# Epigenetics is all you need: A Transformer to decode chromatin structural compartments from the epigenome

Esteban Dodero-Rojas
Center for Theoretical Biological Physics,
Rice University,
Houston, Texas, USA

Vinícius G. Contessoto
Center for Theoretical Biological Physics,
Rice University,
Houston, Texas, USA

Yao Fehlis
Research and Advanced Development,
Advanced Micro Devices,
Austin, Texas, USA

Nicolas Mayala
Research and Advanced Development,
Advanced Micro Devices,
Austin, Texas, USA

José N. Onuchic
Center for Theoretical Biological Physics,
Department of Physics & Astronomy,
Department of Chemistry,
Department of Biosciences,
Rice University,
Houston, Texas, USA

June 11, 2024

### Abstract

Chromatin within the nucleus adopts complex three-dimensional structures that are crucial for gene regulation and cellular function. Recent studies have revealed the presence of distinct chromatin subcompartments beyond the traditional A/B compartments (eu- and hetero-chromatin), each exhibiting unique structural and functional properties. Here, we introduce TECSAS (Transformer of Epigenetics to Chromatin Structural AnnotationS), a deep learning model based on the Transformer architecture, designed to predict chromatin subcompartment annotations directly from epigenomic data. TECSAS leverages information from histone modifications, transcription factor binding profiles, and RNA-Seq data to decode the relationship between the biochemical composition of chromatin and its 3D structural behavior. TECSAS achieves high accuracy in predicting subcompartment annotations and reveals the influence of long-range epigenomic context on chromatin organization. Furthermore, we demonstrate the model's capability to predict the association of loci with nuclear bodies, such as the lamina, nucleoli, and speckles, providing insights into the role of these structures in shaping the 3D genome organization. This study highlights the potential of deep learning models for deciphering the complex interplay between epigenomic features and 3D genome organization, allowing us to better understand genome structure and function.

## 1 Introduction

Within the eukaryotic cell nucleus, the genome folds into three-dimensional structures that vary depending on cell type and stage of development [1]. These architectural features play a crucial role in regulating gene expression, and disruptions in this organization have been linked to various diseases[2, 3, 4, 5]. Over the past decade, DNA-DNA proximity ligation assays, such as Hi-C[6, 7, 8, 9, 10], have enabled the systematic study of genome organization by measuring the frequency of chromatin contacts throughout the genome. Hi-C experiments have revealed that chromatin segregates into regions with preferential long-range interactions, known as compartments[10]. A-type compartments are gene-rich and associated with active and less dense chromatin (euchromatin). These compartments are enriched with proteins like RNA polymerase and specific histone modifications, such as H3K4me3. In contrast, B-type compartments are gene-poor and linked to inactive and more dense chromatin (heterochromatin). They are often associated with the enrichment of different histone modifications, such as H3K9me3 and H3K27me3[11].

High-resolution Hi-C experiments have revealed that chromatin exhibits finer compartmentalization than the A and B[11]. For instance, within B compartments, specific regions are prone to interact with the nuclear lamina or nucleoli. These observations led to the concept of subcompartments, which further classify chromatin based on distinct structural and functional properties. Rao et al. (2014)[11] demonstrated that five subcompartments (A1, A2, B1, B2, and B3) effectively capture the structural heterogeneity observed in Hi-C experiments on the human lymphoblastoid cell line GM12878. Each subcompartment exhibits a unique enrichment profile of epigenetic marks, such as histone modifications. For example, B2 and B3 subcompartments show depletion of most histone modifications, while B1 shows neither depletion nor enrichment of histone modifications except for H3K27me3. Additionally, subcompartment identity correlates with the binding of specific nuclear lamina and nucleoli-associated proteins, suggesting a link between structural diversity and interactions with nuclear bodies[12, 8].

The identification of chromatin compartments and subcompartments has initially relied on the analysis of Hi-C data, which provides information about the spatial proximity of genomic regions. Which has layout the basis of multiple chromatin theoretical models [13, 14, 15, 16, 17]. Several computational methods have been developed to classify regions of the genome into these structural categories based on patterns observed in Hi-C contact maps [18, 19, 20]. For example, the SNIPER method focuses on predicting subcompartments from moderate-coverage Hi-C data by imputing inter-chromosomal contacts[18].

The algorithm Calder uses the Hi-C intrachromosomal interactions to identify multi-scale chromatin subcompartments and compartment domains that enable analysis at variable data resolutions[19]. Recent efforts have focused on linking epigenetic information, such as histone modifications and transcription factor binding, to chromatin compartments and subcompartments labeling. Similar to Calder, a deep learning method, called SLICE, generates subcompartment annotations at 25, 50 and 100 kilobase resolution from Hi-C maps. Interestingly, SLICE provides structural annotations ranging from 2 to 12 possible states [21]. Additionally, the reliance on Hi-C data for training or validation in these methods restricts their applicability to cell types with available Hi-C experiments. Therefore, developing methods to predict chromatin organization directly from epigenomic data is crucial for expanding our understanding of 3D genome structure across diverse cell types.

Though trained partially on Hi-C map information, CoRNN [20], a deep learning model based on recurrent neural networks, utilizes histone modification data to predict A/B compartments in different cell lines. Epiphany tool also employs a deep learning model to predict cell-type-specific Hi-C contact maps from 1D epigenomic signals, which could be used to label compartments and subcompartments for each locus[22]. Additionally, based only on the epigenome data and not using Hi-C maps, PyMEGABASE (PYMB) uses ChIP-Seq from Histone Modification and Transcription factor, and RNA-Seq to predict compartments and subcompartments for hundreds of cell types [23]. PYMB's interpretable predictions and transferability across cell types and species further demonstrate the potential of data-driven models for understanding 3D genome organization. Notwithstanding, PYMB is based on the MEGABASE framework[24]. MEGABASE uses a physics-based approach similar to the Potts Model that builds an energy function focused on the association between epigenetic marks and subcompartments [24, 25, 16, 17]. The potential for using the complex interplay between multiple epigenetic marks to identify structural annotations has not been explored.

This study introduces a novel approach to predicting chromatin's structural annotations based on the epigenome (e.g., histone modifications, transcription factor binding, RNA expression). We introduce TECSAS (Transformer of Epigenetics to Chromatin Structural AnnotationS), a deep learning model that leverages the power of Transformers and Attention layers to capture complex relationships between various epigenetic marks and predict subcompartment annotations with high accuracy[26]. Unlike other methods that rely on Hi-C data for training or validation, our approach focuses solely on epigenetic information. This allows us to predict subcompartment annotations even in cell types where Hi-C data is unavailable. Additionally, our results demonstrate that TECSAS versatility allows for the prediction of additional structural features, such as the association of loci with nuclear bodies like the lamina, nucleoli, and speckles, by simply fine-tuning the final layer of the model. TECSAS flexibility enables the exploration of diverse aspects of 3D genome organization using a single, unified framework.

# 2 Results

## 2.1 TECSAS predicts subcompartments by decoding the loci context of the epigenetic profile

This study introduces TECSAS (Transformer of Epigenetics to Chromatin Structural Annotations), a deep learning model based on the Transformer architecture, to predict structural information from 1D epigenomic data. Figure 1 summarizes the workflow of TECSAS. The model takes as input the signal intensity of various epigenomic features locus-wise, including RNA-seq, histone modification ChIP-seq, and transcription factor ChIP-seq experiments. This diverse data set represents the DNA's biochemical composition and transcriptional activity. To provide context and capture long-range dependencies along the genome sequence, each locus is characterized by the signal intensity of these epigenomic features within a defined neighborhood of N loci upstream and downstream. We refer to this combined input as the ”epigenomic profile” of the locus. TECSAS aims to learn the relationship between a locus' structural annotation and its corresponding epigenomic profile. Initially, we use the subcompartment annotations from GM12878 derived from Hi-C maps [27] as the target structural information to train TECSAS. The genome is segmented into train, test, and validation sets representing 80%, 10%, and 10% of all the loci, respectively (See SI for details).

The primary output of TECSAS is the prediction of chromatin compartments (A and B) and subcompartments (A1, A2, B1, B2, and B3) for each genomic locus, generating genome-wide structure annotations. However, the model's flexibility allows for predicting additional structural features by modifying the target data and fine-tuning the final layer. For example, by utilizing appropriate training datasets, TECSAS can be adapted to predict the association of loci with specific nuclear bodies, such as Lamina-Associated Domains (LADs), Nucleolus-Associated Domains (NADs), and Speckle-Associated Domains (SPADs). This adaptability makes TECSAS a versatile tool for exploring various aspects of 3D genome organization.

As shown in Figure 1B, the TECSAS architecture consists of several key components:

- **Input Embedding:** The input epigenomic profile is first processed through a linear embedding layer, which transforms the data into a higher-dimensional representation suitable for the Transformer encoder.
- **Positional Encoding:** Positional encoding is added to the embedded input to incorporate information about the relative positions of the loci within the epigenomic profile.
- **Transformer Encoder:** The core of the model is a Transformer encoder with multiple attention heads. The encoder uses self-attention mechanisms to learn complex relationships and dependencies between different epigenomic features across the input loci. This allows the model to capture the context and long-range interactions that influence chromatin structure.
- **Linear and SoftMax Output:** The output of the Transformer encoder is passed through a linear layer, followed by a softmax activation function. The softmax layer outputs a probability distribution over the possible structural annotations, allowing the model to assign the most likely annotation to each locus.

To evaluate the performance of TECSAS in predicting chromatin subcompartments, we initially trained the model using the well-characterized subcompartment annotations for the GM12878 cell line derived from Hi-C maps. The model utilized a comprehensive set of epigenomic data from the ENCODE portal, including 11 histone modification ChIP-seq tracks, total and small RNA-seq data, and 140 transcription factor ChIP-seq tracks. For each locus, the input consisted of the signal intensity of these epigenomic features within a 14-locus neighborhood (7 upstream and 7 downstream), capturing the local epigenomic context. The specific hyperparameters used for training TECSAS, such as the number of encoder layers, attention heads, and training epochs, are described in detail in the Methods section. Figure 2A presents the confusion matrix for TECSAS predictions of subcompartments in GM12878, demonstrating high accuracy across all subcompartments. The model achieved an overall accuracy of 0.78, with individual subcompartment accuracies ranging from 0.68 to 0.81. This indicates that each subcompartment possesses a distinct epigenomic signature that TECSAS can effectively learn. Furthermore, the model accurately predicted A/B compartments based on the inferred subcompartments, achieving accuracies of 0.87 and 0.93 for A and B compartments, respectively (Figure 2B). This suggests that the epigenomic profiles of A and B compartments are sufficiently distinct to allow for accurate classification. It is important to mention that the overall accuracy between the compartments extracted from different

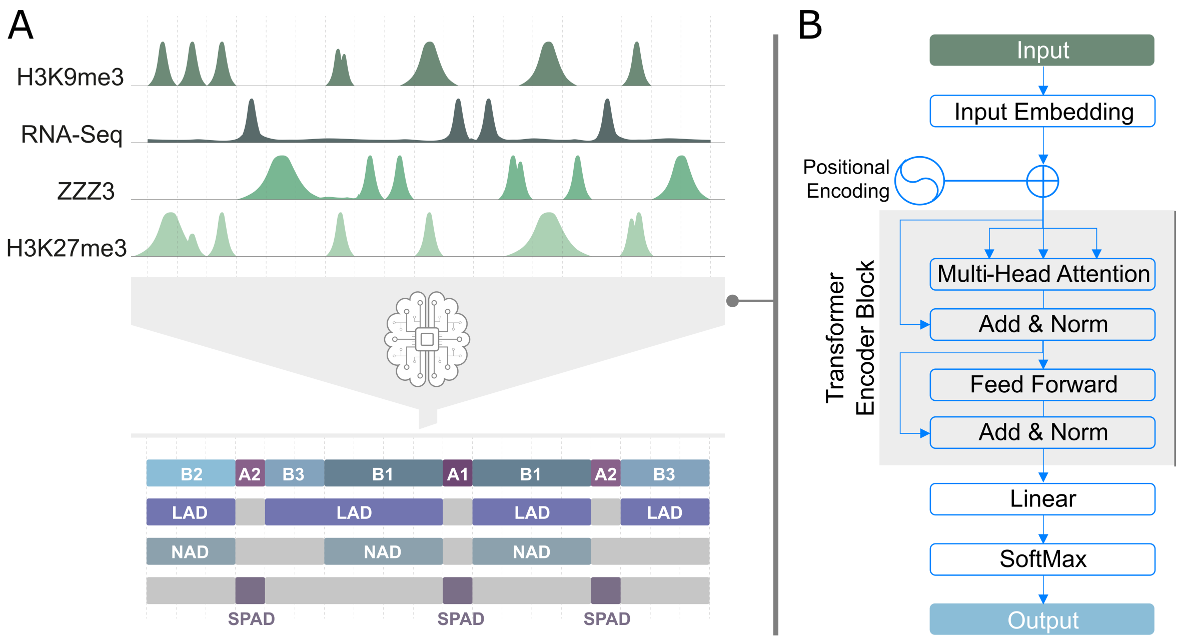

Figure 1: **TECSAS Workflow for Predicting Chromatin Structure from Epigenomic Profiles.** (A) Diverse 1D epigenetic tracks (RNA-seq, histone modifications, transcription factor binding) are extracted from the ENCODE portal and segmented into 50kbp loci. The TECSAS deep learning architecture predicts locus-wise structural annotations, including compartments, subcompartments, and potentially other features like LADS, NADS, and SPADS. Prediction is based on learned correlations within the locus's biochemical composition. (B) The TECSAS architecture begins with an input embedding layer, transforming the epigenomic profile into a higher-dimensional representation. A Transformer encoder then analyzes this representation, capturing complex relationships and long-range dependencies within the epigenomic data to understand the structural context. Finally, the output is decoded through a linear layer and a softmax layer, assigning a probability distribution over possible structural annotations for each locus.

Hi-C methodologies is ≈0.95 (Figure **??**), which means that TECSAS is close to reach the experimental replicate accuracy limit.

TECSAS uses a softmax output layer, which means the output of each node can be related as a probability [28]. Each node represents a different subcompartment. The model predicts the subcompartment for a locus by selecting the node with the highest probability. In this case, a high probability can be interpreted as a proxy for model confidence – we call it the "confidence probability." Figure 2C shows that TECSAS has high confident when predicting the B3 and A1 subcompartments. However, it exhibits lower confidence when predicting the B1 subcompartment. Interestingly, previous research has shown that the B1 subcompartment lacks strong defining characteristics (like specific histone modifications or nuclear body associations [11]). This suggests that B1 has a more complex or less distinct epigenetic profile, making it harder for the model to confidently predict it.

To further assess the model's ability to predict structural annotations at higher resolutions, we trained TECSAS using a set of subcompartments derived from the K562 cell line at 25 kb resolution using the SLICE method. Despite the increased resolution and a smaller set of input features (124 ChIP-seq experiments), TECSAS achieved a higher overall accuracy of 0.80 in predicting the four K562 subcompartments (Figure 2D). This suggests that the structural annotations derived from SLICE possess identifiable epigenomic profiles, further supporting the link between chromatin's biochemical composition and its 3D organization. This finding also highlights the potential of TECSAS to be applied to higher-resolution data, enabling a more detailed analysis of chromatin structure. Additionally, we ob-

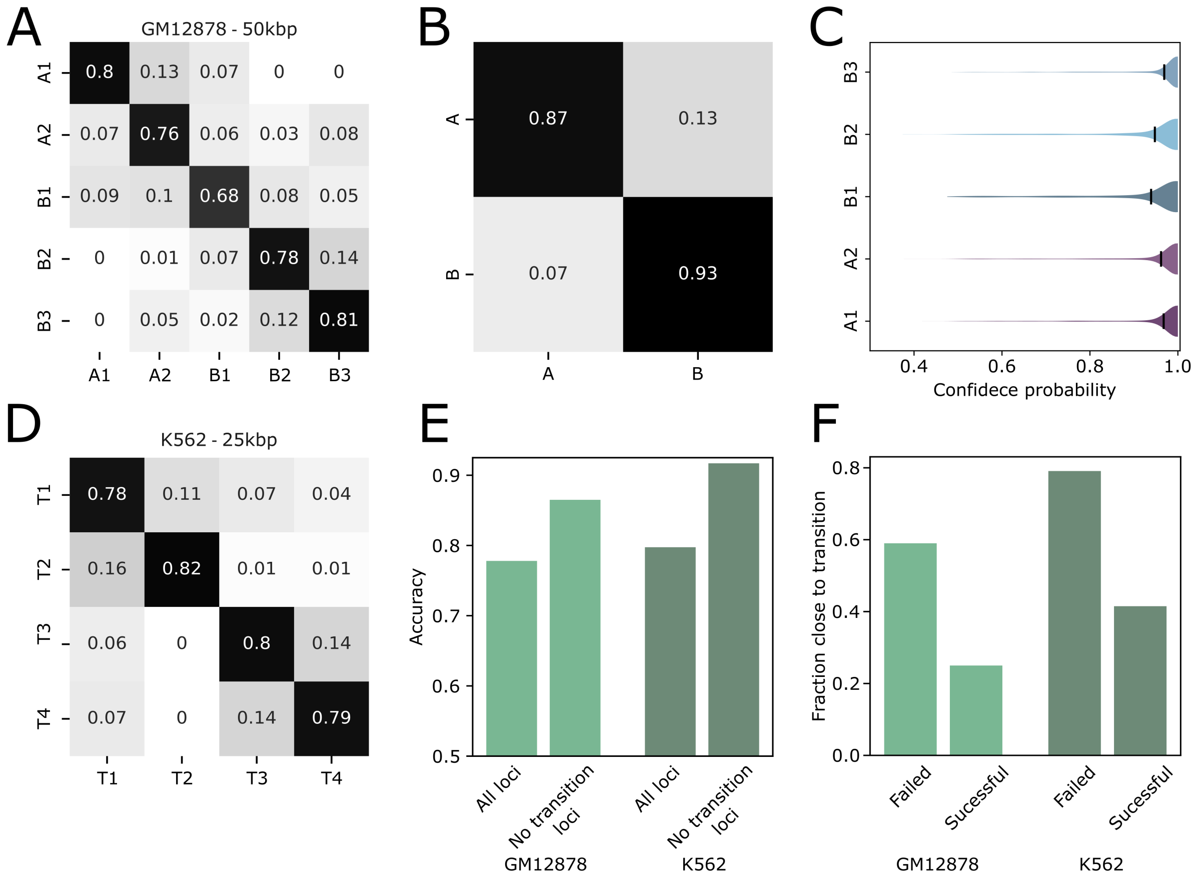

Figure 2: **Assessment of TECSAS prediction at 50kbp and 25kbp resolution for GM12878 and K562 cell lines.** (A) Confusion matrix comparing TECSAS predictions with experimentally derived subcompartment annotations for the GM12878 cell line at 50 kb resolution. The diagonal elements represent the fraction of correctly predicted loci for each subcompartment, highlighting the high accuracy of the model. (B) Confusion matrix for A/B compartment predictions based on the inferred subcompartments in GM12878, demonstrating accurate compartment classification. (C) Distribution of confidence probabilities for each predicted subcompartment in GM12878. B1 and B2 subcompartments exhibit lower average confidence probabilities, reflecting their more complex epigenomic profiles. (D) Confusion matrix comparing TECSAS predictions with subcompartment annotations derived using the SLICE method for the K562 cell line at 25 kb resolution, demonstrating the model's ability to predict subcompartments at higher resolutions. (E) Overall accuracy of TECSAS in predicting subcompartments for GM12878 and K562, comparing performance for all loci and loci excluding transition regions. The exclusion of transition regions significantly improves prediction accuracy for both cell lines. (F) Fraction of successful and failed predictions within transition regions for GM12878 and K562, highlighting the challenges of predicting subcompartments in these regions with mixed epigenomic signatures.

served that TECSAS predictions were less accurate in ”transition regions” between subcompartments, defined as regions within four loci of a subcompartment boundary. These regions likely exhibit a mixed epigenomic signature, making it challenging for the model to assign a definitive subcompartment annotation. When excluding these transition regions from the analysis, the prediction accuracy for both GM12878 and K562 subcompartments increased significantly to 0.87 and 0.92, respectively (Figure 2E and 2F). This highlights the importance of considering the gradual nature of epigenomic changes across the genome and the potential for fuzzy structural behavior in transition regions.

## 2.2 Context of the epigenetic profile contributes to the prediction of subcompartments

To investigate the contribution of the epigenomic context to subcompartment prediction accuracy, we compared the performance of TECSAS with PyMEGABASE (PYMB), another method that also only utilizes epigenomic data for predicting structural annotations. We first adapted TECSAS to use the same input as PYMB, which consists of discretized signals from histone modification ChIP-seq and RNA-seq experiments, including only two neighboring loci upstream and downstream of the target locus. As shown in Figure 3A, this simplified version of TECSAS achieved an accuracy of 0.62, which is lower than the original TECSAS model (0.78) but still higher than the accuracy of PYMB (0.57). Further, we compare the accuracy of TECSAS with more experiments, Figure 3B shows that regardless of the number of experiments used TECSAS outperforms PYMB. This suggests that the Transformer architecture employed by TECSAS contributes to improved prediction accuracy even when using a limited epigenomic context.

Further analysis revealed that increasing the number of neighboring loci included in the input improves the performance of TECSAS. As shown in Figure 3C, the accuracy of the model increases from approximately 0.63 with two neighbors to 0.74 with 18 neighbors, indicating that the structural behavior of a locus is influenced by the epigenomic landscape of a broader genomic region extending up to 900 kb. This observation highlights the importance of capturing long-range dependencies and interactions within the epigenome for accurate prediction of chromatin organization. Notably, the improvement in accuracy with a larger epigenomic context was particularly pronounced for B1 and A1 subcompartments, suggesting that these subcompartments may be more sensitive to the epigenetic state of their surrounding regions. Interestingly, as shown in Figures **????**, the average epigenetic profile of each subcompartment have a different decay over genomic distance from the locus of interest. The difference between epigenetic profiles between A1 and A2 show how these subcompartments can be identified using the decay of some of the epigenetic marks, similar trend is observe between B1 and B2/3 (Figure **??**).

To further understand how TECSAS leverages the epigenomic context for subcompartment prediction, we examined the attention maps generated by the model. Unlike PyMEGABASE, which is based on a Potts model and captures only pairwise relationships between epigenetic marks and subcompartments, TECSAS utilizes a Transformer architecture with self-attention mechanisms. One can consider that the attention layer mechanism can be related to the coupling matrix $J_{ij}$ presented in the Potts model, although a direct comparison is not totally straight[28]. This allows the model to learn complex, higher-order interactions between multiple epigenetic features across different genomic loci. The attention maps provide a visual representation of these interactions, highlighting the regions of the epigenomic profile that are most relevant for predicting the subcompartment annotation of a given locus.

Figure 3D illustrates a subset of the self-attention map for a locus predicted as A1. The attention map reveals the model's focus on specific patterns within the epigenomic profile. For instance, the enrichment of H3K36me3, a histone modification associated with active transcription, is captured by the activation of nodes corresponding to the H3K36me3 signal. Similarly, the attention map also highlights the depletion of repressive histone marks like H3K9me3 and H3K27me3 downstream of the locus. This demonstrates how TECSAS utilizes the interplay between different epigenetic marks at various genomic distances to inform its predictions.

Further examination of the attention maps reveals the ability of TECSAS to capture long-range interactions within the epigenome. Figure 3E illustrates the full attention map for a locus predicted as B1. While the immediate neighboring loci exhibit relatively low enrichment of epigenetic marks, the attention map highlights the importance of more distal regions, particularly loci $L + 7$, $L + 8$, and $L + 9$, which show higher enrichment of specific histone modifications. This suggests that the structural

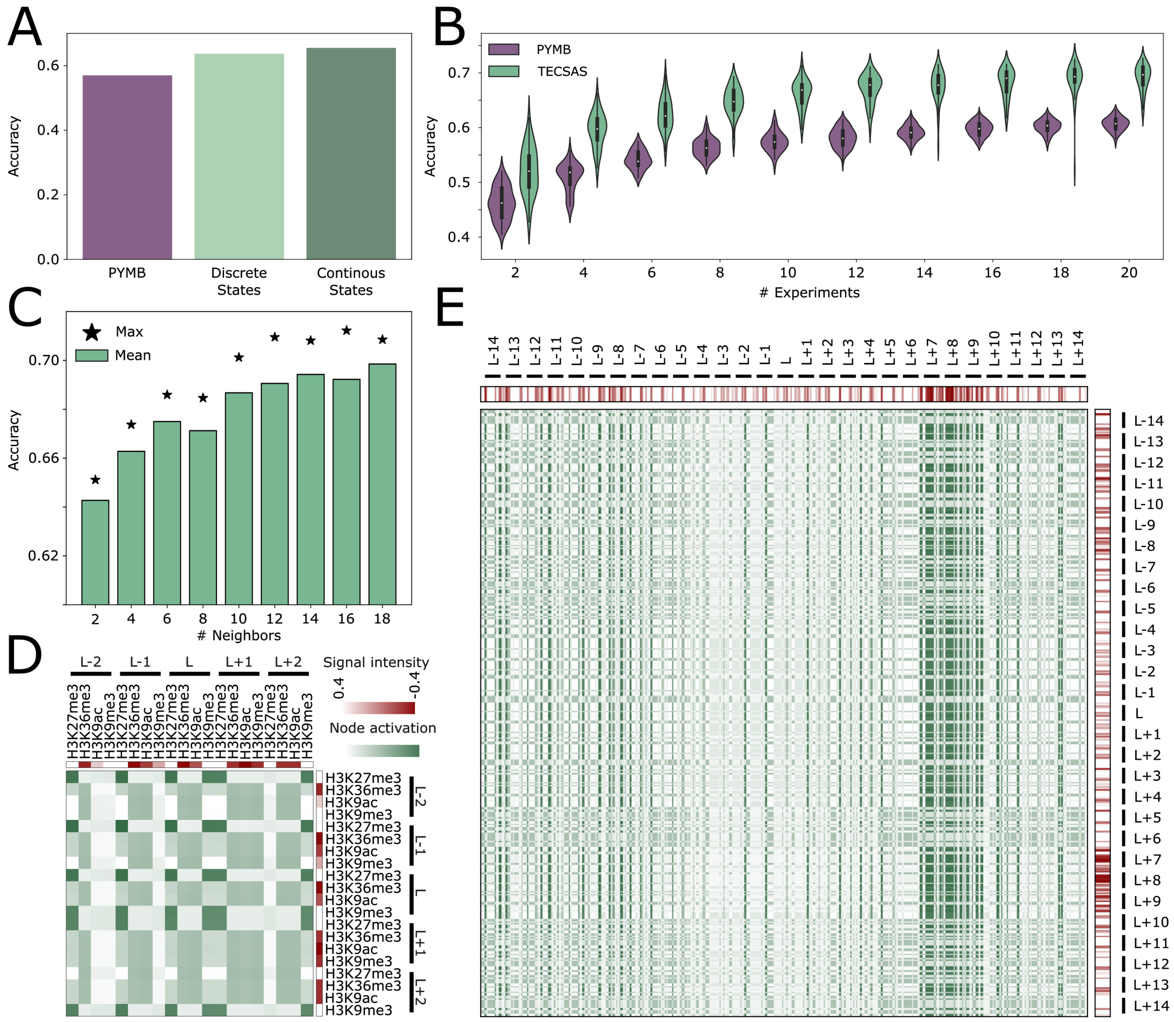

Figure 3: **The importance of epigenomic context and long-range interactions for accurate subcompartment prediction with TECSAS.** (A) Comparison of overall accuracy in predicting subcompartments between PyMEGABASE (PYMB) and TECSAS using both discretized and continuous signal intensities for epigenomic features. TECSAS demonstrates higher accuracy even with a limited epigenomic context. (B) Prediction accuracy as a function of the number of input experiments for both PYMB and TECSAS, highlighting the consistent outperformance of TECSAS regardless of the number of features used. (C) Mean accuracy of subcompartment predictions with increasing numbers of neighboring loci included in the input, demonstrating the significant improvement in accuracy as the epigenomic context expands. The maximum accuracy achieved is indicated by a star. (D) Subset of the attention map for a locus predicted as A1, showing the activation of nodes (green) corresponding to specific epigenomic features (red) and highlighting the model's focus on relevant patterns within the local epigenomic context. (E) Full attention map for a locus predicted as B1, revealing the importance of long-range interactions and the model's attention to distal regions with enriched epigenetic marks, particularly for marks ≈350kbp apart from the loci of interest ($L$).

annotation of a locus can be influenced by the epigenomic landscape of regions located several hundred kilobases apart. The capability to capture these long-range interactions is a key advantage of TECSAS over methods like PYMB, which utilize a more localized epigenomic context and may not fully capture the influence of distal regulatory elements on chromatin organization. By incorporating information from a broader genomic region, TECSAS gains a more comprehensive understanding of the factors that contribute to the 3D structure of chromatin.

Moreover, we expect different 3D behavior of regions where PYMB and TECSAS differ. We explored this possibility by using the OpenMiChroM software [29] to simulate 3D structural ensembles based on the the predicted IMR-90 subcompartments from PYMB and TECSAS. One region where the prediction is different for both methods is the chromatin segment chr4:36-37Mbp segment. As shown in Figure 4A, PYMB predicts it primarily as A-type; in contrast, TECSAS predicts it as B-type, which aligns with the experimental eigenvector. Interestingly, the radial positioning of this segment is significantly different from their respective 3D ensemble of structures (Figure 4B-C). As expected, the global A and B radial distribution of the chromosome is robust for both sets of predicted annotations (Figure **??**), but local motifs are sensitive to their predicted annotations.

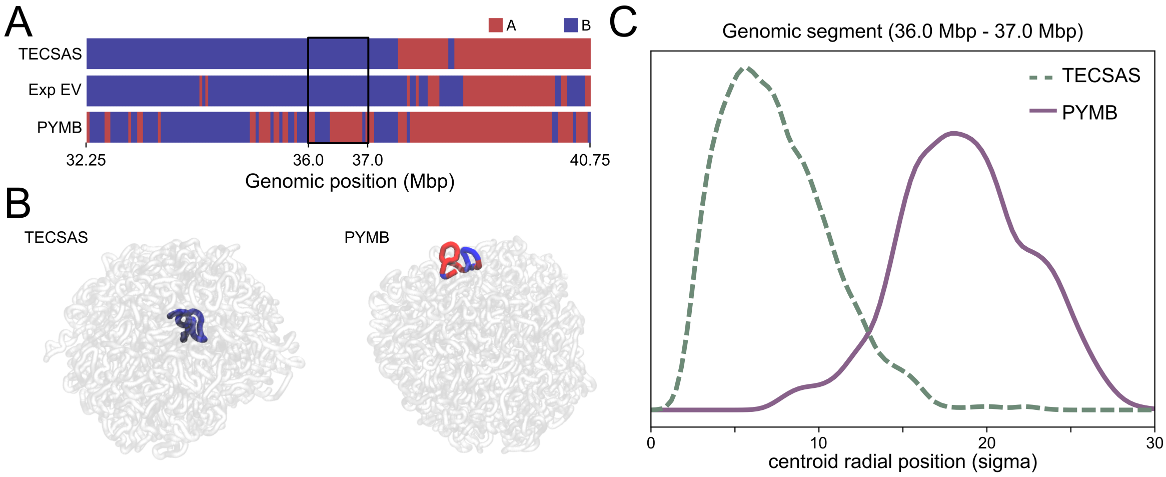

Figure 4: **3D implications of prediction accuracy on IMR-90.** (A) Distribution of radial positioning of the chr4:36-37Mbp segment on the simulated ensemble based on TECSAS and PYMB annotations. (B) Representative structure of chromosome 4 from simulations based on TECSAS and PYMB predictions, highlighting the positioning of the chr4:36-37Mbp segment. (C) Compartment annotations from TECSAS, PYMB and experimental Hi-C around the chr4:36-37Mbp segment.

### 2.3 Fine tuning transformer for functional motifs: NADS, LADS, SPADS, Activity profile

Building upon the ability of TECSAS to learn complex relationships between epigenomic features and chromatin structure, we hypothesized that the model could be adapted to predict additional structural information beyond subcompartments. Specifically, we explored the prediction of associations between genomic loci and specific nuclear bodies, such as the nuclear lamina, nucleoli, and speckles. These associations are often characterized as Lamina-Associated Domains (LADs), Nucleolus-Associated Domains (NADs), and Speckle-Associated Domains (SPADs), respectively. To achieve this, we modified the last linear layer of TECSAS to predict whether a given locus belongs to one of these associated domains or its corresponding negative set (non-LAD, non-NAD, non-SPAD). For training, we utilized LAD and NAD annotations for K562, H1, and HCT116 cells derived from DamID experiments, and SPAD annotations for K562 cells derived from TSA-seq experiments, all obtained from the 4DNucleome Data Portal [30]. All associated domain annotations were provided at a 50 kb resolution.

Given that the Transformer encoder, trained on GM12878 data, effectively interprets epigenomic profiles, we focused on training the last linear layer for each type of associated domain, while freezing the transformer encoder parameters. This involved training the linear layer to map the encoded epigenomic information to the specific structural annotation of interest (e.g., LAD or non-LAD). For this analysis, we used only histone modification ChIP-seq data as input, as these assays are widely available across diverse cell types. Figure 5A presents the prediction accuracy for each associated domain, demonstrating high performance with accuracies ranging from 0.78 for non-NADs to over 0.85 for other categories. Notably, the encoder block was trained on GM12878 data, while the last linear layer was trained on combined data from K562, H1, and HCT116 cells. The high accuracy achieved in predicting associated domains suggests that the relationships between epigenomic features and structural annotations learned by TECSAS in GM12878 are transferable to other cell types. A similar transferability is reported in polymer modeling, where models trained on one cell line, and chromosome can successfully predict experimental Hi-C data in other cell lines.

To further validate the predictions of TECSAS, we applied the model to the IMR-90 cell line, predicting LADs, SPADs, and NADs using histone modification ChIP-seq data as input. We then compared these predictions with A/B compartment annotations derived from the IMR-90 Hi-C data. As expected, LADs and NADs were predominantly found within B compartments, while SPADs were primarily associated with A compartments (Figure 5B). This observation aligns with the reported repressive environment of the nuclear lamina and nucleoli regions and the spatial association of speckles with active transcription. Interestingly, a substantial portion of B compartment regions were predicted to be neither LADs nor NADs. This suggests that these regions are not silenced through association with the lamina or nucleoli, even though they are classified as inactive chromatin. Similarly, a significant fraction of A compartment regions were not predicted as SPADs, indicating that not all active chromatin regions are necessarily localized in spatial proximity with speckles.

In addition to investigating the spatial distribution of predicted associated domains, we projected the annotations onto 3D genome structures of IMR-90 cells obtained from DNA tracing experiments [31]. These structures provide information about the distance of each chromatin segment to the lamina, speckles, and nucleoli. Figure 5C shows the distribution of distances for each associated domain and its corresponding negative set. As expected, LADs exhibited significantly closer proximity to the lamina compared to non-LADs, and SPADs were located closer to speckles than non-SPADs. NADs showed a slight preference for closer proximity to the nucleoli compared to non-NADs, but the difference was less pronounced than for LADs and SPADs. This suggests that the lamina and speckles may exert a stronger influence on the 3D organization of the genome compared to the nucleoli.

Finally, we analyzed the combined distribution of predicted compartments and associated domains, considering all possible combinations of these annotations (Figure 5D). Five combinations were found to be highly populated ($> 4000$ loci), with the most frequent A compartment combinations being A-nonLAD-SPAD-nonNAD and A-nonLAD-nonSPAD-nonNAD. This indicates that A compartment regions are primarily differentiated by their association with speckles, while a significant portion does not appear to interact with any of the analyzed nuclear bodies. Similarly, a substantial fraction of B compartment regions were predicted as not associated with the lamina or nucleoli. These findings suggest that nuclear bodies help to organize and shape the 3D structure of chromosomes within the nucleus, contributing to the variety of ways the genome is arranged.

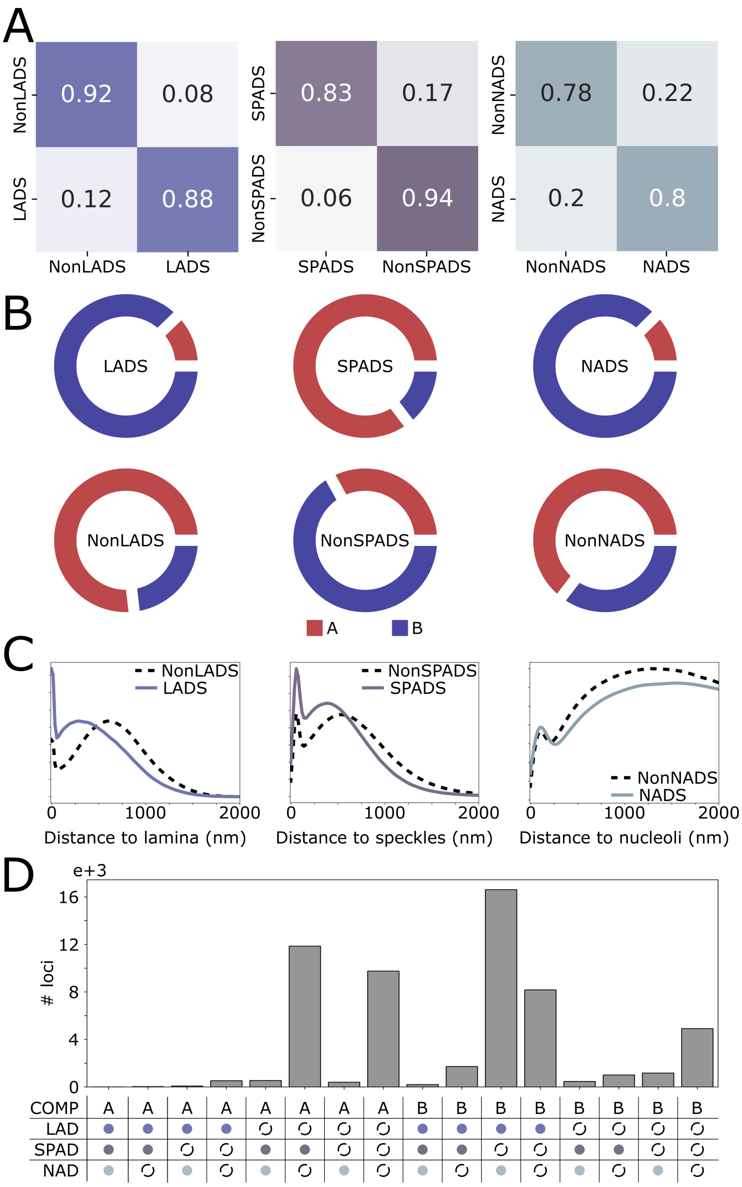

Figure 5: **Prediction of functional structural annotations by TECSAS highlights 3D structural bias due to nuclear body association.** (A) Confusion matrix for predicted LADS, NADS and SPADS against ground truth. (B) Distribution of A and B compartments for IMR-90 for each XAD and nonXAD. (C) Distribution of distance to lamina, speckles and nucleoli for loci predicted as LADS, SPADS and NADS respectively when projected in 3D DNA-tracing experiments[31]. (D) Number of loci in genome predicted as specific combinations of compartment, LAD, SPAD and NAD annotation; solid circles represent XAD and discontinuous circles represents nonXAD.

# 3 Conclusion

This study introduces TECSAS, a deep learning model based on the Transformer architecture, for predicting chromatin structural annotations from one-dimensional epigenomic data. TECSAS utilizes a Transformer encoder to interpret the complex relationships between various epigenetic marks and decode the context of the biochemical composition of the genome. The model achieved high accuracy in predicting subcompartment annotations at both 50 kb and 25 kb resolutions, indicating a strong association between epigenomic profiles and chromatin's structural organization. This finding suggests that changes to the epigenome could be used to directly shape the three-dimensional structure of chromatin.

TECSAS predictions were less accurate in regions transitioning between different subcompartments. These transition regions, characterized by mixed epigenomic signatures, pose a challenge for the model as they do not exhibit a clear association with a single subcompartment. Excluding these transition regions from the analysis significantly improved the prediction accuracy for both GM12878 and K562 subcompartments. This suggests that the epigenomic landscape undergoes gradual changes across the genome, leading to potentially fuzzy or undefined structural behavior in transition regions. It is worthwhile to mention that even the assumed ground truth experimental data may also include some false positives in the annotations, which may create some noise in the TECSAS predictions.

Compared to PyMEGABASE (PYMB), a previously developed method for predicting structural annotations from epigenomic data, TECSAS demonstrated an improvement in performance (Figure 3). This improvement can be attributed to several factors. First, the Transformer architecture with self-attention mechanisms allows TECSAS to capture complex, non-linear relationships between multiple epigenetic marks, while PYMB relies on a simpler Potts model that primarily captures pairwise interactions. Second, TECSAS incorporates information from a larger neighborhood of loci, enabling the model to account for long-range interactions within the epigenome. Finally, TECSAS utilizes continuous signal intensities for epigenomic features, providing a more nuanced representation of the data compared to the discretized approach used in PYMB.

The versatility of TECSAS extends beyond subcompartment prediction. By fine-tuning the final layer of the model, we successfully predicted the association of genomic loci with specific nuclear bodies, including the lamina, nucleoli, and speckles. The model achieved high accuracy in identifying LADs, NADs, and SPADs, demonstrating its ability to learn transferable relationships between epigenomic features and various structural annotations. The agreement between predicted associated domains and the known functional characteristics of nuclear bodies, such as the association of LADs and NADs with inactive chromatin and SPADs with active transcription, further supports the validity of the model's predictions. Additionally, the analysis revealed heterogeneity within chromatin compartments, with a significant portion of regions not exhibiting a clear association with any of the analyzed nuclear bodies. This suggests that nuclear body association contributes to the diversity of chromatin organization within the nucleus.

In conclusion, this study demonstrates that the biochemical composition of the genome, as reflected in epigenomic data, is highly informative for predicting the three-dimensional organization of chromatin. TECSAS, a deep learning model based on the Transformer architecture, effectively captures the complex relationships between various epigenetic marks and accurately predicts chromatin subcompartments and their association with nuclear bodies. The model's ability to account for long-range interactions and its transferability across cell types highlight its potential as a valuable tool for studying 3D genome organization and its role in gene regulation and other nuclear processes. Future research could explore the application of TECSAS to investigate the functional consequences of nuclear body association and the role of 3D genome organization in various biological contexts, such as different organisms, cell phases, and genetic disorders.

# 4 Methods and Materials

## 4.1 Data Acquisition and Preprocessing

Epigenomic data (histone modification ChIP-seq, transcription factor ChIP-seq, and RNA-seq) were acquired from the ENCODE portal for the GM12878 and K562 cell lines. The initial step utilizes the publicly available pyBigWig software [32] for data fetching.

### 4.1.1 Data Processing

- **Resolution:** Data were binned into loci of either 50 kbp (GM12878 subcompartment and associated domain prediction) or 25 kbp (K562 subcompartment prediction).
- **Signal Representation:** ChIP-seq signal intensities were expressed as signal p-values. For experiments with multiple replicates, the average signal track was used.
- **Normalization:**
  - **Min-max normalization (chromosome-wise):** The 5th and 95th percentiles were designated as the minimum and maximum values, respectively. This provides a baseline and mitigates outlier influence.
  - **Z-score normalization (chromosome-wise):** Ensures data standardization.

### 4.1.2 Input Preparation - Preprocess of 1D Experimental Tracks

For each target locus, TECSAS input comprised the normalized signal intensities of all epigenomic features within a window of N neighboring loci (both upstream and downstream). The Results section specifies the N value used in each experiment.

## 4.2 TECSAS Model Architecture and Training

TECSAS is a deep learning model utilizing the Transformer architecture. Its key components include:

- **Input Embedding Layer:** Transforms the epigenomic profile of each locus into a higher-dimensional representation.
- **Positional Encoding:** Injects information about the relative positions of loci within the input window.
- **Transformer Encoder:** Employs self-attention mechanisms to model complex relationships and long-range dependencies across epigenomic features within the input loci. Multiple attention heads are used.
- **Linear Layer:** Maps encoded information to the output space.
- **Softmax Output Layer:** Generates a probability distribution over possible structural annotations (5 nodes for subcompartments, 2 for XADS).

### 4.2.1 Machine Learning Implementation

TECSAS employs a linear layer for the token embedding process, the embbeding dimension is set to 128. This is then followed by a transformer encoder block made of two transformer encoder layers with 8 head each. A linear layer reduces the transformer encoder output to the subcompartment output layers. The dimension of the feedforward layer in the transformer encoder is 64. For training purposes, the dropout rate is set to 1%. PyTorch was used for implementation, and the output layer's activation is processed by a softmax function for activation-to-probability conversion. The dataset is split (80% training, 10% validation, 10% testing). Source code and tutorials are available on GitHub. The model was trained and tested on AMD Radeon Instinct MI50 32GB GPUs. The Stochastic Gradient Descent optimizer was used for training. The learning rate was initially to 2.5, and it was manually reduced every change checkpoint defined as five epochs. If the training loss is lowered after an epoch, the learning rate change checkpoint is reduced by 1 epoch.

## 4.3 Prediction of Associated Domains

To predict the association of loci with nuclear bodies, we fine-tuned the final layer of TECSAS to output probabilities for LADs, NADs, and SPADs, alongside their corresponding negative sets. We obtained annotations for these domains as follows:

- **LADs and NADs (K562, H1, and HCT116):** Acquired from DamID experiments accessible on the 4D Nucleome Data Portal [33, 30]. Relevant experiment IDs include:
  - **K562:** 4DNFIV776O7C
  - **H1:** 4DNFIP6N54B3
  - **HCT116:** 4DNFICCV71TZ
- **SPADs (K562):** Derived from TSA-seq experiments targeting the SON protein (a nuclear speckle marker) [34]. Loci exhibiting signal intensities exceeding the 80th percentile of the genome-wide signal were categorized as SPADs. Experiment ID: 4DNFINI7KVAI.

The fine-tuned model was trained using histone modification ChIP-seq data as input. Performance was assessed on held-out test sets.

# 5 Acknowledgments

We want to thank Peter Wolynes, Sumitabha Brahmachari, Matheus Mello and Antonio B. Oliveira Junior for many useful conversations during the development of this work and for all of their comments and suggestions.

This research was supported by the Center for Theoretical Biological Physics, sponsored by the NSF (Grants PHY-2019745 and PHY-2210291) and by the Welch Foundation (Grant C-1792). JNO is a Cancer Prevention and Research Institute of Texas (CPRIT) Scholar in Cancer Research. We would like to thank AMD (Advanced Micro Devices, Inc.) for the donation of critical hardware and support resources from its HPC Fund that made this work possible.

# References

[1] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292–301, 2001.

[2] Hui Zheng and Wei Xie. The role of 3d genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, 20:535–550, 9 2019.

[3] Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V. Lobanenkov, Joseph R. Ecker, James A. Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518:331–336, 2 2015.

[4] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2:292–301, 2001.

[5] Wendy A. Bickmore. The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics*, 14:67–84, 2013.

[6] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295:1306–1311, 2002. 3C.

[7] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38:1341–1347, 2006. 4C.

[8] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nature Genetics*, 38:1348–1354, 2006. 4C.

[9] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, pages 1299–1309, 2006. 5C.

[10] Erez Lieberman-aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eri S. Landerm, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome erez. *Science*, 326:289–294, 2009.

[11] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian Sanborn, Ido Machol, Arina D Omer, Eric S Lander, Erez Lieberman Aiden, E L A Designed, and M H H Performed Hi-C Experiments. A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping etoc blurb hhs public access. *Cell*, 159:1665–1680, 2014.

[12] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K. Brinkman, Stephen A. Adam, Robert Goldman, Bas Van Steensel, Jian Ma, and Andrew S. Belmont. Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *Journal of Cell Biology*, 217:4025–4048, 11 2018.

[13] Michele Di Pierro, Bin Zhang, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 113:12168–12173, 10 2016.

[14] Antonio B Oliveira Jr, Vinicius G Contessoto, Matheus F Mello, and Jose N Onuchic. A scalable computational approach for simulating complexes of multiple chromosomes. *Journal of Molecular Biology*, 433(6):166700, 2020.

[15] Ryan R Cheng, Vinicius G Contessoto, Erez Lieberman Aiden, Peter G Wolynes, Michele Di Pierro, and Jose N Onuchic. Exploring chromosomal structural heterogeneity across multiple cell lines. *eLife*, 9:e60312, October 2020.

[16] Vinícius G Contessoto, Ryan R Cheng, Arya Hajitaheri, Esteban Dodero-Rojas, Matheus F Mello, Erez Lieberman-Aiden, Peter G Wolynes, Michele Di Pierro, and José N Onuchic. The Nucleome Data Bank: Web-based resources to simulate and analyze the three-dimensional genome. *Nucleic Acids Research*, 49(D1):D172–D182, January 2021.

[17] Vinícius G. Contessoto, Ryan R. Cheng, and José N. Onuchic. Uncovering the statistical physics of 3D chromosomal organization using data-driven modeling. *Current Opinion in Structural Biology*, 75:102418, August 2022.

[18] Kyle Xiong and Jian Ma. Revealing hi-c subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications*, 10, 2019.

[19] Yuanlong Liu, Luca Nanni, Stephanie Sungalee, Marie Zufferey, Daniele Tavernari, Marco Mina, Stefano Ceri, Elisa Oricchio, and Giovanni Ciriello. Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nature Communications*, 12, 12 2021.

[20] Suchen Zheng, Nitya Thakkar, Hannah L. Harris, Megan Zhang, Susanna Liu, Mark Gerstein, Erez Lieberman Aiden, M. Jordan Rowley, William Stafford Noble, Gamze Gürsoy, and Ritambhara Singh. Predicting a/b compartments from histone modifications using deep learning. *bioRxiv*, 2022.

[21] Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, Ulugbek K Baymuradov, Keenan Graham, Casey Litton, Stuart R Miyasato, J Seth Strattan, Otto Jolanki, Jin-Wook Lee, Forrest Y Tanaka, Philip Adenekan, Emma O'Neill, and J Michael Cherry. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1):D882–D889, 11 2019.

[22] Rui Yang, Arnav Das, Vianne R Gao, Alireza Karbalayghareh, William S Noble, Jeffery A Bilmes, and Christina S Leslie. Epiphany: predicting hi-c contact maps from 1d epigenomic signals. *Genome Biology*, 24(1):134, 2023.

[23] Esteban Dodero-Rojas, Matheus F Mello, Sumitabha Brahmachari, Antonio B Oliveira Junior, Vinícius G Contessoto, and José N Onuchic. Pymegabase: Predicting cell-type-specific structural annotations of chromosomes using the epigenome. *Journal of Molecular Biology*, page 168180, 2023.

[24] Michele Di Pierro, Ryan R. Cheng, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 114:12126–12131, 2017.

[25] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[28] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need, April 2021. arXiv:2008.02217 [cs, stat].

[29] Antonio B. Oliveira Junior, Cynthia Perez Estrada, Erez Lieberman Aiden, Vinícius G. Contessoto, and José N. Onuchic. Chromosome Modeling on Downsampled Hi-C Maps Enhances the Compartmentalization Signal. *The Journal of Physical Chemistry B*, 125(31):8757–8767, August 2021.

[30] Sarah B Reiff, Andrew J Schroeder, Koray Kırlı, Andrea Cosolo, Clara Bakker, Luisa Mercado, Soohyun Lee, Alexander D Veit, Alexander K Balashov, Carl Vitzthum, et al. The 4d nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nature communications*, 13(1):2365, 2022.

[31] Jun-Han Su, Pu Zheng, Seon S Kinrot, Bogdan Bintu, and Xiaowei Zhuang. Genome-scale imaging of the 3d organization and transcriptional activity of chromatin. *Cell*, 182(6):1641–1659, 2020.

[32] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, 44(W1):W160–W165, 2016.

[33] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4d nucleome project. *Nature*, 549(7671):219–226, 2017.

[34] Liguo Zhang, Yang Zhang, Yu Chen, Omid Gholamalamdari, Yuchuan Wang, Jian Ma, and Andrew S Belmont. Tsa-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome research*, 31(2):251–264, 2021.