NAR Genomics and Bioinformatics, 2026, **8**, lqag063
https://doi.org/10.1093/nargab/lqag063
Methods Article

OXFORD

Power-law penalties correct distance bias in single-cell co-accessibility and deep-learning chromatin interaction predictions

Luca Schlegel[1], **Fabio Gómez Cano**[2], **Alexandre P. Marand**[2,*], **Frank Johannes**[1,*]

[1]Plant Epigenomics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising 85354, Germany
[2]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, United States

*To whom correspondence should be addressed. Email: f.johannes@tum.de
Correspondence may also be addressed to Alexandre P. Marand. Email: amarand@umich.edu

Abstract

Scalable proxies for 3D genome contacts—such as single-cell co-accessibility and deep learning predictions—have emerged as powerful alternatives to chromatin capture-based methods, but predictions systematically overestimate long-range interactions. Here we show how to correct this bias using distance-based penalty functions informed by Gaussian mixture modeling and polymer-physics scaling. Using Hi-C datasets from maize, rice, and soybean, we derive tissue-specific and global consensus penalties parameterized by multiregime power-law exponents. Applying these corrections to single-cell ATAC sequencing co-accessibility scores improves their distance profiles in concordance with Hi-C and reduces long-range false positives by an average of 73% with tissue-specific penalties and 66% with the global consensus. We provide open-source code and fitted parameters to support adoption in maize, rice, and soybean.

Introduction

The three-dimensional (3D) organization of chromatin is central to gene regulation and genome stability in eukaryotes [1–3]. At fine scales, chromatin loops bring distal regulatory elements, such as enhancers and promoters, into spatial proximity [4–6]. In many plant species, these structures are embedded within larger features, including topologically associating domains [7–9] and active/inactive (A/B) compartments [1,10–13]. While Hi-C has been instrumental in mapping this hierarchy [10], its routine application across many genotypes and tissues remains limited by cost and logistics [14].

Scalable proxy methods, including co-accessibility scores from single-cell ATAC-seq [15] or from deep learning (DL) [16], have recently emerged as powerful alternatives. Co-accessibility inferred from single-cell ATAC sequencing (scATAC-seq) estimates contacts by identifying pairs of loci whose accessibility covaries across cells [17, 18]. Similarly, DL models learn sequence and epigenomic correlates of interaction frequency from experimental training data [19, 20]. Both approaches enable high-throughput generation of interaction-like scores across diverse tissues and species.

A key limitation is that both approaches tend to inflate long-range contacts [21–24]. For co-accessibility, correlations can arise from shared regulation even when loci are not physically proximate [25]. For DL models, predictive signals can be learned without explicitly enforcing the expected distance dependence [19, 21, 26]. This behavior conflicts with the empirically observed distance-decay of contact probability and contributes to very high false positive rates (FPRs) at large separations; in raw crop scATAC-seq co-accessibility, long-range FPR can exceed 0.90 (Supplementary Table S2). While heuristic normalizations exist [27], a broadly usable, data-driven correction that restores distance behavior to Hi-C-like profiles has been lacking.

Here we develop a correction framework that learns species-specific distance penalties from Hi-C and applies them to proxy interaction scores (co-accessibility and DL outputs). Empirical Hi-C distance-decay is often multiregime (piecewise scaling), so a single monotonic penalty is insufficient (see Supplementary Fig. S1). Derived from polymer physics, we assume that the contact probability $P(s)$ follows a power-law scaling $s^{-\alpha}$, where s is the given genomic distance between two chromatin regions and the exponent α reflects the underlying physical state of the chromatin fiber (e.g. fractal globule). [28] We therefore estimate multiregime power-law exponents using Gaussian mixture modeling (GMM) and use these to define tissue-specific and global consensus penalties—derived from an equally weighted aggregation of Hi-C datasets across multiple tissues and conditions—that can be applied to new datasets for each species (see Fig. 1A).

Methods

Data sources

All Hi-C and scATAC-seq datasets used in this study were publicly available and re-analyzed for consistent comparison. We obtained data from multiple tissues, developmental stages, and stress conditions. For soybean (*Glycine max*), we con-

Received: February 11, 2026. **Revised:** April 24, 2026. **Accepted:** May 17, 2026

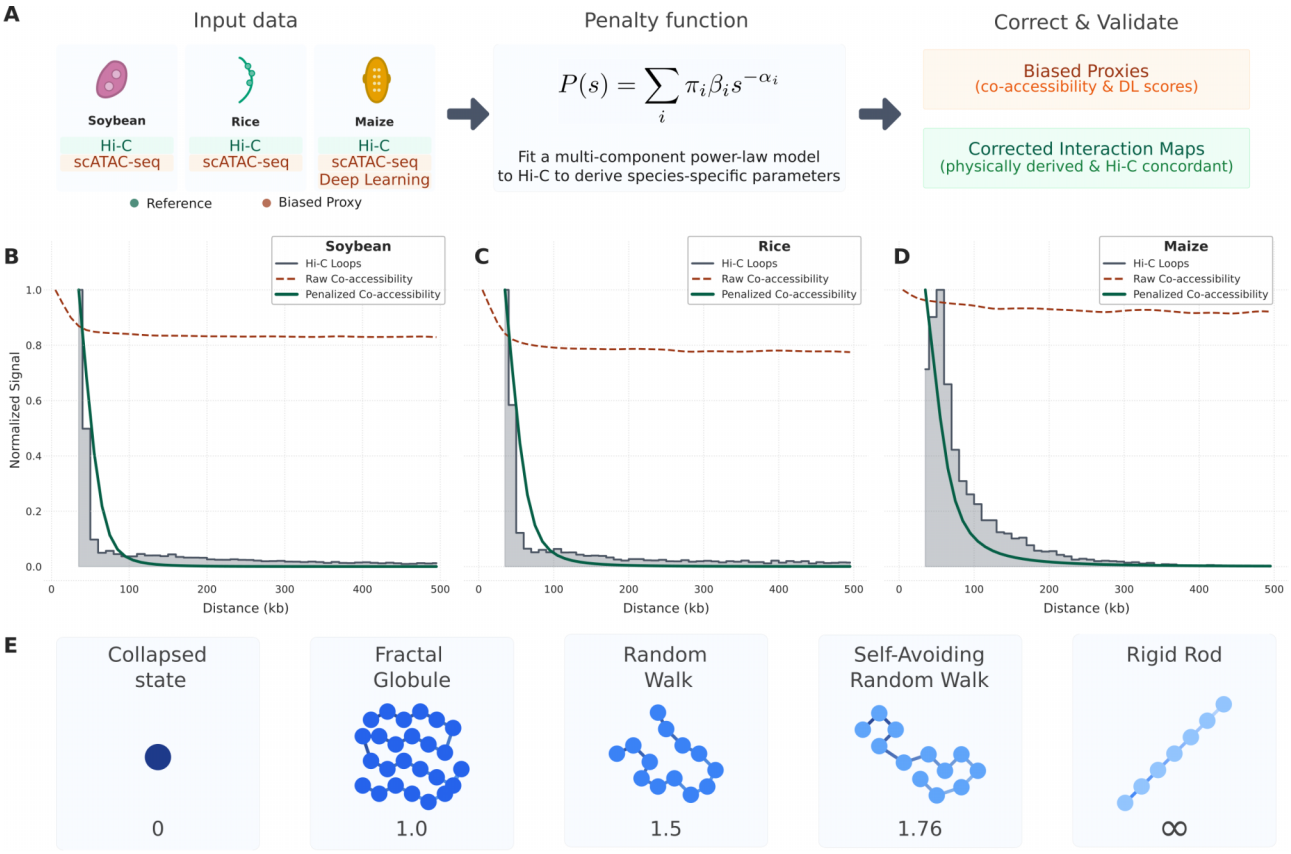

Figure 1. A data-driven penalty corrects distance bias in co-accessibility scores. **(A)** Conceptual workflow: Multicomponent power-law models fitted to species-specific Hi-C data derive penalty functions that correct biased proxies by down-weighting long-range interactions. **(B–D)** Chromatin interaction profiles for soybean, rice, and maize. Co-accessibility scores represent the mean Spearman correlation coefficients (ρ) calculated for accessible chromatin region (ACR) pairs within distance bins; only positive correlations ($\rho > 0$) are considered to characterize interaction decay. Hi-C decay (gray histogram) shows the expected decrease in interaction frequency with genomic distance, while raw co-accessibility scores (dark red dashed line) remain largely distance-independent. The GMM-based penalty corrects this bias, producing penalized profiles (dark green solid line) that closely match the Hi-C reference. All profiles are normalized to 1.0; Hi-C loop data is restricted to interactions > 35 kb. **(E)** The power-law decay exponent α corresponds to distinct theoretical polymer configurations. These idealized models provide a framework for interpreting the varying decay exponents α_i and their regimes.

sidered leaf tissue as well as seedling cotyledons, hooks, and hypocotyls under varying light conditions [18, 29]. For rice (*Oryza sativa*), we utilized data from 10-day-old seedlings, vegetative mesophyll, and reproductive endosperm [30, 31]. For maize (*Zea mays*), we considered leaf tissue from 6-day-old seedlings alongside reproductive tissues such as tassels and ears [27, 32]. Raw Hi-C data were sourced from GEO, SRA, NCBI BioProject (PRJNA), and GSA databases. A full list of accession numbers, key tissue information, and original publication references is provided in Supplementary Table S3. Accessible chromatin region (ACR) is used throughout to denote a peak-defined accessible region derived from scATAC-seq data.

Hi-C data processing, loop calling, and consensus construction

Raw Hi-C data [13, 29, 31–33] were processed using a uniform `HiC-Pro (v3.1.0)` [34] pipeline for raw read alignment, normalization, and matrix generation with nondefault settings. Specifically, reads were aligned with `Bowtie2` in two passes—global followed by local—using `–very-sensitive –score-min L,-0.6,-0.2`, with seed length `-L 30` (global) and `-L 20` (local). Alignments with `MAPQ < 5` were discarded. A *Dpn*II digestion model was used (ligation motif `GATCGATC`) with the corresponding fragment BED; *cis* contacts were defined using `MIN_CIS_DIST = 20` kb. Interaction classes were reported, and singletons, multimappers, and PCR duplicates were removed (`GET_ALL_INTERACTION_CLASSES = 1`; `RM_SINGLETON = 1`; `RM_MULTI = 1`; `RM_DUP = 1`). Valid pairs were aggregated into raw contact matrices at 5 kb, 10 kb, 20 kb, 50 kb, 100 kb, 200 kb, 500 kb, and 1 Mb (upper-triangle format). The 5 kb resolution matrices were used as the basis for all downstream distance-decay profile generation, as this scale maintained sufficient effective resolution across all included datasets (Supplementary Fig. S13). Matrices were then iterative correction and eigenvector decomposition (ICE)-normalized with up to 100 iterations, filtering low-count bins at 2% and disabling high-count filtering (`FILTER_LOW_COUNT_PERC = 0.02`; `FILTER_HIGH_COUNT_PERC = 0`; `EPS = 0.1`). Default `HiC-Pro` settings were used otherwise. `ValidPairs` alignments were converted to the Juicer `.hic` file format using a combination of `HiC-Pro` and `juicer_tools` utilities. Chromatin loops were identified from `.hic`-formatted files using `HICCUPs` from the `Juicer` suite [35] of Hi-C tools with nondefault

parameters (`-m 512 -r 1000,2000,5000,1000 -f 0.5,0.5,0.5,0.5 -p 10,8,4,2 -i 10,8,7,5 -d 5000,10000,20000,20000 –ignore_sparsity`). Genomic distance for all loop calls was defined as the linear separation between anchor midpoints; given the 10 kb Hi-C anchor resolution, this results in a 35 kb minimum midpoint distance (equivalent to a 25 kb end-to-end gap), which accommodates the varying lengths of integrated scATAC-seq and DL anchors. Loop coordinates (BEDPE) were used for distance-decay modeling. For species-level global consensus datasets, unique loop sets were pooled across tissues and developmental stages, with each dataset equally weighted in the final aggregate interaction pool.

scATAC-seq data processing and co-accessibility

We re-analyzed scATAC-seq data from corresponding leaf tissues for soybean [18], rice [30], and maize [27]. ACRs were identified from pseudo-bulk signals and peak calling; a cell-by-peak binary accessibility matrix was then constructed. Co-accessibility scores between ACR pairs (separated by 2–500 kb) were calculated as previously described [21], using pseudocells generated via a k-nearest neighbors strategy [27] to mitigate single-cell sparsity. Scores represent Spearman correlation coefficients (ρ) adjusted for technical confounders such as read depth. To specifically characterize the distance-dependent decay of functional chromatin interactions, we retained only positively co-accessible pairs ($\rho > 0$) for downstream analysis, as these profiles are distributionally representative of the entire dataset (Supplementary Fig. S8). Genic and proximal ACRs were defined as regions overlapping (> 1 bp) gene bodies and located within 2 kb of transcription start sites and 500 bp of transcription termination sites, respectively. Remaining ACRs were classified as distal. For co-expression analyses, pairwise Pearson correlations were calculated among genes expressed (TPM > 1) in at least two independent samples from the maize expression atlas [36].

Derivation of the distance-based penalty function

Data transformation for segmented power-law analysis

Chromatin loop distances were filtered to a maximum genomic separation of 5 Mb to ensure parameter stability and consistent binning. To characterize the heterogeneous nature of contact decay, loop counts were logarithmically binned (150 bins per dataset) using geometric bin midpoints to maintain accuracy in log-space transformations. Both the binned counts and their corresponding distances were transformed into $\log_{10}$ space, where the power-law relationship $P(s) \propto s^{-\alpha}$ is represented as a linear function with a slope of $-\alpha$:

$$\log_{10}(P(s)) = -\alpha \log_{10}(s) + C, \tag{1}$$

where $s \in \mathbb{N}_0$ is the genomic distance between two chromatin regions and $C \in \mathbb{R}$ is a constant. This scaling relationship, particularly for the collapsed and dense states characteristic of chromatin organization, served as the theoretical inspiration for our data-driven fitting method [28, 37].

GMM-based decomposition of interaction profiles

Because empirical contact profiles were frequently nonlinear across the full 5 Mb range, we employed a segmented mixture-of-regressions strategy based on GMM to decompose profiles into n distinct scaling regimes. Specifically, we used the `sklearn.mixture.GaussianMixture` class [38] with `covariance_type='full'` to cluster the two-dimensional distribution of log-transformed data points. This served as a data-driven method to partition the contact points into subpopulations, each corresponding to a segment with a unique decay profile.

The optimal number of components (n) was determined for each dataset using a dual selection process. First, we calculated the Bayesian Information Criterion:

$$BIC = k \ln(N) - 2 \ln(\hat{L}), \tag{2}$$

where k is the number of estimated parameters, N is the number of observations (bins), and $\hat{L}$ is the maximized value of the likelihood function. For each $n \in [1, 10]$, higher complexity was accepted only if it provided a significant statistical improvement (ΔBIC > 20, strong evidence) [39] and simultaneously introduced a new regime with a characteristic decay exponent α differing by at least 0.25 from previously identified components (see Supplementary Fig. S2).

Bootstrapping of GMM transition points

To ensure the stability and relevance of the GMM results, we bootstrapped the loop-distance distributions for species-specific global consensus datasets across 10 000 replicates. Each replicate was re-binned and re-fitted using the selected n-component model to derive robust parameter estimates. We reported medians and 95% confidence intervals (2.5th and 97.5th percentiles) for the transition points T_i and scaling exponents α_i. These parameter distributions, which confirm the structural reliability of the identified regimes, are provided in Supplementary Table S1 and Supplementary Fig. S3.

Construction of the GMM density function

For each of the n clusters identified by the GMM, a linear regression was performed on its constituent log-transformed data points. The slope of this local linear regression, slope_i, was derived directly from the GMM's cluster parameters using the relationship

$$\text{slope}_i = \frac{\Sigma_{i,1,2}}{\Sigma_{i,1,1}}, \tag{3}$$

where Σ_i is the covariance matrix of the i-th component and $\Sigma_{i,j,k}$ denotes the element in the j-th row and k-th column of that matrix. This slope determined the unique power-law exponent α_i for that segment ($\alpha_i = -\text{slope}_i$). The GMM density function $P(s)$ is defined as

$$P(s) = \sum_{i=1}^{n} w_i \cdot \beta_i \cdot s^{-\alpha_i}, \tag{4}$$

where $\sum w_i = 1$ are the weights (derived from the GMM), β_i are the scaling factors, and α_i are the fitted characteristic decay exponents. The specific sets of parameters $\{\alpha_i, w_i, \beta_i\}$ for each species are provided in Supplementary Table S1.

Application and benchmarking

To correct scATAC-seq and DL scores, we derived a continuous piecewise power-law function (linear in $\log_{10}$–$\log_{10}$ space), $P_{\text{applied}}(s)$, from the GMM density function (4). Because the Hi-C loop data used for fitting was restricted to distances $\geq$ 35 kb, we implemented a distance plateau below this threshold where no correction is applied ($P = 1$). For all pairs with

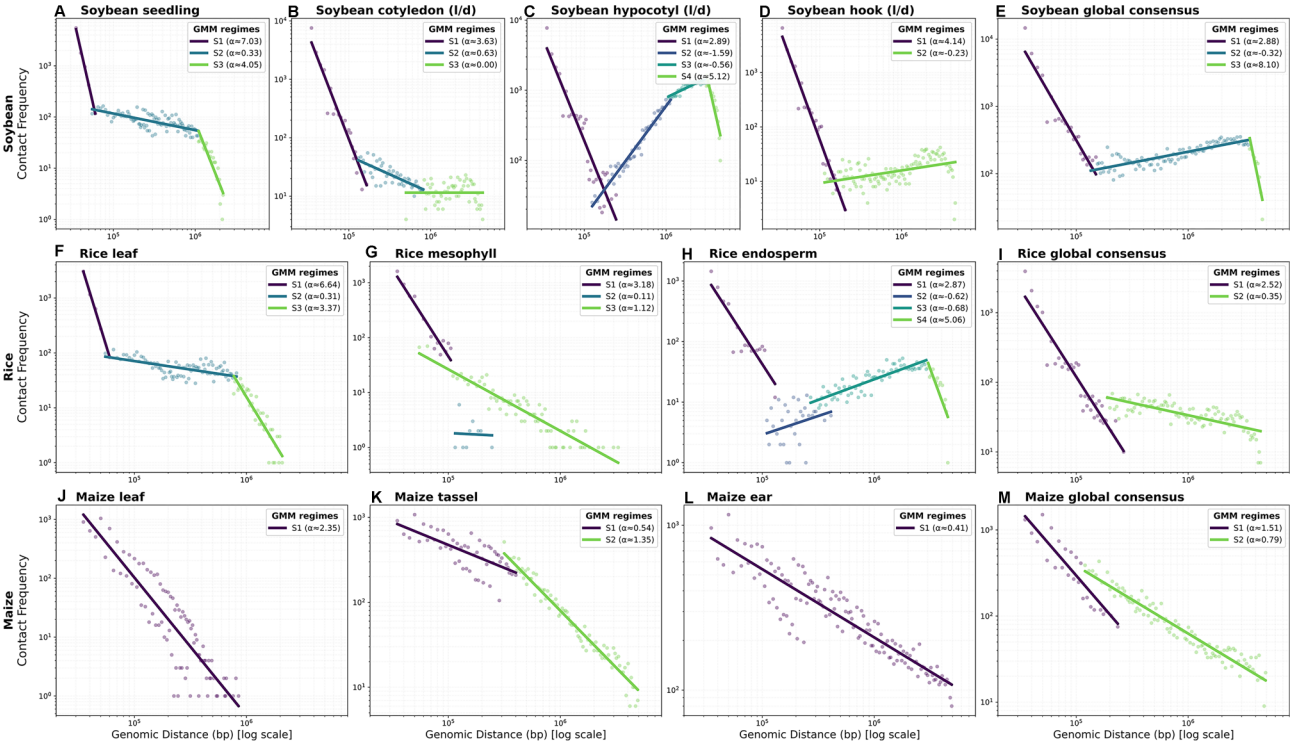

Figure 2. Tissue-specific GMM decomposition of chromatin interaction profiles. Hi-C contact frequencies (points) show that chromatin scaling is piecewise linear in log–log space, requiring multicomponent functions to capture distinct architectural regimes. GMM identifies these interaction subpopulations (S_1 to S_4), with each segment representing a unique power-law decay exponent derived from cluster covariance. These profiles demonstrate that interaction scaling is both species- and tissue-specific. Ten independent datasets are shown: soybean (**A–D**: seedling, cotyledon, hypocotyl, hook), rice (**F–H**: leaf, mesophyll, endosperm), and maize (**J–L**: leaf, tassel, ear), alongside their respective consensus profiles (**E, I, M**).

distance $s \geq 35$ kb, the penalty is calculated as

$$P_{\text{applied}}(s) = \begin{cases} 1 & \text{if } s < 35\,\text{kb} \\ \hat{\beta}_1 s^{-\alpha_1} & \text{if } 35\,\text{kb} \leq s < T_1 \\ \vdots & \vdots \\ \hat{\beta}_n s^{-\alpha_n} & \text{if } s \geq T_{n-1} \end{cases}. \tag{5}$$

The transition points T_i are determined by the intersections of the local regressions in log-log space and verified by bootstrapping. The scaling constants $\hat{\beta}_i$ are recalculated to ensure continuity at each boundary and to normalize the function such that $P_{\text{applied}}(35\,\text{kb}) = 1$. The distance-corrected interaction scores are generated by multiplying the raw co-accessibility scores or DL predictions by the corresponding value of $P_{\text{applied}}(s)$.

To evaluate the GMM-based piecewise model, we benchmarked its performance against a standard exponential ansatz, $P(s) = e^{-\lambda s}$ and Cicero framework which uses $P(s) = (1 - s^{-0.75}) \times \beta$ where β is some constant. Fit quality was assessed using R^2 scores in log-log space across a 5 Mb range. Penalized co-accessibility profiles (P) were compared against Hi-C ground truth (H) by scaling both to a common baseline at 35 kb. Area-based FPR and False Negative Rates (FNR) were calculated across 35–500 kb distance bins (i) as:

$$FPR = \frac{\sum_i \max(0, P_i - H_i)}{\sum_i P_i}, \quad FNR = \frac{\sum_i \max(0, H_i - P_i)}{\sum_i H_i} \tag{6}$$

This formulation measures the integrated positive difference between penalized co-accessibility and physical interactions. Detailed performance metrics, including F_1 scores, Spearman correlations, and Wasserstein distances, are provided in Supplementary Table S2 and Supplementary Figs S3 and S5.

Calculation of DL-based interaction scores

DL-based interaction scores for maize were generated using `GenomicLinks`, a sequence-based model that predicts chromatin looping potential directly from DNA [21]. A schematic overview of the `GenomicLinks` architecture is provided in Supplementary Fig. S11. The model utilizes a combination of convolutional neural networks and long short-term memory layers, previously trained on experimental HiChIP data (Maize leaf tissue, targeting H3K4me3 and H3K27me3) [32]. We applied the pre-trained model to all pairwise combinations of HiChIP anchor regions across the maize genome, utilizing the model's fixed 2500 bp input resolution. The model generates a raw DL score for each pair, representing the predicted interaction potential. To ensure a direct comparison with our other metrics, these predictions were filtered to a genomic distance range of 0–500 kb and corrected using the species-specific penalty function as described above.

Polymer model simulations

To characterize the relationship between the distance-decay exponent α and resulting chromatin topologies, we performed coarse-grained Monte Carlo simulations using the Pivot Algorithm [40]. Chains of $N = 10,000$ steps were simulated on a cubic lattice across four canonical folding regimes: Random Walk (RW, $\alpha \approx 1.5$), self-avoiding walk (SAW, $\alpha \approx 1.76$), and both models under spherical confinement to approxi-

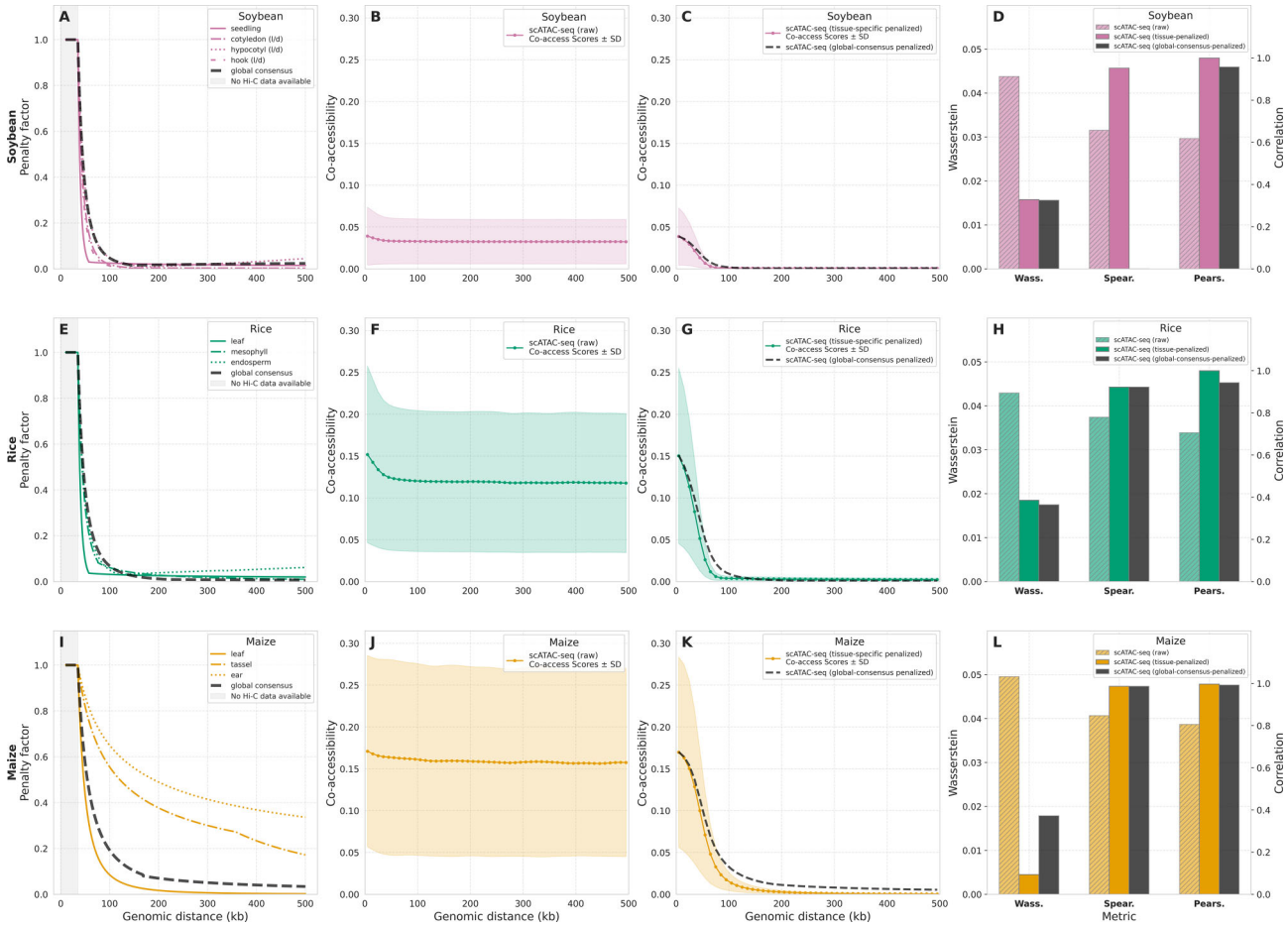

Figure 3. Application and statistical validation of species-specific and global consensus penalty functions. **(A, E, I)** Species-specific and global consensus penalty functions ($P_{\text{applied}}(s)$) derived via GMM decomposition of Hi-C interaction counts. Hi-C data constraints limit the penalty domain to distances > 35 kb. **(B, F, J)** Genomic distance distribution of raw scores with standard deviation, defined as unadjusted scATAC-seq co-accessibility frequencies. Shaded regions indicate the standard deviation across analyzed ACR pairs. These scores exhibit minimal decay, leading to the overestimation of long-range interactions. **(C, G, K)** Transformation into penalized scores, calculated as the product of raw scores and the corresponding distance-dependent penalty [$P_{\text{applied}}(s) \times$ raw scores]. Profiles are shown for tissue-specific (colored solid lines) and global consensus (black dashed lines) models. Shaded areas represent the standard deviation of co-accessibility scores across ACR pairs. **(D, H, L)** Performance benchmarking across species. Compared to raw scores (hatched bars), penalized scores from tissue-specific (colored solid bars) and global consensus (gray bars) models significantly improved concordance with Hi-C ground truth, as measured by reduced Wasserstein distances and increased Pearson correlation coefficients (PCC).

mate Fractal Globule-like scaling ($\alpha \approx 1.0$). Ensembles were equilibrated using a number of pivot moves proportional to the chain length to ensure efficient sampling of the conformational space. The scaling exponent α for each model was empirically derived via linear regression of $\log_{10}[P(s)]$ versus $\log_{10}(s)$ for pairwise genomic separations $s > 10$ steps. Visualizations of the 3D conformations and corresponding $P(s)$ decay curves were generated to validate the physical basis of the scaling exponents (see Supplementary Fig. S4 and Fig. 1E).

Results and discussion

Multicomponent power-law decay of Hi-C contacts reveals a multiscale plant chromatin architecture

We analyzed Hi-C contact–distance decay profiles from soybean, rice, and maize across multiple tissues to quantify how contact probability changes with genomic separation. On log–log axes, Hi-C contact frequency versus distance was consistently piecewise linear, indicating that a single power-law exponent is insufficient to describe chromatin contact scaling across distances [41] (Fig. 2). We therefore fit each profile with a GMM to identify multiple scaling regimes and estimate a set of decay exponents $\{\alpha_i\}$ and their transition points (Figs 2 and 3A, E, and I; see the 'Methods' section). These exponents α_i provide a quantitative measure of chromatin compactness. By mapping empirical values to canonical polymer benchmarks (Fig. 1E), we interpret $\alpha \approx 1$ (Fractal Globule) [10, 28] as a dense, nonequilibrium packing state, while higher values (e.g. $\alpha \approx 1.5$, RW) signify more expanded or stretched configurations. Tissue-specific profiles typically required two to four regimes in soybean and rice but only one or two in maize. For global consensus models (pooled across tissues within species), the GMM identified three regimes in soybean and two regimes in rice and maize (Supplementary Table S1).

The inferred regimes align with known architectural scales. In rice and soybean, the first regime extended to bootstrapped transition points T_1 with medians of 120.4 kb (rice) and 140.2 kb (soybean) and showed steep decay ($\alpha \approx 2.6$–3.0), consistent with strongly partitioned short-range organization (Fig. 1E). These T_1 thresholds coincide with independent loop annotations, encompassing 97% of identified loops in rice [42] and 70% in soybean [29].

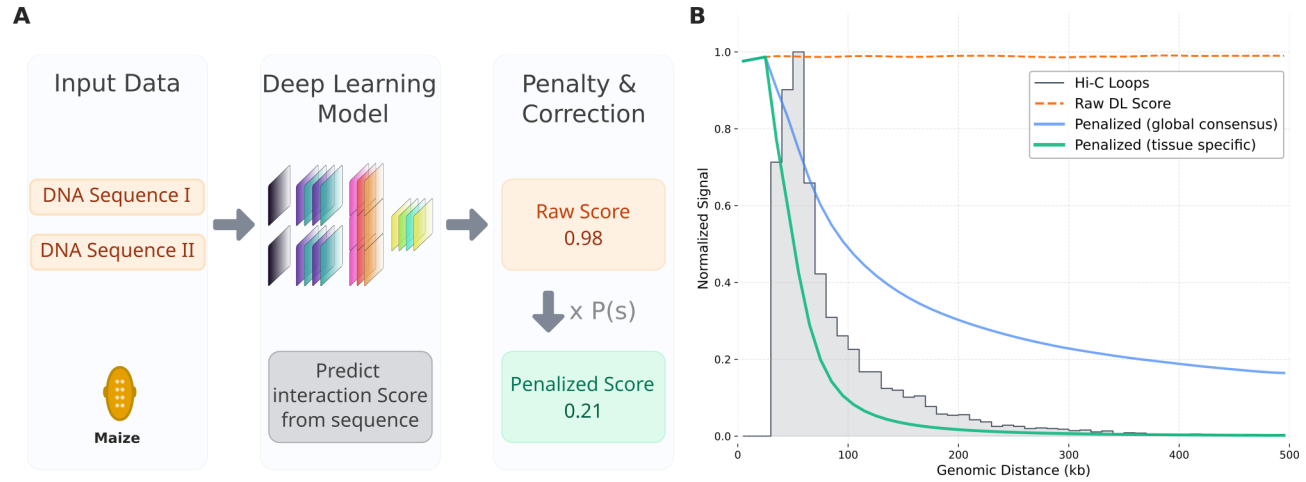

Figure 4. A polymer-based penalty corrects distance bias in DL predictions for maize. **(A)** Workflow for the distance-correction of DL scores. The `GenomicLinks` model predicts raw scores, defined as interaction probabilities derived exclusively from DNA sequence features without distance constraints. These are updated into penalized (corrected) scores by multiplying the raw output by the species-specific and global consensus penalty functions $P_{\text{applied}}(s)$ derived from Hi-C data. **(B)** Distance-bias correction in maize. Raw DL scores (orange dashed line) are mostly independent of genomic distance, leading to the overestimation of long-range contacts compared to the Hi-C reference (gray histogram). Application of the tissue-specific (leaf) or global consensus penalty produces corrected profiles (green solid line, blue solid line, respectively) that recapitulates experimental Hi-C decay. All profiles are normalized to a value of 1.0 at the 35 kb baseline.

Beyond T_1, both rice and soybean entered a second regime with weak or near-flat scaling ($\alpha = 0.37$ in rice; $\alpha = -0.33$ in soybean), consistent with mid-range enrichment and domain-scale organization. A strictly monotonic distance penalty would therefore over-suppress enriched distance ranges, motivating a multiregime penalty design. In contrast, maize exhibited a more monotonic scaling behavior. While the global consensus model supported a transition at ~168 kb, several tissues were best captured by a single power-law (e.g. leaf, $\alpha = 2.35$; Fig. 2).

Co-accessibility scores and deep-learning predictions overestimate long-range interactions

To quantify distance bias in scalable interaction proxies, we re-analyzed leaf scATAC-seq datasets from soybean [18], rice [30], and maize [27] and compared their co-accessibility distance profiles against tissue-matched Hi-C (leaf for rice and maize; seedling for soybean). Co-accessibility inference from scATAC-seq is challenged by sparsity, so we followed standard preprocessing by constructing pseudocells using a data-driven k-nearest-neighbors strategy [27]. We then computed pairwise correlations of normalized accessibility across pseudocells for ACR pairs separated by 2–500 kb. This yielded 6.3 million positively co-accessible ACR pairs in soybean, 3.8 million in rice, and 3.4 million in maize.

A side-by-side comparison with Hi-C revealed a consistent distortion: raw co-accessibility scores show little to no decay with increasing genomic distance, with elevated interaction scores at large separations relative to Hi-C (Fig. 1B–D, and 3B, F, and J). Uncorrected co-accessibility exhibited very high long-range FPR, exceeding 0.90 across datasets (Supplementary Table S2). Here, the FPR is quantified as the area-based excess signal relative to the Hi-C ground truth (see the 'Methods' section for definition).

Distance bias is not specific to co-accessibility. We observed the same pattern in deep-learning (DL) interaction proxies, which often omit explicit genomic distance as a feature during training [16, 21, 43]. Using `GenomicLinks` predictions for maize [21], raw DL scores were largely distance-independent across the 0–500 kb interval (Fig. 4B). While GenomicLinks was trained on HiChIP, the whole maize leaf Hi-C set was used for benchmarking as the regulatory HiChIP subset lacks sufficient statistical power for distributional analysis (Supplementary Fig. S9). This mirrors co-accessibility inflation and overestimating distal interactions relative to the Hi-C reference. Together, these analyses motivate a correction that is distance-aware, learned from empirical Hi-C decay in the relevant species, and portable across proxy score types.

A GMM-based penalty corrects distance bias in co-accessibility and DL predictions

Several methods implement distance-related adjustments for co-accessibility, but many rely on fixed exponents, heuristic cutoffs, or nontransferable background assumptions [17, 44, 45]. In plants, both decay exponents and regime boundaries vary across species and tissues, making universal one-parameter corrections suboptimal. We therefore derived penalties directly from GMM-fitted Hi-C decay profiles. Across the 35–500 kb interval used for scATAC-seq benchmarking, the multiregime model outperformed both the exponential fit and Cicero-based normalizations, with at least an 11.1-fold reduction in residual sum of squares across all datasets (Supplementary Fig. S5).

Applying tissue-specific penalties to raw co-accessibility scores reshaped distance-independent profiles into Hi-C-like decay curves (Fig. 1B–D and 3C, G, and K). Quantitatively, corrected profiles showed improved concordance with Hi-C (higher Pearson, lower Wasserstein) and markedly improved specificity. FPR dropped from above 90% in raw profiles to 2.2% in maize, 39.6% in rice, and 33.4% in soybean, while average F1 increased from 0.15 to 0.83 across tissue-specific models (Supplementary Table S2). The soybean global consensus revealed a statistical uniqueness: while it achieved a strong F1 score (0.814), the Spearman correlation dropped to −0.014. This is localized to a small genomic segment where the penalty slightly upregulates co-accessibility scores due to

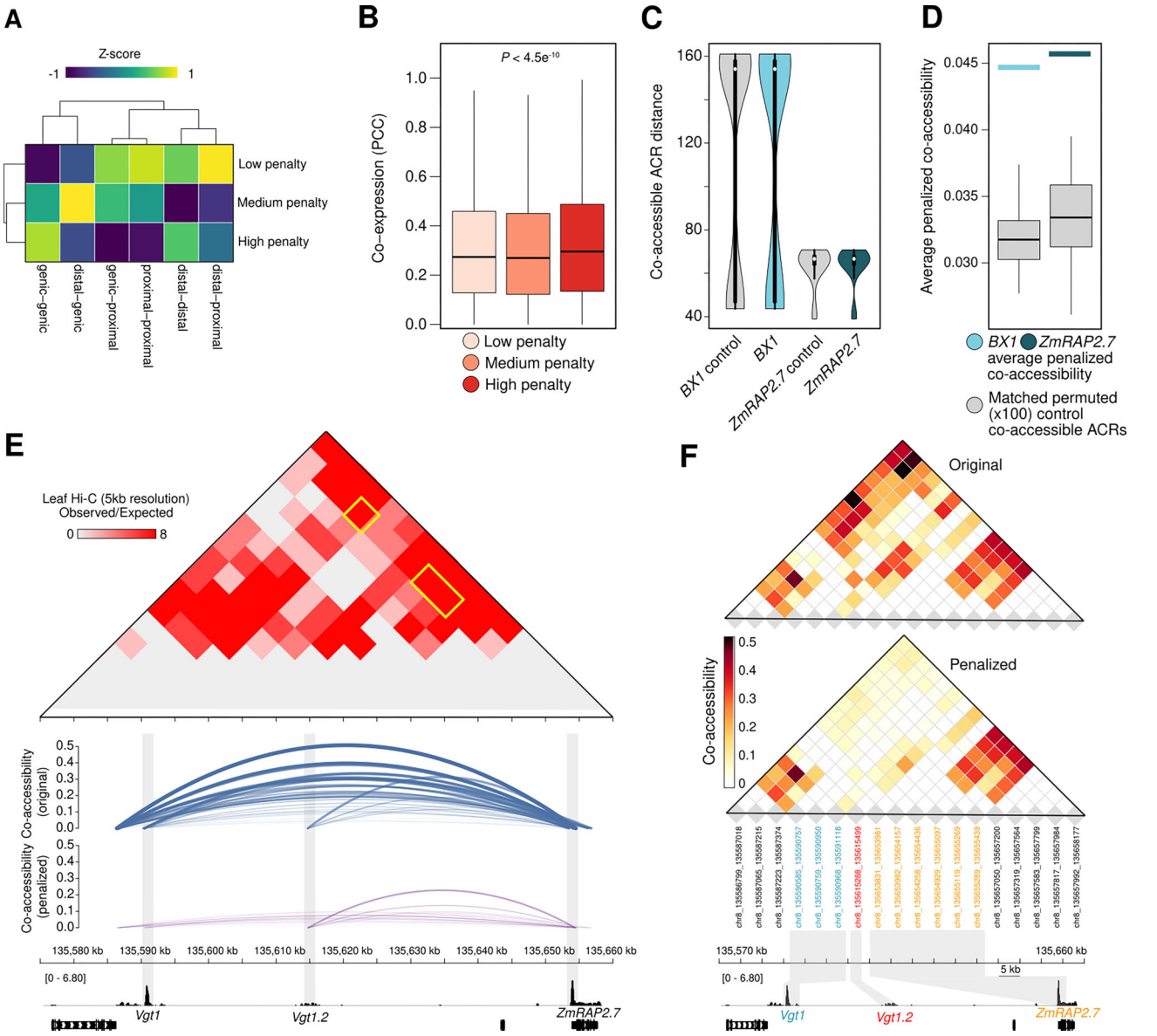

Figure 5. Penalty enriches for functional interactions and separates co-regulation from physical proximity in maize. **(A)** Heatmap of co-accessible ACR classification enrichment (*Z*-score) in low, medium, and high penalized tertiles. Proximal–distal links are enriched in the low penalty group, while genic–genic links are most penalized. **(B)** Boxplots showing that gene pairs with the highest penalty scores also have the highest co-expression (PCC), suggesting the penalty effectively down-weights co-regulated, noninteracting loci. **(C)** Distribution of co-accessible ACR intervening distances for *BX1* control (gray), *BX1* (light blue), *ZmRAP2.7* controls (gray), and *ZmRAP2.7* (dark blue). **(D)** Distribution of average penalized co-accessibility scores for *BX1* and *ZmRAP2.7* loci (solid bars) relative to permuted (×100) controls matched by co-accessible ACR intervening distances. **(E)** Leaf Hi-C chromatin contacts (top), genome browser view of raw and penalized co-accessibility scores (middle), and chromatin accessibility profiles (bottom) at the *Vgt1–ZmRAP2.7* locus. *Vgt1*, Vegetative to generative transition 1. *ZmRAP2.7*, *Zea mays* RELATED TO APETALA2. Yellow boxes showcase *Vgt1* and *Vgt1.2* interactions with the promoter of *ZmRAP2.7*. **(F)** Co-accessibility score pairwise heatmap before (top) and after (bottom) distance penalization. ACR coordinates with respect to the reference genome are shown with gray boxes on the genome browser representation of chromatin accessibility (bottom).

a nonmonotonic scaling regime ($\alpha < 0$), creating a rank-order disagreement (Supplementary Fig. S6B and S10). This reflects a localized violation of the monotonic scaling assumed in idealized polymer models. Global consensus penalties provided a robust default when tissue-matched Hi-C was unavailable. In maize and rice, global consensus performance was close to tissue-specific models across profile and error-rate metrics (Supplementary Table S2). Across species, long-range false positives were reduced by an average of 73% for tissue-specific penalties and 66% for global consensus penalties.

To test biological plausibility, we stratified maize co-accessible ACR pairs by penalty magnitude (Δ co-accessibility). The most penalized tertile was enriched for genic–genic and distal–distal links, while distal–proximal links were enriched among least penalized pairs (Fig. 5A). Genic–genic pairs in the most penalized tertile also showed significantly higher co-expression values (Kruskal-Wallis test, $P < 4.5 \times 10^{-10}$; Fig. 5B), consistent with the depletion of co-regulatory correlations that can inflate long-range proxy scores. At the same time, known long-range cis-regulatory loci associated with *ZmRAP2.7* [46] and *BX1* [47] retained ele-

vated penalized co-accessibility relative to distance-matched background (Permutation test, $P < 1 \times 10^{-3}$; Fig. 5C–F). For example, *Vgt1* and *Vgt1.2* [48] regulatory loci exhibit penalized signal above background with ACRs within 1-kb of *ZmRAP2.7*, though *Vgt1* is down-weighted more than *Vgt1.2* (Fig. 5E and F). *BX1* distal and activator sequence cis-elements (DICE and ASCE) exhibited a similar trend, with interactions from the more distal DICE element downweighted relative to ASCE (Supplementary Fig. S7).

Finally, applying the maize species-specific penalty to `GenomicLinks` predictions converted distance-independent raw DL profiles into curves aligned with Hi-C decay (Fig. 4B). Within the 35–500 kb interval, the Spearman correlation relative to Hi-C ground truth improved from −0.50 to 0.99 (see Supplementary Fig. S12 for distance-bin breakdowns), confirming the method functions as a modular and broadly applicable tool for correcting distance bias in diverse interaction proxies.

Conclusion

We introduced a distance-aware correction framework for scalable chromatin-interaction proxies that systematically inflate long-range contacts. By learning multiregime distance-decay behavior from Hi-C using GMM, we derive tissue-specific and global consensus penalty functions and apply them to co-accessibility and DL interaction scores. This correction restores Hi-C-like distance profiles and reduces long-range false positives by about 70% on average across maize, rice, and soybean, while preserving biologically meaningful interaction signal.

While this framework provides an effective correction, it currently relies on bulk Hi-C data for calibration. However, our global consensus penalties perform nearly as well as tissue-specific models, achieving Spearman correlations as high as 0.98 in maize and 0.92 in rice. Key future directions include integrating these parameters into dynamic polymer models [49] and incorporating distance awareness directly into the DL training process. We release code and fitted parameters (jlab-github or zenodo) to support adoption and extension to additional plant species and tissues.

Acknowledgements

We thank members of the Johannes lab for helpful feedback.

Author contributions: Luca Schlegel (Conceptualization, Methodology, Software, Writing – original draft), Fabio Gómez Cano (Data curation, Resources, Methodology, Writing – review & editing), Alexandre P. Marand (Conceptualization, Resources, Data curation, Supervision, Funding acquisition, Writing – review & editing), Frank Johannes (Conceptualization, Supervision, Funding acquisition, Writing – review & editing).

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by the Technical University of Munich (F.J. and L.S.), and by the National Institute of General Medical Sciences of the National Institutes of Health (1R00GM144742) and start-up funds from the University of Michigan to A.P.M.

Data availability

The open-source code, derived parameters, and processed datasets generated during this study are publicly available on Zenodo (https://doi.org/10.5281/zenodo.18607660). A GitHub version of the code is maintained at https://github.com/jlab-code/polymer-penalty. All Hi-C and scATAC-seq datasets re-analyzed in this study are publicly available; a full list of original sources and accession numbers is provided in Supplementary Table S3.

References

1. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016;17:661–78. https://doi.org/10.1038/nrg.2016.112
2. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. *Nat Rev Genet* 2018;19:789–800. https://doi.org/10.1038/s41576-018-0060-8
3. Domb K, Wang N, Hummel G *et al.* Spatial features and functional implications of plant 3D genome organization. *Annu Rev Plant Biol* 2022;73:173–200. https://doi.org/10.1146/annurev-arplant-102720-022810
4. Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;159:1665–80. https://doi.org/10.1016/j.cell.2014.11.021
5. Gagliardi D, Manavella PA. Short-range regulatory chromatin loops in plants. *New Phytol* 2020;228:466–71. https://doi.org/10.1111/nph.16632
6. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* 2021;22:1–30. https://doi.org/10.1186/s13059-021-02322-1
7. Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485:376. https://doi.org/10.1038/nature11082
8. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. *Nat Genet* 2020;52:8–16. https://doi.org/10.1038/s41588-019-0561-1
9. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv* 2019;5:eaaw1668. https://doi.org/10.1126/sciadv.aaw1668
10. Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326:289–296. https://doi.org/10.1126/science.1181369
11. Harris HL, Gu H, Olshansky M *et al.* Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. *Nat Commun* 2023;14:1–17. https://doi.org/10.1038/s41467-023-38429-1
12. Dong P, Tu X, Li H *et al.* 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol Plant* 2017;10:1497–509. https://doi.org/10.1016/j.molp.2017.11.005
13. Dong P, Tu X, Li H *et al.* Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains. *J Integr Plant Biol* 2020;62:201–17. https://doi.org/10.1111/jipb.12809
14. Nagano T, Lubling Y, Yaffe E *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur

simultaneously in a single cell. *Nat Protoc* 2015;10:1986–2003. https://doi.org/10.1038/nprot.2015.127
15. Marand AP, Chen Z, Gallavotti A *et al.* A cis-regulatory atlas in maize at single-cell resolution. *Cell* 2021;184:3041–55. https://doi.org/10.1016/j.cell.2021.04.014
16. Schwessinger R, Gosden M, Downes D *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;17:1118–24. https://doi.org/10.1038/s41592-020-0960-3
17. Pliner HA, Packer JS, McFaline-Figueroa JL *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;71:858–71. https://doi.org/10.1016/j.molcel.2018.06.044
18. Zhang X, Luo Z, Marand AP *et al.* A spatially resolved multi-omic single-cell atlas of soybean development. *Cell* 2025;188:550–67. https://doi.org/10.1016/j.cell.2024.10.050
19. Piecyk RS, Schlegel L, Johannes F. Predicting 3D chromatin interactions from DNA sequence using Deep Learning. *Comput Struct Biotechnol J* 2022;20:3439–48. https://doi.org/10.1016/j.csbj.2022.06.047
20. Zou J, Huss M, Abid A *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;51:12–18. https://doi.org/10.1038/s41588-018-0295-5
21. Schlegel L, Bhardwaj R, Shahryary Y *et al.* GenomicLinks: deep learning predictions of 3D chromatin interactions in the maize genome. *NAR Genom Bioinform* 2024;6:lqae123. https://doi.org/10.1093/nargab/lqae123
22. Wall BPG, Nguyen M, Harrell JC *et al.* Machine and deep learning methods for predicting 3D genome organization. *Methods Mol Biol* 2025;2856:357–400. https://doi.org/10.1007/978-1-0716-4136-1_22
23. Chen H, Lareau C, Andreani T *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* 2019;20:241. https://doi.org/10.1186/s13059-019-1854-5
24. Lee DS, Luo C, Zhou J *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;16:999–1006. https://doi.org/10.1038/s41592-019-0547-z
25. Zhai J, Jeong DH, De Paoli E *et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev* 2011;25:2540–53. https://doi.org/10.1101/gad.177527.111
26. Wang Y, Kong S, Zhou C *et al.* A review of deep learning models for the prediction of chromatin interactions with DNA and epigenomic profiles. *Brief Bioinform* 2024;26:bbae651. https://doi.org/10.1093/bib/bbae651
27. Marand AP, Jiang L, Gomez-Cano F *et al.* The genetic architecture of cell type-specific cis regulation in maize. *Science* 2025;388:eads6601. https://doi.org/10.1126/science.ads6601
28. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* 2011;19:37–51. https://doi.org/10.1007/s10577-010-9177-0
29. Wang L, Jia G, Jiang X *et al.* Altered chromatin architecture and gene expression during polyploidization and domestication of soybean. *Plant Cell* 2021;33:1430–46. https://doi.org/10.1093/plcell/koab081
30. Yan H, Mendieta JP, Zhang X *et al.* A single-cell rice atlas integrates multi-species data to reveal cis-regulatory evolution. *Nat Plants* 2025;11:2050–71. https://doi.org/10.1038/s41477-025-02106-6
31. Liu C, Cheng YJ, Wang JW *et al.* Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants* 2017;3:742–8. https://doi.org/10.1038/s41477-017-0005-9
32. Ricci WA, Lu Z, Ji L *et al.* Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants* 2019;5:1237–49. https://doi.org/10.1038/s41477-019-0547-0
33. Li Z, Sun L, Xu X *et al.* Light control of three-dimensional chromatin organization in soybean. *Plant Biotechnol J* 2024;22:2596–611. https://doi.org/10.1111/pbi.14372
34. Servant N, Varoquaux N, Lajoie BR *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 2015;16:259. https://doi.org/10.1186/s13059-015-0831-x
35. Durand NC, Shamim MS, Machol I *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;3:95–98. https://doi.org/10.1016/j.cels.2016.07.002
36. Stelpflug SC, Sekhon RS, Vaillancourt B *et al.* An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* 2016;9:plantgenome2015.04.0025. https://doi.org/10.3835/plantgenome2015.04.0025
37. Grosberg AY, Khokhlov AR. *Statistical Physics of Macromolecules*. New York, NY, USA: Amer Inst of Physics, 1994.
38. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
39. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995;90:773–95. https://doi.org/10.1080/01621459.1995.10476572
40. Madras N, Sokal AD. The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk. *J Stat Phys* 1988;50:109–86. https://doi.org/10.1007/BF01022990
41. Sanborn AL, Rao SSP, Huang SC *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 2015;112:E6456–65. https://doi.org/10.1073/pnas.1518552112
42. Zhao L, Wang S, Cao Z *et al.* Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. *Nat Commun* 2019;10:3640. https://doi.org/10.1038/s41467-019-11535-9
43. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;17:1111–7. https://doi.org/10.1038/s41592-020-0958-x
44. Wang C, Sun D, Huang X *et al.* Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol* 2020;21:198. https://doi.org/10.1186/s13059-020-02116-x
45. Granja JM, Corces MR, Pierce SE *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* 2021;53:403–11. https://doi.org/10.1038/s41588-021-00790-6
46. Salvi S, Sponza G, Morgante M *et al.* Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci USA* 2007;104:11376–81. https://doi.org/10.1073/pnas.0704145104
47. Zheng L, McMullen MD, Bauer E *et al.* Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in Zea mays. *J Exp Bot* 2015;66:3917–30. https://doi.org/10.1093/jxb/erv192
48. Engelhorn J, Snodgrass SJ, Kok A *et al.* Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize. *Nat Genet* 2025;57:2313–22. https://doi.org/10.1038/s41588-025-02246-7
49. Vinayak V, Basir R, Golloshi R *et al.* Polymer model integrates imaging and sequencing to reveal how nanoscale heterochromatin domains influence gene expression. *Nat Commun* 2025;16:3816. https://doi.org/10.1038/s41467-025-59001-z

Received: February 11, 2026. **Revised:** April 24, 2026. **Accepted:** May 17, 2026