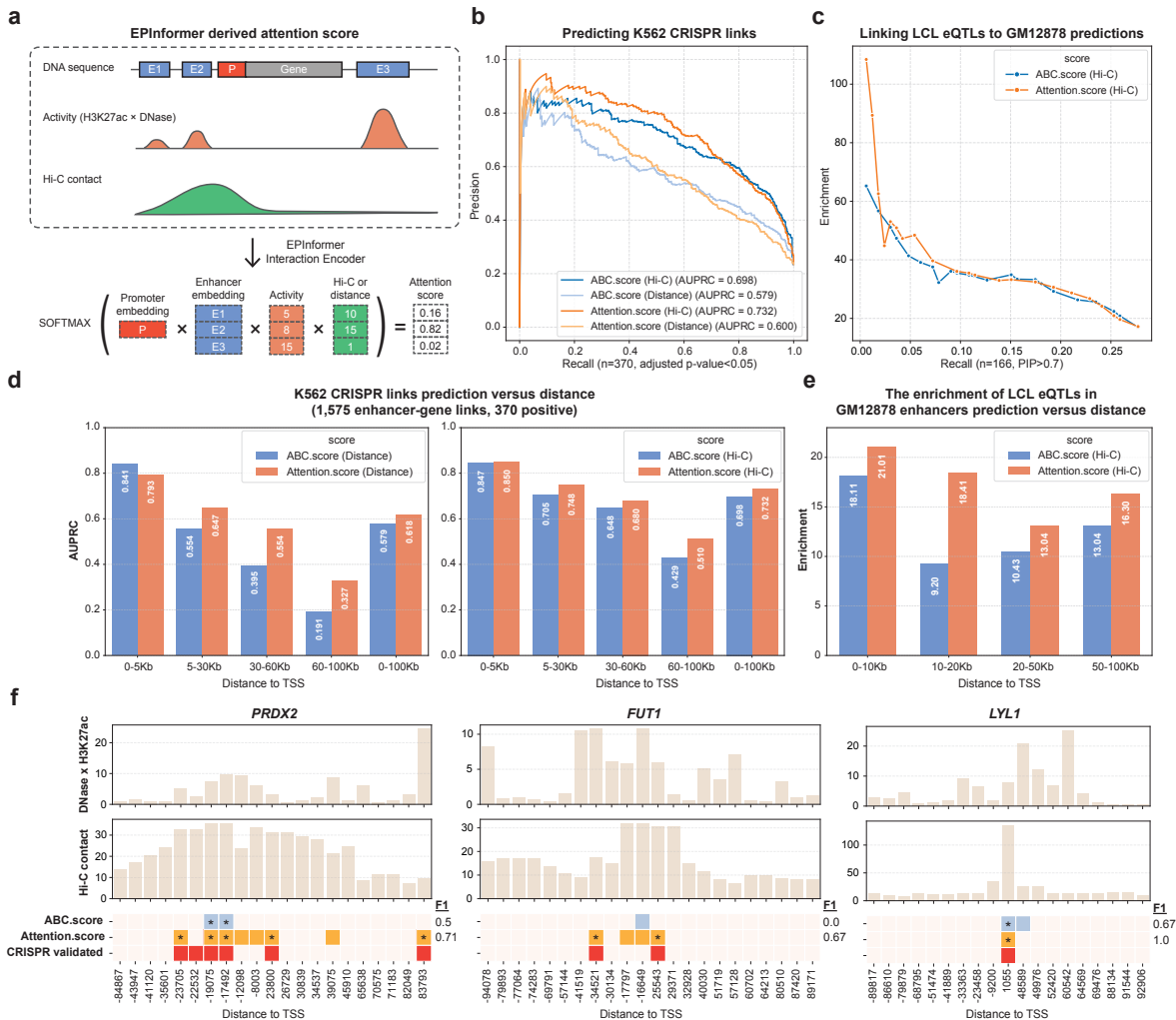


Fig. 3. | EPInformer attention prioritizes cell-type-specific enhancers corroborated by long-range interaction assays. a, Schematic of attention score calculation in EPInformer-PE-Activity(-HiC). The red and blue boxes represent three

candidate enhancers (E1, E2, and E3) located near the promoter (P) of the gene (grey box). The dashed box shows the simplified calculation of the attention score for promoter-enhancer pairs, incorporating DNase, H3K27ac, and Hi-C signals (expressed in arbitrary units, not to scale). b, Precision–recall curves for predicting enhancer–gene (E–G) pairs. Positive E–G pairs are defined as cases where perturbing the candidate enhancer significantly decreases expression of the target gene. The curves compare attention-based scores (Attention.score.Hi-C from EPInformer-PE-Activity-HiC and Attention.score.distance from EPInformer-PE-Activity) with two ABC scores in classifying 1,575 E–G pairs across 244 genes assayed by CRISPR perturbations. c, Enrichment–recall curves for enhancer–gene interaction scores in GM12878 cells. The y-axis shows enrichment—the ratio of fine-mapped distal noncoding eQTLs (PIP > 70% in GM12878) that fall within predicted enhancers relative to all distal noncoding common variants. The x-axis shows recall—the fraction of eQTL variants overlapping enhancers linked to the correct gene. Curves represent varying score thresholds for enhancers predicted by different models. d, Area under the precision-recall curve (AUPRC) values for E-G score performance shown in panel b when classifying enhancer-gene pairs at three different distance ranges. e, Enrichment of putative causal variants (PIP> 70%) within predicted enhancers across different distance ranges. f, Top three genes with the largest per-gene F1 score gains when comparing Attention.score.Hi-C to ABC.score.Hi-C. The first two rows display enhancer activity and Hi-C contact values computed by the ABC pipeline for each candidate enhancer. Light blue boxes indicate enhancers predicted by ABC.score.Hi-C, orange boxes indicate enhancers predicted by Attention.score.Hi-C. Red boxes indicate CRISPR-validated enhancers. Stars indicate correctly predicted enhancers.

5. In the experiment to verify cell-type-specific enhancer prioritization, the authors only used CRISPRi-FlowFISH results from the K562 cell line as labels. However, the role of enhancers is cell-type specific, and we do not know the performance of EPInformer in other cell lines. If possible, at least one other cell line with CRISPR-perturbation experimental labels should be added for verification. (use eQTL data following the ENCODE-E2G)

We thank the reviewer for this excellent suggestion. In response, we have not only expanded our K562 CRISPR perturbation benchmarking using data from more recent studies^{10–13}, but also followed ENCODE-E2G recommendations and adopted their eQTL enrichment pipeline (<https://github.com/EngreitzLab/eQTLEnrichment/>) to evaluate EPInformer's performance in GM12878 cells by linking lymphoblastoid cell line eQTL variants to their target genes.

The recall–enrichment analysis shows that our attention score exhibits significantly higher enrichment than the ABC score at recall <10%, and comparable enrichment at ≥10% (Fig. 3c) in GM12878 enhancer-gene prediction. Moreover, in the distance-based analysis, enhancers nominated by the attention score showed significantly higher enrichment for putative causal variants (PIP > 70%) across all distance ranges compared to those nominated by the ABC score. These new results demonstrate that the EPInformer attention score is generalizable to other cell types.

We have added the following text to the manuscript describing these findings to the manuscript and updated Fig. 3 accordingly:

*“Next, we evaluated how well attention scores can link eQTL variants to target genes in GM12878. We collected distal noncoding eQTL variants with fine-mapping posterior inclusion probability (PIP) > 70% (n = 166). For each score, we quantified (1) recall for eQTL variant-gene links (the fraction of eQTLs that fall within predicted regulatory elements and are assigned to the correct gene), and (2) enrichment of eQTLs within predicted enhancers relative to all distal noncoding variants, across score thresholds. Attention.score.Hi-C showed significantly higher enrichment than ABC.score.HiC at recall <10% and comparable enrichment at ≥10% (**Fig. 3c**). Using a fixed 70% recall threshold (as in K562), Attention.score.Hi-C nominated enhancers showed higher enrichment for putative causal variants (PIP > 70%) across distance ranges than ABC.score.Hi-C (**Fig. 3d**), consistent with the K562 results and reinforcing the attention score’s precision in identifying true enhancer-gene interactions.”*

6. The authors' interpretation of the experimental results is lacking. For example, in Fig2, there are still some housekeeping gene expression predictions that are inaccurate, such as the GADPH gene (predi. Please provide a detailed explanation of the reasons. Additionally, the authors did not elaborate on the role of the red gene names in the manuscript. Fundamentally, please indicate what cell types are present in the K562 cell line and which of these red genes are specific to these cell types; the original text barely explains this. In Fig 4c, the authors did not include a comparison with traditional methods, nor did they further analyze the significance of unknown motifs, which is disappointing.

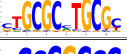
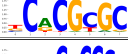
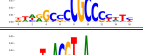
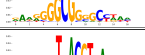
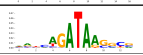
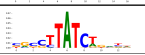
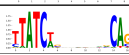
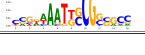
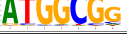
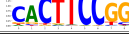
We thank the reviewer for these important points regarding gene annotation in the scatter plot and the lack of comparison with traditional methods.

In response, we have revised our gene annotation in the scatter plot. We now annotate relevant and cell-type-specific genes with high expression in the corresponding cell line in red and genes with low expression in green in Fig. 2a. Regarding K562 cell characteristics, K562 is a chronic myelogenous leukemia cell line, and the highly expressed genes (shown in red) include erythroid-specific genes such as GATA1, TAL1, and HBB, consistent with K562's erythroid differentiation potential.

Regarding the results presented in Fig 4c, and for the suggested comparison with traditional methods we used LS-GKM¹⁸, one of the most widely used SVM-based model for sequence classification, modeling and interpretation. We therefore used it as the baseline for evaluating our EPInformer sequence encoder (EPInformer-seq) in enhancer classification and sequence interpretation. To construct the enhancer classification dataset, we collected 256-bp sequences centered at K562 H3K27ac ChIP-seq summits and two 100-bp overlapping windows flanking each summit, resulting in 273,125 256-bp sequences. We used the ABC pipeline to calculate the enhancer activity (geometric mean of H3K27ac and DNase RPM). Given that LS-GKM is a sequence-classification method, we defined positive enhancers as the top 25% by activity and negative enhancers as the bottom 25% for training both EPInformer-seq and LS-GKM. We used sequences from chr12 and chr2 as the test set, sequences from chr1 and chr11 as the validation set, and the remaining sequences as the training set. Our experiments show that EPInformer-seq outperforms LS-GKM in enhancer classification, achieving an AUROC of 0.961 versus 0.955. Notably, EPInformer-seq substantially reduces training time, completing in 0.5 hours compared to 7 hours for LS-GKM.

Model	AUROC	Training time	Device
EPInformer-seq	0.961	0.5 hrs.	Nvidia V100 GPU
LS-GKM	0.955	7 hrs.	Xeon 6248R CPU

We also thank the reviewer for raising concerns about the analysis of unknown motifs. In response, we integrated EPInformer-seq with Tangermeme¹⁹ and TF-MoDISco²⁰ to produce a comprehensive motif report that enables identification of both known and novel motifs through matching against the JASPAR database. The following figure shows 8 significant motif patterns discovered by our workflow in K562 enhancer sequences. The first column lists the motif pattern ID, and the second column reports the total number of motifs found in the enhancer sequences. The third and fourth columns show the counts of forward and reverse motifs identified by our pipeline. The subsequent columns list the three most similar motifs in the JASPAR 2024 vertebrate dataset, along with their corresponding q-values.

pattern	num_seqs	modisco_cwm_fwd	modisco_cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos_patterns.pattern_0	437			MA1644_2_NFYC	0.096374		MA0060_4_NFYA	0.096374		MA0502_3_NFYB	0.346766	
pos_patterns.pattern_1	277			MA0506_3_Nrf1	0.002534		MA0830_3_TCF4	1.000000		MA0632_3_TCF5	1.000000	
pos_patterns.pattern_2	109			MA1513_2_KLF5	0.001452		MA1959_2_KLF7	0.003862		MA0599_1_KLF5	0.003862	
pos_patterns.pattern_3	48			MA1475_2_CREB3L4	0.000031		MA1140_3_JUNB	0.001345		MA1127_1_FOSB::JUN	0.001345	
pos_patterns.pattern_4	40			MA0645_2_ETV6	0.044739		MA1708_2_ETV7	0.044739		MA1957_1_HOXB2::ELK1	0.044739	
pos_patterns.pattern_5	39			MA0766_3_GATA5	0.008759		MA0140_3_GATA1::TAL1	0.008759		MA0036_4_GATA2	0.021982	
pos_patterns.pattern_6	35			MA0645_2_ETV6	0.001050		MA1708_2_ETV7	0.001196		MA0076_3_ELK4	0.001196	
pos_patterns.pattern_7	33			MA0748_3_YY2	0.217067		MA2329_1_ZBTB11	0.217067		MA0028_3_ELK1	0.217067	

The report shows Patterns 1–6 matched existing TF motifs in the JASPAR 2024 vertebrate dataset with q-values < 0.05. In contrast, Patterns 0 and 7 (q-values > 0.05) appear to be newly discovered motifs and warrant further investigation to understand their potential regulatory roles.

We have added the following text to the manuscript clarifying our motif analysis pipeline and updated Fig. 4a accordingly:

“Having established that EPInformer’s attention scores effectively prioritize enhancers relevant to a target gene, we next examined sequence patterns underlying the functionality of these predicted enhancers. We implemented a base-resolution motif-discovery workflow using EPInformer-seq, TF-MoDISco-lite³⁶ and Tangermeme³⁷ to analyze transcription factor motifs within genome-wide putative enhancers (Fig. 4a).”

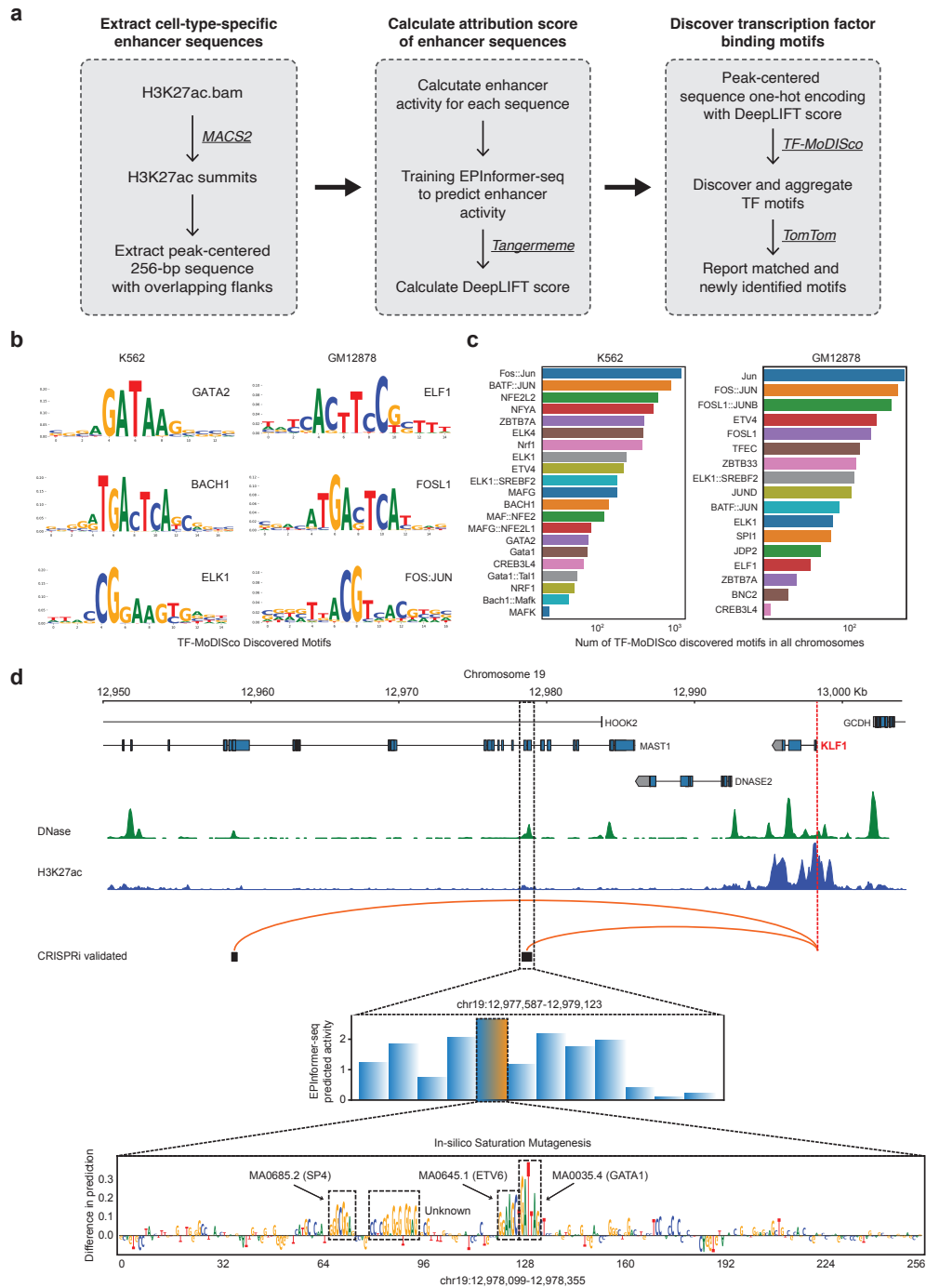
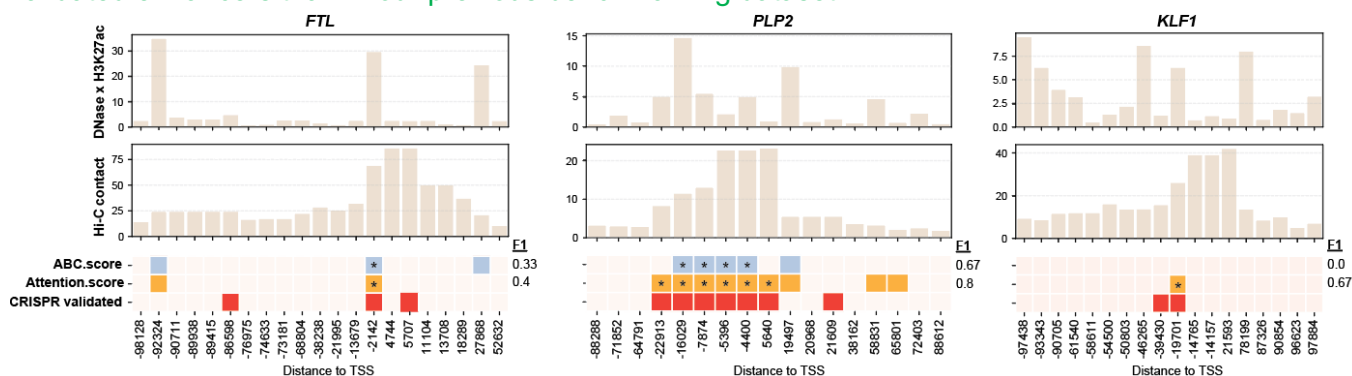


Fig 4. | EPInformer reveals transcription factor motifs at cell-type-specific enhancers. *a*, The workflow of using EPInformer-seq with TF-MoDISco-lite and tangermeme to discover transcription factor binding motifs. *b*, Representative motifs enriched in K562 (left) and GM12878 (right) discovered by TF-MoDISco-lite from recurring sequence patterns learned by EPInformer-seq at cell-type-specific H3K27ac peaks. *c*, Bar plots showing the number of TF-MoDISco-discovered motifs matching JASPAR 2024 CORE vertebrate non-redundant database (q -value < 0.05) in K562 (left) and GM12878 (right). *d*, EPInformer discovers several important transcription factor binding motifs at a putative enhancer of KLF1. The orange dashed line indicates the transcription start site (TSS) of KLF1. The black box with the red arc represents a CRISPRi-validated enhancer, while the grey box denotes an EPInformer-PE-Activity predicted enhancer. The bar plot displays predicted enhancer activities (geometric mean of H3K27ac and DNase signals) for 256-bp sequences tiling the putative enhancer locus (chr19:12,977,587-12,979,123). The bottom panel shows in-silico mutagenesis (ISM) attribution scores for the region (chr19:12,978,099-12,978,355) with the highest predicted enhancer activity. High-scoring sequence motifs are highlighted with dashed boxes, with labels and arrows indicating matching TF motifs from the JASPAR 2024 CORE vertebrate non-redundant database.

7. The argumentation in the article is not sufficient. In Fig3 and Fig4, the authors only used the Klf1 gene for interpretability analysis of the model. The authors did not explain why this gene was chosen, and choosing just one gene is not representative. More valid examples (at least five) should be selected to demonstrate that the model can indeed identify the regulatory relationships of many genes.

Thank you for this constructive suggestion. In response, we have expanded our K562 CRISPR perturbation analysis and now report six genes (PRDX2, FUT1, LYL1, FTL, PLP2 and KLF1) with the greatest F1-score gains of attention.score.Hi-C over ABC.score.Hi-C (three presented in Fig. 3f and three shown below). Note that to ensure a fair and more comprehensive comparison with other methods, we adopted the new K562 CRISPR perturbation dataset from ENCODE-E2G²¹ where regulatory elements overlapping promoter regions (± 500 bp around the TSS) are filtered out. Consequently, the updated KLF1 analysis contains fewer validated enhancers than in our previous benchmarking dataset.



These genes clearly demonstrate EPInformer's superior ability to identify functional enhancer-gene interactions across diverse genomic contexts beyond the *KLF1* locus. We have also added the following text to the manuscript describing these gene-enhancer interaction examples and updated Fig. 3f accordingly:

"To further assess functional relevance, we set a 70% recall threshold for all scores and compared per-gene F1 scores. Remarkably Attention.score.Hi-C outperformed or matched ABC.score.Hi-C for 212 of 244 genes (89%). We highlight PRDX2, FUT1, and LYL1, where Attention.score.Hi-C achieved the largest F1 score gains over ABC.score.Hi-C (Fig. 3f), attributable to more precise identification of distal enhancers, further supporting the effectiveness of the attention score."

Additionally, we chose *KLF1* for detailed motif analysis in Fig. 4d because it is a well-characterized erythroid transcription factor with known regulatory elements, making it an ideal case study for validating our motif discovery pipeline. The CRISPR-validated enhancer for *KLF1* provides ground truth for assessing our predictions.

8. Some concepts are not clearly articulated, such as why mRNA half-life data is used as a promoter feature input to the model to predict expression levels. The paper does not explain the rationale for using this, nor does it further explain what mRNA half-life measurements are and what characteristics they represent. Additionally, why not consider using ATAC-seq data as input for the FUSION LAYER to characterize enhancer accessibility? This is a more widely applicable measurement method, and I believe there is plenty of existing scATAC-seq data for the K562 cell line.

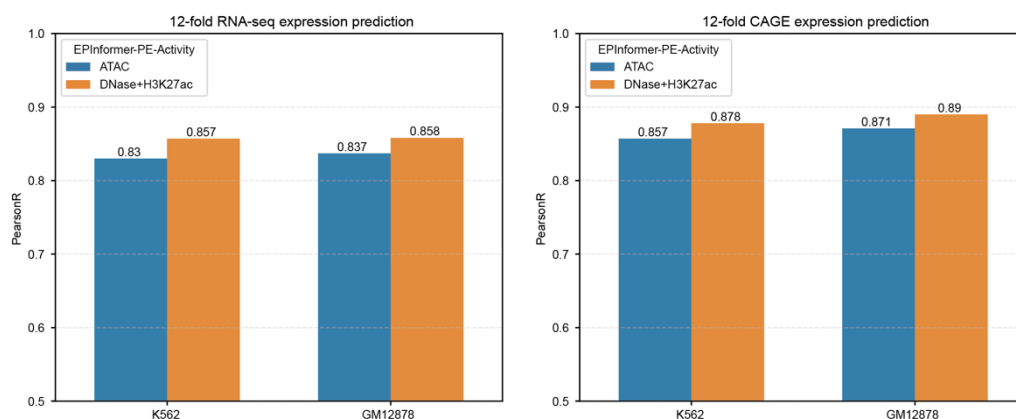
We thank the reviewer for raising these important points about mRNA half-life features and the potential use of ATAC-seq data. We apologize for not clearly explaining the distinction between promoter features and mRNA half-life features.

Promoter features aim to capture the promoter chromatin state and are defined as the geometric mean of H3K27ac and DNase RPM at the 500bp region around the gene's TSS. In contrast, mRNA half-life features, as used by Xpresso^{21,22} capture properties of the coding sequence of a gene related to its expression, including: G/C content, lengths of functional regions (5' UTRs, ORFs, and 3' UTRs), intron length, and exon junction density within the open reading frame. These features capture post-transcriptional regulation and explain an estimated 6%–15% of the variability in steady-state mRNA levels in mammalian cells²¹, highlighting their importance for accurate gene expression prediction.

We have added the following clarification to the Methods section:

“Additionally, we incorporated promoter activity and mRNA half-life features for EPInformer in gene expression modelling. Promoter activity is defined as the geometric mean of H3K27ac and DNase RPM at the 500-bp region around the gene's TSS; mRNA half-life features, as calculated by Xpresso²¹ for gene expression modelling, include G/C content, lengths of functional regions (5' UTRs, ORFs, and 3' UTRs), intron length, and exon junction density within the open reading frame.”

Regarding ATAC-seq, we appreciate the reviewer's excellent suggestion. In response, we evaluated EPInformer-PE-Activity using ATAC-seq for both CAGE and RNA-seq expression modelling. Under 12-fold cross-validation, ATAC-based EPInformer-PE-Activity achieved an average Pearson correlation of 0.843 for K562 and 0.854 for GM12878 across K562 and GM12878, which is comparable to DNase+H3K27ac-based EPInformer-PE-Activity (average Pearson correlation: 0.867 for K562 and 0.874 for GM12878). These results demonstrate that EPInformer can effectively use ATAC-seq for gene expression modeling and is practical for cell types where only ATAC-seq data is available, highlighting our framework's flexibility.



Supplementary Figure 5. EPInformer-PE-Activity performance in predicting gene-level RNA-seq expression from ATAC-seq versus H3K27ac plus DNase.

We have added the following text to the manuscript describing these new results and updated the figure in the Supplementary Information:

*“In addition, we evaluated EPInformer’s ability to leverage ATAC-seq data for modeling gene expression, since ATAC-seq data are widely available across cell types. We adapted the ABC pipeline to obtain potential enhancer regions and their corresponding ATAC-seq signals to train EPInformer-PE-Activity on both CAGE and RNA-seq expression. Under 12-fold cross-validation (**Supplementary Fig. 5**), ATAC-based EPInformer-PE-Activity achieved an average Pearson correlation of 0.843 for K562 and 0.854 for GM12878, which is comparable to DNase+H3K27ac-based EPInformer-PE-Activity (average Pearson correlation: 0.867 for K562 and 0.874 for GM12878). These results demonstrate that EPInformer can effectively use ATAC-seq for gene expression modeling and is practical for cell types where only ATAC-seq data is available, highlighting our framework's flexibility.”*

Minor Concerns:

1. In the section "In EPInformer attention prioritizes cell-type-specific enhancers supported by long-range interaction assays," the authors did not cite Fig3C.

We apologize for this omission and have now updated the text to properly cite Fig 3c in the manuscript.

Reviewer #1 (Remarks on code availability):

Pretty good. there are some ipynb codes.

Thank you for reviewing our code. We appreciate your positive feedback.

Reviewer #2

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank you for your contribution to the review process and appreciate your time and effort in co-reviewing our manuscript. Your participation in this peer review training initiative is valuable, and we hope our revisions have addressed all the concerns raised in the review.

Reviewer #3

This study presents EPInformer, an Enformer-based method for predicting gene expression data, with improvement through the integration of promoter-enhancer interactions, enhancer activity, and Hi-C data. The authors demonstrate that EPInformer outperforms previous models in predicting both mRNA and CAGE expression levels. Furthermore, they show that incorporating the ABC score improves the model's performance. They validate their predictions using CRISPR-based perturbation results in experiments, investigate transcription factor motifs, and assess the predictability and utility of EPInformer.

While this study offers valuable insights, I have the following comments:

Major Comments

1. Definition of TSS

The definition of the transcription start site (TSS), a critical aspect of this study, is not clearly stated. Does the TSS refer to the first base of the 5' UTR or the GENCODE-annotated TSS?

We apologize for not clearly explain the TSS definition in the manuscript. Following Xpresso's procedure²¹, we defined TSSs as the start coordinate of the first exon of the representative transcript (GENCODE v10); for mRNAs with a dominant CAGE peak, we re-centered the TSS to the coordinate of that peak.

We have added the following text to method section to clarify the definition of the TSS:

"Following Xpresso's procedure²³, we defined TSSs as the start coordinate of the first exon of the representative transcript (GENCODE v10); for mRNAs with a dominant CAGE peak, we re-centered the TSS to the coordinate of that peak."

2. Insufficient Transcript Coverage

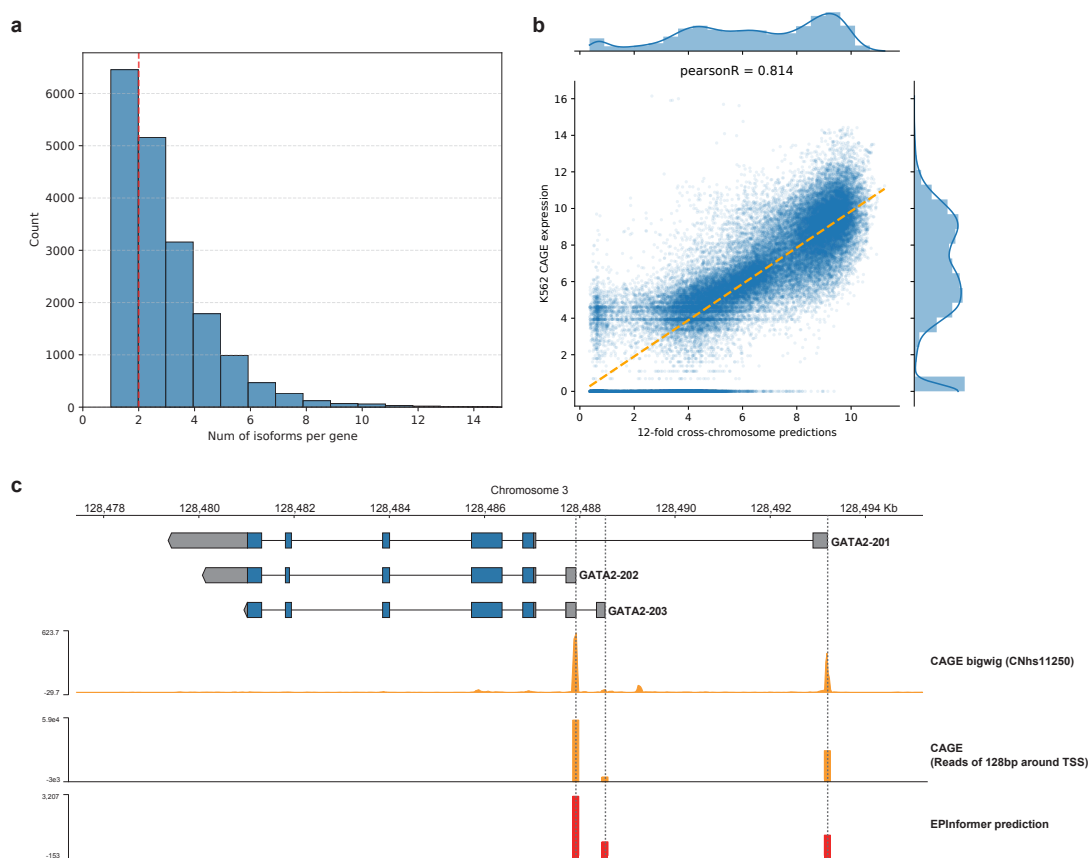
It is unclear whether the model fully captures transcripts originating from the same gene. The authors should evaluate whether EPInformer can be applied to individual transcripts and TSSs, as defined by FANTOM. Given that the ultimate goal is to interpret GWAS associations and understand the genetic basis of complex traits, the model's resolution should be explicitly demonstrated.

We thank the reviewer for highlighting the importance of isoform-specific expression and alternative TSS usage. We agree that incorporating alternative TSSs could improve the modeling of gene expression. In this study, we initially restricted our analyses to the TSS of the canonical (representative) transcript for each gene, as this approach aligns with the gene-level expression data (RNA-seq and CAGE) used for training and evaluation. This approach was also adopted by previous models including Xpresso, Enformer, and for a fair evaluation, we followed the same procedure to ensure direct comparability of results.

However, while EPIInformer currently predicts gene-level expression, the framework could be extended to model isoform-specific expression using isoform-specific TSSs from FANTOM. Therefore, as suggested by the reviewer and to evaluate the effectiveness of EPIInformer on predicting isoform-specific expression, we compiled all alternative TSSs for protein-coding genes from FANTOM and GENCODE v41, yielding 46,616 isoforms with distinct TSSs with a median of two TSSs per gene (**Supplementary Fig. 7a**). Expression of each isoform was quantified using 128-bp CAGE reads counts centered on each TSS. Under 12-fold cross-chromosomal validation, EPIInformer-PE-activity (using DNase and H3K27ac signals) accurately predicted K562 CAGE expression (Pearson $r = 0.814$; **Supplementary Fig. 7b**). As an illustration, EPIInformer predictions closely recapitulated the CAGE signal at GATA2, a key transcription factor expressed in K562 cells (**Supplementary Fig. 7c**).

We have added the following text to the Results section:

*“While the above results demonstrate EPIInformer’s superior performance on gene-level expression prediction, it is important to note that most existing models are evaluated only on canonical transcripts. As many genes produce isoforms with distinct functions and tissue-specific expression, we further assessed EPIInformer-PE-Activity at the isoform level. To evaluate performance, we compiled alternative TSSs for protein-coding genes from FANTOM and GENCODE v41, yielding 46,616 isoforms with distinct TSSs (median two per gene; **Supplementary Fig. 7a**). Isoform-specific expression was quantified using 128-bp CAGE read counts centered on each TSS. Under the same 12-fold cross-chromosomal validation, EPIInformer-PE-Activity (using DNase and H3K27ac signals) accurately predicted K562 CAGE expression (Pearson $r = 0.814$; **Supplementary Fig. 7b**). As an example, predictions closely recapitulated the CAGE signal at GATA2, a key transcription factor expressed in K562 cells (**Supplementary Fig. 7c**).”*



Supplementary Figure 7. Performance of EPIInformer-PE-Activity in predicting CAGE expression for isoforms with distinct transcription start sites (TSSs). **a**, Distribution of the number of isoforms with distinct TSSs per gene; the dashed line denotes the median (two isoforms per gene). **b**, Scatter plot showing

the relationship between EPInformer-PE-Activity–predicted and observed CAGE expression levels in K562 cells; each point represents an isoform with a distinct TSS. c, EPInformer-PE-Activity accurately predicts expression for three GATA2 isoforms in K562. The first row illustrates the three GATA2 isoforms; the second row shows CAGE coverage; the third row shows the sum of 128-bp CAGE read counts centered on each isoform’s TSS; the last row shows EPInformer-PE-Activity predictions for the three isoforms.

3.Comparison with Previous Methods

The comparison with existing methods appears incomplete. How does EPInformer perform relative to Basenji and MENTR? A direct evaluation would strengthen the study's claims of improved accuracy.

We thank the reviewer for this important suggestion. We thoroughly investigated MENTR²⁴ and Basenji²⁵. We found that MENTR is designed for non-coding RNA expression modeling and performs binary classification of genes as highly or lowly expressed, rather than predicting protein-coding gene expression levels as our model does. This fundamental difference in task formulation makes a direct comparison methodologically inappropriate.

Regarding Basenji, this model predates Enformer¹⁵ and was developed by the same authors (Kelley et al.). Enformer has been shown to outperform Basenji on gene expression prediction, so we naturally included Enformer as our baseline from this line of work. To ensure we are comparing against the most current methods, we have now included Borzoi³ (Linder et al., Nature Genetics, 2025), which is the latest iteration from some of the same authors and was published this year. To our knowledge, Borzoi represents the current best-performing approach for gene expression prediction.

Our results show that EPInformer-PE-Activity-HiC outperforms Borzoi on CAGE expression prediction, achieving higher Pearson correlations in K562 ($r = 0.892$ vs. 0.792) and GM12878 ($r = 0.905$ vs. 0.800).

We have updated the manuscript with this important comparison by adding the following text and a new panel in Fig. 2:

*“For CAGE expression prediction, EPInformer-PE-Activity-HiC consistently outperformed all EPInformer variants and competing methods in both K562 and GM12878 (**Fig. 2b–d**). On the held-out test set, it surpassed Enformer in K562 (Pearson’s $r = 0.865$ vs. 0.775) and GM12878 ($r = 0.872$ vs. 0.774 ; **Fig. 2b**). Although Borzoi outperformed Enformer, EPInformer-PE-Activity-HiC achieved higher performance than Borzoi in K562 ($r = 0.892$ vs. 0.792) and GM12878 ($r = 0.905$ vs. 0.800 ; **Fig. 2d**). Notably, even without Hi-C data, EPInformer-PE-Activity exceeded Borzoi, yielding higher Pearson correlations in K562 ($r = 0.841$ vs. 0.792) and GM12878 ($r = 0.847$ vs. 0.800). Under 12-fold cross-chromosome validation in K562 and GM12878 (**Fig. 2c**), EPInformer-PE-Activity-HiC achieved the highest average Pearson correlation ($r = 0.883$), outperforming Seq-GraphReg ($r = 0.744$) and CREaTor ($r = 0.810$).”*

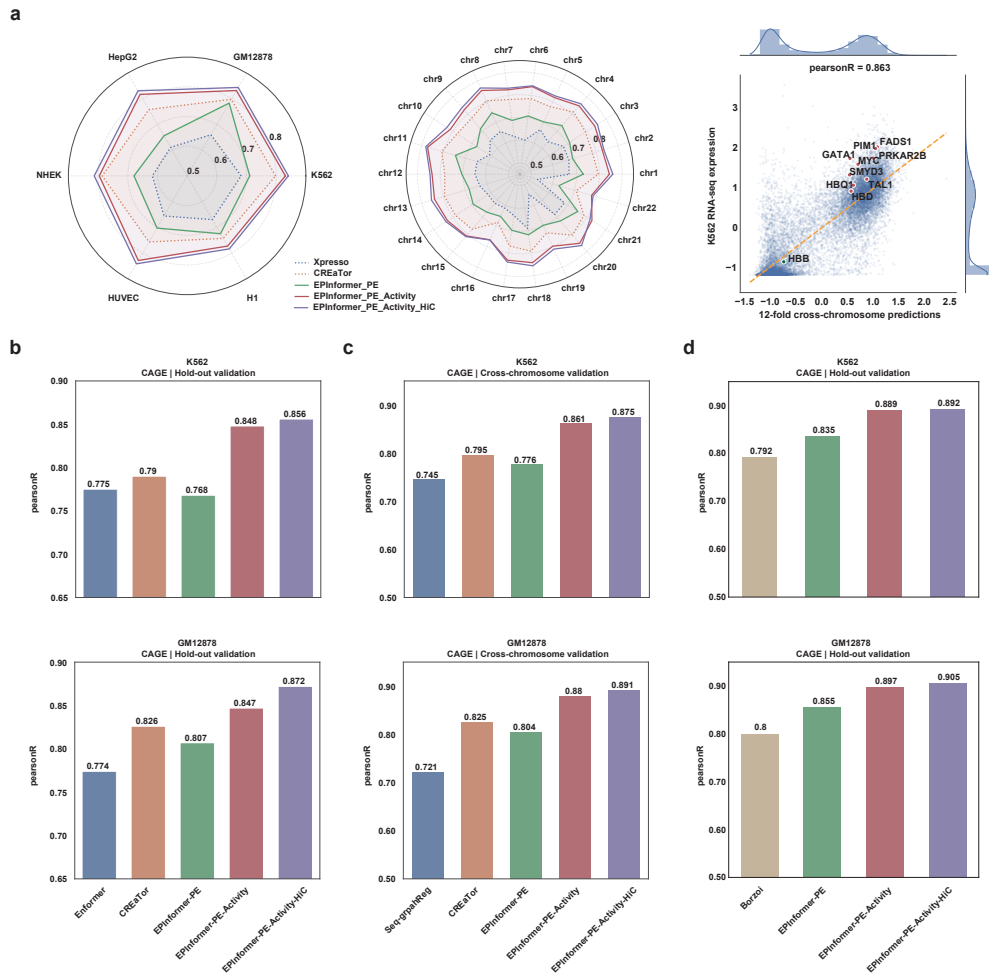


Fig 2. | Performance of EPIInformer models and baseline methods on gene expression prediction. *a*, Pearson correlation between observed and predicted RNA-seq expression levels for 18,377 protein-coding genes across six cell lines. Left: Spider plot showing correlation between observed and predicted expression levels for six cell lines. Middle: Spider plot showing correlation between observed and predicted expression levels for each test chromosome across six cell lines. Right: Scatter plot showing relationship between EPIInformer-PE-Activity-HiC predicted and observed RNA-seq expression levels in K562 cells, with each point representing an individual gene. Genes with high expression in K562 are highlighted in red, and those with low expression are highlighted in green. All evaluations were conducted using 12-fold cross-chromosome validation. *b*, Comparison of EPIInformer models with Enformer for predicting CAGE gene expression levels on hold-out genes in K562 (top) and GM12878 (bottom) cells. *c*, Comparison of EPIInformer models with Seq-GraphReg and CREaTor for CAGE gene expression prediction in K562 (top) and GM12878 (bottom) cells using 12-fold cross-chromosome validation. *d*, Comparison of EPIInformer models with Borzoi for predicting CAGE gene expression levels on hold-out genes in K562 (top) and GM12878 (bottom) cells.

4. Lack of Cell-Type-Specific Hi-C Integration

The authors do not appear to incorporate cell-type-specific Hi-C data. While the ABC model (Fulco et al.) also uses non-cell-type-specific Hi-C, EPIInformer aims to predict cell-type-specific gene expression levels. Therefore, integrating cell-type-specific Hi-C data would be essential for capturing context-dependent regulatory interactions.

We thank the reviewer for this important suggestion. In line with a similar comment from another reviewer, we redesigned the interaction encoder's final layer to better incorporate cell-type-specific Hi-C contact and enhancer activity within the attention module. Briefly, our experiments show that this new attention module significantly outperforms the ABC score, achieving a higher AUPRC (0.732 vs. 0.698).

We have added the following paragraph to the Methods describing this revised attention module, and Fig. 3a illustrates its architecture:

“To better capture distal enhancer-gene interactions, we weight each enhancer embedding in the final encoder layer by Hi-C contact (or inverse distance to the TSS if this assay is not available) and enhancer activity. Formally, the interaction of promoter X_P and the i_{th} enhancer X_{E_i} can be computed as $\alpha_{P-E_i} = softmax(Q_P \cdot K_{E_i}^T / \sqrt{d_k} \cdot F_{A_i} \cdot F_{C_i})$, where F_{A_i} denotes enhancer activity (geometric mean of H3K27ac and DNase RPM) and F_{C_i} denotes the enhancer-gene Hi-C contacts or $1/\text{distance to gene's TSS}$. Each attention head computes a weighted sum over promoter-enhancer pairs: $\alpha_{P-E} \times V$. Multiple heads operate with independent parameters; their outputs are concatenated and passed through a linear projection to produce the final layer output. Consequently, in the interaction encoder's final layer, α_{P-E} quantifies the learned enhancer–gene interaction intensity and is used to prioritize enhancers for each gene.”

The revised attention score (both distance-based and Hi-C–based) outperformed the corresponding ABC scores in terms of AUPRC (**Fig. 3b**). Moreover, Attention.score.Hi-C consistently exceeded the ABC score performance across genomic distance bins, with a notable advantage for very distal links (0.51 vs. 0.43 at 60-100 kb), indicating that our new attention mechanism more effectively captures distal enhancer-gene interactions.

We have revised the Results section titled “EPInformer accurately prioritizes cell-type-specific enhancers-gene interactions” to reflect these improvements and updated Fig. 3 accordingly:

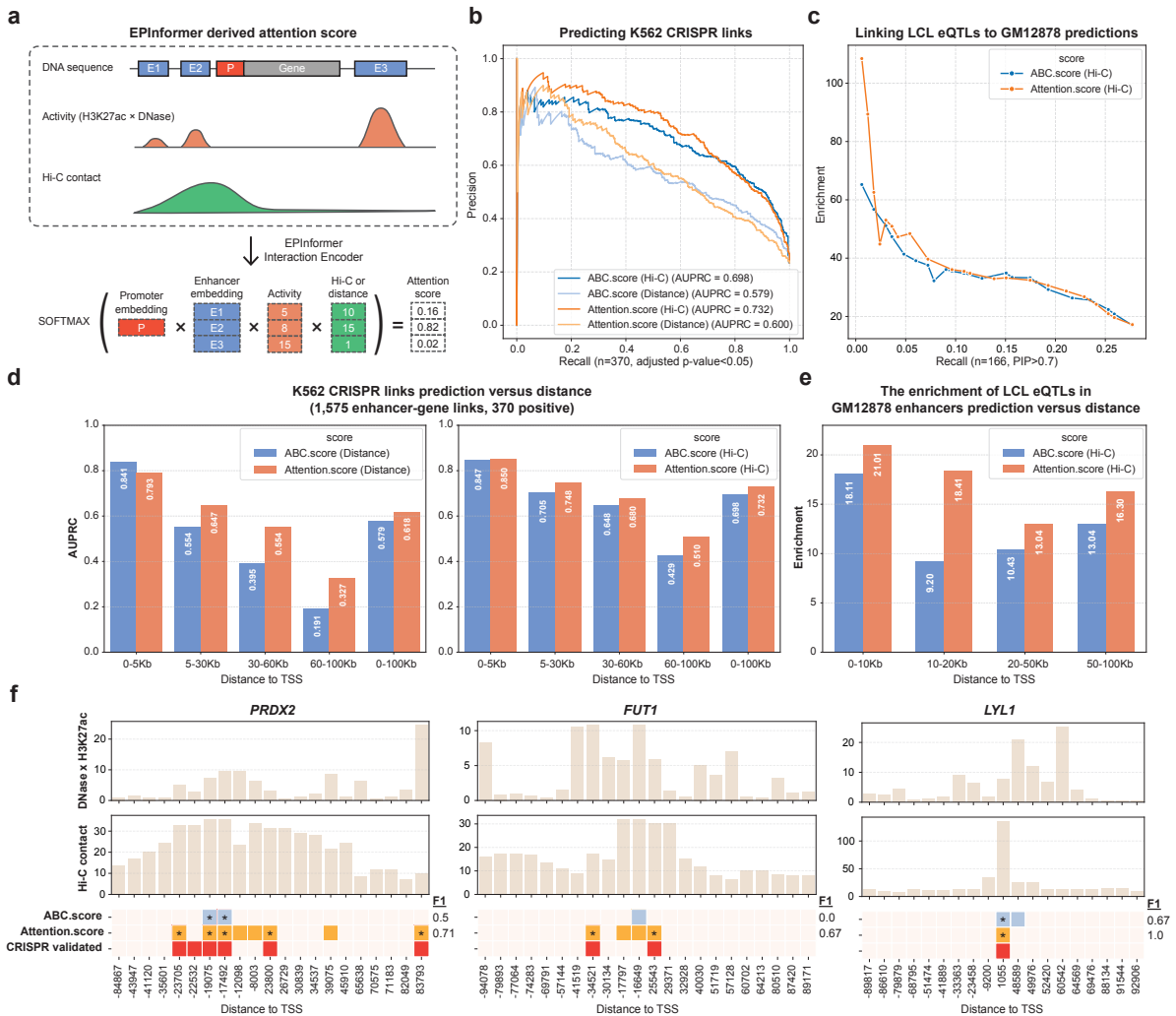


Fig. 3. | EPIInformer attention prioritizes cell-type-specific enhancers corroborated by long-range interaction assays.

a, Schematic of attention score calculation in EPIInformer-PE-Activity(-HiC). The red and blue boxes represent three candidate enhancers (E1, E2, and E3) located near the promoter (P) of the gene (grey box). The dashed box shows the simplified calculation of the attention score for promoter-enhancer pairs, incorporating DNase, H3K27ac, and Hi-C signals (expressed in arbitrary units, not to scale). *b*, Precision–recall curves for predicting enhancer–gene (E–G) pairs. Positive E–G pairs are defined as cases where perturbing the candidate enhancer significantly decreases expression of the target gene. The curves compare attention-based scores (Attention.score.Hi-C from EPIInformer-PE-Activity-HiC and Attention.score.distance from EPIInformer-PE-Activity) with two ABC scores in classifying 1,575 E–G pairs across 244 genes assayed by CRISPR perturbations. *c*, Enrichment–recall curves for enhancer–gene interaction scores in GM12878 cells. The y-axis shows enrichment—the ratio of fine-mapped distal noncoding eQTLs (PIP > 70% in GM12878) that fall within predicted enhancers relative to all distal noncoding common variants. The x-axis shows recall—the fraction of eQTL variants overlapping enhancers linked to the correct gene. Curves represent varying score thresholds for enhancers predicted by different models. *d*, Area under the precision-recall curve (AUPRC) values for E-G score performance shown in panel *b* when classifying enhancer-gene pairs at three different distance ranges. *e*, Enrichment of putative causal variants (PIP > 70%) within predicted enhancers across different distance ranges. *f*, Top three genes with the largest per-gene F1 score gains when comparing Attention.score.Hi-C to ABC.score.Hi-C. The first two rows display enhancer activity and Hi-C contact values computed by the ABC pipeline for each candidate enhancer. Light blue boxes indicate enhancers predicted by ABC.score.Hi-C, orange boxes indicate enhancers predicted by Attention.score.Hi-C. Red boxes indicate CRISPR-validated enhancers. Stars indicate correctly predicted enhancers.

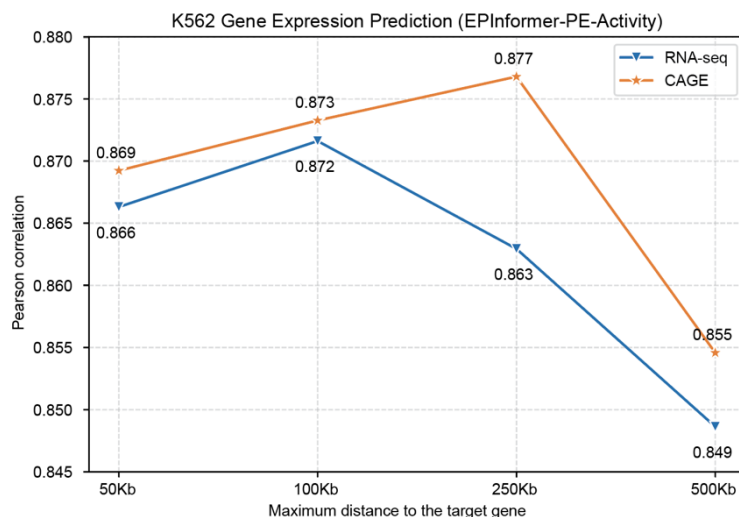
5. Definition of Flanking Distances

The study uses a 200-kb window centered on the TSS to search potential regulatory elements. However, cis-eQTL studies commonly use a 1-Mb window, and trans-eQTLs have been suggested to explain more heritability than cis-eQTLs. The authors should provide justification for using a 200-kb window and assess whether it sufficiently captures the regulatory architecture of gene expression.

We appreciate the reviewer's comments regarding the distance limits used to define candidate enhancers in relation to their target genes for gene expression modeling. Following the reviewer's suggestion, we evaluated EPIInformer-PE-Activity using different distance ranges (from 50 kb to 500 kb from the TSS) for gene expression prediction under 12-fold leave-chromosomes-out cross-validation.

Our experiments revealed that model performance indeed varies with distance range. For RNA-seq expression prediction, the model achieved its highest performance using candidate enhancers within 100 kb of the gene (Pearson $r = 0.873$), while CAGE-based prediction performed best with a 250 kb window (Pearson $r = 0.877$), representing only a modest improvement over the 100 kb window.

Notably, model performance degraded substantially at 500 kb (corresponding to 1-Mb windows centered at the TSS). We attribute this decrease to the inclusion of numerous false-positive enhancer assignments at greater distances, which introduce noise into the model. While our current 200-kb window effectively balances comprehensive enhancer coverage with prediction accuracy, we hypothesize that higher-quality Hi-C data could enable accurate modeling at greater distances by better distinguishing true long-range interactions from spurious associations. These experiments demonstrate both the flexibility of our framework in accommodating different genomic ranges and the importance of selecting appropriate distance thresholds for optimal performance.



Supplementary Figure 6. Effect of genomic distance range on EPInformer gene expression prediction performance.

Performance of EPInformer-PE-Activity for predicting RNA-seq and CAGE expression in K562 cells using different maximum distances from the TSS to define candidate enhancers. The x-axis shows the maximum distance range (50 kb, 100 kb, 250 kb, and 500 kb from the TSS) used to select candidate enhancers. The y-axis shows Pearson correlation between predicted and observed gene expression levels. Blue line represents RNA-seq

expression prediction; orange line represents CAGE expression prediction. Performance peaks at 100 kb for RNA-seq and 250 kb for CAGE, with substantial degradation at 500 kb, suggesting that extending beyond 250 kb introduces noise from false-positive enhancer assignments.

While cis-eQTL studies typically use 1-Mb windows to ensure comprehensive variant capture, gene expression prediction models face different constraints. These analyses suggest that our 200-kb window (100 kb upstream and downstream) captures the majority of functional enhancer-gene interactions while maintaining computational efficiency and avoiding noise from distal false-positive elements. This distance range aligns with recent CRISPR perturbation studies showing that most functional enhancers lie within 100 kb of their target genes.

Regarding trans-eQTL effects, we acknowledge these are important for understanding the full genetic architecture of gene expression. While our current model formulation does not directly capture trans effects, it may indirectly capture some trans-regulatory influences when they manifest through changes in local regulatory elements. For example, a variant affecting a transcription factor in cis may influence another gene by modulating regulatory elements where this factor binds (e.g., enhancers or promoters). We argue that the current architecture will capture these effects when they are reflected in the epigenomic signals at regulatory elements. However, fully modeling trans effects represents an exciting future research direction, where we envision models with genome-wide context that could learn both cis and trans regulatory relationships comprehensively.

We have added the following text to the Results section and created a new supplementary figure documenting these analyses:

"We validated this distance choice by evaluating model performance across different genomic ranges (50 kb to 500 kb from the TSS), finding optimal performance at 100-250 kb with substantial degradation at 500 kb (Supplementary Fig. 6)."

6.Fine-Mapping of Variants

To demonstrate the utility of EPInformer in interpreting GWAS associations, the authors should assess whether their method can aid in fine-mapping causal variants. Specifically, can the model prioritize variants based on enhancer contributions to gene expression?

We thank the reviewer for this valuable suggestion. Following the approach used in recent enhancer-gene linking studies¹⁰, we now demonstrate EPInformer's utility for variant prioritization through our eQTL enrichment analysis in GM12878 cells (**Fig. 3c,e**). We followed ENCODE-E2G¹⁰ to use their eQTL enrichment pipeline (<https://github.com/EngreitzLab/eQTLEnrichment/>) to evaluate GM12878 enhancer-gene prediction in linking lymphoblastoid fine-mapped eQTL variants to target genes.

The recall–enrichment analysis shows that our attention score exhibits significantly higher enrichment than the ABC score at recall <10%, and comparable enrichment at ≥10% (**Fig. 3c**) in GM12878 enhancer-gene prediction. In the distance-based analysis, enhancers nominated by the attention score showed significantly higher enrichment for putative causal variants (PIP > 70%) across all distance ranges compared to those nominated by the ABC score (Fig 3e). These new results demonstrate that EPInformer can effectively prioritize causal variants by identifying the enhancers they disrupt.

We have added the following text to the manuscript describing these findings:

*“Next, we evaluated how well attention scores can link eQTL variants to target genes in GM12878 lymphoblastoid cells. We collected distal noncoding eQTL variants with fine-mapping posterior inclusion probability (PIP) > 70% (n = 166). For each score, we quantified (1) recall for eQTL variant-gene links (the fraction of eQTLs that fall within predicted regulatory elements and are assigned to the correct gene), and (2) enrichment of eQTLs within predicted enhancers relative to all distal noncoding variants, across score thresholds. Attention.score.Hi-C showed significantly higher enrichment than ABC.score.HiC at recall <10% and comparable enrichment at ≥10% (**Fig. 3c**). Using a fixed 70% recall threshold (as in K562), Attention.score.Hi-C nominated enhancers showed higher enrichment for putative causal variants (PIP > 70%) across distance ranges than ABC.score.Hi-C (**Fig. 3d**), consistent with the K562 results and reinforcing the attention score’s precision in identifying true enhancer-gene interactions.”*

We believe future studies could systematically explore whether EPInformer could aid in fine-mapping GWAS variants, building on our eQTL results. Such studies could identify candidate enhancers near GWAS loci using EPInformer's attention scores and prioritize variants within high-scoring enhancers predicted to affect target gene expression. This approach would leverage EPInformer's demonstrated ability to link enhancers to genes and could potentially extend to predicting the functional impact of sequence changes at GWAS loci. We have added the following text to the Discussion section acknowledging our eQTL findings and future directions:

“We have also shown how EPInformer can aid in fine-mapping eQTL variants by accurately linking them to their target genes through enhanced enhancer-gene predictions. Future studies could extend this approach to systematically evaluate EPInformer's utility for fine-mapping GWAS variants, potentially providing insights into the regulatory mechanisms underlying complex traits and diseases.”

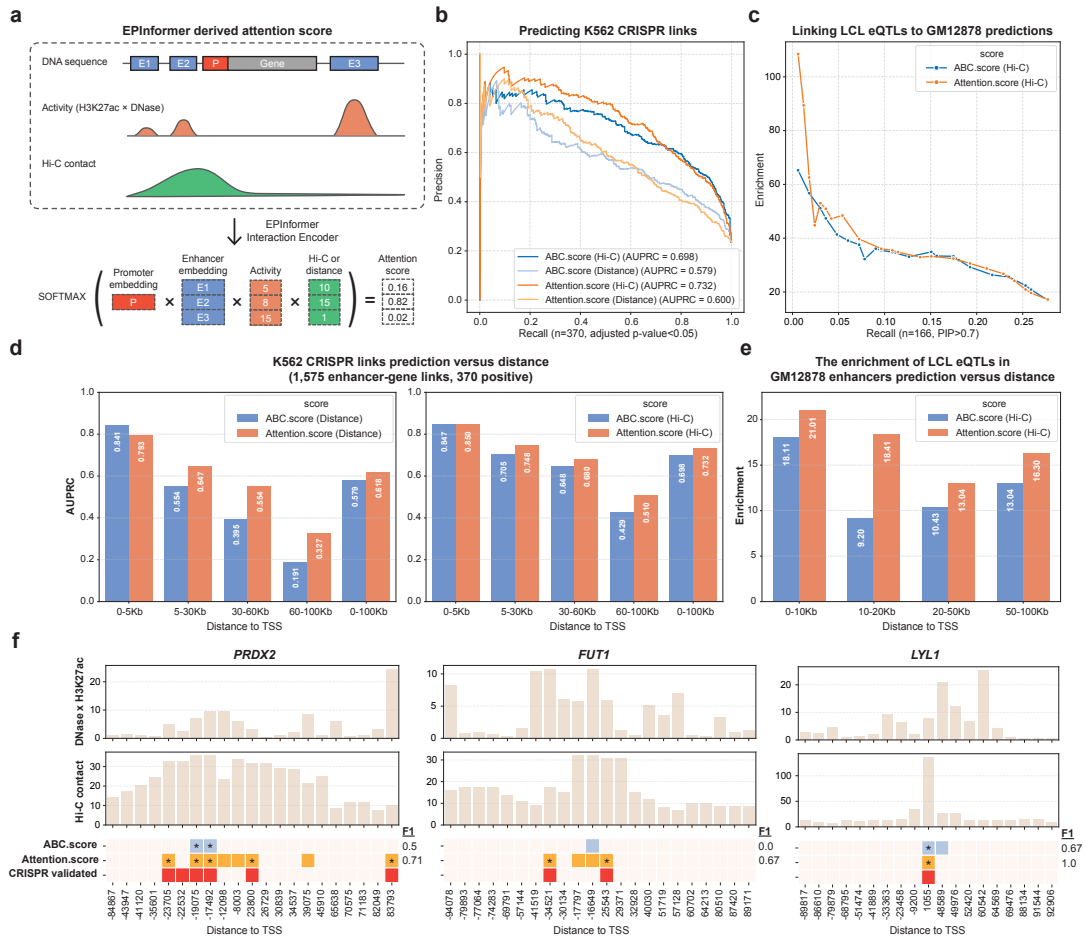


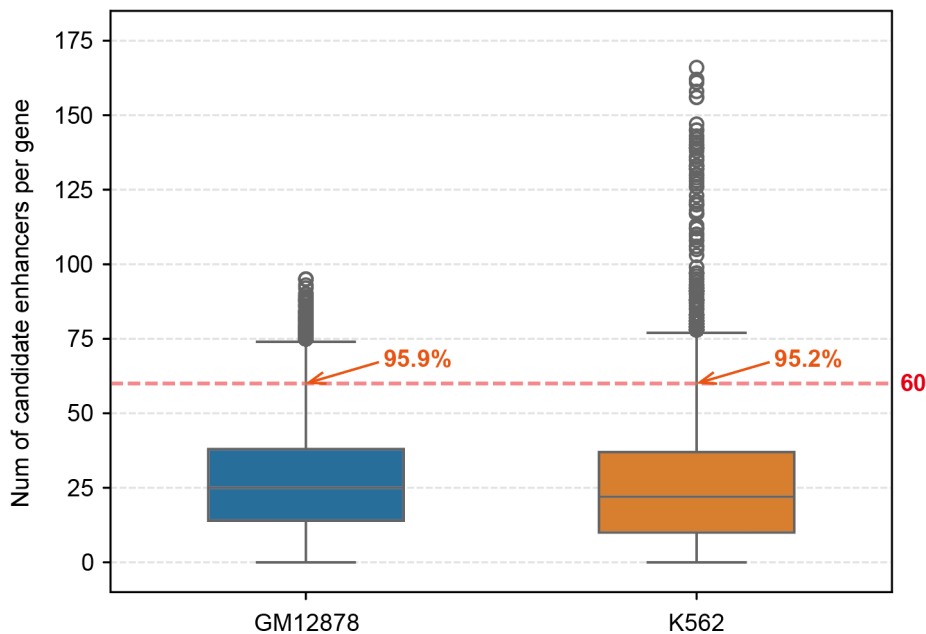
Fig. 3. | EPIInformer attention prioritizes cell-type-specific enhancers corroborated by long-range interaction assays.

a, Schematic of attention score calculation in EPIInformer-PE-Activity(-HiC). The red and blue boxes represent three candidate enhancers (E1, E2, and E3) located near the promoter (P) of the gene (grey box). The dashed box shows the simplified calculation of the attention score for promoter-enhancer pairs, incorporating DNase, H3K27ac, and Hi-C signals (expressed in arbitrary units, not to scale). **b**, Precision–recall curves for predicting enhancer–gene (E–G) pairs. Positive E–G pairs are defined as cases where perturbing the candidate enhancer significantly decreases expression of the target gene. The curves compare attention-based scores (Attention.score.Hi-C from EPIInformer-PE-Activity-HiC and Attention.score.distance from EPIInformer-PE-Activity) with two ABC scores in classifying 1,575 E–G pairs across 244 genes assayed by CRISPR perturbations. **c**, Enrichment–recall curves for enhancer–gene interaction scores in GM12878 cells. The y-axis shows enrichment—the ratio of fine-mapped distal noncoding eQTLs (PIP > 70% in GM12878) that fall within predicted enhancers relative to all distal noncoding common variants. The x-axis shows recall—the fraction of eQTL variants overlapping enhancers linked to the correct gene. Curves represent varying score thresholds for enhancers predicted by different models. **d**, Area under the precision-recall curve (AUPRC) values for E-G score performance shown in panel **b** when classifying enhancer-gene pairs at three different distance ranges. **e**, Enrichment of putative causal variants (PIP > 70%) within predicted enhancers across different distance ranges. **f**, Top three genes with the largest per-gene F1 score gains when comparing Attention.score.Hi-C to ABC.score.Hi-C. The first two rows display enhancer activity and Hi-C contact values computed by the ABC pipeline for each candidate enhancer. Light blue boxes indicate enhancers predicted by ABC.score.Hi-C, orange boxes indicate enhancers predicted by Attention.score.Hi-C. Red boxes indicate CRISPR-validated enhancers. Stars indicate correctly predicted enhancers.

Minor Comments

7.The authors state, "This threshold covers 95% of potential regulatory elements within a 200-kb region centered on the TSS." How was this value determined?

We determined this value by analyzing the distribution of ABC-nominated enhancers across our gene set. Specifically, we counted the number of candidate enhancers within 200 kb of each gene's TSS and found that selecting up to 60 enhancers per gene captured 95% of all potential regulatory elements in this genomic window (see figure below). This threshold balances comprehensive enhancer coverage with computational efficiency.



8.The authors mention, "Across six cell lines and 18,377 genes on test chromosomes." The reported number of genes on test chromosomes seems unexpectedly high. Could they clarify this?

We apologize for the lack of clarity. We performed 12-fold cross-chromosome validation to evaluate predicted expression for all 18,377 genes. In each fold, the model was tested on genes from two held-out chromosomes. After completing all 12 folds, every gene (all 18,377) was evaluated.

We rephrased the following text to avoid misleading to the manuscript:

"Across six cell lines using 12-fold leave-two-chromosomes-out cross-validation, EPInformer-PE-Activity-HiC achieved the highest median Pearson correlation (0.69)."

9.Abbreviations should be defined upon their first appearance in the main text.

We appreciate the reviewer for raising this point. We have carefully reviewed the manuscript and ensured that every abbreviation is defined at its first appearance in the main text.

Reviewer #3 (Remarks on code availability):

The codes seem reasonable

Thank you for reviewing our code. We appreciate your feedback that the code seems reasonable.

Reviewer #4

In this manuscript, the authors present a method called "EPInformer" for inferring gene expression from regulatory element DNA sequences and additional features. While this method has the potential to enhance our understanding of gene regulation, and the transformer-based approach integrates various data sources, I have concerns, especially for its level of innovation, which I would like to have clarified before recommending it for publication.

1. There are some works that predict gene expression. For example, Enformer ('Effective gene expression prediction from sequence by integrating long-range interactions') also uses a transformer framework. In this context, what are the specific contributions of your method, and how does it improve upon Enformer?

We thank the reviewer for raising this question. EPInformer stands out from Enformer and other gene expression prediction methods due to its lightweight design and versatility. While Enformer pioneered the use of transformers for genomic track prediction including gene expression (based on CAGE), EPInformer makes several key contributions that distinguish it from existing approaches.

First and foremost, EPInformer explicitly models promoter-enhancer interactions through dedicated attention mechanisms, unlike Enformer which processes entire genomic sequences without distinguishing these regulatory relationships. This architectural choice allows EPInformer to directly capture enhancer-gene interactions that are fundamental to gene regulation. Furthermore, EPInformer seamlessly integrates DNA sequences with epigenomic signals (DNase-seq, H3K27ac) and Hi-C chromatin contacts if available (which Enformer does not utilize), while Enformer relies primarily on sequence information alone.

With only 0.4 million parameters—compared with Enformer's ~250 million and Borzoi's ~186 million—EPInformer trains substantially faster without sacrificing accuracy. On an A100 GPU, training completes in approximately one hour (Supplementary Table 2), making advanced gene expression modeling more accessible to the community. This efficiency enables rapid adaptation and retraining to new cell types, whereas Enformer's massive size makes such adaptations computationally challenging. Importantly, the model can be trained using only DNase-seq or ATAC-seq data when necessary, providing flexibility for diverse experimental contexts.

Moreover, EPInformer's architecture readily integrates DNA sequence, multiple epigenomic modalities, and chromatin-contact information, enhancing its ability to predict gene expression from diverse assays such as CAGE and RNA-seq. This versatility, combined with our attention mechanism that directly reveals CRISPR-validated enhancer-gene interactions, provides mechanistic insights beyond prediction accuracy. These improvements resulted in EPInformer outperforming Enformer on CAGE expression prediction (Pearson's $r = 0.865$ vs. 0.775 in K562; 0.872 vs. 0.774 in GM12878), demonstrating both its broad applicability and superior performance relative to Enformer, Borzoi, Seq-GraphReg, CREaTor, and Xpresso.

These key advantages are now better highlighted in the manuscript. We have added the following text to the Introduction section to clearly summarize EPInformer's contributions in one place:

"EPInformer offers several key advantages over existing approaches for predicting gene expression: (i) flexible integration of heterogeneous multi-omic data (DNA sequence, epigenomic signals, chromatin contacts); (ii) explicit modeling of long-range promoter–enhancer interactions via dedicated attention module, improving enhancer–gene link prediction and interpretability, as validated by CRISPR perturbations; and (iii) compact, efficient training (0.4M parameters; around 1 hour on a single GPU for 18,377 protein-coding genes). These capabilities enable EPInformer to outperform state-of-the-art methods while remaining computationally efficient."

2. The average precision (AP) is higher for the Attention-ABC score compared to the Attention score with Hi-C, as shown in Figure 3c. From my understanding, the ABC score is a straightforward method that combines features. This raises a concern about the proper utilization of Hi-C data.

We thank the reviewer for this thoughtful concern. In line with a similar comment from another reviewer, we carefully redesigned the interaction encoder's final layer to explicitly and better incorporate cell-type-specific Hi-C contact and enhancer activity directly within the attention module. Our experiments show that this new attention module that better utilize Hi-C significantly outperforms the ABC score, achieving a higher AUPRC (0.732 vs. 0.698).

We have added the following paragraph to the Methods describing this revised attention module, and **Fig. 3** illustrates its architecture:

“To better capture distal enhancer-gene interactions, we weight each enhancer embedding in the final encoder layer by Hi-C contact (or inverse distance to the TSS if this assay is not available) and enhancer activity. Formally, the interaction of promoter X_P and the i_{th} enhancer X_{E_i} can be computed as $a_{P-E_i} = \text{softmax}(Q_P \cdot K_{E_i}^T / \sqrt{d_k} \cdot F_{A_i} \cdot F_{C_i})$, where F_{A_i} denotes enhancer activity (geometric mean of H3K27ac and DNase RPM) and F_{C_i} denotes the enhancer-gene Hi-C contacts or 1/distance to gene's TSS. Each attention head computes a weighted sum over promoter-enhancer pairs: $a_{P-E} \times V$. Multiple heads operate with independent parameters; their outputs are concatenated and passed through a linear projection to produce the final layer output. Consequently, in the interaction encoder's final layer, a_{P-E} quantifies the learned enhancer-gene interaction intensity and is used to prioritize enhancers for each gene.”

The revised attention score (both distance-based and Hi-C-based) outperformed the corresponding ABC scores in terms of AUPRC (**Fig. 3b**). Attention.score.Hi-C achieved the highest AUPRC of 0.732. Moreover, the attention.score.Hi-C consistently exceeded the ABC score across genomic distance bins, with a notable advantage for very distal links (AUPRC 0.51 vs. 0.43 at 60-100 kb), indicating that our new attention mechanism more effectively captures distal enhancer-gene interactions.

We have revised the Result section titled “EPInformer accurately prioritizes cell-type-specific enhancers-gene interactions” with the updated results and updated Fig. 3 accordingly:

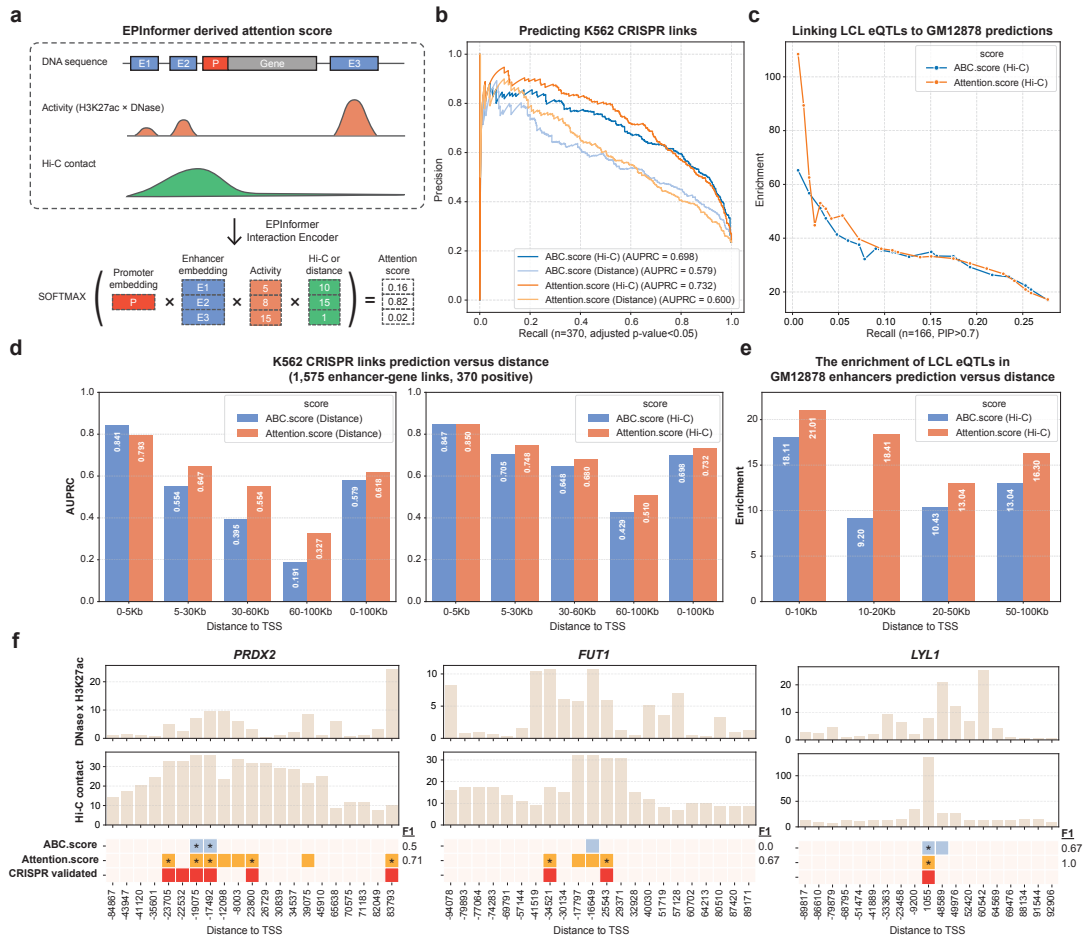


Fig. 3. | EPInformer attention prioritizes cell-type-specific enhancers corroborated by long-range interaction assays. *a*, Schematic of attention score calculation in EPInformer-PE-Activity(-HiC). The red and blue boxes represent three candidate enhancers (E1, E2, and E3) located near the promoter (P) of the gene (grey box). The dashed box shows the simplified calculation of the attention score for promoter-enhancer pairs, incorporating DNase, H3K27ac, and Hi-C signals (expressed in arbitrary units, not to scale). *b*, Precision–recall curves for predicting enhancer–gene (E–G) pairs. Positive E–G pairs are defined as cases where perturbing the candidate enhancer significantly decreases expression of the target gene. The curves compare attention-based scores (Attention.score.Hi-C from EPInformer-PE-Activity-HiC and Attention.score.distance from EPInformer-PE-Activity) with two ABC scores in classifying 1,575 E–G pairs across 244 genes assayed by CRISPR perturbations. *c*, Enrichment–recall curves for enhancer–gene interaction scores in GM12878 cells. The y-axis shows enrichment—the ratio of fine-mapped distal noncoding eQTLs (PIP > 70% in GM12878) that fall within predicted enhancers relative to all distal noncoding common variants. The x-axis shows recall—the fraction of eQTL variants overlapping enhancers linked to the correct gene. Curves represent varying score thresholds for enhancers predicted by different models. *d*, Area under the precision-recall curve (AUPRC) values for E-G score performance shown in panel *b* when classifying enhancer-gene pairs at three different distance ranges. *e*, Enrichment of putative causal variants (PIP > 70%) within predicted enhancers across different distance ranges. *f*, Top three genes with the largest per-gene F1 score gains when comparing Attention.score.Hi-C to ABC.score.Hi-C. The first two rows display enhancer activity and Hi-C contact values computed by the ABC pipeline for each candidate enhancer. Light blue boxes indicate enhancers predicted by ABC.score.Hi-C, orange boxes indicate enhancers predicted by Attention.score.Hi-C. Red boxes indicate CRISPR-validated enhancers. Stars indicate correctly predicted enhancers.

3. Did the study take into account the negative regulation from regulatory elements to target genes?

We thank the reviewer for raising this thoughtful question. The current EPInformer focuses on modeling positive regulation from regulatory elements, as it uses H3K27ac signal—a mark of active enhancers—as an

input for gene expression prediction. However, the model can capture some aspects of negative regulation indirectly, as regions lacking H3K27ac and DNase signals would receive lower activity scores and thus contribute less to predicted expression.

Looking forward, the framework can be extended to explicitly model negative regulation by incorporating repressive histone mark ChIP-seq signals (e.g., H3K9me2, H3K9me3, H3K27me3) alongside H3K27ac. This would enable EPInformer to distinguish between actively repressed regions and simply inactive regions, potentially improving prediction accuracy for genes under active repression.

We have added the following text to the Discussion section acknowledging this important limitation and future direction:

"While the current implementation focuses on activating epigenomic marks (H3K27ac and DNase-seq), the framework can capture some aspects of negative regulation indirectly through the absence of these signals. Future versions could explicitly incorporate repressive histone modifications (e.g., H3K9me3, H3K27me3) to distinguish between actively repressed and simply inactive regions, potentially improving predictions for genes under active repression."

4. 'To this end, we used TF-MoDISco-lite³⁰ and TangerMEME³¹ to examine the TF motifs within enhancers, based on EPInformer-PE-Activity's sequence encoder pre-trained to predict enhancer activity from the sequence.' In the above sentence, does the 'pre-trained' mean the model for predicting the gene expression? How to get this pre-trained model, and why could it be used to predict the enhancer activity?

We thank the reviewer for raising this point, which we had not clearly explained in the manuscript. To clarify: the 'pre-trained' model refers to EPInformer's sequence encoder (EPInformer-seq) that we first trained separately to predict enhancer activity, not gene expression. In response, we added a clearer workflow schematic in **Fig. 4a** and **Supplementary Figure 1** that illustrates how we train and leverage the EPInformer sequence encoder (EPInformer-seq) for motif analysis and gene expression modeling.

EPInformer-seq was pre-trained to predict enhancer activity from DNA sequence. Enhancer activity is defined as the geometric mean of H3K27ac and DNase RPM at 256-bp sequences, in accordance with the ABC score protocol. Specifically, we used 256-bp sequences centered on cell-type-specific H3K27ac peaks and included two flanking sequences with 100-bp overlaps. We employed 12-fold leave-two-chromosomes-out cross-validation (matching EPInformer) to prevent data leakage prior to gene-expression prediction. This pre-training serves two purposes: (1) the convolutional layers learn enhancer-related sequence patterns that improve subsequent gene expression modeling, and (2) the trained model can be analyzed with interpretation tools like TF-MoDISco-lite to discover transcription factor motifs at enhancer regions. As the pre-trained model was optimized to predict enhancer activity, its convolutional layers have learned enhancer-related sequence patterns, which benefit gene-expression modeling when incorporated into the full EPInformer model.

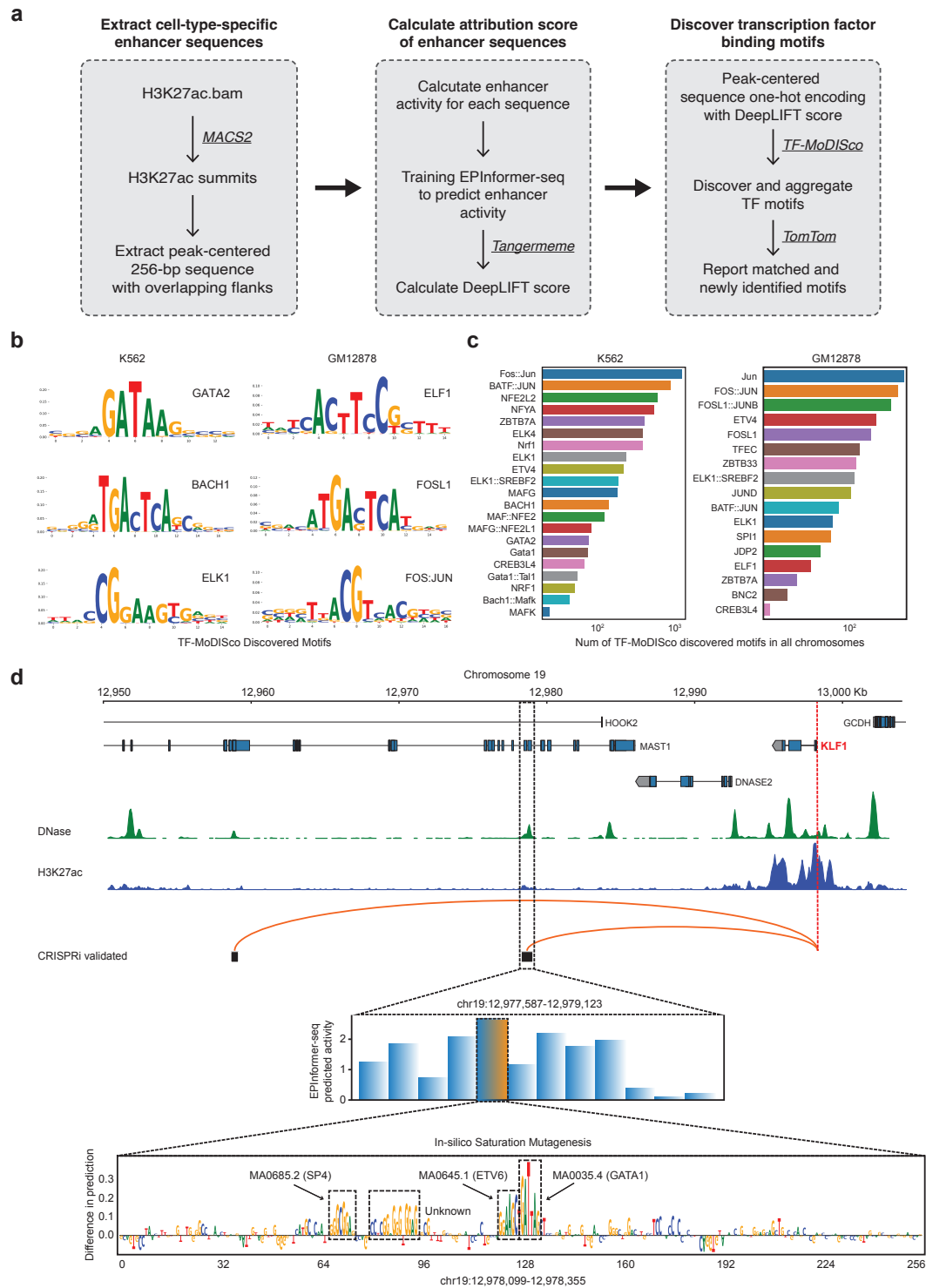


Fig 4. | EPInformer reveals transcription factor motifs at cell-type-specific enhancers. a, The workflow of using EPInformer-seq with TF-MoDISco-lite and tangermeme to discover transcription factor binding motifs. **b**, set of three motifs enriched in K562 (left) and GM12878 (right) discovered by TF-MoDISco-lite by summarizing recurring EPInformer-PE-Activity predictive sequence patterns at all H3K27ac peaks. **c**, bar plot showing the enrichment of motifs discovered by TF-MoDISco-lite that match JASPAR 2024 CORE vertebrate non-redundant database (q -value < 0.05) in K562 (left) and GM12878 (right). **d**, EPInformer discovers several important transcription factor binding motifs at a putative enhancer (in the dotted box) of KLF1. The orange dashed line indicates the transcription start site (TSS) of KLF1. The black box with a red arc represents the CRISPRi-validated enhancer, while the grey box denotes the EPInformer-PE-Activity predicted enhancer. The bar plot displays predicted enhancer activities (geometric mean of the H3K27ac and DNase signals) for 256-bp sequences tiling the putative enhancer locus (chr19:12,977,587-12,979,123).

The bottom panel shows in-silico mutagenesis (ISM) attribution score for the region (chr19:12,978,099-12,978,355) with the highest predicted enhancer activity shown at the bottom. High-scoring sequence motifs are highlighted with dashed boxes, with labels and arrows indicating matching TF motifs from the JASPAR 2024 CORE vertebrate non-redundant database.

5. I did not see the details regarding the methods used for predicting enhancer activity. Is the enhancer activity represented as a single score per base pair or as one score for each enhancer? Additionally, is the input solely the DNA sequence? I apologize if I missed some of these details.

Thank you for seeking clarification on enhancer activity prediction. We apologize for not making these details clearer in the manuscript.

Enhancer activity is represented as one score for each 256-bp enhancer region, not per base pair. The input to EPInformer-seq is solely the DNA sequence (256-bp, one-hot encoded). The model outputs a single activity score for each sequence, which represents the predicted geometric mean of H3K27ac and DNase RPM for that enhancer region.

For the motif analysis shown in **Fig. 4d**, we tile across the enhancer locus using overlapping 256-bp windows to identify the region with highest predicted activity. We then apply DeepLIFT or in-silico mutagenesis (ISM) to that specific window to obtain base-resolution attribution scores for motif discovery.

We have added the following clarification to the Methods section:

"EPInformer-seq takes as input a 256-bp DNA sequence (one-hot encoded) and outputs a single enhancer activity score representing the predicted geometric mean of H3K27ac and DNase signals for that region. For base-resolution analysis, we apply interpretation methods (DeepLIFT or ISM) to derive nucleotide-level attribution scores from the trained model."

6. Why are some of the expression values smaller than 0 in Figure 1a (right)?

We thank the reviewer for this observation. The expression values below 0 in Figure 2a (right panel) occur because we use log-transformed RNA-seq expression values (RPKM: Reads Per Kilobase of transcript per Million mapped reads), following the Xpresso preprocessing protocol. The preprocessing includes also adding a small pseudocount before log transformation to handle zero expression values. When RPKM values plus pseudocount are between 0 and 1, their log₁₀ transformation yields negative values. This is a common practice for handling the wide dynamic range of gene expression data and helps in modeling both lowly and highly expressed genes effectively.

We have clarified this in the Methods section:

"To train and evaluate EPInformer on gene expression prediction, we curated two gene expression datasets, as measured by RNA-seq and CAGE. For CAGE, expression values were determined by aggregating read counts within 384-bp regions centered at each gene's unique TSS, as per Enformer's protocol. RNA-Seq expression data were sourced from Xpresso's training set, quantified by the Roadmap Epigenomics Consortium. To mitigate the right-skewed distribution of gene expression based on raw read count, we applied log₁₀ transformation with one to handle zero values."

Reviewer #4 (Remarks on code availability):

The code is usable. I can install and run the code.

Thank you so much for testing our code. We're glad it was runnable without errors. We have made efforts to ensure the code is well-documented and easy to use, and we appreciate your confirmation that the installation and execution process worked smoothly.

References

1. Marin, F. I. *et al.* BEND: Benchmarking DNA Language Models on biologically meaningful tasks. *arXiv [q-bio.GN]* (2023).
2. Feng, H. *et al.* Benchmarking DNA foundation models for genomic sequence classification. *bioRxiv* 2024.08.16.608288 (2024).
3. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–961 (2025).
4. Zhou, Z. *et al.* DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv [q-bio.GN]* (2023).
5. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 1–11 (2024).
6. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv [cs.LG]* (2023).
7. Wang, Y. *et al.* Genomic Touchstone: Benchmarking genomic language models in the context of the central dogma. *bioRxiv* 2025.06.25.661622 (2025) doi:10.1101/2025.06.25.661622.
8. Tang, Z., Toneyan, S. & Koo, P. K. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nat. Genet.* **55**, 2021–2022 (2023).
9. Karollus, A., Mauormeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).
10. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv* 2023.11.09.563812 (2023) doi:10.1101/2023.11.09.563812.
11. Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390.e19 (2019).
12. Klann, T. S. *et al.* CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
13. Yao, D. *et al.* Multicenter integrated analysis of noncoding CRISPRi screens. *Nat. Methods* **21**, 723–734 (2024).
14. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).
15. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
16. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
17. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
18. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
19. Schreiber, J. tangermeme: A toolkit for understanding cis-regulatory logic using deep learning models. *bioRxiv* 2025.08.08.669296 (2025) doi:10.1101/2025.08.08.669296.
20. Shrikumar, A. *et al.* Technical note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *arXiv [cs.LG]* (2018).
21. Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
22. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).
23. Koido, M. *et al.* Prediction of the cell-type-specific transcription of non-coding RNAs from genome sequences via machine learning. *Nat. Biomed. Eng.* **7**, 830–844 (2023).
24. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).

Reviewer Response

We thank all reviewers for their continued evaluation of our manuscript. We are pleased to see that the reviewers indicated that most concerns from the first round have been satisfactorily addressed. In this revision, we have made the following improvements to address the remaining comments:

- 1. Clarified benchmarking structure:** We have revised Figure 2 and the accompanying text to clearly distinguish between the different benchmarking protocols, explaining why different baseline methods appear in different panels due to distinct train-test splits from the original studies.
- 2. Added Xpresso comparison for CAGE:** We retrained Xpresso on the CAGE dataset and demonstrate that all EPIformer models consistently outperform Xpresso across all benchmarks (new Supplementary Figure 9).
- 3. Added MENTR comparison:** Following Reviewer #3's suggestion, we conducted a direct comparison with MENTR on the FANTOM5 CAGE dataset, demonstrating that EPIformer-PE-Activity-HiC significantly outperforms MENTR (Pearson $r = 0.792$ vs. 0.675 ; new Supplementary Figure 8).

We believe these revisions fully address all remaining concerns. Our detailed responses follow below.

Reviewer #1 (Remarks to the Author)

Thank you for the author's response. Most of the issues have been resolved. However, there is still one confusing question regarding Figure 2. In the benchmark section, why are the benchmark experiments not comprehensive? For example, Figures 2b and 2d appear to represent the same task (at least I cannot discern any differences from the captions and titles).

Why is only Enformer compared in 2b, while only Borzoi is compared in 2d?

Additionally, why do the numerical values for the same method differ between 2b and 2d? For instance, EPIformer_PE_Activity_HiC is 0.856 in 2b but 0.892 in 2d. If they represent the same experimental setup, the two figures should be merged and redrawn. If they represent different setups, the differences between the two figures should be clearly

distinguished, and baseline experiments should be supplemented (e.g., Borzoi should be added to 2b, while Enformer, Seq-graphReg, etc., should be added to 2d).

The same applies to Figures 2a and 2c. Both are cross-chromosome validation tasks, so why is Xpresso included in 2a but replaced by Seq-graphReg in 2c?

Placing them in separate figures is confusing. I recommend that the authors maintain consistency in the baseline methods across all figures. Currently, there appear to be five baseline methods in total, so the results of these five baseline methods should be included in all benchmark experiments (including 2a–2d).

If this is not feasible or inconvenient to implement, the situation should be explicitly explained (e.g., "Method XXX cannot be included in Figure 2a because...").

We thank the reviewer for this important point regarding clarity of the benchmarking structure. The key issue is that Enformer, Borzoi, and Seq-GraphReg each define their own train-test splits, and retraining these models requires substantial computational resources. Therefore, we evaluated their officially released pre-trained models on their original test sets. To ensure fair comparison, we trained EPInformer and CREaTor using the identical splits defined by each benchmark. This is why the same EPInformer model yields different correlation values across panels—the test genes differ.

Following the reviewer's suggestion, we now explicitly explain why certain methods appear in specific panels:

Panel a (RNA-seq, 12-fold cross-chromosome validation): Enformer and Seq-GraphReg cannot be included because they were not designed or trained for mRNA abundance level prediction; they predict CAGE expression.

Panel b (CAGE, Enformer's test set): Borzoi and Seq-GraphReg cannot be included because they use different train-test splits, and retraining these models is not feasible.

Panel c (CAGE, Seq-GraphReg's 12-fold cross-chromosome validation): Enformer and Borzoi cannot be included for the same reason, retraining them on this split is not feasible.

Panel d (CAGE, Borzoi's test set): Enformer and Seq-GraphReg cannot be included because they use different train-test splits, and retraining is not feasible.

To ensure fair comparisons, we trained EPInformer and CREaTor on the identical splits defined by each benchmark, which is why these methods appear consistently across panels.

We have now explicitly annotated each panel in the revised Figure 2 to indicate the data source and testing set.

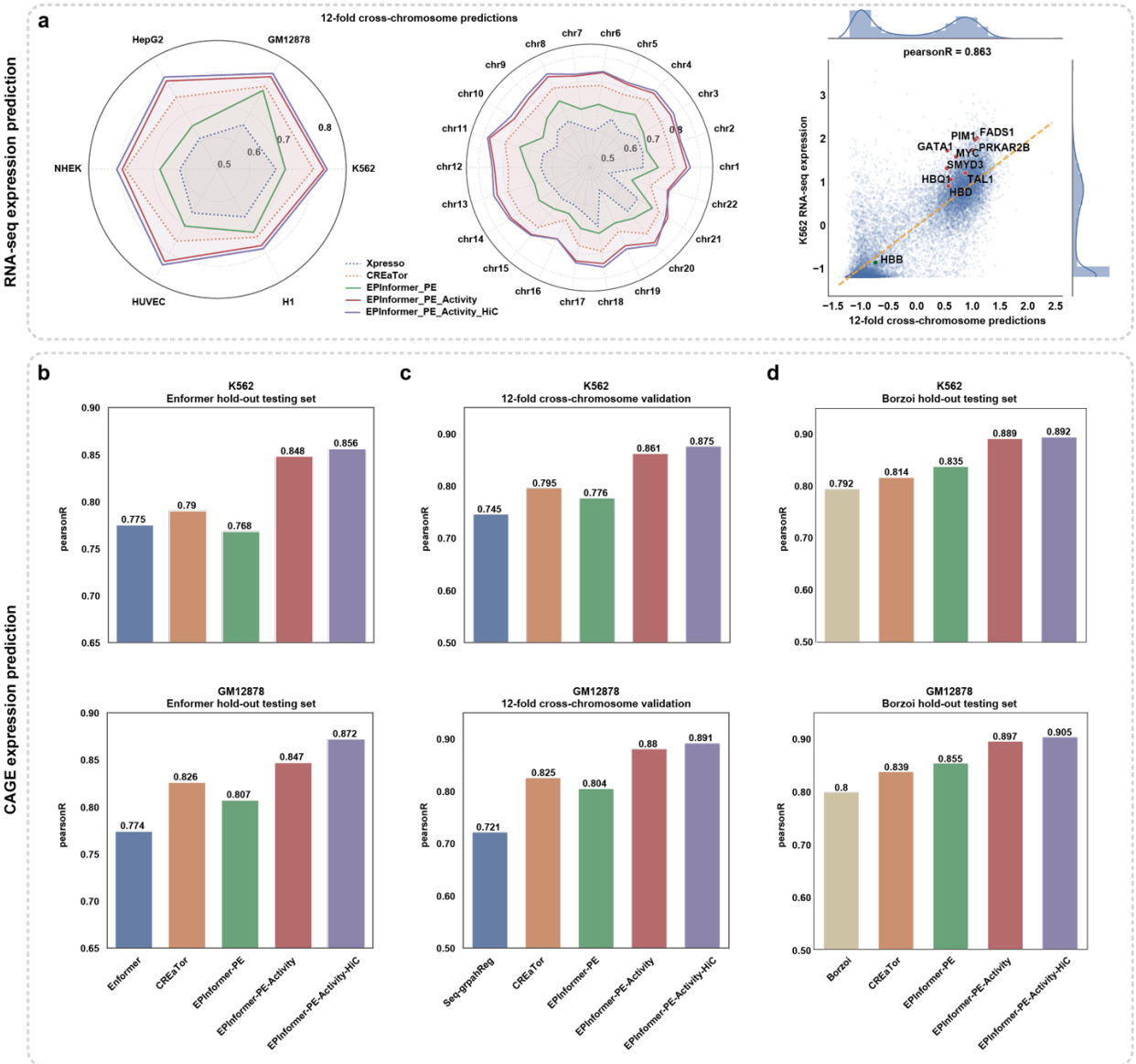
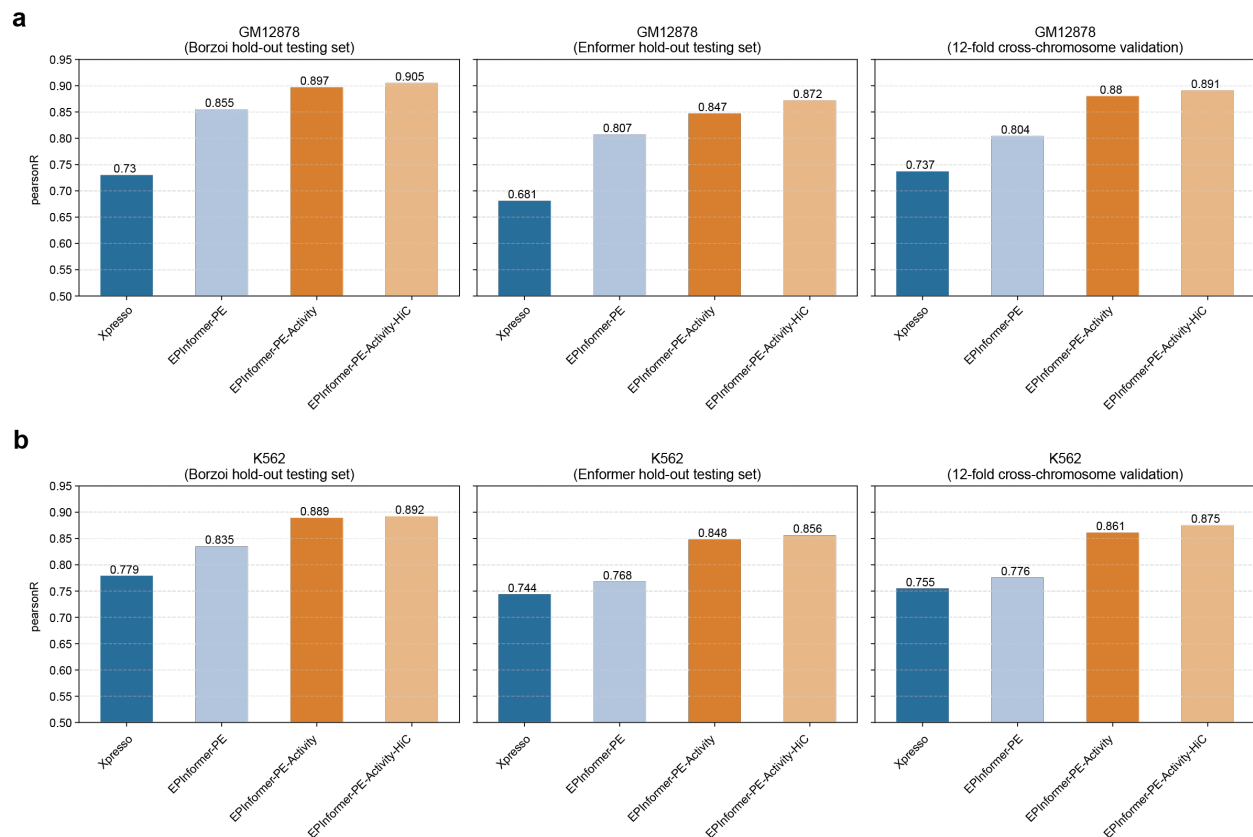


Fig 2. | Performance of EPIInformer models and baseline methods on gene expression prediction. **a**, Pearson correlation between observed and predicted RNA-seq expression levels for 18,377 protein-coding genes across six cell lines. Left: Spider plot showing correlation between observed and predicted expression levels for six cell lines. Middle: Spider plot showing correlation between observed and predicted expression levels for each test chromosome across six cell lines. Right: Scatter plot showing relationship between EPIInformer-PE-Activity-HiC predicted and observed RNA-seq expression levels in K562 cells, with each point representing an individual gene. Genes with high expression in K562 are highlighted in red, and those with low expression are highlighted in green. All

evaluations were conducted using 12-fold cross-chromosome validation. **b**, Comparison of EPIInformer models with Enformer for predicting CAGE gene expression levels **using Enformer's hold-out test set** in K562 (top) and GM12878 (bottom) cells. **c**, Comparison of EPIInformer models with Seq-GraphReg and CREaTor for CAGE gene expression prediction in K562 (top) and GM12878 (bottom) cells using 12-fold cross-chromosome validation. **d**, Comparison of EPIInformer models with Borzoi and CREaTor for predicting CAGE gene expression levels **using Borzoi's hold-out test set** in K562 (top) and GM12878 (bottom) cells.

Additionally, following the reviewer's suggestion, we retrained Xpresso (originally designed for RNA-seq) on the CAGE dataset using the same train-test splits. As shown in Supplementary Figure 9, all EPIInformer models consistently outperform Xpresso across both cell lines and all three CAGE benchmarks.



Supplementary Figure 9. Performance evaluation of EPIInformer variants versus Xpresso across different benchmarks for (a) K562 and (b) GM12878 CAGE prediction.

We have updated the Results section to clearly explain these distinctions:

"To rigorously evaluate the EPIInformer models, we conducted separate benchmarking for CAGE and RNA-seq expression. For RNA-seq prediction, we compared Xpresso and CREaTor against three EPIInformer models using 12-fold cross-chromosome validation (Fig. 2a). In each fold, two chromosomes were designated for testing, two for validation, and the remainder for

training, following protocols established in previous studies. Regarding CAGE expression benchmarking, retraining Seq-GraphReg, Enformer, and Borzoi requires substantial computational resources and complex data preprocessing. Therefore, we evaluated their officially released pre-trained models using the original hold-out test sets defined in their respective studies (Fig. 2b-d). To ensure a fair comparison, the EPInformer models were trained and tested using these identical splits. Performance across all genes was assessed using the Pearson Correlation Coefficient between predicted and observed expression levels."

And:

"Given that the original Xpresso was designed to predict RNA-seq expression, we retrained and evaluated it on the CAGE dataset to enable a comparison with the EPInformer models across the three benchmarks in Fig. 2b-d. As expected, all three EPInformer models significantly outperformed Xpresso in CAGE expression prediction, consistent with our findings in RNA-seq prediction (Supplementary Fig. 9)."

Reviewer #2 (Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for their participation in the peer review process and their contribution to evaluating our manuscript.

Reviewer #3 (Remarks to the Author)

The authors have addressed most of my previous comments satisfactorily. I appreciate their comparison with Borsoi et al. and their decision to extend the analyses to the broader set of FANTOM5 transcription start sites (TSSs), rather than restricting them to main transcripts.

I would like to clarify that MENTR is not a tool specifically designed to infer non-coding RNA expression. It learns from FANTOM5 TSS data, which include both coding (i.e., promoters) and non-coding RNAs. Since the authors have now expanded their analyses to include all FANTOM5 TSSs, a direct comparison between their method and MENTR would be both appropriate and informative.

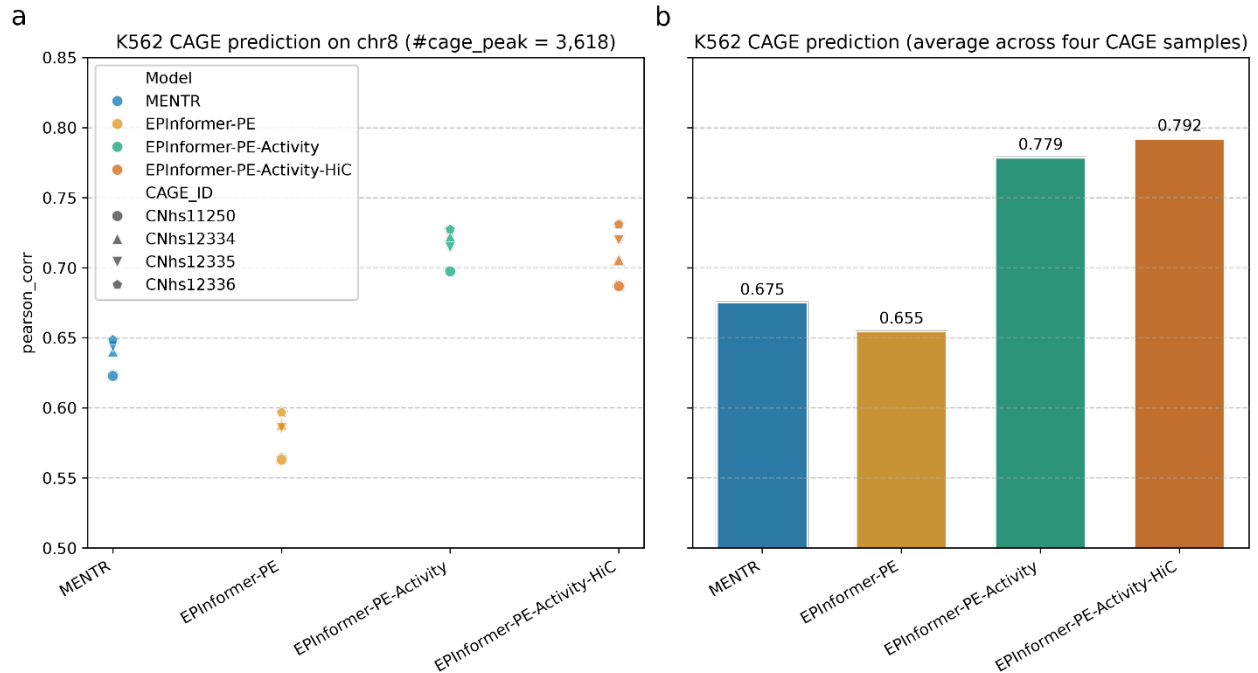
We thank the reviewer for this clarification regarding MENTR and agree that a direct comparison is appropriate. To address this, we communicated with the MENTR authors to obtain their code for CAGE expression prediction and conducted a comprehensive comparison.

We used the identical dataset comprising 92,473 CAGE peaks annotated with coding transcripts. We trained and tested three EPInformer models for each of four individual K562 samples (CNhs11250, CNhs12334, CNhs12335, and CNhs12336) as well as for their averaged signal. Following the MENTR framework, we designated Chromosome 8 for testing, Chromosomes 7 and 17 for validation, and the remaining chromosomes for training. The MENTR model was retrained following the protocols in its official GitHub documentation (<https://github.com/koido/MENTR/wiki/Train-MENTR-ML-models#run-training>).

The results demonstrate that while MENTR outperformed the sequence-only model (EPInformer-PE), the integration of enhancer activity signals (EPInformer-PE-Activity) and Hi-C interactions (EPInformer-PE-Activity-HiC) substantially improved accuracy, enabling these models to surpass MENTR. EPInformer-PE-Activity-HiC achieved the highest Pearson correlation (0.792) compared to MENTR (0.675) when predicting average CAGE signals across 3,618 coding transcripts on chromosome 8.

We have added this comparison to the manuscript and included the results in **Supplementary Figure 8**:

"We also compared our models with MENTR using a larger CAGE dataset comprising 92,473 mRNA coding transcripts. EPInformer-PE-Activity-HiC significantly outperformed MENTR, achieving a Pearson correlation of 0.792 compared to MENTR's 0.675 when predicting average CAGE signals across four K562 samples (Supplementary Fig. 8)."



Supplementary Figure 8. Performance evaluation of EPIInformer variants versus MENTR for K562 CAGE prediction. **a**, Pearson correlation coefficients assessing the accuracy of predicted CAGE expression levels on chromosome 8 (chr8; $n = 3,618$ CAGE peaks). Performance is compared across the baseline MENTR model and three EPIInformer architectures (PE, PE-Activity, and PE-Activity-HiC). Individual data points correspond to four distinct K562 CAGE replicates (CNhs11250, CNhs12334, CNhs12335, and CNhs12336), denoted by unique marker shapes. **b**, Bar chart illustrating the mean Pearson correlation averaged across the four CAGE samples for each model. The proposed EPIInformer-PE-Activity-HiC model demonstrates the highest predictive performance (Pearson $r = 0.792$), outperforming the baseline MENTR (0.675) and other EPIInformer variants.

We thank all reviewers for their careful evaluation of our manuscript. Their insightful comments and suggestions have significantly improved the quality, clarity, and completeness of our work.

Reviewer #1 (Remarks to the Author)

All questions have been resolved and happy to see this paper will be accepted in the future!

We sincerely thank the reviewer for their constructive suggestions and thorough evaluation of our manuscript. We are pleased to see that the reviewers indicated that all concerns have been satisfactorily addressed.

Reviewer #2 (Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for their participation in the peer-review process and for their valuable contribution to the evaluation of our manuscript.