frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 18 April 2023
DOI 10.3389/fgene.2023.1133775

OPEN ACCESS

EDITED BY
Chunyu Wang,
Harbin Institute of Technology, China

REVIEWED BY
Antonio Neme,
National Autonomous University of Mexico, Mexico
Ru Xiaoqing,
University of Tsukuba, Japan

*CORRESPONDENCE
Wen Zhu,
✉ syzhuwen@163.com

RECEIVED 29 December 2022
ACCEPTED 04 April 2023
PUBLISHED 18 April 2023

CITATION
Zheng L, Liu L, Zhu W, Ding Y and Wu F (2023), Predicting enhancer-promoter interaction based on epigenomic signals. *Front. Genet.* 14:1133775.
doi: 10.3389/fgene.2023.1133775

Predicting enhancer-promoter interaction based on epigenomic signals

Leqiong Zheng[1,2,3], Li Liu[2], Wen Zhu[1,3]*, Yijie Ding[3] and Fangxiang Wu[1]

[1]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [3]Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China

Introduction: The physical interactions between enhancers and promoters are often involved in gene transcriptional regulation. High tissue-specific enhancer-promoter interactions (EPIs) are responsible for the differential expression of genes. Experimental methods are time-consuming and labor-intensive in measuring EPIs. An alternative approach, machine learning, has been widely used to predict EPIs. However, most existing machine learning methods require a large number of functional genomic and epigenomic features as input, which limits the application to different cell lines.

Methods: In this paper, we developed a random forest model, HARD (H3K27ac, ATAC-seq, RAD21, and Distance), to predict EPI using only four types of features.

Results: Independent tests on a benchmark dataset showed that HARD outperforms other models with the fewest features.

Discussion: Our results revealed that chromatin accessibility and the binding of cohesin are important for cell-line-specific EPIs. Furthermore, we trained the HARD model in the GM12878 cell line and performed testing in the HeLa cell line. The cross-cell-lines prediction also performs well, suggesting it has the potential to be applied to other cell lines.

KEYWORDS

enhancer-promoter interaction, machine learning, ChIA-PET, random forest, epigenomic signals

1 Introduction

Enhancers and promoters are two of the most critical regulatory elements of gene transcription in the eukaryotic genome (Maston et al., 2006). The physical interactions between them precisely regulate spatiotemporal gene expression, which contributes to complex cellular functions. Aberrant connections between enhancers and promoters may lead to abnormal expression of disease-related genes (Krijger and De Laat, 2016). Therefore, the study of how enhancers and promoters interact can improve our understanding of health and disease. The primary mechanism of enhancer-promoter interaction is chromatin looping (Rubtsov et al., 2006; Miele and Dekker, 2008), which allows distal enhancers to contact the target gene promoters in three-dimensional space (Lv et al., 2021). Such long-range regulatory interactions play a significant role in tissue-specific gene expression (Maston et al., 2006; De Laat and Duboule, 2013) and can link the regulatory element to the target gene (Corradin et al., 2014). In recent decades, the identification of EPIs has relied on high-throughput experimental techniques, such as chromosome conformation capture (3C) (Dekker et al., 2002), 4C (Splinter et al., 2012), 5C (Sanyal et al., 2012), Hi-C (Lieberman-Aiden et al., 2009), Hi-C capture (Schoenfelder et al., 2015), DNase-Hi-C (Ma et al., 2015), and ChIA-PET (Li et al., 2012; Heidari et al., 2014). These experimental approaches are effective in identifying EPIs but are time-consuming and expensive (Ecker et al., 2012). Thus, a more cost-effective method is required for predicting enhancer-promoter interactions. To address this problem, machine learning methods are used to predict EPIs by using available genomic or epigenomic data.

Many deep learning methods have been proposed for predicting EPIs based on DNA sequence, including SPEID, SIMCNN, and EPIVAN. SPEID (Singh et al., 2019) and SIMCNN (Zhuang et al., 2019) employ CNN-based approaches, while EPIVAN (Hong et al., 2020) incorporates an attention mechanism for improved prediction accuracy. Although they achieved good results using only DNA sequences, the cell-line-specific nature of EPIs (Heidari et al., 2014; Ma et al., 2015) presents a challenge (Lv et al., 2021; Ao et al., 2022a). For instance, the same pair of enhancer and promoter contacts in some cell lines, but not in others, despite the DNA sequences have not changed (Schöler and Gruss, 1984). To address this issue, several models have been developed to identify cell-line-specific EPIs using epigenomic signals, including chromatin accessibility, the binding of special transcription factors, and histone modification levels. For example, RIPPLE (Roy et al., 2015) provides a systematic approach for predicting and interpreting

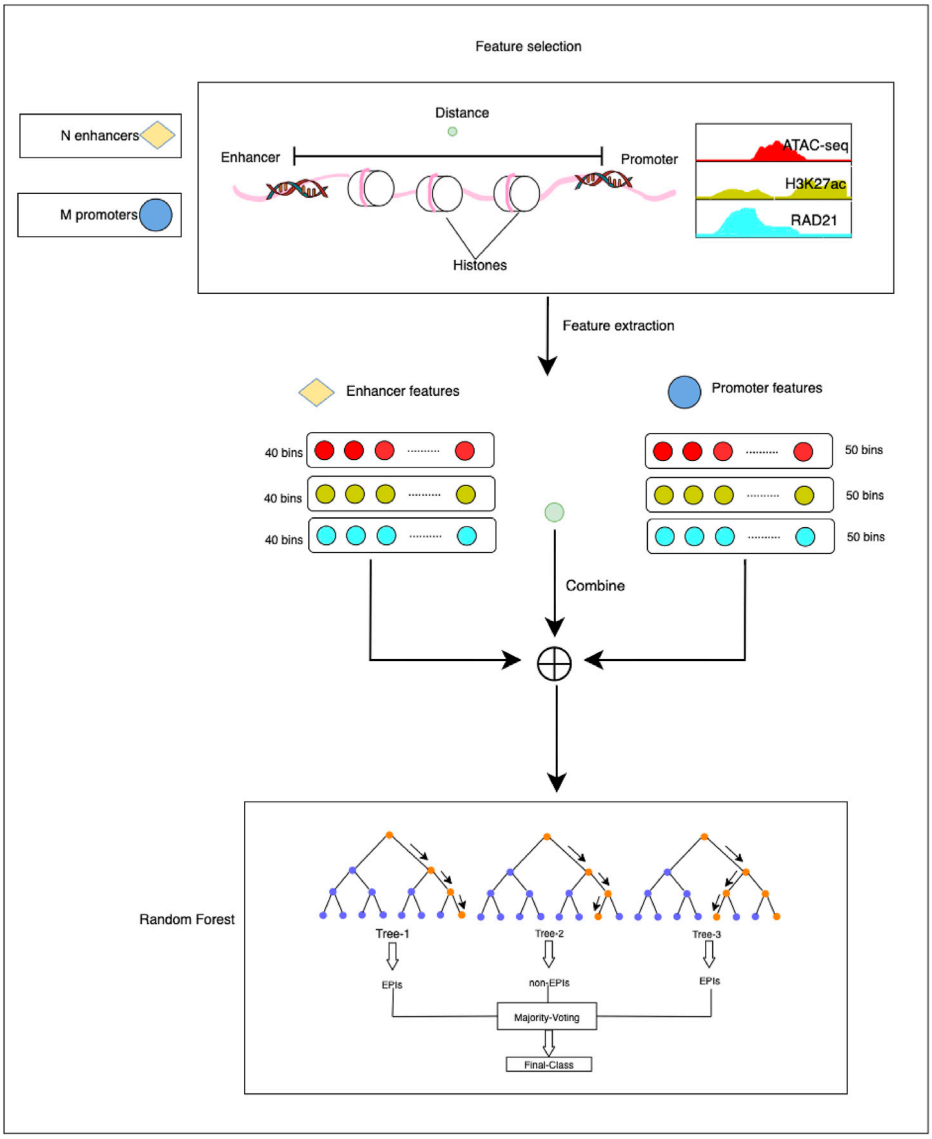

FIGURE 1
The overall framework of the HARD model. First, ATAC-seq, H3K27ac, and RAD21 epigenomic signals were selected as essential features to predict EPIs. Then, the enhancer and promoter were divided into 40 and 50 bins, respectively, with 50 bp per bin. Deeptools was used to extract the epigenomic signals. The epigenomic signal matrix was combined with the distance between the enhancer and promoter. Finally, we input the final feature matrix to the random forest learning machine for training and testing.

TABLE 1 Distribution of samples.

Data set	Positive samples	Negative samples
GM12878 training	6251	25,005
GM12878 test	1563	6251
Hela	347	1388

EPIs in a cell-line-specific manner using a variety of epigenomic features. However, many epigenomic signals are not available for all cell lines.

Based on the aforementioned analyses, we considered using as few epigenomic features as possible to build machine learning models to predict cell-line-specific EPIs. Loose chromatin is a prerequisite for loop formation. The H3K27ac ChIP-seq and ATAC-seq data are often used to represent chromatin accessibility. Chromatin interaction decays with distance. RAD21 is a subunit of cohesin that play important role in a loop formation. Therefore, the four types of features were extracted to train the models. By comparing several machine learning classifiers, the random forest was selected due to the high accuracy. Finally, we compared our HARD model with the sequence-based and other epigenomic features-based models. The results showed that our model outperformed them both in the same cell line and cross-cell-lines.

2 Materials and methods

The HARD model consists of three primary steps: 1) constructing positive and negative sets based on the benchmark database. 2) Extracting epigenomic features that can influence the formation of EPI. 3) predicting EPIs within the same cell line and across cell lines (Figure 1).

2.1 Data collection and processing

The enhancer-promoter interaction data were obtained from the BENGI (Moore et al., 2020) database. To construct a benchmark of enhancer-promoter interactions, BENGI integrated various experimental datasets, such as Hi-C, ChIA-PET, genetic interactions (cis-eQTLs), and CRISPR/Cas9 perturbations. After removing ambiguous pairs, we selected the RNAPII ChIAPET data of GM12878 and HeLa cell lines with a fixed positive and negative sample ratio. Both data have a positive-to-negative sample ratio of 1:4. To ensure the data is more accurate, the ambiguous interaction pairs were removed. The RNAPII ChIAPET data only provides the IDs of cCRE-ELS (cCREs with enhancer-like signatures) and TSS (transcription start site) without the position of cCRE-ELS and TSS. We located the cCRE-ELS and TSS in the genome according to the IDs of hg19-cCREs and GENCODEv19-TSS, respectively. Then, duplicate data was removed to retain unique data.

Next, 2,000 bp upstream and 500 bp downstream of the TSS were defined as the promoter region. For enhancers, upstream 1000 bp and downstream 1000 bp were extracted from the midpoint of the cCRE-ELS region. Ultimately, 39,070 pairs of enhancer-promoter interaction were obtained in the GM12878 dataset, and 1,735 pairs of enhancer-promoter interaction were obtained in the HeLa dataset. Then, the dataset was divided into a training set and a test set for the GM12878 sample. Specifically, 80% of the data was used for training, and the remaining 20% was used as an independent test set. To ensure consistency in data distribution across both datasets, the positive and negative sample ratios of both divided datasets were maintained at a 1:4 ratio. The above data processing part and the subsequent classification experiments were implemented in the python language environment, and the sklearn library is used. The detailed data distribution is shown in Table 1.

We selected three epigenomic signal features as our experimental features, including ATAC-seq, H3K27ac, and RAD21. The epigenomic signal data, which included ATAC-seq, H3K27ac, and RAD21, were obtained from the ENCODE (Ecker et al., 2012) database. The data with IDs ENCFF000XKM, ENCFF051PGV, and ENCFF706HLO corresponded to sequencing data in bigWig format of RAD21, ATAC-seq, and H3K27ac in the HeLa cell line, respectively. Similarly, the data with IDs ENCFF000WCT, ENCFF180ZAY, and ENCFF440GZA

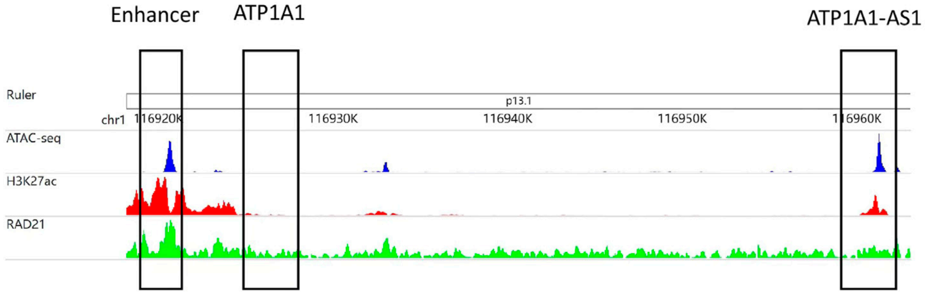

FIGURE 2
The area chr1:116,919,153–116,921,153 selected in the first matrix box is an enhancer subarea. The second matrix box selected region cr1: 116,924,718–116,927,218 is the promoter region of the ATP1A1 gene. The third matrix box selected region chr1:116,959,158–116,961,658 is the promoter region of the ATP1A1-AS1 gene. The three tracks in the figure were generated from the bigWig data of ATAC-seq, H3K27ac and RAD21 of GM12878 cell lines.

TABLE 2 Comparison of the predictive EPI performance of each classifier in the GM12878 cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
RF	**0.578**	**0.964**	**0.799**	**0.887**	**0.919**	**0.773**
Adaboost	0.555	0.947	0.725	0.869	0.881	0.688
GBDT	0.568	0.955	0.759	0.878	0.896	0.739

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

TABLE 3 Comparison of HARD, EPIVAN and RF (10) models in the GM12878 cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
HARD	0.578	0.964	**0.799**	**0.887**	**0.919**	**0.773**
EPIVAN	0.365	**0.971**	0.720	0.850	0.809	0.603
RF (10)	**0.709**	0.730	0.396	0.726	0.799	0.540

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

corresponded to sequencing data in bigWig format of RAD21, ATAC-seq, and H3K27ac in the GM12878 cell line, respectively.

2.2 Feature extraction

The above-mentioned features were extracted through the following steps. First, the genomic site data of EPIs and epigenomic signal data were imported into deeptools (Ramírez et al., 2014), a bioinformatics tool used for feature extraction. Then the enhancer and promoter regions were divided into bins of 50 bp. Each enhancer region was further divided into 40 bins, whereas each promoter region was divided into 50 bins. For ATAC-seq, H3K27ac, and RAD21, it generated a signal value for each bin. Following feature extraction, the enhancers and promoters were represented by 120-dimensional and 150-dimensional feature vectors, respectively. The distance is defined as the number of base pairs from the midpoint of the enhancer to the midpoint of the promoter. The epigenomic feature vector and distance feature vector were concatenated to obtain the final feature matrix. This step involved combining the feature vectors obtained from the enhancer and promoter regions into a single matrix, with each row of the matrix representing a pair of enhancer-promoter interactions. The final feature matrix was then used as input for the classification experiments.

2.3 Classification algorithms

We compared three classifiers, random forest (RF), AdaBoost, and gradient boosting decision tree (GBDT), for predicting EPIs in the GM12878 cell line, which is considered a binary classification problem. All three classifiers proved to be efficient in solving binary classification problems.

Random forest (Breiman, 2001) is an ensemble learning algorithm. It uses multiple decision trees to classify data by randomly selecting data and feature subsets, which helps to reduce the model's variance and overfitting risk. By voting or averaging the outputs of multiple decision trees, the model reduces the error rate and improves accuracy. In the experiment, a large amount of sample data was used, and setting the number of decision trees to 100 produced optimal performance.

AdaBoost (Schapire, 2013) assembles multiple weak classifiers to build a strong classifier, which applies to binary classification problems and has been shown to perform well on complex datasets. The algorithm assigns weights to each instance based on

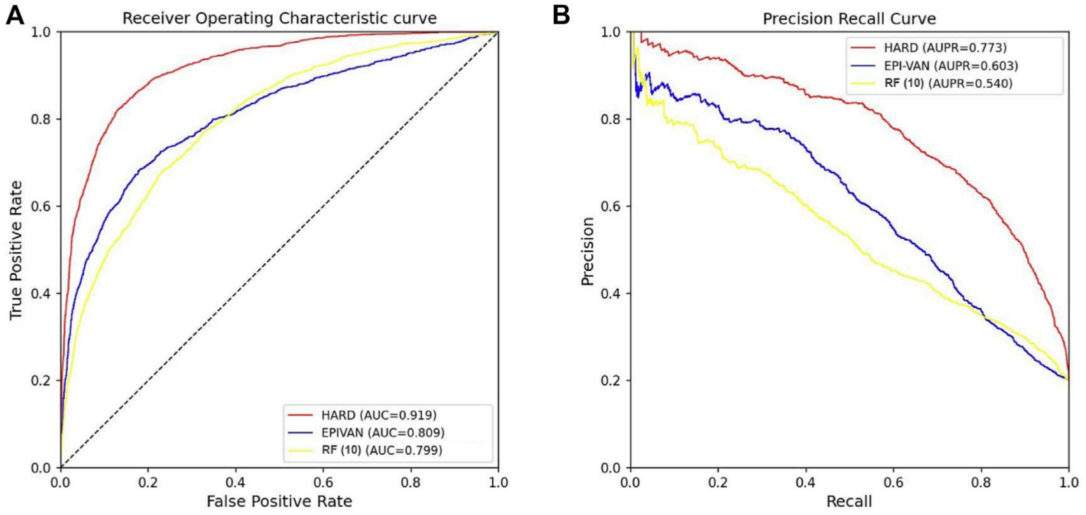

FIGURE 3
Comparison of the AUC and AUPRC performance of the three models tested independently in the GM12878 cell line. **(A)** The red curve is the ROC curve of the HARD model, the blue curve is the ROC curve of the EPIVAN model, and the yellow curve is the ROC curve of the RF (10) model; **(B)** The red curve is the PRC curve of the HARD model, the blue curve is the PRC curve of the EPIVAN model, and the yellow curve is the PRC curve of the RF (10) model.

TABLE 4 Comparison of HARD, EPIVAN and RF (10) models in the HeLa cell line.

Classifier	Sn	Sp	Precision	Acc	AUC	AUPRC
HARD	0.363	**0.953**	**0.660**	**0.836**	**0.831**	**0.601**
EPIVAN	**0.513**	0.890	0.539	0.815	0.795	0.564
RF (10)	0.144	0.949	0.402	0.786	0.572	0.296

The meaning of bold values is the highest value of a specific performance indicator under different classifiers.

its difficulty level and trains weak classifiers on the weighted data. Misclassified instances have increased weight, while correctly classified instances have decreased weight. This process is repeated multiple times until the ensemble classifier reaches a satisfactory level.

Gradient boosting decision tree (Friedman, 2001) builds a model by summing multiple decision trees. It optimizes the model iteratively by adding a new decision tree that reduces the prediction error of the previous trees. The model's accuracy improves with each iteration, making it suitable for binary classification problems. In the experiment, n_estimators, learning_rate, and subsample were set to 100, 0.1, and 1, respectively.

2.4 Performance evaluation

To evaluate the classification performance of the selected features and classifiers, we used six metrics: sensitivity (Sn) (Swift et al., 2020), specificity (Sp) (Swift et al., 2020), precision (Hong et al., 2020; Chen et al., 2023), accuracy (Shao et al., 2020; Yu et al., 2022), the area under the curve (AUC) (Myerson et al., 2001), and the area under the precision-recall curve (AUPRC) (Ozenne et al., 2015). These metrics serve as the basis for evaluation, and the relevant formulas for their calculation are shown below.

$$Sn = recall = TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

In binary classification, there are four possible outcomes: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP corresponds to the cases where the classifier correctly predicts the positive class, while FP corresponds to the instances where the classifier incorrectly predicts the positive class. Similarly, FN refers to the cases where the classifier incorrectly predicts the negative class, and TN refers to the instances where the classifier correctly predicts the negative class. Additionally, TPR (sensitivity/recall) is the ratio of correctly identified positive instances to the actual positive instances, while FPR is the proportion of falsely identified positive instances to the actual negative instances (Zeng et al., 2020). AUC is calculated by plotting TPR against FPR at different thresholds and represents the area under the resulting curve. AUPRC is calculated by plotting precision against recall at different thresholds and represents the area under the resulting curve.

3 Results and discussion

3.1 The features of HARD model are closely related with EPI

The accessibility of chromatin structural regions is associated with the regulation of gene expression. ATAC-seq is commonly

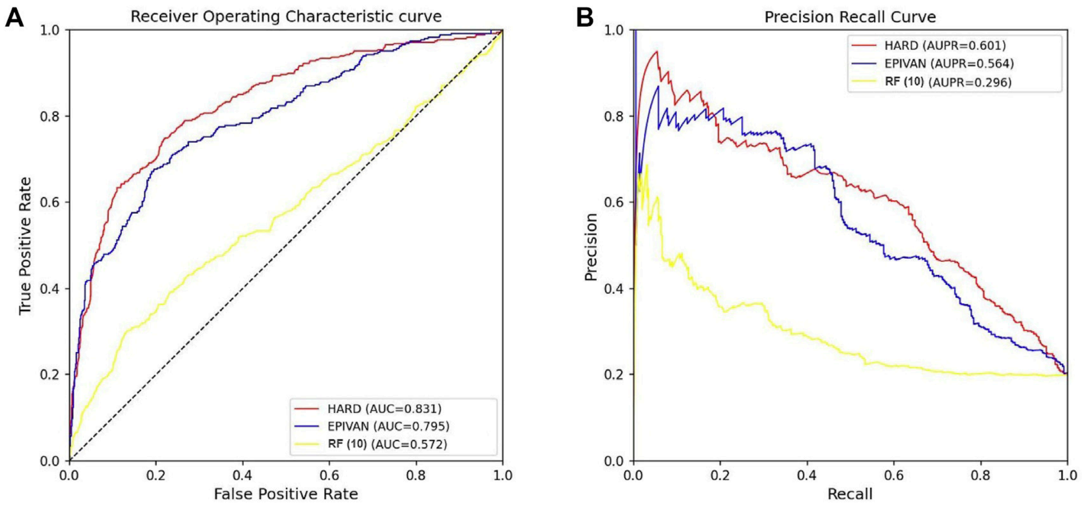

FIGURE 4
Comparison of AUC and AUPRC performance of the three models in the HeLa cell line. **(A)** The red curve is the ROC curve of the HARD model, the blue curve is the ROC curve of the EPIVAN model, and the yellow curve is the ROC curve of the RF (10) model; **(B)** The red curve is the PRC curve of the HARD model, the blue curve is the PRC curve of the EPIVAN model, and the yellow curve is the PRC curve of the RF (10) model.

used to detect open regions of chromatin across the genome. When combined with activated histone modification, such as H3K27ac, ATAC-seq can enable the identification of specific effects on gene expression (Bravo González-Blas et al., 2019). H3K27ac is primarily enriched in enhancer and promoter regions (Herrera-Uribe et al., 2020) and is associated with gene activation (Yan et al., 2019). RAD21 and the insulator-binding protein CTCF bind to highly conserved promoters and distal enhancers, contributing to transcriptional regulation (Whalen et al., 2016; Liu et al., 2021). Numerous studies have shown that distance is a useful factor for studying EPI (Bianco et al., 2018; Al Bkhetan et al., 2019). The distance feature has an essential contribution to many models (Moore et al., 2020; Ao et al., 2022b).

Figure 2 is an example that epigenomic modification influences the formation of EPI. The enhancer region (chr1: 116,919,153–116,921,153) interacts with the ATP1A1-AS1 promoter (chr1:116,959,158–116,961,658) and does not interact with the ATP1A1 promoter (chr1: 116,959,158–116,961,658), according to RNAPII ChIAPET data of GM12878 cell line. In the enhancer region, the signals of ATAC-seq, H3K27ac, and RAD21 are enriched, which indicates that the enhancer is highly activated. The promoter region of ATP1A1-AS1 is enriched in ATAC-seq, H3K27ac modifications, and RAD21 binding, whereas the promoter region of ATP1A1 is not.

3.2 Comparison and selection of classifiers

To select the most accurate classifier, we compared three classifiers, AdaBoost, GBDT, and RF. We trained the model using 31,256 GM12878 samples with ten-fold cross-validation and evaluated its performance on an independent test set of 7,814 GM12878 samples. The classifiers were trained and tested separately, and their performance was compared using different metrics. A comparison of the metrics of the test set is shown in Table 2. Results showed that the RF algorithm outperformed both GBDT and AdaBoost in all metrics. Specifically, the RF algorithm demonstrated higher Sn, Sp, precision, accuracy, AUC, and AUPRC values, at 0.578, 0.964, 0.799, 0.887, 0.919, and 0.773, respectively. Notably, the RF algorithm displayed superior performance in AUPRC and precision metrics. The RF algorithm merges the strengths of ensemble learning and tree models, and it is capable of balancing the error for an unbalanced set of classifiers, making it a suitable choice for the dataset at hand. Consequently, the HARD model was constructed using the RF algorithm.

3.3 Comparison with other models in GM12878 cell line

In order to verify the validity of the HARD model, we next compared the performance of HARD against the sequence-based and other epigenomic features-based models. EPIVAN is a typical representative of sequence-based models, which outperforms the majority of existing models. RIPPLE utilizes many epigenomic features to predict EPI. These epigenomic features include cohesin (RAD21), architectural proteins (CTCF), marks associated with active gene bodies and elongation (H3K36me3, H4K20me1), activating marks of transcription (H3K4me2, H3K27ac, and H3K9ac), open chromatin (DNase I), a repressive mark (H3K27me3), and a general transcription factor (TBP). Here, we used ten available features of RIPPLE to conduct a RF classification model, named RF (10). Then the HARD model was compared with RF (10) and EPIVAN in multiple aspects. We trained the models using 31,256 GM12878 sample data with ten-fold cross-validation and evaluated them using an independent test set of 7,814 GM12878 samples. The comparison results are shown in Table 3. The results indicated that RF (10) performed best in terms of Sn, while EPIVAN produced the best results for Sp. However, each model has its strengths and weaknesses in terms of Sn and Sp. HARD had shown significant improvement in all four performance metrics compared to other models. Specifically, compared to EPIVAN, HARD shows an improvement of 7.9% and 3.7% in precision and Acc, respectively, as well as an increase of 11% and 17% in AUC and AUPRC, respectively. Compared to the RF (10), HARD shows greater improvements, with increases of 40.3%, 16.1%, 12%, and 23.3% in precision, Acc, AUC, and AUPRC, respectively. The comparison of the AUC and ROC curves of the three models is shown in Figure 3.

3.4 Comparison of the HARD, EPIVAN and RF (10) model in cross-cell-lines

To verify the robustness of the models, we conducted a cross-cell-line analysis by training the models on the GM12878 cell line and testing them on the HeLa cell line. We used 39,070 GM12878 samples as the training set for ten-fold cross-validation, and 1,735 HeLa samples as the test set for evaluation. Experiments were implemented for the HARD, EPIVAN and RF (10) models, respectively. Among the three models, HARD achieved the best performance in terms of Sp, precision, accuracy, AUC, and AUPRC, followed by EPIVAN, with RF (10) showing the worst performance. In comparison to EPIVAN, the HARD model slightly improves five metrics, only lower than EPIVAN in Sn. The HARD model outperforms RF (10) by a significant margin (Table 4). The comparison of the AUC and ROC curves of the three models is shown in Figure 4. Results indicated that HARD outperformed EPIVAN and RF (10) in predicting EPIs in cross-cell-lines.

4 Conclusion

The interaction between enhancer and promoter is a complex process. Various genomic and epigenomic features are related to EPI. Many machine learning models have been developed to predict EPI based on a large number of genomic and epigenomic features. The redundancy of features leads to unsatisfactory experimental results and limits the application to more cell lines. In this paper, we developed the HARD model, which employed a minimal number of epigenomic features to predict cell-line-specific EPIs. It is noteworthy that the HARD model is based on benchmark data from the BENGI database, which defined EPI strictly by integrating ChIA-PET, genetic interactions (cis-eQTLs), and CRISPR/Cas9 perturbations. By comparing with two other models, we found HARD outperformed them both in the same cell line and cross-cell-lines. Importantly, our model only used H3K27ac, ATAC-seq, RAD21, and Distance as input, which makes it possible to apply to more cell lines.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LZ: Investigation, Methodology, Writing—Original draft preparation. LL: Conceptualization, Funding acquisition, Writing—Review & Editing. WZ: Conceptualization, Project administration, Funding acquisition. YD: Methodology, Writing—Review & Editing. FW: Investigation, Methodology.

Funding

This work was supported by the National Nature Science Foundation of China (Grant Nos 61863010, 11926205, 11926412, 61873076, and 61961031), National Key R and D Program of China (No. 2020YFB2104400), Natural Science Foundation of Hainan, China (Grant Nos 121RC538, 119MS036, and 120RC588), Key Laboratory of Computational Science and Application of Hainan Province (No. JSKX202201), and the Municipal Government of Quzhou (NO. 2022D017).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al Bkhetan, Z., Kadlof, M., Kraft, A., and Plewczyński, D. (2019). Machine learning polymer models of three-dimensional chromatin organization in human lymphoblastoid cells. *Methods* 166, 83–90. doi:10.1016/j.ymeth.2019.03.002

Ao, C., Jiao, S., Wang, Y., Yu, L., and Zou, Q. (2022a). Biological sequence classification: A review on data and general methods. *Research* 24, 1198. doi:10.1093/bioinformatics/btn089

Ao, C., Zou, Q., and Yu, L. (2022b). NmRF: Identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Briefings Bioinforma.* 23, bbab480. doi:10.1093/bib/bbab480

Bianco, S., Lupiáñez, D. G., Chiariello, A. M., Annunziatella, C., Kraft, K., Schöpflin, R., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50, 662–667. doi:10.1038/s41588-018-0098-8

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., et al. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. methods* 16, 397–400. doi:10.1038/s41592-019-0367-1

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, L., Yu, L., and Gao, L. (2023). Potent antibiotic design via guided search from antibacterial activity evaluations. *Bioinformatics* 39, btad059. doi:10.1093/bioinformatics/btad059

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal, R., et al. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* 24, 1–13. doi:10.1101/gr.164079.113

De Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* 502, 499–506. doi:10.1038/nature12753

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science* 295, 1306–1311. doi:10.1126/science.1067799

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y., and Segal, E. (2012). Genomics: ENCODE explained. *Nature* 489, 52–55. doi:10.1038/489052a

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. statistics* 29, 1189–1232. doi:10.1214/aos/1013203451

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., et al. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res.* 24, 1905–1917. doi:10.1101/gr.176586.114

Herrera-Uribe, J., Liu, H., Byrne, K. A., Bond, Z. F., Loving, C. L., and Tuggle, C. K. (2020). Changes in H3K27ac at gene regulatory regions in porcine alveolar macrophages following LPS or PolyIC exposure. *Front. Genet.* 11, 817. doi:10.3389/fgene.2020.00817

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinforma. Oxf. Engl.* 36, 1037–1043. doi:10.1093/bioinformatics/btz694

Krijger, P. H. L., and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782. doi:10.1038/nrm.2016.138

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98. doi:10.1016/j.cell.2011.12.014

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* 326, 289–293. doi:10.1126/science.1181369

Liu, L., Zhang, L.-R., Dao, F.-Y., Yang, Y.-C., and Lin, H. (2021). A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation. *Mol. Therapy-Nucleic Acids* 23, 347–354. doi:10.1016/j.omtn.2020.11.011

Lv, H., Dao, F.-Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., et al. (2021). A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Briefings Bioinforma.* 22, bbab031. doi:10.1093/bib/bbab031

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. methods* 12, 71–78. doi:10.1038/nmeth.3205

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* 4, 1046–1057. doi:10.1039/b803580f

Moore, J. E., Pratt, H. E., Purcaro, M. J., and Weng, Z. (2020). A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* 21, 17–16. doi:10.1186/s13059-019-1924-8

Myerson, J., Green, L., and Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *J. Exp. analysis Behav.* 76, 235–243. doi:10.1901/jeab.2001.76-235

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* 68, 855–859. doi:10.1016/j.jclinepi.2015.02.010

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids Res.* 42, W187–W191. doi:10.1093/nar/gku365

Roy, S., Siahpirani, A. F., Chasman, D., Knaack, S., Ay, F., Stewart, R., et al. (2015). A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic acids Res.* 43, 8694–8712. doi:10.1093/nar/gkv865

Rubtsov, M. A., Polikanov, Y. S., Bondarenko, V. A., Wang, Y.-H., and Studitsky, V. M. (2006). Chromatin structure can strongly facilitate enhancer action over a distance. *Proc. Natl. Acad. Sci.* 103, 17690–17695. doi:10.1073/pnas.0603819103

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* 489, 109–113. doi:10.1038/nature11279

Schapire, R. E. (2013). "Explaining adaboost," in *Empirical inference: Festschrift in honor of vladimir N. Vapnik* (Berlin, Germany: Springer), 37–52.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* 25, 582–597. doi:10.1101/gr.185272.114

Schöler, H. R., and Gruss, P. (1984). Specific interaction between enhancer-containing molecules and cellular components. *Cell* 36, 403–411. doi:10.1016/0092-8674(84)90233-2

Shao, J., Yan, K., and Liu, B. (2020). FoldRec-C2C: Protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Briefings Bioinforma.* 22, bbaa144. doi:10.1093/bib/bbaa144

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

Splinter, E., Wit, E. D., Werken, H., Klous, P., and Laat, W. D. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* 58, 221–230. doi:10.1016/j.ymeth.2012.04.009

Swift, A., Heale, R., and Twycross, A. (2020). What are sensitivity and specificity? *Evidence-Based Nurs.* 23, 2–4. doi:10.1136/ebnurs-2019-103225

Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48, 488–496. doi:10.1038/ng.3539

Yan, W., Chen, D., Schumacher, J., Durantini, D., Engelhorn, J., Chen, M., et al. (2019). Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat. Commun.* 10, 1705–1716. doi:10.1038/s41467-019-09513-2

Yu, L., Zheng, Y., and Gao, L. (2022). MiRNA–disease association prediction based on meta-paths. *Briefings Bioinforma.* 23, bbab571. doi:10.1093/bib/bbab571

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi:10.1039/c9sc04336e

Zhuang, Z., Shen, X., and Pan, W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics* 35, 2899–2906. doi:10.1093/bioinformatics/bty1050