Text S1. Configurations for different sequence lengths and tokenizer strategies

We optimized the configuration of CLAMP before fine-tuning on chromatin loop prediction by evaluating two key pre-training hyperparameters: input sequence length and tokenizer strategy. For sequence length (Additional file 1: Fig. S13a), we tested ranges from 200 bp to 1,500 bp. While optimal length showed some task dependency (Additional file 1: Fig. S14), longer sequences (1,000 bp and 1,500 bp) generally achieved superior performance across multiple tasks. Since chromatin loop formation involves interactions between distant genomic elements, we selected 1,500 bp as the standard input length to capture long-range dependencies. For the tokenizer strategy (Additional file 1: Fig. S13b), we compared Byte Pair Encoding (BPE) [1] with k-mer methods using various k-values. Although BPE offers an adaptive vocabulary [2], it did not consistently outperform the best k-mer strategies (4-mer, 5-mer, 6-mer). K-mer tokenizer possesses a more direct biological interpretation (e.g., representing motifs) than BPE [3]. Among high-performing k-mer options, we selected 6-mer because it maintained peak performance while representing the longest local sequence pattern, potentially capturing transcription factor binding sites with greater precision.

Based on our systematic evaluation across genomics tasks and considering the requirements of chromatin loop prediction, we established an input sequence length of 1,500 bp and a 6-mer tokenizer as the standard CLAMP configuration for subsequent chromatin loop prediction. This optimized configuration provides a solid foundation for learning sequence features relevant to higher-order chromatin structure.

Text S2. Pre-training on vertebrate data benefits generalization to distant species

To determine whether CLAMP's performance on evolutionarily distant species is a result of generalizable patterns learned during pre-training or from the fine-tuning process alone, we conducted a targeted ablation study. We focused on species from lineages that were not represented in our pre-training dataset: invertebrates, plants, and fungi.

In this experiment, we compared the performance of the standard, pre-trained CLAMP model against a control model. For this control, the DNA Sequence Encoder's weights were not loaded from pre-training but were instead initialized using Xavier initialization [4]. To isolate the impact of the pre-trained weights, all other model components and fine-tuning hyperparameters, including batch size, epochs, and learning rate, were held constant across both conditions. The models were then fine-tuned and evaluated on the respective datasets.

The results demonstrate that pre-training provides a substantial advantage, even for these distant species (Additional file 1: Fig. S15). The pre-trained model outperformed the non-pretrained model across all three groups, achieving higher AUC scores for invertebrates (0.860 vs. 0.801), plants (0.954 vs. 0.889), and fungi (0.742 vs. 0.565). This indicates that the foundational "genomic grammar" learned from vertebrate chromatin accessibility data is transferable and provides a more robust and effective initialization for fine-tuning. These findings validate that the pre-trained model offers benefits beyond vertebrate species, thereby enhancing its ability to predict chromatin architecture across a wide range of eukaryotic organisms.

Text S3. CLAMP performance is dependent on genomic distance

To systematically evaluate CLAMP's performance across different genomic distance scales, we conducted an analysis where we stratified chromatin loops into three distinct categories based on the distance between two anchors. This classification follows the influential operational definition proposed for analyzing statistically significant Hi-C contacts [5]: short-range (<50 kb), medium-range (50 kb–10 Mb), and long-range (>10 Mb) interactions. We then assessed the MCC within each of these categories across the major eukaryotic lineages.

As illustrated in Additional file 1: Fig. S16, the results reveal that CLAMP's capabilities are highly dependent on the interaction distance. The model demonstrates robust and high-accuracy performance for both short-range and medium-range interactions. Across all tested groups, including mammals, other vertebrates, invertebrates, plants, and fungi. The MCC scores for these two categories were consistently high, indicating that the model effectively learns the sequence and epigenetic features that govern local and regional chromatin folding.

Conversely, the analysis also highlighted a significant limitation of the current model. The predictive power for very long-range interactions (>10 Mb) is substantially diminished, with MCC scores approaching zero in all species groups. This suggests that CLAMP is less effective at capturing the principles that govern extremely distant contacts, which may be influenced more by higher-order factors, such as chromosomal territory and subcompartment organization. This finding clarifies the model's optimal application scope and identifies the prediction of long-range interactions as a critical area for future research and model enhancement.

Text S4. Comparative analysis of negative sampling strategies

The construction of a high-quality negative dataset is critical for training robust machine learning models. To evaluate the impact of different negative sampling strategies on CLAMP's performance and generalization, we conducted a comparative experiment. We trained three models on the human CTCF-A673 dataset, each employing a distinct strategy for generating negative samples, while keeping all other parameters and data processing steps identical. The strategies were:

1. Adversarialness: As described in the Methods section (5.1.2), this is the anchor-matched approach used throughout our study. Negative pairs are constructed by sampling from the existing pool of positive anchor regions, creating challenging negative examples that are highly similar to positive loops in terms of genomic and epigenetic context.

2. Random: This method involves randomly selecting two anchor points on the same chromosome from regions outside of the positive loop set to form a negative pair.

3. Distal: This strategy creates negative pairs by randomly selecting two anchor points on the same chromosome that are separated by a large genomic distance (>10 Mb).

After training, we assessed the generalization capability of each model by evaluating its performance on the test set (constructed using the Adversarialness method) across five different human cell lines.

The results, shown in Additional file 1: Fig. S17, indicate that the choice of negative sampling strategy does influence model performance. The models trained with the Distal negative sampling strategy showed a minor decrease in predictive accuracy. This suggests that training on negative examples that are too dissimilar from positive ones may be less effective for learning the rules of loop formation.

Importantly, despite these variations, the overall performance of CLAMP remained high across all conditions. This highlights the robustness of the CLAMP framework, which we attribute primarily to the powerful sequence representations learned during the extensive pre-training stage. This pre-training provides a solid

foundation that stabilizes the fine-tuning process, making the model less sensitive to the specific details of the negative sampling scheme.

References

1. Gallé M: **Investigating the effectiveness of BPE: The power of shorter sequences.** In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2019: 1375-1381.
2. Balde G, Roy S, Mondal M, Ganguly N: **Adaptive BPE Tokenization for Enhanced Vocabulary Adaptation in Finetuning Pretrained Language Models.** *arXiv preprint arXiv:241003258* 2024.
3. Guo Y, Tian K, Zeng H, Guo X, Gifford DK: **A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction.** *Genome Res* 2018, **28:**891-900.
4. Glorot X, Bengio Y: **Understanding the difficulty of training deep feedforward neural networks.** *Journal of Machine Learning Research* 2010, **9:**249-256.
5. Ay F, Bailey TL, Noble WS: **Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.** *Genome Res* 2014, **24:**999-1011.