OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae651

https://doi.org/10.1093/bib/bbae651

Review

A review of deep learning models for the prediction of chromatin interactions with DNA and epigenomic profiles

Yunlong Wang[1], Siyuan Kong[1], Cong Zhou[2,3,4], Yanfang Wang[5,*], Yubo Zhang[1,6,*], Yaping Fang[2,3,4,*], Guoliang Li[2,3,4,*]

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, No. 97 Buxin Road, Dapeng New District, Shenzhen 518120, China
[2]Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, No. 1 Shizishan Street, Hongshan District, Wuhan 430070, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, 3D Genomics Research Center, No. 1 Shizishan Street, Hongshan District, Wuhan 430070, China
[4]College of Informatics, Huazhong Agricultural University, No. 1 Shizishan Street, Hongshan District, Wuhan 430070, China
[5]State Key Laboratory of Animal Biotech Breeding, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), No. 2 West Yuanmingyuan Rd, Haidian District, Beijing 100193, China
[6]Sequencing Facility, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD 21701, United States

*Corresponding authors. Yanfang Wang, State Key Laboratory of Animal Biotech Breeding, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing 100193, China. E-mail: wangyanfang@caas.cn; Yubo Zhang, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. E-mail: ribon_001@163.com; Yaping Fang, Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, China. E-mail: ypfang@mail.hzau.edu.cn; Guoliang Li, Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, China. E-mail: guoliang.li@mail.hzau.edu.cn

Abstract

Advances in three-dimensional (3D) genomics have revealed the spatial characteristics of chromatin interactions in gene expression regulation, which is crucial for understanding molecular mechanisms in biological processes. High-throughput technologies like ChIA-PET, Hi-C, and their derivatives methods have greatly enhanced our knowledge of 3D chromatin architecture. However, the chromatin interaction mechanisms remain largely unexplored. Deep learning, with its powerful feature extraction and pattern recognition capabilities, offers a promising approach for integrating multi-omics data, to build accurate predictive models of chromatin interaction matrices. This review systematically summarizes recent advances in chromatin interaction matrix prediction models. By integrating DNA sequences and epigenetic signals, we investigate the latest developments in these methods. This article details various models, focusing on how one-dimensional (1D) information transforms into the 3D structure chromatin interactions, and how the integration of different deep learning modules specifically affects model accuracy. Additionally, we discuss the critical role of DNA sequence information and epigenetic markers in shaping 3D genome interaction patterns. Finally, this review addresses the challenges in predicting chromatin interaction matrices, in order to improve the precise mapping of chromatin interaction matrices and DNA sequence, and supporting the transformation and theoretical development of 3D genomics across biological systems.

Keywords: 3D genomics; epigenetics; deep learning; multi-omics; chromatin interactions

Introduction

Three-dimensional genomics and chromatin interactions

In eukaryotic organisms such as humans and mice, astonishingly, DNA in a single cell exceeds 2 m in length, yet compacts to fit a nucleus just 10 μm in diameter without losing biological activity. This compaction reveals the genome's intricate three-dimensional (3D) structure within the nucleus [1] (Fig. 1). Highly folded DNA, together with histones, transcription factors, and specific RNA molecules, forms a dynamic and complex chromatin structure [2]. This chromatin structure, with nucleosomes as basic units, regulates life process and coordinates critical biological processes like DNA replication, transcription, mutation, and damage repair [3, 4].

With development in 3D genome sequencing, methods to explore chromatin structure have evolved from the early Fluorescent In Situ Hybridization (FISH) to capture genome like Hi-C [5] and ChIA-PET [6]. These technologies capture and document the complex chromatin interactions. Chromatin interaction is a central to 3D genomics. Chromatin interaction matrices, based on these interactions [7], record the spatial patterns and provide a strong data foundation for analyzing the genome structures [8–10]. These higher-order structures mainly include chromosomal territories, chromatin A/B compartments, TADs and their sub-domains, as well as chromatin loops (Fig. 1A and C). Chromatin A/B compartments are closely related to chromatin activity [11, 12], spanning hundreds of kilobases to several megabases. TADs, as relatively stable structural

Received: September 23, 2024. **Revised:** October 29, 2024. **Accepted:** December 3, 2024

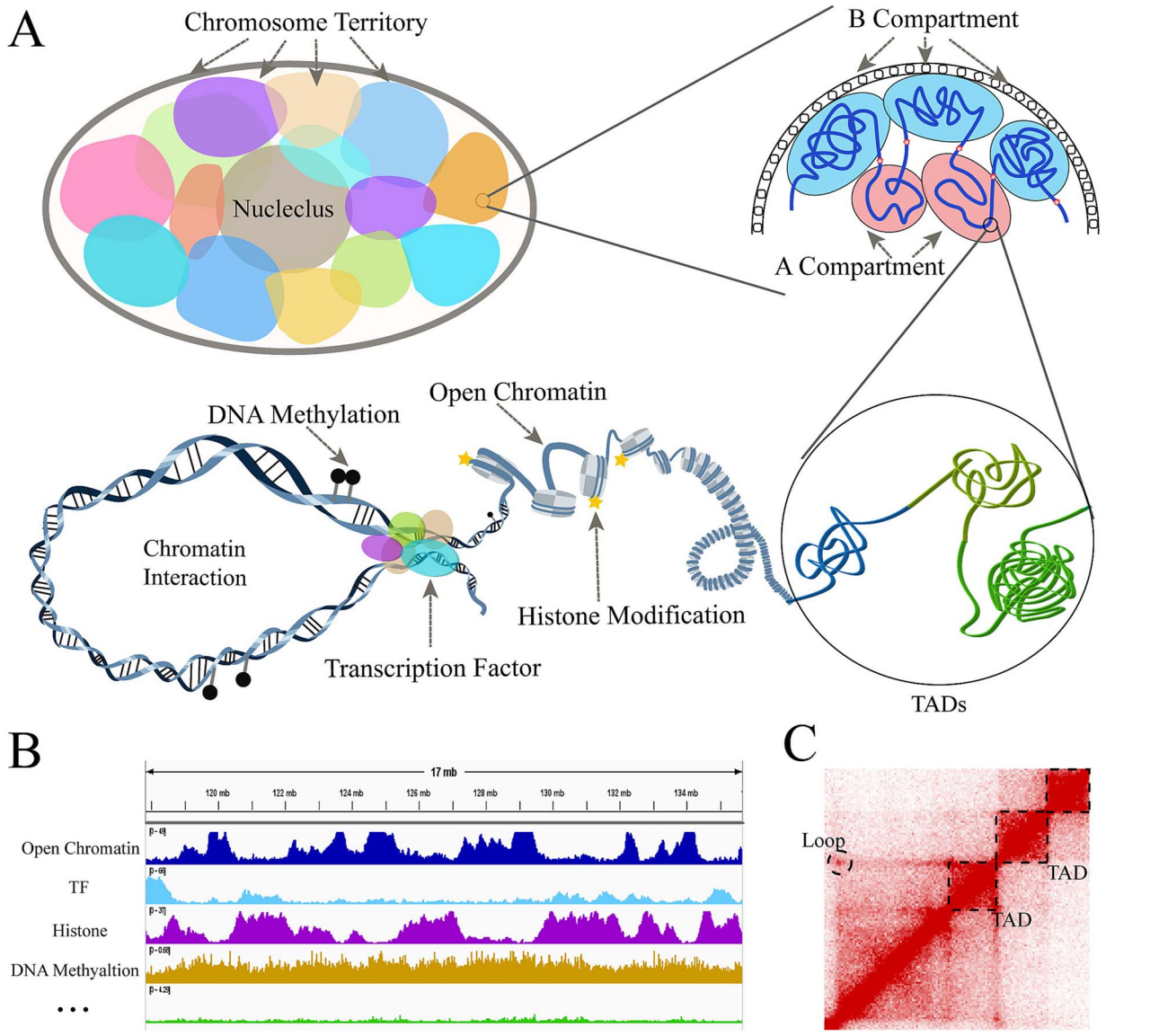

Figure 1. The 3D structure of chromatin interactions. (A) Illustration of the hierarchical structure of chromosomal territories, chromatin a/B compartments, TADs, and chromatin interaction (CI). (B) Characterization patterns of epigenetic modifications related to the 3D structure of the genome. The distribution of epigenetic information, such as open chromatin regions, histone modification, DNA methylation, and transcription factor binding site, is utilized for predicting and analyzing the chromatin interaction matrix. (C) Within the cell nucleus, chromosomes are captured by high-throughput chromatin conformation techniques such as hi-C, to obtain a chromatin interaction matrix. The dashed circle denotes chromatin loops, and the dashed square represents TADs.

units within the chromatin interaction matrix, typically span from hundreds of thousands to millions of base pairs [11, 13]. Disruption of TAD structures can dysregulate gene expression, potentially leading to diseases like cancer [14]. Chromatin loops (CLs), local regions with high-frequency chromatin interactions [15], span hundreds of thousands to millions of base pairs, and are crucial for observing interactions between genes and transcription regulatory elements, promoters and enhancers, directly impacting gene expression [14, 16–18].

Notably, a close relationship exists between chromatin's 3D structure and its epigenetic information (Fig. 1B). This connection is key to cell specificity and essential for understanding gene expression regulation and the normal organismal development [19–22]. Key epigenetic mechanisms, such as histone modifications and DNA methylation, regulate 3D chromatin structure, enabling precise gene expression control [23]. Additionally, higher-order chromatin organization is closely associated with specific epigenetic markers [24–28]. For example, in *Arabidopsis*, A Compartment is enriched with active H3K4me3 marks, while B Compartment is rich in repressive H3K27me3 marks [12, 29]. This distribution pattern directly reflects the close link between chromatin structure and biological function. Similarly, in mammalian cells, key proteins such as CTCFCCCTC-binding factor (CTCF) and cohesin are crucial for forming and maintaining 3D genome structure [30, 31], and the binding sites often align with the epigenetic marker patterns.

According to the principle that "structure dictates function" [15], advancements in 3D genomics have deepened our understanding of how genomic architecture relates to function. Projects like ENCODE, which integrate genetic and epigenetic information, have illuminated key pathways in gene expression regulation and the molecular basis of diseases.

Application of deep learning in the genomics and 3D genomics

Deep learning has shown great potential in genomics and 3D genomics research [32, 33]. For instance, The DeepNull [34] model enhances phenotypic prediction and association analysis of structural variations in cancer genomes, by simulating nonlinear covariate effects. Additionally, the DeepCpG [35] and DeepHistone [36] models accurately predict DNA methylation and histone modifications, both crucial for understanding 3D genome structure. Furthermore, the scDEAL model [37], using transfer learning, integrates bulk and single-cell RNA-seq data to predict drug sensitivity at the single-cell level, opening new avenues for personalized medicine.

To resolve the 3D genome structure efficiently and accurately, high-resolution chromatin interaction reinforcement learning methods have been developed, including HiCPlus (convolutional neural networks [CNNs]) [38], hicGAN (generative learning models) [39], and CAESAR (graph neural networks) [40]. However, these methods still rely on Hi-C data input. Conversely, some research focuses on using DNA sequences or epigenetic information, to predict the genomes's hierarchical structure and chromatin interaction patterns [41–43]. Specifically, models based on DNA sequence information or epigenetic signals, include CoRNN for A/B compartments prediction [44], TAD-Lactuca [45], and pTADS [46] for predicting TADs and their boundaries. ChINN [47], CharID [48], and 3DepiLoop [49] for predicting chromatin loops or interactions in open chromatin regions. Additionally, the EpiMCI model enables multi-way chromatin interaction prediction [50].

Although existing research has summaries specific 3D structures, such as chromatin loops and TADs [8, 10, 41, 42], a systematic review of the core element in 3D genomics—the chromatin interaction matrix—is still lacking. The chromatin interaction matrix provides a global view, clarify the complexity of chromatin point-to-point interactions. Thus, exploring innovative deep learning applications in chromatin interaction matrices could revolutionize genomics, epigenetics, and related biomedical fields, with profound scientific and practical implications.

Overview of the review

This article systematically reviews recent methods for predicting chromatin interaction matrices, which based on DNA sequence and epigenetic information. It explores how the deep learning components in these models work together, and how DNA sequence features and epigenetic markers jointly shape the 3D genome's dynamic structure, focusing on transforming one-dimensional (1D) information into the 3D chromatin interactions. This review also discusses current challenges in predicting chromatin interaction matrices, such as model generalization and biological interpretability, and suggests potential optimization paths for future research. The goal is to improve the precision and practicality of chromatin interaction matrix prediction models, providing new perspectives and powerful tools for genomic research and related fields.

Key modules of chromatin interaction matrices prediction models

High-throughput 3D genome sequencing have generated a vast chromatin interaction data across species, cell lines, and tissue types [51–53]. Traditional analysis methods are becoming inadequate in handling the vast genomic data. Studies in 3D genomics and epigenetics require more efficient and intelligent processing techniques. Consequently, machine learning and deep learning applications have become widespread [54–56]. These advanced algorithms can uncover the potential regulatory mechanisms and patterns in gene expression changes. Table 1 present an overview of the main prediction models for analyzing 3D genome data and predicting chromatin interaction Matrices. The code sources of the models are available in Table 2.

Feature extraction and encoding

Feature encoding for DNA sequence

DNA sequence encoding underpins chromatin interaction prediction, crucial for locating transcription factor binding sites and spatial proximity chromatin positions [67–69]. One-hot encoding represents each base as one of four feature vectors: [1,0,0,0], [0,1,0,0], [0,0,1,0], or [0,0,0,1], preserving sequence integrity and facilitating analysis of base variations and specific motif impacts [70]. This method supports most DNA sequence-based chromatin interaction prediction task [71]. Besides one-hot encoding, GC content [66, 72] and transcription factor binding site information (Motif score) [63] are also widely used in chromatin interaction matrix predictions (Fig. 2A).

Epigenetic information encoding

Epigenetic modification data (such as transcription factor binding profile, histone profile, chromatin openness, etc.) combined with DNA sequences to jointly gene expression, chromatin spatial organization, and cell fate [73–76]. Encoding methods often use the average distribution of epigenetic modification signals (e.g. Reads per million (RPM) and Reads Per Kilobase Million (RPKM)). Additionally, based on the extension of the distribution of epigenetic modifications, correlating enriched epigenetic signals across genomic intervals is another encoding approach for epigenetic data [63] (Fig. 2B).

Feature fusion strategy

DNA sequence features provide fundamental genetic information, indicating potential transcription factor binding sites and gene regulatory sequences, while epigenetic features reflect the genome's regulatory state and functional activity. Combining these features enriches contextual information, enhances the model generalization, and reveals complex regulatory relationships, improving prediction accuracy [77, 78]. The multi-dimensional feature fusion of DNA sequences and epigenetic information discussed in this review mainly includes two methods.

The first method focuses on the initial integration of features, combining DNA sequence features with epigenetic features to create an integrated feature vector using concatenation or stacking techniques. This integrated feature vector then serves as input data for deep learning model to analyze and extract key information related to chromatin interactions patterns [64, 66].

The second method emphasizes the initial extraction of features, processing DNA sequences and epigenetic data with deep learning techniques to extracte core features from each input. Subsequently, concatenation or stacking integrates the signals extracted from different deep learning models. For example, multi-layer perceptions can be used as feature extractors to process the two types of input features separately, and merge the outputs of each model in subsequent stages, enabling joint learning of interaction or shared layer features [63, 65].

Model architecture

Figure 2 illustrates the general model framework and computational process for predicting chromatin interaction matrices.

Table 1. Models for predicting chromatin interaction matrices based on DNA sequence and epigenetic information in the past 5 years.

Feature Classification	Input Features	Model	Dimension Transformation Method	Bulk data	Cite
DNA	DNA sequence	Akita	1D to 2D	Hi-C	*Nat Methods*, (2020) [57]
	DNA sequence	DeepC	zig-zag stripe	Hi-C	*Nat Methods* (2020) [58]
	DNA sequence	Orca	1D to 2D	Micro-C	*Nat Genet*, (2022) [59]
	DNA sequence	HiCDiffusion	1D to 2D	Hi-C	bioRxiv,2024 [60]
Epigenetic related	Histones+TFs + DNase I	HiC-Reg	anchor regions	Hi-C	*Nat Commun*, (2019) [61]
	Histones+CTCF+DNase I	Epiphany	zig-zag stripe	Hi-C	*Genome Biol*, (2023) [62]
DNA+ Epigenetic related	Motif+scATAC	ChromaFold	V-stripe	Hi-C	bioRxiv,2023 [63]
	DNA + ATAC	EPCOT-COP	1D to 2D	Hi-C,Micro-C, ChIA-PET	*Nucleic Acids Res*, (2023) [64]
	DNA + ATAC+ CTCF	C.Origami	1D to 2D	Hi-C	*Nat Biotechnol*, (2023) [65]
	Motif+GC + TF+ Histones	ChIPr	anchor regions	ChIA-PET	*Genome Biol*, (2024) [66]

Table 2. The code resources of the model.

Model	Link
Akita	https://github.com/calico/basenji/tree/master/manuscripts/akita
DeepC	https://github.com/rschwess/deepC
Orca	https://github.com/jzhoulab/orca
HiCDiffusion	https://github.com/SFGLab/HiCDiffusion
HiC-Reg	https://github.com/Roy-lab/HiC-Reg
Epiphany	https://github.com/arnavmdas/epiphany
ChromaFold	https://github.com/viannegao/ChromaFold/tree/main
EPCOT-COP	https://github.com/liu-bioinfo-lab/EPCOT
C.Origami	https://github.com/tanjimin/C.Origami
ChIPr	https://git.biohpc.swmed.edu/s206442/chipr

The core framework consists of two modules: the Encoder and the Decoder (Fig. 2C). The Encoder extracts structural features related to chromatin interactions from 1D feature information. This section includes five commonly used modules, such as CNNs and attention mechanisms. In the Decoder section, we present three commonly used methods that transform 1D information into two-dimensional (2D) information. Ultimately, using actual chromatin interaction data, the model computes the loss function, updates its weights, and outputs predictions of chromatin interactions.

"CNNs" are a feature extraction models that uses multiple layers of nonlinear processing units for feature learning and pattern recognition. They are suitable for static data with grid structures and can effectively extract spatial features. In deep learning models for predicting chromatin interaction matrices, they are suitable for capturing features of proximal chromatin interaction features [79].

"Long Short-Term Memory Networks (LSTMs)," including their bidirectional counterparts (Bi-LSTMs), are specialized recurrent neural networks designed for processing sequence and dynamic data, enabling them to encode the order, context, and long-range dependencies. In deep learning models designed to predict chromatin interaction matrices, LSTM/Bi-LSTM networks are particularly effective at elucidating the interaction patterns between genomic regions and their flanking sequences [80].

"Transformer" is a model architecture widely used in the field of natural language processing [81], leverages multiple layers of self-attention module and Multi-head attention. Self-attention modules, which include Query (Q), Key (K), and Value (V) components, generate contextually rich representations. Multi-head attention is an extension of the self-attention, enhances model's ability to extract features and operate with high parallelism. This design enables the transformer to effectively handle long-distance dependencies and large-scale parallel computing [82].

"Generative Adversarial Networks (GANs)" are powerful framework for unsupervised learning used to generate new data samples. They typically operate through the competitive interaction between two components: a generator and a discriminator, which together produce high-quality data. In predicting chromatin interaction matrices, the ultimate goal is to generate matrices indistinguishable from actual data by the discriminator [83].

"Dilated Convolutions" are a special type of convolutional operation. Compared to regular convolutions, they increase the receptive field while maintaining resolution by inserting dilated spaces (or skipping certain pixels/ features) between the convolution kernels. This allows them to extract a broader range of contextual information without increasing the number of parameters [84].

"Complex structures including cascade structures, transfer learning, and diffusion processes." Cascade structures employ a hierarchical multi-resolution encoder design to explore genomic interaction matrix information at multiple scales [59] (Fig. 2D). As the input scale increases and the resolution becomes coarser, the broader visual angle enhances the prediction of long-range chromatin interactions.

Transfer learning allows a model to apply knowledge learned from one task (source task) to another related but different task (target task). Typically, a model is trained on a data-rich task and then fine-tuned on the target task to adapt to the specific needs of the less data-abundant target task [58] (Fig. 2E).

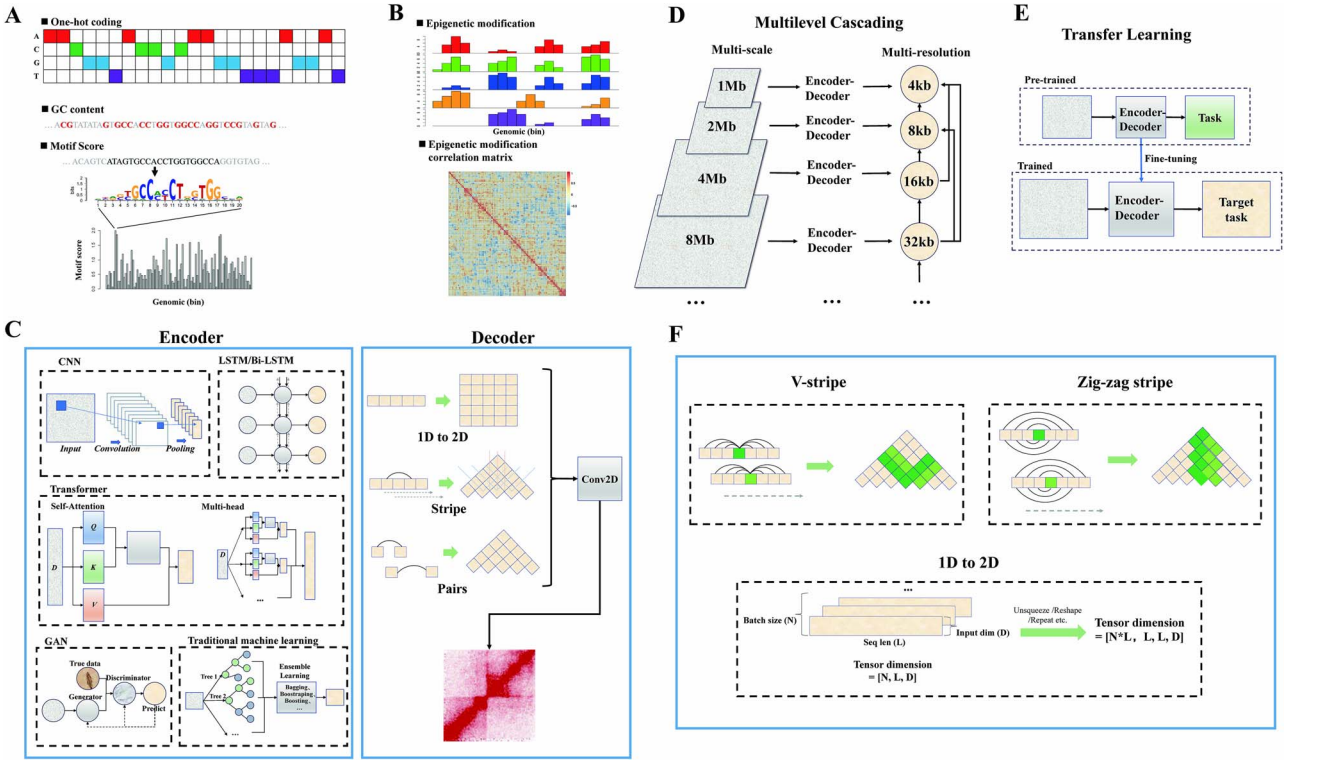

Figure 2. Feature encoding and key components in chromatin interaction prediction models. (A) Feature encoding based on DNA sequences, including GC content, one-hot encoding, and motif scores. (B) Feature encoding related to epigenetic modifications, involving transcription factor binding sites, histone modifications, and chromatin accessibility. (C) Model architecture, showing the functions of encoders and decoders and how they process 1D information to predict 3D chromatin interactions. (D) Multilevel decoding mode. (E) Schematic diagram of transfer learning. (F) Data dimension transformation methods, including stripe methods and 1D to 2D conversion techniques.

The diffusion process is a generative model in deep learning that introduces noise incrementally and learns to remove this noise during training to generate data [60].

"Traditional Machine Learning Models" are algorithms widely used before the rise of deep learning. Random Forest (RF) is a commonly used traditional machine learning models that solves classification, regression, and clustering problems through ensemble learning methods like Bagging and Bootstrapping (Fig. 2C). Due to their superior feature interpretability and stronger generalization capabilities with limited data, traditional machine learning methods are more suitable for specific functional chromatin interaction predictions [61, 66].

Data dimension transformation

Traditional machine learning methods, such as RF, commonly utilize anchor regions that form chromatin interaction information as input samples. In contrast, deep learning models typically use large-scale fixed windows as input samples. These two types of input samples are essentially based on 1D feature information. To predict chromatin interaction matrices from one dimension to two dimensions, the decoder must transform the linear 1D signal into a 2D signal. Deep learning models mainly use the following two data transformation methods (Fig. 2F).

"Stripe methods" utilize the partitioning concept from traditional machine learning, dividing 1D genomic information into bins to capture interactions between two bins. The main difference is that the Stripe method focuses on chromatin interactions along specific paths. As the window slides, this path generates a series of continuous 2D data. In deep learning models, various Stripe methods have developed based on specific pathway patterns. The V-stripe approach focuses on interactions between the center bin of the window and its upstream and downstream regions. The Zig-zag stripe mode begins at the center of the window, moving equidistantly toward both ends. It focuses on interactions between bins at corresponding positions and their adjacent bins, ultimately forming a Z-shaped path perpendicular to the center of the window.

"1D to 2D Conversion Methods" uses an encoder to extract features and perform operations like tiling and transposition to transform tensor dimensions. This conversion helps models capture local features and global relationships within linear sequences, enhancing the model's learning ability and generalization performance. Specifically, operations like *unsqueeze* and *Reshape* are typically used to insert new axes at specific tensor dimensions, thus accomplishing the transition from 1D to 2D data.

Performance evaluation criteria for predicting chromatin interaction matrices

Evaluating the model performance is a critical for ensuring the accuracy and reliability of predicted chromatin interactions. This process not only guides model optimization but also enhancers the understanding of new gene regulatory mechanisms, advancing biological knowledge. To reflect the degree of consistency between model predictions and actual data, the predictive performance of chromatin interaction matrices is typically assessed based on the following five criteria:

Contact matrices correlation

The chromatin interaction matrix, the core data form of technologies like Hi-C, displays the frequency of chromatin interactions across the genome or local regions. By calculating the correlation between the interaction matrix predicted by the model and that from the actual data, the model's accuracy in the global view and specific structural units of the 3D genome can be evaluated. Pearson and Spearman coefficients are commonly used to quantify the correlation, to comprehensively assessing the model's predictive ability for the complexity of genomic spatial organization.

Distance-stratified correlation

This method divides chromatin interaction matrices into different levels based on genomic distance, revealing distance-dependent regulatory patterns. By calculating the correlation between the model's predictions and the actual data at various distance levels, the model's predictive accuracy at different spatial scales is assessed. In addition to Pearson and Spearman coefficients, Area Under the Curve is also an important quantification metric.

Chromatin loop correlation

Chromatin loop refer to interaction events in the chromatin interaction matrices that are significant and biologically meaningful compared to the random background. Calculating the correlation between chromatin loops predicted by computational models and those found in actual data, this helps to verify the model's ability to capture interactions that actually affect gene regulation.

Insulation score correlation

The 3D structure of the genome is closely related to chromatin interactions, determining cell fate and biological processes. The insulation score, a metric calculated from chromatin interaction matrices, is crucial for identifying topologically associating domains (TADs). Therefore, evaluating the consistency of the insulation scores between the chromatin interaction matrices predicted by models and the actual data, or verifying the accuracy of the detected TADs, constitutes another important criterion for measuring model performance.

Model loss rate

The model loss rate, commonly known as the output value of the loss function during model training, serves as an indicator of the difference between the model's predictions and the actual data. The loss function is a core component of the model training process, guiding the optimization of model parameters and improving predictive performance. Common loss functions include Mean Squared Error, Cross Entropy Loss, etc. Choosing the appropriate loss function is critical for building an effective machine learning model, as it directly impacts the model's learning objectives and performance.

Deep learning in predicting chromatin interaction matrices

Using DNA sequence information and multi-dimensional epigenetic features to predict chromatin interaction matrices, the chosen of features and their encoding methods significantly impacts the prediction results. Based on the input features, existing models can be categorized into three groups (Table 1): those that use only DNA as input, those that use only epigenetic information, and those that combine both DNA sequences and epigenetic features. The following will introduce examples of these three groups of models.

Models based on DNA sequence

Table 3 shows that, during the encoder stage, deep learning models using DNA features as input commonly employ CNNs and dilated convolutions to extract features. These architectures excel at capturing both local and long-range dependencies within sequences. Compared to earlier models like Akita [57], DeepC [58], Orca [59], and HiCDiffusion [60] introduced more complex structures (Table 3).

DeepC first developed a pre-trained model for chromatin epigenetic features. Then, through transfer learning, this model was fine-tuned for predicting chromatin interactions. This innovative approach significantly enhanced the model's ability to recover chromatin interaction matrices from low-depth sequencing data.

Orca adds a cascading structure to CNNs and dilated convolutions, enhancing the model's performance in feature extraction and generalization. This design strategy allows the model to perform excellently when handling chromatin interaction data, and processing genomic data at various scales simultaneously [59].

The HiCDiffusion model builds on CNNs and dilated convolutions, further integrating a Transformer structure to enhance feature extraction and generalization capabilities. The uniqueness of this model lies in its introduction of a diffusion model, a generative model that aids in producing higher-resolution Hi-C matrices [60].

For data dimensionality transformation, while DeepC employs Z-stripe, the other three models utilize 1D to 2D methods. This suggests that 1D to 2D transformation may be more widely applicable in most cases.

Models based on epigenetic information

HiC-Reg [61] and Epiphany [62] are predictive models for chromatin interaction matrices that use epigenetic features as input. Both models incorporate histone modifications, transcription factor binding profiles, and chromatin accessibility as sample features. However, the key difference is that HiC-Reg uses anchor regions of chromatin interaction as samples, employing traditional machine learning method like RF to learn epigenetic information from these samples. Epiphany uses fixed windows as samples and employs CNN, Bi-LSTM, and GAN to learn epigenetic information from the samples, utilizing Z-stripe for dimensionality transformation (Table 3).

HiC-Reg [61] can operate with less training data and a limited number of features, allowing it to construct a model that predicts the interaction strength between any two regions. Epiphany [62] enhances the resolution of chromatin interaction predictions through its GAN architecture. Although this does not significantly improve the correlation between predicted and actual data, the GAN module reconstructs the chromatin interaction matrix with greater precision.

Models combining DNA and epigenetic information

EPCOT-COP [64], ChIPr [66], C.Origami [65], and ChromaFold [63] are predictive models for chromatin interaction matrices that combine both DNA sequences and epigenetic information as inputs, utilizing different feature fusion methods and model frameworks (Table 3).

EPCOT-COP and ChIPr employ a preliminary feature integration method. EPCOT-COP concatenates DNA sequences and chromatin

Table 3. Framework composition of chromatin interaction matrix prediction model based on deep learning.

Input	Model	Model framework in Encoder							Data dimension conversion method		
		CNN	Dilated	LSTM/ BI-LSTM	Transformer	Transfer Learning	Multilevel Cascading	Diffusion	GAN	1D to 2D	Z-stripe	V-stripe
DNA	Akita	√	√							√		
	DeepC	√	√			√					√	
	Orca	√	√				√			√		
	HiCDiffusion	√	√		√			√		√		
Epi	Epiphany	√		√					√		√	
DNA + Epi	C.Origami	√	√		√					√		
	ChromaFold	√		√								√
	EPCOT-COP	√	√		√					√		

Note: √ indicates that the model includes this module.

accessibility data into a feature vector within a 1 Mb region. This vector is then fed into modules such as CNN and self-attention mechanisms. The ChIPr model integrates seven distinct feature sets: RAD21, H3K27ac, H3K27me3 ChIP-Seq signals, genomic distance, GC content, and CTCF motif orientation. By utilizing dense neural networks, RFs, and gradient boosting, it effectively models the interaction strength between anchor regions by extracting and integrating these features.

C.Origami and ChromaFold employ a preliminary feature extraction method. C.Origami quantifies CTCF narrow peaks and chromatin accessibility as genomic features while using one-hot encoding as DNA features. Two CNN modules process DNA features and genomic features separately. Subsequently, they concatenate these features column-wise to enhance the model's ability to recognize long-distance interactions. Similarly, ChromaFold utilizes DNA information (CTCF motif score) computed by FIMO and peak information detected in scATAC-seq data as input features. Two independent CNN modules process DNA information and scATAC-seq data separately, and concatenate the extracted information before using linear regression to predict chromatin interaction matrices in Hi-C data.

Model performance and feature importance analysis

Impact of model structure on performance

Research results indicate that when input sample format and features are consistent, and the model output resolution is <10 kb, the Orca model with a cascade structure and the HiCDiffusion model with a diffusion structure achieve higher prediction accuracy (Fig. 3A). This finding highlights the positive impact of model complexity on predictive performance, when handling large-scale datasets and complex feature sets. Further comparative studies indicate that (Fig. 3B), under the same resolution and sample format, Transformer architectures outperform LSTM networks, demonstrating their superiority in capturing complex interactions between sequences.

Impact of feature encoding on performance

Based on reported model performances, we can preliminarily evaluate the impact of different feature encodings on the prediction performance of chromatin interaction matrices. In the same cell line, the EPCOT-COP model outperforms the Orca and Akita models in both GM12878 and Human foreskin fibroblasts (HFF) cell lines (Fig. 3C). At the same resolution for chromatin interaction matrix prediction, we observed that the Epiphany model exhibits similar performance to the DeepC model, while the EPCOT-COP model demonstrates the highest prediction accuracy (Fig. 3D). This suggests that combining DNA sequence features with epigenetic information can enhance the model's performance in predicting chromatin interaction matrices. In cross-cell line predictions, models that integrate multi-dimensional features achieve performance above 0.4. In contrast, models using only DNA information as input achieve an average performance of 0.26, indicating that multi-dimensional feature integration yields significantly higher cross-cell line prediction performance (Fig. 3E).

Analysis of key features influencing chromatin interaction matrices

Chromatin interactions are essential for gene expression regulation, cellular differentiation, and disease development within

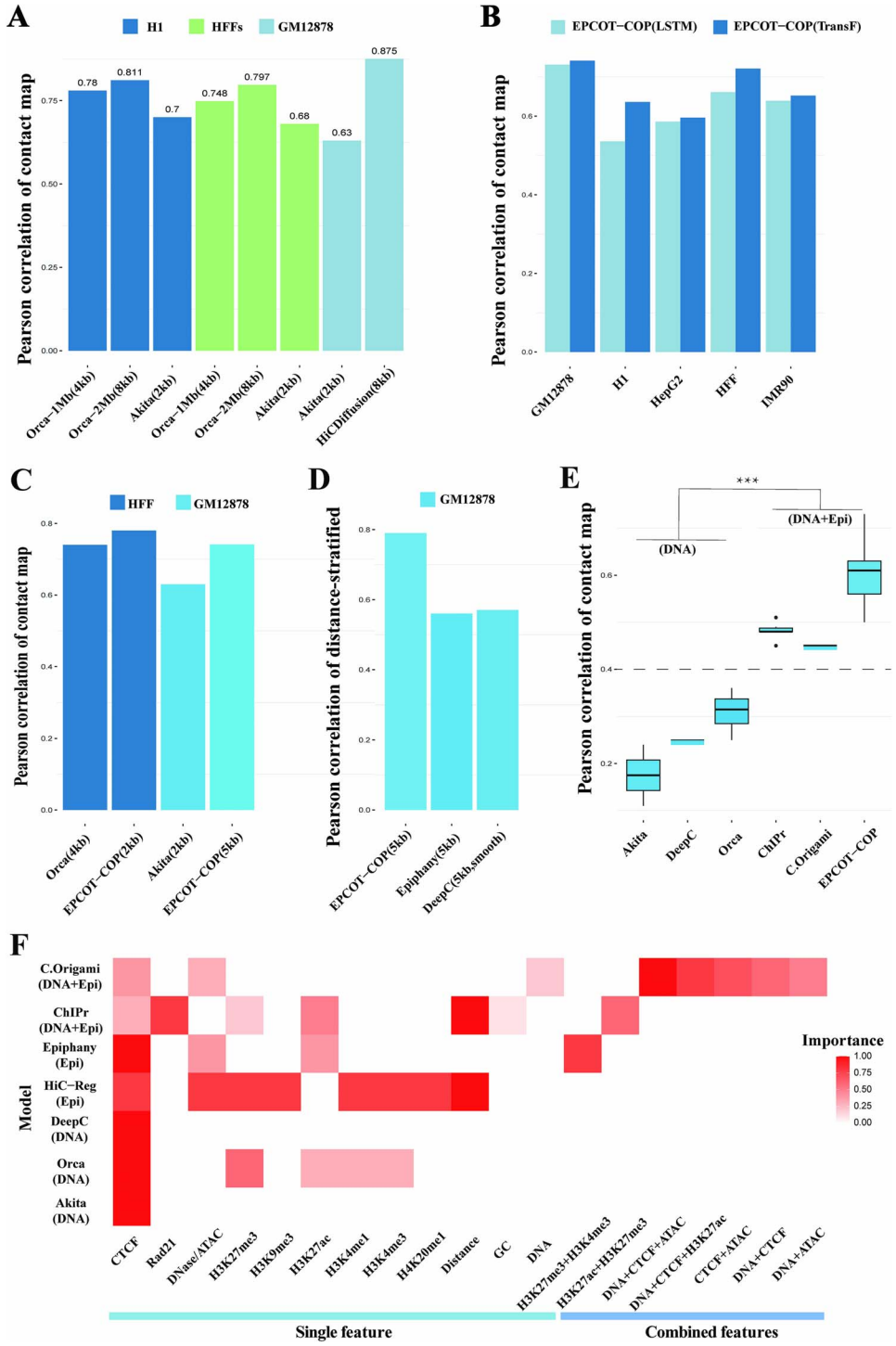

Figure 3. Performance evaluation for predicting chromatin interaction matrices. (A) Comparison of Pearson correlation coefficients for different model designs in predicting chromatin interaction matrices. Akita employs CNNs and dilated convolutions, while orca and HiCDiffusion add more complex structures. (B) Comparison of model prediction accuracy between LSTM and transformer structures in the EPCOT-COP model. (C) Comparison of Pearson correlation coefficients for different feature encoding in predicting chromatin interaction matrices. Orca and Akita use DNA sequences as the sole input, while EPCOT-COP utilizes a combination of DNA sequences and epigenetic information. The numbers in the parentheses indicate the resolution of the predicted chromatin interaction matrices. (D) Comparison of Pearson correlation coefficients for different feature encodings in predicting chromatin interaction matrices. Epiphany uses epigenetic information as the input, and DeepC uses DNA sequences as the input. (E) Comparison of Pearson correlation coefficients for cross-cell lines predictions. Akita, DeepC, and orca represent models using DNA sequences as the input, showing predictions in GM12878 after training in IMR90. ChIPr, C. Origami, and EPCOT-COP represent models using both DNA and epigenetics as inputs, demonstrating cross-cell line predictions across multiple cell lines. *** represents $P < 0.001$, the result of t-test. (F) The importance of selecting key features in chromatin interaction matrix prediction models. The vertical axis displays the various models used for key feature analysis, while the horizontal axis lists all key features identified by seven models, sorted according to individual transcription factor binding profiles, histone modifications, and feature combinations. The blank areas in the figure indicate that the feature was either not mentioned in a particular model or its impact on the model's predictive ability is negligible. To quantify the importance of features, we assigned each feature selected by the models a value between 0 and 1, with values closer to 1 indicating the importance of the feature. The model performance comparisons and key features mentioned in the article are sourced from previously reported literature, including Refs. 57–66.

the field of 3D genomics. Analyzing key features of chromatin interactions, such as CTCF binding sites, chromatin accessibility, and specific histone modifications, is crucial for understanding the 3D structure of the genome and its influence on gene regulation and cellular functions. Deep learning models can simulate these complex interactions, enabling predictions of chromatin interaction matrices and identification of features that influence chromatin structure. Key features analyze methods primarily fall into two categories: model selection methods and gradient-based methods.

Model selection methods, such as forward selection, feature ablation, and feature perturbation, identify essential features by optimize model performance through varying feature counts (Fig. 3F).

For DNA sequence-based models, base perturbation is commonly used, due to the unique one-hot encoding. Models like Akita and Orca use this strategy to quantify how base changes affect predictions, highlighting CTCF (CCCTC-binding factor) binding sites in chromatin interactions. Orca model further reveals that, after removing CTCF-enriched regions, H3K27me3 has a greater impact on chromatin interaction matrices than H3K4me3 or combinations of H3K27ac and H3K4me1.

Epigenetic signal-based models frequently use feature ablation techniques to assess each feature's contribution. This method involves setting specific epigenetic features to zero and comparing predictions to actual data. Notably, the Epiphany, C.Origami, and ChromaFold models have successfully applied this strategy, with finding showing that CTCF is most significant in single-feature ablation experiments. Epiphany further reveals a synergistic effect of the combination of H3K27me3 and H3K4me3, while C.Origami and ChromaFold emphasize the importance of ATAC-seq signals. ChIPr combining feature ablation and permutation testing, identifies genomic distance and RAD21 as the critical factors in GM12878 cell line predictions, while the combination of H3K27ac and H3K27me3 is relatively more significant than CTCF alone. The Hi-C-Reg model uses various methods to assess feature importance, including single-feature permutation and feature counting, ranks genomic distance, CTCF, and H4K20me1 in decreasing importance for chromatin interaction predictions. Additionally, Hi-C-Reg employs co-occurrence counts and Non-negative Matrix Factorization (NMF)-based clustering analysis to assess preferences for different feature combinations across categories of chromatin interactions (Fig. 3F).

Gradient-based methods reveal the importance of features within neural networks through forward and backward propagation. DeepC uses gradient-based image visualization techniques to evaluate the significance scores of CTCF and DNase peaks, revealing their important impact on chromatin interaction matrix (Fig. 3F).

Integrating these feature selection methods, clarify key features the importance and preferences in genomics, unraveling the complexity of chromatin interaction matrices. Specifically, features related to CTCF binding profile are widely recognized as critical for predicting chromatin interactions (Fig. 3F). Additionally, the chromatin accessibility along with epigenetic marks such as H3K27ac and H3K27me3, recognized by the three models as having an important impact on model performance. Notably, DNA-noly models show limitations in identifying key features for chromatin interactions, suggesting that incorporating epigenetic data can provide a more comprehensive understanding of the chromatin complexity and dynamics.

Challenges and limitations

Despite significant advancements in deep learning for predicting chromatin interaction matrices, single-cell sequencing technology has underscored the importance of genome's 3D structure for cellular function and gene regulation. However, research in this field still faces several challenges and limitations.

First, combining single-cell sequencing with chromosome conformation capture techniques (such as scHi-C [85] and sci DLO Hi-C [86]), has enabled chromatin structure and function studies at the single-cell level. Multi-omics techniques, such as scCARE-seq [87], LiMCA [88], and GAGE-seq [89], detect both genomic structure and gene expression simultaneously, revealing complex regulatory relationships between 3D chromatin structure and gene expression. Additionally, long-read sequencing technologies like scNanoHi-C [90], has further advanced the study of multi-directional interactions between enhancers and promoters. However, single-cell data have greater intrinsic variability and sparsity than traditional methods [91]. This variability, along with spatiotemporal dynamics of cellular states, demands models capable of capturing chromatin conformation changes at different developmental stages or under varying environmental conditions. Thus, modeling dynamic changes in 3D chromatin structure across time and tissue types remains a key challenge.

Second, the growing volume of experimental data, along with variations in experimental conditions, technological platforms, and sequencing depth, adds to data dimensionality and complexity, increasing computational demands. These factors may introduce biases or batch effects that obscure biological signals, limiting chromatin interaction model training and impact downstream analysis reliability. Consequently, data standardization, batch effect correction, large-scale epigenomic dataset integration and computational resource management, have become essential for integrating DNA sequence and epigenetic data, particularly in single-cell data.

Finally, Deep learning models have been widely applied in bioinformatics because of their powerful predictive capabilities. However, the black-box nature of these models hampers biological interpretability, complicating the extraction of underlying biological mechanisms. Although recent methods have been proposed to improve model transparency and interpretability, few solutions or concrete research examples exist for addressing these challenges in model interpretability.

Conclusion and prospects

In the field of 3D genomics, the collaborative use of deep learning and traditional machine learning has advanced chromatin interaction matrix prediction models, providing powerful tools for exploring gene expression regulation and molecular biology mechanisms. This review systematically examines innovations, which leverage multi-dimensional omics features, such as DNA sequences or epigenetic signals. Through various machine learning and deep learning frameworks, these models have transformed 1D information into 3D representations, significantly enhancing the model's ability to capture feature relationships on a global scale. Notably, neural networks and transformer frameworks have significantly improved prediction accuracy. Under conditions with fewer chromatin interaction features, traditional machine learning methods like RF show distinct advantages. Moreover, integrating multi-dimensional features, such as DNA sequences, chromatin open regions, histone

modifications, and transcription factor binding information, can considerably enhance predictive performance compared to single features. In-depth analyses reveal that besides CTCF binding sites, ATAC signals, H3K27ac, H3K27 methylation, RAD21 and chromatin interaction distance are critical factors in regulating chromatin interactions.

To address challenges in single-cell research, future studies should focus on improving model generalization to maintain high predictive accuracy even in new cell lines. For example, toolboxes for the analysis of the heterogeneity in single-cell data, such as SCRAT [92], along with data imputation algorithms like DeepImpute (neural network frameworks) [93]. Additionally, data imputation algorithms scHiCluster (linear convolution and random walks) [94] and Higashi (hypergraph neural networks) [95], offer effective solutions for processing single-cell Hi-C data. The SNN-Cliq (Shared Nearest Neighbor, SNN) [96] method for dimensionality reduction and clustering, aids in identifying single-cell types. These algorithms, provide a stronger foundation for understanding chromatin interaction data at the single-cell level, advancing single-cell chromatin interaction research.

In large-scale genomics, high-performance computing and optimized algorithms address data integration and computational challenges. Methods like ICE, HiCNorm, KR, and SCN [97–99] standardize data and correct batch effects reducing technical variability and enhancing comparability. Algorithms such as BatchI [100] and ComBat-seq [101], adjust for batch effects to ensure that integrated data accurately reflect biological phenomena accurately. RUVseq [102] corrects for library preparation and sequencing depth variations, minimizing technical biases. In the analysis of single-cell data, the Seurat v3 [103] strategy provides an integration method for scRNA-seq and scATAC-seq data, exploring chromatin activity and transcription relationships. CellHint [104] employs clustering tree to address differences in cell type annotation and technical biases across various datasets, while PRPS [105] eliminates tumor purity and batch effects.

Deep learning models, given their capacity for large-scale data processing, require significant computational resources influenced by data volume, model complexity, batch size, learning rate, hyperparameters, and GPU performance. For instance, the Akita model, with its 500 * 2000 * 4 encoded vector per sample, requires at least a 16GB Graphics processing unit (GPU), while training Orca model necessitates more substantial computational resources, specifically four NVIDIA Tesla V100 GPUs (32GB of memory). Thus, we recommend high-performance GPUs (such as the NVIDIA Tesla or RTX series) and ample memory, enabling parallel computing and distributed storage systems to accelerate analysis. Deep learning also aids in data dimensionality reduction, simplifying subsequent analyses and thereby reducing computational demands. Large models encoding DNA sequences are an emerging trend, with methods like DeepLncLoc [106], dna2vec [107] and GP-GCN [108], which are based on sequence embedding or graph models of K-mer feature encoding, as well as techniques like byte pair encoding and DNABERT-2 [109], effectively extract the order, position, and inter-feature DNA relationships.

To improve model interpretability, recent methods focus on feature importance and transparency, facilitating better understanding of predictive results. Attribution algorithms, which assess input importance to outputs for enhancing model interpretability [110], such as layer-wise relevance propagation [111], DeepLIFT [112], and SHAP tools [113]. Occlusion-based methods like Shapley value [114] assess output variations, while gradients-based methods, such as Gradient*Input [115], calculating the impact of input features on outputs. The LIME [116] algorithm is utilized to approximate the local interpretability of complex models and explain the structure of each model layer. Attention mechanisms are widely used to identify sequence fragments linked to protein functions, enhancing model interpretability in tasks like allele-specific activity prediction [117]. Advances in model interpretability provide readers with valuable insights and practical tools for chromatin interaction prediction.

The combination of 3D genomics and deep learning promises significant advancements in the research of animals, plants, and microorganisms. As datasets grow in scale and diversity, more refined and comprehensive predictive models will be developed. Besides chromatin interactions based on DNA–DNA, these models will extend to predict interaction matrices involving RNA–DNA and RNA–RNA interactions. A multimodal data approach, incorporating DNA sequences, epigenetic information, and transcription factor binding sites, will enable comprehensive 3D genome modeling, shedding light on biological developmental processes. Model interpretability combined with experimental validation will enhance trust and understanding of predictions, grounding genomic research in a solid theoretical framework.

In summary, the role of 3D genomics in genomic research is increasingly prominent. Chromatin interaction matrix prediction models based on deep learning will continue to be essential for elucidating biology processes. As technology and methodology advance, these models will play an even more critical role in future research.

Key Points

- Deep learning models playing a crucial role in predicting chromatin interaction matrices from DNA sequences and epigenomic profiles.
- Integrating multi-omics features such as DNA sequences, chromatin accessibility, histone modifications, and transcription factor binding information significantly improves the prediction accuracy compared to using single features.
- Modules that can capture long-distance dependencies, such as those using Transformer architectures, are more effective at predicting chromatin interactions.
- The field of chromatin interactions still faces challenges in model generalization, biological interpretability, and encoding high-dimensional features.
- The following features are critical for predicting chromatin interactions: CTCF binding sites, ATAC signals, Histone modifications H3K27ac and H3K27me3, Chromatin interaction distances.

A Glossary of Key Terms

AUC (Area Under the Curve): Indicates the area enclosed by the ROC (Receiver Operating Characteristic) curve and the coordinate axis, and represents the model's ability to distinguish between classes.

3D Genomics: The field studying the three-dimensional structure of the genome, which influences gene expression and cellular function.

ChIA-PET: A technique for identifying protein-mediated chromatin-chromatin interactions genome-wide, by combining chromatin immunoprecipitation and proximity ligation with paired-end tagging sequencing.

ChIP-seq: A technique combining chromatin immunoprecipitation with massively parallel DNA sequencing to identify the profile of DNA-associated proteins on the genome.
Chromatin Interactions: The physical contacts between genomic regions that can affect gene expression and regulation through chromatin structure changes.
Chromatin loop: A DNA loop formed by chromatin fibers, bringing distant regulatory elements and genes into spatial proximity.
CNN (Convolutional Neural Network): It is a deep learning model. It has powerful feature extraction ability to process image data by imitating human visual system.
Deep Learning: An AI technology, that uses neural networks with multiple layers to model and learn complex patterns, enabling tasks such as speech recognition and image classification.
Dilated Convolutions: A technique in convolutional neural networks, which expand the receptive field by inserting gaps between kernel units, allowing the capture of broader contextual information.
Epigenetics: The study of heritable gene function changes without DNA sequence changes, such as DNA methylation and histone modifications.
GAN (Generative Adversarial Network): A deep learning framework consisting of a generator and discriminator, which compete to improve the generator's ability to produce realistic data.
Hi-C: A chromosome conformation capture technique that maps 3D genome organization.
Insulation score: A measure of the degree to which a genomic region is insulated from interactions with other regions, indicating potential regulatory boundaries.
LSTM (Long Short-Term Memory): A recurrent neural network architecture capable of learning long-term dependencies, suitable for sequential data.
Random Forest: An ensemble learning method. It is operated by constructing multiple decision trees.
TAD (Topologically Associating Domain): Local regions in the genome. In three-dimensional space, they tend to be closer to each other and interact with each other to form structural and functional units.
Transformer: A neural network architecture based on self-attention mechanisms, effective for processing sequential data.
Transfer learning: An approach applying a pre-trained model to a new but related problem, leveraging learned features to improve performance.

Conflict of interest: The authors declared that there are no conflicts of interest or disclosures to report regarding this manuscript. All authors have reviewed and agree to the statement that there are no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 32370630, 32250710678, 32202653), National Key Research and Development Program of China (Grant No. 2021YFC2701201), and the China Postdoctoral Science Foundation (Grant No. 2023 M733830, BX2021367), National Key R&D Program of China (2023ZD04061) and Yingzi Tech & Huazhong Agricultural University Intelligent Research Institute of Food Health (IRIFH202209). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Misteli T. The self-organizing genome: principles of genome architecture and function. *Cell* 2020;**183**:28–45. https://doi.org/10.1016/j.cell.2020.09.014.
2. Cheng J, Cao X, Wang S. *et al.* 3D genome organization and its study in livestock breeding. *J Integr Agric* 2024;**23**:39–58. https://doi.org/10.1016/j.jia.2023.04.007.
3. Wang W, Gao R, Yang D. *et al.* ADNP modulates SINE B2-derived CTCF-binding sites during blastocyst formation in mice. *Genes Dev* 2024;**38**:168–88. https://doi.org/10.1101/gad.351189.123.
4. Karoutas A, Akhtar A. Functional mechanisms and abnormalities of the nuclear lamina. *Nat Cell Biol* 2021;**23**:116–26. https://doi.org/10.1038/s41556-020-00630-5.
5. Lieberman-Aiden E, van Berkum NL, Williams L. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93. https://doi.org/10.1126/science.1181369.
6. Fullwood MJ, Liu MH, Pan YF. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009;**462**:58–64. https://doi.org/10.1038/nature08497.
7. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. *Genome Biol* 2015;**16**:183. https://doi.org/10.1186/s13059-015-0745-7.
8. Liu L, Han K, Sun H. *et al.* A comprehensive review of bioinformatics tools for chromatin loop calling. *Brief Bioinform* 2023;**24**:bbad072. https://doi.org/10.1093/bib/bbad072.
9. Tang L, Hill MC, Ellinor PT. *et al.* Bacon: a comprehensive computational benchmarking framework for evaluating targeted chromatin conformation capture-specific methodologies. *Genome Biol* 2022;**23**:30. https://doi.org/10.1186/s13059-021-02597-4.
10. Tao H, Li H, Xu K. *et al.* Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. *Brief Bioinform* 2021;**22**:bbaa405. https://doi.org/10.1093/bib/bbaa405.
11. Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. *Mol Cell* 2016;**62**:668–80. https://doi.org/10.1016/j.molcel.2016.05.018.
12. Bickmore WA, van Steensel B. Genome architecture: domain Organization of Interphase Chromosomes. *Cell* 2013;**152**:1270–84. https://doi.org/10.1016/j.cell.2013.02.001.
13. Dixon JR, Selvaraj S, Yue F. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80. https://doi.org/10.1038/nature11082.
14. Akdemir KC, le VT, Chandran S. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat Genet* 2020;**52**:294–305. https://doi.org/10.1038/s41588-019-0564-y.
15. Zhang Y, Li G. Advances in technologies for 3D genomics research. *Sci China Life Sci* 2020;**63**:811–24. https://doi.org/10.1007/s11427-019-1704-2.
16. Wang X, Yue F. Hijacked enhancer–promoter and silencer–promoter loops in cancer. *Curr Opin Genet Dev* 2024;**86**:102199. https://doi.org/10.1016/j.gde.2024.102199.

17. Oh S, Shao J, Mitra J. *et al.* Enhancer release and retargeting activates disease-susceptibility genes. *Nature* 2021;**595**:735–40. https://doi.org/10.1038/s41586-021-03577-1.
18. Consortium EP. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;**583**:699–710. https://doi.org/10.1038/s41586-020-2493-4.
19. Ouyang WZ, Cao Z, Xiong D. *et al.* Decoding the plant genome: from epigenome to 3D organization. *J Genet Genomics* 2020;**47**:425–35. https://doi.org/10.1016/j.jgg.2020.06.007.
20. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. *Nat Rev Genet* 2018;**19**:789–800. https://doi.org/10.1038/s41576-018-0060-8.
21. Jerkovic I, Cavalli G. Understanding 3D genome organization by multidisciplinary methods. *Nat Rev Mol Cell Biol* 2021;**22**:511–28. https://doi.org/10.1038/s41580-021-00362-w.
22. Tang Z, Luo OJ, Li X. *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 2015;**163**:1611–27. https://doi.org/10.1016/j.cell.2015.11.024.
23. Jingwen Zhang QX, Li G. Epigenetics in the genesis and development of cancers. *Hereditas (Beijing)* 2019;**41**:567–81.
24. Pei LL, Li G, Lindsey K. *et al.* Plant 3D genomics: the exploration and application of chromatin organization. *New Phytol* 2021;**230**:1772–86. https://doi.org/10.1111/nph.17262.
25. Ouyang WZ, Xiong D, Li G. *et al.* Unraveling the 3D genome architecture in plants: present and future. *Mol Plant* 2020;**13**:1676–93. https://doi.org/10.1016/j.molp.2020.10.002.
26. Dubois F, Sidiropoulos N, Weischenfeldt J. *et al.* Structural variations in cancer and the 3D genome. *Nat Rev Cancer* 2022;**22**:533–46. https://doi.org/10.1038/s41568-022-00488-9.
27. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* 2019;**20**:535–50. https://doi.org/10.1038/s41580-019-0132-4.
28. Li G. *et al.* Emergence of 3D genomics. *Chinese Sci Bull (Chinese Version)* 2014;**65**:1165–72. https://doi.org/10.1360/TB-2019-0885.
29. Deng L, Zhou Q, Zhou J. *et al.* 3D organization of regulatory elements for transcriptional regulation in. *Genome Biol* 2023;**24**:181. https://doi.org/10.1186/s13059-023-03018-4.
30. Li MF, Gan J, Sun Y. *et al.* Architectural proteins for the formation and maintenance of the 3D genome. *Sci China-Life Sci* 2020;**63**:795–810. https://doi.org/10.1007/s11427-019-1613-3.
31. Barshad G, Lewis JJ, Chivu AG. *et al.* RNA polymerase II dynamics shape enhancer-promoter interactions. *Nat Genet* 2023;**55**:1370–80. https://doi.org/10.1038/s41588-023-01442-7.
32. Eraslan G, Avsec Ž, Gagneur J. *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403. https://doi.org/10.1038/s41576-019-0122-6.
33. Zou J, Huss M, Abid A. *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12–18. https://doi.org/10.1038/s41588-018-0295-5.
34. McCaw ZR, Colthurst T, Yun T. *et al.* Deep null models non-linear covariate effects to improve phenotypic prediction and association power. *Nat Commun* 2022;**13**:241. https://doi.org/10.1038/s41467-021-27930-0.
35. Angermueller C, Lee HJ, Reik W. *et al.* DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* 2017;**18**:67. https://doi.org/10.1186/s13059-017-1189-z.
36. Yin QJ, Wu M, Liu Q. *et al.* DeepHistone: a deep learning approach to predicting histone modifications. *BMC Genomics* 2019;**20**:193. https://doi.org/10.1186/s12864-019-5489-4.
37. Chen JY, Wang X, Ma A. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun* 2022;**13**:6494. https://doi.org/10.1038/s41467-022-34277-7.
38. Zhang Y, An L, Xu J. *et al.* Enhancing hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun* 2018;**9**:750. https://doi.org/10.1038/s41467-018-03113-2.
39. Liu Q, Lv H, Jiang R. hicGAN infers super resolution hi-C data with generative adversarial networks. *Bioinformatics* 2019;**35**:i99–107. https://doi.org/10.1093/bioinformatics/btz317.
40. Feng F, Yao Y, Wang XQD. *et al.* Connecting high-resolution 3D chromatin organization with epigenomics. *Nat Commun* 2022;**13**:2054. https://doi.org/10.1038/s41467-022-29695-6.
41. Piecyk RS, Schlegel L, Johannes F. Predicting 3D chromatin interactions from DNA sequence using deep learning. *Comput Struct Biotechnol J* 2022;**20**:3439–48. https://doi.org/10.1016/j.csbj.2022.06.047.
42. Yang M, Ma J. Machine learning methods for exploring sequence determinants of 3D genome organization. *J Mol Biol* 2022;**434**:167666. https://doi.org/10.1016/j.jmb.2022.167666.
43. Belokopytova P, Fishman V. Predicting genome architecture: challenges and solutions. *Front Genet* 2020;**11**:617202.
44. Zheng S, Thakkar N, Harris HL. *et al.* Predicting a/B compartments from histone modifications using deep learning. *iScience* 2024;**27**:109570. https://doi.org/10.1016/j.isci.2024.109570.
45. Gan W, Luo J, Li YZ. *et al.* A computational method to predict topologically associating domain boundaries combining histone marks and sequence information. *BMC Genomics* 2019;**20**:980. https://doi.org/10.1186/s12864-019-6303-z.
46. Wang Y, Liu Y, Xu Q. *et al.* TAD boundary and strength prediction by integrating sequence and epigenetic profile information. *Brief Bioinform* 2021;**22**:bbab139. https://doi.org/10.1093/bib/bbab139.
47. Cao F, Zhang Y, Cai Y. *et al.* Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome Biol* 2021;**22**:226. https://doi.org/10.1186/s13059-021-02453-5.
48. Shen Y, Zhong Q, Liu T. *et al.* CharID: a two-step model for universal prediction of interactions between chromatin accessible regions. *Brief Bioinform* 2022;**23**:bbab602. https://doi.org/10.1093/bib/bbab602.
49. Al Bkhetan Z, Plewczynski D. Three-dimensional Epigenome statistical model: genome-wide chromatin looping prediction. *Sci Rep* 2018;**8**:5217. https://doi.org/10.1038/s41598-018-23276-8.
50. Hong Y, Liu L, Feng Y. *et al.* mHapBrowser: a comprehensive database for visualization and analysis of DNA methylation haplotypes. *Nucleic Acids Res* 2024;**52**:D929–37. https://doi.org/10.1093/nar/gkad881.
51. Liu Z, Chen Y, Xia Q. *et al.* Linking genome structures to functions by simultaneous single-cell hi-C and RNA-seq. *Science* 2023;**380**:1070–76. https://doi.org/10.1126/science.adg3797.
52. Tian H, Yang Z, Xu X. *et al.* Three-dimensional chromosome conformation capture and its derived technologies. *Sheng Wu Gong Cheng Xue Bao* 2020;**36**:2040–50. https://doi.org/10.13345/j.cjb.200112.
53. Liu Z, Zhang Z. Mapping cell types across human tissues. *Science* 2022;**376**:695–96. https://doi.org/10.1126/science.abq2116.
54. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**:851–69. https://doi.org/10.1093/bib/bbw068.
55. Zhang P, Wu H. IChrom-deep: An attention-based deep learning model for identifying chromatin interactions. *IEEE J Biomed Health Inform* 2023;**27**:4559–68. https://doi.org/10.1109/JBHI.2023.3292299.

56. Zhang SS, Plummer D, Lu L. *et al.* DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell hi-C data at kilobase resolution. *Nat Genet* 2022;**54**:1013–25. https://doi.org/10.1038/s41588-022-01116-w.
57. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**:1111–17. https://doi.org/10.1038/s41592-020-0958-x.
58. Schwessinger R, Gosden M, Downes D. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;**17**:1118–24. https://doi.org/10.1038/s41592-020-0960-3.
59. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022;**54**:725–34. https://doi.org/10.1038/s41588-022-01065-4.
60. Chiliński M, Plewczynski D. HiCDiffusion - diffusion-enhanced, transformer-based prediction of chromatin interactions from DNA sequences. *BMC Genomics* 2024;**25**:964. https://doi.org/10.1186/s12864-024-10885-z.
61. Zhang S, Chasman D, Knaack S. *et al.* In silico prediction of high-resolution hi-C interaction matrices. *Nat Commun* 2019;**10**:5449. https://doi.org/10.1038/s41467-019-13423-8.
62. Yang R, das A, Gao VR. *et al.* Epiphany: predicting hi-C contact maps from 1D epigenomic signals. *Genome Biol* 2023;**24**:134. https://doi.org/10.1186/s13059-023-02934-9.
63. Gao VR. *et al.* ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. *Nat Commun* 2024;**15**:9432. https://doi.org/10.1038/s41467-024-53628-0.
64. Zhang Z, Feng F, Qiu Y. *et al.* A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Res* 2023;**51**:5931–47. https://doi.org/10.1093/nar/gkad436.
65. Tan J, Shenker-Tauris N, Rodriguez-Hernaez J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol* 2023;**41**:1140–50. https://doi.org/10.1038/s41587-022-01612-8.
66. Abbas A, Chandratre K, Gao Y. *et al.* ChIPr: accurate prediction of cohesin-mediated 3D genome organization from 2D chromatin features. *Genome Biol* 2024;**25**:15. https://doi.org/10.1186/s13059-023-03158-7.
67. Avsec Z, Agarwal V, Visentin D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–1203. https://doi.org/10.1038/s41592-021-01252-x.
68. Ji Y, Zhou Z, Liu H. *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083.
69. Chen KM, Wong AK, Troyanskaya OG. *et al.* A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* 2022;**54**:940–49. https://doi.org/10.1038/s41588-022-01102-2.
70. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107. https://doi.org/10.1093/nar/gkw226.
71. Sokolova K, Chen KM, Hao Y. *et al.* Deep learning sequence models for transcriptional regulation. *Annu Rev Genomics Hum Genet* 2024;**25**:105–22. https://doi.org/10.1146/annurev-genom-021623-024727.
72. Bohlin J, Eldholm V, Brynildsrud O. *et al.* Modeling of the GC content of the substituted bases in bacterial core genomes. *BMC Genom* 2018;**19**:589. https://doi.org/10.1186/s12864-018-4984-3.
73. Braghini MR, Lo Re O, Romito I. *et al.* Epigenetic remodelling in human hepatocellular carcinoma. *J Exp Clin Cancer Res* 2022;**41**:107. https://doi.org/10.1186/s13046-022-02297-2.
74. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. *Trends Genet* 2021;**37**:1012–27. https://doi.org/10.1016/j.tig.2021.05.002.
75. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;**20**:207–20. https://doi.org/10.1038/s41576-018-0089-8.
76. Gao T, Diaz-Hirashi Z, Verdeguer F. Metabolic Signaling into chromatin modifications in the regulation of gene expression. *Int J Mol Sci* 2018;**19**:4108. https://doi.org/10.3390/ijms19124108.
77. Belokopytova PS, Nuriddinov MA, Mozheiko EA. *et al.* Quantitative prediction of enhancer-promoter interactions. *Genome Res* 2020;**30**:72–84. https://doi.org/10.1101/gr.249367.119.
78. Füllgrabe J, Gosal WS, Creed P. *et al.* Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat Biotechnol* 2023;**41**:1457–64. https://doi.org/10.1038/s41587-022-01652-0.
79. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the Acm* 2017;**60**:84–90. https://doi.org/10.1145/3065386.
80. Siami-Namini S, Tavakoli N, Namin AS. The performance of LSTM and BiLSTM in forecasting time series. *IEEE Int Conf Big Data (Big Data)* 2019;**2019**:3285–92.
81. Vaswani A. *et al.* Attention is all you need. *Adv Neural Inform Processing Syst* 2017;**30**:5998–6008.
82. Niu ZY, Zhong GQ, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing* 2021;**452**:48–62. https://doi.org/10.1016/j.neucom.2021.03.091.
83. Li YC, Wang Q, Zhang J. *et al.* The theoretical research of generative adversarial networks: An overview. *Neurocomputing* 2021;**435**:26–41. https://doi.org/10.1016/j.neucom.2020.12.114.
84. Fang J, Ning X, Mao T. *et al.* A multi-focus image fusion network combining dilated convolution with learnable spacings and residual dense network. *Comput Electrical Eng* 2024;**117**:109299. https://doi.org/10.1016/j.compeleceng.2024.109299.
85. Nagano T, Lubling Y, Stevens TJ. *et al.* Single-cell hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**:59–64. https://doi.org/10.1038/nature12593.
86. Lin D, Xu W, Hong P. *et al.* Decoding the spatial chromatin organization and dynamic epigenetic landscapes of macrophage cells during differentiation and immune activation. *Nat Commun* 2022;**13**:5857. https://doi.org/10.1038/s41467-022-33558-5.
87. Qu J, Sun J, Zhao C. *et al.* Simultaneous profiling of chromatin architecture and transcription in single cells. *Nat Struct Mol Biol* 2023;**30**:1393–1402. https://doi.org/10.1038/s41594-023-01066-9.
88. Wu H, Zhang J, Jian F. *et al.* Simultaneous single-cell three-dimensional genome and gene expression profiling uncovers dynamic enhancer connectivity underlying olfactory receptor choice. *Nat Methods* 2024;**21**:974–82. https://doi.org/10.1038/s41592-024-02239-0.
89. Zhou T, Zhang R, Jia D. *et al.* GAGE-seq concurrently profiles multiscale 3D genome organization and gene expression in single cells. *Nat Genet* 2024;**56**:1701–11. https://doi.org/10.1038/s41588-024-01745-3.
90. Li W, Lu J, Lu P. *et al.* scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. *Nat Methods* 2023;**20**:1493–1505. https://doi.org/10.1038/s41592-023-01978-w.

91. Zhen C, Wang Y, Geng J. *et al.* A review and performance evaluation of clustering frameworks for single-cell hi-C data. *Brief Bioinform* 2022;**23**:bbac385. https://doi.org/10.1093/bib/bbac385.
92. Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT. *Bioinformatics* 2017;**33**:2930–32. https://doi.org/10.1093/bioinformatics/btx315.
93. Arisdakessian C, Poirion O, Yunits B. *et al.* DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol* 2019;**20**:211. https://doi.org/10.1186/s13059-019-1837-6.
94. Zhou J, Ma J, Chen Y. *et al.* Robust single-cell hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci USA* 2019;**116**:14011–18. https://doi.org/10.1073/pnas.1901423116.
95. Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell hi-C analysis with Higashi. *Nat Biotechnol* 2022;**40**:254–61. https://doi.org/10.1038/s41587-021-01034-y.
96. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 2015;**31**:1974–80. https://doi.org/10.1093/bioinformatics/btv088.
97. Imakaev M, Fudenberg G, McCord RP. *et al.* Iterative correction of hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 2012;**9**:999–1003. https://doi.org/10.1038/nmeth.2148.
98. Knight PA, Ruiz D. A fast algorithm for matrix balancing. *Ima J Num Anal* 2013;**33**:1029–47. https://doi.org/10.1093/imanum/drs019.
99. Cournac A, Marie-Nelly H, Marbouty M. *et al.* Normalization of a chromosomal contact map. *BMC Genomics* 2012;**13**:436. https://doi.org/10.1186/1471-2164-13-436.
100. Papiez A, Marczyk M, Polanska J. *et al.* BatchI: batch effect identification in high-throughput screening data using a dynamic programming algorithm. *Bioinformatics* 2019;**35**:1885–92. https://doi.org/10.1093/bioinformatics/bty900.
101. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020;**2**:lqaa078. https://doi.org/10.1093/nargab/lqaa078.
102. Risso D, Ngai J, Speed TP. *et al.* Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;**32**:896–902. https://doi.org/10.1038/nbt.2931.
103. Stuart T, Butler A, Hoffman P. *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
104. Xu C, Prete M, Webb S. *et al.* Automatic cell-type harmonization and integration across human cell atlas datasets. *Cell* 2023;**186**:5876–5891.e20. https://doi.org/10.1016/j.cell.2023.11.026.
105. Molania R, Foroutan M, Gagnon-Bartsch JA. *et al.* Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat Biotechnol* 2023;**41**:82–95. https://doi.org/10.1038/s41587-022-01440-w.
106. Zeng M, Wu Y, Lu C. *et al.* DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. *Brief Bioinform* 2022;**23**:bbab360. https://doi.org/10.1093/bib/bbab360.
107. Hong ZY, Zeng X, Wei L. *et al.* Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 2020;**36**:1037–1043. https://doi.org/10.1093/bioinformatics/btz694.
108. Wei SB, Zhu G, Sun Z. *et al.* GP-GCN: global features of orthogonal projection and local dependency fused graph convolutional networks for aspect-level sentiment classification. *Connection Science* 2022;**34**:1785–1806. https://doi.org/10.1080/09540091.2022.2080183.
109. He JS, Zhang S, Fang C. Prediction of DNA enhancers based on multi-species genomic base model DNABERT-2 and BiGRU network. In: *Proceedings of 2024 4th International Conference on Bioinformatics and Intelligent Computing*. New York, NY, USA: Association for Computing Machinery, 2024, 375–79. https://doi.org/10.1145/3665689.3665752.
110. Huiqi Deng NZ, Mengnan D, Chen W. *et al.* Understanding and unifying fourteen attribution methods with Taylor interactions. Preprint, bioRxiv. 2023. arXiv:2303.01506. https://doi.org/10.48550/arXiv.2303.01506.
111. Bach S, Binder A, Montavon G. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 2015;**10**:e0130140. https://doi.org/10.1371/journal.pone.0130140.
112. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning*, Vol. **70**. JMLR.org, 2017, 3145–53.
113. Zhang J. *et al.* Recommendation with causality enhanced natural language explanations. In: *Proceedings of the ACM Web Conference 2023*. New York, NY, USA: Association for Computing Machinery, 2023, 876–86. https://doi.org/10.1145/3543507.3583260.
114. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Vol. **30**. Red Hook, NY, USA: Curran Associates Inc., 2017, 4768–77.
115. Hechtlinger Y. Interpretation of prediction models using the input gradient. Preprint, bioRxiv. 2016. arXiv:1611.07634. https://doi.org/10.48550/arXiv.1611.07634.
116. Ribeiro MT, Singh S, Guestrin C. "why should I trust you?" explaining the predictions of any classifier. In: *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, 1135–44. https://doi.org/10.1145/2939672.2939778.
117. Rozowsky J, Gao J, Borsari B. *et al.* The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell* 2023;**186**:1493–1511.e40. https://doi.org/10.1016/j.cell.2023.02.018.

Briefings in Bioinformatics, 2025, **26(1)**, bbae651
https://doi.org/10.1093/bib/bbae651
Review