Journal Title Here, XXXX, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Capturing large genomic contexts for accurately predicting enhancer-promoter interactions

Ken Chen,[1] Huiying Zhao[2] and Yuedong Yang[1,3,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, 510000, Guangzhou, China, [2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, 510000, Guangzhou, China and [3]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, Country

[*]Corresponding author. yangyd25@mail.sysu.edu.cn

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Accurately identifying enhancer-promoter interactions (EPIs) is challenging because enhancers usually act on the promoters of distant target genes. Although a variety of machine learning and deep learning models have been developed, many of them are not designed to or could not be well applied to predict EPIs in cell types different from the training data. In this study, we develop the TransEPI model for EPI prediction based on datasets derived from Hi-C and ChIA-PET data. TransEPI compiles genomic features from large intervals harboring the enhancer-promoter pair and adopts a Transformer-based architecture to capture the long-range dependencies. Thus, TransEPI could achieve more accurate prediction by addressing the impact of other genomic loci that may competitively interact with the enhancer-promoter pair. We evaluate TransEPI in a challenging scenario, where the independent test samples are predicted by models trained on the data from different cell types and chromosomes. TransEPI robustly predicts cross-cell-type EPI prediction by achieving comparable performance in cross-validation and independent test. More importantly, TransEPI significantly outperforms the state-of-the-art EPI models on the independent test datasets, with the Area Under Precision-Recall Curve (auPRC) score increasing by 48.84 % on average. Hence, TransEPI is applicable for accurate EPI prediction in cell types without chromatin structure data. Moreover, we find the TransEPI framework could also be extended to identify the target gene of non-coding mutations, which may facilitate studying pathogenic non-coding mutations.

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

Introduction

Enhancers are functional DNA fragments acting as cis-regulatory elements on the genome, which regulate gene expressions through chromatin interactions with the promoter of target genes [Maston et al., 2006, Plank and Dean, 2014]. The enhancer-promoter interactions (EPIs) vary highly with cell types and thus play a critical role in cell development and differentiation [Plank and Dean, 2014, Schoenfelder and Fraser, 2019]. 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 [Lupiáñez et al., 2015, Li et al., 2018]. Accordingly, linking enhancer mutations to the promoter of target genes could not only help to interpret a substantial number of non-coding mutations but also provide implications for therapeutic approaches [Javierre et al., 2016, Chen and Tian, 2016, Fadason et al., 2018]. However, it remains challenging to identify EPIs accurately because enhancers and their target promoters are typically separated by thousands of base pairs [Schoenfelder and Fraser, 2019, Sanyal et al., 2012].

Previous studies have utilized expression quantitative trait loci (eQTLs) to infer EPIs indirectly [Wang et al., 2013, Wu et al., 2020], while eQTL-based methods are limited to investigate only loci containing variants with high minor allele frequencies due to the requirement of a large number of samples [Consortium, 2015]. Over the last decade, high-throughput chromatin conformation capture-based (3C-based) techniques (e.g., Hi-C [Lieberman-Aiden et al., 2009], ChIA-PET [Fullwood et al., 2009]) have enabled to detect chromatin interactions directly, which could be applied to identify EPIs [Rao et al., 2014, Lu et al., 2020]. However, these 3C-based methods are costly and laborious, experimentally identified EPIs data are available in a few cell types.

To mitigate the problem of identifying EPIs, a variety of computational methods have been proposed to predict EPIs. Though enhancers are sometimes assumed to interact with the nearest promoter, such a method is not reliable because enhancers do not regulate the nearest gene in most cases [Sanyal et al., 2012]. Hence, correlation-based methods were developed later to decipher the underlying rules of EPIs using the correlations of genomic signals at enhancers and genes (or promoters) across a series of cell types [Thurman et al., 2012, Sheffield et al., 2013, Fishilevich et al., 2017]. These methods are of low performance because enhancers are usually cell-type-specific, and EPIs vary across cell types [Moore et al., 2020]. To unveil the complex determinants of EPIs, machine learning (including deep learning) approaches were adopted. Sequence-based methods were developed by making predictions from enhancer and promoter sequences through machine learning or deep learning techniques. Though many methods successfully predict using DNA sequences, such as PEP [Yang et al., 2017], SPEID [Singh et al., 2019], Zhuang's method [Zhuang et al., 2019], EPIVAN [Hong et al., 2020], the sequence-based methods are inherently cell-type-agnostic and thus are not useful for predicting cell-type-specific EPIs. In parallel, other methods trained machine learning models based on genomic features derived from ChIP-seq and DNase-seq to capture the cell-type-specific characteristics, such as RIPPLE [Roy et al., 2015], TargetFinder [Whalen et al., 2016], JEME [Cao et al., 2017], EPIP [Talukder et al., 2019], EAGLE [Gao and Qian, 2019]. They are more effective than sequence-based models, while they tend to become less inaccurate for predicting EPIs in cell types different from the training data. Besides, the performance of the machine learning models is often exaggerated because the test datasets are not rigorously independent from the training data[Cao and Fullwood, 2019, Moore et al., 2020]. Recently, several methods, like 3DPredictor [Belokopytova et al., 2020] and DeepC [Schwessinger et al., 2020], highlight the importance of employing features of large genomic contexts for chromatin structure modeling. Inspired by these models, we speculate that integrating additional genomic features from large genomic contexts would improve EPI prediction, given that the EPIs are inherently determined by chromatin conformation.

In this study, we present a novel deep learning model named TransEPI for EPI prediction using the Transformer architecture [Vaswani et al., 2017]. TransEPI directly takes the input of genomic signals from large genomic contexts harboring the enhancer-promoter pairs to predict the EPIs. We expected that the features outside enhancers and promoters could enable the model to address the impact of other genomic loci that may competitively interact with the enhancers or the promoters. Given the success of Transformer in protein structure modelling[Jumper et al., 2021, Baek et al., 2021], we adopt a Transformer-based framework in TransEPI for capturing the long-range dependencies between enhancers and promoters. To the best of our knowledge, it is the first model that applies Transformer to predict chromatin interactions. To make an EPI model robust for cross-cell-type prediction, we developed and evaluated TransEPI in a challenging scenario. The TransEPI is trained by a chromosome-split cross-validation scheme, where the training data were split by chromosomes. Then, we tested it on samples where the cell type and chromosome are both different from the training data, as we expected the evaluation could reliably reflect the actual performance of models. Our TransEPI method is shown robust by achieving comparable performance in cross-validation and the independent test datasets. More importantly, TransEPI significantly outperforms two state-of-the-art EPI models, with an average auPRC increase by 48.84 % on the test datasets. Additionally, we found our TransEPI framework is also helpful for predicting cell-type-specific target genes affected by non-coding mutations.

The implementation of TransEPI and the datasets are available at `https://github.com/biomed-AI/TransEPI`.

Materials and Methods

Datasets

We employed the BENGI dataset to develop TransEPI for predicting EPIs and the Hi-C loop dataset to extend TransEPI for identifying the target gene of non-coding mutations.

The BENGI dataset We recruit data from the "Benchmark of candidate Enhancer-Gene Interactions (BENGI)" dataset [Moore et al., 2020] to develop the TransEPI model. BENGI is a collection of enhancer-target gene pairs from several cell lines or tissues identified by 3C-based methods or genetic approaches. Since we aimed to predict the physical interactions between enhancers and promoters, only the samples identified by Hi-C (GM12878, HeLa-S3, HMEC, IMR90, K562, and NHEK) and ChIA-PET (GM12878 and HeLa-S3) were utilized.

Because the samples curated by BENGI are enhancer-gene interactions, we first mapped the genes to transcripts based on GENCODE annotation (GRCh37/hg19) [Frankish et al., 2019]. Then, by defining the 1500-base pair (bp) upstream and the 500-bp downstream of transcript start site (TSS) as the promoter, we obtained the enhancer-promoter(EP) pairs we need for developing our TransEPI model. Since even the TSSs of the same gene may be thousands of base pairs apart, the EP pairs derived from positive samples but with promoters residing outside the anchor region of chromatin loops were discarded. Besides, we also removed the samples with low-expressed transcripts (transcript per million (TPM) < 1) from both positive and negative samples as they were less likely to be regulated by enhancers (We did this to eliminate false-positive samples as many as possible). Finally, we obtained 45,182 positive and 307,135 negative samples (Table 1) from 6 cell lines.

We combined the samples from GM12878 and HeLa-S3 to construct a training set, namely BENGI-train, which contains 36,843 positive and 186,967 negative ones. The samples from the other 4 cell lines are all reserved for independent test, namely BENGI-HMEC, BENGI-IMR90, BENGI-K562, and BENGI-NHEK, respectively.

The HiC-loop dataset Because most of the non-coding mutations reside outside the putative enhancer regions, the original TransEPI model trained on BENGI fails to help find target genes of mutations. Therefore, we extended the TransEPI model for predicting general chromatin interactions by training it on a novel dataset consisted of Hi-C loops. To this end, we obtained Hi-C loops in 7 cell lines (GM12878, HeLa-S3, HMEC, HUVEC, IMR90, K562, and NHEK)

Table 1. Summary of the dataset

cell line	source	positive sample	negative sample
GM12878	Hi-C	2695	46,212
GM12878	CTCF 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

from the Gene Expression Omnibus (GEO) database under the accession number of GSE63525 [Rao et al., 2014].

Positive samples are defined as the pairs of Hi-C loop anchors. The negative samples were generated by randomly pairing the anchor regions of loops and the other randomly selected regions from the genome. Notably, the distances between the loop anchors are not in the uniform distribution (Figure S2). Therefore, we intentionally sampled more samples (about 50 %) matching the E-P distribution of positive samples to avoid the model predicting EPIs by simply using the E-P distance.

Finally, we compiled the Hi-C loop dataset consisting of $38,608$ positive and $272,397$ negative samples (Details shown in Table S2)

Input features

TransEPI is designed to take the input of features from large genomic regions harboring the candidate E-P pairs (Figure 1a). We set the length of large genomic regions to 2,500,000 bp because the maximum value of E-P distance in BENGI is $2,246,878$ and there are only 30 samples with E-P distance longer than $2,000,000$ bp. The start and the end of the large genomic regions are determined by extending 1,250,000 bp from the midpoint of the E-P pair up- and down-stream. When the enhancer or promoter is close to the ends of chromosomes, we will shift the region to keep the region within the range of chromosomes.

Inspired by Belokopytova et al[Belokopytova et al., 2020], we partitioned each 2.5Mbp region into 5000 consecutive bins using a bin size of 500 bp and averaged the genomic and epigenomic signals within each bin to represent the chromatin states. Here, the genomic and epigenomic signals include CTCF binding sites, chromatin accessibility (DNase-I signals), and 5 histone modification marks (H3K4me1, H3K4me3, H3K27me3, H3K36me3, and H3K9me3). The CTCF binding sites in narrowPeak format for each cell line were obtained from the Encyclopedia of DNA Elements (ENCODE) project [Davis et al., 2018]. The DNase-I and histone marks data in bigWig format were taken from the Roadmap Epigenomics Project [Kundaje et al., 2015]. Apart from the genomic features, we also encoded the relative distance to the enhancer or the promoter for each bin as an additional feature to make the model aware of the locations of the E-P. Details about feature preparation are described in Supplementary Methods.

The TransEPI model

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

Firstly, TransEPI utilizes a one-dimensional convolutional neural network (1D CNN) to exact 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 (The output sequence will be denoted by $\boldsymbol{X} = [x_1, x_2, \cdots, x_l]$, where $x_i \in \mathcal{R}^h$, h is the dimension of the hidden states).

Next, we employ the Transformer encoder module to capture the long-range dependencies between enhancers and promoters within the sequential signals $\boldsymbol{X}$. $\boldsymbol{X}$ is first transformed into a key, a query, and a value sequence, respectively:

$$\begin{aligned} \boldsymbol{K} &= \boldsymbol{X}\boldsymbol{W}_k, \\ \boldsymbol{Q} &= \boldsymbol{X}\boldsymbol{W}_q, \\ \boldsymbol{V} &= \boldsymbol{X}\boldsymbol{W}_v, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_k \in \mathcal{R}^{h\times d}, \boldsymbol{W}_q \in \mathcal{R}^{h\times d}, \boldsymbol{W}_v \in \mathcal{R}^{h\times h}$ are learnable weight matrices. $\boldsymbol{K}$ and $\boldsymbol{Q}$ are then used to construct the attention matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \text{softmax}\left(\frac{\boldsymbol{K}\boldsymbol{Q}^T}{\sqrt{d}}\right), \tag{2}$$

where the attention coefficient $a_{i,j}$ in $\boldsymbol{A}$ could be understood as the correlation between the i-th and the j-th position in $\boldsymbol{X}$. By multiplying $\boldsymbol{V}$ by $\boldsymbol{A}$, the hidden states at different positions are exchanged and updated based on $\boldsymbol{A}$. To make the Transformer model deeper, multiple Transformer encoder layers could be stacked sequentially. The output from the Transformer module is denoted by $\boldsymbol{M}_0 \in \mathcal{R}^{l\times h}$.

Thereafter, we employ a self-attention-based sequence embedding module [Lin et al., 2017] to obtain a low-dimensional embedding for each sequence. Specifically, we feed $\boldsymbol{M}_0$ into a two-layer fully-connected (FC) network to obtain the attention coefficients (weights) for different positions in $\boldsymbol{M}_0$ and then multiply $\boldsymbol{M}_0$ by the weights to obtain a weighted embedding $\boldsymbol{M}_1 \in \mathcal{R}^{l\times h}$. Then, we apply the average and the maximum pooling over all the hidden states in $\boldsymbol{M}_1$ and concatenate them with the hidden states corresponding to the location of the enhancer ($\boldsymbol{h}_e$) and the promoter ($\boldsymbol{h}_p$):

$$\boldsymbol{M} = \text{AvgPool}(\boldsymbol{M}_1)||\text{MaxPool}(\boldsymbol{M}_1)||\boldsymbol{h}_e||\boldsymbol{h}_p, \tag{3}$$

where $\boldsymbol{M} \in \mathcal{R}^{4h}$. It includes the global features from the whole sequence and the local features from the enhancer and the promoter.

Finally, the final sequence embedding $\boldsymbol{M}$ is passed to a two-layer FC module to predict the EPI:

$$p = \text{sigmoid}(\boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{M} + b_1) + b_2), \tag{4}$$

where $\boldsymbol{W}_1 \in \mathcal{R}^{4h\times f}$, $\boldsymbol{W}_2 \in \mathcal{R}^{f\times 1}, b_1, b_2$ are all the weights in the FC module. As a Sigmoid function is used, p ranges from 0 to 1 ($p \in (0, 1)$), representing the probability that the input enhancer and the promoter interact with each other. Meanwhile, in order to make TransEPI sensitive to the location of the enhancer and the promoter, we use another FC module to predict the E-P distance:

$$d_{\text{pred}} = \boldsymbol{W}_4(\boldsymbol{W}_3\boldsymbol{M} + b_3) + b_4. \tag{5}$$

Evaluation metrics

We evaluated the model using the Area Under the Precision-Recall (PR) curve (auPRC) and the Area Under the Receiver Operating Characteristic (ROC) curve [Hanley and McNeil, 1982] (AUC). The PR curve is a plot of precision against the recall at a series of thresholds. Similarly, the ROC curve is a plot of true-positive rate (TPR) against the false-positive curve (FPR). Precision, recall, TPR, and FPR are defined as:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}, \end{aligned} \tag{6}$$

where TP, FP, TN, and FN are short for the True Positives (correctly predicted interacting pairs), False Positives (falsely predicted non-interacting pairs), True Negatives (correctly predicted non-interacting pairs), and False Negatives (falsely predicted interacting pairs).

Since the auPRC is associated with the ratio of the number of positive and negative samples, we also used auPRC-ratio (dividing auPRC by the proportion of positive samples [Pratapa et al., 2020]) as a metric for comparing performance across datasets.

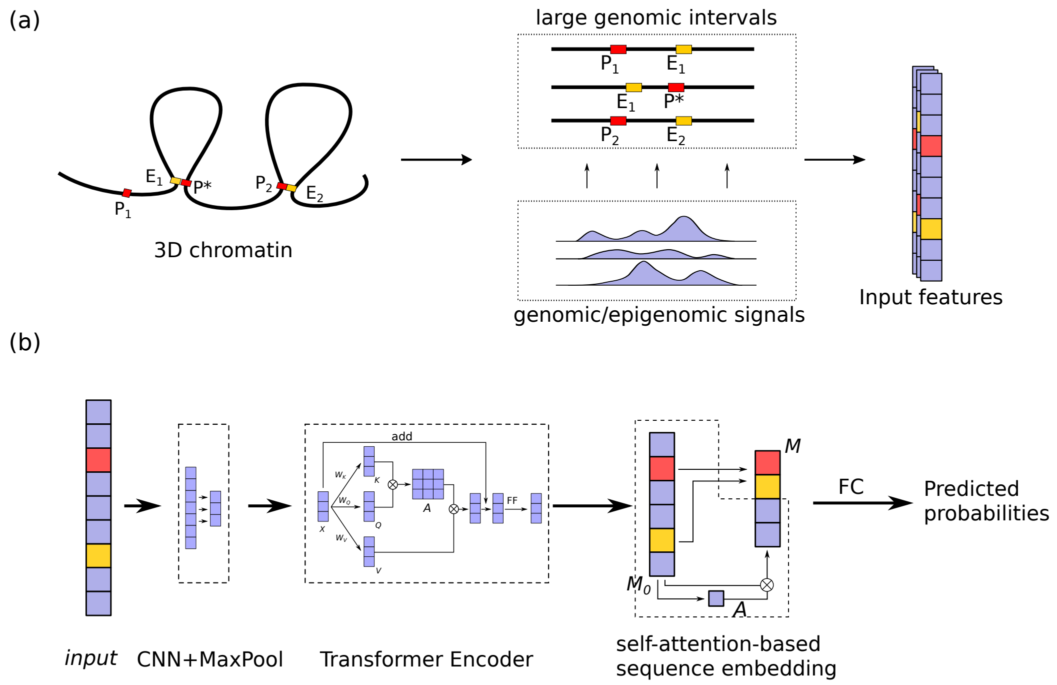

Fig. 1. The TransEPI framework. (a) Feature preparation. Genomic features (CTCF, DNase I, H3K27me3, H3K4me1, H3K4me3, H3K36me3, and H3K9me1) are extracted from large intervals harboring the candidate enhancer-promoter pairs (enhancers are marked in yellow and promoters are marked in red); (b) The architecture of the TransEPI. TransEPI are mainly consisted of 3 modules, a CNN+MaxPooling module extracting features from the input sequences, a Transformer Encoder module capturing the long-range dependencies between enhancers and promoters, and a self-attention-based sequence embedding module encoding the sequential features into a low-dimensional embedding. Finally, the embedding of the whole sequence and the hidden states of the enhancer bin and the promoter bin are concatenated together to predict the probability with a fully-connected layer. (⊗: matrix multiplying operation: CNN: convolutional neural network; MaxPool: max-pooling; FC: fully-connected)

Model training and evaluation

The TransEPI model is implemented with PyTorch (version 1.9.0) [Paszke et al., 2019] in Python 3.8. It was trained to minimize the binary cross entropy loss for EPI prediction and the mean squared error (MSE) loss for E-P distance prediction simultaneously:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\mathrm{EPI}} + \mathcal{L}_{\mathrm{EP-distance}} \\ =& -\frac{1}{N}\sum_{i}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \\ & + \frac{1}{N}\sum_{i}^{N}(d_{\mathrm{pred},i} - d_{\mathrm{true},i})^2,\end{aligned} \tag{7}$$

where p_i, y_i, $d_{\mathrm{pred},i}$, and $d_{\mathrm{true},i}$ are the the predicted EPI probability, true EPI label (0 or 1), the predicted E-P distance, the true E-P distance of the i-th sample, respectively. We used the Adam optimizer [Kingma and Ba, 2017] to update the weights in the neural network. The early stopping strategy was utilized for regularization.

In order to avoid over-fitting, we adopt a cross-chromosome 5-fold cross-validation scheme to fine-tune the hyper-parameters (Figure S1) in TransEPI. We divided the samples in BENGI-train into 5 folds by chromosomes, ensuring that the samples from the same chromosome would also be put into the same fold (chromosomes assigned to each fold are listed in Table S1). It is a critical approach to avoid over-fitting for building machine learning models in genomics [Schreiber et al., 2020]. In each training epoch, we trained the model on 4 folds and validated it on the remaining fold in turn. The average AUC and AUPR on the 5 folds are used to measure the performance. For evaluating our method in a challenging way, the independent test sets are also split into 5 folds by chromosomes according to the partition table in Table S1. The test samples are only predicted by the models trained on different chromosomes, as illustrated in Figure S1.

State-of-the-art methods for comparison

We compared TransEPI to two state-of-the-art methods: TargetFinder and 3DPredictor.

TargetFinder The TargetFinder model is a Gradient Boosting model using various genomic features [Whalen et al., 2016]. As shown in two studies [Xi and Beer, 2018, Cao and Fullwood, 2019], the reported performance in the original paper was inflated because the test data was not rigorously independent from the training data. However, it still outperforms many other EPI models, according to a benchmark study by Moore et al [Moore et al., 2020]. Thus, we used TargetFinder as the first baseline model for comparison. Given that no official implementation of TargetFinder is available, we implemented it in Python using the XGBoost package, which is a state-of-the-art gradient boosting model [Chen and Guestrin, 2016]. The genomic features used in TargetFinder are the same as our TransEPI model.

3DPredictor The 3DPredictor model [Belokopytova et al., 2020] is an XGBoost model which was originally developed to identify enhancer-promoter contacts via predicting the Hi-C contact map. It demonstrates that integrates only integrating the oriented CTCF binding peaks within and around the pair of chromatin loci could achieve accurate predictions. Here, we adapted it to directly predict EPIs.

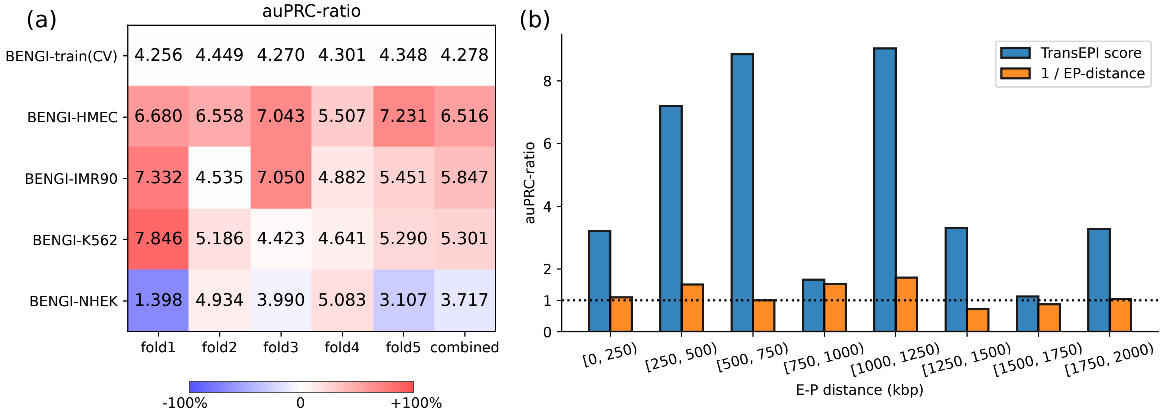

Fig. 2. Evaluating TransEPI on independent test datasets. (a) The auPRC-ratio scores of TransEPI on BENGI-train(CV) and 4 independent tests. (b) The auPRC-ratio scores of TransEPI on independent test samples stratified by distance. TransEPI consistently outperforms the naive distance-based method in every group.

For a fair comparison, the baseline models and TransEPI were trained on the same datasets. The hyper-parameters in the baseline models are rigorously refined using the random grid search strategy.

The source codes associated with data preparation and model training of the baseline models are available in our Github repository, as well.

Identifying target genes for non-coding mutations

We extend the TransEPI framework to predict the target gene of non-coding mutations. As the original TransEPI is built for predicting EPIs, it is not directly applicable to predict mutation-gene pairs because we expect it to be applicable for all non-coding mutations, even the mutations residing outside enhancer regions. Hence, we trained TransEPI on the Hi-C loop dataset, using HMEC, HUVEC, and the remaining cell lines as the validation, test, and training datasets, respectively.

For identifying the target genes, a candidate mutation-transcript list is first curated for each mutation by pairing it with the transcripts (transcript's TSS) within 1M bp. Subsequently, TransEPI is applied on the mutation-TSS pairs using genomic features from two neural cells and 3 brain tissues (Table S10). The mutation-TSS pair with a predicted probability above a certain threshold will be kept as interacting pairs.

Results

TransEPI accurately predicts EPIs in different cell types

EPIs are cell-type-specific chromatin interactions that vary significantly across cell types. To make TransEPI applicable for cross-cell-type EPI prediction, it is critical to avoid over-fitting when we develop and evaluate the model [Schreiber et al., 2020]. To this end, we not only adopt the chromosome-split 5-fold cross-validation scheme to train and fine-tune the model but also test it on samples from different cell types and chromosomes (Figure S1).

We first compared the results of TransEPI on cross-validation and independent test datasets to investigate whether it could achieve consistent performance across cell types. As shown in Figure 2a, the auPRC-ratios achieved by TransEPI on 4 independent test datasets are 6.516, 5.847, 5.301, and 3.717, 3 out of which are higher (in red) than that on the cross-validation dataset (auPRC-ratio=4.278). Moreover, such a trend is consistently observed in the results on each fold. As such, the results suggest that TransEPI trained on BENGI-train can be well applied to new cell types.

Previous studies have found that the distance from enhancer to the promoter (EP-distance) may have a strong predictive power on some datasets [Gao and Qian, 2019, Moore et al., 2020]. However, in fact, the predictive power of the EP-distance is determined by the distribution of EP-distance in negative samples, which could be controlled by the way of negative sampling. In order to eliminate potential bias caused by EP-distance, we additionally evaluated TransEPI on test datasets stratified by EP-distance. Specifically, we merged the samples in BENGI-HMEC, BENGI-IMR90, BENGI-K562, and BENGI-NHEK and stratified the samples by E-P distance into 8 groups using the bin size of 250,000 bp. In this way, we ensured that using only EP-distance can hardly predict EPIs in each group, as the auPRC-ratio ranges from 0.7222 to 1.727 (a random model is expected to achieve the auPRC-ratio of 1). In contrast, TransEPI achieves much higher auPRC-ratios, which range from 1.125 to 9.036, on the E-P distance-stratified datasets (Figure 2b), demonstrating that TransEPI does capture the underlying determinants of EPI instead of just relying on E-P distance alone.

Taken together, we could conclude that TransEPI is capable of deciphering the mechanisms of enhancer-promoter interaction that are universal to different cell types.

TransEPI significantly outperforms the baseline methods

To further evaluate TransEPI, we compared it with two state-of-the-art methods, TargetFinder [Whalen et al., 2016] and 3DPredictor [Belokopytova et al., 2020], on the independent test datasets.

In Figure 3a, the ROC plots show that TransEPI significantly outperforms TargetFinder and 3DPredictor on BENGI-HMEC and BENGI-IMR90 (all the P-values $< 1 \times 10^{-5}$, by McNeil & Hanley's test on AUC, as listed in Table S3). On BENGI-K562 and BENGI-NHEK, although the AUC of TransEPI does not surpass that of TargetFinder and 3DPredictor, TransEPI tends to have a higher TPR when the FPR scores are close to 0. This means that TransEPI can find more interacting E-P pairs when controlling false positive rate at a low level, implying that TransEPI is more helpful for practical use. When we compare these models by auPRC (Figure 3b), TransEPI consistently outperforms TargetFinder and 3DPredictor on all the test datasets. The average auPRC of TransEPI increases by 48.84 % compared with the second best method TargetFinder, further demonstrating the superior performance of TransEPI to the other methods.

Compared to our TransEPI model, TargetFinder employs only the average genomic signals within the genomic windows between enhancers and promoters. It does not only lack a fine representation of the features in genomic windows but also completely ignores the features within the outer regions of the enhancer-promoter pairs. Therefore, we believe that it is important to use a finer features representation strategy and leverage features from larger genomic contexts.

As for 3DPredictor, its authors employ only the CTCF binding site as a feature for chromatin structure modelling as they found additional features could not provide a significant performance improvement. However, in our study, additional features are found beneficial (See Supplementary Methods and Table S4). We speculate that this is because the EPIs may have distinct characteristics from the common chromatin interactions, as the activity of promoters and enhancers are usually inferred from chromatin accessibility and histone modification marks [Tsompana and Buck, 2014, Gao and Qian, 2020].

TransEPI benefits from the features outside enhancers and promoters

In this section, we quantitatively studied the contribution of features from the window regions between the enhancers and promoters ("window features") and the neighboring regions outside the enhancer-promoter pairs ("neighbor features"). To this end, we masked the window or the neighbor features by setting the genomic features within the window (the w/o W setting) or the neighbor (the w/o N setting) regions to 0, respectively. In both settings, the features in the enhancer and promoter bin along with the 2 bins up- and down-stream of them are kept. Additionally, we also conducted a w/o NW setting by masking both the window and the neighbor features.

As shown in Figure 4a, masking window and neighbor features decrease the auPRC scores on all 4 independent test datasets. The exclusion of window features has the most prominent effect, with the average auPRC score decreasing by 41.45 % (Table S5). Such an observation is in concordance with previous studies [Whalen et al., 2016, Gao and Qian, 2019], in which they demonstrate the importance of window features. More importantly, we found that the neighbor features are non-trivial, as well. The absence of neighbor features decreases the average auPRC score by 13.08 % and masking both neighbor and window features results in the average auPRC score decreasing by 54.96 %, on the 4 independent test datasets (Table S5).

A case study under the w/o N setting intuitively explains why the features outside enhancers and promoters are indispensable. In cell line NHEK, the enhancer E_0 (EH37E0265448,chr12:47,070,581-47,071,150) interacts with the promoter P_1 (ENSEMBL transcript ID: ENST00000548870, TSS=chr12:46,761,193), but not the promoter P_0 (ENSEMBL transcript ID: ENST00000550413, TSS=chr12:47,473,425). The promoter P_1 resides at the neighboring region of E_0-P_1. When we use all the features, TransEPI correctly identifies E_0-P_1 as an interacting pair ($P(E_0\text{-}P_1) = 0.7248$) and E_0-P_0 as a non-interacting pair ($P(E_0\text{-}P_0) = 0.4732$), respectively (taking 0.5 as the threshold) (Figure 4b). When we mask the neighbor features, TransEPI is not able to aware of the existence of promoter P_0. As a result, TransEPI mistakenly regards E_0-P_0 as an interacting pair, with $P(E_0\text{-}P_0)$ increasing from 0.4732 to 0.7254 (Figure 4c). The case study suggests the other regulatory elements (elg.: enhancers, promoters) around the enhancer-promoter pair may competitively interact with them. Considering only the chromatin states of the enhancer-promoter pair is far from enough for predicting EPIs accurately.

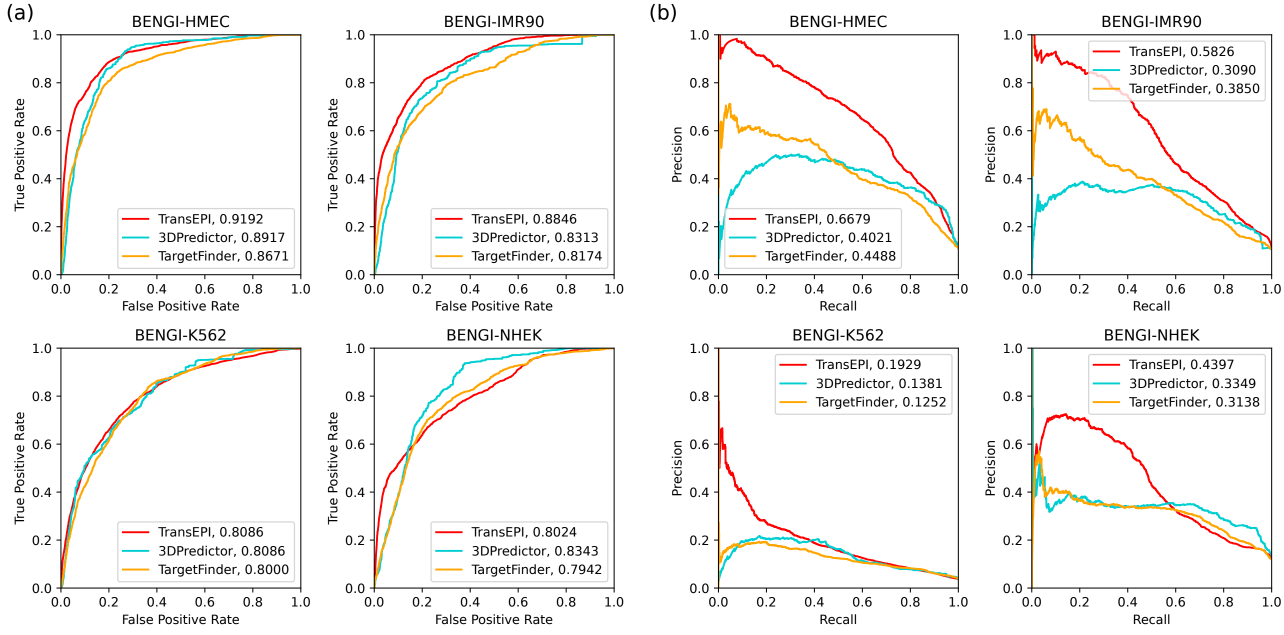

Fig. 3. Comparing TransEPI with TargetFinder and 3DPredictor on 4 independent test sets. (a) Receiver Operating Characteristic (ROC) curves and (b) Precision-recall (PR) curves are plotted.

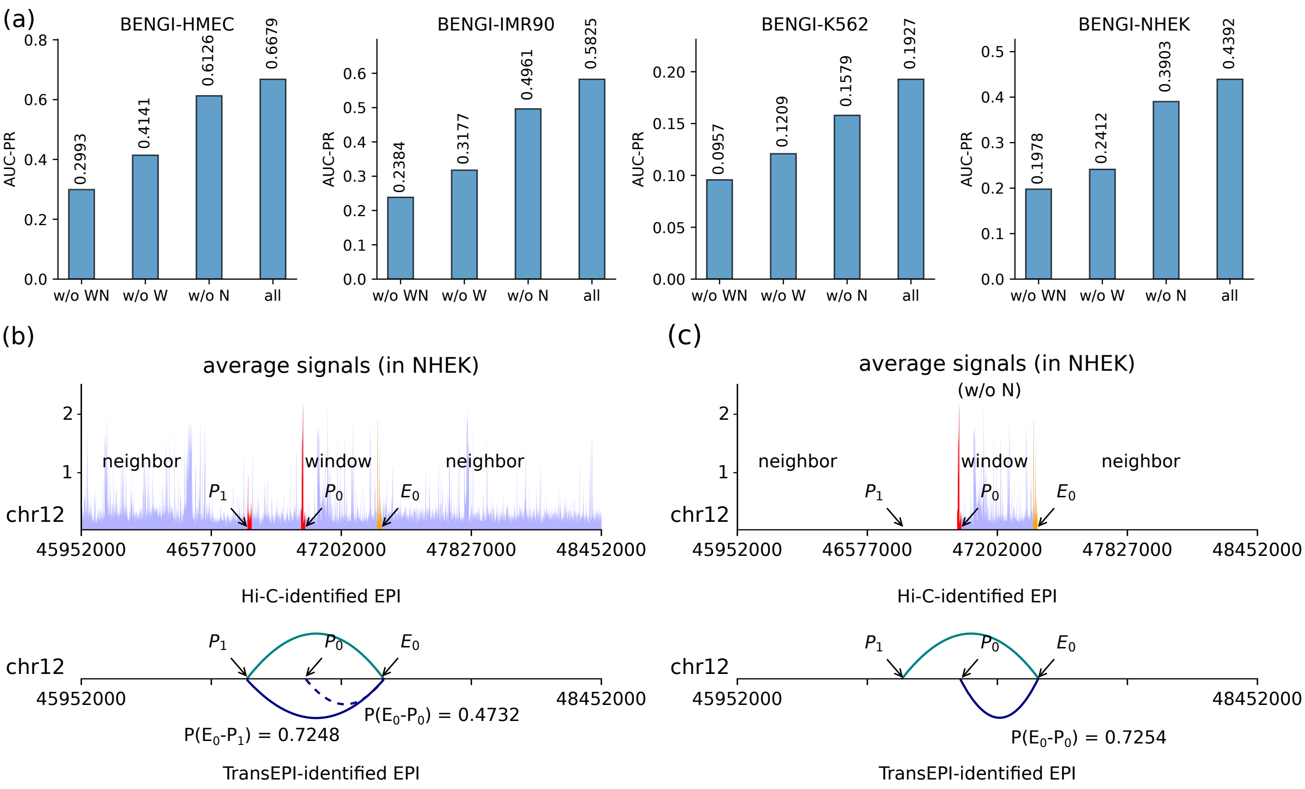

Fig. 4. Ablation study on features outside enhancers and promoters. (a) We reported the average auPRC of TransEPI model using all the features (all), masking features in the neighbor regions (w/o N), masking features in window regions (w/o W), and masking feature in both neighbor and window regions (w/o NW) on 4 independent test sets, where the error bar stands for the standard deviation; (b) E_0 (EH37E0265448) is an enhancer in the NHEK cell line. P_0 and P_1 are promoters of transcripts ENST00000548870 and ENSG00000139211, respectively. When we use both the neighbor and the window features, TransEPI correctly predicts E_0-P0 and E_0-P_1 as non-interacting (marked with dotted line) and interacting pairs (marked with solid line), respectively. (c) When we mask the neighbor features, TransEPI mistakenly identifies E_0-P_0 as an interacting pair. The genomic signals shown in (b) and (c) are the average over different features.

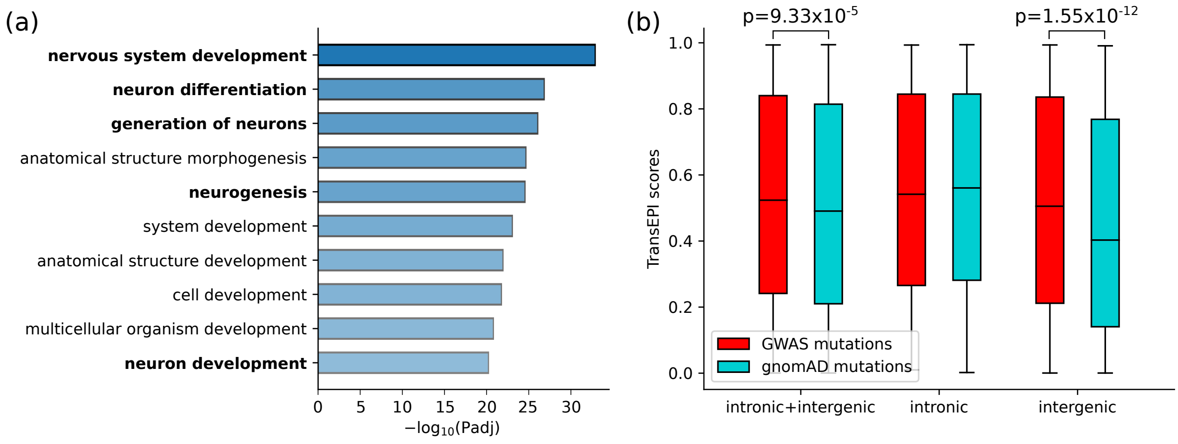

Fig. 5. Applying TransEPI to find target genes for non-coding mutations. (a) The top-10 Gene Ontology Biological Process (GO:BP) items enriched by TransEPI-predicted protein-coding genes for a group of non-coding variants related to neural diseases or brain disorders. There are 5 GO items associated with neural functions among the top-10 GO items. (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 is assessed by *t*-test.

TransEPI facilitates identifying target genes of non-coding mutations

Explaining GWAS results may be a challenging task because a large number of the risk variants reside in non-coding regions whose functions are not well characterized. An effective solution is to link the variants to target genes and previous studies have successfully employed Hi-C or eQTL data for explaining GWAS results [Sey et al., 2020, Chen and Tian, 2016, Lu et al., 2020]. However, the Hi-C and eQTL data are tissue-specific and may not be available in the tissue or the cell line studied by researchers. Here, given the outstanding performance of TransEPI for predicting EPIs, we extended to predict tissue-specific target genes of non-coding variants.

We collected the mutations associated with neural diseases or brain disorders sorted by Lu et al [Lu et al., 2020], which are taken from the GWAS Catalog [Buniello et al., 2019]. Only the non-coding mutations residing in intergenic or intronic regions are kept, by which we obtained 3943 non-coding mutations (Table S6). Using TransEPI, we identified 5131 mutation-target gene pairs associated with 3034 genes and 2571 mutations (Table S7), only protein coding genes are included). We firstly conducted Gene Ontology (GO) analysis on the target genes using gProfiler [Raudvere et al., 2019]. As shown in Table S8, the target genes significantly enrich 400 GO terms, which suggests that TransEPI-predicted genes are of biological significance. More importantly, we found various neural function-associated GO terms and 5 out of the top-10 GO biological process (GO:MF) are relevant to neural functions (Figure 5a). As a case study, TransEPI correctly identifies the target gene of two mutations: rs10153620 (NRP2, TransEPI-score = 0.9100) [Ebejer et al., 2013] and rs10457592 (POU3F2, TransEPI-score = 0.9400)[Hyde et al., 2016], which have been validated using Hi-C[Lu et al., 2020].

The above analysis implies that TransEPI-predicted genes may be functionally associated with neural functions. However, the statistical significance of our observations can not 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 first extracted the highest TransEPI-predicted probability, namely top-1 score, among all the candidate target genes for each mutation. The top-1 score represents the probability that a mutation interacts with at least one target gene. We believe that the disease-related mutations are expected to be more likely to interact with target genes than the disease-irrelevant ones. Accordingly, the top-1 scores of the disease-related mutations should be higher than those of the disease-irrelevant ones.

Specifically, we used the above-mentioned GWAS mutations as disease-related mutations (GWAS-mutations) and compiled a group of disease-irrelevant non-coding mutations (n = 19,715, 5 times that of GWAS mutations) from the gnomAD database (gnomAD-mutations) (Table S9). As shown in Figure 5b, the disease-related mutations have significantly higher top-1 scores than the disease-irrelevant ones (P-value = 9.33×10^{-5}, by t-test). Next, we split non-coding mutations into an intronic and an intergenic group and compared the predictions for them, respectively. We observed a more significant difference in the intergenic group (P-value=1.55×10^{-12}), while no significant difference is found in the intronic group. This is likely because that 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 identify 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. Instead of considering only the states of an individual pair of enhancer and promoter (EP), TransEPI takes the whole environment where they locate into account. In this way, TransEPI could be aware of the impact of other regulatory elements that may competitively interact with the EP and hence achieve the state-of-the-art performance for EPI prediction.

A variety of EPI models have been proposed yet, while many of them suffer from exaggerated performance and are not well applicable for cross-cell-type EPI prediction [Xi and Beer, 2018, Cao and Fullwood, 2019, Moore et al., 2020]. This is because the datasets used for training and validation (or test) are randomly separated. Therefore, samples with sharing features may be included in both the training and the validation data, resulting in severe over-fitting caused by data leakage. To alleviate the problem of over-fitting, we train and fine-tune the TransEPI model through 5-fold cross-validation (CV), where the data are split by chromosomes to ensure that the samples in different folds do not overlap with each other. Besides, we evaluate TransEPI on independent datasets derived from 4 cell lines to assess whether it could predict EPIs in cell lines different from the training data. As TransEPI achieves comparable performance in CV and on independent test datasets, the chromosome-split cross-validation scheme is shown to be effective to avoid over-fitting.

Since EPIs are inherently determined by the conformation of chromatin, we speculated that additional enhancers and promoters around the E-P pair to be predicted are all critical for accurate EPI prediction. To effectively capture the long-range dependencies between the enhancers and promoters, we present the TransEPI framework which is mainly based on the Transformer encoder architecture. By ablation study, we found TransEPI is sensitive to the additional regulatory elements that may competitively interact with the E-P pair to be predicted, demonstrating the importance of using large genomic contexts in EPI models.

Given that TransEPI enables 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 ratio to act on target genes than those irrelevant to diseases.

Although TransEPI has achieved the 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 [Vaswani et al., 2017] are quadratic to the length of the input sequence, which are computationally expensive. It is infeasible for us to take larger genomic contexts into consideration in our model. In the future, we may leverage more light-weight Transformer architectures [Zhou et al., 2020, Choromanski et al., 2020, Katharopoulos et al., 2020] to alleviate the problem. As for datasets, we currently leverage only EPIs identified by Hi-C and ChIA-PET. Future versions of TransEPI could consider using EPIs identified by the other 3C-based methods like capture Hi-C [Mifsud et al., 2015] and HiChIP[Mumbach et al., 2016], which are more suitable to detect EPIs. Besides, due to the resolution of 3C-based experiments, the assignment of enhancers to target promoters may be ambiguous and thus make the training data noisy. It is of interest to explore how to integrate additional evidence (e.g.: eQTL mapping) or employ the models enhancing the resolution of Hi-C data [Zhang et al., 2018, Li and Dai, 2020]to improve the quality of training data.

Competing interests

There is NO Competing Interest.

Acknowledgments

This study has been supported by the National Key R&D Program of China (2020YFB020003), 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

M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. v. Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, July 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abj8754. URL `https://science.sciencemag.org/content/early/2021/07/19/science.abj8754`. Publisher: American Association for the Advancement of Science Section: Research Article.

P. S. Belokopytova, M. A. Nuriddinov, E. A. Mozheiko, D. Fishman, and V. Fishman. Quantitative prediction of enhancer–promoter interactions. *Genome Res.*, 30(1):72–84, Jan. 2020. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.249367.119. URL `http://genome.cshlp.org/lookup/doi/10.1101/gr.249367.119`.

A. Buniello, J. A. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, Jan. 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1120. URL `https://doi.org/10.1093/nar/gky1120`.

F. Cao and M. J. Fullwood. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nat Genet*, 51(8):1196–1198, Aug. 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0434-7. URL `https://www.nature.com/articles/s41588-019-0434-7`.

Q. Cao, C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng, and K. Y. Yip. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nature Genetics*, 49 (10):1428–1436, Oct. 2017. ISSN 1546-1718. doi: 10.1038/ng.3950. URL `https://www.nature.com/articles/ng.3950`.

J. Chen and W. Tian. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. *Nucleic Acids Res*, 44(18):8641–8654, Oct. 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw519. URL `https://academic.oup.com/nar/article/44/18/8641/2468323`.

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`. event-place: San Francisco, California, USA.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller. Rethinking Attention with Performers. *arXiv:2009.14794 [cs, stat]*, Sept. 2020. URL `http://arxiv.org/abs/2009.14794`. arXiv: 2009.14794.

T. G. Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348 (6235):648–660, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1262110. URL `https://science.sciencemag.org/content/348/6235/648`.

C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, 46(D1):D794–D801, 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx1081.

J. L. Ebejer, D. L. Duffy, J. van der Werf, M. J. Wright, G. Montgomery, N. A. Gillespie, I. B. Hickie, N. G. Martin, and S. E. Medland. Genome-wide association study of inattention and hyperactivity-impulsivity measured as quantitative traits. *Twin Res Hum Genet*, 16(2):560–574, Apr. 2013. ISSN 1832-4274. doi: 10.1017/thg.2013.12.

T. Fadason, W. Schierding, T. Lumley, and J. M. O'Sullivan. Chromatin interactions and expression quantitative trait loci reveal genetic drivers of multimorbidities. *Nature Communications*, 9(1):5198, Dec. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07692-y. URL `https://www.nature.com/articles/s41467-018-07692-y`. Number: 1 Publisher: Nature Publishing Group.

S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet, and D. Cohen. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, 2017, Jan. 2017. doi: 10.1093/database/bax028. URL `https://academic.oup.com/database/article/doi/10.1093/database/bax028/3737828`. Publisher: Oxford Academic.

A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773, Jan. 2019. ISSN 1362-4962. doi: 10.1093/nar/gky955.

M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K. Sung, E. T. Liu, C.-L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, Nov. 2009. ISSN 1476-4687. doi: 10.1038/nature08497.

T. Gao and J. Qian. EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLOS Computational Biology*, 15 (10):e1007436, Oct. 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007436. URL `https://journals.plos.org/ploscompbiol/`

article?id=10.1371/journal.pcbi.1007436.

T. Gao and J. Qian. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res*, 48(D1):D58–D64, Jan. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz980. URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz980/5628925.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, Apr. 1982. ISSN 0033-8419. doi: 10.1148/radiology.143.1.7063747. URL https://doi.org/10.1148/radiology.143.1.7063747.

H. Hong, S. Jiang, H. Li, G. Du, Y. Sun, H. Tao, C. Quan, C. Zhao, R. Li, W. Li, X. Yin, Y. Huang, C. Li, H. Chen, and X. Bo. DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. *PLOS Computational Biology*, 16(2):e1007287, Feb. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007287. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007287. Publisher: Public Library of Science.

C. L. Hyde, M. W. Nagle, C. Tian, X. Chen, S. A. Paciga, J. R. Wendland, J. Y. Tung, D. A. Hinds, R. H. Perlis, and A. R. Winslow. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet*, 48(9):1031–1036, Sept. 2016. ISSN 1546-1718. doi: 10.1038/ng.3623.

B. Javierre, O. Burren, S. Wilder, R. Kreuzhuber, S. Hill, S. Sewitz, J. Cairns, S. Wingett, C. Várnai, M. Thiecke, F. Burden, S. Farrow, A. Cutler, K. Rehnström, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett, F. Wang, H. G. Stunnenberg, J. A. Todd, D. R. Zerbino, O. Stegle, W. H. Ouwehand, M. Frontini, C. Wallace, M. Spivakov, and P. Fraser. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, Nov. 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.09.037. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5123897/.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, pages 1–11, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biophysics;Machine learning;Protein structure predictions;Structural biology Subject_term_id: computational-biophysics;machine-learning;protein-structure-predictions;structural-biology.

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2006.16236.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Jan. 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb. 2015. ISSN 1476-4687. doi: 10.1038/nature14248. URL https://www.nature.com/articles/nature14248/.

R. Li, Y. Liu, Y. Hou, J. Gan, P. Wu, and C. Li. 3D genome and its disorganization in diseases. *Cell Biol Toxicol*, 34(5):351–365, Oct. 2018. ISSN 1573-6822. doi: 10.1007/s10565-018-9430-4. URL https://doi.org/10.1007/s10565-018-9430-4.

Z. Li and Z. Dai. SRHiC: A Deep Learning Model to Enhance the Resolution of Hi-C Data. *Front. Genet.*, 11, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00353. URL https://www.frontiersin.org/articles/10.3389/fgene.2020.00353/full. Publisher: Frontiers.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289, Oct. 2009. doi: 10.1126/science.1181369. URL http://science.sciencemag.org/content/326/5950/289.abstract.

Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A Structured Self-attentive Sentence Embedding. *arXiv:1703.03130 [cs]*, Mar. 2017. URL http://arxiv.org/abs/1703.03130. arXiv: 1703.03130.

L. Lu, X. Liu, W.-K. Huang, P. Giusti-Rodríguez, J. Cui, S. Zhang, W. Xu, Z. Wen, S. Ma, J. D. Rosen, Z. Xu, C. F. Bartels, R. Kawaguchi, M. Hu, P. C. Scacheri, Z. Rong, Y. Li, P. F. Sullivan, H. Song, G.-l. Ming, Y. Li, and F. Jin. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Molecular Cell*, page S1097276520303920, June 2020. ISSN 10972765. doi: 10.1016/j.molcel.2020.06.007. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276520303920.

D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, May 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.04.004.

G. A. Maston, Sara K. Evans, and M. R. Green. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genom. Hum. Genet.*, 7(1):29–59, Sept. 2006. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev.genom.7.080505.

115623. URL http://www.annualreviews.org/doi/10.1146/annurev.genom.7.080505.115623.

B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. LeProust, G. A. Follows, P. Fraser, N. M. Luscombe, and C. S. Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, June 2015. ISSN 1546-1718. doi: 10.1038/ng.3286. URL https://www.nature.com/articles/ng.3286. Number: 6 Publisher: Nature Publishing Group.

J. E. Moore, H. E. Pratt, M. J. Purcaro, and Z. Weng. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biology*, 21(1):17, Jan. 2020. ISSN 1474-760X. doi: 10.1186/s13059-019-1924-8. URL https://doi.org/10.1186/s13059-019-1924-8.

M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13(11):919–922, Nov. 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3999. URL https://www.nature.com/articles/nmeth.3999. Number: 11 Publisher: Nature Publishing Group.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

J. Plank and A. Dean. Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. *Molecular Cell*, 55(1):5–14, July 2014. ISSN 10972765. doi: 10.1016/j.molcel.2014.06.015. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276514005231.

A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*, 17(2):147–154, Feb. 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0690-6. URL http://www.nature.com/articles/s41592-019-0690-6.

S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, Dec. 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.11.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867414014974.

U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, July 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz369. URL https://doi.org/10.1093/nar/gkz369.

S. Roy, A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson, and R. Sridharan. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res*, 43(18):8694–8712, Oct. 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv865. URL https://academic.oup.com/nar/article/43/18/8694/2414464. Publisher: Oxford Academic.

A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, Sept. 2012. ISSN 1476-4687. doi: 10.1038/nature11279. URL https://www.nature.com/articles/nature11279. Number: 7414 Publisher: Nature Publishing Group.

S. Schoenfelder and P. Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet*, 20(8):437–455, Aug. 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0128-0. URL https://www.nature.com/articles/s41576-019-0128-0.

J. Schreiber, R. Singh, J. Bilmes, and W. S. Noble. A pitfall for machine learning methods aiming to predict across cell types. *Genome Biology*, 21(1):282, Nov. 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02177-y. URL https://doi.org/10.1186/s13059-020-02177-y.

R. Schwessinger, M. Gosden, D. Downes, R. C. Brown, A. M. Oudelaar, J. Telenius, Y. W. Teh, G. Lunter, and J. R. Hughes. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, pages 1–7, Oct. 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0960-3. URL https://www.nature.com/articles/s41592-020-0960-3. Publisher: Nature Publishing Group.

N. Y. A. Sey, B. Hu, W. Mah, H. Fauni, J. C. McAfee, P. Rajarajan, K. J. Brennand, S. Akbarian, and H. Won. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*, 23(4):583–593, Apr. 2020. ISSN 1546-1726. doi: 10.1038/s41593-020-0603-0. URL https://www.nature.com/articles/s41593-020-0603-0. Number: 4 Publisher: Nature Publishing Group.

N. C. Sheffield, R. E. Thurman, L. Song, A. Safi, J. A. Stamatoyannopoulos, B. Lenhard, G. E. Crawford, and T. S. Furey. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.*, 23(5):777–788, May 2013. ISSN 1549-5469. doi: 10.1101/gr.152140.112.

S. Singh, Y. Yang, B. Póczos, and J. Ma. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant Biol*, 7(2):122–137, June 2019. ISSN 2095-4697. doi: 10.1007/s40484-019-0154-0. URL https://doi.org/10.1007/s40484-019-0154-0.

A. Talukder, S. Saadat, X. Li, and H. Hu. EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction. *Bioinformatics*, 35(20):3877–3883, Oct. 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz641. URL https://academic.oup.com/bioinformatics/article/35/20/3877/5549495. Publisher: Oxford Academic.

R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sept. 2012. ISSN 1476-4687. doi: 10.1038/nature11232. URL https://www.nature.com/articles/nature11232. Number: 7414 Publisher: Nature Publishing Group.

M. Tsompana and M. J. Buck. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, 7(1):33, Nov. 2014. ISSN 1756-8935. doi: 10.1186/1756-8935-7-33. URL `https://doi.org/10.1186/1756-8935-7-33`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

D. Wang, A. Rendon, and L. Wernisch. Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Res*, 41(3):1450–1463, Feb. 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1339. URL `https://academic.oup.com/nar/article/41/3/1450/2903273`. Publisher: Oxford Academic.

S. Whalen, R. M. Truty, and K. S. Pollard. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, 48(5):488–496, May 2016. ISSN 1546-1718. doi: 10.1038/ng.3539. URL `https://www.nature.com/articles/ng.3539`.

Z. Wu, N. M. Ioannidis, and J. Zou. Predicting target genes of non-coding regulatory variants with IRT. *Bioinformatics*, 36 (16):4440–4448, Aug. 2020. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btaa254. URL `https://academic.oup.com/bioinformatics/article/36/16/4440/5824790`.

W. Xi and M. A. Beer. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLOS Computational Biology*, 14(12): e1006625, Dec. 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006625. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006625`. Publisher: Public Library of Science.

Y. Yang, R. Zhang, S. Singh, and J. Ma. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics*, 33(14):i252–i260, July 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx257. URL `https://academic.oup.com/bioinformatics/article/33/14/i252/3953972`. Publisher: Oxford Academic.

Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun*, 9(1):1–9, Feb. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03113-2. URL `https://www.nature.com/articles/s41467-018-03113-2`.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *arXiv:2012.07436 [cs]*, Dec. 2020. URL `http://arxiv.org/abs/2012.07436`. arXiv: 2012.07436.

Z. Zhuang, X. Shen, and W. Pan. A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data. *Bioinformatics*, 35(17):2899–2906, Sept. 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty1050. URL `https://academic.oup.com/bioinformatics/article/35/17/2899/5289332`.