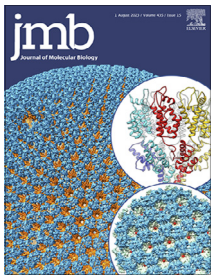

PyMEGABASE: Predicting Cell-Type-Specific Structural Annotations of Chromosomes Using the Epigenome

Esteban Dodero-Rojas [1], Matheus F. Mello [1], Sumitabha Brahmachari [1], Antonio B. Oliveira Junior [1], Vinícius G. Contessoto [1*] and José N. Onuchic [1,2,3,4*]

*1 - **Center for Theoretical Biological Physics,** Rice University, Houston, TX, USA*
*2 - **Department of Physics & Astronomy,** Rice University, Houston, TX, USA*
*3 - **Department of Chemistry,** Rice University, Houston, TX, USA*
*4 - **Department of Biosciences,** Rice University, Houston, TX, USA*

Correspondence to Vinícius G. Contessoto and José N. Onuchic: Center for Theoretical Biological Physics, Rice University, Houston, TX, USA (J. Onuchic).*vinicius.contessoto@rice.edu (V.G. Contessoto), jonuchic@rice.edu (J.N. Onuchic)*
https://doi.org/10.1016/j.jmb.2023.168180

Abstract

The folding patterns of interphase genomes in higher eukaryotes, as obtained from DNA-proximity-ligation or Hi-C experiments, are used to classify loci into structural classes called compartments and subcompartments. These structurally annotated (sub) compartments are known to exhibit specific epigenomic characteristics and cell-type-specific variations. To explore the relationship between genome structure and the epigenome, we present PyMEGABASE (PYMB), a maximum-entropy-based neural network model that predicts (sub) compartment annotations of a locus based solely on the local epigenome, such as ChIP-Seq of histone post-translational modifications. PYMB builds upon our previous model while improving robustness, capability to handle diverse inputs and user-friendly implementation. We employed PYMB to predict subcompartments for over a hundred human cell types available in ENCODE, shedding light on the links between subcompartments, cell identity, and epigenomic signals. The fact that PYMB, trained on data for human cells, can accurately predict compartments in mice suggests that the model is learning underlying physicochemical principles transferable across cell types and species. Reliable at higher resolutions (up to 5 kbp), PYMB is used to investigate compartment-specific gene expression. Not only can PYMB generate (sub) compartment information without Hi-C experiments, but its predictions are also interpretable. Analyzing PYMB's trained parameters, we explore the importance of various epigenomic marks in each subcompartment prediction. Furthermore, the predictions of the model can be used as input for OpenMiChroM software, which has been calibrated to generate three-dimensional structures of the genome. Detailed documentation of PYMB is available at https://pymegabase.readthedocs.io, including an installation guide using pip or conda, and Jupyter/Colab notebook tutorials.

1. Introduction

The three-dimensional organization of the eukaryotic genome within the cell nucleus is related to biological-function-determining characteristics like gene expression and cell identity.[1–4] Chromosome conformation capture (3C) techniques have been crucial in our understanding of the three-dimensional architecture of the genome. Observations from these DNA-DNA

ligation experiments, such as Hi-C maps,[5–9] report the frequency that any pair of chromatin loci are observed to be in spatial proximity. Hi-C maps endorse the existence of chromosome territories, also observed via microscopy.[3,9,10] In addition, the overall genome organization can be described by two major compartments, A and B. The structural type, annotated compartment A, is gene-rich and is related to euchromatin. In contrast, compartment B is gene-poor and related to inactive heterochromatin.[9]

Upon investigation at higher resolution interchromosomal Hi-C maps of human GM12878 cell line, Rao and coworkers[11] proposed that compartments can be further divided into subcompartments. A-type contains two subcompartments (A1 and A2), and compartment B has four (B1, B2, B3, and B4). Each subcompartment exhibits distinct contact patterns in Hi-C maps reflective of structural signatures of these loci. Interestingly, each subcompartment can be associated with local biochemical profiles.[11] For example, A1 and A2 correlate positively with histone modification H3K36me3, while B1 correlates with H3K27me3.[11] Furthermore, B2 shows strong associations with the nuclear lamina and the nucleoli. On the other hand, B3 is mainly present at the nuclear lamina.[11] Recently, tyramide signal amplification sequencing (TSA-Seq) was employed to measure distances of chromatin relative to nuclear speckles and lamina.[12] As a result, A1 loci and nuclear speckles present close spatial proximity. Whereas B2/B3 loci present short distances to nuclear lamina.[12] These measures suggest that subcompartments may be associated with specific positioning inside the nucleus. Further, subcompartment annotations may indicate preferential loci interactions with nuclear bodies and transcription activity.[12] Currently, GM12878 (human lymphoblastoid) is the only cell line in which the subcompartments are annotated.

Previously, we had developed a neural network named MEGABASE (Maximum Entropy Genomic Annotation from Biomarkers Associated to Structural Ensembles) that predicts structural annotations based on epigenetic markers (i.e. biochemical information). This method predicted subcompartment information from chemical information represented by ChIP-Seq data.[13] Also, it takes advantage of the fact that ChIP-Seq tracks are widely available assays for many samples, including different cell lines and tissues.[14] It is important to notice that, in contrast to other models to call structural annotations, MEGABASE predicts structure from biochemical composition of the genome alone, without relying on input structural information like Hi-C maps.[13]

Various tools for calling compartments and subcompartments have been developed since MEGABASE.[13,15–22] These methods can be classified into two main groups: those that call subcompartments based on Hi-C maps and those that use an existing set of annotations to predict subcompartments in other datasets. The first group includes methods such as SCI,[16] Calder,[17] and the clustering from Rao et al. 2014.[11] Each method proposes a different number of subcompartments and assigns unique sets of subcompartment labels. Despite these differences, all report an association between subcompartments and characteristic epigenetic features. The second group contains methods like SNIPER,[15] which predicts subcompartments from moderate coverage Hi-C maps based on the annotations from Rao et al. 2014. Other methods in this group, such as MEGABASE,[13] SCI-DNN,[16] and CoRNN,[18] use biochemical information to infer subcompartment or compartment annotations. Table S1 presents the main characteristics of these different methods. It is essential to recognize that although all these methodologies provide the same output (subcompartment or compartment annotations), their performances should not be compared as if they were equivalent methods. Some methods rely on structural information to identify/predict the annotations, while others use the relationship between biochemical information and subcompartment identities to predict the annotations.

In this work, we introduce PyMEGABASE (PYMB), an automated and intuitive software for calling structural annotations (compartments or subcompartments) based only on 1D epigenetic tracks. PYMB derives from the MEGABASE model,[13] while expanding its capabilities and broadening its accessibility. PYMB introduces several improvements in the computational pipeline. For example, PYMB runs automatic preprocessing steps and fast data fetching from the ENCODE (Encyclopedia of DNA Elements) portal.[14] These implemented functionalities allow the users to predict subcompartment annotations for any cell line or sample in the ENCODE portal in minutes. In addition to the ChIP-Seq tracks of histone modifications and transcription factors, PYMB also allows the usage of RNA-Seq experiments to predict subcompartments. To demonstrate the capability of PYMB to call structural annotations, we employed the methodology in different conditions. We compared the predictions of compartments and subcompartments in the GM12878 cell line using many data sets to train and test the model. Further, we present the accuracy of PYMB predictions at the compartment level on other systems, such as different cell lines, tissue samples, and other organisms. We predicted subcompartment annotations for over 150 human cell types, tissue samples, and single donor samples whose ChIP-Seq or RNA-Seq data were available on the ENCODE database.[14] This massive generation of annotations shed light on genomic sequence patterns related to cell differentiation. Further, we evaluate the predictions of PYMB when used at finer resolutions (5 kbp). At 5

kbp resolution, it allows us to investigate the relationship between the subcompartments and the epigenetic features at the scale relevant to a single gene. Finally, by analyzing the trained model parameters, we suggest molecular characteristics associated with the composition of each subcompartment.

2. Materials and Methods

Based on the Maximum Entropy (MaxEnt) approach, MEGABASE predicts compartment and subcompartment annotations based on genomic biomarkers.[13] We implemented the MEGABASE model into an intuitive and user-friendly tool called PyMEGABASE (PYMB). PYMB is a neural network that decodes the relationship between epigenetic markers and structural annotations from 1D experimental data tracks. It is worth mentioning that PYMB receives only 1D data as input, *i.e.*, PYMB does not require a 2D Hi-C map to call structural annotations. This is an advantage especially when there are no Hi-C maps available for that cell line or tissue sample. PYMB consists of a Potts model[13,23] where one node is associated with the structural annotations, called S-node. The states of the S-node correspond to structural annotations (A1, A2, B1, B2, and B3). We exclude the B4 subcompartment from PYMB training since it is only observed in a segment of chromosome 19.[11] The other nodes, called D-nodes, represent the enrichment of biomarkers signals integrated at loci resolution. Then, for a given locus I, we construct a state vector $\sigma(I)$ for the Potts model. The state vector $\sigma(I)$ includes the structural annotation of the locus I and the signal intensity values for each experiment in the positions $I-2, I-1, I, I+1$, and $I+2$. The addition of the neighboring loci was demonstrated to enhance the prediction accuracy by filtering noise in the data processing.[13] As a result, we can represent the state vector of the Potts model as:

$$\begin{aligned}\vec{\sigma}(I) = (&C(I), \mathrm{Exp}_1(I-2), \ldots, \mathrm{Exp}_L(I-2),\\ &\mathrm{Exp}_1(I-1), \ldots, \mathrm{Exp}_L(I-1), \mathrm{Exp}_1(I), \ldots, \mathrm{Exp}_L(I),\\ &\mathrm{Exp}_1(I+1), \ldots, \mathrm{Exp}_L(I+1), \mathrm{Exp}_1(I+2), \ldots, \mathrm{Exp}_L(I+2))\end{aligned} \quad (1)$$

where $C(I)$ is the structural annotation (S-node). $\mathrm{Exp}_i(m)$ is the data signal intensity of the i-th experiment (ChIP-Seq or RNA-Seq) at locus m (D-nodes). The Hamiltonian describing the energy of the system is defined by:

$$H\left(\vec{\sigma}\right) = -\sum_{i<j} J_{ij}(\sigma_i, \sigma_j) - \sum_i h_i(\sigma_i), \quad (2)$$

where the coupling term J_{ij} captures the correlation between epigenetic marks and chromatin annotations. h_i is the self-activation energy term that correlates with the frequencies of chromatin annotations and markers enrichment.

Figure 1 shows PYMB computational pipeline for predicting subcompartments. The first step uses the publicly available pyBigWig software[24] to fetch data from the ENCODE Portal.[14] This step may include experimental data of histone modifications (ChIP-Seq), transcription factors (ChIP-Seq), small RNA-Seq, and total RNA-Seq. The user also has the option of uploading custom tracks into PYMB and running the predictions using their data as well. PYMB supports multiple signal formats for ChIP-Seq tracks, such as signal p-values and fold change-over-control of the experiments. In addition, PYMB receives.bed and.bigwig file formats as input tracks. This initial step also includes signal processing. Each experimental track is partitioned into 50 kbp loci, followed by a min–max normalization for each chromosome. The 5th percentile is assigned as the minimal value, and the 95th percentile as the maximum. Assigning the min–max value sets up a data baseline and avoids outliers dominating the signal integration. When several replicas exist for each experimental target, we set the average track as the representative track for the respective target.

The second step involves neural network optimization.[23,27] PyMEGABASE energy function is built following the Maximum Entropy principle (MaxEnt), where the probability of observing the state $\vec{\sigma}$ on the system follows a Boltzmann distribution, given by $\mathrm{P}\left(\vec{\sigma}\right) = \frac{1}{Z}\exp\left(-\mathrm{H}\left(\vec{\sigma}\right)\right)$. The energy terms are obtained by maximizing the log-pseudo-likelihood of a set made of every locus in the training set (see SI for details). We maximize the log-pseudo-likelihood function as described in,[23] using publicly available tools implemented in Python.[27] Furthermore, this process allows us to have a Hamiltonian that correlates the structural annotations with epigenetic marks.

After training the model parameters, we use the Hamiltonian presented in Eq. 2 to predict the chromatin subcompartment annotations of a particular set of chromosomes (see SI for details). The prediction is equivalent to finding the subcompartment annotation that minimizes the energy function. This step is performed for a given set of experimental measures for each locus. PYMB obtains the subcompartment annotations as the state of the S-node that, when interacting with the rest of the D-nodes, will lead to the highest probability in the Boltzmann distribution (lowest energy value). This process is repeated for each locus in the target cell genome. In summary, the prediction pipeline identifies the intersecting set of experiments (ChIP-Seq targets and RNA-Seq) between GM12878-hg19 and the target cell. Next, PYMB is trained on GM12878-hg19's structural annotations and the experimental tracks from GM12878-hg19 within the intersecting set. Finally, the model uses the target cell's experimental 1D tracks to predict structural annotations for each locus in the target cell's genome. For example, if the target cell has the following experiments: [*RNA-Seq, CTCF Chip-Seq, H3K9me3 Chip-Seq, H4K20me2 Chip-Seq*]

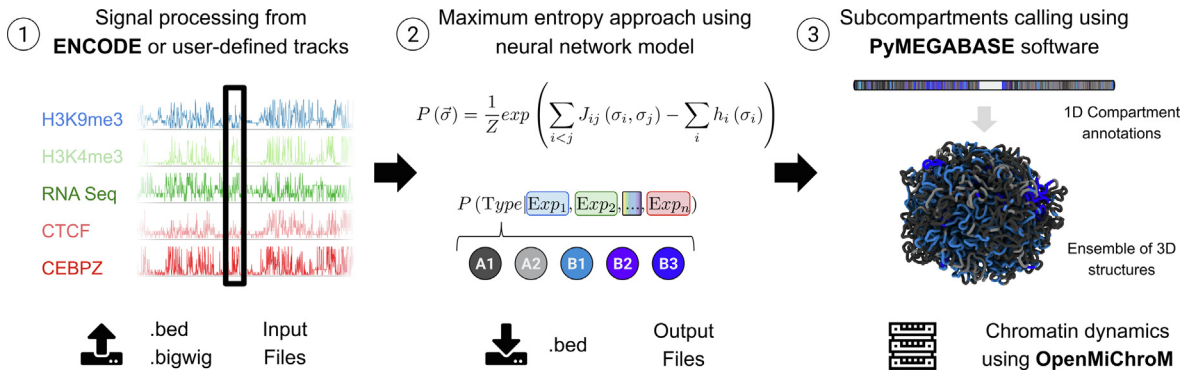

Figure 1. PyMEGABASE computational pipeline. The model uses as input 1D data sets such as RNA-Seq, Histone Modification ChIP-Seq, and Transcription Factor ChIP-Seq experiments. The data is extracted from the ENCODE portal[14] or submitted by the user. Then, the neural network is trained to capture the relationship between structural annotations and epigenetic marks. Finally, PYMB is used call compartments and subcompartments using only the RNA-Seq and ChIP-Seq experiments across the genome. These annotations can be then used for further analysis such as the prediction of 3D simulated structure ensembles using the calibrated force fields of OpenMiChroM.[25,26]

and GM12878-hg19 has [*RNA-Seq, CTCF Chip-Seq, H3K9me3 Chip-Seq, ZZZ3 ChIP-Seq,* ...], then the intersection set consists of [*RNA-Seq, CTCF Chip-Seq, H3K9me3 Chip-Seq*]. PYMB will be trained on GM12878-hg19 structural annotations and GM12878-hg19's [*RNA-Seq, CTCF Chip-Seq, H3K9me3 Chip-Seq*]; the trained model would predict the target cell's annotations using the target cell's [*RNA-Seq, CTCF Chip-Seq, H3K9me3 Chip-Seq*] tracks.

Then, based on the subcompartment predictions, we generate compartment predictions by setting A1 and A2 as A and B1, B2, or B3 as B. PYMB outputs the predictions in.bed file format that can be loaded into 1D Juicebox Hi-C visualization platform.[28] Also, PYMB subcompartment predictions also serve as input for OpenMiChroM chromatin dynamics software.[25,26] OpenMiChroM employs the Minimal Chromatin Model (MiChroM) polymer physics energy function to generate an ensemble of 3D structures that are consistent with 2D Hi-C data from the 1D structural annotations generated from PYMB.[29,13,22,30,20] To evaluate the prediction efficiency of PYMB we employed several measures to assess the capability of PYMB to predict structural annotations correctly. Hi-C-based annotations and PYMB predictions at the compartment level are compared as True or False hits. Therefore, we analyzed the accuracy and the Area Under the Receiver Operating Characteristic Curve (AUCROC) score to measure the similarity between PYMB and Hi-C-based annotations.[31] We also obtained information about mispredictions by building a confusion matrix between the experimental annotations and PYMB predictions. This analysis informs on PYMB accuracy and shows which subcompartments are prone to mislabeling. Further, we analyzed the multiclass AUROC using the one-vs-one comparison to assess the prediction capability of the model at predicting subcompartments.[31] This analysis uses the mean of the binary AUCROC score for all combination pairs between subcompartments and averages over the scores. It is important to emphasize that given that the cell GM12878-hg19 is the only one with (A1, A2, B1, B2 and B3) subcompartments experimentally annotated, this is the only system where the model can be tested for the predictive capability at the subcompartment level.

3. Results

3.1. PYMB predicts compartment and subcompartment annotations at 50 kbp resolution from epigenetic data on Gm12878-hg19

As mentioned, subcompartment annotations are determined only for the cell line GM12878 in the assembly hg19 using a Hi-C-based approach.[11] Also, this cell line has multiple epigenetic marks experiments available such as RNA-Seq, histone modification ChIP-Seq, transcription factor ChIP-Seq, among others. Given that this cell line is the only cell with both structural annotations and chemical information, we used it as the training set for any prediction using PYMB. We show the accuracy of PYMB predicting subcompartment annotations at 50 kbp using odd-numbered chromosomes as training data and even-numbered chromosomes as the test data on GM12878-hg19. We also employed PYMB in different sets of experiments to explore the prediction capabilities limit as detailed below.

3.1.1. Compartment annotation predictions using only RNA-seq yield highly accurate results. A/B compartments have been reported to be correlated with transcription activity. A-type compartments have higher gene expression levels when compared to B-type loci.[11,9,15] It is expected

that the distinction between A/B can be learned from assays that correlate with transcription activity of a genomic locus, such as RNA-Seq, ATAC-Seq, and DNAse-Seq.[32–34] To assess PYMB's capability to predict compartments using only RNA-seq data, the model was trained using small and total RNA-Seq of the odd chromosomes at 50 kb resolution. Figure 2(A) presents the confusion matrix between PYMB and ground truth (Hi-C-based). PYMB demonstrated high compartment-predicting efficiency using only RNA-seq data across even chromosomes. We compare this prediction with a null model only based on the intensity of the signal of the RNA-Seq tracks. PYMB outperforms the null model, regardless of the threshold on either small or total RNA-Seq (Figure S1). Even though some correlation was expected between transcription and structural information, this result shows that PYMB captures the non-trivial relationship between RNA-Seq and compartment annotations.

3.1.2. PYMB predicts subcompartment annotations using histone post-translational modifications. Histone Modifications (HistMod) are considered a useful set of epigenetic marks to characterize several aspects of genomic loci, such as transcriptional activity and nuclear positioning.[35] Correlations between HistMod and the subcompartment annotations have been suggested in the literature.[11,12] For example, H3K4me1 and H3K36me3 are highly enriched on active subcompartments (A1 and A2).[35,36] We used

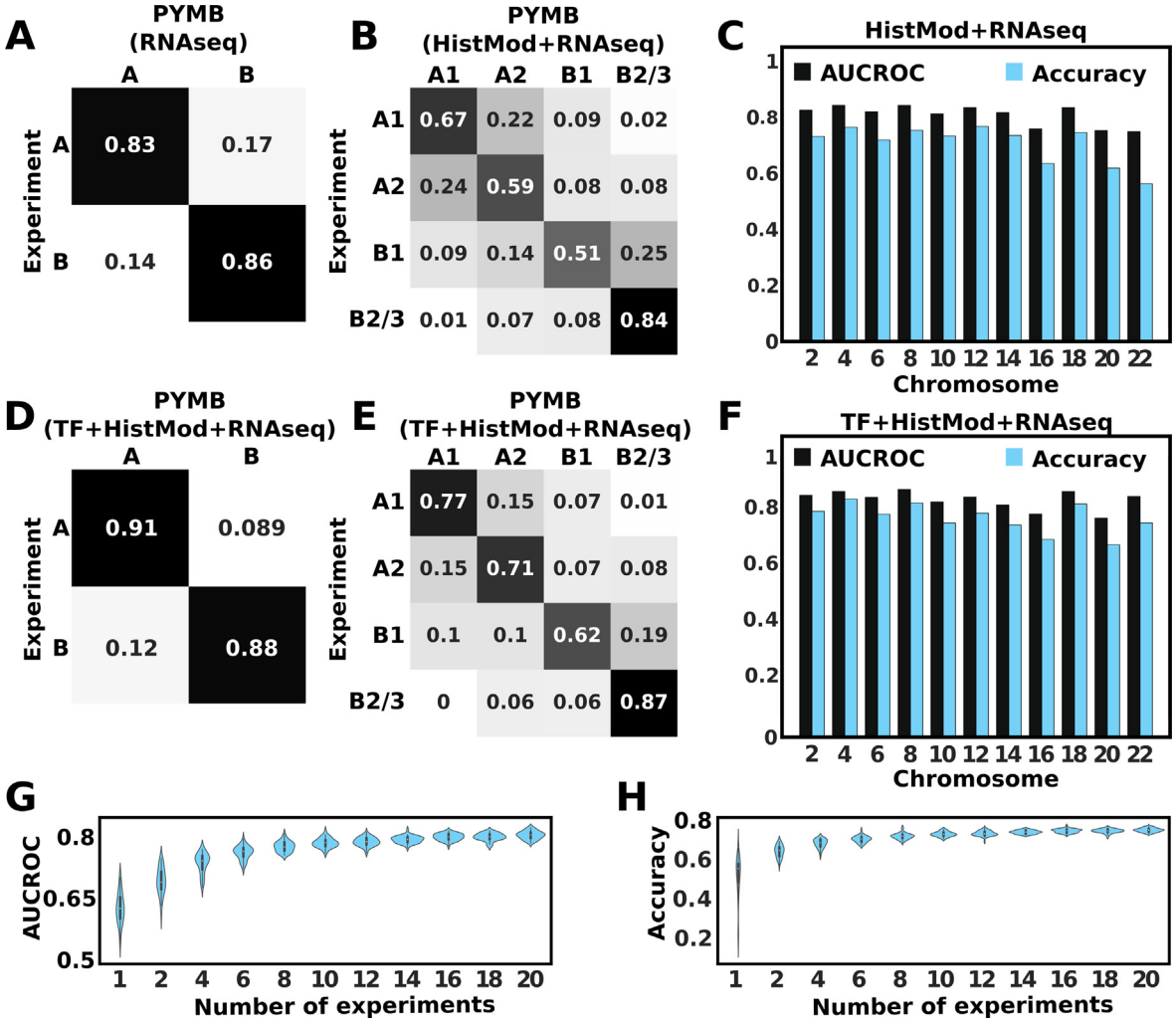

Figure 2. PYMB prediction performance on even-numbered chromosomes in GM12878-hg19 at 50 kbp resolution. (A) Comparison between compartment annotation predicted with RNA-Seq, and the experiment-derived compartment annotations. (B) Confusion matrix comparing HistMod ChIP-Seq and RNA-Seq based subcompartment predictions from PYMB and the experimental annotations from Rao et al.[11] (C) Performance results across even chromosomes at predicting subcompartment annotations measured by multiclass AUCROC score (black) and accuracy (blue). (D-E-F) Prediction performance at the compartment and subcompartment level of PYMB using RNA-Seq, HistMod, and TF ChIP-Seq data. (G-H) AUROC scores and accuracy of subcompartment predictions as a function of the number of unique experiments used as input for PYMB.

HistMod ChIP-Seq experiments and the RNA-Seq data for GM12878-hg19 to optimize PYMB parameters, then capturing the relationship between the chemical composition of the genome with the structural annotations. PYMB training was performed on the odd chromosomes and subcompartments predicted for even chromosomes at 50 kbp resolution. In this scenario, PYMB used 11 different histone modification ChIP-Seq, small RNA-Seq, and total RNA-Seq directly extracted from the ENCODE database.[14] Figure 2(B) shows the accuracy and confusion matrix of PYMB predictions of each subcompartment based on HistMod and RNA-Seq data. PYMB has higher efficiency when predicting subcompartments A1 and B2/3 than A2 and B1. However, the majority of the mislabeled predictions come from annotating B1 as B2/3, and A1 as A2. In other words, PYMB mislabeling falls within the same compartment. This shows the robustness of this methodology in predicting compartments. Figure 2 also demonstrates that the accuracy and multiclass AUCROC for subcompartment prediction is higher than 0.75 for most chromosomes while always remaining above 0.70 for all chromosomes. Therefore, the model achieves significant predictive efficacy when using chemical information from HistMod and RNA-Seq.

3.1.3. Incorporating binding profiles of transcription factors significantly increases prediction accuracy. The binding affinity of transcription factors (TF) shows the enrichment of certain TFs in specific regions of the genome. For example, UBTF colocalizes with the chromatin that is associated with the nucleolus.[37] On the other hand, ARID3A and STAT1 usually bind on active chromatin.[38,39] These examples present possibilities of using TF enrichment as input to call subcompartments. We incorporated HistMod, RNA-Seq, and TF ChIP-Seq data in the PYMB model. This data set contains a total of 155 experimental tracks from the GM12878-hg19 cell line available in ENCODE. Figure 2(D) shows that including the TF and HistMod increases the accuracy of the prediction for the compartment annotations when compared to PYMB using only RNA-Seq (Figure 2(A)). Figure 2 (E) shows the confusion matrix indicating a significant increase in accuracy for all the subcompartments. These results suggest that TF tracks used in PYMB help to differentiate subcompartments and compartments. Figure 2(F) presents the multiclass AUCROC and the accuracy of subcompartment annotations for each even chromosome. As expected, when PYMB uses additional information from TF tracks and non-redundant data, both the accuracy scores are higher than PYMB using only HistMod and RNA-Seq data.

Interestingly, as shown in,[11] the epigenetic profile of B2 and B3 are highly similar. It was expected that PYMB would have the lowest performance in distinguishing between these subcompartments. Figure S2 confusion matrix shows that the model predicts most of the B2 subcompartment as B3. The introduction of TF tracks increases PYMB's ability to predict B2 subcompartments.To further assess the prediction ability of PYMB, we employed machine learning approaches to predict subcompartments using the same training and test sets as PYMB. Figure S3 reveals that PYMB outperforms all the tested machine-learning approaches. Although personalized machine learning architectures such as SCI-DNN[16] perform similarly to PYMB in predicting subcompartments from epigenetic data as shown in Figure S3. In contrast, PYMB parameters can be further used to analyze the relationship between structural annotations and their biochemical profiles. This then suggests that the complexity of the Potts model implemented within PYMB captures the interplay between epigenome composition and structure.

Next, we performed a data reduction test to investigate how many experiments are necessary for high-accuracy predictions. In context, there is a lack of available tracks in some investigated biosamples. This adds a new challenge of predicting (sub) compartments in samples with only a few ChIP-Seq experiments. In this scenario, the model is trained using the experiments available for the training (GM12878-hg19) and the predicted cell line. We generated the prediction on even chromosomes in GM12878-hg19 using a different number of experiments. The experiments were selected randomly from the 155 set of TF, HistMod, and RNA-Seq data. Figure 2(G–H) shows the accuracy scores at predicting subcompartment annotations across chromosomes using data sets with the number of experiments ranging between 2 to 20. As expected, there is an increase in the accuracy score as more data are used to predict the subcompartments. Both performance measures show that data sets containing less than 4 experiments present lower accuracy. To reach higher accuracy in the predictions, PYMB should be trained using at least 4 experiments. However, certain data sets containing less than 8 tracks have similar accuracy/AUCROC scores than sets with a high number of experiments. This suggests that some epigenetic marks may carry information that helps PYMB to differentiate subcompartments. For example, the best set of 4 experiments consists of EZH2, HDGF, KDM1A, and TCF7. We include in Table S2 the sets of experiments that yield the best results for each number of experiments tested.

3.2. PYMB predicts compartment annotations with high accuracy on human cells

Given that GM12878-hg19 is the only sample with subcompartment annotations, comparing the predictions at that level on other cell types or samples is not possible. However, using Principal

Component Analysis (PCA), compartment annotations are extracted from the experimental Hi-C maps of various cell lines and tissue samples lacking subcompartment annotations. These annotations are generated based on the first eigenvector of the correlation matrix from Hi-C maps.[9] Now that we have experimentally determined compartment annotations to compare, we used PYMB to predict the annotations for the following cell lines (GRCh38 assembly): GM12878, IMR-90, K562, A549, and HepG2. We anticipate that epigenetic marks exhibit similar, if not identical, functions across various cell types, such as gene silencing or activation, enhancement of transcription, and other regulatory processes.[40] Consequently, we hypothesize that the observed correlation between local biochemical composition and structural information in GM12878-hg19 is also present in other cells. This underlying assumption enables PYMB to predict structural annotations effectively by utilizing the respective epigenomes of different cell types. In order to predict the structural annotations on these cells, we use all the chromosomes from GM12878-hg19 as the training set while using the biochemical markers of the target cell to predict the structural annotations. Figure 3 (A) shows the performance of PYMB at predicting compartment annotations measured by AUCROC and accuracy. All cell lines show accuracy scores higher than 0.80. This suggests that the correlations between the epigenome (i.e. biochemical composition of the loci) and the structure learned by PYMB are transferable across cell lines.

Figure 3(B) shows the comparison between the predicted and experimental compartments on chromosome 2. The compartment annotations extracted from the sign of the principal eigenvector may present ambiguity whether a locus corresponding to a value close to zero should be A or B. This ambiguous label is called "weak" compartments,[18] whereas the ones with a high value are termed "strong" compartments. Here we used the definition of strong compartments by Zheng et al.[18] The strong compartment has eigenvector values greater than the mean subtracted one standard deviation of the unsigned eigenvector values. As a result, PYMB accuracy is significantly higher for strong compartments than weak ones (Figure S4). This suggests that correlations between epigenetics and structure can be learned for strong compartments. However, these correlations are ambiguous for weak compartments. Furthermore, we tested the accuracy of PYMB in predicting compartments on tissue samples from single donors. We employed PYMB to generate the A/B annotations using the single donor ChIP-Seq available in the ENCODE database (accession codes: ENCDO845WKR and ENCDO451RUA). Predictions were computed for a transverse colon sample from a 37 year-old male and a gastrocnemius medialis sample from a 54 year-old male. The annotations were obtained at 50 kbp and 100 kbp resolution. Figure S5 shows that at 50 kbp and 100 kbp resolutions, PYMB reaches high scores on AUCROC and accuracy when predicting the annotations on both samples, outperforming ML learning methods such as CORNN[18] for single donor samples at 100kbp resolution. This indicates that PYMB can be applied to extract compartment annotations for single donor samples. This also indicates that PYMB is transferable across cell lines, tissue samples, and different human donors. Moreover, PYMB uses only local information at each locus to predict subcompartment and compartment annotations. These results also indicate that structural annotations correlate to the local biochemical composition of the epigenome.

3.3. Human epigenetic signatures captured by PYMB are transferable to other organisms

PYMB can also use epigenetic data and run structural annotation predictions on other organisms. Data on HistMod, TF, and RNA-Seq are widely available for several other systems. It has been reported that some marks have similar functions across organisms.[41,42] We expanded PYMB to evaluate whether the model trained on human cells (GM12878-hg19) would hold predictive power on other living systems. We predicted the subcompartment and compartment annotations for the mouse cell line CH12.LX. CH12.LX cell line has available Hi-C maps and multiple ChIP-Seq and RNA-seq experiments. Therefore, it was possible to compare the accuracy of PYMB at the compartment level using the first eigenvector of the correlation matrix of the Hi-C.[9] Subsequently, we predicted the structural annotations for this mouse cell, and the distribution of subcompartments across chromosomes is depicted in Figure S6.

Figure 3(C) shows that PYMB is able to predict accurate compartments on this mouse cell line. PYMB reaches accuracy scores around 0.85 or higher across chromosomes. The efficient performance of PYMB in predicting compartment annotations for both mouse and human cells underscores the concept that epigenetic marks, such as transcription factor binding, histone modifications, and transcription, exhibit comparable functions across different organisms. This similarity in function is observed as these epigenetic marks consistently correlate with structural annotations, regardless of the organism. Further, we used OpenMiChroM to generate an ensemble of 3D chromosomal structures using the mice compartment annotations.[25,26] We used the set of structures to compute *in silico* Hi-C maps that are consistent with the experiments. Figure 3(D) shows the (*in silico*) and experimental Hi-C maps, lower and upper triangle, respectively. These analyses are presented for chromosomes 2 and 8 with a representative 3D structural configuration. Combining PYMB and OpenMiChrom tools allows gen-

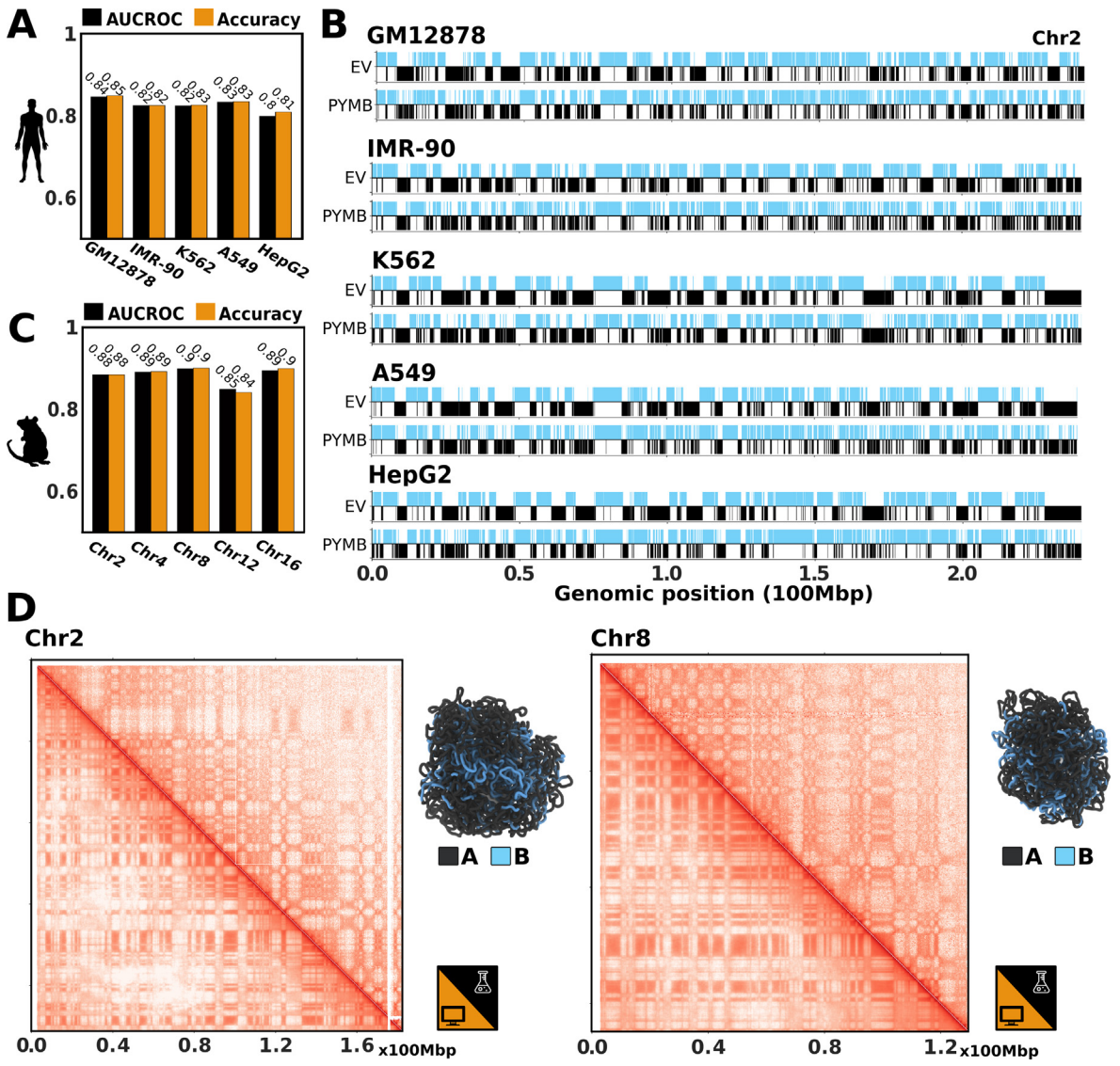

Figure 3. Compartment prediction capability of PYMB across human and mice cells at 50kbp resolution. (A) AUCROC score and accuracy measures between predicted compartment and compartment derived from experimental Hi-C for multiple human cell lines. (B) Visual comparison between the eigenvector derived from Hi-C (EV) and PYMB compartment predictions for the human cell lines: GM12878, IMR-90, K562, A549, and HepG2. (C) AUCROC score and accuracy measures between predicted and experimental-derived compartments for the mice cell line CH12.LX. (D) Mice cell line CH12.LX experimental Hi-C (upper triangle) and in silico Hi-C (bottom triangle) from chromosome 2 and 8 along with a representative structure of each chromosome generated by OpenMiChroM[25,26] using as input PYMB compartment annotations predictions.

erating the ensemble of 3D chromosomal structures across organisms[43] using as input the chemical composition of the epigenome alone.

3.4. Prediction of subcompartment annotations on more than a hundred cell types shows features on cell identity

As presented, PYMB predicts compartment and subcompartment annotations across different cell types. Therefore, we massively employed PYMB to predict subcompartment annotations on all human cell types (tissue samples and cell lines) available in ENCODE database. The predictions were performed for samples with ChIP-Seq or RNA-Seq data using GRCh38 as the reference genome. In addition, the following analysis was performed for samples with at least four experiments. Figure 4(A) shows the subcompartment predictions on chromosome 2 for more than 150 human cell types. Each color represents a different subcompartment, while the subpanel shows the frequency of subcompartments at each locus across samples.

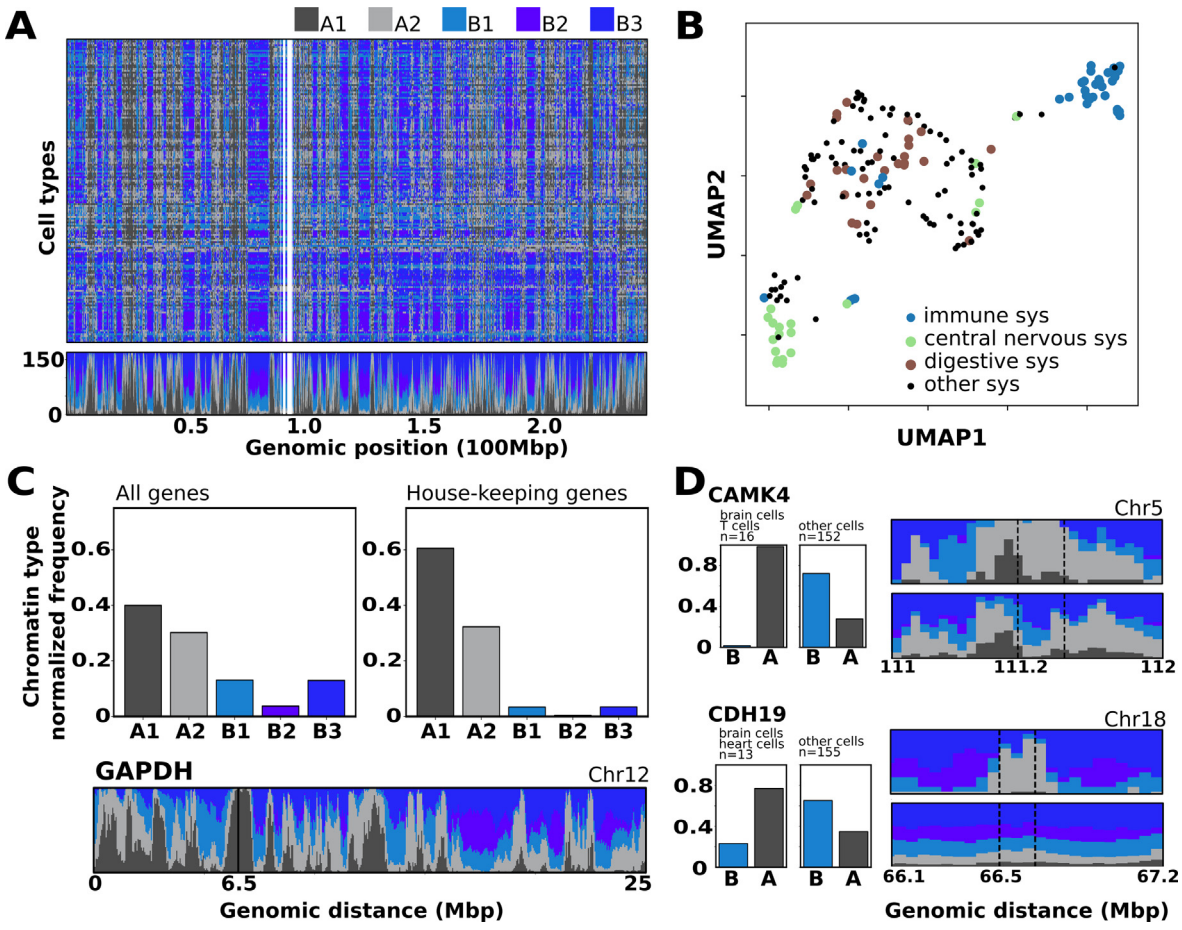

Figure 4. Subcompartment predictions for more than hundred human cell types highlight phylogenetic information. (A) Chromosome 2 subcompartment annotations for cell types with at least 4 unique experiments found in the ENCODE database (top), and the frequency of each subcompartment at each locus along the chromosome (bottom). (B) Genomic sequence landscape projected using UMAP method for dimension reduction.[44] (C) Normalized frequency of subcompartment occupied by all human genes and housekeeping genes. Genomic placement of the GAPDH gene across predicted cells (D) Cell specific genes found on A1 and A2 subcompartments for subset of cells. CAMK4 gene is mostly found on A-type subcompartments for brain and T-cells. Similarly, CHD19 is mainly found in A compartments for brain and heart cells. While these genes are in B-type loci for the rest of the cells.

Interestingly, there are shared patterns between cells, for example, some regions are annotated as active subcompartments (A1 or A2) in all cell lines. Also, there are loci where inactive subcompartments (B1, B2, or B3) are predominant.

Figure 4(B) shows the genomic sequence space projected using Uniform Manifold Approximation and Projection (UMAP).[44] Cell types and tissues are clustered related to their biological sample. This indicates that those cells have similar subcompartment annotations. Figure 4(C) shows that most human genes are in A1 or A2 loci in all the predicted cells. A significant set of genes is located on A-type loci only in a subset of cells. This is consistent with cell-specific gene activity observed experimentally.[45] Moreover, genes associated with the survival of cells, i.e., house-keeping genes, are expected to be active in all cells.[45] We identified the predicted type of loci where house-keeping genes are located. Figure 4(C) shows that house-keeping genes ubiquitously reside on A-type loci across all the cells. Also, there is a high likelihood for this set of genes to occupy the A1 locus. For example, the house-keeping gene encoding GAPDH (a protein needed for glycolysis[46]) is in the A1 subcompartment across cells.

On the other hand, B compartment loci have a low density of genes. Furthermore, B compartments also align with the G-bands showing heterochromatin (Figure S7). These results suggest that PYMB may capture the activation profile of the genome across the cell lines. Further, we reviewed regions on the genome active on only a subset of cell types. Therefore, we expected these regions to be associated with cell-specific genes. To explore this possibility, we analyzed the activation of two different genes: CAMK4 and CHD19. Figure 4(D) shows that the

region containing CAMK4 has a high frequency of A-type subcompartments on brain cells and T-cells, while high B-type subcompartments frequency on other cell lines. Similarly, the region with the CDH19 gene was mostly active on brain and heart cells while often inactive on the rest of the cell types. This observation is consistent with the experimental expression profiles for human tissues found on the Human Protein Atlas database (proteinatlas.org).[47,48] This suggests that PYMB can also capture cell-specific active and inactive regions based on the local biochemical composition. Given that the subcompartment annotation profile was obtained for more than 150 cell types, we organized all the cells based on the similarity of these annotations across the genome. Figure S8 shows a dendrogram based on the subcompartment annotation of each cell type. Together with the cell clustering presented in Figure 4(B), this indicates that the subcompartment annotations can assess cell similarity and may serve as a phylogeny indicator.

3.5. Prediction of annotations at 5 kbp suggests a relationship between cell identity and compartments

Initially, PYMB was optimized to run subcompartment predictions at 50 kbp resolution. Next, we expanded the PYMB method to run at a higher resolution of 5 kbp, relevant to single genes. We generated A/B annotations for GM12878-hg19 for training. Compartments A and B are determined using POSSUMM software.[49] To assess PYMB prediction accuracy, we used GM12878-hg19 chromosome 2 as the training set and ran the prediction to the others. For 5 kbp predictions, we used only HistMod data. Figure 5(A) shows the A/B annotation predictions for multiple chromosomes in GM12878-hg19. PYMB reaches an accuracy score of 0.80 across chromosomes at 5 kbp. We also observed that PYMB prediction on strong compartments is higher, yielding an accuracy of 0.84. Further, to test the transferability of the 5 kbp model, we used PYMB to predict the compartment annotations for the chromosomes of the K562 cell line (GRCh38 assembly). Figure 5(B) show that PYMB predictions on K562 reach similar accuracy to the predictions on GM12878-hg19. This indicates that the correlations between structural characteristics and the local epigenome are transferable at 5 kbp resolution. Predictions at higher resolution have the advantage of being at the genomic scale as single genes. Therefore, gene quantification from RNA-seq experiments was used to obtain FPKM (fragments per kilobase of exon per million mapped fragments). First, we aligned the predicted compartment annotations at 5 kbp resolution with the genes. Then, the gene label was assigned to either A or B based on most predicted compartment annotations within the gene. Figure 5(B) shows the average of FPKM for all the genes assigned either A or B. The genes predicted as being A compartment show an increased expression (higher FPKM). This suggested that structural motifs correlate with the gene expression level. Moreover, we expect more transcribed genes in a specific sample to reside in compartment A. We also extracted roughly 100 genes with higher and lower relative expression (predicted sample compared to other cell types) from the Roadmap Epigenomics Cell and Tissue Gene Expression Profiles data set.[50–52] When each gene for both lists was mapped to the predictions from PYMB for GM12878-hg19, we found most of the high relative expression genes within A-compartment loci. On the other hand, the low relative expression genes did not show a strong preference for any specific compartment, as shown in Figure 5. This was also observed on other cell lines and tissue samples such as K562, A549, HMEC, and thymus cells (Figure 5(C)). While highly expressed genes are likely to be associated with A compartments, genes with low expression are equally likely to be in either compartment.

3.6. Inferring epigenetic marks enrichment on subcompartments from PyMEGABASE parameters

We have demonstrated the capability of PYMB to predict annotations across chromosomes and cell types. Here, we explore the information that can be extracted from the model itself after the training. During training, PYMB fits the energy-coupling terms, $J_{\mathrm{exp}}(\mathrm{Sub}, \mathrm{signalvalue})$, in the information space spanned by each subcompartment and the discrete values of the signal intensity for each experiment. For example, the energy term associated with the combination of the subcompartment A1 and signal intensity of 7 for the ChIP-Seq track of H3K9me3 is denoted as $J_{\mathrm{H3K9me3}}(\mathrm{A1}, 7)$. We then extracted the coupling term of each subcompartment with every signal intensity value (0–9), which depicts the learned correlation between them. A high J value means a high likelihood of that signal intensity value being associated with the respective subcompartment and vice versa. Figure 6(A) shows the coupling term between each subcompartment annotation for every signal intensity value for some experimental tracks. Starting with H3K36me3, we noticed that high signal intensity values have high energy coupling with A1 subcompartments. In contrast, low signal intensity values have a high energy coupling term with B2 and B3. This is interpreted as H3K36me3 being likely enriched on A1 subcompartments and depleted on B2 and B3 subcompartments. Similarly, Figure 6(A) shows that SKIL is likely enriched on A2 subcompartments, CEBPB is enriched in B1 and B2, CBX5 is enriched in B2, and SUZ12 is enriched in B2 and B3. Finally, E2F8 has high energy coupling terms on high signal intensity values with A-type subcompartments. This

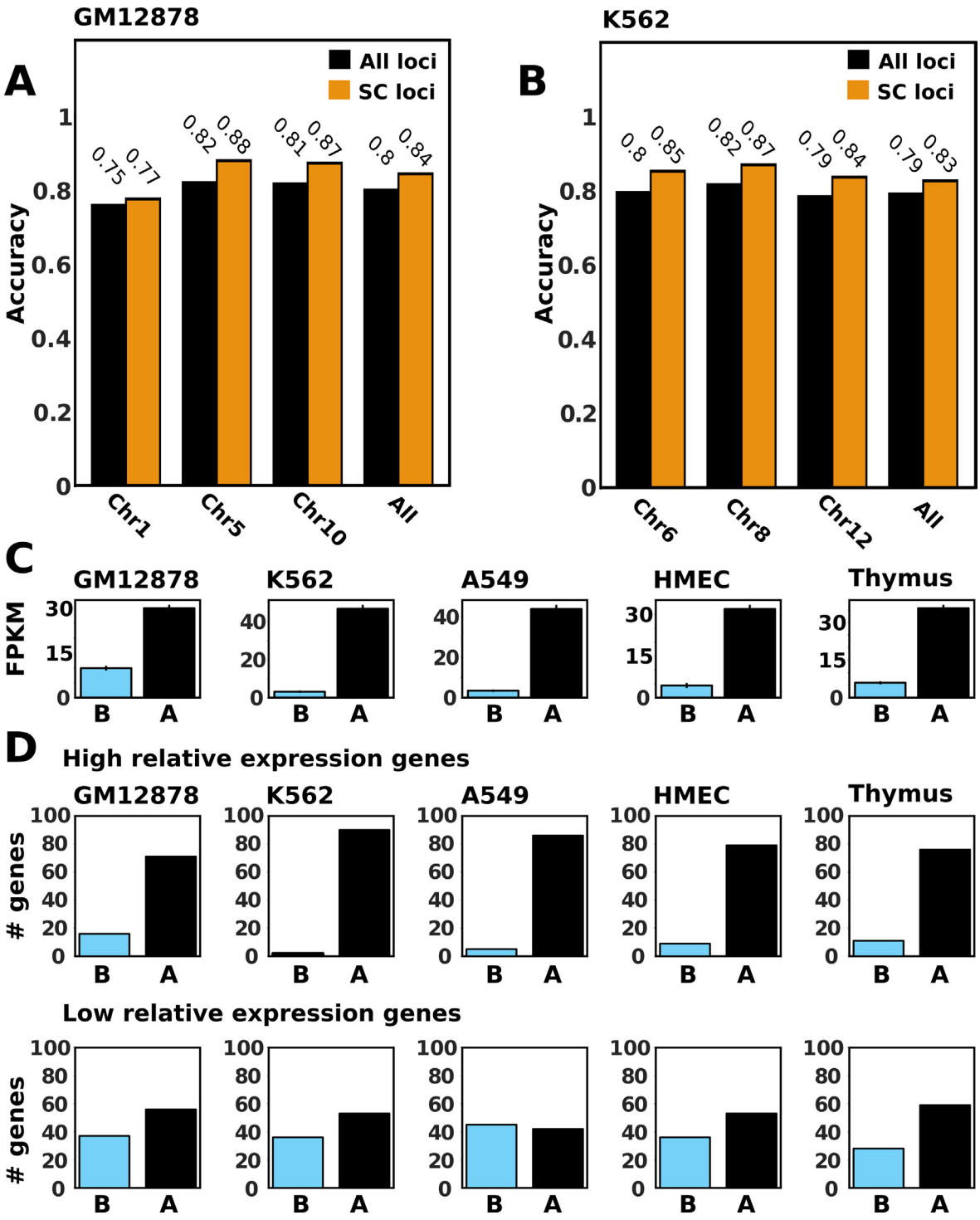

Figure 5. PYMB compartment accurate predictions at 5 kbp demonstrate a correlation between gene activation and expression with structural motifs. (A) Accuracy of PYMB predictions with the first eigenvector derived from the experimental Hi-C on GM12878-hg19 for all loci and strong-compartment loci. (B) PYMB performance at predicting compartments in K562 for all loci and strong compartment loci. (C) Gene expression measured by fragments per kilobase of exon per million mapped fragments (FPMK) for each compartment in GM12878-hg19, K562, A549, HMEC, and thymus cells. (D) Predicted compartment localization of 100 genes with higher-relative gene expression and 100 genes with lower-relative expression for GM12878-hg19, K562, A549, HMEC, and thymus cells.

transcription factor is likely found in active chromatin while depleted in inactive chromatin. The predicted enrichment profiles are consistent with experimental observations. For example, HistMod H3K36me3 is enriched on active chromatin.[53] At the same time, TF CBX5 and TF

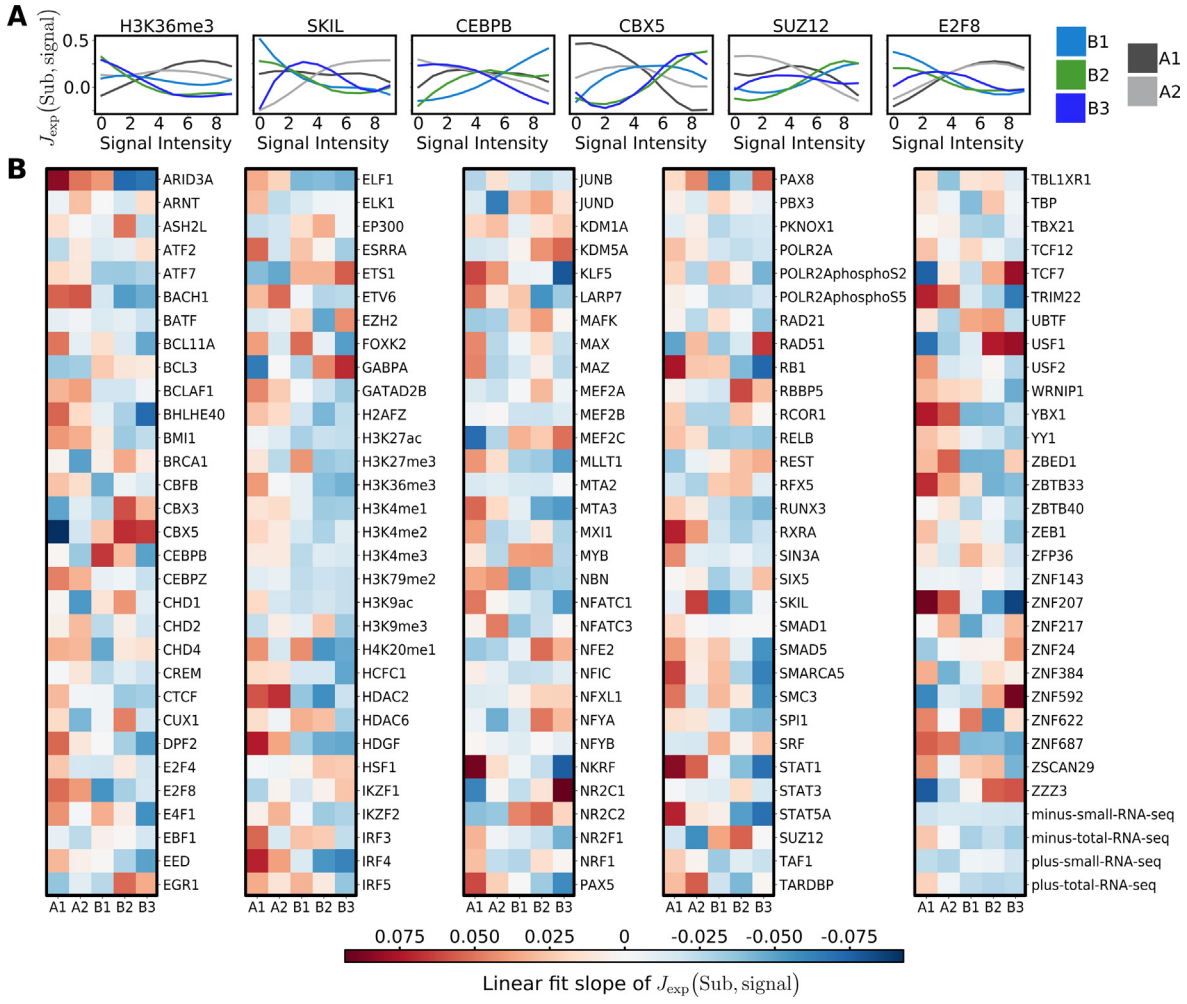

Figure 6. Model coupling energy terms demonstrate the relationship between subcompartment and epigenomic marks enrichment. (A) Energy coupling versus experiment signal intensity, plotted for various subcompartments. The panels show the curves for different epigenetic marks, including HistMod and TF (H3K36me3, SKIL, CEBPB, CEBX5, SUZ12, and E2F8), derived from training on GM12878-hg19 at 50 kbp resolution. (B) Subcompartment enrichment profile of each epigenetic mark. Each row of the heat maps corresponds to an epigenetic factor, each column represents a subcompartment, and the color depicts the intensity of coupling obtained from the linear slope of the energy coupling curves. A red hue indicates a strong association, while blue represents anticorrelation.

SUZ12 are likely to be found on inactive chromatin.[54,55] A linear fit was performed on the energy coupling curves for each subcompartment to better visualize the relationship between the enrichment of epigenetic marks and subcompartments. The positive slope of the fit indicates enrichment of the epigenetic mark on that subcompartment. In contrast, a negative slope is associated with depletion. Figure 6(B) presents the slopes for all the possible marks found in ENCODE for GM12878-hg19. Most of the Histone Modifications are enriched in A-type subcompartments while depleted in B-type regions, except for H3K27me3 and H3K9me3.[56] On the other hand, the transcription factors have diverse profiles of enrichment for different subcompartments. For example, CUX1 is predicted to be enriched in A1 and B2 subcompartments, while SIX5 is only expected in the B3 subcompartments. This enrichment analysis becomes highly relevant when characteristics of the subcompartments are incorporated into the discussion. For example, B2 and B3 subcompartments are correlated with lamina-associated domains.[11,12] Suppose an epigenetic mark is enriched exclusively on those subcompartments. In that case, we expect it to be located close to the nuclear envelope. Hence, from Figure 6(B) we expect that transcription factors such as EGR1, CBX5, USF1, and ZNF592 would be close to lamina due to enrichment on B2 and B3 compartments. In contrast, marks enriched on A1 and B1, such as H3K36me3, SMARCA5, BACH1, and H3K27me3, would be positioned towards the interior of the nucleus, as these subcompartments lack strong association with the

nuclear lamina.[11,12] Similarly, B2 is found close to nucleolus-associated domains,[11] which is consistent with the high slope found on two nucleolar transcription factors: UBF1 and ZZZ3.[47,48] Together, these results show how the parameters of PYMB may be interpreted to elucidate the correlations between epigenomic factors and the structural aspects, including features like nuclear positioning.

4. Conclusions

We present a neural-network model called PyMEGABASE (PYMB), capable of predicting subcompartment and compartment annotations from 1D epigenomic data, which reflects the biochemical composition of the epigenome. PYMB encodes the relationship between structural annotations and biochemical signals at a locus, enabling interpretable and transferable predictions. Building upon our previous work, MEGABASE,[13] we have significantly improved preprocessing and integrating multi-modal data, resulting in more reliable predictions.

Implemented in Python, PYMB software is user-friendly and includes automated steps for downloading user-defined data, processing, de novo training, and making predictions based on available data. PYMB can predict (sub) compartment annotations across cell lines and species (Figs. 2 and 3). This suggests that PYMB, in decoding the relationship between epigenetic enrichment and structural annotations, learns physicochemical rules transferable across cell types and organisms.

PYMB-predicted subcompartment sequences for a wide variety of human cell types and samples available in ENCODE, when projected in lower dimensions (two UMAP dimensions), show clustering consistent with their cellular identities (Figure 3). These predictions, yet to be tested experimentally, suggest that subcompartment annotation sequences may serve as markers of cellular identity. This work paves the way for investigating cell differentiation through the lens of changing subcompartments, raising questions about the relationship between these structural changes and cell-fate transitions.

Analyzing the chromosomal positioning of genes, we found that most human genes lie in A-type subcompartments, while B-type regions have low gene density across all cells (Figure 3). Furthermore, housekeeping genes are highly enriched in A1-type loci, indicating a strong correlation between gene density and subcompartment annotations. By using PYMB to predict compartment annotations (A and B) at finer resolutions (5 kbp) relevant to single genes, we observed higher average gene expression levels in the A compartment (Figure 4). Genes with high relative expression in a particular cell type are more likely to reside in the A compartment of that cell type. However, the subset of genes with low relative expression does not show a preference for any compartment, warranting future investigation.

PYMB's interpretability is one of its advantages. By investigating the trained parameters, we can elucidate the underlying rules that PYMB learns from the correlations between the epigenome and structural annotations (Figure 5). For example, PYMB suggests that most Histone Modifications are enriched in A-type compartments and depleted in B-type compartments, while the enrichment profile across subcompartments is highly diverse for Transcription Factors. Additionally, using the enrichment profile derived from the model's parameters, we can discover specific characteristics of epigenetic marks, such as nuclear positioning or activation levels of the regions where these motifs are bound or located.

In conclusion, PYMB serves as proof of principle that the epigenome contains enough information to predict the structural motif of a chromosomal locus. We explored the relationship across cell types, tissues, and species. However, much remains to be discovered about chromatin structural motifs at the compartment and subcompartment levels. Further research on the connection between gene expression, typically associated with biological function, and structural compartments present an intriguing future endeavor. The user-friendly nature of PYMB, combined with its high predictive power, will enable researchers to incorporate structural aspects of chromosomes into their investigations. Our software is publicly available at https://github.com/ed29rice/PyMEGABASE, including examples and tutorials written on Colab Jupyter Notebooks to facilitate easy access to annotation prediction using PyMEGABASE for the scientific community.

CRediT authorship contribution statement

Esteban Dodero-Rojas: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Matheus F. Mello:** Software, Writing – original draft, Writing – review & editing. **Sumitabha Brahmachari:** Conceptualization, Writing – original draft, Writing – review & editing. **Antonio B. Oliveira Junior:** Software, Writing – original draft, Writing – review & editing. **Vinícius G. Contessoto:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **José N. Onuchic:** Conceptualization, Supervision, Project administration, Writing – original draft, Writing – review & editing.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We want to thank Peter Wolynes, Michele Di Pierro, and Ryan Cheng 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, PHY-2014141, 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. ABOJ acknowledges the Robert A. Welch Postdoctoral Fellow Program. We would like to thank AMD for the donation of critical hardware and support resources from its HPC Fund that made this work possible.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jmb.2023.168180.

Received 15 December 2022;
Accepted 6 June 2023;
Available online 09 June 2023

Keywords:
Genome organization;
Epigenetics;
chromatin subcompartments;
maximum entro

References

[1]. Zheng, H., Xie, W., (2019). The role of 3d genome organization in development and cell differentiation, https://doi.org/10.1038/s41580-019-0132-4.

[2]. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., et al., (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336. https://doi.org/10.1038/nature14222.

[3]. Cremer, T., Cremer, C., (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2**, 292–301. https://doi.org/10.1038/35066075.

[4]. Bickmore, W.A., (2013). The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84. https://doi.org/10.1146/annurev-genom-091212-153515.

[5]. Dekker, J., Rippe, K., Dekker, M., Kleckner, N., (2002). Capturing chromosome conformation. *Science* **295**, 1306–1311. https://doi.org/10.1126/science.1067799. 3C.

[6]. Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., et al., (2006). Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347. https://doi.org/10.1038/ng1891, 4C.

[7]. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., Wit, E.D., Steensel, B.V., Laat, W.D., (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet.* **38**, 1348–1354. https://doi.org/10.1038/ng1896, 4C.

[8]. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., et al., (2006). Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 1299–1309. https://doi.org/10.1101/gr.5571506.1. 5C.

[9]. Lieberman-aiden, E., Berkum, N.L.V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., et al., (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome erez. *Science* **326**, 289–294.

[10]. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., et al., (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, 0826–0842. https://doi.org/10.1371/journal.pbio.0030157. chromosome territories.

[11]. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E. K., Bochkov, I.D., Robinson, J.T., Sanborn, A., Machol, I., et al., (2014). 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. https://doi.org/10.1016/j.cell.2014.11.021.

[12]. Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., Steensel, B.V., et al., (2018). Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *J. Cell Biol.* **217**, 4025–4048. https://doi.org/10.1083/jcb.201807108.

[13]. Pierro, M.D., Cheng, R.R., Aiden, E.L., Wolynes, P.G., Onuchic, J.N., (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc. Nat. Acad. Sci. USA* **114**, 12126–12131. https://doi.org/10.1073/pnas.1714980114.

[14]. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., et al., (2019). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889. https://doi.org/10.1093/nar/gkz1062. arXiv:https://academic.oup.com/nar/article-pdf/48/D1/D882/32874907/gkz1062.pdf.

[15]. Xiong, K., Ma, J., (2019). Revealing hi-c subcompartments by imputing inter-chromosomal

chromatin interactions. *Nat. Commun.* **10** https://doi.org/10.1038/s41467-019-12954-4.

[16]. Ashoor, H., Chen, X., Rosikiewicz, W., Wang, J., Cheng, A., Wang, P., Ruan, Y., Li, S., (2020). Graph embedding and unsupervised learning predict genomic sub-compartments from hic chromatin interaction data. *Nat. Commun.* **11** https://doi.org/10.1038/s41467-020-14974-x.

[17]. Liu, Y., Nanni, L., Sungalee, S., Zufferey, M., Tavernari, D., Mina, M., Ceri, S., Oricchio, E., et al., (2021). Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat. Commun.* **12** https://doi.org/10.1038/s41467-021-22666-3.

[18]. Zheng, S., Thakkar, N., Harris, H.L., Zhang, M., Liu, S., Gerstein, M., Aiden, E.L., Rowley, M.J., et al., (2022). Predicting a/b compartments from histone modifications using deep learning. *bioRxiv*. https://doi.org/10.1101/2022.04.19.488754. URL: https://www.biorxiv.org/content/early/2022/09/19/2022.04.19.488754. arXiv: https://www.biorxiv.org/content/early/2022/09/19/2022.04.19.488754.full.pdf.

[19]. Qi, Y., Zhang, B., (2019). Predicting three-dimensional genome organization with chromatin states. *PLOS Comput. Biol.* **15**, e1007024. https://doi.org/10.1371/journal.pcbi.1007024.

[20]. Cheng, R.R., Contessoto, V.G., Lieberman Aiden, E., Wolynes, P.G., Di Pierro, M., Onuchic, J.N., (2020). Exploring chromosomal structural heterogeneity across multiple cell lines. *eLife* **9**, e60312. https://doi.org/10.7554/eLife.60312.

[21]. Contessoto, V.G., Cheng, R.R., Hajitaheri, A., Dodero-Rojas, E., Mello, M.F., Lieberman-Aiden, E., Wolynes, P. G., Di Pierro, M., et al., (2021). The Nucleome Data Bank: Web-based resources to simulate and analyze the three-dimensional genome. *Nucleic Acids Res.* **49**, D172–D182. https://doi.org/10.1093/nar/gkaa818.

[22]. Contessoto, V.G., Cheng, R.R., Onuchic, J.N., (2022). Uncovering the statistical physics of 3D chromosomal organization using data-driven modeling. *Curr. Opin. Struct. Biol.* **75**, 102418. https://doi.org/10.1016/j.sbi.2022.102418.

[23]. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E., (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E* **87**, 012707.

[24]. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., et al., (2016). deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165.

[25]. Oliveira Jr, A.B., Contessoto, V.G., Mello, M.F., Onuchic, J.N., (2020). A scalable computational approach for simulating complexes of multiple chromosomes. *J. Mol. Biol.* **433**, 166700. https://doi.org/10.1016/j.jmb.2020.10.034.

[26]. Oliveira Junior, A.B., Estrada, C.P., Aiden, E.L., Contessoto, V.G., Onuchic, J.N., (2021). Chromosome modeling on downsampled Hi-C maps enhances the compartmentalization signal. *J. Phys. Chem. B* **125**, 8757–8767. https://doi.org/10.1021/acs.jpcb.1c04174.

[27]. Zerihun, M.B., Pucci, F., Peter, E.K., Schug, A., (2020). pydca v1. 0: a comprehensive software for direct coupling analysis of rna and protein sequences. *Bioinformatics* **36**, 2264–2265.

[28]. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Aiden, E.L., (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101. https://doi.org/10.1016/j.cels.2015.07.012.

[29]. Pierro, M.D., Zhang, B., Aiden, E.L., Wolynes, P.G., Onuchic, J.N., (2016). Transferable model for chromosome architecture. *Proc. Nat. Acad. Sci. USA* **113**, 12168–12173. https://doi.org/10.1073/pnas.1613607113. URL: https://www.pnas.org/content/113/43/12168.

[30]. Contessoto, V.G., Dudchenko, O., Aiden, E.L., Wolynes, P.G., Onuchic, J.N., Pierro, M.D., (2022). Interphase chromosomes of the Aedes aegypti mosquito are liquid crystalline and can sense mechanical cues. *Biophys. J.*. https://doi.org/10.1101/2022.02.01.478655.

[31]. Hand, D.J., Till, R.J., (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learn.* **45**, 171–186.

[32]. Wang, Z., Gerstein, M., Snyder, M., (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

[33]. Doganli, C., Sandoval, M., Thomas, S., Hart, D., (2017). Assay for transposase-accessible chromatin with high-throughput sequencing (atac-seq) protocol for zebrafish embryos. *In: Eukaryotic Transcriptional and Post-Transcriptional Gene Expression Regulation*. Springer, pp. 59–66.

[34]. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E., (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322.

[35]. Bannister, A.J., Kouzarides, T., (2011). Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395.

[36]. Kouzarides, T., (2007). Chromatin modifications and their function. *Cell* **128**, 693–705.

[37]. Jantzen, H.-M., Admon, A., Bell, S.P., Tjian, R., (1990). Nucleolar transcription factor hubf contains a dna-binding motif with homology to hmg proteins. *Nature* **344**, 830–836.

[38]. Ramana, C.V., Chatterjee-Kishore, M., Nguyen, H., Stark, G.R., (2000). Complex roles of stat1 in regulating gene expression. *Oncogene* **19**, 2619–2627.

[39]. Liu, G., Huang, Y.J., Xiao, R., Wang, D., Acton, T.B., Montelione, G.T., (2010). Solution nmr structure of the arid domain of human at-rich interactive domain-containing protein 3a: a human cancer protein interaction network target. *Proteins* **78**, 2170–2175.

[40]. Su, X., Wellen, K.E., Rabinowitz, J.D., (2016). Metabolic control of methylation and acetylation. *Curr. Opin. Chem. Biol.* **30**, 52–60.

[41]. Ngo, V., Chen, Z., Zhang, K., Whitaker, J.W., Wang, M., Wang, W., (2019). Epigenomic analysis reveals dna motifs regulating histone modifications in human and mouse. *Proc. Nat. Acad. Sci.* **116**, 3668–3677.

[42]. Pugacheva, E.M., Kubo, N., Loukinov, D., Tajmul, M., Kang, S., Kovalchuk, A.L., Strunnikov, A.V., Zentner, G. E., et al., (2020). Ctcf mediates chromatin looping via n-terminal domain-dependent cohesin retention. *Proc. Nat. Acad. Sci.* **117**, 2020–2031.

[43]. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V.G., et al., (2021). 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* **372**, 984–989. https://doi.org/10.1126/science.abe2218.

[44]. McInnes, L., Healy, J., Melville, J., (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. doi:10.48550/arXiv.1802.03426. arXiv:1802.03426.

[45]. Eisenberg, E., Levanon, E.Y., (2013). Human housekeeping genes, revisited. *TRENDS Genet.* **29**, 569–574.

[46]. Li, X.-B., Gu, J.-D., Zhou, Q.-H., (2015). Review of aerobic glycolysis and its key enzymes–new targets for lung cancer therapy. *Thoracic Cancer* **6**, 17–24.

[47]. Berglund, L., Björling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C.A.-K., Persson, A., Ottosson, J., et al., (2008). A genecentric human protein atlas for expression profiles based on antibodies. *Mol. Cell. Proteom.* **7**, 2019–2027.

[48]. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., et al., (2015). Tissue-based map of the human proteome. *Science* **347**, 1260419.

[49]. Gu, H., Harris, H., Olshansky, M., Eliaz, Y., 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. *BioRxiv*.

[50]. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., et al., (2010). The nih roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048.

[51]. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., et al., (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330.

[52]. Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., Ma'ayan, A., (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**

[53]. Huang, C., Zhu, B., (2018). Roles of h3k36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. *Biophys. Rep.* **4**, 170–177.

[54]. Lomberk, G., Wallrath, L., Urrutia, R., (2006). The heterochromatin protein 1 family. *Genome Biol.* **7**, 1–8.

[55]. Cao, R., Zhang, Y., (2004). Suz12 is required for both the histone methyltransferase activity and the silencing function of the eed-ezh2 complex. *Mol. Cell* **15**, 57–67.

[56]. Sharakhov, I.V., Sharakhova, M.V., (2015). Heterochromatin, histone modifications, and nuclear architecture in disease vectors. *Curr. Opinion Insect Sci.* **10**, 110–117.