bioSBM: A Random Graph Model to Integrate Epigenomic Data in Chromatin Structure Prediction

Alex Chen Yi Zhang [1,2,*] Angelo Rosa [2,†] and Guido Sanguinetti [2,‡]
[1] *Institute of Science and Technology Austria (ISTA), Klosterneuburg AT-3400, Austria*
[2] *Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

(Received 26 September 2024; accepted 9 September 2025; published 21 October 2025)

The spatial organization of chromatin within the nucleus plays a crucial role in gene expression and genome function. However, the quantitative relationship between this organization and nuclear biochemical processes remains under debate. In this study, we present a graph-based generative model, bioSBM, designed to capture long-range chromatin interaction patterns from Hi-C data and, importantly, simultaneously link these patterns to biochemical features. Applying bioSBM to Hi-C maps of the GM12878 lymphoblastoid cell line, we identified a latent structure of chromatin interactions, revealing seven distinct communities that strongly align with known biological annotations. Additionally, we infer a linear transformation that maps biochemical observables, such as histone marks, to the parameters of the generative graph model, enabling accurate genome-wide predictions of chromatin contact maps on out-of-sample data, both within the same cell line and on the completely unseen HCT116 cell line under RAD21 depletion. These findings highlight bioSBM's potential as a powerful tool for elucidating the relationship between biochemistry and chromatin architecture and predicting long-range genome organization from independent biochemical data.

DOI: 10.1103/gy1p-4256

I. INTRODUCTION

A characteristic feature of a eukaryotic cell, as opposed to archaea and eubacteria, is the sequestration of the cellular genome in a tight cellular space called the nucleus. In humans, the approximately two-meter-long chromatin filament is tightly packed in a nucleus of 5–10 μm in diameter. This packing is highly organized, as demonstrated by immunofluorescence microscopy experiments that show the heterogeneous subnuclear localization patterns of various proteins and histone marks thus hinting at the existence of functionally distinct compartments within the nucleus [1–4].

The advent of Chromatin Conformation Capture (3C) [5], particularly its now popular derivative Hi-C [6], allowed the mapping of chromatin contacts at a genome-wide level. Due to the fixation of nuclei with formaldehyde, which preserves information about the spatial proximity of linearly distal DNA loci, Hi-C experiments generate contact frequency maps that display nontrivial interaction motifs. One notable feature of the organization of chromatin unveiled by Hi-C is the segregation of the genome into two classes of domains dubbed A and B compartments, which are characterized by distinct interaction patterns. On top of these connectivity differences, A and B compartments have been shown to correlate with epigenetic marks associated with active (A) and silenced (B) transcriptional states [6]. Despite these correlations, more detailed microscopy studies and information from epigenetics (e.g., histone modifications and preferential binding of transcription factors [7–12]) suggest that the transcriptional state of the genome is more nuanced and the binary classification into A/B compartments may be excessively oversimplified. In particular, while A/B compartments and the subcompartments defined by Rao *et al.* [9] (further studied in numerous other works [13–15]) seem to point out the existence of a direct statistical correlation between epigenetic marks and 3D chromatin interaction patterns, there is comparably much less quantitative understanding of the "microscopic" processes at the origin of these correlations. In general, several attempts to explain the emergence and spatial organization of 3D compartments have invoked polymer-based models [16–19], which aim at rationalizing the observed structure as the consequence of direct, sequence-specific interactions between distal chromatin loci. While such models connecting the biochemistry of epigenetics to genome folding provide valuable insights, their reliance on polymer simulations results in high computational cost, which becomes especially problematic when one tries to scale simulations up to the size of a typical chromosome [19]. In contrast, more recent deep learning models scale efficiently and have achieved impressive success in predicting Hi-C contact maps. Early models make predictions using DNA sequence alone, but these predictions do not account for cell-type-specific variability [20–22]. More recently, deep learning approaches have started to incorporate 1D epigenetic signals [23], improving predictive accuracy across different cellular conditions. However, these models remain largely

*Contact author: alexchenyi.zhang@ist.ac.at
†Contact author: anrosa@sissa.it
‡Contact author: gsanguin@sissa.it

uninterpretable, making it difficult to connect their predictive power with underlying biological mechanisms.

In recent years, ideas from the field of network or graph theory [24] have emerged as a promising paradigm to study chromatin organisation at the mesoscopic level [25]. These methods avoid the computational overheads of microscopic polymer-based models by abstracting chromatin structure as a network of interactions, where DNA loci are treated as nodes and their contacts as edges. Such graph-based approaches not only were successfully used to reveal structural patterns [26–28] but also provided interpretable insights into the relationship between chromatin architecture and biological function [29–31].

In this paper, we propose bioSBM, an interpretable network model that directly links chromatin structure with biochemical features. bioSBM is based on the stochastic block model (SBM) [32,33], a class of generative network models that partition the network into communities based on interaction patterns, making them highly suitable for uncovering latent structures in chromatin interaction maps. The first SBM attempt to model long-range chromatin contacts was proposed in 2015 by Cabreros *et al.* [26]. bioSBM builds on this previous study by modulating this community structure by considering biochemical covariates such as histone modifications and transcription factor binding, therefore constructing a quantitative framework to understand the relationship between 3D chromatin organization and biochemical processes, a problem already studied in previous works such as [34]. Unlike traditional SBM's, which assign each node to a single community, bioSBM allows for mixed memberships, enabling genomic regions to participate in multiple communities simultaneously, thus capturing the context-dependent nature of chromatin interactions.

We apply our model to Hi-C data from the GM12878 lymphoblastoid cell line, where we identify interpretable community structures that include, but go beyond, the conventional A/B compartmentalization and subcompartments [9]. In addition to community detection, we can infer the map from biochemical features to the community composition of entire chromosomes, and we learn the interaction patterns that link the various communities. Finally, the results of our inference allow us to show that bioSBM can serve as a generative model capable of predicting Hi-C maps for unseen chromosomes and cellular conditions, further demonstrating its robustness and utility.

The paper is organized as the following: In Sec. II we provide a detailed overview of the *vanilla* stochastic block model and discuss how it can be adapted to better describe different types of analyzed data. In particular, we provide details of the data we utilize (distance-corrected, or *observed-over-expected*, Hi-C maps) in Sec. II A, and our customized version of the SBM in Sec. II B. Then in Sec. II C we introduce the main features of the inference algorithm specifically developed to compute posterior probabilities for our model, while leaving the mathematical details of its derivation to the Supplemental Material [35]. We present our main results in Sec. III, demonstrating the biological relevance and predictive power of our bioSBM model. Finally, in Sec. IV we discuss our results in the context of chromosome organization and conclude by highlighting, in particular, possible future applications.

II. MODEL AND METHODS

A. SBM and its generalization

SBMs are a particular class of random graphs. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices $\mathcal{V}$, representing entities $1, \ldots, N$ and a set of edges $(i, j) \in \mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, denoting pairwise interactions between these entities. Edges can be binary, indicating the presence or the absence of a link, or they can be weighted, reflecting the *strength* of interactions. Random graphs [36,37] denote graphs whose edges (or edge values) are generated according to a probability distribution, making the graph structure itself random.

SBMs have their roots in the world of social sciences [38,39], where they were used to model populations divided into subpopulations or *communities*. The central idea is that interactions between individuals are influenced by their community, creating a nontrivial structure in the interaction graph. This simple yet powerful idea made block models into popular models to study general types of relational data.

Numerous algorithms have been developed to detect community structures in complex networks and to make sense of them [22,23,26,40]. The box provides a schematic overview of the main ideas behind SBM's and of some of their most common extensions.

In our case, the relational data are derived from Hi-C contact frequency maps. Early Hi-C experiments demonstrated that contact frequency between genomic loci is strongly influenced by *linear proximity* or genomic distance, with the contact probability exhibiting a power-law decay as a function of this distance [6]. This scaling behavior can be explained based on fundamental polymer physics mechanisms that shape the three-dimensional folding of chromosomes [41–46]. Instead of using raw Hi-C data, it is often more insightful to study the so-called *observed-over-expected* (OE) Hi-C maps. These OE maps are derived as the logarithmic ratio between the actual contact frequencies recorded in Hi-C matrices and the expected contact frequencies based on genomic distances. These maps effectively highlight interaction patterns of chromatin by accounting for and removing the global polymeric effects that contribute to the power-law scaling. Significant interactions can be observed across large genomic scales, sometimes spanning entire chromosomes. To uncover a latent structure in these interaction patterns, we employ a weighted SBM, and to capture the complexity of community structures we use a so-called *mixed membership* version of the SBM (MMSBM; see Sec. II B) that allows the same genomic regions to belong to multiple communities.

Box 1: Flavors of stochastic block models (SBMs)

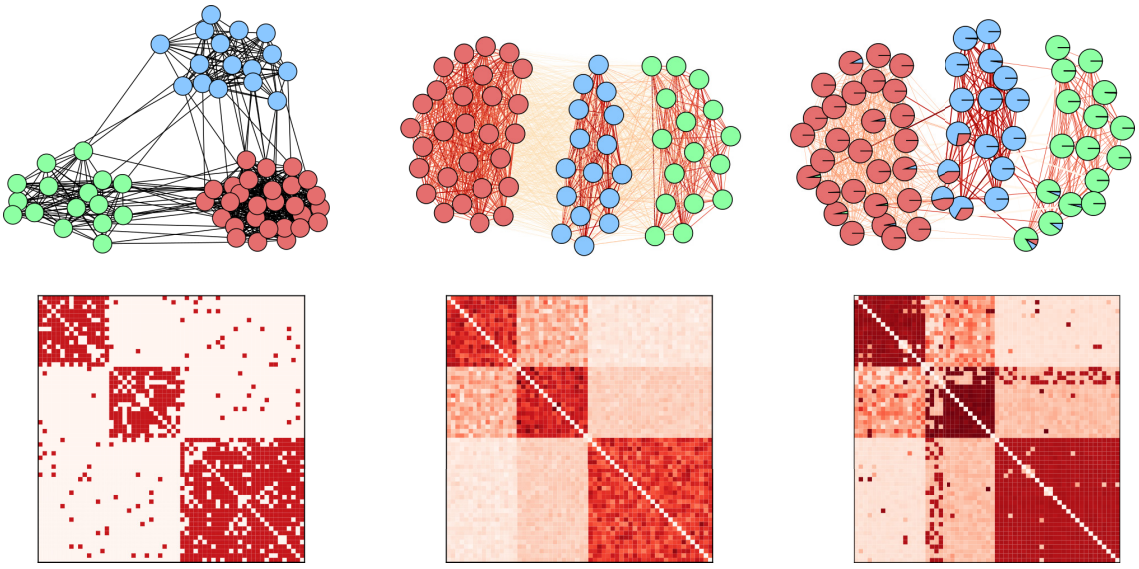
SBMs are a particular type of generative models used in network theory to describe the structure of networks by dividing nodes into communities or blocks. Each block represents a group of nodes that have a similar pattern of connections. The SBM assumes that the probability of a connection between any two nodes depends only on the blocks to which the nodes belong. This model helps us to understand the network's underlying structure and is commonly used for community detection [38–40]. The generative process defined by an SBM is as follows:

(1) Determine the communities and their total number, K.

(2) Assign each node to one of the K communities.

(3) For each pair of nodes, generate an edge with a probability that depends on the communities of the nodes. Specifically, generate a Bernoulli random variable with the parameter of the distribution depending on the colors of the two nodes.

The left panel of the figure shows a stochastic block model with $K = 3$ communities. In this example, the graph is an *assortative* SBM, meaning that intracommunity edge probabilities are higher than intercommunity ones.



In its basic version, the SBM is binary, i.e., the edges between nodes are either present or absent. However, many real-world networks involve weighted edges, where the connections between nodes have different strengths or capacities. To adapt the binary SBM for valued (weighted) graphs, we can modify the probability distribution of the edges given the colors or communities to which the two involved nodes belong. Instead of a Bernoulli random variable, we might use Poisson random variables for integer-value edges, or Gaussian random variables for real-value edges by specifying the means and variances of the distributions for each pair of distinct communities [47]. The central panel shows an instance of weighted SBM with Gaussian edges.

Another aspect we can tweak is the fact that in the traditional SBM, each node belongs to a single community or block. In many networks, nodes may exhibit characteristics of multiple communities. The *mixed membership* SBM (MMSBM) [48] addresses this by allowing nodes to belong to multiple communities by specifying a probability distribution over communities or membership proportions. In this paper, we will work with an SBM that has real-value edges and nodes with mixed membership proportions. The right panel shows an example of weighted MMSBM.

B. bioSBM: A covariate-dependent MMSBM for long-range chromatin contacts

A key difference between standard relational data and the biological setting we consider is the availability of a wealth of additional data in biology. While Hi-C measures contact patterns between chromosomal regions, a variety of biochemical assays, such as ChIP-seq [49], provide 1D genomic maps of specific epigenomic marks at such regions. In our model, we integrate the information from ChIP-seq data as a vector of biochemical covariates associated with each node, and which modulates the probability of each node belonging to the different communities.

Formally, bioSBM is a hierarchical Bayesian model [50,51] over weighted graphs. Graph nodes $i \in \{1, \ldots, N\}$ represent a set of contiguous genomic regions of fixed length. The observed weighted network has adjacency matrix Y, with Y_{ij} representing the logarithmic OE Hi-C contact frequency (Sec. II A) between region i and j. The observed weighted network is assumed to be generated according to the latent distributions of group memberships for each node/genomic region, as well as the matrices that specify group-group interaction strengths. Each node i has an associated membership proportions' vector θ_i, where θ_{ig} denotes the probability of node i belonging to group g, allowing nodes to belong

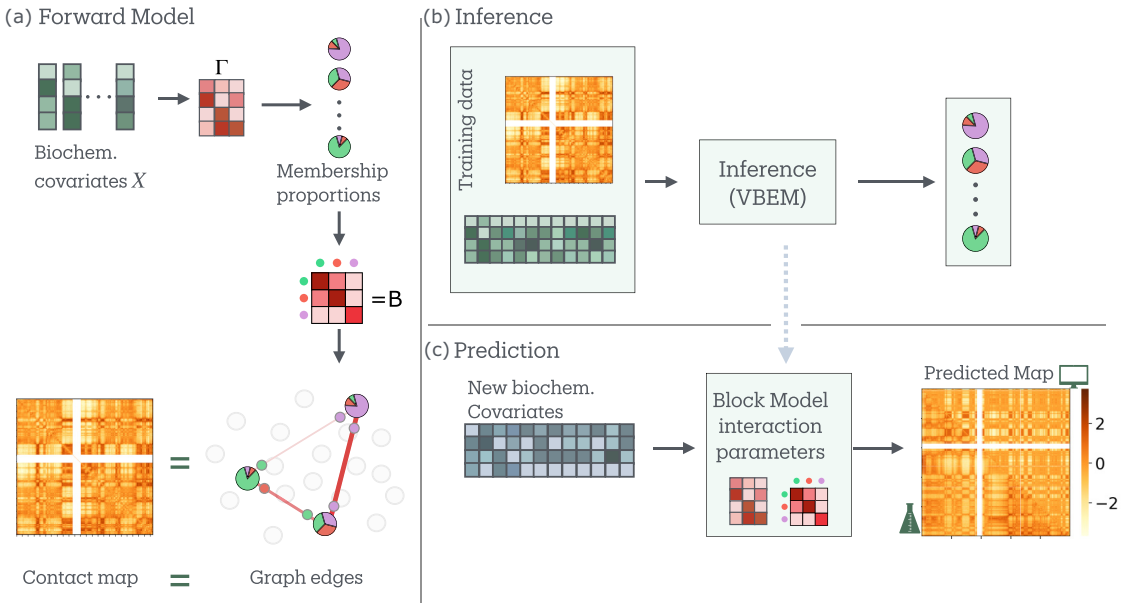


FIG. 1. (a) Forward model of the bioSBM model. Biochemical covariates associated with each node or genomic region are mapped into membership proportions. The membership proportions, together with the interaction strengths between pairs of communities (encoded in the block matrix **B**), determine the connectivity patterns of the graph. (b) Schematic representation of the inverse inference problem. Given a set of contact maps and covariates matrices, our inference procedure generates estimates of the latent membership proportion for the studied genomic loci, and infers the interaction parameters that characterize the model. (c) The model parameters inferred on training data are then used to make *de novo* predictions of contact maps, from biochemical features, for unseen data.

to multiple communities and display interactions that are context-dependent.

The group-group interaction strengths are defined by matrices B and σ^2, where B_{kg} and σ^2_{kg} represent, respectively, the mean and variance of the strength of interaction between community k and community g. For each pair of nodes (i, j), discrete variables z_{ij} and z_{ji} denote the group membership of i when interacting with j, and vice versa.

Then the edge weight is sampled from a Gaussian distribution parameterized by B and σ^2 matrices. Putting everything together, the generative process for bioSBM proceeds as follows:

(1) Each node has an associated vector of P features x_i, with $X \in \mathbb{R}^{P\times N}$ denoting the covariate matrix. The covariates correspond to biochemical data that can be associated with the different genomic regions, such as data from ChIP-seq assays.

(2) For every node i, we sample the distribution θ_i from the logistic normal distribution [52,53] with mean $\mu_i = \Gamma(x_i)$ and global covariance Σ, i.e.,

$$\eta_i \sim \mathcal{N}(\Gamma(x_i), \Sigma), \tag{1}$$

$$\theta_{ik} = \frac{\exp(\eta_{ik})}{\sum_{k'=1}^{K} \exp(\eta_{ik'})}. \tag{2}$$

$\Gamma : \mathbb{R}^P \to \mathbb{R}^K$ is the parametric function that maps biochemical features to probabilities over group memberships. In our specific implementation, Γ is simply a linear transformation encoded in a $K \times P$ matrix. Notice that Γ and Σ are global parameters shared among all nodes.

(3) For every pair of nodes (i, j) with $i = 1, \ldots, N$ and $j = 1, \ldots, i-1$, we define

$$z_{ij} \sim \text{Mult}(\theta_i), \quad z_{ji} \sim \text{Mult}(\theta_j), \tag{3}$$

with z_{ij} being the membership of node i interacting with node j and vice versa, sampled from the multinomial distribution with probabilities θ_i and θ_j respectively.

(4) Once the memberships of i and j are sampled, the weight Y_{ij} of the corresponding edge is sampled from a Gaussian whose parameters are encoded in the global B and σ^2 matrices:

$$P(Y_{ij}|z_{ij} = k, z_{ji} = g, B, \sigma^2) = \mathcal{N}\big(Y_{ij}|B_{kg}, \sigma^2_{kg}\big), \tag{4}$$

where the notation (used throughout the whole text) "$P(\cdot|\cdot)$" stands for the conditional probability of the variable(s) on the left given the variable(s) on the right. Schematically, the bioSBM model is illustrated as a graphical model in Fig. 1.

C. Posterior inference

To uncover the latent structure of chromatin interactions, we developed a posterior inference algorithm tailored to bioSBM. This algorithm estimates the latent parameters that best explain the observed Hi-C interaction data, integrating both chromatin interaction frequencies and biochemical covariates.

Our approach is based on variational inference, a method well suited for complex probabilistic models like the bioSBM, where exact inference is intractable. We optimize a variational lower bound on the model evidence, commonly referred to

as the Evidence Lower Bound (ELBO), which enables us to approximate the posterior distribution of the latent variables.

The variational inference procedure optimizes the ELBO, defined as

$$\mathcal{L}(q, \Psi) = \mathbb{E}_q[\log P(Y, \eta_{1:N}, Z|\Psi, X) - \log q(\eta_{1:N}, Z)], \quad (5)$$

where Y and X are for the OE Hi-C interaction data and the biochemical covariates, respectively, $\eta_{1:N}$ are the latent membership vectors ($\theta_{1:N}$ are the normalized versions), Z represents the community assignments for edges, Ψ includes the global model parameters, and the symbol $\mathbb{E}_q(\cdot)$ denotes the expectation value of the bracketed quantity with respect to the variational distribution q. The variational distribution $q(\eta_{1:N}, Z)$ approximates the true posterior distribution $P(\eta_{1:N}, Z|Y, X, \Psi)$.

Then the algorithm proceeds in two main steps:

(1) *Variational E-step*. We update the variational distributions of the latent variables, η_i and z_{ij}, by maximizing the ELBO with respect to the variational parameters. The factorized variational distributions take the following form:

$$q(\eta_i) \propto \exp\{\log P(\eta_i|\mu_i, \Sigma) + \mathbb{E}_{q(Z)}[\log P(Z|\eta_i)]\}, \quad (6)$$

$$q(z_{ij}) \propto \exp\left\{\mathbb{E}_{q(z_{ji})}[\log P(Y_{ij}|z_{ij}, z_{ji}, B)] + \mathbb{E}_{q(\eta_i)}[\log P(z_{ij}|\eta_i)]\right\}. \quad (7)$$

Here η_i are the continuous latent membership vectors, and z_{ij} are the discrete community assignments for edges.

(2) *Variational M-step*. This step involves optimizing the model parameters $\Psi \equiv (\Sigma, \Gamma, B, \sigma^2)$ with the current estimates of the variational distributions. The matrix Γ maps the biochemical covariates to the latent space, while Σ is the covariance matrix capturing the variability in the latent memberships. The matrices B and σ^2 describe the mean interaction strengths and variances between communities.

The iterative process of alternating between the E-step and the M-step continues until convergence, at which point the model parameters and variational distributions jointly provide an interpretation of the chromatin interaction patterns. For the detailed mathematical derivations and specific parameter update rules, refer to Sec. S1 [35].

III. RESULTS

To train the model, we have performed posterior inference using pairs of biochemical covariates and Hi-C matrices $(X^\mu, Y^\mu)_{\mu=1}^M$ for M chromosomes ($M = 11$ was chosen as the number of chromosomes in each training set in the experiments). More specifically, we employ the two sets of odd-numbered human chromosomes from 1 to 21 and the even-numbered human chromosomes from 2 to 22, and we use the model trained on one set to make predictions on the other and vice versa computing approximate posterior distributions over per-node latent membership vectors θ_i and the model parameters. We have applied the inference algorithm to data from the GM12878 lymphoblastoid cell line at a resolution of 100 kilo-basepairs; then, through Bayesian model selection, using evidence lower bound (ELBO) as a criterion, we determined that the optimal number of communities for our model is $K = 7$ (see Fig. S1 [35]). These results align with recent efforts to describe chromatin organization extending beyond the conventional A/B compartmentalization [6] and encompassing more nuanced frameworks, including subcompartments [9] and Interaction Profile Groups (IPGs) from Spracklin *et al.* [15]. Notably, a recent orthogonal approach based on polymer modeling by Esposito *et al.* [54] showed that their model could recapitulate Hi-C contact patterns with a set of "binding domains," which could be clustered into nine statistically significant, epigenetically distinct groups. The similarity in the number of inferred domain types, despite the methodological differences, highlights the consistency and biological relevance of both approaches.

The *maximum-a-posteriori* (MAP) estimates of the vectors θ_i provide the most plausible community membership proportions for each node, based on the observation of experimental Hi-C maps and associated biochemical covariates. Along with the θ_i values, the inference process also estimates the global parameters that characterize the generative model. A key parameter is the matrix Γ (Sec. II C), which represents the linear transformation mapping biochemical features to the probabilities of belonging to each community, offering insights into how biochemical factors shape chromatin structure. Additionally, the matrix B encodes the interaction strengths between all pairs of communities.

A. bioSBM explains the hierarchical organization of the chromatin in terms of epigenomic marks

The first output of the inference procedure is a set of K-dimensional ($K = 7$ for GM12878) *mixed-membership* probability vectors θ_i for all the genomic bins in the training data. These vectors represent the probabilistic community membership of each genomic region.

To facilitate biological interpretation, we perform a separate clustering step on these vectors using k-means clustering [55] to obtain discrete labels. Clustering into two groups allowed us to compare these clusters to known A/B compartments, while setting the number of clusters to six enabled comparisons with both the subcompartments defined by Rao *et al.* [9] and the Interaction Profile Groups (IPGs) computed with the algorithm by Spracklin *et al.* [15] (see Sec. S2.E [35]).

The results of the clustering showed a significant overlap with the established biological annotations. Figure 2 illustrates this comparison for chromosomes 16 and 19. The binary subdivision in A/B compartments as well as the more granular classification in subcompartments or IPGs can be captured from the full mixed membership vectors inferred by our model (see Fig. S9 [35] a comparison of centrome distances of the various clusters). Further validation of the inferred communities was performed by assessing the enrichment (see Sec. S2.C [35] for details) of each k-mean-derived cluster in the biochemical features x_i. These enrichments were compared (see Fig. 3) to those observed in A/B compartments, subcompartments, and annotations from [15], where we can see a near-perfect agreement of the discretized bioSBM results with the enrichment of the binary A/B classification and a very good agreement with the enrichment of the subcompartments defined by [9] and the IPGs.

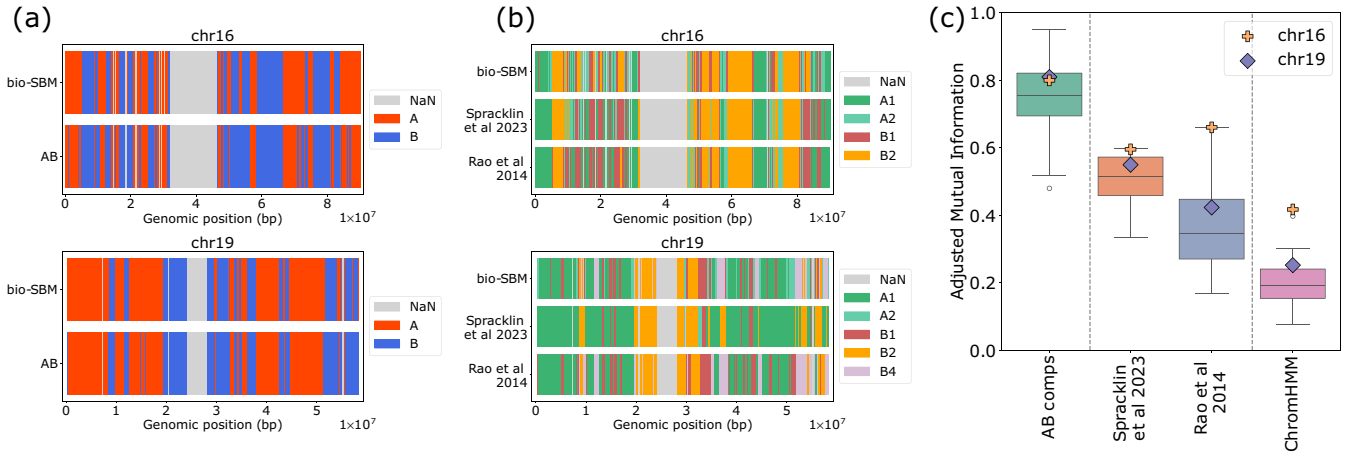


FIG. 2. The latent representation found through inference of the bioSBM model is biologically meaningful. The top rows in (a) and (b) represent the clusters obtained by applying k-means on the MAP membership vectors inferred through our algorithm. The bottom rows are biological annotations. (c) Adjusted Mutual Information score between the clustering obtained by bioSBM's membership vectors and annotations previously reported in the literature for the example chromosomes 16 and 19 and for all other chromosomes from 1 to 22.

We extended this analysis to other chromosomes (see Sec. S2.A, in particular Table S1 [35], for details on datasets used and Figs. S3 and S4 [35] for enrichment results for all chromosomes) and computed the similarity between the subdivision found by clustering the membership vectors and those based on the biological annotations using the Adjusted Mutual Information (AMI) score (see Sec. S2.D [35]). We obtained a median AMI score (which ranges from 0 to 1) of $\mathrm{AMI}^{\mathrm{A/B}} \simeq 0.76$ for the binary clustering, $\mathrm{AMI}^{\mathrm{IPG}} \simeq 0.52$ for the comparison with Spracklin *et al.* [15] and $\mathrm{AMI}^{\mathrm{subcomp}} \simeq 0.35$ for the comparison with Rao *et al.* [9]. These results suggest that the community structure inferred by bioSBM more closely resembles IPGs [15] than subcompartments, though it does not correspond one-to-one with either. Additionally, we performed a comparison with the 15-state annotations by ChromHMM [11] [Fig. 2(c)]. While there is a degree of overlap between the community structure detected by our model and the ChromHMM segmentation (median $\mathrm{AMI}^{\mathrm{chromHMM}} \simeq 0.18$), the level of accordance is notably lower than the comparison we made between bioSBM and other labelings based on Hi-C data. This difference is not entirely surprising, as bioSBM does incorporate biochemical covariates, but one of its core components is community detection based on chromatin contacts. In contrast, ChromHMM performs a genome segmentation based solely on epigenomic data.

Importantly, the model goes beyond merely segmenting chromatin regions based on their epigenetic features; it also illustrates how these different communities interact, as represented by the matrix B where each entry B_{kg} encodes the interaction strength between community k and g. Figure 4(a) shows that each of the seven inferred communities is associated both with distinct epigenomic patterns and with interaction patterns between communities. Interestingly, despite the apparent redundancy for some of the communities

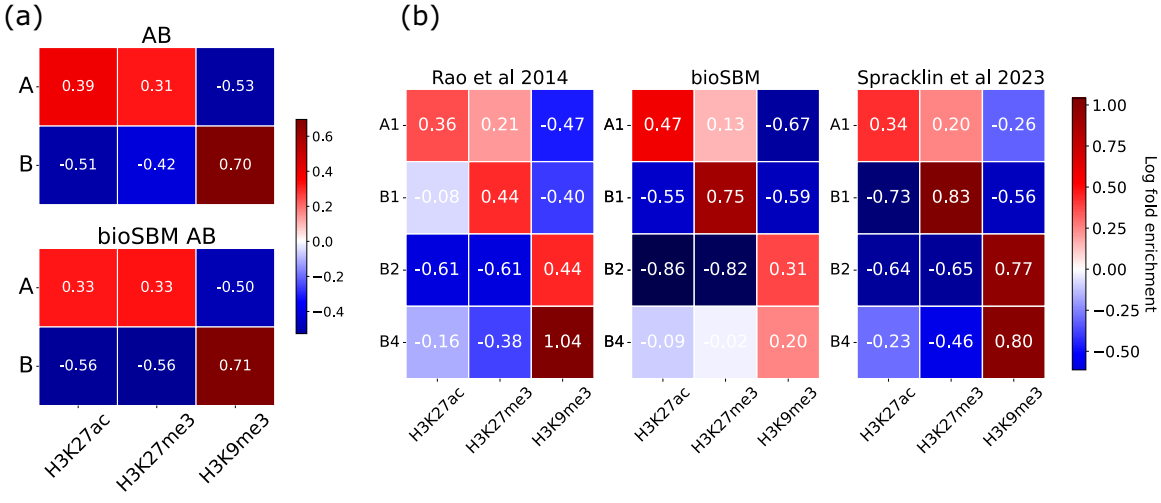


FIG. 3. Log-fold enrichment in biochemical features (such as the presence of histone marks) for chromosome 19. The patterns of enrichment for our clusters are in good agreement with the enrichment found for A/B compartments, subcompartments, and IPGs from [15]. The naming for the cluster labels for the clustering obtained applying k-means to bioSBM membership vectors, and for IPGs, have been mapped to match the nomenclature of subcompartments by [9].

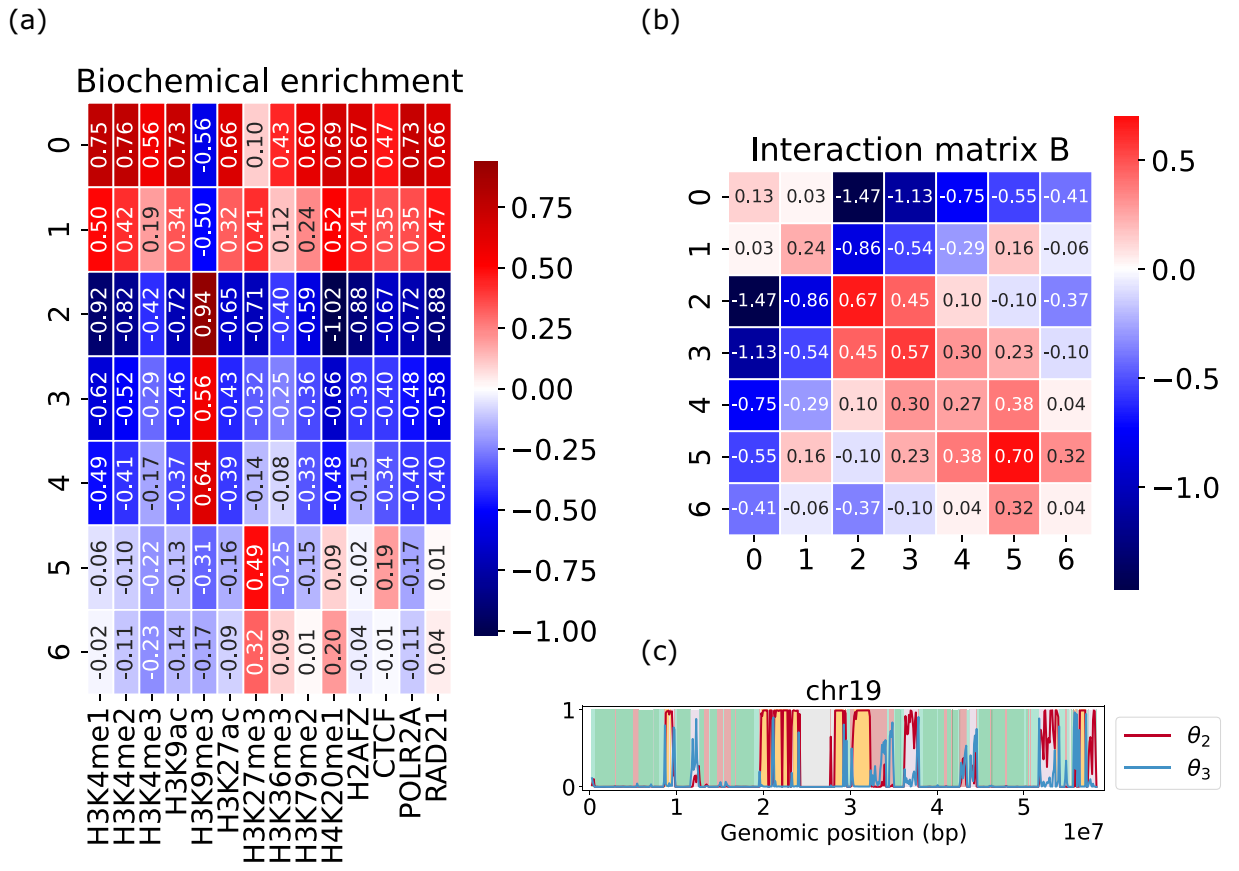


FIG. 4. (a) Enrichment of mixed membership communities in the biochemical features used in the model. Each community correlates with distinct epigenetic features. (b) Interaction intensities between all pairs of communities. (c) Example of membership proportions of communities 4 and 5, for whole chromosome 19.

at the level of epigenomic profiles, there are clear distinctions between the interaction patterns of different communities. For instance [see Fig. 4(c)], communities 0 and 1 have similar epigenetic profiles but regions predominantly associated with community 0 interact mainly with community 0 and community 1, while nodes predominantly associated with community 1 interact also with community 5. Another notable observation is that community 2 interacts positively only with the epigenetically similar communities 3 and 4, but community 3 interacts positively also with the very different community 5. The inset of Fig. 4(c) shows that regions with nonzero probabilities of belonging to communities 2 and 3 overlap with clusters corresponding to the B2 and B4 compartments. Here, however, nodes do not need to be categorized into one community or another, as they can share properties of multiple communities in different proportions (see also Fig. S2 [35] for some results characterizing the importance and degree of the nodes' *mixed membershipness*).

Altogether, these observations show that the bioSBM representation of chromatin interaction patterns provides a more nuanced description than the one provided by simple segmentation of the genome in different clusters.

B. bioSBM's predictive power

The previous section focused on an analysis of the interpretability of bioSBM, showing that the inferred model parameters recapitulate and extend previous observations on the epigenetic state of the chromatin and its compartments. In this section, we leverage the generative structure of bioSBM to test whether the simple representation of the genome contacts in terms of interactions between a limited number of communities and the fact that these communities can be determined starting from independent measurements of biochemical covariates is enough to reproduce long-range genome-wide chromatin contact patterns.

Specifically, given matrices B and Γ, inferred from some training chromosomes in some conditions, and biochemical measurements for test chromosomes (either different chromosomes from the same cell line, or chromosomes from a different cell line), we can compute their community structure as $\theta_i = \Gamma x_i$. With the predicted community structure, we use the matrix B to sample contact maps and compute their expectations, which can then be compared to experimentally obtained Hi-C maps. Refer to Sec. S2.B [35] for technical details and Figs. 1(b) and 1(c) for a schematic representation of the pipeline.

1. bioSBM assessment strategy

Testing machine learning models in biology is nontrivial, because the meaning of generalization is often unclear (see, e.g., Schreider *et al.* [56] for an in-depth discussion): one may seek to predict for unseen regions in the same chromosome

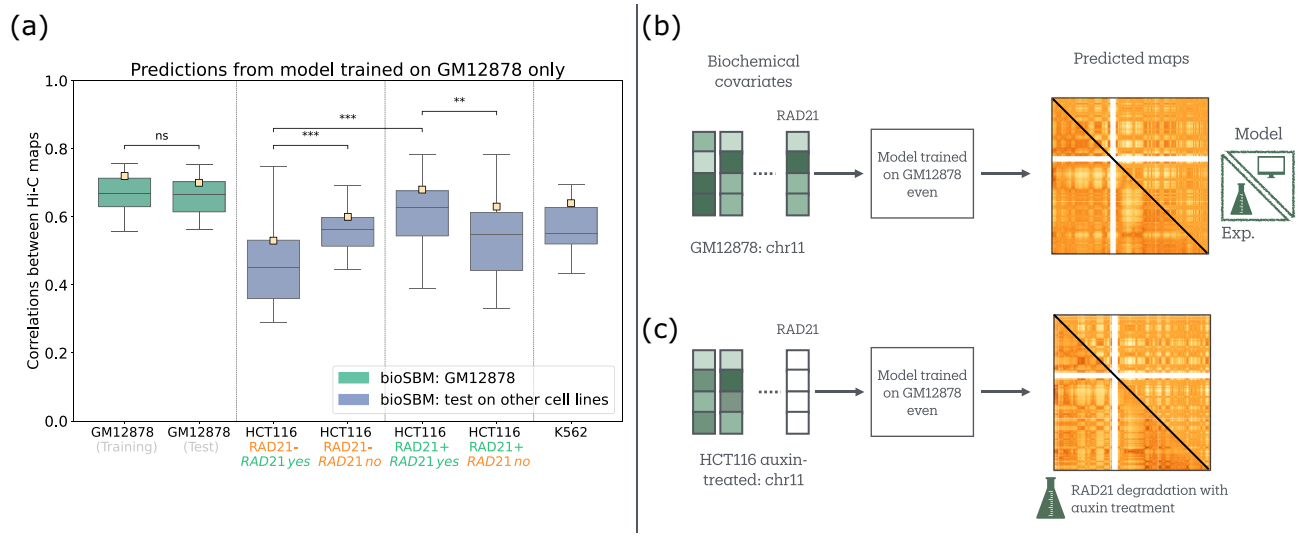


FIG. 5. (a) Box plots of correlations between experimental maps and model generated maps, where the individual data points are single chromosomes. Ivory squares are the correlation values for an example chromosome 11. To test whether the training performance was significantly higher or not than the test on different chromosomes on the same cell line we performed a one-tailed Mann-Whitney U test that yielded a negative answer (p value $= 0.82$). To test whether the improvement of predictions granted by removal of the RAD21 feature, for predictions on the AID-treated HCT116 cell line, was significant, we used the same test, which yielded a p value $= 1.20 \times 10^{-4}$ (significant difference). Also the predictive accuracy on HCT116 without AID treatment is significantly higher (p value $= 2.33 \times 10^{-5}$). See also Fig. S6 [35] for a comparison with the results obtained with a neural network parametrization of Γ. (b), (c) Schematization of the test performed on the AID-treated HCT116 cell line.

(hence, with similar local environment), or for different chromosomes in the same biological condition (cell line, tissue, etc.), or seek to extrapolate to completely new unseen conditions. In our experiments, we stress test bioSBM under the two most stringent conditions: first, we learn model parameters using ChIP-seq and Hi-C data from a subset of chromosomes in the lymphoblastoid cell line GM12878. Then we test model performance in predicting Hi-C data on the remaining chromosomes, using as input the relative ChIP-seq data [Figs. 1(b) and 1(c)]. We then seek to assess the performance of bioSBM on the task of predicting Hi-C data in two completely unseen cell lines: the K562 leukemia cell line, using ENCODE ChIP-seq data, and the colorectal carcinoma line HCT116, with and without a major external perturbation (the rapid degradation of the RAD21 loop extrusion factor [57]). On the HCT116 RAD21-cell line, we initially perform the prediction task using the parameters learned from the GM12878 dataset and ENCODE ChIP-seq data from HCT116, including tracks for RAD21. We then seek to simulate the ability of bioSBM to predict perturbation by repeating our prediction experiment (still with the same parameters), but simply setting to zero the ChIP-seq tracks for RAD21 [Figs. 5(b) and 5(c)]. This enables us to assess bioSBM predictions both under covariate shift (moving from GM12878 to HCT116 RAD21 cells), and its ability to model perturbations by changing the inputs in a deterministic way (setting to zero the RAD21 tracks).

2. *bioSBM predicts the structure of the chromatin in unseen cell lines*

The results of the tests described above are reported in Fig. 5(a). The first two columns show Pearson correlation between experimental log O/E Hi-C maps, and model-generated maps, for the training chromosomes, and for test chromosomes on the same cell line. Interestingly, the difference in accuracy between the training and test sets was marginal, with a one-tailed Mann-Whitney U test yielding a p value of 0.82 (meaning the difference in accuracy is not statistically significant); see Fig. S5 [35] for a comparison between predicted and real maps for all chromosomes. Therefore, bioSBM effectively generalizes across different chromosomes within the same cell line, suggesting that the inferred associations B_{kg} between communities and the biochemistry-to-structure map Γ reflect genuine chromatin interactions. Despite some differences in methodology and evaluation metrics, the correlation values we observe align well with the "distance-corrected" Pearson correlation values reported by Esposito *et al.* [54]. Their polymer-based approach, grounded in physical modeling of chromatin structure, offers valuable and direct mechanistic insight. While their model performs well on training data, it exhibits a noticeable decrease in predictive accuracy on unseen chromosomes. In contrast, bioSBM maintains robust predictive performance across chromosomes, highlighting its strong generalization capability. These complementary strengths illustrate the potential for integrating diverse modeling strategies to better understand chromatin organization. Additionally, we apply the model trained on half the chromosomes of GM12878, to make predictions on contacts on the other half of the chromosomes of the cancer cell lines HCT116 and K562 [last and third to last columns in Fig. 5(a)], from covariate tracks downloaded from the ENCODE [58] database. The results show that the model trained on GM12878 chromosomes is able to explain a large fraction of the variance in the data from

HCT116 and K562. Finally, we tested the model on a dataset for the HCT116 cell line, where auxin-inducible degron (AID) technology was used to degrade RAD21 [57].

The tests on HCT116 are especially informative, as Rao *et al.* [57] showed that RAD21 depletion in HCT116 cells preserves the global compartmentalization pattern—quantified in terms of the first eigenvector of the Hi-C correlation maps—but introduces *quantitative* changes, such as a marked increase in compartment strength. To quantify this effect, we computed chromosome-wise correlations between treated and untreated Hi-C maps for the HCT116 cell line (data from Rao *et al.* 2017) [57], yielding an average Pearson correlation of $\sim$0.78. This confirms that while the plaid pattern remains globally consistent, substantial quantitative differences exist that a predictive model like bioSBM should aim to capture.

We make predictions of contact frequencies with the model trained on GM12878 and independent covariates from the ENCODE Project [58] obtained from experiments on the HCT116 cell line (without AID-induced RAD21 degradation). We tested two scenarios: one where all covariates were used (denoted *RAD21 yes,* and one where the RAD21 track was zeroed out to simulate degradation (*RAD21 no*). We then compare these predicted maps to experimental maps of HCT116 cells without AID treatment (and therefore still have RAD21) or cells after 6 h of treatment, leading to RAD21 depletion. We denote the first experimental dataset as RAD21+, and the second one as RAD21− (see Table S1 [35] for data accession codes).

In the RAD21 yes vs RAD21+ setup, we observe a decrease in predictive accuracy, compared to predictions made on the same cell line as the training data, likely due to the mismatch between the learned map Γ, which includes RAD21, and a biological sample where the latter has been degraded. Additionally, RAD21 deletion could have induced slight changes also in the other covariates, further affecting predictive performance. Notably, when the RAD21 covariate was excluded from the input (RAD21 no), performance on the RAD21− experimental data improved significantly (p value $= 1.20 \times 10^{-4}$) (see Fig. S8 [35] for the analysis stratified by genomic distance). Conversely, the *in silico* RAD21 exclusion reduced the predictive accuracy for the RAD21+ data (see Fig. 5).

These results underscore the flexibility and robustness of bioSBM in capturing chromatin interactions under different cellular conditions (see also Fig. S7 [35] for an example of predictions with a model retrained on a subset of covariates).

IV. DISCUSSION AND CONCLUSIONS

The bioSBM model introduced in this study offers a novel approach for modeling long-range chromatin interactions by integrating Hi-C data with biochemical covariates such as histone modifications and binding of transcription factors. By employing a mixed membership stochastic block model, we capture a more refined and nuanced view of chromatin structure, extending beyond the traditional binary A/B and subcompartments framework. Our results show that the seven communities identified by bioSBM correlate with known epigenetic features, reinforcing the idea that chromatin interactions are closely tied to the biochemical landscape of the genome. The partial agreement with subcompartments suggests that our model may capture additional layers of chromatin interaction complexity that may be missed by conventional classification methods.

Although more abstract than polymer models, which explicitly take into account (to varying degrees of detail) the physical nature of the linear chromosomes, the latent representation learned by bioSBM remains biologically interpretable. The inferred associations between biochemistry and structure encoded in the linear map Γ and the communities interactions contained in B provide a starting point for systematically exploring more mechanistic descriptions of how chromatin folding is affected by nuclear biochemical processes. Importantly, bioSBM does not incorporate the underlying polymeric nature of chromatin, which is known to influence 3D genome architecture. Factors such as the dynamics of loop extrusion, maintained by SMC complexes, and the local mechanical state of the chromatin fiber can lead to distinct contact patterns that are not always explained by epigenomic marks alone. This has been demonstrated both experimentally (e.g., [59]) and via polymer simulations (e.g., [60]). These properties—absent from our current modeling framework—may account for some of the structural distinctions inferred by bioSBM, even among communities with seemingly similar epigenomic profiles.

The predictive power of the model is another important contribution. By leveraging biochemical covariates, bioSBM can accurately predict chromatin contacts across different chromosomes and cell lines, comparing favorably with state-of-the-art polymer approaches [54]. Notably, the model's robustness is highlighted by its performance on the HCT116 RAD21-cell line, where the removal of the RAD21 input covariate improved predictive accuracy, indicating that bioSBM can adapt to different chromatin environments and capture interactions under varying conditions, such as RAD21 depletion. Interestingly, bioSBM also compares favorably with recent deep-learning models, though their objectives and designs differ. While some deep learning approaches rely solely on DNA sequence input [20,21], limiting their ability to model cell-type specific variations, bioSBM explicitly integrates epigenetic features for this purpose. A recent model, Epiphany [23], also incorporated epigenomic data and reports predictive performance on held-out chromosomes in GM12878 that is in a similar range to what we observe for bioSBM. However, we note that a direct comparison is not straighforward, as Epiphany operates at a higher resolution and focuses on fine-scale perturbations. Additionally, its evaluation in other cell types is based on structural features such as insulation scores rather than whole-map correlation.

While our model cannot operate at this fine-grained scale, it remains effective in predicting the effects of larger-scale manipulations such as the AID-induced RAD21 reported above. Despite these differences, both approaches highlight the value of integrating epigenomic information into 3D genome prediction models. Crucially, bioSBM stands out for its interpretability: every parameter has a clear probabilistic semantic, allowing their analysis to uncover a clear biological meaning. This makes bioSBM a valuable, flexible tool for studying chromatin organization in diverse cellular contexts.

While this study focuses on Hi-C data, the so-called enrichment methods [61], such as Promoter Capture Hi-C (PCHi-C) [62] or ChIA-PET [63,64], may offer more functionally relevant perspectives on chromatin structure. PCHi-C, for example, highlights promoter region interactions, while ChIA-PET focuses on interactions involving specific proteins. These methods present challenges for polymer models because they capture interactions between noncontiguous regions. However, bioSBM, being graph-based, could accommodate these noncontiguous interactions with some adaptation. Though some graph-based studies exist that link nuclear biochemistry to chromatin structure as assayed by these enrichment-based data [29,30], a fully generative predictive model mapping biochemistry to the chromatin interaction patterns, such as that provided by bioSBM for Hi-C, has yet to be developed. With appropriate adjustments, bioSBM could readily be extended to accommodate these data types.

In conclusion, the bioSBM model successfully balances interpretability and scalability, offering a valuable tool for understanding the relationship between chromatin structure and its biochemical underpinnings. By incorporating biochemical features and allowing for mixed memberships, bioSBM provides a more flexible and biologically meaningful representation of chromatin interactions. The model's predictive power and adaptability across different cellular contexts underscore its potential for further applications, such as exploring chromatin dynamics in different developmental stages or disease states.

ACKNOWLEDGMENTS

G.S. acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 - Project FAIR "Future Artificial Intelligence Research". This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22]. A.R. acknowledges financial support from PNRR Grant CN 00000013 CN-HPC, M4C2I1.4, spoke 7, funded by Next Generation EU.

DATA AVAILABILITY

An implementation of the code for the inference algorithm described in this paper is available at [65].

[1] A. G. Matera, M. Izaguire-Sierra, K. Praveen, and T. K. Rajendra, Nuclear bodies: Random aggregates of sticky proteins or crucibles of macromolecular assembly? Dev. Cell **17**, 639 (2009).

[2] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, Biomolecular condensates: Organizers of cellular biochemistry, Nat. Rev. Mol. Cell Biol. **18**, 285 (2017).

[3] L. Wang, Y. Gao, X. Zheng, C. Liu, S. Dong, R. Li, G. Zhang, Y. Wei, H. Qu, Y. Li *et al.*, Histone modifications regulate chromatin compartmentalization by contributing to a phase separation mechanism, Mol. Cell **76**, 646 (2019).

[4] A.-M. Ladouceur, B. S. Parmar, S. Biedzinski, J. Wall, S. G. Tope, D. Cohn, A. Kim, N. Soubry, R. Reyes-Lamothe, and S. C. Weber, Clusters of bacterial RNA polymerase are biomolecular condensates that assemble through liquid–liquid phase separation, Proc. Natl. Acad. Sci. USA **117**, 18540 (2020).

[5] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, Capturing chromosome conformation, Science **295**, 1306 (2002).

[6] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom *et al.*, Comprehensive mapping of long range interactions reveals folding principles of the human genome, Science **326**, 289 (2009).

[7] M. H. Nichols and V. G. Corces, Principles of 3D compartmentalization of the human genome, Cell Rep. **35**, 109330 (2021).

[8] M. J. Rowley and V. G. Corces, Organizational principles of 3D genome architecture, Nat. Rev. Genet. **19**, 789 (2018).

[9] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell **159**, 1665 (2014).

[10] L. Hedström, R. Metzler, and L. Lizana, Enhancer-insulator pairing reveals heterogeneous dynamics in long-distance 3D gene regulation, PRX Life **2**, 033008 (2024).

[11] J. Ernst and M. Kellis, Chromatin-state discovery and genome annotation with chromhmm, Nat. Protocols **12**, 2478 (2017).

[12] C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, and M. Kellis, Regulatory genomic circuitry of human disease loci by integrative epigenomics, Nature (London) **590**, 300 (2021).

[13] Y. Wang, Y. Zhang, R. Zhang, T. van Schaik, L. Zhang, T. Sasaki, D. Peric-Hupkes, Y. Chen, D. M. Gilbert, B. van Steensel *et al.*, SPIN reveals genome-wide landscape of nuclear compartmentalization, Genome Biol. **22**, 36 (2021).

[14] K. Xiong and J. Ma, Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions, Nat. Commun. **10**, 5069 (2019).

[15] G. Spracklin, N. Abdennur, M. Imakaev, N. Chowdhury, S. Pradhan, L. A. Mirny, and J. Dekker, Diverse silent chromatin states modulate genome compartmentalization and loop extrusion barriers, Nat. Struct. Mol. Biol. **30**, 38 (2023).

[16] S. Bianco, D. G. Lupiáñez, A. M. Chiariello, C. Annunziatella, K. Kraft, R. Schöpflin, L. Wittler, G. Andrey, M. Vingron, A. Pombo *et al.*, Polymer physics predicts the effects of structural variants on chromatin architecture, Nat. Genet. **50**, 662 (2018).

[17] M. Di Pierro, R. R. Cheng, E. Lieberman Aiden, P. G. Wolynes, and J. N. Onuchic, De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture, Proc. Natl. Acad. Sci. USA **114**, 12126 (2017).

[18] D. Jost, P. Carrivain, G. Cavalli, and C. Vaillant, Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains, Nucl. Acids Res. **42**, 9553 (2014).

[19] A. C. Y. Zhang, A. Rosa, and G. Sanguinetti, Bottom-up data integration in polymer models of chromatin organization, Biophys. J. **123**, 184 (2024).

[20] G. Fudenberg, D. R. Kelley, and K. S. Pollard, Predicting 3D genome folding from DNA sequence with Akita, Nat. Methods **17**, 1111 (2020).

[21] R. Schwessinger, M. Gosden, D. Downes, R. C. Brown, A. M. Oudelaar, J. Telenius, Y. W. Teh, G. Lunter, and J. R. Hughes, DeepC: Predicting 3D genome folding using megabase-scale transfer learning, Nat. Methods **17**, 1118 (2020).

[22] J. Zhou, Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale, Nat. Genet. **54**, 725 (2022).

[23] R. Yang, A. Das, V. R. Gao, A. Karbalayghareh, W. S. Noble, J. A. Bilmes, and C. S. Leslie, Epiphany: Predicting Hi-C contact maps from 1D epigenomic signals, Genome Biol. **24**, 134 (2023).

[24] M. E. J. Newman, *Networks: An Introduction*, reprint ed. (Oxford University Press, Oxford, 2016).

[25] H. Ashoor, X. Chen, W. Rosikiewicz, J. Wang, A. Cheng, P. Wang, Y. Ruan, and S. Li, Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data, Nat. Commun. **11**, 1173 (2020).

[26] I. Cabreros, E. Abbe, and A. Tsirigos, Detecting community structures in Hi-C genomic data, arXiv:1509.05121.

[27] H. K. Norton, D. J. Emerson, H. Huang, J. Kim, K. R. Titus, S. Gu, D. S. Bassett, and J. E. Phillips-Cremins, Detecting hierarchical genome folding with network modularity, Nat. Methods **15**, 119 (2018).

[28] L. Hedström, A. C. Martínez, and L. Lizana, Identifying stable communities in Hi-C data using a multifractal null model, arXiv:2405.05425.

[29] V. Pancaldi, E. Carrillo-de Santa-Pau, B. M. Javierre, D. Juan, P. Fraser, M. Spivakov, A. Valencia, and D. Rico, Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity, Genome Biol. **17**, 152 (2016).

[30] V. Pancaldi, Chromatin network analyses: Towards structure-function relationships in epigenomics, Front. Bioinform. **1**, 742216 (2021).

[31] V. Pancaldi, Network models of chromatin structure, Curr. Opin. Genet. Dev. **80**, 102051 (2023).

[32] Y. J. Wang and G. Y. Wong, Stochastic blockmodels for directed graphs, J. Am. Stat. Assoc. **82**, 8 (1987).

[33] T. A. Snijders and K. Nowicki, Estimation and prediction for stochastic blockmodels for graphs with latent block structure, J. Classif. **14**, 75 (1997).

[34] X. Lan, H. Witt, K. Katsumura, Z. Ye, Q. Wang, E. H. Bresnick, P. J. Farnham, and V. X. Jin, Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages, Nucl. Acids Res. **40**, 7690 (2012).

[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/gy1p-4256 for details on the mathematical derivation of the variational algorithm and for supplemental figures and table.

[36] A. Frieze and M. Karoński, *Introduction to Random Graphs* (Cambridge University Press, Cambridge, 2015).

[37] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2001).

[38] C. Lee and D. J. Wilkinson, A review of stochastic block models and extensions for graph clustering, Appl. Netw. Sci. **4**, 122 (2019).

[39] J.-J. Daudin, F. Picard, and S. Robin, A mixture model for random graphs, Stat. Comput. **18**, 173 (2008).

[40] E. Abbe, Community detection and stochastic block models: Recent developments, J. Mach. Learn. Res. **18**, 1 (2018).

[41] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Crumpled globule model of the three-dimensional structure of DNA, Europhys. Lett. **23**, 373 (1993).

[42] C. Münkel and J. Langowski, Chromosome structure predicted by a polymer model, Phys. Rev. E **57**, 5888 (1998).

[43] A. Rosa and R. Everaers, Structure and dynamics of interphase chromosomes, PLoS Comput. Biol. **4**, e1000153 (2008).

[44] A. Rosa, N. B. Becker, and R. Everaers, Looping probabilities in model interphase chromosomes, Biophys. J. **98**, 2410 (2010).

[45] A. Y. Grosberg, How two meters of DNA fit into a cell nucleus: Polymer models with topological constraints and experimental data, Polym. Sci. Ser. C **54**, 1 (2012).

[46] J. D. Halverson, J. Smrek, K. Kremer, and A. Y. Grosberg, From a melt of rings to chromosome territories: The role of topological constraints in genome folding, Rep. Prog. Phys. **77**, 022601 (2014).

[47] M. Mariadassou, S. Robin, and C. Vacher, Uncovering latent structure in valued graphs: A variational approach, Ann. Appl. Stat. **4**, 715 (2010).

[48] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. **9**, 1981 (2008).

[49] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, Genome-wide mapping of in vivo protein-DNA interactions, Science **316**, 1497 (2007).

[50] G. M. Allenby, P. E. Rossi, and R. E. McCulloch, Hierarchical Bayes Models: A Practitioner's Guide, SSRN working paper, Ohio State University, University of Chicago, 2025, https://ssrn.com/abstract=655541.

[51] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., Texts in Statistical Science Series (CRC Press, Boca Raton, FL, 2014).

[52] J. Aitchison, The statistical analysis of compositional data, J. R. Stat. Soc.: Ser. B **44**, 139 (1982).

[53] D. M. Blei and J. D. Lafferty, A correlated topic model of science, Ann. Appl. Stat. **1**, 17 (2007).

[54] A. Esposito, S. Bianco, A. M. Chiariello, A. Abraham, L. Fiorillo, M. Conte, R. Campanile, and M. Nicodemi, Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states, Cell Rep. **38**, 110601 (2022).

[55] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory **28**, 129 (1982).

[56] J. Schreiber, R. Singh, J. Bilmes, and W. S. Noble, A pitfall for machine learning methods aiming to predict across cell types, Genome Biol. **21**, 282 (2020).

[57] S. S. P. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov *et al.*, Cohesin loss eliminates all loop domains, Cell **171**, 305 (2017).

[58] ENCODE (2024), https://www.encodeproject.org/.

[59] I. F. Davidson and J.-M. Peters, Genome folding through loop extrusion by SMC complexes, Nat. Rev. Mol. Cell Biol. **22**, 445 (2021).

[60] S. Leidescher, J. Ribisel, S. Ullrich, Y. Feodorova, E. Hildebrand, A. Galitsyna, S. Bultmann, S. Link, K. Thanisch, C. Mulholland *et al.*, Spatial organization of transcribed eukaryotic genes, Nat. Cell Biol. **24**, 327 (2022).

[61] I. Jerković and G. Cavalli, Understanding 3D genome organization by multidisciplinary methods, Nat. Rev. Mol. Cell Biol. **22**, 511 (2021).

[62] S. Schoenfelder, B.-M. Javierre, M. Furlan-Magaril, S. W. Wingett, and P. Fraser, Promoter capture Hi-C: High-resolution, genome-wide profiling of promoter interactions, J. Vis. Exp. **136**, e57320 (2018).

[63] G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H.-S. Ooi, C. Tennakoon *et al.*, ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing, Genome Biol. **11**, R22 (2010).

[64] G. Li, L. Cai, H. Chang, P. Hong, Q. Zhou, E. V. Kulakova, N. A. Kolchanov, and Y. Ruan, Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application, BMC Genomics **15**, S11 (2014).

[65] https://github.com/alex-chenyi-zhang/bioSBM-code.