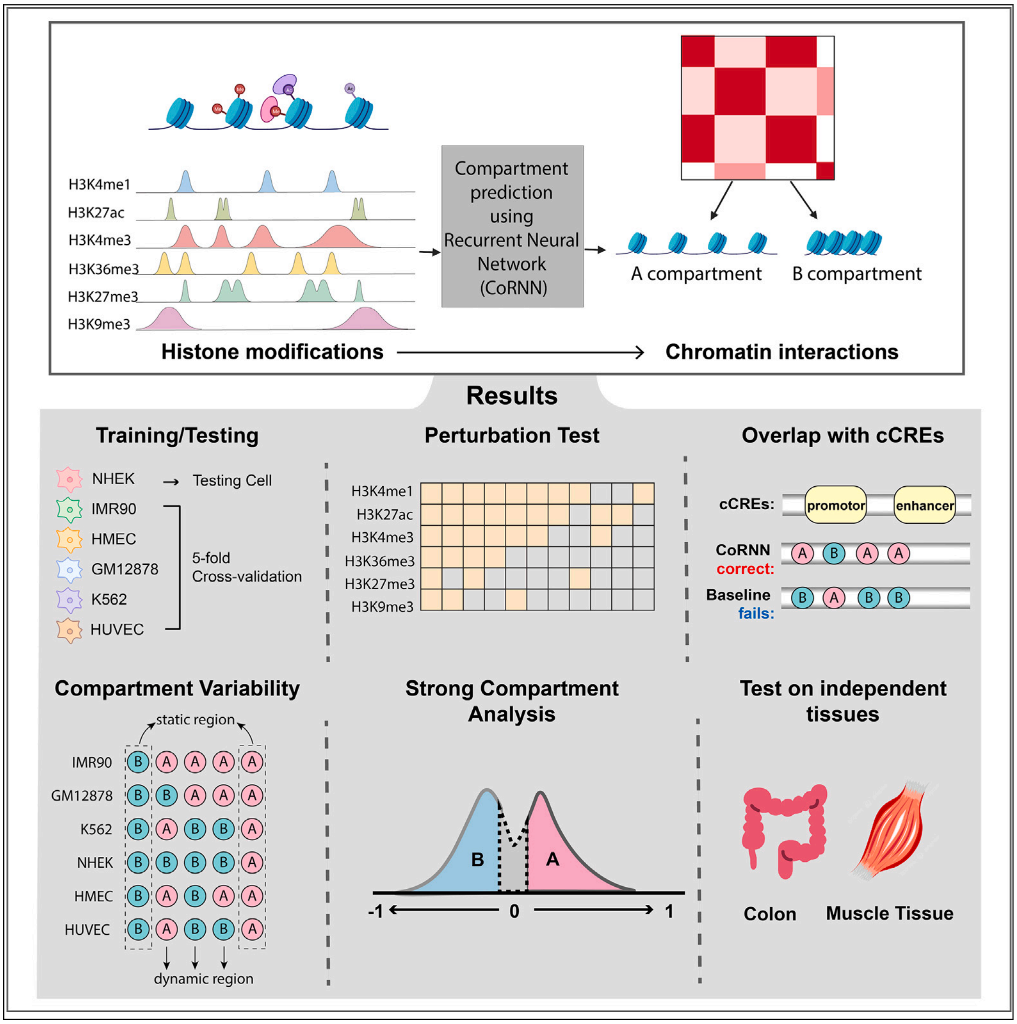
iScience

CellPress
OPEN ACCESS

Article

Predicting A/B compartments from histone modifications using deep learning



Suchen Zheng, Nitya Thakkar, Hannah L. Harris, ..., William Stafford Noble, Gamze Gürsoy, Ritambhara Singh

gg2845@cumc.columbia.edu (G.G.)
ritambhara@brown.edu (R.S.)

Highlights

CoRNN model predicts A/B compartments using histone modifications

High accuracy across cell types surpasses traditional methods

Key histone modifications identified for accurate predictions: H3K27ac and H3K36me3

Applicable to tissue samples, broadening genomic research scope

Zheng et al., iScience *27*, 109570
May 17, 2024
https://doi.org/10.1016/j.isci.2024.109570

iScience

CellPress
OPEN ACCESS

Article

Predicting A/B compartments from histone modifications using deep learning

Suchen Zheng,[1,10] Nitya Thakkar,[1,10] Hannah L. Harris,[2] Susanna Liu,[3] Megan Zhang,[3] Mark Gerstein,[4] Erez Lieberman Aiden,[5] M. Jordan Rowley,[6] William Stafford Noble,[7] Gamze Gürsoy,[8,*] and Ritambhara Singh[9,11,*]

SUMMARY

The three-dimensional organization of genomes plays a crucial role in essential biological processes. The segregation of chromatin into A and B compartments highlights regions of activity and inactivity, providing a window into the genomic activities specific to each cell type. Yet, the steep costs associated with acquiring Hi-C data, necessary for studying this compartmentalization across various cell types, pose a significant barrier in studying cell type specific genome organization. To address this, we present a prediction tool called compartment prediction using recurrent neural networks (CoRNN), which predicts compartmentalization of 3D genome using histone modification enrichment. CoRNN demonstrates robust cross-cell-type prediction of A/B compartments with an average AuROC of 90.9%. Cell-type-specific predictions align well with known functional elements, with H3K27ac and H3K36me3 identified as highly predictive histone marks. We further investigate our mispredictions and found that they are located in regions with ambiguous compartmental status. Furthermore, our model's generalizability is validated by predicting compartments in independent tissue samples, which underscores its broad applicability.

INTRODUCTION

The physical organization of DNA inside the cell nucleus directly impacts the function and biology of the genome. DNA organization has been implicated in numerous biological processes from differentiation to oncogenesis.[1] Genome-wide chromosome conformation capture (Hi-C) and related techniques enable the characterization of this organization by capturing the long-range pairwise interactions among different genomic elements.[2–5] Hi-C data have revealed that the genome is organized into many organizational units such as compartments, topologically associating domains (TADs), and loops. Of particular interest to this study is the observation that the genome is organized into two distinct compartments, labeled ''A'' (active) and ''B'' (inactive).[3] Each of these compartments corresponds to distinct properties of the associated genomic regions. For example, there are preferential interactions within compartment types, such that loci in the A compartment tend to interact with loci in the same compartment. A/B compartment boundaries are typically identified by applying principal components analysis (PCA) to the correlation matrix obtained from the Hi-C interaction frequency matrix, in which the sign of the first principal component corresponds to the A/B compartments. Conventionally, loci associated with A compartments (active) are assigned positive values, while those in B compartments (inactive) are assigned negative values in the eigenvector.

While Hi-C is a powerful experimental technique to detect chromosomal compartments, the high cost and technical difficulties make obtaining Hi-C data for many different cell lines and types challenging. Therefore, there are still many cell lines, cell types, and tissue types for which Hi-C data are not yet available. Fortunately, chromosomal compartments have been found to correlate with histone modification patterns.[3,6] For example, it was shown that there is a high concordance between the chromatin immunoprecipitation sequencing (ChIP-seq) signal of active histone mark enrichment such as H3K4me1 in regions that are located in A compartments.[3] Thus, predicting chromosomal compartments via more abundantly available data types, such as ChIP-seq, can remedy the lack of Hi-C data and enable the discovery of 3D genome organization without the need for expensive experiments. Furthermore, such prediction methods can provide insight into the

[1] Department of Computer Science, Brown University, Providence, RI, USA
[2] Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA
[3] Data Science and Statistics, Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT, USA
[4] Computational Biology and Bioinformatics, Molecular Biophysics & Biochemistry, Data Science and Statistics, Computer Science, Yale University, New Haven, CT, USA
[5] Department of Genetics, Baylor College of Medicine, Department of Computer Science, Computational and Applied Mathematics, Rice University, Houston, TX, USA
[6] Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA
[7] Department of Genome Sciences, Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[8] Department of Biomedical Informatics, Columbia University, New York Genome Center, New York, NY, USA
[9] Department of Computer Science, Center for Computational Molecular Biology, Brown University, Providence, RI, USA
[10] These authors contributed equally
[11] Lead contact
*Correspondence: gg2845@cumc.columbia.edu (G.G.), ritambhara@brown.edu (R.S.)

https://doi.org/10.1016/j.isci.2024.109570

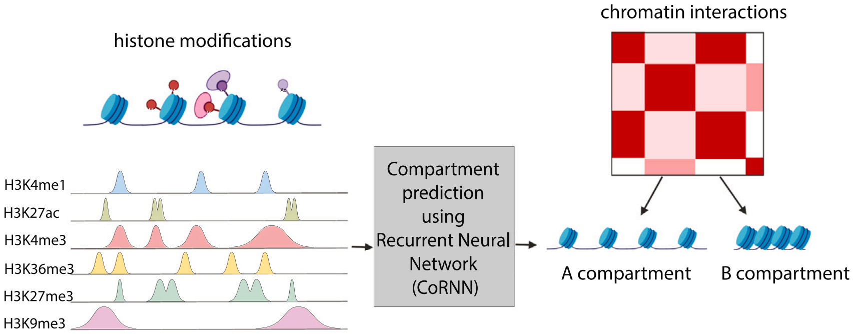


Figure 1. Overview of the approach

The A/B compartment prediction task is formulated as a binary classification problem. We use six histone modification ChIP-seq experiments as our inputs to predict A/B compartments via an RNN model.

interplay between the 3D organization of the genome and its 1D activity level. Therefore, it is crucial to (1) find ways to infer cell-type specific compartments without needing Hi-C data generation and (2) discover relationships between compartments and chromatin marks to better understand the connections between the spatial organization and biology of the genome.

Previous methods have used epigenomic signals (see review paper Tao H. et al.[7]) to predict different organizational units of the 3D genome. For example, Fortin et al.[8] used the eigenvectors calculated from correlation matrices of DNA methylation experiments and reported correlation values of $\sim 0.56 - 0.71$ with the A/B compartments. Moore et al.[9] developed a random forest model (3DGenome) to predict A/B compartment values from co-localization of 22 transcription factor binding sites (TFBSs) and 10 histone marks. Raineri et al.[10] used a linear regression model to predict compartments from GC-content and DNA methylation experiments and reported a mean absolute error of 0.9. Jenkinson et al.[11] showed that entropy blocks calculated from DNA methylation data correspond well to the TAD boundaries obtained from Hi-C data. They reported a correlation score of ~ 0.80 across three cell lines. Other recent TAD prediction methods[12] also report high TAD prediction accuracy ($\sim$ 90%). Al Bkhetan et al.[13] used a random forest model to predict contact loops obtained from chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) experiments, transcription factors (TFs), and histone modification experiments. They reported an accuracy of 0.87 for their model (3DEpiLoop). However, most of these prediction methods require a large number of or diverse types of data, which may not be readily available for a cell type of interest. Since the state of the chromatin directly relates to its activity (e.g., gene expression, TF binding, etc.),[14,15] we hypothesize that we can accurately predict A/B compartments using a few histone modification signals, thereby enabling prediction from abundantly available data.

Here we propose a deep learning framework for **co**mpartment prediction using recurrent neural networks (CoRNN) to predict chromosome compartments using histone modification ChIP-seq data (Figure 1). We perform this prediction from existing datasets to gain insights into 3D genome organization without the need for expensive Hi-C data generation efforts. For instance, ENCODE data portal contains > 3000 histone modification ChIP-seq data from many cell lines, types, and tissues and only $\sim$ 120 Hi-C datasets. Our RNN-based framework allows us to learn directly from a high-resolution and granular ChIP-seq signal (binned at 1000bp) to predict compartments at a coarse resolution (100 kbp). Thus, to prevent the dilution of the input signal, we refrain from conducting any averaging or aggregation at the input level. We restrict our study to only six histone modifications to demonstrate the effectiveness of these few input signals in predicting A/B compartments, unlike previous studies[9,16] that incorporate a large number ($\sim 30 - 80$) of diverse input signals like different transcription factors along with histone modifications. Our proposed framework, CoRNN (detailed in Figure 2), performs better than state-of-the-art baselines, achieving an average accuracy improvement of 10% across all cell lines. More importantly, it outperforms the mean compartment value baseline (or the "mean baseline") that predicts the compartment assignment of a genomic region based on the average compartment values across cell lines in the training set. Since most compartments are conserved across cell types,[3] this baseline is quite challenging to beat. Next, we show that CoRNN correctly predicts compartments for genomic regions whose compartment values vary across cell lines. These highly variable regions, correctly predicted by CoRNN but missed by the mean baseline, are biologically relevant and overlap with known candidate *cis*-regulatory elements. We further find that the regions that CoRNN mispredicts tend to have highly ambiguous compartment scores that vary between different Hi-C biological replicates of the same cell line (human lymphoblastoid cells). Additionally, our perturbation analysis on the trained model shows that H3K27ac and H3K36me3 are the most informative histone marks for CoRNN. Finally, we run CoRNN on two independent out-of-distribution test datasets from human muscle and colon tissues to assess the model's generalizability. We show that CoRNN outperforms the mean baseline by 13.9% for these previously unseen datasets. Overall, this study and our new framework CoRNN enable the assignment of A/B compartments to genomic regions for cell lines and tissues with no available Hi-C data using existing ChIP-seq datasets as inputs.

RESULTS

CoRNN gives state-of-the-art compartment prediction performance with cell line specificity

Figure 3 presents the A/B compartment classification performance of CoRNN across six selected cell lines using the AUROC score (See Table S3 for performance scores). In functional genomics, especially in measurements of 3D genome configuration, most of the signal can

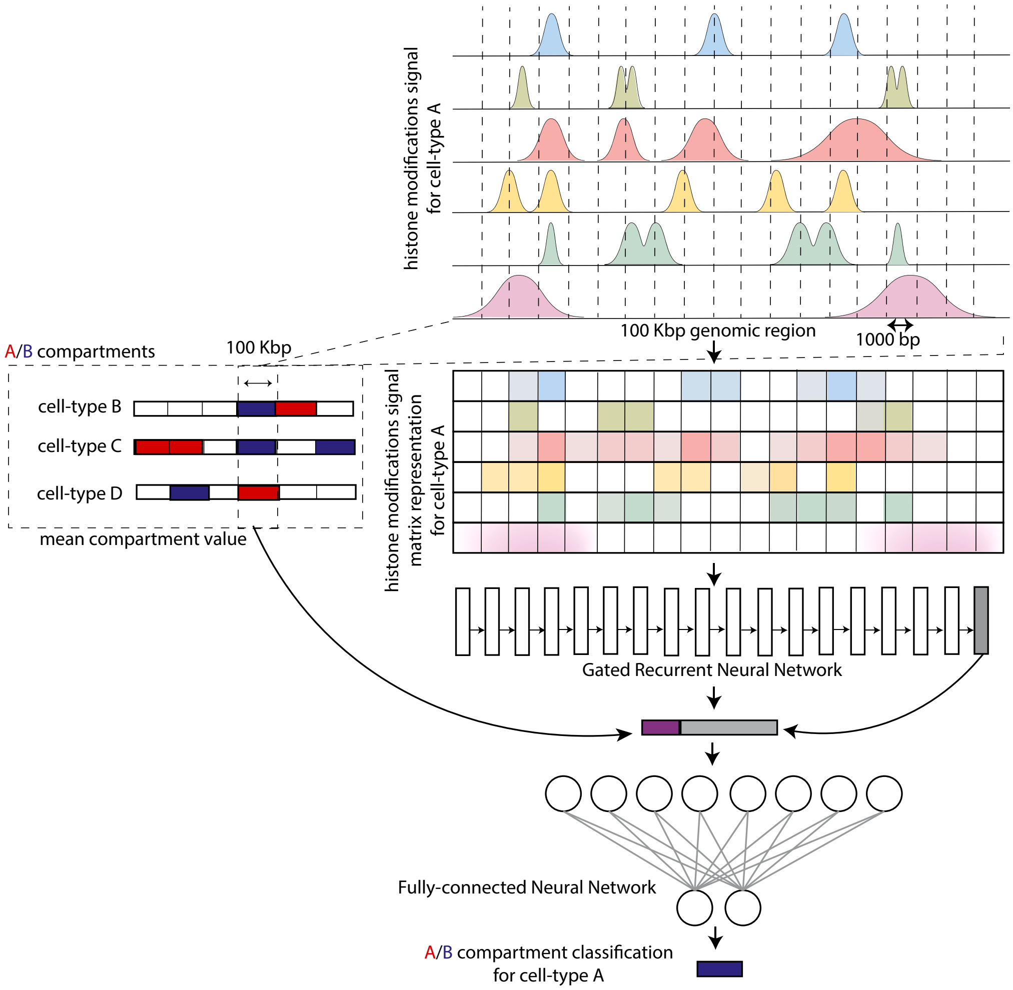


Figure 2. CoRNN architecture

The model consists of gated recurrent unit layers to capture the sequential information of the histone modification signals across the genomic region. The output of the GRU is then concatenated with the mean compartment value c for the training cell lines and fed into fully connected layers. The output $y \in [0, 1]$ represents the binarized compartment value for the input genomic region.

be highly conserved across cell types.[17] For example, if we look at the correlation of compartment values among the six cell lines, we find generally high correlation values (minimum correlation is 73 % and maximum correlation is 96 %; Figure S6). This observation indicates that 100 kbp compartment labels across the genome are largely consistent among different cell types. Interestingly, our predictions are more accurate than the mean compartment value baseline for five of the six cell lines (a similar trend for AUPRC scores in Figure S7). Note that inclusion of mean compartment value as input in CoRNN boosts the performance over a GRU model (labeled as "GRU") that does not leverage this information.

Many genomic regions have the same Hi-C-derived compartment labels across different cell lines, while some are more variable (or dynamic) due to the cell type specificity. We divided all the genomic regions with associated compartment values into sub-groups based on their label concordance across the five training cell lines. Figure 4A plots these sub-groups on the x axis and reports the accuracy of the mean baseline and CoRNN for these regions in the sixth test cell line on the y axis. We cannot obtain a ranking for the mean baseline to calculate the AUROC score for this analysis as it predicts only one value (0, 0.2, 0.4, 0.6, 0.8, or 1.0) for each sub-group. Therefore, we use the accuracy metric instead. These accuracy scores have been averaged across all the test cases. A value of 0 on the x axis represents genomic regions with B compartment labels consistent across all five cell lines, and a value of 5 represents the same for the A compartment. Similarly, 1 and 4 represent regions with A or B label concordance in four out of five cell lines, and 2 and 3 for three out of five. We call the genomic regions with low A or B compartment concordance in labels across the cell lines "dynamic regions" (values 1–4). As expected, our CoRNN model exhibits a performance gain over the mean baseline for these regions, which is especially significant for regions with A compartment variability across 3 and 4 cell lines. This result indicates that CoRNN can leverage the histone modification profiles to make more accurate predictions for cell-type-specific compartments. Next, we investigated the regions that are correctly predicted by CoRNN but are missed by the mean baseline. These

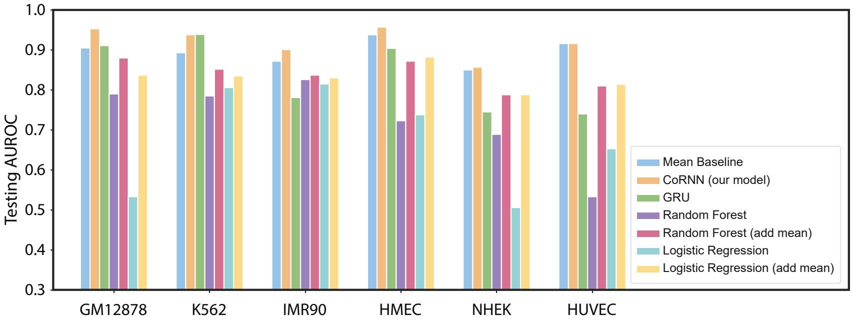


Figure 3. Prediction performance of CoRNN and baseline methods

Our model gives the best prediction performance across all cell lines and outperforms the mean baseline for five out of six cell lines.

regions tend to be cell-type specific; hence, one will not be able to assign compartment labels based on the mean compartment values of all cell lines. We calculated the enrichment of candidate *cis*-regulatory elements[18] (cCREs, ENCODE:ENCFF788SJC) on regions that are predicted by CoRNN but missed by the mean baseline and, conversely, on regions that are predicted accurately by the mean baseline but missed by CoRNN. This overlap is calculated by annotating every 100kb bin in the genome as containing cCREs if there is at least one cCRE overlapping with the bin. This is done because cCREs and compartments are defined with different resolutions.

We found that almost all of the regions (99%) CoRNN correctly predicts but the mean baseline misses overlap with cCREs. In contrast, only $20 - 30\%$ of the complementary regions overlap with cCREs (Figure 4B). This trend holds for all cell lines, even though the predicted regions differ across cell lines. This observation suggests that most of the regions that CoRNN exclusively predicts correctly are functionally important.

CoRNN has high accuracy for regions with unambiguous compartment assignments

When we look at the eigenvalues used to assign compartments, there are genomic regions that show weak signals (small absolute values). Therefore, it is unclear which compartments these ambiguous regions belong to and whether this is an issue with the resolution of the Hi-C data. Because we frame the A/B compartment prediction task as binary classification, we hypothesize that CoRNN would show better prediction performance for regions with "strong" compartments (that is, regions with high eigenvalues and, thus, more confident compartment assignments). Therefore, we select a subset of these strong compartment regions as those with absolute eigenvalues $>(mean - std.deviation)$ (Figure S8). We observe a marked increase in AUROC scores for both CoRNN and the mean baseline for these strong compartment regions (Figure 4C). Interestingly, our model is very good at predicting these regions, achieving AUROC scores ~ 0.98 for four out of six cell lines.

Regions with ambiguous chromatin marks or compartment values are difficult to predict

We explored the regions that CoRNN struggled to predict accurately to understand the functionality of these regions and why histone modification information may be insufficient to classify their compartment values accurately. We also included Pol2Ser2 and Pol2RA signals in this analysis to understand these regions' activity profiles better as Pol2Ser2 and Pol2RA signals indicate the RNA polymerase activity, which is indicative of gene expression. Unsurprisingly, when we looked at the mispredictions, we found that regions predicted as A by CoRNN but annotated as B by Hi-C eigenvalues generally had histone marks more similar to that expected by the regions that are in the B compartment (Figure 5A). However, there was an exception with histone mark H3K27me3, whose values were more similar to

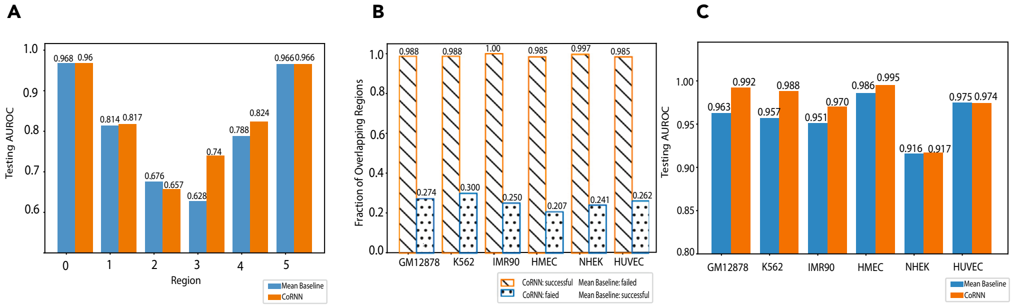


Figure 4. Investigation of CoRNN's predictive performance

(A) Comparison of prediction accuracy between CoRNN and mean baseline when the regions are categorized by A/B compartment variability across cell lines.

(B) Fraction of overlap between candidate *cis*-regulatory elements (cCREs) and predicted compartment for regions that are correctly predicted by CoRNN but missed by mean baseline and vice versa.

(C) Comparison of AUROC scores between CoRNN and mean baseline for all cell lines for genomic regions with strong compartment values.

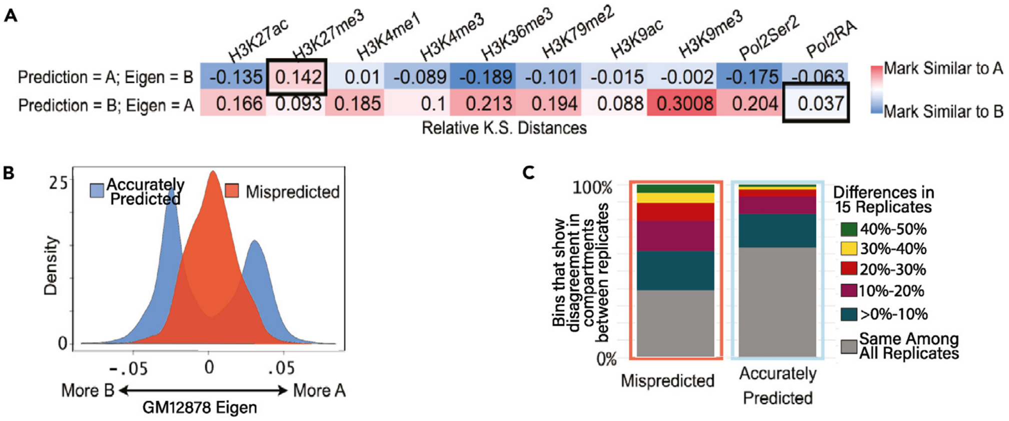


Figure 5. Investigation of ambiguous regions that are difficult to predict by CoRNN

(A) Similarity of chromatin marks in mispredicted bins to that of correctly predicted A or B compartments. Values represent the subtraction of Kolmogorov-Smirnov distances for each. Black rectangles highlight marks that may have contributed to the misprediction.

(B) The distribution of the eigenvector in the GM12878 map for bins that were mispredicted vs. accurately predicted.

(C) Examination of the eigenvector for 15 biological replicates of the GM12878 Hi-C map and the percentage of bins that show disagreement between individual maps.

that of regions correctly assigned as the A compartment. This observation indicates that regions residing in the B compartment might be challenging to classify if their H3K27me3 status is similar to that of regions in the A compartment. This also means that it is difficult to predict the compartmental status of loci with both active and repressive chromatin marks, such as bivalent enhancers. Indeed, it is likely that these types of regulatory elements form unique chromatin interaction patterns.[19] In contrast, regions that were predicted as B by CoRNN but annotated as A by Hi-C eigenvalues had somewhat intermediate levels of active marks (Figure 5A). Pol2RA levels were especially low compared to A compartment regions. Altogether, these results indicate that regions that are mispredicted by CoRNN often exhibit unusual chromatin activity mark enrichments for the compartment status designated by Hi-C. This could also be indicative of the limitations imposed by a two-state compartment model,[20] suggesting that sub-compartment calling can provide valuable additional information for these difficult-to-predict regions. We further analyzed the difficult-to-predict genomic regions by examining the eigenvalues from the Hi-C map of GM12878 cell line.[21] Regions that CoRNN has trouble predicting have values closer to 0 in the eigenvector, indicating a more ambiguous compartment status compared to those that are accurately predicted (Figure 5B). Because the GM12878 Hi-C map represents a combination of 15 independent replicates, the ambiguous compartment status in the combined map may be due to variability between individual replicates. Using the Hi-C maps of the individual replicates, we annotated compartments from the eigenvector and examined the compartment status of each for mispredicted versus accurately predicted bins. From this analysis, we found that bins mispredicted by CoRNN often represent sites with poor agreement among replicates (Figure 5C). These analyses suggest that CoRNN is highly accurate for most regions. However, some sites are difficult to predict due to an ambiguous compartment status from unexpected chromatin marks or variability in the sampled population.

Perturbation analysis reveals the most predictive histone marks for CoRNN

Histone modifications can provide redundant information because many types of histone modifications are highly correlated with one another. We perform a perturbation analysis to determine which histone modifications have the best predictive power and are the most relevant for our accurate A/B compartment classification for the test cell lines. We take our trained CoRNN model for each cell line and mask out all possible combinations of histone modification signals one by one by replacing the input matrix rows with zeros and recording the AUROC scores. Figure 6A presents the performance results for the combinations of histone marks that result in a similar or higher AUROC score compared to the mean baseline. The list of histone modifications is ranked based on their frequency of occurrence for such combinations. We observe that H3K27ac and H3K36me3 are the most important histone marks for CoRNN to make accurate A/B compartment classification. These are followed by H3K4me1 and H3K9me3, respectively. H3K4me3 and H3K27me3 seem to be the least relevant for the CoRNN predictions. These results align with recent studies connecting histone modifications to A/B compartments. A recent study[22] found that H3K4me1, H3K9me3, and H3K27ac are some of the histone marks that are significantly predictive of most Hi–C interactions in human ES cells. Another study,[19] using the ultra-resolution Hi-C contact map in lymphoblastoid cell lines, observed the correlation of low H3K4me1 and H3K36me3 signals with the presence of discordant compartmentalization. Therefore, by modeling the relationship of histone modifications with the A/B compartments, CoRNN can capture the relevance of these marks in highlighting the properties of genomic compartmentalization. This perturbation analysis also helps us identify the minimum amount of information required for predicting A/B compartments. For example, if only H3K27ac and H3K36me3 ChIP-seq experiments are available for a given cell line, we can still make accurate A/B compartment classifications using CoRNN.

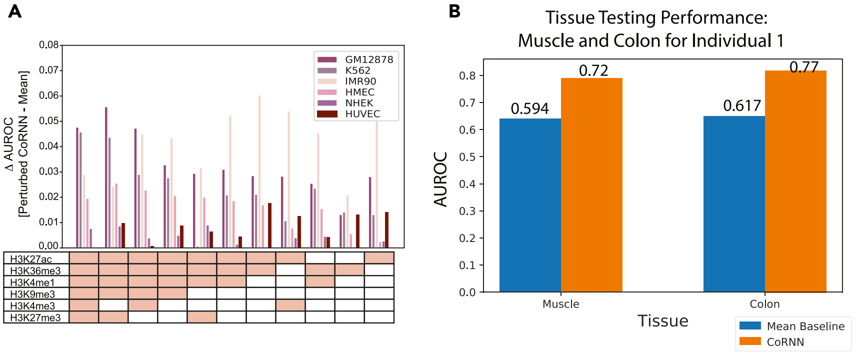


Figure 6. Further investigation of CoRNN's predictive power using histone mark subset and on independent tissues

(A) Perturbing different combinations of histone marks results in a similar or higher AUROC score compared to the mean baseline.

(B) CoRNN predicts A/B compartments for both muscle and colon tissue samples with higher AUROC scores compared to the mean baseline.

CoRNN accurately predicts A/B compartments in independent tissues

Finally, we demonstrate the generalizability of a trained CoRNN model by testing it on new tissue samples. Note that the CoRNN model is trained on datasets from six cell lines, and we test it on colon and muscle tissue samples taken from a consented individual (an out-of-distribution test set). These datasets are entirely unseen by the model during training. We selected them because they had all the necessary ChIP-seq inputs and Hi-C experiments as ground truth for evaluating the model predictions. We present the prediction performance of CoRNN and compare it to the mean baseline in Figure 6B. We observe that CoRNN predicts the A/B compartments more accurately than the mean baseline with AUROC scores of 0.72 and 0.77 and AUPRC scores of 0.79 and 0.82 (Figure S9) for muscle and colon tissue, respectively. However, the scores, in general, are lower for the tissue samples as compared to the cell lines. We hypothesize that this observation is due to the heterogeneity of the tissue samples from the cell lines that the model is trained on. Nevertheless, our results indicate that CoRNN is a useful predictive tool for new out-of-distribution cell lines or tissues with missing Hi-C data, and a better alternative than using mean values as proxies for compartment values.

DISCUSSION

Our proposed CoRNN method takes one-dimensional ChIP-seq signals of histone modification enrichment and accurately predicts the chromosomal compartments that otherwise require Hi-C data for their computation. This method will enable obtaining compartment designations for cell types that do not have Hi-C data available and will also allow investigation of the relationship between the epigenomic landscape and its three-dimensional shape in the nucleus. In this study, we compare CoRNN to a common baseline used in genomics when an experimental dataset for a cell line, cell type, or tissue type is missing. That is, we average all the compartment scores across available cell lines and use them as proxies for compartment values for the held-out test cell line. We show that CoRNN predicts regions with cell-type specific compartments better than this mean baseline. When we analyzed the regions that CoRNN mispredicted, we found that they represent regions with unusual marks for their Hi-C annotated compartment. Our perturbation analysis shows that highly accurate predictions can be made even when using only 2–3 histone modification ChIP-seq datasets, thereby enabling inference of large-scale genome organization from experimental data that is easier and cheaper to obtain than Hi-C. Finally, we demonstrated CoRNN's generalizability to out-of-distribution samples. In summary, CoRNN offers a cost-effective alternative to expensive Hi-C experiments for inferring A/B compartments. By utilizing only a few histone modification signals, it makes the analysis of 3D genome organization more accessible.

Limitations of the study

Due to the limited availability of high-coverage Hi-C datasets across multiple cell lines, we performed our predictions and analyses at a low resolution (100 kbp). As more high-quality Hi-C experiments become available, CoRNN can be easily extended to perform predictions of compartments at high resolutions. Note that high coverage Hi-C data from Rao et al.[21] also identified chromosomal subcompartments that show distinct and more detailed associations with various features such as gene expression, active and repressive histone marks, DNA replication timing, and specific subnuclear structures. Therefore, future work would include the prediction of sub-compartments (as done previously in Ashoor H. et al.[23]) using CoRNN architecture and investigation of the role of epigenomic landscape toward modifying the finer sub-structures of the chromatin.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109570.

ACKNOWLEDGMENTS

We would like to thank ENCODE Consortium's Nuclear Architecture Working Group for helpful discussions and suggestions. This work was funded by National Institutes of Health awards U24 HG009446 (William S. Noble's effort) and R35 HG011939 (Ritambhara Singh's effort).

AUTHOR CONTRIBUTIONS

G.G. conceptualized the idea. R.S. led the method development. R.S., G.G., and W.S.N. advised the project. S.Z. implemented the method and N.T. implemented the baselines. S.Z. and N.T. worked on the experimental setup and results. H.L.H. and M.J.R. performed the A/B compartment analysis for GM12878 cell line. S.L. and M.Z. helped with data analysis. M.G. and E.L.A. provided useful insights during project discussions. All the authors contributed toward manuscript preparation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 4, 2024
Revised: February 28, 2024
Accepted: March 22, 2024
Published: March 27, 2024

REFERENCES

1. Zheng, H., and Xie, W. (2019). The role of 3D genome organization in development and cell differentiation. Nat. Rev. Mol. Cell Biol. *20*, 535–550. https://doi.org/10.1038/s41580-019-0132-4.
2. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311. https://doi.org/10.1126/science.1067799.
3. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293. https://doi.org/10.1126/science.1181369.
4. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367. https://doi.org/10.1038/nature08973.
5. Montefiori, L., Wuerffel, R., Roqueiro, D., Lajoie, B., Guo, C., Gerasimova, T., De, S., Wood, W., Becker, K.G., Dekker, J., et al. (2016). Extremely long-range chromatin loops link topological domains to facilitate a diverse antibody repertoire. Cell Rep. *14*, 896–906. https://doi.org/10.1016/j.celrep.2015.12.067.
6. Spracklin, G., Abdennur, N., Imakaev, M., Chowdhury, N., Pradhan, S., Mirny, L., and Dekker, J. (2021). Heterochromatin diversity modulates genome compartmentalization and loop extrusion barriers. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.05.455340.
7. Tao, H., Li, H., Xu, K., Hong, H., Jiang, S., Du, G., Wang, J., Sun, Y., Huang, X., Ding, Y., et al. (2021). Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. Brief. Bioinform. *22*, bbaa405. https://doi.org/10.1093/bib/bbaa405.
8. Fortin, J.P., and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. *16*, 1–23. https://doi.org/10.1186/s13059-015-0816-9.

9. Moore, B.L., Aitken, S., and Semple, C.A. (2015). Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. *16*, 110–114. https://doi.org/10.1186/s13059-015-0730-1.
10. Raineri, E., Serra, F., Beekman, R., García Torre, B., Vilarrasa-Blasi, R., Martin-Subero, I., Martí-Renom, M.A., Gut, I., and Heath, S. (2018). Inference of genomic spatial organization from a whole genome bisulfite sequencing sample. Preprint at bioRxiv. https://doi.org/10.1101/384578.
11. Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A.P. (2017). Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat. Genet. *49*, 719–729. https://doi.org/10.1038/ng.3811.
12. Stilianoudakis, S.C., Marshall, M.A., and Dozmorov, M.G. (2022). preciseTAD: a transfer learning framework for 3D domain boundary prediction at base-pair resolution. Bioinformatics *38*, 621–630. https://doi.org/10.1093/bioinformatics/btab743.
13. Al Bkhetan, Z., and Plewczynski, D. (2018). Three-dimensional Epigenome Statistical Model: Genome-wide Chromatin Looping Prediction. Sci. Rep. *8*, 5217. https://doi.org/10.1038/s41598-018-23276-8.
14. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. https://doi.org/10.1038/nmeth.1906.
15. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476. https://doi.org/10.1038/nmeth.1937.
16. Di Pierro, M., Cheng, R.R., Lieberman Aiden, E., Wolynes, P.G., and Onuchic, J.N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. SA *114*, 12126–12131. https://doi.org/10.1073/pnas.1712577114.
17. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380. https://doi.org/10.1038/nature11082.
18. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.
19. Gu, H., Harris, H., Olshansky, M., Mohajeri, K., Eliaz, Y., Kim, S., Krishna, A., Kalluchi, A., Jacobs, M., Cauer, G., et al. (2021). Fine-mapping of nuclear compartments using ultra-deep Hi-C shows that active promoter and enhancer elements localize in the active A compartment even when adjacent sequences do not. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.05.455340.
20. Nichols, M.H., and Corces, V.G. (2021). Principles of 3D compartmentalization of the human genome. Cell Rep. *35*, 109330. https://doi.org/10.1016/j.celrep.2021.109330.
21. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.
22. Sefer, E. (2021). Hi–C interaction graph analysis reveals the impact of histone modifications in chromatin shape. Appl. Netw. Sci. *6*, 1–19. https://doi.org/10.1007/s41109-021-00405-1.
23. Ashoor, H., Chen, X., Rosikiewicz, W., Wang, J., Cheng, A., Wang, P., Ruan, Y., and Li, S. (2020). Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. Nat. Commun. *11*, 1173–1184. https://doi.org/10.1038/s41467-020-19832-7.
24. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Preprint at arXiv. https://doi.org/10.1101/2021.08.05.455340.
25. Schreiber, J., Singh, R., Bilmes, J., and Noble, W.S. (2020). A pitfall for machine learning methods aiming to predict across cell types. Genome Biol *21*, 1–6. https://doi.org/10.1186/s13059-020-02177-y.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Hi-C data	GEO	GEO:GSE63525
Software and algorithms		
CoRNN	This paper	https://github.com/rsinghlab/CoRNN
Python v3.7.1	Python	https://www.python.org/
NumPy v1.17.0	NumPy	https://numpy.org/
tqdm v4.56.0	Tqdm	https://tqdm.github.io/
PyTorch v1.2.0	PyTorch	https://pytorch.org/
scikit-learn v0.0	scikit-learn	https://scikit-learn.org/stable/
Matplotlib v3.3.4	Matplotlib	https://matplotlib.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact - Ritambhara Singh, ritambhara@brown.edu.

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data: All Hi-C data are publicly available under GEO:GSE63525- GSE63525_NHEK_combined_30.hic, GSE63525_K562_combined_30.hic, GSE63525_IMR90_combined_30.hic, GSE63525_HUVEC_combined_30.hic, GSE63525_HMEC_combined_30.hic, GSE63525_GM12878_*insitu*_primary_replicate_combined_30.hic.
- Code: Our code is publicly accessible at https://github.com/rsinghlab/CoRNN.
- Other: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data preprocessing

For training the model, we selected Hi-C and histone modification ChIP-seq experiments for six cell lines: (1) NHEK, (2) IMR90, (3) HMEC, (4) GM12878, (5) K562, and (6) HUVEC. We predicted the A/B compartments for each cell line (as a test set) by training the CoRNN model on the other five cell lines (training set). To generate the input using histone modification signals, we divided the chromosome into 100 kilo-base pair (kbp) regions and binned each region into 100 bins of size 1000 bp. For each bin, we calculated the average histone modification ChIP-seq signal. We chose the following six histone modification marks: H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K9me3, and H3K27me3. Next, we obtained the A/B compartment values by calculating the first-order eigenvectors of the Hi-C matrix (at 100 kbp resolution) for each cell line. We formulated the A/B compartment prediction as a binary classification problem. Therefore, we assigned output labels 1 (A compartment) and 0 (B compartment) to positive and negative compartment values, respectively. Finally, we included the histone modification marks and compartment values from the colon and muscle tissue samples taken from a consented individual (accession code ENCDO845WKR) in the ENCODE portal (entex.encodeproject.org) as an independent test set to demonstrate the generalizability of our model.

Input and output formulation for the prediction task

Figure 2 shows an example input sample (representing a 100 kbp genomic region) denoted as a matrix $X \in \mathbb{R}^{m \times t}$. Here, $m = 6$ denotes the number of histone marks, and $t = 100$ is the number of genomic bins. Note that while we are making predictions at a low resolution (100kbp), we incorporate the histone modification signal at a higher resolution (1kbp) through our binning strategy and choice of architecture. Therefore, the model looks at a high-resolution histone modification signal (instead of a mean average value in the 100kbp region) for making A/B compartment predictions. We input matrix X and scalar c, which is the mean compartment value for the training cell lines for each genomic region, and predict its compartment. Our results show that including the mean compartment value (only from training cell lines with available Hi-C information) boosts the performance of our model on the test cell line (without available Hi-C information). The output $y \in [0, 1]$ represents the binarized compartment value for the input genomic region.

CoRNN architecture

CoRNN is an end-to-end A/B compartment prediction model (Figure 2). It consists of three main components.

Gated recurrent units (GRUs)

Gated recurrent units (GRUs) are a variation of the traditional recurrent neural network.[24] GRUs can capture long-range sequential information from the input samples. We also tested a convolutional neural network (CNN), but it did not perform as well as the GRU (Figure S4). Therefore, in our setting, we hypothesize that a GRU layer effectively models the sequential dependency of the histone marks in consecutive bins across the genome, resulting in better performance. Given our input matrix X, GRUs take in one input column x_t (with all six histone marks) at a time. Together with the hidden state h_{t-1} from the previous time step, GRUs generate the current hidden state h_t as the input to the next time step.

GRUs first calculate the update gate z_t for time step t using $z_t = \sigma(W^{(z)}x_t + U^z h_{t-1})$,where current input x_t is multiplied by its weight $W^{(z)}$, and the hidden state h_{t-1} from the previous time step is multiplied by its weight U^z. These two values are added and input to a sigmoid activation function, σ, to constrain the result between 0 and 1. The update gate function acts as the long-term memory of the network. It determines how much past information will be passed down to the next step: $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$. GRUs also have a reset gate to determine the short-term memory of the network, that is, how much information to discard, using the following formula: $r_t = \sigma(W^{(r)}x_t + U^r h_{t-1})$. Next, GRUs determine the current memory content by applying the output of the reset gate r_t to the hidden state from the previous time step h_{t-1}. This step uses an element-wise product between r_t and Uh_{t-1}. The current input x_t is multiplied by weight W. These values are added together and inputted to the *tanh* activation function: $h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1})$. Finally, GRUs calculate the h_t using the following formula: $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t$. When the GRU sets z_t close to 1, it will retain most of the information from the previous hidden state h_{t-1}. Since $(1 - z_t)$ will be close to 0, the model will ignore most of the current content from h'_t. We use GRUs to learn the representation of the histone modification signals. The number of GRU layers and the size of the hidden units are hyperparameters of the model (Table S1). When incorporating multiple layers, only the first GRU layer takes the original histone modification signals as input. The subsequent layers take the hidden state outputs from the previous layer as input. The output of the last hidden unit, h_{100}, of the final GRU layer, concatenated with mean compartment value c, goes into the next component of the model, the fully connected network.

Fully connected network (FCN)

This network consists of two fully connected layers. It takes the last hidden state of the GRU and the mean compartment value as inputs and generates an output vector of size two. By concatenating the mean compartment value to the GRU's output, we enable CoRNN to leverage information from histone modification signals and the compartment consensus of other cell lines in the training set.

Concatenating histone modifications with mean compartment values results in a vector of size $h_{100} + 1$ as the input to the fully connected network. Therefore, the output of this network can be written as: $h_{fc} = W_2(W_1[h_{100} \parallel c] + b_1) + b_2$. Here, $\parallel$ represents concatenation, and c represents the mean compartment value. W_1 and b_1 represent the learnable weight and bias parameters of the first fully connected layer, and W_2 and b_2 represent the learnable weight and bias parameters of the second fully connected layer.

Softmax function

Finally, a softmax function is applied to h_{fc}. We formulate the A/B compartment prediction as a binary classification task with classes $y \in \{0,1\}$, corresponding to whether a chromosome region is in A (active) compartment ($y = 1$) or B (inactive) compartment ($y = 0$). The softmax function takes in the output value from the fully connected network and computes the probability of each class y. We use the cross-entropy loss for the predicted probability of the true label to train the model's weights. Details of experimental setup

Correct compartment value assignments We calculated the correlation coefficient between the compartment values and the H3K4me3 ChIP-seq signals to correct the signs of the compartments. H3K4me3 has been observed to be positively correlated with the A/B compartment values.[3] Therefore, we flipped the sign of the compartments if the correlation coefficient was negative.

Handling the missing signals

We eliminated regions with a missing compartment value and imputed input for regions with missing histone modification signals. For example, for NHEK, the H3K27me3 and H3K36me3 experiments were missing. Similarly, for GM12878, the H3K9me3 experiment was missing. We imputed the missing histone modification values using the average ChIP-Seq signal across other cell lines.

End-to-end training

Out of the six selected cell lines, we iteratively chose one cell line as the test set and the other five cell lines as the training set. For each iteration, we performed hyperparameter tuning over the following grid of values to pick the best model architecture: the size of hidden state $\in \{32, 64, 128\}$ and the number of GRU layers $\in \{1,2,3,4\}$. We selected the best-performing hyperparameters using a five-fold cross-validation scheme. We selected one cell line from the training set as a validation cell line for the current fold. Then we trained the model on the remaining four cell lines. We did this for each fold and obtained the average validation performance from the five folds. Finally, the best average validation performance model was used to make final predictions for the held-out test cell line for that iteration. We present our cross-validation scheme in Figure S3. The same setup was used for the baseline models that we describe in the next section.

QUANTIFICATION AND STATISTICAL ANALYSIS

Baseline method details

We chose the following baseline methods for our evaluations. It was previously shown that in genomic activity prediction tasks, a machine learning model might falsely appear to perform well by effectively memorizing the average activity associated with each locus across the training cell types.[25] The mean compartment value baseline (or the mean baseline) uses the average compartment values across cell lines in a training set as a proxy for the A/B compartment prediction in the test cell line. Since most of the compartments are conserved across different cell lines, the mean baseline's predictions can achieve a performance that is difficult to beat. Previous studies[9,13] use a random forest model to predict chromatin structures using a variety of histone modifications and transcription factors. We included a similar Random Forest model as one of the baselines. Rainer et al.[10] proposed a logistic regression framework to predict A/B compartments from the GC content of the sequence and DNA methylation. Following their setup, we trained a logistic regression model to include as one of our baselines. See details below for implementation of baseline methods.

Mean compartment value baseline

First, we binarize the compartment values to 1 or 0 based on positive and negative values. Next, we take the average of the five binarized compartment values. Since the training set comprises five cell lines, the predictions made by the mean baseline will have the following values: 0, 0.2, 0.4, 0.6, 0.8, and 1.0. A mean compartment value close to 0 or 1 for a genomic bin indicates that the compartment value is more consistent in this region across all five training cell lines, showing that this is a more conserved region. Similarly, a mean compartment value of around 0.5 means the compartment value varies across different cell lines and represents a less conserved region.

Random forest

The hyperparameter tuning was performed on the number of trees in the forest, the maximum depth of the tree, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. For the model input, we calculated the mean and standard deviation for each of the six histone modification signals in the 100 kbp region for input features. Mean and standard deviation values from six histone modification signals made up an input vector of length 12. To keep this model consistent with our framework and for a fair comparison, we concatenate the mean compartment value at the end of the input vector. We also tried using all of the 6×100 features as input to train the Random Forest model. However, the performance of this model was worse than using mean and standard deviation values (Figure S5).

Logistic regression

The input to the model was the mean and standard deviation of six histone modification signals in the 100 kbp region combined with the mean compartment value of the region across the training cell lines. We performed hyperparameter tuning of the model for the norm of the penalty ($l1, l2$), the C value (inverse of regularization strength), type of solver ($newton - cg, lbfgs, liblinear, sag, saga$), and the maximum number of iterations taken for the solvers to converge. We also tried using all of the 6×100 features as input to train the logistic regression model. However, similar to the Random Forest model, the performance using all 600 features was not as good as using the mean and standard deviation of the signals (Figure S5).

Evaluation metrics

Area under the receiver operating curve (AUROC)

We use the area under the receiver operating characteristic (AUROC) score as our evaluation metric since the number of samples in our two classes is roughly balanced (Table S2). The AUROC score evaluates the classifier's ability to distinguish two classes. It measures the probability that a random positive sample will be ranked higher than a randomly selected negative sample. The AUROC score ranges between 0 and 1, where values closer to 1 indicate a more successful classifier. However, we have also included the area under the precision-recall curve (AUPRC) scores to evaluate the classification performance in the Appendix for completeness.