Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DeepCis: A Transformer-based Multi-Omics Framework for Cross-Scale Explainability in Gene Regulatory Element Identification

Xinrong Li
Harbin Medical University
Wei Feng
Harbin Medical University
Junjie Duan
Harbin Medical University
Yujie Liu
Harbin Medical University
Yusen Zhao
Harbin Medical University
Shuangshuang Wang
Harbin Medical University
Hongbo Zhu
Harbin Medical University
Manyi Xu
Harbin Medical University
Yongle Xu
Harbin Medical University
Mengmeng Liu
Harbin Medical University
Wenbo Dong
Harbin Medical University
Peng Wang
Harbin Medical University
Shangwei Ning
Harbin Medical University
Hui Zhi
zhihui@ems.hrbmu.edu.cn

Harbin Medical University

Article

Keywords: DeepCis, Cis-regulatory elements, Multi-omics, Gene regulation

Posted Date: March 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8988269/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Dysregulation of gene expression is a hallmark of cancer, wherein cis-regulatory elements(CREs) in non-coding genomic regions play pivotal roles in precise transcriptional control. Although numerous candidate CREs(cCREs) have been identified, the complex regulatory relationships between cCREs and their target genes remain poorly characterized, and current computational approaches are inadequate for large-scale prediction of functional cCREs. Here, we introduce DeepCis, a Transformer-based multi-omics framework that integrates base-level genomic sequence and epigenomics features, as well as region-level 3D chromatin organization to capture regulatory information across multiple scales. DeepCis quantitatively assesses the contribution of individual cCREs to gene regulation and reveals the cooperative interactions among cCREs. Our framework demonstrates robust performance in gene expression prediction, effectively identifying key cCREs and their specific contributions to transcriptional output. The model captures both proximal dominance and distal activity of cCREs and elucidates how 3D chromatin organization constrains gene regulation. Furthtermore, DeepCis successfully identifies transcription factor binding sites associated with active regulation. Through comprehensive multi-level interpretability analyses, we validate the reliability and effectiveness of the DeepCis in deciphering gene regulation mechanisms, providing a powerful foundation for advancing our understanding of transcripgional regulation and enabling applications in precision medicine.

1. Introduction

Gene expression dysregulation represents a fundamental characteristic of cancer pathogenesis. CREs including promoters, enhancers, silencers, and insulators—reside in non-coding genomic regions and precisely orchestrate transcription by recruiting transcription factors (TFs) and chromatin-associated proteins in a spatiotemporal manner[1,2,3]. Disruption of these CREs through genetic mutations or epigenetic alterations can lead to aberrant transcriptional programs and cancer progression[4,5]. Gene regulation is inherently multifactorial, involving synergistic contributions from DNA sequence, chromatin accessibility, histone modifications, and higher-order 3D chromatin organization[6,7]. Recent advances in high-throughput sequencing and large-scale consortia such as ENCODE and 4D Nucleome have identified millions of cCREs[8]. However, determining which cCREs functionally influence specific genes, and quantifying their individual contributions to transcriptional output, remains a fundamental challenge in genomics[9,10].

Experimental approaches such as CRISPR-Cas9 perturbations can validate regulatory activity at individual loci with high precision, but suffer from limitions in scalability, cost, and throughput, and fail to capture cooperative or compensatory interactions among cCREs across large genomic distances[11]. Chromatin conformation techniques (Hi-C[12], ChIA-PET[13], HiChIP[14], and PLAC-seq[15]) provide physical contact maps between regulatory regions and promoters, but physical proximity does not necessarily equate to functional regulation.

Computational models have attempted to infer enhancer-gene relationships from sequence, chromatin accessibility, or 3D chromatin organization. The Activity-by-Contact (ABC) model conceptualizes enhancer influence as the product of enhancer activity and contact frequency but assumes linear additivity and may fail to account for non-linear cooperative regulation[16]. Deep learning models trained on binary enhancer-promoter interactions or CRISPR screens are constrained by limited positive samples and poor cross-cell-type generalizability[17–19]. Basenji2 employs convolutional neural networks to capture local dependencies, but it remains limited in scope and may not effectively capture long-range dependencies[20]. Enformer captures long-range interactions from sequence alone but struggles with consistent performance in unseen cellular contexts[21]. GraphReg integrates chromatin interactions with multi-omics features but retains bias toward local connectivity[22]. CREaTor introduces zero-shot modeling of regulatory patterns from sequence or epigenomics but lacks integration of spatial chromatin organization and provides insufficient clarity explaining the biological interpretability of key cCREs[23].

Thus, a critical unmet need exists for a unified, quantitative, and interpretable framework that (i) integrates genomic, epigenomic, and 3D chromatin organization across multiple scales, (ii) operates without predefined positive or negative regulatory labels, (iii) quantifies individual and cooperative contribution of each cCRE to gene expression, and (iv) maintains biological interpretability.

Here, we present DeepCis, a Transformer-based multi-omics framework that integrates DNA sequence, base-resolution epigenomic signals, and condensed spatial chromatin organization features derived from Hi-C data. DeepCis employs a zero-shot learning strategy and a two-stage encoders architecture to capture both local sequence determinants and global cCRE-gene interactions. Crucially, we introduce a Regulatory Contribution Score (RCS) to quantitatively assess each cCRE's contribution to transcriptional output. Through multi-level interpretability analyses, we demonstrate that DeepCis robustly identifies key cCREs across cell types, captures both proximal dominance and distal enhancer activity, identifies transcription factor binding sites and cooperative regulation, reveals 3D chromatin constraints on gene regulation, and discovers hub cCREs overlapping super-enhancers. Collectively, DeepCis provides a scalable and interpretable strategy for decoding cis-regulatory architectures, offering mechanistic insights and potential applications in disease research and precision medicine.

2. Results

2.1. Overview of the DeepCis Framework

DeepCis comprises two cascaded Transformer encoders that integrate genomic sequence, epigenomic signals, and 3D chromatin organization features to quantitatively model cCRE-gene regulatory relationships and coordinative interactions among cCREs in a supervised learning framework. Initially, the base information of cCREs is processed through numerical and one-hot encoding. Epigenomic signals are then incorporated to extract base-level signal features, which are then normalized

accordingly. For 3D chromatin organization, we construct cCREs contact matrices and perform singular value decomposition (SVD) for dimensionality reduction, obtaining condensed spatial features for each cCRE. This process preserves core spatial information, capturing the relative positional relationships and synergistic regulatory concacts of cCREs in 3D chromatin organization (Fig. 1a, Methods).

The DeepCis architecture(Fig. 1b, Methods) incorporates four core modules: input module, feature mapping module, feature fusion encoding module, and the interaction representation encoding module. This architecture enables extraction and fusion of complementary information types to decipher complex gene regulatory mechanisms. The feature mapping module processes each omics data type, converting raw features into high-dimensional representations suitable for integration. The feature fusion encoding module processes cCREs and target gene features, focusing on cross-scale feature fusion to generate abstract representations for each cCRE or gene. We employ distinct fusion strategies for features at different granularities. At the base pair level, DeepCis concatenates encoded genomic sequence features with corresponding epigenomic features, as they provide complementary information, with genomic sequence suppling genetic encoding information while epigenomic features indicating active state and regulatory potential. This concatenated feature is downsampled and passed through the first encoder to obtain initial representations of cCREs and gene regions. Relative positional bias is incorporated to capture relative base positions within the sequence, enhancing the model's understanding of internal sequence information. Spatial embedding representations (Methods), are then fused with the initial representation leveraging the complementary nature of spatial, sequence, and epigenomic information.

The interaction representation encoding module integrates features from each target gene and its corresponding cCREs for global regulation, modeling relationships between different cCREs and between cCREs and genes. This module identifies key functional cCREs (enhancers and promoters) and transcription factor binding sites by quantifying cCREs contributions through the Regulatory Contribution Score (RCS) (Methods), providing a basis for biological interpretation. Based on abstract features extracted in previous modules, features of cCREs related to the target gene are integrated with the gene’s features to create a unified representation optimized for predicting gene expression. To capture precise regulatory relationships, we introduce dual masks based on biological priors (Methods).All DeepCis modules are jointly trained in an end-to-end framework.

DeepCis offers several advantages: First, it integrates comprehensive multi-omics data (genomic sequences, epigenomic signals, and 3D chromatin organization), enhancing accuracy and richness of cCREs representations. Second, it employs a flexible feature fusion mechanism that adapts strategies based on data scale to effectively combine various omics features, improving the model expressiveness. Through spatial representations, DeepCis captures the relative positioning of cCREs in 3D chromatin organization, identifying potential synergistic regulatory relationships. The multi-head attention mechanism of Transformer enables capture of long-range interactions between cCREs and between cCREs and genes. The model also introduces dual mask constraints (Methods). Finally, DeepCis excels

in identifying key CREs, recognizing known enhancers, promoters, and transcription factor binding sites, while uncovering potential novel cCREs, providing insights into gene regulatory mechanisms (Fig. 1c).

2.2. Robust gene expression prediction by DeepCis

We evaluated DeepCis performance across cell lines and chromosomes. Independent tests were conducted on the test set from chromosome 6 in four cell lines and the unseen HCT116 cell line. DeepCis achieved an overall Pearson correlation coefficient of 0.752 on the test set, with a mean absolute error of 0.708 (Fig. 2b). In the K562 cell line, the Pearson correlation for chromosome 6 reached 0.795(Fig. 2c). In the unseen HCT116 cell line, the average Pearson correlation across all chromosomes was 0.720, with the highest correlation of 0.747 on chromosome 22 (Fig. 2d). These results demonstrate strong generalization capability of DeepCis across different biological contexts.

To assess advantages of DeepCis, we first tested models trained with single-omic data on the test set. Pearson correlation coefficients for single-omic models were consistently lower than those of DeepCis (Supplementary Fig. S1a). Notably, Pearson correlation coefficients for single-omic models all exceeded 0.50, indicating that each individual omic data type contains significant predictive information for gene expression, validating the necessity and effectiveness of multi-omics itegration. We compared different dimensionalities of condensed spatial features and found that a 32-dimensional features achieved the optimal balance between retaining key information and reducing training complexity (Fig. 2b). Different fusion strategies were compared, with concatenating genomic and epigenomic features before adding spatial embedding features yielding the highest model accuracy (Fig. 2a). Model predictive performance remained stable even after removing subsets of features, with only minor changes in Pearson correlation coefficient (Fig. 2e, Supplementary Fig. S1b), indicating robustness in handling epigenomic features.

2.3. Quantification of cCREs contributions and identification of key cCREs by DeepCis

To validate whether high-RCS cCREs predicted by DeepCis correspond to key cCREs in actual regulatory processes, we examined enrichment of enhancers and promoters in high and low regulatory contribution groups (Methods). We obtained known enhancer data from FANTOM5 database for GM12878, HepG2, K562, and HeLaS3 cell lines and identified promoter regions (TSS ± 1 kb) for each gene. Results show significant enrichment of enhancers and promoters in the high regulatory contribution group ($p < 0.001$) (Fig. 3a), with average fold enrichment of 2 in high-RCSs group compared to low-RCSs group (Fig. 3b).This confirms that the attention scores learned by DeepCis strongly correlated with biological functions of gene regulation, providing a solid basis for identifying key cCREs.

To provide direct functional evidence, we evaluated DeepCis on an established benchmark dataset of CRISPR/Cas9-validated enhancers and non-enhancers from K562 cell line, assessing the association between RCSs and functional cCREs, with performance quantified by the AUC. Prior to evaluation, we normalized predicted RCSs within the same gene to eliminate inter-gene variability and highlight relative regulatory strength. Results show close alignment between model-predicted RCSs and functional

enhancers, with AUC of 0.88 (95% CI: 0.87–0.89) (Fig. 3c). Model performance across different chromosomes showed highest AUC reaching 0.96 (chr4) (Supplementary Fig. S2a), demonstrating effective identification of functionally relevant cCREs. We further tested the selected optimal epigenomic feature combinations (Fig. 3d, Methods), integrating DNA sequence and spatial features. Performance remained stable and high across different numbers of retained epigenomic features (AUC: 0.86–0.90) (Fig. e-i), demonstrating that DeepCis is robust and not heavily dependent on epigenomic features. Notably, H3K79me2, POLR2A, and DNase emerged as key markers influencing predictive accuracy of gene expression regulation in our model (Supplementary Fig. S2b, Methods), with performance improving as more feature combinations were included in selection process (Supplementary Fig. S2c). DNase hypersensitive regions mark open chromatin states, indicating enhanced accessibility for transcription factor binding. Previous studies established H3K79me2 functional regulatory role through co-transcriptional splicing mechanism[24]. The importance of POLR2A in cancer is well-established-identified as a "synthetic vulnerability" target in triple-negative breast cancer[25] and driving abnormal tumor cell proliferation in esophageal cancer in cooperation with Bclaf1[26]. Comparison of DeepCis with other methods for identifying key cCREs demonstrated superior performance of DeepCis (Fig. 3j, Methods).

2.4. DeepCis captures proximal dominance and distal activity of cCREs

Previous studies established that regulatory efficacy of cCREs is significantly influenced by proximity to target genes, exhibiting clear distance effect[27–29]. To systematically evaluate whether DeepCis captures distance dependency of cCREs regulation on gene expression, we analyzed relationship between RCSs of cCREs learned by the model and distance to the target gene across five cell lines. Results consistently showed significant negative correlation between RCSs and distance to target gene (Spearman correlation coefficient from − 0.723 to -0.698, $p < 0.01$) (Fig. 4a), indicating that the model captures distance decay effects commonly observed in genomic regulation, consistent with prior studies.

However, accumulating evidence supports critical roles of distal enhancers in complex gene regulatory networks[30,31]. To assess the model's ability to identify distal regulation, we further analyze types of cCREs in high regulatory contribution group. In GM12878 cell line, proximal enhancer-like signatures (pELS) accounted for 32.2%, promoter-like signatures (PLS) for 19.0%, distal enhancer-like signatures (dELS) for 15.6%, and CTCF-only and CTCF-bound cCREs for 12.6%. To minimize potential for random variation, we validated results across all chromosomes in each cell line, observing consistent trends(Fig. 4b, Supplementary Fig. S3a-S3d). These findings demonstrate the model integrates both proximal regulatory dominance and key long-range regulatory effects, overcoming simple distance bias.

2.5. DeepCis accounts for 3D chromatin organization impact on gene regulation

The genome is organized in intricate three-dimensional structures through complex chromatin folding, imposing critical constraints on precise gene regulation[6]. Topologically Associating Domains (TADs) represent fundamental functional units in the genome 3D organization, where cCREs within same domain interact more frequently, while across-TAD interactions are relatively rare[32]. We systematically analyzed high regulatory contribution group (Methods) located within same or different TADs (Methods)from their target genes. Results show RCSs of cCREs within same TAD as target gene are significantly higher than those located in different TADs (Mann-Whitney U test, $p < 0.0001$). This significant TAD co-localization between high contribution cCREs and genes was consistently observed across the analyzed cell lines (Fig. 4c, Supplementary Fig. S3e), strongly suggesting that DeepCis effectively accounts for the impact of the 3D chromatin organization on gene regulation.

2.6. DeepCis identifies core transcription factor binding sites associated with active regulation

To investigate the association between high RCSs cCREs and the binding of active TFs, we integrated ChIP-seq data for 108 known TFs across five cell lines with target gene information (Methods). We analyzed the correlation between RCSs predicted by DeepCis and enrichment of transcription factor binding sites for these target genes. Comparison of TFs binding site counts between high and low regulation contribution groups (Methods) revealed significant enrichment of 87 TFs on high RCSs cCREs (Fisher's Exact Test, $p < 0.05$), with an average enrichment factor of 2.393 (Fig. 5a, Supplementary Fig. S4a), demonstrating that DeepCis effectively identifies core regulatory regions highly associated with transcription factor binding. Notably, some TFs were not enriched in high-contribution cCREs, possibly due to shared regulation of adjacent genes by neighboring cCREs. Additionally, TFs such as CUX1 and FOXP1, implicated in transcriptional repression in existing studies[33,34], may suppress gene expression or maintain dormant state upon DNA binding, potentially explaining the lower contribution scores for their associated cCREs.

To quantify the relationship between RCSs of cCREs and TF activity, we focused on the cCREs overlapping with TF binding peaks. We analyzed the correlation between RCSs of these cCREs and corresponding TF ChIP-seq signal intensities. Two distinct correlation patterns emerged at gene level: first group showed significant positive correlation between RCSs of cCREs and TF binding signal intensity, with an average Spearman correlation coefficient of 0.7, representing activation mode(Fig. 5b). Second group exhibited negative correlation, with an average Spearman coefficient of -0.77 (Fig. 5b), indicating regulation by transcriptional repression mechanisms. We validated this correlation using tissue-specific data and corresponding cell lines, finding DeepCis predictions consistent with observed directions (Fig. 5c, Supplementary Fig. S4b, Methods). For instance, activation of *ROCK2* by transcription factor NRF1 in the K562 cell line showed significant positive correlation (Spearman correlation = 0.655, $p = 0.001$), orroborated in acute myeloid leukemia (LAML) cohort (Spearman correlation = 0.372, $p = 2.40 \times 10\text{e-}6$). Conversely, repression of *TYROBP* by NRF1 in K562 showed a significant negative correlation (Spearman correlation = -0.354, $p = 0.004$), supported by LAML data (Spearman correlation = -0.240, $p = 0.003$)(Fig. 5c, 5d). However, Negative correlations may reflect complex inhibitory

mechanisms, such as competitive binding or recruitment of chromatin remodeling complexes that indirectly suppress gene expression. Overall, these findings suggest that DeepCis captures multi-level regulatory patterns.

2.7. DeepCis predicts hub CREs in gene regulatory systems

Beyond quantifying regulatory relationships between genes and cCREs, DeepCis quantifies the synergistic interactions between cCREs, revealing global regulatory architectures. We hypothesize that cCREs frequently forming high-intensity synergistic relationships with other cCREs in different gene regulatory networks may serve as core hubs genome-wide. To test this, we summarized and analyzed the high regulatory contribution group (Methods) identified for each gene as high contribution cCREs to identify key nodes with global influence.

Genome-wide statistical analysisrevealed that only a few cCREs were repeatedly selected across different gene regulatory networks with high occurrence frequency. Some high-frequency cCREs overlapped known super-enhancer regions (Supplementary Table 2), critical for regulating genes involved in cell identity, development, and disease processes[35,36]. For example, in K562, chr9:35732261–35732610 region overlapped with a super-enhancer associated with key genes including *CREB3* (transcription factor involved in stress response and cell growth regulation[37]), *TLN1* (key molecule in cell adhesion, migration, and signal transduction[38]), and *CA9* (carbonic anhydrase family member promoting tumor growth and metastasis in acidic tumor environments[39]). In the HCT116, chr1:1308334–1308678 region also overlapped a super-enhancer associated with key genes including *ISG15* (important in antiviral immunity[40]), and *ATAD3A* (related to mitochondrial function[41]) (Fig. 5d). These findings suggest that these cCREs function as key "hub" cCREs orchestrating regulation of multiple genes through extensive interactions, providing new insights into complex epigenetic regulatory network.

2.8. Causal impact of key cCREs confirmed by perturbation analysis

To confirm causal role of high RCS cCREs identified by the model in gene expression regulation, we performed progressive genome perturbation experiments in each cell line across chromosomes. Specifically, for each gene's associated cCREs, we simulated knockout of 5%, 10%, 15%, 20%, 25%, and 30% of high RCS cCREs, ranked high to low by RCS, comparing the results to knockouts of the same proportions from low RCS cCREs. We then predicted gene expression following these perturbations.The perturbation analysis demonstrated decisive impact of high RCS cCREs on gene expression: removing low RCS cCREs did not significantly affect gene expression correlation, indicating minimal contribution. In contrast, removal of high RCSs cCREs caused significant drop in prediction correlation across all chromosomes (Fig. 6a). Significant difference between high and low RCS cCREs was observed (Wilcoxon test, $p < 0.0001$), confirming that the high RCS cCREs identified by the model have direct causal impact on gene expression (Fig. 6a).

For further validation, we examined the changes in gene expression values. From unseen HCT116 cell line, we selected 100 genes with the smallest prediction error on chromosome 6 as stable genes. Gradually removing 10%, 20%, and 30% of high RCS cCREs resulted in a noticeable decrease in the predicted expression values of most stable genes (Fig. 6b). EEF1E1[42] is a key factor in protein synthesis with a critical role in cellular physiology and pathology and showed significantly decreased expression levels when cCREs with top 10% RCSs were knocked out. Notably, these top 10% cCREs were predominantly located in enhancer and promoter regions.This indicates correlation reduction reflects the substantive contribution of key cCREs identified by DeepCis, consistent with experimental findings where enhancer knockouts decrease target gene expression. This perturbation experiment provides strong evidence for the causal impact of key cCREs and highlights the model's ability to accurately identify core functional cCREs in complex gene regulatory networks.

3. Conclusion

This study presents DeepCis, a novel framework for understanding gene expression regulation mechanisms through integration genomic, epigenomic, and 3D chromatin organization. By combining data across different scales, we capture the complex nature of gene regulation and identify potential CREs and their modes of action. Introduction of "regulatory contribution score (RCS)" as a quantitative measure enables clear assessment of each regulatory element's contribution to gene expression, enhancing the model interpretability and providing valuable biological insights into gene regulation mechanisms.

Using cCREs as a basic unit and constructing their contact matrices with dimensionality reduction strategy effectively reduces the computational burden when processing large-scale Hi-C data, improving the analysis efficiency. In capturing regulatory patterns, the multi-head attention mechanism captures both traditional short-range regulatory effects and the influence of distal cCREs. It successfully reveals the constraints of chromatin organization on gene regulation, indicating that the model understands not only simple linear distances but also how complex spatial organization restricts gene regulation.

DeepCis associate high RCSs cCREs with transcription factor binding sites, revealing which TFs play key roles in activation or repression of gene expression. This provides in-depth insights into how TFs regulate gene expression under different conditions, helping identifying and understanding critical TFs. DeepCis also identifies core regulatory hub cCREs overlapping super-enhancer regions that frequently appear in multiple gene regulatory networks and regulate the expression of key genes. Super-enhancers play a crucial role in cell identity, development, and disease processes. Therefore, these findings provide new insights into function of super-enhancer functions in gene expression regulation and offer important clues for future epigenetic research.

We validated the biological functions of key cCREs through enrichment analysis and CRISPR experiments, providing comprehensive biological validation of the model effectiveness and reliability. Additionally, we confirmed the key role of high RCS cCREs predicted by the model in gene expression

regulation through genomic perturbation experiments from computational perspective. These combined validation results demonstrate that the model achieves high accuracy at both biological and computational levels, providing strong theoretical and experimental support for deeper understanding of gene regulation mechanisms.

Despite the progress in improving prediction accuracy and interpretability of gene regulation mechanisms, this study still has certain limitations. First, while dimensionality reduction reduces the computational burden of Hi-C data through, processing extremely large datasets and higher-resolution Hi-C data may still face computational resources and time limitations. Second, although our model performs well in capturing long-range interactions between cCREs, in some cases, particularly with complex regulatory networks, the model may not fully capture all subtle interactions and regulatory mechanisms. Future research could further optimize the model structure to enhance handling of complex regulatory networks.

In summary, the predictive and interpretive capabilities of the DeepCis offer broad potential applications in clinical medicine, particularly in precision medicine and disease diagnosis. The model can accurately analyze the impact of non-coding region cCREs on gene regulation, helping to unravel how genes are influenced by multiple regulatory factors and capturing their dynamic interactions, which can assist clinicians in better understanding the mechanisms of diseases. By learning gene regulatory patterns in specific disease states and normal cell lines, the model can identify key cCREs driving diseases. These key cCREs help reveal the underlying regulatory mechanisms behind gene expression abnormalities and can serve as potential drug targets, enabling precise targeted therapies. Despite some limitations, with the continuous accumulation of data and optimization of algorithms, the accuracy and applicability of the model will be further enhanced, offering more personalized and efficient solutions in precision medicine and advancing medical technology.

4. Methods

4.1. Data collection and preprocessing

Data were obtained from publicly available databases, including candidate cis-regulatory elements (cCREs) in five human cell lines (GM12878, HepG2, K562, HeLaS3 and HCT116), chromatin accessibility and epigenomic modification datasets, 3D chromatin interaction maps(Hi-C), and transcription factor ChIP-seq peaks for each cell line. The details are as follows:

4.1.1. Selection of cCREs.

To derive a high-confidence dataset of cCREs from the SCREEN (https://screen.encodeproject.org/) database and to enhance computational efficiency in subsequent analyses, we applied a filtering strategy. Specifically, two classes of potentially low-quality cCREs were removed: (i) Low_DNase cCREs,

which may correspond to non-functional open chromatin regions; and (ii) DNase_only cCREs, characterized by DNase accessibility in the absence of corroborating epigenomic modifications.

4.1.2. DNA sequence extraction.

Genomic sequences of cCREs were extracted from the GRCh38 reference genome using bedtools getfasta (v2.31.1). For compatibility with the deep learning model, nucleotide bases were numerically encoded (A = 1, T = 2, C = 3, G = 4, N = 5). All extracted sequences were subsequently padded to a uniform length of 350 bp (P = 0), corresponding to the maximum cCREs length across cell types.

4.1.3. ChIP-seq data processing.

In this study, ChIP-seq data corresponding to 16 different assays for cCREs (CTCF, DNase, EP300, H2AFZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1, POLRZA, POLR2AphosphoS5) (Supplementary Table 1) were downloaded from the ENCODE database(https://www.encodeproject.org/) and processed to extract signal features and perform normalization. First, based on the genomic coordinates of the cCREs, signal values at single-base resolution were extracted from BigWig files using the pyBigWig library, generating a site-level signal matrix. The signal results from all chromosomes were then merged, followed by a log(1 + x) transformation. Subsequently, the features were normalized to the range [0, 1] using Min-Max scaling, preparing the data for downstream feature analysis and modeling.

4.1.4. Hi-C data processing.

Hi-C data from the 4D Nucleome(https://4dnucleome.org/) and ENCODE databases (Supplementary Table 1) were processed according to a standardized protocol. For the raw .hic files from the 4D Nucleome, we used the hic2cool (v0.1.1) tool to directly convert them into .cool format files with a 10 kb resolution. For the raw .hic files from ENCODE, we first extracted the 10 kb raw contact frequency matrices using the hicstraw library, and then generated .cool files at the same 10 kb resolution using the cooler tool. To eliminate systematic biases introduced by sequencing depth and technical preferences, all converted .cool format data were normalized using the Iterative Correction and Eigenvector Decomposition (ICE) method.

4.1.5. Construction of cCRE contact matrices based on Hi-C data.

To quantitatively analyze the spatial concact frequencies between cCREs, we constructed contact matrices with cCREs as the basic unit using Hi-C data at a 10 kb resolution. Each cCRE segment was mapped to the coordinate system of the Hi-C matrix. For any given cCRE segment, we identified all bins that overlap with its genomic location. If a segment spans multiple genomic bins, these bins collectively represent the position of the cCRE in the Hi-C contact matrix. The contact strength between two cCREs segments was defined as:

$$I_{ij} = \frac{1}{|B_i| \cdot |B_j|} \sum_{u \in B_i} \sum_{v \in B_j} M_{normalized}(u, v)$$

Where $M_{normalized}(u,v)$ represents the normalized Hi-C concact frequency between bin *i* and bin *j*, B_i is the set of Hi-C bins covered by $cCRE_i$, and B_j is the set of Hi-C bins covered by $cCRE_j$.$|B_i|$and $|B_j|$ represent the number of bins covered by each respective cCRE.

For each chromosome, cCREs contact matrix $M_{CRE} \in R^{n \times n}$was generated, where *n* denotes the number of valid cCREs successfully mapped on the chromosome.

4.1.6. Acquisition of condensed spatial features for cCREs.

To obtain the spatial features of each cCRE in three-dimensional space, we applied the Truncated Singular Value Decomposition (SVD) dimensionality reduction algorithm to map the high-dimensional contact matrix of cCREs to a lower-dimensional latent space. The top k principal components (where k = 32,64) were retained, ensuring that the cumulative explained variance exceeded 50%, preserving the essential topological and distance relationships. The resulting features were then scaled to the range [0, 1] using Min-Max scaling, ensuring consistent feature scaling and allowing the reduced vectors to capture spatial proximity information for the cCREs. In Hi-C data, concact frequencies reflect the proximity of chromatin fragments in three-dimensional space. cCREs with similar interaction patterns correspond to low-dimensional vectors with high similarity, indicating that these segments likely have spatial proximity in the three-dimensional genome and reflect co-localization trends. Therefore, the low-dimensional vectors also approximate the relative positioning of cCREs in three-dimensional space. Additionally, cCREs with similar features in certain principal component dimensions may share similar regulatory functions, reflecting cooperative regulatory trends. In summary, this spatial condensation method not only condenses the high-dimensional contact information between cCREs but also retains the core features of their spatial structure.

4.1.7. RNA data processing.

For each gene, the transcriptional start site (TSS) of the transcript with the highest expression was selected as the representative TSS. Additionally, the expression values of all transcripts corresponding to a gene were summed to represent the gene's overall expression level. To address data skewness and stabilize variance, a logarithmic transformation ($\log_{10}$ (TPM + 1)) was applied.

4.1.8. Selection of cCREs in the proximal regions of genes.

The TSS of each gene was used to identify cCREs in its upstream and downstream regions. The distance between all cCREs on the same chromosome and the TSS was calculated, and the 90 closest cCREs upstream and downstream were selected. To focus on those most likely involved in gene regulation, only cCREs within 1.5 Mb of the TSS were retained, a distance sufficient to capture typical cCREs.

4.1.9. Transcription factor and target genes retrieval.

Transcription factors (TFs) were retrieved by performing motif enrichment analysis on all cCREs for each cell line using Homer. Significant motifs ($P < 0.05$) were selected and mapped to corresponding TFs. To ensure data validity, only TFs with available ChIP-seq data in the ENCODE database were retained for analysis. We integrated the target genes of TFs from the TRRUST (https://www.grnpedia.org/trrust/) and hTFtarget (https://guolab.wchscu.cn/hTFtarget/#!/) databases.

4.1.10. Enhancer and Super-Enhancer Data Retrieval.

The enhancer data for GM12878, HepG2, K562, and HeLaS3 cell lines were downloaded from the FANTOM5 database (https://fantom.gsc.riken.jp/5/), while the super-enhancers for the GM12878, HepG2, K562, HeLaS3 and HCT116 cell lines were retrieved from the dbSUPER database (https://asntech.org/dbsuper/).

4.2 Dataset Construction

For each cell line, each gene was treated as a sample. The model's input for each gene comprised a concatenated sequence of its cCREs, which were individually represented by a 350-base sequence. Each base was encoded with a 6-dimensional one-hot vector and 16-dimensional epigenetic features. Spatial information was captured by a 32-dimensional spatial features for each element, supplemented by linear chromosomal positions and gene-level positional data. A mask vector was used to handle variable numbers of cCREs per gene. The target labels for the model were the transformed gene expression values.

To ensure the model focused solely on the impact of cCREs, gene sequences and their corresponding epigenetic signals at the gene locus were intentionally masked. This approach prevented information leakage and allowed the model to learn the regulatory relationship through the “pathway” between cCREs and their target genes.

To evaluate the model’s generalization and robustness across different genomic contexts, a chromosome-based partitioning strategy was employed. Chromosome 16 was designated as the validation set due to its enrichment in low-copy repeat sequences and structural variation hotspots[43,44], presenting a challenge for the model. Chromosome 6 was chosen as the test set, as its moderate length, high gene density, and complex regions [45–47]make it a representative and robust benchmark for evaluating performance on unseen genomic contexts. All other autosomes were used for training.

4.3 Model Design

4.3.1. Model Architecture Design.

DeepCis integrates information from cCREs and genes to predict target gene expression levels and model regulatory relationships using two cascaded Transformer encoders. The core of the Transformer encoder is the multi-head attention mechanism, where attention weights are calculated for each token in each head relative to other tokens in the sequence, capturing global contextual relationships[48]. The

outputs of all heads are concatenated and passed through a linear transformation to obtain the final multi-head output.

The model first maps genomic sequence features, epigenetic features, and spatial embedding features into continuous 256-dimensional vectors. These mapped features are concatenated along their dimensions and passed through a linear layer and layer normalization to form a unified 256-dimensional joint feature. In the feature extraction phase, this joint feature is input into a 1D convolutional module with a stride of 5 to extract local patterns and downsample the sequence length. The extracted feature sequence is then fed into the first Transformer encoder, where relative positional bias is applied to incorporate base positional information. The 256-dimensional representation of the final token from the encoder output is used as an initial feature for the cCREs or genes. The spatial embedding features are then added to this initial feature to form the final feature representation for each cCRE and gene.

In subsequent modeling, the target gene and its 180 corresponding cCREs serve as input sequences to the second Transformer encoder. Multi-scale relative positional biases are constructed based on the genomic coordinates of the cCREs and target gene, enabling the model to effectively incorporate the impact of three-dimensional spatial constraints on gene regulation at different resolutions. Finally, the output of the second encoder is layer-normalized, and the feature representation corresponding to the target gene token is extracted and mapped to a one-dimensional output via a fully connected layer to predict the target gene's expression.

We extracted the attention scores from the final layer of the second encoder and aggregated the attention values across all positions by averaging. The resulting score, termed the Regulatory Contribution Score (RCS), quantifies the degree of attention a gene gives to cCREs or the attention one cCRE gives to another, thereby quantifying the contribution of each cCRE to gene expression regulation.

4.3.2 Dual masks design.

To reflect the biological directionality of regulation, a base mask is introduced in the encoder to restrict information flow, allowing cCREs to focus on genes while limiting the reverse flow from cCREs to genes, modeling the directionality of regulation. Another mask in DeepCis allows for the selective masking of one or more cCREs within the input, enabling precise investigation of the role of cCREs.

4.3.3 Model Training.

The model was trained using the AdamW optimizer with a ReduceLROnPlateau learning rate scheduler, which dynamically adjusts the learning rate based on the validation loss. The loss function employed was mean squared error (MSE). To prevent overfitting, early stopping was applied, halting training if the validation loss did not improve for 15 consecutive epochs. Dropout with a rate of 0.1 was used during training. In the initialization phase, the model was pre-initialized with the CREaTor[49] pretrained model, with incompatible parameters initialized using Xavier and Kaiming random initialization. The DNA embedding layer was frozen, while all other layers were updated during training.

4.3.4. Model Comparison.

To further evaluate the performance of our model, we referred to the K562 functional enhancer dataset, validated through CRISPR/Cas9 gene editing, which is provided in CREaTor[23]. Using the results from CREaTor, we compared the performance of DeepCis, ABC, Enformer, GraphReg, and CREaTor in predicting enhancers in the K562 cell line.

4.4. Definition of Regulatory Contribution Element Grouping

For each gene, the corresponding cCREs are ranked by their contribution scores RCS from high to low. The top 10% of cCREs are defined as the "high regulatory contribution group," while the bottom 10% are defined as the "low regulatory contribution group."

4.5. Identification of Optimal Epigenomic Feature Combinations

We performed an exhaustive search to identify the epigenomic features combinations that maximize the Pearson correlation coefficient on the test set for chromosome 6, retaining 2 to 6 features for each cell line.

4.6. TAD Identification

In this study, we used 40 kb resolution Hi-C data and the hicFindTADs tool to identify TADs in various cell lines.

4.7. Validation of regulatory direction using TCGA data

To validate the regulatory direction predicted by DeepCis, we acquired TCGA datasets from the Genomic Data Commons (GDC) Data Portal(https://portal.gdc.cancer.gov/), encompassing multiple cancer types: diffuse large B-cell lymphoma (TCGA-DLBC), liver hepatocellular carcinoma (TCGA-LIHC), acute myeloid leukemia (TCGA-LAML), cervical squamous cell carcinoma and endocervical adenocarcinoma (TCGA-CESC), and colon adenocarcinoma (TCGA-COAD).For samples in these datasets, we calculated the correlation between the expression levels of TFs and their corresponding target genes. We then compared these correlations with the relationships between the predicted RCSs and the ChIP-seq signal intensities of the corresponding TFs. By examining whether the signs of the correlations were consistent, we validated the consistency of the regulatory direction. A match in sign between the correlation of transcription factor-target gene expression and the model's predicted RCS scores indicates that the regulatory direction is aligned, supporting the accuracy of our model.

Declarations

Conflicts of Interest

The authors declare no conflicts of interest.

Funding:

This work was supported by the National Natural Science Foundation of China (32170674).

Author Contribution

H.Z., S.W.N. and P.W. conceived the research direction and guided the research design. X.R.L. designed the study, drafted the initial manuscript and performed data analysis. W.B.D., J.J.D., M.M.L. and S.S.W. conducted data collection. M.Y.X., Y.L.X., H.B.Z. and Y.S.Z. verified the research results. W.F. and Y.J.L. were responsible for data visualization.All authors critically revised the manuscript for important intellectual content and approved the final version for publication.

Acknowledgements

The author thanked the National Natural Science Foundation of China (32170674).

Data Availability

The data that support the findings of this study are available upon request. The following databases were used in this research: SCREEN ([http://screen.encodeproject.org/](http:/screen.encodeproject.org)), ENCODE ([https://www.encodeproject.org/](https:/www.encodeproject.org)), 4D Nucleome ([https://4dnucleome.org/](https:/4dnucleome.org)), FANTOM5 ([https://fantom.gsc.riken.jp/5/](https:/fantom.gsc.riken.jp/5)), TRRUST ([https://www.grnpedia.org/trrust/](https:/www.grnpedia.org/trrust)), hTFtarget ([https://guolab.wchscu.cn/hTFtarget/#!/](https:/guolab.wchscu.cn/hTFtarget)), Genomic Data Commons (GDC) ([https://portal.gdc.cancer.gov/](https:/portal.gdc.cancer.gov)), and dbSUPER ([https://asntech.org/dbsuper/](https:/asntech.org/dbsuper)).

References

1. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).
2. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
3. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).
4. Cis-Regulatory Alterations in FOXP4 Modulate Esophageal Cancer Susceptibility Induced by Chronic Alcohol Exposure - PubMed. https://pubmed.ncbi.nlm.nih.gov/40228145/.

5. Castro-Mondragon, J. A. *et al.* Cis-regulatory mutations associate with transcriptional and post-transcriptional deregulation of gene regulatory programs in cancers. *Nucleic Acids Res.* **50**, 12131–12148 (2022).
6. Kim, K. L. *et al.* Dissection of a CTCF topological boundary uncovers principles of enhancer-oncogene regulation. *Mol. Cell* **84**, 1365-1376.e7 (2024).
7. Zhao, S. G. *et al.* Integrated analyses highlight interactions between the three-dimensional genome and DNA, RNA and epigenomic alterations in metastatic prostate cancer. *Nat. Genet.* **56**, 1689–1700 (2024).
8. ENCODE data at the ENCODE portal - PubMed. https://pubmed.ncbi.nlm.nih.gov/26527727/.
9. Thomas, H. F. *et al.* Enhancer cooperativity can compensate for loss of activity over large genomic distances. *Mol. Cell* **85**, 362-375.e9 (2025).
10. Costa, J. R. *et al.* Transcription factor cooperativity at a GATA3 tandem DNA sequence determines oncogenic enhancer-mediated activation. *Cell Rep.* **44**, 115705 (2025).
11. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning | Nature Communications. https://www.nature.com/articles/s41467-021-23576-0#ethics.
12. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
13. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing | Genome Biology. https://link.springer.com/article/10.1186/gb-2010-11-2-r22.
14. HiChIP: efficient and sensitive analysis of protein-directed genome architecture | Nature Methods. https://www.nature.com/articles/nmeth.3999.
15. Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq) | SpringerLink. https://link.springer.com/protocol/10.1007/978-1-0716-1597-3_10.
16. Fulco, C. P. *et al.* Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
17. Ahmed, F. S., Aly, S. & Liu, X. EPI-Trans: an effective transformer-based deep learning model for enhancer promoter interaction prediction. *BMC Bioinformatics* **25**, 216 (2024).
18. Zhang, T., Zhao, X., Sun, H., Gao, B. & Liu, X. GATv2EPI: Predicting Enhancer-Promoter Interactions with a Dynamic Graph Attention Network. *Genes* **15**, 1511 (2024).
19. KansformerEPI: a deep learning framework integrating KAN and transformer for predicting enhancer-promoter interactions - PubMed. https://pubmed.ncbi.nlm.nih.gov/40515390/.
20. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
21. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
22. Chromatin interaction-aware gene regulatory modeling with graph attention networks - PubMed. https://pubmed.ncbi.nlm.nih.gov/35396274/.

23. Li, Y. *et al.* CREaTor: zero-shot cis-regulatory pattern modeling with attention mechanisms. *Genome Biol.* **24**, 266 (2023).
24. Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing | Genome Medicine. https://link.springer.com/article/10.1186/s13073-018-0538-1.
25. Precise targeting of POLR2A as a therapeutic strategy for human triple negative breast cancer | Nature Nanotechnology. https://www.nature.com/articles/s41565-019-0381-6.
26. Bclaf1 mediates super-enhancer-driven activation of POLR2A to enhance chromatin accessibility in nitrosamine-induced esophageal carcinogenesis. *J. Hazard. Mater.* **492**, 138218 (2025).
27. Why the activity of a gene depends on its neighbors: Trends in Genetics. https://www.cell.com/trends/genetics/abstract/S0168-9525(15)00129-8.
28. Schindler, M. *et al.* Induction of kidney-related gene programs through co-option of SALL1 in mole ovotestes. *Development* **150**, dev201562 (2023).
29. Young LINE-1 transposon 5' UTRs marked by elongation factor ELL3 function as enhancers to regulate naïve pluripotency in embryonic stem cells | Nature Cell Biology. https://www.nature.com/articles/s41556-023-01211-y.
30. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation | Nature Structural & Molecular Biology. https://www.nature.com/articles/s41594-020-00539-5.
31. SMARCB1 loss activates patient-specific distal oncogenic enhancers in malignant rhabdoid tumors | Nature Communications. https://www.nature.com/articles/s41467-023-43498-3.
32. TAD boundary deletion causes PITX2-related cardiac electrical and structural defects | Nature Communications. https://www.nature.com/articles/s41467-024-47739-x.
33. CUX1 restrains latent hematopoietic stem cell plasticity by suppressing stem cell-intrinsic inflammatory pathways. *Blood* https://doi.org/10.1182/blood.2024026815 (2025) doi:10.1182/blood.2024026815.
34. FOXP1 directly represses transcription of proapoptotic genes and cooperates with NF-κB to promote survival of human B cells | Blood | American Society of Hematology. https://ashpublications.org/blood/article/124/23/3431/33456/FOXP1-directly-represses-transcription-of.
35. Mechanism of ERBB2 gene overexpression by the formation of super-enhancer with genomic structural abnormalities in lung adenocarcinoma without clinically actionable genetic alterations | Molecular Cancer. https://link.springer.com/article/10.1186/s12943-024-02035-6.
36. Super-enhancer-driven SLCO4A1-AS1 is a new biomarker and a promising therapeutic target in glioblastoma | Scientific Reports. https://www.nature.com/articles/s41598-024-82109-z.
37. Dysregulated CREB3 cleavage at the nuclear membrane induces karyoptosis-mediated cell death | Experimental & Molecular Medicine. https://www.nature.com/articles/s12276-024-01195-1.
38. TLN1: an oncogene associated with tumorigenesis and progression | Discover Oncology. https://link.springer.com/article/10.1007/s12672-024-01593-x.

39. Targeting CA9 restricts pancreatic cancer progression through pH regulation and ROS production | Cellular Oncology. https://link.springer.com/article/10.1007/s13402-024-01022-9.
40. The diverse repertoire of ISG15: more intricate than initially thought | Experimental & Molecular Medicine. https://www.nature.com/articles/s12276-022-00872-3.
41. ATAD3A: A Key Regulator of Mitochondria-Associated Diseases. https://www.mdpi.com/1422-0067/24/15/12511.
42. A Novel HCC Prognosis Predictor EEF1E1 Is Related to Immune Infiltration and May Be Involved in EEF1E1/ATM/p53 Signaling - PubMed. https://pubmed.ncbi.nlm.nih.gov/34282404/.
43. Clinical Implications of Chromosome 16 Copy Number Variation | Molecular Syndromology | Karger Publishers. https://karger.com/msy/article-abstract/13/3/184/825184/Clinical-Implications-of-Chromosome-16-Copy-Number?redirectedFrom=fulltext.
44. Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988–994 (2004).
45. Copy Neutral LOH Affecting the Entire Chromosome 6 Is a Frequent Mechanism of HLA Class I Alterations in Cancer. https://www.mdpi.com/2072-6694/13/20/5046.
46. DNA Methylation and Transcription of HLA-F and Serum Cytokines Relate to Chinese Medicine Syndrome Classification in Patients with Chronic Hepatitis B | Chinese Journal of Integrative Medicine. https://link.springer.com/article/10.1007/s11655-021-3279-8.
47. Thomas, C. *et al.* TERT promoter mutation and chromosome 6 loss define a high-risk subtype of ependymoma evolving from posterior fossa subependymoma. *Acta Neuropathol. (Berl.)* **141**, 959–970 (2021).
48. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
49. Li, Y. *et al.* CREaTor: zero-shot cis-regulatory pattern modeling with attention mechanisms. *Genome Biol.* **24**, 266 (2023).

Figures

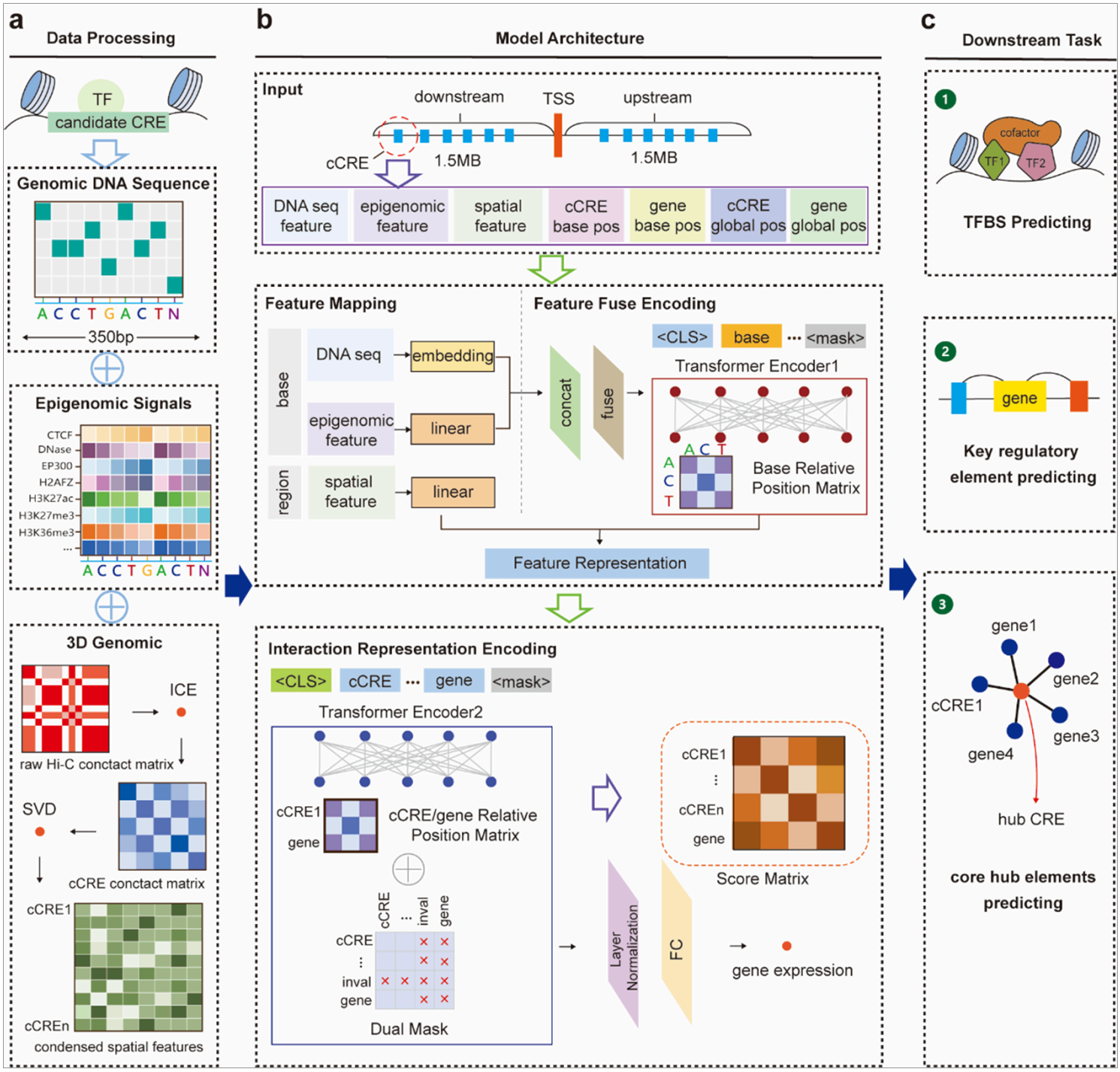


Figure 1

Overview of the DeepCis framework. a Data preprocessing.Fine-grained features of cCREs are extracted at the base pair level from genomic sequence and epigenomic signals, and condensed spatial features of cCREs are obtained at the region level. **b** DeepCis architecture. The framework consists of the sample input module, feature mapping module, feature fusion encoding module, and interaction representation encoding module, with input features from cCREs within 1.5 MB upstream and downstream of the TSS. **c** Tasks achievable with DeepCis.

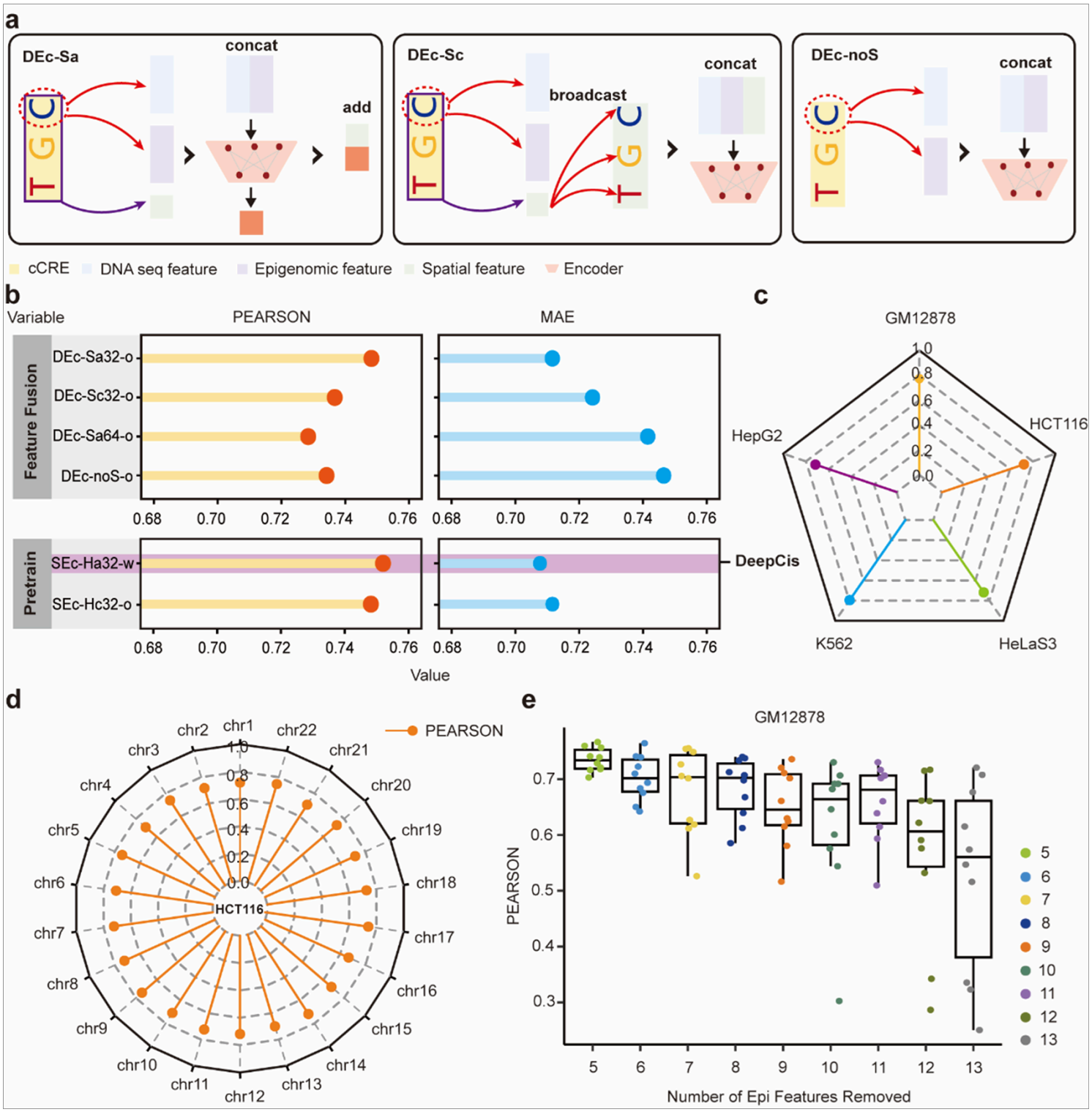


Figure 2

Robustness of DeepCis in gene expression prediction. **a** The impact of different feature fusion strategies on model performance. **b** Performance of different fusion strategies and spatial feature dimensions (32 and 64) on the test set, comparing the difference between initializing with a pretrained model (w) and without (o). **c** Prediction results on chromosome 6 of the test set across different cell lines. **d** Prediction performance on various chromosomes in the unseen HCT116 cell line. **e** Prediction results after

randomly removing a subset of epigenomic features (removal range: 5-13 features, iterated 10 times) and evaluating robustness using the Pearson correlation coefficient.

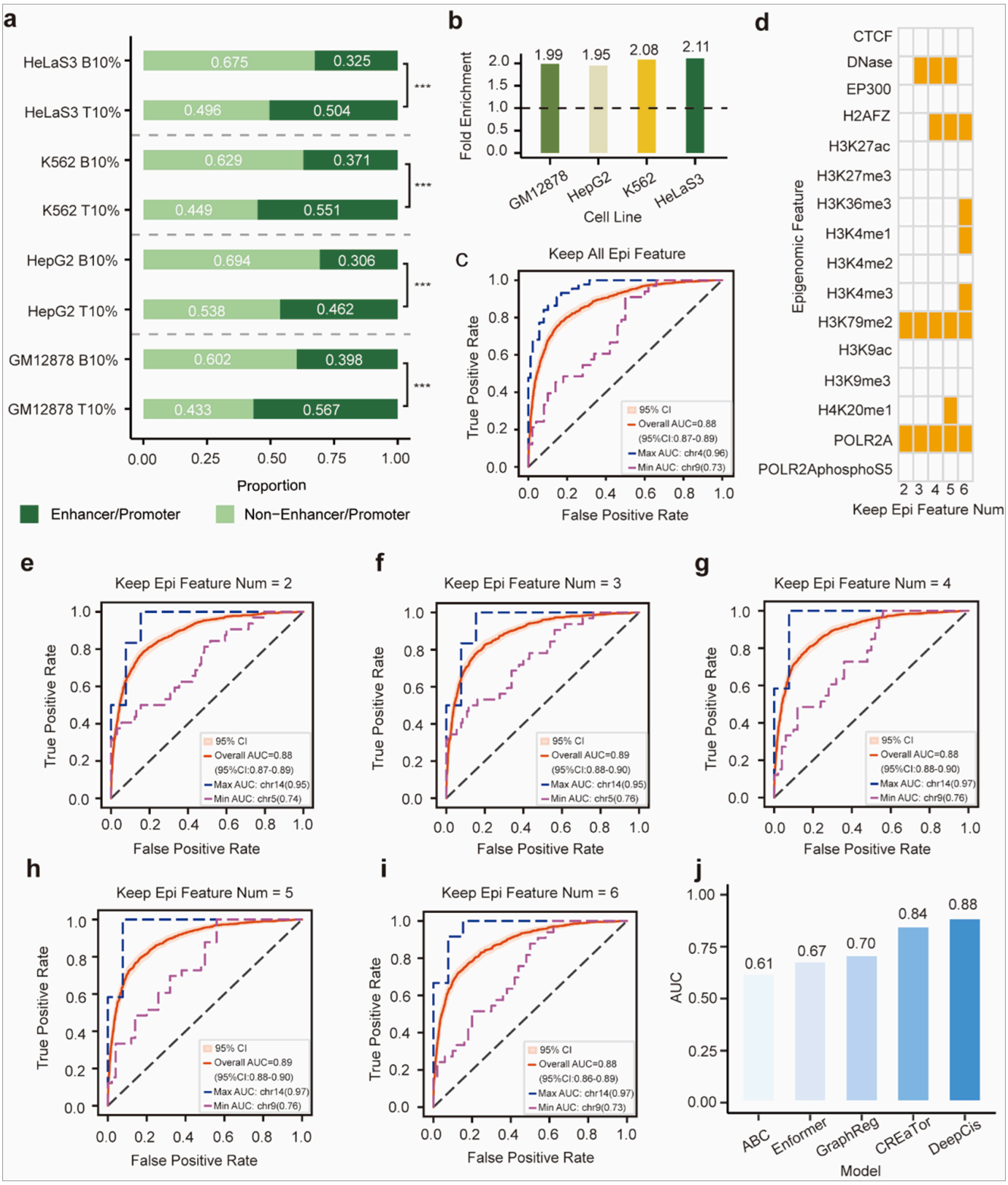


Figure 3

Quantification and identification of key cCREs by DeepCis. **a** Enrichment distribution of enhancers and promoters in high and low regulatory contribution group. **b** Fold enrichment of enhancers and promoters in high regulatory contribution group relative to low regulatory contribution group across cell lines. **c** AUC

performance of the model on experimentally validated K562 enhancer data with all input Epigenomic features. **d** Epigenomic features that correspond to the maximum Pearson correlation coefficient in K562 prediction when retaining 2–6 features. **e–i** Overall AUC of the model across all chromosomes and the maximum and minimum AUC values for each chromosome when retaining the optimal number (2–6) of features. **j** Performance comparison of different models based on AUC.

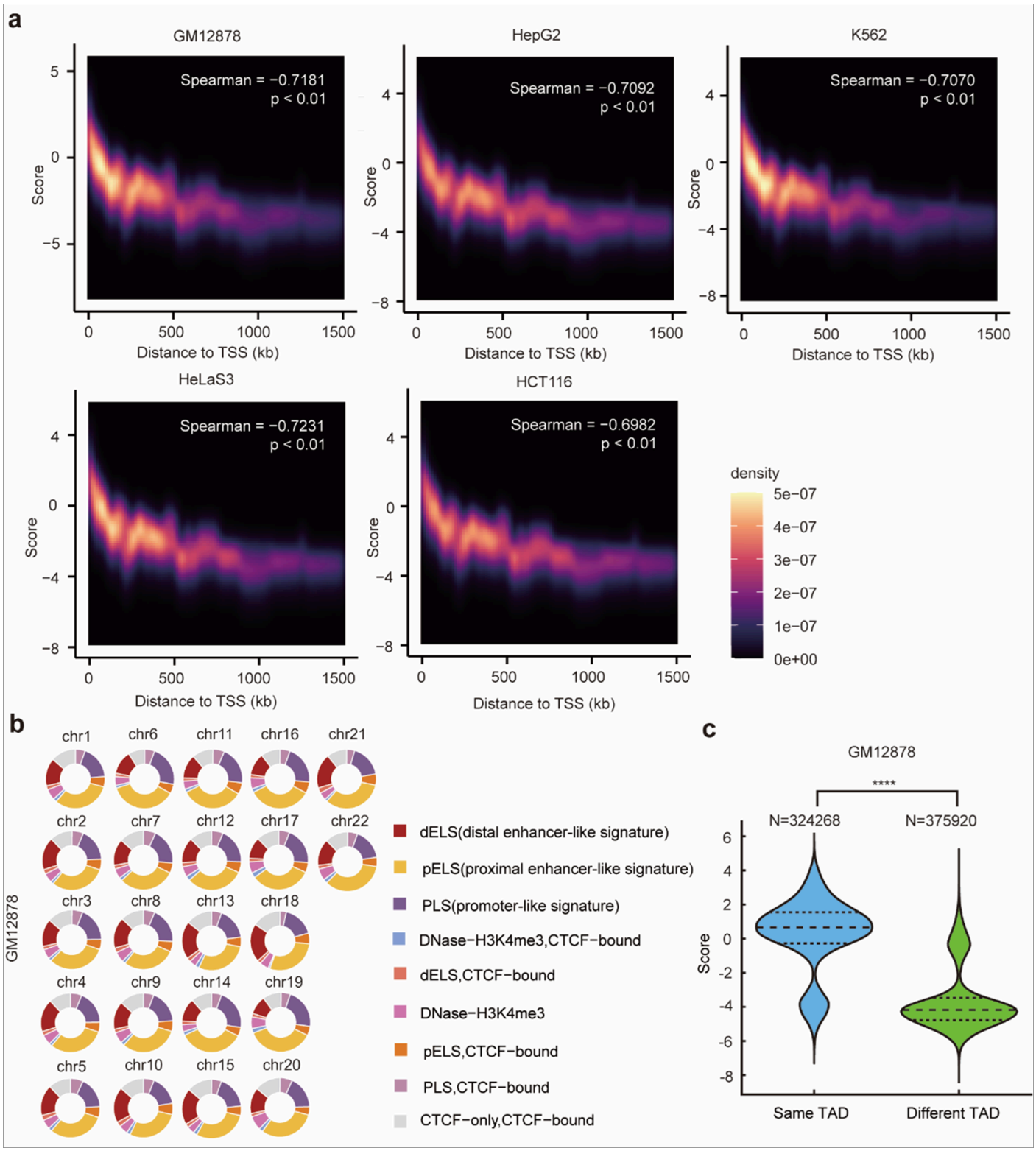


Figure 4

DeepCis's ability to capture different cCREs. a Spearman correlation between the regulatory contribution scores of cCREs and their distance to the gene TSS. **b** Distribution of the types of the top ten high-contribution cCREs for genes across chromosomes in the GM12878 cell line. c Differences in the co-localization of genes and their top ten high-contribution cCREs within TADs in the GM12878 cell line.

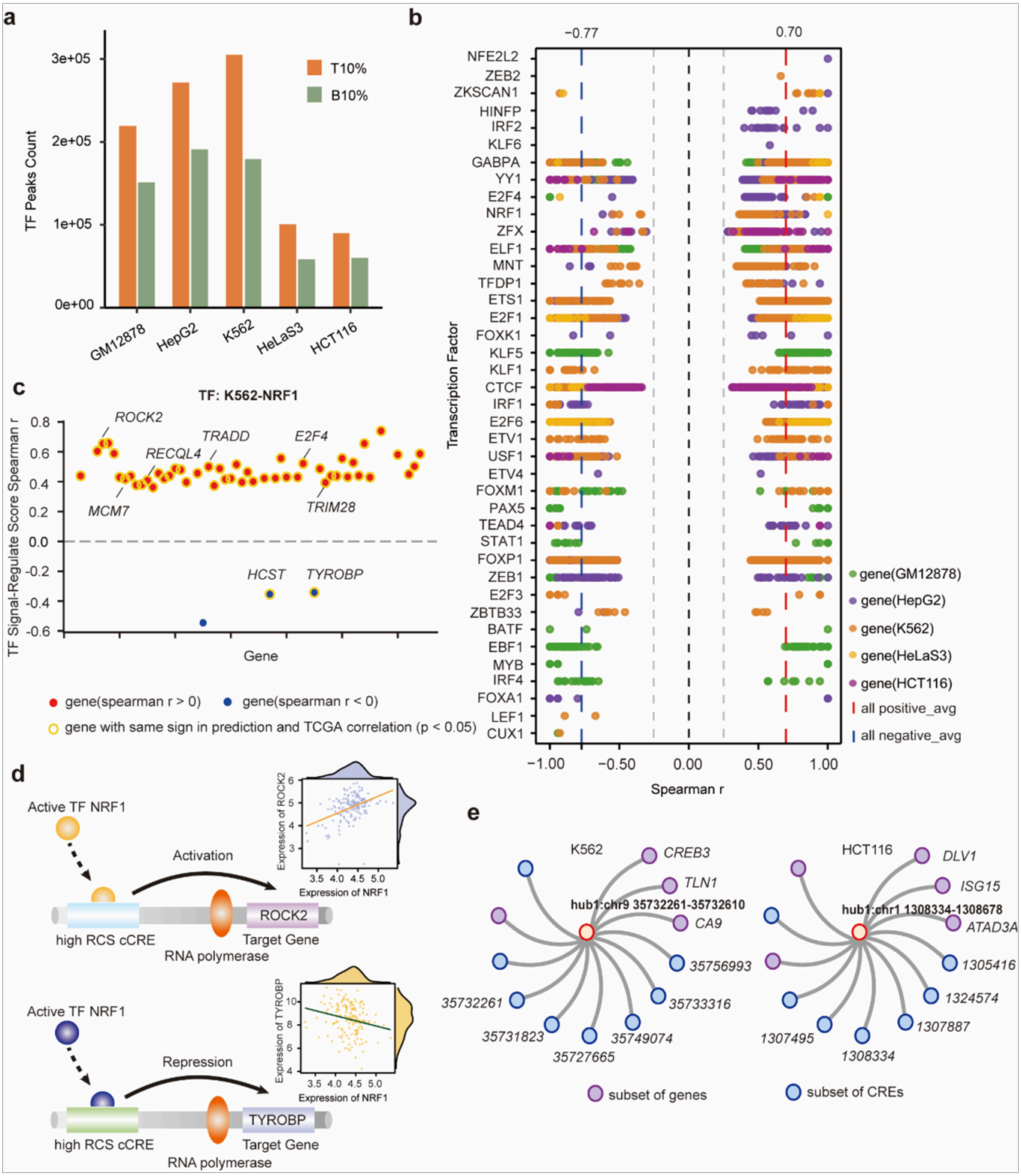


Figure 5

Identification of core TFs and prediction of key cCREs by DeepCis. a Comparison of the enrichment of transcription factor binding peaks on cCREs in the high and low regulatory contribution groups across different cell lines.**b** For each transcription factor (y-axis), each point represents a target gene associated with that factor. The x-axis shows the Spearman correlation coefficient, measuring the correlation between the model-predicted regulatory scores and TF ChIP-seq signal intensity for cCREs with binding peaks on the gene. Positive correlations indicate that the transcription factor tends to promote the target gene's transcriptional activity, while negative correlations suggest a repressive effect. **c** Comparison of the predicted regulatory directions from DeepCis with the correlation directions between TFs and target gene expression levels in TCGA samples, using NRF1 in the K562 cell line as an example. **d** Two distinct regulatory modes, transcriptional activation and transcriptional repression, and their regulatory mechanisms.**e** Example of DeepCis predicting core hub cCREs, showing that such cCREs can simultaneously influence both the target gene and other cCREs.

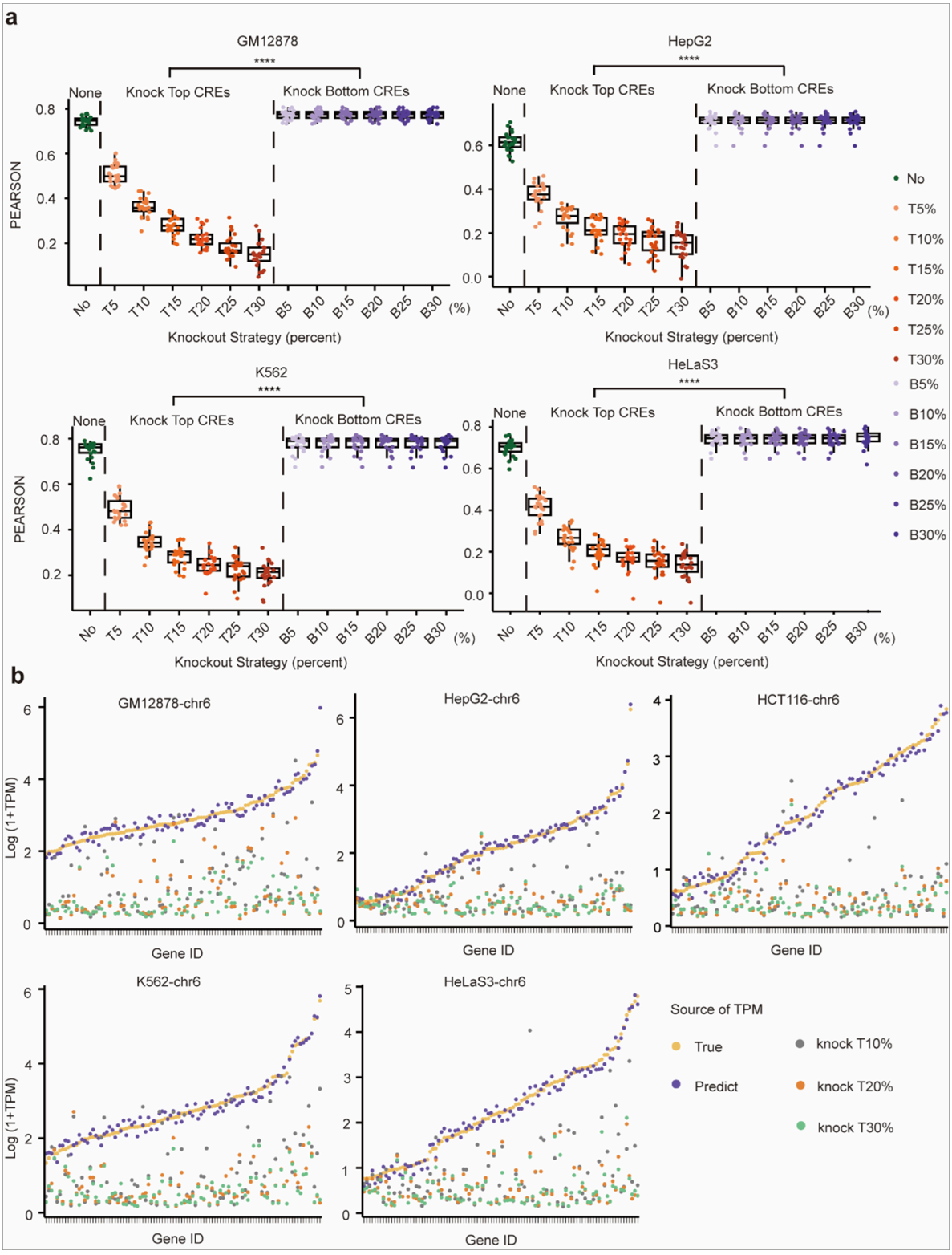


Figure 6

Perturbation Analysis Results. a Model performance across different cell lines. Each data point represents a chromosome, with the y-axis showing the Pearson correlation coefficient between predicted values and gene knockdown experimental observations. The x-axis represents percentile ranges of regulatory contribution scores, where "No" indicates no element knockdown, “T” refers to the top 5–30% of cCREs ranked by DeepCis predicted RCSs, and “B” refers to the bottom 5–30%. **b**

Comparison of gene expression levels before and after gene knockdown for specific genes on chromosome 6 across different cell lines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supportingInformation.docx