Nucleic Acids Research, 2026, **54**, gkag485
https://doi.org/10.1093/nar/gkag485
Methods

UniversalEPI: robust prediction of cell type-specific and differential chromatin interactions from DNA sequence and chromatin accessibility

Aayush Grover[1,2], Lin Zhang[3], Till Muser[3], Simeon Häfliger[1], Minjia Wang[1,10], Josephine Yates[1,2], Marie-Claire Indilewitsch[1], Yizhen Wang[1], Eliezer M. Van Allen[4,5,6], Fabian J. Theis[7,8], Ignacio L. Ibarra[7,*], Ekaterina Krymova[3,*], Valentina Boeva[1,2,9,*]

[1]Department of Computer Science, ETH Zurich, Zurich 8092, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
[3]Swiss Data Science Center, EPF Lausanne and ETH Zurich, Zurich 8092, Switzerland
[4]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, United States
[5]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, United States
[6]Division of Medical Sciences, Harvard University, Boston, MA 02115, United States
[7]Institute of Computational Biology, Helmholtz Center Munich, Neuherberg 85764, Germany
[8]School of Computation, Information and Technology, Technical University of Munich, Garching bei München 85748, Germany
[9]Université Paris Cité, Institut Cochin, INSERM U1016, Paris 75014, France
[10]Present address: School of Engineering and Applied Sciences, Harvard University, Boston, MA 02134,United States

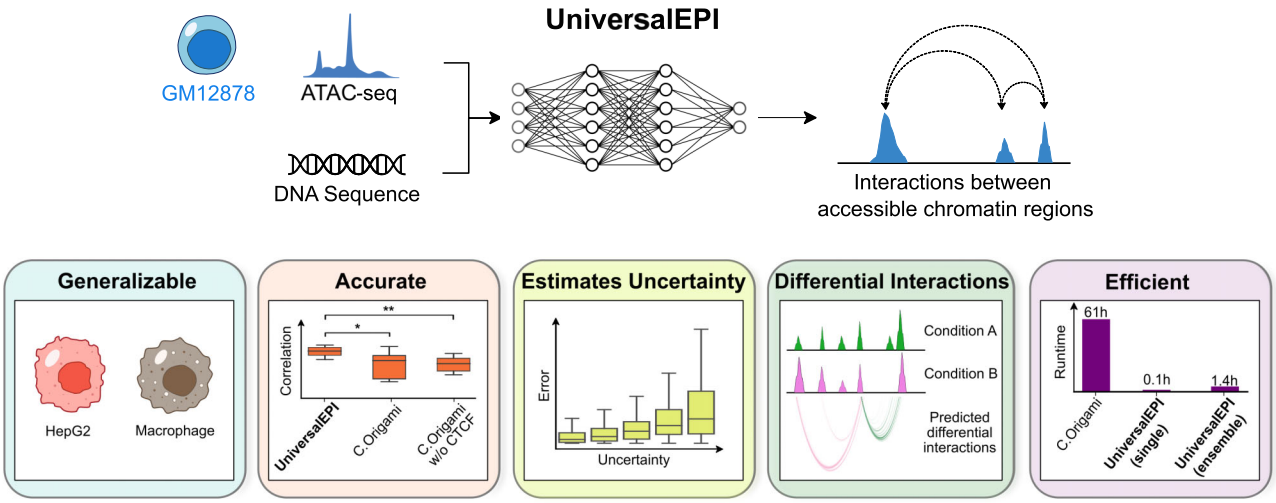
*To whom correspondence should be addressed. Email: valentina.boeva@inf.ethz.ch
Correspondence may also be addressed to Ignacio L. Ibarra. Email: ignacio.ibarra@helmholtz-muenchen.de
Correspondence may also be addressed to Ekaterina Krymova. Email: ekaterina.krymova@sdsc.ethz.ch

Abstract

Enhancer–promoter interactions (EPIs) play a central role in gene regulation, but experimental techniques such as Hi-C for mapping these interactions remain costly and labor-intensive. Computational methods have been developed to predict EPIs *in silico* from DNA sequence and chromatin information; however, there are major challenges with the generalizability and accuracy of predictions by existing methods across cell types and conditions unseen during model training. We developed and validated UniversalEPI, an attention-based deep ensemble model that predicts EPIs up to 2 Mb apart using only DNA sequence and chromatin accessibility (ATAC-seq) data. Unlike models that reconstruct full Hi-C contact maps, UniversalEPI focuses on biologically relevant, sparse chromatin interactions between accessible regulatory elements. It generalizes across both bulk and single-cell ATAC-seq-derived pseudo-bulk datasets, delivering state-of-the-art performance while using fewer input modalities than existing approaches. By modeling predictive uncertainty, UniversalEPI enables statistically robust differential analysis of chromatin interactions across conditions. We demonstrate its utility by tracking dynamic EPIs during human macrophage activation and identifying regulatory differences between cancer cell states in esophageal adenocarcinoma. By providing precalculated Hi-C predictions for 157 ENCODE datasets, UniversalEPI expands the scope and applicability of *in silico* 3D genome modeling for studying gene regulation in development and disease.

Received: July 3, 2025. **Revised:** April 7, 2026. **Accepted:** April 26, 2026

Graphical abstract



Introduction

In complex organisms, different cell types have distinct transcriptional programs that lead to their specific functions. Chromatin interactions between gene promoters and *cis*-regulatory elements, specifically enhancers, modulate gene expression through the formation of three-dimensional chromatin architecture. These interactions facilitate the recruitment of ubiquitously expressed or tissue-specific transcription factors (TFs), coactivators, and the basal transcription machinery to gene promoters, enabling precise spatial and temporal control of gene expression [1]. Enhancer–promoter looping is mediated by architectural proteins (e.g. CTCF, YY1) and protein complexes (e.g. cohesin and mediator), which help establish and stabilize these chromatin contacts. CTCF facilitates long-range chromatin interactions by organizing topologically associating domains (TADs), boosting enhancer–promoter interactions (EPIs) within the same domain [2, 3]. The zinc-finger TF YY1 acts as a structural tether by forming DNA loops that bring enhancers and promoters into close proximity [4]. Several lines of evidence support that the TF SP1 contributes to chromatin architecture at gene promoters [5–7].

The dynamic nature of EPIs allows cells to respond to developmental cues, environmental signals, and other regulatory inputs, ensuring appropriate gene expression that is necessary for cellular identity and function [8, 9]. Disruptions in chromatin interactions, such as through mutations or chromosome instability, are associated with various genetic diseases, including cancer [10, 11]. Chromatin interactions can be profiled experimentally using Hi-C, an unbiased but costly and labor-intensive high-throughput chromosome conformation capture (3C) technique. Because Hi-C is resource-intensive, attention has shifted to computational inference; the challenge is to have a universal model that generalizes across cell types, uses minimal input modalities, and yields quantitative outputs for differential interaction testing across DNA-sequence-level or chromatin-state changes.

Several recent studies have proposed deep learning-based architectures to predict the strength of EPIs. These include DeepC, Akita, and Orca, which predict chromatin interactions in a specific cell type solely from the DNA sequence [12–14]. Because these methods rely only on DNA sequence, they can be used to assess the effects of genomic variants on chromatin structure in the training cell type. Although they generate realistic Hi-C maps, their reliance on DNA sequence alone means they lack cell-type-specific regulatory information, limiting generalization to unseen cell types without retraining.

In contrast, methods such as HiC-Reg, TargetFinder, Epiphany, and TransEPI do not use DNA sequence as input but instead rely on profiles of several epigenetic features, such as chromatin accessibility and histone post-translational modifications [15–18]. These additional epigenetic signals enable these models to generalize across cell types by providing information on cell-type-specific regulatory activity. However, such epigenetic inputs are not always readily available, and because these methods do not use DNA sequence directly, they cannot capture sequence-encoded mechanisms of chromatin folding and therefore cannot predict the effects of genomic variants on 3D genome organization.

Some hybrid EPI models, such as DeepTACT, DeepPHiC, and ChINN, incorporate both DNA sequence and cell-type-specific chromatin accessibility at the interacting anchors, as well as the distance between the anchors [19–21]. A potential limitation of this anchor-only formulation is that it does not explicitly model the intervening genomic context between the two anchors, which can contain insulators or additional regulatory elements (e.g. CTCF or YY1 binding sites) that modulate looping and enhancer–promoter communication.

The C.Origami and EPCOT methods currently provide state-of-the-art solutions for predicting cell-type-specific chromatin interactions from bulk data by leveraging DNA sequence and chromatin accessibility within large genomic windows, thereby incorporating the intervening regulatory context between interacting loci [22, 23]. Both approaches combine convolutional neural networks (CNNs) with transformer architectures but differ in the required input modalities and predicted outputs. C.Origami additionally requires CTCF ChIP-seq for modeling, whereas EPCOT relies only on DNA sequence and chromatin accessibility profiles. Notably, EPCOT generates observed-over-expected contact maps, which do not provide information on absolute interaction frequencies, whereas C.Origami predicts contact intensities on a quantitative scale (i.e. not distance-normalized), enabling

comparisons across conditions and more direct interpretation of biologically meaningful differences in chromatin interactions. Due to their complexity and model size, C.Origami was initially trained on a single cell line (IMR90), whereas EPCOT was trained on data from four cell lines (K562, MCF-7, GM12878, and HeLa-S3).

More recently, ChromaFold was introduced to predict chromatin interactions using single-cell ATAC-seq (scATAC-seq) data and occurrences of CTCF motifs derived from the DNA sequence [24]. ChromaFold addresses several limitations of the earlier methods: it eliminates the requirement for CTCF ChIP-seq data, significantly reduces training time, and efficiently incorporates data from multiple cell types. However, ChromaFold is unlikely to fully capture the effects of other regulatory TFs on chromatin folding, as it considers only CTCF-binding motifs. Furthermore, ChromaFold predicts distance- and GC-content–normalized Z-scores rather than quantitative contact intensities. It was also developed specifically for single-cell inputs. Importantly, neither ChromaFold nor other state-of-the-art methods output uncertainty estimates along with predictions; such estimates enable the identification of statistically significant changes in EPIs across conditions, *e.g.* variation in the strength of EPIs due to DNA variants or epigenetic differences across transcriptional cell states. Thus, there remains a need for a fast and accurate method that (i) generalizes to unseen cell types using widely available inputs, (ii) predicts quantitative interaction strengths (retaining distance effects), and (iii) enables statistically robust differential analysis through uncertainty estimates.

In this work, we introduce UniversalEPI, a lightweight and accurate deep-ensemble model consisting of CNN layers and transformer blocks. This method addresses key limitations of prior work by training on automatically extracted DNA-binding motifs of three ubiquitously expressed TFs, CTCF, YY1, and SP1 [25–27], complemented with the ATAC-seq profile from chromatin-accessible regions. This choice of architecture allows for reducing noise and model size while still capturing all functionally relevant interactions in large genomic domains (up to 2 Mb). Furthermore, we incorporate uncertainty estimation by integrating both aleatoric (data) and epistemic (model) uncertainty using stochastic training loss [28] and deep ensemble [29], enhancing the predictive reliability of the model. Therefore, unlike earlier methods that provide only point estimates, our approach also quantifies prediction confidence. Using the reported maximum-confidence fold-change (FC), a unique feature of UniversalEPI that removes inherent data noise, the user can focus on EPIs that are significantly altered across conditions.

After training and testing UniversalEPI on four human cell lines with all necessary input data, we benchmarked UniversalEPI against the state-of-the-art methods. We found its performance in predicting chromatin interactions in unseen human cell types was significantly superior to methods such as C.Origami, Akita, EPCOT, and ChromaFold. We showed that UniversalEPI can be used to assess chromatin dynamics across conditions by using data generated by Reed *et al.* on human macrophage activation [30]. The uncertainty-aware UniversalEPI predictions agreed with the ground-truth Hi-C measurements with Spearman's correlation above 0.9. Finally, we showed that UniversalEPI can predict Hi-C interactions from pseudo-bulks of scATAC-seq data profiles and reveal chromatin dynamics across undifferentiated and differentiated cell states in human esophageal adenocarcinoma (EAC), specifically in promoters of genes encoding master transcriptional regulators. Together, these findings show that UniversalEPI is a lightweight, generalizable model that accurately predicts EPIs using bulk or scATAC-seq and DNA sequencing data in unseen cell types while indicating a level of uncertainty. This model now makes it possible to carry out *in silico* experiments to assess changes in chromatin folding under diverse biological scenarios.

Materials and methods

Data processing

Hi-C data processing

The raw hg38-aligned Hi-C files for four cell lines (GM12878, K562, IMR90, and HepG2) were obtained directly from the 4D Nucleome Data Portal [31] under the accession codes listed in Supplementary Table S1. We then extracted the 5-kb-resolution contact matrices. To remove any systematic biases, we applied ICE normalization [32] using HiCExplorer v2.2.1 [33] with filter thresholds of −1.1 and 4.5. Since we are only interested in intra-chromosomal contacts, the ICE normalization was also done independently on each chromosome of each cell line.

The Hi-C contact matrices must be normalized across cell lines for the model predictions to be comparable to the training data. We achieved this by applying a distance-stratified robust z-score normalization using the healthy GM12878 cell line as a reference. GM12878 was selected as a reference cell line due to the high sequencing depth of its Hi-C matrix and, hence, better data quality (∼759M interactions as compared to ∼275M interactions in the other cell lines). For each chromosome of the GM12878 cell line, we stored the median and median absolute deviation (MAD) of all measured Hi-C interactions corresponding to regions that are d bins apart where $d \in [0, 800]$ (each bin being 5 kb of the genome). In some chromosomes, there were too few long-range interactions for accurate calculations of median and MAD. Hence, we fitted a B-spline, using SciPy's splrep function [34], setting the smoothing condition s to 5. This spline was then used to smooth our median and MAD curves across the genomic distance. We denote the smoothed median and MAD as $m_d^{GM12878}$ and $\text{mad}_d^{GM12878}$, respectively. For a new cell type, e.g. the K562 cell line, the normalized Hi-C score ($\hat{y}$) between regulatory elements in bins i and j is then calculated as done in equation (1),

$$\hat{y}_{i,j}^{K562} = \left(\frac{y_{i,j}^{K562} - m_{|j-i|}^{K562}}{\text{mad}_{|j-i|}^{K562}} \right) \text{mad}_{|j-i|}^{GM12878} + m_{|j-i|}^{GM12878}, \quad (1)$$

where y is the ICE-normalized Hi-C value. Supplementary Fig. S1 depicts the effect of this normalization on all cell lines.

Fudenberg *et al.* [13] applied Gaussian smoothing with unit variance to fill in the missing interactions and reduce the noise of the Hi-C experiment. However, this can lead to a significant reduction in strong Hi-C contacts in the presence of missing values at shorter distances. To mitigate this issue, we applied Gaussian smoothing with distance-dependent variance, which increased with distance (Supplementary Fig. S2). Particularly, we used four kernels with variances of 0.5, 0.65, 0.8, and 1 for interactions between regions that are closer than 25 kb, between 25 and 100 kb, between 100 and 250 kb, and >250 kb, respectively. Finally, we applied a logarithmic transformation to the interaction scores. Several methods, like Akita [13] and

ChromaFold [24], also remove the effect of distance from the model by applying z-scores or dividing the observed interactions by expected interactions. While this approach highlights long-range interactions, we aim to capture the true biological interactions and hence, retain the effect of distance in our Hi-C matrices.

Micro-C data processing

The hg38-aligned Micro-C data for GM12878, K562, and HCT116 cell lines were obtained from Hong *et al.* [35]. The accession codes can be found in Supplementary Table S1. These files were mapped at 50 bp resolution. Hence, we used the *zoomify* function from cooler v0.10.4 to create a 1-kb-resolution .cool file for each of the cell lines. This was followed by the normalization and smoothing steps as described above for Hi-C.

ATAC-seq data processing

In this work, we used the GRCh38/hg38 human reference genome. ATAC-seq data from five human cell lines was used in this study, consisting of healthy (GM12878 and IMR90) and cancer (K562, A549, and HepG2) cells. The carefully processed signal *P*-value bigwig profiles and the pseudoreplicated peak files were downloaded directly from the ENCODE portal (http://www.encodeproject.org/) under the accession codes listed in Supplementary Table S1. For multiple ATAC-seq peaks within the 500 bp genomic region, we retain the peak with the highest read count and remove the remaining peaks. This resulted in ∼175K peaks for each cell line.

To ensure consistency between the ATAC-seq bigwig profiles of different cell types, we normalized the bigwig signal using GM12878 as the reference cell line. We used conserved CTCF sites as a stable set of regions expected to show similar ATAC-seq signals across all cell types. First, we identified a conserved set of CTCF sites by selecting the common peaks from CTCF ChIP-seq tracks of all five cell lines. For each identified peak, we then obtained the signal values from the .bigwig files of each cell line using deepTools multiBigwigSummary v3.5.3 [36]. Finally, we applied the trimmed mean of M-values (TMM) normalization of EdgeR [37] using the extracted ATAC-seq signal values on these conserved CTCF sites to obtain the scaling factor for each cell line. GM12878 cell line was used as a reference while TMM normalization was applied. The normalized bigwig was then constructed by multiplying the scaling factor with the original signal track. The effect of normalization for all the cell lines can be observed in Supplementary Fig. S3.

ChIP-seq data processing

The "narrow peak" files of each TF from the list of target TFs (CTCF, YY1, and SP1) for all five cell lines (GM12878, K562, IMR90, HepG2, and A549) were obtained from the ENCODE portal. The accession codes can be found in Supplementary Table S1. To account for the role of CTCF orientation in the formation and stability of TADs, we split CTCF ChIP-seq peaks based on their orientation. Specifically, we utilized the consensus motif MA0139.1 from the JASPAR database [38] to derive a position-specific scoring matrix, which was subsequently convolved over the forward and reverse strands of the reference genome. To determine the optimal cutoff for the resulting convolved signal, we employed a false negative rate of 1%, based on empirical estimates [39] of the intergenic nucleotide frequency background distribution. Signals larger than the cutoff indicated the presence of the consensus motif on the respective strand. Since the resolution of the ChIP-seq signal was lower than the motif length, a peak was labeled as *forward* and *backward* facing if both signals exceeded the cutoff value (18% of all peaks). Overall, 49% of CTCF peaks contained at least one forward-strand motif hit, and 48% contained at least one reverse-strand motif hit. For 20% of the peaks, no corresponding strand could be determined. Since these peaks were also highly correlated with low enrichment, they were dropped from the dataset during training.

Model architecture

The architecture of UniversalEPI only accounts for the information coming from accessible chromatin regions (ATAC-seq peaks); all information about genomic regions between accessible regions is provided to the model via the value of the distance between ATAC-seq peaks. The model, therefore, predicts the entries in the interaction matrices only for the pairwise interactions between genomic regions containing ATAC-seq peaks. We propose a two-stage deep learning model to learn the mapping $F: \mathbb{X} \mapsto \mathbb{Y}$, where $\mathbb{X}$ is the space of genomic information about ATAC-seq peaks and $\mathbb{Y}$ is the space of interactions. Namely, for each $X \in \mathbb{X}$, $X = \{x_1,x_2,..,x_{401}\}$ with $x_i \in \mathbb{R}^{1000\times 5}$ being stacked one-hot encoded DNA sequences and ATAC-seq signal of one ATAC-seq peak, the mapping should produce $Y = \{y^{(i,j)} \mid y^{(i,j)} \in \mathbb{R}, i \leq j, i,j = 101,\ldots,300\} \in \mathbb{Y}$ with elements representing the interaction between the pair of peak locations i and j in the central 200 ATAC-seq peaks. Note that the entry $y^{(i,j)}$, which represents the interaction between the pair of peaks locations i and j, depends not only on the data at those locations but also on the information across other peaks in their neighborhood.

The first stage learns genomic features for the peak regions, relevant to predicting generalized chromatin features across cell types, whereas the second stage predicts Hi-C values by modeling the interactions between the learned features at peak locations. A schematic overview of the proposed method is depicted in Supplementary Fig. S4. Below, we describe the two stages in detail.

Stage 1: representation learning

In the first stage, we train a representation network $f_\theta(x)$, where $x \in \mathbb{R}^{1000\times 5}$, to predict binding affinity of the target TFs $y_{TF} \in \mathbb{R}^{N_{TF}}$ using the data across different cell lines, where N_{TF} is the number of the target TFs. This approach ensures that the learned features capture information that is both invariant across different cell lines and relevant for predicting the target Hi-C. Inspired by DeepC [12], we adopt a convolution-based model for this purpose. The model consists of five convolutional layers with {30, 60, 60, 90, 90} channels, and kernel sizes of {11,11,11, 5, 5}, respectively. After each convolutional layer, max pooling with widths of {4, 5, 5, 4, 2} is applied to aggregate the learned features, which is followed by Leaky Rectified Linear Unit (LeakyReLU) [40] with a slope of 0.2 to introduce non-linearity. A fully connected layer takes the output of the convolutional layers and maps to the target TFs. To prevent overfitting, a 20% dropout is applied during training. The model is optimized by minimizing the empirical version of the mean squared error (MSE) defined as in equation (2),

$$\min_{\theta} \mathbb{E}_{(x,y_{\text{TF}})\sim P_{\text{TF}}} \left[\| f_\theta(x) - y_{\text{TF}} \|^2 \right], \quad (2)$$

where paired x and y_{TF} are assumed to be sampled from a joint data distribution P_{TF}.

The use of convolutional layers enables the network to effectively extract hierarchical features from the input data. After applying the non-linear activations, the outputs of convolutional blocks serve as intermediate representations and are used as the input for the second stage. We use a linear projection head for each feature layer to map it to the space of dimension C, which is set to 180. For one peak x, we obtain a set of N_L projected features $h(x) = \{h_1(x),h_2(x),\ldots,h_{N_L}(x)\}$, $h_i(x) \in \mathbb{R}^C$, where $N_L = 5$ is the total number of convolutional blocks. Thus, for the subsequent 401 ATAC-seq peaks, we get the stacked projected features of the dimension $401 \times C \times N_L$, which we denote by $h(X)$. The feature selection module is then employed to select from N_L features, as described in details below.

Stage 2: Hi-C prediction

- Automatic feature selection is enabled by adopting the stochastic gating mechanism proposed by Yamada *et al.* [41] to automatically select deep features learned during the first stage for predicting Hi-C interactions. Unlike traditional deterministic approaches, which rely on predefined criteria to select features, stochastic gating allows the model to explore a broader range of potential feature subsets and identify those crucial for the training task. Stochastic gating approximates the ℓ_0 sparsity constraint by continuous relaxation of Bernoulli distribution.

We jointly optimize the feature selection function s_μ and the target network g_ϕ, which predicts the Hi-C values. The total objective can be written as equation (3),

$$\min_{\phi,\nu} \mathbb{E}_{(X,Y)\sim P_{HiC}} \mathbb{E}_Z [L_T(g_\phi(s_\nu(h(X), Z)), Y) + \lambda L_R(Z)] \tag{3}$$

where L_T and L_R denote the target loss function and feature sparsity regularization, respectively, parameter λ controls the regularization strength, Z is a random sparsity-defining vector of size N_L, X and Y are sampled from a joint data distribution P_{HiC}. The components of Z provide the gating mechanism: For k-th feature $z_k = \max(0,\min(1,\nu_k + \epsilon_k))$, where ν_k is learned and ϵ_k is sampled from $\mathcal{N}(0,\sigma^2)$ with a predefined σ during training. Following Yamada *et al.* [41], we set $\sigma = 0.5$. Herein we propose to use the stochastic gating $s_\nu(h(X),Z) = \{s_\nu(h(x_1), Z), s_\nu(h(x_2), Z), \ldots, s_\nu(h(x_{401}), Z)\}$ at the feature level instead of the node level (the latter was initially proposed by Yamada *et al.* [41]) and aggregate the selected features using average pooling to reduce the model complexity. The gating mechanism is applied for each ATAC-seq peak x_i : $s_\nu(h(x_i), Z) = \sum_{k=1}^{N_L} h_k(x_i) \cdot z_k$. The sparsity regularization is the sum of the probabilities that the gates are active, which is equal to $\mathbb{E}_Z L_R(Z) = \sum_{k=1}^{N_L} P(z_k > 0) = \sum_{k=1}^{N_L} \phi(\frac{\nu_k}{\sigma})$, where ϕ is the Gaussian cumulative distribution function. During the inference stage, we set $\hat{z}_k = \max(0,\min(1,\nu_k))$.

- Hi-C prediction network g_ϕ comprises a transformer-based encoder g_ϕ^{ENC} that extracts information about the local and global context of the input sequence followed by a task-specific lightweight multilayer perceptron (MLP) decoder g_ϕ^{DEC}. The encoder consists of four multi-head attention blocks as proposed by Vaswani *et al.* [42]. We used four attention heads, a dropout rate of 0.1, and set the hidden dimension d_{model} equal to 32. To more accurately encode the genomic distance between input tokens, which is crucial for predicting interaction level, here we propose a genomic-distance-aware positional encoding. We compute the relative genomic distance between each token and a reference token and encode the distance using the sine-cosine positional encoding scheme [42]. Each token corresponds to a region centered around a single ATAC-seq peak. We designate the position of the highest ATAC-seq peak value as the token position. The middle token is chosen as the reference. We encode a maximum distance of 3 Mbp with a resolution of 500 bp. Note that the orientation is considered by assigning negative distances to the left peaks from the center and positive distances to the right peaks. We add a constant of 3 Mbp to all the distances to ensure they remain positive before the sine-cosine encoding.

The encoder extracts a representation for each token, capturing its relation with the other tokens in the input sequence. The output of g_ϕ^{ENC} is defined as $E_{\phi,\nu} = \{e_1,e_2,\ldots,e_{401}\}$ with the i-th token representation $e_i \in \mathbb{R}^{d_{model}}$, i.e. $E_{\phi,\nu} = g_\phi^{ENC}(s_\nu(h(X), Z))$. To compute the Hi-C interaction value between two peaks locations i and j, the encoder output of the i-th and j-th token are concatenated and fed into the decoder predicting Hi-C value, *i.e.* $\hat{y}_{\phi,\nu}^{(i,j)} = g_\phi^{DEC}(e_i, e_j)$ and e_j depend on ϕ,ν. We use two MLP layers with ReLU activation for the decoder. During training, sequences are randomly flipped to ensure that the decoder remains invariant to the order of the tokens. The target loss L_T between the true interactions Y and the estimated interactions $\hat{Y}_{\phi,\nu}$ in equation (3) is defined in equation (4),

$$L_T\left(\hat{Y}_{\phi,\nu},\ Y\right) = \frac{1}{N}\sum_{i,j}\left(\hat{y}_{\phi,\nu}^{(i,j)} - y^{(i,j)}\right)^2, \tag{4}$$

where i,j are encoded following the DeepC approach [12] in a vertical, zigzag pole over the center of the sequence window (Fig. 1). N is the total number of interaction pairs. During training the empirical version of the loss [equation (3)] is optimized in both ν and ϕ.

- Auxiliary information can be optionally used for Hi-C prediction. Since the Hi-C values between two bins are also dependent on the mappability of the independent bins [43], we also include the 36 bp-mappability track, obtained from the UCSC Genome Browser [44, 45], as input to the Stage 2 of UniversalEPI. First, we use a linear head to embed auxiliary information. We set the embedding dimension the same as the hidden dimension of the transformer. The embedded vector is then concatenated with the input embeddings, resulting in a new input to the attention layers. This method allows the use of any auxiliary information, such as mappability and ATAC-seq.

Uncertainty estimation

We incorporate both aleatoric and epistemic uncertainty estimation into UniversalEPI, providing information on the reliability of model predictions. To capture the aleatoric uncertainty, we assume that each observed log Hi-C value is a sample from a Gaussian distribution. Instead of the point estimate $\hat{Y}_{\phi,\nu}$ comprised of $\hat{y}_{\phi,\nu}^{(i,j)}$ provided by the entire Hi-C prediction

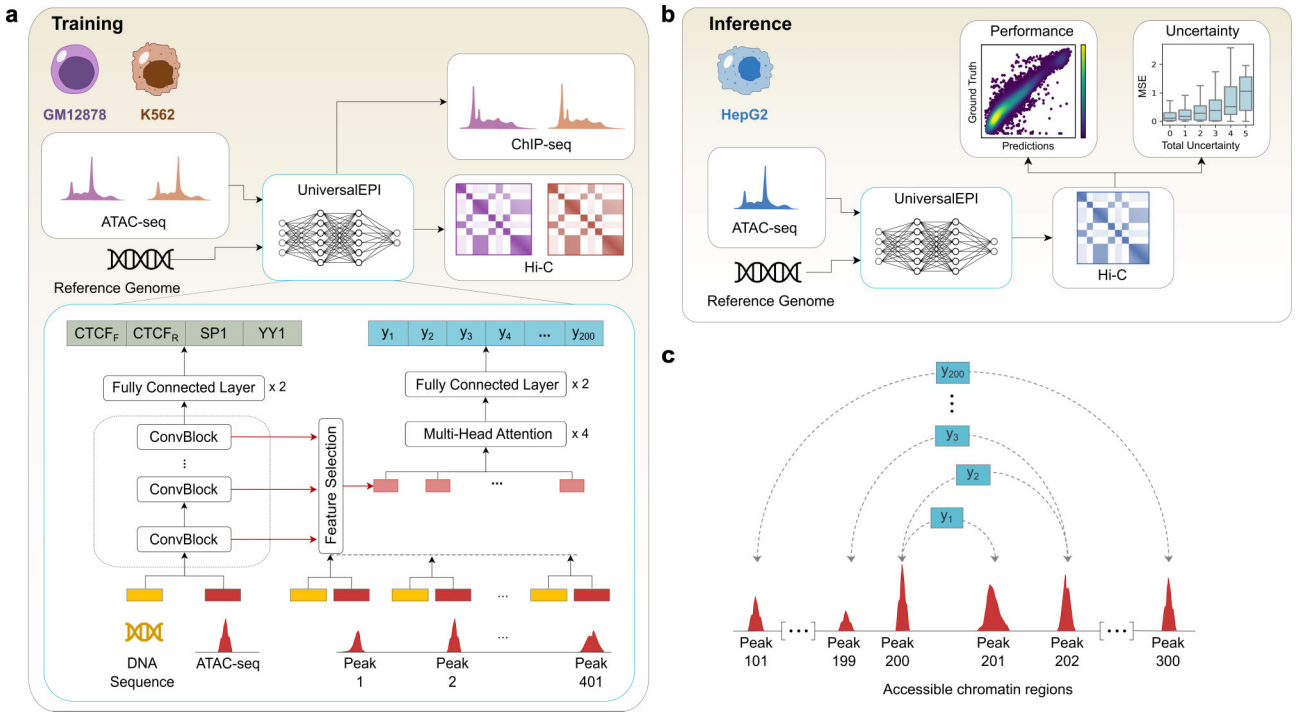


Figure 1. The UniversalEPI model architecture. (**a**) UniversalEPI uses one-hot-encoded DNA sequences and ATAC-seq *P*-value tracks as input to predict cell-type-specific TF binding and chromatin interactions. The first stage of UniversalEPI consists of a convolutional architecture that predicts TF (CTCF forward and reverse, SP1 and YY1) binding affinity from DNA sequence and ATAC-seq data. The second stage automatically selects the features from the pretrained convolutional layers from the first stage using a stochastic gating mechanism. The multi-head attention block then uses these features to predict cell-type-specific Hi-C interactions between accessible regions. (**b**) The trained UniversalEPI can then be used to predict Hi-C interactions for an unseen cell type. The model also outputs uncertainty associated with each prediction, which increases with the MSE. (**c**) UniversalEPI predicts pairwise Hi-C interactions (y_i) between the central 200 accessible regions corresponding to ATAC-seq peaks.

component $g_{\phi^\circ} s_v$, we model the parameter of the distribution: the mean $M_{\phi,v} = \{\mu_{\phi,v}^{(i,j)} | \mu_{\phi,v}^{(i,j)} \in \mathbb{R}\}$ and the variance $\Sigma_{\phi,v} = \{[\sigma_{\phi,v}{}^{(i,j)}]^2 \mid [\sigma_{\phi,v}{}^{(i,j)}]^2 \in \mathbb{R}_{\geq 0}\}$ between two peaks locations i and j, which are structured the same way as Y. We apply an exponential activation function to the variance output to guarantee non-negativity. Instead of the MSE loss L_T defined in equation (4), herein we minimize in ϕ and μ the negative log-likelihood loss weighted by the β-exponentiated variance estimates [28], referred to as $\beta-$NLL loss, which is defined in equation (5),

$$L_{\beta-\text{NLL}}\left(M_{\phi,v}, \Sigma_{\phi,v}, Y\right) = \frac{1}{2N}\sum_{i,j} \text{sg}\left(\left[\sigma_{\phi,v}^{(i,j)}\right]^2\right)^{\beta} \times \left(\log\left(\left[\sigma_{\phi,v}^{(i,j)}\right]^2\right) + \frac{\left(y_{\phi,v}^{(i,j)} - \mu_{\phi,v}^{(i,j)}\right)^2}{\left[\sigma_{\phi,v}^{(i,j)}\right]^2}\right), \quad (5)$$

with the stop-gradient operator sg, and the parameter β controlling the trade-off between the regression accuracy and log-likelihood estimation. We set $\beta = 0.5$ following [28], which has been empirically found to provide the best trade-off. To estimate epistemic uncertainty, we utilize the deep ensemble method [29] by training K models with different random initializations, resulting in an ensemble $\{(\phi_k,\nu_k),\ k = 1,..,K\}$. The total predictive uncertainty $[\sigma_T^{(i,j)}]^2$ is a sum of aleatoric $[\sigma_a^{(i,j)}]^2$ and epistemic $[\sigma_e^{(i,j)}]^2$ uncertainties. It is quantified by combining these components as described in equation (6),

$$\left[\sigma_T^{(i,j)}\right]^2 = \left[\sigma_a^{(i,j)}\right]^2 + \left[\sigma_e^{(i,j)}\right]^2 = \frac{1}{K}\sum_k \left[\sigma_{\phi_k,\nu_k}^{(i,j)}\right]^2 + \frac{1}{K}\sum_k \left(\mu_{\phi_k,\nu_k}^{(i,j)} - \mu_T^{(i,j)}\right)^2, \quad (6)$$

with the empirical estimate of the mean $\mu_T^{(i,j)} = \frac{1}{K}\sum_k \mu_{\phi_k,\nu_k}^{(i,j)}$. We use an ensemble size of $K = 10$ for the reported results.

Training details

Our method is implemented with the PyTorch framework [46]. The training of the proposed method involves a two-stage model, where each stage was trained separately with different training setups. The TF prediction network was trained with Adam [47] optimizer using a learning rate of 10^{-4}, a weight decay of 10^{-4}, with a batch size of 1024 over 100 epochs. The best model was identified using Pearson's correlation score on the validation dataset. For each active binding site, we select a 1 kb segment of DNA sequence around the peak center and the corresponding ATAC-seq as input. TFs CTCF, YY1, and SP1 were empirically chosen as target TFs based on their ability to assist in chromatin organization prediction (Supplementary Fig. S5). ZNF143 was initially considered another target TF candidate [48–50]. However, a recent study [51] found the commonly used antibody of ZNF143 cross-reacting with CTCF, challenging reported associations between ZNF143 binding and chromatin looping. This, along with the fact that including ZNF143 did not improve Hi-C prediction accuracy (Supplementary Fig. S5), led us to exclude

this TF from further consideration in our work. The first-stage model was trained on the GM12878 and K562 cell lines. We randomly split the chromosomes into training, validation, and test sets. Chromosomes 5, 12, 13, and 21 were used for validation, 2, 6, and 19 for testing; and the remaining chromosomes for training. After training, the backbone of the TF prediction network was used as a feature extractor, and its model parameters were frozen.

In the second stage, the Hi-C prediction network was trained using the AdamW [52] optimizer with a learning rate of 10^{-3}. We set the feature sparsity regularization λ to 0.01, which was tuned to select the minimum number of feature layers necessary. Since distance between elements is highly predictive of the Hi-C values, we employed a position embedding upscaling by a factor of $\sqrt{d_{\text{model}}}$. We trained two sets of models, one set uses GM12878 and K562 cell lines for training, and IMR90 and HepG2 cell lines for testing whereas the other uses HepG2 and IMR90 cells for training and GM12878 and K562 cells were held out for testing. The second-stage model was trained using the same chromosome split as the first stage. We found that the Hi-C data at 5 kb resolution also contains arbitrary bins with very few reads mapped to them (Supplementary Fig. S6). To ensure that these noisy samples do not affect our analysis, we removed these and their one-hop neighboring bins from all datasets. To increase robustness, we augmented the accessible regions in the training dataset with 10% of inaccessible regions of length 1 kb each that were chosen randomly from the genome ensuring that the model can also predict accurately for various inaccessible regions, which can arise due to mutations. The combined regions were then used to form the 401 consecutive peaks, which are the input of UniversalEPI. For 1 kb Micro-C, we used a similar training procedure as described above, only replacing 5 kb Hi-C with 1 kb Micro-C. The Micro-C model was trained using the smoothed contact matrix, as the high-resolution Micro-C data is inherently noisier than the Hi-C data.

In all cases, we trained the network for 20 epochs and chose the model for evaluation with the highest Spearman's correlation on the validation data. Training the Hi-C and Micro-C prediction network took 12 h on 0.2 of an A100 GPU with a batch size of 32, and 24 h on a single RTX2080Ti with a batch size of 16.

Evaluation details

UniversalEPI was evaluated on unseen chromosomes of unseen cell types. To ensure that reliable targets were used for evaluation, we first flagged the bins that overlap with unmappable (https://storage.googleapis.com/basenji_barnyard2/umap_k36_t10_l32_hg38.bed) and blacklisted (https://storage.googleapis.com/basenji_barnyard2/hg38.blacklist.rep.bed) regions as reported by Kelley *et al.* [53]. We then removed the interactions containing a flagged endpoint. Finally, we calculated Spearman's correlation and Pearson's correlation using the smoothed log Hi-C of the remaining interactions.

To understand the biological information captured by our model, we removed the strong influence of distance on Hi-C prediction by calculating distance-stratified correlation. Specifically, we calculated the correlation using all interactions that lie at a particular distance d away from each other, where d lies between 0 and 2 Mb with a step size of 5 kb.

Results

UniversalEPI: a transformer-based model to predict genomic interactions from DNA sequence and chromatin accessibility

UniversalEPI is a two-stage deep learning model that comprises two sequentially trained neural networks that use DNA sequences and signals from ATAC-seq from training cell lines as input and ground-truth ChIP-seq and Hi-C data as targets (Fig. 1a). To infer the Hi-C signal in a test cell type, UniversalEPI requires DNA sequence and cell-type-specific ATAC-seq profiles as input (Fig. 1b).

The first neural network of UniversalEPI was trained to predict the genome-wide binding occupancy of the TFs SP1, CTCF, and YY1 using maximal intensities from ChIP-seq data in human cancerous and non-cancerous cell lines GM12878 and K562 (Model 1) and IMR90 and HepG2 (Model 2). Since the orientation of CTCF binding to chromatin plays an important role in determining the stability of TADs [54], we split the set of CTCF ChIP-seq peaks by motif orientation (forward or reverse strand) and predicted the corresponding maximal ChIP-seq intensities as separate targets. The prediction of TF occupancy was achieved by training a five-layer one-dimensional CNN that produced a four-dimensional output, corresponding to each target TF, on the concatenation of one-hot-encoded 1-kb DNA sequences and their corresponding ATAC-seq *P*-value signals. The pre-trained convolutional layers from the first stage thus generated sequence embeddings, indicating the presence or absence of the three TFs in each accessible region of the genome. This design of sequence embeddings was chosen to prevent the model from capturing DNA motifs of other, potentially cell-type-specific, TFs and thereby enable the model to generalize predictions across different cell types in the second stage.

In the second stage, we trained the model to predict Hi-C values between accessible regions based on the DNA sequence embeddings calculated in the first stage (Fig. 1c). By focusing on accessible regions, we retain most active regulatory elements, ensuring that the chromatin interactions between key regions involved in gene regulation are predicted by the model (Supplementary Fig. S7). We employed a transformer-based encoder for its ability to capture long-range interactions and to create large-context embeddings of genomic elements. This allowed the model to efficiently learn the interdependency between the input regions and, consequently, predict Hi-C interactions between regulatory elements with high accuracy in a cell-type-specific manner. Specifically, the model input in the second stage corresponded to 401 consecutive accessible regions (or peaks) spanning in total ~4 Mb. Each region was 1 kb in length, centered around the position of the maximum ATAC-seq signal. Skipping the information about the DNA sequence between accessible regions allowed for a lightweight architecture of UniversalEPI (2.5M parameters) that can run on standard GPUs while maintaining a large receptive field and making accurate predictions for interactions between regulatory elements that are up to 2 Mb apart.

To provide uncertainty estimations and make the model robust to the initialization parameters, we implemented a deep ensemble of 10 single models [29]. This approach allowed the estimation of uncertainty for each prediction, facilitating confident differential predictions between two conditions, i.e. when inputs (DNA or ATAC-seq signal) change. The total predictive uncertainty reported was a combination of two types:

aleatoric uncertainty, which stems from inherent data variability, and epistemic uncertainty, which reflects model limitations. The former was estimated by minimizing the negative log-likelihood loss to incorporate the predicted variance, whereas the latter was quantified through the variability across the deep ensemble.

To ensure the stability of the proposed architecture, we trained both stages of UniversalEPI in a cross-validation setting on human cancerous and non-cancerous cell lines. Model 1 was trained on GM12878 and K562 cells and tested on unseen chromosomes of IMR90 and HepG2 cells. Model 2 was trained on IMR90 and HepG2 cells and tested on unseen chromosomes of GM12878 and K562 cells. Approximately 15% of the human genome was set aside for validation (chromosomes 5, 12, 13, and 21) and ∼15% for testing (chromosomes 2, 6, and 19). The first stage of UniversalEPI predicted TF binding on unseen chromosomes and unseen cell types for CTCF, YY1, and SP1 with an average Pearson's correlation of 0.78, 0.62, and 0.47, respectively (Supplementary Fig. S8). We further validated that the first stage of our model learned biologically relevant information in the input DNA sequence by analyzing the model gradients using the DeepLIFT attributions scores [55]. These attribution scores revealed important DNA motifs associated with the binding of each TF, closely resembling their respective known consensus motifs (Supplementary Fig. S8).

UniversalEPI can predict changes in TAD organization in response to *in silico* modifications

Experimental Hi-C values, which quantify chromatin interactions between each pair of genomic regions, are inversely correlated with the distance between regions and follow the exponential decline resulting from polymer physics [56]. Similarly, UniversalEPI's single model predictions exponentially declined with distance on the test chromosomes of the unseen cell lines, highlighting the model's ability to accurately capture the effect of distance on chromatin structure (Fig. 2a). Further, to show that UniversalEPI was sensitive to variability in chromatin accessibility and thus is capable of making predictions for unseen cell types based on ATAC-seq inputs, we computed the distance-stratified correlation between the predicted and ground-truth Hi-C signal between pairs of accessible regions. Using cell-line-specific ATAC-seq profiles as inputs, UniversalEPI obtained equally high values of distance-stratified correlation between predicted and ground-truth Hi-C values for all four test cell lines, IMR90 and HepG2 (Model 1), and GM12878 and K562 (Model 2) (Fig. 2b).

Next, we evaluated how the deep ensemble implemented in UniversalEPI improved prediction accuracy compared to a single model. As expected, the deep ensemble approach significantly outperformed the single model in generalizing to test chromosomes from previously unseen cell lines (Fig. 2c). Notably, higher total uncertainty was generally associated with a greater mean-squared error between predicted and ground-truth Hi-C values, indicating that the uncertainty estimates were well calibrated (Fig. 2d). Moreover, in both IMR90 and HepG2, higher total uncertainty was aligned with increased biological variation between replicates, indicating that the model's uncertainty partly reflects underlying biological variability (Supplementary Fig. S9).

We then assessed the sensitivity of the deep-ensemble UniversalEPI model to changes in the DNA sequence by implementing *in silico* inversions of CTCF binding sites located at TAD boundaries (Supplementary Sections S1 and S2). Disruptions in TADs have been linked to diseases such as cancer, making chromatin interactions within and across TADs important for exploring disease mechanisms [57–59]. For 293 TADs across the four cell lines, we inverted one endpoint of a TAD boundary containing a CTCF site in a convergent orientation, which was expected to significantly reduce the strength of chromatin interactions at the original TAD boundary (Fig. 2e). Indeed, when we provided the original and modified inputs to UniversalEPI, we predominantly observed negative values of the log FC across all four test cell lines, indicating a decrease in predicted interactions following CTCF motif inversion (Fig. 2f). The maximum-confidence FC values, estimated by our uncertainty-aware model to represent the largest changes with 90% probability, aligned with the original predictions (Fig. 2g and Supplementary Section S3). This consistency can be attributed to the strong effect of the *in silico* change to the TAD boundary.

The reported prediction uncertainty played an important role in determining other significantly altered interactions predicted by UniversalEPI after the *in silico* inversion of CTCF binding motifs at TAD boundaries. We illustrated this effect using a randomly selected TAD in IMR90 cells (chr6:17 280 000–17 585 000). UniversalEPI predicted a substantial reduction in interaction between the original TAD boundaries as well as between the 3′ TAD boundary and several enhancer regions upstream of the original 5′ TAD boundary (Fig. 2h). Moreover, we observed a significant increase in predicted interactions with a reverse CTCF motif located nearly 200 kb away, downstream of the original 3′ TAD boundary. Importantly, the estimated uncertainty allowed the calculation of the maximum-confidence FC and retention of only the highly reliable differential interactions (Fig. 2h and Supplementary Section S3). This result confirmed that UniversalEPI is capable of capturing the effects of CTCF binding motif orientation on the stability and organization of TADs, and this is maintained when filtering for significant differential interactions using the estimated prediction uncertainty.

To further evaluate UniversalEPI's sensitivity to genomic variation, specifically assessing whether UniversalEPI can detect allele-specific chromatin interactions driven by single-nucleotide variation, we tested the model on a well-characterized case involving SNP rs1800734. This variant was experimentally demonstrated to enhance *DCLK3* expression in colorectal cancer by strengthening interactions between a distal regulatory element encompassing rs1800734 and the 3′ UTR of the gene [60]. Using ATAC-seq data from isogenic COLO320 cell lines homozygous for either the reference (G) or alternate (A) allele generated by Liu *et al.* [60], we computed UniversalEPI's Hi-C predictions. Consistent with experimental findings, UniversalEPI predicted stronger interactions between the SNP-containing region and the 3′ UTR of *DCLK3* in the alternate allele background, as shown by maximum-confidence FC of predicted Hi-C values (Fig. 2i). Notably, UniversalEPI also predicted other differential interactions driven by this SNP that had not been discussed in the previous work. These novel predictions open avenues for future functional exploration of mechanisms underlying gene expression in these regions. Overall, these results demonstrate that UniversalEPI can detect allele-specific chromatin interaction differences when paired with corresponding chromatin accessibility data.

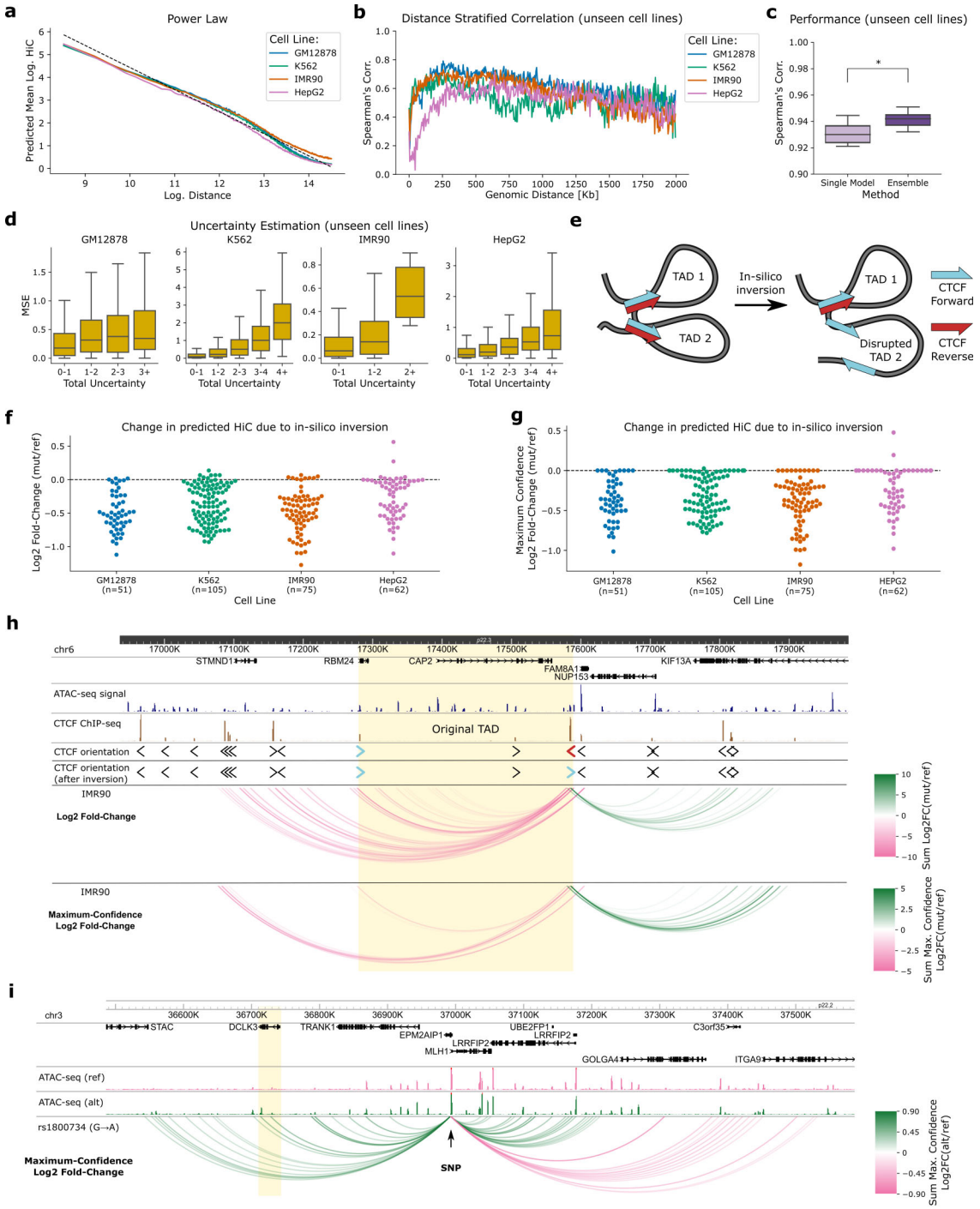


Figure 2. UniversalEPI learns information from biologically relevant variables and provides well-calibrated uncertainty estimates. (**a**) The predicted Hi-C signal by UniversalEPI follows the power law resulting from polymer physics. The average predicted signal is shown for test chromosomes of test cell lines (i.e. cell lines unseen during training): GM12878, K562, IMR90, and HepG2. The theoretical relationship between the log-distance and log-Hi-C values is depicted as the dotted line. (**b**) Distance-stratified Spearman's correlation for all the unseen cell lines during training displaying UniversalEPI's sensitivity to the input ATAC-seq. (**c**) Spearman's correlation is compared between the predictions made by a single model as compared to the ensemble. The predictions are made on the test chromosomes of test cell lines. A Mann–Whitney U test is used to compare the methods; $^{*}P \leq .05$. (**d**) Relationship between ensemble's predictive error, calculated as a MSE, and prediction uncertainty is illustrated on the test chromosomes of test cell lines. (**e**) Graphical illustration depicting the effect of *in silico* inversion of a CTCF binding site on genome organization. (**f**) Log_2 FC between the UniversalEPI predictions before and after the *in silico* inversion. The experiment is done for all TADs in test chromosomes in unseen cell lines: GM12878, K562, IMR90, and HepG2. (**g**) Maximum-confidence Log_2 FC between the UniversalEPI predictions before and after the *in silico* inversion (Supplementary Section S3), same regions as in panel (f). (**h**) An example of UniversalEPI predictions for a TAD in chromosome 6 of IMR90 cells (unseen cell line and unseen chromosome). The original TAD location is highlighted. The change in Hi-C predicted by an ensemble model is calculated using log_2 FC before and after the *in silico* inversion (top track) and maximum-confidence log_2 FC (bottom track). In the presence of multiple ATAC-seq peaks within the same 5-kb bin, the log_2 FC is summed among all pairs of interactions contained in the two bins. (**i**) Maximum-confidence log_2 FC of Hi-C measurements predicted by UniversalEPI for the regulatory element encompassing the single-nucleotide polymorphism (SNP) rs1800734 (G → A) as one of the endpoints. The *DCLK3* gene, experimentally shown to exhibit stronger interactions with the alternate allele (A) [60], is highlighted.

Benchmarking UniversalEPI on unseen cell types using bulk inputs

We compared UniversalEPI against the leading method predicting quantitative bulk Hi-C contact maps, C.Origami [22] and EPCOT [23]. While demonstrating high reported accuracy, C.Origami relies on additional CTCF ChIP-seq input [22]. To enable a fair comparison under limited-input settings, we additionally retrained C.Origami using only DNA sequence and ATAC-seq data (Supplementary Section S4). We further benchmarked UniversalEPI against purely sequence-based models such as Akita [13]. We also introduced three baselines that capture the effect of the distance between interacting elements and the variability between the train and test Hi-C interaction matrices (Supplementary Section S5). As the central goal of this work is to develop a model that generalizes to unseen cell types, we benchmarked the models in four cell lines not used for model training, using the same train-test split as above. Performance was evaluated based on the ability to predict experimental Hi-C interactions between accessible regions within 1 Mb, which was the maximum common receptive field across all existing methods. To enable a fair comparison with EPCOT, we employed the official pre-trained model provided by the authors, as we were unable to fully reproduce the training performance reported in the original study. This model was benchmarked only on the IMR90 cell line, which served as a held-out cell type in the EPCOT training set (Supplementary Section S4).

We observed that UniversalEPI significantly outperformed the original version of C.Origami, which used CTCF ChIP-seq as an additional input, in predicting chromatin interactions in cell lines not seen during training ($P < .05$; Fig. 3a, Supplementary Fig. S10). Performance declined further when C.Origami was retrained without CTCF input ($P < .01$; Fig. 3a). UniversalEPI also outperformed the DNA sequence-only method Akita, which is inherently limited in its ability to generalize to cell types not seen during training ($P < .001$; Fig. 3a). Notably, UniversalEPI was the only method to achieve Spearman's correlations above 0.90 between predicted and experimental Hi-C values across all test cell lines (Fig. 3b). On IMR90 and K562, UniversalEPI achieved comparable results to C.Origami, and it substantially outperformed C.Origami on HepG2 and GM12878 cell lines. UniversalEPI also outperformed EPCOT on the common unseen cell line (IMR90).

Since the distance between the interacting elements is a major factor in determining the strength of interaction, we also compared the existing methods after removing the distance effect. This was done by calculating the distance-stratified correlation (Fig. 3c), which assessed the ability of tested methods to capture biological features across several distance ranges. We observed that UniversalEPI predicted the strength of chromatin interactions at a given distance more accurately than all existing methods, including C.Origami. The possible reason for the lower performance of C.Origami could be overfitting on cell lines used in training. This was also consistent with the higher performance of C.Origami on the test chromosomes of the training cell lines compared to UniversalEPI (Supplementary Fig. S11). In addition to better performance than existing methods on cell types unseen during training, UniversalEPI offers an efficient solution with a small model size and a short inference runtime, making our method a more scalable alternative for large-scale applications (Fig. 3d).

We compared UniversalEPI and C.Origami (with and without CTCF ChIP-seq as input) on their ability to predict specific types of EPIs and PPIs in the same test cell lines as above using experimental Hi-C as ground-truth. To study the ability of the models to capture distance-dependent interactions, we split EPIs into two categories: proximal EPIs (pEPIs, 200 bp–2 kb) and distal EPIs (dEPIs, >2 kb). The annotations for pEPIs, dEPIs, and PPIs were obtained from the ENCODE library [61]. We also compared the two methods based on their ability to predict interactions between promoters and sEPIs, which are clusters of active enhancers that often regulate cell identity genes [62] and are defined for each cell line in the dbSuper database [63]. UniversalEPI outperformed both versions of C.Origami on all but the K562 cell line, where C.Origami with additional CTCF binding information showed slightly higher prediction accuracy than UniversalEPI (Fig. 3e). Overall, we conclude that UniversalEPI generally captures important *cis*-regulatory interactions more accurately than C.Origami and that DNA sequence and bulk ATAC-seq data are sufficient to accurately predict cell-type-specific chromatin interactions.

Finally, we also trained a UniversalEPI model using 1-kb Micro-C data from the GM12878 and K562 cell lines and compared its ability to generalize to unseen cell types against the EPCOT method [23], recently used to predict the Micro-C signal. Compared to EPCOT, on the held-out HCT116 cell line, UniversalEPI resulted in higher Spearman's correlation between predicted and ground-truth Micro-C values (Fig. 3f). This increase was predominantly observed for shorter interactions (<100 kb) while achieving comparable performance for longer-distance interactions (150–200 kb) (Fig. 3g). Because Akita and EPCOT output observed-over-expected contact frequencies, we normalized UniversalEPI predictions to observed-over-expected by dividing each value by the distance-stratified mean prediction across the genome. Under this standardized setting, UniversalEPI consistently outperformed both Akita and EPCOT (Supplementary Fig. S12). This highlights UniversalEPI's flexibility to train on different high-resolution chromatin contact maps and its consistently improved performance across resolutions.

UniversalEPI captures chromatin dynamics during macrophage activation

UniversalEPI can be used to predict changes in regulatory chromatin interactions upon cell reprogramming. To demonstrate this, we analyzed the model's predictions of chromatin interactions during activation of human macrophages *in vitro*. ATAC-seq and ground-truth Hi-C data were collected at eight time points following macrophage activation with lipopolysaccharide (LPS) and interferon-γ (IFN$_\gamma$) [30] (Fig. 4a and Supplementary Section S6). In a zero-shot setting, Model 1 of UniversalEPI, which was trained on GM12878 and K562 cell lines, demonstrated high prediction accuracy, with Pearson's and Spearman's correlation with ground-truth Hi-C values exceeding 90% (Fig. 4b). Upon the introduction of LPS + IFNγ treatment, Reed *et al.* observed the formation of a 570 kb loop connecting the *GEM* promoter and a distal enhancer [30]. UniversalEPI predicted a similar interaction, highlighting the reliability and biological relevance of the model's predictions (Supplementary Fig. S13).

Reed *et al.* identified specific chromatin loops that exhibited differential interactions upon macrophage activation—lost, gained, or remaining static (Fig. 4c). For these selected chro-

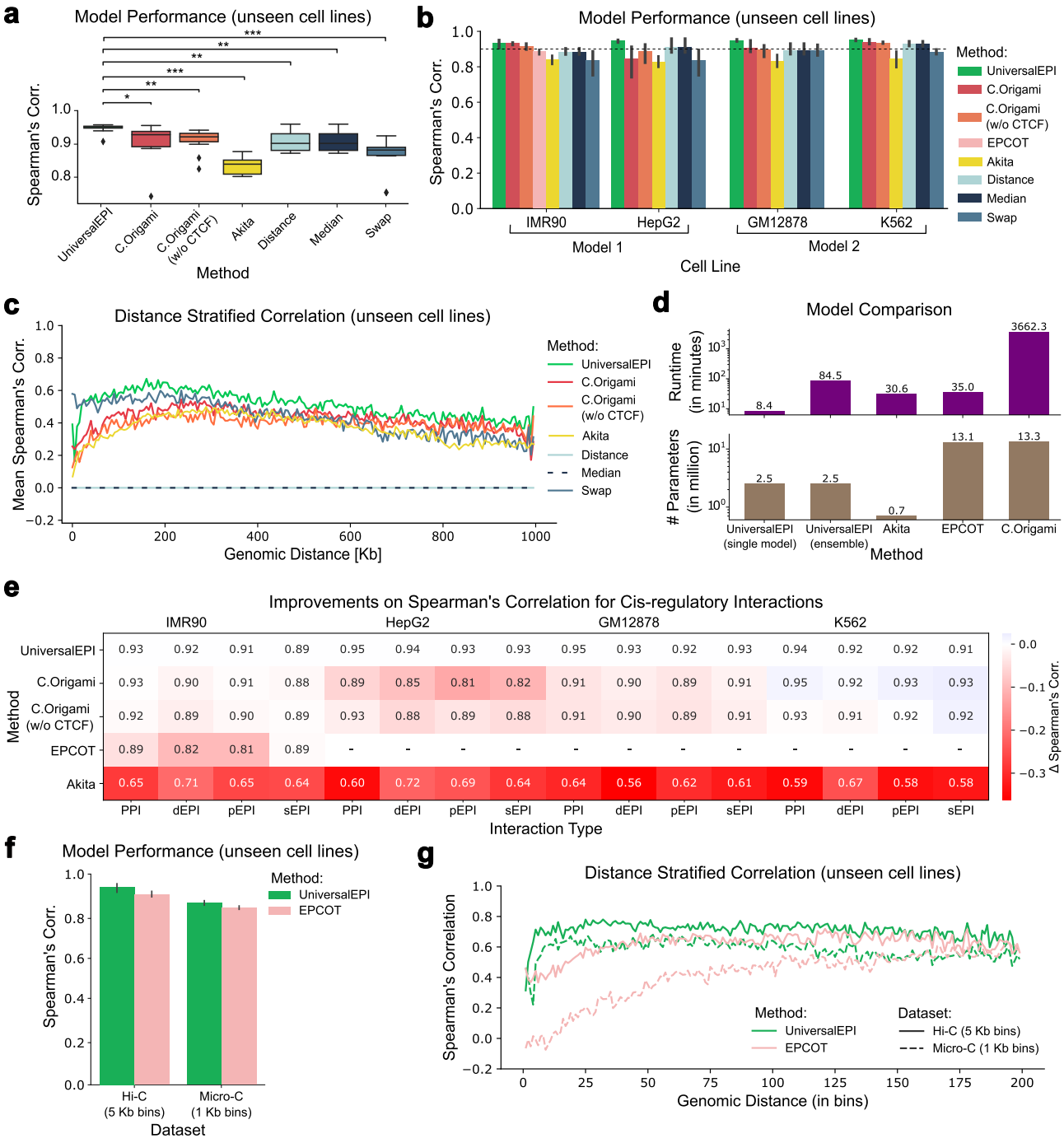


Figure 3. Benchmarking of the UniversalEPI method. (**a**) Aggregated model performances for prediction of chromatin interactions in unseen cell types. Each point corresponds to Spearman's correlation coefficient between predictions and ground-truth Hi-C for an unseen chromosome in unseen cell lines, IMR90 and HepG2 (Model 1), or GM12878 and K562 (Model 2). For each method, Model 1 is trained on GM12878 and K562 cells, whereas Model 2 is trained on IMR90 and HepG2 cells. A Wilcoxon signed rank test is used to compare the methods; $^{***}P < .001$, $^{**}P < .01$, $^{*}P < .05$. (**b**) Model performances for prediction of chromatin interactions in unseen cell types stratified by test cell line. The dotted line indicates Spearman's correlation of 0.9. (**c**) Line plot showing the mean distance-stratified Spearman's correlation across four cell lines. (**d**) Inference runtime and total number of model parameters are compared between different methods. The inference runtime is calculated for making the predictions using the test chromosomes of HepG2 cell line. (**e**) Method performances for prediction of the Hi-C signal for different types of *cis*-regulatory interactions. The change in Spearman's correlation is calculated with respect to UniversalEPI. Positive values suggest that C.Origami or its variant have higher scores than UniversalEPI. dEPI: distal Enhancer–Promoter Interactions (further than 2 kb), pEPI: proximal Enhancer–Promoter Interactions (within 2 kb), PPI: promoter–promoter interaction, sEPI: super enhancer–promoter interactions. (f, g) Benchmarking UniversalEPI against EPCOT on high-resolution chromatin contact maps in unseen cell types. (**f**) Overall performance (Spearman's correlation). (**g**) Performance stratified by genomic distance.

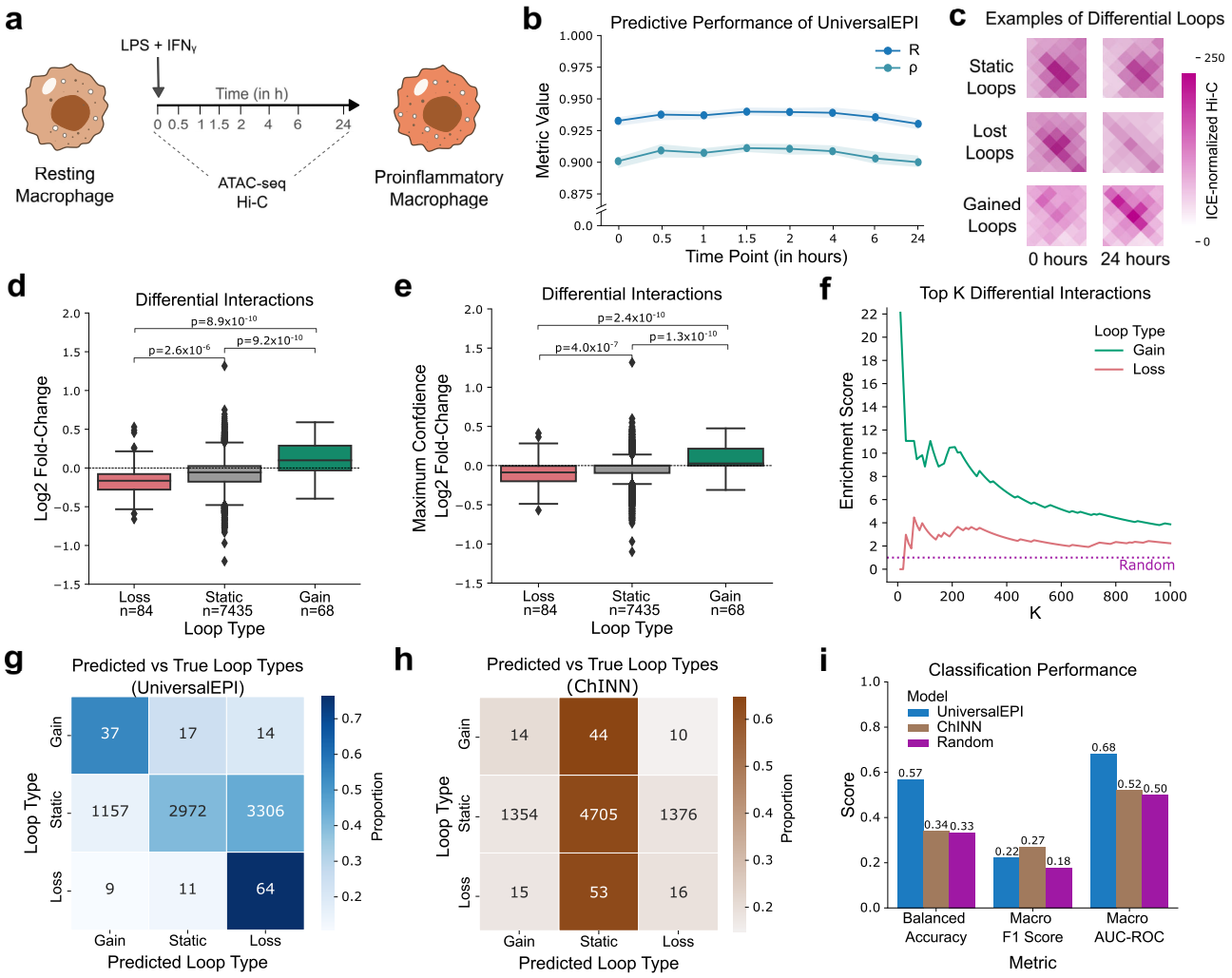


Figure 4. UniversalEPI captures chromatin dynamics during cell reprogramming. (**a**) Experimental layout defined in Reed *et al.* [30]. Macrophages derived from the human THP-1 monocytic cell line were activated by treating with LPS + IFN_γ *in vitro*; ATAC-seq and Hi-C were subsequently measured at eight time points. (**b**) Performance of UniversalEPI ensembles (trained on GM12878 and K562 cells) using ATAC-seq for each of the eight time points, quantified by Pearson's (***R***) and Spearman's (ρ) correlation. (**c**) Examples of static, lost, and gained loops in the chromatin between 0 and 24 h identified by Reed *et al.* (**d**) Differential interactions predicted by the UniversalEPI model between the 24-h time point and the 0-h for the differential loops. The significance is measured using a Mann–Whitney-Wilcoxon test with Bonferroni correction. (**e**) The maximum-confidence $\log_2$ FC between Hi-C values measured by UniversalEPI ensemble predictions for the differential loops. The significance is measured using a Mann–Whitney–Wilcoxon test with Bonferroni correction. (**f**) Gain and loss enrichment score over random for selecting a gained or lost loop, respectively, in the K most differential interactions as predicted by the UniversalEPI model. The pink dotted line represents the enrichment score of labeling a random interaction as gained or lost. (**g**) Confusion matrix between experimental loop types and predictions by UniversalEPI. (**h**) Confusion matrix between experimental loop types and predictions by ChINN. (**i**) Comparison of UniversalEPI against ChINN and the random baseline across three key classification metrics: balanced accuracy, macro F1 score, and macro AUC-ROC.

matin loops, UniversalEPI effectively identified positive and negative changes in the Hi-C signal during macrophage activation at 24 and 0 h time points solely based on the variation in ATAC-seq inputs (Fig. 4d). The application of the maximum-confidence $\log_2$ FC, based on the estimated prediction uncertainty, further improved the significance of predicted differential interactions (Fig. 4e and Supplementary Section S7). Moreover, using the maximum-confidence $\log_2$ FC, UniversalEPI accurately distinguished gained and lost loops from static loops with a true positive rate that is at least four times higher than a random classification (Fig. 4f). Interactions were subsequently classified as gained, lost, or static based on the predicted maximum-confidence $\log_2$ FC. The confusion matrix for UniversalEPI (Fig. 4g) shows that the model correctly classified gained and lost loops more frequently than static loops, demonstrating high precision for dynamic regulatory changes. In contrast, the corresponding confusion matrix for ChINN [21], an anchor-based deep-learning classifier that predicts the presence or absence of chromatin interactions from DNA sequence and ATAC-seq, revealed substantially lower performance on the same task (Fig. 4h and Supplementary Section S8). Comprehensive benchmarking confirmed that UniversalEPI substantially outperformed ChINN and a random baseline on balanced accuracy and macro AUC-ROC, while achieving comparable performance on macro F1-score (Fig. 4i). Therefore, we conclude that the unique feature of UniversalEPI—estimation of the maximum-confidence FC between conditions based on the reported uncertainty values—allows for accurate prediction of differential EPIs across conditions.

UniversalEPI predicts differential chromatin interactions using single-cell ATAC-seq data

To demonstrate the applicability to single-cell studies, we used UniversalEPI to predict chromatin structure using as inputs pseudo-bulks of scATAC-seq data (Fig. 5a). We compared the performance of UniversalEPI with ChromaFold [24], the only other method that predicts bulk Hi-C from scATAC-seq and DNA sequence as inputs. Because ChromaFold was specifically designed to work on single-cell inputs, we did not include it in the above benchmarking on bulk ATAC-seq. scATAC-seq and Hi-C data from GM12878, K562, HepG2, and IMR90 cell lines were used in this analysis, with GM12878 and HepG2 cells employed for training both models. Keeping the experimental design similar to the ChromaFold study by Gao *et al.* [24], we used Hi-C at 10 kb resolution with Hi-C-DC + z-scores as targets [64] (Supplementary Section S9).

The two methods were compared based on their ability to predict Hi-C values on unseen chromosomes in the unseen cell lines IMR90 and K562. UniversalEPI significantly outperformed ChromaFold on both unseen cell lines (Fig. 5b and c). UniversalEPI also obtained a higher distance-stratified correlation, especially for short-to-medium-range chromatin interactions (50–600 kb), highlighting the ability of our model to maintain its generalization capacity while extending to pseudo-bulk scATAC-seq data (Fig. 5d).

Further, we used 10× multiome data (scATAC-seq and scRNA-seq) from eight human EAC [65] to assess potential differences in chromatin structure between two cancer cell states: differentiated and undifferentiated malignant cells expressing cNMF4 and cNMF5 transcriptional programs, respectively, as described in Yates *et al.* [65] (Supplementary Section S10). Pseudo-bulk ATAC-seq profiles were obtained from merging scATAC-seq data from all cells across patients expressing a given transcriptional program. These pseudo-bulk ATAC-seq profiles were then used to predict Hi-C interactions between accessible regions using the UniversalEPI model 1, pre-trained on GM12878 and K562 cell lines (Fig. 5e).

Based on predicted Hi-C data, we defined values of promoter activity for each gene and compared them with measured gene expression from pseudo-bulk scRNA-seq. Promoter activity acts as a representative for gene expression and was defined as a function of ATAC-seq and the predicted Hi-C. We modeled promoter activity as a combination of ATAC-seq promoter signal, ATAC-seq enhancer signals, and the predicted Hi-C interaction strengths between the promoter and enhancers (Supplementary Section S11). The derived promoter activity not only showed a high correlation with gene expression (Supplementary Fig. S14) but also a positive correlation with differential gene expression between undifferentiated and differentiated cells in EAC (Pearson's $R = 0.224$, Fig. 5f). Thus, UniversalEPI-derived differential chromatin interactions can indicate the strength of regulatory effects on gene expression. Yates *et al.* also identified mTFs for each program, including *LCOR* and *MECOM* in differentiated and undifferentiated EAC cells, respectively [65]. These TFs play a pivotal role in defining and maintaining malignant cell states. Specifically, *LCOR* expression drives tumor cell differentiation and reduces tumor growth, whereas low *LCOR* expression is associated with suppression of antigen processing and presentation, enabling resistance to immune checkpoint blockade therapy in breast cancer [66, 67]. Conversely, overexpression of *MECOM* has been linked with activating oncogene expression [68], promoting stem cell-like properties, and reducing apoptosis in different cancers [69, 70]. The UniversalEPI-predicted Hi-C signal for the two malignant cell states of EAC was in agreement with these findings, showing higher estimated promoter activity of *LCOR* in differentiated cells relative to the undifferentiated cells, in line with the increased *LCOR* gene expression, and vice versa for *MECOM* (Fig. 5f).

Furthermore, using UniversalEPI-estimated uncertainty values, we identified significant differences in Hi-C interactions for the *LCOR* and *MECOM* promoters between the two malignant transcriptional states. In differentiated EAC cells, the *LCOR* promoter exhibited stronger interactions with proximal and distal enhancers both upstream and downstream of the gene (Fig. 5g). Conversely, in undifferentiated EAC cells, the *MECOM* gene promoter formed strong interactions with downstream enhancers in its vicinity, presumably contributing to gene activation (Fig. 5h). The uncertainty estimates enabled the model to filter low-confidence interactions, which may assist in prioritizing candidate regulatory elements for further investigation (Supplementary Fig. S15). This example illustrates the unique feature of UniversalEPI, which enables accurate identification of differential EPIs across conditions.

UCSC Genome Browser hub of UniversalEPI-predicted Hi-C interactions

We applied UniversalEPI to the ENCODE ATAC-seq compendium to generate genome-wide predicted Hi-C interaction maps for 157 datasets spanning 116 cell lines and 41 primary cells. To make these maps broadly usable, we provided a public UCSC Genome Browser [71] track hub (https://genome.ucsc.edu/cgi-bin/hgHubConnect). The hub offers two track types: (i) log-transformed, ICE-normalized predictions for quantitative analyses and (ii) z-score–normalized tracks that highlight strong long-range interactions (Supplementary Section S12). As an example, all regions with strong interactions in the neighborhood of the *MYC* gene in the breast cancer MCF-7 cell line can be readily identified using this resource (Fig. 6a), highlighting the previously reported functional interaction with the 67-kb upstream enhancer region [72].

We performed bioinformatics analysis of the predicted Hi-C tracks. Using the predicted quantitative Hi-C values together with experimental ATAC-seq signal for the 116 cell lines, we computed a promoter-activity score for each protein-coding gene (Supplementary Section S11) and performed hierarchical clustering of cell lines based on these scores (Supplementary Section S13). Four major clusters emerged (Fig. 6b), primarily separating samples by malignancy and lineage. Non-malignant cell lines formed two lineage-specific clusters, one containing induced pluripotent stem cells (iPSCs), such as WTC-11 and GM23338, and another comprising B-cell lymphoblastoid lines. Malignant cell lines were grouped mainly together, with DND-41 forming a cluster on its own. This stratification indicates that the resource captures biologically meaningful variation and supports downstream functional analysis.

Next, we assessed the functional significance of predicted Hi-C interactions by comparing malignant and non-malignant cell lines, focusing on two gene categories with contrast-

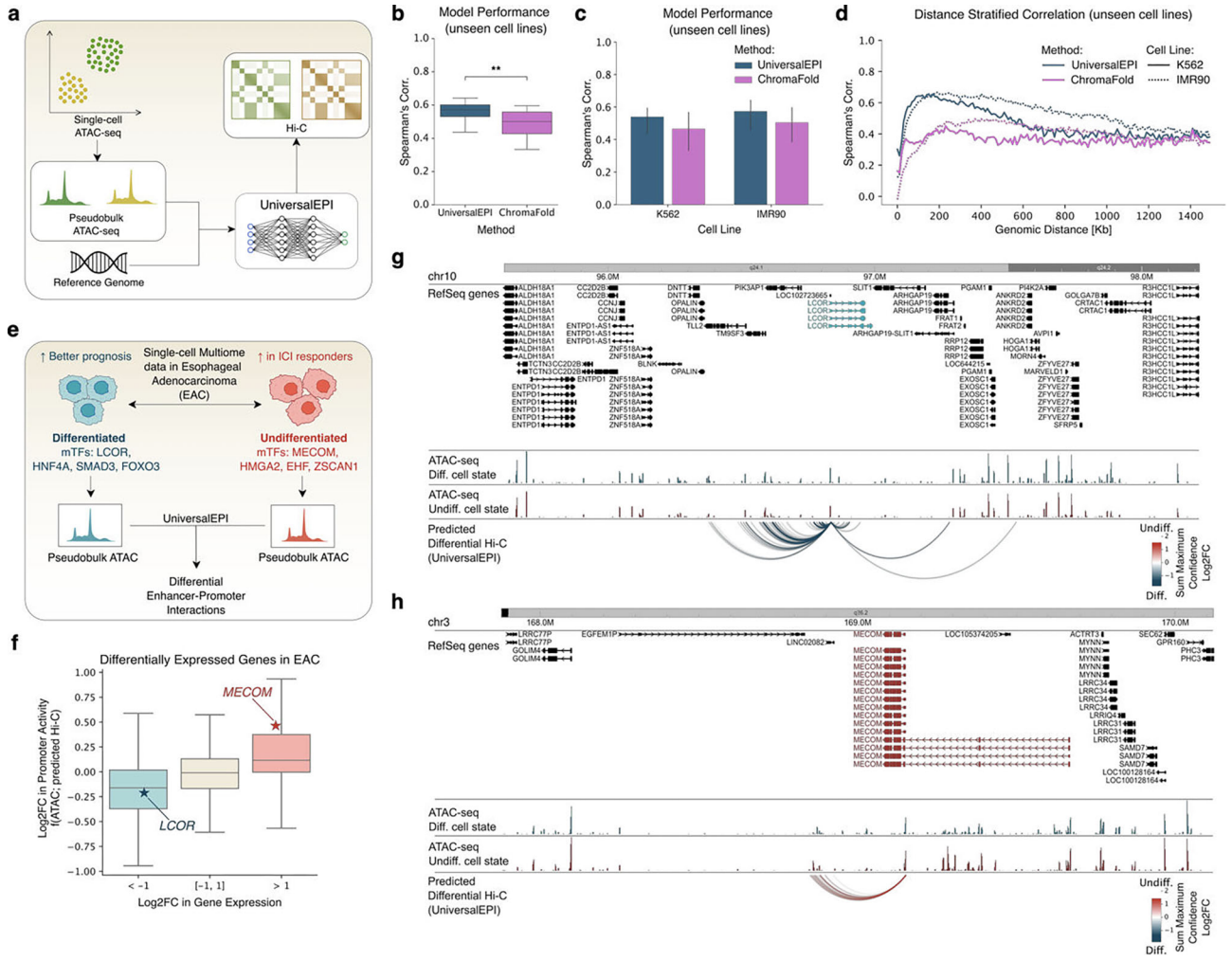


Figure 5. UniversalEPI reveals changes in chromatin structure in complex tissues using pseudo-bulk scATAC-seq data. (**a**) UniversalEPI can predict cell-type-specific Hi-C-derived chromatin interactions using scATAC-seq data. (**b**) Comparison between UniversalEPI and ChromaFold on their ability to predict bulk Hi-C from scATAC-seq using GM12878 and HepG2 cell lines as training. Spearman's correlation is calculated on unseen chromosomes (chr5, chr18, chr20, chr21) of unseen cell lines (K562, IMR90). Wilcoxon signed-rank test is used to test significance. $^{**}P \leq .01$. (**c**) Performance comparison on unseen chromosomes stratified by unseen cell lines. (**d**) Distance-stratified Spearman's correlation based on the unseen chromosomes and cell lines. (**e**) The 10× multiome data (scRNA-seq and scATAC-seq) of differentiated and undifferentiated malignant cells are obtained from human EAC tumors [65]. Master transcription factors (mTFs) for each program are identified in Yates *et al.* [65]. The scATAC-seq data from the candidate cells expressing each program are merged from multiple patient tumors to obtain transcriptional program-specific pseudo-bulk ATAC-seq profiles. These are then used as inputs to the pretrained UniversalEPI model to obtain Hi-C interactions for each program. (**f**) A positive correlation is observed between the $\log_2$ FC in bulkified gene expression profiles and log2FC in estimated promoter activity. The blue star highlights the *LCOR* gene, which is a mTF for differentiated cells, whereas the red star highlights the *MECOM* gene, which is a mTF for undifferentiated cells. (**g**, **h**) The change in predicted Hi-C interactions between accessible regions, measured by the sum of $\log_2$ FC between undifferentiated and differentiated cells, is shown for the *LCOR* and *MECOM* gene promoters.

ing roles in malignancy: oncogenes and tumor suppressors [73]. The overall number of strong long-distance EPIs (z-score > 2) showed significant differences between malignant and non-malignant cell lines (oncogenes: avg. 448 versus 553, Mann–Whitney U test $P < 10^{-4}$; tumor suppressors: avg. 519 versus 594, Mann–Whitney U test $P < 10^{-7}$). Moreover, we identified a distinct subset of 18 genes with significantly different long-distance interactions (FDR-corrected [74] $P < .01$, absolute $\log_2$ FC > 1). Notably, most of these differentially connected genes (13 of 18) showed a gain in promoter connectivity in malignant cells, whereas a smaller subset (five genes) showed a corresponding loss. This subset was enriched for oncogenes: 13 of the 18 genes (72%), including *TERT*, *FGFR3*, and *PDCD1LG2* (two-sided Binomial test $P = .098$) (Fig. 6c).

To determine whether the observed differences in predicted promoter connectivity between malignant and non-malignant cell lines could be explained solely by changes in promoter accessibility, we repeated the same analyses using promoter ATAC-seq signal instead of predicted Hi-C connectivity (Supplementary Fig. S16). While promoter accessibility also differed between malignant and non-malignant cell lines, the set of genes identified was not fully concordant with those observed using predicted chromatin interactions, indicating that long-range interaction patterns may provide complementary regulatory information beyond local chromatin openness.

We next examined representative genes illustrating the functional impact of these directional changes predicted by our model. For instance, the *TERT* oncogene, which encodes the telomerase catalytic subunit, exhibited significantly

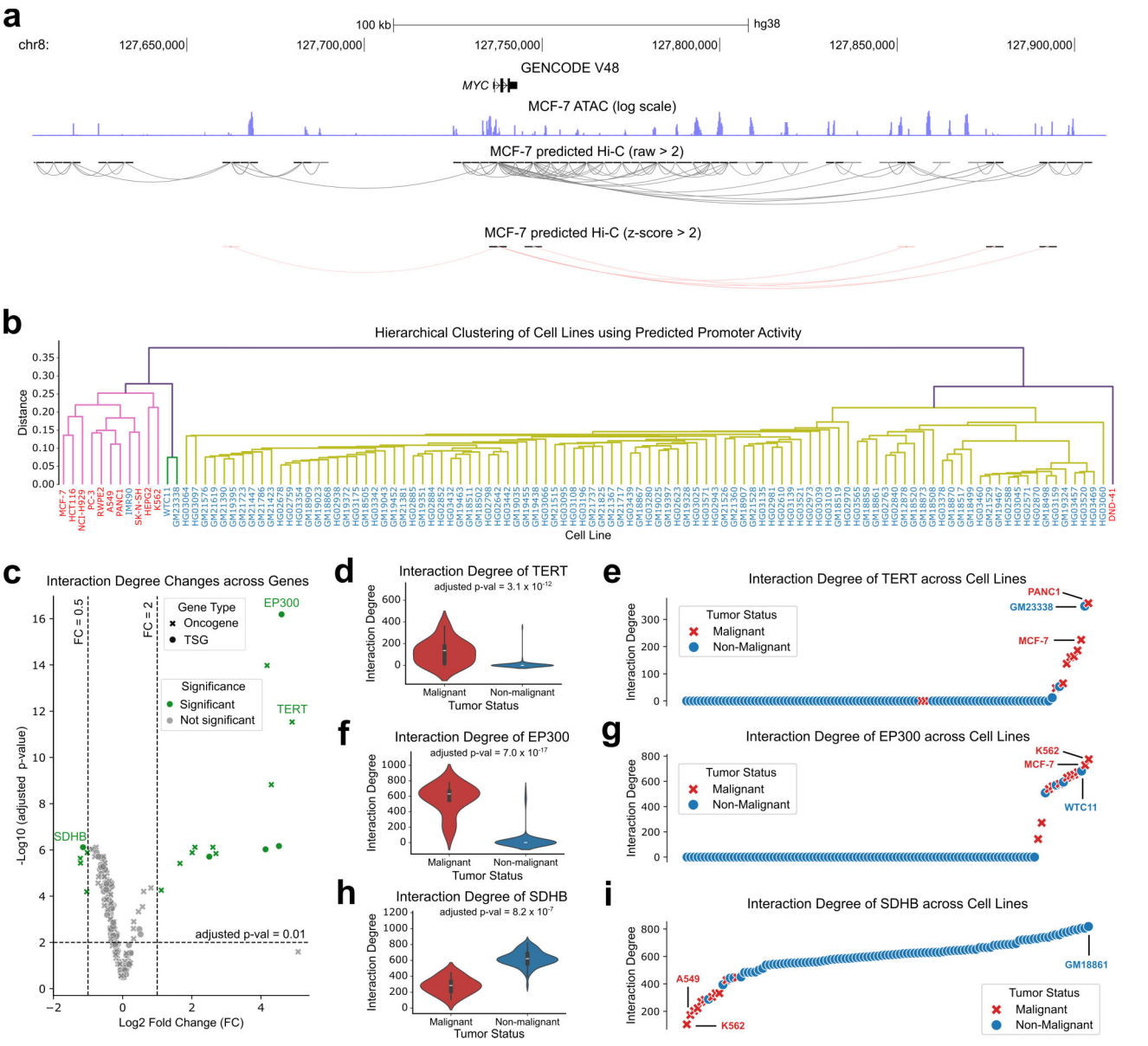


Figure 6. UniversalEPI-derived promoter activity and genomic connectivity reveal differential regulation across ENCODE cell lines. (**a**) An example from the UCSC Genome Browser highlighting strong interactions (ICE-normalized Hi-C values and z-scores) around the *MYC* gene in the MCF-7 cell line. (**b**) Dendrogram showing hierarchical clusters of cell lines based on promoter activity across all protein-coding genes. 1—Pearson's correlation is used as a distance. Malignant cell lines are shown in red. (**c**) Volcano plot highlighting change in interaction degree (number of interactions with the promoter) between malignant and non-malignant cell lines for all oncogenes and tumor suppressor genes. The FC is calculated as the ratio of the average number of interactions in malignant cell lines to non-malignant cell lines. (**d–i**) Interaction degree between malignant and non-malignant cell lines for *TERT* (d), *EP300* (f), and *SDHB* (h) genes. The interaction degrees in each cell line for these genes are highlighted in (e), (g), and (i), respectively.

higher promoter interaction scores in malignant cell lines compared to non-malignant ones (Fig. 6d). This effect was prominent in several cancer lines (e.g. PANC1, MCF-7, and A549) and, notably, also in the iPSC line GM23338 (Fig. 6e). Promoter accessibility was also elevated in malignant contexts (Supplementary Fig. S16). This observation aligns with the known augmentation of telomerase expression and activity in both cancer and pluripotent cells [75–78]. Similarly, the *EP300* gene, which encodes the p300 histone acetyltransferase, also showed significantly increased promoter connectivity in malignant cell lines, including K562 and MCF-7 (Fig. 6f and g). While this is consistent with *EP300*'s role as a transcriptional coactivator frequently dysregulated in malignancy [79, 80], the corresponding promoter accessibility signals were similar between malignant and non-malignant cell lines (Supplementary Fig. S16), suggesting that changes in long-range connectivity are not fully captured by local chromatin accessibility. In contrast, the *SDHB* tumor suppressor gene exhibited a significant loss of promoter connectivity in malignant cell lines (Fig. 6h and i), mirroring the reduced expression and function often observed for this gene in familial and sporadic tumors [81, 82]. However, promoter accessibility at the SDHB locus was comparable between malignant and non-malignant cell lines (Supplementary Fig. S16), again indicating a disconnect between 3D promoter connectivity and local accessibility.

Taken together, these results indicate that predicted long-range chromatin interactions capture regulatory alterations associated with malignancy that are only partially reflected by promoter accessibility alone. This suggests that modeling 3D genome connectivity provides complementary insight beyond local chromatin state when characterizing gene regulatory changes in cancer.

Discussion

In this work, we present UniversalEPI, an attention-based deep ensemble model that is trained on chromatin accessibility (ATAC-seq), DNA sequence, and TF binding data (ChIP-seq) from multiple cell lines to accurately predict chromatin interactions on unseen cell types without retraining solely based on DNA sequence and chromatin accessibility. Other models require several data modalities as inputs and, because of their large size, can only be trained on one or two cell types. UniversalEPI overcomes these drawbacks by operating on the DNA sequence within open chromatin. By ignoring information in closed chromatin, other than assessing the distance between open chromatin regions, UniversalEPI is lightweight and demonstrates notable scalability (Fig. 1). In contrast to other techniques, such as C.Origami, which was originally trained on a single cell line [22], this enables our model to be trained on several cell lines concurrently in less than a day on a standard GPU and to have vastly improved inference time (Fig. 3d). The latter will allow our model to be used on large datasets such as The Cancer Genome Atlas (TCGA) in the future, e.g. to assess the effects of non-coding variants on EPIs.

Using CNN layers and transformer blocks, UniversalEPI effectively captures the essential TF binding motifs and spatial dependencies underlying interactions between regions of open chromatin. The model operates robustly across multiple resolutions, maintaining consistent performance without retraining (Fig. 3f and g). Consequently, UniversalEPI consistently outperforms the state-of-the-art methods, including C.Origami and EPCOT, and shows Spearman's correlation between the predicted and ground-truth Hi-C values on unseen cell types above 0.9 in all experimental settings (Figs 3b and 4b).

By applying UniversalEPI to predict changes in chromatin architecture in human macrophages stimulated with LPS and IFN_{γ}, we demonstrated it can perform in a zero-shot setting to find differential EPIs. Thus, UniversalEPI functions as a generalizable model that can identify modified chromatin interactions across conditions in cell types that are unseen during training (Fig. 4). The model's ability to detect dynamic changes in chromatin organization, which can be crucial in examining how different cell types respond to drugs and environmental stimuli, was demonstrated by its accurate identification of differential chromatin loops. UniversalEPI also displayed high performance against the only other model, ChromaFold, which was designed to predict bulk Hi-C using pseudo-bulk scATAC-seq data (Fig. 5b and c). Furthermore, different chromatin looping patterns were identified by UniversalEPI for mTF genes in EAC cells in different states of differentiation (Fig. 5e–g). This feature means that UniversalEPI can be used to accurately investigate the characteristics and drivers of heterogeneity in chromatin organization that occurs in complex tissues and cancer cell states. Finally, we generated a comprehensive public resource of predicted Hi-C contact maps for 157 ENCODE cell lines and primary cells (Fig. 6). This resource, available as a UCSC Genome Browser hub, provides both ICE-normalized and z-score-normalized tracks, allowing researchers to explore predicted 3D chromatin architecture across a wide range of cell types.

Its predictive power will enable UniversalEPI to be used to study changes in chromatin organization in the presence of non-coding mutations and structural variations, thereby helping to decipher the regulatory mechanisms of genetic diseases. In this work, we performed *in silico* inversions of CTCF binding motifs to reveal changes in chromatin interactions using UniversalEPI; future work could apply the same principles to large cancer datasets such as TCGA, which comprises data on chromatin accessibility and DNA sequence variation.

As UniversalEPI is trained to predict the Hi-C signal, the strong EPIs predicted by UniversalEPI are not strictly functional. It is known that enhancers and promoters can gain chromatin accessibility and interact even before gene expression is activated [9]. Additional data, such as profiling of histone H3 lysine K27 acetylation (H3K27ac) or enhancer RNA (eRNA) transcription, may be needed to assess the functionality of predicted strong EPIs.

Finally, the computational architecture of UniversalEPI sheds light on the ubiquitousness of chromatin folding mechanisms between promoters and enhancers in the cell types used for validation (lung fibroblasts, B lymphocytes, macrophages, chronic myelogenous leukemia cells, and hepatocellular carcinoma cells): the method demonstrates top performance while discarding all information about non-accessible chromatin and cell-type-specific TF binding motifs. Additional validation experiments may be required to check whether the model can show equally high performance in all cell types and whether highly specialized, cell-type-specific mechanisms that regulate chromatin interactions exist. As the numbers of ATAC-seq and Hi-C datasets increase, UniversalEPI can be used to explore these questions further.

In the future, UniversalEPI could benefit from being trained on more datasets to further improve its generalizability. Moreover, the first stage of UniversalEPI could be enhanced by incorporating techniques such as adversarial training [83–85], which would additionally help prevent the model from capturing cell-type-specific information and improve its ability to generalize across cell types. Nonetheless, UniversalEPI achieves state-of-the-art performance in predicting EPIs in unseen cell types from only DNA sequence and chromatin accessibility profiles, and because of its unique feature of estimating prediction uncertainty, significant differential interactions can be detected across conditions. Therefore, UniversalEPI brings a major advancement in being able to predict and experiment on chromatin interactions *in silico* that can be widely applied to study complex regulatory landscapes across the genome.

Acknowledgements

Author contributions: Aayush Grover (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal]), Lin Zhang (Formal analysis [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal]), Till Muser (Formal analysis [equal], Methodology [equal], Software [equal], Writing—review & editing [equal]), Simeon Häfliger (Formal analysis [equal], Methodology [equal], Software [equal]), Minjia Wang (Software [equal],

Validation [equal]), Josephine Yates (Data curation [equal], Formal analysis [equal]), Marie-Claire Indilewitsch (Validation [equal]), Yizhen Wang (Formal analysis [equal], Software [equal]), Eliezer Mendel Van Allen (Supervision [equal], Writing—review & editing [equal]), Fabian J. Theis (Supervision [equal], Writing—review & editing [equal]), Ignacio L. Ibarra (Methodology [equal], Supervision [equal]), Ekaterina Krymova (Methodology [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]), and Valentina Boeva (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Project administration [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd, and Omniscope Ltd, and has ownership interest in Dermagnostix GmbH and Cellarity. I.L.I. currently works at Bioptimus.

Funding

This project is partially funded by the Swiss Data Science Center (SDSC) collaborative projects grant (C22-09); A.G. is funded by the Swiss Government Excellence Scholarship (ESKAS-Nr: 2021.0468). Funding to pay the Open Access publication charges for this article was provided by ETH Zurich core funding to V.B.

Data availability

All the raw and processed datasets corresponding to the cell lines were obtained from publicly accessible databases, including ENCODE (https://www.encodeproject.org/), 4DNucleome (4DN) (https://data.4dnucleome.org/), and Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), with accession codes mentioned in Supplementary Sheet 1. This includes the THP-1 cell line used for macrophage activation, generated in Reed *et al*. The raw single-cell ATAC-seq and single-cell RNA-seq datasets for cell differentiation in esophageal carcinoma were obtained from the Database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/) with accession number phs003438.v1.

The precomputed Hi-C predictions for the 157 ENCODE ATAC-seq tracks of cell lines and primary cells are made available as a track hub for the UCSC Genome Browser at https://genome.ucsc.edu/cgi-bin/hgHubConnect. The mapping between the generated interaction files and ENCODE datasets can be found in Supplementary Table S2.

The source code of UniversalEPI is made available at https://github.com/BoevaLab/UniversalEPI and deposited to https://doi.org/10.5281/zenodo.14622040. The tutorial is available at https://github.com/BoevaLab/UniversalEPI/wiki.

References

1. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 2013;14:390–403. https://doi.org/10.1038/nrg3454
2. Cavalheiro GR, Pollex T, Furlong EE. To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? *Curr Opin Genet Dev* 2021;67:119–29. https://doi.org/10.1016/j.gde.2020.12.015
3. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* 2019;20:437–55. https://doi.org/10.1038/s41576-019-0128-0
4. Weintraub AS, Li CH, Zamudio AV *et al*. YY1 Is a structural regulator of enhancer-promoter loops. *Cell* 2017;171:1573–88.e28. https://doi.org/10.1016/j.cell.2017.11.008
5. Deshane J, Kim J, Bolisetty S *et al*. Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells. *J Biol Chem* 2010;285:16476–86. https://doi.org/10.1074/jbc.M109.058586
6. Yang Y, Zhang R, Singh S *et al*. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics* 2017;33:i252–60. https://doi.org/10.1093/bioinformatics/btx257
7. Hsieh T-HS, Cattoglio C, Slobodyanyuk E *et al*. Enhancer-promoter interactions and transcription are maintained upon acute loss of CTCF, cohesin, WAPL, or YY1. *Nat Genet* 2022;54:1919–1932. https://doi.org/10.1038/s41588-022-01223-8
8. Uyehara CM, Apostolou E. 3D enhancer-promoter interactions and multi-connected hubs: organizational principles and functional roles. *Cell Rep* 2023;42:112068. https://doi.org/10.1016/j.celrep.2023.112068
9. Pollex T, Rabinowitz A, Gambetta MC *et al*. Enhancer-promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. *Nat Genet* 2024;56:686–96. https://doi.org/10.1038/s41588-024-01678-x
10. 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
11. Feng Y, Pauklin S. Revisiting 3D chromatin architecture in cancer development and progression. *Nucleic Acids Res* 2020;48:10632–47. https://doi.org/10.1093/nar/gkaa747
12. 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
13. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;17:1111–7. https://doi.org/10.1038/s41592-020-0958-x
14. 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
15. 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
16. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;48:488–96. https://doi.org/10.1038/ng.3539
17. 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
18. Chen K, Zhao H, Yang Y. Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. *Brief Bioinform* 2022;23:bbab577. https://doi.org/10.1093/bib/bbab577
19. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* 2019;47:e60. https://doi.org/10.1093/nar/gkz167
20. Agarwal A, Chen L. DeepPHiC: predicting promoter-centered chromatin interactions using a novel deep learning approach.

Bioinformatics 2023;39:btac801. https://doi.org/10.1093/bioinformatics/btac801
21. 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
22. 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
23. 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
24. Gao VR, Yang R, Das A *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
25. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014;15:234–46. https://doi.org/10.1038/nrg3663
26. Makhlouf M, Ouimette J-F, Oldfield A *et al.* A prominent and conserved role for YY1 in Xist transcriptional activation. *Nat Commun* 2014;5:4878. https://doi.org/10.1038/ncomms5878
27. O'Connor L, Gilmour J, Bonifer C. The role of the ubiquitously expressed transcription factor Sp1 in tissue-specific transcriptional regulation and in disease. *Yale J Biol Med* 2016;89:513–25.
28. Seitzer M, Tavakoli A, Antic D *et al.* On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*. 2022. https://doi.org/10.48550/arXiv.2203.09168
29. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. 30, 2017.https://doi.org/10.48550/arXiv.1612.01474
30. Reed KSM, Davis ES, Bond ML *et al.* Temporal analysis suggests a reciprocal relationship between 3D chromatin structure and transcription. *Cell Rep* 2022;41:111567. https://doi.org/10.1016/j.celrep.2022.111567
31. Reiff SB, Schroeder AJ, Kırlı K *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* 2022;13:2365. https://doi.org/10.1038/s41467-022-29697-4
32. 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
33. Wolff J, Rabbani L, Gilsbach R *et al.* Galaxy HiCExplorer 3: a web server for reproducible hi-C, capture hi-C and single-cell hi-C data analysis, quality control and visualization. *Nucleic Acids Res* 2020;48:W177–84. https://doi.org/10.1093/nar/gkaa220
34. Virtanen P, Gommers R, Oliphant TE *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2
35. Hong CKY, Feng F, Ramanathan V *et al.* Genome structure mapping with high-resolution 3D genomics and deep learning. bioRxiv, https://doi.org/10.1101/2025.05.06.650874, 7 May 2025, preprint: not peer reviewed.
36. Ramírez F, Ryan DP, Grüning B *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160–165. https://doi.org/10.1093/nar/gkw257
37. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;11:R25. https://doi.org/10.1186/gb-2010-11-3-r25
38. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;52:D174–82. https://doi.org/10.1093/nar/gkad1059
39. Swindell WR, Johnston A, Sun L *et al.* Meta-profiles of gene expression during aging: limited similarities between mouse and human and an unexpectedly decreased inflammatory signature. *PLoS One* 2012;7:e33204. https://doi.org/10.1371/journal.pone.0033204
40. Khalid M, Baber J, Kasi MK *et al.* Empirical evaluation of activation functions in deep convolution neural network for facial expression recognition. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. 2020, pp. 204–7. https://doi.org/10.1109/TSP49548.2020.9163446
41. Yamada Y, Lindenbaum O, Negahban S *et al.* Feature selection using stochastic Gates. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 10648–59. https://doi.org/10.48550/arXiv.1810.04247
42. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. 30, 2017. https://doi.org/10.48550/arXiv.1706.03762
43. Carty M, Zamparo L, Sahin M *et al.* An integrated model for detecting significant chromatin interactions from high-resolution hi-C data. *Nat Commun* 2017;8:15454. https://doi.org/10.1038/ncomms15454
44. Nassar LR, Barber GP, Benet-Pagès A *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* 2023;51:D1188–95. https://doi.org/10.1093/nar/gkac1072
45. Karimzadeh M, Ernst C, Kundaje A *et al.* Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* 2018;46:e120. https://doi.org/10.1093/nar/gky677
46. Ansel J, Yang E, He H *et al.* PyTorch .2. In :faster machine learning through dynamic Python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. ACM, La Jolla CA USA, 2024, pp. 929–47. https://doi.org/10.1145/3620665.3640366
47. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *International Conference on Learning Representations*. 2015. https://doi.org/10.48550/arXiv.1412.6980
48. Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the Human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;159:1665–80. https://doi.org/10.1016/j.cell.2014.11.021
49. Bailey SD, Zhang X, Desai K *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* 2015;2:6186. https://doi.org/10.1038/ncomms7186
50. Zhou Q, Yu M, Tirado-Magallanes R *et al.* ZNF143 mediates CTCF-bound promoter-enhancer loops required for murine hematopoietic stem and progenitor cell function. *Nat Commun* 2021;12:43. https://doi.org/10.1038/s41467-020-20282-1
51. Magnitov MD, Maresca M, Alonso Saiz N *et al.* ZNF143 is a transcriptional regulator of nuclear-encoded mitochondrial genes that acts independently of looping and CTCF. *Mol Cell* 2024;85:24–41.e11. https://doi.org/10.1016/j.molcel.2024.11.031
52. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations*. 2019. https://doi.org/10.48550/arXiv.1711.05101
53. Kelley DR, Reshef YA, Bileschi M *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* 2018;28:739–50. https://doi.org/10.1101/gr.227819.117
54. de Wit E, Vos ESM, Holwerda SJB *et al.* CTCF binding polarity determines chromatin looping. *Mol Cell* 2015;60:676–84. https://doi.org/10.1016/j.molcel.2015.09.023
55. 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*, ICML'17. JMLR.org, Sydney, NSW, Australia, 2017, pp. 3145–53. https://doi.org/10.48550/arXiv.1704.02685
56. 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
57. Valton A-L, Dekker J. TAD disruption as oncogenic driver. *Curr Opin Genet Dev* 2016;36:34–40. https://doi.org/10.1016/j.gde.2016.03.008
58. Hnisz D, Weintraub AS, Day DS *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 2016;351:1454–8. https://doi.org/10.1126/science.aad9024
59. Kloetgen A, Thandapani P, Ntziachristos P *et al.* Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat Genet* 2020;52:388–400. https://doi.org/10.1038/s41588-020-0602-9
60. Liu NQ, Ter Huurne M, Nguyen LN *et al.* The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression. *Nat Commun* 2017;8:14418. https://doi.org/10.1038/ncomms14418
61. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE *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
62. Tang F, Yang Z, Tan Y *et al.* Super-enhancer function and its application in cancer targeted therapy. *NPJ Precis Oncol* 2020;4:2. https://doi.org/10.1038/s41698-020-0108-z
63. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* 2016;44:D164–D171. https://doi.org/10.1093/nar/gkv1002
64. Sahin M, Wong W, Zhan Y *et al.* HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. *Nat Commun* 2021;12:3366. https://doi.org/10.1038/s41467-021-23749-x
65. Yates J, Mathey-Andrews C, Park J *et al.* Cell states and neighborhoods in distinct clinical stages of primary and metastatic esophageal adenocarcinoma. *Cell Rep Med* 2025;6:102188. https://doi.org/10.1101/2024.08.17.608386
66. Celià-Terrassa T, Liu DD, Choudhury A *et al.* Normal and cancerous mammary stem cells evade interferon-induced constraint through the miR-199a-LCOR axis. *Nat Cell Biol* 2017;19:711–23. https://doi.org/10.1038/ncb3533
67. Pérez-Núñez I, Rozalén C, Palomeque JÁ *et al.* LCOR mediates interferon-independent tumor immunogenicity and responsiveness to immune-checkpoint blockade in triple-negative breast cancer. *Nat Cancer* 2022;3:355–70. https://doi.org/10.1038/s43018-022-00339-4
68. Bleu M, Mermet-Meillon F, Apfel V *et al.* PAX8 and MECOM are interaction partners driving ovarian cancer. *Nat Commun* 2021;12:2442. https://doi.org/10.1038/s41467-021-22708-w
69. Ma Y, Kang B, Li S *et al.* CRISPR-mediated MECOM depletion retards tumor growth by reducing cancer stem cell properties in lung squamous cell carcinoma. *Mol Ther* 2022;30:3341–57. https://doi.org/10.1016/j.ymthe.2022.06.011
70. Lou M, Zou L, Zhang L *et al.* MECOM and the PRDM gene family in uterine endometrial cancer: bioinformatics and experimental insights into pathogenesis and therapeutic potentials. *Mol Med* 2024;30:190. https://doi.org/10.1186/s10020-024-00946-0
71. Raney BJ, Barber GP, Benet-Pagès A *et al.* The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res* 2024;52:D1082–D1088. https://doi.org/10.1093/nar/gkad987
72. Wang C, Mayer JA, Mazumdar A *et al.* Estrogen induces c-myc gene expression via an upstream enhancer activated by the estrogen receptor and the AP-1 transcription factor. *Mol Endocrinol* 2011;25:1527–38. https://doi.org/10.1210/me.2011-1037
73. Vogelstein B, Papadopoulos N, Velculescu VE *et al.* Cancer genome landscapes. *Science* 2013;339:1546–58. https://doi.org/10.1126/science.1235122
74. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 2006;93:491–507. https://doi.org/10.1093/biomet/93.3.491
75. Huang D-S, Wang Z, He X-J *et al.* Recurrent TERT promoter mutations identified in a large-scale study of multiple tumour types are associated with increased TERT expression and telomerase activation. *Eur J Cancer* 2015;51:969–76. https://doi.org/10.1016/j.ejca.2015.03.010
76. Dratwa M, Wysoczańska B, Łacina P *et al.* TERT-regulation and roles in cancer formation. *Front Immunol* 2020;11:589929. https://doi.org/10.3389/fimmu.2020.589929
77. Wang F, Yin Y, Ye X *et al.* Molecular insights into the heterogeneity of telomere reprogramming in induced pluripotent stem cells. *Cell Res* 2012;22:757–68. https://doi.org/10.1038/cr.2011.201
78. Günes C, Rudolph KL. The role of telomeres in stem cells and cancer. *Cell* 2013;152:390–3. https://doi.org/10.1016/j.cell.2013.01.010
79. Gronkowska K, Robaszkiewicz A. Genetic dysregulation of EP300 in cancers in light of cancer epigenome control – targeting of p300-proficient and -deficient cancers. *Molecular Therapy: Oncology* 2024;32:200871.
80. Lenoir WF, McKeown MR, Giorgetti G *et al.* Catalytic inhibition of p300 preferentially targets IRF4 oncogenic activity and tumor growth in multiple myeloma. *Cancer Res* 2025;86:1010–34. https://doi.org/10.1158/0008-5472.CAN-25-3440
81. Astuti D, Latif F, Dallol A *et al.* Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am Hum Genet* 2001;69:49–54. https://doi.org/10.1086/321282
82. Tseng P-L, Wu W-H, Hu T-H *et al.* Decreased succinate dehydrogenase B in human hepatocellular carcinoma accelerates tumor malignancy by inducing the Warburg effect. *Sci Rep* 2018;8:3081. https://doi.org/10.1038/s41598-018-21361-6
83. Madry A, Makelov A, Schmidt L *et al.* Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations*. 2018. https://doi.org/10.48550/arXiv.1706.06083
84. Shafahi A, Najibi M, Ghiasi MA *et al.* Adversarial training for free! In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. https://doi.org/10.48550/arXiv.1904.12843
85. Bai T, Luo J, Zhao J *et al.* Recent advances in adversarial training for adversarial robustness. In: *International Joint Conference on Artificial Intelligence*, Vol. 5, 2021. pp. 4312–21. https://doi.org/10.24963/ijcai.2021/591

Received: July 3, 2025. **Revised:** April 7, 2026. **Accepted:** April 26, 2026