OXFORD

Briefings in Bioinformatics, 2022, 1–11
https://doi.org/10.1093/bib/bbab577
Problem Solving Protocol

Capturing large genomic contexts for accurately predicting enhancer-promoter interactions

Ken Chen, Huiying Zhao and Yuedong Yang

Corresponding author: Yuedong Yang, School of Computer Science and Engineering, Key Laboratory of Machine Intelligence and Advanced Computing (MOE), Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-37106046; Fax: +86 020-37106020; E-mail: yangyd25@mail.sysu.edu.cn

Abstract

Enhancer-promoter interaction (EPI) is a key mechanism underlying gene regulation. EPI prediction has always been a challenging task because enhancers could regulate promoters of distant target genes. Although many machine learning models have been developed, they leverage only the features in enhancers and promoters, or simply add the average genomic signals in the regions between enhancers and promoters, without utilizing detailed features between or outside enhancers and promoters. Due to a lack of large-scale features, existing methods could achieve only moderate performance, especially for predicting EPIs in different cell types. Here, we present a Transformer-based model, TransEPI, for EPI prediction by capturing large genomic contexts. TransEPI was developed based on EPI datasets derived from Hi-C or ChIA-PET data in six cell lines. To avoid over-fitting, we evaluated the TransEPI model by testing it on independent test datasets where the cell line and chromosome are different from the training data. TransEPI not only achieved consistent performance across the cross-validation and test datasets from different cell types but also outperformed the state-of-the-art machine learning and deep learning models. In addition, we found that the improved performance of TransEPI was attributed to the integration of large genomic contexts. Lastly, TransEPI was extended to study the non-coding mutations associated with brain disorders or neural diseases, and we found that TransEPI was also useful for predicting the target genes of non-coding mutations.

Keywords: enhancer-promoter interaction, chromatin structure, Transformer, non-coding mutation

Introduction

Enhancers are functional deoxyribonucleic acid (DNA) fragments acting as *cis*-regulatory elements on the genome, which regulate gene expression through the interactions with the promoters of target genes [1, 2]. There are millions of fragments that have the potential to act as enhancers in the mammalian genome, while their activity varies greatly in different cell types [3]. Therefore, the enhancer-promoter interactions (EPIs) are also cell type-specific and play a critical role in cell development and differentiation [2, 4]. EPIs may be disrupted by genetic variations and lead to the dysfunction of genes underlying the potential pathogenicity of mutations occurring in non-coding regions [5, 6]. Accordingly, linking enhancer mutations to the promoter of target genes could help to interpret a substantial number of non-coding mutations [7–9]. However, it remains challenging to accurately identify EPIs because enhancers and their target promoters are typically separated by thousands of base pairs (bps) [4, 10].

Previous studies have utilized expression quantitative trait loci (eQTL) to infer EPIs [11, 12], while eQTL mapping requires a large number of samples and eQTL-identified EPIs are mostly short-range ones [13, 14]. Over the last decade, chromatin conformation capture-based (3C-based) techniques (e.g. Hi-C [15] and ChIA-PET [16]) have enabled the direct detection of long-range chromatin interactions, which could be applied to identify EPIs [17]. However, high-resolution 3C-based experiments are costly so that experimentally identified EPI data are only available in a few cell types.

To mitigate the high cost of identifying EPIs, a variety of computational methods have been proposed to predict EPIs. Early studies attempted to decipher the determinants of EPIs using the correlations of genomic signals at enhancers and genes (or promoters) over a series of cell types [18–20], while these methods were usually of low performance because enhancers are usually cell-type-specific and EPIs vary across cell types [21]. Subsequently, machine learning (including deep learning models) models for EPI prediction have been proposed. Methods like RIPPLE [22], TargetFinder [23], JEME [24], EPIP [25] and EAGLE [26] employed genomic and epigenomic signals for EPI prediction. Meanwhile, other groups built methods

Ken Chen is a PhD student in the School of Computer Science and Engineering at Sun Yat-Sen University. His research interests lie in deep learning, gene regulation and genetic variation analysis.
Huiying Zhao is an associate research fellow in the Sun Yat-sen Memorial Hospital at Sun Yat-sen University. Her research interests include pathogenic gene analysis, protein function and RNA function prediction.
Yuedong Yang is a professor in the School of Computer Science and Engineering and the National Super Computer Center at Guangzhou, Sun Yat-sen University, China. His research group emphasizes on developing HPC and AI algorithms for protein function prediction, multi-omics data integration and intelligent drug design. He is also responsible for constructing the HPC platform for biomedical applications based on the Tianhe-2 supercomputer.
Received: September 15, 2021. **Revised:** December 13, 2021. **Accepted:** December 15, 2021
© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

that use DNA sequences for EPI prediction, including SPEID [27], Zhuang's method [28], EPIVAN [29] and EPI-DLMH [30]. However, these methods were mainly evaluated on randomly split samples and thus may suffer from inflated performance evaluation, as reported by several studies [21, 31–33].

Recently, several methods that do not suffer from the above issue have been proposed. 3DPredictor [34] employs gene expression and CTCF-binding sites within and around pairs of genomic loci to model Hi-C contact maps for inferring EPIs. DeepC [35] and Akita [36] present that deep convolutional neural networks (CNNs) could efficiently predict Hi-C contact maps using megabase-scale genome sequences. Another CNN-based model, ChINN [37], takes a further step to predict chromatin interactions at a genome-wide scale. Though progress has been made, there is still much room for improvement. First, DeepC and Akita benefit from the features extracted from megabase-scale DNA sequences using CNN. However, CNN is less efficient than recurrent neural networks or the Transformer architecture [38] in capturing long-range dependencies [39]. Given the success of Transformer in modeling protein structure [40, 41] and predicting gene expression [42], it is promising to apply Transformer to build EPI models. Second, in contrast to DeepC and Akita, 3DPredictor compiles the features within and around genomic loci in a relatively coarse manner, and the ChINN model uses only the DNA sequences from pairs of chromatin interaction anchors (or open chromatin regions). They both lack fine representations of the features from large-scale genomic contexts. Since no direct comparison has been made between these methods, it remains unclear whether incorporating large genomic contexts could boost EPI prediction.

In this study, we present a novel deep learning model, entitled TransEPI, for EPI prediction using the Transformer architecture. TransEPI directly takes the input of genomic signals from large genomic contexts harboring the enhancer-promoter (E-P) pairs and employs Transformer encoders to capture the long-range dependencies. To avoid over-fitting, we trained the model under the cross-validation (CV) scheme, where the training data were rigorously split by chromosomes [32, 33]. TransEPI achieved a consistent performance across different cell types and outperformed the state-of-the-art methods on independent test datasets. Moreover, we confirmed that the integration of large genomic context features was critical for TransEPI to make predictions. Lastly, we extended TransEPI to predict the target genes of mutations associated with brain disorders or neural diseases and found it could find a variety of distant target genes which enrich neural function pathways. It implied the potential ability of TransEPI for studying the pathogenicity of non-coding mutations.

The codes and datasets are available at https://github.com/biomed-AI/TransEPI.

Table 1. Summary of the dataset

Cell line	Source	Positive sample	Negative sample
GM12878	Hi-C	2695	46 212
GM12878	CTCT ChIA-PET	4817	36 028
GM12878	RNAPII ChIA-PET	24 985	70 670
HeLa-S3	Hi-C	2256	21 086
HeLa-S3	CTCF ChIA-PET	1346	10 789
HeLa-S3	RNAPII ChIA-PET	744	2182
HMEC	Hi-C	2286	20 019
IMR90	Hi-C	1468	13 268
K562	Hi-C	2765	73 299
NHEK	Hi-C	1820	13 582

Methods

Datasets

The BENGI dataset

We employed the 'Benchmark of candidate Enhancer-Gene Interactions (BENGI)' dataset [21] to develop TransEPI for EPI prediction. BENGI is a collection of enhancer-gene interactions from various biosamples, which were identified by 3C-based or genetic (e.g. eQTL and CRISPR/dCAS9 perturbations) approaches. The positive samples were defined as the enhancer-gene pairs identified by 3C-based or genetic experiments. The negative samples were generated by pairing enhancers with non-interacting genes within the 95th percentile of the enhancer-gene distances of positive samples [21]. Since we aimed to predict the physical interactions between enhancers and promoters, only the samples identified by Hi-C and ChIA-PET were utilized.

We first mapped the genes in the BENGI dataset to transcripts based on GENCODE annotation (v19) [43] to convert the enhancer-gene pairs in BENGI to E-P pairs. Here, the promoter was defined as the 1500 bp upstream and the 500 bp downstream of a transcript start site (TSS). Because the TSSs of some genes differ by more than thousands of bps, the promoters in some E-P pairs derived from positive samples may reside outside the chromatin interaction anchors. These E-P pairs were then excluded from our datasets. Besides, we removed the samples (including positive and negative samples) with low-expressed transcripts [transcript per million < 1] as they were less likely to be regulated by enhancers. In this way, we obtained 45 182 positive and 307 135 negative samples in six cell lines, as listed in Table 1.

We combined the samples from GM12878 and HeLa-S3 to construct a dataset for developing the model, namely BENGI-train, which contains 36 843 positive and 186 967 negative samples. The other datasets from four cell lines, namely BENGI-HMEC, BENGI-IMR90, BENGI-K562 and BENGI-NHEK, were reserved for independent tests.

Hi-C data from GSE63525

We collected the chromatin interactions identified by Hi-C in a mouse and seven human cell lines from Gene Expression Omnibus under the accession ID

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

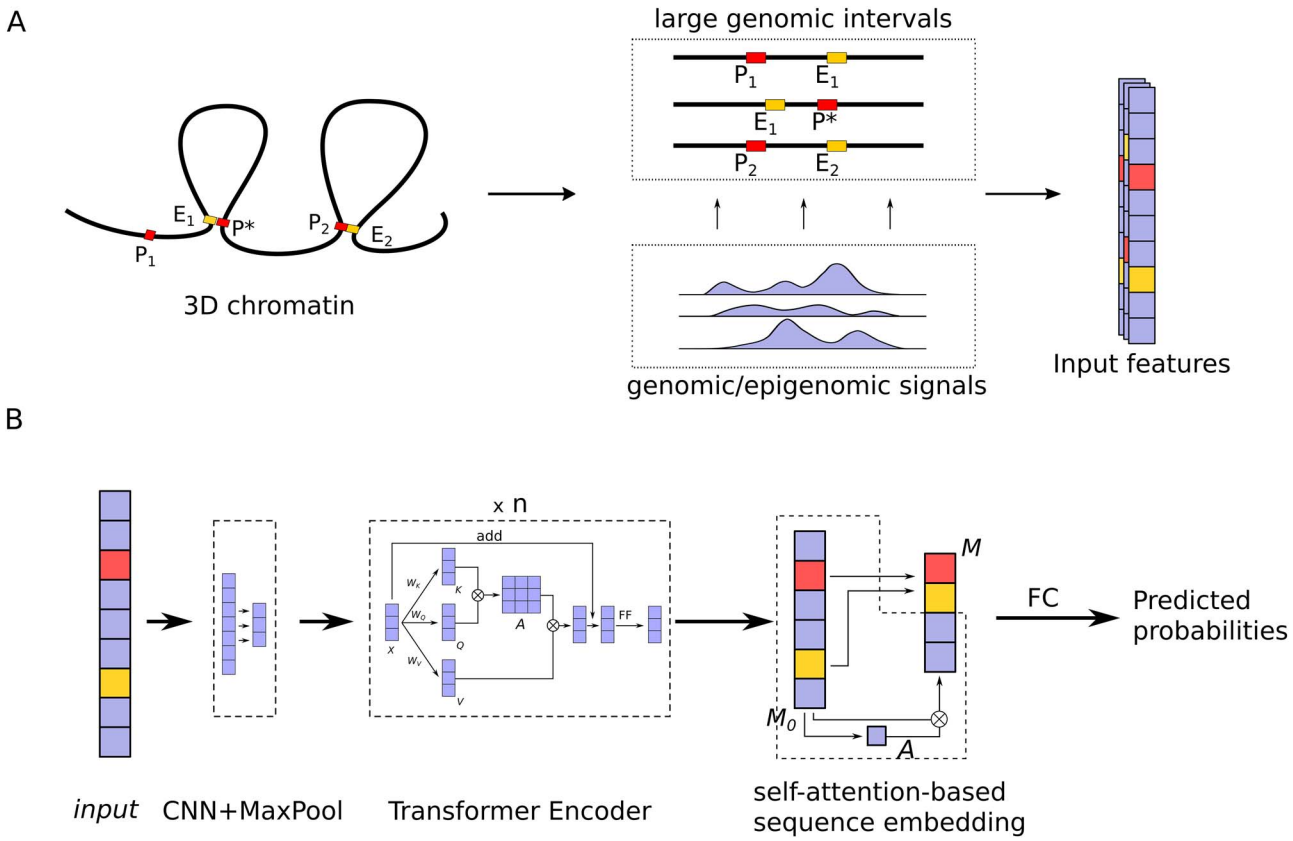

Figure 1. The TransEPI framework. (**A**) Feature preparation. Genomic features (CTCF, DNase I, H3K27me3, H3K4me1, H3K4me3, H3K36me3 and H3K9me3) are extracted from large intervals harboring the candidate E-P pairs (enhancers are in yellow and promoters are in red). (**B**) The architecture of the TransEPI. (⊗, matrix multiplying operation; MaxPool, max-pooling).

GSE63525 [17]. They were used to evaluate TransEPI in a mouse cell line and extend the original model for predicting non-coding variants in brain tissues. We provided the details about the data in the Supplementary Notes and Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

Input features

For each candidate E-P pair, TransEPI extracts genomic features from a large genomic region of 2.5 Mbp that centers on the E-P pair (Figure 1**A**). The reason for using the size of 2.5 Mbp is that the maximum E-P distance in BENGI is about 2 Mbp and we found that using a size >2.5 Mbp could not improve the performance (see Supplementary Notes for details). To represent the chromatin states of the large region, a total of seven types of genomic features were recruited, including CTCF-binding sites, chromatin accessibility (DNase-I signals) and five histone marks (H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3; the 'core marks' in the Roadmap project [44]). TransEPI compiles these genomic data for each 2.5 Mbp region as a multi-dimensional (multi-channel) array, where each channel corresponds to a genomic feature type. In this way, the chromatin states of the entire 2.5 Mbp region could be fed into TransEPI to make predictions. In practice, it is infeasible for TransEPI to directly process signal arrays of 2.5 Mbp due to the limitation of Graphics Processing Unit's memory. Meanwhile, the resolution of the genomic data is typically 10s–1000s of bps. Thus, there is no need to compile the features at bp resolution. Inspired by Belokopytova *et al.* [34], we partitioned each 2.5 Mbp region into 5000 consecutive bins with the bin size of 500 bp and averaged the genomic signals in each bin for each feature, respectively. In addition, we need to mark the location of the enhancer and promoter in each region. To this end, we added an additional channel to each input array, where the value in each bin is the shortest distance from it to the enhancer or the promoter. The technical details of feature preparation are provided in Supplementary Notes.

The TransEPI model

The architecture of the TransEPI model is illustrated in Figure 1**B**.

First, TransEPI utilizes a one-dimensional CNN to extract features from the input signals. A max-pooling layer is then used to down-sample the features and shrink the length of each input sequence.

Next, TransEPI adopts a stack of multiple Transformer encoders to capture the long-range dependencies. In each Transformer encoder layer, the input sequence $\boldsymbol{X} \in R^{l \times h}$ is first transformed into a key, a query and a

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

value sequence, respectively:

$$\boldsymbol{K} = \boldsymbol{X}\mathbf{W_k}, \mathbf{Q} = \boldsymbol{X}\mathbf{W_q}, \mathbf{V} = \boldsymbol{X}\mathbf{W_v},$$

where $\mathbf{W_k} \in R^{h\times d_k}, \mathbf{W_q} \in R^{h\times d_k}$ and $\mathbf{W_v} \in R^{h\times h}$ are learnable weights. $\boldsymbol{K}$ and $\mathbf{Q}$ are then used to obtain the attention matrix $\mathbf{A}$:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right),$$

where the coefficient a_{ij} in $\mathbf{A}$ could be understood as the association between the ith and the jth positions in $\boldsymbol{X}$. By multiplying $\boldsymbol{V}$ by $\mathbf{A}$, the hidden states at different positions are updated. We denote the output from the Transformer module by $\mathbf{M}_0 \in R^{l\times h}$.

Subsequently, a self-attention sequence embedding method [45] is used to learn a low-dimensional representation for $\mathbf{M}_0$. We first pass $\mathbf{M}_0$ to a two-layer fully connected (FC) network:

$$\mathbf{A}_s = \text{softmax}\left(\mathbf{W}_{s2}\tanh\left(\mathbf{W}_{s1}\mathbf{M}_0^T\right)\right),$$

where $\mathbf{W}_{s1} \in R^{s\times h}, \mathbf{W}_{s2} \in R^{r\times s}, \mathbf{A}_s \in R^{r\times l}$. Each row in $\mathbf{A}_s$ sums up to 1, representing a group of attention coefficients for the different positions in $\mathbf{M}_0$. We multiply $\mathbf{M}_0$ by $\mathbf{A}_s$ and acquire a weighted embedding $\mathbf{M}_1 \in R^{r\times h}$ ($\mathbf{M}_1 = \mathbf{A}_s\mathbf{M}_0$). Then, the average and the maximum values along the second dimension of $\mathbf{M}_1$ are concatenated as a low-dimensional embedding of $\mathbf{M}_1$. Additionally, we concatenate the low-dimensional embedding with the hidden states at the enhancer ($\boldsymbol{h}_e$) and the promoter ($\mathbf{h}_p$) bin to integrate both the global information of the whole sequence and the local information of the enhancer and the promoter:

$$\mathbf{M} = \text{AvgPool}(\mathbf{M}_1) \parallel \text{MaxPool}(\mathbf{M}_1) \parallel \boldsymbol{h}_e \parallel \boldsymbol{h}_p,$$

where $\mathbf{M} \in R^{4h}$.

Finally, the prediction could be made using a multilayer perception (MLP):

$$p = \sigma(\text{MLP}_1(\mathbf{M})),$$

where σ represents the sigmoid function; p ranges from 0 to 1 ($p \in (0, 1)$), representing the probability that the input enhancer and the promoter interact with each other. In parallel, to make the model sensitive to the locations of the enhancer and the promoter, we use a second MLP module to predict the distance from enhancer to promoter:

$$d_{\text{pred}} = \text{MLP}_2(\mathbf{M}).$$

Evaluation metrics

We evaluated the model with the area under the precision-recall (PR) curve (auPRC) and the area under the receiver operating characteristic (ROC) curve [46] (AUC). The PR curve is a plot of precision against recall at a series of thresholds. Similarly, the ROC curve is a plot of true-positive rate (TPR) against false-positive curve (FPR). Precision, recall, TPR and FPR are defined as follows:

$$precision = \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN},$$
$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN},$$

where TP, FP, TN and FN are short for the true positives, false positives, true negatives and false negatives.

Since auPRC is associated with the proportion of positive samples in the dataset, we used auPRC-ratio (dividing auPRC by that of a random predictor, which equals the proportion of positive samples in the dataset [47]) as a metric for comparing the performance of TransEPI across different datasets.

Model training and evaluation

The TransEPI model is implemented with PyTorch (version 1.9.0) [48] in Python 3.8. It was trained to minimize a combined loss for EPI prediction (binary cross entropy loss) and E-P distance (mean squared error loss) prediction:

$$\begin{aligned}\mathcal{L}(\mathbf{y}, \boldsymbol{p}, \boldsymbol{d}_{pred}, \boldsymbol{d}_{true}) &= \mathcal{L}_{EPI}(\mathbf{y}, \boldsymbol{p}) + \mathcal{L}_{EP-distance}(\boldsymbol{d}_{pred}, \boldsymbol{d}_{true}) \\ &= -\frac{1}{N}\sum_i^N \left[\mathbf{y}_i \log(\boldsymbol{p}_i) + (1-\mathbf{y}_i)\log(1-\boldsymbol{p}_i)\right] \\ &\quad + \frac{1}{N}\sum_i^N (\boldsymbol{d}_{pred,i} - \boldsymbol{d}_{true,i})^2,\end{aligned}$$

where p_i, y_i, $d_{\text{pred},i}$ and $d_{\text{true},i}$ are the predicted EPI probability, true EPI label (0 or 1), the predicted E-P distance and the real EP distance of the ith sample, respectively. The Adam optimizer [49] is recruited to update the learnable weights in the neural network (e.g. the parameters like W_k, W_q and W_v in the model) using a learning rate of 0.0001.

To avoid over-fitting, we adopted a CV scheme to fine-tune the hyper-parameters in TransEPI. We divided the samples in BENGI-train into 5-fold by chromosomes (chromosome-split CV), ensuring that the samples from the same chromosome would also be put into the same fold (chromosomes assigned to each fold are listed in Supplementary Table S2 available online at http://bib.oxfordjournals.org/). In each training epoch, we trained the model on 4-fold and validated it on the remaining fold in turn. The average AUC and auPRC over the 5-fold were used to evaluate the performance. For independent tests, we also split the test data by chromosomes to evaluate the models on data from different cell types and chromosomes. The pipeline of model training and evaluation is depicted in Supplementary Figure S1 available online at http://bib.oxfordjournals.org/.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

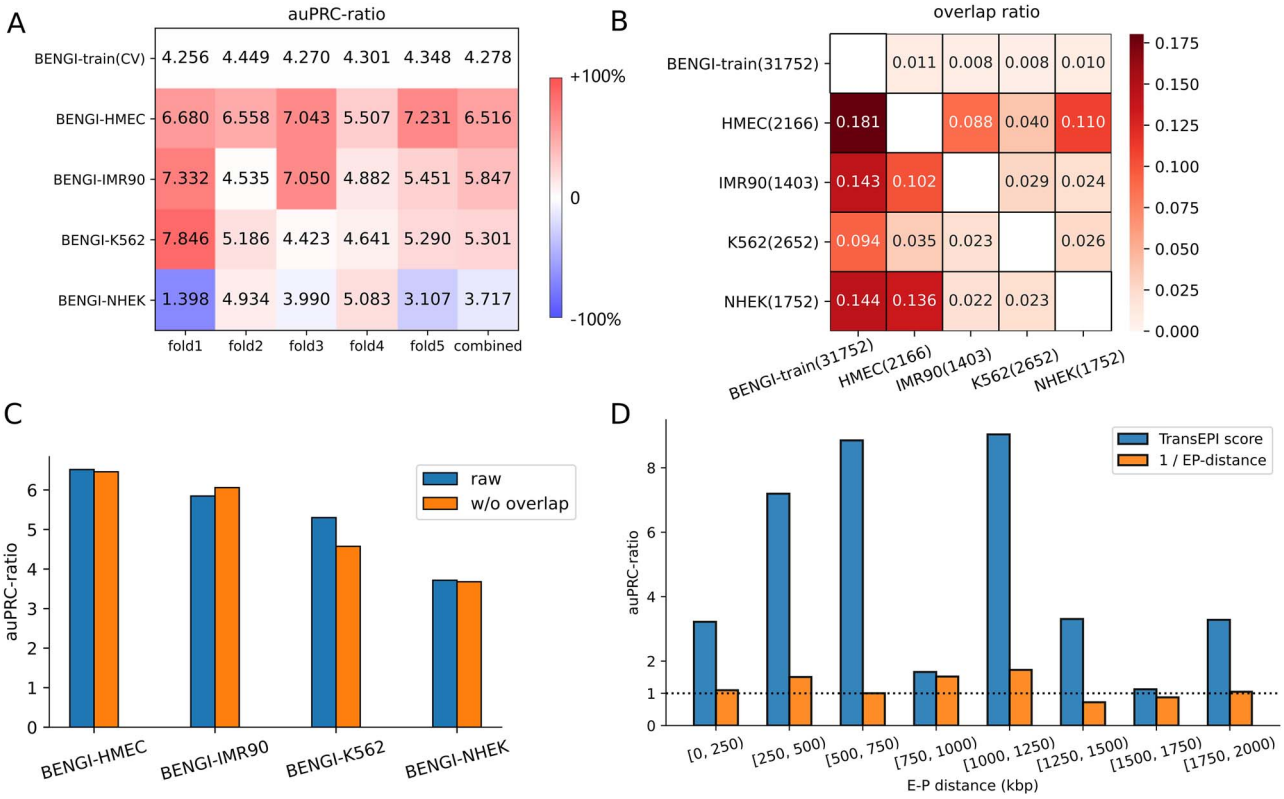

Figure 2. Evaluating TransEPI on independent test datasets. (**A**) The auPRC-ratio scores of TransEPI on BENGI-train (CV) and four independent tests. The results on each fold and the combined dataset are shown. (**B**) The ratio of the EPIs shared by different cell types. The ratios are the numbers of overlapping EPIs divided by the size of the datasets in the rows (bottom left) or the columns (top right). (**C**) The auPRC-ratio of TransEPI on the raw test datasets or the test datasets with overlapping EPIs excluded. (**D**) The auPRC-ratio scores achieved by using TransEPI score or 1/EP-distance on the samples stratified by EP-distance.

Extending TransEPI to predict target genes for non-coding mutations

We further applied TransEPI to predict the target genes of non-coding mutations. We collected 3943 non-coding mutations associated with neural diseases or brain disorders from Lu *et al.*'s work [14] (Supplementary Table S6 available online at http://bib.oxfordjournals.org/). However, only 198 out of the 3943 mutations were found in the enhancers of human brain tissue (the annotation was taken from the Enhancer Atlas 2.0 database [50]). To make TransEPI compatible with the mutations outside enhancers, we extended TransEPI by training it on the Hi-C loop dataset (details in Supplementary Notes). We reserved the samples in HMEC and HUVEC for validation and test, respectively, and trained the model using the other samples.

For each mutation, we paired it with the transcripts within its 1 000 000 bp up- and downstream to curate a list of candidate mutation-transcript pairs. We applied TransEPI on the pairs in two human brain tissues and three neural cell lines (Supplementary Table S3 available online at http://bib.oxfordjournals.org/), respectively. The mutation-transcript pairs with TransEPI-score above a certain threshold were kept as interacting pairs. Here, we set the threshold to 0.33 because the FPR on the test dataset at this threshold is <0.05.

In addition, we use 'top-1 score' to denote the maximum TransEPI-predicted probability among all the candidate pairs for each mutation, which could reflect how likely a mutation interacts with at least one target gene.

Results

TransEPI is capable of predicting EPIs in different cell types

We first investigated whether TransEPI could predict the EPIs in different cell lines in the BENGI datasets. The TransEPI model was developed on BENGI-train (GM12878 + HeLa-S3) and we thus evaluated it on independent test datasets from four different cell lines (HMEC, IMR90, K562 and NHEK).

We first compared the performance of TransEPI on the CV and independent test datasets to investigate whether it could achieve a consistent performance across different cell types. As shown in Figure 2**A**, the auPRC-ratios achieved by TransEPI on four independent test datasets are 6.516, 5.847, 5.301 and 3.717, respectively, three out of which are higher than that on the CV dataset (auPRC-ratio = 4.278). Similar trends can also be observed on almost all 5-fold. Since the overlap between different datasets is <20% (Figure 2**B**), the consistent performance across cell types can reflect TransEPI's ability to predict cell type-specific EPIs. In addition, we filtered the test

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

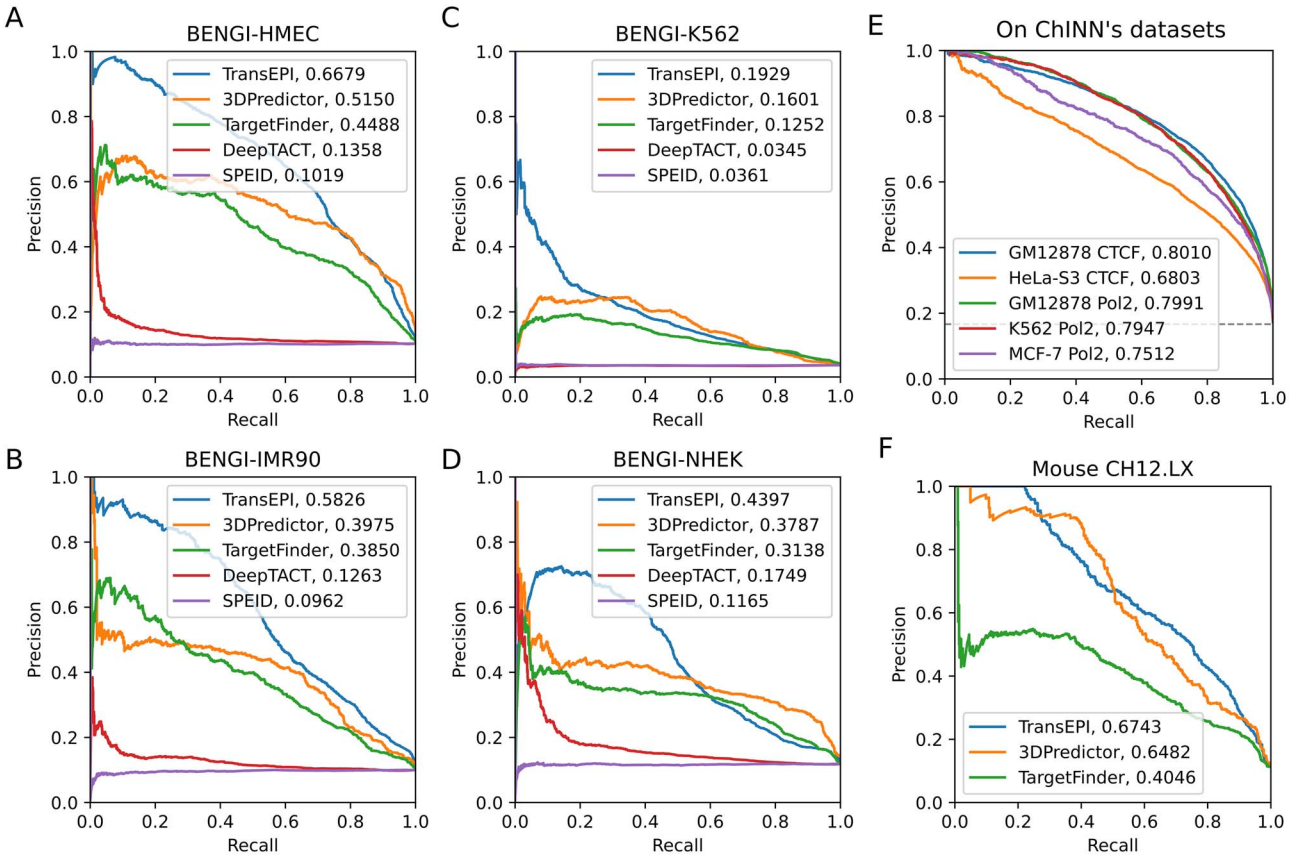

Figure 3. The PRC of TransEPI and the other models on independent test datasets. (**A**) On BENGI-HMEC. (**B**) BENGI-IMR90. (**C**) BENGI-K562. (**D**) BENGI-NHEK. (**E**) The PRC of TransEPI on ChINN's distance-matched datasets. (**F**) On an EPI dataset in mouse CH12.LX cell line.

datasets by excluding the samples that were included in BENGI-train. As shown in Figure 2**C**, excluding the overlapping samples has almost no impact on the performance of TransEPI.

Previous studies have found that the distance between enhancer and promoter (EP-distance) may have strong predictive power in some datasets [21, 26] (Supplementary Figure S2 available online at http://bib.oxfordjournals.org/). However, the predictive power of the EP-distance is determined by the distribution of EP-distance in the positive and the negative samples, which could be manually controlled when constructing the datasets. Hence, we further evaluated TransEPI on test samples stratified by EP-distance to eliminate the potential bias caused by EP-distance distribution. Specifically, we merged the four independent test datasets and used a bin size of 250 000 bp to stratify the samples into eight groups (Figure 2**D**). In each group, the EP-distance is not informative for predicting EPIs, as the auPRC-ratio ranges from 0.7222 to 1.727 (a random model is expected to achieve an auPRC-ratio of 1). In contrast, TransEPI achieves much higher auPRC-ratios, which range from 1.125 to 9.036 (Figure 2**D**), demonstrating that TransEPI captures the underlying determinants of EPI in addition to EP distance.

Taken together, we could conclude that TransEPI is capable of predicting cell type-specific EPIs.

TransEPI outperforms the other models on independent test data

To further evaluate TransEPI, we compared it with TargetFinder [23], SPEID [27], DeepTACT [51], 3DPredictor [34] and ChINN [37] (Supplementary Table S4 available online at http://bib.oxfordjournals.org/). For a fair comparison, we trained the methods (apart from ChINN) on BENGI-train through the same chromosome-split 5-fold CV scheme as our model. The comparison with ChINN was made on ChINN's dataset. There are other models like DeepC [35] and Akita [36] for chromatin interaction prediction, while they only predict the contacts within 1 Mb so that they are not suitable for comparison in our study.

As shown in Figure 3**A**–**D**, TransEPI outperforms SPEID, DeepTACT, TargetFinder and 3DPredictor on all the four test datasets. The auPRC of TransEPI increases by an average of 28.1% compared to the state-of-the-art method 3DPredictor. Although the AUC of TransEPI is lower than that of 3DPredictor on BENGI-NHEK (Supplementary Figure S5 available online at http://bib.oxfordjournals.org/), TransEPI achieves a higher TPR than 3DPredictor when the FPR is close to 0. This means that TransEPI can detect more interacting E-P pairs at a low FPR, which is helpful in practical use. Another state-of-the-art method, ChINN, requires to use DNA sequences of the entire chromatin anchors as input,

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

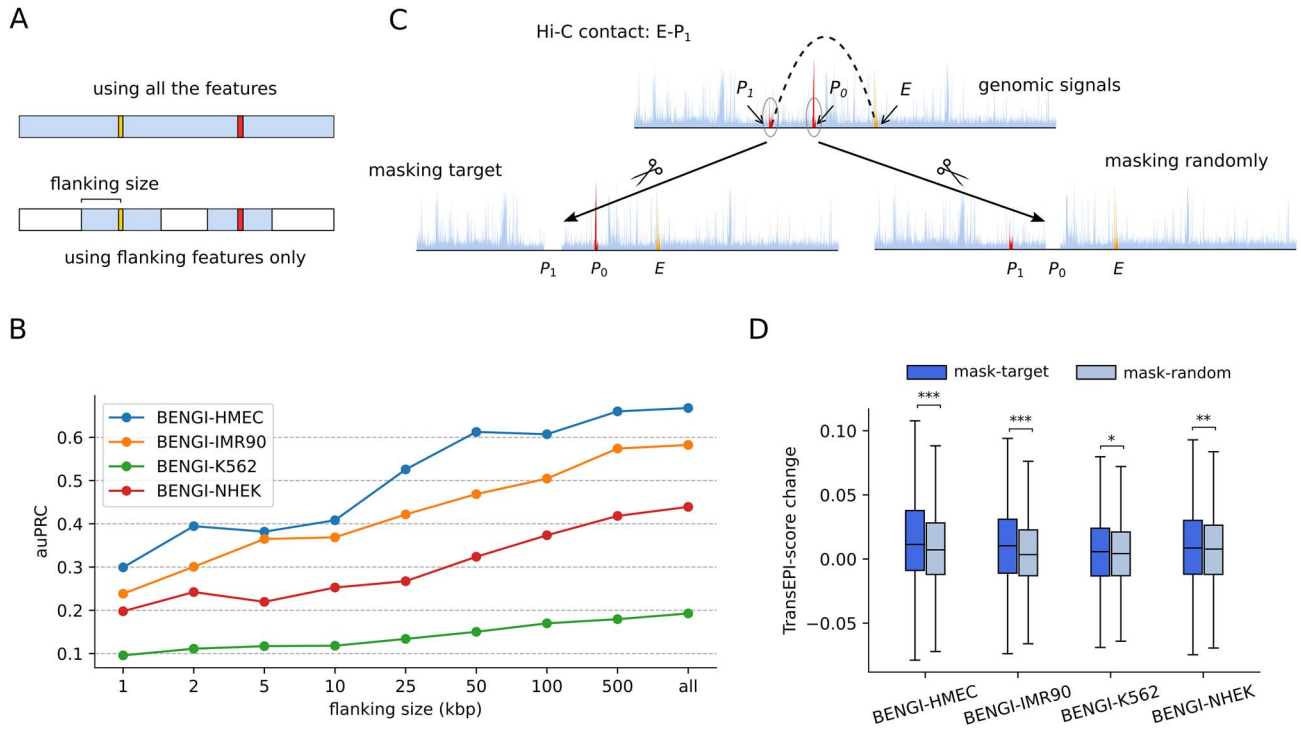

Figure 4. Analyzing the contribution of large-scale features. (**A**) Using all the features or using features in the enhancer (in yellow), the promoter (red) and the flanking regions (light blue). (**B**) The auPRC of TransEPI grows as the flanking size increases. (**C**) Masking the target (P_1) or a randomly selected promoter (P_0) for the enhancer E. (**D**) The distribution of the change of TransEPI-scores in the mask-target group and mask-random group in four BENGI test datasets. (***: *P*-value < 1E-6, **: *P*-value < 1E-3, * *P*-value < 0.05, by one-sided *t*-test).

which is quite distinct from the samples in BENGI. Hence, we evaluated TransEPI on the datasets constructed by ChINN. As shown in Figure 3**E**, TransEPI achieves the auPRC of 0.8010, 0.6803, 0.7991, 0.7947 and 0.7512 on the five ChINN datasets, respectively, outperforming ChINN by an average of 28.7% (details in Supplementary Notes and Supplementary Table S5 available online at http://bib.oxfordjournals.org/) [37]. Compared with the genomic feature-based model presented in ChINN's study, TransEPI still achieves a higher auPRC by an average of 5.87%.

In addition, we applied the TransEPI model trained on human cell lines to an EPI dataset in a mouse cell line (CH12.LX). TransEPI achieved the auPRC and AUC of 0.6743 and 0.9040 (Figure 3**F** and Supplementary Figure S6 available online at http://bib.oxfordjournals.org/), respectively, outperforming 3DPredictor and TargetFinder. This implies that TransEPI has the potential to predict EPIs in different species.

TransEPI benefits from the features outside enhancers and promoters

The major difference between TransEPI and most of the previous models is that it incorporates the genomic features from large contexts. Hence, it is of interest to explore how TransEPI can benefit from large-context features, especially those from the genomic loci far from the E-P pair.

To this end, we evaluated TransEPI using only the features in E-P and their flanking regions within a custom range (Figure 4**A**), and the features in the other regions were masked by setting their values to 0. We tested a series of flanking sizes from 1 kb to 500 kb, with training and evaluating the models for each flanking size, respectively. As shown in Figure 4**B**, a general trend is that the auPRC of the model grows as the flanking size increases, indicating that using large genomic contexts is beneficial for TransEPI.

In addition, inspired by Xi and Beer [31], we particularly wondered whether TransEPI could address the impact from other regulatory elements which may also interact with the E or P in the E-P pair. More specifically, can TransEPI employ not only the status of a pair of enhancer and promoter but also the potential contacts between them and other regulatory elements to make a prediction? To explore this hypothesis, we focus on the negative samples in the datasets. For each negative sample, we masked the target promoters of the enhancer (mask-target) and the same number of randomly selected non-interacting promoters (mask-random) of the enhancer, respectively (Figure 4**C**). If the hypothesis holds, we would observe a larger change of TransEPI-score induced by mask-target than that by mask-random. Figure 4**D** displays the distribution of the change in TransEPI-score of the negative samples in test datasets for the mask-target and the mask-random group (only the samples with TransEPI-score change >0.01 are shown. See Supplementary Notes for details). By one-sided *t*-test, we found the change in TransEPI-score of the mask-target group was significantly higher than that of the mask-random group in the test datasets (the *P*-values are 1.05E-28, 2.22E-11, 1.00E-3 and 1.57E-5, respectively). It implies that when

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

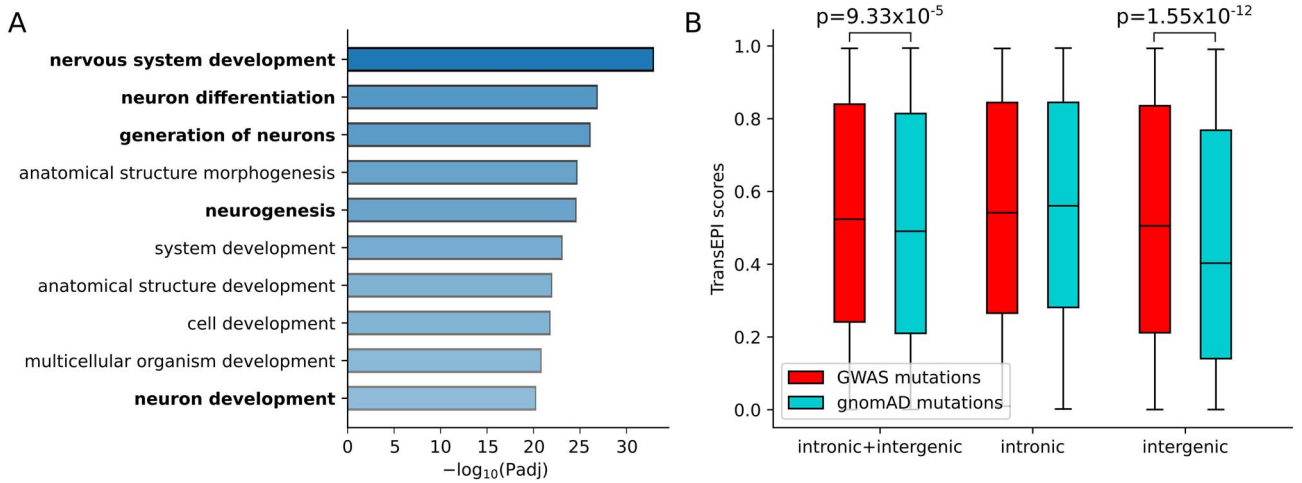

Figure 5. Applying TransEPI to find target genes for non-coding mutations. (**A**) The top-10 GO biological process (GO:BP) terms enriched by TransEPI-predicted protein-coding genes for non-coding variants related to neural diseases or brain disorders. There are five GO terms associated with neural functions in the top-10 GO:BP terms. (**B**) Comparison of the TransEPI-predicted top-1 scores of GWAS mutations and gnomAD mutations in the non-coding (intronic + intergenic), the intronic and the intergenic group. Statistical significance was assessed by *t*-test.

the features of the loci interacting with the enhancer are masked, the probability that the enhancer interacts with the other promoters will likely be overestimated. The observation is consistent with our hypothesis, potentially explaining the improved performance of TransEPI.

Taken together, we could conclude that TransEPI benefits from the genomic features outside enhancers and promoters.

TransEPI facilitates identifying target genes of non-coding mutations

Genome-wide association study (GWAS) has identified a large number of risk mutations related to diseases. However, most of the risk mutations reside in non-coding regions so that it is hard to understand the mechanisms of pathogenicity [8, 52]. One solution is to link non-coding mutations to target genes through chromatin interactions, while high-resolution 3D chromatin data are only available in a few cell types. Hence, we extended TransEPI to facilitate explaining GWAS results via predicting the target genes of non-coding mutations.

We collected 3943 non-coding (intronic and intergenic) mutations associated with neural diseases or brain disorders from Lu *et al.*'s work [14] (Supplementary Table S6 available online at http://bib.oxfordjournals.org/). Using TransEPI, we identified 5131 mutation-target gene pairs between 3034 genes and 2571 mutations (Supplementary Table S7 available online at http://bib.oxfordjournals.org/, only protein-coding genes are included). As a case study, we found that TransEPI correctly predicted the target genes of two mutations that have been validated by Hi-C [14]: rs10153620 (NRP2, TransEPI-score = 0.9100) [53] and rs10457592 (POU3F2, TransEPI-score = 0.9400) [54]. Next, we performed Gene Ontology (GO) analysis on the predicted genes using g:Profiler [55]. As shown in Supplementary Table S8 available online at http://bib.oxfordjournals.org/, the target genes significantly enrich 400 GO terms, including various neural function-associated GO terms. Notably, 5 out of the top-10 GO biological process (GO:BP) terms are relevant to neural functions (Figure 5**A**).

The above analysis implies the TransEPI-predicted genes may be functionally associated with neural functions. However, the statistical significance cannot be assessed since we lack the ground-truth target genes for most mutations. To further evaluate the predictions, we adopted an alternative approach by comparing the predicted target genes of disease-related mutations to those of disease-irrelevant mutations. We hypothesized that the disease-related mutations were more likely to interact with target genes than the disease-irrelevant mutations. Accordingly, the top-1 scores of the disease-related mutations should be higher than those of the disease-irrelevant ones.

To this end, we randomly sampled 19715 (five times that of GWAS mutations) non-coding mutations from the gnomAD database (v2) as disease-irrelevant mutations (Supplementary Table S9 available online at http://bib.oxfordjournals.org/). As shown in Figure 5**B**, the disease-related GWAS mutations have significantly higher top-1 scores than the disease-irrelevant ones (*P*-value = 9.33×10^{-5}, by *t*-test). When we set apart the intronic and intergenic mutations, we observed a more significant difference in the intergenic group (*P*-value = 1.55×10^{-12}), while no significant difference was found in the intronic group (*P*-value = 0.0990). This is likely because the intergenic mutations are more likely to affect distal target genes than intronic mutations.

Taken together, we could conclude that the TransEPI framework is also helpful to predict the target gene of non-coding mutations and thus could potentially facilitate explaining GWAS results.

Discussion

In this study, we present a novel deep learning model, TransEPI, for predicting EPIs by capturing large genomic contexts using the Transformer architecture. Unlike previous methods, we take the large genomic interval

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

where the E-P pair locates into consideration. In this way, TransEPI could address the impact of other distant regulatory elements that may potentially interact with the E or P [31].

Given the fact that the many EPI models suffer from over-fitting [21, 31, 32], we trained and fine-tuned the TransEPI model under the 5-fold CV scheme, where the data were split by chromosomes to ensure that the samples in different folds do not overlap. Besides, we evaluated TransEPI on independent datasets derived from four cell lines different from the training data. As TransEPI achieves a consistent performance on the CV and the independent test datasets, TransEPI is shown robust for predicting EPIs in different cell types in the BENGI datasets. Since TransEPI enabled accurate EPI prediction, we extended the framework to find the target genes of non-coding mutations. By applying the model on mutations associated with neural diseases or brain disorders, TransEPI found target genes that are functionally associated with neural functions. Moreover, these disease-associated non-coding mutations are found to have a higher probability to act on target genes than those irrelevant to diseases.

Although TransEPI has achieved state-of-the-art performance, there is still much room for improvement. For example, the time complexity and memory usage required by the standard Transformer module are quadratic to the length of the input sequence, which are computationally expensive. It is infeasible for us to process much larger genomic contexts in our model. In the future, we may adopt more lightweight Transformer architectures [56–58] to alleviate the problem. A recent study on predicting gene expression using Transformer demonstrates that the attention weights in the Transformer model may facilitate inferring chromatin interactions [42]. Thus, we can investigate whether the attention weights learned in TransEPI can provide more biological implications in future studies. An attempt to apply TransEPI to predict the enhancer-gene pairs derived from CRISPRi perturbations [59] only achieved relatively low performance (Supplementary Figure S7 available online at http://bib.oxfordjournals.org/). It is likely because the Hi-C identified contacts and CRISPRi-inferred enhancer-gene pairs are poorly overlapped [21]. The future versions of TransEPI could consider using EPIs identified by the other 3C-based methods like capture Hi-C [60] and HiChIP [61], or directly using the enhancer-gene pairs identified by genetic approaches (eQTL, CRISPRi perturbations, etc.). So far, we could only test TransEPI on several cell lines with high-quality chromatin interaction data. With genomic data accumulating, we are expected to be able to evaluate TransEPI in more tissues and cell lines in the future.

Data availability

The datasets and models are available at https://github.com/biomed-AI/TransEPI.

Key Points

- We present a Transformer-based model, TransEPI, using genomic data from large genomic contexts to predict EPIs in different cell types.
- TransEPI compares favorably to state-of-the-art methods on independent datasets from different cell types and chromosomes.
- TransEPI largely benefits from the large-scale context features outside enhancers and promoters.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This study has been supported by the National Key R&D Program of China (2020YFB0204803), National Natural Science Foundation of China (61772566), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), Introducing Innovative and Entrepreneurial Teams (2016ZT06D211), Guangzhou S&T Research Plan (202007030010).

References

1. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006;**7**:29–59.
2. Plank JL, Dean A. Enhancer function: mechanistic and genome-wide insights come together. *Mol Cell* 2014;**55**:5–14.
3. Heinz S, Romanoski CE, Benner C, *et al.* The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* 2015;**16**:144–54.
4. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* 2019;**20**:437–55.
5. Lupiáñez DG, Kraft K, Heinrich V, *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 2015;**161**:1012–25.
6. Li R, Liu Y, Hou Y, *et al.* 3D genome and its disorganization in diseases. *Cell Biol Toxicol* 2018;**34**:351–65.
7. Javierre BM, Burren OS, Wilder SP, *et al.* Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 2016;**167**:1369–1384.e19.
8. Chen J, Tian W. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. *Nucleic Acids Res* 2016;**44**:8641–54.
9. Sey NYA, Hu B, Mah W, *et al.* A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci* 2020;**23**:583–93.
10. Sanyal A, Lajoie BR, Jain G, *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–13.
11. Wang D, Rendon A, Wernisch L. Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Res* 2013;**41**:1450–63.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

12. Wu Z, Ioannidis NM, Zou J. Predicting target genes of non-coding regulatory variants with IRT. *Bioinformatics* 2020;**36**:4440–8.
13. Võsa U, Claringbould A, Westra H-J, *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021;**53**:1300–10.
14. Lu L, Liu X, Huang W-K, *et al.* Robust Hi-C maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. *Mol Cell* 2020;**79**:521–534.e15.
15. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289.
16. Fullwood MJ, Liu MH, Pan YF, *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009;**462**:58–64.
17. Rao SSP, Huntley MH, Durand NC, *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.
18. Thurman RE, Rynes E, Humbert R, *et al.* The accessible chromatin landscape of the human genome. *Nature* 2012;**489**:75–82.
19. Sheffield NC, Thurman RE, Song L, *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* 2013;**23**:777–88.
20. Fishilevich S, Nudel R, Rappaport N, *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* 2017;**2017**:bax028.
21. Moore JE, Pratt HE, Purcaro MJ, *et al.* A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol* 2020;**21**:17.
22. Roy S, Siahpirani AF, Chasman D, *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res* 2015;**43**:8694–712.
23. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96.
24. Cao Q, Anyansi C, Hu X, *et al.* Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* 2017;**49**:1428–36.
25. Talukder A, Saadat S, Li X, *et al.* EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction. *Bioinformatics* 2019;**35**:3877–83.
26. Gao T, Qian J. EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLoS Comput Biol* 2019;**15**:e1007436.
27. Singh S, Yang Y, Póczos B, *et al.* Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant Biol* 2019;**7**:122–37.
28. Zhuang Z, Shen X, Pan W. A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data. *Bioinformatics* 2019;**35**:2899–906.
29. Hong Z, Zeng X, Wei L, *et al.* Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 2020;**36**:1037–43.
30. Min X, Ye C, Liu X, *et al.* Predicting enhancer-promoter interactions by deep learning and matching heuristic. *Brief Bioinform* 2021;**22**:bbaa254.
31. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput Biol* 2018;**14**:e1006625.
32. Cao F, Fullwood MJ. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nat Genet* 2019;**51**:1196–8.
33. Schreiber J, Singh R, Bilmes J, *et al.* A pitfall for machine learning methods aiming to predict across cell types. *Genome Biol* 2020;**21**:282.
34. Belokopytova PS, Nuriddinov MA, Mozheiko EA, *et al.* Quantitative prediction of enhancer–promoter interactions. *Genome Res* 2020;**30**:72–84.
35. Schwessinger R, Gosden M, Downes D, *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;**17**:1118–24.
36. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**:1111–7.
37. Cao F, Zhang Y, Cai Y, *et al.* Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome Biol* 2021;**22**:226.
38. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:5998–6008.
39. Chang S, Zhang Y, Han W, *et al.* Dilated recurrent neural networks. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA: Curran Associates, Inc., 2017.
40. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
41. Baek M, DiMaio F, Anishchenko I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
42. Avsec Ž, Agarwal V, Visentin D, *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1–8.
43. Frankish A, Diekhans M, Ferreira A-M, *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;**47**:D766–73.
44. Kundaje A, Meuleman W, Ernst J, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
45. Lin Z, Feng M, Santos CN DOS, *et al.* A structured self-attentive sentence embedding. *arXiv:1703.03130* 2017.
46. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
47. Pratapa A, Jalihal AP, Law JN, *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**:147–54.
48. Paszke A, Gross S, Massa F, *et al.* PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019;**32**:8024–35.
49. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv:1412.6980* 2017.
50. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020;**48**:D58–64.
51. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* 2019;**47**:e60–0.
52. Edwards SL, Beesley J, French JD, *et al.* Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 2013;**93**:779–97.
53. Ebejer JL, Duffy DL, van der Werf J, *et al.* Genome-wide association study of inattention and hyperactivity-impulsivity measured as quantitative traits. *Twin Res Hum Genet Off J Int Soc Twin Stud* 2013;**16**:560–74.
54. Hyde CL, Nagle MW, Tian C, *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet* 2016;**48**:1031–6.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022

55. Raudvere U, Kolberg L, Kuzmin I, *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;**47**: W191–8.
56. Zhou H, Zhang S, Peng J, *et al.* Informer: beyond efficient transformer for long sequence time-series forecasting. *arXiv:2012.07436* 2020.
57. Choromanski K, Likhosherstov V, Dohan D, *et al.* Rethinking attention with Performers. *arXiv:2009.14794* 2020.
58. Katharopoulos A, Vyas A, Pappas N, *et al.* Transformers are RNNs: fast Autoregressive transformers with linear attention. *Proceedings of the 37th International Conference on Machine Learning*, 2020;**119**:5156–65.
59. Gasperini M, Hill AJ, McFaline-Figueroa JL, *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 2019;**176**:377–390.e19.
60. Mifsud B, Tavares-Cadete F, Young AN, *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 2015;**47**:598–606.
61. Mumbach MR, Rubin AJ, Flynn RA, *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016;**13**:919–22.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab577/6513727 by Sun Yat-Sen University user on 05 February 2022