

## Supplementary Sections

### S1. TAD Annotations

TADs were identified using boundary calls from the 4DN data portal(1, 2). 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 10 kb 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.

### S2. 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 and 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 5 kb 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 5 kb of each endpoint.

### S3. Maximum Confidence Fold Change

The maximum-confidence log2 fold-change between the predictions for two conditions with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$  respectively, a parameter  $d$  was defined as  $d = (\mu_2 - \mu_1) - 1.282 \sqrt{\sigma_1^2 + \sigma_2^2}$  for  $\mu_2 > \mu_1$  and  $d = (\mu_2 - \mu_1) + 1.282 \sqrt{\sigma_1^2 + \sigma_2^2}$  otherwise. Here,  $d$  quantifies whether the observed difference exceeds the expected variability, with 1.282 representing the z-score for a one-tailed 90% confidence level. The maximum-confidence log2 fold-change between the two conditions was set to 0 if  $d < 0$  and  $\mu_2 > \mu_1$  or  $d > 0$  and  $\mu_2 \leq \mu_1$ , indicating the change was not statistically significant. In case of no overlap, the maximum-confidence log2 fold-change was given by  $d$ . This corresponds to log2 fold-change being at least  $d$  with 90% model confidence.

#### **S4. State-of-the-art Methods**

In this section, we describe the configurations of the state-of-the-art methods C.Origami(3), Akita(4) and EPCOT(5), against which we compare our approach. We did not compare against DeepC(6), as it has approximately 10 times more parameters than Akita and the two models are shown to perform comparably(4).

Various existing methods follow different Hi-C preprocessing techniques, predict Hi-C up to different distances and at different resolutions. For example, Akita predicts observed over expected Hi-C at 2048bp resolution up to 1 Mb, EPCOT also predicts observed over expected Hi-C at 5 kb resolution up to 1 Mb, whereas C.Origami retains the effect of distances and predicts Hi-C at 10 kb resolution capturing interactions up to 2 Mb away. Moreover, Akita employs 2D Gaussian smoothing which is not used by C.Origami.

To mitigate these differences, we re-trained C.Origami on our preprocessing of 5 kb Hi-C data and tuned its hyperparameters based on its performance on validation chromosomes of seen cell lines. We used the open-sourced code as a starting point. We did not modify the loss functions, optimizers, or learning rate schedulers. In case of C.Origami, we reduced the input size from 2 Mb to 1 Mb to compensate from increase in Hi-C resolution from 10 kb to 5 kb. 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.

We trained two Akita models: (1) trained on GM12878 and K562 cells and evaluated on IMR90 and HepG2 cell lines, and (2) trained on IMR90 and HepG2 cells and evaluated on GM12878 and K562 cell lines. The average prediction across both training cell lines is used for all unseen cell lines. As Akita predicts the Hi-C contact map at 2048bp resolution, we apply adaptive average pooling to rescale the contact maps at 5 kb resolution, allowing us to make a fair comparison between UniversalEPI and Akita.

Finally, we used the pre-trained EPCOT models, one trained using the data from GM12878 cell line and the other using HFF cell line. Both models were trained using ATAC-seq and hg38 DNA sequence. Like Akita, EPCOT also predicts observed over expected Hi-C using a 1-Mb DNA sequence and chromatin accessibility as input. Since the sequence encoder of EPCOT is trained using data from GM12878, K562, and HepG2 cell lines, we compared the model predictions on the IMR90 cell line. Using IMR90's ATAC-

seq as input, the average prediction of the two pre-trained EPCOT models was considered as the final prediction.

For the task of 1 kb Micro-C prediction, we used the pre-trained EPCOT model trained on the HFF cell line. A 600 kb DNA sequence was used as input, and the interactions up to 200 kb were compared between the UniversalEPI and EPCOT models.

Both Akita and EPCOT, predict observed-over-expected contact maps. To convert them to observed maps, we multiply the predictions with expected values, which are calculated using the training data.

## S5. 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  $j$  in equation (1),

$$\hat{y}_{i,j} = -\log(d_{i,j}) \quad (1)$$

where  $d_{i,j}$  is the genomic distance (in 5 kb 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 equation (2),

$$\hat{y}_{i,j} = \text{Median}([y_1, y_2, \dots, y_n]) \quad (2)$$

where  $y_1 \dots y_n$  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 equation (3).

$$\hat{y}_{i,j} = \frac{1}{2} (y_{i,j}^{GM12878} + y_{i,j}^{K562}) \quad (3)$$

## S6. 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.*(7) 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(8) 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(9) 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(10) was done with the GM12878 cell line as the reference (Supplementary Fig. S3).

The .hic files, for each of the 8 different time stamps, were directly obtained from Reed *et al.*(7). Nearly 370M interactions were observed on average for each of the time points. ICE normalization(11) was applied to each .hic file followed by the z-score normalization using the GM12878 cell line as the reference (as described in Main).

## S7. 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.* (7). 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 maximum-confidence log2 fold-change was calculated between the predictions at these two time points.

Each of the differential loops was then classified based on the maximum-confidence log2 fold-change ( $mcLFC$ ) predicted by UniversalEPI. Specifically, if  $mcLFC = 0$ , then the loop is labelled as static, if  $mcLFC < 0$ , then the loop is labelled as lost, and if  $mcLFC > 0$ , then the loop is labelled as gained.

## S8. Comparison with ChINN

To benchmark our model, we trained the ChINN framework using the official codebase(12). Specifically, we trained the distance-matched extended classifier variant on ATAC-seq peaks and 5-kb-resolution Hi-C data from the GM12878 cell line. Positive examples were defined as Hi-C chromatin interactions (5-kb bin resolution); negative examples were sampled in a distance-matched manner (positive-to-negative ratio  $\approx 1:5$ ) and an extended negative set was used to train the final gradient-boosted tree classifier, exactly as described in the original study. The CNN feature extractor was trained for 40 epochs on the distance-matched data, after which its weights were frozen for extended-classifier training.

The resulting model was applied to predict interaction probabilities for the chromatin loops reported by Reed et al.(7). A probability threshold of 0.08 was used to classify pairs as interacting (label = 1) or non-interacting (label = 0). This threshold was selected on the validation set to approximately match the class imbalance ratio observed in the classifier training data. The predicted binary interaction labels at the 0th and 24th hour time points were then used to classify each loop into one of three categories: lost (label changed from 1 at 0th to 0 at 24th hour), gained (label changed from 0 at 0th to 1 at 24th hour), or static (label remained 0 or 1 at both time points).

## S9. Comparison with ChromaFold

To compare UniversalEPI's performance against ChromaFold(13), we trained and evaluated both 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 the sections above. These bulkified ATAC-seq signals and peaks were used to train the UniversalEPI model and the evaluation of both 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.*(13) *i.e.* we used ICE-normalized Hi-C at 10 kb resolution. Z-scores were then calculated using HiC-DC+(14) 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.*(13). 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 up to 1.5 Mb away from each other.

## S10. Esophageal Adenocarcinoma Data Processing

The 10x multiome (single-nuclei ATAC-seq and single-nuclei RNA-seq) profiles of 8 esophageal adenocarcinoma (EAC) patients were obtained from Yates *et al.*(15). Using the cell scores from Yates *et al.*, we selected the top 20% unique cells in each of the differentiated (cNMF4) and undifferentiated (cNMF5) programs as the representative cells.

The representative cell barcodes are extracted from the individual patients' .bam file, and the resulting patient-specific .bam files are merged to obtain the pseudo-bulk ATAC-seq .bam. For each program, ATAC-seq bam files were then used to obtain the signal  $p$ -values bigwig and narrowpeak files using ENCODE's ATAC-seq pipeline(9) as done before. The ATAC-seq peaks were merged for the two programs to obtain a common set of peaks for downstream differential analysis. This was followed by deduplication (as done for datasets before) which resulted in a total of 280,368 peaks. The signal  $p$ -value bigwigs were normalized using edgeR's TMM normalization(10) was done with the GM12878 cell line as the reference (Supplementary Fig. S3).

The pseudo-bulk gene expression is obtained by summing the gene counts from all representative cells. Finally, the gene expression counts for each program were converted to counts-per-million (CPM) to mitigate the library-size bias.

## S11. Promoter Activity

The promoter activity ( $A_P$ ) was obtained using the accessibility of the promoter ( $ATAC_P$ ), accessibility of all the interacting enhancers ( $ATAC_E$ ), and the interaction strength between the enhancer and promoter ( $Hi-C_{E,P}$ ). The accessibility of each enhancer was extracted as the maximum signal in the vicinity ( $\pm 50$ bp) of the gene transcriptional start site or the point of maxima called by peak calling of MACS2 for the enhancers. Specifically, we defined the promoter activity as a linear combination of the promoter accessibility and the Hi-C weighted by the total accessibility of the enhancer and promoter. Inspired by Loubiere *et al.*(16), the total accessibility of enhancer and promoter was defined using the multiplicative model. This is mathematically given by equation (4),

$$A_P = \beta_0 + \beta_1 \log_2(ATAC_P + 1) + \beta_2 \sum_E (ATAC_P \cdot ATAC_E \cdot Hi-C_{E,P}) \quad (4)$$

where  $\beta_0$  is the intercept and  $\beta_1$ , and  $\beta_2$  are the coefficients, all of which are learned by a lasso regressor on randomly selected 80% of protein-coding genes in IMR90 cells. To mitigate the effect of

weak interactions, only those enhancers were considered that had a predicted log Hi-C interaction of greater than 1 with the promoter.

### S12. Availability of Hi-C Predictions on UCSC Genome Browser

We generated the Hi-C predictions corresponding to all possible ATAC-seq data from the ENCODE portal. We included all human cell lines and primary cells that were not genetically perturbed and had sufficient read depth. All the datasets are summarized in Supplementary Table 2. This resulted in 116 ATAC-seq profiles for different cell lines and 41 ATAC-seq profiles for primary cells.

For each profile, the Hi-C predictions were made using the pretrained UniversalEPI models. The predictions were exported as *bigInteract* files with ensemble mean as interaction strength. The estimated uncertainty was reported in the interaction label. Along with the log-transformed Hi-C, we also generated the z-score Hi-C interactions as a separate track, highlighting the strong long-range interactions. For each predicted interaction  $y_{i,j}$  between regions  $i$  and  $j$ , we calculate the z-score normalized Hi-C by equation (5), where  $\mu_d$  and  $\sigma_d$  are mean and standard deviation of all interactions that are at  $d = |j - i|$  distance away in the same chromosome.

$$hic_z(i,j) = \frac{y_{i,j} - \mu_d}{\sigma_d} \quad (5)$$

To ensure that only the strong interactions are viewed by the users, we set a threshold of 5 for ICE-normalized Hi-C predictions (or 1.8 after the log-transformation) and 0 for z-score normalized Hi-C predictions. A total of 314 tracks (157 log-transformed ice-normalized Hi-C + 157 z-score normalized Hi-C) were then submitted to UCSC Genome Browser as a public track hub enabling interactive visualization of predicted chromatin interactions in different cellular contexts. The URL to each track is also reported in Supplementary Table 2.

### S13. Hierarchical Clustering of Cell Lines based on Promoter Activity

For each of the cell lines in the ENCODE database, promoter activity of each protein-coding gene was calculated using predicted Hi-C and experimental ATAC-seq. Using these promoter activity scores as features, the cell lines were hierarchically clustered with  $distance = 1 - R$ , where  $R$  is Pearson's correlation. Linkage was calculated using SciPy's linkage function(17) with *method = 'complete'*.

## References

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

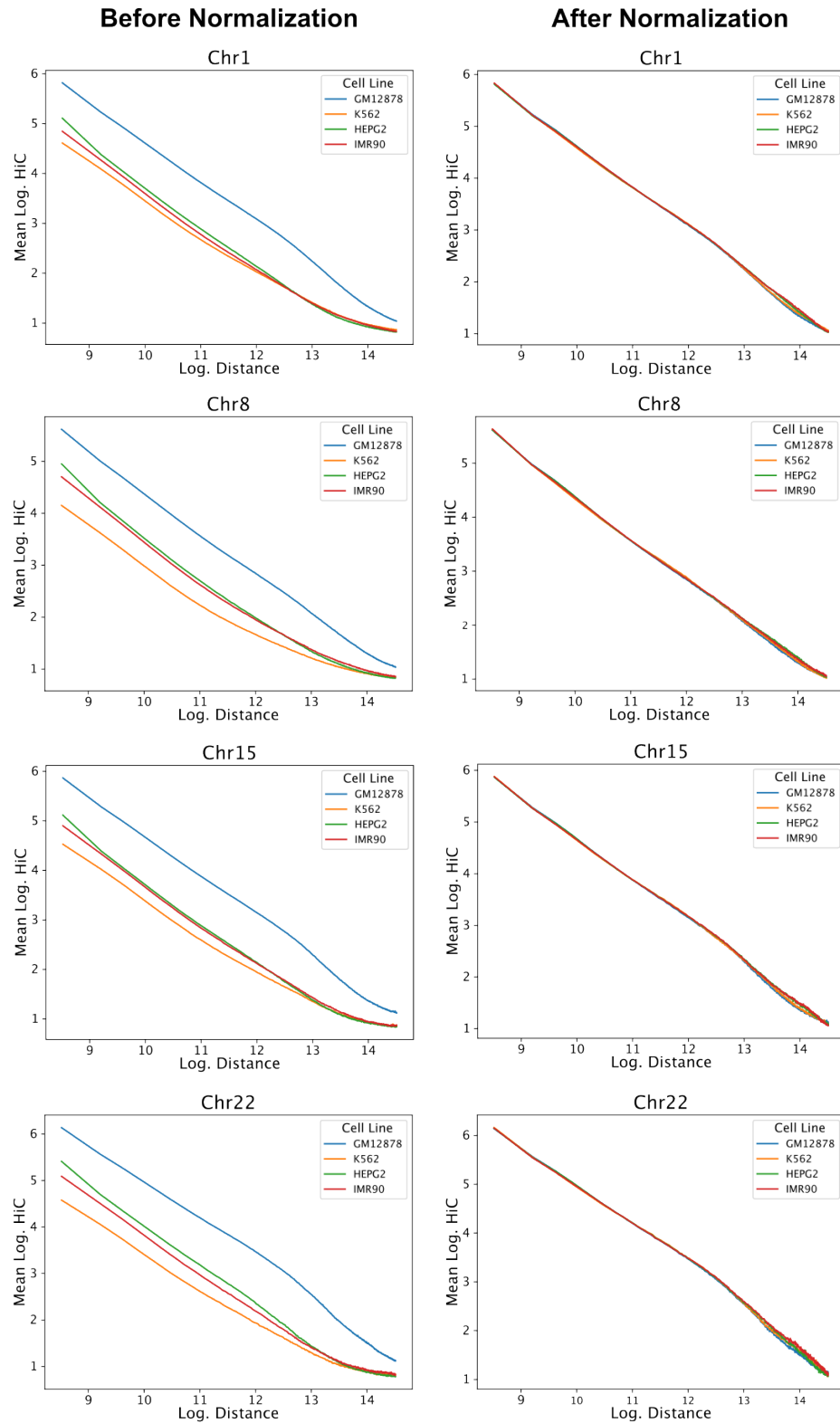
2. Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., *et al.* (2017) The 4D nucleome project. *Nature*, **549**, 219–226.
3. Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., Thandapani, P., Skok, J., Aifantis, I., Fenyő, D., *et al.* (2023) Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol*, **41**, 1140–1150.
4. Fudenberg, G., Kelley, D.R. and Pollard, K.S. (2020) Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods*, **17**, 1111–1117.
5. Zhang, Z., Feng, F., Qiu, Y. and Liu, J. (2023) A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Research*, **51**, 5931–5947.
6. Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G. and Hughes, J.R. (2020) DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods*, **17**, 1118–1124.
7. Reed, K.S.M., Davis, E.S., Bond, M.L., Cabrera, A., Thulson, E., Quiroga, I.Y., Cassel, S., Woolery, K.T., Hilton, I., Won, H., *et al.* (2022) Temporal analysis suggests a reciprocal relationship between 3D chromatin structure and transcription. *Cell Rep*, **41**, 111567.
8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
9. Lee, J., Grey Christoforo, C.S., Foo, P., Probert, C., Anshul Kundaje, B., N., Kohpangwei, D., Dacre, M. and Kim, D. (2016) kundajelab/atac\_dnase\_pipelines: 0.3.3. 10.5281/ZENODO.596029.
10. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.
11. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, **9**, 999–1003.
12. Cao, F., Zhang, Y., Cai, Y., Animesh, S., Zhang, Y., Akincilar, S.C., Loh, Y.P., Li, X., Chng, W.J., Tergaonkar, V., *et al.* (2021) Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome Biology*, **22**, 226.
13. 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.* (2024) ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. *Nat Commun*, **15**, 9432.
14. Sahin, M., Wong, W., Zhan, Y., Van Deynze, K., Koche, R. and Leslie, C.S. (2021) HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. *Nat Commun*, **12**, 3366.
15. Yates, J., Mathey-Andrews, C., Park, J., Garza, A., Gagné, A., Hoffman, S., Bi, K., Titchen, B., Hennessey, C., Remland, J., *et al.* (2024) Cell states and neighborhoods in distinct clinical



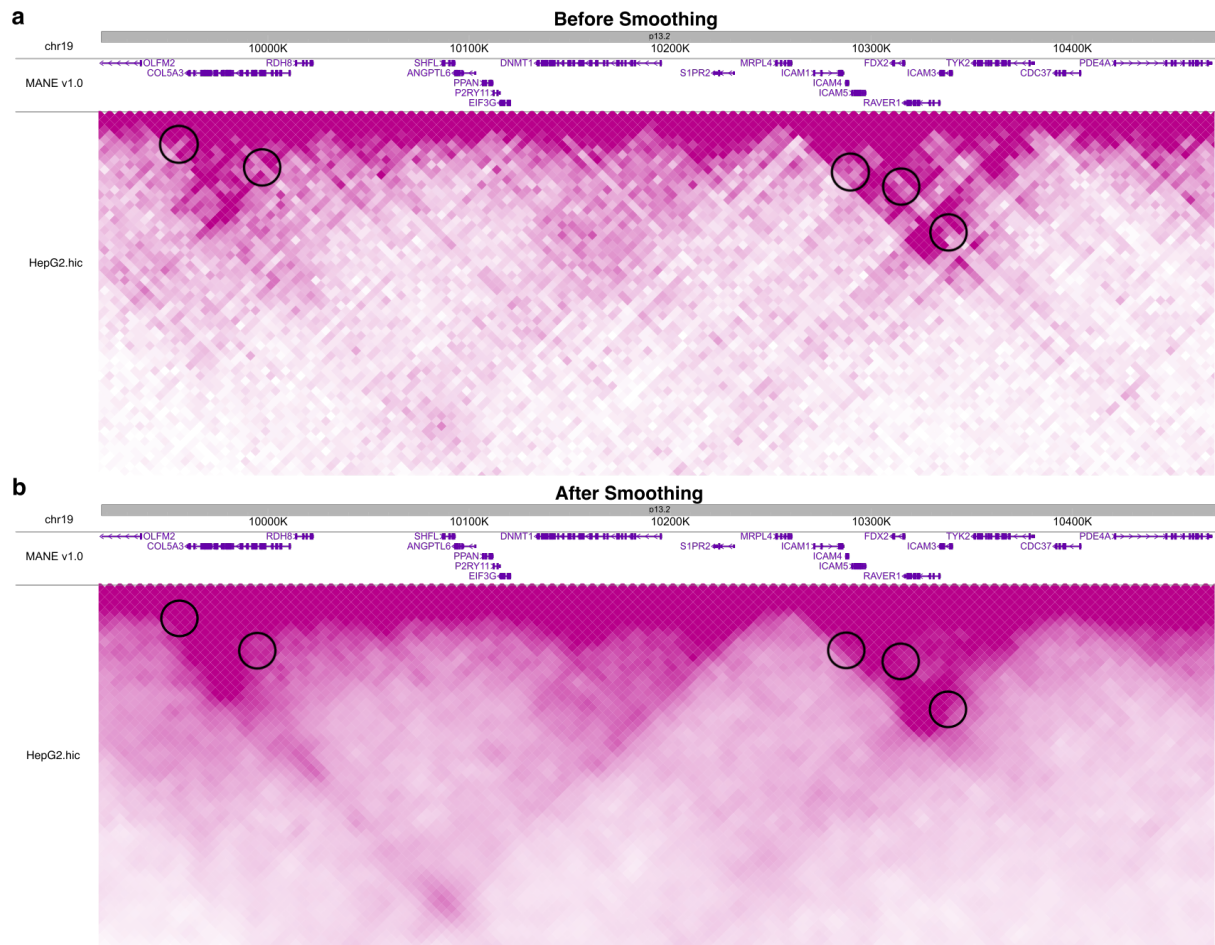
stages of primary and metastatic esophageal adenocarcinoma. *bioRxiv*,  
10.1101/2024.08.17.608386.

16. Loubiere,V., de Almeida,B.P., Pagani,M. and Stark,A. (2024) Developmental and housekeeping transcriptional programs display distinct modes of enhancer-enhancer cooperativity in *Drosophila*. *Nat Commun*, **15**, 8584.
17. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, **17**, 261–272.

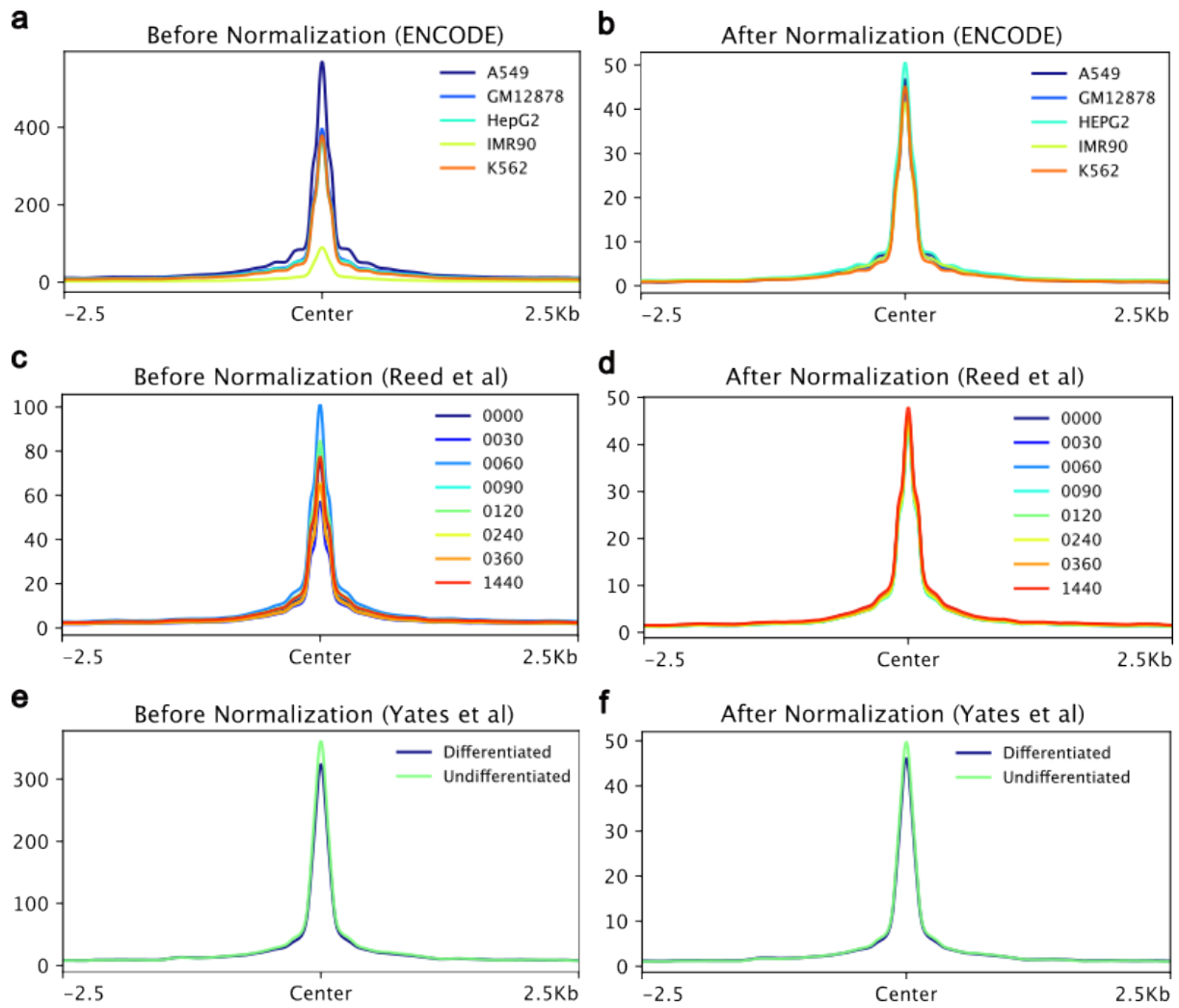
## Supplementary Figures



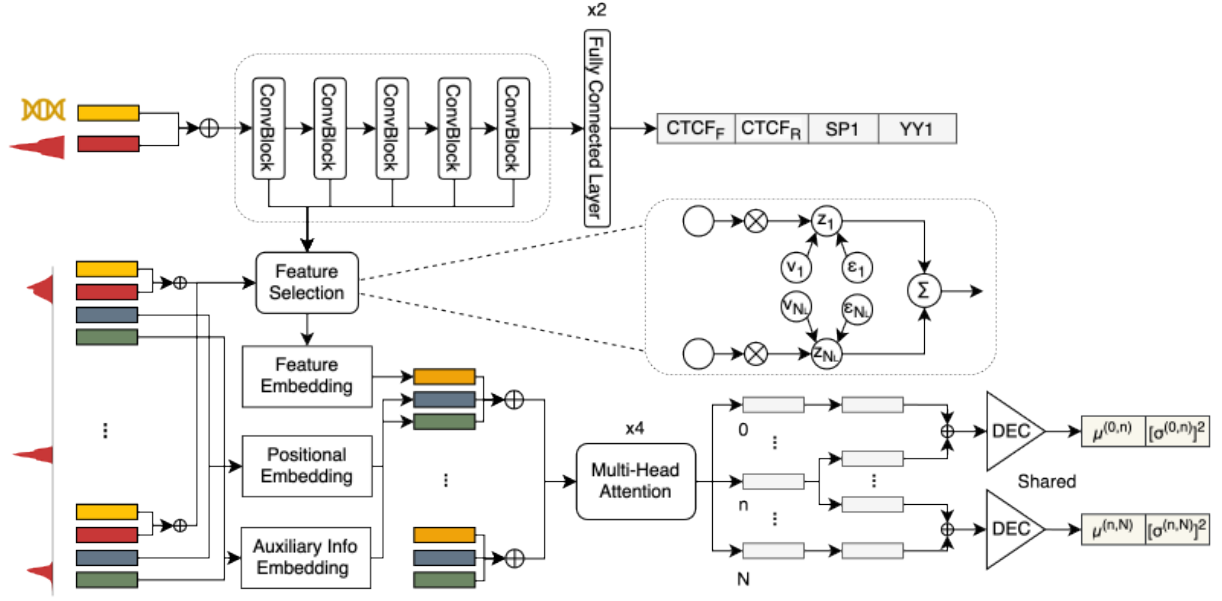
Supplementary Fig. S1: **Cross-cell-type normalization of Hi-C.** Before and after distance-stratified robust z-score normalization on four randomly selected chromosomes (chromosomes 1, 8, 15, and 22) on all cell lines. GM12878 is used as the reference cell line due to its high sequencing depth.



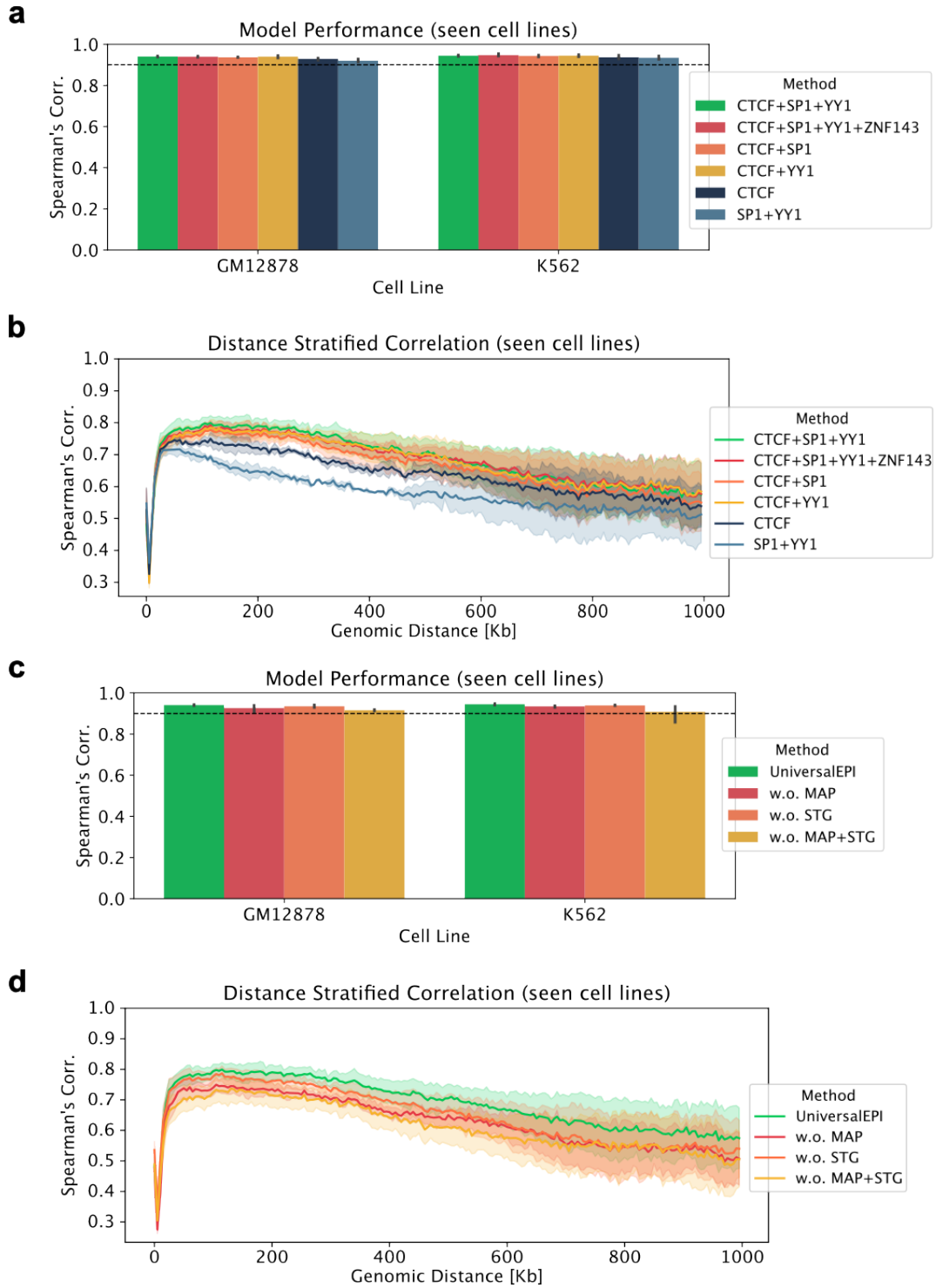
Supplementary Fig. S2: **Smoothing of Hi-C matrices.** The effect of our variable Gaussian smoothing on a randomly selected region of chromosome 19 in the HepG2 cell line. **a**, We can observe gaps(circled) in the original Hi-C matrix that can be attributed to the technical effects. **b**, These gaps are filled using our variable smoothing thereby giving a more realistic Hi-C matrix. The figure is generated using the WashU Epigenome Browser.



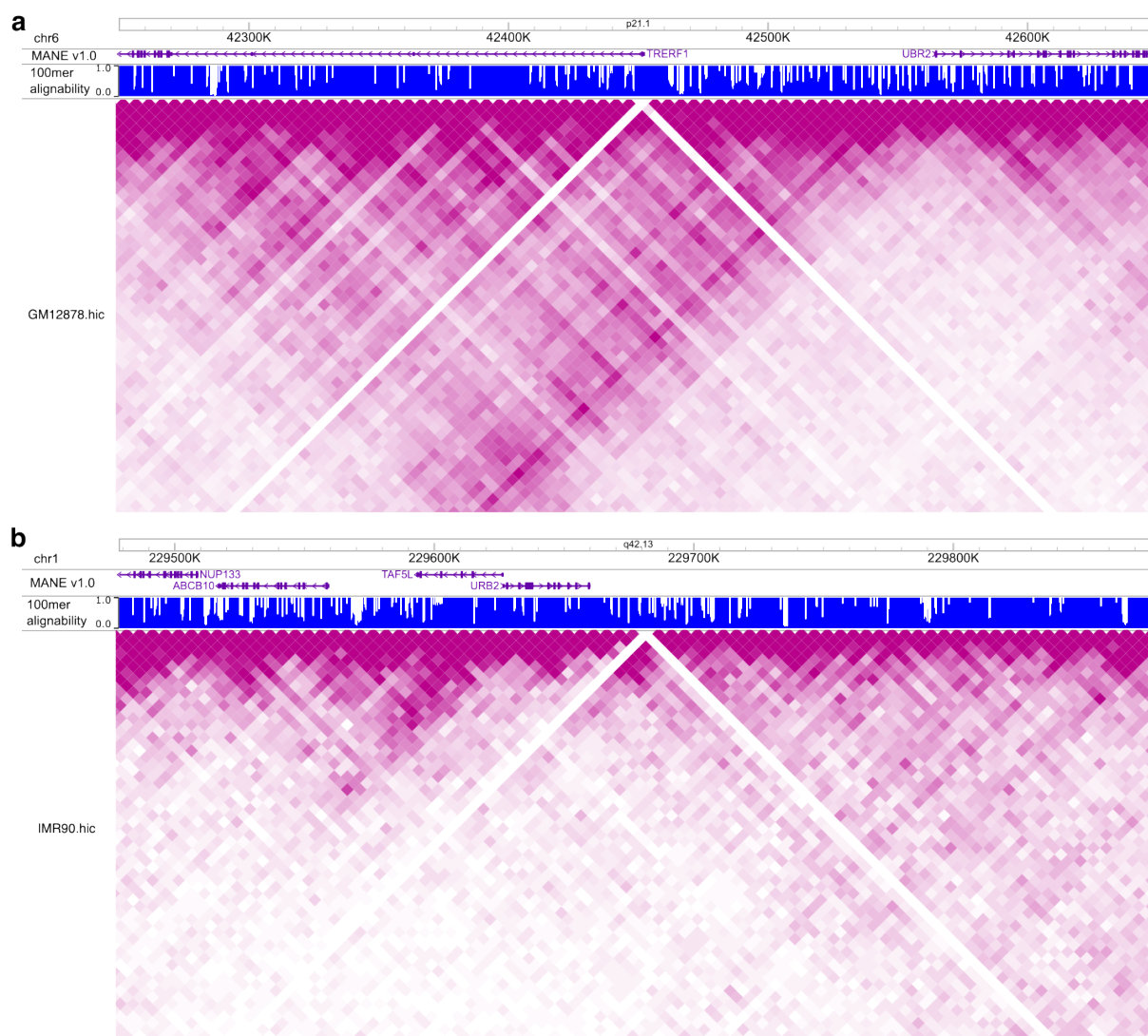
Supplementary Fig. S3: **Cross-cell-type normalization of ATAC-seq signal.** **a-b**, Before and after applying edgeR's trimmed mean of M-values normalization on ENCODE cell lines with GM12878 as the reference cell line. The conserved active CTCF sites are used as reference regions. **c-d**, Same normalization is applied to the macrophage differentiation data (Reed *et al*) with GM12878 as reference. **e-f**, The ATAC-seq from esophageal carcinoma (Yates *et al*) is normalized using the same normalization with GM12878 as reference.



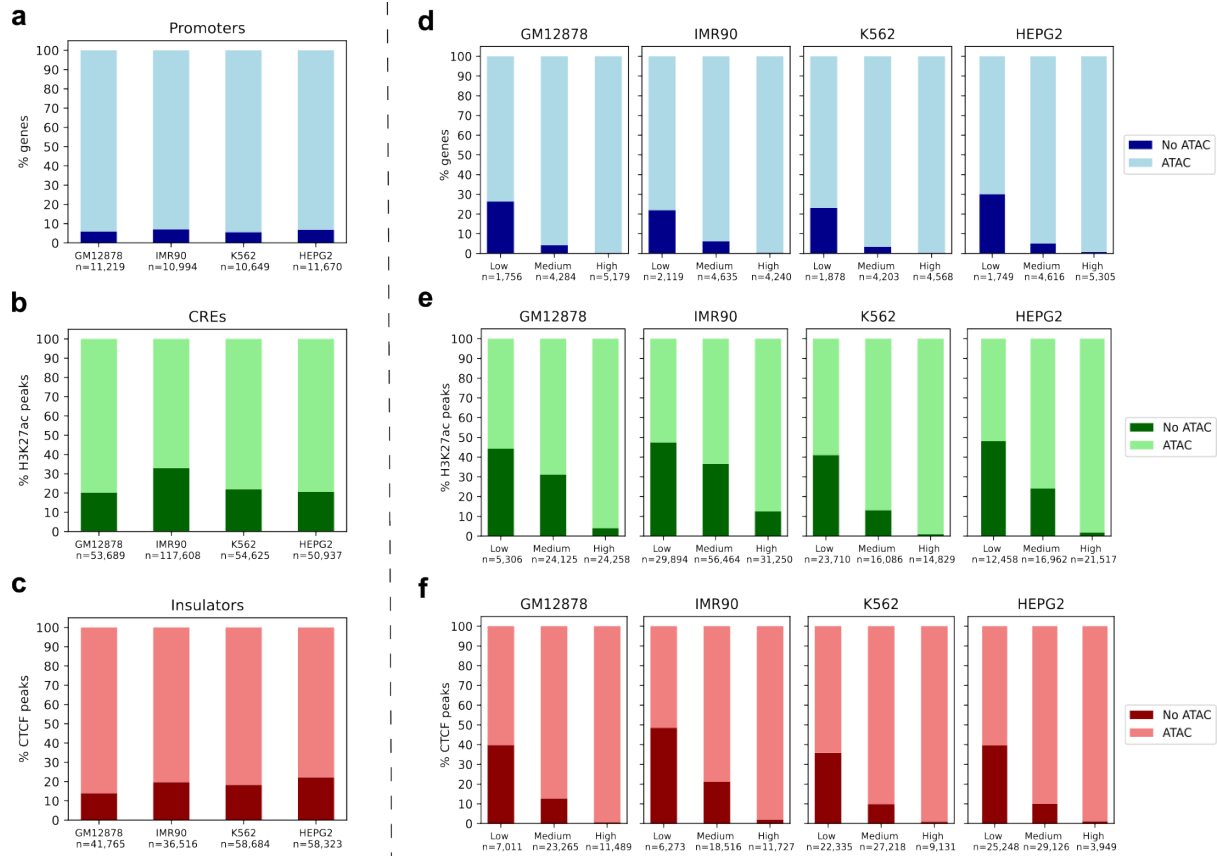
Supplementary Fig. S4: **A graphical illustration of UniversalEPI.** UniversalEPI consists of a CNN based representation network, a feature selection component leveraging stochastic gating (STG), and a transformer based Hi-C prediction network. The representation network takes one-hot encoded DNA sequences and ATAC signals as input, and predicts the binding affinity of four transcription factors. It comprises five convolutional blocks, followed by two fully connected layers. The Hi-C prediction network takes a sequence of ATAC peak regions. Each region is representation by the features from learned convolutions selected using the STG mechanism, its genomic distance and auxilliary information, such as mappability. Before being fed into the multi-head attention layers, each component is embedded via a linear projection layer. The decoder (DEC) operates on pair of output tokens at positions  $i$  and  $j$  of the attention layers, to estimate the corresponding mean  $\mu^{(i,j)}$  and variance  $[\sigma^{(i,j)}]^2$  of the Hi-C interaction value.  $\oplus$  and  $\otimes$  denote concatenation and multiplication, respectively.



Supplementary Fig. S5: **Ablation studies are performed on UniversalEPI.** Unseen chromosomes of the training cell lines (GM12878 and K562) are used to perform all the ablations. Moreover, the ablations are performed on UniversalEPI without uncertainty estimation and ATAC-seq as input. **a**, Different sets of transcription factors that are used for training the first stage of UniversalEPI are compared based on Spearman's correlation. **b**, Distance-stratified Spearman's correlation is used to reflect differences between models that have similar overall performance. The shaded region represents the variation across the two cell lines. **c**, The effect of using mappability tracks as auxiliary information in the second stage of UniversalEPI and the effect of applying stochastic gating is compared. **d**, Distance-stratified Spearman's correlation is studied as above.

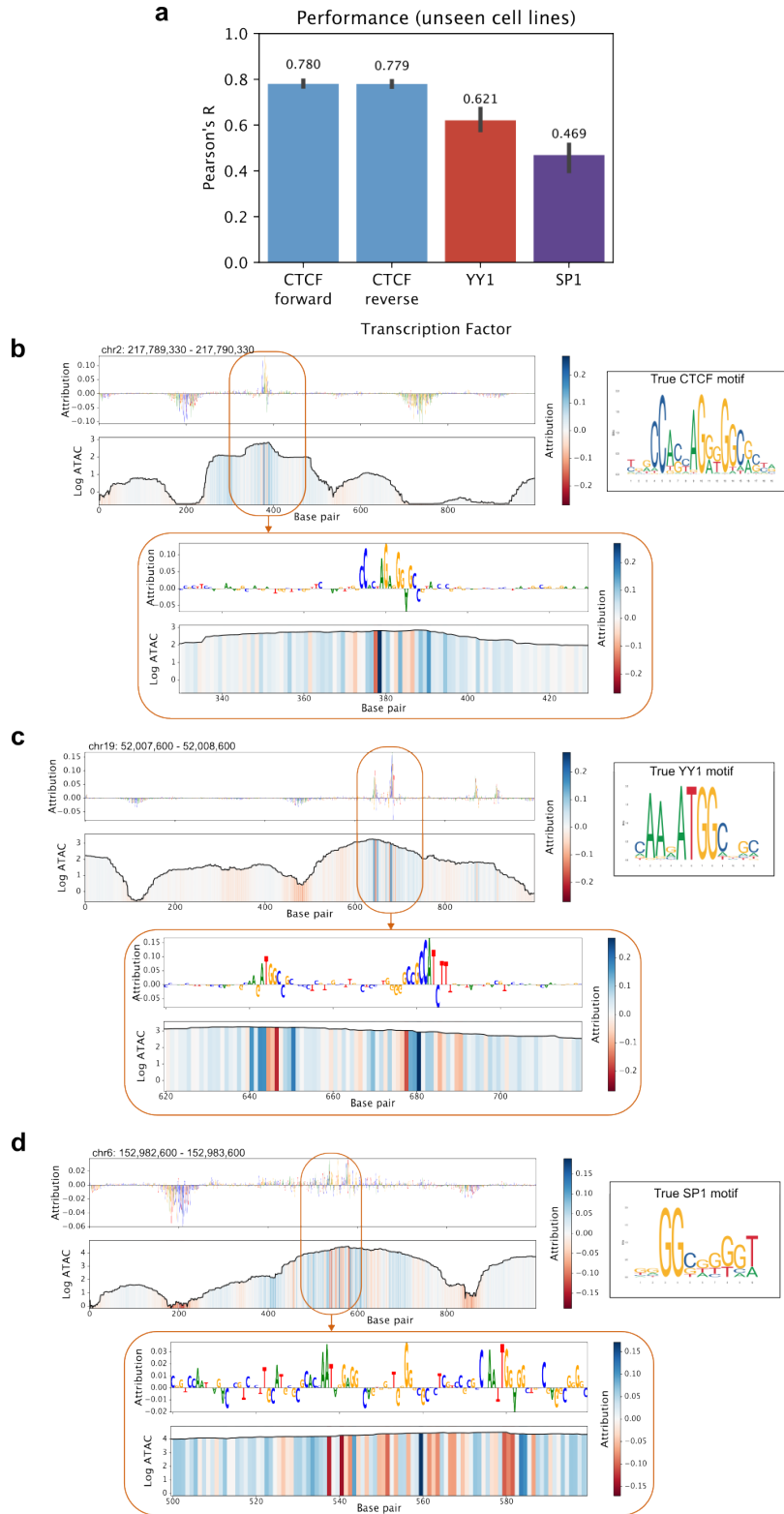


Supplementary Fig. S6: **Examples of blacklisted regions.** Several bins in the Hi-C data of each cell line arbitrarily has no interactions despite being mappable. **a**, An example of such a region in the GM12878 cell line is observed at chromosome 6 for the bin 42,450,000-42,455,000. **b**, Another example can be seen in the IMR90 cell line at chromosome 1 between 229,680,000 and 229,685,000. The figure is generated using the WashU Epigenome Browser.

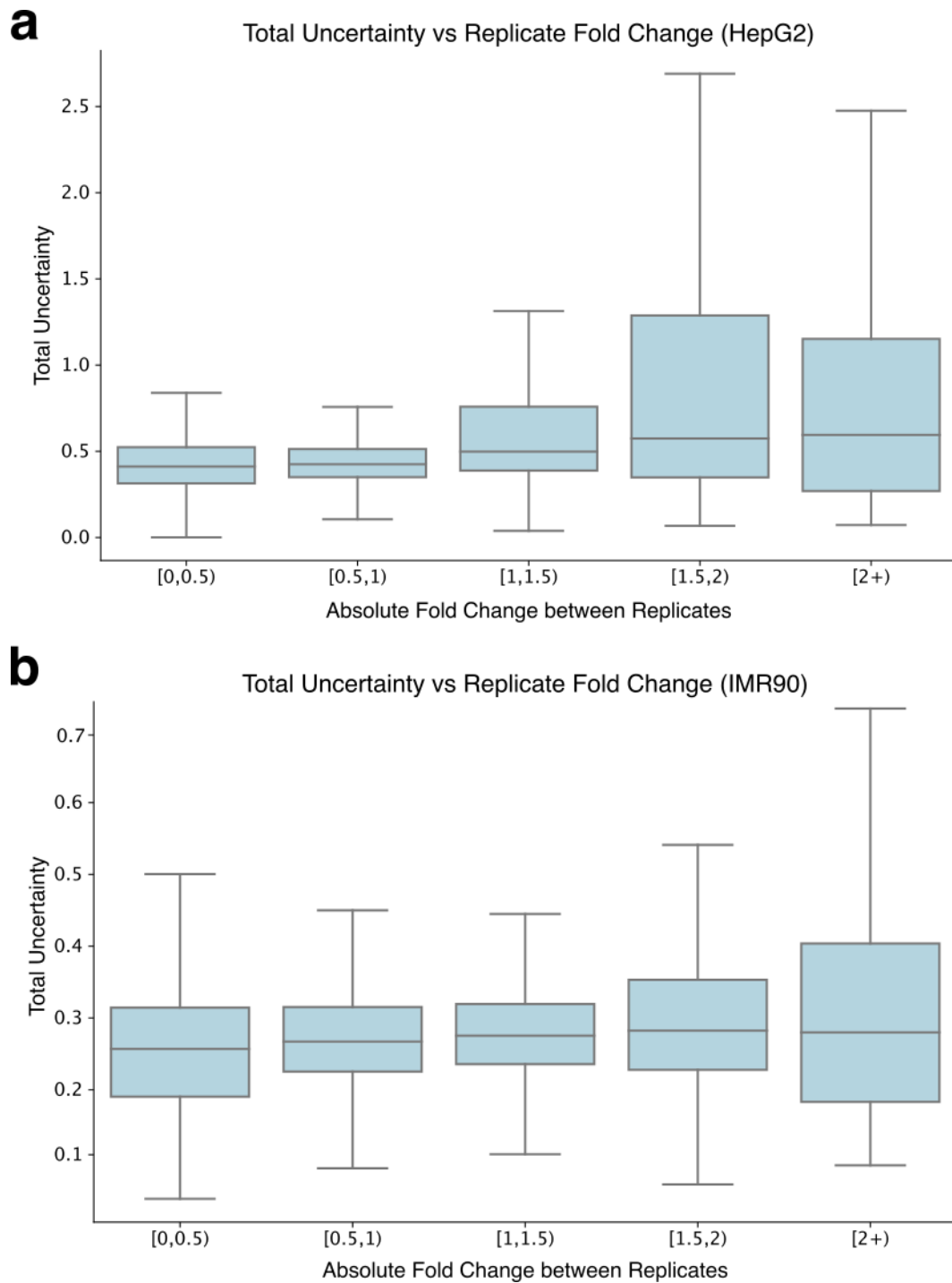


**Supplementary Fig. S7: Information loss by retaining only accessible chromatin regions.** Information lost due to the selection of accessible chromatin regions as measured by the overlap of ATAC-seq peaks with the transcription start site of the **a**, expressed genes (promoter-like), **b**, H3K27ac histone modification (cis-regulatory elements or CREs), and **c**, CTCF sites (insulator-like) in four cell lines: GM12878, IMR90, K562, and HepG2. These cell lines have 186,423, 174,095, 179,240, and 175,487 peaks respectively. Nearly 95% of active promoters overlap with ATAC-seq peaks whereas approximately 80% of active CREs and insulators are captured by ATAC-seq peaks. **d-f**, The promoters, CREs, and insulators not captured by ATAC-seq peaks often have low coverage.

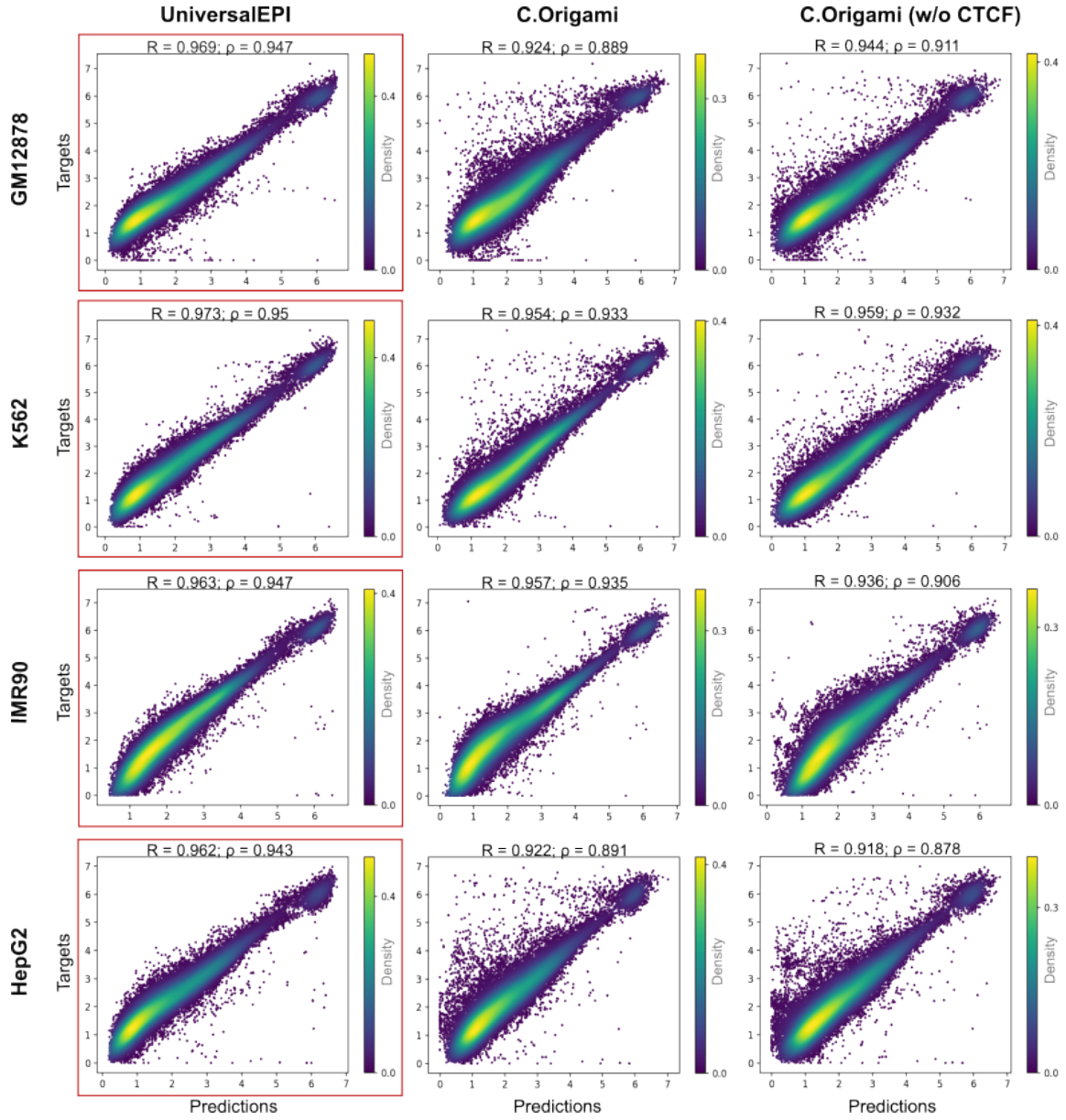




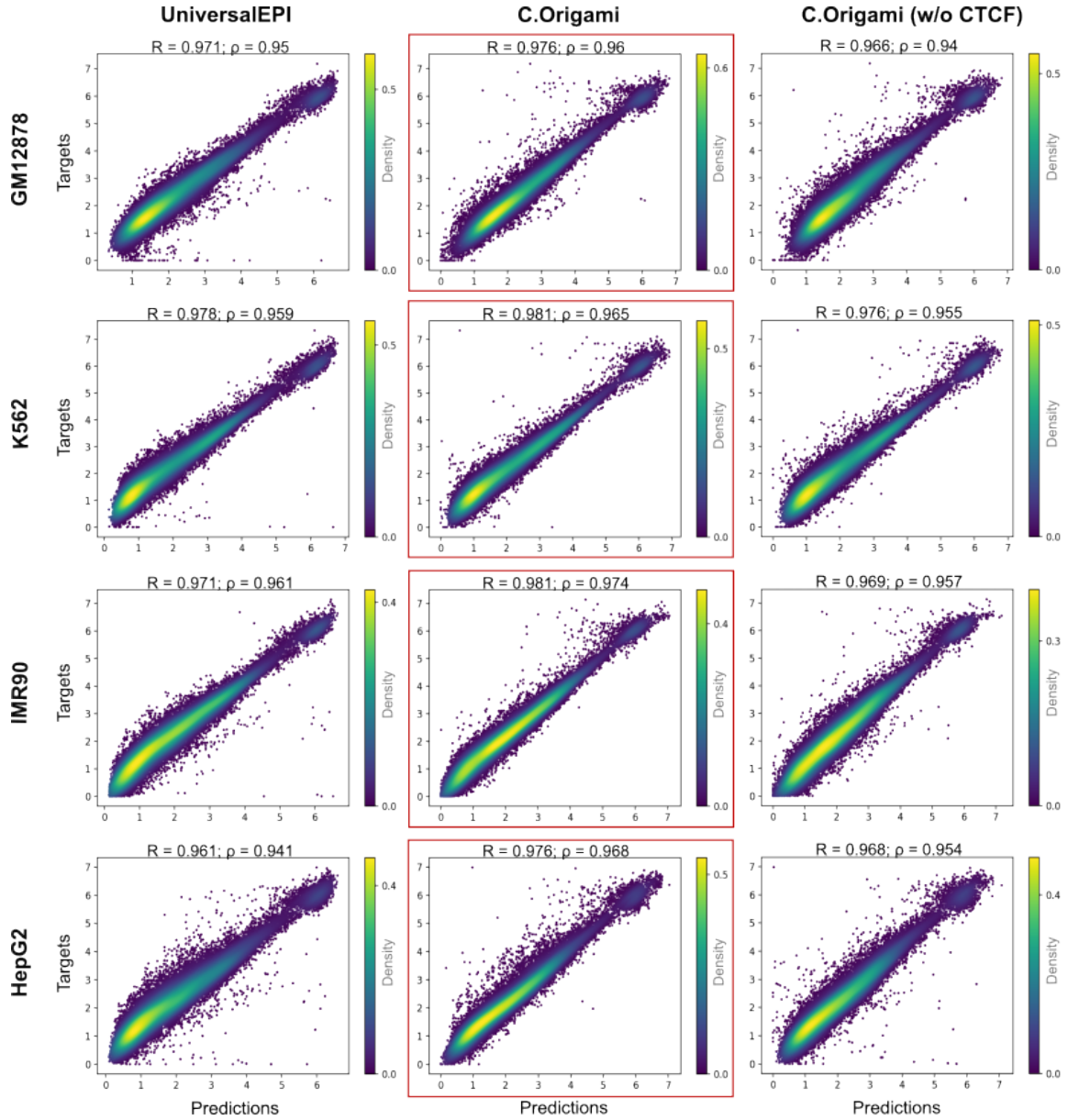
Supplementary Fig. S8: **Performance of first stage of UniversalEPI.** **a**, Predictive performance of the first stage of UniversalEPI on unseen chromosomes of cell lines that are not seen during training (GM12878, K562, HepG2, A549). The mean score across the four cell lines is reported and the range is represented above the bar. The true motif is obtained from JASPAR. **b**, DeepLIFT attribution scores on unseen chromosome of unseen cell line (HepG2) matches the CTCF motif. **c**, DeepLIFT attribution score is matched with YY1 motif. Two YY1 motifs are identified in this example (one in reverse complement). **d**, SP1 motif is also identified by the model.



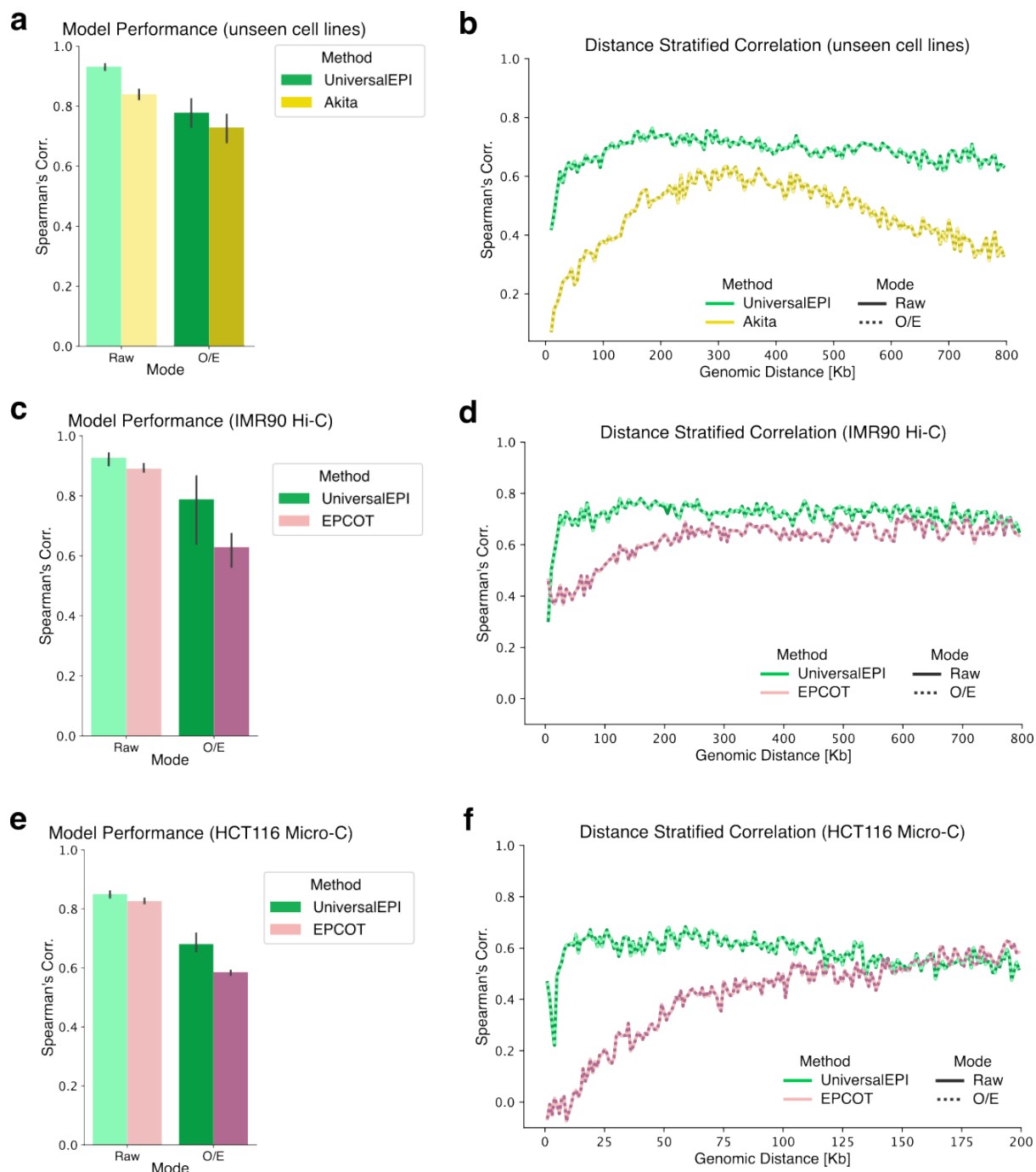
Supplementary Fig. S9: **Comparison between UniversalEPI's estimated uncertainty and biological variance in unseen cells.** Relationship between biological replicate variance, calculated as a absolute fold change between two replicates, and prediction uncertainty is illustrated on the test chromosomes of unseen an cell lines: **a**, HepG2, and **b**, IMR90.



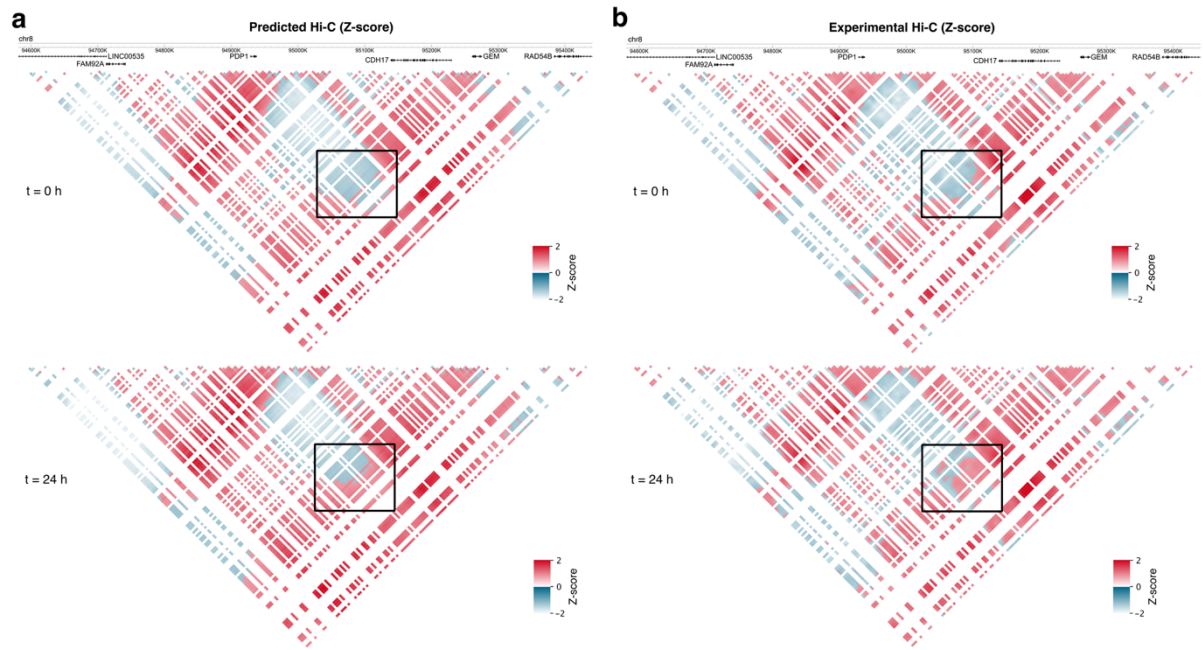
Supplementary Fig. S10: **Performance comparison between UniversalEPI and C.Origami on unseen cell lines.** Comparison of UniversalEPI against state-of-the-art methods like C.Origami, and C.Origami (when CTCF ChIP-seq is removed from the input) on the test chromosomes (chromosomes 2, 6, and 19) of cell lines that were not seen by the model during training. Two versions of models are used here: one is trained on GM12878 and K562 cell lines and the other is trained on IMR90 and HepG2 cell lines. Pearson's correlation ( $R$ ) and Spearman's correlation ( $\rho$ ) are calculated for each method and cell line. The best performing method for each cell line (based on  $\rho$ ) is highlighted in red.



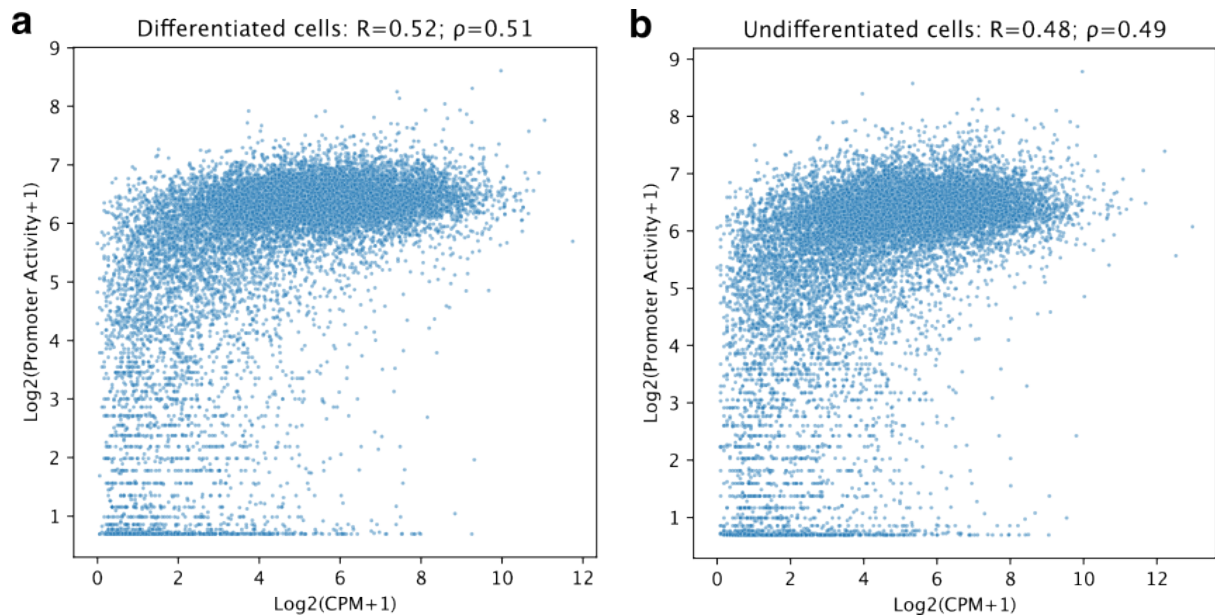
Supplementary Fig. S11: **Performance comparison between UniversalEPI and C.Origami on seen cell lines.** Comparison of UniversalEPI against state-of-the-art methods like C.Origami, and C.Origami (when CTCF ChIP-seq is removed from the input) on the test chromosomes (chr2, chr6, and chr19) of cell lines that were seen by the model during training. Two versions of models are used here: one is trained on GM12878 and K562 cell lines and the other is trained on IMR90 and HepG2 cell lines. Pearson's correlation ( $R$ ) and Spearman's correlation ( $\rho$ ) are calculated for each method and cell line. The best performing method for each cell line (based on  $\rho$ ) is highlighted in red.



Supplementary Fig. S12: **Normalization-independent comparison of UniversalEPI against Akita and EPCOT.** **a-b**, Comparison between UniversalEPI and Akita on unseen cell lines (GM12878, K562, IMR90, and HepG2) for predicting 5-kb-resolution Hi-C signal on test chromosomes (chr2, chr6, and chr19). Since Akita outputs observed/expected (O/E) Hi-C matrices while UniversalEPI outputs raw contact maps, the models are compared in both settings. UniversalEPI's predictions are converted to O/E by dividing each prediction with distance-stratified mean whereas Akita's O/E predictions are converted to raw by multiplying each prediction by expected contact, calculated using training data. **b**, Distance-stratified Spearman's correlation between predictions and ground truth provides a normalization-independent basis for comparison. **c-d**, Similar comparison is done between UniversalEPI and EPCOT on the task of predicting 5-kb-resolution Hi-C signal in unseen cell line (IMR-90) **e-f**, Comparison of UniversalEPI and EPCOT on Micro-C data on the unseen HCT116 cell line at 1-kb resolution.

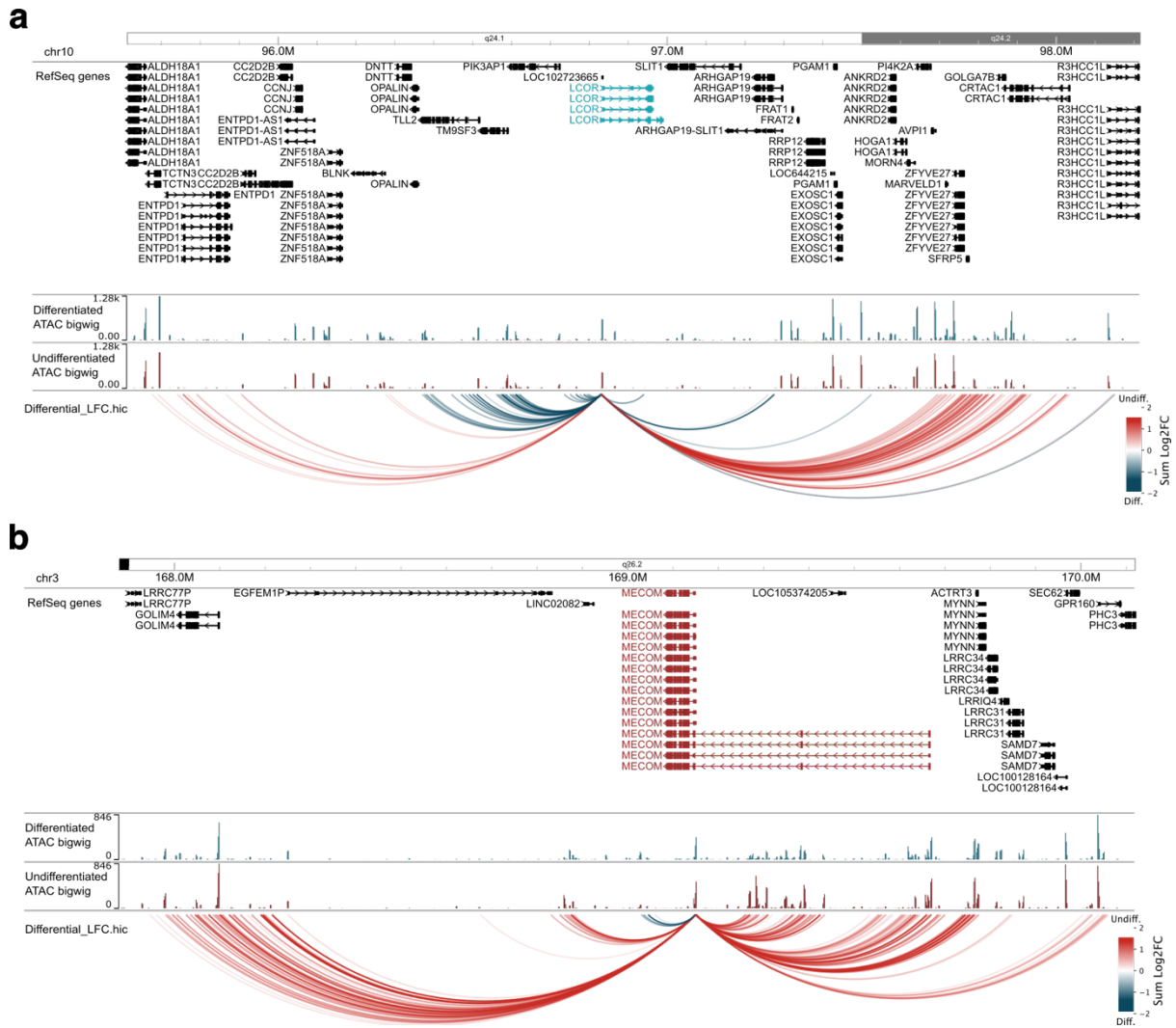


Supplementary Fig. S13: **Example of predicted z-score Hi-C interactions compared to the experimental interactions during macrophage activation.** **a**, Z-score Hi-C interactions predicted by UniversalEPI at time points 0h and 24h after the LPS + IFN $\gamma$  treatment in resting macrophages. **b**, Experimental z-score Hi-C interactions as generated in Reed *et al.* The interactions in the neighborhood of a distal enhancer with the GEM promoter are highlighted in the square box. The figure is generated using the WashU Epigenome Browser.

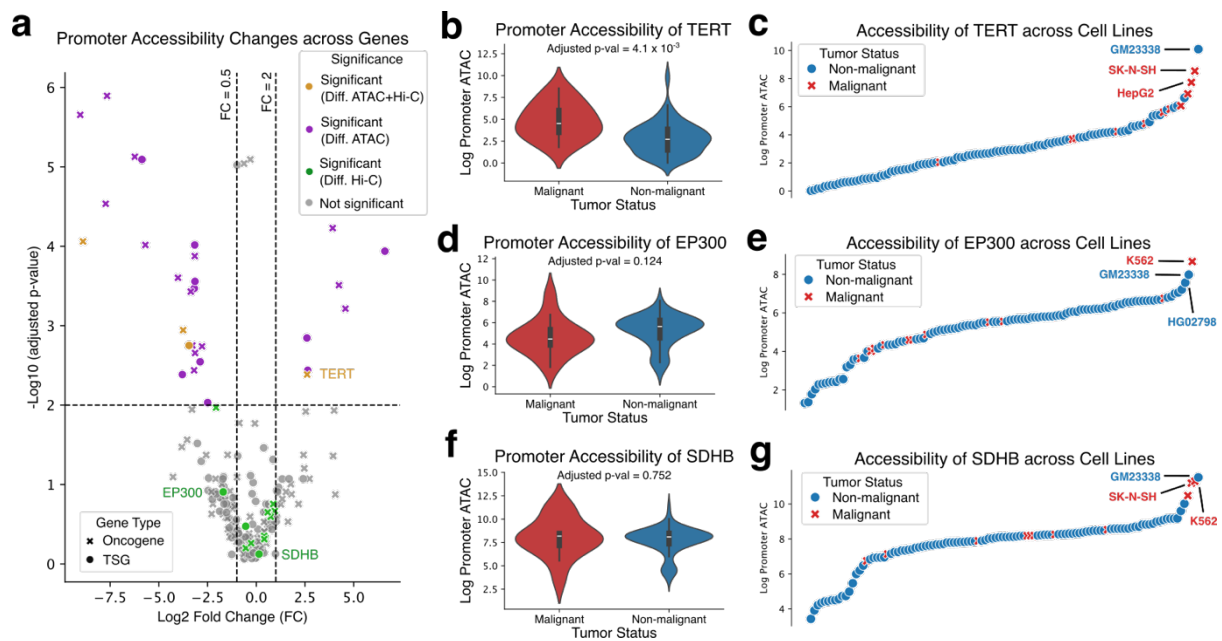


Supplementary Fig. S14: **Correlation of promoter activity with gene expression in EAC.** **a-b**, Pearson's and Spearman's correlations ( $R$  and  $\rho$  respectively) of the estimated promoter activity using experimental ATAC-seq and Hi-C derived from UniversalEPI in differentiated and undifferentiated cells in EAC patients.





Supplementary Fig. S15: **Differential interactions in EAC without using uncertainty.** Differential interactions between undifferentiated and differentiated cells in EAC are predicted using UniversalEPI. No filtering is applied based on estimated uncertainty. **a**, Different sets of strong interactions are identified in favor of differentiated and undifferentiated cells with LCOR promoter. LCOR is a master transcription factor in differentiated cells. **b**, Similar interactions are shown for MECCOM, which is a master transcription factor in undifferentiated cells. The figure is generated using the WashU Epigenome Browser.



Supplementary Fig. S16: **Analysis of promoter accessibility across malignant and non-malignant cell lines.** **a**, Volcano plot highlighting change in promoter ATAC-seq between malignant and non-malignant cell lines for all oncogenes and tumor suppressor genes. The fold change (FC) is calculated as the ratio of the average promoter ATAC-seq in malignant cell lines to non-malignant cell lines. **b-g**, Promoter accessibility between malignant and non-malignant cell lines for TERT (**b**), EP300 (**d**), SDHB (**f**) genes. The promoter accessibility in each cell line for these genes are highlighted in (**c**), (**e**), and (**g**), respectively.