Supplemental Files

Supplemental File 1.
Predictions across the test set for the human-trained model. Top row shows predictions, middle row shows target log(observed/expected) contact maps, and bottom row shows log10(observed) contact maps. Columns show results from left to right for: 0. HFF Micro-C, 1. H1hESC Micro-C, 2. GM12878 Hi-C, 3. IMR90 Hi-C, 4. HCT116 Hi-C. Note the different colormap for log10(observed) contact maps. Regions filtered due to insufficient coverage shown as white columns and rows on the log10(observed) maps; these rows are interpolated when preparing target log(observed/expected) contact maps (Online Methods).

Supplemental Notes

Supplemental Note 1: Previous approaches
Previous predictive modeling approaches for Hi-C matrices and features fall in two broad categories: machine learning and polymer-based. Some approaches draw on features of both. We focus on machine learning approaches in the main text because these are closer in spirit to our approach.

When we began this project, machine learning approaches for Hi-C prediction focused on predicting derived features of contact maps, including the positions of TAD boundaries and peaks\textsuperscript{9,10}. More recently, machine learning approaches have been adapted to predict quantitative contact frequencies\textsuperscript{7,8}. Importantly, due to the strong dependence of contact frequency on genomic distance in all Hi-C datasets, we and others focus on performance metrics for distance-corrected observed/expected maps. Zhang et al., 2019 report predictions with observed/expected correlations 0.3–0.5 at 5kb resolution on held-out chromosomes. Belokopytova et al., 2020 report observed/expected correlations between 0.53–0.72 at 25kb resolution on held-out chromosomes. Crucially, these approaches solve a distinct problem to the one we address: the goal of these approaches is to make the best predictions starting from epigenomic information (including ChIP-seq, expression, and open chromatin). These approaches cannot make predictions on unseen DNA sequences. In contrast, Akita starts with DNA sequence alone as an input and learns generalizable relationships between sequences and distance-corrected observed/expected contact maps. Hence, Akita can be used to predict contact maps for any DNA sequence, including sequences the model has never seen as inputs. Despite these differences in approaches, Akita achieves competitive \textasciitilde0.6 observed/expected correlations at higher 2048bp resolution.

Polymer approaches have also proven useful for predicting features of locus-specific genome folding. These approaches vary in the required inputs for the model, their resolution, and their intended use. One of the earliest approaches\textsuperscript{62} modelled a 324kb region at 3kb resolution, fitting to 5C data. More recently, MEGABASE+MiChroM\textsuperscript{63} modeled contact patterns of whole chromosomes from Hi-C at the 50kb scale, using epigenomic features as input. Qi and Zhang\textsuperscript{64}
modeled contact patterns from Hi-C for 20Mb and 50Mb regions at 5kb resolution, using chromatin states as input. PRISMR\textsuperscript{66} modeled a 6Mb region at 476bp resolution, achieving observed/expected correlations 0.4-0.6 at 10kb resolution. Importantly, while PRISMR allows for predicting the impacts of large-scale deletions and duplications, PRISMR parameters are specific to the region modeled and cannot be used to make predictions for unseen DNA sequences. Indeed, despite the insights gained from polymer approaches, there are still no approaches that can directly leverage DNA sequence as an input.

To our knowledge, only one other approach currently predicts Hi-C maps from DNA sequence, deepC\textsuperscript{35}, which we discuss in detail below. We highly recommend the discussion in\textsuperscript{7,8} for additional detail and insight into the many predictive modelling approaches for Hi-C data.

Supplemental Note 2: Lessons from previous architectures
Before arriving at the model described above, we considered various alternative designs in the space of possible model architectures and data pre-processing schemes.

Some of our early attempts at predicting 3D genome folding from sequence involved predicting slices of the Hi-C matrix (‘virtual-4C’) directly from the outputs of the trunk, with the idea that predicting a contact vector of length N could require fewer parameters in the ‘head’ than predicting a contact map of size NxN. While such virtual-4C models readily learned boundaries, we found they failed to learn sharp peaks. We also found that predictions of these models were often asymmetric (i.e. pred\textsubscript{i,j} \neq pred\textsubscript{j,i}), likely because the virtual-4C architecture we considered did not encode a symmetry constraint.

We also tested the performance of models that replace the dilated convolution layers in the trunk with bidirectional LSTMs, popular layers for capturing long-range dependencies in natural language processing, while preserving roughly the same time per training epoch. This architecture performed slightly worse on both training and validation data, and we did not pursue it further. We also explored the utility of separable convolutions to reduce the number of parameters in the convolutional tower, but found little benefit in the rate of learning and a slight loss in accuracy.

Supplemental Note 3: DeepC comparison
In Schwessinger et al.\textsuperscript{35}, the authors report successful predictions of Hi-C maps at 10kb resolution using a similar deep convolutional neural network approach, deepC. While deepC has a similar ‘trunk’ to Akita, it differs greatly in the architecture of the ‘head’, data pre-processing, and training schemes. First, deepC uses non-linearly quantile normalized targets, rather than log(observed/expected) targets, which may emphasize learning peaks rather than insulation. Second, deepC predicts a ‘zig-zag pole’ target directly with a dense layer from the output of their model’s trunk, which implicitly encodes distance but requires a large number of parameters, rather than predicting a dense patch of a map. Third, we focused on higher-resolution predictions (2048bp bins vs. 10kb bins). Finally, deepC currently requires pre-training a model on a large set of epigenomic profiles, and transferring weights to their full model. It is possible that strict transfer learning could limit the richness of representations that a deep CNN
can learn for 3D genome folding; for example, a CTCF profile may not contain information about the directionality of motifs under its peaks, which is important for predicting genome folding.

References

62. Giorgetti, L. et al. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. Cell 157, 950–963 (2014).
63. Di Pierro, M., Cheng, R. R., Lieberman Aiden, E., Wolynes, P. G. & Onuchic, J. N. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proc. Natl. Acad. Sci. U. S. A. 114, 12126–12131 (2017).
64. Qi, Y. & Zhang, B. Predicting three-dimensional genome organization with chromatin states. PLoS Comput. Biol. 15, e1007024 (2019).
65. Bianco, S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. 50, 662–667 (2018).