Computational and Structural Biotechnology Journal 21 (2023) 4759–4768

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method Article

TOAST: A novel method for identifying topologically associated domains based on graph auto-encoders and clustering ☆

Haiyan Gong [a,b,c,*], Dawei Zhang [a,c], Xiaotong Zhang [b,c,**]

[a] Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, 100083, China
[b] School of Computer and Communication Engineering, Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing, 100083, China
[c] Shunde innovation School, University of Science and Technology Beijing, Foshan, 528399, Guangdong, China

ARTICLE INFO

Keywords:
Chromatin structure
Topological associated domains
Hi-C
Graph auto-encoders
Cluster

ABSTRACT

Topologically associated domains (TADs) play a pivotal role in disease detection. This study introduces a novel TADs recognition approach named TOAST, leveraging graph auto-encoders and clustering techniques. TOAST conceptualizes each genomic bin as a node of a graph and employs the Hi-C contact matrix as the graph's adjacency matrix. By employing graph auto-encoders, TOAST generates informative embeddings as features. Subsequently, the unsupervised clustering algorithm HDBSCAN is utilized to assign labels to each genomic bin, facilitating the identification of contiguous regions with the same label as TADs. Our experimental analysis of several simulated Hi-C data sets shows that TOAST can quickly and accurately identify TADs from different types of simulated Hi-C contact matrices, outperforming existing algorithms. We also determined the anchoring ratio of TAD boundaries by analyzing different TAD recognition algorithms, and obtained an average ratio of anchoring CTCF, SMC3, RAD21, POLR2A, H3K36me3, H3K9me3, H3K4me3, H3K4me1, Enhancer, and Promoters of 0.66, 0.47, 0.54, 0.27, 0.24, 0.12, 0.32, 0.41, 0.26, and 0.13, respectively. In conclusion, TOAST is a method that can quickly identify TAD boundary parameters that are easy to understand and have important biological significance. The TOAST web server can be accessed via http://223.223.185.189:4005/. The code of TOAST is available online at https://github.com/ghaiyan/TOAST.

1. Introduction

With the continuous improvement of three-dimensional genomics, researchers have begun to apply it to temporal studies, known as the 4D Nucleome Project [1], aiming to understand how nuclear organization influences nuclear function across both spatial and temporal dimensions. In higher eukaryotes, including humans, the genome is arranged in intricate and dynamic three-dimensional conformations, whereby approximately 2 meters of genomic DNA is intricately condensed through elaborate folding to seamlessly reside within the nucleus, which spans a diameter of roughly 10 μm [2]. The interactions between different genomic loci are critical as a blueprint for gene expression. The principles underlying the hierarchical folding of chromatin structure have offered significant contributions to the field of genomics and were established through the utilization of the so-called "C-method" [3,4]. This method depends on the proximity ligation of DNA fragments to understand the physical proximity of specific genomic regions within the cell nucleus. While it was proposed by Cullen et al. as early as the 1990s [3,4], this method did not gain widespread application until the development of chromatin conformation capture techniques [5]. The advancement of such techniques, combined with high-throughput sequencing, resulted in the emergence of the Hi-C technology [6]. Hi-C allows for the genome-wide assessment of physical proximity between various DNA segments, allowing for substantial insights into the general principles of chromatin folding.

Existing research suggests that interphase chromatin in humans is arranged into multi-layered hierarchical structures, including 3D chromatin structures [7], A/B compartments, subcompartments, topologically associated domains (TADs), and chromatin loops [8]. The A/B compartments are linked to open and transcriptionally active chromatin (A compartment) as well as closed and transcriptionally inactive chromatin (B compartment) [6]. TADs are structural domains within

☆ This work is supported by the Foshan Higher Education Foundation [BKBS202203].
* Corresponding author at: Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, 100083, China.
** Principal corresponding author at: Shunde innovation School, University of Science and Technology Beijing, Foshan, 528399, Guangdong, China.
E-mail addresses: ghaiyan@ustb.edu.cn (H. Gong), zxt@ies.ustb.edu.cn (X. Zhang).

https://doi.org/10.1016/j.csbj.2023.09.019
Received 31 May 2023; Received in revised form 16 September 2023; Accepted 16 September 2023
Available online 27 September 2023

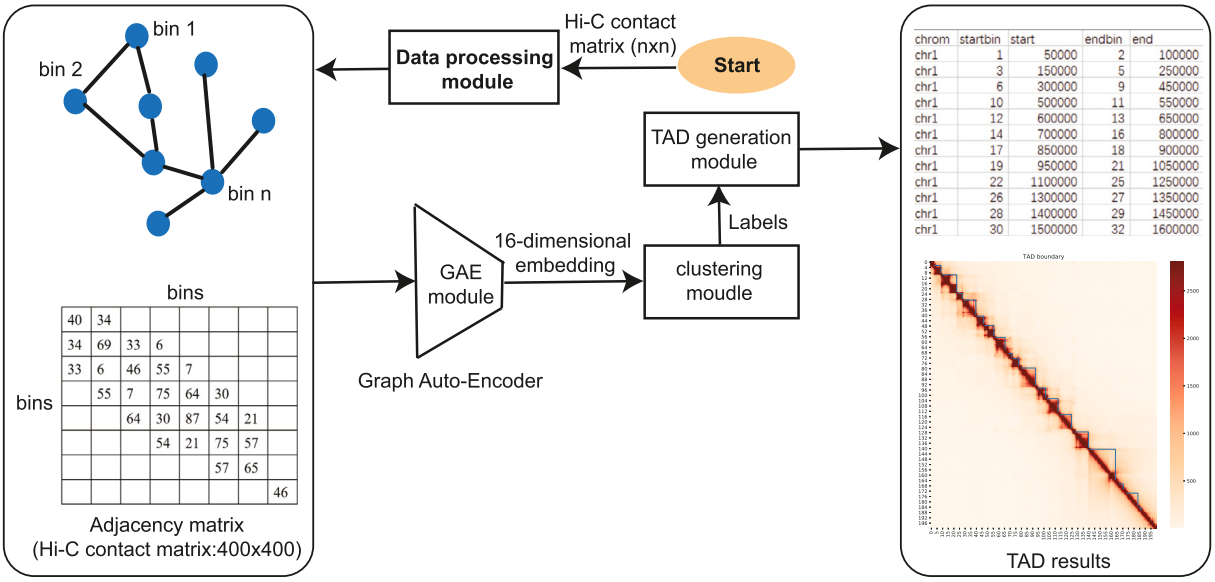

Fig. 1. The flowchart of TOAST.

chromatin regions in which interactions occur more frequently than between distinct chromatin regions. These domains typically range in size from 200 kilobases (Kb) to 5 megabases (Mb) [9] and exhibit high conservation across different species and cell types. TADs are required components in chromatin's three-dimensional structure and are believed to function in gene regulation, playing crucial roles in cell differentiation, development, and disease treatment [10]. Additionally, TADs are essential for transcriptional regulation, DNA replication, and VDJ recombination, a critical process for producing diverse antigen receptor genes in immune cells [11]. Disruptions in TAD boundaries have been implicated across genetic disorders and cancer, highlighting their potential significance in disease [12]. Hence, developing a method capable of swiftly and accurately identifying TAD boundaries is critical in advancing the studying of chromatin structure and its implications in diseases.

The majority of TAD identification methods directly detect TAD boundaries from the Hi-C contact frequency matrix. Currently, there are several categories of methods are available for TAD identification: (1) metric-based methods, such as the Directionality Index (DI) [13] method proposed by Dixon et al. in 2012, as well as techniques like Armatus [14] that synergize dynamic programming and multi-scale modeling, Arrowhead [15], TopDom [16], among others [17–19], offer accurate computations while necessitating the configuration of relevant metric parameters. (2) Statistical methods, exemplified by HiC-Seg [20], which harmonizes maximum likelihood estimation with one-dimensional segmentation, as well as TADtree [21], PSYCHIC [22], TADbit [23], among others [24], often result in hierarchical structures containing numerous inaccurate small-scale TADs. (3) Clustering-based methods, including ClusterTAD [25], using K-means clustering, IC-Finder [26] relying on hierarchical clustering, TADpole [27] employing constrained hierarchical clustering, and CASPIAN [28], confront issues where the quality of TAD boundaries is contingent on cluster count otherwise resulting in diminished accuracy. (4) Network-based methods, such as MrTADFinder [29], which capitalizes the similarity between TADs in chromatin contact map and densely connected modules in the network for modularization, and 3DNetMod [30]. However, the methods mentioned above are impaired by complex parameter settings, lengthy computation time, or low accuracy in TAD identification.

Graph Auto-Encoders (GAE) [31,32], using their encoder-decoder structure, have been employed to obtain node embeddings in graphs for various tasks, including miRNA-disease association prediction [33] and the identification of system-level anomalies in nuclear power plants [34]. Given the advantages of GAE in obtaining embeddings, to address the issues above, we introduce a TAD recognition algorithm known as TOAST, combining the power of GAE and clustering. Firstly, Leveraging GAE's encoder-decoder structure, we trained it on Hi-C contact matrices, divided into non-overlapping 400×400 matrices and deriving 16-dimensional embeddings. Subsequently, the resultant embeddings served as inputs for an unsupervised clustering algorithm, effectively assigning labels to each bin. TAD boundaries were later delineated based on the assigned labels, and this information is harnessed for visualization. We conducted rigorous validation using simulated Hi-C contact matrices at different noise levels and real Hi-C datasets from the GM12878 cell line spanning resolutions of 50 Kb, 25 Kb, and 5 Kb. The results demonstrate that TOAST outperforms other algorithms in terms of accurately identifying TAD boundaries in simulated Hi-C data sets at different resolutions. Furthermore, by calculating the TAD boundaries identified by different TAD recognition algorithms, we could quantify the proportions of different transcription factors and histone modifications at recognized TAD boundaries, elucidating their roles in shaping genomic organization.

2. Methods

2.1. Overview of the TOAST framework

This paper proposes a TAD recognition algorithm (TOAST) based on the amalgamation of GAE and clustering techniques. As outlined in Fig. 1, TOAST primarily consists of four parts: A data processing module, a GAE module, a clustering module, and a TAD generation module. The data processing module primarily deals with multiple 400×400 Hi-C contact matrices. The GAE module employs GAE neural network training to obtain 16-dimensional embedding. The clustering module utilizes the unsupervised clustering method to coordinate the generated embeddings and label each bin. The TAD generation module processes the generated labels to deliver the TAD boundary file and performs the heat map visualization featuring TAD boundaries. More comprehensive information can be found in sections pertaining to the Data processing module, GAE module, clustering module, and TAD generation module.

2.2. Data preprocessing module

Hi-C sequencing data is commonly generated in .hic format. To extract the matrices at different resolutions (5000 base pairs (5 Kb), 25 Kb, and 50 Kb), we utilized Juicer tools [35]. The resulting matrices could be either raw or KR-normalized [36] Hi-C contact matrices, represented

as NxN matrices. Each value in the matrix represented the interaction strength between bins, and N denoted the number of bins. Treating each bin as a vertex, the Hi-C contact matrix could be viewed as the graph's adjacency matrix. To facilitate the training of GAE networks, in this study, we divided the Hi-C contact matrix into m $n \times n$ submatrices by partitioning along the diagonal using a step of n. For Hi-C contact matrices with resolutions below 5 Kb (e.g., 50 Kb, 25 Kb, 10 Kb), we set $n = 400$. If the resolution exceeded 5 Kb (e.g., 1 Kb), $n = 2000000/res$, where res represented the resolution of the Hi-C contact matrix. Given that TAD identification predominantly involves Hi-C matrices at resolutions under 5 Kb, we maintained $n = 400$ for this study.

2.3. GAE module

Graph $\varsigma = (v, \xi)$ can be represented as v, where v is the node-set, ξ is the edge set. The 400×400 sub-matrix, denoted as adjacency matrix A of GAE network training, is a crucial component for training the GAE network. $N = 400$ represents the number of nodes, while $X \in R^{N \times N}$ signifies the node's feature matrix. As the Hi-C contact matrix functions solely as the input, it is initialized as a matrix containing all ones on the diagonal and zeros elsewhere. f represents the dimension of embedding, $Z \in R^{N \times f}$ is the embedding of the node.

Graph Auto-Encoders (GAE) module is divided into two parts: the encoder and the decoder. The encoder obtains the GCN function of node feature X and adjacency matrix A as input and node embedding Z as output to obtain node embedding. Users are then able to change the number of layers and the dimension f of embedding Z. In this work, we compare the TAD identification results under different network architecture parameters and determine that the optimal TAD identification result is achieved when the network has 2 layers, an embedding dimension of 16, and a hidden layer dimension of 64. (Details can be found in the section "Determination of Network Architecture Parameters") GCN can be expressed using Eq. (1).

$$Z = GCN(A) = W_2 \times \mathrm{Re}\,LU(W_1 \times A + b1) + b2 \tag{1}$$

Where W_2 and W_1 represent the weight matrices, $b2$ and $b1$ are the bias vectors, and ReLU is the Rectified Linear Unit.

The Decoder portion also uses the GCN function to reconstruct the original graph, and takes the embedded Z obtained from the encoder as input and the 400×400 reconstructed adjacency matrix $\hat{A}$ as output. The decoder can be represented by Eq. (2).

$$\hat{A} = W_4 \times \mathrm{Re}\,LU(W_3 \times Z + b3) + b4 \tag{2}$$

Where W_4 and W_3 are the weight matrices, while $b4$ and $b3$ are the bias vectors.

Because the adjacency matrix preserves the structure of the graph, to ensure that the reconstructed adjacency matrix $\hat{A}$ is as similar as possible to the original adjacency matrix A, as Eq. (3) shows, we employed the Mean Squared Error (MSE) as the loss function.

$$MSE = \frac{1}{1600}\sum_{i=1}^{1600}(y_i - \hat{y}_i)^2 \tag{3}$$

Where y_i represents the value of a specific element in the adjacency matrix A, and $\hat{y}_i$ represents the value of the corresponding component of the reconstructed adjacency matrix $\hat{A}$.

Throughout the training process, we utilized a learning rate of $lr = 0.001$, trained the model for 200 epochs, and employed the Adam optimizer to obtain the final embedding Z.

2.4. Clustering module

The clustering module employed the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm [37], a density-based hierarchical clustering algorithm. Using the embedding Z as input, HDBSCAN conducted unsupervised clustering, assigning labels to each node. Notably, HDBSCAN automatically discovered clusters of diverse densities within the data, representing them as a hierarchical tree structure. Compared to traditional clustering algorithms such as K-Means and DBSCAN, HDBSCAN offers enhanced flexibility and robustness, enabling it to handle noise, sparse data, and clusters of varying densities. It does this through the following steps:

(1) Calculate Kernel Density: Kernel density estimation is performed on the input data to obtain the local density for each sample point.
(2) Build Minimum Spanning Tree (MST): An MST is constructed among the core samples, utilizing either Prim's or Boruvka's algorithms. Core samples denote data points with densities over a designated threshold.
(3) Assign Cluster Memberships: The cluster membership for each sample point is determined based on the edge weights of the MST. A lower edge weight indicates higher similarity between sample points and a higher probability of belonging to the same cluster. Specifically, beginning from the maximum edge of the MST, extend downwards, and for each subtree, assign all sample points to the same cluster. If the density of a subtree is below a certain threshold, consider it as noise.
(4) Determine the Hierarchical Structure of Clusters: Clusters are arranged in descending order based on densities and depicted as a hierarchical tree structure. The root node represents the entire dataset, while the leaf nodes represent individual sample points.
(5) Extracting Final Clustering Results: Pruning and trimming of the hierarchical tree is conducted as mentioned in [37], and the final clustering results of genomic bins are obtained.

2.5. TAD generation module

Genomic bins with the same label were grouped to form contiguous regions. Studies have typically indicated TAD intervals ranging from 200 Kb to 5 Mb [9]. Therefore, the number of genomic bins and the resolution can be used to determine if a contiguous region should be considered as a TAD. If the length of a contiguous region exceeds 200 Kb, it can be identified as a TAD. For example, when analyzing Hi-C contact matrix data with a resolution of 50 Kb, a continuous region must contain at least four genomic bins with the same label to be considered a TAD. This ensures that we only identify TADs of sufficient length and avoids mistakenly identifying short contiguous regions as TADs. TADs are succinctly represented in the format of (start bin, start location, end bin, end location).

2.6. Datasets

2.6.1. Simulated Hi-C data

Two simulated datasets were utilized in our experiments. One dataset was generated by Wang et al. [38] in their investigations. The simulated contact matrix data with five noise levels (4, 8, 12, 16, and 20) was generated using the Directionality Index (DI) method [13] in the IMR-90 Human Embryonic Lung Fibroblast (IMR90) cell line. The other dataset [39] provided by Trussart et al. used polymer modeling to simulate six artificially generated genomes with different structures, referred to as 'toy genomes'. A total of 168 simulated interaction matrices were extracted from these toy genomes, with increasing noise and structural diversity levels. The simulated Hi-C matrices used in this study encompass noise levels of 50, 100, 150, and 200, denoting an increasing presence of noise in the data. Additionally, these matrices possess linear densities of 40, 75, and 150 base pairs per nanometer (bp/nm), signifying variations in the structural arrangement.

2.6.2. Real Hi-C data

The experimental data within this study were derived from a genuine dataset from the human B-lymphoblastoid cell line (GM12878). To obtain the raw data, the .hic file with the unique accession number

4DNFI1UEG1HD was downloaded from the 4D Nucleome Data Portal (https://data.4dnuc-leome.org/). Subsequently, the file was processed using Juicer [15], resulting in KR-normalized Hi-C contact matrices with resolutions of 5 Kb, 25 Kb, and 50 Kb, and raw Hi-C contact matrices.

2.6.3. ChIP-seq data

ChIP-seq data targeting CTCF, POLR2A, RAD21, SMC3, H3K4me3, H3K36me3, and H3K9me3 in the GM12878 cell line were downloaded from the ENCODE project platform [40] (www.encodeproject.org). The corresponding accession numbers for these datasets are as follows: ENCFF74 -9HDD, ENCFF002GST, ENCFF822QJA, ENCFF775OOS, ENCFF480KNX, ENCFF537KDM, ENCFF174RRQ, and ENCFF218YZR.

The above ChIP-seq data were used for the following: (1) The replicated peaks files, formatted in bed narrow peak file format, were utilized to identify the peaks of various transcription factors. The number and proportion of peaks anchored to the TAD boundaries within a 20 Kb genomic region for each transcription factor were calculated. (2) The signal p-value files, formatted in bigWig file format, facilitated the generation of average p-values from ChIP-seq data within a 40 Kb genomic interval proximate to TAD boundaries.

2.6.4. Enhancer and promoter data

Previous studies [41,42] have demonstrated that some TAD boundaries coincide with the anchor points of chromatin loops, such as Enhancer-Promoter interactions. Therefore, in our study, the EPI dataset of the GM12878 cell line from https://github.com/wgmao/EPIANN/tree/master/GM12878 was downloaded to qualify the ratio of anchoring Enhancer or Promoter within a 20 Kb genomic region near the TAD boundaries.

2.7. Evaluation metric

(1) A notable characteristic of TAD structures is that regions within a given TAD possess a higher contact frequency distribution, while regions outside of TADs have fewer contacts. Based on this property, the average contact frequency between each bin within TAD_i (denoted as $intra(i)$) and the average contact frequency between TAD_i and its adjacent TAD TAD_j (denoted as $inter(i,j)$) are calculated, where $|i-j|=1$. As shown in Fig. S1, a triangle represents a TAD. For example, TAD i has adjacent TADs j to the left and right. $inter(i,j)$ is the average number of contacts between TAD i and TAD j. $intra(i)$ is the average number of contacts within TAD i (i.e., the blue region). Therefore, the TAD quality score ($TAD_i\_quality$) can be represented by the following equation. Eq. (4) is used to compute for each TAD defined in the dataset. The overall quality score for a set of TADs defined by the contact matrix is the average of their quality scores.

$$TAD_i\_quality = intra(i) - inter(i,j) \tag{4}$$

Similar to a previous study [28], we employed the following metrics to evaluate the TAD quality.

(2) Visualization comparison: For simulated Hi-C contact matrices with real TAD data, the quality of identified TADs is evaluated by comparing their visualization with the actual TAD boundaries. For real Hi-C contact matrices without known TAD data, the effectiveness of different algorithms is assessed by comparing the strength of signals between visualized TADs, based on the definition of TADs (strong interactions within TADs and weak interactions between TADs).

(3) For real Hi-C contact matrices, the effectiveness of different algorithms is evaluated by assessing the rationality of the number of identified TADs, TAD quality, and average TAD length.

(4) Comparing the similarity between TADs identified by TOAST and true TADs obtained from simulated data (or TADs identified by other existing algorithms). In this study, the genomic bins were divided into three clusters based on the TAD results: TAD boundaries, bins within TADs, and bins between TADs. The consistency between these two clusters is evaluated using the Fowlkes-Mallows score (FMS) and the Rand Index (RI). As shown in Eq. (5), FMS is the geometric mean of precision and recall.

$$FMS = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{5}$$

Where TP represents the number of true positive sample pairs, FP is the number of false positive sample pairs, and FN represents the number of false negative sample pairs. The Fowlkes-Mallows score (FMS) ranges from 0 to 1, with higher values indicating higher similarity between the two clusters.

The Rand Index (RI) is defined as $RI = N_1/N$, which measures the similarity between two clusters. A higher RI value indicates a higher similarity between the two clusters. Here, N_1 represents the number of genome bin pairs with consistent TAD cluster labels, and N represents the total number of genome bin pairs.

(5) Within a 40 Kb genomic distance range surrounding TAD boundaries, the average p-value distribution of different transcription factors or histone modifications from ChIP-seq data. If the p-value distribution of transcription factors such as CTCF, associated with transcriptional activation, exhibits peaks, it indicates the overall accuracy of the identified TAD distribution.

(6) The quality of the identified TADs is evaluated using the ratio of anchored CTCF, SMC3, H3K4me3, Enhancer, Promoter, and other TFs or histone modifications.

3. Results and discussion

3.1. Experiment setting

In this paper, we conducted experiments using the following experimental software: Ubuntu 18.04.4 LTS (Bionic Beaver), Pytorch 1.7.1, Python 3.6.12, Numpy 1.19.5, Matplotlib 3.3.3, Cooltools 0.4.1, Fanc 0.9.24, Jupyter 1.0.0, Pandas 1.1.5, pyBigWig 0.3.18, seaborn 0.11.1. Code for TOAST is available online at https://github.com/ghaiyan/TOAST.

3.2. Baselines

This study compares our proposed TOAST method with several TAD partitioning algorithms, including CASPIAN [28], Directionality Index (DI) [13], Insulation score [43], TopDom [16], IC-Finder [26], HiCSeg [20], and ClusterTAD [25]. These algorithms share the following characteristics: 1) They all identify non-hierarchical TADs. 2) Among these methods, Insulation Score, DI, TopDom, IC-Finder, and HiCSeg are five standard TAD identification tools, while ClusterTAD and CASPIAN are TAD identification tools based on clustering algorithms. The scripts for calling TADs using other methods can be obtained from https://github.com/ghaiyan/TOAST.

3.3. Determination of network architecture parameters

To determine the dimensions for generating embeddings and the number of network layers, we designed six different networks: one-layer networks and two-layer networks with embedding dimensions of 16, 64, and 128 (i.e. $onelayer\_dim = 16$, $onelayer\_dim = 64$, $onelayer\_dim = 128$, $hidden\_dim = 64$ and $output\_dim = 16$, $hidden\_dim = 256$ and $output\_dim = 64$, $hidden\_dim = 256$ and $output\_dim = 128$). These networks were applied to TAD identification on simulated Hi-C contact matrices provided by Wang et al. [38].

First, we compared the reconstruction loss over epochs for simulated Hi-C data with different noise levels and GAE network architectures (Fig. S2). As shown in Fig. S2, when using the same embedding dimension, the one-layer GAE network consistently resulted in higher reconstruction loss than the network with an additional hidden layer,

Table 1
Performance of TOAST on simulated Hi-C contact matrix at different noise level.

noise level	FMS	TAD number	True TAD number	avg_length (Kb)	avg_true_length (Kb)
4	0.990	175	171	970.1	997.4
8	0.991	172	171	1038.1	1048.6
12	0.975	178	171	963.4	1017
16	0.981	170	171	999.3	1002
20	0.979	165	171	106.9	1037.7

indicating that the network with an added hidden layer generates more effective embeddings.

Next, we compared the TAD identification results on simulated Hi-C data with varying noise levels, using different GAE network architectures and embedding dimensions. Tables S1–S5 indicate that the best TAD identification results were achieved when using $hidden\_dim = 64$ and $output\_dim = 16$, or when employing a single-layer network with an output embedding dimension of 128 ($onelayer\_dim = 128$). However, upon comparing the results, the network structure with $hidden\_dim = 64$ and $output\_dim = 16$ showed higher robustness. Consequently, in this study, we select $hidden\_dim = 64$ and $output\_dim = 16$ as the optimal network configuration.

3.4. Performance of TOAST on simulated datasets

(1) Performance of TOAST on Regularized Simulated Hi-C Datasets. To evaluate the performance of TOAST on simulated datasets, we selected simulated datasets provided by Wang et al. [38] and Trussart et al. [39]. The simulated Hi-C contact matrices from Wang et al. [38] were generated at a resolution of 40 Kb, with different noise levels ranging from 4 to 20, exhibiting well-defined TAD boundaries. The ground truth TAD boundaries were also provided. Table 1 presents the Fowlkes-Mallows Score (FMS), a measure for quantifying the similarity between TOAST and the ground truth TADs calculated across different noise levels (4, 8, 12, 16, and 20). We observed a gradual decrease in FMS as the noise level increased, but the overall FMS remained around 0.98. Additionally, we computed the number of TADs and the average TAD length (avg_length) across different noise levels. As shown in Table 1, compared to the ground truth TADs, the difference in the number of TADs was at most 8 and at least 1, while the difference in average TAD length (avg_length) compared to the average true TAD length (avg_true_length) was less than 1 Kb in the best case and up to 5 Kb in the worst case. Furthermore, a visual comparison of TOAST's TAD boundaries with the ground truth TADs, illustrated in Fig. S3, underscored a remarkable alignment.

(2) Performance of TOAST on Irregular Simulated Hi-C Datasets. To assess the performance of TOAST on non-regularized simulated datasets, we utilized TOAST on simulated Hi-C matrices provided by Trussart et al. [39]. These datasets featured varying noise levels (50, 100, 150, and 200) and linear densities (40, 75, and 150 bp/nm). Each combination of noise level and linear density corresponded to a distinct resolution, with higher noise levels and linear densities resulting in lower resolution and denser matrices. Table S6 (in Supplementary Files) and Fig. 2 present the results of the TAD boundary identified by TOAST on datasets with various noise levels and linear densities.

Table S6 (in Supplementary Files) illustrates that datasets with lower noise levels and linear densities yielded increased TAD boundaries. The visual results show that datasets with higher noise levels and linear densities exhibit more pronounced TAD divisions. This can be observed in Fig. 2, where the blue TAD boundary lines align closely with the white divisions in the heatmaps. While datasets with lower noise levels and linear densities might not manifest explicitly defined TAD boundaries, carefully examining the leftmost columns in the heatmaps reveals an overlap between the identified TAD boundaries and the white division lines. Consequently, we can conclude that TOAST can achieve favorable TAD partitioning results on simulated Hi-C contact matrices with different resolutions.

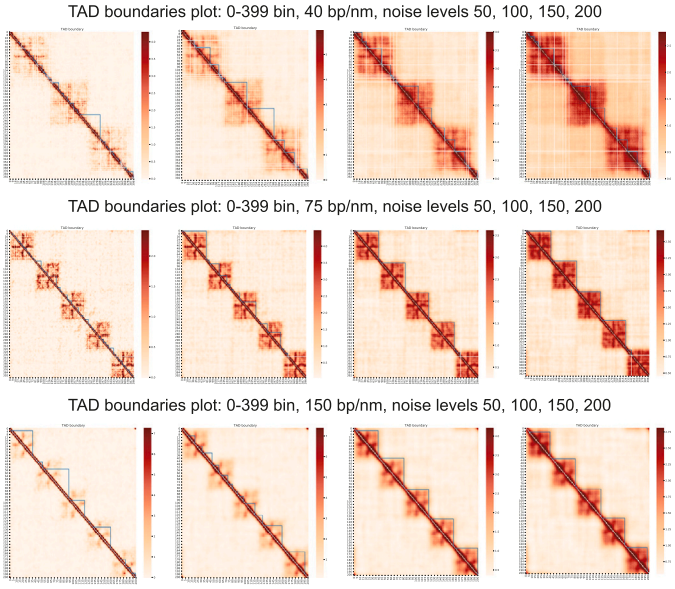

Fig. 2. Comparison of TAD visualization for different Hi-C contact matrices at different linear densities and noise level.

(3) TOAST Outperforms Other Algorithms on Simulated Hi-C Datasets. To benchmark the performance of TOAST against other TAD identification algorithms on simulated Hi-C datasets, we selected CASPIAN [28], Insulation score [43], TopDom [16], IC-Finder [26], HiCSeg [20], and ClusterTAD [25]. We evaluated simulated Hi-C contact matrices featuring 4, 8, 12, 16, and 20 noise levels. Table 2 presents the outcomes from the datasets with a noise level of 16, including the computed similarity values (FMS), TAD quality scores, TAD number, and average TAD length compared to the true TAD boundaries. The results indicate that TOAST and TopDom achieved the highest FMS value of 0.97. TOAST, TopDom, and HiCSeg yielded TAD quality scores surpassing 10.

However, TOAST exhibited the lowest deviation regarding the TAD number and average TAD length relative to the true TAD boundaries. Visual comparison of the TAD boundaries identified using different algorithms, as depicted in Fig. S7 (in Supplementary Files), distinctly illustrates TOAST's impeccable alignment with the true TAD boundaries. TopDom, CASPIAN, and Insulation Score exhibited relatively good performance, while HiCSeg and IC-Finder identified significantly higher TAD boundaries, leading to substantial discrepancies from the true TADs. ClusterTAD showed the least favorable performance among the algorithms.

Furthermore, to validate TOAST's excellent performance under different noise levels, we compared the results of TAD identification for noise levels 4, 8, 12, and 20 (Figs. S4–S6, Fig. S6, Tables S7–S10 in Supplementary Files). The results consistently demonstrated that TOAST outperformed the other algorithms quantitatively and visually across different noise levels.

3.5. Performance of TOAST on real Hi-C dataset

In this section, we experiment sequentially on the 50 Kb, 25 Kb, and 5 Kb Hi-C contact matrices of the GM12878 cell line for all chromosomes (from 1 to 22). Experiments show that the run time for computing 50 Kb Hi-C contact matrices sequentially is 2 hours, for the 25 Kb matrices is 4 hours, and for the 5 Kb matrices it is 11 hours. In the other words, we could obtain all TAD results for the GM12878 cell line within 0.5 hours through parallel computing.

(1) Comparison of TAD Partitioning Results by TOAST across Different Chromosomes. To comprehensively evaluate the perfor-

Table 2
The results of detecting TADs from simulated Hi-C data when noise level = 16.

Method	FMS	TAD_quality	TAD number	True TAD number	avg_length (Kb)	avg_true_length (Kb)
TOAST	**0.98**	10.2	**170**	171	966.9	1002
CASPAIN	0.95	7.1	154	171	1099.5	1002
TopDom	0.97	10.4	219	171	773.7	1002
HiCSeg	0.69	10.3	239	171	418.7	1002
IC_Finder	0.86	9.6	281	171	594	1002
IS	0.83	9.2	204	171	826.7	1002
ClusterTAD	0.89	3.5	64	171	2645	1002

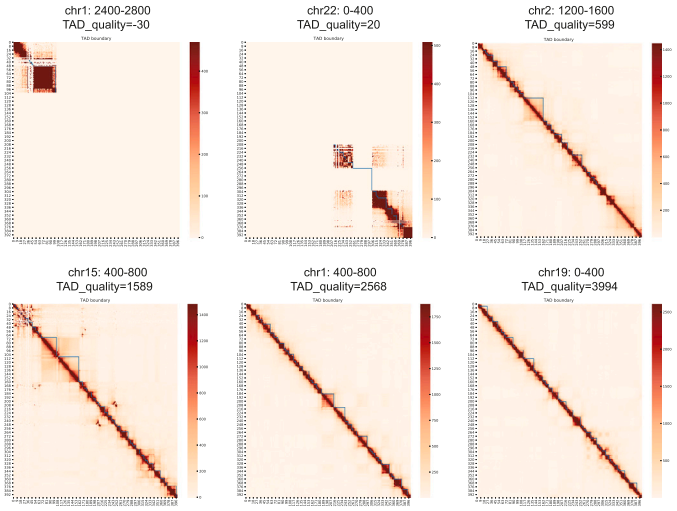

Fig. 3. Comparison for TAD visualization at different TAD_quality level.

mance of TOAST on real Hi-C datasets, we applied TOAST to 50 Kb Hi-C contact matrices of the GM12878 cell line for chromosomes 1 to 22. As the TOAST method partitions the Hi-C contact matrix along the diagonal before the TAD identification process, we calculated TAD quality for each 400×400 diagonal matrix in the GM12878 cell line for chromosomes 1 to 22. As shown in Table S11 (in Supplementary Files), the distribution of TAD quality values exhibited unevenness, with the majority falling within the 1000–4000 range. However, certain outliers were present. Consequently, we inspected the Hi-C contact heatmaps and their corresponding TAD partitioning results across distinct fields of TAD quality values, encompassing 0–100, 100–1000, 1000–2000, 2000–3000, and 3000–4000.

Fig. 3 illustrates that we visualized the diagonal matrices corresponding to TAD quality values of -30, 20, 599, 1589, 2568, and 3994. Notably, for TAD quality values below 100, blank regions were conspicuous in the diagonal matrix heatmap. Upon investigating the gene structure within the genomic area of chr1:120000000–125000000, it was found near the telomere between 1p31.1 and 1q12, q41, characterized by minimal chromatin interactions. Therefore, a low TAD quality value was obtained for this region. Analyzing the TAD partitioning results for TAD quality values of 1589, 2568, and 3994 shows that when TAD quality surpassed 1000, the partitioning results aligned well with visual inspection. Furthermore, higher TAD quality values resulted in more TAD structures being identified, including those at smaller scales. This indicates that the TAD quality values obtained through the TOAST method not only distinguish the effectiveness of TAD partitioning but also provide valuable insights into the corresponding genomic location, such as proximity to the telomere.

(2) Robustness Analysis of TOAST Method on Real Hi-C Datasets. To assess the performance of the TOAST method across varied Hi-C contact matrices obtained from different normalization algorithms and resolutions, we compared TAD similarities (RI and FMI) between the raw and KR-normalized 50 Kb-resolution Hi-C contact matrices, as well as the TAD results derived from Hi-C contact matrices at different resolutions (50 Kb, 25 Kb, and 5 Kb). Fig. 4(a, b) compares the number of TADs and TAD_quality obtained from the raw and KR-normalized Hi-C contact matrices. The results show that both methods exhibit a decreasing trend in TAD numbers and fluctuation in TAD_quality with increasing chromosome numbers. Fig. 4(c) calculates the TAD similarities (FMI and RI) obtained from both methods, demonstrating that they exhibit the same changing pattern with increasing chromosome numbers. Furthermore, Fig. 4(f) visually presents the TAD results within positions 0–400 bins of chromosome 1, demonstrating a remarkable similarity in TAD boundaries between the two methods.

Fig. 4(d, e) compares the number of TADs and TAD quality derived from Hi-C contact matrices at different resolutions (50 Kb, 25 Kb, and 5 Kb), and the results similarly demonstrate that the trends in TAD numbers and TAD_quality align across the different resolutions with increasing chromosome numbers. These results indicate that TOAST can effectively identify TADs from Hi-C contact matrices obtained using other methods.

(3) Performance of TOAST Compared to Other Algorithms on Real Hi-C Datasets. To comprehensively compare the performance of the TOAST algorithm with other prominent TAD identification algorithms on real Hi-C datasets, we selected several TAD identifying algorithms, including CASPIAN [28], Directionality Index (DI) [13], IS [43], TopDom [16], IC-Finder [26], HiCSeg [20], and ClusterTAD [25]. The comparative analysis was carried out using the real GM12878 cell line Hi-C contact matrix encompassing chromosomes 1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, and 22 at a resolution of 50 Kb.

Initially, we focused on chromosome 1 of the GM12878 cell line and visualized the TADs with in the range of bins 3600 to 3800 at 50 Kb resolution. Fig. 5 shows the visual results. Through visual inspection, it is evident that both TOAST and TopDom effectively captured TAD boundaries with a higher degree of completeness and accuracy, aligning closely with the actual boundaries. Although HiCSeg and IC-Finder identified more TADs, visual examination revealed that many of their detected TAD boundaries were false. The other algorithms were only able to identify partial TAD boundaries.

Furthermore, we analyzed the distribution of TAD number and TAD quality across multiple chromosomes within the GM12878 cell line. Notably, TOAST and TopDom demonstrated relatively minor disparities in both TAD number and TAD quality distributions. In contrast, HiCSeg exhibited a significantly higher number of TADs than other methods. However, visualization results in Fig. 5 showed that HiCSeg identified many false TAD boundaries. On the other hand, Insulation score (IS) had higher TAD_quality values than other algorithms, yet its visualization results were comparatively subpar. Table S12 (in Supplementary Files) provides a concise overview of the TAD quantities and TAD_quality values identified by different methods. HiCSeg exhibited markedly higher TAD number and TAD quality than most other algorithms. However, upon considering the visualization results in Fig. 5 for HiCSeg and to ensure fairness, we calculated the average TAD number and TAD quality values after excluding HiCSeg, which were 3826 and 1468.68, respectively. In comparison, TOAST achieved an average TAD number of 4736 and a TAD_quality value of 1654.03, surpassing the average

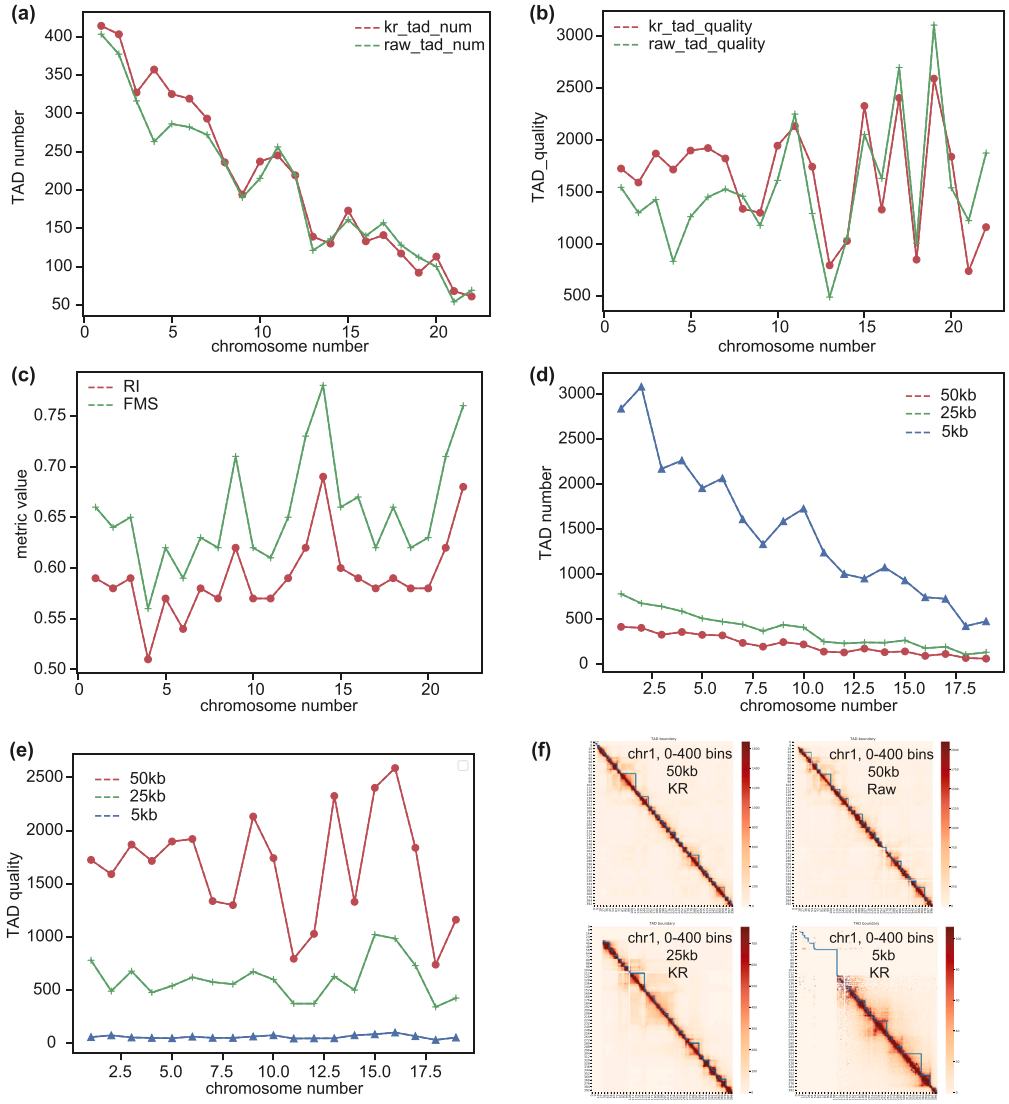

Fig. 4. Comparison of 50 Kb Hi-C contact matrices extracted from various normalization methods and resolutions. (a, b) The comparison of TAD number and TAD_quality changes with the chromosome number. (c) The similarity between TADs called from raw and KR-normalized Hi-C contact matrix. (d) The comparison of TAD number and TAD_quality changes with the chromosome number, where TADs are called from 5 Kb, 25 Kb, and 50 Kb resolution Hi-C contact matrix, separately. (f) The TAD visualization of TADs called from different resolution or normalized Hi-C contact matrix.

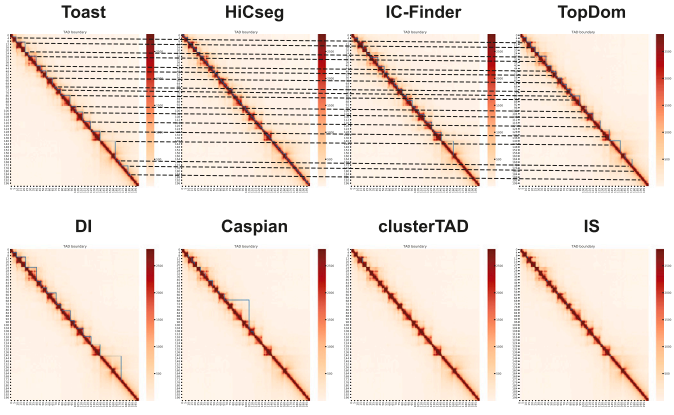

Fig. 5. Comparison for TAD visualization from bin 3600 to bin 3800 on GM12878 cell line, chromosome 1, 50 Kb Hi-C contact matrix.

values. Therefore, we conclude that the TOAST method performs well in both quantitative and visual indicators.

(4) TOAST can Reliably Identify a Significant Proportion of TADs Obtained by Other Algorithms. Lastly, we compared the similarity (RI and FMS) between the TADs identified by TOAST and those obtained by other algorithms. As evident from Fig. 6, both TOAST and Insulation score achieved the highest RI value (approximately 0.7) and the highest FMS value (approximately 0.8). By comparing the p-values of RI (or FMS), it was found that the similarity distribution between TOAST and CASPIAN was similar to the similarity distribution between TOAST and TopDom. Specifically, the RI values hovered around 0.6, and the FMS values were around 0.7, yielding a p-value of 1. These p-values were calculated using a two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. Consequently, these findings underscore TOAST's consistent capability to reliably identify a substantial proportion of TADs identified by other state-of-the-art algorithms.

(5) TOAST can Detect Significant TADs on the Real Hi-C Dataset. A previous study [37] has shown that TAD boundaries are enriched with CTCF, H3K4me3, and transcription start sites. To validate the biological relevance of TAD boundaries, we employed TOAST to identify TADs in the Hi-C contact matrices of GM12878 at a 50 Kb resolution for all chromosomes. We then visualized the signals of CTCF, H3K4me3, POLR2A, and H3K9me3 (a histone modification associated with transcriptional

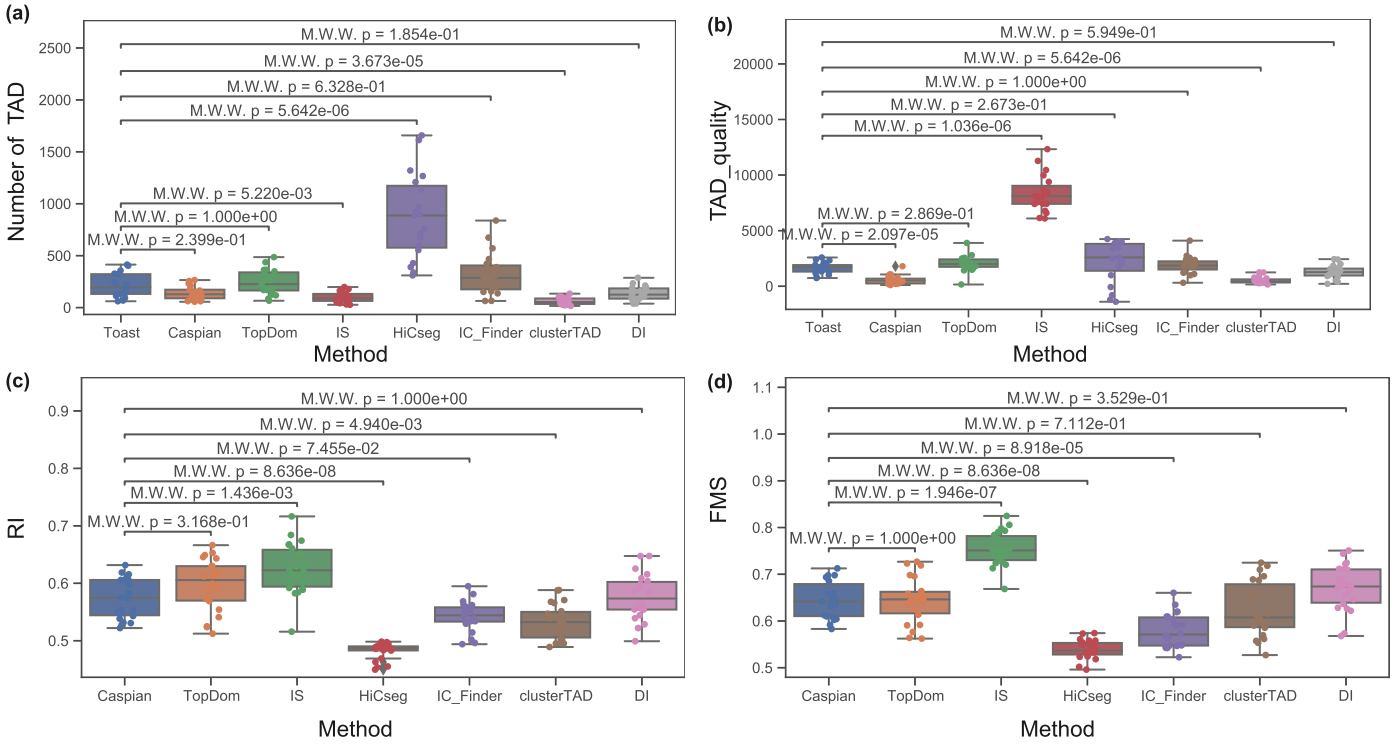

Fig. 6. The comparison for TAD number, TAD_quality and similarity with other TAD identification methods.

repression) within a 40 Kb region around the TAD boundaries. As shown in Fig. 7, the signed distance could be positive or negative. For example, -40 Kb indicates a distance of 40000 genomic loci to the left of the TAD boundary position, while 40 Kb represents a distance of 40000 genomic loci to the right of the TAD boundary position.

Fig. 7 illustrates the signal profiles of CTCF, H3K4me3, POLR2A, and H3K9me3 at the TAD boundaries, visualized in terms of their distance from the boundary. CTCF, H3K4me3, and POLR2A exhibit peaks at the TAD boundaries (when the distance from the boundary is 0), while H3K9me3 intriguingly shows a valley-like pattern at the TAD boundaries. To further observe the intricate relationship between TAD boundaries and the various transcription factors and histone modifications, we visualized the ChIP-seq tracks and heatmap of CTCF, POLR2A, H3K27ac, H3K4me3, H3K9me3 in the GM12878 cell line, chromosome 19, at the 0–400 bin location (Fig. S9 in Supplementary Files). Fig. S9 demonstrates that the majority of TAD boundaries are associated with strong signals for CTCF and POLR2A. Interestingly, a small fraction of TAD boundaries display elevated signals for H3K9me3. These observations robustly indicate that the TAD boundaries identified by TOAST elegantly align with prior knowledge, showcasing a noticeable enrichment of CTCF, POLR2A, H3K4me3, and other related factors at the TAD boundaries.

To quantify the proportions of various transcription factors, histone modification peaks, Enhancer, and Promoter at TAD boundaries, we calculated the ratios of CTCF, SMC3, RAD21, POLR2A, H3K36me3, H3K9me3, H3K4me3, H3-K4me1, Enhancer, and Promoter peaks within a 20 Kb range around the TAD boundaries (Table 3). We found that the average ratios of these factors anchored by TAD boundaries identified by different methods were 0.66, 0.47, 0.54, 0.27, 0.24, 0.12, 0.32, 0.41, 0.26, and 0.13, respectively. Interestingly, the ratios of these factors anchored by TAD boundaries identified by TOAST were also close to the values mentioned above. Therefore, we conclude that TOAST can effectively identify TAD boundaries with biological significance, as it exhibits similar proportions of these factors as observed in the analysis.

(6) TOAST can perform well on 1 Kb Hi-C contact matrices. To verify whether TOAST can operate effectively at high resolution (1 Kb), we downloaded the hic format data with accession number 4DNFI2WSZPG9 from https://data.4dnucleome.org/. We then extracted the contact matrix for Chromosome 1 at 1 Kb resolution and performed TAD identification. Since the matrix partition size of 400 was determined based on the 5 Kb resolution standard, for the 1 Kb resolution matrix, we set the matrix partition size to 2000, which is calculated as 2000000/1000, to ensure the sufficient retention of TAD features along the diagonal. With an input matrix size of 2000×2000, to fully preserve embedding features, we set $hidden\_dim = 512$, $output\_dim = 256$, and trained the GAE network for 500 epochs. We selected the embedding with the minimum reconstruction loss as input for clustering. The results showed that the calculation for all TADs on Chromosome 1 at 1 Kb resolution was completed in 81 minutes (including the time for reading matrix data).

As TADs were computed on submatrices (a total of 125 submatrices), we compared the distribution of TAD qualities for each submatrix (Fig. S10a). We then selected the submatrix with the highest TAD quality (Submatrix 4: 51.21) and the one with the lowest TAD quality (Submatrix 1: -2.02) for heatmap visualization (Fig. S10b and Fig. S10c). These heatmaps reaffirmed that regions with negative TAD quality values lack chromatin interactions, while regions with higher TAD quality exhibit distinct TAD structures. Based on these results, we conclude that TOAST can also compute high-quality TAD results for high-resolution Hi-C contact matrices, such as 1 Kb resolution.

4. Discussion and conclusion

scHiCEmbed [44] is a method used to analyze single-cell Hi-C data, sharing a similar approach to our practice for detecting TADs from single-cell Hi-C data. Both TOAST and scHiCEmbed [44] employ a two-step approach to identify TADs from chromatin bin. Firstly, GAE is used to obtain embedded representations for each chromatin bin. Subsequently, clustering methods are used to cluster chromatin bins, resulting in labeled regions. Regions with contiguous labels are considered as TADs. However, these two methods have distinct differences. Firstly, scHiCEmbed primarily targets single-cell Hi-C contact matrices, which require imputation before using GAE to obtain embedded representations. At the same time, TOAST is designed for the bulk Hi-C (Hi-C

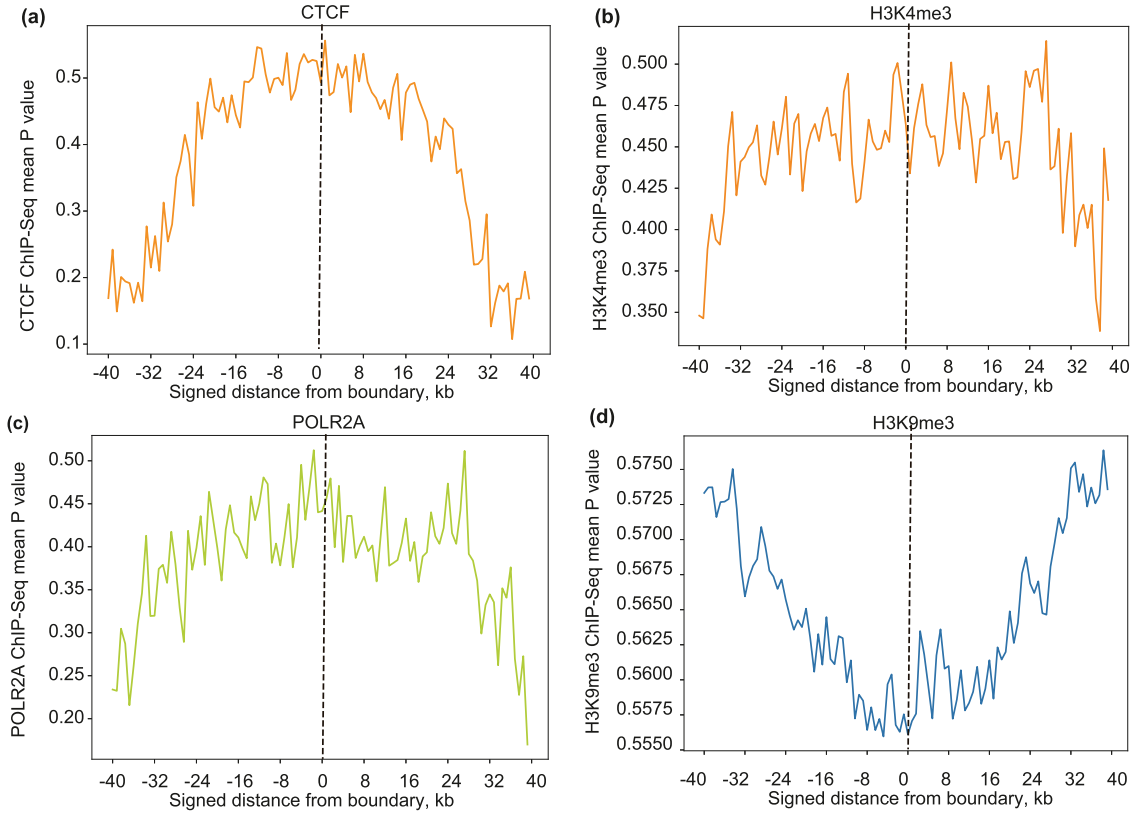

Fig. 7. Density distribution of different TFs or Histone modification around TAD's boundaries.

Table 3
The anchor ratio of TF or histone modification ChIP-seq peaks around TAD boundaries.

	Toast	Caspian	ClusterTAD	HiCSeg	IC-Finder	IS	TopDom	DI	Average
CTCF	0.65	0.61	0.53	0.58	0.67	0.73	0.73	0.75	0.66
SMC3	0.47	0.46	0.35	0.39	0.47	0.52	0.57	0.56	0.47
RAD21	0.53	0.50	0.41	0.46	0.55	0.59	0.63	0.62	0.54
POLR2A	0.25	0.28	0.20	0.20	0.27	0.32	0.30	0.34	0.27
H3K36me3	0.23	0.25	0.19	0.18	0.26	0.29	0.25	0.29	0.24
H3K9me3	0.13	0.13	0.10	0.08	0.15	0.12	0.10	0.10	0.12
H3K4me3	0.31	0.34	0.24	0.23	0.34	0.36	0.34	0.39	0.32
H3K4me1	0.41	0.44	0.35	0.30	0.45	0.43	0.41	0.45	0.41
Enhancer	0.26	0.28	0.26	0.20	0.27	0.28	0.26	0.29	0.26
Promoter	0.12	0.14	0.13	0.09	0.13	0.14	0.13	0.13	0.13

from a population of cells) contact matrix. Due to the high dimensionality of the bulk Hi-C contact matrix, directly using the original matrix as input would result in significant feature loss. Therefore, in this study, we divided the bulk Hi-C contact matrix into smaller submatrices of size 400×400 along the diagonal region. The selection of 400×400 size was based on considerations of covering genomic regions above 2 Mb, as it has been reported in previous studies that normal TAD sizes fall within 2 Mb. Of course, if a higher resolution than 5 Kb is targeted, we recommend setting the submatrix size to be $N = 2000000/\text{resolution size}$ to preserve more features.

Furthermore, the description of the TAD identification method in scHiCEmbed is unclear; it merely mentions using GAE to obtain an embedded representation (p-dimensional), but does not specify how the value of p is determined. In contrast, TOAST performs clustering by comparing the optimal embeddings (determined by the lowest MSE loss) for different p values, namely 16, 64, and 128. The selection of the optimal p value is based on the comparison of TAD quality.

Moreover, TOAST utilizes the HDBSCAN algorithm, which employs the concept of density-reachable graphs to cluster the generated optimal embeddings. This method considers the spatial density characteristics of chromatin loci and can effectively identify clusters with high local densities in high-dimensional data. On the other hand, scHiCEmbed uses a constrained hierarchical clustering approach to form a clustering tree but does not consider the spatial density of chromatin loci.

Lastly, the TOAST method is a comprehensive approach to TAD identification, encompassing GAE training, TAD generation, TAD quality assessment, TAD visualization, and other functionalities. In contrast, scHiCEmbed focuses solely on generating TADs without conducting a quality assessment or visualization of the TADs.

In conclusion, this paper presents a novel TAD recognition method called TOAST, which is based on graph auto-encoders and clustering. TOAST can quickly, and robustly identify different types of Hi-C contact matrices and outperform existing algorithms by comparing them with other TAD callers using simulated Hi-C contact and true TADs. Additionally, by analyzing different TAD recognition algorithms, we determined the average anchoring ratio of TAD boundaries with various transcription factors and histone modifications, such as CTCF, SMC3, RAD21, POLR2A, H3K36me3, H3K9me3, H3K4me3, H3K4me1, Enhancer, and Promoters, to be 0.66, 0.47, 0.54, 0.27, 0.24, 0.12, 0.32, 0.41, 0.26, and 0.13, respectively. As Table S13 shows, compared with other methods,

TOAST offers simplicity in parameters, provides more complete outputs (TAD files and all visualization files), supports ultra-high resolutions (such as 1 Kb resolution).

CRediT authorship contribution statement

All authors have contributed to drafting the work and revising it critically for intellectual content.

Declaration of competing interest

The authors declare no conflicts of interest.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2023.09.019.

References

[1] Dekker J, Belmont AS, Guttman M, et al. The 4D nucleome project. Nature 2017;549:219–26.
[2] Razin SV, Ulianov SV. Gene functioning and storage within a folded genome. Cell. Mol. Biol. Lett. 2017;22:18.
[3] Gothard LQ, Hibbard JC, Seyfred MA. Estrogen-mediated induction of rat prolactin gene transcription requires the formation of a chromatin loop between the distal enhancer and proximal promoter regions. Mol Endocrinol 1996;10:185–95.
[4] Cullen KE, Kladde MP, Seyfred MA. Interaction between transcription regulatory regions of prolactin chromatin. Science 1993;261:203–6.
[5] Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. Science 2002;295:1306–11.
[6] 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;326:289–93.
[7] Gong H, Yang Y, Zhang X, et al. NeRV-3D-DC: a nonlinear dimensionality reduction visualization method for 3D chromosome structure reconstruction with high resolution Hi-C data. In: 2022 IEEE international conference on bioinformatics and biomedicine. IEEE; 2022. p. 422–9.
[8] Gong H, Li M, Ji M, et al. MINE is a method for detecting spatial density of regulatory chromatin interactions based on a MultI-modal NEtwork. Cell Rep Methods 2023:100386.
[9] Rocha PP, Raviram R, Bonneau R, et al. Breaking TADs: insights into hierarchical genome organization. Epigenomics 2015;7(4):523–6.
[10] Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Mol Cell 2016;62(5):668–80.
[11] Hu J, Zhang Y, Zhao L, et al. Chromosomal loop domains direct the recombination of antigen receptor genes. Cell 2015;163(4):947–59.
[12] Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 2015;161(5):1012–25.
[13] Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485:376–80.
[14] Filippova D, Patro R, Duggal G, et al. Identification of alternative topological domains in chromatin. Algorithms Mol Biol 2014;9:14.
[15] Durand NC, Robinson JT, Shamim MS, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;3:99–101.
[16] Shin H, Shi Y, Dai C, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res 2016;44:e70.
[17] Zhan Y, Mariani L, Barozzi I, et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res 2017;27:479–90.
[18] Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. Nat Commun 2017;8:535.
[19] Wang XT, Cui W, Peng C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. Nucleic Acids Res 2017:45:e163.
[20] Levy-Leduc C, Delattre M, Mary-Huard T, et al. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics 2014;30:i386–92.
[21] Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics 2016;32(11):1601–9.
[22] Ron G, Globerson Y, Moran D, et al. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun 2017;8:2237.
[23] Serra F, Baù D, Goodstadt M, et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol 2017;13(7):e1005665.
[24] Xing H, Wu Y, Zhang MQ, et al. Deciphering hierarchical organization of topologically associated domains through change-point testing. BMC Bioinform 2021;22:183.
[25] Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC Bioinform 2017;18:480.
[26] Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res 2017;45:e81.
[27] Soler-Vila P, Cusco P, Farabella I, et al. Hierarchical chromatin organization detected by TADpole. Nucleic Acids Res 2020;48:e39.
[28] Gong H, Yang Y, Zhang X, et al. CASPIAN: a method to identify chromatin topological associated domains based on spatial density cluster. Comput Struct Biotechnol J 2022;20:4816–24.
[29] Yan KK, Lou S, Gerstein M. MrTADFinder: a network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS Comput Biol 2017;13:e1005647.
[30] Norton HK, Emerson DJ, Huang H, et al. Detecting hierarchical genome folding with network modularity. Nat Methods 2018;15:119–22.
[31] Kipf TN, Welling M. Variational graph auto-encoders. arXiv preprint. arXiv:1611.07308, 2016.
[32] Hasanzadeh A, Hajiramezanali E, Narayanan K, et al. Semi-implicit graph variational auto-encoders. Adv Neural Inf Process Syst 2019;32.
[33] Ding Y, Tian L-P, Lei X, et al. Variational graph auto-encoders for miRNA-disease association prediction. Methods 2021;192:25–34.
[34] Zhang L, Cheng W, Liu X, et al. System-level anomaly detection for nuclear power plants using variational graph auto-encoders. In: 2021 IEEE international conference on sensing, diagnostics, prognostics, and control. IEEE; 2021. p. 180–5.
[35] Robinson JT, Turner D, Durand NC, et al. Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst 2018;6:256–8. e251.
[36] Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal 2013;33:1029–47.
[37] Campello RJ, Moulavi D, Zimek A, et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans Knowl Discov Data 2015;10:1–51.
[38] Wang Y, Li Y, Gao J, et al. A novel method to identify topological domains using Hi-C data. Quant Biol 2015;3:81–9.
[39] Trussart M, Serra F, Baù D, et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res 2015;43:3465–77.
[40] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57.
[41] Tang Z, Luo OJ, Li X, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell 2015;163:1611–27.
[42] Anania C, Acemel RD, Jedamzick J, et al. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. Nat Genet 2022;54:1026–36.
[43] Crane E, Bian Q, McCord RP, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 2015;523:240.
[44] Liu T, Wang Z. scHiCEmbed: bin-specific embeddings of single-cell Hi-C data using graph auto-encoders. Genes 2022;13(6):1048.