Hi-BDiSCO: Folding 3D Mesoscale Genome Structures from Hi-C Data using Brownian Dynamics

Supporting Information

Zilong Li[1,4] and Tamar Schlick[1,2,3,4]

[1]Department of Chemistry, 100 Washington Square East, Silver Building, New York University, New York, NY 10003 U.S.A.

[2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012 U.S.A.

[3]New York University-East China Normal University Center for Computational Chemistry, New York University Shanghai, Shanghai 200122 China.

[4]Simons Center for Computational Physical Chemistry, 24 Waverly Place, Silver Building, New York University, New York, NY 10003 U.S.A.

October 6, 2023

Nucleosome Interaction We have studied the nucleosome interactions in the section "Reconstruction of known structures demonstrates biophysically sound fibers", Figure 4 of the main text. The nucleosome interaction is calculated as follows:

Definition of contact: A contact is defined when any element (i.e., core, linker DNA, histone tails) of core i and any element of core j are within a cutoff distance, here set as cutoff = 2 nm.

Calculate contact probability matrices: We normalize the contact probabilities across all frames for a single trajectory and obtain the contact probability matrix. We sum all the contact probability matrices across all trajectories to obtain the final contact probability matrix.

1D projection of the contact matrix: To create a one-dimension projection $I(k)$ of the contact matrix $I'(i,j)$, we sum across each row k of the matrix, and $I(k)$ is given by:

$$I(k) = \frac{\sum_{i=1}^{N_C} I'(i, i \pm k)}{\sum_{j=1}^{N_C} I(j)} \tag{1}$$

where N_C is the total number of cores. This provides information on the fraction of configurations between nucleosomes separated by k nucleosomes (i.e., $i \pm k$ nucleosome neighbors are interacting).

Volume and Radius of Gyration We calculate the volume and radius of gyration of chromatin fibers in the following sections of the main text: "Increasing N_{rep} improves the accuracy of the reconstruction" (Figure 2), "Subsequent MC resolves clashes while maintaining the contacts" (Figure 3), and "Reconstruction of Fbn2 Gene Reveals Cohesin and RNA Pol II Roles on Chromatin Architecture" (Figure 6). The calculation of these terms is as follows:

Volume: The volume of each chromatin fiber is calculated using Matlab's AlphaShape protocol. That is, all the elements of the fiber are enclosed in a convex bounding surface. Then the volume of this convex surface is calculated and treated as the volume of the chromatin fiber.

Radius of Gyration: The radius of gyration is calculated by:

$$R_g^2 = \frac{1}{N} \sum_{j=1}^{N} (r_j - r_{mean})^2 \tag{2}$$

where r is the center of each nucleosome core, and r_{mean} is the average of all core position [1]

Life-like Fibers We have used the term "life-like fibers" in the section "Reconstruction of known structures demonstrates biophysically sound fibers", Figure 4 of the main text. A "life-like chromatin fiber" was defined by Bascom et al. [2] to resemble nucleosome distributions in living cells. We model life-like distributions of linker lengths [2] based on the analyses of Brogaard et al. [3], which results in the following linker length distributions: 70% 26 bp, 20% 35 bp, and 10% 44 bp. To set up such fibers, we first calculate the number of each linker length for a 50-nucleosome fiber, which results in 35, 10, and 5 linkers for 26 bp, 35 bp, and 44 bp, respectively. Then we randomly assign these linkers to define a life-like fiber (i.e., place one nucleosome followed by a 26 bp linker, then a second nucleosome followed by a 26 bp fiber, then a third nucleosome followed by a 44 bp fiber... until all linkers are used); we generated one distribution of such life-like fiber as the target fiber and reconstructed N_{sim} replicas of 3D structures with the same nucleosome placements.

Distance/Force Parameter Choices To determine reasonable values of adjustable model parameters for the distance constraints, we run trajectories with $h_m = \frac{h}{50}, \frac{h}{20}, \frac{h}{5}$, and h, where h is the stretching constant for the connecting DNA beads and nucleosome cores in our mesoscale model. It is given by $h = 100 k_B T / l_0^2$, where $l_0 = 3$ nm, k_B is the Boltzmann constant, and T is the temperature. When the stretching force is strong ($h_m = h$), the contacted beads will be pulled

together strongly and the overall result is a very condensed structure, Figure S1; when the stretching force is weak $\left(h_m = \frac{h}{50}\right)$, the chromatin fiber folds more smoothly and decondensed, see Figure S1.

Figure S2 shows the sedimentation, packing ratio, volume, and radius of gyration with different force constants. As the force increases, sedimentation and packing ratios are also higher, and the volume and radius of gyration are lower.

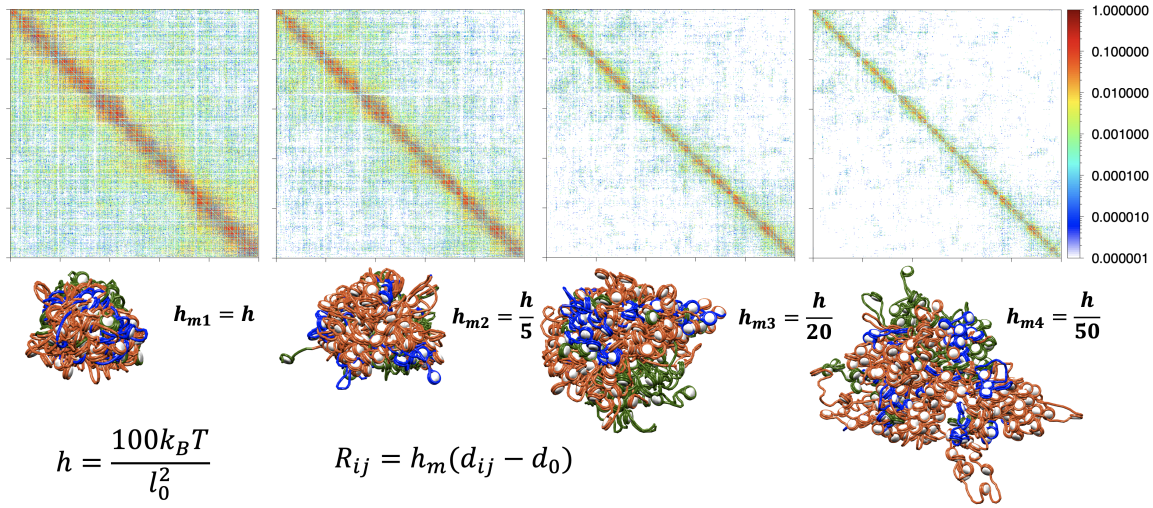

Figure S1: Assess the force parameters through the reconstruction of the HOXA gene using various sets of force parameters. Top: Hi-BDiSCO reconstructed contact maps (HOXA) by applying different force parameters (left to right: $h_{m_1} = h$, $h_{m_2} = \frac{h}{5}$, $h_{m_3} = \frac{h}{20}$ and $h_{m_4} = \frac{h}{50}$, respectively, strong to soft). Strong force produces more contacts. Bottom: 3D structures examples with different forces. The stronger the force is, the more condensed the structure is. The formula to calculate the stretching parameter h, and the restraint R_{ij} is also shown.

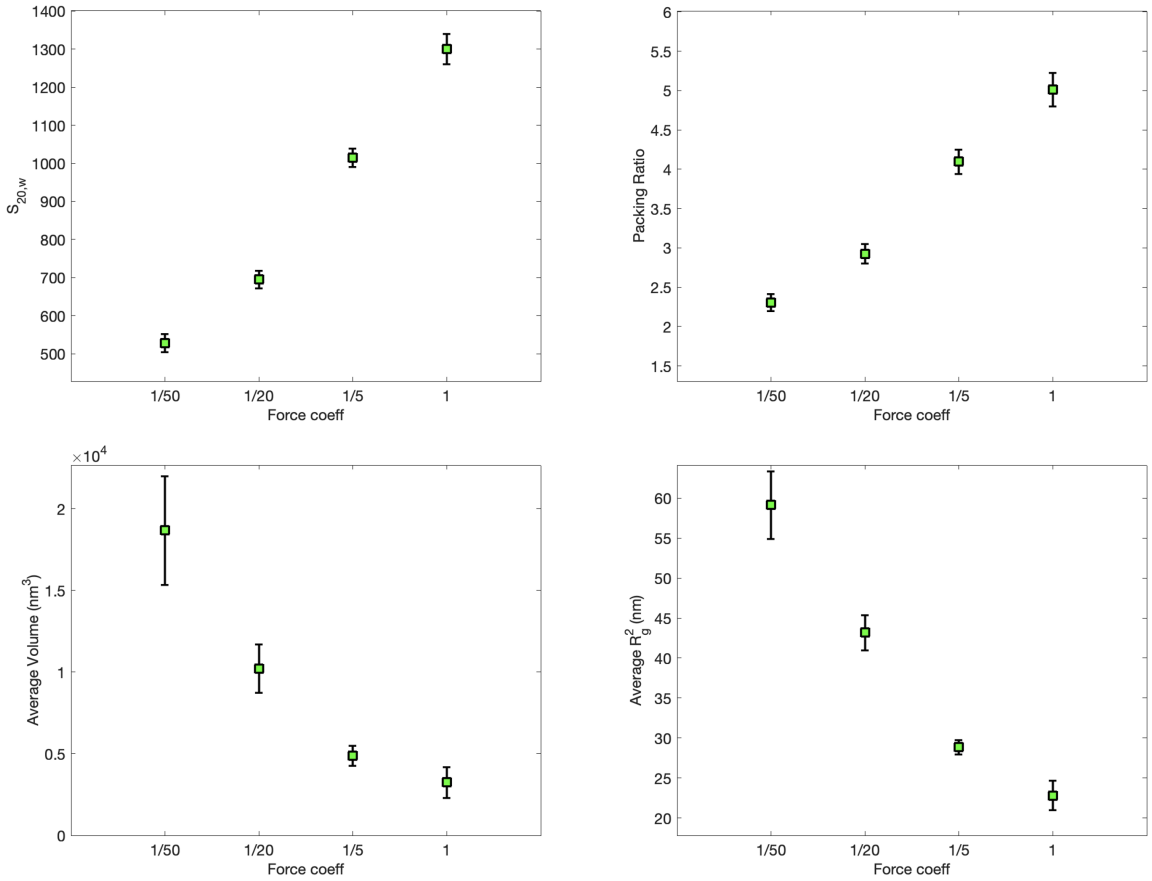

Figure S2: The sedimentation, packing ratio, volume and radius of gyration of 100 ns simulation of HOXA shown in Figure S1 with different stretching force constant for the restraints.

Hi-BDiSCO predicted structures yield agreement with MC predicted structures We have demonstrated that our Hi-BDiSCO can predict the folding of gene-level fibers with Micro-C restraints and MC modifications. A separate aspect is how these predictions compare with our previous MC predictions de novo, without any Micro-C data. Bascom et al. have studied the HOXC gene with our mesoscale MC simulation [4] that simulated 8 million steps with full details of wild-type and acetylated histone tails, linker histones in 50 replicas. Here, we used Hi-BDiSCO instead and simulated the same system with the same nucleosome positions, acetylated tails, and LH occupation for 100 replicas. We compare the results from these two methods by calculating the RMSD between each replica. As shown in Figure S3A, the SCC coefficient, 0.665, shows tentative correlation, possibly not as high because the sample size for analysis is not sufficiently large. Both methods represent a small portion of a population (1M) of structures. Even so, the SCC is positive, indicating the two ensembles of structures are correlated. There are visually more data points on the MC map than on the Hi-BDiSCO map because this contact map is an accumulation of contact counts among the MC (80 million) steps, but Hi-BDiSCO uses 20 million MC steps subsequent to BD refinement. As shown in Figure S3B, we calculated three sets of RMSD, which are between each replica of i) Hi-BDiSCO, ii) MC, and iii) both. By comparing i) and ii), we see that the Hi-BDiSCO

structures have overall higher RMSD than the MC structures between each replica, which indicates that Hi-BDiSCO yields a greater variety of structures. From iii), we also found two outlier structures with high RMSDs compared with all other structures: replica #5 of Hi-BDiSCO results and replica #36 of MC results. We found it is also true for these two structures in i) and ii). Besides these two structures, most of the structures has small RMSD (under 100 nm) for the two methods, indicating that the two methods have similar predictive results.

In addition, Bascom et al. [4] found a notable contact hub between acetylation-rich and LH-rich region (highlight square part in Figure S3A). Here we calculate the ratio of the sum of the contact frequencies of the highlight region against the whole map $\left(ratio = \frac{\sum f_{highlight}}{\sum f_{total}}\right)$. The MC predicted map has a $ratio_{MC} = 0.445$, and the Hi-BDiSCO predicted map has a $ratio_{Hi-BDiSCO} = 0.503$.

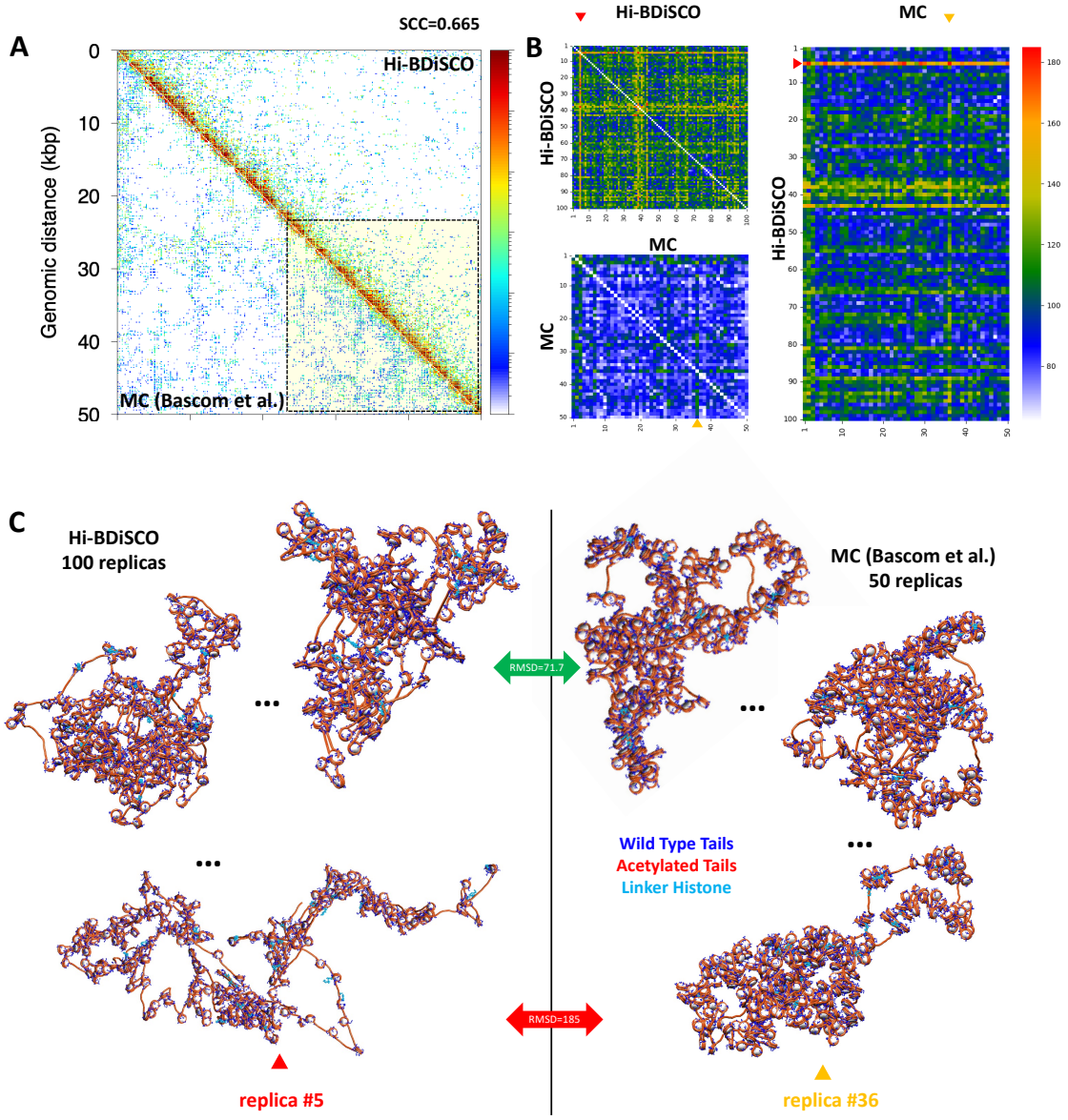

Figure S3: Comparison between MC predicted and Hi-BDiSCO reconstructed structures. A: The contact maps for MC predictions of Bascom et al. [4] (lower triangle) are compared to Hi-BDiSCO results (upper triangle) along with the Spearman correlation coefficient (SCC). The highlight region is the acetylated tail-rich region that comes into contact with an LH-rich region. B: The RMSD plot between each pair of structures between 1) Hi-BDiSCO reconstructed structures (left top); 2) MC predicted structures (left bottom); and Hi-BDiSCO reconstructed structures against MC predicted structures (right). C: Representative structures obtained from Hi-BDiSCO after MC (100 replicas in total) verses MC simulation [4] (50 replicas in total). Similar structures (RMSD=71.7 nm), outlier structures (RMSD=185 nm), and randomly chosen structures for both methods are shown.

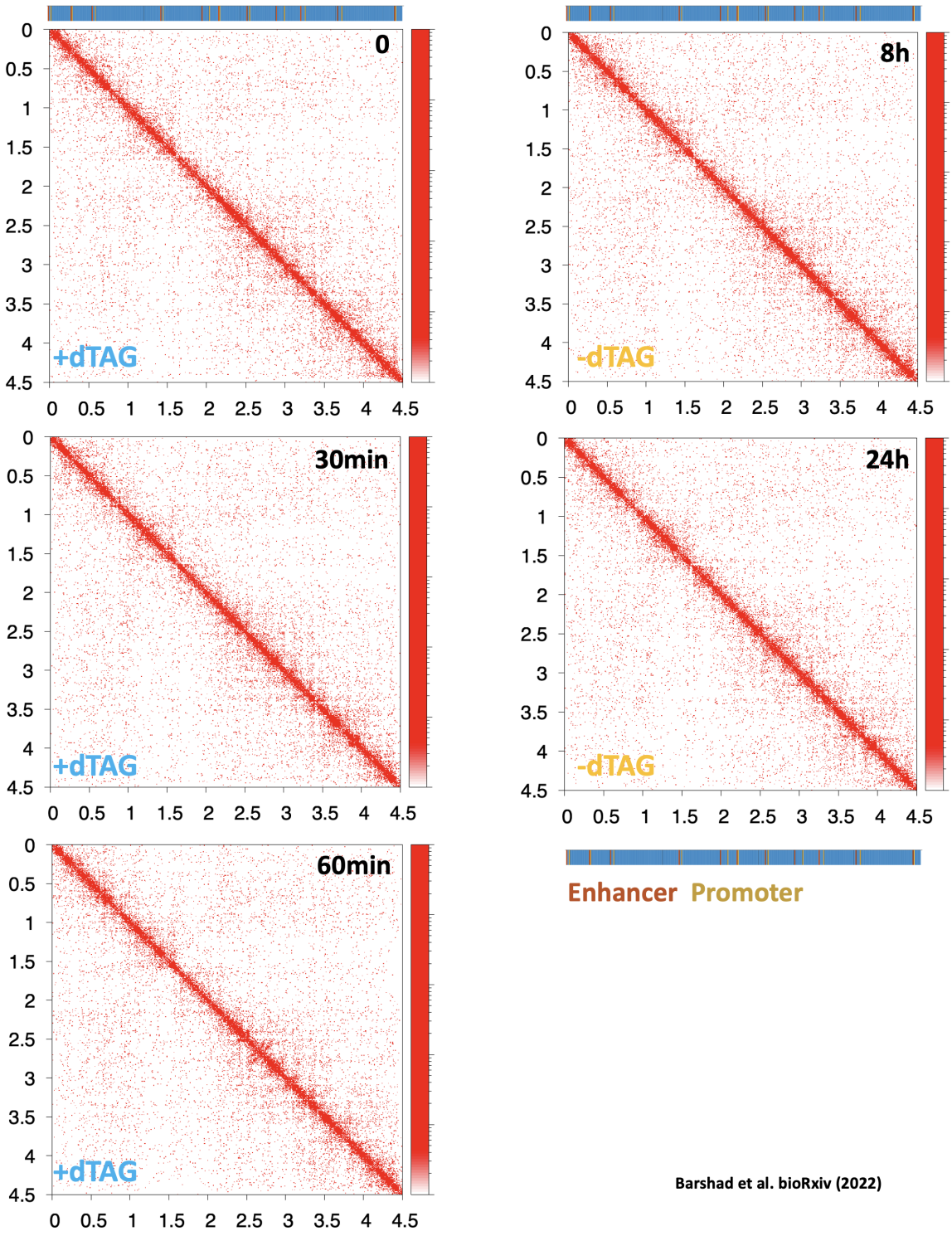

Figure S4: Micro-C experimental data at different time snapshots for HOXA gene region from Barshad et al. [5]. The experiment explores the role of Pol II pausing by introducing (dTAG) a Pol II inhibitor at time 0, and removing the inhibitor at time 60 min.

Summary for data sources of the studied gene systems In Table S2, we summarize the experimental data used to define each gene system. To clarify the source of average Micro-C data: in GSE130275 [6], cells were derived from 38 replicates of wildtype mouse embryonic stem cells, then merged WT data into two different levels of sequencing coverage with 1.3B and 2.6B reads;

Gene	Micro-C	MNase-seq	Chip-seq	description
HOXA	GSE130275 [6]	GSM2083105 [8]	GSE46134 [9]	
HOXC	GSE130275	GSM2083105	GSE46134	Figures 4,5,6 of the main text
NXN	GSE130275	GSM2083105	GSE46134	
HOXA	GSE206133 [5]	GSM2083105	GSE46134	Figure 8 of the main text
fbn2	GSE207225 [7]	GSM2083105	GSE46134	Figure 9 of the main text

Table S1: Summary for the experimental data, such as MNase and Chip-seq, used to set up each gene system, that is, to place nucleosomes, linker histones, and epigenetic marks along the gene sequence. All data are for Mouse embryonic stem cells (mESCs).

in GSE206133 [5], the authors used two biological replicates with two technical replicates, where samples were prepared for each time point of dTAG-13 incubation for mECSs with a homozygous endogenous NELFB-FKBP12F36V fusion protein (0, 30 and 60 minutes as well as for 8 hours and 24 hours after dTAG-13 washout); in GSE207225 [7] the authors applied Region Capture Micro-C to reveal contact patterns between enhancers and promoters; they mapped genomic structures at 0.4-1.9 Mb-sized regions at the Klf1, Ppm1g, Fbn2, Sox2, and Nanog loci across four conditions: wildtype and transcriptionally inhibited mouse Embryonic Stem Cells (mESCs), and cohesin-depletion and control treatments in a RAD21-AID genome-edited mESC line.

Single Cell Hi-C Hi-BDiSCO can also be implemented for single-cell Hi-C (scHi-C) data. Although there are scHi-C data available, these have low resolution. To demonstrate how Hi-BDiSCO can reconstruct scHi-C maps, we have created and tested an artificial scHi-C map based on one of the generated 3D structures of the NXN gene, as shown in Figure S5. The scHi-C maps are sparse matrices (1 for contacts and 0 for no contacts). We set all the contacts as restraints and used Hi-BDiSCO to reconstruct the 3D structure. The resulting structure (Figure S5 top right) is similar to the target structure (bottom left). The contact map of the resulting structure also has a strong correlation with the target one, with SCC=0.868. Thus, Hi-BDiSCO could be used to reconstruct scHi-C as long as there are high-resolution data available. Specific algorithms could be applied to sparse matrices, such as Cholesky factorizations with the Brownian Dynamics refinement.

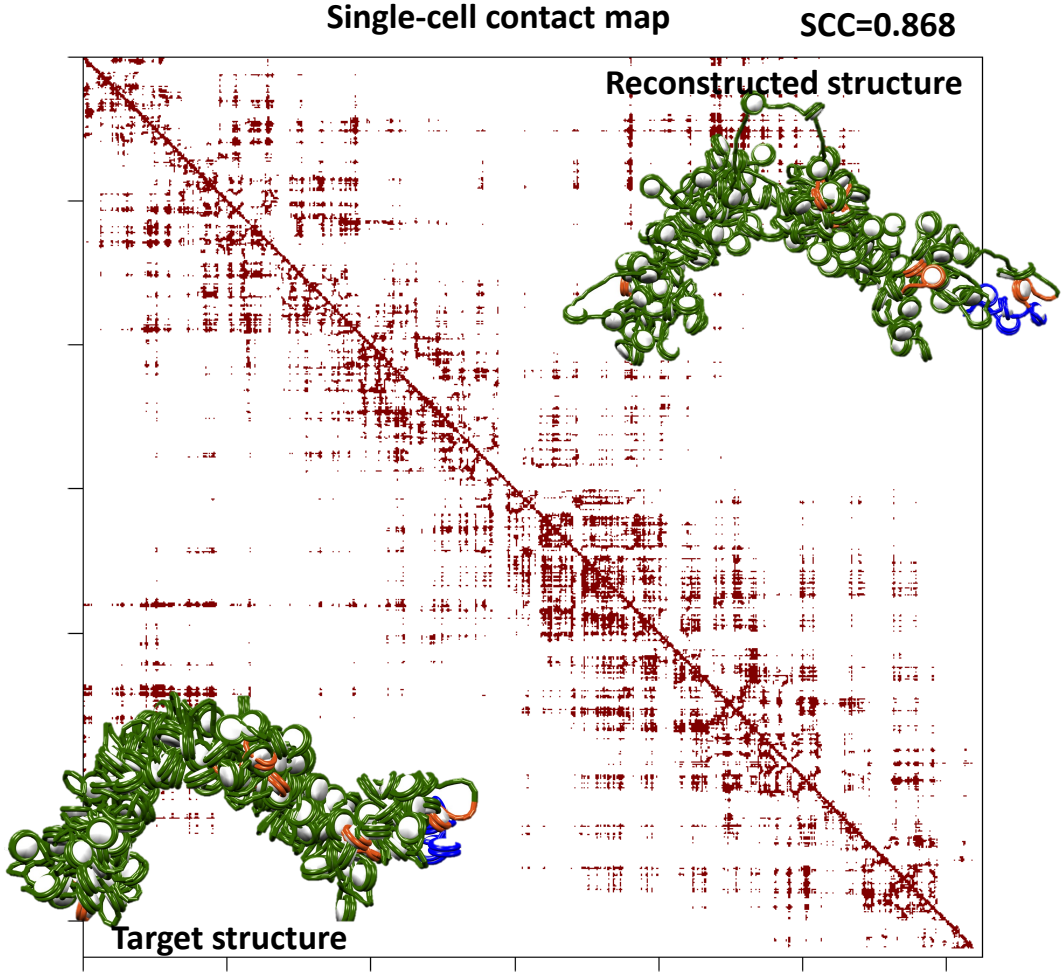

Figure S5: Contact map and single structure demonstrate a single cell Hi-C reconstruction for the NXN gene. Bottom left: the target structure and related contact map. Top right: reconstructed structure and contact map.

Unsupervised Clustering Algorithm We perform unsupervised clustering algorithm for the 100 reconstructed 3D structures of the wildtype fbn2 gene. Specifically, we extract the nucleosome positions of each structure as 3D coordinates and flatten them onto a 1D feature vector. We cluster these 1D feature vectors using a K-means clustering algorithm in the Python scikit-learn package and evaluate the number of distinct clusters obtained. We obtained 8 distinct clusters from 100 structures as shown in Figure S6. Calculated cluster global properties (volume, radius of gyration, average number of clutches, and the average number of nucleosomes per clutch) show some variation among clusters. Other properties of interest could similarly be calculated from such clusters.

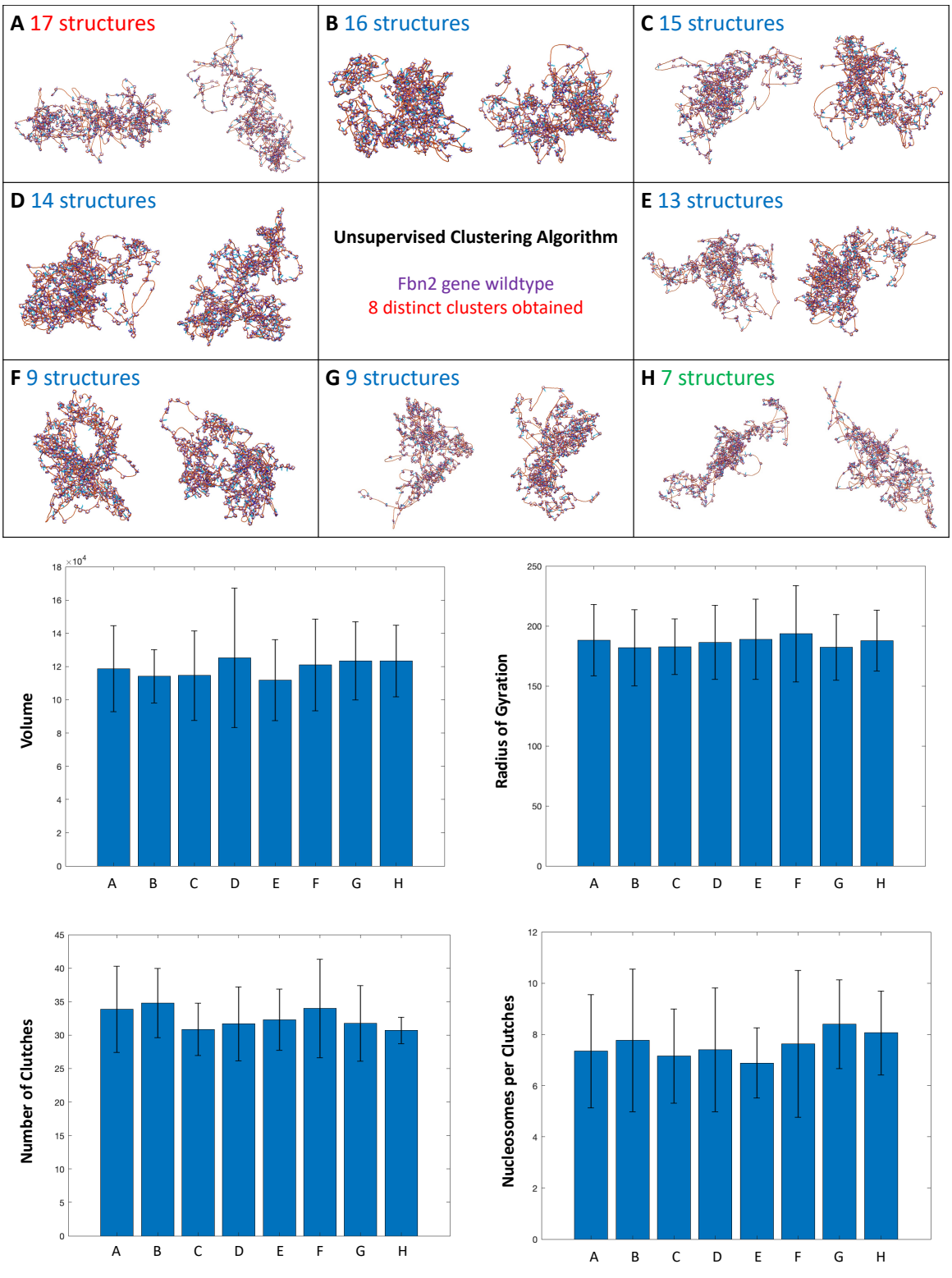

Figure S6: Implementation of an unsupervised clustering algorithm for the reconstructed 3D structures of wildtype fbn2 gene. Resulting 8 distinct clusters among a total 100 structures are shown (A to H). Top: two representative structures for each cluster, with the number of structures obtained in each cluster. Cluster A (marked in red) is the most populated, and cluster H (marked in green) the less populated. Bottom, volume, radius of gyration, average number of clutches and the average number of nucleosomes in each clutch for each cluster.

	System	N_{rep}	CVC
Simulation	HOXA	100	31.4–66.1%
	HOXC	1M	12.1–49.9%
	NXN	2M	11.7–45.9%
Experimental	interphase chromatin		12–52%

Table S2: Comparison of CVC values for our simulated structures with different N_{rep}.

Chromatin volume concentrations (CVC) We compare calculated CVC values for the three gene systems to experimental CVC values for interphase chromatin [10]. The CVC for HOXA, HOXC, and NXN are in the ranges of 31.4–66.1%, 12.1–49.9%, and 11.7–45.9%, respectively. CVC values for the HOXC reconstructed structures are closest to the experimental values (12–52% [10]), supporting the choice of N_{rep} equal to the cell population, although values span a broad range.

References

[1] O. Perišić, R. Collepardo-Guevara, and T. Schlick, "Modeling Studies of Chromatin Fiber Structure as a Function of DNA Linker Length," *J. Mol. Biol.*, vol. 403, pp. 777–802, nov 2010.

[2] G. D. Bascom, T. Kim, and T. Schlick, "Kilobase Pair Chromatin Fiber Contacts Promoted by Living-System-Like DNA Linker Length Distributions and Nucleosome Depletion," *J. Phys. Chem. B*, vol. 121, pp. 3882–3894, apr 2017.

[3] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom, "A map of nucleosome positions in yeast at base-pair resolution," *Nature*, vol. 486, pp. 496–501, jun 2012.

[4] G. D. Bascom, C. G. Myers, and T. Schlick, "Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub," *Proc. Natl. Acad. Sci.*, vol. 116, pp. 4955–4962, mar 2019.

[5] G. Barshad, J. J. Lewis, A. G. Chivu, A. Abuhashem, N. Krietenstein, E. J. Rice, Y. Ma, Z. Wang, O. J. Rando, A.-K. Hadjantonakis, and C. G. Danko, "RNA polymerase II dynamics shape enhancer–promoter interactions," *Nat. Genet.*, vol. 55, pp. 1370–1380, aug 2023.

[6] T. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, and X. Darzacq, "Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding," *Mol. Cell*, vol. 78, no. 3, pp. 539–553.e8, 2020.

[7] V. Y. Goel, M. K. Huseyin, and A. S. Hansen, "Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments," *bioRxiv*, 2022.

[8] J. Mieczkowski, A. Cook, S. K. Bowman, B. Mueller, B. H. Alver, S. Kundu, A. M. Deaton, J. A. Urban, E. Larschan, P. J. Park, R. E. Kingston, and M. Y. Tolstorukov, "MNase titration

reveals differences between nucleosome occupancy and chromatin accessibility," *Nat. Commun.*, vol. 7, p. 11485, may 2016.

[9] K. Cao, N. Lailler, Y. Zhang, A. Kumar, K. Uppal, Z. Liu, E. K. Lee, H. Wu, M. Medrzycki, C. Pan, P. Y. Ho, G. P. Cooper, X. Dong, C. Bock, E. E. Bouhassira, and Y. Fan, "High-resolution mapping of h1 linker histone variants in embryonic stem cells," *PLoS Genet.*, vol. 9, no. 4, p. e1003417, 2013.

[10] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O'Shea, "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells," *Science (80-.).*, vol. 357, jul 2017.