Supplementary Information

EpiGePT: a Pretrained Transformer model for epigenomics

Zijing Gao[1,#], Qiao Liu[2,#,*], Wanwen Zeng[2], Wing Hung Wong[2,3,*] and Rui Jiang[1,*]

[1] Ministry of Education Key Laboratory of Bioinformatics, Research Department of Bioinformatics at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China;

[2] Department of Statistics, Stanford University, Stanford, CA 94305, USA;

[3] Department of Biomedical Data Science, Bio-X Program, Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA;

* To whom correspondence should be addressed.

The first two authors contributed equally.

E-mail: liuqiao@stanford.edu, whwong@stanford.edu, ruijiang@tsinghua.edu.cn

Contents

Supplementary Texts 3

Text S1. Data splitting strategy for model training. 3

Text S2. System design and implementation of the web server. 4

Text S3. Running time of the EpiGePT and baseline methods. 5

Text S4. Implementation of Enformer model and Enformer+. 6

Supplementary Figures 7

Fig. S1 7

Fig. S2 8

Fig. S3 10

Fig. S4 12

Fig. S5 14

Fig. S6 15

Fig. S7 17

Fig. S8 18

Supplementary Tables 19

References 20

Supplementary Texts

Text S1. Data splitting strategy for model training.

To comprehensively validate the performance of EpiGePT in predicting chromatin accessibility, we adopted three different data splitting strategies in the DNase[1] prediction experiment to verify the model's prediction ability when facing new genomic regions and cell types, which can meet researchers' usage needs to the maximum extent. Firstly, cross-cell type prediction refers to splitting the training and testing sets according to cell types in the same genomic region, where the cell types in the testing set have not appeared in the training set (Figs. S1B). Secondly, cross-genomic region prediction refers to splitting the training and testing sets according to genomic regions in the same cell type (Figs. S1A). Thirdly, simultaneous cross-cell type and genomic region prediction, where the prediction can be performed in completely novel cell types and genomic regions with the expression of transcription factors in that cell type. The training set needs to subset both cell types and genomic regions (Figs. S1C). To complete the latter two auxiliary predictions, we also split the data into 5 folds according to both cell types and genomic regions, so that both cross-validation can be performed in one round of training, but this will also reduce the amount of training and testing data.

Text S2. System design and implementation of the web server.

EpiGePT-online runs on a Linux-based Apache web server (https://www.apache.org) and utilizes the Bootstrap v3.3.7 framework (https://getbootstrap.com/docs/3.3/) for its web-frontend display. The backend of the server uses PHP v7.4.5 (http://www.php.net). The platform is compatible with the majority of mainstream web browsers, including Google Chrome, Firefox, Microsoft Edge, and Apple Safari.

Text S3. Running time of the EpiGePT and baseline methods.

To demonstrate the computational efficiency of our model, we recorded the runtime of EpiGePT and baseline methods for one epoch on two sets of experiments, with different data sizes and input sequence lengths. Firstly, in the DNase signal prediction experiment on 129 cell types, with an input sequence length of 10kbp and using the same training data, Enformer requires approximately 3 hours and 4 minutes to complete one epoch, while EpiGePT only takes 2 hours and 17 minutes. In contrast, ChromDragoNN[2], which uses a genomic bin rather than a long region as the model input, requires 24 hours for pre-training and 8 hours for fine-tuning. In this case, the batch size of ChromDragoNN was set to 1024, which is equivalent to EpiGePT using a batch size of around 20. This modeling and computation approach presents challenges in terms of computational efficiency when dealing with large amounts of data. DeepCAGE[3] faces similar efficiency issues using the same approach. Even with a batch size of 256 on a single GPU, it still takes nearly 10 hours to complete one epoch of training. Secondly, we also recorded the running time of the models under larger-scale data and longer input sequences. When the number of input genomic bins increased from 50 to 1000, which corresponds to an input sequence length of approximately 128k, EpiGePT took approximately 3 hours to complete one epoch of training on 20 cell lines and 13,300 genomic regions, while Enformer required approximately 27 hours to train one epoch, as it required a longer input sequence of approximately 190kbp. Furthermore, EpiGePT without TF module (EpiGePT-seq) had approximately 1/4 of the parameters of Enformer and took approximately 2 hours and 40 minutes to train. In terms of performance, EpiGePT-seq performed similarly to Enformer on this dataset. This also explains why we chose to simplify the pure sequence model rather than directly adding a TF module to Enformer.

Text S4. Implementation of Enformer model and Enformer+.

To ensure a fair comparison between models and prevent the possibility of information leakage, we implemented the Enformer[4] model ourselves and trained it on our own collected data. Due to differences in dataset size and partitioning compared to Enformer, we reduced the number of encoder layers in Enformer to prevent overfitting. Thus, we reduced the number of encoder layers in Enformer to 3. Additionally, we introduced Enformer+ to enable a fair comparison between EpiGePT and Enformer in locus-level prediction. As Enformer takes only the DNA sequence as input, it tends to predict the same values for the same locus in different cell types, resulting in a loss of locus-level prediction ability. To address this, we incorporated the binding status and expression of the same transcription factors in Enformer+, and compared it to EpiGePT's performance on the same tasks.

Supplementary Figures

Fig. S1

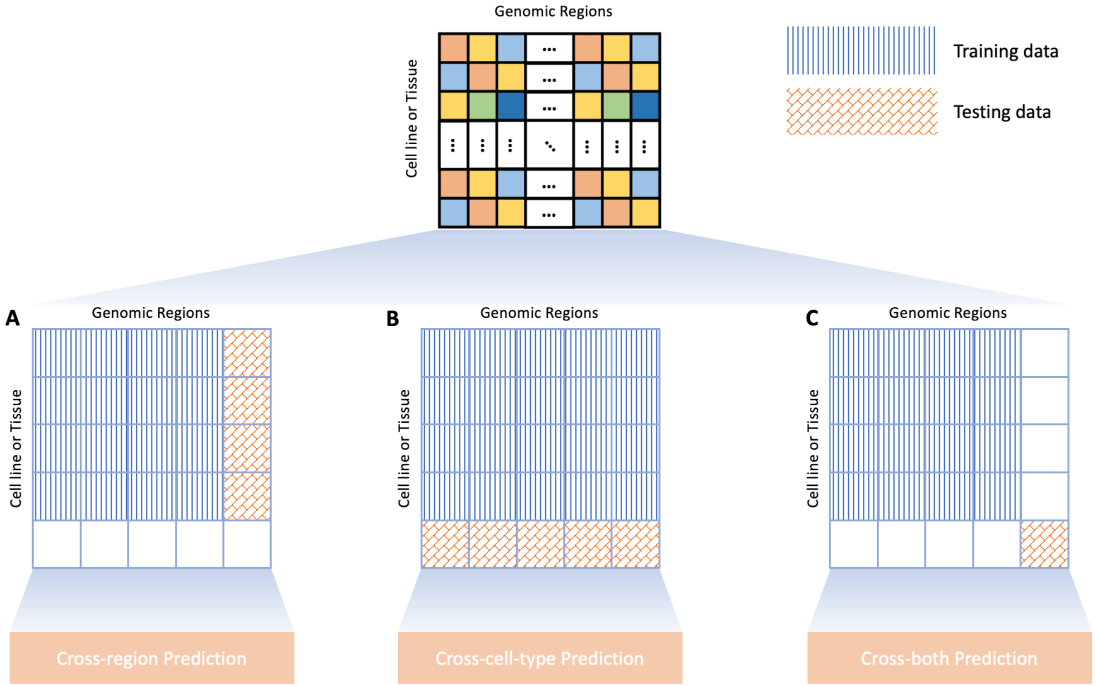


Fig. S1. Three data partitioning strategies for model training and testing. (A) Cross genomic region prediction. The training and testing datasets utilized the expression profiles of identical cell types, but were evaluated on novel genomic regions for prediction. (B) Cross cell type prediction. The training and testing datasets utilized the same genomic regions, but were evaluated on novel cell types for prediction. (C) Cross genomic region and cell type prediction. The cell types and genomic regions used in the training and test sets were both different.

Fig. S2

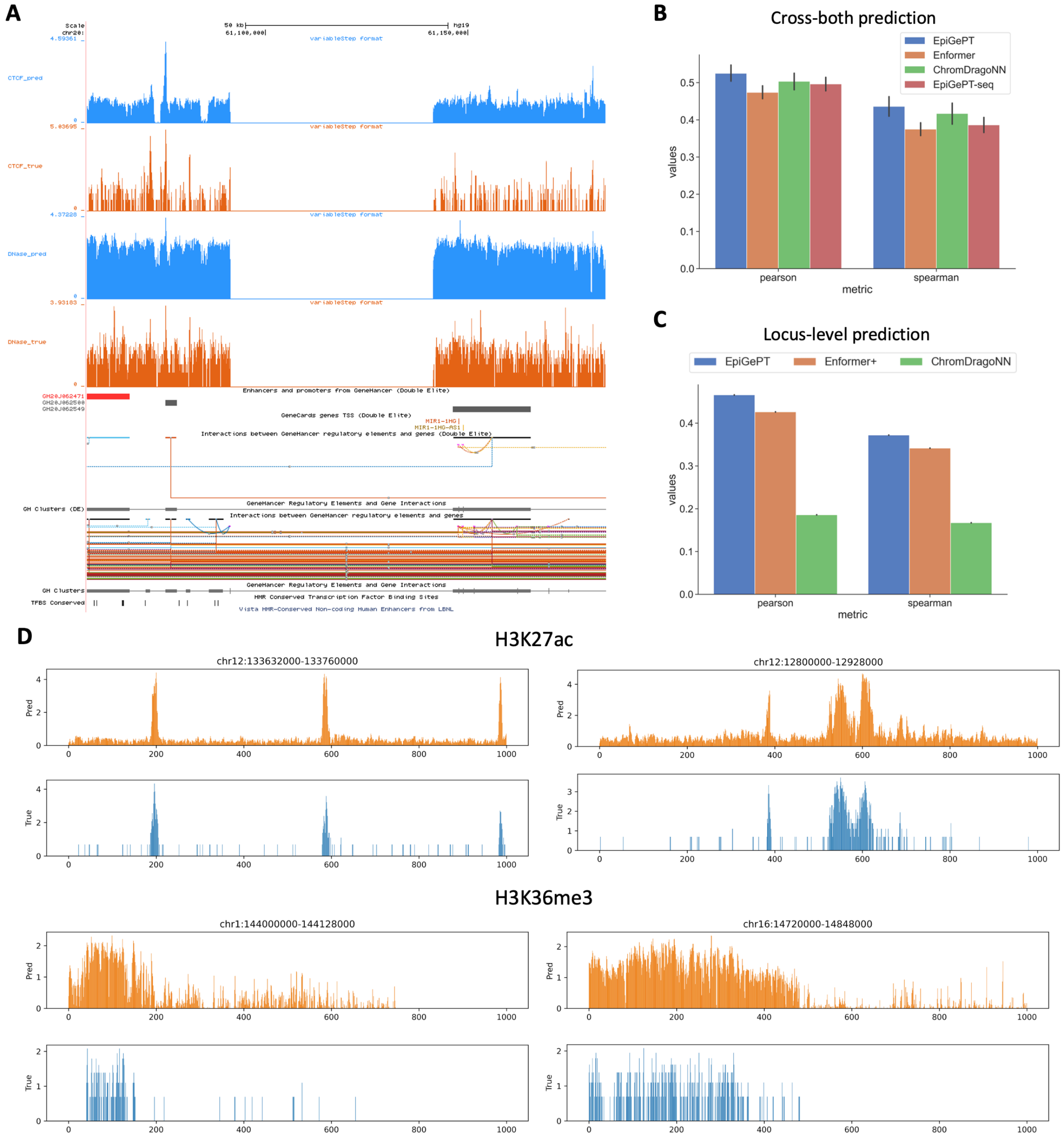


Fig. S2. EpiGePT's performance in predicting DNase-seq and other epigenetic signals is demonstrated in (A) through visualization of predicted results for DNase and CTCF signals. EpiGePT is able to make accurate predictions for these signals, as well as for the regulatory relationships within a genomic region of 20th chromosome ranging from 61,100,000 to 61,150,000. (B) EpiGePT and baseline methods were compared for their performance in predicting epigenetic signals in new cell types and genomic regions (cross-both prediction). The left panel shows the Pearson correlation coefficient, and the right panel shows the

Spearman correlation coefficient. (C) Locus level prediction of DNase signal. We predicted a value for each genomic locus, and calculated the correlation coefficient between the predicted values and true values for the same locus in different cell types. (D) Visualization of predicted signals, such as the comparison between predicted and true values in a 128kbp region (from 133,632,000 to 133,760,000) on chromosome 12, shows that the presence of a large number of zeros in both the true and predicted signals can limit the correlation between the two signals.

Fig. S3

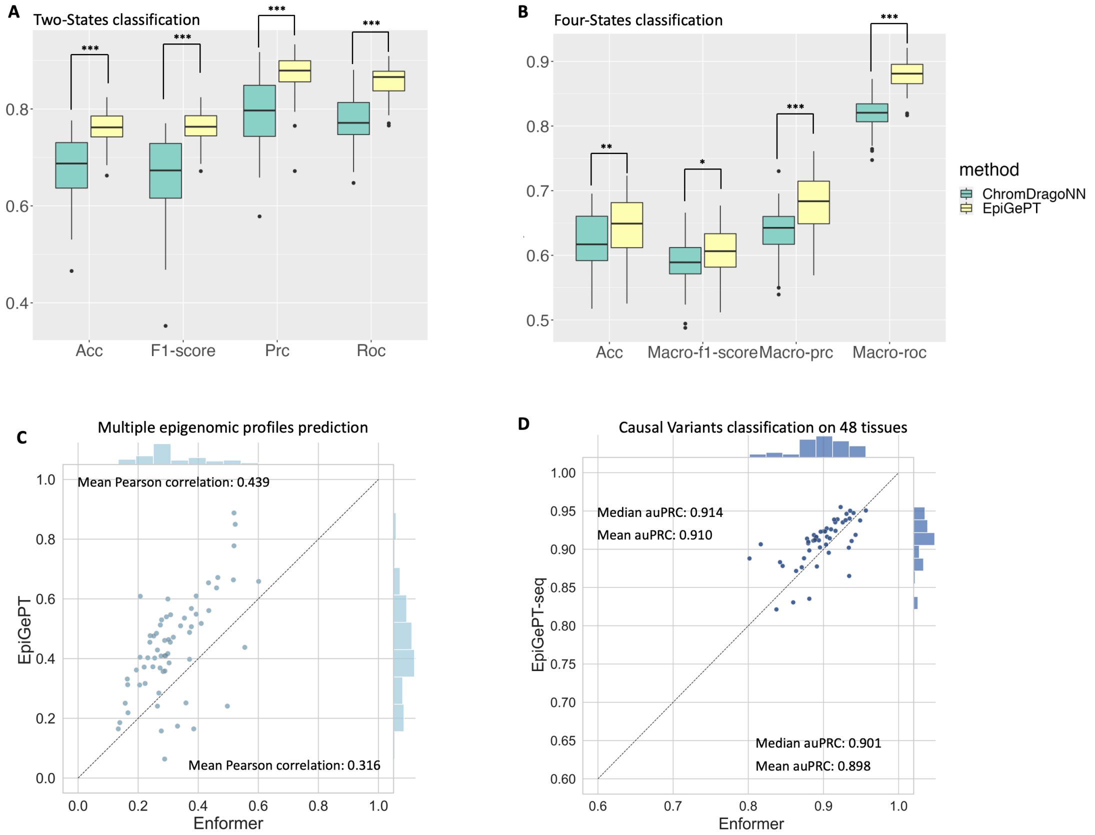


Fig. S3. Performance of EpiGePT and baseline methods on chromatin states classification, multiple epigenomic profiles prediction and causal variants classification. (A) Binary classification of chromatin states for distinguishing functional regions on the chromatin based on the annotation data from ChromHMM-15-states. (B) Four-class chromatin state classification is used to distinguish functional regions on the chromatin, including TSS, potential enhancers, other functional regions, and non-functional regions based on the annotation data from ChromHMM-15-states. * indicates that the *p*-value is less than 0.005 under a one-sided Wilcoxon hypothesis test , ** indicates that the *p*-value is less than 0.005 under a one-sided Wilcoxon hypothesis test, and *** indicates that the *p*-value is less than 1e-3 under this hypothesis test. (C) Cross-cell-type prediction of 8 epigenomic signals at 8 test cell types. Each dot denotes the Pearson correlation coefficient of the predicted signals and true signals at the specific cell types on a specific epigenomic signal. (D) The performance

of EpiGePT and Enformer in discriminating causal eQTLs across 48 tissues, each dot representing the average auPRC obtained from 5-fold cross-validation on a specific tissue.

Fig. S4

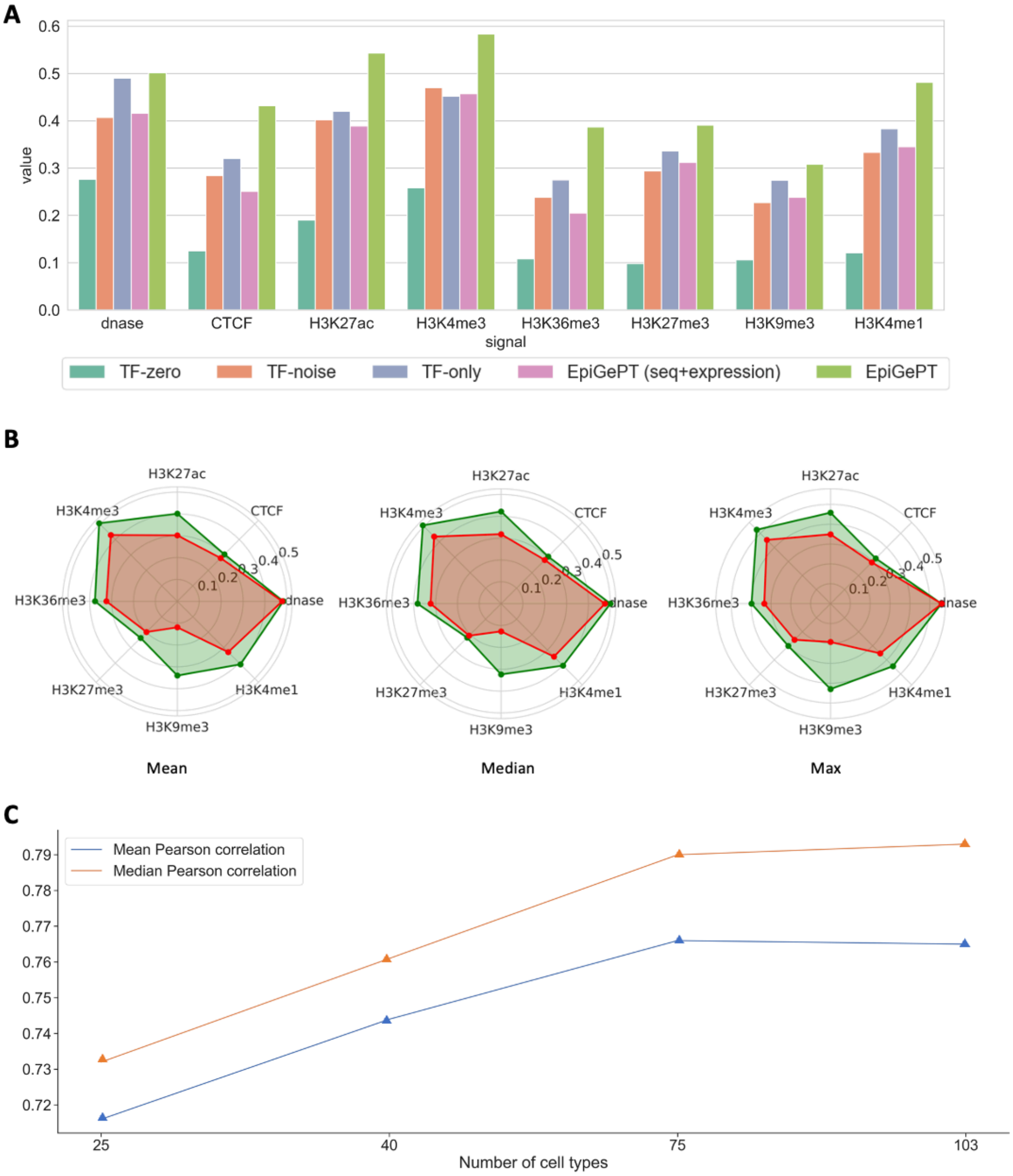
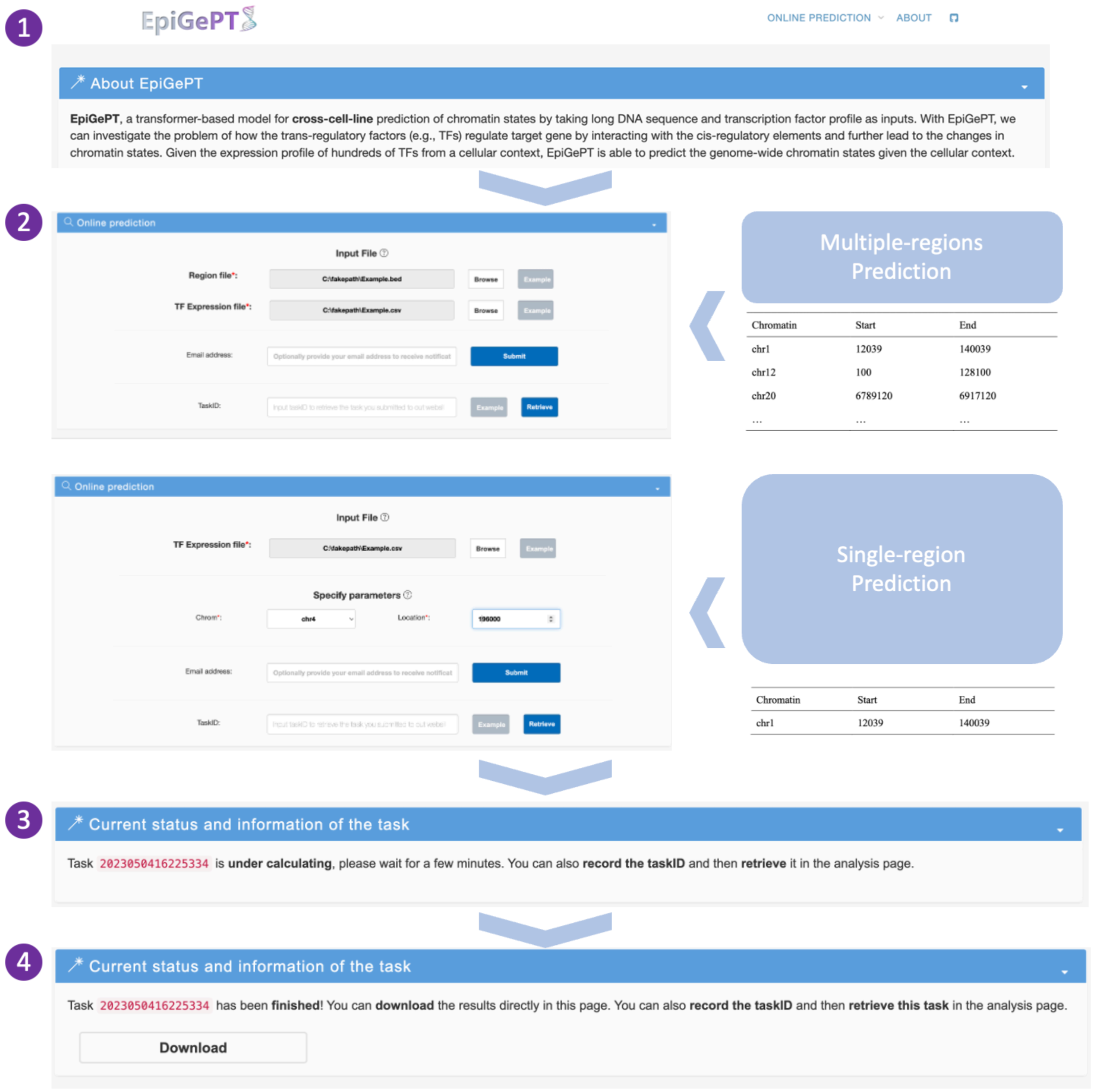


Fig. S4. Ablation analysis of the EpiGePT model. (A) Ablation analysis on the TF module and the Sequence module, we observed a decrease in predictive performance for each module across eight chromatin epigenetic signals, as evidenced by a reduction in Pearson correlation coefficient. (B) Ablation analysis on the Multi-task module. The green shaded area in the figure represents the results of multi-signal cross-cell-type predictions, while the red shaded area represents the results of training and predicting on each signal individually. It can

be observed that the multi-task module has a positive effect on the model performance across all signals. (C) Ablation analysis of the number of the training cell types. When the number of training cell types increases while the number of testing cell types remains constant, there is an increasing trend in performance as the number of training cell types increases.

Fig. S5



About EpiGePT

EpiGePT, a transformer-based model for **cross-cell-line** prediction of chromatin states by taking long DNA sequence and transcription factor profile as inputs. With EpiGePT, we can investigate the problem of how the trans-regulatory factors (e.g., TFs) regulate target gene by interacting with the cis-regulatory elements and further lead to the changes in chromatin states. Given the expression profile of hundreds of TFs from a cellular context, EpiGePT is able to predict the genome-wide chromatin states given the cellular context.

Region file*: C:\fakepath\Example.bed

TF Expression file*: C:\fakepath\Example.csv

Email address: Optionally provide your email address to receive notificat

TaskID: Input taskID to retrieve the task you submitted to out webs

Chromatin	Start	End
chr1	12039	140039
chr12	100	128100
chr20	6789120	6917120
…	…	…

TF Expression file*: C:\fakepath\Example.csv

Chrom*: chr4 Location*: 196000

Email address: Optionally provide your email address to receive notificat

TaskID: Input taskID to retrieve the task you submitted to out webs

Chromatin	Start	End
chr1	12039	140039

Current status and information of the task

Task 2023050416225334 is **under calculating**, please wait for a few minutes. You can also **record the taskID** and then **retrieve** it in the analysis page.

Current status and information of the task

Task 2023050416225334 has been **finished**! You can **download** the results directly in this page. You can also **record the taskID** and then **retrieve this task** in the analysis page.

Fig. S5. Case application of the EpiGePT-online. Users can choose either single locus annotation or multi-region annotation on EpiGePT-online, and each genomic region requires a length of 128kbp. Users need to upload the TPM values of transcription factors expression simultaneously. After annotation, users can enter the result page and download the predicted files. The prediction accuracy is 128bp genomic bin, and users can obtain the predicted signals of these eight epigenetic modifications.

Fig. S6

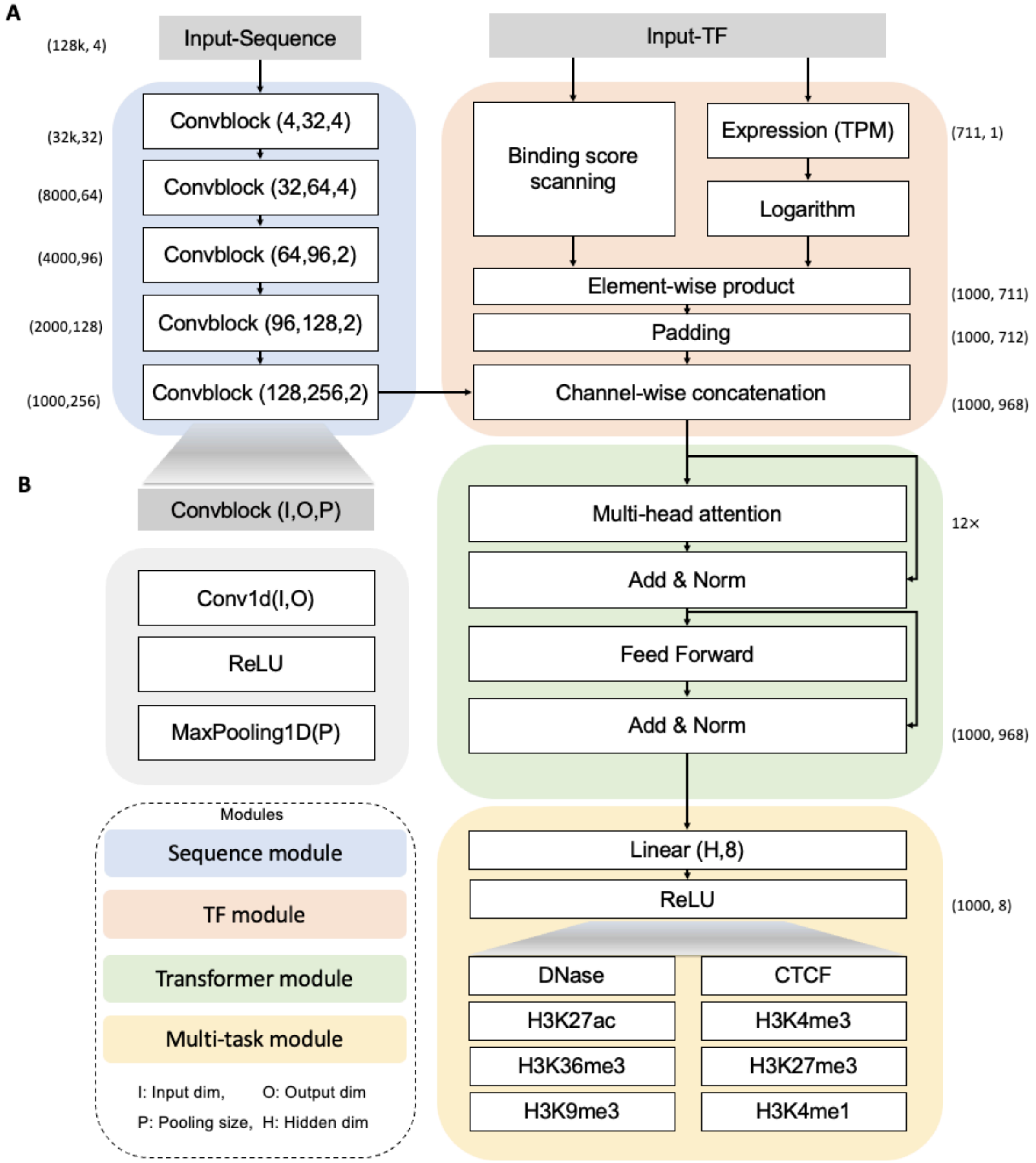


Fig S6. Model architecture of EpiGePT for multiple epigenomic signals prediction. (A) The computational process of EpiGePT. The sequence module employs a stack of five convolutional layers followed by pooling operations, resulting in representations that capture sequence patterns. The TF module integrates motif binding information and gene expression data to represent cell-specific information. The Transformer module takes the genomic bin

sequences mentioned above as input and learns the interaction relationships between bins, capturing the interactions among them. Finally, the obtained embeddings are mapped to the eight types of epigenomic signals through a fully connected layer. (B) Specific details of the convolutional block involve the fusion of 1D convolution, ReLU activation function, and max pooling operation to achieve changes in the feature dimension O and extract bin-level features.

Fig. S7

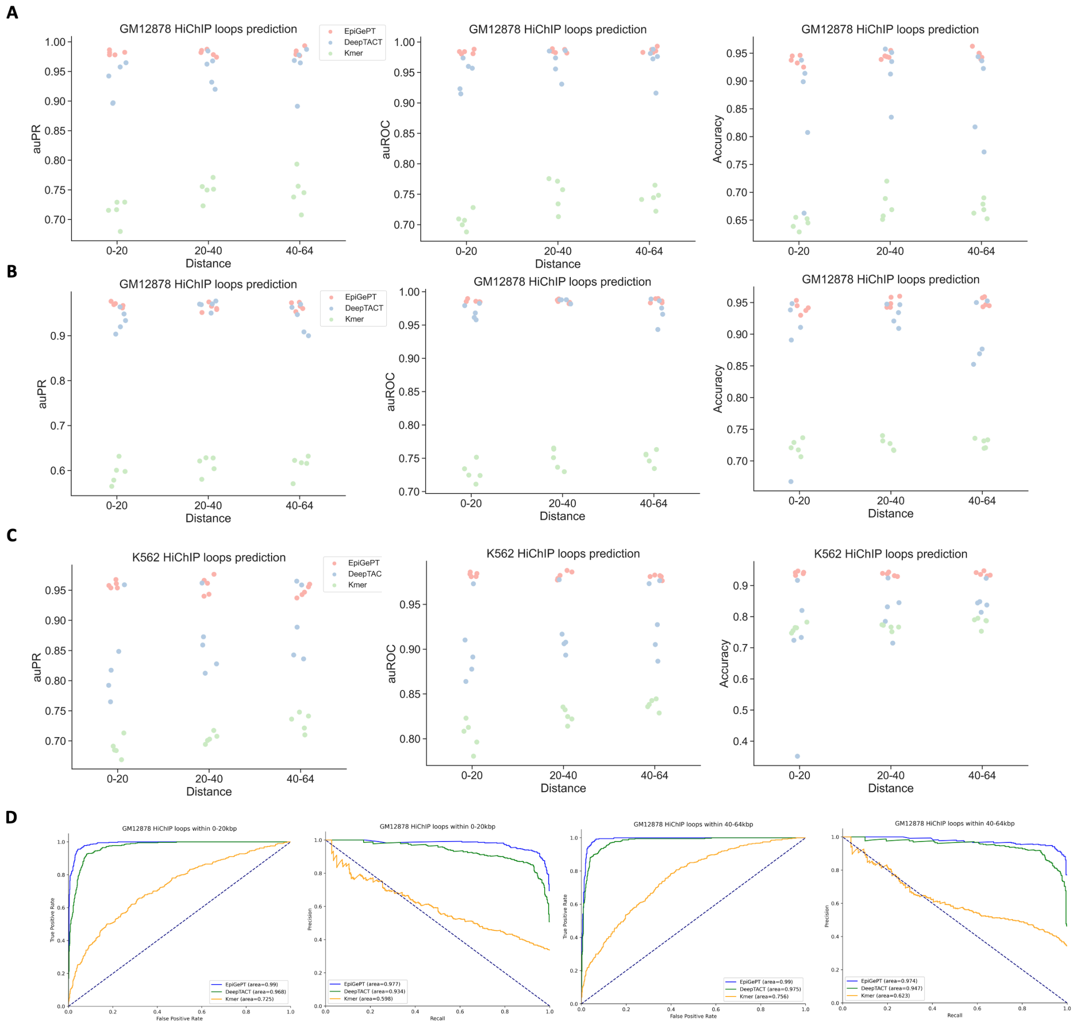


Fig. S7. The fine-tuning performance of the EpiGePT model on predicting potential enhancer-promoter regulatory networks. (A) The performance (measured by auROC and auPRC) of the fine-tuned EpiGePT model and baseline methods (DeepTACT and Kmer) on HiChIP loops data in distinguishing enhancer-gene pairs at various distance ranges (0-20 kbp, 20-40 kbp and 40-64 kbp). (B) The performance of the fine-tuned EpiGePT model and baseline methods on HiChIP loops data in distinguishing enhancer-gene pairs under 1:2 positive-negative sample ratio on GM12878 cell line. (C) The performance of the fine-tuned EpiGePT model and baseline methods on HiChIP loops data in distinguishing enhancer-gene pairs under 1:2 positive-negative sample ratio on K562 cell line. (D) The ROC and PR curves of EpiGePT and baseline methods for predicting HiChIP loops from the GM12878 cell type.

Fig. S8

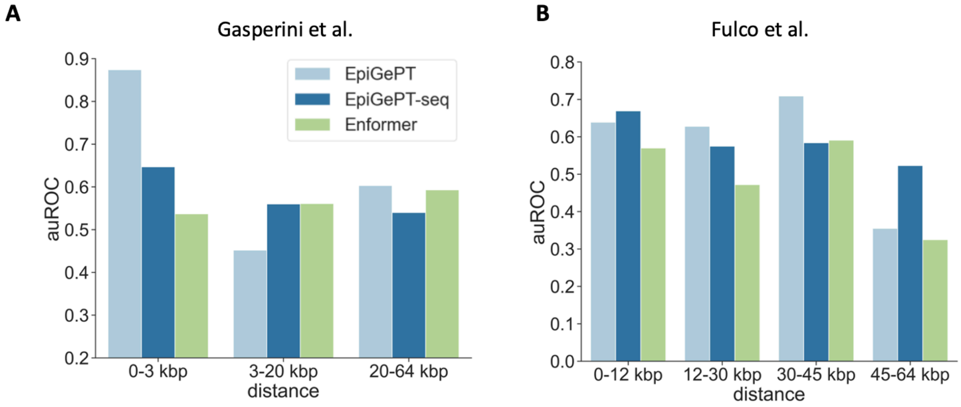


Fig. S8. The performance (auROC) of attention score of EpiGePT in distinguishing regulatory element-gene pairs at different distance ranges. (A) The performance of EpiGePT in distinguishing enhancer-gene pairs at different distance ranges on the data from Gasperini et al[5]. (B) The performance of EpiGePT in distinguishing enhancer-gene pairs at different distance ranges on the data from Fulco et al[6].

Supplementary Tables

Table S1. The information of DNase-seq bam file across 129 biosamples from the ENCODE[7] project.

Table S2. The information of RNA-seq tab-separated values (tsv) file across 129 biosamples from the ENCODE project.

Table S3. The information of DNase-seq, CTCF and other six Histone markers bam file across 28 cell lines or tissues from the ENCODE project.

Table S4. The information of RNA-seq tab-separated values (tsv) file across 28 cell lines or tissues from the ENCODE project.

Table S5. The preprocessed expression data of 711 human transcription factors from the ENCODE project across 129 biosamples.

Table S6. The preprocessed expression data of 711 human transcription factors from the ENCODE project across 28 cell lines or tissues.

Table S7. The order and names of epigenomes of the expression matrices across 56 epigenomes from the ROADMAP[8] project.

Table S8. The preprocessed expression data of 642 human transcription factors across 56 epigenomes from the ROADMAP project.

References

1. Song, L. & Crawford, G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols* **2010**, pdb. prot5384 (2010).
2. Nair, S., Kim, D.S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108-i116 (2019).
3. Liu, Q., Hua, K., Zhang, X., Wong, W.H. & Jiang, R. DeepCAGE: incorporating transcription factors in genome-wide prediction of chromatin accessibility. *Genomics, Proteomics & Bioinformatics* **20**, 496-507 (2022).
4. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196-1203 (2021).
5. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377-390. e319 (2019).
6. Fulco, C.P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature genetics* **51**, 1664-1669 (2019).
7. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
8. Bernstein, B.E. et al. The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045-1048 (2010).