UniversalEPI: Harnessing Attention Mechanisms to Decode Chromatin Interactions in Rare and Unexplored Cell Types

Aayush Grover[1,2], Lin Zhang[3], Till Muser[3], Simeon Häfliger[1], Minjia Wang[1], Fabian J. Theis[4,5], Ignacio L. Ibarra[4*], Ekaterina Krymova[3*], Valentina Boeva[1,2,6*]

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland.
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland.
[3]Swiss Data Science Center, EPF Lausanne and ETH Zurich, Zurich, Switzerland.
[4]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany.
[5]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany.
[6]Université Paris Cité, Institut Cochin, INSERM U1016, Paris, France.

*Corresponding author(s). E-mail(s): ignacio.ibarra@helmholtz-muenchen.de; ekaterina.krymova@sdsc.ethz.ch; valentina.boeva@inf.ethz.ch;

Abstract

Chromatin interactions between cis-regulatory elements (CREs) and specific gene promoters offer insights into the mechanisms of gene activation and repression. This knowledge is crucial for evaluating the effects of non-coding variants and chromatin alterations in cell differentiation or disease. Experimental techniques Hi-C and single-cell Hi-C provide information on CRE-promoter interactions, but the high costs and labor requirements limit their widespread use. We introduce UniversalEPI, an attention-based neural network designed to predict CRE-promoter interactions up to 2 Mb apart across unseen cell types without model retraining and using only DNA sequence and chromatin accessibility data as input. In UniversalEPI, chromatin accessibility profiles with masked non-accessible regions and DNA sequence data get passed through several convolution layers to identify binding motifs of three key transcription factors ubiquitously involved in promoter activation and CRE-promoter looping; a transformer block then takes the information from the convolutional levels passed through stochastic gating to predict chromatin interactions. While being a relatively lightweight model because of its architecture, UniversalEPI outperforms the state-of-the-art methods in inferring CRE-promoter interactions in unseen cell types across human cancerous and non-cancerous cell lines (Spearman's Rho > 0.92). The model can also incorporate pseudo-bulk chromatin accessibility data from single-cell ATAC-seq experiments to pinpoint cell subtype-specific interactions. We illustrate the ability of UniversalEPI to track the dynamics of CRE-promoter interactions during macrophage reprogramming and the transition between cancer transcriptional states in esophageal adenocarcinoma. UniversalEPI thus advances in-silico 3D chromatin modeling, supporting studies of chromatin dynamics in cell differentiation and impacts of non-coding variants.

Keywords: Chromatin, Cis-regulatory elements, Enhancer-promoter interactions, deep learning, transformer

1 Introduction

In complex organisms like humans, different cell types have different transcriptional programs leading to varying cellular functions which are important for the organism's survival. These differences are linked to specific chromatin interactions between gene promoters and cis-regulatory elements (CRE), often referred to as enhancers, which regulate gene expression [1, 2]. Disruptions in chromatin interactions are associated with various genetic diseases, including cancer, making their study crucial for understanding disease initiation and progression. Additionally, chromatin interactions provide insights into the functions of non-coding genomic regions which are pivotal in studying cell differentiation during organism development and disease mechanisms. Chromatin interactions can be profiled experimentally via Hi-C, an unbiased but costly high-throughput chromosome conformation capture (3C) technique. Due to the high time and resource demands of Hi-C, interest has increasingly shifted to *in-silico* methods for inferring these interactions.

Several recent studies have proposed deep learning-based architectures to predict chromatin architecture. This includes methods predicting chromatin interaction in a specific cell type for hold-out genomics regions from the DNA sequence alone [3–5]. These methods can be used to assess effects of genomic variants on chromatin structure in a cell type of interest. While these methods generate realistic Hi-C maps, by design, they do not generalize to cell types unseen during the model training. Methods that can predict cell-type-specific chromatin interactions across cell types require additional epigenetic features as inputs such as accessible chromatin regions and histone marks [6–8]. Nevertheless, these methods do not truly allow for a generalization across cell types, as by construction they only capture the information about mechanisms of transcription factors (TFs) expressed in the cell type used for training. Moreover, these methods require diverse input features which are not always readily available. Some other methods focus only on predicting interactions between a given enhancer and a promoter using epigenetic features [9, 10]. However, such methods cannot be used to understand the mechanisms of non-coding mutations as these methods cannot capture the effect of mutations in the insulators, which can affect an enhancer-promoter interaction.

The recently published method C.Origami provides the state-of-the-art solution to predicting cell-type-specific 3D-chromatin interactions using bulk data as input [11]. Specifically, C.Origami, based on convolutional neural networks (CNNs) combined with transformers, accurately predicts chromatin interactions across cell types from DNA sequence, profiles of accessible chromatin, and CTCF binding affinities. While

C.Origami was designed to work with bulk inputs, the ChromaFold method was recently developed to predict chromatin interactions from single-cell chromatin accessibility profiles leveraging co-accessibility across metacells [12]. However, despite the progress made so far, there is still room for significant improvements including increasing model transferability to cell types unseen during training by limiting the model's visual scope to ubiquitously functional genomic elements and creating a lightweight model that can work with both single-cell and bulk chromatin accessibility data.

In this work, we propose UniversalEPI, a lightweight deep learning model consisting of CNN layers and transformer blocks. It overcomes the shortcomings of earlier methods by (*i*) using, for training, DNA binding information of ubiquitously expressed TFs CTCF, YY1, and SP1 that are involved in determining enhancer-promoter architecture across cell types [13–15], (*ii*) using DNA sequence and chromatin accessibility features as inputs to make the model cell-type-specific without heavily depending on large datasets, and (*iii*) training solely on DNA accessible regions to reduce noise and model size while still capturing all functionally relevant interactions in large genomic domains (up to 2Mb). UniversalEPI is evaluated based on its ability to predict cell-type-specific Hi-C interactions between accessible genomic regions in four human cell lines. As expected, UniversalEPI demonstrates superior performance to methods that rely solely on DNA sequence; it performs comparably to or outperforms the state-of-the-art C.Origami method even when the latter uses experimental CTCF affinities in addition to bulk chromatin accessibility for inference. As further validation, we show the applicability of our method in assessing chromatin dynamics during cell reprogramming on the macrophage activation data generated by Reed *et al* [16]. Finally, we benchmark the ability of UniversalEPI to predict Hi-C from pseudobulks of single-cell chromatin accessibility profiles.

2 Results

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

We developed UniversalEPI, a two-stage deep learning model that accurately predicts chromatin interactions across cell types. UniversalEPI consists of two sequentially trained neural networks that use DNA sequences and signals from assays for transposase-accessible chromatin using sequencing (ATAC-seq) as input and chromatin immunoprecipitation followed by sequencing (ChIP-seq) and Hi-C data as target

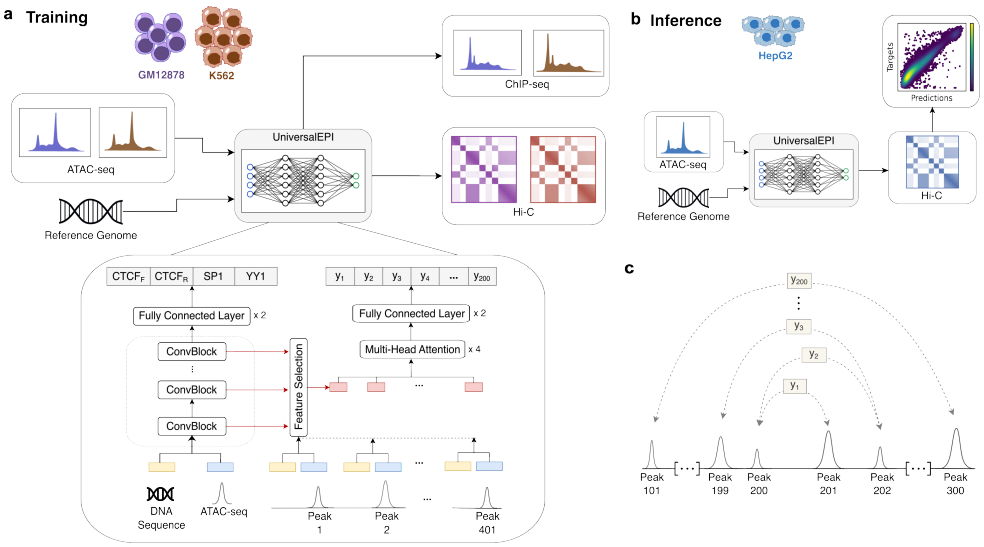

Fig. 1 Outline of the UniversalEPI model architecture. a, UniversalEPI uses one-hot encoded DNA sequence and ATAC-seq p-value tracks as input to predict cell-type-specific transcription factor binding and chromatin interactions. It can be trained on data from multiple cell lines. The first stage of UniversalEPI consists of a convolutional architecture that predicts transcription factor binding affinity from DNA sequence and ATAC-seq. The second stage automatically selects the features from the pre-trained convolutional layers from the first stage using 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. **c,** UniversalEPI predicts pairwise Hi-C interactions (y_i) between the central 200 accessible regions in a zigzag manner.

(**Figure 1a**). For the inference of Hi-C signal in a test cell type, UniversalEPI requires DNA sequence and cell-type-specific ATAC-seq profiles as input (**Figure 1b**).

In the first stage, we trained UniversalEPI to predict the genome-wide binding occupancy of TFs ubiquitously involved in promoter activation and chromatin looping: maximal intensities of SP1, CTCF, and YY1 ChIP-seq [13–15]. This was achieved by training a one-dimensional 5-layer CNN producing a 4-dimensional output on the concatenation of one-hot encoded 1Kb DNA sequences and its corresponding ATAC-seq p-value signals (Methods). Since the orientation of CTCF plays an important role in determining the stability of topologically associated domains (TADs), we split the CTCF ChIP-seq based on orientation (forward or reverse) and predicted the corresponding maximal ChIP-seq intensities as separate targets. 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 in the genome; this design of sequence embeddings was chosen to prevent capturing DNA motifs of other,

potentially cell-type-specific TFs and enable the model to generalize predictions across different cell types at the second stage.

In the second stage, we trained a model to predict Hi-C values between accessible regions based on the DNA sequence embeddings calculated in the first stage (**Figure 1c**). We employed a transformer-based encoder because of its ability to capture long-range interactions and create large-context embeddings of genomic elements (Methods). This allowed us to efficiently learn the inter-dependency between the input regions and consequently, predict the 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 about 4Mb. Each region was 1Kb long, centered around the position of the maximum ATAC-seq signal. Skipping the information about the DNA sequence in between accessible regions allowed for a rather lightweight architecture of UniversalEPI (2.5M parameters) while maintaining a large receptive field and making accurate predictions for interactions between regulatory elements located up to 2Mb away from each other.

To ensure the model observed sufficient variability during training, both stages of UniversalEPI were trained using the GM12878 and K562 cell lines. Approximately 15% of the genome was set aside for validation (chromosomes 5, 12, 13, and 21) and approximately 15% for testing (chromosomes 2, 6, and 19). The remaining chromosomes were used to train both stages of UniversalEPI. The first stage of UniversalEPI was tested on the unseen chromosomes of the A549 and HepG2 cell lines. The first stage of UniversalEPI predicted the transcription factor binding on unseen chromosomes and unseen cell types for CTCF, YY1, and SP1 with average Pearson's correlation of 0.75, 0.65, and 0.55 respectively (**Figure S1**). We further validated that the model learnt biologically relevant information by analyzing the gradients using the DeepLIFT attributions scores [17]. The first stage of UniversalEPI captured the binding motif of all the three TFs on the test set (**Figure S1**). The second stage of UniversalEPI was tested on the IMR90 and HepG2 cell lines. To demonstrate the robustness of our method, we also trained the second stage model on IMR90 and HepG2, and tested it on GM12878 and K562.

2.2 UniversalEPI learns information from biologically-relevant variables

Experimental Hi-C values, quantifying chromatin interactions between each pair of genomic regions, are inversely correlated with the distance between regions and

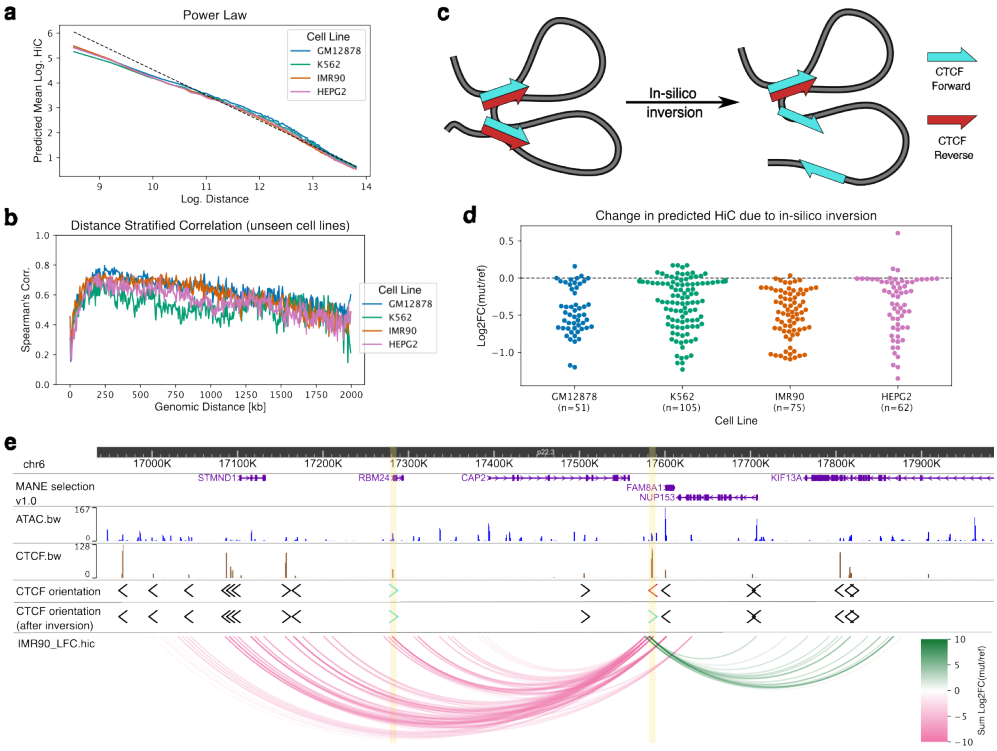

Fig. 2 UniversalEPI learns information from biologically-relevant variables. a, Predicted Hi-C signal by UniversalEPI follows the power law, resulting from polymer physics. The average predicted signal is shown for test chromosomes of test (*i.e.*, unseen during training) cell lines: GM12878, K562, IMR90, 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 displaying UniversalEPI's sensitivity to the input ATAC-seq. **c,** Graphical illustration depicting the effect of in-silico inversion of a CTCF site on genome organization. **d,** Log-fold change between the UniversalEPI predictions before and after the inversion. The experiment is done for all TADs in test chromosomes in unseen cell lines: GM12878, K562, IMR90, and HepG2. **e,** An example of UniversalEPI predictions for a TAD in IMR90 in chromosome 6 (unseen cell line and unseen chromosome). The original TAD is highlighted. The change in predicted Hi-C is calculated using $\log_2$ fold change (logFC) between before and after the in-silico inversion. In the presence of multiple ATAC-seq peaks within the same 5Kb bin, the logFC is summed among all pairs of interactions contained in the two bins.

strongly follow the power law resulting from polymer physics [18]. Similarly, UniversalEPI's predictions followed the power law on the test chromosomes of the unseen cell lines highlighting the model's ability to accurately capture the effect of distance on chromatin structure (**Figure 2a**). Further, to show that UniversalEPI was sensitive to the variability in the ATAC-seq signal, we computed distance-stratified correlation between the predicted and ground-truth Hi-C signal between pairs of accessible regions

(Methods). Using as inputs cell-line-specific ATAC-seq profiles, UniversalEPI obtained equally high distance-stratified correlation for all the four cell lines (**Figure 2b**).

At last, we assessed the sensitivity of the UniversalEPI model to the changes in the DNA sequence through in silico inversions of CTCF binding sites located at the TAD boundaries. TADs influence which cis-regulatory elements interact with gene promoters and are typically bordered by the CTCF protein bound in the convergent orientation [19]. Disruptions in TADs have been linked to diseases like cancer, making the information about chromatin interactions within and across TADs essential for exploring disease mechanisms [20–22]. In-silico inversions of one of the endpoints of a TAD boundary containing a CTCF site in a convergent orientation were expected to reduce the strength of chromatin interactions between the original TAD boundary (**Figure 2c**). Indeed, when we provided the original and modified inputs to UniversalEPI, we predominantly observed negative values of the log-fold change across the four test cell lines (**Figure 2d**), indicating a decrease in the predicted interaction between the original TAD boundaries after the inversion.

We further illustrated the effect of in-silico TAD boundary inversion on the predictions of UniversalEPI with a randomly selected TAD in IMR90 (chr6:17,280,000-17,585,000). UniversalEPI predicted a strong reduction in interactions for the TAD boundary (highlighted) as well as between the regions within the original TAD (**Figure 2e**). Moreover, we observed a significant increase in predicted interactions with a reverse CTCF site located nearly 200Kb away, outside of the original TAD boundary, and a new interaction formed across the original TAD with the promoter of the *NUP153* gene. This result confirmed that UniversalEPI was capable of capturing effects of CTCF-orientation on the stability of TADs.

2.3 Benchmark of the UniversalEPI method on the unseen cell types

Since predicting chromatin interactions across cell types based on cell-type-specific inputs is the main goal of this work, we compared UniversalEPI against the existing methods in terms of their ability to predict chromatin interactions in cell types not used for model training. We also introduced three baselines that captured the effect of distance between interacting elements and the variability between the train and test Hi-C interaction matrices (Methods). We calculated method performances for all interactions within 1Mb regions, which was the maximum common receptive field across all existing methods. We observed that the performance of UniversalEPI was comparable to the original version of the C.Origami method, which used CTCF

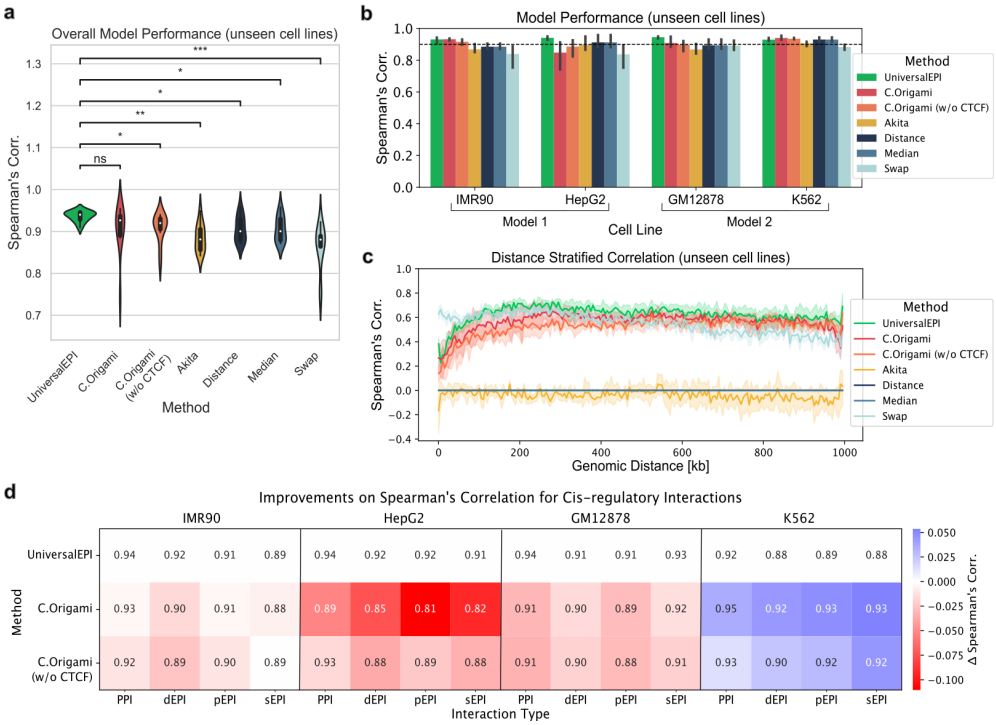

Fig. 3 Benchmark of the UniversalEPI method on the unseen cell types. a, UniversalEPI predicts chromatin interactions in unseen cell types with comparable performance to C.Origami, which additionally uses CTCF ChIP-seq as input. UniversalEPI outperforms C.Origami when the latter uses the same inputs and Akita, which makes predictions solely based on DNA sequence. Each point corresponds to the Spearman's correlation coefficient between predictions and ground-truth Hi-C for an unseen chromosome in an unseen cell line, GM12878, K562, IMR90, and HepG2. A Wilcoxon signed rank test is used to compare the methods; ***: $P < 0.001$, **: $P < 0.01$, *: $P < 0.05$. **b,** UniversalEPI is the only method that achieves Spearman's correlation greater than 0.9 (depicted by dotted line) for all unseen cell types. For each method, Model 1 is trained on GM12878 and K562 cells whereas Model 2 is trained on IMR90 and HepG2 cells. **c,** Distance-stratified correlation shows that UniversalEPI captures the biological information better than other methods. Each line demonstrates the mean distance-stratified Spearman's correlation across four cell lines; the shaded area marks the 95% confidence interval. **d,** UniversalEPI outperforms C.Origami and its variant without CTCF ChIP-seq in predicting different types of cis-regulatory interactions in all but K562 cell line. The change in Spearman's correlation is calculated with respect to UniversalEPI. Positive values suggest that C.Origami or its variant has higher score than UniversalEPI and vice-versa. PPI: Promoter-Promoter Interaction, dEPI: distal Enhancer-Promoter Interactions (further than 2Kb), pEPI: proximal Enhancer-Promoter Interactions (within 2Kb), and sEPI: Super Enhancer-Promoter Interactions.

ChIP-seq as an additional input (**Figure 3a** and **Figure S2**). When trained without CTCF as input, C.Origami showed significantly worse performance than UniversalEPI (**Figure 3a**). UniversalEPI also outperformed sequence-only methods like Akita as these, by design, were not expected to generalize across cell types. When stratified

by cell lines, UniversalEPI consistently obtained a Spearman's correlation between predicted and ground-truth Hi-C values above 90% (**Figure 3**b). On IMR90 and K562, UniversalEPI achieved comparable results to C.Origami, and it substantially outperformed C.Origami on HepG2 and GM12878 cell lines.

Since distance between the interacting elements is a major driving 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 (**Figure 3c**), which assessed the ability of tested methods to capture biological features across several distance ranges. We observed that UniversalEPI captured the biological information better than all the existing methods, including C.Origami. This might be because C.Origami, being an end-to-end model, which, in contrast to UniversalEPI, does not limit the model's visual scope to specific TF motifs, learns cell-type-specific information that is not transferable to all cell types.

Finally, we compared the top three methods – UniversalEPI and C.Origami (with and without CTCF ChIP-seq used as input), on their ability to predict enhancer-promoter interactions (EPIs) and promoter-promoter interactions (PPIs). To study the ability of the models to capture distance-dependent interactions, we split EPIs into two categories – proximal EPI (pEPI, 200bp–2Kb) and distal EPI (dEPI, >2Kb). The annotations for pEPIs, dEPIs, and PPIs were obtained from ENCODE [23]. We also compared the three methods based on their ability to predict interactions between promoters and super-enhancers, corresponding to clusters of active enhancers often regulating cell identity genes [24] and being defined for each cell line in the dbSuper database [25] (sEPIs). UniversalEPI outperformed C.Origami and its counterpart without CTCF binding in the input on all but the K562 cell line (**Figure 3e**), thereby better capturing important cis-regulatory interactions. Overall, we concluded that DNA sequence and ATAC-seq were sufficient to accurately predict cell-type-specific chromatin interactions.

2.4 UniversalEPI captures chromatin dynamics during cell reprogramming

To validate the applicability of UniversalEPI for predicting changes in regulatory interactions upon cell reprogramming, we analyzed the model's predictions upon reprogramming of macrophages, for which ATAC-seq and Hi-C data were collected at eight different time points following the macrophage activation with a mixture of Lipopolysaccharide (LPS) and Interferon gamma (IFN_γ) [16] (**Figure 4a**). In a zero-shot setting, our model trained on GM12878 and K562 cell lines demonstrated

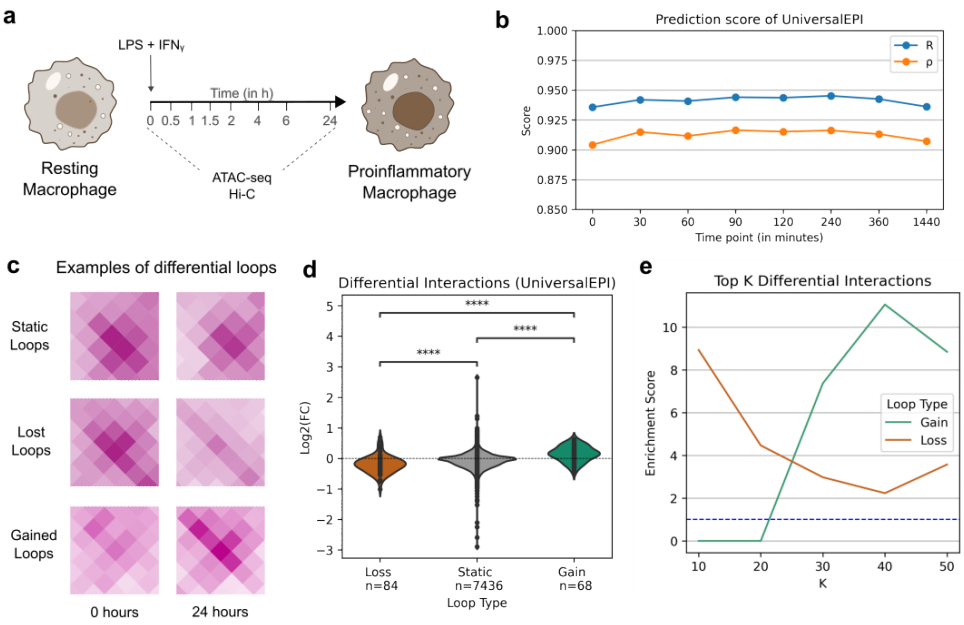

Fig. 4 UniversalEPI captures chromatin dynamics during cell reprogramming. a, Experimental layout defined in Reed *et al.* [16]. Macrophages derived from the THP-1 monocytic cell line are activated using LPS + IFN_γ treatment. ATAC-seq and Hi-C are measured at 8 time points. **b,** Performance of UniversalEPI (trained on GM12878 and K562 cells) using ATAC-seq for each of the 8 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 hours identified by Reed *et al.* **d,** The log2 fold change between Hi-C values at the 24-hour time point over the 0th time point measured by UniversalEPI predictions for the differential loops. The significance is measured using a Mann-Whitney-Wilcoxon test with Bonferroni correction. ****: $P \leq 10^{-4}$. **e,** Gain and loss enrichment score over random for selecting a gained or lost loop respectively in the K most differential interactions as predicted by UniversalEPI. The blue dotted line represents the enrichment score of labelling a random interaction as gained or lost.

a high degree of prediction accuracy, with Pearson's and Spearman's correlation with ground-truth Hi-C values exceeding 90% (**Figure 4b**).

Reed *et al* identified specific chromatin loops that exhibited differential interactions upon macrophage activation—lost, gained, or ones remaining static (**Figure 4c**). For these selected chromatin loops, UniversalEPI effectively identified treatment-induced changes in chromatin architecture (**Figure 4d**). Moreover, UniversalEPI's predictive power allowed it to distinguish gained and lost loops from static loops with a performance level that is at least twice as accurate as random classification (**Figure 4e**).

2.5 UniversalEPI unravels chromatin structure in complex tissues using pseudo-bulk single-cell ATAC-seq data

To show the applicability of our computational method to single cell studies, we used UniversalEPI to predict chromatin structure of cell types using as inputs pseudo-bulks of single cell ATAC-seq (scATAC-seq) data (**Figure 5**a). We compared UniversalEPI with ChromaFold [12], the only other method that predicts bulk Hi-C from scATAC-seq and DNA sequence as inputs. scATAC-seq and Hi-C data from GM12878, K562, HepG2, and IMR90 cell lines were used in this analysis. Healthy and cancer cell lines GM12878 and HepG2 cells were employed for training both the models. Keeping the study design similar to the one proposed in Gao *et al* [12], we used Hi-C at 10Kb resolution with HiC-DC+ z-scores as targets [26].

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 outperformed ChromaFold on both unseen cell lines (**Figure 5b**). UniversalEPI also obtained a higher distance-stratified correlation, especially for short to medium ranged interactions (50-600Kb) (**Figure 5c**), highlighting the ability of our model to maintain its generalization capacity while extending to pseudo-bulk scATAC-seq data.

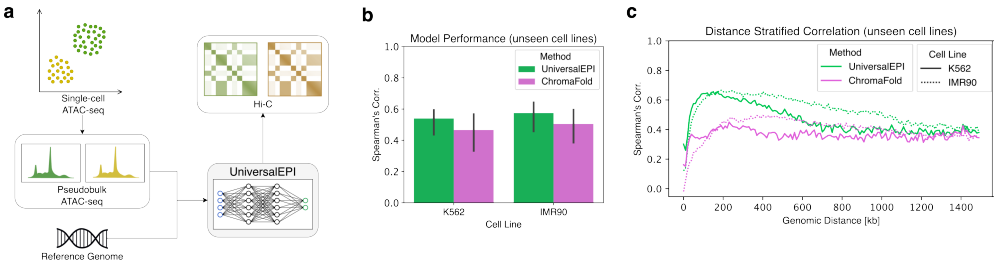

Fig. 5 UniversalEPI unravels chromatin structure in complex tissues using pseudo-bulk single-cell ATAC-seq data. a, UniversalEPI can predict cell-type-specific Hi-C from different cell types identified using scATAC-seq data. **b,** Comparison between UniversalEPI and ChromaFold on their ability to predict bulk Hi-C from scATAC-seq. Spearman's correlation is calculated on unseen chromosomes (chr5, chr18, chr20, chr21) of unseen cell lines. **c,** Distance-stratified Spearman's correlation based on the unseen chromosomes and cell lines.

3 Discussion

In this work, we presented UniversalEPI, a novel attention-based model that accurately predicts chromatin interactions across unseen cell types. One major drawback of current approaches, which require several data modalities as inputs, was solved

by UniversalEPI, which solely uses DNA sequence and chromatin accessibility data. Furthermore, since our model learns only from accessible regions of the genome, it demonstrates notable scalability (**Figure 1**). In contrast to other techniques, such as C.Origami, which was originally trained on a single cell line [11], this enables our model to be trained on several cell lines concurrently in a fair amount of time.

By using CNN layers and transformer blocks, UniversalEPI effectively captures the essential transcription factor binding motifs and spatial dependencies underlying chromatin interactions. The model's high prediction accuracy demonstrated by the strong distance-stratified correlation between the predicted and ground-truth Hi-C values confirms its sensitivity to biologically relevant information, which is critical for accurate modeling of chromatin architecture (**Figure 2**). This enables UniversalEPI's predictions to consistently achieve high correlations with experimental Hi-C and even outperform existing models, including C.Origami, for unseen cell types (**Figure 3**).

By applying UniversalEPI to predict treatment-induced chromatin changes in macrophages stimulated with LPS and INF_{γ}, we were able to further validate the model performance in a zero-shot setting (**Figure 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 when predicting bulk Hi-C using pseudo-bulk single-cell ATAC-seq data (**Figure 5**). This feature increases the method's utility for providing insights into the heterogeneity in chromatin organization observed in complex tissues or across cell states.

Despite showing robust performance across cell types, UniversalEPI was only trained on a limited number of cell lines. Therefore, additional training may be required for its use in rare and highly specialized cell types with unique regulatory mechanism. However, this issue can be mitigated in the near future with the increasing production of ChIP-seq and Hi-C data in new cell types. Additionally, the first stage of UniversalEPI could further improve by applying techniques like adversarial training [27–29] further preventing the model from capturing cell-type-specific information and consequently, enhancing its ability to generalize across cell types.

In conclusion, UniversalEPI achieves state-of-the-art performance in predicting chromatin interactions in unseen cell types from only DNA sequence and chromatin accessibility profiles. This brings a major advancement in chromatin interaction predictions that can be widely applied to study complex regulatory landscapes of the genome. Its predictive power would allow UniversalEPI to be used for studying the changes

in chromatin organization in the presence of non-coding mutations and structural variations and thereby, deciphering the regulatory mechanisms of genetic diseases.

Methods

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 [30, 31] under the accession codes listed in Supplementary Table 1. We then extracted the 5Kb 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 (approximately 759M interactions as compared to approximately 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 5Kb 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 $\mathrm{m}_d^{\mathrm{GM12878}}$ and $\mathrm{mad}_d^{\mathrm{GM12878}}$ respectively. For a new cell type, e.g., the K562 cell line, the normalized Hi-C score ($\bar{y}$) between regulatory elements in bins i and j is then calculated as follows:

$$\bar{y}_{i,j}^{\mathrm{K562}} = \left(\frac{y_{i,j}^{\mathrm{K562}} - \mathrm{m}_{|j-i|}^{\mathrm{K562}}}{\mathrm{mad}_{|j-i|}^{\mathrm{K562}}} \right) \mathrm{mad}_{|j-i|}^{\mathrm{GM12878}} + \mathrm{m}_{|j-i|}^{\mathrm{GM12878}}$$

where y is the ICE-normalized Hi-C value. Figure S3 depicts the effect of this normalization on all cell lines.

Fundenberg *et al* [4] 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 increases with distance (**Figure S7**). Particularly, we used four kernels with variances of 0.5, 0.65, 0.8, and 1 for interactions between regions that are closer than 25Kb, between 25Kb and 100Kb, between 100Kb and 250Kb, and greater than 250Kb respectively. Finally, we applied logarithmic transformation to the interaction scores. Several methods like Akita [4] and ChromaFold [12] also remove the effect of distance from the model by applying z-scores. 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.

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 1. For multiple ATAC-seq peaks within the 500bp genomic region, we retain the peak with the highest read count and remove the remaining peaks. This resulted in approximately 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. 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 bigwigs of each cell line using deepTools multiBigwigSummary v3.5.3 [35]. Finally, we applied the TMM normalization of EdgeR [36] 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 Figure S4.

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

1. 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 [37] 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 [38] 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 the CTCF peaks were assigned to a motif on the forward strand and 48% to a motif on the reverse strand. 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

Herein we focus on modelling accessible regions (ATAC-seq peaks) only; the information about genomic regions in between accessible regions is provided to the model via the distance between peaks. We therefore predict 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 $\mathrm{F} : \mathbb{X} \mapsto \mathbb{Y}$, 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 $Y = \{y^{(i,j)} \in \mathbb{R},\, i \leq j,\, i, j = 101, \ldots, 300\} \in \mathbb{Y}$ with elements representing the interaction between the pair of peak locations i and j in the central 200 ATAC-seq. 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 theirs 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 modelling the interactions between the learned features at peak locations. A schematic overview of the proposed method is depicted in Figure S5. 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 $y_{\mathrm{TF}} \in \mathbb{R}^{N_{\mathrm{TF}}}$ 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 [3], 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) [39] 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 over-fitting, a 20% dropout is applied during training. The model is optimized by minimizing the empirical version of the mean squared error (MSE) defined as

$$\min_\theta \mathbb{E}_{(x,y_{\mathrm{TF}})\sim P_{\mathrm{TF}}}[||f_\theta(x) - y_{\mathrm{TF}}||^2], \tag{1}$$

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. The outputs of convolutional blocks, after applying the non-linear activations, serve as intermediate representations and are used the input for the second stage. We use a linear projection head for each feature layer to map it to the space of the 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.* [40] 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

$$\min_{\phi,\mu} \mathbb{E}_{(X,Y)\sim P_{\text{HiC}}} \mathbb{E}_Z \left[L_T \left(g_\phi(s_\mu(h(X), Z)), Y \right) + \lambda L_R(Z) \right], \tag{2}$$

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 $z_k = \max(0, \min(1, \mu_k + \epsilon_k))$, where μ_k is learned and ϵ_k is sampled from $\mathcal{N}(0, \sigma^2)$ with a predefined σ during training. Following Yamada *et al.* [40], we set $\sigma = 0.5$. Herein we propose to use the stochastic gating $s_\mu(h(X), Z) = \{s_\mu(h(x_1), Z), s_\mu(h(x_2), Z), \ldots, s_\mu(h(x_{401}), Z)\}$ at the feature level instead of the node level (the latter was initially proposed by Yamada *et al.* [40]) 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: $s_\mu(h(x_i), Z) = \sum_{k=1}^{N_L} h_k(x_i) \cdot z_k$. The sparsity regularization is the sum of the probabilities that the gates are active, which is equal to $\mathbb{E}_Z L_R(Z) = \sum_{k=1}^{N_L} P(z_k > 0) = \sum_{k=1}^{N_L} \Phi\left(\frac{\mu_k}{\sigma}\right)$, where Φ is the Gaussian cumulative distribution function. During the inference stage, we set $\hat{z}_k = \max(0, \min(1, \mu_k))$.

- *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* [41]. 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, herein 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 [41]. 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. 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

$E = \{e_1, e_2, ..., e_{401}\}$ with the i-th token representation $e_i \in \mathbb{R}^{d_{\text{model}}}$, *i.e.* $E = g_\phi^{\text{ENC}}(s_\mu(h(X), Z))$. 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.*, $\hat{y}^{(i,j)} = g_\phi^{\text{DEC}}(e_i, e_j)$. 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 $\hat{Y}$ in (2) is defined as

$$L_T(\hat{Y}, Y) = \sum_{i,j} (\hat{y}^{(i,j)} - y^{(i,j)})^2. \tag{3}$$

where i, j are encoded following the DeepC approach [3] in a vertical, zig-zag pole over the center of the sequence window (**Figure 1c**). During training the empirical version of the loss (2) is optimized in both μ and ϕ.

- *Auxiliary information* can be optionally used for Hi-C prediction. Since the Hi-C value between two bins are also dependent on the mappability of the independent bins [42], we also include the 36bp-mappability track, obtained from the UCSC Genome Browser [43, 44], 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, and resulting in a new input to the attention layers. This method allows the use of any auxiliary information, such as mappability and ATAC-seq.

Training details

Our method is implemented with the PyTorch framework [45]. 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 [46] 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 1Kb segment of DNA sequence around the peak center and the corresponding ATAC-seq as input. Transcription factors CTCF, YY1, and SP1 were empirically chosen as target TFs based on their ability to assist in chromatin organization prediction (**Figure S7**). ZNF143 is another TF that might ubiquitously play a role in chromatin looping [47].

However, a recent study [48] found the commonly used antibody of ZNF143 cross-reacting with CTCF, thereby making most of its existing ChIP-seq profiles unusable. Therefore, ZNF143 is excluded in this work. The first stage model was trained on the GM12878 and K562 cell lines. We randomly split the chromosomes into training, validation and test set. 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 [49] optimizer with a learning rate of 10^{-3}, and a batch size of 32. 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 $\sqrt{d_{model}}$. 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 heldout 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 5Kb resolution also contains arbitrary bins with very few reads mapped to them (**Figure S8**). 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 1Kb 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 is the input of UniversalEPI. We trained the network for 20 epochs and choose the model for evaluation with the highest Spearman's correlation on the validation data.

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[1] and blacklisted[2] regions as reported by Kelley *et al* [50]. 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.

[1] https://storage.googleapis.com/basenji_barnyard2/umap_k36_t10_l32_hg38.bed
[2] https://storage.googleapis.com/basenji_barnyard2/hg38.blacklist.rep.bed

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 2Mb with a step size of 5Kb.

State-of-the-art methods

In this section, we describe the configurations of the state-of-the-art methods C.Origami and Akita, against which we compare our approach. Various existing methods follow different Hi-C preprocessing techniques, predict Hi-C upto different distances and at different resolutions. For example, Akita [4] predicts Hi-C at 2Kb resolution upto 1Mb after removal of effect of distance whereas C.Origami [11] retains the effect of distances and predicts Hi-C at 10Kb resolution capturing interactions upto 2Mb away. Moreover, Akita employs 2D Gaussian smoothing which is not used by C.Origami.

To mitigate these differences, we re-trained the existing methods on our preprocessing of 5Kb Hi-C data and tuned their hyperparameters based on their performance on validation chromosomes of seen cell lines. We used the open-sourced codes as a starting point. For both methods, we did not modify the loss functions, optimizers, or learning rate schedulers. In case of C.Origami, we reduced the input size from 2Mb to 1Mb to compensate from increase in Hi-C resolution from 10Kb to 5Kb. This majorly led to changes in the hidden layer sizes and number of convolution layers in the encoder. To obtain the C.Origami model without the CTCF ChIP-seq as input, we simply excluded this track from the inputs without changing the configuration of the model. Finally, both versions of C.Origami models were trained for a maximum of 80 epochs. The top performing model was selected based on the validation loss and used for inference.

On the other hand, for Akita, we kept the input size of 1Mb intact. Since the Hi-C resolution is reduced from 2Kb to 5Kb and consequently, the target length is reduced from 512 to 201, we reduced the complexity of the convolution trunk. Specifically, we replaced the 11 convolution layers with 7 convolution layers. Moreover, the pooling size was increased from 2 to 5 for the final 4 convolution layers. This configuration led to the best performing Akita model. The Akita model was trained with two heads, corresponding to the two training cell lines. The model was trained for a maximum of 10000 epochs. The training was stopped early if the validation loss did not improve for 12 epochs. This resulted in training stopping after 35 epochs on

average for the two models (one trained on GM12878 and K562 cells and the other on HepG2 and IMR90 cells). Adjusting these hyperparameters enabled each state-of-the-art method to achieve performance levels comparable to those reported in their respective publications.

Baseline methods

We define 3 baseline methods, (1) Distance, (2) Median, and (3) Swap. These are computationally very simple as compared to the complex deep learning methods.

As the Hi-C interactions between two genomic segments heavily depend on the distance between the segments, we introduce the Distance baseline which only depends on the genomic distance between the open regions. In particular, we define the prediction between regions i and j as follows:

$$\hat{y}_{i,j} = -\log(d_{i,j})$$

where $d_{i,j}$ is the genomic distance (in 5Kb bins) between regions i and j.

Similarly, we define the Median baseline where prediction is the median of all the interactions between the open regions in the training data that are at a particular distance away. Mathematically, the prediction for interaction between regions i and j, that are $d_{i,j}$ bins away, is given by:

$$\hat{y}_{i,j} = \text{Median}([y_1, y_2, \ldots, y_n])$$

where $y_1 \ldots y_n$ are Hi-C interactions from the training data between all pairwise open regions that are $d_{i,j}$ bins apart.

Finally, we define the third baseline called Swap to capture the similarity in Hi-C across cell types. Here, we take the mean of Hi-C interactions from all the cell lines in the training data and use this as the predicted Hi-C matrix. If GM12878 and K562 are taken as training cell lines, then the predicted Hi-C interaction for the new unseen cell line would be given by:

$$\hat{y}_{i,j} = \frac{1}{2}(y_{i,j}^{\text{GM12878}} + y_{i,j}^{\text{K562}})$$

TAD annotations

TADs were identified using boundary calls from the 4DN data portal [30, 31]. We defined a consecutive pair of boundary calls (i, j where $i < j$) as a TAD if there was

an overlapping ATAC-seq peak and a CTCF ChIP-seq peak within 10Kb of boundary calls i and j. Additionally, we required that the CTCF orientation is forward for i and reverse for j. The CTCF orientation was determined following the same approach used to prepare target data for the first stage of UniversalEPI. Applying these criteria yielded approximately 200 TADs per cell line.

In-silico inversion of TAD boundary

For each identified TAD, we modified the ATAC-seq peak at boundary j by reversing the ATAC-seq signal, mappability profile, and by taking the reverse complement of the DNA sequence, resulting in a modified peak. We then extracted features from this modified peak using the first stage of UniversalEPI. These features, along with those from the 400 neighboring peaks, were used as input to the second stage of UniversalEPI.

Since interactions between two 5Kb bins can be influenced by multiple accessible chromatin regions within the two bins, it is challenging to isolate the effect of inverting a TAD endpoint in such cases. To address this, we focused our analysis on TADs with a unique ATAC-seq peak within 5Kb of each endpoint.

Macrophage activation data processing

We validated UniversalEPI on the data collected for activated macrophages derived from the THP-1 monocytic cell line. Since we require the ATAC-seq peak calls and the signal p-value bigwig track, we used the raw ATAC-seq paired-end fastq files provided by Reed *et al* [16] and performed preprocessing similar to the ones done by the authors. Adaptors and low-quality reads were first trimmed using Trim Galore! v0.6.10. Reads were then aligned using BWA mem. Samtools v1.13 [51] was then used to sort the aligned reads. Duplicated reads were removed using PicardTools whereas mitochondrial reads were filtered using Samtools. The replicates for each timestamp were merged using Samtools. Finally, ENCODE's ATAC-seq pipeline [52] was used to obtain the peaks and signal p-value bigwig track. Deduplication was done as described before for the generated peak files. This resulted in approximately 200K peaks for each time point. For the generated bigwig tracks, edgeR's TMM normalization [36] was done with the GM12878 cell line as the reference (**Figure S5**).

The .hic files, for each of the 8 different time stamps, were directly obtained from Reed *et al* [16]. Nearly 370M interactions were observed on average for each of the time points. ICE normalization [32] was applied to each .hic file followed by the z-score normalization using the GM12878 cell line as the reference (as described above).

Differential loop prediction

For the validation of the ability of UniversalEPI to identify differential loops, we used the selected loops provided by Reed *et al.* The *gain.early* and *gain.late* were merged to form the set of gained loops. A similar approach was used for lost loops. To ensure a reliable set of differential loops, the sets of gained, static, and lost loops were further filtered based on fold change (FC) of ICE-normalized Hi-C between the 24 hours and 0 hour time points. Specifically, lost loops with $FC < 0.67$, gained loops with $FC > 1.5$, and static loops with $FC \in [0.67, 1.5]$ were retained.

The ATAC-seq peak sets at 0th and 24th hour time points were merged to get a set of regions for which Hi-C signal was predicted for the two time-points using UniversalEPI. If multiple peaks existed within 500bp of each other, the peak with the maximum overall enrichment was retained resulting in a total of 235,690 peaks. UniversalEPI predictions were then independently calculated on this merged peak set using ATAC-seq bigwig signals for the two time points. Finally the log-fold change was calculated between the predictions at these two time points.

Comparison against ChromaFold

To compare UniversalEPI's performance against ChromaFold [12], we trained and evaluated both the models using the same input single-cell ATAC-seq (scATAC-seq) data. Since UniversalEPI is designed to work with bulk ATAC-seq signal and ATAC-seq peaks, we first bulkified the scATAC-seq data. The bam files were directly downloaded from ENCODE (Supplementary Table 1) and converted to signal and peak files using the techniques explained in sections above. These bulkified ATAC-seq signals and peaks were used to train the UniversalEPI model and the evaluation of both the models.

Both the models were trained using data from GM12878 and HepG2 cell lines and evaluated on K562 and IMR90 cell lines. We followed the same processing as introduced in Gao *et al* [12] *i.e.* we used ICE-normalized Hi-C at 10Kb resolution. Z-scores were then calculated using HiC-DC+ [26] and the resulting values were clipped between -16 and 16. Chromosomes 5, 18, 20, and 21 were used for testing whereas chromosomes 3 and 15 for validation, and the remaining chromosomes for training, as done in Gao *et al* [12]. This allowed us to directly use the preprocessing, model, and training scripts from the source code of ChromaFold.

The models were compared using interactions between accessible chromatin regions. The interactions were also filtered based on unmappable and blacklisted

regions. UniversalEPI and ChromaFold were also evaluated based on distance-stratified correlation, evaluating the performance of these methods in predicting interactions between genomic regions located upto 1.5Mb away from each other.

Acknowledgments. This project is partially funded by the Swiss Data Science Center (SDSC) collaborative projects grant (C22-09); Swiss Government Excellence Scholarship (ESKAS-Nr: 2021.0468) to A.G.

Declarations

Competing interests

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.

References

[1] Uyehara, C.M., Apostolou, E.: 3D enhancer-promoter interactions and multi-connected hubs: Organizational principles and functional roles. Cell Reports **42**(4) (2023)

[2] Pollex, T., Rabinowitz, A., Gambetta, M.C., Marco-Ferreres, R., Viales, R.R., Jankowski, A., Schaub, C., Furlong, E.E.: Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. Nature Genetics **56**(4), 686–696 (2024)

[3] Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., Hughes, J.R.: DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nature Methods **17**(11), 1118–1124 (2020)

[4] Fudenberg, G., Kelley, D.R., Pollard, K.S.: Predicting 3D genome folding from DNA sequence with Akita. Nature Methods **17**(11), 1111–1117 (2020)

[5] Zhou, J.: Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. Nature Genetics **54**(5), 725–734 (2022)

[6] Li, W., Wong, W.H., Jiang, R.: DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Research **47**(10), 60–60 (2019)

[7] Zhang, S., Chasman, D., Knaack, S., Roy, S.: In silico prediction of high-resolution Hi-C interaction matrices. Nature Communications **10**(1), 5449 (2019)

[8] Yang, R., Das, A., Gao, V.R., Karbalayghareh, A., Noble, W.S., Bilmes, J.A., Leslie, C.S.: Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. Genome Biology **24**(1), 134 (2023)

[9] Whalen, S., Truty, R.M., Pollard, K.S.: Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics **48**(5), 488–496 (2016)

[10] Chen, K., Zhao, H., Yang, Y.: Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. Briefings in Bioinformatics **23**(2), 577 (2022)

[11] Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., Thandapani, P., Skok, J., Aifantis, I., Fenyö, D., *et al.*: Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. Nature Biotechnology **41**(8), 1140–1150 (2023)

[12] Gao, V.R., Yang, R., Das, A., Luo, R., Luo, H., McNally, D.R., Karagiannidis, I., Rivas, M.A., Wang, Z.-M., Barisic, D., *et al.*: ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. Nature Communications **15**(1), 9432 (2024)

[13] Ong, C.-T., Corces, V.G.: CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics **15**(4), 234–246 (2014)

[14] Makhlouf, M., Ouimette, J.-F., Oldfield, A., Navarro, P., Neuillet, D., Rougeulle, C.: A prominent and conserved role for YY1 in Xist transcriptional activation. Nature Communications **5**(1), 4878 (2014)

[15] O'Connor, L., Gilmour, J., Bonifer, C.: Focus: Epigenetics: The role of the ubiquitously expressed transcription factor Sp1 in tissue-specific transcriptional regulation and in disease. The Yale journal of biology and medicine **89**(4), 513 (2016)

[16] Reed, K.S., Davis, E.S., Bond, M.L., Cabrera, A., Thulson, E., Quiroga, I.Y., Cassel, S., Woolery, K.T., Hilton, I., Won, H., et al.: Temporal analysis suggests

a reciprocal relationship between 3D chromatin structure and transcription. Cell reports **41**(5) (2022)

[17] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153 (2017). PMLR

[18] Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., *et al.*: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**(5950), 289–293 (2009)

[19] Wit, E., Vos, E.S., Holwerda, S.J., Valdes-Quezada, C., Verstegen, M.J., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H., Laat, W.: CTCF binding polarity determines chromatin looping. Molecular Cell **60**(4), 676–684 (2015)

[20] Valton, A.-L., Dekker, J.: TAD disruption as oncogenic driver. Current opinion in genetics & development **36**, 34–40 (2016)

[21] Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., *et al.*: Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science **351**(6280), 1454–1458 (2016)

[22] Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., Chen, X., Hu, H., Bakogianni, S., Wang, J., *et al.*: Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. Nature Genetics **52**(4), 388–400 (2020)

[23] Moore, J., Purcaro, M., Pratt, H., Epstein, C., Shoresh, N., Adrian, J., Kawli, T., Davis, C., Dobin, A., Kaul, R., et al.: & Weng, Z.(2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature **583**(7818), 699–710

[24] Tang, F., Yang, Z., Tan, Y., Li, Y.: Super-enhancer function and its application in cancer targeted therapy. NPJ precision oncology **4**(1), 2 (2020)

[25] Khan, A., Zhang, X.: dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Research **44**(D1), 164–171 (2016)

[26] Sahin, M., Wong, W., Zhan, Y., Van Deynze, K., Koche, R., Leslie, C.S.: HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. Nature Communications **12**(1), 3366 (2021)

[27] Madry, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)

[28] Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! Advances in Neural Information Processing Systems **32** (2019)

[29] Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. arXiv preprint arXiv:2102.01356 (2021)

[30] Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'shea, C.C., Park, P.J., Ren, B., *et al.*: The 4D nucleome project. Nature **549**(7671), 219–226 (2017)

[31] Reiff, S.B., Schroeder, A.J., Kırlı, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A.D., Balashov, A.K., Vitzthum, C., *et al.*: The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. Nature Communications **13**(1), 2365 (2022)

[32] Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., Mirny, L.A.: Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods **9**(10), 999–1003 (2012)

[33] Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R., Grüning, B.A.: 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 Research **48**(W1), 177–184 (2020)

[34] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods **17**(3), 261–272 (2020)

[35] Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T.: deepTools2: a next generation web server for

deep-sequencing data analysis. Nucleic Acids Research **44**(Web Server issue), 160 (2016)

[36] Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology **11**, 1–9 (2010)

[37] Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J.A., Ferenc, K., Kumar, V., Lemma, R.B., Lucas, J., Chèneby, J., Baranasic, D., *et al.*: JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Research **52**(D1), 174–182 (2024)

[38] Swindell, W.R., Johnston, A., Sun, L., Xing, X., Fisher, G.J., Bulyk, M.L., Elder, J.T., Gudjonsson, J.E.: Meta-profiles of gene expression during aging: limited similarities between mouse and human and an unexpectedly decreased inflammatory signature. PloS One **7**(3), 33204 (2012)

[39] Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)

[40] Yamada, Y., Lindenbaum, O., Negahban, S., Kluger, Y.: Feature selection using stochastic gates. In: International Conference on Machine Learning, pp. 10648–10659 (2020). PMLR

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[42] Carty, M., Zamparo, L., Sahin, M., González, A., Pelossof, R., Elemento, O., Leslie, C.S.: An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nature Communications **8**(1), 15454 (2017)

[43] Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., *et al.*: The UCSC genome browser database: 2023 update. Nucleic Acids Research **51**(D1), 1188–1195 (2023)

[44] Karimzadeh, M., Ernst, C., Kundaje, A., Hoffman, M.M.: Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Research **46**(20),

120–120 (2018)

[45] Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24) (2024). https://doi.org/10.1145/3620665.3640366 . ACM. https://pytorch.org/assets/pytorch2-2.pdf

[46] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[47] Zhou, Q., Yu, M., Tirado-Magallanes, R., Li, B., Kong, L., Guo, M., Tan, Z.H., Lee, S., Chai, L., Numata, A., *et al.*: ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function. Nature Communications **12**(1), 43 (2021)

[48] Magnitov, M.D., Maresca, M., Alonso Saiz, N., Teunissen, H., Braccioli, L., Wit, E.: ZNF143 is a transcriptional regulator of nuclear-encoded mitochondrial genes that acts independently of looping and CTCF. bioRxiv, 2024–03 (2024)

[49] Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[50] Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., Snoek, J.: Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research **28**(5), 739–750 (2018)

[51] Li, H.: The sequence alignment/map (SAM) format and SAMtools 1000 genome project data processing subgroup. Bioinformatics **25**, 1 (2009)

[52] Lee, J., Christoforo, G., Foo, C., Probert, C., Kundaje, A., Boley, N., Kim, D., Dacre, M., et al.: kundajelab/atac_dnase_pipelines: 0.3. 3. Zenodo (2016)