nature communications

Article

https://doi.org/10.1038/s41467-024-53628-0

ChromaFold predicts the 3D contact map from single-cell chromatin accessibility

Received: 18 October 2023

Accepted: 14 October 2024

Published online: 01 November 2024

Check for updates

Vianne R. Gao[1,2], Rui Yang[1,2], Arnav Das[3], Renhe Luo[4], Hanzhi Luo[5], Dylan R. McNally[6], Ioannis Karagiannidis[7], Martin A. Rivas[7,8], Zhong-Min Wang[9], Darko Barisic[7], Alireza Karbalayghareh[1], Wilfred Wong[1,2], Yingqian A. Zhan[10], Christopher R. Chin[7], William S. Noble[3], Jeff A. Bilmes[3], Effie Apostolou[11,13], Michael G. Kharas[5,13], Wendy Béguelin[7,13], Aaron D. Viny[12,13], Danwei Huangfu[4,13], Alexander Y. Rudensky[9,13], Ari M. Melnick[7,13] & Christina S. Leslie[1] ✉

Identifying cell-type-specific 3D chromatin interactions between regulatory elements can help decipher gene regulation and interpret disease-associated non-coding variants. However, achieving this resolution with current 3D genomics technologies is often infeasible given limited input cell numbers. We therefore present ChromaFold, a deep learning model that predicts 3D contact maps, including regulatory interactions, from single-cell ATAC sequencing (scATAC-seq) data alone. ChromaFold uses pseudobulk chromatin accessibility, co-accessibility across metacells, and a CTCF motif track as inputs and employs a lightweight architecture to train on standard GPUs. Trained on paired scATAC-seq and Hi-C data in human samples, ChromaFold accurately predicts the 3D contact map and peak-level interactions across diverse human and mouse test cell types. Compared to leading contact map prediction models that use ATAC-seq and CTCF ChIP-seq, ChromaFold achieves state-of-the-art performance using only scATAC-seq. Finally, fine-tuning ChromaFold on paired scATAC-seq and Hi-C in a complex tissue enables deconvolution of chromatin interactions across cell subpopulations.

Genome-wide chromosome conformation capture techniques such as Hi-C, HiChIP, and ChIA-PET[1–3] provide powerful tools for mapping cell-type-specific regulatory interactions that can link enhancers to genes and enable the interpretation of non-coding disease-associated variants[4,5]—at least when there is sufficient input material to generate high-complexity libraries and allow for very deep sequencing. Indeed, the use of these assays is often impeded by their substantial costs, time requirements, and technical difficulty, especially when studying rare

[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Tri-Institutional Program in Computational Biology and Medicine, New York, NY, USA. [3]University of Washington, Seattle, WA, USA. [4]Developmental Biology Program, Sloan Kettering Institute, New York, NY, USA. [5]Molecular Pharmacology Program, Experimental Therapeutics Center and Center for Stem Cell Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Caryl and Israel Englander Institute for Precision Medicine, Institute for Computational Biomedicine, Weill Cornell Medicine, Cornell University, New York, NY, USA. [7]Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medical College, New York, NY, USA. [8]Department of Biochemistry & Molecular Biology; Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA. [9]Howard Hughes Medical Institute and Immunology Program, Sloan Kettering Institute and Ludwig Center at Memorial Sloan Kettering Cancer Center, New York, NY, USA. [10]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [11]Joan and Sanford I. Weill Department of Medicine, Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [12]Departments of Medicine, Division of Hematology & Oncology, and of Genetics & Development, Columbia Stem Cell Initiative, Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA. [13]These authors contributed equally: Effie Apostolou, Michael G. Kharas, Wendy Béguelin, Aaron D. Viny, Danwei Huangfu, Alexander Y. Rudensky, Ari M. Melnick. ✉e-mail: lesliec@mskcc.org

cell populations where obtaining a sufficient number of cells for a high-quality contact map becomes impractical[6,7]. On the other hand, single-cell chromosome conformation mapping technologies, such as single-cell Hi-C or ChIA-Drop, although exciting, are experimentally challenging and produce sparse datasets that are typically analyzed at 100 kb–1 Mb resolution[8–11]. By contrast, single-cell chromatin accessibility (scATAC-seq) datasets can be readily generated from small amounts of input material due to the availability of commercial kits[12]. Genome-wide chromatin accessibility profiles reflect the extent to which nuclear molecules, including transcription factors, chromatin remodelers, histones, and other chromatin-associated proteins, can physically interact with chromatinized DNA, and single-cell chromatin accessibility contains subtle information about pairwise 3D interactions[13]. This raises the question of whether one can predict chromatin interactions and connect regulatory elements to their target genes using scATAC-seq data alone.

Several models have been proposed to predict chromatin interactions from genomic sequence and easier-to-obtain bulk or single-cell epigenomic data[14–19]. For instance, Cicero was the first method to leverage the co-accessibility structure between accessible elements ('peaks') in scATAC-seq data to infer chromatin interactions in an unsupervised fashion[18]. DeepC[19], Akita[14], and Orca[15] are supervised deep neural network-based models that predict chromatin contact maps from genomic DNA sequences. Epiphany, a model we introduced recently for cell-type-specific contact map prediction, uses a collection of bulk 1D epigenomic input tracks to enable generalization to novel cell types[17]. Another recent model, C.Origami, is also capable of making cell-type-specific predictions using DNA sequence together with bulk ATAC-seq and CTCF ChIP-seq in the target cell type[16]. However, these existing models for chromatin interaction prediction have practical limitations. Unsupervised models like Cicero offer modest accuracy, whereas sequence-based models such as DeepC, Akita, and Orca fail to generalize effectively to new cell types and, indeed, tend to predict similar contact maps across training cell types[14,16]. Meanwhile, C.Origami and Epiphany both require multiple input data modalities, which are not always available, and C.Origami, in particular, employs a more complex model that may be susceptible to overfitting[20].

In this study, we introduce ChromaFold, a supervised deep-learning model that predicts the 3D contact map from scATAC-seq data and CTCF motif tracks as input features. Given the linkage between the accessibility landscape of regulatory elements and 3D genome organization, our underlying hypothesis is that we can leverage the covariation in accessibility stemming from asynchronous chromatin looping events across single cells. This assumption is further substantiated by prior studies showing that pairs of genomic bins with high co-accessibility are enriched for chromatin looping events[18,21]. Additionally, given the crucial role of the CTCF protein in shaping 3D chromatin structure, the inclusion of CTCF-associated signals is expected to enhance the model's predictive power[22–24]. For wider adaptability, we do not require CTCF ChIP-seq as an input and offer two versions of ChromaFold. *ChromaFold + CTCF motif* uses CTCF motif score, a measure of the likelihood that a genomic region contains a binding site for the CTCF protein, as a proxy for CTCF binding[25]. *ChromaFold + CTCF ChIP* uses the actual CTCF ChIP-seq track as input (unless otherwise noted, ChromaFold refers to *ChromaFold + CTCF motif*).

The key advantages of ChromaFold include its requirement of only scATAC-seq data as experimental input data, its ability to make cell-type-specific predictions in new cell types, and its lightweight architecture, making it compatible with standard GPUs. Importantly, ChromaFold can also be employed to deconvolve bulk chromatin interaction data across constituent cell types—resolving the cell-type-specificity of chromatin interactions—by fine-tuning bulk Hi-C and scATAC-seq data from the same complex tissue.

We evaluated ChromaFold on five human and three mouse test cell types and tissues. ChromaFold was able to make accurate cell-type-specific predictions of 3D contact maps (as evaluated by distance-stratified Pearson correlation) and peak-level interactions (as evaluated by receiver operating characteristic and precision-recall analysis) in new cell types and species. In particular, ChromaFold predictions at important lineage-defining loci in murine germinal center B cells (GCBs), regulatory T (Treg) cells, and hematopoietic stem cells (HSCs) recovered correct cell-type-specific 3D interactions. Interestingly, despite its smaller model and reduced information requirements, ChromaFold's performance was comparable to C.Origami when using CTCF motif information as input and outperformed C.Origami when using CTCF ChIP-seq track as input on new cell types. Finally, using paired Hi-C and scATAC-seq in human pancreatic islets, ChromaFold successfully deconvolved chromatin interactions into those specific to alpha cells and beta cells.

Overall, ChromaFold achieves state-of-the-art generalization to novel cell types while requiring only a single input modality to enable accurate Hi-C contact map predictions, including regulatory interaction predictions, in any setting where scATAC-seq can be generated.

Results

ChromaFold is a deep-learning model that predicts 3D contact maps from scATAC-seq data

To enable fast and accurate prediction of chromatin contacts from scATAC-seq data alone, we developed ChromaFold, a lightweight convolutional neural network-based model that makes cell-type-specific predictions. ChromaFold is trained on paired scATAC-seq and Hi-C data from a panel of training cell types. ChromaFold takes three input tracks—pseudobulk chromatin accessibility and correlation structures in accessibility (co-accessibility) profiles across cells, both computed from scATAC-seq, and predicted CTCF motif scores—all processed for a 4.01 Mb genomic region (Fig. 1a). These processed inputs are passed through two feature extractors in the ChromaFold architecture. The first feature extractor takes the pseudobulk accessibility and CTCF motif score tracks with a bin size of 50 bp as input, while the second takes the co-accessibility with a bin size of 500 bp as input. For memory efficiency, we only compute the co-accessibility between the genomic bins in the center 10 kb region with the rest of the bins in the 4.01 Mb region as input. These extractors produce a latent representation of the genomic region, which is then passed through the linear predictor to predict the chromatin interactions between the center genomic bin and its neighboring bins within a 2 Mb distance (V-stripe) at 10 kb resolution, using the HiC-DC + Z-score[26] normalized Hi-C contact map for the corresponding region and cell type as the target (Fig. 1b and Supplementary Fig. 1a).

To process the input data, the CTCF motif score track is generated by scanning a set of CTCF position weight matrices[27,28] (Supplementary Fig. 1b) across the DNA sequence. The pseudobulk chromatin accessibility is obtained by aggregating the accessibility profile across single cells in a population. The co-accessibility is derived by first generating metacells to combat sparsity, then calculating the Jaccard similarity[29] between binarized accessibility profiles across metacells. During training, we randomly subsample single cells and metacells from the population per iteration to generate pseudobulk accessibility and co-accessibility input data, respectively. This data augmentation step is critical for improving model generalizability to datasets of varying quality and size[30–32]. As a sanity check, we observed an enrichment of CTCF occupancy as measured by ChIP-seq in genomic bins with high CTCF motif score (Supplementary Fig. 1c), and an enrichment of chromatin interactions as measured by Hi-C in co-accessible genomic bins for datasets with greater variability (Supplementary Fig. 1d). These results suggest that our input tracks provide valuable information for predicting chromatin contacts that can be harnessed by ChromaFold when trained across sufficiently diverse training cell types.

https://doi.org/10.1038/s41467-024-53628-0

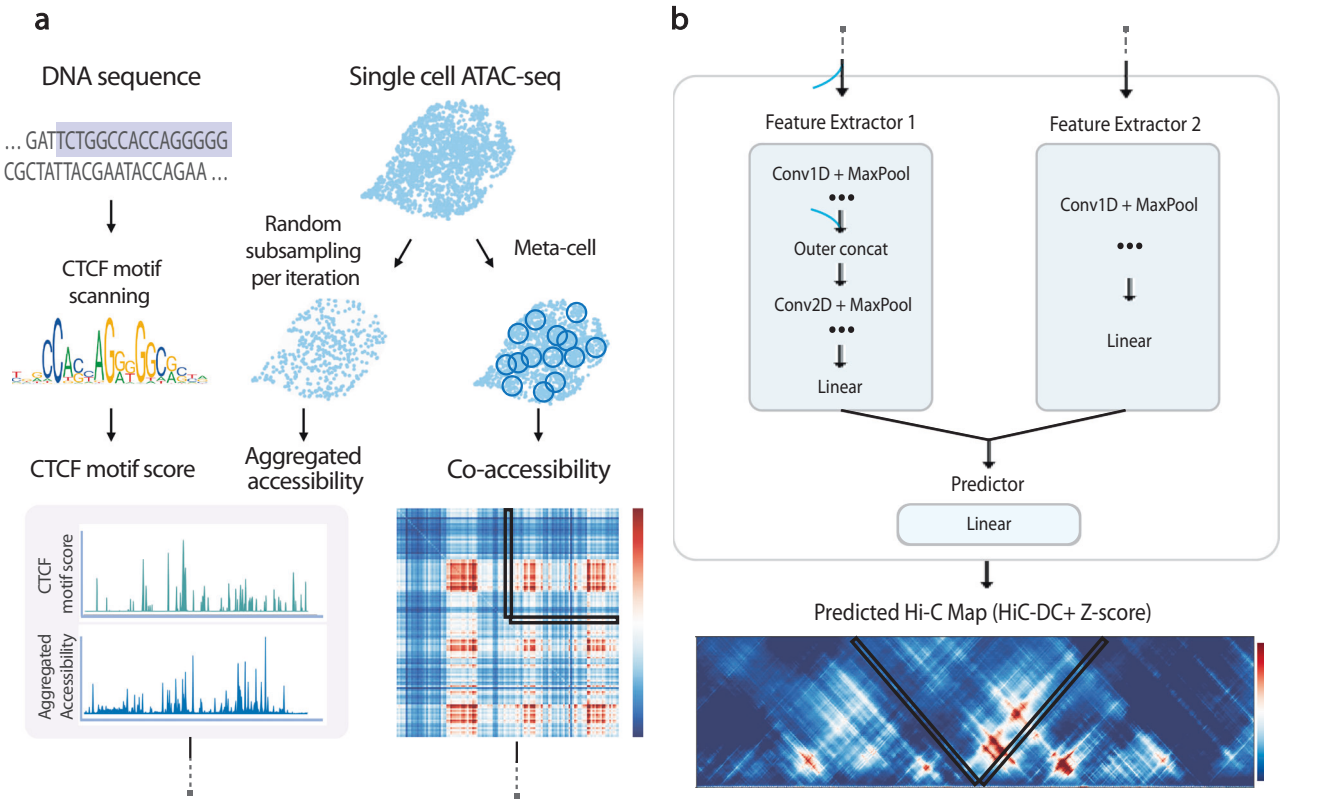


Fig. 1 | ChromaFold predicts the 3D contact map from scATAC-seq alone. ChromaFold is a deep-learning model that enables the prediction of 3D contact maps solely from scATAC-seq data, using pseudobulk chromatin accessibility and co-accessibility from scATAC-seq as well as predicted CTCF motif tracks as input features. **a** Schematic of the ChromaFold input data processing framework. **b** ChromaFold model architecture. The model consists of two feature extractors: feature extractor 1 for the aggregated accessibility and CTCF motif score tracks with a bin size of 50 bp, and feature extractor 2 for the co-accessibility extracted from a V-stripe region with a bin size of 500 bp. The feature extractors produce a latent representation of the 4 Mb genomic region. The Z-score predictor then takes this latent representation and predicts the chromatin interactions between the center genomic tile and its neighboring bins within a 2 Mb distance, annotated by the V-shaped black box. Each genomic tile is 10Kb in length.

We trained ChromaFold on three human cell types (IMR-90, GM12878, and HUVEC) to improve model generalizability to novel test cell types. Fifteen chromosomes were used for training, two for validation, and four were held out for testing and evaluating model performance. We held out three other human cell types (K562, hESC, and activated CD4 + T cells) to test how well ChromaFold can generalize to new cell types. The full contact map was obtained by combining the V-stripe predictions along the chromosome (Methods). To evaluate ChromaFold's performance, we assessed both the chromosome-wide contact map and significant interaction prediction (based on HiC-DC+ top-scoring interactions) on held-out chromosomes for both training and held-out cell types (Fig. 2a, b). ChromaFold achieved an average distance-stratified Pearson correlation of 0.55–0.60 and 0.45–0.47 and an average area under the ROC curve (AUROC) of 0.84–0.85 and 0.77–0.79 in training and held-out cell types, respectively. These results demonstrate ChromaFold's ability to effectively predict the 3D contact map in unseen data and capture significant interactions.

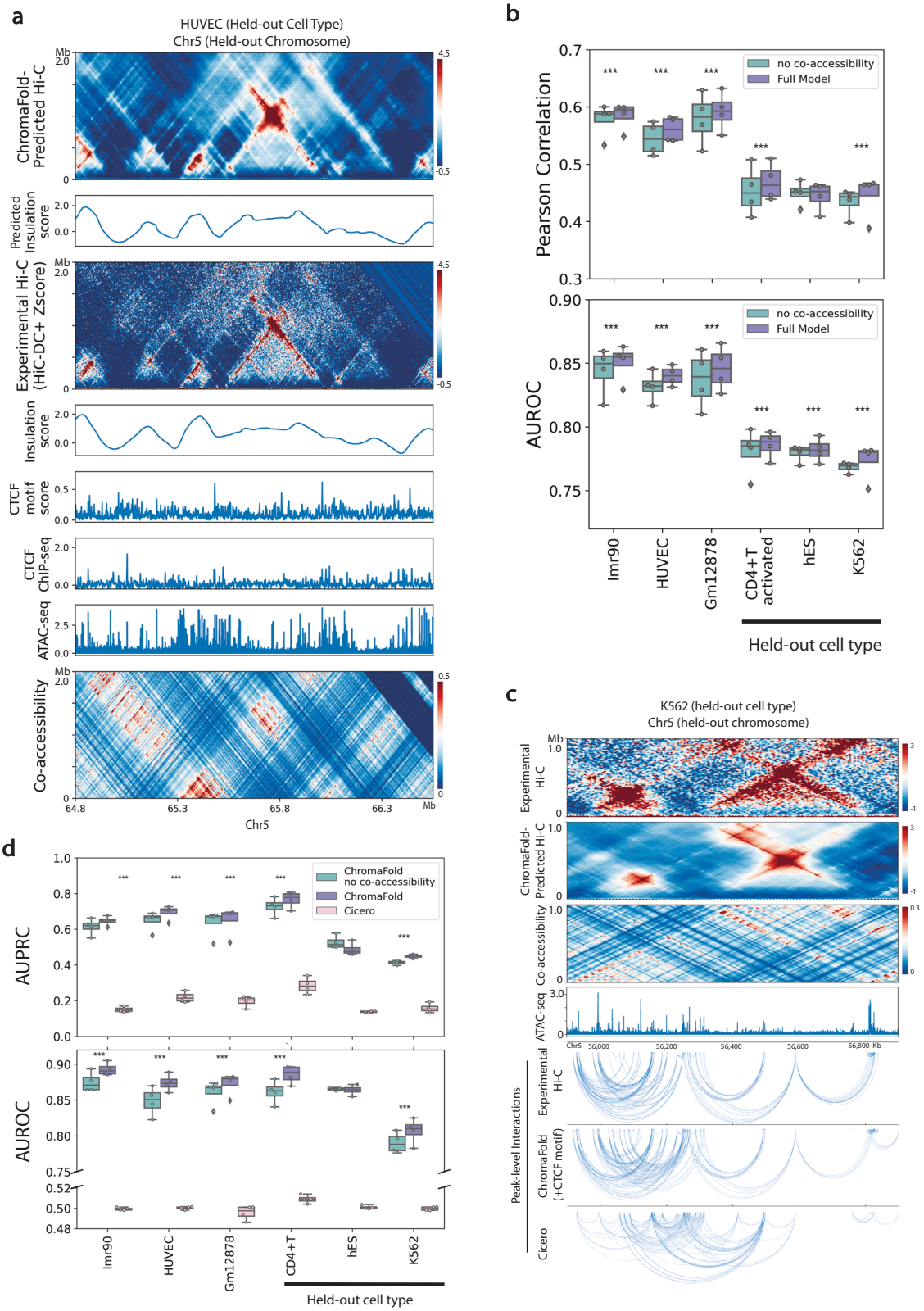
Our choice of 10 kb resolution for prediction of the contact map matches the typical resolution of current Hi-C datasets. However, finer-grained resolution is feasible given suitable Hi-C training targets. In particular, we confirmed that a variant of the ChromaFold model that predicts at 5 kb resolution achieved comparable performance to the 10 kb model on held-out chromosomes when trained on GM12878, the highest-quality training Hi-C dataset (Supplementary Fig. 2a). We further confirmed that ChromaFold produced robust predictions over biological replicate scATAC-seq inputs and substantially different contact maps between cell types (Supplementary Fig. 2b). Finally, we found that only mild decreases in prediction accuracy were incurred by sampling down to about 3000 test cells (Supplementary Fig. 2c), again confirming the robustness of the model.

Co-accessibility and CTCF information improve contact map and peak-level interaction prediction

A key goal of ChromaFold is to predict chromatin interactions that connect regulatory elements to their target genes. To this end, we examined the interactions between accessible peaks by associating ATAC-seq peaks with the overlapping genomic bin and calling peak-level interactions based on the experimental/predicted bin-level contact map (Fig. 2c, Methods). On held-out chromosomes, ChromaFold achieves an average area under the precision-recall curve (AUPRC) of 0.65–0.7 and 0.45–0.75 and an average AUROC of 0.87–0.89 and 0.81–0.89 in training and testing cell types, respectively (Fig. 2d). It should be noted that the diminished performance in K562 is likely attributable to the inferior quality of the Hi-C contact map used for evaluation.

We also compared ChromaFold against Cicero, an unsupervised model that first introduced the idea of using co-accessibility to infer chromatin interactions between accessible peaks[18]. Cicero identifies co-accessible pairs of genomic regions based on their correlation in accessibility across metacells, then uses a graphical lasso regularization to predict a sparser contact map. While peaks with high Cicero co-accessibility are indeed enriched for chromatin interactions compared to peaks with co-accessibility <0, the unsupervised nature of Cicero limits the accuracy of the model, resulting in low precision and recall (Fig. 2c, d). Spurious interaction calls are frequently made, since pairs of genomic regions can be correlated in accessibility without

https://doi.org/10.1038/s41467-024-53628-0



representing true 3D interactions (Fig. 2c). On the other hand, we also observed numerous examples where interacting regions are uncorrelated across metacells, leading to false negative predictions (Supplementary Fig. 4c, d). Additionally, Cicero does not take into account the pseudobulk accessibility profile of peaks and relies solely on correlation structures over metacells, which are heavily influenced by the level of variability in the scATAC-seq dataset (Supplementary Fig. 1d). Nevertheless, we did observe a significant improvement in both 3D contact map and peak-level interaction prediction when we incorporated co-accessibility as an input to ChromaFold (Fig. 2b, d), suggesting that the supervised model can extract useful information from the co-accessibility signal.

While ChromaFold yields peak-level interactions that include regulatory interactions, we caution that neither ground truth nor

Fig. 2 | Co-accessibility information improves contact map prediction in new cell types. a Visualization of real vs. ChromaFold-predicted Hi-C contact map, insulation scores, epigenetic tracks, and co-accessibility on held-out chromosome 5 in HUVEC. **b** Quantitative evaluation of Hi-C map prediction performance by ChromaFold, with and without the co-accessibility input, across training and held-out human cell types/tissues. Box plots show (top) the averaged distance-stratified Pearson correlation for each of $n = 4$ held-out chromosomes between the experimental and predicted contact map and (bottom) the averaged distance-stratified AUROC for each held-out chromosome of significant interactions (top 10% in Z-score). Performance comparisons were assessed by one-sided paired *t*-tests on the distance-stratified Pearson correlation across four test chromosomes from 10 Kb to 2 Mb incrementing by 10 Kb, consisting of $n = 796$ pairs. The *p* value for the Pearson correlation of the full model vs. no co-accessibility model from left to right is $<10^{-16}$ for IMR-90, $<10^{-16}$ for HUVEC, $<10^{-16}$ for GM12878, $<10^{-16}$ for CD4+ activated T cells, 0.999 for hESC and $<10^{-16}$ for K562 (top); the *p* value for the AUROC is $<10^{-16}$ for IMR-90, $<10^{-16}$ for HUVEC, $<10^{-16}$ for GM12878, $<10^{-16}$ for CD4+ activated T cells, 1.95×10^{-7} for hESC and $<10^{-16}$ for K562 (bottom); legend *: <0.05, **: <0.01, ***: <0.001. **c** Visualization of ChromaFold-predicted Hi-C contact map and significant peak-level interactions and Cicero-predicted peak-level interactions in held-out cell type K562 on held-out chromosome 5. **d** Quantitative evaluation of significant peak-level prediction performance by ChromaFold and Cicero. Box plots show the AUPRC (top) and AUROC (bottom) of significant peak-level interaction prediction for each of $n = 4$ held-out chromosomes. Performance comparisons were assessed by one-sided paired *t*-tests on the distance-stratified AUROC and AUPRC across four test chromosomes from 10 to 500 Kb incrementing by 10 Kb, consisting of $n = 196$ pairs. The *p* value for the AUPRC of ChromaFold vs. ChromaFold no co-accessibility from left-to-right is $<10^{-16}$ for IMR-90, $<10^{-16}$ for HUVEC, $<3.69 \times 10^{-5}$ for GM12878, $<10^{-16}$ for CD4+ T cells, 0.782 for hESC and $<1.35 \times 10^{-9}$ for K562 (top). The *p* value for the AUROC is $<10^{-16}$ for IMR-90, $<10^{-16}$ for HUVEC, $<10^{-16}$ for GM12878, $<10^{-16}$ for CD4+ T cells, 3.41×10^{-4} for hESC and 3.20×10^{-7} for K562 (bottom). The *p* values for both ChromaFold models vs. Cicero are $<10^{-16}$. In **b**, **d**, boxes show the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points greater or less than 1.5 times the inter-quartile range from the first or third quartile respectively. Source data are provided as a Source Data file.

predicted Hi-C contact maps alone are sufficient to infer functional enhancer-promoter interactions as validated by assays such as CRISPRi-FlowFISH[33]. In particular, recent work on the activity-by-contact model[34] and the supervised ENCODE-E2G model[33] suggests that H3K27ac data—in addition to chromatin accessibility and 3D interactions—is required to accurately predict functional enhancer-gene links. For example, examining both ground truth and ChromaFold-predicted promoter-anchored 3D interactions at the *MYC* locus in K562 cells, we find that there is reasonable concordance between high-scoring true and predicted interactions and that both recover some of the CRISPRi-FlowFISH-validated *MYC* enhancers (Supplementary Fig. 3a). However, many ground truth promoter-anchored Hi-C interactions do not validate as significant functional enhancers by FlowFISH. Looking more generally at chromosome-wide results (chr8), most FlowFISH-validated enhancers are close (<50 kb) to the TSS, and despite good concordance between true and predicted HiC-DC Z-scores across candidate enhancer-promoter interactions, the 3D interaction strength did not discriminate between significant and insignificant FlowFISH open chromatin regions (Supplementary Fig. 3b-d). Therefore, ChromaFold accurately predicts 3D interactions between chromatin-accessible regions but does not directly infer their regulatory activity.

We next compared ChromaFold's performance when using different types of CTCF information. A qualitative examination of the predicted contact maps in hESC revealed that CTCF information—either predicted binding tracks via motif scores or occupancy from ChIP-seq—is crucial for accurate prediction of the contact map (Supplementary Fig. 4a). A quantitative analysis of the predicted Hi-C maps and peak-level interactions confirmed this observation, as there was a significant decline in performance when ChromaFold operated without any CTCF information across all tested cell types. The most pronounced performance degradation occurred in hESC, which suggests a potential differential mapping between accessibility, CTCF binding, and chromatin interactions in this cell type. As expected, in the majority of cell types examined, ChromaFold performed optimally when it utilized cell-type-specific CTCF ChIP-seq data in the majority of cell types examined. It should be noted, however, that supplying ChromaFold with predicted CTCF motif information alone was sufficient to significantly enhance its accuracy in predicting both the contact map and significant interactions (Supplementary Fig. 4b, c).

ChromaFold is able to predict 3D interactions that are not associated with CTCF binding, although performance metrics do differ on interactions that are occupied by CTCF at both anchors, one anchor, or neither anchor (Supplementary Fig. 5a). In particular, the vast majority of interaction bins in the contact matrix have no CTCF binding at either anchor, and ROC performance is strongest on this class of candidate interactions, while precision-recall is weakest due to strong negative class bias. Meanwhile, ROC performance on CTCF-associated candidate interactions is poorer, but precision-recall is much stronger. Interestingly, when we modified the model to use both forward and reverse motif tracks in order to capture the known orientation bias of CTCF-mediated loops, we did not find consistent improvement across test cell types (Supplementary Fig. 5b). Potentially, the topologically associating domain structure associated with convergent CTCF motifs is already well captured through accessibility and co-accessibility.

ChromaFold competes with state-of-the-art models that use multiple bulk epigenomic tracks

We next benchmarked ChromaFold against C.Origami, a recent model that uses bulk ATAC-seq, DNA sequence, and CTCF ChIP-seq as inputs to predict the 3D contact map[16]. To ensure a fair comparison, we re-trained ChromaFold and C.Origami on the same cell type, IMR-90, with HiC-DC + Z-score normalized Hi-C contact maps as the target and used the same chromosomes for training, validation (Chr10), and testing (Chr15). While ChromaFold and C.Origami achieved similar performance on the held-out chromosome in the training cell type (Supplementary Fig. 6a–c), ChromaFold models outperformed C.Origami on a new cell type, GM12878 (Fig. 3). Further expanding our comparison to include two additional cell types used in C.Origami's cross-cell-type prediction evaluation, K562, and hESC, we found that the ChromaFold model consistently surpassed C.Origami across all metrics when CTCF ChIP-seq data was provided, and achieved comparable performance when using CTCF motif information. Given that HiC-DC+ normalization employs negative binomial regression to control for genomic distance as well as other covariates such as GC content and mappability to identify statistically significant interactions, we propose that this normalization makes contact map prediction more challenging than other normalization methods, such as ICE[35]. Consequently, more heavily parameterized models, like C.Origami, may be more susceptible to overfitting, thereby compromising generalizability.

For completeness, we repeated the comparison of ChromaFold with C.Origami when both models were trained and evaluated against ICE-normalized Hi-C target contact maps, maintaining the same training, validation, and test chromosomes as above. With ICE normalization, TAD structures dominate the target Hi-C contact map, with little visible structure within TADs (Supplementary Fig. 7a–d). Relative to this smoother normalization, Pearson correlation with the target was higher for both ChromaFold and C.Origami, and we found that ChromaFold with CTCF ChIP-seq achieved the same performance as C.Origami across most test cell types (Supplementary Fig. 7e–h). Potentially, CTCF ChIP-seq is needed for optimal prediction with a TAD-dominated normalization.

https://doi.org/10.1038/s41467-024-53628-0

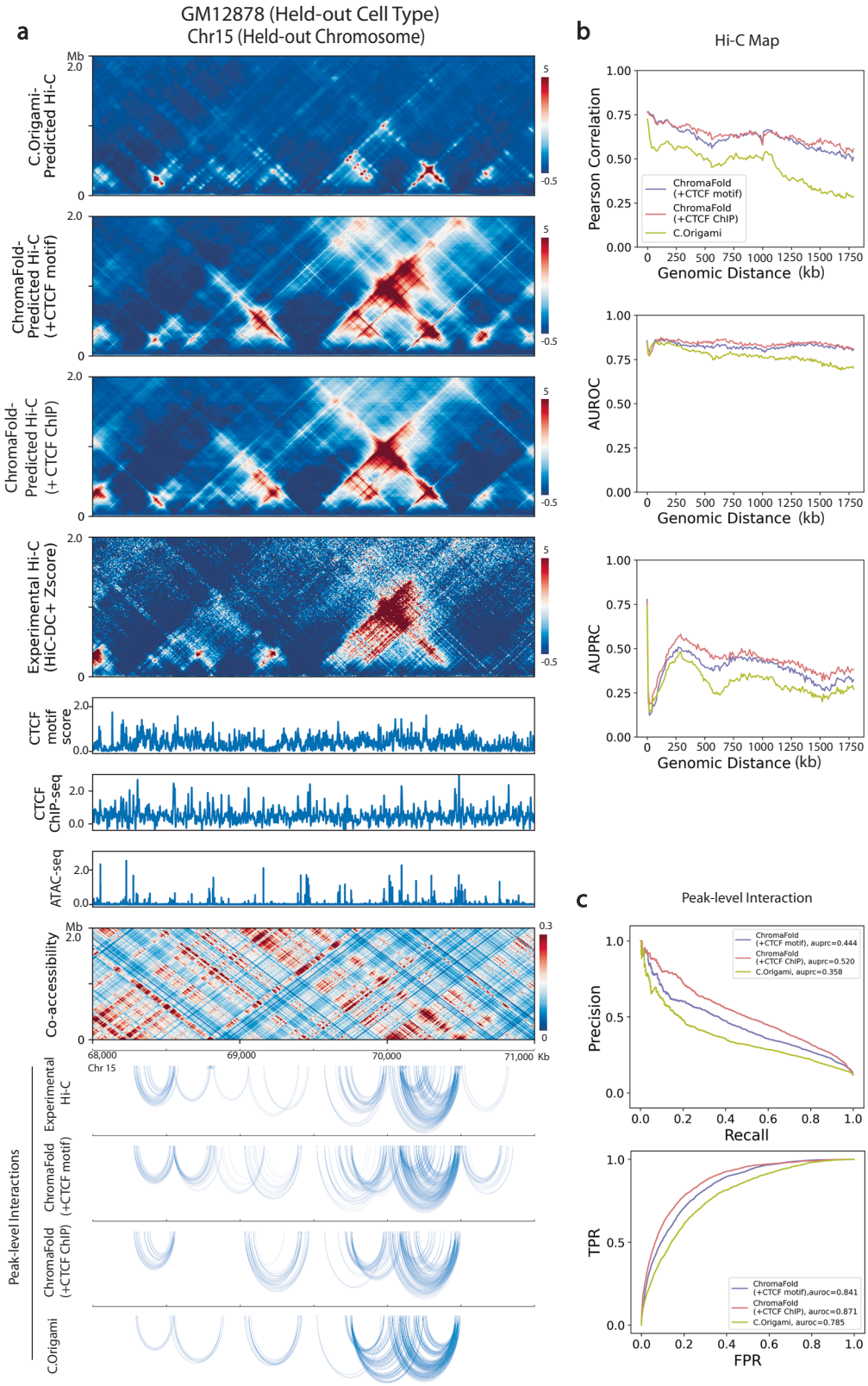


Fig. 3 | ChromaFold achieves state-of-the-art performance for predicting significant Hi-C interactions in new cell types. C.Origami and ChromaFold were trained using the same training/test chromosomes on IMR-90 to predict contact maps normalized by HiC-DC+ Z-score. **a** Visualization of C.Origami and ChromaFold-predicted Hi-C contact maps and peak-level interactions in held-out cell type GM12878. **b** Line plots show distance-stratified (top) Pearson correlation between the experimental and predicted contact map, (middle) AUROC and (bottom) AUPRC of significant interactions (top 10% in Z-score) for ChromaFold and C.Origami on held-out chromosome 15. **c** Line plots show (top) PR curves and (bottom) ROC curves for peak-level interaction prediction on held-out chromosome 15. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-53628-0

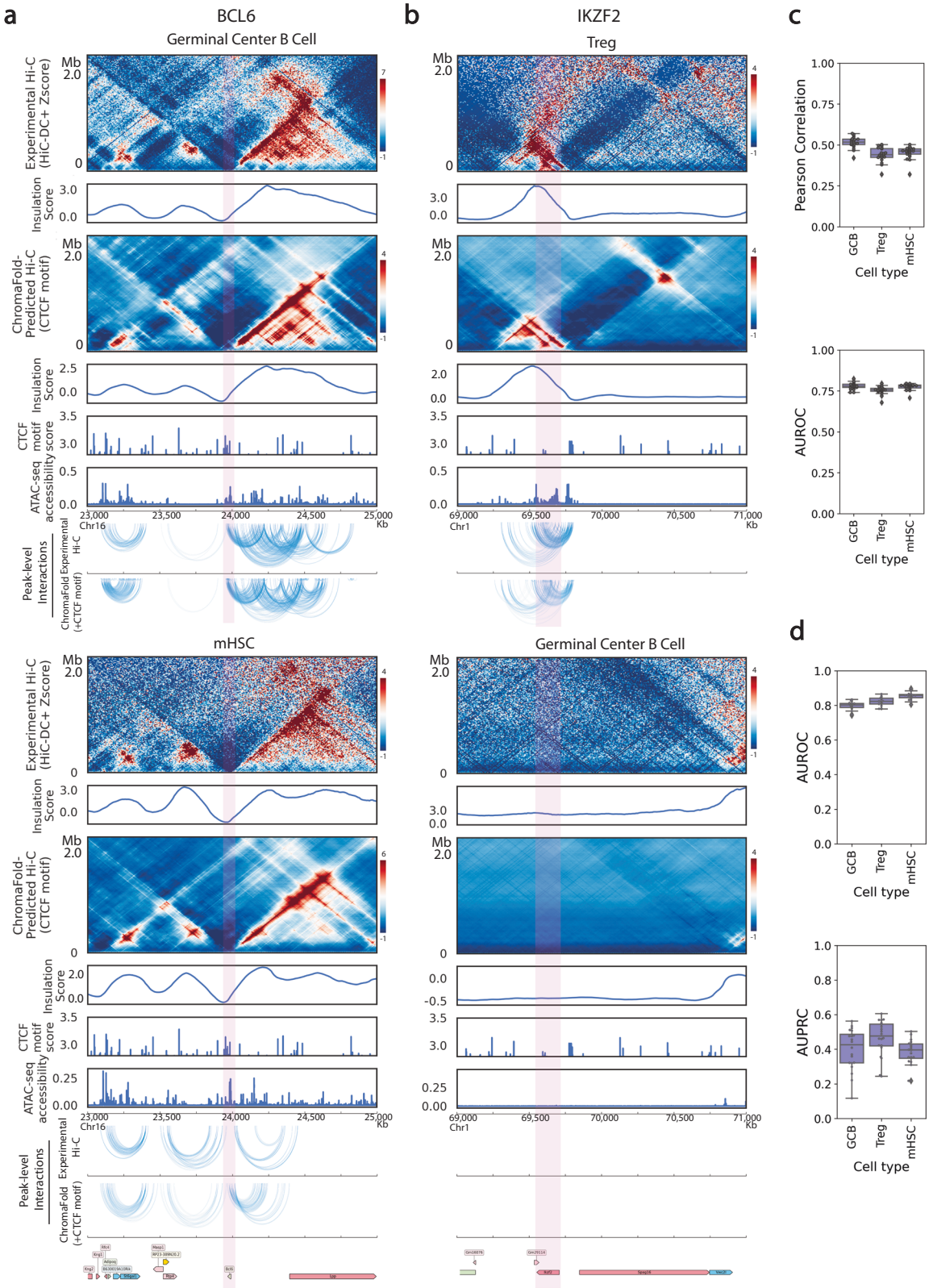


Fig. 4 | ChromaFold accurately generalizes across cell types and species. **a**, **b** Comparison of experimental vs. ChromaFold-predicted Hi-C contact map and peak-level interactions at different loci in the mouse genome across different murine cell types: the *Bcl6* gene locus in mouse germinal center B cells (**a**, top) and in mHSC (**a**, bottom) and the *Ikzf2* gene locus in regulatory T cells (**b**, top) and germinal center B cells (**b**, bottom). **c** Box plots show (top) the averaged distance-stratified Pearson correlation between the experimental and predicted contact map and AUROC of predicted significant interactions (bottom; top 10% in Z-score) from 10 kb to 2 Mb for $n = 20$ chromosomes. **d** Box plots show the distance-stratified AUROC(top) and AUPR (bottom) of significant peak-level interaction prediction from 10 to 500 kb for $n = 20$ chromosomes across mouse cell types. In **c**, **d**, boxes show the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points greater or less than 1.5 times the inter-quartile range from the first or third quartile respectively. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-53628-0

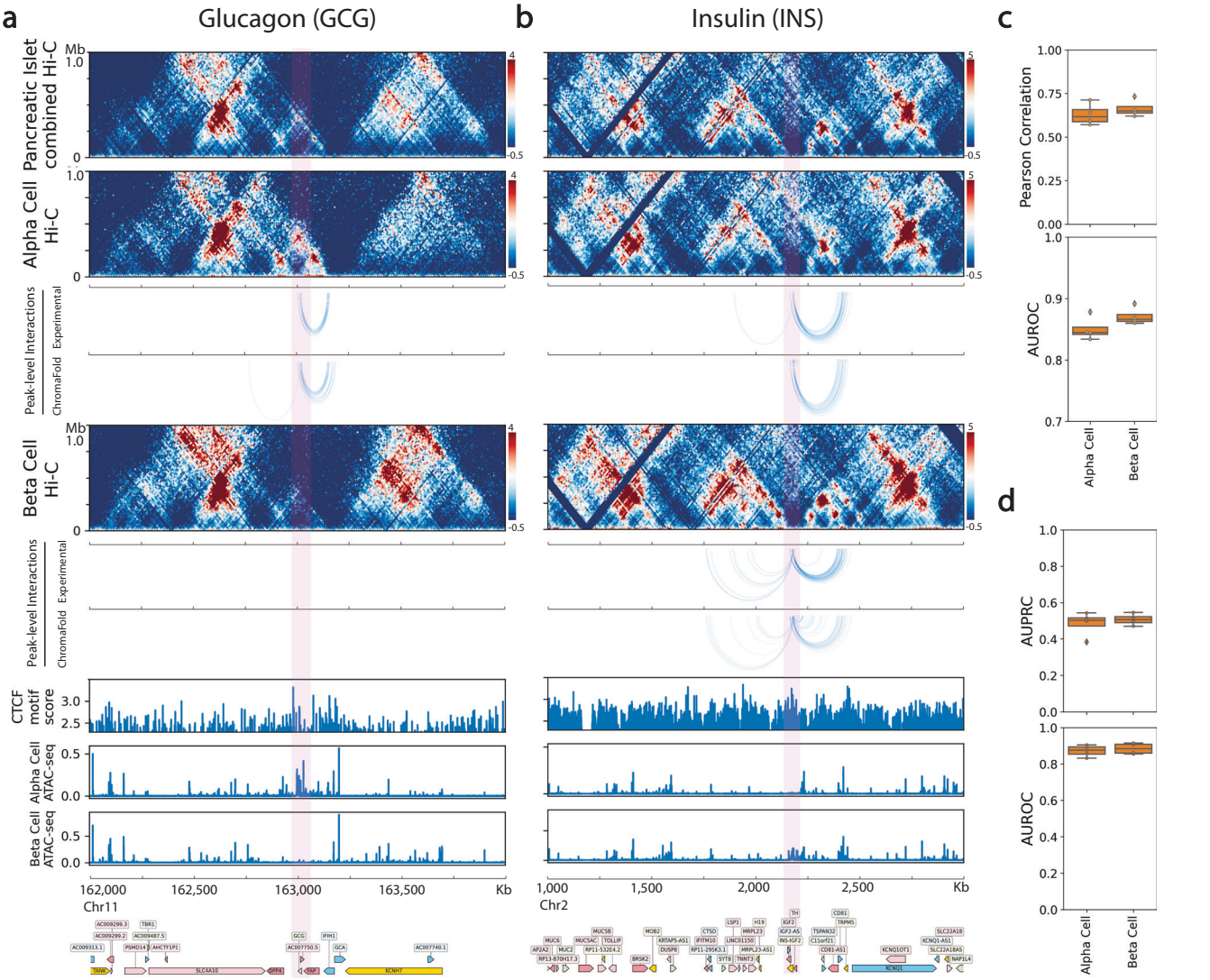


Fig. 5 | ChromaFold enables deconvolution of Hi-C interactions in pancreatic islet cells. a, **b** Visualization of peak-level interactions derived from experimental Hi-C data and ChromaFold-predicted Hi-C map in alpha cells and beta cells near the TSS of (**a**) glucagon (*GCG*) and (**b**) insulin (*INS*). **c** Box plots show (top) the averaged distance-stratified Pearson correlation and AUROC of significant interactions (top 10% in Z-score), for $n = 4$ test chromosomes from 10 Kb to 2 Mb in alpha and beta cells. **d** Box plots show the AUPRC (top) and AUROC (bottom) of significant peak-level interaction prediction for $n = 4$ test chromosomes from 10 Kb to 2 Mb in alpha and beta cells. In **c**, **d** boxes show the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points greater or less than 1.5 times the inter-quartile range from the first or third quartile respectively.

Finally, we compared against another deep-learning method we developed, Epiphany[17], which uses multiple bulk epigenomic tracks with a bi-LSTM model to predict the Hi-C contact map. Here we trained Epiphany to predict the HiC-DC Z-score normalized contact map in GM12878 using the full five-track model (ATAC, H3K27ac, H3K4me3, H3K27me3, and CTCF) and a two-track model (ATAC, CTCF) and evaluated on the same held-out chromosomes as ChromaFold (Supplementary Fig. 7i). ChromaFold achieved comparable performance to the two-track Epiphany predictions, with the five-track Epiphany model giving slightly higher Pearson correlation. Interestingly, when we benchmarked ChromaFold with Epiphany in K562, a held-out cell type, ChromaFold outperformed the 5-track Epiphany model while achieving comparable performance to the 2-track Epiphany model. This poorer generalization for the five-track Epiphany model may be due to technical differences across the epigenomic tracks between cell types, or potentially due to overfitting to training data. Overall, we can conclude that ChromaFold using only scATAC-seq achieves state-of-the-art performance compared to models that use bulk chromatin accessibility and CTCF ChIP-seq.

ChromaFold can generalize across species and make cell type-specific predictions

Having shown that ChromaFold can generalize to new human cell types, we proceeded to test whether the model could generalize to a different mammalian genome, since we expect evolutionarily conserved rules governing the mapping between chromatin accessibility and 3D interactions. We therefore directly applied ChromaFold, trained on three human cell types/tissues, to mouse germinal center B cells (GCBs), hematopoietic stem cells (HSCs), and regulatory T (Treg) cells, and evaluated both the predicted contact maps and peak-level interactions. We observed performance comparable to that in human cell types, despite evaluating in a different genome and against lower quality ground-truth Hi-C contact maps in mouse cell types (Fig. 4c, d and Supplementary Fig. 8a, b). Similar to observations in human test cell types, ChromaFold predictions in mouse are compromised when we ablate the co-accessibility or CTCF motif score input (Supplementary Fig. 8a). Notably, we achieve good performance on GCBs with only ~1500 cells in the scATAC-seq dataset, whereas the smallest training cell type contains ~3300 cells. These findings suggest that

ChromaFold, trained on human data, can generalize to mouse and potentially to other mammalian genomes.

Next, we sought to confirm ChromaFold's ability to make cell-type-specific predictions at loci of interest. Although the predicted CTCF motif score is not cell-type-specific, we expected that the accessibility inputs would confer cell-type-specificity. To illustrate this, we zoomed in on two genes of interest in these cell types: B cell lymphoma 6 (*Bcl6*) and Helios (*Ikzf2*). The *Bcl6* gene encodes a transcription factor that is critical for GCB development[36,37]. Upon comparing the 3D contact maps at the *Bcl6* locus in GCBs and in HSCs, we observed various conformation changes upstream of the *Bcl6* gene, specifically around the region Chr16:24,250,000–24,600,000 bp. These differences were accurately captured by ChromaFold-predicted contact maps and insulation scores (Fig. 4a). The *Ikzf2* gene is a transcription factor that is essential for the development and function of thymically-derived Treg cells[38,39]. ChromaFold can predict the presence of chromatin interactions or lack thereof near the *Ikzf2* locus in Treg cells and GCBs, respectively (Fig. 4b). Taken together, we conclude that ChromaFold is able to leverage cell-type-specific single-cell chromatin accessibility data and make cell-type-specific contact map predictions.

ChromaFold can deconvolve chromatin interactions in complex tissue

The ability to study chromatin interactions in fine-grained cell populations can help dissect cell-type-specific gene regulatory programs and contribute to elucidating the pathogenesis of complex genetic diseases. However, the application of experimental techniques such as Hi-C is challenging in rare cell populations due to the difficulty of acquiring sufficient cells for the assay. Although single-cell Hi-C sequencing has made significant advances, the associated experiments remain difficult and expensive, and the sparse contact maps produced are typically analyzed at coarse resolution (100 kb–1 Mb bins)[11,40].

We therefore sought to use ChromaFold to deconvolve chromatin interactions in complex tissues. In scenarios where we possess scATAC-seq and a bulk Hi-C contact map of a tissue or cell population with diverse cell types/states, we decided to fine-tune the pretrained ChromaFold model using input and output data from the mixed population to adapt to the dataset. We then applied the fine-tuned model to individual cell populations (clusters) to predict cluster-specific contact maps and thus achieve bulk Hi-C deconvolution.

To evaluate this approach, we applied ChromaFold to deconvolve chromatin interactions in alpha and beta cells within pancreatic islet cell populations using scATAC and bulk Hi-C from non-diabetic islet donors[21] (Supplementary Fig. 9a). The predictions were validated against an independent dataset containing Hi-C in sorted alpha and beta cells[41]. Our results show that ChromaFold can accurately deconvolve chromatin interactions in the held-out chromosomes (Fig. 5 and Supplementary Fig. 9b). Further, we visualized the predicted interactions at alpha and beta cell marker genes glucagon (*GCG*) and insulin (*INS*). Notably, we predicted a large number of contacts between the *GCG* gene and distal chromatin regions in the alpha cells but not the beta cells, consistent with ground truth data in sorted populations (Fig. 5a). On the other hand, we predicted an increased number of contacts between the *INS* gene and both the upstream and downstream chromatin regions in beta cells compared to alpha cells, again matching ground truth contact maps (Fig. 5b).

Discussion

Our study demonstrates the utility and potential of ChromaFold for predicting chromatin contacts and mapping putative regulatory elements to their target genes. ChromaFold's performance, as validated across several metrics and cell types, surpasses previous models such as Cicero and C.Origami, confirming its robustness and versatility. We also found that ChromaFold accurately generalized across species by making cell-type-specific predictions at important loci in diverse mouse cell types from scATAC-seq alone. These findings underscore the shared rules governing the mapping from chromatin accessibility to 3D interaction in mammalian genomes. Furthermore, the ability of ChromaFold to operate on scATAC-seq datasets with ~1000 cells and the application of ChromaFold for deconvolving bulk contact maps in complex tissues enables the study of chromatin interactions in fine-grained cell populations, providing a novel window into cell-type-specific gene regulatory programs and the dysregulation of these programs in complex genetic diseases.

ChromaFold enables the inference of peak-level interactions between accessible elements, which include regulatory interactions such as enhancer-promoter (E-P) interactions. However, we caution that the presence of a promoter-anchored peak-level 3D interaction is not sufficient to guarantee a functional E-P interaction. Indeed, models to predict functional E-P links generally use the active histone mark H3K27ac as well as accessibility and 3D connectivity[33]. In reanalyzing published CRISPRi-FlowFISH data, we found that most of the validated E-P interactions were promoter-proximal, and that 3D interaction strength alone did not discriminate functional from non-functional candidate E-P interactions (Supplementary Fig. 5). Therefore, ChromaFold's peak-level interactions provide useful cell-type-specific predictions about the connectivity of gene promoters and accessible elements but do not guarantee the regulatory activity of these interactions.

Our analyses point to several still-unresolved questions for the prediction of the 3D contact map: what epigenomic data is most useful for achieving good generalization in new cell types, and what information is captured by DNA sequence models beyond CTCF motif information? Ablation experiments with ChromaFold demonstrated that co-accessibility from scATAC-seq gave a significant performance improvement over pseudobulk accessibility alone. While a number of models, including EPCOT[42] and C.Origami, have relied on bulk ATAC-seq as an input signal to help generalization across cell types, our results suggest that covariation in scATAC-seq provides additional information that can be leveraged for contact map prediction. ChromaFold prediction accuracy improved when cell-type-specific CTCF ChIP-seq data was provided as an input. However, using predicted CTCF motif tracks in place of CTCF ChIP-seq data performed comparably to C.Origami, a state-of-the-art model that uses both a full DNA sequence model as well as ATAC-seq and CTCF ChIP-seq. This result suggests that an improved method for predicting cell-type-specific CTCF ChIP-seq occupancy—in place of the fixed CTCF motif tracks currently used as input—could increase ChromaFold's accuracy. Interestingly, including CTCF orientation information does not significantly or consistently improve the model's prediction (Supplementary Fig. 5b). We hypothesize that the signal from scATAC-seq co-accessibility, together with non-oriented CTCF motif data, may already capture sufficient information about CTCF-mediated looping, and therefore that motif orientation does not provide addition predictive value. Furthermore, we note that the performance advantage or disadvantage of adding CTCF motif information depends on (i) the overall similarity of the test cell type Hi-C to that of the training cell types and (ii) the quality/resolution of the test cell type.

However, it remains unclear what biological information is captured by introducing a full deep sequence model for contact map prediction, or whether overfitting to spurious sequence signals may be masking relevant information beyond CTCF-associated binding motifs. These questions may be addressed in the coming years through advances in deep-learning model interpretation and through ongoing modeling efforts in regulatory genomics. For now, ChromaFold provides a highly favorable trade-off between model complexity,

performance, and ease of use, through a lightweight deep-learning model that achieves state-of-the-art chromatin map prediction from scATAC-seq alone.

Methods

Ethics statement

Generation of human ESC scATAC-seq data: Experiments were conducted per National Institute of Health (NIH) guidelines and approved by the Tri-SCI Embryonic Stem Cell Research Oversight Committee. Generation of mouse regulatory T cell Hi-C data: Animals were housed at the Memorial Sloan Kettering Cancer Center (MSKCC) animal facility under specific pathogen-free (SPF) conditions on a 12-h light/dark cycle. All studies were performed under protocol 08–10-023, approved by the MSKCC Institutional Animal Care and Use Committee. Generation of mouse germinal center B cell scATAC-seq data: The experimental procedures involving animals were executed in stringent accordance with the institutional guidelines delineated by Weill Cornell Medicine, as per the Guide for the Care and Use of Laboratory Animals, and standards established by the Association for Assessment and Accreditation of Laboratory Animal Care International. The Research Animal Resource Center, the Institutional Animal Care and Use Committee of Weill Cornell Medicine and Cornell Institutional Animal Care and Use Committee, having vetted all procedures, duly approved the entirety of the study involving mice under protocols #2011-0031 and #2017-0035. Generation of mouse hematopoietic stem cell scATAC-seq: All animal studies were performed on animal protocol #11-10-025 approved by the Institutional Animal Care and Use Committee (IACUC) at Memorial Sloan Kettering Cancer Center.

Preprocessing of Hi-C and Micro-C data

We used nine human and three mouse datasets (Supp. Table 1). For datasets provided in this study and those where a processed.hic file is not available online, Hi-C FASTQ files were aligned to hg38, hg19, or mm10 genomes, and reads that are duplicates or invalid ligation products were filtered out using the HiC-Pro[43] pipeline (v3.1.0) with default settings. Hi-C contact matrices were binned at 10 kb resolution and normalized using the following approaches. ICE-normalized contact maps were calculated using the HiCExplorer[44] package. The counts were log2 normalized using a pseudocount of 1. Z-score normalization was calculated by the HiC-DC+[26] package. Specifically, HiC-DC+ models observed raw counts for interaction bins using negative binomial regression to estimate the expected count based on genomic distance, GC content, mappability, and effective bin size based on RE sites in the corresponding pair of genomic intervals.

Preprocessing of scATAC-seq data

For datasets provided in this study and those where the processed scATAC-seq fragment file was not available online, scATAC-seq FASTQ files were aligned to hg38, hg19, or mm10 and counted by Cell Ranger ATAC v1.2.0[45] with default parameters. Arrow files were created from the scATAC-seq fragments using ArchR v1.0.1[46]. Specifically, we binarized sparse accessibility matrices binned into 500 bp tiles across the genome. Cells with fewer than 1000 fragments and TSS <4 were filtered out. Latent Semantic Indexing (LSI) was performed on the 25,000 top variable tiles identified after two iterations of "IterativeLSI" by ArchR. Tiles from non-standard chromosomes, chrM and chrY, were not included. Cells were clustered (method=Seurat, k.param = 30, resolution = 1) and visualized with UMAP[47] (nNeighbors = 30) using 30 LSI components. For datasets with multiple cell types, we annotated and extracted the cell type of interest by computing the mean gene score of marker genes per cluster. This was cross-checked with cell type annotations provided by the original sources, if available.

Peak calling

For peak calling of the scATAC-seq data, filtered fragments for cells in each dataset/cell population were aggregated and used as input to the MACS2[48] peak caller (parameters -f BED, -g 2.7e9, -no-model, -shift −75, -extsize 150, -q 0.05). Peaks were filtered using an IDR[49] cutoff of 0.05. Peaks within 500 bp of each other were merged. A peak-by-cell count matrix was then created by ArchR.

Bulk ATAC-seq data processing

Bulk ATAC-seq data were obtained from ENCODE[50] in the form of bam files. Bam files from replicates were merged using samtools[51], binned at 1 bp resolution for C.Origami, and RPKM normalized using the bamCoverage function in deepTools[52] to generate bigwig files.

CTCF ChIP-seq and motif score data processing

We obtained the CTCF motif scores from the CTCF R package[27], an AnnotationHub resource that represents genomic coordinates of FIMO-predicted CTCF binding sites for human and mouse genomes. Specifically, CTCF motif scores were generated by scanning for all three JASPAR[28] CTCF PWMs in genomic DNA sequence using FIMO[25]. CTCF ChIP-seq data were obtained from ENCODE in the form of bam files. Bam files from replicates were merged using samtools, binned at 50 bp resolution for ChromaFold and 1 bp resolution for C.Origami, and RPKM normalized using the bamCoverage function in deepTools to generate bigwig files. The log2 fold change from the control ChIP-seq in the corresponding cell types were computed using the bigwigCompare function in deepTools.

ChromaFold input data processing

ChromaFold takes three inputs: pseudobulk chromatin accessibility, co-accessibility profiles across cells, and predicted CTCF motif score/CTCF ChIP-seq. The pseudobulk chromatin accessibility is obtained by aggregating the accessibility profile across single cells in a population binned at 50 bp, library-size normalizing, and log transforming with a pseudocount of 1. The co-accessibility is derived by first generating metacells to combat sparsity in scATAC-seq datasets, then calculating the Jaccard similarity between binarized accessibility profiles across metacells, binned at 500 bp. Metacells are generated using the same algorithm used by Cicero[18]. Specifically, to generate the co-accessibility input corresponding to the V-stripe region, we directly compute the co-accessiblity between the 500 bp genomic bins in the center 20 kb region of the input window with all 500 bp genomic bins flanking the center 10 kb region. The CTCF motif score for each 50 bp bin in the genome is defined as the maximum score assigned to any genomic region that overlaps at least 10 bp with the 50 bp bin.

ChromaFold model architecture

The ChromaFold model consists of two feature extractors and a linear predictor module. The first feature extractor takes the pseudobulk accessibility and the CTCF motif score or ChIP-seq signal as two channels. This feature extractor consists of fifteen 1D convolutional layers followed by batch normalization and ReLU activation. Next, we perform outer-concatenation where the model transforms the resulting $L \times C$ matrix, where L is the length of the output vector and C is the number of channels, into a $L \times L \times 2C$ by performing point-wise concatenation of the output features. This operation allows the information from pairs of genomic bins to be joined together. We implement a skip connection with the input layer by average-pooling the input and transforming it into a 3D tensor via outer concatenation. After concatenation, the data is passed through three 2D convolutional layers followed by a linear layer to consolidate the extracted features, producing a latent representation of the two input tracks.

The second feature extractor takes the co-accessibility data as input. For memory efficiency, we only compute the co-accessibility between the bins in the center 10 kb region with the rest of the bins in

the 4.01 Mb region as input. We use three 1D convolutional layers followed by two residual blocks and three additional 1D convolutional layers. Finally, a linear layer consolidates the extracted features and produces a latent representation of the co-accessibility input. These latent representations of the genomic region are concatenated and passed through a final linear layer to predict the contact between genomic bin t and its neighboring bins within a 2 Mb distance, which corresponds to a V-shaped stripe (V-stripe) in the contact map centered at t.

ChromaFold model training

We trained ChromaFold using data pooled from three cell types, IMR-90, GM12878, and HUVEC. Chromosomes 3 and 15 were used for validation, chromosomes 5, 18, 20, 21 were held out for testing and evaluating model performance, and the rest were used for training. For each V-stripe prediction centered at genomic bin t, the input is the 4.01 Mb region centered at t. During training, we randomly subsampled 500–5000 single cells and 400–1000 metacells from the population per iteration for pseudobulk accessibility and co-accessibility computation, respectively. This data-augmentation step was critical for improving model generalizability to datasets of varying quality and size. We injected additional variation into the input by randomly shifting by −50 or 50 bp. Since neither our input nor output data contain directionality information, we further reduce redundancies in our model by predicting only one side of the V-stripe, and we simply reversed the input to predict the other side (shared model weights). To improve model stability, we used a two-step approach and first train ChromaFold's feature extractor 1 to predict the target contact map by appending a dummy linear predictor at the end. After convergence, we froze the weights for this part of the network while training feature extractor 2 and the final linear module. Genomic regions with low mappability were masked from training based on the total signal for each bin in the contact map. We took the training window to start and end 4 and 5 Mb after the chromosome starting location and before the ending location, respectively, to create buffer regions since ChromaFold requires 4.01 Mb windows as inputs. The prediction target is the HiC-DC+ normalized Z-score, with outlier target values clipped to lie between −16 and 16 to avoid training bias. We optimized the MSE loss using stochastic gradient descent. We trained the model for 30 epochs and implemented early stopping with a patience of 10 epochs, the learning rate of 1e-6 and weight decay 1e-3. The model was trained on a single NVIDIA Tesla V40 GPU for ~5 h when using one training cell type and ~14 h when using three training cell types.

De novo contact map prediction in a new cell type

The ChromaFold model trained on IMR-90, GM12878, and HUVEC can be directly applied to other cell types and species without retraining. To perform de novo contact map prediction, we supplied scATAC-seq data of the new cell type and predicted CTCF motif scores in the corresponding genome to ChromaFold. If CTCF ChIP-seq data was available for the test cell type, we could alternatively use the *ChromaFold + CTCF ChIP-seq* model.

ChromaFold Hi-C contact map prediction

To generate the complete predicted contact map for each chromosome, we first performed inference and predicted the interaction between each genomic bin t and all its neighboring bins within a 2 Mb distance, producing a V-stripe. Since the input region is 4.01 Mb centered at the bin t, we zero-padded the input vectors if they extended beyond the chromosome edges. We combined the predicted V-stripes and averaged the two predictions for each genomic bin. Contact map prediction for one full chromosome took on average ~1.5 min on a standard GPU like NVIDIA Tesla V40.

Distance-stratified correlation

To evaluate the overall performance of genome-wide chromatin contact map prediction, we computed the distance-stratified correlation between the experimental and predicted contact maps. The rationale for distance-stratification is to remove any remaining genomic distance effect and avoid inflating the correlation. Specifically, we computed the Pearson correlation for all interactions with genomic distance d for d from 0 to 2 Mb, for each chromosome. We then used a paired t-test[53] to compare the performance between models. In the boxplot visualizations, each point represents the Pearson correlation averaged across genomic distance, per chromosome.

Topologically associated domain (TAD) annotations

We called TADs at 10 kb resolution using TopDom[54] (v0.0.2) using w = 30 on normalized Hi-C contact maps and predicted contact maps and used the insulation scores to evaluate ChromaFold's ability to predict TAD structures.

Significant interactions

We defined significant interactions at the genomic bin level as interactions with the top 10% HiC-DC + Z-scores per chromosome. For each chromosome and at each genomic distance (incrementing by 10 kb), we used AUROC and AUPRC to evaluate how well significant interactions are captured by ChromaFold's predicted contact map. We used a paired t-test to compare the performance between models. In the boxplot visualizations, each point represents the corresponding metric averaged across genomic distance, per chromosome. To define significant peak-level interactions, we first mapped each peak to the overlapping genomic bin(s) at 10 kb resolution. If a peak overlapped two bins, it was assigned to both. Next, we labeled pairs of peaks as significantly interacting if the corresponding HiC-DC + FDR-corrected p value is less than 0.25. The distance-stratified AUROC and AUPRC were computed in a similar fashion as described above.

Benchmarking against Cicero

We used Cicero to calculate co-accessibility for pairs of peaks. The same metacell groupings used for ChromaFold training/inference were used for running Cicero. We then used Cicero to calculate co-accessibility using a window size of 1 Mb and a distance constraint of 500 kb. We evaluated the performance of peak-level significant interaction prediction using Cicero co-accessibility at various cutoffs and compared that using ChromaFold-predicted contact maps. All evaluations of peak-level significant interactions were distance-constrained to 500 kb for comparison with Cicero.

Benchmarking against C.Origami

To ensure a fair comparison, we re-trained ChromaFold (with CTCF motif score or with CTCF ChIP-seq) and C.Origami on the same cell type, IMR-90, towards HiC-DC+ normalized Hi-C contact maps and used the same chromosomes for training, validation (Chr10) and testing (Chr15) as specified in C.Origami[16]. The training procedure for ChromaFold was the same as described above, and that for C.Origami was the same as described in the original paper. C.Origami converged after training for 45 epochs. After training, we evaluated the performance of both models on the test chromosome in IMR-90, as well as in three held-out cell types GM12878, K562, and hES. For held-out cell types, we used the IMR-90-trained models but used GM12878/K562/hESC inputs to make de novo contact map predictions. For both models, we merged predictions into a chromosome-wide Hi-C contact map and evaluated the following metrics: (1) distance-stratified Pearson correlation, (2) distance-stratified bin-level significant interaction prediction, and (3) peak-level significant interaction prediction.

Deconvolution of chromatin interactions in alpha and beta cells in the pancreatic islet

ChromaFold can be used for deconvoluting chromatin interactions in complex tissues. Using the scATAC-seq and bulk 3D contact map for pancreatic islet cells, we fine-tuned the pretrained ChromaFold model for 1 epoch on the training chromosomes to better adapt the model predictions to the dataset. We then applied the fine-tuned model to alpha and beta cell populations to achieve deconvolution. Specifically, we extracted the alpha and beta cell clusters from the scATAC-seq to use as input to ChromaFold to generate deconvolved contact map predictions. Next, we used the deconvolved contact maps to generate peak-level interaction predictions as described in the section above. We evaluated the deconvolved chromatin interaction predictions using an independent dataset with Hi-C of sorted human alpha and beta cell populations. For peak-level interaction visualization, we restricted to only interactions involving peaks that lie within 10 Kb of the TSS of the highlighted genes. The overall contact map prediction quality was evaluated using distance-stratified Pearson correlation. Significant bin- and peak-level interaction predictions were evaluated using distance-stratified AUROC and AUPRC.

Single-cell ATAC sequencing data collection

Human embryonic stem cells were harvested for single-cell multiome analysis with a targeted collection of ~7000 cells. Nuclei were isolated with Demonstrated Protocol Nuclei Isolation for Single-Cell Multiome ATAC+Gene Expression Sequencing_RevA. 500 K cells underwent lysis in 500 µl lysis buffer in ice for 3 min, then were subjected to the standard protocol for wash and counting. Single-cell Multiome libraries were generated with the 10x Genomics Chromium Next GEM Single-Cell Multiome ATAC + Gene Expression Kit following the manufacturer's guidelines. The libraries were sequenced on the NovaSeq 6000 platform following the manufacturer's guidelines.

To collect scATAC-seq data in mouse hematopoietic stem cells (Lin-Kit+ cells), bone marrow cells were harvested from a total of $n = 3$ C57BL6 wildtype mice and subjected to red blood cell lysis. Bone marrow cells were then incubated with MACS beads (CD117, Miltenyi Biotec, 130-091-224). Then enriched c-Kit+ cells were collected by running AutoMACS (Miltenyi Biotec) according to the manufacturer's instructions. The cells were then stained with a cocktail: Lineage marker (CD3, CD8, Gr1, B220, CD19, and Ter119)- PE-Cy5 (dilution 1:100), cKit-APC-Cy7 (1:100), and DAPI (1:5000). Live Lin-cKit+ cells were sorted on BD Aria machine. Freshly sorted cells were then resuspended in PBS + 0.04% BSA at around 300 k/250 ul, followed by scATAC-seq protocol.

Hi-C data collection

Isolation of murine regulatory T cells was conducted as previously described[55]. The cell suspension was made from pooled secondary lymphoid organs (spleen; peripheral and mesenteric lymph nodes) of Foxp3-GFP mice[56], and CD4 T cells were enriched using the Dynabeads Flowcomp Mouse CD4 Kit (Thermo Fisher, 11461D) according to manufacturer's instructions. The resulting cells were stained with antibodies, washed extensively, and resuspended in isolation buffer (PBS w/ 2% FBS, 10 mM HEPES buffer, 1% L-glutamine, and 2 mM EDTA) containing 0.01% SYTOX Blue dead cell stain (Thermo Fisher, S34857) to facilitate dead cell exclusion, and sorted on a FACSAria (BD) instrument. Treg cells (TCRβ + CD4 + Foxp3-GFP+) and naïve conventional CD4 T cells (TCRβ + CD4 + Foxp3-GFP−CD44loCD62Lhi) were sorted by gating on the respective populations. Hi-C was performed as previously described[57]. Briefly, sorted T cell populations (~1×10^5) were cross-linked in 1% formaldehyde for 10 min and quenched in 125 mM glycine. Cross-linked cells were lysed, and chromatin was restriction enzyme digested using restriction enzymes that digest chromatin at ^GATC and G^ANTC, where N can be any of the four genomic bases (Arima Genomics, San Diego, CA). Digested chromatin was reverse cross-linked using NaCl and eluted in 20 uL 2X Shearing buffer (Covaris, Woburn, MA) and fragmented to 350 base pair fragments using a Covaris LE220Rsc sonicator (Covaris, Woburn, MA). Sheared genomic material was biotinylated and enriched using streptavidin beads. Genomic libraries were prepared to streptavidin-bound DNA using Arima protocol modifications for Accel-NGS 2S DNA plus library kit (IDT, Coralville, IA). After end repair and ligation, libraries were quantified using the KAPA library quantification kit (Roche, Indianapolis, IN) and PCR amplified for the number of cycles required to generate >4 nM per library. Hi-C libraries were sequenced on an Illumina NovoSeq at 500 M read depth, and raw sequencing data in the Fastq format were obtained.

Germinal center B cell centrocytes and centroblast cells were sorted from the spleens of mice immunized with SRBCs for 8 days. Briefly, single-cell suspensions were stained with antibodies against B220 (BV786, BD 563894), CD95/Fas (BUV805, BD 741968), GL7 (AF647, BD 561529), CXCR4 (PE, BD 561734), and CD86 (PECy7, BioLegend 105014). Centrocytes (Live B220 + CD95/Fas+GL7 + CXCR4−CD86+) and centroblasts (Live B220 + CD95/Fas+GL7 + CXCR4 + CD86−) were FACS sorted. All antibodies were used at 1/500 dilution, except CXCR4 and CD86, which were used at 1/250 dilution in PBS + 2% FBS + 0.5 mM EDTA. DAPI was used at 1 µg/mL for the exclusion of dead cells. Cell sorting was performed in a BD Influx cell sorter in the Weill Cornell Medicine Flow Cytometry Core Facility. Flow-sorted CB and CC were fixed in 2% formaldehyde for 10 min. Fixation was quenched by the addition of 0.125 M glycine for 10 min. In situ Hi-C was performed as described (Rao et al. Cell 2014). Nuclei were permeabilized, and DNA was digested overnight with 100 U DpnII (New England BioLabs). The ends of the restriction fragments were labeled using biotin-14-dATP and ligated in a 1-ml final volume. After reversal of cross-links, ligated DNA was purified and sheared to a length of ~400 bp, at which point ligation junctions were pulled down with streptavidin beads, DNA fragments were repaired, and dA-tailed and Illumina adapters were ligated. The library was produced by 6–10 cycles of PCR amplification. Sequencing (paired-end, 50 bp) was performed in a HiSeq 2500 Illumina sequencer in the Weill Cornell Medicine Epigenomics Core.

Statistics and reproducibility

No statistical method was used to predetermine the sample size. In all cases, we held out chromosomes during training of the ChromaFold model and reported the model's performance on the previously held-out test chromosomes. Cell-type-specific ChromaFold predictions were performed on pre-clustered cells using scATAC-seq data. Additionally, we conducted a down-sampling analysis and observed robust performance of ChromaFold with as few as 3000 randomly selected test cells.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Some of the scATAC-seq, Hi-C, and CTCF ChIP-seq data used for training and evaluation were obtained from publicly available repositories, and the remainder were generated for this study and deposited to NCBI Gene Expression Omnibus (GEO) database under accession code GSE246859. The accession numbers for publicly available datasets are listed in the Supplementary Information and described below. IMR-90 data were obtained from the ENCODE data portal with accession numbers ENCSR778RZT (scATAC-seq), ENCSR345VTI (Hi-C), and ENCSR000EFI (CTCF ChIP-seq). HUVEC data were obtained from ENCODE with accession numbers ENCSR516MHK (scATAC-seq), ENCSR788FBI (Hi-C), and ENCSR000ALA (CTCF ChIP-seq). GM12878 scATAC-seq data were

obtained from 10X Genomics (https://www.10xgenomics.com/resources/datasets/10-k-1-1-mixture-of-fresh-frozen-human-gm-12878-and-mouse-a-20-cells-next-gem-v-1-1-1-1-standard-2-0-0) and from ENCODE (ENCSR680NPV), Hi-C data from the 4DN data portal with accession numbers 4DNFI1UEG1HD and 4DNESCIHJOXA, and CTCF ChIP-seq data from ENCODE (ENCSR000AKB). K562 scATAC-seq data were obtained from ENCODE (ENCSR308ZGJ, ENCSR217VXJ), Hi-C data from the 4DN data portal (4DNFITUOMFUQ), 4DNES9J6QJQS [https://data.4dnucleome.org/higlass-view-configs/2fb04ff2-b951-4f3d-857c-40a7e22ec56e/], CTCF ChIP-seq data from ENCODE (ENCSR000AKO), and IDR thresholded peak data from ENCODE (ENCFF598YSU). Human ESC Hi-C data were obtained from the 4DN data portal (4DNFI2TK7L2F) and CTCF ChIP-seq data from ENCODE (ENCSR000AMF), and scATAC-seq data were generated in this study and deposited to GEO (GSE246859). Human CD4 + T cell scATAC-seq data were obtained from ENCODE (ENCSR628NXO) and Hi-C data from ENCODE (ENCSR421CGL). Mouse germinal center B cells Hi-C data was obtained from GEO (GSE143853), and scATAC-seq data were generated in this study and deposited to GEO (GSE246859). Mouse regulatory T cell scATAC-seq data were obtained from GEO (GSE156112), and Hi-C data were generated in this study and deposited to GEO (GSE246859). Mouse hematopoietic stem cell Hi-C data were obtained from GEO (GSE135031), and scATAC-seq data were generated in this study and deposited to GEO (GSE246859). Human pancreatic islet cell, sorted alpha cell, and sorted beta cell scATAC-seq data were obtained from GEO (GSE160472). Human pancreatic islet cell Hi-C data were obtained from the Accelerating Medicines Partnership data portal under accession number (DFF064KIG). Sorted human alpha and beta cell Hi-C were obtained from GEO (GSE188311). CRISPRi-FlowFISH data were obtained from EPCrispr-Benchmark_ensemble_data_GRCh38.tsv.gz (https://github.com/EngreitzLab/CRISPR_comparison/tree/main/resources/crispr_data). A minimum dataset of processed input data and normalized Hi-C contact maps for IMR-90 (hg38) is available at Zenodo[58] [10.5281/zenodo.13362537]. Source data for generating the figures are provided at Zenodo[58] [10.5281/zenodo.13362537].

Code availability

The ChromaFold model code has been deposited into GitHub under the MIT license and is publicly accessible at ChromaFold Github[59] [10.5281/zenodo.13862915] (https://github.com/viannegao/ChromaFold/tree/main). The data preprocessing code incorporates components from external software packages ArchR, which is used under the terms of the GNU General Public License (GPL) version 2 or later. The original ArchR code and its associated copyright information can be found at https://www.archrproject.com/index.html.

References

1. Van Berkum, N. L. et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **6**, e1869 (2010).
2. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
3. Fullwood, M. J. et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
4. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
5. Krijger, P. H. L. & De Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. cell Biol.* **17**, 771–782 (2016).
6. Liu, Q., Lv, H. & Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **35**, i99–i107 (2019).
7. Zhang, Y. et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat. Commun.* **9**, 750 (2018).
8. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
9. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
10. Kim, H.-J. et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput. Biol.* **16**, e1008173 (2020).
11. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat. Biotechnol.* **40**, 254–261 (2022).
12. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
13. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
14. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).
15. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* **54**, 725–734 (2022).
16. Tan, J. et al. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat. Biotechnol.* **1**, 11 (2023).
17. Yang, R. et al. Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. *Genome Biol.* **24**, 1–26 (2023).
18. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
19. Schwessinger, R. et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* **17**, 1118–1124 (2020).
20. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2 (Springer, 2009).
21. Chiou, J. et al. Single-cell chromatin accessibility identifies pancreatic islet cell type– and state-specific regulatory programs of diabetes risk. *Nat. Genet.* **53**, 455–466 (2021).
22. Hsieh, T.-H. S. et al. Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat. Genet.* **54**, 1919–1932 (2022).
23. Grubert, F. et al. Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* **583**, 737–743 (2020).
24. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
25. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
26. Sahin, M. et al. HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. *Nat. Commun.* **12**, 3366 (2021).
27. Dozmorov, M. G. et al. CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinform. Adv.* **2**, vbac097 (2022).
28. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
29. Choi, S.-S., Cha, S.-H. & Tappert, C. C. A survey of binary similarity and distance measures. *J. Syst. Cybern. Inform.* **8**, 43–48 (2010).
30. Lal, A. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. *Nat. Commun.* **12**, 1507 (2021).
31. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48 (2019).
32. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. *Proc. AAAI Conf. Artif. Intell.* **34**, 13001–13008 (2020). vol.

33. Gschwind, A. R. et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. Preprint at *bioRxiv* (2023).
34. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 12 (2019).
35. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
36. Basso, K. & Dalla-Favera, R. Roles of BCL6 in normal and transformed germinal center B cells. *Immunol. Rev.* **247**, 172–183 (2012).
37. Kitano, M. et al. Bcl6 protein expression shapes pre-germinal center B cell dynamics and follicular helper T cell heterogeneity. *Immunity* **34**, 961–972 (2011).
38. Hahm, K. et al. Helios, a T cell-restricted Ikaros family member that quantitatively associates with Ikaros at centromeric heterochromatin. *Genes Dev.* **12**, 782–796 (1998).
39. Kim, H.-J. et al. Stable inhibitory activity of regulatory T cells requires the transcription factor Helios. *Science* **350**, 334–339 (2015).
40. Galitsyna, A. A. & Gelfand, M. S. Single-cell Hi-C data analysis: safety in numbers. *Brief. Bioinform.* **22**, bbab316 (2021).
41. Su, C. et al. 3D chromatin maps of the human pancreas reveal lineage-specific regulatory architecture of T2D risk. *Cell Metab.* **34**, 1394–1409.e4 (2022).
42. Zhang, Z., Feng, F., Qiu, Y. & Liu, J. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Res.* **51**, 5931–5947 (2023).
43. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
44. Wolff, J. et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).
45. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
46. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
47. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
48. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
49. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
50. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
51. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
53. David, H. A. & Gunnink, J. L. The paired t test under artificial pairing. *Am. Stat.* **51**, 9–12 (1997).
54. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70–e70 (2016).
55. Hu, W. et al. Regulatory T cells function in established systemic inflammation and reverse fatal autoimmunity. *Nat. Immunol.* **22**, 1163–1174 (2021).
56. Fontenot, J. D. et al. Regulatory T cell lineage specification by the forkhead transcription factor Foxp3. *Immunity* **22**, 329–341 (2005).
57. Viny, A. D. et al. Cohesin members Stag1 and Stag2 display distinct roles in chromatin accessibility and topological control of HSC self-renewal and differentiation. *Cell Stem Cell* **25**, 682–696.e8 (2019).
58. Gao, V. R. et al. ChromaFold minimum dataset [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.13362537 (2024).
59. Gao, V. R. et al. viannegao/ChromaFold: Initial release of ChromaFold (v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.13862915 (2024)

Acknowledgements

This work was supported in part by NIH U01 awards HG012103 (C.S.L. and A.Y.R.) and DK128852 (C.S.L., D.H., and E.A.). M.A.R. is supported by a Junior Faculty Scholar Award from the American Society of Hematology. H.L. is supported by NYSTEM training award contract C32599GG, and K99 DK128602-01. W.B. is an American Society of Hematology Junior Faculty Scholar Awardee, supported by NIH R01CA270245 and grants from The Leukemia & Lymphoma Society, Lymphoma Research Foundation, and The Follicular Lymphoma Foundation. M.G.K. is a Scholar of the Leukemia and Lymphoma Society and supported by Starr Cancer Consortium, NIDDK NIH R01-DK101989-01A1; NCI 1R01CA193842-01, R01HL135564, R01 CA274249-01A1, R01CA186702, R01CA283578, and R01CA225231-01.

Author contributions

V.R.G. and C.S.L. developed the model. V.R.G., R.Y., and C.S.L. designed and conducted analyses and wrote the manuscript. R.Y., A.D., W.S.N., J.A.B., A.K., and W.W. contributed to the model conception. A.K. and W.W. processed Hi-C data and conducted computational modeling. Y.A.Z. and C.R.C. contributed to dataset processing and annotation. R.L. H.L., D.R.M., I.K., M.A.R., Z.M.W., D.B., E.A., M.G.K., W.B., A.D.V., D.H., A.Y.R., and A.M.M. generated and contributed new datasets.

Competing interests

C.S.L. is an SAB member and co-inventor of IP with Episteme Prognostics, unrelated to the current study. M.G.K is a member of the scientific advisory board of 858 Therapeutics and the laboratory gets research support from AstraZeneca and Transition Bio. A.D.V. is an SAB member of Arima Genomics. A.Y.R. is an SAB member and has equity in Sonoma Biotherapeutics, Santa Ana Bio, RAPT Therapeutics, and Vedanta Biosciences. He is an SEB member of Amgen and BioInvent and is a co-inventor or has IP licensed to Takeda that is unrelated to the content of the present study. A.M.M. has research funding from Janssen, Epizyme, and Daiichi Sankyo. A.M.M. has consulted for Exo Therapeutics, Treeline Biosciences, and AstraZeneca. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53628-0.

Correspondence and requests for materials should be addressed to Christina S. Leslie.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.