Efficient Hi-C inversion facilitates chromatin folding mechanism discovery and structure prediction

Greg Schuette[1], Xinqiang Ding[1], Bin Zhang[1,*],

1 Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*** binz@mit.edu**

Abstract

Genome-wide chromosome conformation capture (Hi-C) experiments have revealed many structural features of chromatin across multiple length scales. Further understanding genome organization requires relating these discoveries to the mechanisms that establish chromatin structures and reconstructing these structures in three dimensions, but both objectives are difficult to achieve with existing algorithms that are often computationally expensive. To alleviate this challenge, we present an algorithm that efficiently converts Hi-C data into contact energies, which measure the interaction strength between genomic loci brought into proximity. Contact energies are local quantities unaffected by the topological constraints that correlate Hi-C contact probabilities. Thus, extracting contact energies from Hi-C contact probabilities distills the biologically unique information contained in the data. We show that contact energies reveal the location of chromatin loop anchors, support a phase separation mechanism for genome compartmentalization, and parameterize polymer simulations that predict three-dimensional chromatin structures. Therefore, we anticipate that contact energy extraction will unleash the full potential of Hi-C data and that our inversion algorithm will facilitate the widespread adoption of contact energy analysis.

Significance Statement

The three-dimensional organization of the genome is essential to many DNA-templated processes, and numerous experimental techniques have been introduced to characterize its features. High-throughput chromosome conformation capture experiments, or Hi-C, have proven particularly useful, reporting the interaction frequency between pairs of DNA segments *in vivo* and genome-wide. However, the polymer topology of chromosomes complicates Hi-C data analysis, which often employs sophisticated algorithms without explicitly accounting for the disparate processes affecting each interaction frequency. In contrast, we introduce a computational framework based on polymer physics arguments that efficiently removes the correlation between Hi-C interaction frequencies and quantifies how each local interaction influences genome folding globally. This framework facilitates the identification of mechanistically important interactions and the prediction of three-dimensional genome structures.

Introduction

The three-dimensional organization of the genome is an essential component of the processes that regulate gene expression and encode cellular diversity [1–4]. This biological importance has motivated the development of many experimental techniques to characterize genome structures and understand the mechanisms leading to their formation. Hi-C experiments [5, 6] have proven particularly insightful, as the high-resolution and genome-wide information they report has revealed structural features including chromatin loops [7], topologically associating domains [8, 9], and A/B compartments [6].

Hi-C reports the interaction frequency between each pair of genomic loci *in vivo*, and the experimental output is typically presented in two-dimensional contact probability maps whose axes index the genetic sequence [10, 11].

This data contains the information needed to identify non-random contacts of functional significance, including the promoter-enhancer pairs involved with gene regulation [7, 10]. However, interpreting Hi-C data can be complicated, as the polymer structure of chromosomes introduces several features that obscure the functionally relevant information [10–12]. Notably, configuration entropy penalizes the interaction between genomic loci far apart in sequence and couples contact formation between distinct genomic pairs [5, 12–16]. Existing approaches often address the genomic distance effect by normalizing contact probabilities to the average value of all contact pairs having the same genomic separation [11, 17–19], but they rarely account for the correlation between contact probabilities [14, 15] that obscures the location of specific interactions [20–23]. Therefore, it remains difficult to determine whether a genomic pair exhibiting high contact probability is energetically favorable due to a specific interaction between them (driving) or entropically favorable due to topological constraints associated with a driving interaction occurring elsewhere (passive). Differentiating between driving and passive interactions is necessary to identify which are biologically significant, deepen mechanistic insight, and improve upon existing approaches to predict three-dimensional genome structures using Hi-C [16, 20, 21, 24–35].

The information theoretic approach introduced by Zhang and Wolynes decouples the contribution of specific interactions from topological constraints in Hi-C data by converting contact probabilities into interaction energies that measure the attraction between genomic loci brought into proximity [20, 21]. Their algorithm performs this task by inverting a polymer model of chromatin derived to reproduce Hi-C data while obeying the maximum entropy principle [36]. The resulting interaction energies are independent of one another because the energy function of the model explicitly separates the effects of local interactions and polymer topology, which couples contact formation. Therefore, interaction energies directly correspond to the biologically important, specific interactions that cause the unique features in Hi-C probability maps to form, providing direct insight to the forces that drive genome folding.

While the interpretation of interaction energies is biologically meaningful, extracting them from Hi-C contact probabilities is nontrivial. The existing solutions [20–23] to this problem require many time-consuming molecular dynamics (MD) simulations to be performed sequentially, so the algorithms suffer from high computational complexity. Additionally, interaction energies extracted by these approaches suffer from statistical noise because adequately sampling conformations of long polymers during each algorithmic step is computationally infeasible using MD simulations .

In this study, we introduce a new method to address the challenges associated with converting Hi-C probabilities into interaction energies. This new approach utilizes a contact-space representation of polymers, in which a set of binary variables represents polymer conformations by indicating which monomers are in contact. An approachable, Ising-like Hamiltonian accounts for the topology-driven coupling between contacts and the distance effect in this space, and we introduce analytical expressions that parameterize the Hamiltonian using contact statistics. Extending this model to chromatin systems, another analytical expression converts Hi-C contact probabilities into contact energies, which are the contact-space analogue of interaction energies with an identical interpretation. Testing this extraction algorithm with an *in silico* model showed that the extracted contact energies accurately identify the mechanistically important contacts. Furthermore, we found that contact energies extracted from micro-C data [37] of hESC and HFF cells successfully reveal the location of anchors in chromatin loops. Contact energies extracted from compartment-scale data support a phase-separation model, as interactions between loci in the same compartment are almost universally more favorable than interactions between loci in different compartments. Finally, we show that contact energies can parameterize polymer models that reasonably reproduce Hi-C data at different resolutions. We believe these results indicate that the algorithm introduced herein will greatly simplify the interpretation of Hi-C data and facilitate the prediction of genome structures using computer simulations.

Theory

Maximum entropy inversion algorithm

As mentioned in the Introduction, the contact probability between two genomic loci is influenced by specific interactions that occur both locally and at distant interaction sites because the polymer structure of chromosomes couples all contact pairs [14, 15]. As a consequence, Hi-C contact probabilities are not a reliable

metric for locating biologically significant interactions, and converting from contact probabilities to interaction energies is essential to identify which pairwise interactions drive chromatin folding.

This conversion can be performed by inverting a polymer model of chromatin to reproduce Hi-C contact probabilities while obeying the maximum entropy principle [20, 21]. For example, starting with a homopolymer model that has the energy function $U_{\mathrm{hp}}(\boldsymbol{r})$, the least-biased model that reproduces Hi-C data can be derived by maximizing the excess entropy as [38–42]

$$U_{\mathrm{ME}}\left(\boldsymbol{r}\right) = U_{\mathrm{hp}}\left(\boldsymbol{r}\right) + \sum_{\substack{k,l=1\\k<l}}^{M} \alpha_{kl} f(r_{kl}), \tag{1}$$

where k and l index all M monomers, $\boldsymbol{r}$ contains their Cartesian coordinates, f is a threshold function that determines whether two monomers are in contact given their Euclidean distance r_{kl}, and α_{kl} is an interaction energy such that the unique set $\boldsymbol{\alpha} \equiv \{\alpha_{kl}\}$ can be determined from experimental contact probabilities. Physically, $\boldsymbol{\alpha}$ captures the influence of specific interactions by quantifying the interaction strength between each pair of genomic loci, whereas $U_{\mathrm{hp}}(\boldsymbol{r})$ captures the polymer topological effects. Thus, inverting this model separates the physical phenomena that contribute to Hi-C contact probabilities, and $\boldsymbol{\alpha}$ readily identifies structurally significant interactions.

Although this approach is theoretically sound, existing inversions [20–23] solve $\boldsymbol{\alpha}$ by iteratively performing polymer simulations, which are computationally expensive and suffer from statistical noise. Thus, further developing robust and efficient inversions would expand the number of feasible use cases for interaction energies and simplify Hi-C data analysis.

Modeling chromatin and extracting energies in contact space

Rather than inverting the polymer model in three dimensions, we approach the problem in a binary contact space. The polymer is mapped from its Cartesian representation $\boldsymbol{r}$ to its binary representation $\boldsymbol{q}$ via the transformation

$$\boldsymbol{q} \equiv \{q_{kl} = \theta(r_c - r_{kl})\, \forall\, (k,l) \mid k,l \in \{1,\ldots,M\},\ k < l-1\}\,, \tag{2}$$

where $q_{kl} = 1$ ($q_{kl} = 0$) indicates that monomer pair (k,l) is (not) in contact. In place of the general threshold function f, this conversion determines contact with the Heaviside step function θ and contact distance r_{c}, meaning q_{kl} is equal to

$$\theta(r_c - r_{kl}) = \begin{cases} 1 & \text{if } r_{kl} \leq r_{\mathrm{c}} \\ 0 & \text{Otherwise.} \end{cases} \tag{3}$$

Imposing $k < l-1$ ensures that each q_{kl} is unique while also ignoring nearest neighbors, which are trivially in contact.

To capture topological effects in contact space, we approximate $U_{\mathrm{hp}}(\boldsymbol{r})$ with the Ising-like expression

$$H_{\mathrm{hp}}\left(\boldsymbol{q}\right) = \sum_i h_i q_i + \sum_i \sum_{j>i} J_{ij} q_i q_j\,, \tag{4}$$

where $i,j \in \{1,\ldots,N\}$ ($N = (M-1)(M-2)/2$) index the unique contact pairs contained in $\boldsymbol{q}$. This Hamiltonian has found success in studies of protein folding [43, 44], and the linear and quadratic model parameters account for distinct effects of polymer topology contributing to contact probabilities [45]. In particular, $h_i \in \boldsymbol{h} \equiv \{h_i\}$ increases with the sequence separation between the monomers involved with contact i, while the quadratic parameters $J_{ij} \in \boldsymbol{J} \equiv \{J_{ij}\}$ capture the correlation between contact pairs.

Extending H_{hp} to a chromatin model that recreates Hi-C probabilities while maximizing information entropy, we find

$$H_{\mathrm{ME}}\left(\boldsymbol{q}\right) = \sum_i \left(h_i + \epsilon_i\right) q_i + \sum_i \sum_{j>i} J_{ij} q_i q_j. \tag{5}$$

In analogy to $\boldsymbol{\alpha}$, ϵ_i represents a *contact energy*, and $\boldsymbol{\epsilon} \equiv \{\epsilon_i\}$ captures all system-specific, biologically interesting contributions to Hi-C probabilities. Conveniently, the ensemble average $\langle q_i \rangle_{\mathrm{ME}}$ and the Hi-C contact probability

p_i are equal, where the subscript in $\langle\cdot\rangle_{\mathrm{A}}$ denotes the Hamiltonian used for the average. The simple form of the Hamiltonian allows $\boldsymbol{\epsilon}$ to be extracted from Hi-C contact probabilities $\boldsymbol{p} \equiv \{p_i\}$ with the following expression:

$$\epsilon_i = \ln\left(\frac{1-p_i}{p_i}\right) + \sum_{j\neq i} \ln\left[\frac{\left((1-p_j)+p_j e^{+J_{ij}}\right)^{p_i}}{\left((1-p_j)+p_j e^{-J_{ij}}\right)^{1-p_i}}\right] - h_i. \tag{6}$$

The Supporting Material details our derivation of Eq. 6, which relies on the inverse equations we developed to determine the topological parameters $\boldsymbol{h}$ and $\boldsymbol{J}$.

Determining the topological parameters of contact-space models

In order to solve the contact energies $\boldsymbol{\epsilon}$ using Eq. 6, we must first determine the values of the topological parameters $\boldsymbol{h}$ and $\boldsymbol{J}$. Their exact solution is not known, but the parameters in a generalized Ising model can often be well-approximated with inversions that use equilibrium statistics or a set of configurations. These inversions have been widely used to analyze protein sequence statistics and infer which amino acid pairs are in direct contact [46, 47]. For homopolymers, however, we found that existing approaches, including the mean-field [48, 49] and optimization-based [50–53], provide unphysical (Fig. S1) and noisy (Fig. S2) results, respectively. To overcome these obstacles, we developed inverse equations well-suited to homopolymer models in contact space.

The Supporting Material provides our derivation of the inverse equation we used to compute the coupling matrix $\boldsymbol{J}$,

$$J_{ij} = \sum_{\ell\neq i} \ln\left(\frac{\left(\widetilde{p}_i - \widetilde{\mathscr{C}}_{i\ell}\right)\left(\widetilde{p}_\ell - \widetilde{\mathscr{C}}_{i\ell}\right)}{\widetilde{\mathscr{C}}_{i\ell}\left(1-\widetilde{p}_i-\widetilde{p}_\ell+\widetilde{\mathscr{C}}_{i\ell}\right)}\right)\widetilde{p}_\ell\left(1-\widetilde{p}_\ell\right)\left(\left(\widetilde{C}_{\overline{i}\overline{i}}\right)^{-1}\right)_{j\ell}, \tag{7}$$

where $\widetilde{p}_i \equiv \langle q_i\rangle_{\mathrm{hp}}$ is the contact probability of contact i, $\widetilde{\mathscr{C}}_{ij} \equiv \langle q_i q_j\rangle_{\mathrm{hp}}$ is the *connected correlation* between contacts i and j, $\widetilde{C}$ is the $N\times N$ covariance matrix such that $\widetilde{C}_{ij} \equiv \widetilde{\mathscr{C}}_{ij} - \widetilde{p}_i\widetilde{p}_j$, and $\widetilde{C}_{\overline{i}\overline{i}}$ is the $(N-1)\times(N-1)$ submatrix of $\widetilde{C}$ in which the ith row and column have been removed. Similarly, the field parameters can be computed with

$$h_i = \ln\left(\frac{1-\widetilde{p}_i}{\widetilde{p}_i}\right) + \sum_{j\neq i}\ln\left[\frac{\left((1-\widetilde{p}_j)+\widetilde{p}_j e^{-J_{ij}}\right)^{1-\widetilde{p}_i}}{\left((1-\widetilde{p}_j)+\widetilde{p}_j e^{+J_{ij}}\right)^{\widetilde{p}_i}}\right], \tag{8}$$

whose derivation is also introduced in the Supporting Material. Therefore, both $\boldsymbol{h}$ and $\boldsymbol{J}$ can be determined using the first- and second-order contact statistics of a homopolymer, which we approximated with polymer simulations.

Methods

Simulating the homopolymer to obtain contact statistics

We carried out MD simulations in reduced units, using the software LAMMPS [54] to obtain an ensemble of homopolymer configurations that provided contact statistics. The homopolymer consists of 500 monomers, and nearest-neighbor monomers are connected by class2 bonds, implemented as

$$U_{\mathrm{bond}}(r_{kl}) = \kappa_2(r_{kl}-r_0)^2 + \kappa_3(r_{kl}-r_0)^3 + \kappa_4(r_{kl}-r_0)^4, \tag{9}$$

with bond length $r_0 = 1$ and force constants $\kappa_2 = \kappa_3 = \kappa_4 = 20$. Excluded volume effects were not included to facilitate the relaxation of polymer configurations. Langevin dynamics maintained the system temperature of 1.0 using a damping coefficient of 10. One billion total time steps of length $\Delta t = 0.005$ were performed, and configurations were saved every 5000 time steps to provide 200,000 total configurations.

We mapped each simulated conformation to contact space by implementing Eq. 2 in Python with contact distance $r_{\mathrm{c}} = 1.76$. The distance between monomers, r_{kl}, was computed with mdtraj 1.9.6 [55], then the simulated contact probabilities and covariances were computed with PyTorch 1.11.0 [56]. All further analysis used NumPy 1.19.2 [57].

We computed statistics for contacts separated by less than 200 monomers. Simulating a much longer polymer helps minimizes edge effects, which alter the statistics of contacts involving monomers near the end of the polymer [14, 15]. We further minimized noise in these simulated statistics by considering polymer topology; specifically, many contact probabilities and correlations should be identical in an infinite homopolymer because the corresponding interactions are topologically identical [14, 15]. Accordingly, we used the average value of topologically identical statistics to minimize uncertainty in their values, and the Supporting Material contains a detailed discussion of this process.

Simulating polymers with specific interactions

In addition to the above homopolymer simulation, we performed simulations to generate *in silico* contact probability maps and to evaluate the quality of contact energies extracted from Hi-C data. These simulations again featured a polymer model consisting of 500 monomers connected by class2 bonds. We also applied non-bonded contact potentials between monomers k and l in the following form

$$U_{\text{contact}}(r_{kl}) = \frac{\alpha_{kl}}{2}\left[1 + \tanh\left(\eta(r_{0.5} - r_{kl})\right)\right] \quad \text{for } r_{kl} < r_c^{\text{con}} \tag{10}$$

using decay rate $\eta = 5$, interaction distance $r_{0.5} = 1.75$, cutoff distance $r_c^{\text{con}} = 3$. Our analysis focused on the central 200 monomers in the polymer to minimize edge effects, so we indexed the monomers with $k, l \in \{-149, \ldots, 350\}$ to maintain consistency between sections.

When using the extracted contact energies to recreate Hi-C probability maps, we also applied a Lennard-Jones potential between all monomer pairs to account for the excluded volume effect of chromatin and better match the separation dependence of contact probabilities observed in chromatin. The potential is defined as

$$U_{\text{LJ}}(r_{kl}) = 4\varepsilon\left[\left(\frac{\sigma}{r_{kl}}\right)^{12} - \left(\frac{\sigma}{r_{kl}}\right)^{6}\right] \qquad r_{kl} < r_c^{\text{LJ}} \tag{11}$$

with $\varepsilon = 0.35$, $\sigma = 1.0$, and cutoff $r_c^{\text{LJ}} = 2.6$. In addition, we reduced simulation length to 120,000,000 time steps when recreating Hi-C maps, performed 10 simulations in parallel, and increased save frequency to once per 1000 time steps. This provided 6,000,000 conformations to analyze while improving computation time.

Preparing Hi-C data

Micro-C experiments, a Hi-C variant, provided the *in vivo* contact probabilities used throughout this work. The .mcool files containing the Micro-C dataset for hESC and HFF cells were downloaded from the 4DN Data Portal with accession numbers 4DNFI9GMP2J8 and 4DNFI9FVHJZQ, respectively [37]. We loaded balanced Hi-C contacts probabilities from these files at the resolution of interest for the entire chromosome of interest using Cooler 0.8.11 [58]. The average of all real-valued probabilities from contacts with the same loop size, $K \equiv |l - k|$, served as the unnormalized, background-averaged contact probability $\bar{p}_K^{(\circ)}$. We further processed the 200-bin region of interest on its own, using cooltools 0.5.1 [59] to remove zero-valued probabilities via adaptive coarse-graining then undefined probabilities via linear interpolation. Finally, we divided all contact probabilities by $\bar{p}_1^{(\circ)}$ to normalize the data, satisfying our assumption that nearest neighbors have contact probability 1. The normalized data are denoted $\bar{\boldsymbol{p}} \equiv \{\bar{p}_K\}$ and $\boldsymbol{p} \equiv \{p_i\}$ for loop size-averaged and full Hi-C contact probabilities, respectively.

Obtaining compartment profiles

To aid the analysis of compartment-level contact energies, we obtained compartment profiles using the cooltools 0.5.1 [59] implementation of eigenvector decomposition [60] in Python. Positive and negative values in the first eigenvector, E1, correspond to genomic regions in the A- and B-compartment, respectively. In addition to the aforementioned Micro-C datasets [37], this process required the GC content associated with each genomic bin in the Hi-C data. To obtain this information, we downloaded the UCSC version of the December 2013 assembly of the human genome (GRCh38 Genome Reference Consortium accession number GCA_000001405.15) [61] in

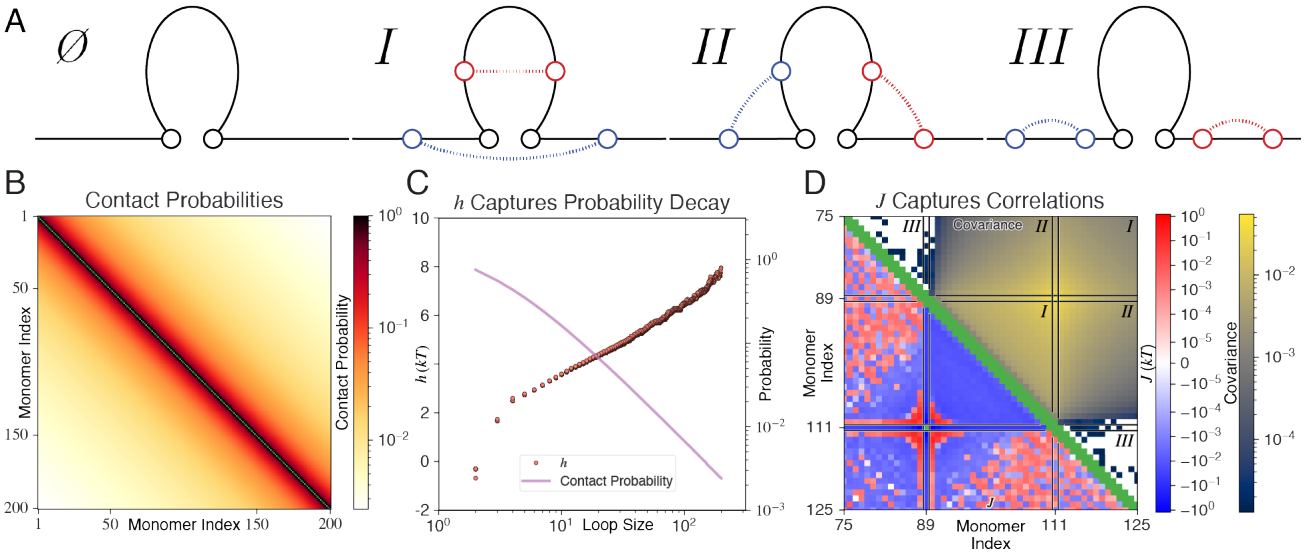

Figure 1. Parameters in the Ising-like Hamiltonian capture polymer topological effects in contact space. (a) Illustration of various topological relationships between loops. $\emptyset$ illustrates the loop associated with one contact. This loop resides within or fully contains loops in region I, partially overlaps with loops in region II, and is independent of loops in region III. (b) Contact probabilities between pairs of monomers calculated from MD simulations of a homopolymer model. (c) Dependence of the field parameters $\boldsymbol{h}$ (left axis) and average contact probabilities (right axis) as a function of the loop size. (d) The correlation (upper triangle) and J_{ij} coupling parameters (lower triangle) between contacts $i = (89, 111)$ and $j = (k, l)$ are compared. Black lines indicate the border between regions I, II, and III, and green indicates nonexistent values.

FASTA format from the UCSC Genome Browser database [62], loaded it using bioframe 0.3.3 [63], and determined the fraction of GC content in each genomic bin using the bioframe function frac_gc. Finally, we used the eigs_cis function in cooltools to perform eigenvector decomposition for the region of interest only.

Results and Discussion

In the Theory section, we introduced an efficient algorithm that uses Hi-C contact probabilities to infer the strength of non-bonding contact potentials between each pair of monomers in polymer models of chromatin. This required representing polymers in contact space, where a set of binary variables, $\boldsymbol{q}$, indicates which pairs of monomers are in contact. The Ising-like Hamiltonian in Eq. 4 approximates the free energy function of a homopolymer in this space and, therefore, captures the generic effects of polymer topology. Specifically, $\boldsymbol{h}$ accounts for the entropic penalty associated with forming each contact, while $\boldsymbol{J}$ combines the higher-order effects of polymer connectivity to capture contact correlations. Following a maximum entropy argument, Eq. 5 incorporates contact energies, $\epsilon_i \in \boldsymbol{\epsilon}$, to reproduce Hi-C contact probabilities and account for the local interactions between monomer pairs. The strength of these local interactions is independent of the polymer topological effects that couple contact formation and correlate Hi-C contact probabilities, so contact energies directly relate pairs of genomic loci to the biologically significant interactions that drive chromatin folding. Therefore, extracting contact energies from Hi-C contact probabilities distills the biologically significant information contained in Hi-C data and should simplify its analysis.

Contact-space parameters capture polymer topological effects

The Ising-like Hamiltonian of homopolymers in contact space accounts for the effects of polymer topology on contact probability and the coupling between contact sites with $\boldsymbol{h}$ and $\boldsymbol{J}$, respectively. We computed these parameters using contact statistics obtained from MD simulations of a homopolymer, thus ensuring that

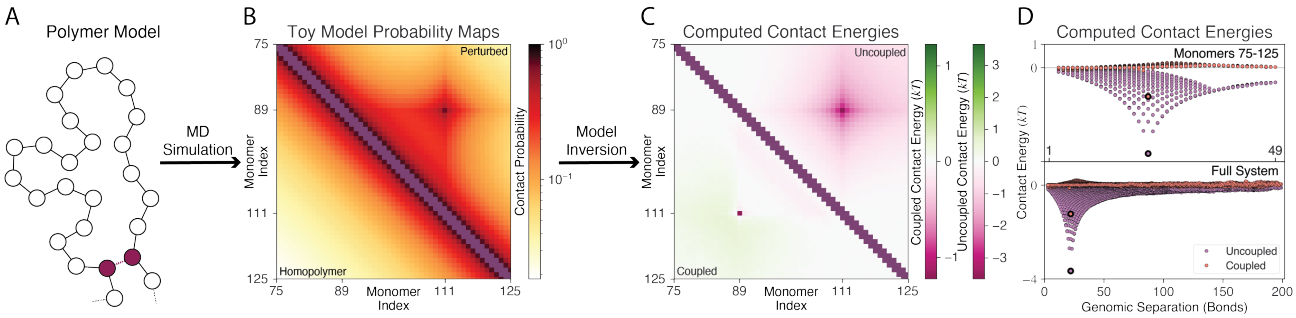

Figure 2. The energy extraction algorithm successfully identifies the specific contact of an *in silico* chromatin model. A flowchart illustrates our proposed algorithm using data from the *in silico* experiment. (A) Starting with the initial homopolymer model, we applied an attraction between monomers 89 and 111, illustrated in purple. (B) Comparing the simulated contact probabilities for this model (upper triangle) to those of the homopolymer model (lower triangle) reveals the effects of loop formation. (C) Comparing the contact energies extracted from the simulated contact probabilities using the coupled ($\boldsymbol{\epsilon}$, lower triangle) and uncoupled ($\boldsymbol{\epsilon}^{\mathrm{uc}}$, upper triangle) models illustrates the effectiveness of $\boldsymbol{J}$ in accounting for coupling in the polymer system. (D) Plotting contact energies against their corresponding separation further emphasizes the uncorrelated nature of contact energies extracted using $\boldsymbol{J}$.

topological effects alone contribute to their value. The Methods section provides simulation details, and Fig. 1 presents the results. As expected, the average contact probabilities, $\langle q_i \rangle_{\mathrm{hp}}$, decrease with distance from the diagonal in Fig. 1*B* since larger loops are less likely to form. For illustrative purposes, we also plotted the correlation between a specific contact, that involving monomers 89 and 111, and other contacts in the upper triangle of Fig. 1*D*. However, the qualitative trends are not affected by this particular choice (see Fig. S3).

Three topologies determine the correlation between various contacts in a homopolymer [14, 15], and Fig 1*A* illustrates them as regions in which one contact can reside with respect to another. For example, in region *I*, the larger loop completely encloses the smaller one, so the formation of one loop facilitates the formation of the other by reducing its configurational entropy penalty. The partially overlapping loops in region *II* are correlated for the same reason. On the other hand, the non-overlapping loops in region *III* are generally uncorrelated because the configuration entropy of one loop is unaffected by the other. Indeed, the simulated pairwise correlations in Fig. 1*D* agree with these topological predictions.

We used the first- and second-order contact statistics, i.e. all $\langle q_i \rangle_{\mathrm{hp}}$ and $\langle q_i q_j \rangle_{\mathrm{hp}}$, to compute $\boldsymbol{J}$ and $\boldsymbol{h}$ with Eqs. 7 and 8, respectively. Fig. 1*C* shows that contact probability and loop size are linearly related on a log scale, as anticipated [13–15]. h_i accounts for this effect by increasing with loop size. Meanwhile, the coupling parameters, J_{ij}, track the correlation between contacts. As with pairwise correlations, we plotted coupling parameters associated with the contact between monomers 89 and 111 in the lower triangle of Fig. 1*D*. Most of the J_{ij} are negative in regions *I* and *II*, supporting the correlation of the corresponding loops. The values are much smaller in magnitude in region *III* due to the independence of the loops.

An *in silico* model validates the energy extraction algorithm

With the Ising-like Hamiltonian of homopolymers parameterized, we extended the contact-space model to a chromatin-like model by introducing a set of contact energies, $\boldsymbol{\epsilon} \equiv \{\epsilon_i\}$, in the Hamiltonian in Eq. 5. This model can exactly reproduce Hi-C contact probability maps, and each map is associated with a unique set of contact energies. Our algorithm extracts this set of energies from Hi-C contact probabilities by applying the inverse Eq. 6 to all contacts. The resulting energies are local quantities because $\boldsymbol{h} \equiv \{h_i\}$ and $\boldsymbol{J} \equiv \{J_{ij}\}$ separately account for the effects of polymer topology, so they lack the cross-correlation that obscures biologically important information in Hi-C contact probabilities.

We assessed the quality of extracted contact energies by investigating an *in silico* chromatin model consisting of 200 monomers. In addition to the homopolymer interactions considered in the prior subsection, this model included a contact potential to promote loop formation. For simplicity, we only applied one non-bonded

interaction, and it used the potential defined in Eq. 10 with $\alpha_{kl} = -4k_BT$ between monomers 89 and 111. Conceptually, this represents a biologically significant interaction stabilizing a chromatin loop, e.g. a promoter-enhancer interaction. Fig. 2*A* portrays this step.

Polymer simulations of this model provided the *in silico* Hi-C contact probabilities, which Fig. 2*B* compares to the homopolymer probabilities. As anticipated, the favorable interaction between monomers 89 and 111 directly increased the contact probability at that site and indirectly increased the contact probability at many mechanistically insignificant contact sites. With no pairwise interaction between the monomers at these alternative sites, the coupling effect of polymer topology is singularly responsible for the change in their interaction frequency.

Next, we extracted contact energies using Eq. 6, where p_i represents the *in silico* Hi-C contact probabilities. As shown in the lower triangle of Fig. 2*C* and with orange markers in Fig. 2*D*, a prominent, negative-valued contact energy is associated with the attraction between monomers 89 and 111. In excellent agreement with the simulated model, near-zero contact energies appear at all other sites.

For comparison, the upper triangle of Fig. 2*C* and purple markers in Fig. 2*D* plot the equivalent contact energies provided by an uncoupled model, denoted $\boldsymbol{\epsilon}^{\mathrm{uc}} \equiv \{\epsilon_i^{\mathrm{uc}}\}$, meaning that all $J_{ij} = 0$ when solving Eq. 6 and Eq. 8. Because $\boldsymbol{J}$ accounts for the coupling effect of polymer topology, the contribution of coupling between contacts is not removed in this case. Thus, unlike contact energies extracted while accounting for coupling, these maintain the correlations observed in the contact probability map. This crucial difference emphasizes the importance of $\boldsymbol{J}$ when extracting contact energies and illustrates the quality of J_{ij} computed with our inversion.

Contact energies facilitate the identification of mechanistically important interactions

Chromatin looping and gene regulation are intricately connected, as persistent loops often bring promoter and enhancer regions into proximity [64–66]. Analytical techniques that use Hi-C data have expanded our understanding of the regulation of chromatin loops themselves by identifying the location of loop anchors and relating them to sequence information. However, identifying these interactions often requires requires sophisticated algorithms [7, 67–72], as the aforementioned influence of polymer topology complicates the relationship between the functional importance of an interaction and its probability. Because our algorithm separates these effects from the influence of local interactions, we anticipate that contact energy extraction will facilitate the identification of functionally important interactions.

To demonstrate their usefulness in this task, we extracted contact energies from Hi-C contact probabilities covering several regions of HFF and hESC cells; these were preprocessed according to the procedure described in the Methods section. In addition to extracting $\boldsymbol{\epsilon}$ from the Hi-C contact probabilities $\boldsymbol{p}$, we extracted $\bar{\boldsymbol{\epsilon}}$ from the loop size-averaged contact probabilities $\bar{\boldsymbol{p}}$. While the former represents the contact energies required to reproduce Hi-C probabilities, the latter represents the contact energies required to reproduce the probability decay of the relevant chromosome, meaning they capture nonspecific effects. Therefore, we analyzed the difference $\Delta\boldsymbol{\epsilon} \equiv \boldsymbol{\epsilon} - \bar{\boldsymbol{\epsilon}}$ to further isolate the biologically interesting pairwise interactions. For comparison, we further computed contact energies extracted in the uncoupled case, in which $J_{ij} = 0$ everywhere. Notably, the uncoupled energies simplify to

$$\Delta\epsilon_i^{\mathrm{uc}} = \epsilon_i^{\mathrm{uc}} - \bar{\epsilon}_i^{\mathrm{uc}} = \ln\left(\frac{\bar{p}_i(1-p_i^*)}{p_i^*(1-\bar{p}_i)}\right) \approx -\ln\left(\frac{p_i^*}{\bar{p}_i}\right) \tag{12}$$

in the low probability limit, representing the normalization used to address sequence separation-related probability decay in other methods, e.g. [17–19].

Fig. 3*A* shows the contact energy map for a 1 Mb region of chromosome 1 in hESC cells at 5 kb resolution, with $\Delta\boldsymbol{\epsilon}$ and $\Delta\boldsymbol{\epsilon}^{\mathrm{uc}}$ shown in lower and upper triangle, respectively. The strong correlation between uncoupled energies in each non-overlapping loop illustrates the topology-driven coupling of contact formation. $\boldsymbol{J}$ accounts for this effect when extracting energies with the coupled model, so the vast majority of loop-related contacts exhibit near-zero contact energy in that case despite their altered probability, leaving a small minority of contacts predicted to be strongly attractive. To emphasize this difference, Fig. 3*B* plots the energies enclosed by each box in Fig. 3*A* against the genomic separation between the relevant Hi-C bins. The most negative coupled contact energy associated with each loop is circled in black for easy identification, and the uncoupled energy of the same

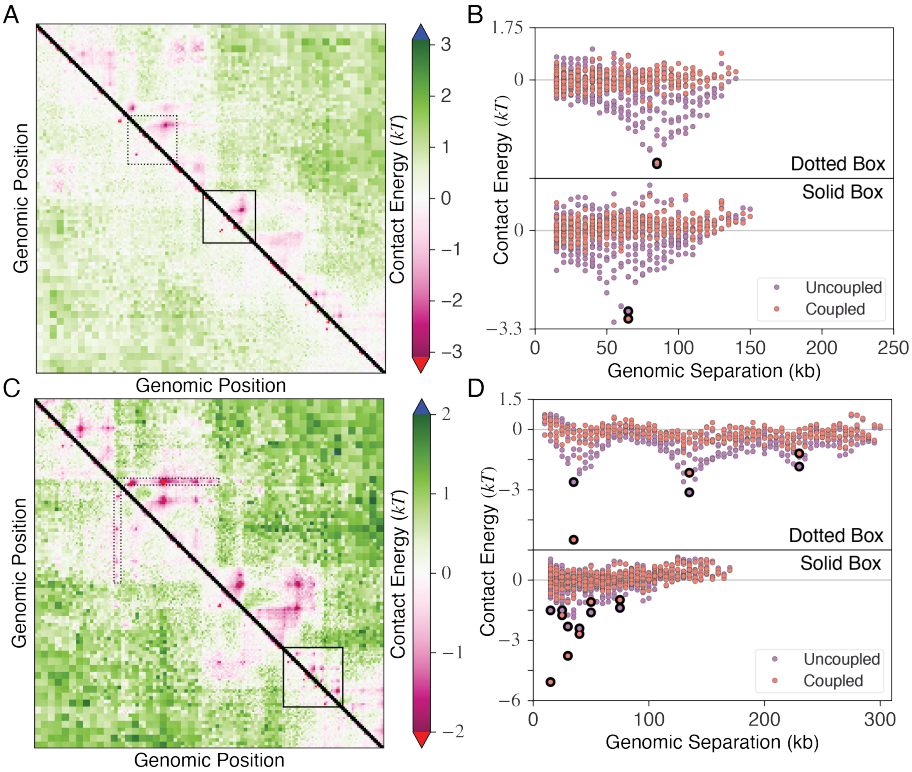

Figure 3. Contact energy maps contain prominent low-energy interactions, revealing the location of chromatin loop anchors. (A) Comparison of contact energies extracted with the coupled ($\Delta\boldsymbol{\epsilon}$, lower triangle) and uncoupled ($\Delta\boldsymbol{\epsilon}^{\mathrm{uc}}$ upper triangle) model. Each extraction used Hi-C contact probabilities from a 1 Mb region of chromosome 1, genomic positions $155,858,000 - 156,858,000$, at 5 kb resolution, and the final data are displayed as such. Each boxed region contains a chromatin loop. (B) Contact energies from the boxed regions in (A) are plotted against their corresponding genomic separation (distance from the map's diagonal), with $\Delta\boldsymbol{\epsilon}$ in orange and $\Delta\boldsymbol{\epsilon}^{\mathrm{uc}}$ in purple. (C) Same plot as (A) but for HFF cells. (D) Same plot as (B) but for the boxed regions in (C).

interaction is circled for comparison. These energies overlap significantly, whereas coupled energies are greatly reduced at nearby interaction sites. This clearly illustrates the uncorrelated nature of contact energies and indicates that our algorithm can facilitate the identification of mechanistically important contacts.

We note that, to highlight the contribution of each interaction to the contact probability map, we multiplied all p_i and $\bar{p}_i$ by 3 prior to computing contact energies with Eq. 6. Contact energies extracted without the multiplication factor are qualitatively similar (Fig. S4). However, increasing Hi-C contact probabilities, which are significantly smaller than the values found in homopolymer simulations, enhances the difference between the specific interaction energies for loop anchors and the non-specific ones. Scaling probabilities in this way could serve as a practical strategy to facilitate the identification of chromatin loops when analyzing contact energies. However, this additional multiplication factor is not necessary and was not used in later sections.

We analyzed the same region of HFF cells, with the corresponding results shown in Figs. 3*C* and 3*D*. Our algorithm performs well despite the overlapping loops present in this differentiated cell type, which presents a greater challenge. However, the apparent prominence of significant interactions depends on loop size in these features, which we attribute to the complex interplay between nested interactions, the simplified $\boldsymbol{J}$ used to analyze their effects, and the scaling of contact probabilities. (A similar distance dependence does not appear in the contact energies analyzed in Fig. S4.)

Analyzing other regions with (Fig. S5) and without (Fig. S6) scaling p_i and $\bar{p}_i$ by 3 supports the ability of contact energies to facilitate the identification of structure-driving interactions. Similarly, extracting contact

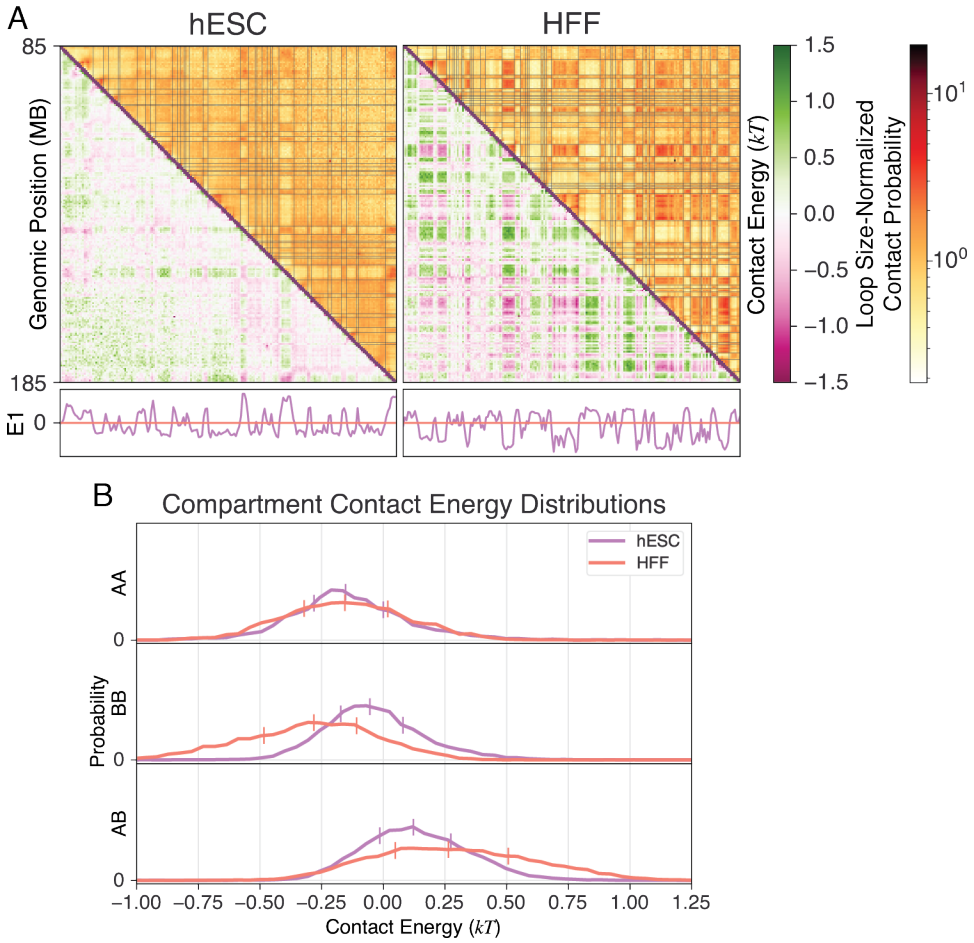

Figure 4. Contact energies indicate that intra-compartment interactions are more favorable than inter-compartment interactions. (A) Using the coupled model, contact energies ($\Delta\epsilon$, lower triangle) were extracted at 500 kb resolution in the 100 Mb region spanning genomic positions $85,000,000$-$185,000,000$ in chromosome 4 of H1 hESC (left) and HFF (right) cells. Using eigenvector decomposition, the sign of the first eigenvector (E1) indicates compartment identity for each genomic locus, and each fine black line on the loop size-normalized Hi-C contact probability maps (upper triangle) indicates a boundary between loci in different compartments. (B) Contact energies that correspond to intra- and inter-compartment interactions maintain different probability distributions.

energies from Hi-C data processed at 25 kb resolution provides largely uncorrelated results both with (Fig. S7) and without (Fig. S8) scaling p_i and $\bar{p}_i$ by 3, illustrating the robustness of our algorithm.

Contact energies support the microphase separation of A/B compartments

At large scales, the genome is organized into self-interacting and characteristically distinct compartments [6]. Specifically, A compartments are enriched with euchromatin and active histone modifications, while B compartments are enriched with heterochromatin and silencing histone modifications [40]. These compartments may emerge from a phase separation mechanism driven by the interaction between histone modifications and potentially mediated by additional protein molecules [12, 73–95]. This mechanism requires interactions within a compartment to be more favorable than interactions between compartments, i.e. lower in energy. Contact energies extracted from Hi-C data allow a direct test of the phase separation hypothesis.

We extracted contact energies for a 100 MB region of chromosome 4 from Hi-C data of hESC and HFF cells

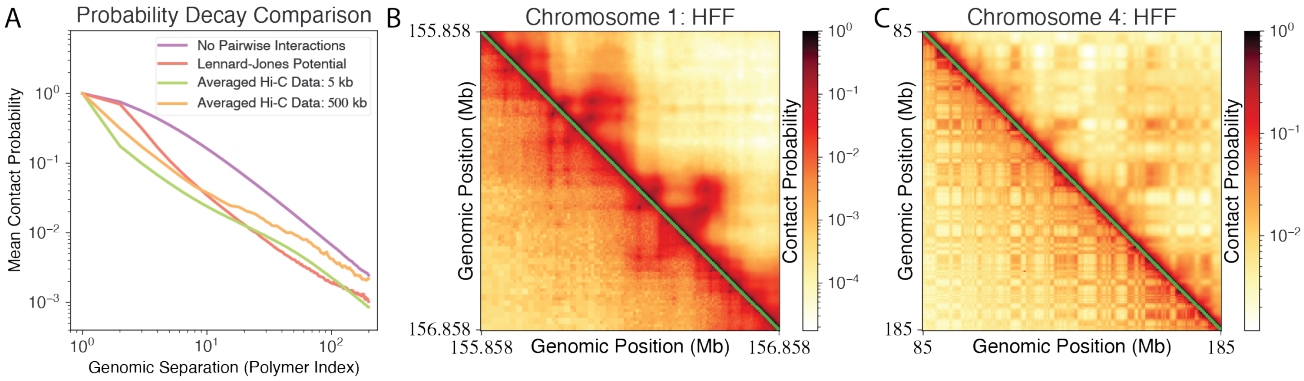

Figure 5. Polymer models parameterized with contact energies produce three-dimensional structures that recreate Hi-C contact maps. (A) This compares the separation dependence of the homopolymer model used to compute $\boldsymbol{h}$ and $\boldsymbol{J}$, the same model with Lennard-Jones potentials applied uniformly, and the separation-averaged Hi-C contact probabilities from chromosome 1 and 4 of HFF cells at 5 and 500 kb resolution, respectively. (B, C) These compare the experimental (lower triangle) and simulated (upper triangle) contact probabilities for HFF cells in the genomic regions (B) $155,858,000 - 156,858,000$ in chromosome 1 at 5 kb resolution and (C) $85,000,000 - 185,000,000$ in chromosome 4 at 500 kb resolution.

at 500 kb resolution. In addition, we classified each interaction as AA, AB, or BB according to the compartment associated with the relevant genomic loci (see Methods section). Fig. 4*B* shows the distribution of contact energies for each interaction type, which clearly indicate that intra-compartment interactions (AA and BB) are generally favorable while inter-compartment interactions (AB) are generally unfavorable. This difference penalizes the mixing of different types of chromatin and supports a phase separation mechanism in compartment formation. Compared to hESC cells, BB interactions are more favorable, AB interactions are less favorable, and AA interactions are similarly favorable in HFF cells. Therefore, interactions involving heterochromatin appear to underlie the increased compaction of heterochromatin and stricter separation between compartment types in differentiated cells.

Compartment-level contact energies in chromosome 2 (Fig. S9) and at 250 kb resolution (Fig. S10) support the same conclusion.

Parameterizing MD simulations with extracted energies

The results presented so far suggest that extracting contact energies from Hi-C data is useful to visualize interactions and analyze structural features. At the same time, contact energies are the contact-space analogue of the interaction energies needed to parameterize accurate polymer models of chromatin, denoted $\alpha_{kl} \in \boldsymbol{\alpha}$ in Eq. 1, so the contact energies extracted by our algorithm approximate $\boldsymbol{\alpha}$. Existing approaches [20–23] to determine $\boldsymbol{\alpha}$ suffer from computational complexity, so parameterizing polymer models with contact energies offers a valuable alternative when using computer simulations to predict the three-dimensional structures that contribute to Hi-C probability maps. To showcase the usefulness of contact energies in this task, we parameterized polymer models with contact energies, simulated their contact probability maps, and compared the results to the initial Hi-C data.

Unlike the homopolymer model we simulated to obtain the contact statistics, these polymer models of chromatin apply the Lennard-Jones potential defined in Eq. 11 between all monomers. As shown in Fig. 5*A*, including the excluded volume effect yields better agreement between probability decay in the *in silico* and *in vivo* contact statistics, so these models require interaction energies that account for local perturbations alone. We encoded this information in the model by using Eq. 10 with $\alpha_{kl} = 0.2\Delta\epsilon_i$ to apply specific interactions. Scaling contact energies by 0.2 helps correct for the approximations made by the contact-space representation of the polymer model; for example, Eq. 5 implicitly assumes that the Heaviside function in Eq. 2 defines the shape of interaction potentials, while Eq. 10 defines them with a hyperbolic tangent function. Fig. S11 analyzes the accuracy of simulated contact probability maps when contact energies are scaled by other factors.

Fig. 5 presents the simulated contact probabilities for two 200-monomer models representing HFF cells at 5

(Fig. 5B) and 500 kb (Fig. 5C) resolution. These span the same regions analyzed in Fig. 3C and Fig. 4A, respectively. The first model highlights chromatin loops while the second highlights compartments, two prominent features of chromatin organization. Visually, each model's statistics maintain remarkable agreement with the initial Hi-C probability map, and all major features appear in the simulated data. The Pearson correlation coefficient (PCC) between the simulated and *in vivo* Hi-C contact probabilities quantifies their agreement, and the high- and low-resolution models maintain a PCC of 0.9136 and 0.8621, respectively, corroborating the visual analysis.

These results indicate that contact energies can effectively parameterize polymer models of chromatin. Simultaneously, they confirm that the contact energies analyzed in prior subsections are physically meaningful. If more accurate predictions of chromatin structures are desired, the preexisting optimization algorithms [20–23] can use contact energies as an intelligent first guess. Given the quality of the predictions shown in Fig. 5, we anticipate that these algorithms will converge to a solution quickly when models are initialized with contact energies, thus minimizing their primary drawback.

Conclusions

We introduced an efficient algorithm to convert correlated Hi-C contact probabilities to uncorrelated contact energies, which quantify the strength of local interactions. Our approach simplified this task by utilizing a binary contact space, where an easily-invertible Ising-like model represents polymer models of chromatin. Importantly, the Ising-like Hamiltonian separately accounts for the local interactions of interest and the system-wide effects of polymer topology. We showed that interpreting contact energies is highly intuitive, and that they can facilitate the identification of functionally important interactions. Furthermore, simulating polymer models parameterized with these energies can predict an ensemble of three-dimensional configurations that reproduce Hi-C contact probabilities.

Several aspects of the model can be improved. For example, we ignored excluded volume while parameterizing the contact-space model, which likely oversimplifies the topological conditions of chromatin. Therefore, using models that better approximate chromatin systems should better account for polymer topology while extracting contact energies. In addition, the Ising-like Hamiltonian may be too simple to reproduce the higher-order interactions affecting Hi-C contact probabilities. Neural network-based generative models may better approximate the probability distribution in contact space [96–98] and could improve the quality of the extracted contact energies. Finally, the size of $\boldsymbol{J}$ is $\mathcal{O}(M^4)$, where M is the number of monomers. This presents memory challenges for larger systems, limiting the algorithm's ability to analyze large regions at high resolution.

Author Contributions

Designed research: G.S., X.D., and B.Z.; developed new equations: G.S.; performed simulations: G.S. and B.Z.; analyzed data: G.S.; interpreted results: G.S., X.D., and B.Z.; drafted manuscript: G.S. and B.Z.

Acknowledgments

This work was supported by the National Institutes of Health (Grant R35GM133580). We thank A. Sood for many helpful conversations.

Supporting Material

A supplement to this article can be found by visiting bioRxiv and uses several addition references [99–101].

References

1. Xingcheng Lin, Yifeng Qi, Andrew P. Latham, and Bin Zhang. Multiscale modeling of genome organization with maximum entropy optimization. *The Journal of Chemical Physics*, 155(1):010901, July 2021.

2. Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

3. Lu Bai and Alexandre V. Morozov. Gene regulation by nucleosome positioning. *Trends in Genetics*, 26(11):476–483, November 2010.

4. William K. M. Lai and B. Franklin Pugh. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology*, 18(9):548–562, September 2017.

5. Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (80-.).*, 295(5558):1306–1311, 2002.

6. Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).*, 326(5950):289–293, 2009.

7. Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, December 2014.

8. Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. Van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012.

9. Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

10. Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403, June 2013.

11. Bryan R. Lajoie, Job Dekker, and Noam Kaplan. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, 72:65–75, January 2015.

12. Geoffrey Fudenberg and Leonid A Mirny. Higher-order chromatin structure: bridging physics and biology. *Current Opinion in Genetics & Development*, 22(2):115–124, April 2012.

13. Homer Jacobson and Walter H. Stockmayer. Intramolecular Reaction in Polycondensations. I. The Theory of Linear Systems. *The Journal of Chemical Physics*, 18(12):1600–1606, December 1950.

14. Hue Sun Chan and Ken A. Dill. Intrachain loops in polymers: Effects of excluded volume. *The Journal of Chemical Physics*, 90(1):492–509, January 1989.

15. Hue Sun Chan and Ken A. Dill. The effects of internal constraints on the configurations of chain molecules. *The Journal of Chemical Physics*, 92(5):3118–3135, March 1990.

16. Luca Giorgetti, Rafael Galupa, Elphège P. Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell*, 157(4):950–963, May 2014.

17. Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13(1):436, December 2012.

18. Jian Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature Genetics*, 54(5):725–734, May 2022.

19. Nicolas Servant, Nelle Varoquaux, Edith Heard, Emmanuel Barillot, and Jean-Philippe Vert. Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics*, 19(1):313, December 2018.

20. Bin Zhang and Peter G. Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067, May 2015.

21. Bin Zhang and Peter G. Wolynes. Shape Transitions and Chiral Symmetry Breaking in the Energy Landscape of the Mitotic Chromosome. *Physical Review Letters*, 116(24):248101, June 2016.

22. Guang Shi and D. Thirumalai. From Hi-C Contact Map to Three-Dimensional Organization of Interphase Human Chromosomes. *Phys. Rev. X*, 11(1):11051, 2021.

23. Xiakun Chu and Jin Wang. Quantifying Chromosome Structural Reorganizations during Differentiation, Reprogramming, and Transdifferentiation. *Phys. Rev. Lett.*, 129(6):068102, aug 2022.

24. Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98, January 2012.

25. Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Research*, 22(7):1295–1305, July 2012.

26. Dario Meluzzi and Gaurav Arya. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Research*, 41(1):63–75, January 2013.

27. Gamze Gürsoy, Yun Xu, Amy L. Kenter, and Jie Liang. Spatial confinement is a major determinant of the folding landscape of human chromosomes. *Nucleic Acids Research*, 42(13):8223–8230, July 2014.

28. Michele Di Pierro, Bin Zhang, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113(43):12168–12173, October 2016.

29. Michele Di Pierro, Ryan R. Cheng, Erez Lieberman Aiden, Peter G. Wolynes, and José N. Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, 114(46):12126–12131, November 2017.

30. Gamze Gürsoy, Yun Xu, and Jie Liang. Spatial organization of the budding yeast genome in the cell nucleus and identification of specific chromatin interactions from multi-chromosome constrained chromatin model. *PLOS Computational Biology*, 13(7):e1005658, July 2017.

31. Jonas Paulsen, Monika Sekelja, Anja R. Oldenburg, Alice Barateau, Nolwenn Briand, Erwan Delbarre, Akshay Shah, Anita L. Sørensen, Corinne Vigouroux, Brigitte Buendia, and Philippe Collas. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biology*, 18(1):21, December 2017.

32. Yifeng Qi and Bin Zhang. Predicting three-dimensional genome organization with chromatin states. *PLOS Computational Biology*, 15(6):e1007024, June 2019.

33. Alan Perez-Rathke, Qiu Sun, Boshen Wang, Valentina Boeva, Zhifeng Shao, and Jie Liang. CHROMATIX: computing the functional landscape of many-body chromatin interactions in transcriptionally active loci from deconvolved single cells. *Genome Biology*, 21(1):13, December 2020.

34. Qiu Sun, Alan Perez-Rathke, Daniel M. Czajkowsky, Zhifeng Shao, and Jie Liang. High-resolution single-cell 3D-models of chromatin ensembles during Drosophila embryogenesis. *Nature Communications*, 12(1):205, January 2021.

35. Nelle Varoquaux, William S Noble, and Jean-Philippe Vert. Inference of 3D genome architecture by modeling overdispersion of Hi-C data. *Bioinformatics*, 39(1):btac838, January 2023.

36. E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957.

37. Nils Krietenstein, Sameer Abraham, Sergey V. Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S. Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A. Mirny, Job Dekker, and Oliver J. Rando. Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*, 78(3):554–565.e7, May 2020.

38. Jed W Pitera and John D Chodera. On the Use of Experimental Observations to Bias Simulated Ensembles. *J Chem Theory Comput*, 8(10):3445–3451, 2012.

39. Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys*, 138(8):84107, 2013.

40. Wen Jun Xie and Bin Zhang. Learning the Formation Mechanism of Domain-Level Chromatin States with Epigenomics Data. *Biophysical Journal*, 116(10):2047–2056, May 2019.

41. Andrew P. Latham and Bin Zhang. Consistent Force Field Captures Homologue-Resolved HP1 Phase Separation. *J. Chem. Theory Comput.*, 17(5):3134–3144, may 2021.

42. Andrew P. A.P. Latham and Bin Zhang. Improving Coarse-Grained Protein Force Fields with Small-Angle X-ray Scattering Data. *J. Phys. Chem. B*, 123(5):1026–1034, feb 2019.

43. B A Shoemaker, J Wang, and P G Wolynes. Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. *J Mol Biol*, 287(3):675–694, 1999.

44. Steven S. Plotkin and José N. Onuchic. Understanding protein folding with energy landscape theory Part I: Basic concepts. *Quarterly Reviews of Biophysics*, 35(2):111–167, May 2002.

45. Amogh Sood, Greg Schuette, and Bin Zhang. Coupling chromatin folding with histone modifications reveals dynamical asymmetry in the epigenetic landscape. preprint, Biophysics, November 2022.

46. Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, 108(49):E1293–E1301, 2011.

47. Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.

48. D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'Solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, March 1977.

49. H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261, July 2017.

50. Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, January 1974.

51. Erik Aurell and Magnus Ekeberg. Inverse Ising Inference Using All the Data. *Physical Review Letters*, 108(9):090201, March 2012.

52. Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):012707, January 2013.

53. Andrea Pagnani and Giovanni Catania. pagnani/PlmIsing v0.3.0. `https://github.com/pagnani/PlmIsing`, 2019. [Online; accessed 7-Jul-2022].

54. Steve Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1):1–19, March 1995.

55. Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.

56. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

57. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

58. Nezar Abdennur and Leonid A Mirny. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1):311–316, January 2020.

59. Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A. Oksuz, and Sergey V. Venev. Cooltools: enabling high-resolution Hi-C analysis in Python. preprint, Bioinformatics, November 2022.

60. Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, October 2012.

61. Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, May 2017.

62. Kate R. Rosenbloom, Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Glenn Hickey, Angie S. Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T. Lee, Chin H. Li, Karen H. Miga, Ngan Nguyen, Benedict Paten, Brian J. Raney, Arian F. A. Smit, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, January 2015.

63. Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, and Sergey V. Venev. Bioframe: Operations on Genomic Intervals in Pandas Dataframes. preprint, Bioinformatics, February 2022.

64. Stephan Kadauke and Gerd A. Blobel. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1):17–25, January 2009.

65. Antonio Mora, Geir Kjetil Sandve, Odd Stokke Gabrielsen, and Ragnhild Eskeland. In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, page bbv097, November 2015.

66. Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678, November 2016.

67. Zheng Xu, Guosheng Zhang, Cong Wu, Yun Li, and Ming Hu. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics*, 32(17):2692–2695, September 2016.

68. Taylor M. Lagler, Armen Abnousi, Ming Hu, Yuchen Yang, and Yun Li. HiC-ACT: improved detection of chromatin interactions from Hi-C data via aggregated Cauchy test. *The American Journal of Human Genetics*, 108(2):257–268, February 2021.

69. François Serra, Davide Baù, Mike Goodstadt, David Castillo, Guillaume J. Filion, and Marc A. Marti-Renom. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Computational Biology*, 13(7):e1005665, July 2017.

70. Lei Liu, Bokai Zhang, and Changbong Hyeon. Extracting multi-way chromatin contacts from Hi-C data. *PLOS Computational Biology*, 17(12):e1009669, December 2021.

71. John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics*, 35(17):2916–2923, September 2019.

72. Erhu Liu, Hongqiang Lyu, Qinke Peng, Yuan Liu, Tian Wang, and Jiuqiang Han. TADfit is a multivariate linear regression model for profiling hierarchical chromatin domains on replicate Hi-C data. *Communications Biology*, 5(1):608, June 2022.

73. Adam G. Larson, Daniel Elnatan, Madeline M. Keenen, Michael J. Trnka, Jonathan B. Johnston, Alma L. Burlingame, David A. Agard, Sy Redding, and Geeta J. Narlikar. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, 547(7662):236–240, jul 2017.

74. Amy R. Strom, Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen. Phase separation drives heterochromatin domain formation. *Nature*, 547(7662):241–245, jul 2017.

75. Roubina Tatavosian, Samantha Kent, Kyle Brown, Tingting Yao, Huy Nguyen Duc, Thao Ngoc Huynh, Chao Yu Zhen, Brian Ma, Haobin Wang, and Xiaojun Ren. Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. *J. Biol. Chem.*, 294(5):1451–1463, 2019.

76. Sumitabha Brahmachari, Vinícius G Contessoto, Michele Di Pierro, and José N Onuchic. Shaping the genome via lengthwise compaction, phase separation, and lamina adhesion. *Nucleic Acids Res.*, 50(8):4258–4271, may 2022.

77. Yifeng Qi and Bin Zhang. Chromatin network retards nucleoli coalescence. *Nat. Commun.*, 12(1):6824, dec 2021.

78. Kartik Kamat, Yifeng Qi, Yuchuan Wang, Jian Ma, and Bin Zhang. Genome Compartmentalization with Nuclear Landmarks: Random yet Precise. *bioRxiv*, 2021.

79. Martin Falk, Yana Feodorova, Natalia Naumova, Maxim Imakaev, Bryan R. Lajoie, Heinrich Leonhardt, Boris Joffe, Job Dekker, Geoffrey Fudenberg, Irina Solovei, and Leonid A. Mirny. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, 570(7761):395–399, jun 2019.

80. Aaron J. Plys, Christopher P. Davis, Jongmin Kim, Gizem Rizki, Madeline M. Keenen, Sharon K. Marr, and Robert E. Kingston. Phase separation of polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. *Genes Dev.*, 33(13-14):799–813, jun 2019.

81. Liangqi Xie, Peng Dong, Yifeng Qi, Tsung-Han S. Hsieh, Brian P English, SeolKyoung Jung, Xingqi Chen, Margherita De Marzio, Rafael Casellas, Howard Y Chang, Bin Zhang, Robert Tjian, and Zhe Liu. BRD2 compartmentalizes the accessible genome. *Nat. Genet.*, 54(4):481–491, apr 2022.

82. Rachel Leicher, Adewola Osunsade, Gabriella N. L. Chua, Sarah C. Faulkner, Andrew P. Latham, John W. Watters, Tuan Nguyen, Emily C. Beckwitt, Sophia Christodoulou-Rubalcava, Paul G. Young, Bin Zhang, Yael David, and Shixin Liu. Single-stranded nucleic acid binding and coacervation by linker histone H1. *Nat. Struct. Mol. Biol.*, 29(5):463–471, may 2022.

83. Xingcheng Lin, Rachel Leicher, Shixin Liu, and Bin Zhang. Cooperative DNA looping by PRC2 complexes. *Nucleic Acids Res.*, 49(11):6238–6248, jun 2021.

84. Sirui Liu, Ling Zhang, Hui Quan, Hao Tian, Luming Meng, Lijiang Yang, Huajie Feng, and Yi Qin Gao. From 1D sequence to 3D chromatin dynamics and cellular functions: A phase separation perspective. *Nucleic Acids Res.*, 46(18):9367–9383, 2018.

85. Rabia Laghmach, Michele Di Pierro, and Davit A Potoyan. Mesoscale Liquid Model of Chromatin Recapitulates Nuclear Order of Eukaryotes. *Biophys. J.*, 118(9):2130–2140, 2020.

86. Gaurav Bajpai, Daria Amiad Pavlov, Dana Lorber, Talila Volk, and Samuel Safran. Mesoscale phase separation of chromatin in the nucleus. *Elife*, 10, may 2021.

87. Ruggero Cortini, Maria Barbi, Bertrand R. Caré, Christophe Lavelle, Annick Lesne, Julien Mozziconacci, and Jean-Marc Victor. The physics of epigenetics. *Reviews of Modern Physics*, 88(2):025002, April 2016.

88. Mariliis Tark-Dame, Roel van Driel, and Dieter W. Heermann. Chromatin folding – from biology to polymer models and back. *Journal of Cell Science*, 124(6):839–845, March 2011.

89. J. Langowski. Polymer chain models of DNA and chromatin. *The European Physical Journal E*, 19(3):241–249, March 2006.

90. Gavin Bascom and Tamar Schlick. Linking Chromatin Fibers to Gene Folding by Hierarchical Looping. *Biophysical Journal*, 112(3):434–445, February 2017.

91. Renjie Wang, Julien Mozziconacci, Aurélien Bancaud, and Olivier Gadal. Principles of chromatin organization in yeast: relevance of polymer models to describe nuclear organization and dynamics. *Current Opinion in Cell Biology*, 34:54–60, June 2015.

92. Fabian Erdel and Karsten Rippe. Formation of Chromatin Subcompartments by Phase Separation. *Biophysical Journal*, 114(10):2262–2270, May 2018.

93. Shin Fujishiro and Masaki Sasai. Generation of dynamic three-dimensional genome structure through phase separation of chromatin. *bioRxiv*, 2021.

94. Daniel Jost, Pascal Carrivain, Giacomo Cavalli, and Cédric Vaillant. Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.*, 42(15):9553–9561, 2014.

95. Quinn MacPherson, Bruno Beltran, and Andrew J Spakowitz. Bottom–up modeling of chromatin segregation due to epigenetic modifications. *Proc. Natl. Acad. Sci.*, 115(50):12739 LP – 12744, dec 2018.

96. Hoda Akl, Brooke Emison, Xiaochuan Zhao, Arup Mondal, Alberto Perez, and Purushottam D. Dixit. GENERALIST: An efficient generative model for protein sequence families. preprint, Biophysics, December 2022.

97. Xinqiang Ding and Bin Zhang. Computing Absolute Free Energy with Deep Generative Models. *J. Phys. Chem. B*, 124(45):10166–10172, nov 2020.

98. Wen Jun Xie, Yifeng Qi, and Bin Zhang. Characterizing chromatin folding coordinate and landscape with deep learning. *PLOS Comput. Biol.*, 16(9):e1008262, sep 2020.

99. David Sherrington and Scott Kirkpatrick. Solvable Model of a Spin-Glass. *Physical Review Letters*, 35(26):1792–1796, December 1975.

100. Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

101. E. Juárez Ruiz, R. Cortes Maldonado, and F. Pérez Rodríguez. Relationship between the Inverses of a Matrix and a Submatrix. *Computación y Sistemas*, 20(2), July 2016.