Biophysical Journal, Volume *115*

Supplemental Information

A Polymer Model for the Quantitative Reconstruction of Chromosome Architecture from HiC and GAM Data

Guillaume Le Treut, François Képès, and Henri Orland

CONTENTS

1. SUPPLEMENTARY FIGURES

TABLE S1 – Application of the GEM reconstruction method to several experimental data sets.

Figure	**Reference**	**Data**	**Cell type**	**Genomic range**	**Resolution**	**N**	**Normalization**	**LSE$^{1/2}$**
Fig. S1	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 7 137:138 Mbp	5 kbp	200	uniform	0.023
Fig. S2	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 7 130:140 Mbp	10 kbp	1000	uniform	0.013
Fig. S3	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 8 133.6:134.6 Mbp	5 kbp	200	uniform	0.022
Fig. S4	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 10 90.5:91.5 Mbp	5 kbp	200	uniform	0.023
Fig. S5	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 94:96 Mbp	10 kbp	200	uniform	0.022
Fig. S6	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 86:96 Mbp	10 kbp	1000	uniform	0.014
Fig. S7	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 19:107.2 Mbp	100 kbp	882	uniform	0.013
Fig. S8	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 16 85.5:87.5 Mbp	5 kbp	400	uniform	0.019
Fig. S9	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 7 137:138 Mbp	5 kbp	200	matrix balancing	0.057
Fig. S10	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 7 130:140 Mbp	10 kbp	1000	matrix balancing	0.026
Fig. S11	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 8 133.6:134.6 Mbp	5 kbp	200	matrix balancing	0.056
Fig. S12	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 10 90.5:91.5 Mbp	5 kbp	200	matrix balancing	0.059
Fig. S13	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 94:96 Mbp	10 kbp	200	matrix balancing	0.056
Fig. S14	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 86:96 Mbp	10 kbp	1000	matrix balancing	0.026
Fig. S15	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 14 19:107.2 Mbp	100 kbp	882	matrix balancing	0.026
Fig. S16	Rao *et al.* (2014)	Hi-C	GM12878 (human)	Chr. 16 85.5:87.5 Mbp	5 kbp	400	matrix balancing	0.042
Fig. S17	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 19 30:60 Mbp	30 kbp	1000	GAM	0.032
Fig. S18	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 19 3:61.2 Mbp	100 kbp	582	GAM	0.028
Fig. S19	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 19 3:60 Mbp	1 Mbp	57	GAM	0.021
Fig. S20	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 12 40:70 Mbp	30 kbp	1000	GAM	0.033
Fig. S21	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 12 30:120 Mbp	100 kbp	900	GAM	0.029
Fig. S22	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 12 3:120 Mbp	1 Mbp	117	GAM	0.025
Fig. S23	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 1 135:165 Mbp	30 kbp	1000	GAM	0.032
Fig. S24	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 1 90:190 Mbp	100 kbp	1000	GAM	0.029
Fig. S25	Beagrie *et al.* (2017)	GAM	mouse 46C line embryonic stem cells	Chr. 1 3:196 Mbp	1 Mbp	193	GAM	0.026

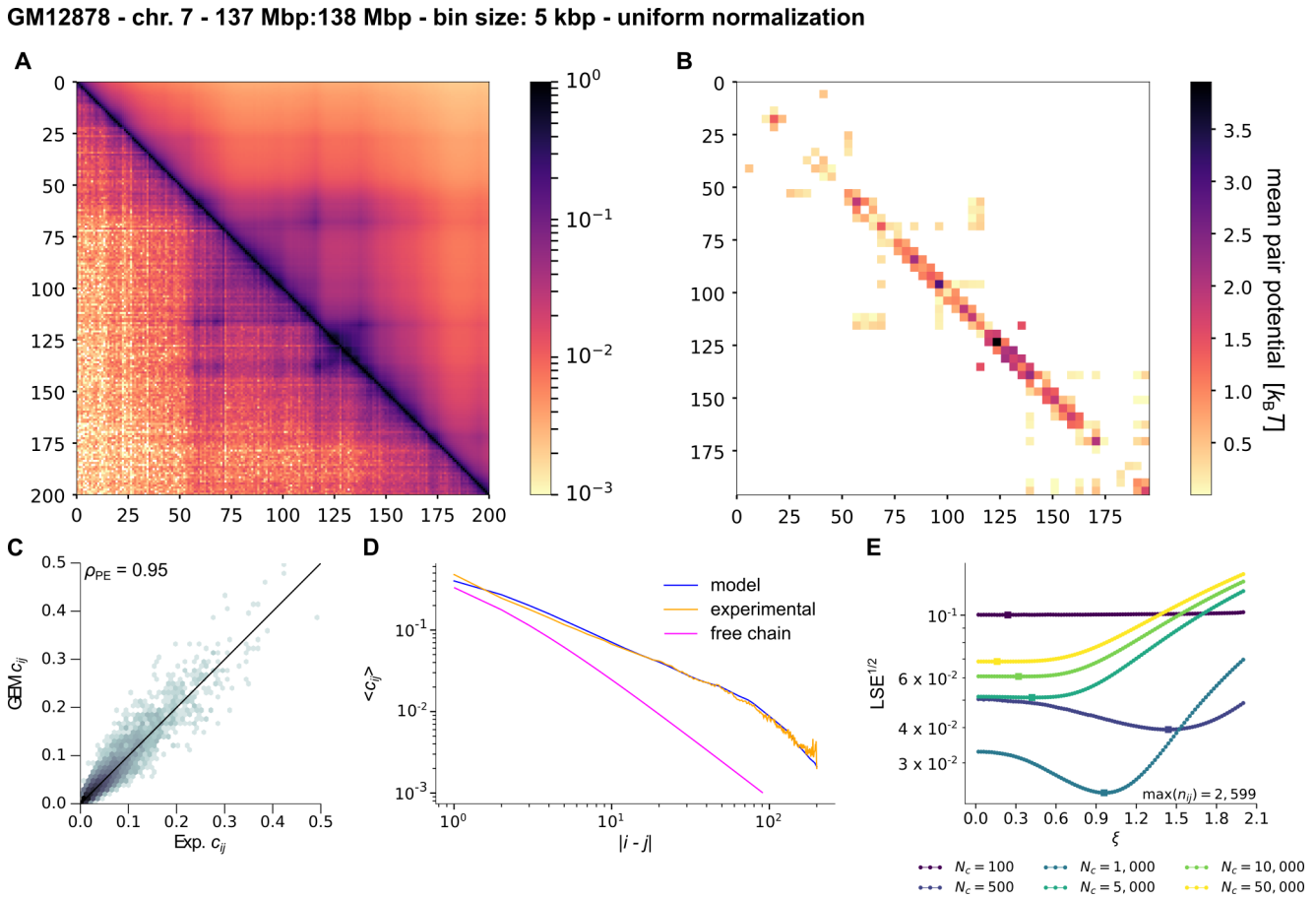

FIGURE S1 – GEM reconstruction for Hi-C data of human chromosome 7 (1) (5 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

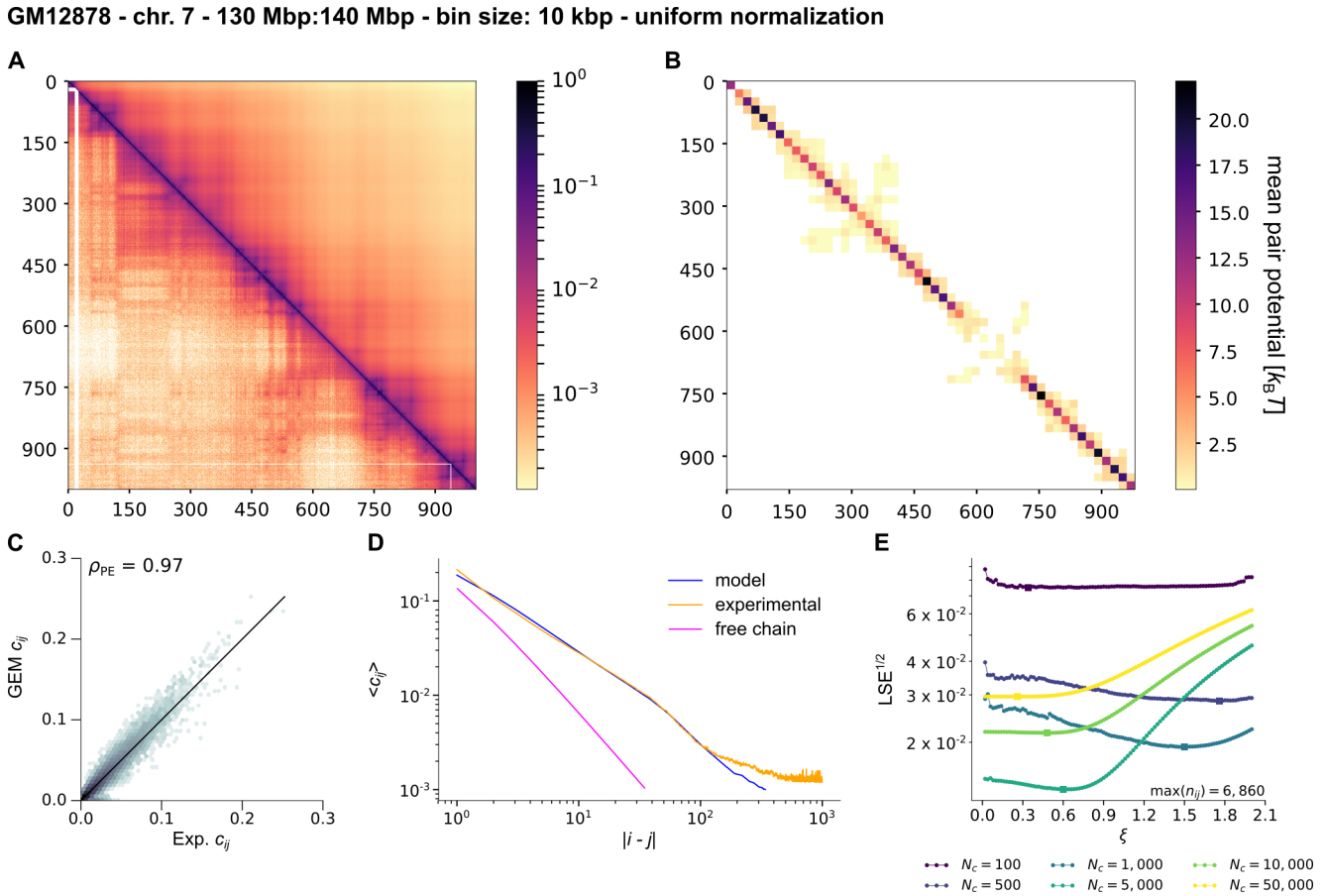

FIGURE S2 – GEM reconstruction for Hi-C data of human chromosome 7 (1) (10 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

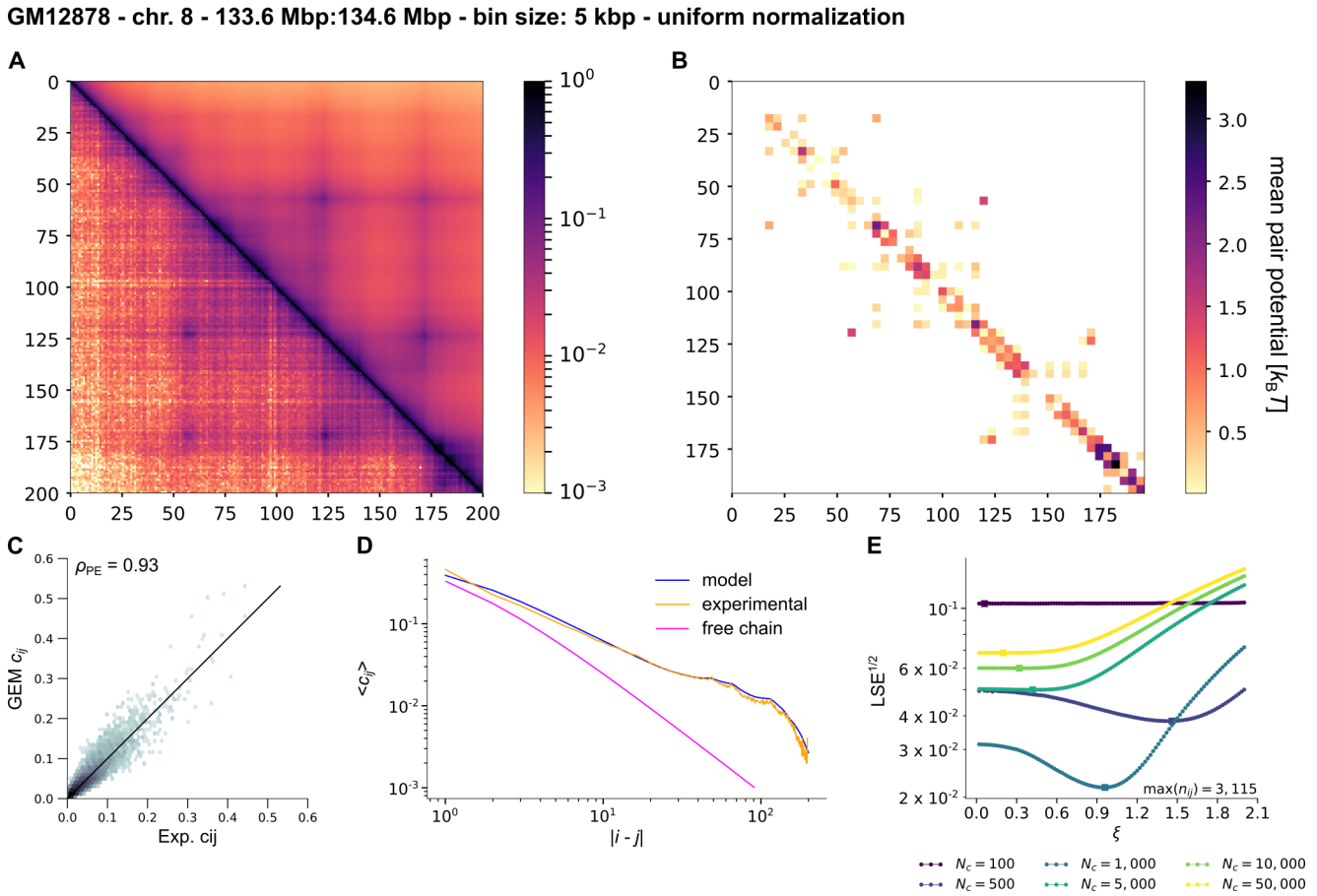

FIGURE S3 – GEM reconstruction for Hi-C data of human chromosome 8 (1) (5 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

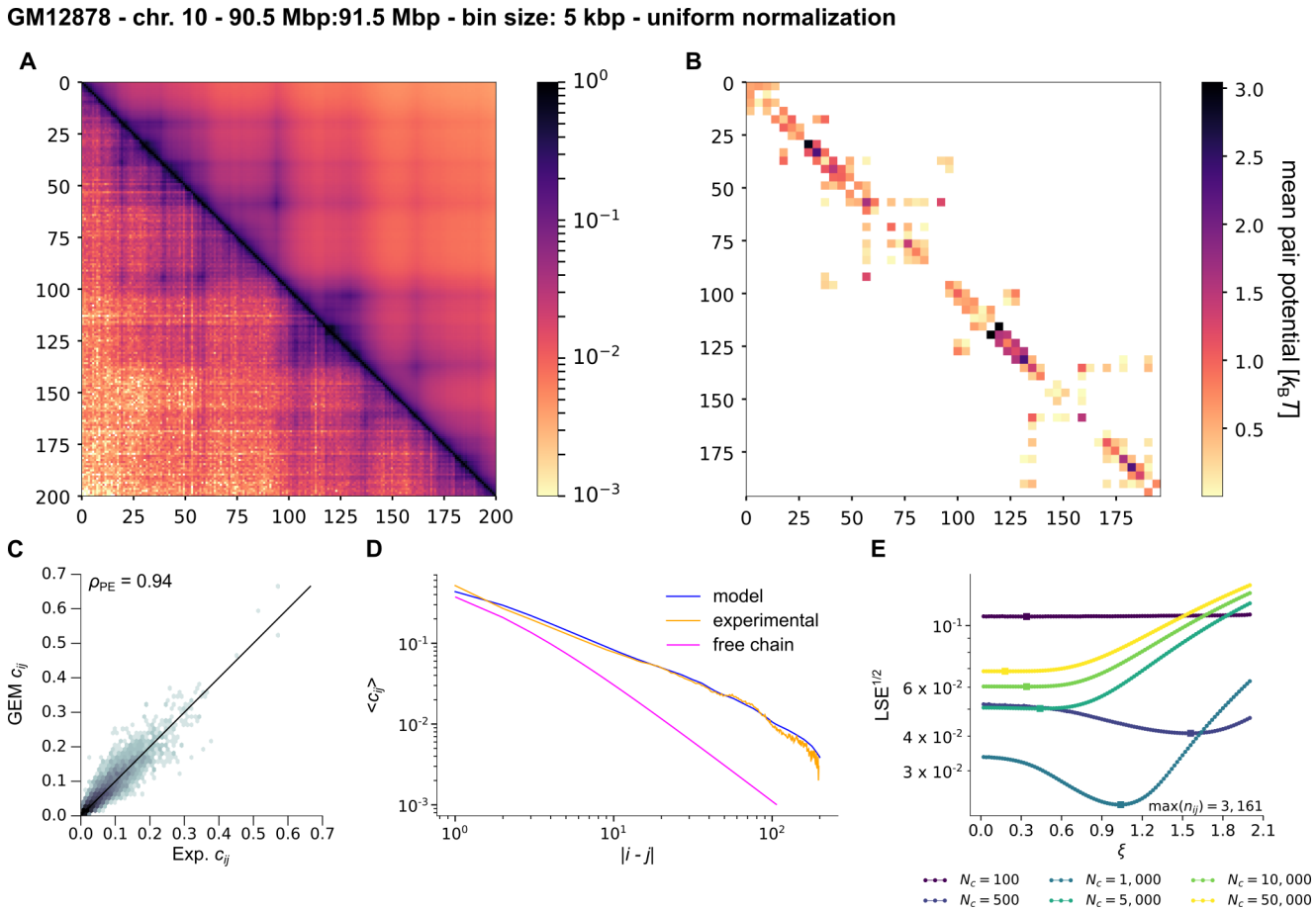

FIGURE S4 – GEM reconstruction for Hi-C data of human chromosome 10 (1) (5 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

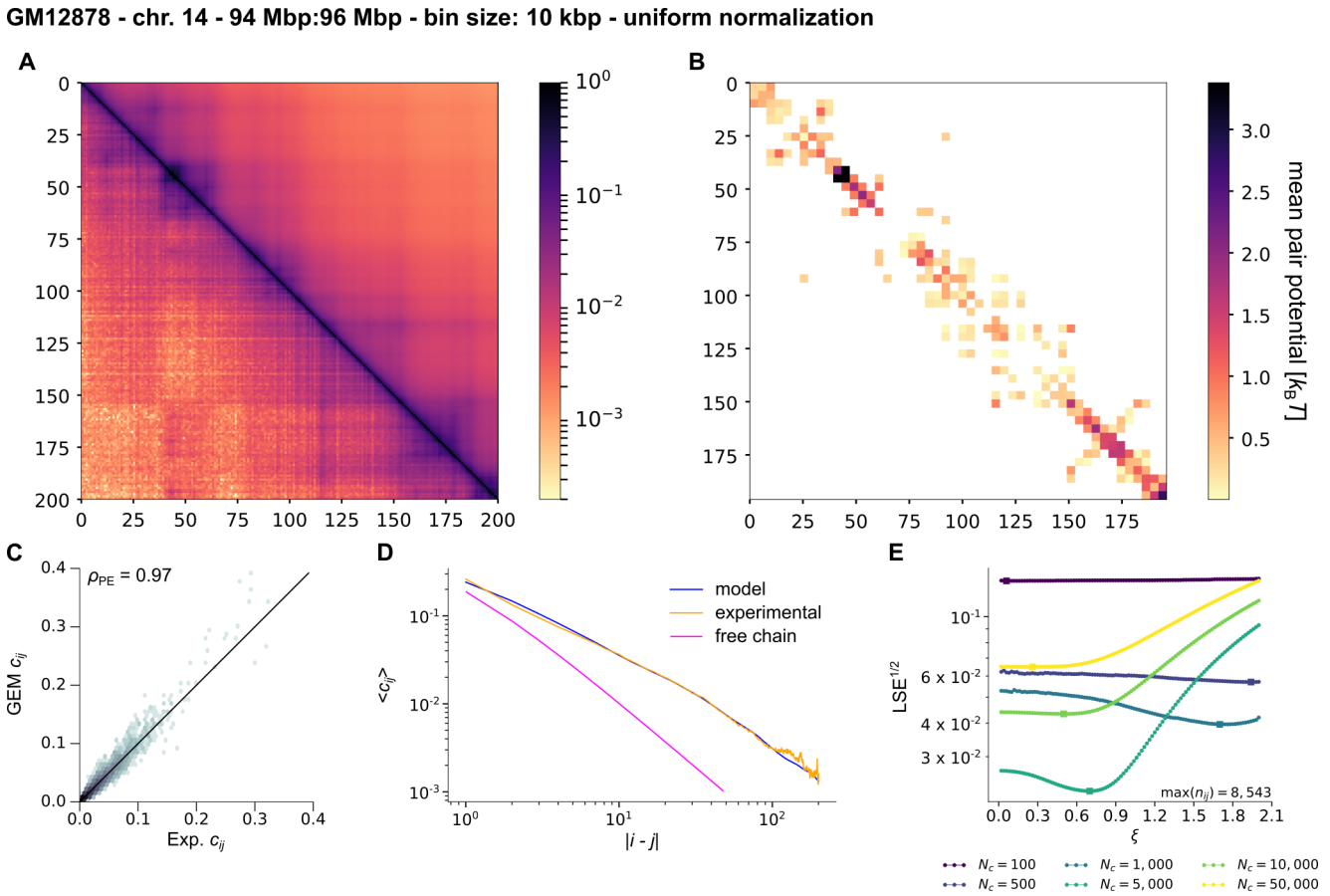

FIGURE S5 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (10 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

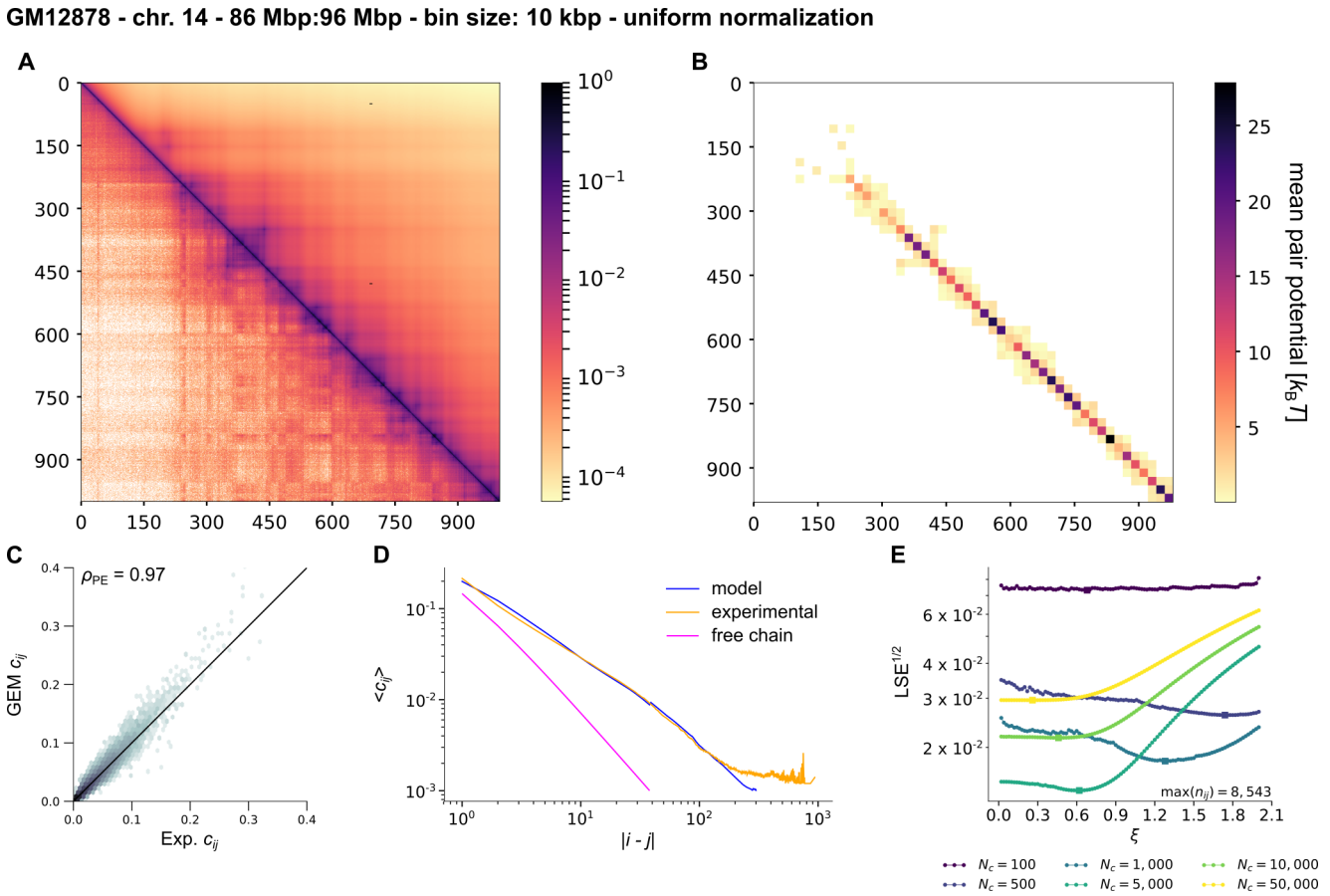

FIGURE S6 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (10 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

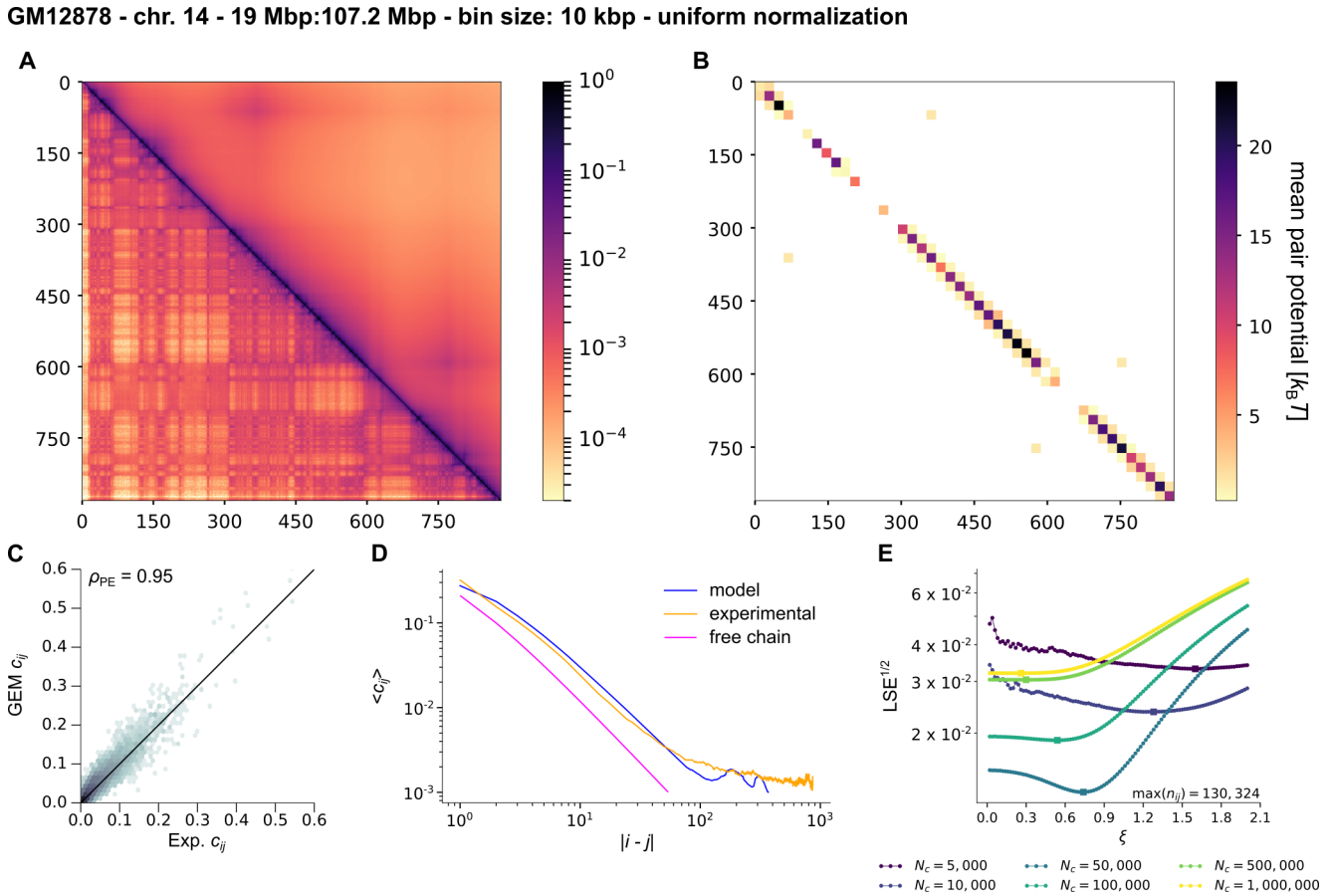

FIGURE S7 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (100 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

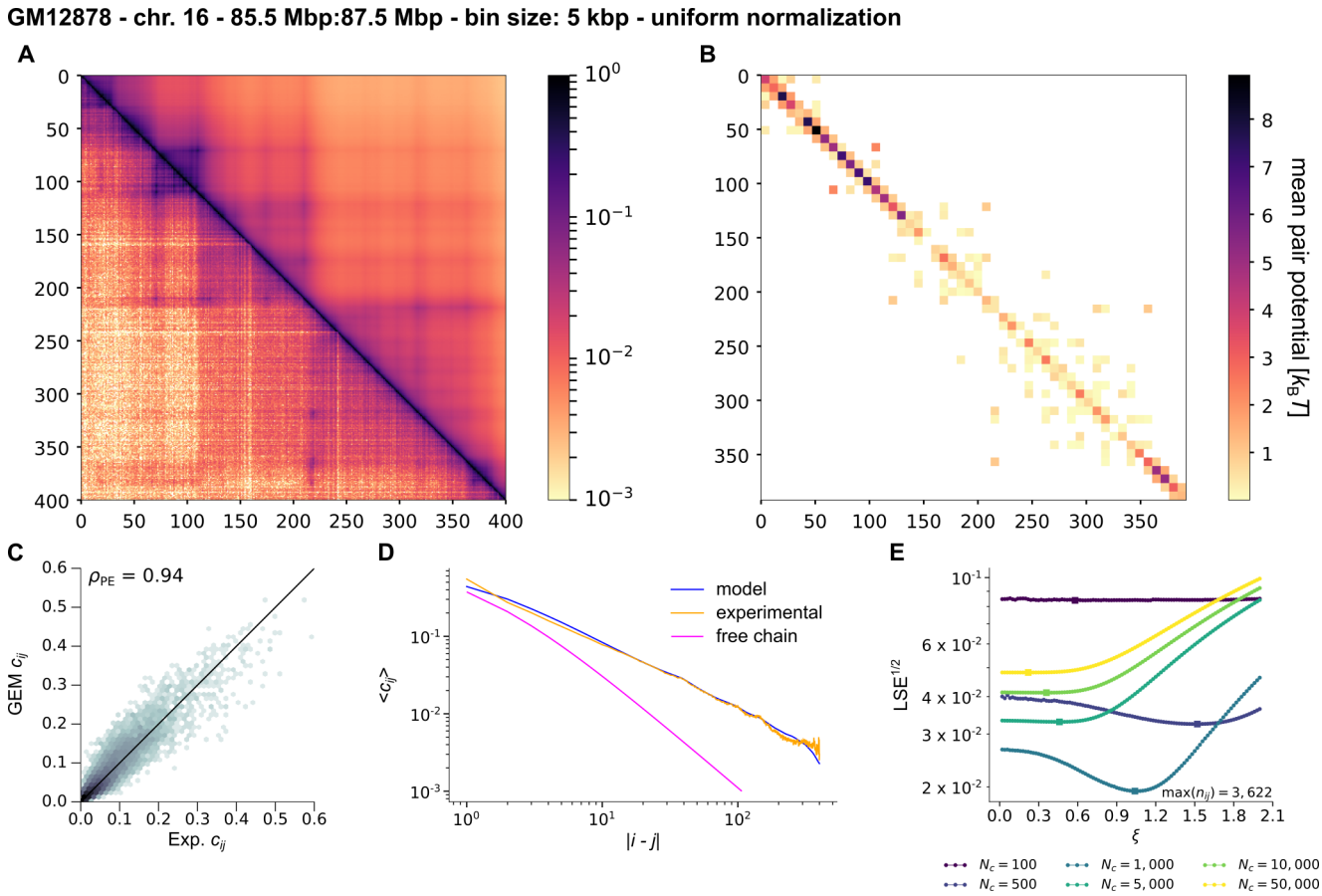

FIGURE S8 – GEM reconstruction for Hi-C data of human chromosome 16 (1) (5 kbp resolution), normalized by applying a global factor. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:8. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping, and for different normalizations N_c of the Hi-C counts.

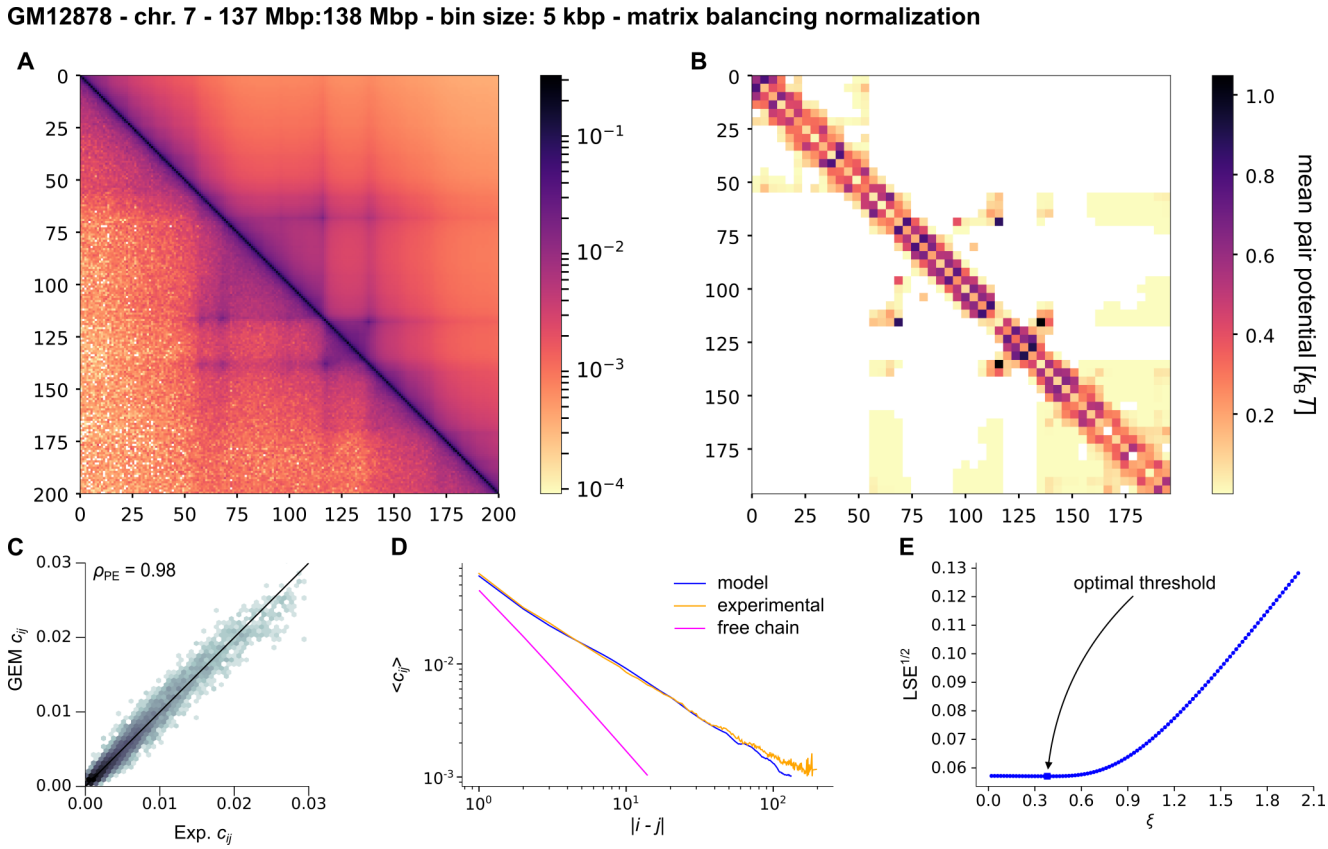

FIGURE S9 – GEM reconstruction for Hi-C data of human chromosome 7 (1) (5 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

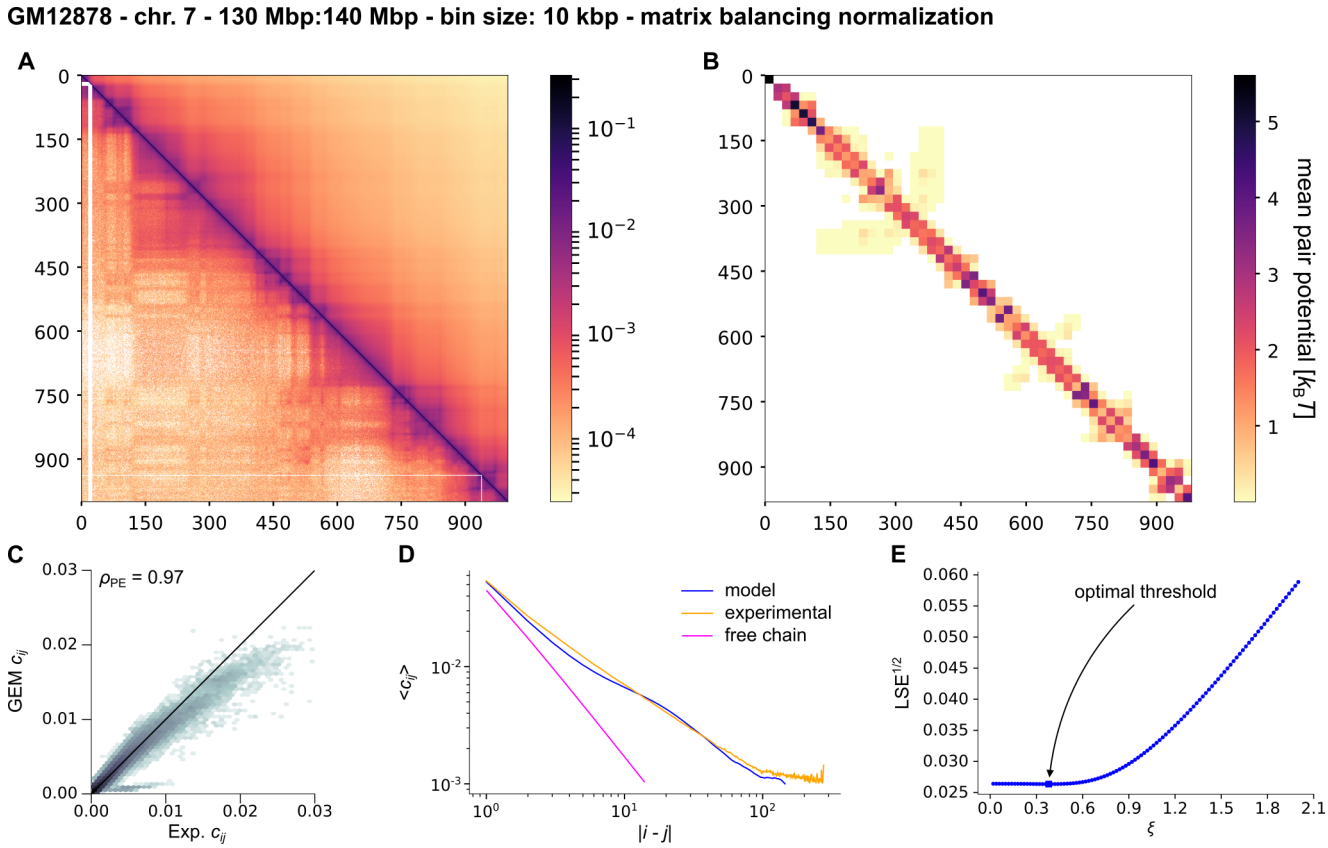

FIGURE S10 – GEM reconstruction for Hi-C data of human chromosome 7 (1) (10 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

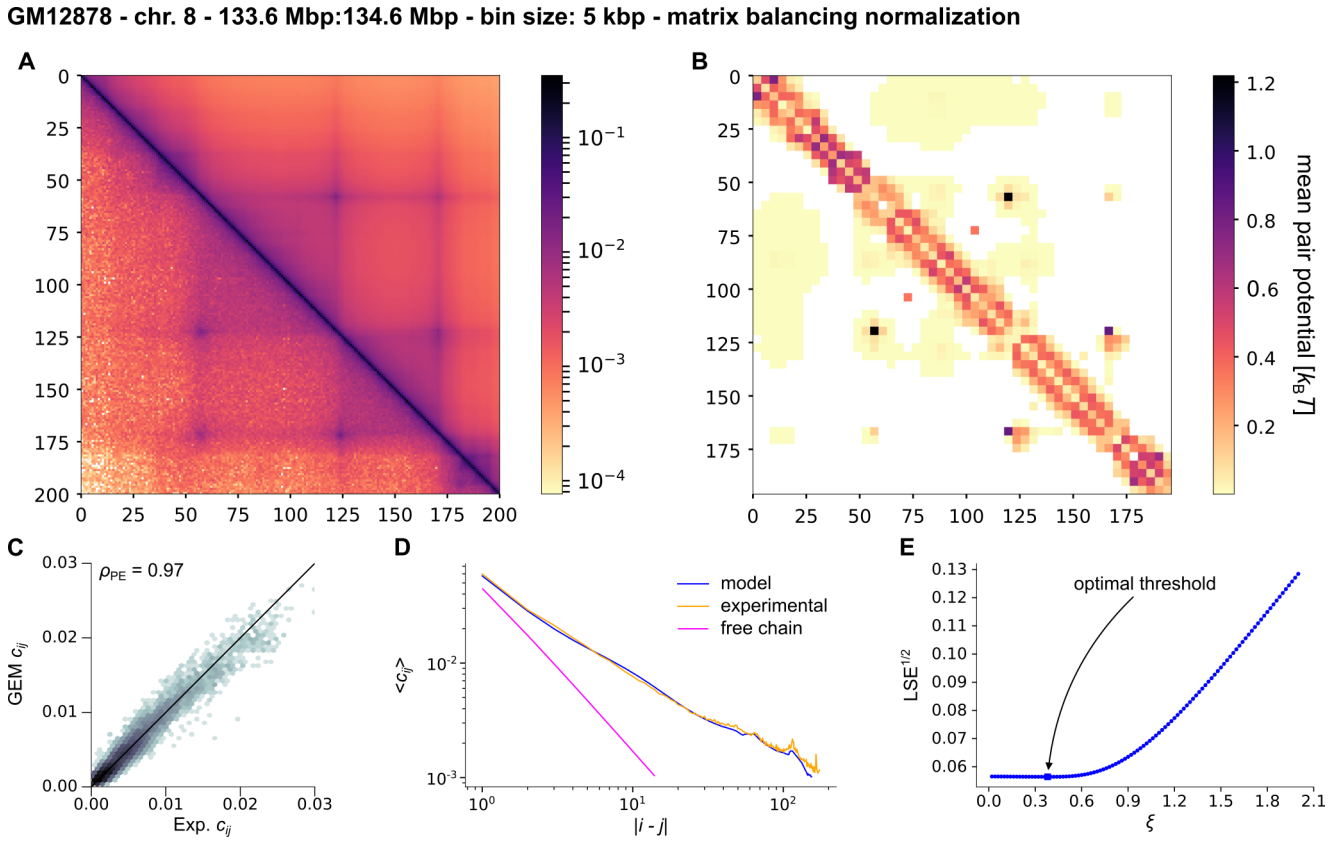

FIGURE S11 – GEM reconstruction for Hi-C data of human chromosome 8 (1) (5 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

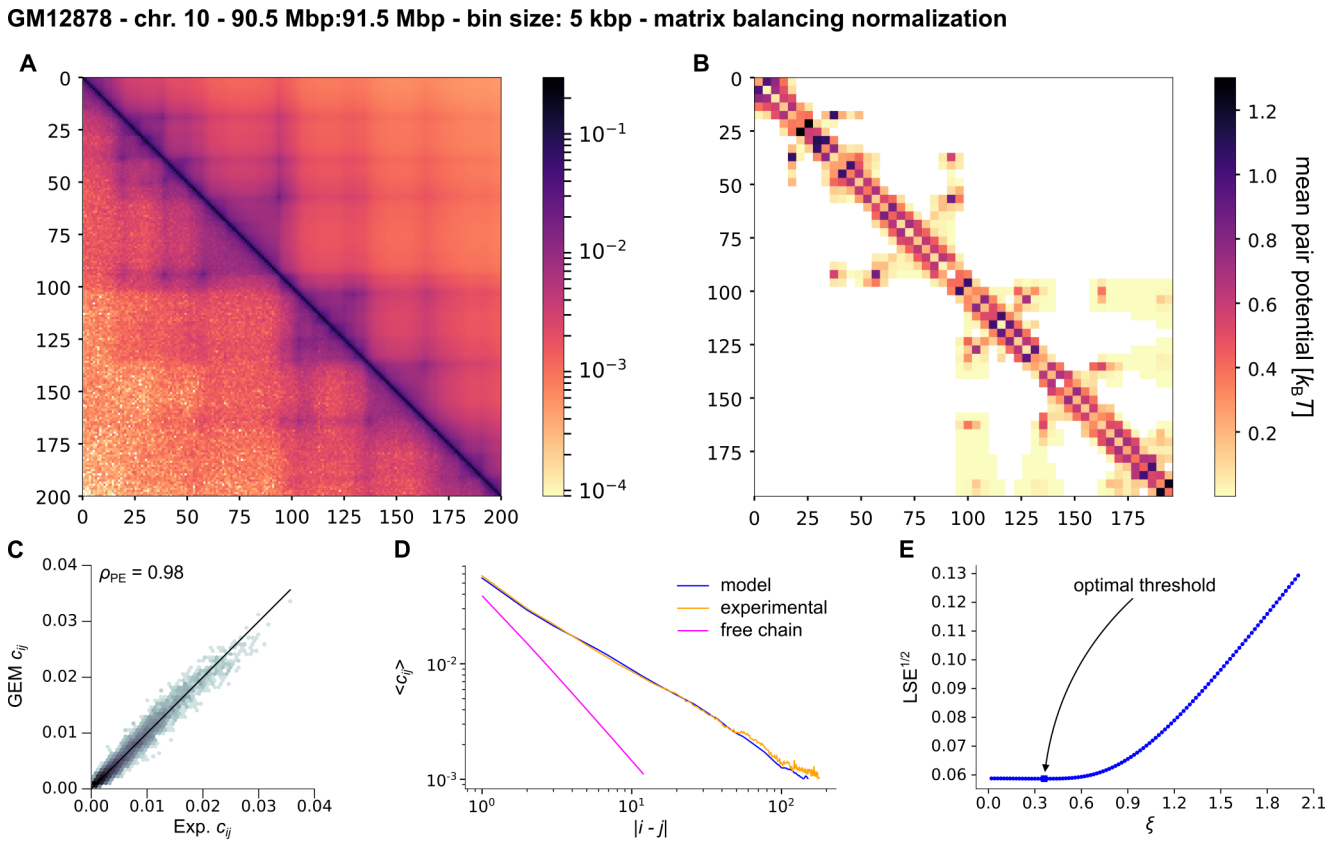

FIGURE S12 – GEM reconstruction for Hi-C data of human chromosome 10 (1) (5 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

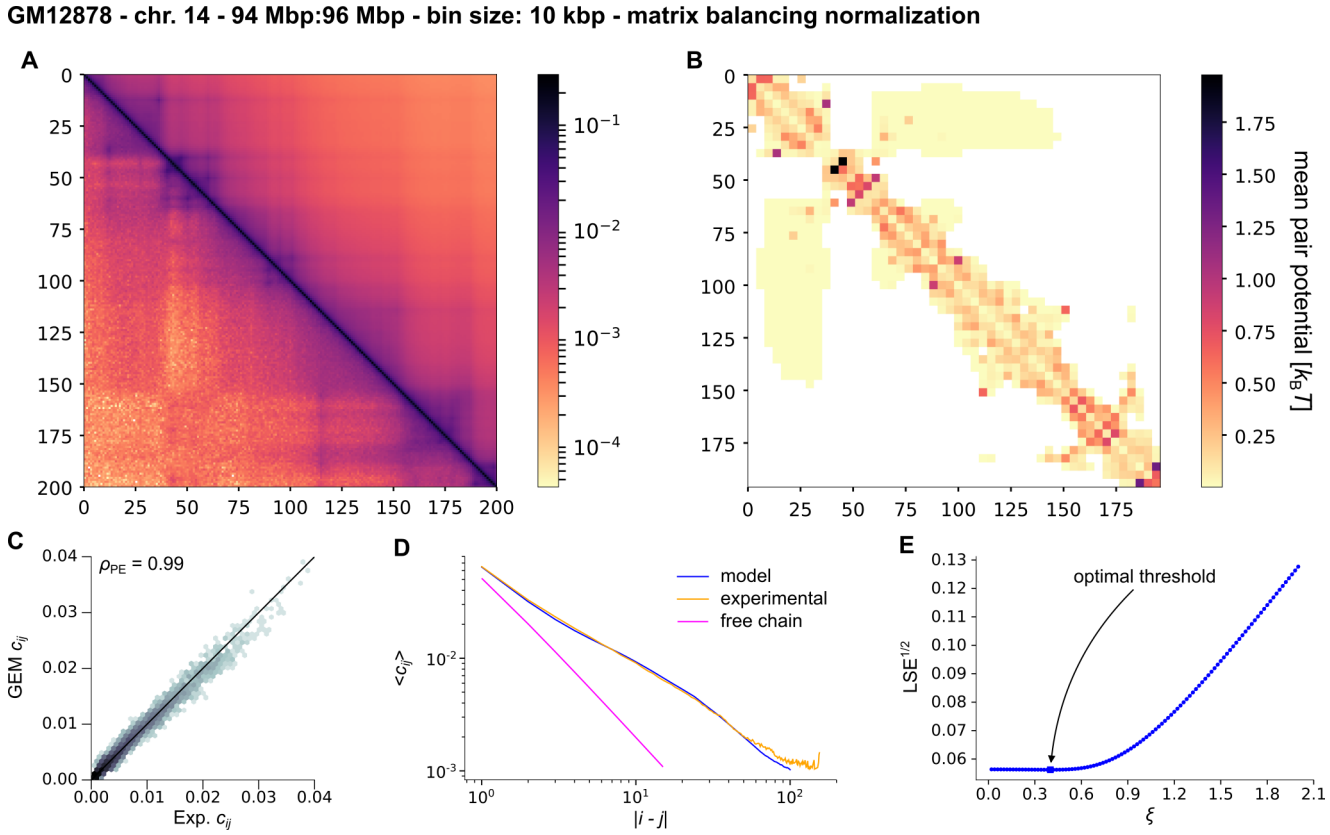

FIGURE S13 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (10 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:4. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

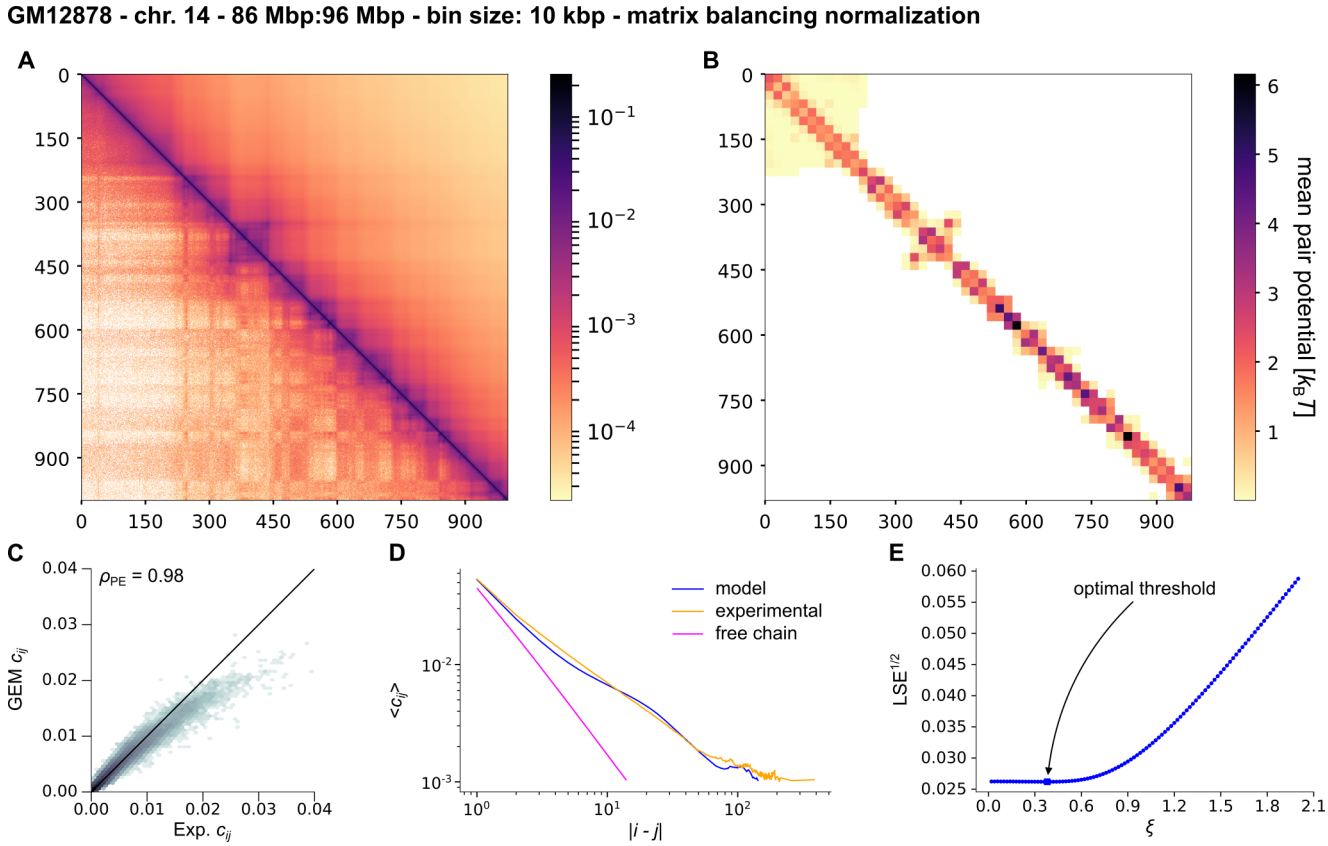

FIGURE S14 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (10 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

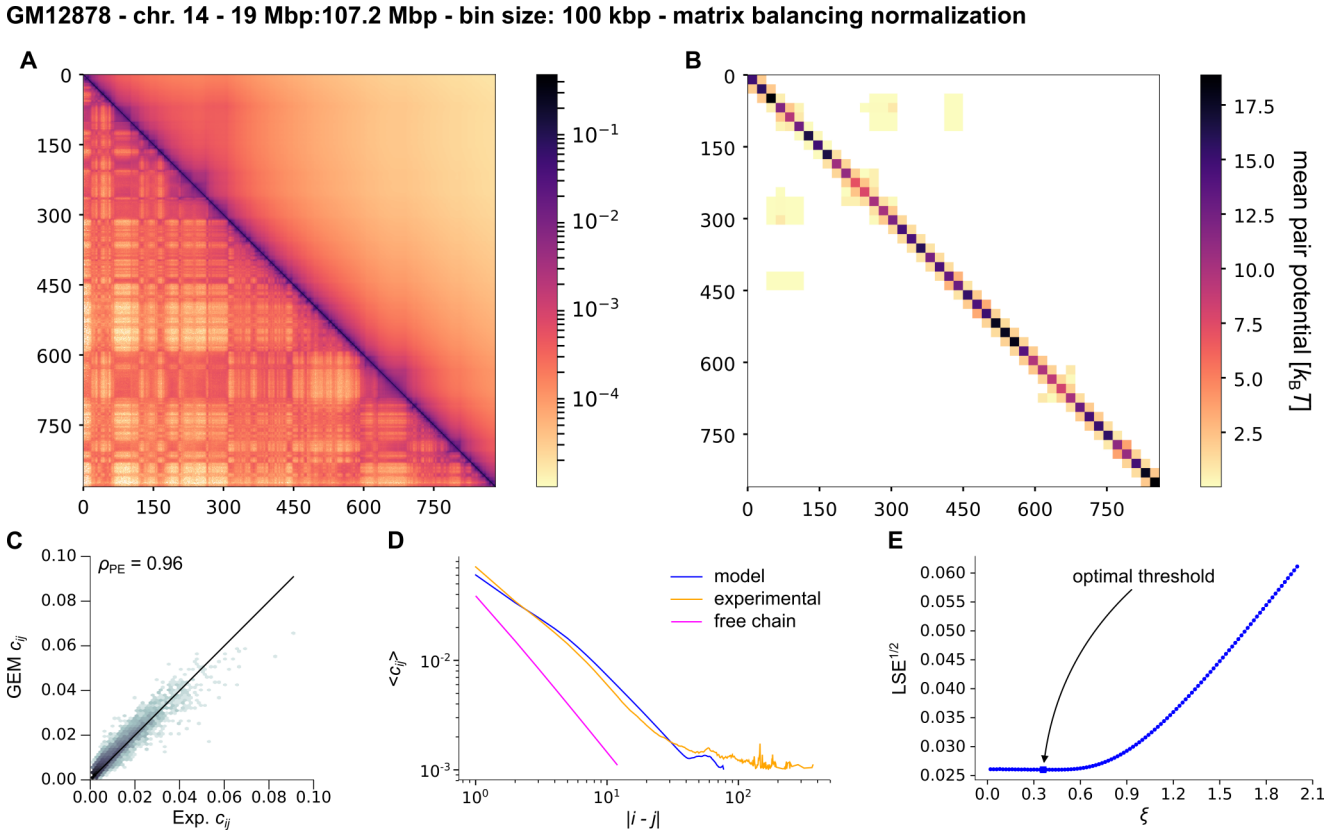

FIGURE S15 – GEM reconstruction for Hi-C data of human chromosome 14 (1) (100 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

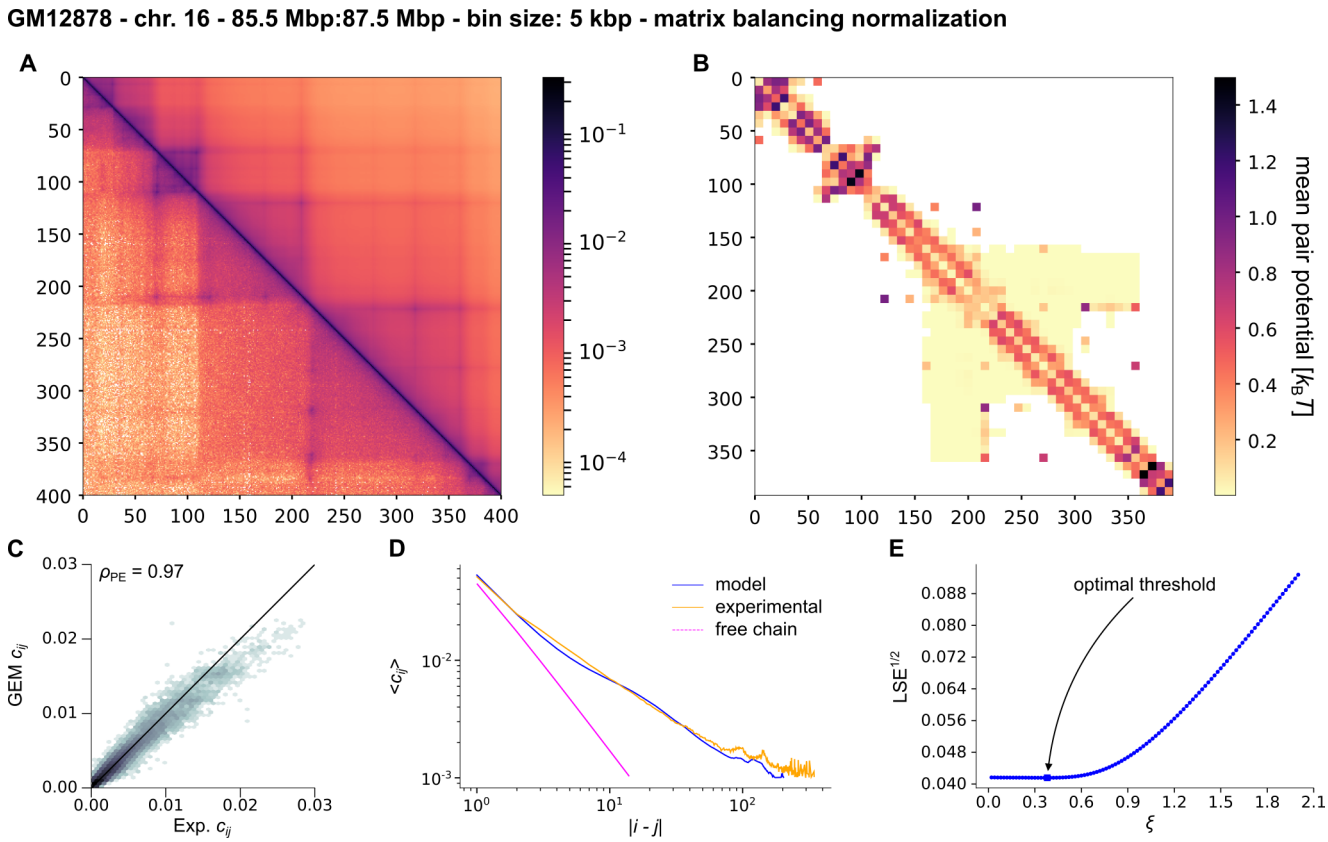

FIGURE S16 – GEM reconstruction for Hi-C data of human chromosome 16 (1) (5 kbp resolution), normalized by matrix balancing. **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:8. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

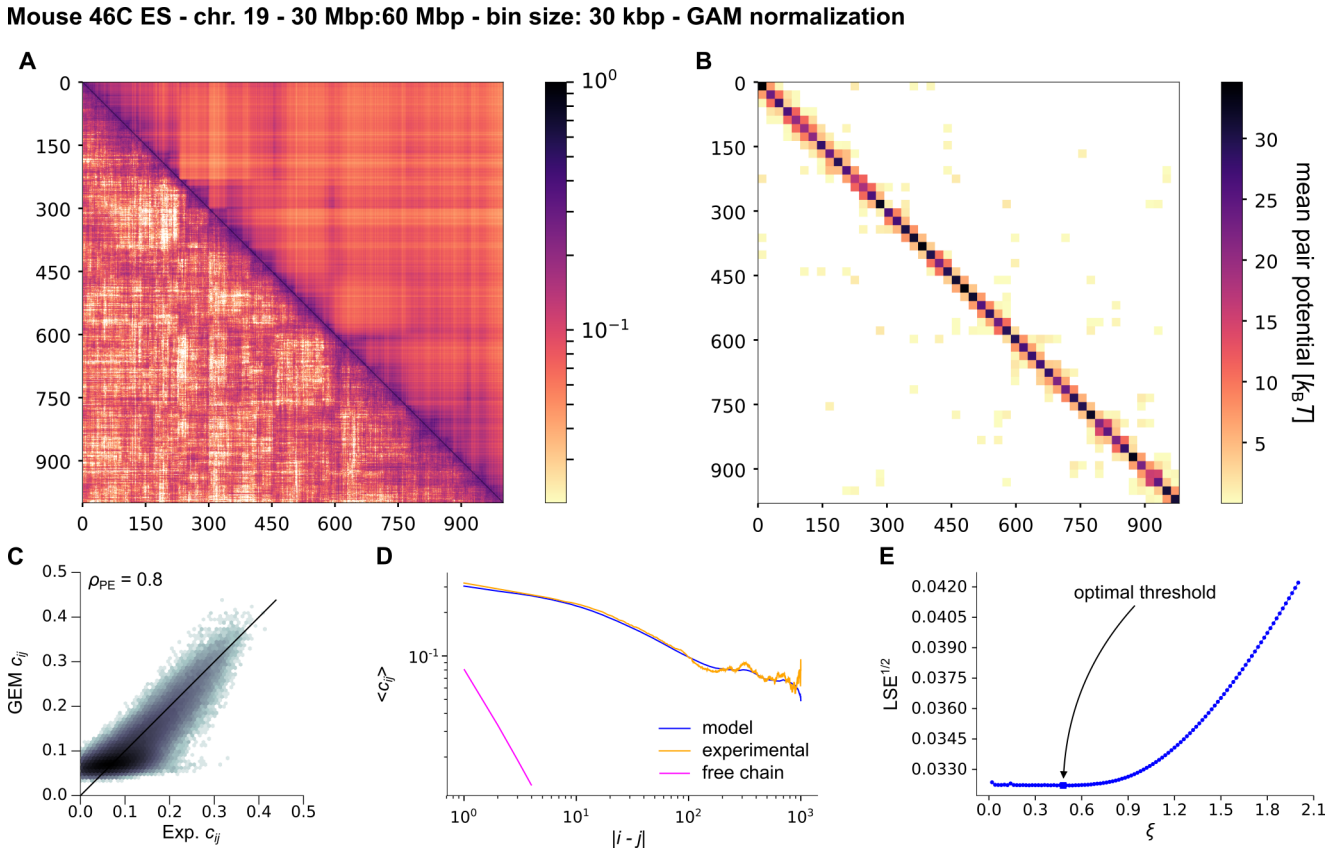

FIGURE S17 – GEM reconstruction for GAM data of mouse chromosome 19 (2) (30 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

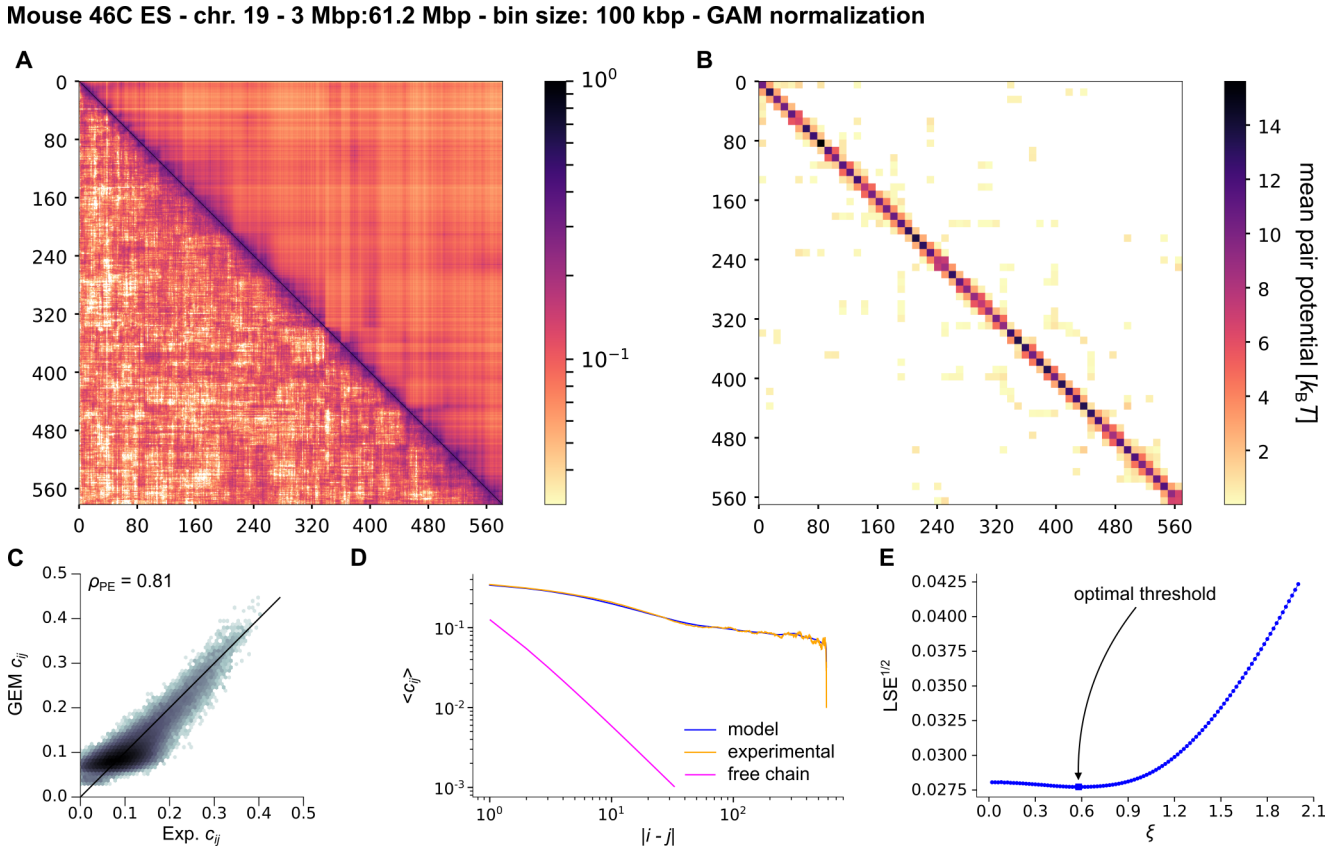

FIGURE S18 – GEM reconstruction for GAM data of mouse chromosome 19 (2) (100 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:10. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

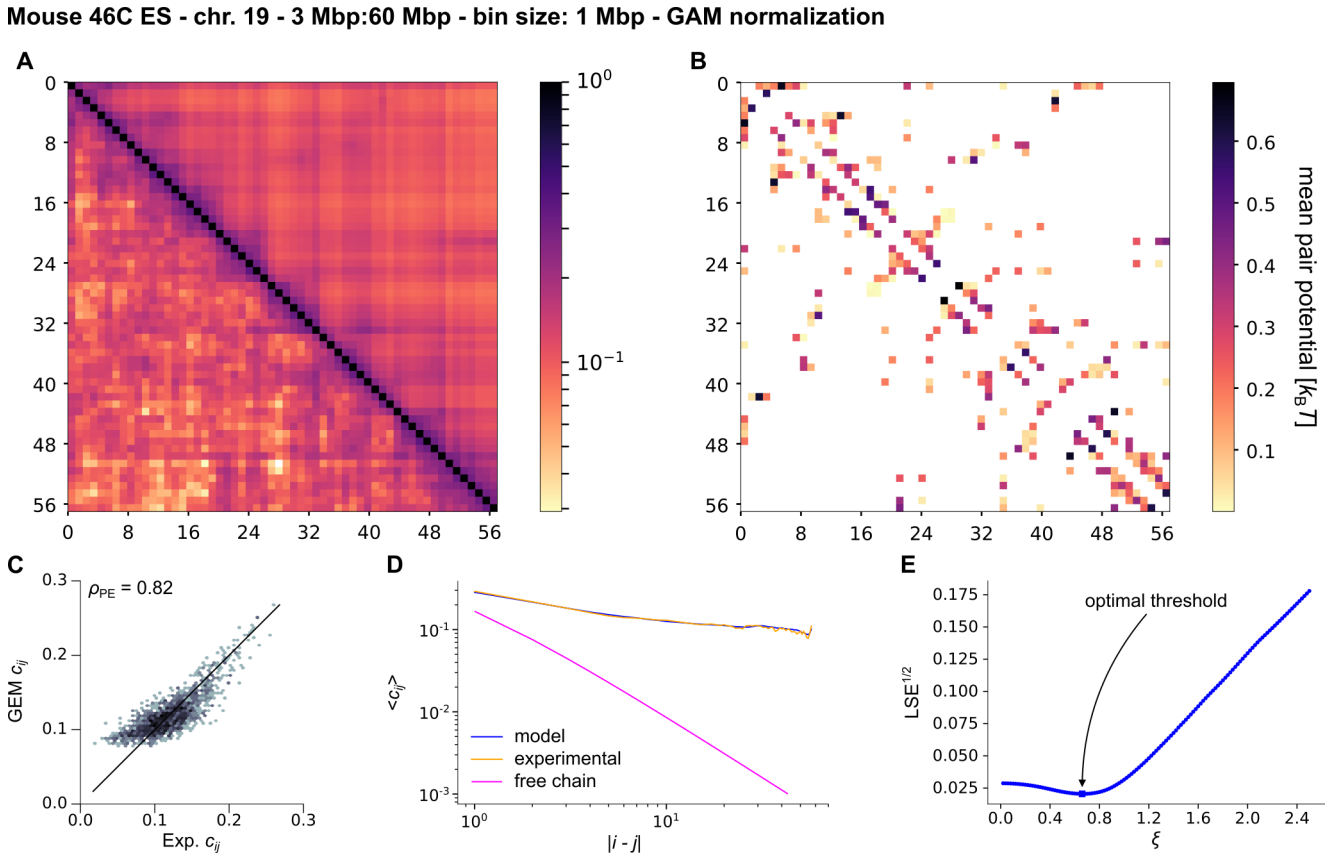

FIGURE S19 – GEM reconstruction for GAM data of mouse chromosome 19 (2) (1 Mbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

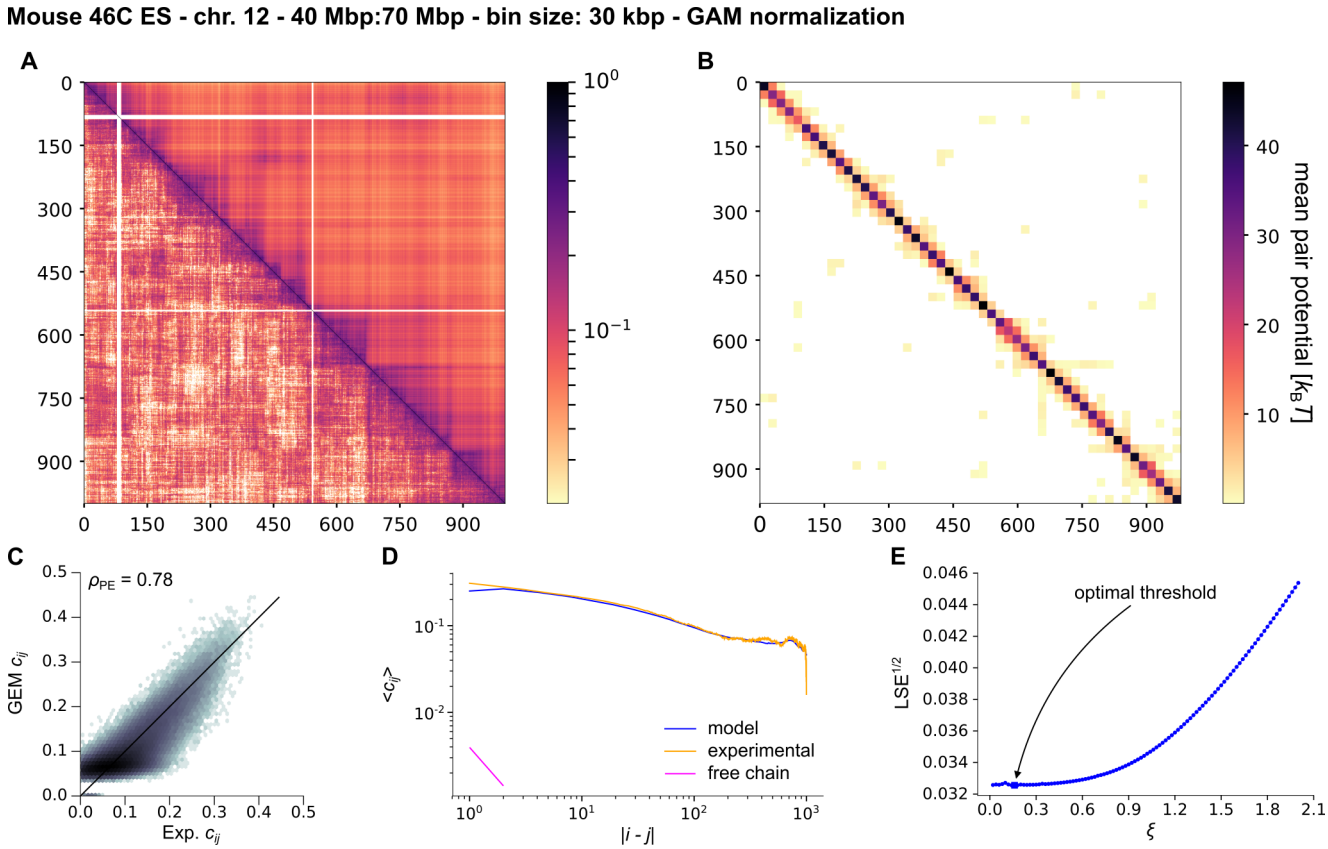

FIGURE S20 – GEM reconstruction for GAM data of mouse chromosome 12 (2) (30 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

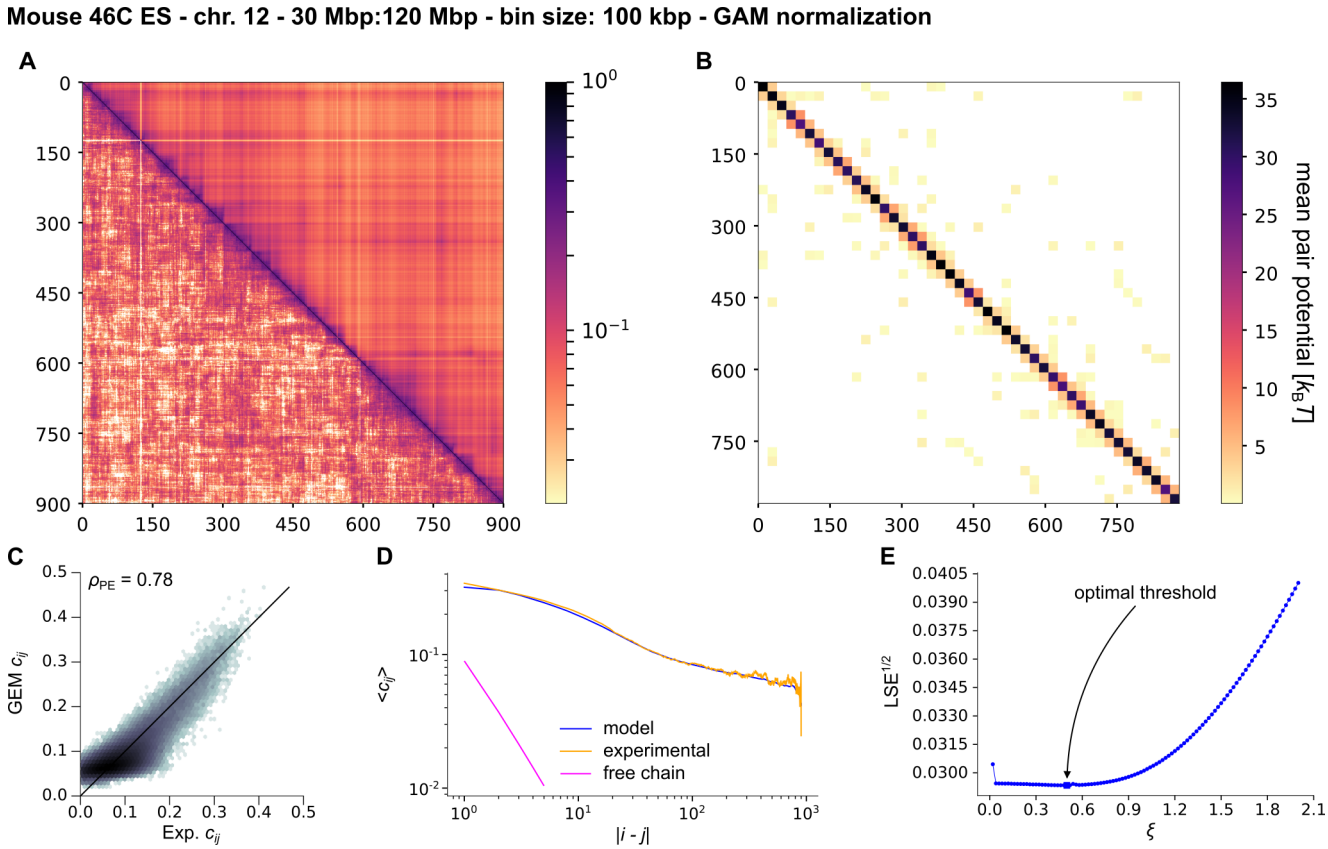

FIGURE S21 – GEM reconstruction for GAM data of mouse chromosome 12 (2) (100 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

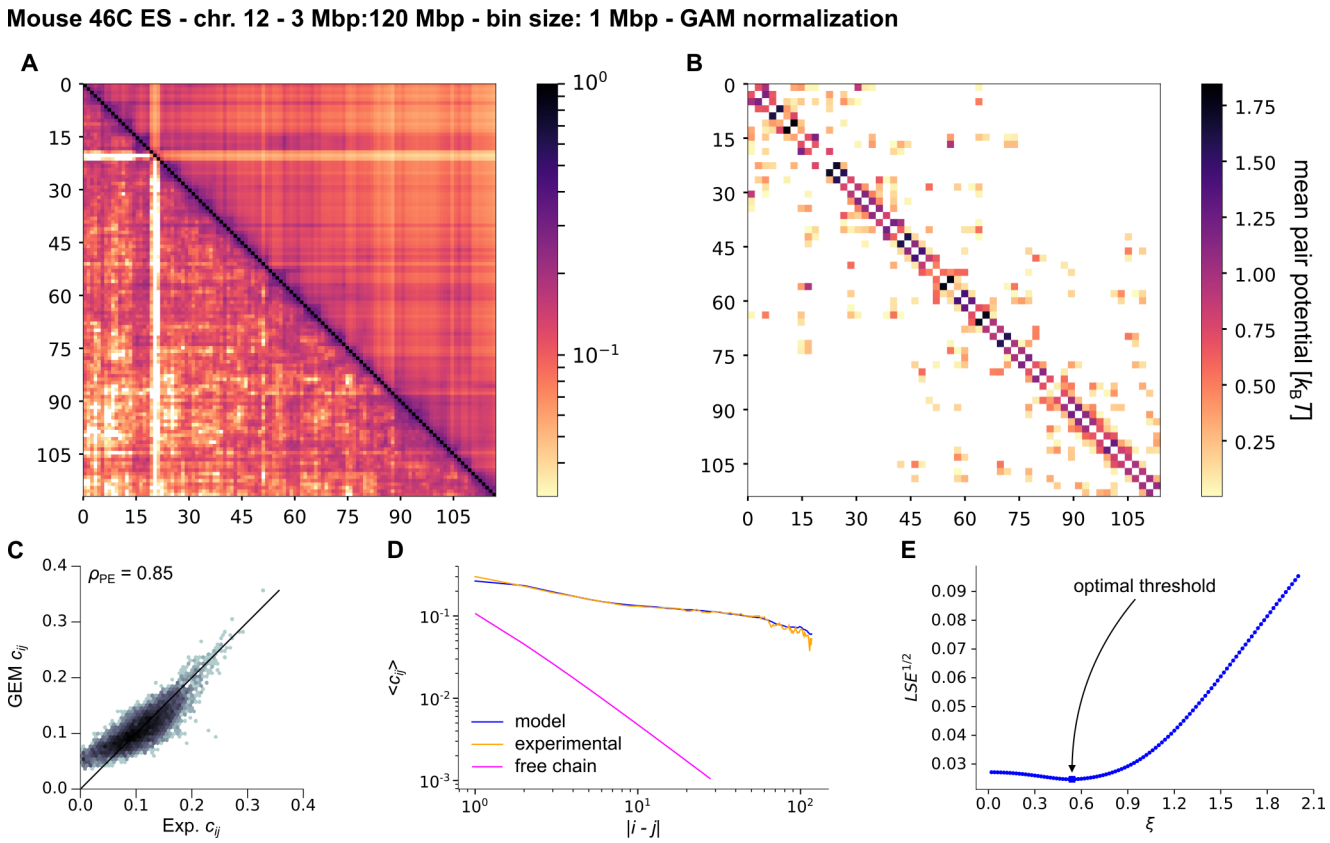

FIGURE S22 – GEM reconstruction for GAM data of mouse chromosome 12 (2) (1 Mbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

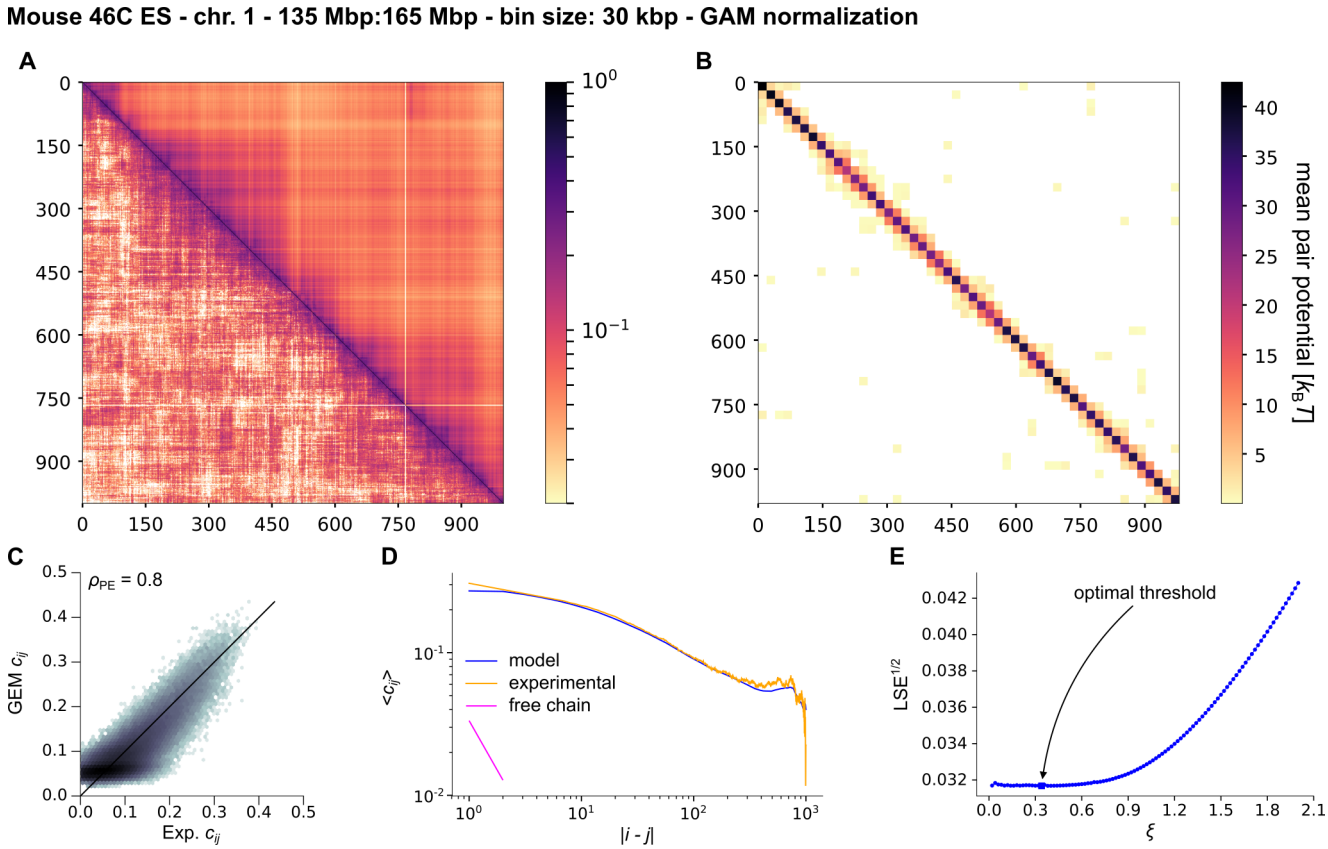

FIGURE S23 – GEM reconstruction for GAM data of mouse chromosome 1 (2) (30 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

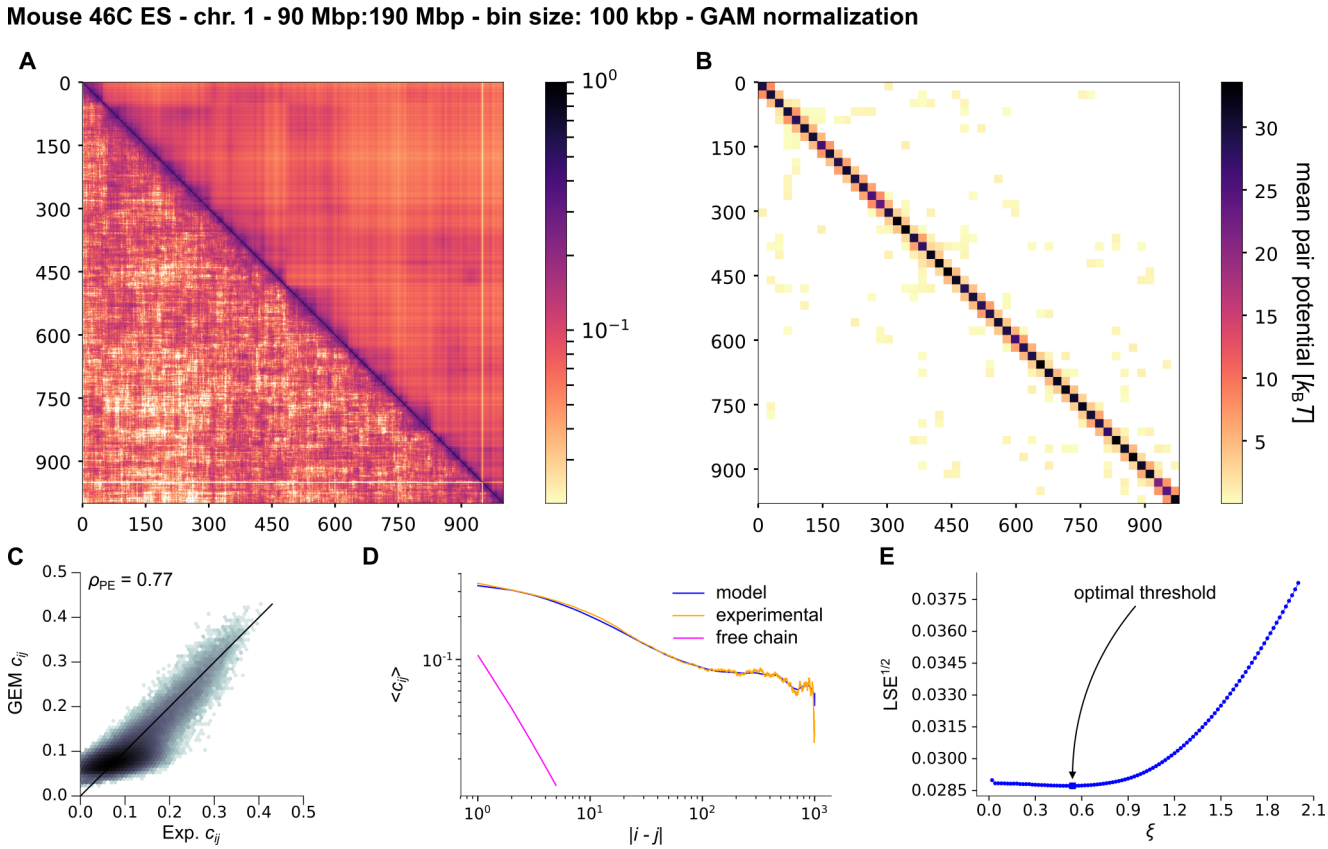

FIGURE S24 – GEM reconstruction for GAM data of mouse chromosome 1 (2) (100 kbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials, binned with a ratio 1:20. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

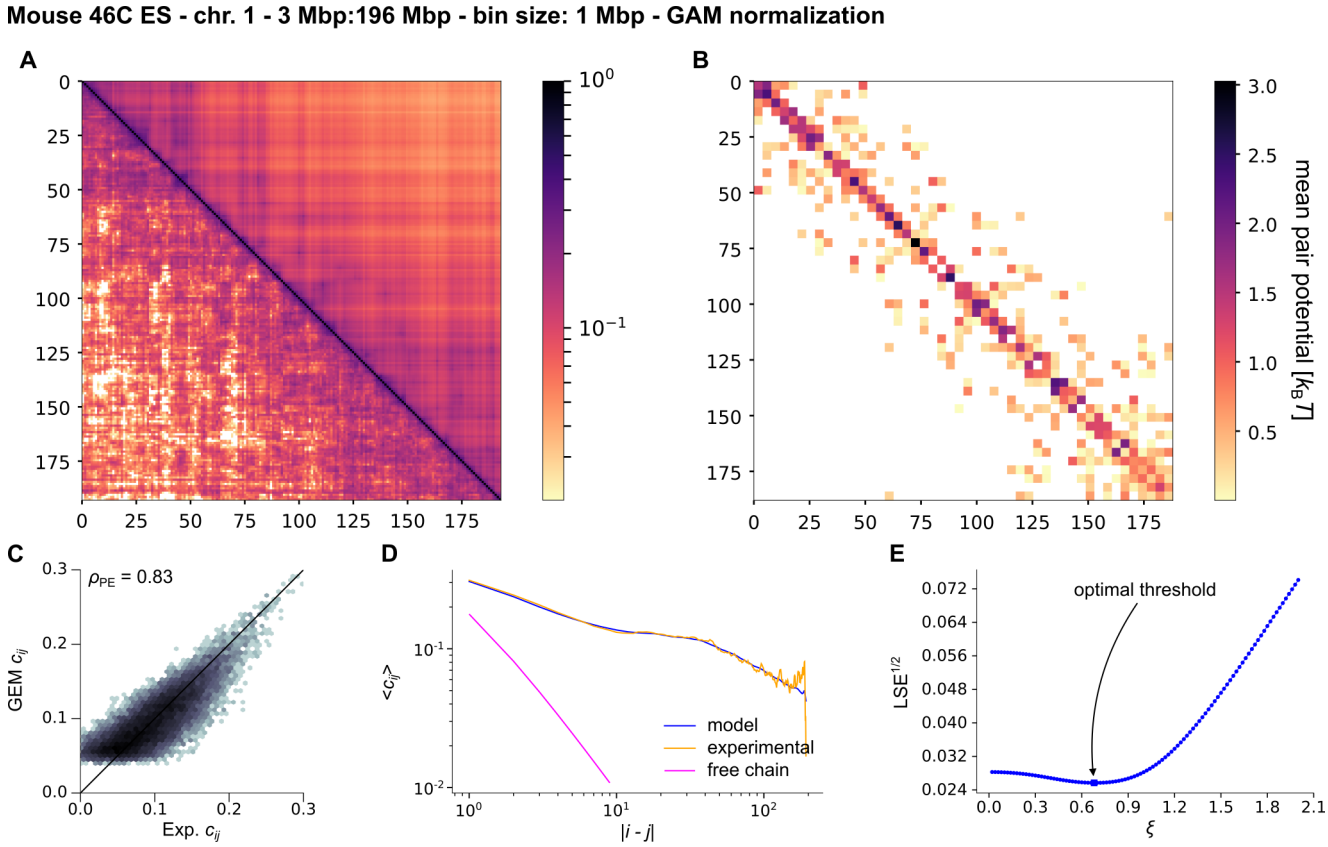

FIGURE S25 – GEM reconstruction for GAM data of mouse chromosome 1 (2) (1 Mbp resolution). **(A)** Comparison between experimental (lower left) and GEM (upper right) contact probabilities. **(B)** Matrix of mean pair potentials. **(C)** Comparison of experimental and GEM contact probabilities (2d-histogram). We give the Pearson correlation coefficient. **(D)** Average contact probability as a function of the contour length. **(E)** LSE as a function of the threshold ξ used for the GEM mapping.

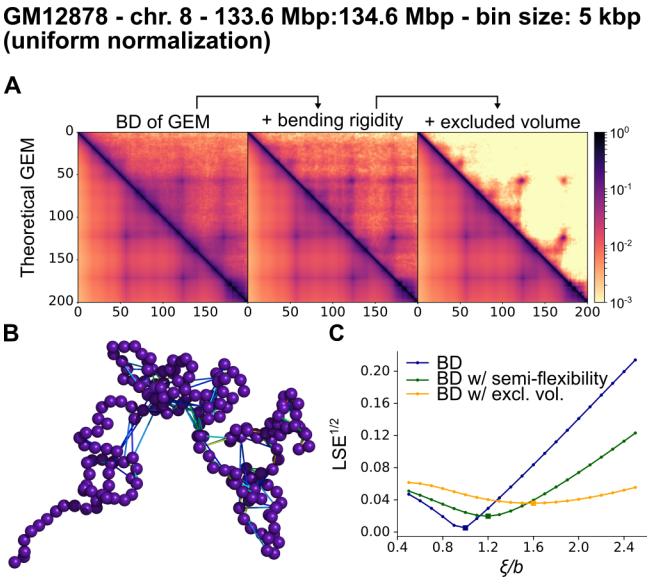

FIGURE S26 – Brownian dynamics (BD) of the reconstructed GEM for Hi-C data of human chromosome 8 (1) (5 kbp resolution). **(A)** Contact probability matrices obtained through BD simulation of: (i) the GEM, (ii) the GEM with bending rigidity, and (iii) the GEM with bending rigidity and with excluded volume. The contact probabilities were computed from BD trajectories and are compared with the theoretical values for the GEM. **(B)** Snapshot of a configuration obtained by BD of the reconstructed GEM with bending rigidity and excluded volume. The couplings are represented by tie lines, from weak couplings (in blue) to strong couplings (in red). **(C)** LSE as a function of the threshold ξ between contact probabilities computed from the BD trajectory and the theoretical values.

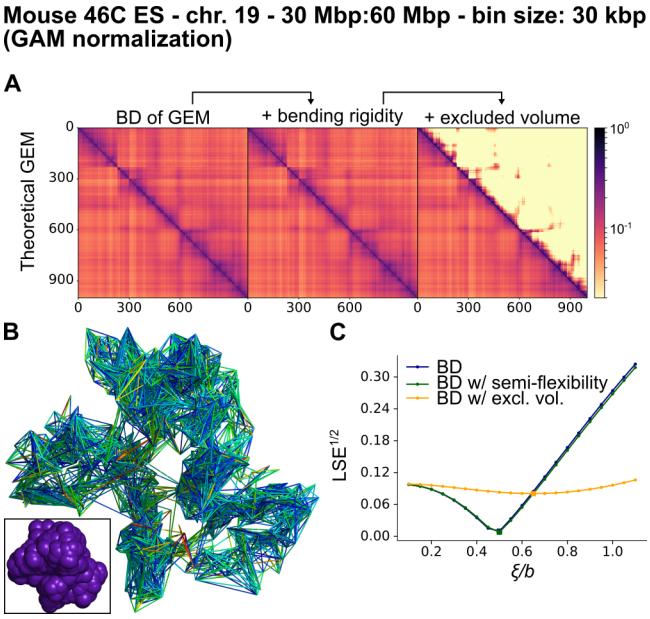

FIGURE S27 – Brownian dynamics (BD) of the reconstructed GEM for GAM data of mouse chromosome 19 (2) (30 kbp resolution). **(A)** Contact probability matrices obtained through BD simulation of: (i) the GEM, (ii) the GEM with bending rigidity, and (iii) the GEM with bending rigidity and with excluded volume. The contact probabilities were computed from BD trajectories and are compared with the theoretical values for the GEM. **(B)** Snapshot of a configuration obtained by BD of the reconstructed GEM with bending rigidity and excluded volume. The couplings are represented by tie lines, from weak couplings (in blue) to strong couplings (in red). The inset shows the same configuration with the monomers. Note that the hard-core distance is $\sigma = 1$ whereas the bond length is $b = 8$. **(C)** LSE as a function of the threshold ξ between contact probabilities computed from the BD trajectory and the theoretical values.

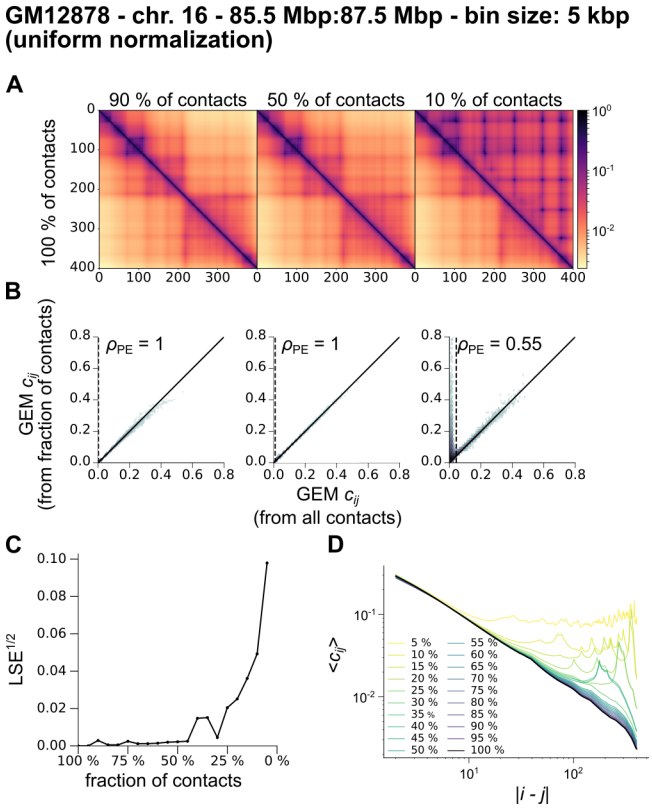

FIGURE S28 – Robustness of GEM reconstruction for Hi-C data of human chromosome 16 (1) (5 kbp resolution). For all GEM reconstructions we used a threshold $\xi = 1$ and a normalization factor $N_c = 10^3$. **(A)** Comparison of the contact probabilities of the reconstructed GEM with those of a GEM obtained by performing the minimization only on the top 90 %, 50 % and 10 % experimental contacts. **(B)** 2d-histograms corresponding to the matrices shown in **(A)**. We give the Pearson correlation coefficients. The thresholding quantiles are represented by vertical dashed lines. **(C)** Comparison of the GEMs reconstructed from a decreasing fraction of the experimental contacts with the original GEM. LSE$^{1/2}$ is the Euclidean distance between contact probabilities divided by $(N + 1)$. **(D)** Average contact probability as a function of the contour length for GEMs reconstructed from a decreasing fraction of the experimental contacts.

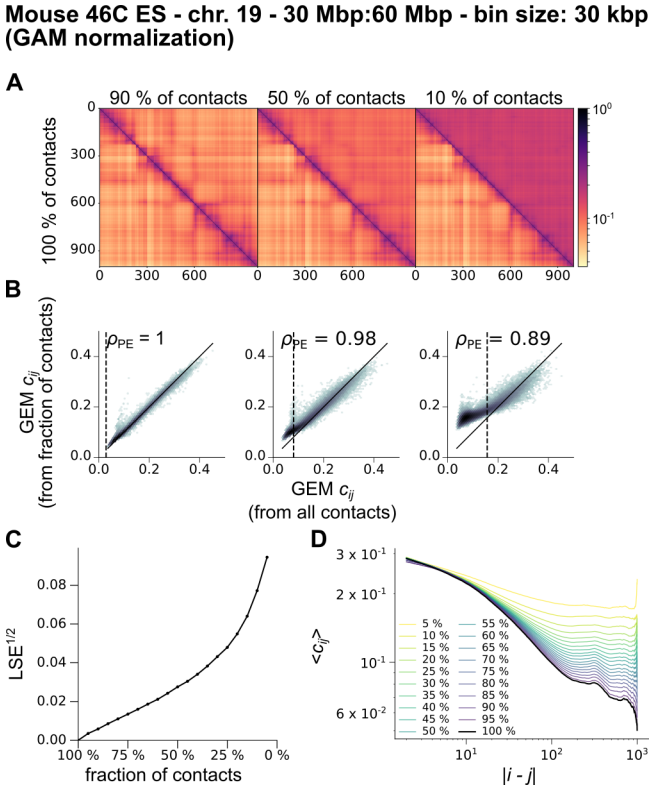

FIGURE S29 – Robustness of GEM reconstruction for GAM data of mouse chromosome 19 (2) (30 kbp resolution). For all GEM reconstructions we used a threshold $\xi = 0.5$. **(A)** Comparison of the contact probabilities of the reconstructed GEM with those of a GEM obtained by performing the minimization only on the top 90 %, 50 % and 10 % experimental contacts. **(B)** 2d-histograms corresponding to the matrices shown in **(A)**. We give the Pearson correlation coefficients. The thresholding quantiles are represented by vertical dashed lines. **(C)** Comparison of the GEMs reconstructed from a decreasing fraction of the experimental contacts with the original GEM. LSE$^{1/2}$ is the Euclidean distance between contact probabilities divided by $(N + 1)$. **(D)** Average contact probability as a function of the contour length for GEMs reconstructed from a decreasing fraction of the experimental contacts.

2. EXISTING METHODS TO RECONSTRUCT CHROMOSOME ARCHITECTURE

Let us review some of the models which have been proposed in the past to address the reconstruction of chromosome architecture from 3C data. Our aim is not to review thoroughly the available methods, but rather to emphasize essential differences with our own approach. For a more detailed review of the existing methods for reconstructing chromosome architecture we refer the interested reader to (3).

2.1. Non-polymer models

2.1.1. Harmonic model

A numerical procedure relying on the introduction of harmonic potentials has been proposed to reconstruct the equilibrium configurations of the chromosome from the experimental contact probabilities (4, 5). Harmonic interactions are introduced between each chromosomal bin pair (i, j), such that the contribution to the internal energy is:

$$U(\{\mathbf{r}_i\}) = \sum_{i<j} \frac{k}{2} \left(r_{ij} - r_{ij}^0\right)^2, \tag{1}$$

in which $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$ is the distance between loci i and j, k is an arbitrarily chosen elastic constant and r_{ij}^0 is the length of the isolated spring. A Monte-Carlo simulation is then performed to sample equilibrium configurations of the system defined in Eq. 1. These configurations are used to represent the chromosome configurations.

In this method, the elastic constant was assigned arbitrarily to $k = 5\, k_\mathrm{B}T$. The fact that this elastic constant is the same for all (i, j) is a first limitation in this approach. The spring lengths are taken such that $r_{ij}^0 = d_{ij}$, where d_{ij} is the distance desired between beads i and j. The authors assumed that the equilibrium distance between two chromosomal loci is inversely proportional to the contact probability, $d_{ij} = 1/c_{ij}$. We will come back to this assumption.

2.1.2. Constraint satisfaction

Another approach is to cast the problem of reconstituting chromosome architecture into a constraint satisfaction problem (6). The reformulated problem then consists in finding the coordinates $\{\mathbf{r}_i\}$ such that the distances between any pair of chromosomal bins (i, j) is bounded from below and from above:

$$a_{ij} < r_{ij} < b_{ij}. \tag{2}$$

In Eq. 2 the upper bound is taken inversely proportional to the experimental contact probability, $b_{ij} \propto 1/c_{ij}$, and the proportionality coefficient is a parameter of the method. The lower bound a_{ij} is introduced to take into account excluded volume between any pair of chromosomal loci, and to penalize contacts between adjacent loci due to the chromosome bending rigidity. This is a constraint satisfaction problem, which can be solved with the simplex method. The obtained solution is then used to represent a chromosome configuration.

The main limitation of this approach is clearly that the choice of the lower and upper bounds must be adjusted by the user and adapted to each data set. Beside, this is not a physical model of the chromosome architecture.

2.1.3. Singular value decomposition of the spatial correlation matrix

Let us consider the matrix R of size $d \times N$, where $d = 3$ is the space dimension and N is the number of bins in the Hi-C contact matrix. The matrix element $r_{\alpha i}$ is therefore the spatial coordinate of loci i along the α-axis ($\alpha = x, y, z$). Next we consider the Singular Value Decomposition (SVD) of R:

$$r_{\alpha i} = \sum_{\gamma=1}^{d} \lambda_\gamma u_{\alpha\gamma} v_{i\gamma}, \tag{3}$$

where $U = [u_{\alpha\gamma}]$ and $V = [v_{\gamma i}]$ are two orthogonal matrices, and $\left\{\lambda_\gamma\right\}_{\gamma=1,\ldots,d}$ are the singular values of R. Then $C = R^T R$ and $\tilde{C} = RR^T$ have the same non-zero eigenvalues, which are λ_1^2, λ_2^2 and λ_3^2 (if $d = 3$). Finally we introduce the matrix of distances, D, with elements:

$$d_{ij} = \sqrt{\sum_{\alpha=1}^{d} \left(r_{\alpha i} - r_{\alpha j}\right)^2}. \tag{4}$$

It turns out that the correlation matrix C can be obtained from the distance matrix D (7, 8). Therefore, from the knowledge of the distances, one can infer the singular values of the coordinates matrix, and obtain an approximation for R.

2.2. Polymer models

Models presented in section 2.1 lack a physical model of the chromosome. In clear, the Hi-C bins define a gas of particles with coordinates $\{\mathbf{r}_i\}$ and minimizing Eq. 1 (resp. solving Eqs. 2 and 4) can result in configurations that violate topological constraints of the polymer chain representing the chromosome. Therefore, subsequent improvements have consisted in incorporating a polymer model of the chromosome when attempting to reconstruct chromosome architecture.

2.2.1. Random walk backbone with tethered loops

Another way to look at Hi-C data is to consider that when the contact probability between loci i and j is high enough, it defines a DNA loop. This is the approach taken in (9). In short, whenever

$$c_{ij} > c_{min}, \tag{5}$$

with an arbitrary lower bound c_{min} on the contact probability, the authors considered that the DNA subchain in the interval $[i, j]$ constitutes a loop, with $\mathbf{r}_i = \mathbf{r}_j$. The chromosome is then represented by a backbone polymer with Gaussian statistics on which are tethered polymer loops with varying sizes. Numerical simulations are then performed on the basis of this polymer model of the chromosome.

2.2.2. First-principle approach

In (10, 11), the authors start from a polymer representation of the chromosome, and add interactions between different regions of the chromosome. However, due to the complexity of chromosome interactions with proteins, this kind of studies can only be made under strong simplifying assumptions. For example, a unique generic type of protein is included and/or the variety in the binding energies with different loci on the chromosome is replaced by a single binding energy (or just a few). For this reason comparisons with experimental contact matrices have been rather qualitative.

2.2.3. Inverse approach

As mentioned in the main text, chromosome architecture might be well described with an effective model in which microscopical details, such as proteins and sequence effects, are coarse-grained. In particular, the effect of structuring proteins can be taken into account implicitly by introducing an effective potential $V_{ij}(r)$ between each (i, j) monomer pair. In other words, each location on the genome experiences an effective interaction with the other loci on the genome, which mimics the effect of multivalent proteins. This type of approach was used, in which such potentials

are considered to be short-range square potentials (12):

$$V_{ij}(r) = \begin{cases} +\infty & \text{if } r < \sigma \\ -\varepsilon_{ij} & \text{if } \sigma < r < \xi \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where σ is the hard-core distance and ξ is a threshold which defines at the same time the range of the potential and the distance below which monomers i and j are said to be in contact. By performing MC simulations on a polymer model with the pair potentials in Eq. 6, one can obtain equilibrium configurations and use them to compute contact probabilities between monomer pairs.

Let us note c_{ij}^{exp} the experimental contact probability between restriction fragments i and j obtained from Hi-C experiments, and c_{ij} the contact probability between monomers i and j obtained from MC simulations of a polymer model with potentials as in Eq. 6. We define the least-square estimator between the experimental and the predicted contact matrices:

$$d(c_{ij}, c_{ij}^{exp}) = \frac{2}{N(N+1)} \sum_{i<j} \left(c_{ij} - c_{ij}^{exp}\right)^2, \tag{7}$$

Finding a good model for chromosome architecture now consists in finding a collection of potentials $V_{ij}(r)$ that minimize $d(c_{ij}, c_{ij}^{exp})$. The solution is achieved at the optimal values for σ, ξ and the matrix of binding energy ε_{ij}. In (12), a MC simulation was performed at each step of the minimization procedure, in order to re-sample equilibrium configurations of the chromosome and compute the c_{ij} values. Therefore the computational burden is high.

3. SCALING OF CONTACT PROBABILITIES OF A POLYMER

Several of the methods we have presented (4, 6, 7) have the inconvenience to rely on an estimate of the average distances between loci on the chromosome taken to be inversely proportional to the contact probabilities:

$$d_{ij} \propto 1/c_{ij}. \tag{8}$$

While Eq. 8 may appear to be a reasonable assumption, there is no fundamental reason to support it. As pointed out in (3), a more general functional dependence would be $d_{ij} \sim c_{ij}^{-\gamma}$. For instance, if we model the chromosome as a polymer with scaling exponent ν, we have (13):

$$\begin{aligned} &\Pr\left(\mathbf{r}_{ij}\right) \simeq \frac{1}{\langle r_{ij}\rangle^d} f_p\left(\frac{r_{ij}}{\langle r_{ij}\rangle}\right), \qquad f_p(x) \underset{x\sim 0}{\sim} x^g \\ &\langle r_{ij}\rangle \simeq b \mid i-j \mid^{\nu}. \end{aligned} \tag{9}$$

Let us consider that the contact probabilities are given by $c_{ij} = \Pr\left(r_{ij} = b\right)$, and write $d_{ij} = \langle r_{ij}\rangle$. Then, we obtain the relation:

$$d_{ij} \sim 1/c_{ij}^{1/(d+g)}. \tag{10}$$

For a Gaussian chain, we have $g = 0$, and for a self-avoiding chain, $g = 1/3$. Hence we obtain ($d = 3$), $d_{ij} \sim 1/c_{ij}^{0.33}$ and $d_{ij} \sim 1/c_{ij}^{0.3}$, in direct contradiction with Eq. 8.

Reducing chromosome architecture to a mere conformation characterized by the average pair distances d_{ij} is probably unrealistic. Indeed, co-localization of loci on the chromosome results from the effect of divalent (or multivalent) proteins. We may estimate the strength of the binding by considering contributions of about one $k_{\mathrm{B}}T$ per significant contact (14). Thus, we may consider that structuring proteins have a binding energy with DNA in the range $\varepsilon = 3 - 20\, k_{\mathrm{B}}T$. Consequently, the probability to form a DNA loop between monomers i and j should read:

$$\Pr\left(r_{ij} = b\right) \simeq \frac{1}{\mid i-j \mid^{\nu(d+g)}} e^{\beta\varepsilon} \qquad (b = 1), \tag{11}$$

where $\nu(d+g) = 2$ for a self-avoiding polymer chain with scaling exponent $\nu = 3/5$. For example, considering a relatively strong transcription factor, with $\varepsilon = 10\, k_{\mathrm{B}}T$, the contact probability $c_{ij} \approx 1$ when $\mid i - j \mid = 150$ monomers and falls quickly to zero for larger contour distances. Here a monomer typically represents the diameter of the DNA fiber. In eukaryotes, a monomer typically represents 3000 bp. Therefore, it is very unlikely that chromosome loops are stable for contour length beyond 500 kbp approximatively. In other words, thermodynamic fluctuations may provide the chromosome folding with a non negligible conformational entropy.

4. CONVERSION OF HI-C AND GAM DATA INTO CONTACT PROBABILITIES

In this section, we present the methods that have been used in this article to estimate experimental contact probabilities from the experimental measurements.

4.1. Hi-C

After sequencing, the read-pairs obtained in Hi-C experiments are mapped to a reference genome. Provided that the genome is divided into bins of equal size, each read can then be associated to a unique bin, say i, on the genome. Therefore, each read-pair defines a contact between the corresponding bin-pair. *In fine*, a contact count matrix $[n_{ij}]$ can be constructed, where each entry n_{ij} represents the number of times bins i and j were found in contact in the experiment. From this count matrix, the matrix of contact probabilities can be estimated. In the sequel we present the two methods that have been used in this article to compute the contact probability matrix $[c_{ij}]$ from the count matrix $[n_{ij}]$.

4.1.1. Uniform normalization

In first approximation, it seems reasonable to consider that n_{ij} represents the number of cells in which bins i and j were found in contact. Assuming that N_c is the number of cells in the experiment sample, the contact probability between bins i and j is simply:

$$c_{ij} = \frac{n_{ij}}{N_c}. \tag{12}$$

The previous expression suggests that the matrix of contact probabilities can be obtained from the count matrix by applying a global normalization factor. In practice however, the number of cells in the sample is unknown. Therefore, when using this normalization method to reconstruct the optimal Gaussian effective model, we have tried several values for N_c and chosen the value giving the smallest distance between contact probabilities of the model and of the experiment.

4.1.2. Matrix balancing

Although intuitive, the "uniform normalization" presented above suffers from several pitfalls inherent to the Hi-C protocol. Sources of bias in the n_{ij} counts comprise: chromatin accessibility to the restriction enzyme, alignability (*e.g.* one bin containing many repeats may result in very few detected contacts because reads cannot be aligned uniquely) and restriction site density on the chromosome. For example, if one bin i suffers from a bias leading to undersampling, the entry n_{ij} will underestimate the contact frequency between bis i and j.

The problem of count matrix normalization has been thoroughly studied (1, 15, 16). In short, these methods apply a different normalization factor to each entry of the count matrix $[n_{ij}]$. Among them, matrix balancing can be used to construct a corrected count matrix $[\tilde{n}_{ij}]$ such that the number of interactions with other bins on the chromosome is the same for every bin. To be more accurate, matrix balancing yields two vectors $\mathbf{U}$ and $\mathbf{V}$ such that:

$$\begin{aligned} n_{ij} &= U_i \tilde{n}_{ij} V_i \\ N_c &= \sum_j \tilde{n}_{ij} \quad = \quad \sum_i \tilde{n}_{ij} \ . \end{aligned} \tag{13}$$

The matrix of contact probabilities is then computed as: $c_{ij} = \tilde{n}_{ij}/N_c$. The resulting matrix, $[c_{ij}]$, is bistochastic: each row and column sums to one.

As pointed out in (1), the problem of matrix balancing has been well studied. In particular, an efficient algorithm is available to balance any non-negative matrix with total support (17). Other implementations of matrix balancing dedicated to Hi-C data sets are also available (see for instance (18)).

In this article, we considered the contact probability matrix obtained by matrix balancing for the Hi-C data coming from (1). The normalized matrices, using the algorithm from (17), were readily available.

4.2. GAM

Genome Architecture Mapping (GAM) is a recent experimental technique which has been proposed as an alternative to the Hi-C technique to collect information on chromosome architecture (2). The procedure may be summarized as follows:

1. Collect slices of a cell population by cryosectioning.
2. Sequence DNA contained in each slice.
3. Map reads to genomic coordinates by aligning to a reference genome.
4. Assign genomic coordinates to bins corresponding to a regular subdivision of the genome.

Each slice collected contains thin layers of many nuclei with random orientations. Such a slice is represented in Fig. S30. Let us stress that a pair of DNA sequences detected in the same slice are not necessarily in contact. However, given that cells have been sliced in different orientations, if this pair is repeatedly found in the same slices, it means that these sequences belong to regions of the chromosome with a high contact probability. We now present the method used in this article to infer contact probabilities c_{ij} from the GAM experimental data.

The main output of GAM experiments is a segregation matrix $[s_{ia}]$ in which: rows correspond to bins on the genome, columns correspond to slices collected and each entry $s_{ia} = 1$ if bin i was detected in slice a and $s_{ia} = 0$ otherwise. Assuming that there are P slices, we define following reference (2):

- The segregation frequency for bin i:

$$f_i = \frac{1}{P}\sum_{a=1}^{P} s_{ia}. \tag{14}$$

- The co-segregation frequency for bins i and j:

$$f_{ij} = \frac{1}{P}\sum_{a=1}^{P} s_{ia}s_{ja}. \tag{15}$$

We now relate the segregation and co-segregation frequencies to actual contact probabilities. The probability that bins i and j are detected in a slice S_a (*i.e.* f_{ij}) can be decomposed according to the law of total probability as:

$$\begin{aligned}\Pr(i \text{ and } j \text{ in } S_a) &= \Pr(i \text{ and } j \text{ in } S_a | i \text{ and } j \text{ in contact})\Pr(i \text{ and } j \text{ in contact}) \\ &+ \Pr(i \text{ and } j \text{ in } S_a | i \text{ and } j \text{ not in contact})\Pr(i \text{ and } j \text{ not in contact})\end{aligned} \tag{16}$$

The probability that bins i and j are detected in a slice, conditioned to the fact that they are in contact (first term in the right hand side of the previous equation), is the probability that at least one of the bins is detected in the slice. Therefore, the previous expression is expressed in terms of the segregation frequencies, co-segregation frequencies and contact probabilities as:

$$f_{ij} = (1 - (1 - f_i)(1 - f_j))c_{ij} + f_i f_j (1 - c_{ij}). \tag{17}$$

We finally obtain for the contact probability between bins i and j:

$$c_{ij} = \frac{f_i f_j - f_{ij}}{f_i + f_j - 2 f_i f_j} \tag{18}$$

In this article, we have used the above equation to estimate the contact probability matrix from the experimental segregation matrix. Actually, Eq. 18 ensures that $c_{ij} < 1$. However, the nominator can be negative, in which case we set $c_{ij} \leftarrow \max(c_{ij}, 0)$.

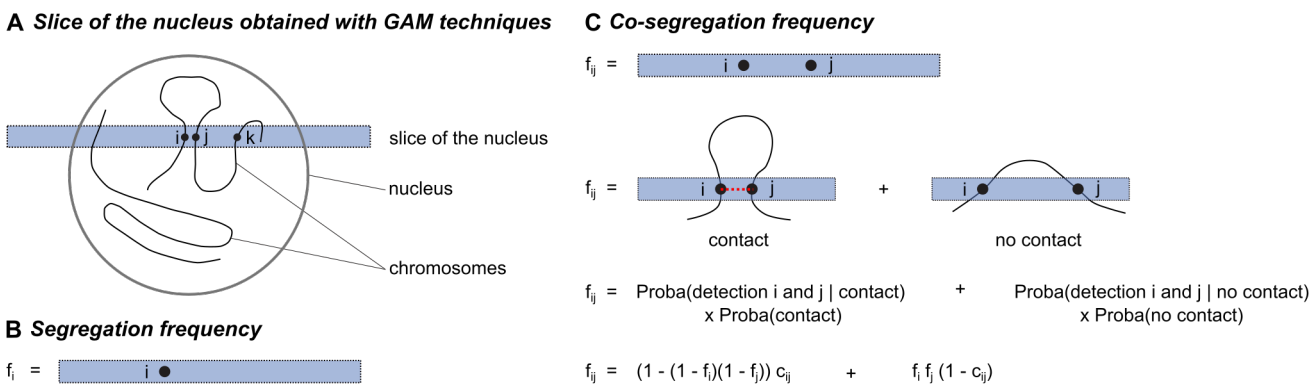

FIGURE S30 – Estimation of the contact probability matrix from GAM data sets. **(A)** GAM experiments use cryosectioning to obtain thin slices of a cell population sample. Each slice (or nuclear profile) cuts many nuclei in random orientations. The genomic content of each slice is sequenced. **(B)** The segregation frequency f_i is the fraction of nuclear profiles containing a specific genomic locus i. **(C)** The co-segregation frequency f_{ij} is the fraction of nuclear profiles containing a pair of specific loci i and j. The segregation and co-segregation frequencies can be used to estimate the contact probability of c_{ij} (see Eq. 18).

5. THE GAUSSIAN EFFECTIVE MODEL

5.1. Partition function

We consider the Gaussian effective model (GEM) with energy defined in the main text. To break the translational invariance, we attach the first monomer to the origin: $\mathbf{r}_0 = 0$. We can now write the GEM partition function as a Gaussian integral:

$$\begin{aligned} Z &= \int \prod_{i=1}^{N} \mathrm{d}^3\mathbf{r}_i \, \exp\left(-\beta U[\{\mathbf{r}_i\}]\right) \\ &= \int \prod_{i=1}^{N} \mathrm{d}^3\mathbf{r}_i \, \exp\left(-\frac{3}{2b^2} \sum_{i,j} \mathbf{r}_i \cdot \mathbf{r}_j \sigma_{ij}^{-1}\right), \end{aligned} \tag{19}$$

where we have introduced the inverse covariance matrix Σ^{-1} with elements σ_{ij}^{-1} and formally expressed as:

$$\Sigma^{-1} = T + W, \tag{20}$$

with:

$$T = \begin{pmatrix} 2 & -1 & \dots & 0 & 0 \\ -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}, \qquad W = \begin{pmatrix} \sum\limits_{\substack{j=0 \\ j\neq 1}} k_{1j} & -k_{12} & \dots & -k_{1N-1} & -k_{1N} \\ -k_{21} & \sum\limits_{\substack{j=0 \\ j\neq 2}} k_{2j} & \dots & -k_{2N-1} & -k_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -k_{N-11} & -k_{N-12} & \dots & \sum\limits_{\substack{j=0 \\ j\neq N-1}} k_{N-1j} & -k_{N-1N} \\ -k_{N1} & -k_{N2} & \dots & -k_{NN-1} & \sum\limits_{\substack{j=0 \\ j\neq N}} k_{Nj} \end{pmatrix}. \tag{21}$$

The partition function can be conveniently computed by separating the integration along each dimension:

$$\begin{aligned} Z &= \prod_{a=x,y,z} \left[\int \prod_{i=1}^{N} \mathrm{d}r_i^a \, \exp\left(-\frac{3}{2b^2} \sum_{i,j} r_i^a r_j^a \sigma_{ij}^{-1} \right) \right] \\ &= \left[\int \prod_{i=1}^{N} \mathrm{d}x_i \, \exp\left(-\frac{3}{2b^2} \sum_{i,j} x_i x_j \sigma_{ij}^{-1} \right) \right]^3 \\ &= z^3, \qquad \text{with } z = \left(\frac{2\pi b^2}{3} \right)^{N/2} \det \Sigma^{1/2}. \end{aligned} \tag{22}$$

For any function of the monomer coordinates, $A(\{\mathbf{r}_i\})$, we can therefore define the thermodynamical average:

$$\langle A(\{\mathbf{r}_i\}) \rangle = \frac{1}{Z} \int \prod_{i=1}^{N} \mathrm{d}^3\mathbf{r}_i \, A(\{\mathbf{r}_i\}) \exp\left(-\beta U[\{\mathbf{r}_i\}]\right). \tag{23}$$

5.2. Pair correlation function

Let us introduce the vector $\mathbf{r} = (r^x, r^y, r^z)$ and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$. The pair correlation function $\langle \delta(\mathbf{r} - \mathbf{r}_{ij}) \rangle$ can be expressed as:

$$\begin{aligned} \langle \delta(\mathbf{r} - \mathbf{r}_{ij}) \rangle &= \frac{1}{Z} \int \prod_{m=1}^{N} \mathrm{d}^3\mathbf{r}_m \, \delta(\mathbf{r} - \mathbf{r}_{ij}) \exp\left(-\frac{3}{2b^2} \sum_{m,n} \mathbf{r}_m \cdot \mathbf{r}_n \sigma_{mn}^{-1} \right) \\ &= \prod_{a=x,y,z} \underbrace{\left[\frac{1}{z} \int \prod_{m=1}^{N} \mathrm{d}r_m^a \, \delta(r^a - r_{ij}^a) \exp\left(-\frac{3}{2b^2} \sum_{m,n} r_m^a r_n^a \sigma_{mn}^{-1} \right) \right]}_{I(r^a)}, \end{aligned} \tag{24}$$

The integral $I(x)$ can be computed by exponentiating the δ-function:

$$\begin{aligned} I(x) &= \frac{1}{z} \int \prod_{m=1}^{N} \mathrm{d}x_m \int \frac{\mathrm{d}k}{2\pi} \exp\left(ik(x - x_{ij}) - \frac{3}{2b^2} \sum_{m,n} x_m x_n \sigma_{mn}^{-1} \right) \\ &= \frac{1}{z} \int \frac{\mathrm{d}k}{2\pi} \exp(ikx) \int \mathrm{d}^N X \, \exp\left(-\frac{3}{2b^2} X^T \Sigma^{-1} X - ikX^T E_{ij} \right), \end{aligned} \tag{25}$$

where the vector $E_{ij} = E_j - E_i$ and $E_i = (0, \cdots, 0, 1, 0, \cdots, 0)$, with the non-zero element being at the index i. By performing a first Gaussian integration we obtain:

$$I(x) = \int \frac{\mathrm{d}k}{2\pi} \exp(ikx) \exp\left(-\frac{b^2}{6} k^2 (\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})\right). \tag{26}$$

Finally, by performing a second Gaussian integration and by substituting this result into Eq. 24, we obtain the expression for the pair correlation function:

$$\langle \delta(\mathbf{r} - \mathbf{r}_{ij}) \rangle = \left(\frac{3}{2\pi \langle r_{ij}^2 \rangle}\right)^{3/2} \exp\left(-\frac{3}{2} \frac{r^2}{\langle r_{ij}^2 \rangle}\right), \tag{27}$$

where $\langle r_{ij}^2 \rangle = (\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}) b^2$.

5.3. Form factor dependent contact probability

The contact probability between monomers i and j can be expressed as:

$$\begin{aligned} c_{ij} &= \langle \mu(r_{ij}) \rangle, \\ &= \int \mathrm{d}^3\mathbf{r}\, \mu(r) \langle \delta(\mathbf{r}_{ij} - \mathbf{r}) \rangle, \end{aligned} \tag{28}$$

where $\mu(r)$ is a form factor. An intuitive choice of form factor is to consider a theta function:

$$\mu_T(r) = \theta(\xi - r). \tag{29}$$

In the context of Hi-C experiments, this is equivalent to consider that every restriction fragment pair separated by a distance $r < \xi$ can be cross-linked. Or in other words, the probability that restriction fragments separated by a distance r cross-link is

$$\Pr\left(\text{cross-link between i and j} \mid r_{ij} = r\right) = \begin{cases} 1 & \text{if } r < \xi \\ 0 & \text{otherwise .} \end{cases} \tag{30}$$

However, formaldehyde, the cross-linking agent used in most Hi-C experiments, can polymerize. It is present in aqueous solution in the form of methylene glycol $HOCH_2OH$ monomers, but it also exists in the form of oligomers $HO(CH_2O)_nH$, where n is a polymerization index. The equilibrium of the polymerization reaction depends on the formaldehyde concentration. For instance, in an aqueous solution with 40 % mass fraction of formaldehyde at 35 °C, the proportion of monomers in solution is only 26.80 %, the rest being oligomers with $n > 1$ (19, 20). This suggests that cross-links between restriction fragments have varying size depending on the formaldehyde oligomer that made the cross-link.

For that reason, the cross-linking probability may be more accurately represented by a function which ensures that most of the cross-links occur for distances $r < \xi$, but which also allows for few cross-links to occur when $r > \xi$. Based on these considerations, it seems natural to consider a Gaussian form factor:

$$\mu_G(r) = \exp\left(-\frac{3}{2} \frac{r^2}{\xi^2}\right), \tag{31}$$

or an exponential form factor:

$$\mu_E(r) = \exp\left(-\frac{r}{\xi}\right). \tag{32}$$

Let us emphasize that the form factor $\mu(r)$ is not a probability distribution function, so it does not need to be normalized. It should rather be considered as the probability for a Bernoulli random variable. For a pair of restriction fragments separated by a distance r, the probability to cross-link is $\mu(r)$ and the probability not to cross-link is $1 - \mu(r)$. Note that $\mu(0) = 1$.

5.4. Contact probabilities of the Gaussian effective model

From Eqs. 27 and 28, we can compute the contact probability c_{ij} for monomers i and j. Substituting $\mu(r)$ by the expression in Eqs. 29, 31 and 32 we obtain:

- For the Gaussian form factor:

$$\begin{aligned} c_{ij} &= F_G(\langle r_{ij}^2 \rangle) \\ &= \left(1 + \frac{\langle r_{ij}^2 \rangle}{\xi^2}\right)^{-3/2}, \end{aligned} \tag{33}$$

- For the theta form factor:

$$\begin{aligned} c_{ij} &= F_T(\langle r_{ij}^2 \rangle) \\ &= \mathrm{erf}\left(\frac{X}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} X \exp\left(-\frac{X^2}{2}\right), \qquad X = \sqrt{\frac{3\xi^2}{\langle r_{ij}^2 \rangle}}. \end{aligned} \tag{34}$$

  where we have introduced the standard error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}t \, e^{-t^2}. \tag{35}$$

- For the exponential form factor:

$$\begin{aligned} c_{ij} &= F_E(\langle r_{ij}^2 \rangle) \\ &= (1 + Y^2)\left(1 - \mathrm{erf}\left(\frac{Y^2}{2}\right)\right) \exp\left(\frac{Y^2}{2}\right) - Y\sqrt{\frac{2}{\pi}}, \qquad Y = X^{-1} = \left(\sqrt{\frac{3\xi^2}{\langle r_{ij}^2 \rangle}}\right)^{-1}. \end{aligned} \tag{36}$$

The functional dependence of the contact probability c_{ij} on the average square pair-distance $\langle r_{ij}^2 \rangle$ depends therefore on the choice of the form factor (Fig. S31).

5.5. Equilibrium properties

5.5.1. Radius of gyration

The radius of gyration of the GEM can be computed from the covariance matrix Σ. It has the expression:

$$\langle R_g^2 \rangle = \frac{1}{2(N+1)^2} \sum_{i,j=0}^{N} \langle r_{ij}^2 \rangle. \tag{37}$$

It can be used to characterize the swelling of the underlying polymer. For instance, we may monitor the ratio $\langle R_g^2 \rangle / \langle R_{g,0}^2 \rangle$ of the square radius of gyrations of the GEM with respect to the free Gaussian chain (all $k_{ij} = 0$).

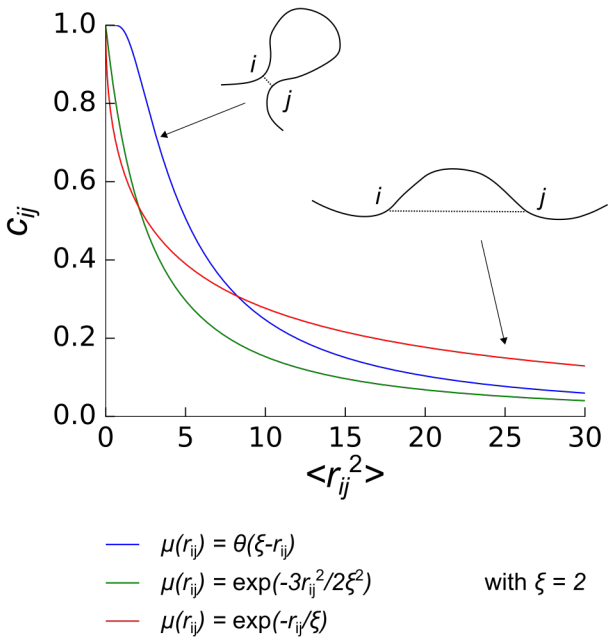

FIGURE S31 – In the GEM, the contact probability c_{ij} is expressed as a function of the mean square distance $\langle r_{ij}^2 \rangle$. This function depends on the form factor $\mu(r_{ij})$ used in the model.

5.5.2. Mean potentials of interaction

Other quantities of interest include the mean potentials of interaction at equilibrium. For any pair of monomers i and j, it is defined as:

$$\begin{aligned} \langle e_{ij} \rangle &= \frac{3}{2b^2} k_{ij} \langle r_{ij}^2 \rangle, \\ &= -\frac{\partial \ln Z}{\partial \ln k_{ij}}. \end{aligned} \tag{38}$$

The quantity defined in Eq. 38, expressed in $k_{\mathrm{B}}T$, reflects the state of the polymer. While high energy states are not favoured, they can however occur at thermal equilibrium if they are associated with large conformational entropy.

In addition, the mean potentials of interaction are extensive quantities. For instance, the mean potential of interaction between two groups $A = \{i_1, i_2, \ldots, i_M\}$ and $B = \{j_1, j_2, \ldots, j_{M'}\}$ of monomers is given by:

$$\langle e_{AB} \rangle = \sum_{(i,j) \in A \times B} e_{ij}. \tag{39}$$

5.6. Illustration

As an example, we considered an arbitrary coupling matrix $[k_{ij}]$, specifying the interactions for a polymer of $N + 1 = 100$ monomers. The coupling matrix was constructed by choosing randomly $M = 10$ pairs (i, j) and by assigning to each coupling a random number $k_{ij} = U$ between 0 and 1. Considering a Gaussian form factor with a threshold $\xi = 1.5$, we computed the contact probability of the GEM. We then sampled with Brownian Dynamics simulation configurations in the Boltzmann ensemble for this GEM. To compute the simulated contact probabilities, the average in Eq. 28 was carried over the sampled configurations. As can be seen in Fig. S32, the simulated contact probabilities converge to the model prediction when the number of sampled configuration increases.

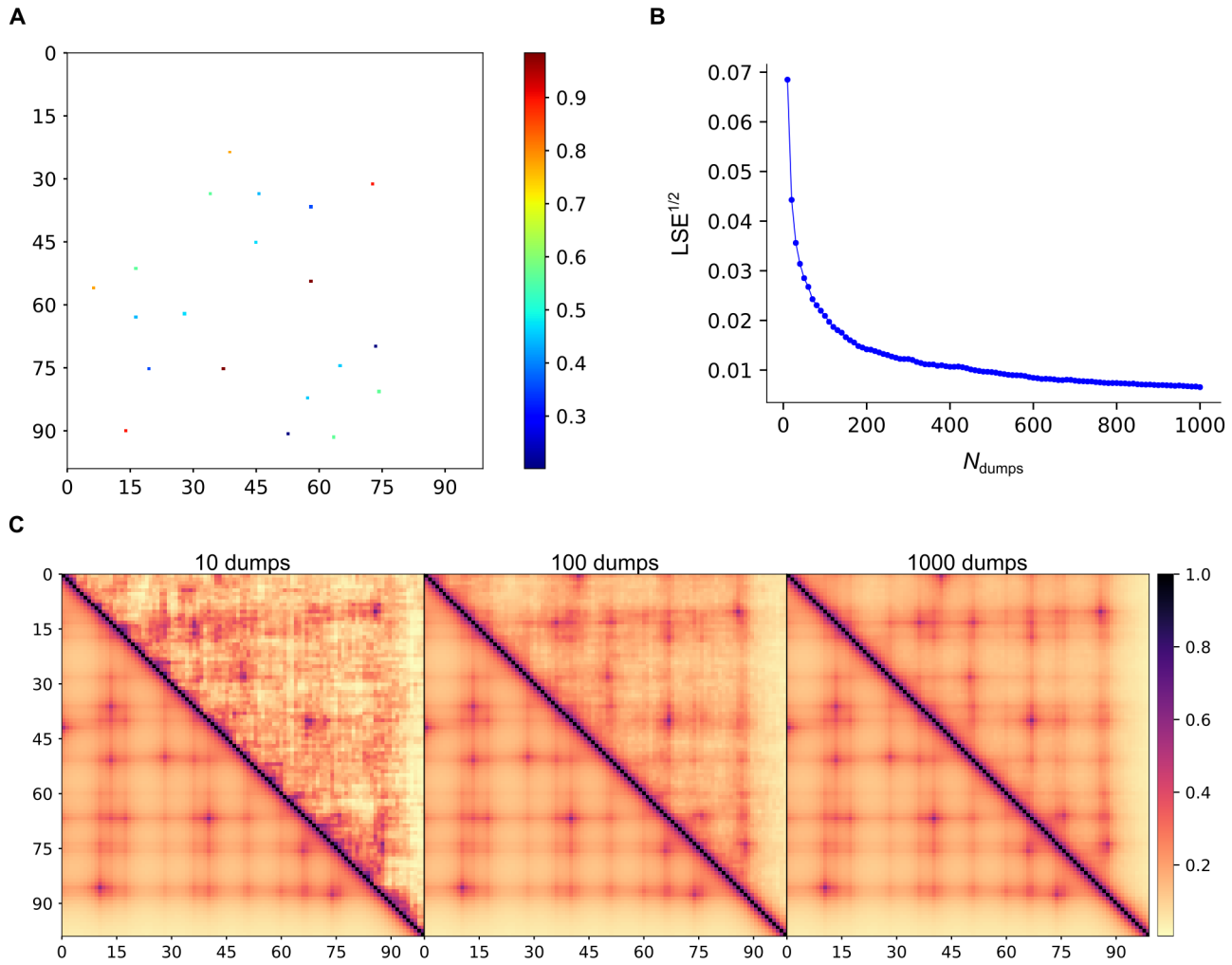

FIGURE S32 – **(A)** Arbitrary coupling matrix defining a GEM with $N + 1 = 100$ monomers. **(B)** Convergence of the contact probability matrix c_{ij}^{exp} computed from a Brownian Dynamics simulations to the GEM contact probability matrix c_{ij}, as a function of the number of sampled configurations (we used a threshold $\xi = 1.5$ and a Gaussian form factor). **(C)** Comparison of the contact probability matrices c_{ij}^{exp} and c_{ij}, for 10, 100 and 1000 configurations sampled by BD.

6. RECONSTRUCTION BY DIRECT MAPPING

6.1. Method

In section 5, we have shown that for any GEM, the matrix of contact probabilities is uniquely determined by the matrix of couplings. Reciprocally, for any contact probability matrix $[c_{ij}^{exp}]$ obtained from Hi-C experiments, one can reconstruct the GEM with the same contact probability matrix, $[c_{ij} = c_{ij}^{exp}]$, by computing the corresponding coupling matrix. This can be done as follows:

1. Compute the matrix of mean-square distances of the GEM, $[\langle r_{ij}^2 \rangle]$, using the relation:

$$\langle r_{ij}^2 \rangle = F^{-1}(c_{ij}), \tag{40}$$

   where F^{-1} is the inverse of one of the maps in Eqs. 33, 34 and 36.

2. Invert the covariance matrix $\Sigma = [\langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle]$ and compute the coupling matrix from Eqs. 20 and 21.

In this method, the threshold ξ used in the map F is a free parameter that needs to be adjusted. We chose ξ such that the Euclidean norm of the coupling matrix, $\|K\|$, is a minimum. This ensures that we select the GEM with the smallest perturbations compared to the free Gaussian chain case.

As an example, we have applied the reconstruction method by direct mapping to contact probability matrices computed from Brownian Dynamics trajectories of an arbitrary GEM. Namely, we simulated the GEM defined by the coupling matrix $[k_{ij}^{th}]$ in Fig. S32A. The experimental contact probability matrix were computed by carrying the thermodynamical average $c_{ij} = \langle \mu(r_{ij}) \rangle$ over the sampled configurations. We used a threshold $\xi^{exp} = 2$ and either a Gaussian or an exponential form factor. We therefore obtained two "artificial" contact probability matrices (see also Fig. S33):

Contact matrix	**Form factor**	ξ^{exp}
A	Gaussian	2.0
B	Exponential	2.0

In this specific scenario, the true coupling matrix is known, and we can therefore compute the distance between those couplings and the reconstructed ones by monitoring the quantity $\|K - K^{th}\|$. As can be seen in Fig. S33, both $\|K\|$ and $\|K - K^{th}\|$ are minimum for the same value of the threshold ξ so we use one or the other as proxies to determined the optimal value of the threshold, even when the true coupling matrix is not known or when the input contact probability matrix was not generated from a GEM.

Note that for contact matrix A, the optimal threshold is the same as the threshold used to compute the "experimental" contact probabilities, $\xi = \xi^{exp}$. This is because the form factors used for computing the "experimental" contact probabilities and for the reconstruction are both Gaussian. For matrix B, the form factor used to compute the "experimental" contact probabilities is exponential, and is therefore different from the Gaussian form factor used in the reconstruction. In this case, $\|K\|$ has several local minima. Yet at the global minimum, the coupling matrix is still reconstructed to a good accuracy.

6.2. Unphysical GEM and effect of the noise

In Fig. S33, there is a region where the reconstructed GEM has a covariance matrix Σ with negative eigenvalues. When this happens, the corresponding GEM has a non-finite free energy and does not represent a physical system. Unfortunately, when applying this reconstruction by direct mapping to contact probabilities obtained from Hi-C experiments (1, 21), this situation was almost systematic. It is therefore desirable to better understand under which conditions such instabilities occur. In particular, we may expect that Hi-C contact matrices contain some noise due

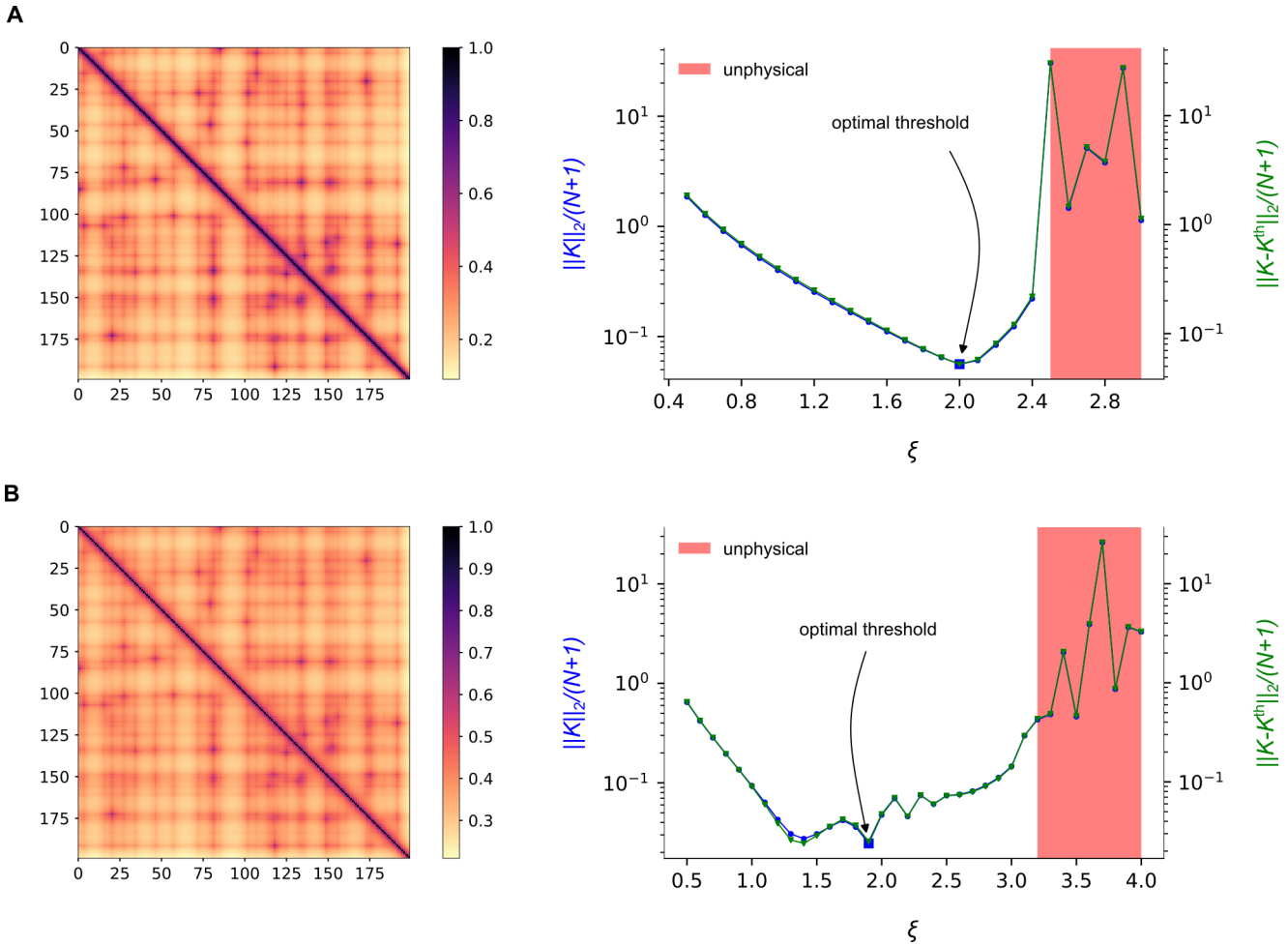

FIGURE S33 – Reconstruction of a GEM by direct mapping applied to an "artificial" contact probability matrix obtained from a Brownian Dynamics (BD) trajectory of an arbitrary GEM with a threshold $\xi^{exp} = 2$ and: **(A)** a Gaussian or **(B)** an exponential form factor. The reconstructed coupling matrix is very close to the original one used for the BD simulation. The red area denotes values of the threshold where the reconstructed GEM has a covariance matrix Σ with negative eigenvalues.

to inaccuracies in the measures or biases inherent to the experimental procedure, that lead to such effects.

Let us start from an artificial GEM with arbitrary couplings $K^{th} = [k_{ij}^{th}]$. We compute the associated contact matrix $[c_{ij}^{th}]$, using a threshold ξ^{th} and a form factor μ^{th}. When we perform Brownian Dynamics simulations of this system, we obtain configurations from which we compute the experimental contact matrix c_{ij}^{exp}, using a threshold ξ^{exp} and a form factor μ^{exp}. We take $\mu^{th} = \mu^{exp}$ as Gaussian form factors, and we chose $\xi^{exp} = 3.00$ to compute the experimental contact probabilities from Brownian Dynamics trajectories. Thermal fluctuations, together with the finite number of such configurations obtained from Brownian Dynamics simulations results in $c_{ij}^{exp} \neq c_{ij}^{th}$. We may therefore write the experimental contact probabilities as:

$$c_{ij}^{exp} = c_{ij}^{th} + \eta_{ij}, \tag{41}$$

where η_{ij} is a noise with unknown distribution, corrupting the "true" contact probabilities. For a chain with $N + 1 = 200$ monomers and $M = 20$ non-zeros couplings drawn from a uniform distribution in the interval $[0, 1]$, we computed the probability distribution function (pdf) of the difference $c_{ij}^{th} - c_{ij}^{exp}$. We tried different values for the threshold ξ^{th} used in the GEM mapping (Fig. S34) and obtained that when $\xi^{th} = \xi^{exp}$ the pdf of η_{ij} fits well a centered Gaussian distribution.

Consequently, instead of running Brownian Dynamics simulations in order to compute experimental contact

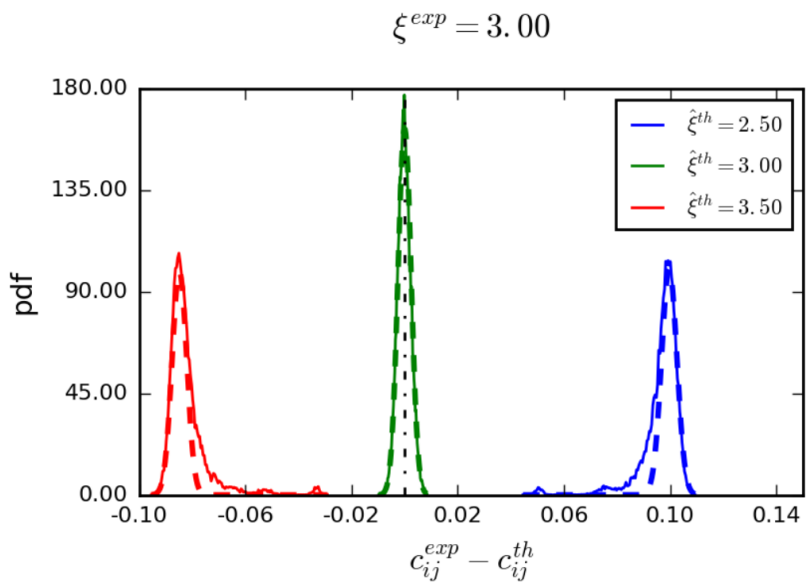

FIGURE S34 – Distribution of the noise $\eta_{ij} = c_{ij}^{exp} - c_{ij}^{th}$, fitted to a Gaussian distribution.

matrices c_{ij}^{exp}, we may construct pseudo-experimental contact matrices by adding a Gaussian noise with mean and variance given by

$$\langle \eta_{ij} \rangle = 0, \qquad \langle \eta_{ij}^2 \rangle = \varepsilon^2, \tag{42}$$

to the theoretical contact matrix c_{ij}^{th}. This trick allows us to investigate the stability of the reconstruction method by direct mapping as a function of the noise amplitude ε. Furthermore, it also allows us to explore more values for M than if we had to run systematically a Brownian Dynamics simulation.

Following this observation, we explored the stability of the reconstruction method by direct mapping in the (ε, M) plane. We considered a large size of polymer with $N + 1 = 1000$. For each value of M, we generated a random coupling matrix k_{ij}^{th} by drawing M random variables from a uniform distribution in the interval $[0, 1]$ and computed the theoretical contact probabilities c_{ij}^{th} of the corresponding GEM. Then we computed a pseudo-experimental contact probability matrix c_{ij}^{exp} by adding to the theoretical contact probabilities a centered Gaussian noise with standard deviation ε. Following our previous observation, we assumed that the contact probabilities obtained are a good approximation for the experimental contact probabilities that would be obtained by performing a Brownian Dynamics simulation of the GEM. Then we applied the reconstruction procedure to c_{ij}^{exp} using $\xi = \xi^{th}$. We therefore obtained a predicted GEM with couplings $[k_{ij}]$ that we compared to the theoretical couplings by computing the distance:

$$d(\hat{k}_{ij}, k_{ij}^{th}) = \frac{1}{(N+1)} \left[\sum_{ij} (\hat{k}_{ij} - k_{ij}^{th})^2 \right]^{1/2}, \tag{43}$$

The result of this analysis is shown in Fig. S35, in which we shaded in grey the region where the reconstructed couplings $[k_{ij}]$ define an unstable GEM with a correlation matrix Σ having negative eigenvalues. We observe that for each value of the number of constraints, M, there is an upper bound $\overline{\varepsilon}$ on the noise amplitude such that for $\varepsilon > \overline{\varepsilon}$, the direct reconstruction method fails, in the sense that the predicted GEM is unstable. It is remarkable that for $\varepsilon < \overline{\varepsilon}$ the direct reconstruction methods perform very well, with $d(\hat{k}_{ij}, k_{ij}^{th}) \lesssim 10^{-2}$ in the worse cases. Therefore, the reconstruction by direct mapping appears to be robust to noise until some critical value of the noise amplitude is reached. Then the method suddenly starts to fail. We also note that the value of $\overline{\varepsilon}$ seems to depend on the number of constraints of the underlying GEM. In particular, it is clear that the performances of the direct reconstruction method

get worse when $M \to 0$. Specifically, for $M = 0$, we observe that even blurring the theoretical contacts with a noise of very small amplitude is sufficient to make the reconstruction fail. On the contrary, the value of $\overline{\varepsilon}$ seems to be maximum for a number of constraints in a range between $M = 0.1N$ and $M = N$.

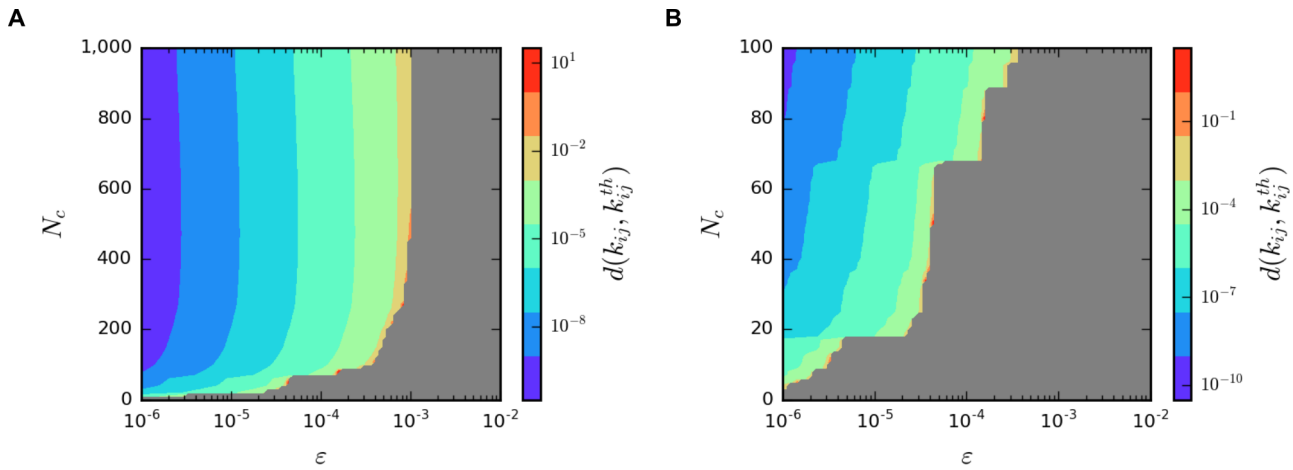

FIGURE S35 – Performance of the direct reconstruction method when the theoretical contact probabilities c_{ij}^{th} are blurred with a Gaussian noise such that $\langle \eta_{ij} \rangle = 0$ and $\langle \eta_{ij}^2 \rangle = \varepsilon^2$. We used $N + 1 = 1000$. The region in which the predicted couplings $[k_{ij}]$ define an unstable GEM was shaded in grey. **(A)** $M = 0, \ldots, 1000$. **(B)** Zoom for $M = 0, \ldots, 100$.

7. RECONSTRUCTION BY LSE MINIMIZATION

7.1. Steepest descent approach

As emphasized in the main text, some coupling matrices can lead to an unstable GEM. More precisely, the covariance matrix Σ has negative eigenvalues, so that it does not define a physically admissible model. In order to restrain our study to admissible GEMs, we have used a minimization scheme to find the admissible GEM reproducing as closely as possible an experimental contact probability matrix. The function to minimize is:

$$J(K) = \frac{1}{2}\|C(K) - E\|^2, \tag{44}$$

where the matrix $C(K) = [c_{ij}]$ is the matrix of contact probabilities of the Gaussian effective model, and $E = [e_{ij}]$ is the matrix of experimental contact probabilities. The contact probability matrix $C(K)$ is a function of the matrix of couplings $K = [k_{ij}]$. Note that C, E and K are indexed with $0 \leq i, j \leq N$, *i.e.* they are $(N+1) \times (N+1)$ matrices. Here, we used the Frobenius norm, such that for any matrix A, $\|A\|^2 = Tr(A^T A) = \sum_{i,j} a_{ij}^2$.

In order to minimize J as a function of K, under the constraint $K \geq 0$ (*i.e.* all k_{ij} are positive), we implemented a steepest descent method with projection (Fig. S36). At each iteration n, the matrix of couplings K^n is updated according to:

$$K' = K^n - h\frac{\nabla J^n}{\|\nabla J^n\|}, \tag{45}$$

$$K^{n+1} = p_{\mathbb{R}_+}(K'), \tag{46}$$

where the scalar h is a small time step, and the projection operator $p_{\mathbb{R}_+}$ applies the operation $x \leftarrow \max(x, 0)$ to all entries of its matrix argument. In practise, the time step was adjusted at each iteration. Namely, if $J^{n+1} > J^n$, then we decreased the time step according to: $h \leftarrow 0.1 \times h$. Otherwise, we increased h for the next iteration according to $h \leftarrow 2 \times h$.

We stopped the minimization when the relative variation in the cost function became sufficiently small:

$$\frac{2|J^{n+1} - J^n|}{|J^{n+1}| + |J^n|} < \varepsilon_r, \tag{47}$$

with typically $\varepsilon_r = 1 \times 10^{-9}$.

The minimization scheme that we just described requires to compute the gradient as a function of the k_{ij} variables.

7.2. Expression of the gradient of the least-square estimator

We will express J as the composition of several maps, and then use rules of differential calculus to find its differential form dJ. Since J takes scalar values, we will then find its gradient as the matrix such that: $dJ_K(H) = Tr(\nabla J(K)^T H)$.

Let us first consider the matrix of reduced couplings $W = [w_{pq}]$, as defined in Eq. 21, which is indexed with $1 \leq p, q \leq N$. We may introduce the linear map $\mathcal{A}$ which transforms a coupling matrix in its reduced coupling matrix:

$$\begin{array}{rccc} \mathcal{A}: & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} & \rightarrow & \mathbb{R}^N \times \mathbb{R}^N \\ & K & \mapsto & \mathcal{A}(K) = W. \end{array} \tag{48}$$

Actually, the matrix elements of the reduced couplings can be expressed as:

$$w_{pq} = \sum_{i,j} a_{pqij} k_{ij}, \tag{49}$$

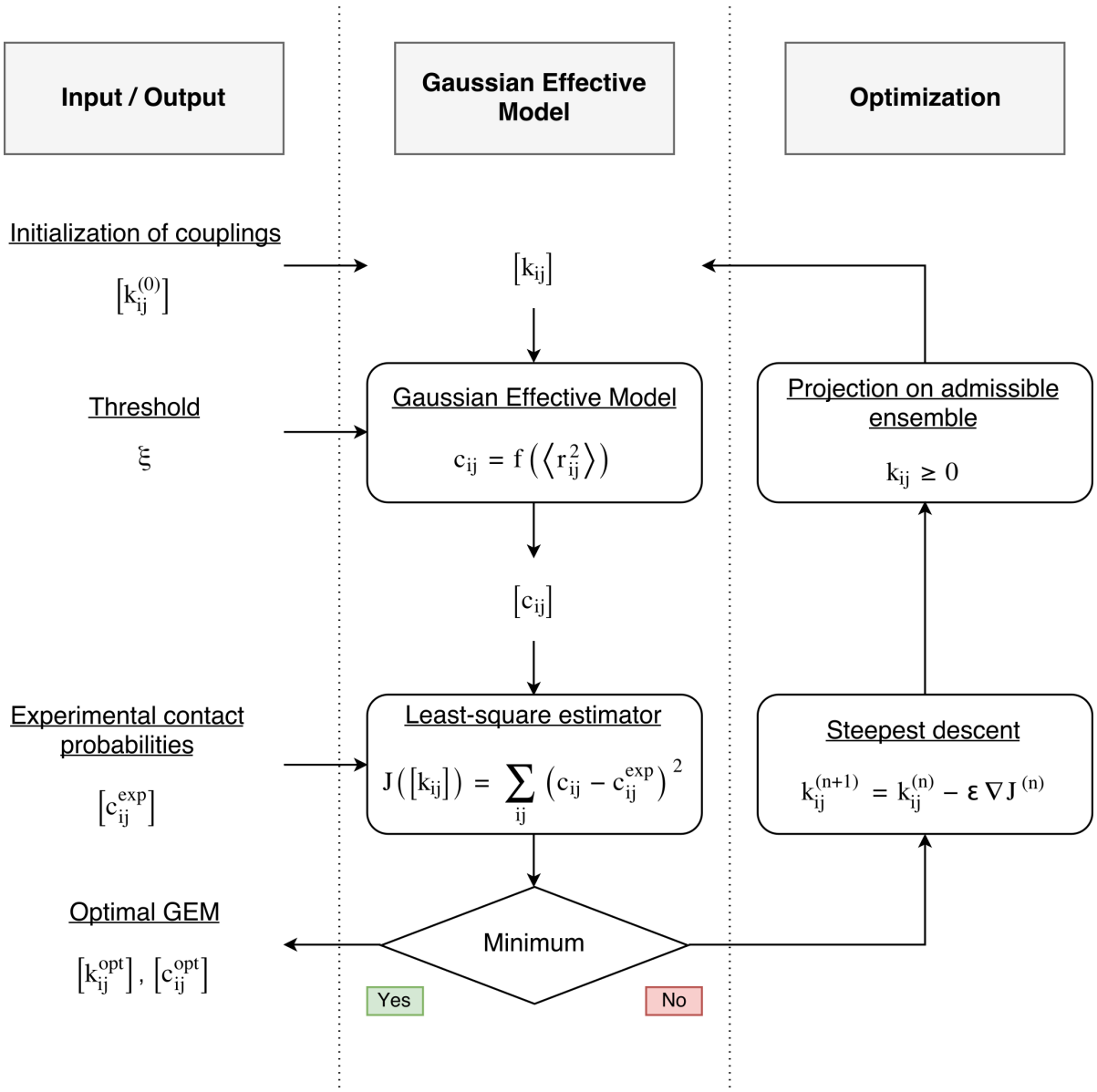

FIGURE S36 – Minimization algorithm to find the Gaussian effective model with the closest contact probability matrix to an experimental contact probability matrix.

where:

$$a_{pqij} = \begin{cases} (1-\delta_{pq})\left[-\frac{\delta_{pi}\delta_{qj}+\delta_{pj}\delta_{qi}}{2}\right] + \delta_{pq}\left[\frac{\delta_{pi}+\delta_{pj}}{2}\right] & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{50}$$

Here, $\delta_{pq} = 1$ if $p = q$ and $\delta_{pq} = 0$ otherwise. The previous expression ensures that W is a symmetrical matrix. The expression obtained suggests to introduce the tensor $\underline{A} = [a_{pqij}]$ and to use the matrix-vector notation:

$$\mathcal{A}(K) = \underline{A}K, \tag{51}$$

where K is seen at a vector of $\mathbb{R}^{2(N+1)}$ and $\underline{A}$ as a matrix of $\mathbb{R}^{2(N+1)} \times \mathbb{R}^{2(N+1)}$. The differential of $\mathcal{A}$ is expressed as:

$$\begin{array}{rccc} d\mathcal{A}_K: & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} & \rightarrow & \mathbb{R}^N \times \mathbb{R}^N \\ & H & \mapsto & d\mathcal{A}_K(H) = \underline{A}H. \end{array} \tag{52}$$

Actually, we may define the map associating to any coupling matrix the associated inverse covariance matrix $\Sigma^{-1} = \tilde{\mathcal{A}}(K) = \mathcal{A}(K) + T$ of a GEM, with T as in Eq. 21. It is straightforward that $d\tilde{\mathcal{A}} = d\mathcal{A}$.

Next, following Eq. 20, we can express the covariance matrix as $\Sigma = \mathcal{I}(W + T)$, where we introduced the inversion operator:

$$\begin{array}{rccc} \mathcal{I}: & \mathbb{R}^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^N \times \mathbb{R}^N \\ & X & \mapsto & \mathcal{I}(X) = X^{-1}. \end{array} \tag{53}$$

The differential of $\mathcal{I}$ at the matrix X is:

$$\begin{array}{rccc} d\mathcal{I}_X: & \mathbb{R}^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^N \times \mathbb{R}^N \\ & H & \mapsto & d\mathcal{I}_X(H) = -X^{-1}HX^{-1}. \end{array} \tag{54}$$

Then, we introduce the matrix of mean square distances $\Gamma = [\gamma_{ij}]$ of a GEM, with $\gamma_{ij} = \langle r_{ij}^2 \rangle$, indexed with $0 \leq i, j \leq N$. By definition, it is related to the matrix of covariance $\Sigma = [\sigma_{pq}]$:

$$\begin{array}{rcccl} \gamma_{ij} & = & \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij} & & \text{for } 0 < i, j \leq N, \\ \gamma_{0j} & = & \sigma_{jj} & & \text{for } 0 < j \leq N. \end{array} \tag{55}$$

We now introduce the map:

$$\begin{array}{rccc} \mathcal{B}: & \mathbb{R}^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \\ & \Sigma & \mapsto & \mathcal{B}(\Sigma) = \Gamma. \end{array} \tag{56}$$

Similarly as before, we may express this map in a matrix-vector notation, $\mathcal{B}(\Sigma) = \underline{B}\Sigma$, where the tensor $\underline{B}$ has the elements:

$$b_{ijpq} = \left(\delta_{ip} + \delta_{jq}\right)\delta_{pq} - 2\delta_{ip}\delta_{jq}. \tag{57}$$

The differential of $\mathcal{B}$ in Σ is then expressed as:

$$\begin{array}{rccc} d\mathcal{B}_\Sigma: & \mathbb{R}^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \\ & H & \mapsto & d\mathcal{B}_\Sigma(H) = \underline{B}H. \end{array} \tag{58}$$

The final step of the Gaussian effective model mapping is to express the matrix of contact probabilities C as a function of Γ. To this end, we introduce the map:

$$\begin{array}{rccc} \mathcal{F}: & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} & \rightarrow & \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \\ & \Gamma & \mapsto & \mathcal{F}(\Gamma) = C. \end{array} \tag{59}$$

In the previous expression, the matrix elements of C are given by:

$$c_{ij} = F(\gamma_{ij}), \tag{60}$$

where F is one of Eqs. 33, 34 and 36, depending on the form factor used. We can then identify the differential of $\mathcal{F}$ by performing an expansion around Γ. We obtain:

$$\begin{array}{rccc} d\mathcal{F}_\Gamma: & \mathbb{R}^{N+1}\times\mathbb{R}^{N+1} & \rightarrow & \mathbb{R}^{N+1}\times\mathbb{R}^{N+1} \\ & H & \mapsto & d\mathcal{F}_\Gamma(H) = F'(\Gamma)\circ H, \end{array} \tag{61}$$

where we introduced the Hadamard product such that for any two matrices $(A\circ B) = [a_{ij}b_{ij}]$, and the short-hand notation $F'(\Gamma) = [F'(\gamma_{ij})]$.

Finally, we introduce the linear form:

$$\begin{array}{rccc} \mathcal{G}: & \mathbb{R}^{N+1}\times\mathbb{R}^{N+1} & \rightarrow & \mathbb{R} \\ & C & \mapsto & \mathcal{G}(C) = \frac{1}{2}\|C-E\|^2. \end{array} \tag{62}$$

By definition of the Frobenius norm, $\|A\|^2 = Tr(A^T A)$, we obtain for the differential of $\mathcal{G}$ in C:

$$\begin{array}{rccc} d\mathcal{G}_C: & \mathbb{R}^{N+1}\times\mathbb{R}^{N+1} & \rightarrow & \mathbb{R} \\ & H & \mapsto & d\mathcal{G}_C(H) = Tr\left[(C-E)^T H\right]. \end{array} \tag{63}$$

In summary, we have introduced several maps and expressed the cost function to minimize as $J(K) = \mathcal{G}\circ\mathcal{F}\circ\mathcal{B}\circ\mathcal{I}\circ\tilde{\mathcal{A}}(K)$. Using the rules of composition for differential calculus, we obtain the differential of J in K:

$$\begin{aligned} dJ_K(H) &= d\mathcal{G}_{\mathcal{F}\circ\mathcal{B}\circ\mathcal{I}\circ\tilde{\mathcal{A}}(K)}\circ d\mathcal{F}_{\mathcal{B}\circ\mathcal{I}\circ\tilde{\mathcal{A}}(K)}\circ d\mathcal{B}_{\mathcal{I}\circ\tilde{\mathcal{A}}(K)}\circ d\mathcal{I}_{\tilde{\mathcal{A}}(K)}\circ d\mathcal{A}_K(H), \\ &= Tr\left[\nabla J(K)^T H\right], \end{aligned} \tag{64}$$

After calculations, the gradient of J in K reads:

$$\begin{array}{rcl} \nabla J(K) & = & -\underline{A}^*\left[(X^{-1})^T Y (X^{-1})^T\right], \\ \text{with:} & & \\ X & = & \underline{A}K + T \\ Y & = & \underline{B}^*\left[(C-E)\circ F'(\Gamma)\right]. \end{array} \tag{65}$$

To obtain the last expression, we introduced the adjoint tensors $\underline{A}^* = [a^*_{ijpq} = a_{pqij}]$ and $\underline{B}^* = [b^*_{pqij} = b_{ijpq}]$. Or writing explicitely all the summations we have:

$$\begin{array}{rcl} \frac{\mathrm{d}J}{\mathrm{d}k_{ij}} & = & -\sum\limits_{m,n=1}^{N} a_{mn,ij}\sum\limits_{p,q=1}^{N} x^{-1}_{pm}y_{pq}x^{-1}_{nq}, \\ \text{with:} & & \\ y_{pq} & = & \sum\limits_{k,l=0}^{N} b_{kl,pq}(c_{kl}-e_{kl})F'(\gamma_{kl}). \end{array} \tag{66}$$

7.3. Computational burden

The main computational burden in evaluating the cost function J as well as its gradient ∇J resides in the matrix inversion $\Sigma^{-1}\rightarrow\Sigma$, with $O(N^3)$ complexity. In this work, we have used the routines of the Intel®Math Kernel Library to perform the algebra operations and the matrix inversion. We used the parallel implementation to distribute the computation over 12 processors.

As an alternative to the cost function in Eq. 44, we have also considered minimizing:

$$J' = \frac{1}{(N+1)^2}\sum_{i,j}\left(\sum_k \sigma^{-1}_{ik}s_{kj} - \delta_{ij}\right)^2, \tag{67}$$

where $[s_{ij}]$ is the covariance matrix of the GEM reproducing exactly the experimental contacts $[c_{ij}^{exp}]$, and $[\sigma_{ij}]$ is the covariance matrix of a candidate (stable) GEM with couplings k_{ij}. The advantage of this form over the previous one is that it does not require any matrix inversion. More accurately, it is a quadratic function of the k_{ij} variables. Therefore the existence of a minimum satisfying $k_{ij} \geq 0$ is guaranteed and it is unique. Consequently, it is less computationally intensive and the minimum can be found efficiently with descent methods using conjugate directions. We found this form to work very well with contact probability matrix generated from predefined GEM by Brownian Dynamics simulations. However, for Hi-C contact probabilities, we found that it was much less successful. More precisely, the least-square estimator between the contact probabilities of the Hi-C experiment and of the optimal model was not as low.

8. BROWNIAN DYNAMICS

8.1. Physical model

In this article, we have performed two types of Brownian Dynamics simulation. The potentials used for each of them are summarized in the following table and discussed in further details below.

Potential	**BD of GEM**	**BD of GEM with semi-flexibility and excluded volume**
Chain structure	U_e	U_{fene}
GEM couplings	U_I	U_I
Bending rigidity	-	U_b
Excluded Volume	-	U_{ev}
Total	$U_e + U_I$	$U_{fene} + U_b + U_{ev} + U_I$

8.1.1. Chain structure

We modeled the chromosome as a beads-on-string polymer with monomers of size b and coordinates $\mathbf{r}_i$. The index varies between $i = 0$ and $i = N$. The bond vectors are $\mathbf{u}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$.

In the absence of excluded volume, we considered a Gaussian chain for the polymer structure, with potential:

$$\beta U_e\left[\{\mathbf{r}_i\}\right] = \frac{3}{2b^2}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{i-1})^2. \tag{68}$$

An important property of Gaussian chains is that the mean-square value of the end-to-end vector $\mathbf{R}_e = \mathbf{r}_N - \mathbf{r}_0$ scales linearly with the contour length:

$$\langle R_e^2\rangle = b^2 N. \tag{69}$$

In reality, approximating a polymer to a Gaussian chain is only valid for weak perturbations, $R_e \ll Nb$. Besides, a Gaussian polymer allows the bond distance to fluctuate quite a lot ($\langle u_i^2\rangle = b^2$). This is problematic in Brownian Dynamics simulations with excluded volume interactions because this would result in possible crossings between different bonds. Therefore, for Brownian Dynamics with excluded volume interactions, we have preferred instead the finitely-extensible non-linear elastic potential (FENE):

$$U_{fene}\left[\{\mathbf{r}_i\}\right] = -\frac{3k_e r_0^2}{2b^2}\sum_{i=1}^{N}\ln\left(1 - \frac{u_i^2}{r_0^2}\right), \tag{70}$$

where r_0 is a distance above which non-linear effects start to appear in the bonds elasticity and k_e is the rigidity constant of the non-linear spring. Note that for $u_i \ll r_0$ we recover the Gaussian chain potential, *i.e.* a linear spring (with $k_e = 1\, k_\mathrm{B}T$). In practical applications we have taken $r_0 = 1.5\, b$ and $k_e = 10\, k_\mathrm{B}T$ (22).

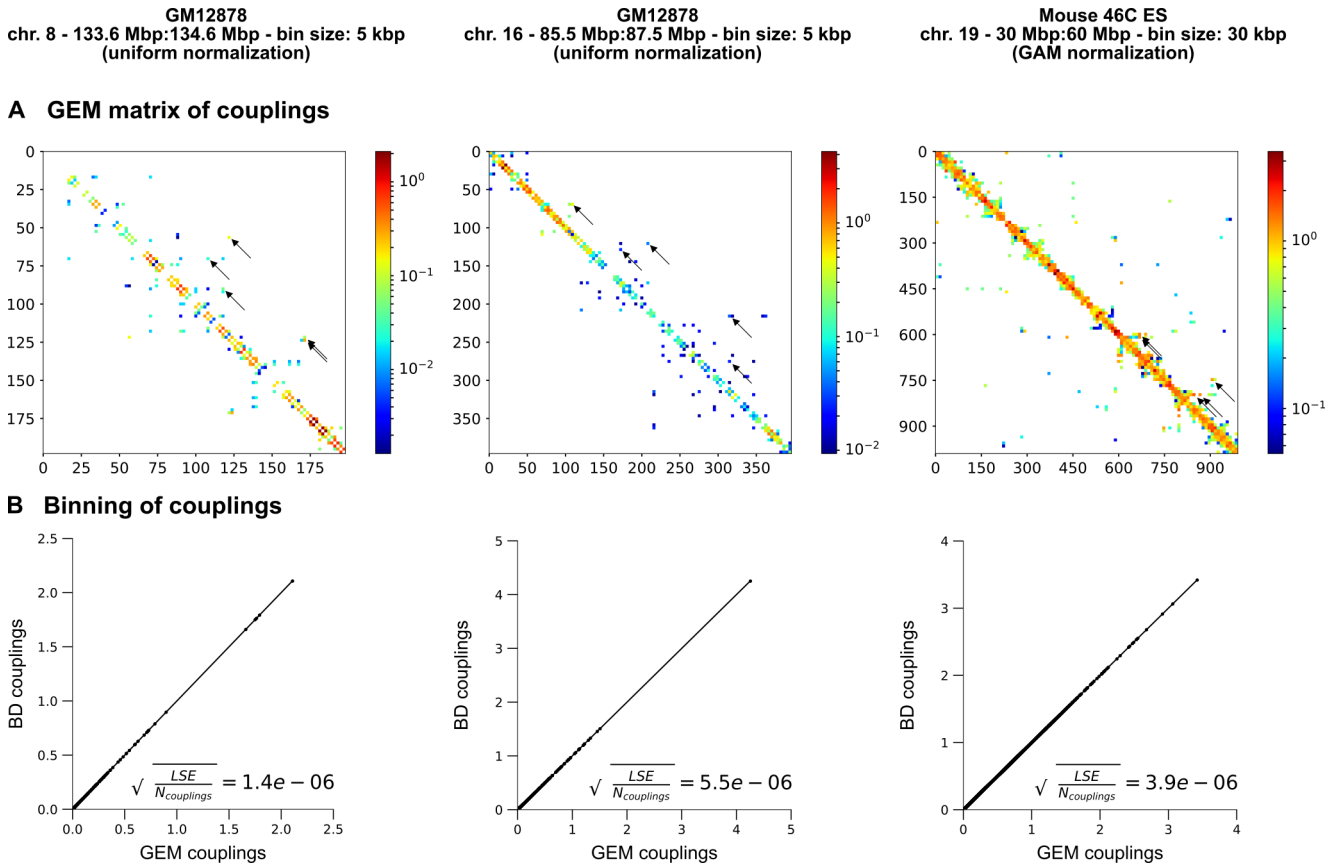

FIGURE S37 – Comparison of GEM couplings with binned couplings used in Brownian Dynamics (BD) simulations.

8.1.2. Gaussian effective model interactions

Following the model described in the main text, we introduced the GEM interaction potential:

$$\beta U_I\left[\{\mathbf{r}_i\}\right] = \frac{3}{2b^2} \sum_{0 \le i < j \le N} k_{ij} \left(\mathbf{r}_i - \mathbf{r}_j\right)^2, \tag{71}$$

where the k_{ij} are the couplings from a GEM. In order to have a reasonable amount of distinct couplings values in the implementation of BD simulations, we binned the GEM couplings. Specifically, we considered 1000 bins of same length in the interval $[k_{min}, k_{max}]$ where k_{min} (resp. k_{max}) is the minimum (resp. maximum) of the reconstructed GEM couplings. Note that we discarded all couplings $k_{ij} < 0.001$. Despite this binning procedure, the couplings used in the BD simulations remained very close to the reconstructed GEM ones (see Fig. S37).

8.1.3. Bending rigidity

In reality, the DNA fiber opposes a certain resistance to bending. To model this effect, we used a Kratky-Porod potential:

$$\beta U_b\left[\{\mathbf{r}_i\}\right] = l_p \sum_{i=1}^{N-1} \left(1 - \cos\theta_i\right), \tag{72}$$

where θ_i is the angle between bonds $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$.

For a polymer with a Gaussian chain potential plus a bending rigidity potential as defined above, the linear

scaling of the mean-square end-to-end distance with the contour length still holds:

$$\langle R_e^2 \rangle \approx l_K^2 \frac{N}{g_K}, \tag{73}$$

where $l_K = 2l_p$ is the Kuhn length and g_K is the number of original monomers per Kuhn length. Thus a semi-flexible polymer behaves like a Gaussian chain at large scales, with $N' = N/g_K$ and $b' = l_K$.

8.1.4. Excluded volume

A commonly used two-parameter empirical form for describing non-bonded interactions between two neutral (but possibly polarized) particles is the Lennard-Jones, or "6-12", potential. For two monomers separated by a distance r, it reads:

$$V_{LJ}(r) = 4\varepsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right), \tag{74}$$

where ε is an energy scale in $k_\mathrm{B}T$ and σ is the hard core distance. Here, the interaction still decays as a power law of the distance r. A standard method to make this interaction short-range, is to introduce a threshold r^{th} such that for distances $r > r^{th}$ the interaction vanishes. Therefore, in simulations, we have considered the truncated Lennard-Jones potential:

$$U_{ev}(r) = \begin{cases} V_{LJ}(r) - V_{LJ}(r^{th}) & \text{if } r < r^{th}, \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

We have considered take $\varepsilon = 1\, k_\mathrm{B}T$, but the hard-core distance may be different from the monomer size (see next below). To model excluded volume interactions, we set $r^{th} = 2^{1/6}\sigma$, resulting in $U_{ev}(r) > 0$ for $r < r^{th}$. In particular, this ensures that the repulsive force, $-\partial U_{ev}/\partial r$, vanishes precisely for $r = r^{th}$.

8.1.5. Numerical values

In eukaryotes, the interphase chromosome is packed into a fiber with a diameter of 30 nm, which is usually designated as chromatin. It has a linear packing fraction $\nu \approx 100\,\mathrm{bp\,nm^{-1}}$ and persistence length $l_p = 90\,\mathrm{nm}$ (23). Therefore, the appropriate size for monomers is $\sigma = 30\,\mathrm{nm}$, which correspond to $g = 3000\,\mathrm{bp}$. The persistence length expressed in units of these monomers gives $l_p = 3\sigma$, and σ is also the hard-core distance for excluded volume interactions between monomers.

In the Brownian Dynamics simulations performed in this article, the natural unit of monomer is the Hi-C bin resolution. We have considered specifically $g_1 = 5000\,\mathrm{bp}$ and $g_2 = 30\,000\,\mathrm{bp}$ with corresponding monomer sizes b_1 and b_2. The persistence lengths for each case thus read $l_p = \nu l_p/g_1\, b_1 = 1.8\, b_1$ and $l_p = \nu l_p/g_2\, b_2 = 0.3\, b_2$.

For the first resolution, we may consider that $g_1 \approx g$, meaning that monomers can be represented as impenetrable beads. We thus take for the hard-core distance $\sigma_1 = b_1$. The second resolution however defines monomers much larger than the chromatin fiber diameter. Following the scaling relations introduced above, we may express the monomer sizes as:

$$b_2^2 \approx \frac{g_2}{g_K} l_K^2, \tag{76}$$

where $g_K = 18\,000\,\mathrm{bp}$ is the number of monomers per Kuhn length. We obtain that $b_2 \approx 8\sigma$. Therefore, we have considered a hard-core distance $\sigma_2 = 0.125\, b_2$.

We summarize in the following table the values of the different parameters we took for our Brownian Dynamics simulations.

Data set	GM12878 chromosome 8 133.6 Mbp:134.6 Mbp bin size: 5 kbp uniform normalization	GM12878 chromosome 16 85.5 Mbp:87.5 Mbp bin size: 5 kbp uniform normalization	Mouse 46C ES chromosome 19 30 Mbp:60 Mbp bin size: 30 kbp GAM normalization
Gaussian chain	$b = 1$	$b = 1$	$b = 1$
FENE chain	$b = 1$ $r_0 = 1.5$ $k_e = 10\,\mathrm{k_B T}$	$b = 1$ $r_0 = 1.5$ $k_e = 10\,\mathrm{k_B T}$	$b = 8$ $r_0 = 12$ $k_e = 10\,\mathrm{k_B T}$
Bending rigidity	$l_p = 1.8$	$l_p = 1.8$	$l_p = 2.4$
Excluded volume	$\sigma = 1$ $r_c = 1.1225$	$\sigma = 1$ $r_c = 1.1225$	$\sigma = 1$ $r_c = 1.1225$
GEM couplings	1000 equal sized bins $\min(k_{ij}) \geq 10^{-3}$	1000 equal sized bins $\min(k_{ij}) \geq 10^{-3}$	1000 equal sized bins $\min(k_{ij}) \geq 10^{-3}$

8.2. Implementation of Brownian Dynamics

Brownian dynamics simulations are molecular dynamics simulations in which many molecular details are coarse-grained. The classical framework to describe the Brownian motion of a particle is the Langevin equation. For a bead with coordinates $x(t)$ it reads:

$$m\ddot{x}(t) = -\gamma\dot{x} - \frac{\partial U}{\partial x}(x(t)) + \gamma\eta(t), \tag{77}$$

in which m is the mass of the bead, γ is a damping term and $-\partial U/\partial x$ is the force applied to the bead, deriving from a potential U. The first two terms in the right-hand side of the above dynamics are deterministic. In addition there is a stochastic term, $\eta(t)$ which represents energy exchanges between the bead and a bath at temperature T. More accurately, η is an uncorrelated Gaussian random process with two first moments:

$$\langle\eta(t)\rangle = 0, \qquad \langle\eta(t)\eta(t')\rangle = 2D\delta(t - t'), \tag{78}$$

where D is the diffusion coefficient of the bead. It can be shown that the above dynamics converges to the Boltzmann equilibrium provided that D satisfies the Stokes-Einstein relation:

$$D = k_\mathrm{B}T/\gamma, \tag{79}$$

where finally from the Stokes' law applied to a bead of diameter b we get $\gamma = 3\pi b\mu$, with μ being the fluid viscosity.

In order to produce Brownian Dynamics trajectories, the Langevin equation Eq. 77 was applied to each bead of our polymer model and integrated numerically with the LAMMPS simulation package (24). It uses a standard velocity Verlet integration scheme (25). In practise, this requires the choice of an integration time step, and we chose the value $dt = 0.001$. We also set $\gamma = 1$ (in simulation dimensionless units).

The choice of the initial configuration is important, especially when excluded volume is included. Although we can start from an arbitrary configuration respecting excluded volume constraints, the relaxation to the Boltzmann equilibrium can be very slow. To circumvent this problem and generate quickly an initial configuration for a polymer with excluded volume interactions we have used the following procedure.

First, perform a relaxation run without excluded volume nor short-range attractive interactions. This corresponds to the dynamics of an ideal chain and aims at sampling rapidly a large number of configurations to loose the memory of the initial condition.

Second, perform an intermediate run with few iterations (generally 10^6 iterations) with a soft pair potential:

$$U_{soft}(r) = A\left(1 + \cos\left(\frac{\pi r}{r^{th}}\right)\right), \tag{80}$$

where r^{th} is the same cutoff as in the truncated Lennard-Jones potential from Eq. 75. The magnitude A is progressively increased from 1 to 60 during the run (22), so that we obtain in the end a configuration with no overlaps between the beads.

Third, the main run with excluded volume and short-range interactions is performed starting from the configuration without overlaps. Several configurations (generally 1000) are extracted from the resulting trajectory, which sample the Boltzmann ensemble. These configurations can be used to compute equilibrium averages according to the ergodic property of the Boltzmann equilibrium.

It is possible to map the simulation time to the real time. Let us write the diffusion coefficient as $D = b^2/\tau_B$. During the time τ_B, a bead typically travels through a distance b, which is its own size. Consequently τ_B is the natural unit of time for this diffusive process and is called the Brownian time. In Brownian Dynamics simulations we take $b = 1$ and $D = 1$ (in dimensionless units), therefore a unit of simulation time correspond to the Brownian time. The diffusion coefficient in the bacterial nucleoid was found to be $D = 10\,\mu\text{m}^2\,\text{s}^{-1}$ (26). Therefore, for $b = 30\,\text{nm}$ we find $\tau_B = 90\,\mu\text{s}$.

SUPPORTING REFERENCES

1. Rao, S. S., 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. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159:1665–1680.

2. Beagrie, R. A., A. Scialdone, M. Schueler, D. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. d. de Santiago, L.-M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. Edwards, M. Nicodemi, and A. Pombo. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543:519–524.

3. Serra, F., M. D. Stefano, Y. G. Spill, Y. Cuartero, M. Goodstadt, D. Baù, and M. A. Marti-Renom. (2015). Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Letters* 589:2987–2995.

4. Umbarger, M. A., E. Toro, M. A. Wright, G. J. Porreca, D. Baù, S.-H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker, and G. M. Church. (2011). The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. *Molecular Cell* 44:252–264.

5. Baù, D., and M. A. Marti-Renom. (2012). Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods* 58:300–306.

6. Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. (2010). A three-dimensional model of the yeast genome. *Nature* 465:363–367.

7. Lesne, A., J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci. (2014). 3D genome reconstruction from chromosomal contacts. *Nature Methods* 11:1141–1143.

8. Marbouty, M., A. Le Gall, D. I. Cattoni, A. Cournac, A. Koh, J.-B. Fiche, J. Mozziconacci, H. Murray, R. Koszul, and M. Nollmann. (2015). Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Molecular Cell* 59:588–602.

9. Jhunjhunwala, S., M. C. van Zelm, M. M. Peak, S. Cutchin, R. Riblet, J. J. van Dongen, F. G. Grosveld, T. A. Knoch, and C. Murre. (2008). The 3D Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions. *Cell* 133:265–279.

10. Jost, D., P. Carrivain, G. Cavalli, and C. Vaillant. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Research* 42:9553–9561.

11. Brackley, C. A., J. M. Brown, D. Waithe, C. Babbs, J. Davies, J. R. Hughes, V. J. Buckle, and D. Marenduzzo. (2016). Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biology* 17:59.

12. Giorgetti, L., R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard. (2014). Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell* 157:950–963.

13. de Gennes, P. (1979). Scaling Concepts in Polymer Physics. Cornell University Press.

14. Sheinman, M., O. Bénichou, Y. Kafri, and R. Voituriez. (2012). Classes of fast and specific search mechanisms for proteins on DNA. *Reports on Progress in Physics* 75:026601.

15. Yaffe, E., and A. Tanay. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* 43:1059–1065.

16. Imakaev, M., G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* 9:999–1003.

17. Knight, P. A., and D. Ruiz. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* 33:1029–1047.

18. Mirny Lab. (n.d.). Cooler package. https://github.com/mirnylab/cooler.

19. Jackson, V. (1999). Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods* 17:125–139.

20. Reuss, G., W. Disteldorf, A. O. Gamer, and A. Hilt. (2000). Formaldehyde. Wiley-VCH Verlag GmbH & Co. KGaA.

21. Lieberman-Aiden, E., N. L. v. Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326:289–293.

22. Kremer, K., and G. S. Grest. (1990). Dynamics of entangled linear polymer melts: A molecular dynamics simulation. *The Journal of Chemical Physics* 92:5057–5086.

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

24. Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* 117:1–19.

25. Press, W. H. (2007). Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press.

26. Elowitz, M. B., M. G. Surette, P.-E. Wolf, J. B. Stock, and S. Leibler. (1999). Protein Mobility in the Cytoplasm of *Escherichia coli*. *Journal of Bacteriology* 181:197–203.