


METHOD

Open Access



# HTAD: a human-in-the-loop framework for supervised chromatin domain detection

Wei Shen<sup>1,2,3</sup>, Ping Zhang<sup>1,2</sup>, Yiwei Jiang<sup>1,2</sup>, Hailin Tao<sup>1,2</sup>, Zhike Zi<sup>3\*</sup>  and Li Li<sup>1,2\*</sup>

\*Correspondence:  
zk.zi@siat.ac.cn; li.li@mail.hzau.edu.cn

<sup>1</sup> College of Informatics, Huazhong Agricultural University, Wuhan, China

<sup>2</sup> Hubei Hongshan Laboratory, Hubei Key Laboratory of Agricultural Bioinformatics, Wuhan, China

<sup>3</sup> Shenzhen Key Laboratory of Synthetic Genomics, Guangdong Provincial Key Laboratory of Synthetic Genomics, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

## Abstract

Topologically associating domains (TADs) are essential units of genome architecture, influencing transcriptional regulation and diseases. Despite numerous methods proposed for TAD identification, it remains challenging due to complex background and nested TAD structures. We introduce HTAD, a human-in-the-loop TAD caller that combines machine learning with human supervision to achieve high accuracy. HTAD begins with feature extraction for potential TAD border pairs, followed by an interactive labeling process through active learning. Performance assessments using public curation and synthetic datasets demonstrate HTAD's superiority over other state-of-the-art methods and reveal highly hierarchical TAD structures, offering a human-in-the-loop solution for detecting complex genomic patterns.

## Background

Chromosome conformation capture techniques have reshaped our ability to address the spatial organization of the genome, reframing it as a global chromatin contact frequency problem. With the discovery of topologically associating domains (TADs) through Hi-C [1] and 5C [2] experiments, numerous computational approaches have emerged to identify these local modular structures from contact matrices [1, 3–7]. TADs, ubiquitous across eukaryotic species, play essential roles in transcriptional regulation with implications for phenotypic outcomes, including various diseases [8–10].

Despite the established methodologies for accurate signal detection and quantification in ChIP-seq and RNA-seq, the detection of TADs in nucleome data remains a challenging task [3, 4]. This difficulty stems from the intricate nature of nucleome data, which, unlike RNA-seq and ChIP-seq data, is two-dimensional and exhibits substantial heterogeneity in count distribution across genomic distances [11]. Moreover, Hi-C data is inherently noisy and contains intricate background. TADs cannot be viewed as isolated patterns due to potential interference from other features such as compartments, loops, strips, or even nested TADs, which makes their identification challenging due to large variations in interaction intensities.



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Existing TAD detection methods basically fall into two categories: those centered on local feature extraction and those based on statistical modeling of global features [1, 6, 12, 13]. Additionally, there are methods that leverage network modularity or clustering approaches [7, 14]. Despite the diversity in these unsupervised methods, their ability to identify TADs consistently across studies is limited [3], leading to a lack of consensus regarding the number of TADs present in the human genome [15]. This lack of agreement hinders our understanding of the regulatory role of TADs in gene expression and their underlying formation mechanisms. In addition, the numerous parameters in most TAD callers are often too complex for researchers to fine-tune, thereby limiting their broad applicability.

Supervised learning, often involving manual labeling, has proven indispensable for achieving optimal model performance in complex pattern recognition tasks, as demonstrated by various scenarios of imaging data analysis [16, 17]. Machines excel by leveraging knowledge acquired from human annotations. Furthermore, human-in-the-loop (HITL) strategy, which integrates human knowledge and experience into the learning process [18–20], outperforms random sample selection for labeling in machine learning endeavors [21]. In this context, we introduce HTAD (human-in-the-loop TAD caller) as a novel solution to the TAD identification problem. HTAD integrates Discriminative Active Learning (DAL), an effective supervised learning approach that trains a binary classifier to discriminate between the labeled or unlabeled samples [22]. This implementation is complemented by a web-based labeling tool, collectively forming an effective strategy for accurately identifying TADs with enhanced efficiency and adaptability.

The evaluation of HTAD highlights a notable improvement in performance achieved through active learning-based sample selection compared to the random sampling method. HTAD outperforms state-of-the-art methods by accurately capturing TAD structure features and identifying border signals associated with chromatin architecture protein bindings and epigenetic marks. Furthermore, when evaluated against an independent annotated dataset, HTAD demonstrates superior performance relative to other tools, highlighting the importance of manual labeling in TAD identification.

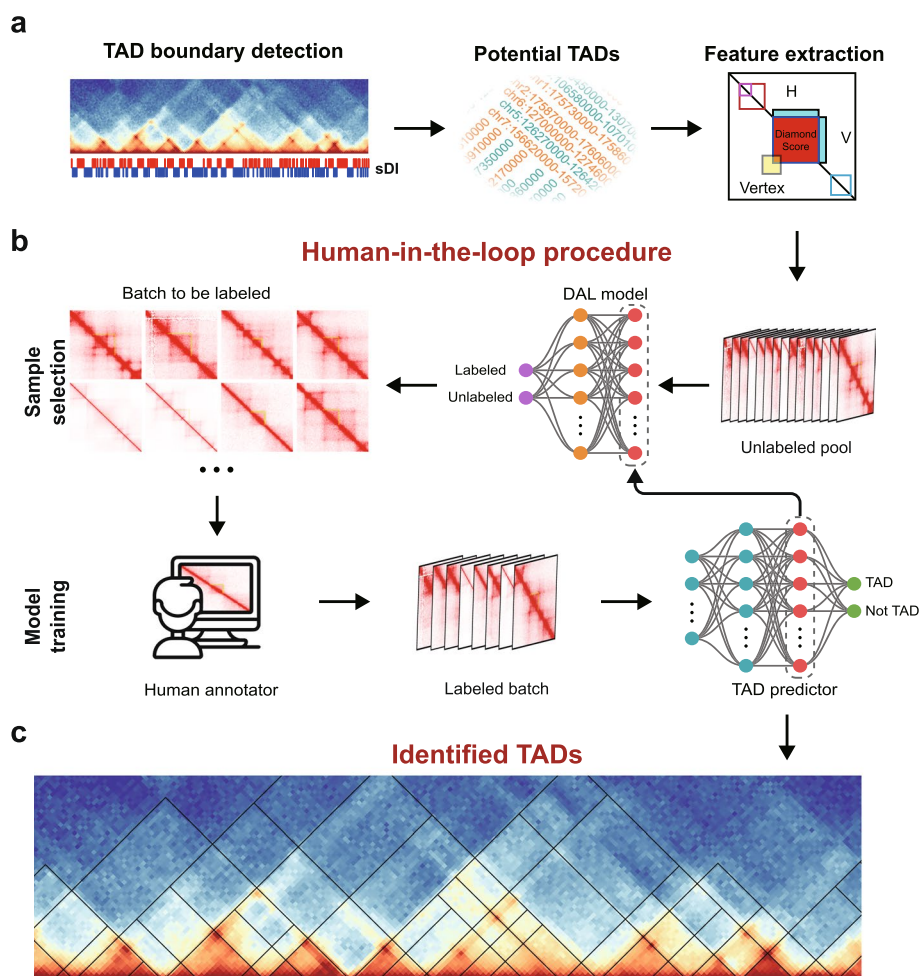
In summary, HTAD represents a novel HITL approach for identifying chromatin architecture based on contact information and manual labeling. The active learning scheme effectively enhances model performance for TAD identification with minimal labeling efforts, showcasing the superiority of supervised learning over unsupervised methods in complex pattern recognition tasks involving sequencing data.

## Results

### A framework for human-in-the-loop style TAD identification

The concept behind HTAD is to align the machine learning model with manual judgment by extracting TAD features and utilizing active learning. Thus, we integrate manual labeling into the TAD identification process of HTAD. Figure 1 illustrates the workflow of HTAD. In contrast to existing methods, this approach generates “what you see is what you get” (WYSIWYG) TAD results.

Before manual labeling and model training, it is necessary to construct an unlabeled sample pool containing predominantly all the potential TADs. To narrow down the range of TAD detection, HTAD initially identifies numerous potential TADs using



**Fig. 1** Schematic of the HTAD framework. **a** HTAD generates potential TADs based on the simplified Directionality Index (sDI). TAD features defined in the right box are then extracted from potential TADs. **b** A binary classification TAD model is trained using human-in-the-loop procedure. **c** After interactive labeling of potential TADs, HTAD identifies TADs at different resolutions by the well-trained TAD model

a simplified Directionality Index (sDI) (Fig. 1a, left). The sDI value only indicates the interaction tendency of each bin. Comparing with the original Directionality Index (DI), sDI increases the sensitivity on TAD boundary detection while sacrificing some accuracy (Additional File 1, Fig. S1). The enhanced sensitivity ensures the inclusion of nearly all positive TADs in the sample pool. Subsequently, the boundaries pairs within a 100-bin distance are combined to construct potential TADs (Fig. 1a, middle). These potential TADs serve as candidates for further model training and identification, significantly enhancing the computational efficiency and detection sensitivity of HTAD. HTAD then extracts features from these potential TADs based on the Hi-C matrices (Fig. 1a, right). These features are defined based on our understanding of TAD structure, including horizontal border strength ( $H$ ), vertical border strength ( $V$ ), vertex area ( $V_a$ ), and diamond score ( $DS$ ).

Following feature extraction, a TAD identification model is trained through iterations using the DAL method, which involves manual labeling via HTAD’s web-based interface

(Fig. 1b). The manual labeling process is straightforward, as users only need to select either “YES” or “NO” based on the provided potential TAD. For each round following the initial one, the DAL model selects the 50 most valuable unlabeled samples to be labeled based on the current TAD identification model. In our experiments, we trained the model over 11 rounds: one random round for initial training and ten DAL rounds with each round labeling 50 samples. The trained TAD identification model is then employed to filter potential TADs at resolutions of 10 kb, 20 kb, and 40 kb respectively. Finally, a merging strategy based on sDI is applied to integrate multi-resolution TAD results (Fig. 1c). Overall, this procedure enables HTAD to produce hierarchical WYSI-WYG TAD results with high sensitivity and accuracy.

### **Active learning enhances the training of TAD model**

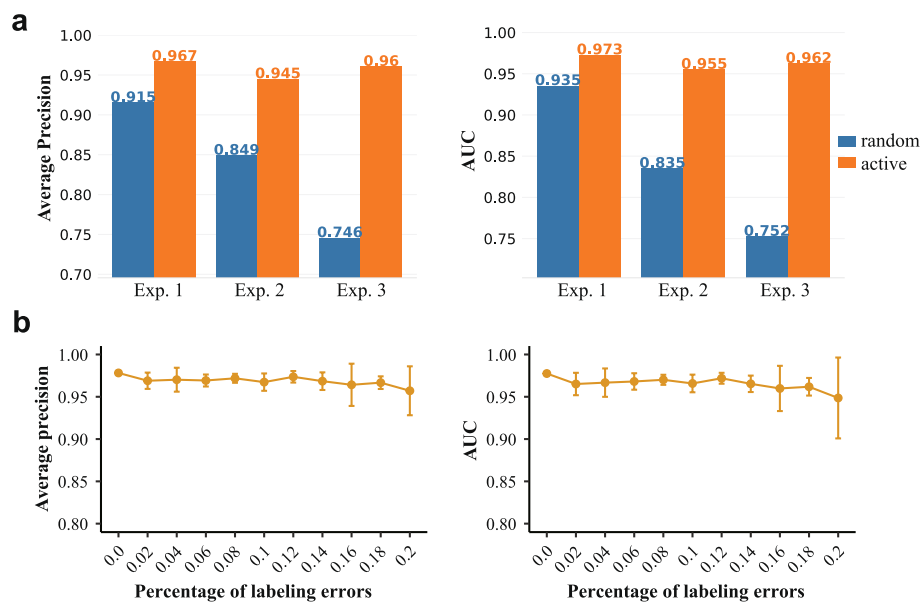
The web-based HTAD labeler can be used to curate potential TADs. However, achieving high precision through manual labeling of randomly selected candidates is impractical due to the extreme imbalance between positive and negative samples in a huge sample space. Therefore, it is necessary to selectively label the most representative potential TADs, aligning with the concept of active learning. In this work, we implemented DAL to enhance the training of the TAD model and achieved better performance than random sampling.

Using GM12878 Hi-C dataset, we compared the performance of models trained by random sampling against those trained with the DAL strategy based on our manually labeled test set, which contains 1550 annotated samples (841 positive samples and 709 negative samples). For each model comparison, both active learning and random sampling were initialized with the same random seed to ensure consistency in the zero round of training. In the three comparison experiments (Exp. 1–Exp. 3), the active learning strategy consistently outperformed random sampling, as evidenced by improved metrics for the area under the receiver operating characteristic curve (AUC) and average precision (Fig. 2a). These results suggest that HTAD achieves high performance through active learning. Moreover, by reducing the need for extensive human curation, active learning is more efficient than random sampling in supervising the model to achieve optimal performance.

To assess HTAD’s resilience to mislabeling errors, we intentionally introduced random manual errors into the labeled data for the GM12878 dataset. We assessed the model performance across varying mislabeling rates ranging from 0 to 20%. Remarkably, both average precision and AUC remained consistently stable (Fig. 2b), although a slight decline and increased variability were observed as the mislabeling rate increased. These results suggest that HTAD is robust against low to moderate level of labeling errors.

### **Performance comparison between HTAD and state-of-the-art methods**

We conducted an overlap comparison between the predicted TADs from HTAD and several state-of-art methods, including 3DNetMod [7], hicexplorer [23], TopDom [24], arrowhead [5], and rGMAP [6], using the GM12878 Hi-C dataset. To validate the effectiveness of sDI in HTAD, we also implemented a DI method in this analysis. Each method used the Hi-C data at 10 kb, 20 kb, and 40 kb resolution to identify TADs, which were then combined through BEDTools for overlap analysis [25]. The initial analysis of

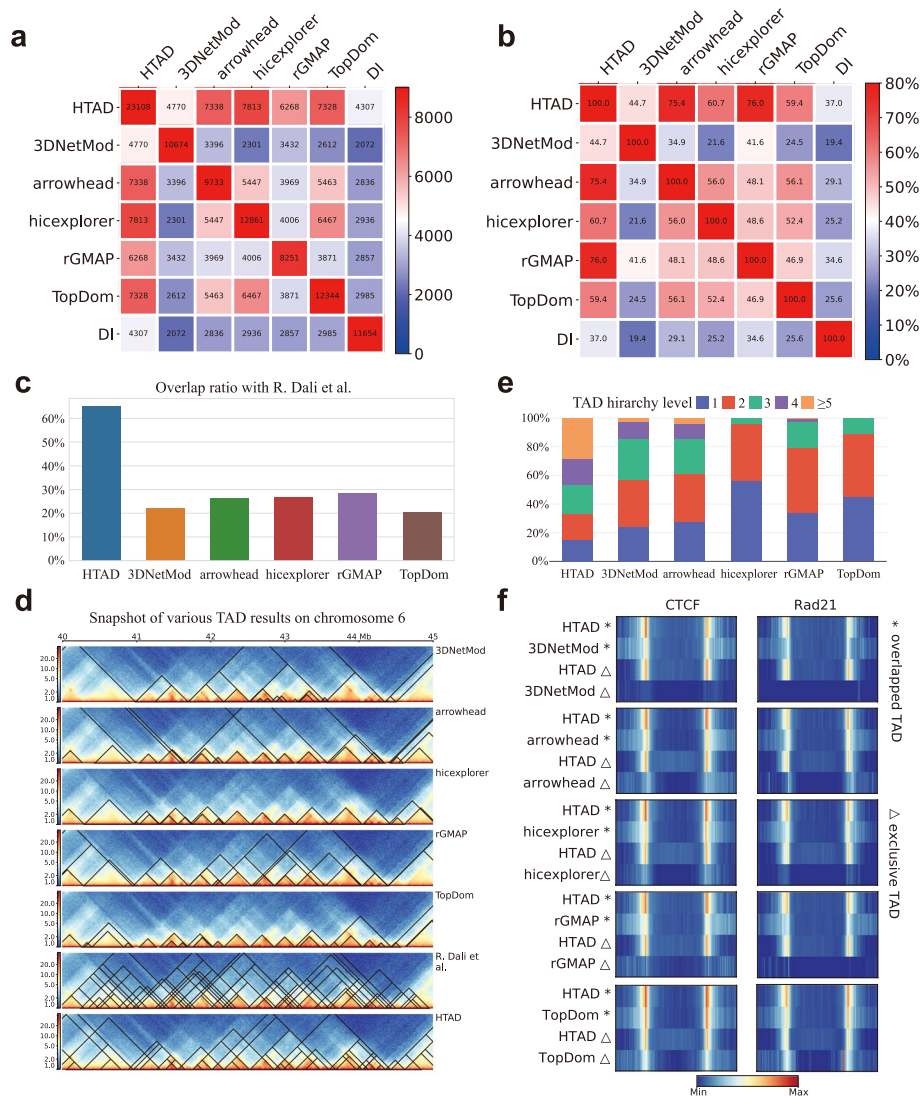


**Fig. 2** DAL facilitates TAD model training. **a** Average precision and AUC (area under the receiver operating characteristic curve) scores from 3 instances of random sampling and DAL learning. **b** Average precision and AUC scores cross different rates of labeling errors

the identified TADs revealed that HTAD identified a total of 23,108 TADs, substantially more than those identified by other methods (Fig. 3a). In addition, HTAD predicted TADs exhibited the highest overlap with those identified by other methods (Fig. 3b). To further assess the performance of the TAD callers, we compared the overlaps of the identified TADs against an independently annotated GM12878 TAD dataset from previous research [15]. As shown in Fig. 3c, HTAD achieved a 65% overlap, while the maximum overlap for other methods was just 28%. Figure 3d showcases one of the manually annotated regions together with results of various TAD callers, highlighting HTAD's superior performance in TAD identification.

In mammalian cells, TADs often exhibit nested structures, defined as high-level (or high-order) TADs encompassing low-level ones. Existing methods show a rapidly decrease in the number of identified TADs as their hierarchy level increases [4], which hinders accurate identification of higher-level TAD structures. Comparing distributions of TAD levels among these methods revealed that HTAD maintains a considerable proportion of TADs at 3 or more boundary levels (Fig. 3e), while other methods show a sharp decline in the number of identified TADs within the second or third levels. This result underscores HTAD's capability to detect TAD across different hierarchical levels. Additionally, the size distribution of identified TADs varies among methods and HTAD exhibited a TAD size distribution similar to that of TopDom (Additional File 1, Fig. S2).

Next, we analyzed the intersections and differences between HTAD and the other methods by examining enrichment levels of chromatin associated protein CTCF and Rad21 (Fig. 3f). HTAD showed the strongest signals at common TAD boundaries compared to other methods, indicating high accuracy in TAD boundary positioning. Furthermore, TADs uniquely identified by HTAD displayed strong signal enrichment, whereas those uniquely identified by other methods exhibited weak or no enrichment of



**Fig. 3** Comparison analysis of different TAD callers on GM12878. **a** The number of overlapped TADs identified by different methods. **b** The percentage of overlapped TADs between two TAD callers. The overlap percentage is defined as the number of overlapped TADs divided by the minimum number of identified TADs between two methods. **c** The percentage of overlapped TADs between different TAD callers and the manual annotation result from R. Dali et al. [15]. **d** Snapshot of TADs identified by different methods on chromosome 6: 40–45 Mb. **e** TAD order distribution of different methods. **f** CTCF and Rad21 enrichment over different TAD sets for comparison between HTAD and other methods. \* denotes the overlapped TAD set of corresponding methods,  $\Delta$  denotes the exclusive TAD set of corresponding methods

CTCF and Rad21 (Fig. 3f). We also calculated the enrichment (average peak number) of various regulatory elements around TAD boundaries using a 20-kb window. Among the tested methods, HTAD demonstrated the highest enrichment of CTCF, Rad21, SMC3, and H3K4me3, which are known as positive indicators of TAD boundaries (Additional file 2: Table S1).

To further evaluate HTAD’s performance, we extended our analysis to 7 additional Hi-C datasets generated from human lymphoblastoid K562 cells, human testis and ovary tissues, mouse CH12.LX, CH12F3, C57BL6 ESC cell lines, and *Xenopus tropicalis* brain

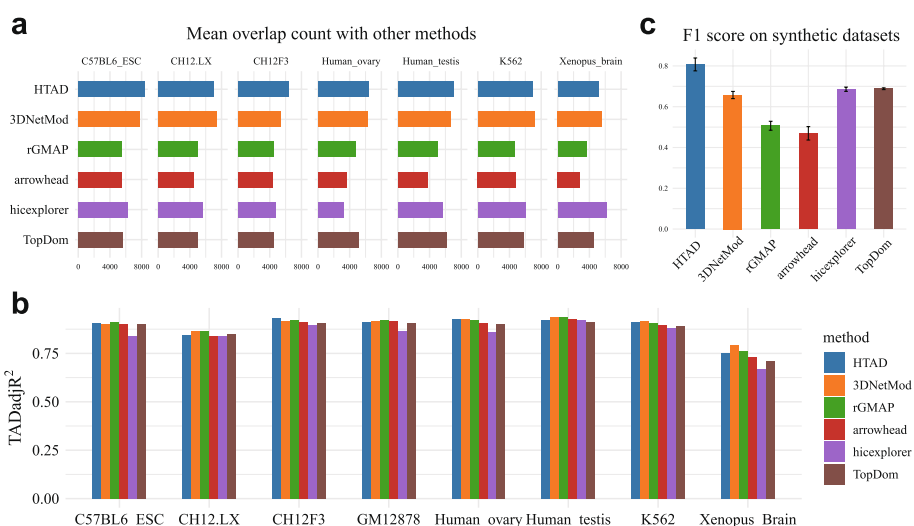
samples. Across these datasets, HTAD consistently showed high overlap counts with other methods (Fig. 4a). We propose that the accuracy of a TAD caller can be gauged by the degree of predictions overlap with other methods. The consistent robustness of HTAD’s overlap across diverse datasets highlights its reliability in TADs’ prediction. Furthermore, we utilized the TADadjR<sup>2</sup> score to evaluate TAD prediction accuracy, with a higher TADadjR<sup>2</sup> score indicates a stronger correlation between the TAD boundaries and the spatial decay of contact frequencies, suggesting more accurate TAD identification [26]. Notably, HTAD achieved high TADadjR<sup>2</sup> scores across various datasets (Fig. 4b).

Given the lack of ground-truth TAD structures in experimental Hi-C data, we benchmarked HTAD and other methods using simulated Hi-C data with nested TADs. Evaluations on four simulated Hi-C datasets demonstrated that HTAD outperformed all tested methods, achieving the highest F1 score (Fig. 4c). Collectively, our analyses indicate that HTAD can accurately predict TAD structures across multiple datasets.

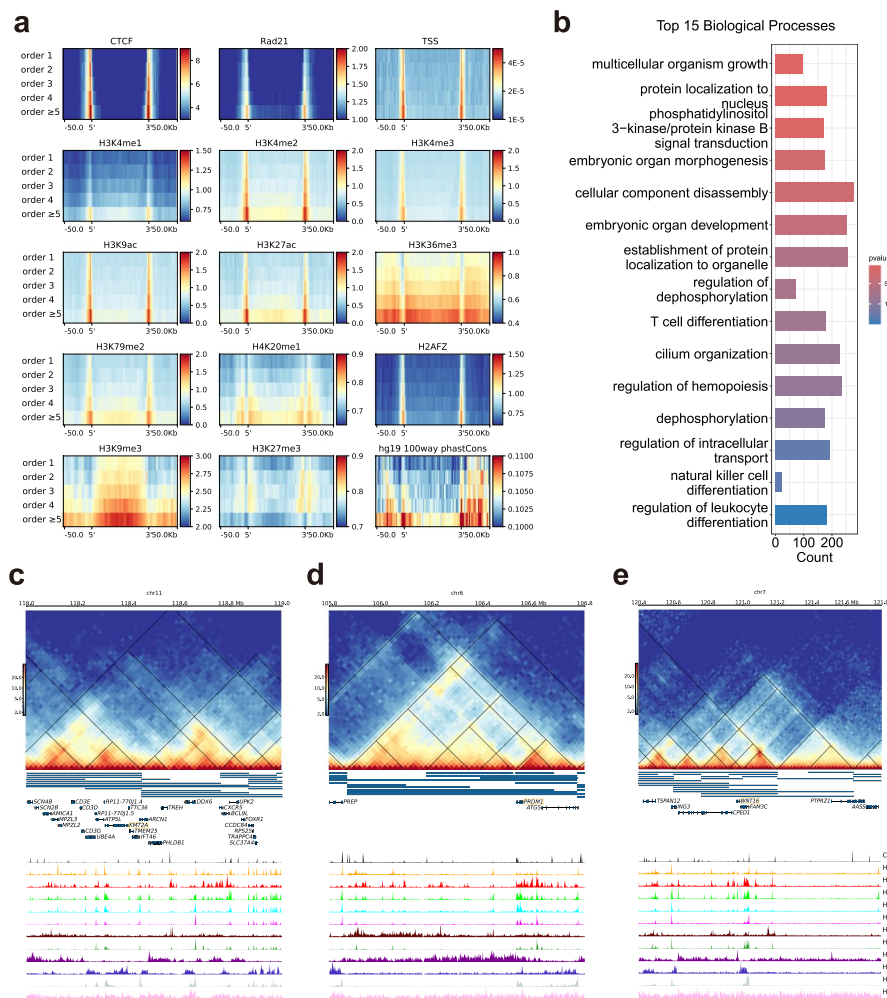
Finally, we investigated the cross-sample performance of HTAD by predicting TADs for new Hi-C datasets with the model pre-trained on the GM12878 dataset. This model produced approximately 70% overlap in TAD predictions with those predicted from models trained independently on the new datasets (Additional File 1, Fig. S3), indicating HTAD’s strong cross-sample performance and low risk of overfitting. For optimal TAD identification, we recommend that users train the model on their specific datasets.

### The correlation between TAD hierarchy and chromatin features

Previous studies have suggested that TADs with higher hierarchy tend to exhibit more active epigenetic signals compared to those with lower hierarchy [27]. In our investigation using GM12878 dataset, we examined signals of various histone modifications to validate this observation (Fig. 5a).



**Fig. 4** Comparison analysis of different TAD callers on various datasets. **a** Mean overlap of TAD numbers among different TAD prediction methods for seven Hi-C datasets from three species. **b** TADadjR<sup>2</sup> values for eight Hi-C datasets on TADs identified by each TAD prediction method. **c** F1 score achieved by different methods on simulated Hi-C data



**Fig. 5** Characteristics and related genes of high-order TADs. **a** Enrichment of CTCF, RAD21, and histone modifications cross different orders of TADs. **b** Biological process enrichment of genes near high-order ( $\geq 5$ ) TADs. **c–e** Example regions of genes (KMT2A, PRDM1, WNT16) in high-order TADs. TADs are outlined as black lines in the heatmap, also as stacked horizontal blue bars below the heatmap. The gene annotation and signals of CTCF, RAD21, and histone modifications are depicted below the heatmap

Firstly, we observed that the enrichment of chromatin associated factors (CTCF and RAD21) at TAD boundaries increases with the TAD order (the innermost one is referred to as order of 1 in a hierarchy), suggesting a relationship between hierarchical TADs’ formation and the binding strength of CTCF and cohesin. Similarly, the density of transcription start sites (TSS) and the enrichment of active histone modification markers (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, and H3K79me2) also show a similar trend. In contrast, the enrichment of repressed markers (H3K9me3 and H3K27me3) decreases near TAD boundaries as the TAD order increases. Interestingly, while in H3K9me3 levels within TADs increase with the TAD order, this pattern is not observed for H3K27me3. Additionally, the H4K20me signal, a key regulator of genomic integrity [28], exhibits an increasing trend in its enrichment at both the boundaries and within the body of TADs as their order grows. Moreover, the histone variant H2A.Z, which facilitates the activation of early replication



origins and promote the deposition of H4K20me2 [29, 30], exhibits enrichment at high-order TADs.

We then analyzed the conservation of sequences across TADs of different orders using the PhastCons conservation scores (obtained from UCSC 100-way [31]). Our analysis reveals that higher-order TADs tend to maintain a more conserved sequence than lower-order TADs (Fig. 5a). Considering the results of histone modifications, we speculate that genes covered by high-order TADs hold functional significance. To validate this hypothesis, we conducted biological process Gene Ontology (GO) enrichment analysis of these genes (covered by 5- or higher-order TADs). GO analysis indicated that several biological processes, including “multicellular organism growth,” “protein localization to nucleus,” “embryonic organ morphogenesis,” and “embryonic organ development,” were enriched with genes surrounding high-order TADs (Fig. 5b).

Interestingly, we found that gene *KMT2A* is typically wrapped into high-order TAD (Fig. 5c), whose copy gains and break aparts are associated with CTCF depletion and reduced binding [32]. *KMT2A* amplifications and translocations are prevalent in infant, adult, and therapy-induced leukemia. HTAD effectively delineates the intricate chromatin interaction context surrounding the *KMT2A* gene, as shown in Fig. 5c. This observation implies that the instability of *KMT2A* associated with CTCF may be attributed to subsequent alterations in chromatin conformation.

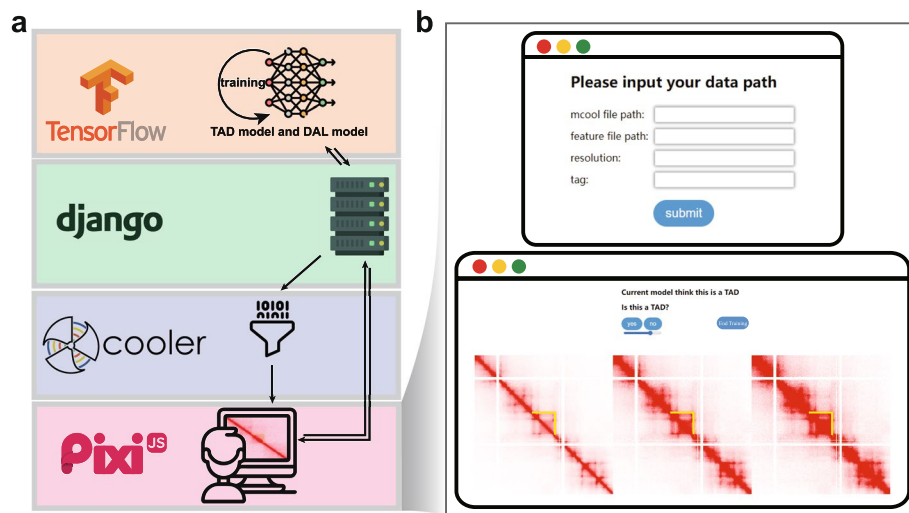
We next analyzed the chromatin conformation and epigenetic modifications context over other functionally important genes (*PRDM1*, *WNT16*, *RC3H2*, *TBC1D23*, *ALX4*, *KCNQ1*, *FOXC1*) (Fig. 5d, e and Additional File 1, Fig. S4-S8). These results are important for understanding the regulation relationship between genes and other functional elements such as enhancers. Notably, recent research indicates that genes sharing the same CTCF TADs do not exhibit increased transcriptional similarity beyond what is predicted by linear genome proximity, and functionally important genes are often situated within TADs independently [33]. Our results suggest that highly hierarchical TADs may shield the encompassed sequences enriched with functionally important genes from mutations or facilitate intricate gene regulatory processes, which aligns with observed association between hierarchical TADs and transcriptional abnormalities in cancer [34].

In summary, our HTAD analysis using the GM12878 dataset confirms that higher-order TADs exhibit more active epigenetic signals. In addition, higher-order TADs exhibit greater sequence conservation and are associated with genes involved in key biological processes, suggesting a potential role in gene regulatory processes.

### The implementation and user interface of HTAD

The implementation and user interface of HTAD are organized into four functional modules: data extraction, front-end labeler, back end, and machine learning, rendering HTAD a comprehensive tool and potential framework for identifying various Hi-C data-related structures (Fig. 6a).

For data extraction, we utilized the widely used cooler library to efficiently retrieve Hi-C data from a cooler format file [35]. We improved computational steps, including sDI calculation and feature extraction from extensive potential TADs, by converting the pixel format data of cooler into Dask’s parallel DataFrame object [36].



**Fig. 6** Architecture diagram and user interface of HTAD. **a** Architecture diagram depicts the modules employed in HTAD. **b** User interfaces of HTAD web-based labeler. After the submission of data path, the browser will present a labeler interface with button, color bar slider, and heatmaps corresponding to three different saturations. The yellow line in heatmap marks the TAD position

In the machine learning module, we used Tensorflow [37] as the framework to implement the TAD multilayer perceptron (MLP) model and DAL method. Rendering Hi-C data fetched from back-end to zoomable and draggable heatmap was facilitated by PixiJS [38]. The back end of web labeler, constructed using the django [39] python library and its extension library Channels [40], communicates via WebSocket protocol. During the DAL process, potential TADs are labeled using a simple “yes/no” button on the webpage, enabling users to train the TAD model effectively (Fig. 6b, c). In addition, HTAD provides three images at different saturation levels and incorporates a color slider, enhancing users’ intuitive judgement (Fig. 6c).

## Discussion

In this study, we introduced HTAD, a novel approach for identification of the TAD architecture from Hi-C data. Our approach integrates preliminary screening, manual feature extraction, deep learning, and active learning. Through comparison with several state-of-the-art methods, we demonstrated that HTAD is accurate and reliable, outperforming existing methods in four key metrics: sensitivity, accuracy, ability to detect hierarchical TADs, and boundary precision.

Unlike other methods that rely on artificially setting thresholds for parameter determination, our approach uses manual labeling and training, which alleviates issues related to parameter adjustment. HTAD employs Discriminative Active Learning (DAL) strategy, which iteratively selects the most informative samples for annotation, thus minimizing the time required for manual labeling and model training. As shown in Additional file 3: Table S2, each round of manual labeling takes approximately 1–2 min across our tested datasets, with the entire process of labeling and model training taking less than 20 min. Our feature extraction method, similar to arrowhead [5], adjusts the extraction space based on TAD size, mitigating potential bias in resultant

TAD sizes. Furthermore, our feature extraction and TAD model structure allow for expansion and optimization to achieve more accurate and refined TAD identification in future studies.

The “easy in, strict out” TAD identification strategy we implemented minimizes invalid calculations while maintaining sensitivity. This strategy provides insights that can be applied to other analyses on the three-dimensional genome, such as compartment, differential interaction sites, or structural variations across conditions. Through empirical testing, we found that 11 labeling rounds yield optimal performance for HTAD. While 11 rounds are recommended, users should evaluate the model’s prediction quality to determine whether more rounds may need. Future work could explore adaptive stopping mechanisms to further enhance the usability and efficiency of HTAD. Moreover, our tool has the potential to simplify the training process further through the recent work of zero-round active learning [41]. In addition, Our TAD merging strategy effectively maintains TAD boundary accuracy at high resolutions while integrating multi-resolution TAD sets. This approach may serve as a reference for methods aiming at detecting TAD structures across multiple resolutions.

HTAD establishes a framework for TAD identification, enabling researchers to fine-tune the TAD model with its labeling facilities to suit specific data contexts. Given the absence of a gold standard dataset with labeled TADs for benchmarking, it will be valuable to accumulate a large, verified collection of TADs to create a comprehensive training set and provide additional measurement indices for thorough evaluation of TAD methods. These efforts will facilitate the development of new methods for identifying TADs and enhance understanding of the role of chromatin conformation in biological processes.

While the HITL approach has been applied in image recognition for biomedical engineering [42, 43], its application in genomic data analysis is still underexplored. Here, we demonstrated how human involvement in model training can improve the feature recognition ability in sequencing data. More broadly, integrating HITL into all aspects of TAD identification process, such as feature extraction, model training, and result feedback, holds promise for improving the accuracy of TAD identification. The in-depth embedding of human prior knowledge into machine learning models will provide effective guidance and supervision for the generalization of these models and generate more accurate TAD identification results.

## Conclusions

In this study, we introduce HTAD, a new HITL framework that enhances the supervised detection of TAD structure from Hi-C data through the integration of active learning and manual labeling. Our evaluation across multiple datasets indicates that HTAD performs well in terms of sensitivity and accuracy while providing a user-friendly experience that mitigates the challenges of parameter tuning found in traditional unsupervised approaches. Overall, HTAD underscores the importance of integrating human-machine interaction to enhance machine learning outcomes for complex biological challenges. We anticipate that HTAD will become a valuable resource for researchers aiming for more accurate and profound analyses of chromatin organization.

## Methods

### Feature extraction for TAD

To determine whether a region of the genome qualifies as a TAD, we defined four representative TAD features: horizontal border strength ( $H$ ), vertical border strength ( $V$ ), vertex area ( $Va$ ), and diamond score (Fig. 1). These features are extracted from the observed/expected matrix, computed using the Hi-C matrix normalized by iterative correction and eigenvector decomposition (ICE) [44]. Horizontal border strength represents the difference between mean values of the inner and outer side of a potential TAD's horizontal boundary (boundary width: 2). Similarly, vertical border strength is defined for the vertical boundary. Vertex area refers to a  $5 \times 5$  square area centered on the vertex of a potential TAD. When extracting  $Va$ , the  $5 \times 5$  matrix will be normalized by subtracting the contact value of the corresponding vertex. Hence  $Va$  has a length of 24 after discarding the vertex point. The diamond score was calculated following a previous study [45]. Each TAD feature has a total length of 28 units ( $H(1) + V(1) + Va(24) + DS(1) + \text{TAD size}(1) = 28$ ). By employing manual feature extraction, as opposed to automated methods like convolutional neural network (CNN), we streamlined the general framework with a deep understanding of TAD structure.

### Generation of potential TAD

The directionality index (DI) is a classic signal for effectively identifying TAD borders, while the sign of the DI value indicates the contact tendency of corresponding bin. In this study, we introduce a simplified version of DI ( $sDI$ ), which solely represents the interaction tendency ( $-1$  for upstream,  $0$  for neutral,  $1$  for downstream) of corresponding bin without normalization (see formula 1, where  $A$  and  $B$  refer to the interaction between corresponding bin and its upstream and downstream). This improves the sensitivity for calling potential TAD borders (Additional File 1, Fig. S1). To ensure robustness in border detection and enhance noise immunity, HTAD calculates  $sDI$  using five window sizes (ranging from 7 to 12 bins). Results obtained from different window sizes are merged to represent overall contact tendencies for potential border detection.

$$sDI = \frac{(B - A)}{|B - A|} \quad (1)$$

Based on our best practice, a bin is considered as a potential border when two conditions are met: the contact tendency changed from upstream ( $1$ ) to downstream ( $-1$ ), and the previous 2 bins should both tend to contact more with upstream bins.

To expedite these calculations, we employed the Python library Dask to parallelly compute the  $sDI$ s based on Hi-C contact data extracted from a cooler format file.

Following the identification of potential TAD boundaries, we generated combinations of these boundaries within a distance range of 7–100 bins to form potential TADs. Subsequently, potential TADs exhibiting negative values for  $H$  or  $V$  (refer to manual feature extraction for TAD) were filtered. While most potential TADs may be

inauthentic, they tend to encompass all correct results and provide an initial screening outcome. This not only ensures heightened sensitivity of HTAD but also yields an appropriate number of TAD samples (around 140,000 TADs for the human genome) suitable for active learning.

#### Active learning-based sample selection

Our TAD MLP model is a  $28 \times 64 \times 16 \times 2$  fully connected neural network: 28 input features through two hidden layers with 64 nodes and 16 nodes and then spread to the output layer for classification. The sigmoid activation function governs the TAD model, with all machine learning algorithms within HTAD developed using the Tensorflow framework [37]. The number of epochs is set to 1000, and a batch size of 20 is assigned for each round of model training.

During each round of model training, excluding the first random round, we used the DAL [22] algorithm to fetch unlabeled samples. DAL poses active learning as a binary classification task, attempting to choose examples to label in such a way as to make the labeled set and the unlabeled pool indistinguishable. After each round of model training, the DAL model will be trained based on the current TAD identification model and the pool of labeled/unlabeled samples. Although there exists a multitude of potential TADs, DAL has the capability to meticulously choose the most valuable samples for manual labeling in each subsequent round.

Formally, with TAD identification model  $\Psi : \mathcal{X} \rightarrow \hat{\mathcal{X}}$  being a mapping from the original input space to some learned representation, DAL is defined as a binary classification model with  $\hat{\mathcal{X}}$  as input space and  $\mathcal{Y} = \{l, u\}$  as its label space, where  $l$  is the label for a potential TAD being in the labeled set and  $u$  is the label for the unlabeled set. As shown in the active learning part of Fig. 1, the layer of TAD model with red nodes was extracted as learned representation for DAL model.

For every iteration of the active learning process, the classification problem is solved by minimizing the log loss and obtaining a model  $\hat{P}(y|\Psi(x))$ . The top 50 potential TADs that satisfy  $\underset{x \in \mathcal{U}}{\operatorname{argmax}} \hat{P}(y = u|\Psi(x))$  are selected (Fig. 1, active learning part).

For each potential TAD, its mirrored counterpart (a TAD with reversed 5' and 3' ends) should possess the same label. Consequently, we incorporate these mirrored TADs into the training set to expedite the model training process.

#### Web-based interactive labeler

HTAD features a fast and lightweight interactive labeler, accessible via the web and powered by the WebSocket protocol. This protocol allows for bidirectional communication, ensuring real-time communication, and surpasses http and https protocols for efficient and seamless data labeling processes.

The HTAD labeler server is built on Python's Django framework [39]. The WebSocket communication between the server and client labeler is established through the utilization of the Django Channels package [40]. Upon connection establishment, the server initializes a MLP model for TAD detection and a DAL model for unlabeled sample selection. Initially, 50 random potential TADs are sent from the server to the client for manual labeling. After each labeling round, the server trains both the TAD detection model and updates the DAL model based on this trained TAD model. The refreshed DAL

model then identifies 50 most valuable unlabeled potential TADs for subsequent manual labeling.

The client web page utilizes PixiJS [38] for Hi-C heatmap rendering and pixiViewPort [46] for draggable and zoomable heatmap functionality. Additionally, we have incorporated a slider into the labeler to regulate the saturation level of the heatmap. Users can designate whether the current potential TAD is a genuine by either clicking on the button or typing the corresponding keyboard shortcut: Q for positive and W for negative.

### TAD calling

We evaluated our model using Hi-C matrices at resolutions of 10 kb, 20 kb, and 40 kb, revealing its robust performance across these resolutions. Following best practices, we employed HTAD to identify TADs in datasets with resolutions of 10 kb, 20 kb, and 40 kb using the model trained on 10 kb resolution. To ensure performance robustness, we introduced a hyperparameter, denoted as “ $n$ ,” which dictates the number of TADs to be extracted from a genome at specific resolution. For the human genome analysis, we extract 20,000, 10,000, and 5000 TADs from matrices with resolutions of 10 kb, 20 kb, and 40 kb, respectively, based on TADs’ probabilities ranking. A higher “ $n$ ” value means more TADs to be detected. Subsequently, the multi-resolution outcomes are merged using HTAD’s TAD merging strategy (see below).

### Strategy for merging multi-resolution TADs

TAD calling at low resolution can compromise the accuracy of TAD boundaries compared to high-resolution outcomes. To address this challenge, we propose a novel and highly effective strategy for merging TAD calling results from multiple resolutions.

Firstly, we introduce the concept of the DI check value (DCV) for the highest resolution of 10 kb. The DCV represents a quantification measure in our merging process and is defined as:

$$DCV = \sum_{k=i}^{i+2} sDI[k] - \sum_{k=i-3}^{i-1} sDI[k] \quad (2)$$

When there is a stronger tendency for upstream or downstream bins to interact in their respective directions, a bin is more likely to serve as a TAD boundary and will exhibit an increased DCV.

Secondly, we merged the boundaries of 5’ and 3’ separately. When two or more boundaries are in proximity (within 6 bins), they will be consolidated into a single optimal boundary. The bin with the highest DCV is selected as the most favorable boundary.

In this way, HTAD’s TAD merging approach maintains precise boundaries at the highest resolution while seamlessly integrating results from multiple resolutions.

### The generation of simulated Hi-C data

We generated simulated Hi-C data for a single chromosome (length 100Mbp) at 10 Kb resolution, using a modified method based on Lun and Smyth [47]. The simulation

included two types of interactions between bins of 10 kb resolution, including TAD interactions and non-specific background interactions.

For TAD interactions, we generated consecutive TADs with width randomly distributed within the intervals [8, 30]. To establish nested TADs, we randomly merged 50% of two adjacent TADs to form larger TADs. This process was repeated twice to form a TAD set with up to 3 layers. The size of TAD was defined with an upper limit of 100. Each entry  $(x, y)$  within TAD was assigned to a value of  $k_t(x - y + p)^c$ , where  $k_t = 80/3$ ,  $p = 1$ , and decay rate  $c = -0.8$ .

Non-specific interactions were simulated by randomly selecting 200000 bin pairs from all possible combinations  $((x, y))$ , where  $x \neq y$ . The selection probability of each pair  $(x, y)$  was proportional to  $(x - y + p)^c$ , maintaining the distance-abundance relationship. Each selected entry was assigned a constant value 5 to represent the non-specific ligation. Additionally, the diagonal of interaction matrix was set to a value of 100.

After constructing the simulated Hi-C matrix, we utilized cooler [35] to generate a normalized .cool file for further analysis. To evaluate the performance of each method, we calculated F1 score using the following formula:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

#### Analysis on the impact of labeling errors in the GM12878 dataset

We introduced a fraction of randomly labeling errors, ranging from 2 to 20%, into the manually labeled data of the GM12878 dataset. To evaluate the impact of these errors, we utilized the GM12878 test set to assess the area under the curve (AUC) and average precision scores across different fractions of labeling errors. Each fraction of labeling errors was tested in ten independent repetitions to ensure statistical reliability.

#### Comparison of TAD sets from different methods

To assess the overlap ratio between TAD sets generated by various methods, we used the intersect command of bedtools (v2.26.0) [25]. TADs were considered as overlapping if their shared region accounts for at least 80% of each domain's size, as determined by the bedtools "intersect -f 0.8 -r" command. This criterion was applied to generate the intersection sets and differences of sets. For signal enrichment analysis of these TAD sets, we employed deepTools [48] to compute and visualize the enrichment heatmap.

To compute the average number of regulatory element peaks around TAD boundaries, we used intersect command of bedtools with "-c" option. Each TAD boundary was expanded to a 20-Kb region. For the index of  $TADadjR^2$ , we applied the R script from Liu et al. [26] to measure each TAD set using 40 Kb resolution interaction data within a 1.5 Mb distance.

#### Gene ontology analysis

Genes within 5- or higher-order TADs were identified using the intersect command of bedtools. Biological process enrichment analysis was performed by clusterProfiler [49]

with annotation org.Hs.eg.db [50]. PyGenome Tracks [51] was used for the plotting of multivariate genomic datasets for genes covered by high-order TADs.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03445-x>.

Additional file 1. Supplementary figures.

Additional file 2: Table S1. The average peak number of ten regulatory elements surrounding TAD boundaries detected by various methods.

Additional file 3: Table S2. Time cost of simulated dataset and human dataset.

Additional file 4: Table S3. Accession number of public datasets (ENCODE and UCSC).

Additional file 5. Review history.

### Acknowledgements

We thank Kai Li and Jinsheng Xu for their discussions and suggestions on this work.

### Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Review history

The review history is available as Additional file 5.

### Authors' contributions

Conceptualization: L. L. W.S. and P.Z.; investigation: W.S.; methodology: W.S.; software development: W.S.; software test: W.S., Y.J. and H.T.; visualization: W.S. and H.T.; project administration: L.L.; resources: Z.Z. and L.L.; supervision: Z.Z. and L.L.; funding acquisition: Z.Z. and L.L.; writing—original draft: W.S.; writing—review and editing: Z.Z. and L.L.

### Funding

This work was supported by National Natural Science Foundation of China (Grant No. 32470661 to L.L.), Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (to L.L.), Guangdong Provincial Key Laboratory of Synthetic Genomics (2023B1212060054), and Shenzhen Key Laboratory of Synthetic Genomics (ZDSYS201802061806209).

### Data availability

HTAD is available in Zenodo repository (<https://doi.org/10.5281/zenodo.13822061>) [52] and public GitHub repository (<https://github.com/shenscore/HTAD>) under MIT license. Manually annotated TADs for GM12878 dataset is available at Zenodo (<https://zenodo.org/records/14186235>) [53].

Processed Hi-C contacts of GM12878 and K562 cell lines were downloaded from Gene Expression Omnibus under accession number: GSE63525 [54]. Additionally, Hi-C data for human testis, human ovary, mouse CH12LX, mouse CH12F3, and mouse C57BL6 ESC were downloaded from ENCODE project [55] (detailed see Additional file 4: Table S3). Hi-C data for the *Xenopus tropicalis* brain tissue was obtained from the BioProject database under accession number PRJNA606649 [56].

The fold change over control signals of CTCF, Rad21, and histone modifications were downloaded from ENCODE project (detailed see Additional file 4: Table S3). ChIP-seq peaks for GM12878 cell line were downloaded from UCSC Genome Browser [31] (detailed see Additional file 4: Table S3). BED files for TSSs and SINEs of hg19 genome were downloaded from UCSC Genome Browser. Housekeeping genes [57] were aligned to the hg19 genome. The PhastCons score file in bigwig format was downloaded from UCSC Genome Browser.

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Competing interests

The authors declare no competing interests.

Received: 19 April 2024 Accepted: 22 November 2024

Published online: 02 December 2024

### References

1. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485:376–80.
2. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485:381–5.



3. Sefer E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics*. 2022;23:127.
4. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol*. 2018;19:217.
5. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159:1665–80.
6. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nat Commun*. 2017;8:535.
7. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, et al. Detecting hierarchical genome folding with network modularity. *Nat Methods*. 2018;15:119–22.
8. Deng S, Feng Y, Pauklin S. 3D chromatin architecture and transcription regulation in cancer. *J Hematol Oncol* *Hematol Oncol*. 2022;15:49.
9. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet TIG*. 2016;32:225–37.
10. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161:1012–25.
11. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res*. 2017;27:1939–49.
12. Chen F, Li G, Zhang MQ, Chen Y. HiCDB: a sensitive and robust method for detecting contact domain boundaries. *Nucleic Acids Res*. 2018;46:11239–50.
13. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*. 2015;523:240–4.
14. Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. *BMC Bioinformatics*. 2017;18:480.
15. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res*. 2017;45:2994–3005.
16. Pachitariu M, Stringer C. Cellpose 2.0: how to train your own model. *Nat Methods*. 2022;19:1634–41.
17. Anzani M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. *Med Phys*. 2020;47:e185–202.
18. Amershi S, Cakmak M, Knox WB, Kulesza T, Lau T. IUI workshop on interactive machine learning. *Proc Companion Publ 2013 Int Conf Intell User Interfaces Companion*. Santa Monica California USA: ACM; 2013 [cited 2024 Mar 7]. p. 121–4. Available from: <https://dl.acm.org/doi/https://doi.org/10.1145/2451176.2451230>.
19. Amershi S, Cakmak M, Knox WB, Kulesza T. Power to the people: the role of humans in interactive machine learning. *AI Mag*. 2014;35:105–20.
20. Kumar V, Smith-Renner A, Findlater L, Seppi K, Boyd-Graber J. Why didn't you listen to me? Comparing user control of human-in-the-loop topic models. *Proc 57th Annu Meet Assoc Comput Linguist*. Florence, Italy: Association for Computational Linguistics; 2019 [cited 2024 Mar 7]. p. 6323–30. Available from: <https://www.aclweb.org/anthology/P19-1637>.
21. Chandler C, Foltz PW, Elvevåg B. Improving the applicability of ai for psychiatric applications through human-in-the-loop methodologies. *Schizophr Bull*. 2022;48:949–57.
22. Gissin D, Shalev-Shwartz S. Discriminative active learning. 2019 [cited 2023 Nov 28]; Available from: <https://arxiv.org/abs/1907.06347>.
23. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 2018;9:189.
24. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*. 2016;44: e70.
25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl*. 2010;26:841–2.
26. Liu K, Li H-D, Li Y, Wang J, Wang J. A comparison of topologically associating domain callers based on Hi-C data. *IEEE/ACM Trans Comput Biol Bioinform*. 2023;20:15–29.
27. An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, et al. OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol*. 2019;20:282.
28. Jorgensen S, Schotta G, Sorensen CS. Histone H4 Lysine 20 methylation: key player in epigenetic regulation of genomic integrity. *Nucleic Acids Res*. 2013;41:2797–806.
29. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. *Nature*. 2020;577:576–81.
30. Lee CSK, Weiß M, Hamperl S. Where and when to start: regulating DNA replication origin activity in eukaryotic genomes. *Nucleus*. 2023;14:2229642.
31. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res*. 2023;51:D1188–95.
32. Gray ZH, Chakraborty D, Duttweiler RR, Alekbaeva GD, Murphy SE, Chetal K, et al. Epigenetic balance ensures mechanistic control of MLL amplification and rearrangement. *Cell*. 2023;186:4528–4545.e18.
33. Long HS, Greenaway S, Powell G, Mallon A-M, Lindgren CM, Simon MM. Making sense of the linear genome, gene function and TADs. *Epigenetics Chromatin*. 2022;15:4.
34. Du G, Li H, Ding Y, Jiang S, Hong H, Gan J, et al. The hierarchical folding dynamics of topologically associating domains are closely related to transcriptional abnormalities in cancers. *Comput Struct Biotechnol J*. 2021;19:1684–93.
35. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Wren J, editor. *Bioinformatics*. 2020;36:311–6.
36. Dask Development Team. Dask: library for dynamic task scheduling. 2016. Available from: <https://dask.org>.

37. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. Proc 12th USENIX Symp Oper Syst Des Implement OSDI '16. 2016. p. 265–83.
38. Pixijs | The HTML5 Creation Engine | Pixijs. Available from: <https://pixijs.com/>.
39. Django [Internet]. Django Proj. [cited 2023 Dec 6]. Available from: <https://www.djangoproject.com/>.
40. django/channels: Developer-friendly asynchrony for Django [Internet]. Available from: <https://github.com/django/channels>.
41. Chen S, Wang T, Jia R. Zero-round active learning. 2021 [cited 2023 Dec 12]; Available from: <https://arxiv.org/abs/2107.06703>.
42. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal.* 2021;71: 102062.
43. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev.* 2023;56:3005–54.
44. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods.* 2012;9:999–1003.
45. Niu L, Shen W, Shi Z, Tan Y, He N, Wan J, et al. Three-dimensional folding dynamics of the *Xenopus tropicalis* genome. *Nat Genet.* 2021;53:1075–87.
46. davidfig/pixi-viewport: A highly configurable viewport/2D camera designed to work with pixijs. Available from: <https://github.com/davidfig/pixi-viewport>.
47. Lun ATL, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics.* 2015;16:258.
48. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44:W160–5.
49. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation.* 2021;2:100141.
50. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. 2023.
51. Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinforma Oxf Engl.* 2021;37:422–3.
52. Shen W, Zhang P, Jiang Y, Hailin T, Zhike Z, Li L. HTAD. Zenodo; 2024. Available from: <https://doi.org/10.5281/zenodo.13822061>.
53. Shen W. Test dataset of HTAD. Zenodo; 2024 [cited 2024 Nov 19]. Available from: <https://zenodo.org/doi/10.5281/zenodo.14186235>.
54. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell. Datasets. Gene Expression Omnibus.* 2014. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525>.
55. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* 2020;48:D882–9.
56. Niu L, Shen W, Shi Z, Tan Y, He N, Wan J, et al. Three-dimensional folding dynamics of the *Xenopus tropicalis* genome. *BioProject.* 2021. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA606649>.
57. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet.* 2013;29:569–74.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.