Supplementary tables

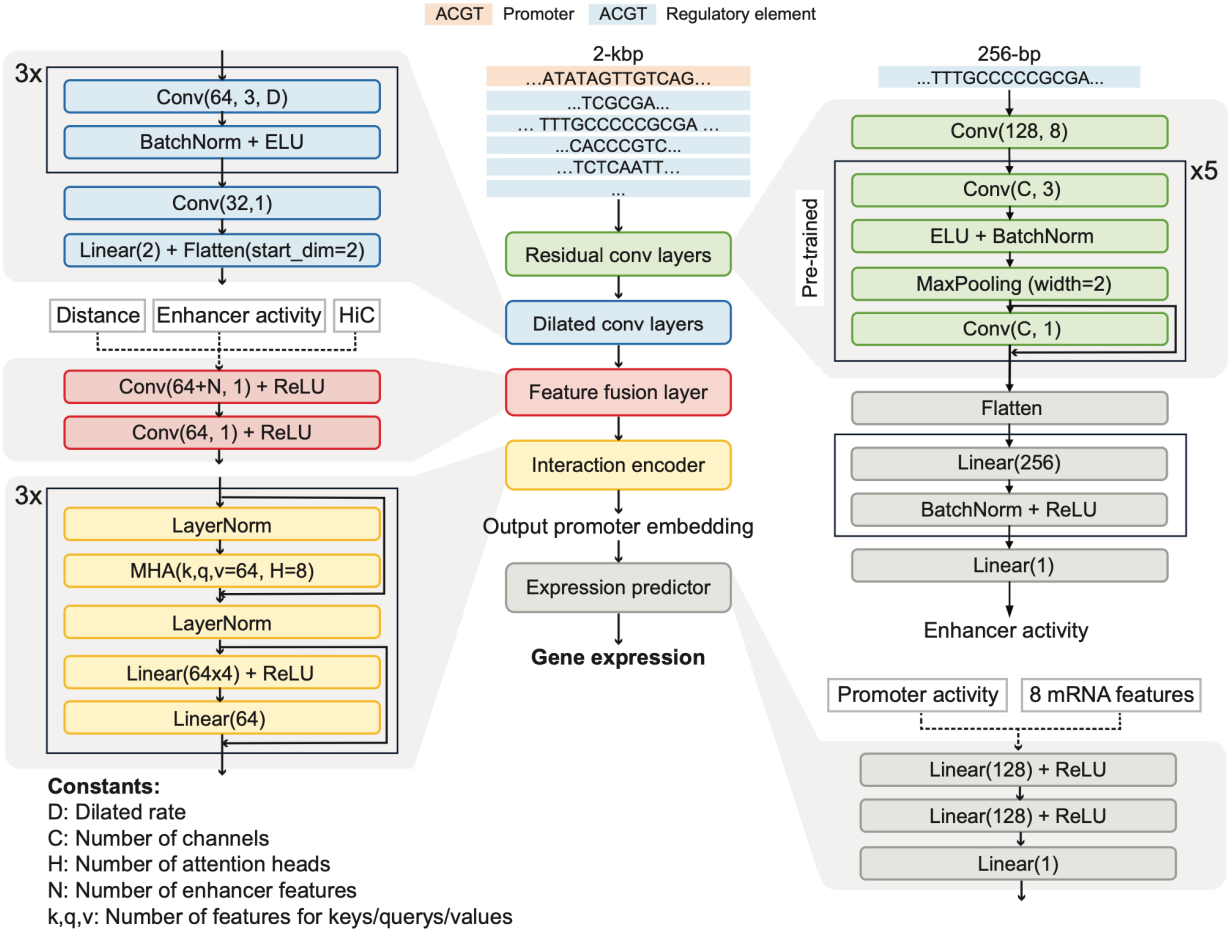
Supplementary Table 1. The required input data for predicting the gene expression. The row highlighted in orange are EPInformer models.

Model	Promoter	Enhancer	DNase	H3K27ac	Hi-C
Xpresso	10kb around TSS		No	No	No
GraphReg	1 Mb		Yes	Yes	Yes
CREaTor	The closest 200 CREs around TSS		Yes	Yes	No
Enformer	196 kb		Yes	Yes	No
Borzoi	524 kb		Yes	Yes	No
PE	1kb around TSS	100k around TSS	Yes	No	No
PE-Activity	1kb around TSS	100k around TSS	Yes	Yes	No
PE-Activity-HiC	1kb around TSS	100k around TSS	Yes	Yes	Yes

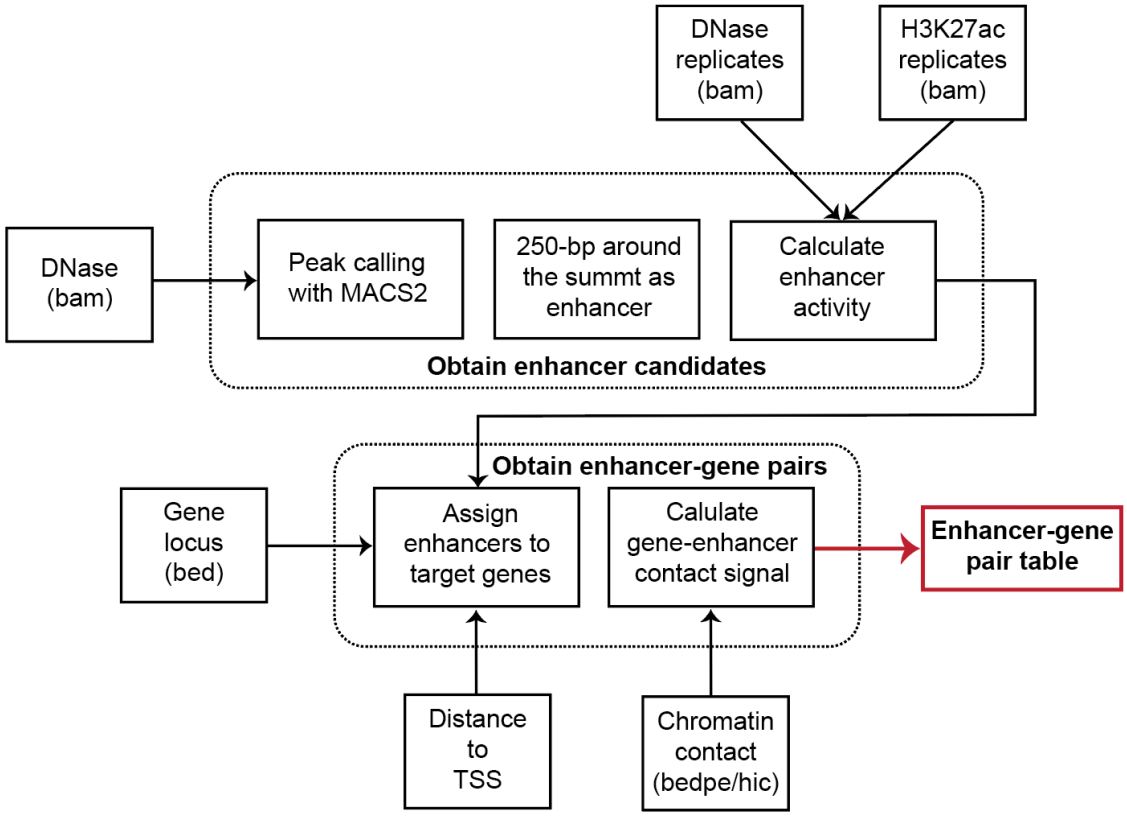
Supplementary Table 2. The parameters and training time of gene expression

Model	#Parameters	Computing resources	Training time
Enformer	250M	64 TPU v3 cores	~3 days
Borzoi	186M	2 Tesla A100	~25 days
CREaTor	5M	1 Tesla A100	~3 hours
Xpresso	0.1M	1 Tesla A100	~5 mins
EPInformer	0.4M	1 Tesla A100	~1 hour

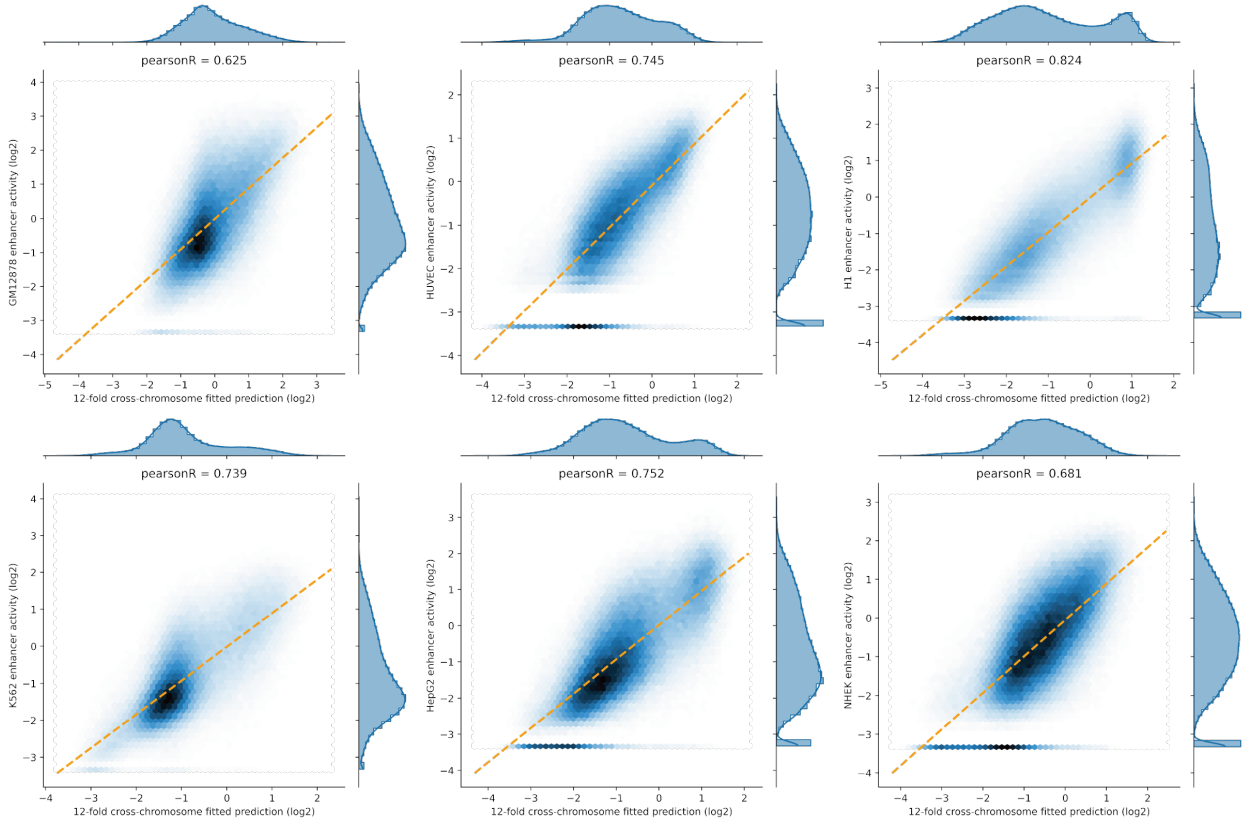
Supplementary figures



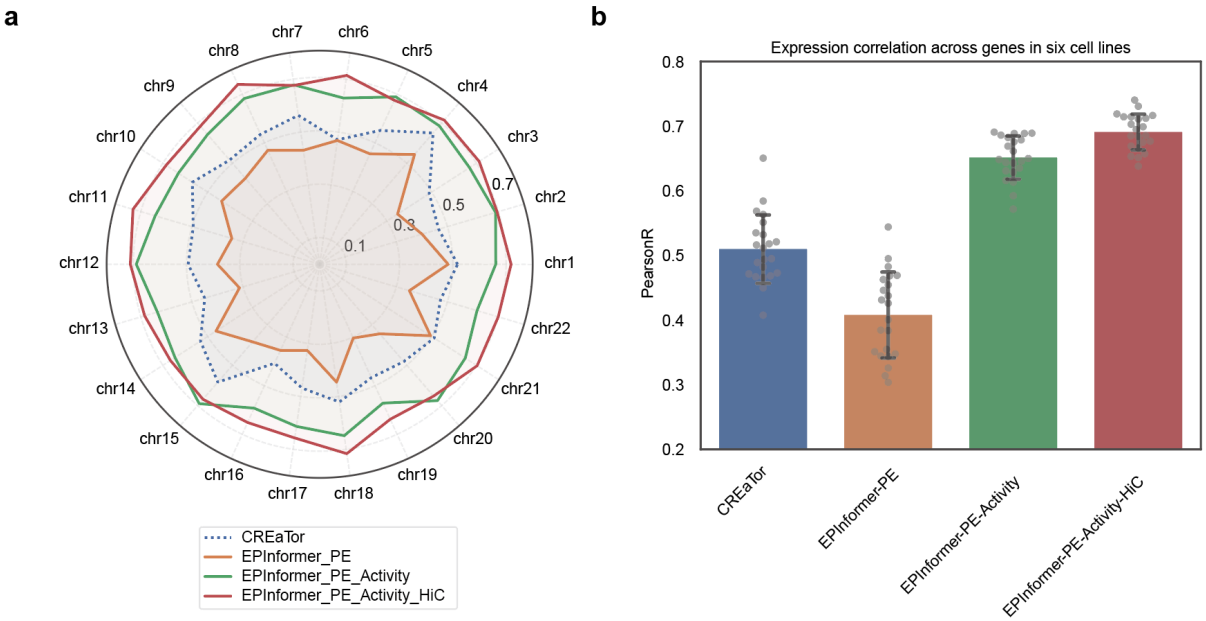
Supplementary Figure 1. Structure of EPInformer models. All proposed EPInformer models in our study are based on this core architecture. EPInformer-promoter masks all enhancer sequences with a zero vector and does not use any extra features; EPInformer-PE is a sequence-only model, using distance as the only additional feature. EPInformer-PE-Activity employs pre-trained residual convolution layers and takes promoter-enhancer sequences, distance, enhancer activity, promoter activity, and mRNA half-life features as inputs; EPInformer-PE-Activity-HiC includes also HiC contact data; MHA stands for Multi-Head Attention.



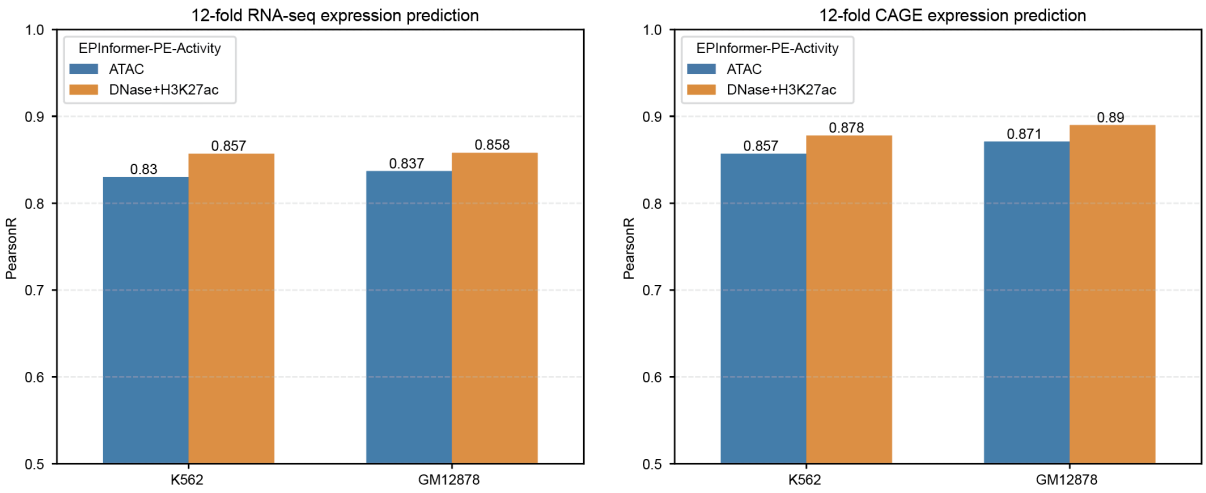
Supplementary Figure 2. Schematic overview of ABC Pipeline for nominating enhancer-gene pairs



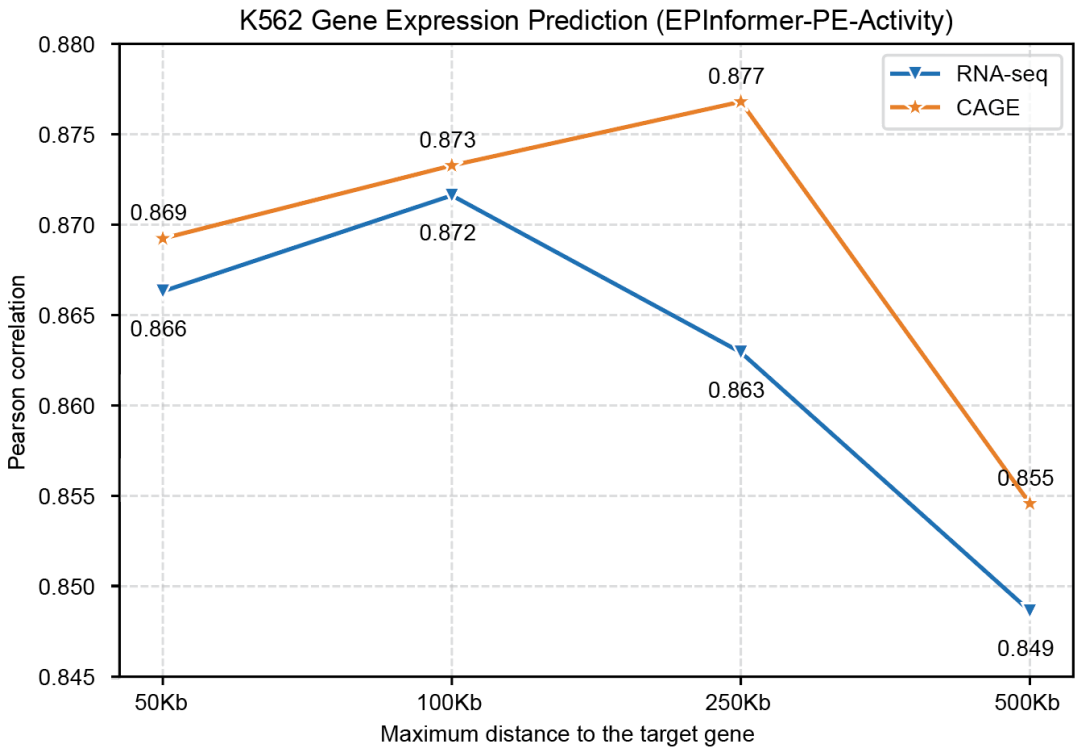
Supplementary Figure 3. EPInformer sequence encoder predicts enhancer activity quantitatively. The pre-trained sequence encoder quantitatively predicts enhancer activity (geometric mean of H3K27ac and DNase signals) at all H3K27ac peaks genome wide for six cell lines. The scatter plots show observed versus predicted enhancer activity across all DNA sequences in 12-fold cross-chromosome validation for six cell lines, with point density indicated by color. Source data are provided as a Source Data file.



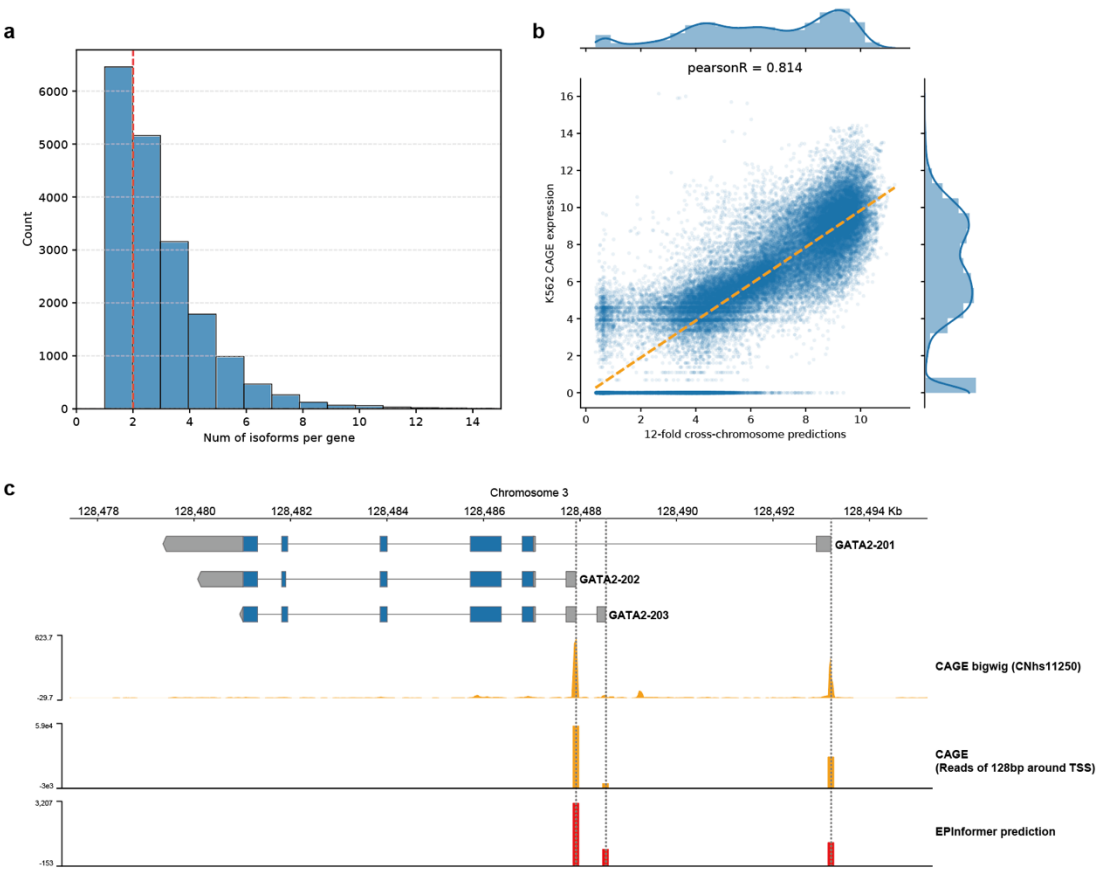
Supplementary Figure 4. EPInformer-PE-Activity-(HiC) outperforms CREaTor in predicting RNA-seq expression levels across genes. a) Comparison of 12-fold cross-chromosome validation performance (median Pearson correlation between observed and predicted RNA-seq expression across cell lines for each testing chromosome) of CREaTor and EPInformer models across 18,377 protein-coding genes in six cell lines (K562, GM12878, HepG2, NHEK, HUVEC, H1). b) Distribution of Pearson correlations for EPInformers and CREaTor between observed and predicted RNA-seq expression in each testing chromosome. Error bars represent the 25th and 75th percentiles (interquartile range). Source data are provided as a Source Data file.



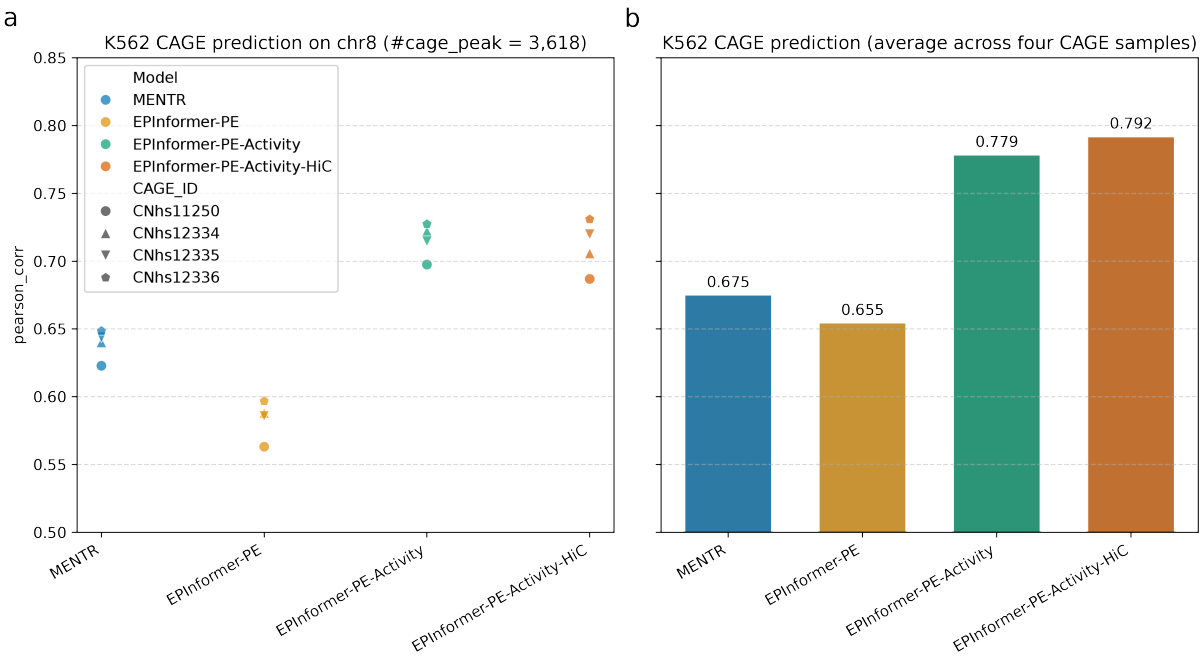
Supplementary Figure 5. EPInformer-PE-Activity performance using ATAC-seq versus DNase-seq and H3K27ac. Comparison of 12-fold cross-chromosome validation performance for EPInformer-PE-Activity trained with ATAC-seq signals versus DNase-seq and H3K27ac signals for predicting gene expression in K562 and GM12878 cells. ATAC-based EPInformer-PE-Activity achieved comparable performance to DNase+H3K27ac-based EPInformer-PE-Activity, demonstrating the framework's flexibility for cell types where only ATAC-seq data is available.



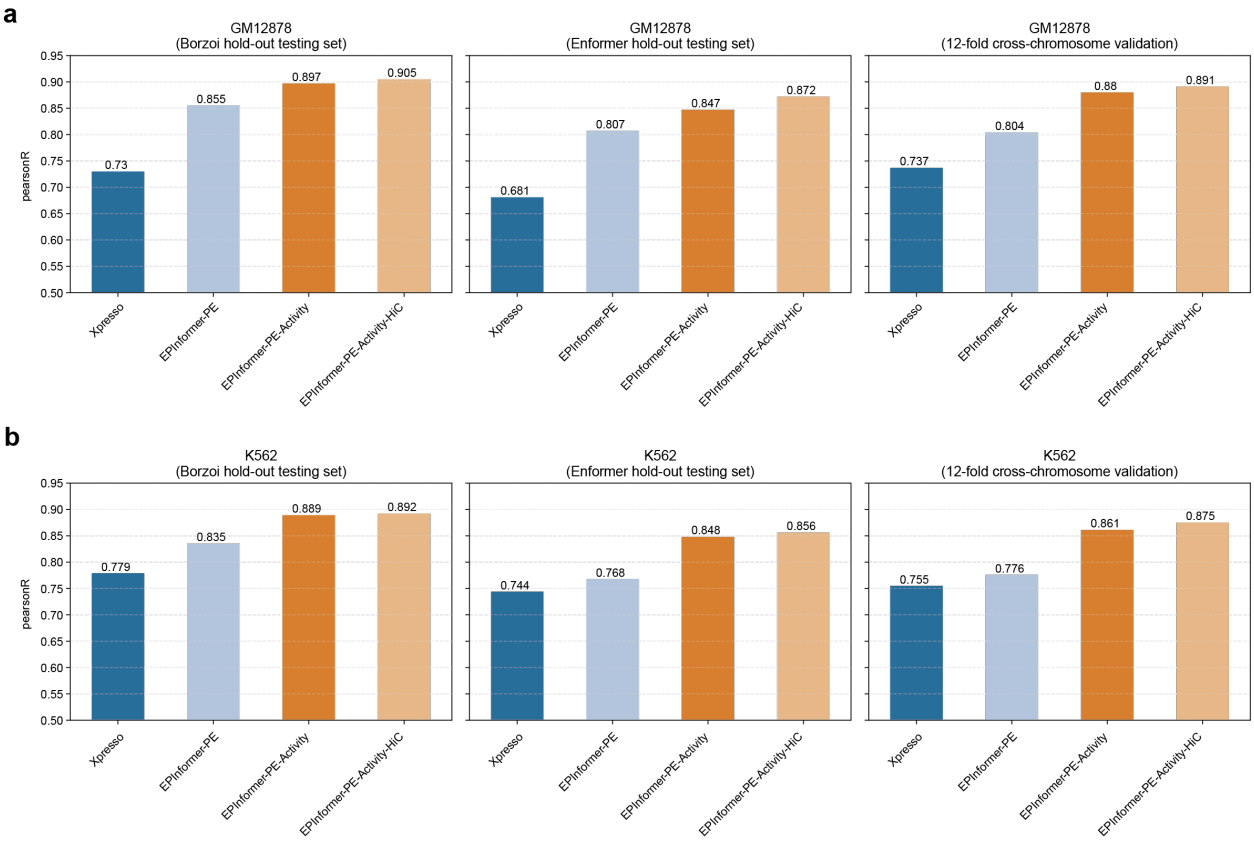
Supplementary Figure 6. Effect of genomic distance range on EPInformer gene expression prediction performance. Performance of EPInformer-PE-Activity for predicting RNA-seq and CAGE expression in K562 cells using different maximum distances from the TSS to define candidate enhancers. The x-axis shows 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.



Supplementary Figure 7. Performance of EPInformer-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. Source data are provided as a Source Data file.



Supplementary Figure 8. Performance evaluation of EPInformer 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 EPInformer 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 EPInformer-PE-Activity-HiC model demonstrates the highest predictive performance (Pearson r = 0.792), outperforming the baseline MENTR (0.675) and other EPInformer variants. Source data are provided as a Source Data file.



Supplementary Figure 9. Performance evaluation of EPInformer variants versus Xpresso across different benchmarks for (a) K562 and (b) GM12878 CAGE prediction. Source data are provided as a Source Data file.