MinCutTAD: Interpretable graph neural network - driven TAD prediction from Hi-C chromatin interactions and chromatin states

- Heidelberg Team -

Team members:
Lucas Arnoldt, Paul Kittner, Stefanie Mantz and Charlotte Westhoven

Supervised by: Carl Herrmann

Partner team: TAD team 2 from the University of Sorbonne

Table of contents

Introduction

In recent years, the continuous effort to unravel the details of all DNA functionalities has led to many technological advancements, and these advancements have shifted the spotlight onto the three-dimensional structure of DNA. For example, DNA folding has been determined to be crucial in fully understanding the process of gene regulation (Kempfer *et al.*, 2020). Chromosome conformation capture techniques, such as Hi-C, provide insights into the genomic architecture by accurately detecting chromatin interactions within the nucleus, as well as inter- and intra-chromosomal interactions (Belton *et al.*, 2012) at a high resolution. This Hi-C genomic conformation data is then further analyzed computationally to discover the substructures of the three-dimensional folding of DNA. One such substructure are topologically associated domains (TADs). As part of the hierarchical folding, they display strong intra-actions and boundary regions separate neighboring TADs. They have been found in various mammalian species and can consist of a minimum of 10,000 and up to a maximum of 5,000,000 individual bases (Kempfer *et al.*, 2020; Dixon *et al.*, 2012). They have also been shown to play a crucial role in a multitude of genomic functions in addition to gene expression, such as replication timing, enhancing, and promoter interactions. The boundary regions between two TADs are strongly enriched with many genomic markers, such as insulator proteins, e.g. CTCF, active transcriptions markers, and housekeeping genes (Rao *et al.*, 2014; Spiro *et al.*, 2022). Given the interactions captured by the Hi-C contact maps and the enrichment information for genomic markers, many different approaches and algorithms have been deployed to obtain the exact location of TADs. These approaches include the Arrowhead caller (Rao *et al.*, 2014), spectralTAD (Cresswell *et al.*, 2020), preciseTAD (Spiro *et al.*, 2022), TopDom (Shin *et al.*, 2015) and the Sub-Compartment Identifier (Ashoor *et al.*, 2020). Although each algorithm claims to be precise, there are differences in the genomic locations they determine as TADs. These differences in location stem from a multitude of difficulties with determining TADs: the Hi-C matrices can be compromised by noise, as well as varying coverage. Additionally, the lack of ground truth requires the creation and development of individual metrics to evaluate the precision and accuracy of these models (Dali *et al.*, 2017). For this project, we have developed an algorithm that utilizes both genomic annotations for enriched genomic markers as well as Hi-C matrices to determine TAD locations. From the input data, a graph is generated, which is then used in our neural network. The output is a list of TAD regions for each chromosome. Our Graph Neural Network (GNN) approach is driven by spectral clustering, inspired by (Bianchi *et al.*, 2020), and leverages their MinCut pooling layer (Fig. 1).

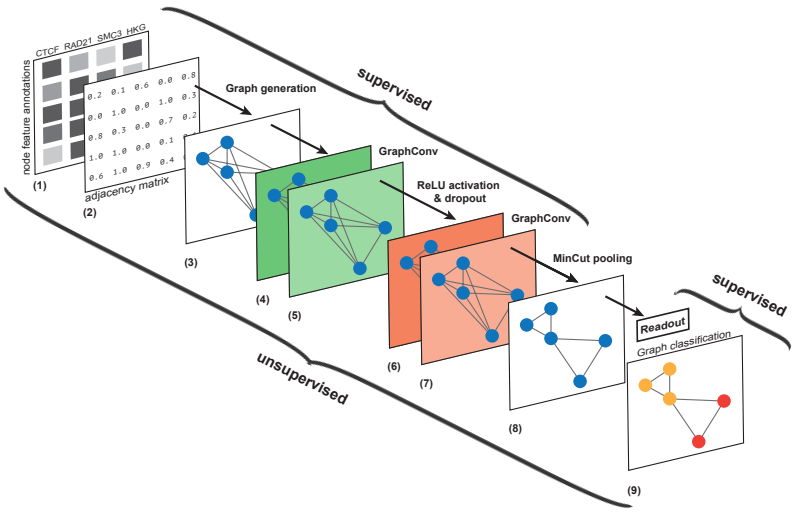

***Figure 1: Workflow of the MinCutTAD TAD detection algorithm**. First, the values of the genomic marker annotations for each bin on the chromosome are calculated (1). Then, the Hi-C interaction strength is assigned to the genomic bin in a chromosome (2). The information is aggregated into a graph domain (3). The subsequent four slides represent the neural net within the GNN and function as the message passing ((4)-(7)). If the model is trained in unsupervised training mode the intermediate layer is MinCut pooling (8). Lastly, the output of the GNN depends on the training mode: in supervised training the nodes are classified as either TAD or no TADs; while they are clustered into TAD regions in the unsupervised method (9).*

Materials & Methods

Data

The data used for our algorithm is comprised of the Hi-C matrix data in both 25kb and 100kb resolution for all chromosomes, except the Y chromosome, of the GM12878 and IMR-90 cell lines, as well as the genomic annotations for the individual base positions provided by the ENCODE database (Sloan *et al.*, 2016). The specific annotations used are CTCF, an insulator protein, RAD21, a double-strand-break repair protein, and SMC3, a structural maintenance protein for chromosomes, which interacts with RAD21 and others to form a highly conserved cohesion complex (Dixon *et al.*, 2012; Pati *et al.*, 2002). These annotations represent molecular markers of 3D chromatin structures and have been found to be highly up-regulated in the boundary areas of previously defined TAD regions (Rao *et al.*, 2014; Dixon *et al.*, 2012). In addition to these three markers, the housekeeping genes are also integrated into our data structure, as their signal has been shown to be enriched within TAD boundaries (Spiro *et al.*, 2022). The input graph domains for each chromosome were split for each cell line into a train, test and validation set, split into 0.6, 0.2 and 0.2 of the chromosome graphs, respectively. For the supervised prediction, we also generated the binary labels for each genomic bin, whether the bin is a TAD, 1, or no-TAD, 0, using the Arrowhead algorithm (Rao *et al.*, 2014). Our partner team generated their labels by combining the outputs of several TAD calling algorithms and various Hi-C matrix resolutions. The precise explanation to the generation of TAD labels in that manner can be found in their report.

To generate the graph domain from our described data, we combined the Hi-C matrices data with the genomic annotations. This generation is achieved by connecting and linking the individual genomic bins according to the information provided by the Hi-C matrix to form the nodes of the graph structure. Subsequently, each node is subsidized with genomic information. As the genomic annotations are provided on a base-level resolution, this subsidization is accomplished by aggregating the information within the genomic bins provided by the Hi-C matrices to become the node feature annotations.

Code and data availability

Our code can be found in our GitHub repository[1]

A sample dataset, which has been generated with our preprocessing pipeline, can be found here: https://1drv.ms/u/s!AvJpVBIXzqzAiNQ8xRJap-PdlQOtgQ?e=wEwyH5

The genomic annotations were downloaded from ENCODE. For cell line GM12878: CTCF[2], RAD21[3] SMC3[4]. For cell line IMR-90: CTCF[5], RAD21[6] and SMC3[7].

The housekeeping genes were published in the HRT Atlas (Hounkpe *et al.*, 2021). A list of housekeeping genes is provided in our GitHub repository[8].

Methods

Deep Learning, a machine learning method, has become ever more crucial for the analysis and discovery of patterns in highly complex data (Schmidt *et al.*, 2019). Convolution Neural Networks (CNN) have been shown to be able to extract principal information features from highly complex, non-linear data. This approach down-scales the feature complexity by smoothening local neighborhoods with pooling, i.e. removing redundant information from the feature space, and therefore, restraining

[1] https://github.com/meet-eu-21/Team-HA1/
[2] https://www.encodeproject.org/annotations/ENCFF074FXJ/
[3] https://www.encodeproject.org/annotations/ENCFF110OBQ/
[4] https://www.encodeproject.org/annotations/ENCFF049WIK/
[5] https://www.encodeproject.org/annotations/ENCFF276MRX/
[6] https://www.encodeproject.org/annotations/ENCFF374EXW/
[7] https://www.encodeproject.org/annotations/ENCFF476RFS/
[8] https://github.com/meet-eu-21/Team-HA1/blob/main/ressources/Housekeeping_GenesHuman.csv

the model size (Albawi *et al.*, 2017). This approach is limited by the ordered structure of the input data. While GNNs build upon the same convolution method to summarize local neighborhoods, they are not restricted to the strict ordering of the input and can process unordered structures such as graph domains (Zhou *et al.*, 2020).
We developed a GNN algorithm driven by spectral clustering to detect TADs from Hi-C interaction matrices. The approach is guided by the work of (Bianchi *et al.*, 2020) and their use of a MinCutPool layer, inspiring the name of our algorithm MinCutTAD.

Our generated graph data is processed by two algorithms. Initially, the supervised learning algorithm is trained to map the input data (the graph nodes) to the specific label obtained predictions of the Arrowhead Caller for this genomic bin. Throughout the training, a scoring function measures the ability of the algorithm to map the data correctly with the aim to obtain minimal error values. As the Arrowhead algorithm is recognized as a well performing model, we wanted to replicate this level of performance with our supervised model, by providing evidence that our MinCutTAD model can correctly classify between TAD and no TAD regions. By obtaining a similar performance, we justify the application of the unsupervised learning method. Our unsupervised algorithm does not require any additional labels and its goal is not to classify the individual genomic bins as TAD or no TAD regions. Instead, the unsupervised model is given a number of clusters and tries to learn abstractions of the data by which it is able to determine differences among the samples and distinguish and assign them accordingly into subclusters (Zhao *et al.*, 2007; Cresswell *et al.*, 2020). These subclusters represent the aggregated TAD regions. Therefore, the unsupervised model does not only focus on predicting whether the genomic bin is a TAD or no TAD, but also predicts entire TAD regions.

For both types of learning, we utilize individually fitted models and we have incorporated two different message passing layers. Preceding the application of the Graph convolutional layer, the nodes are made up of the weighted sum of their own and adjacent node features. We have applied GATConv and GraphConv as convolutional elements in our model (Fig. 2). GraphConv is a basic convolutional layer, which, as previously mentioned, combines the current nodes' features $\boldsymbol{x}_i$ with the surrounding node features $\boldsymbol{x}_j$ to smoothen the gradient between them and exchange important information. Because involving every node in the convolution or message passing for one node would lead to an over smoothing and loss of information, a threshold is set, to define up to which order of connections are included. Only nodes that are directly connected to $\boldsymbol{x}_i$ are considered first-order connections, and the threshold was set at first-order connections. In our dataset for the 25kb GM12878 cell line a single genomic bin node is interconnected with approximately 3,100 other nodes. The set of nodes that fulfill this requirement is denoted $\mathcal{N}_i$, the local neighborhood of $\boldsymbol{x}_i$. The edge weight from the source node to the adjacent nodes is denoted as $\omega_{\mathrm{j,i}}$ and Θ represents a weight matrix. Once the convolution is complete the node $\boldsymbol{x}_i$ is replaced by the newly, neighbor-aggregated $\boldsymbol{x}'_i$ vector. This convolution is described in equation 1 (Morris *et al.*, 2019):

$$\boldsymbol{x}'_i = \Theta_1 \boldsymbol{x}_i + \Theta_2 \sum_{j \in \mathcal{N}(i)} \omega_{j,i} \times \boldsymbol{x}_j \qquad [1]$$

In addition to the GraphConv Layer, we also utilized the GATConv layer, which extends the GraphConv layer by adding attention to the node aggregation.

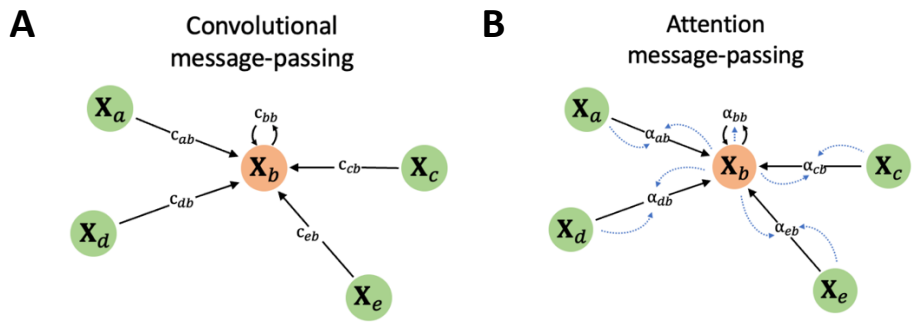

Figure 2: Schematic representations of convolutional and attention message passing layers*. Subfigure A displays the convolutional layer. The current node (orange) is updated with the information it obtains from its surrounding nodes via convolution and passes on its information into its updated node. Subfigure B exhibits the attention layer. Here, the current node is not just updated by the information of the surrounding nodes, but also scales the magnitude of the given information for its updates node features and weights the incoming information accordingly (Adapted from Bronstein et al., 2021).*

Given these smoothening operations of the node features by the GNN, it becomes interesting to obtain insights into the importance each node feature plays to the labeling of the node as TAD or no-TAD. Therefore, we have also utilized the GNNExplainer, which is initialized with the model and is applied directly following the training stage. The explainer reports the importance of the node features in the attribution of one node to a class. The GNNExplainer provides an explanation of the relevance of each of the 4 different node features incorporated in the data; how they compare within one class and in-between the two classes: TAD and no-TAD region (Ying *et al.*, 2019).

After the message passing, the pooling operation is applied to the graph domain MinCutPool (Bianchi *et al.*, 2020). Spectral clustering can be used to identify similarities between the nodes and aggregate closely related notes into sub-clusters. The base concept of this pooling strategy is to apply a tailored loss term, as well as the continuous relaxation of the MinCut problem, representing a cut within a graph that is minimal in a metric in the pooling layer of the GNN. As a result, the GNN must learn how to determine minimum cut in the graph domain to combine the clusters and thereby decrease the domain graph size (Bianchi *et al.*, 2020). In the case of TADs, this pooling strategy corresponds to finding and combining connected TAD clusters and aggregating them to reduce the graph size.

Given the differences between the two learning approaches, the evaluation metrics for them vary. The supervised approach learns to classify the given nodes, the genomic bins, in a binary manner into either TAD or no TAD. For binary models, it is common practice to evaluate them using the area under the receiver operating characteristics curve (AUROC). The trained model is applied to the withheld testing set to compute the scores for each input. The ROC plot shows the true positive rate against the false positive rate for different thresholds of the predicted score. To condense this information, the area under the curve is computed, thereby representing the performance of the binary model in a single value, the AUROC, which lies between 0 and 1, where 0.5 represents random classification (Parker *et al.*, 2011).

For the unsupervised learning approach, we utilize the output of the supervised learning to validate the functionality of the model's set-up and learning capabilities. The loss of the unsupervised model is made up of two loss functions. The first loss term, L_c, animates the model to cluster closely related nodes into one subcluster and the second loss term, L_o, animates the model to generate equally sized clusters. These loss terms are then added up to equal the overall loss of the model. To score the model, we compute the silhouette coefficient for one sample as follows:

$$s(i) = \frac{\mathbf{b}(i) - \mathbf{a}(i)}{\max\{\mathbf{a}(\mathrm{i}), \mathbf{b}(\mathrm{i})\}} \quad [2]$$

The silhouette coefficient is a metric that calculates the mean distance of nodes corresponding to the same cluster (**a**) and the mean distance to the nodes of the nearest cluster (**b**) for every node. Using these computed values, the silhouette coefficient provides an insight into the power of the model to distinguish among clusters and aggregate the graph domain into a specified number of clusters. A higher value indicates a better separation of the clusters (Shutaywi *et al.*, 2021).

Table 1: Hyperparameters of the evaluated models and training times for various model set-ups.

Hyperparameter Table

Applies to both models

Learning rate	0.5, with LR scheduler	Message passing layer	GraphConv
Max. Epochs	100	Number layers	2
Dropout	0.5		

Training times

Method	Data	Training time
Supervised	25kb, GM12878, unfiltered Hi-C/genomic data	30-45 min
Supervised	25kb, GM12878, 100 Hi-C threshold	30-45 min
Supervised	100kb, GM12878, unfiltered Hi-C/genomic data	20-40 min
Unsupervised	100kb, GM12878, unfiltered Hi-C/genomic data	500 min

Predictions for all chromosomes using a trained model	5 min

Results

The first experiment was to evaluate the performance of the supervised GNN model. This model was initialized with the GraphConv message passing layer and without the MinCut pooling, only considering the first-order neighbors of the nodes. The following results are obtained by testing on the chromosome 15, which was chosen at random from the testing set. First, the supervised model was tested on the 25kb and 100kb Hi-C resolution of the GM12878 cell line. The 100kb resolution displayed higher performance with an AUROC of 0.91 compared to the performance of the model on the 25kb resolution, with an AUROC of 0.84 (Fig. 3A).

Further experiments were performed to investigate the influence of the data on the model's performance. Both the 25kb resolution Hi-C matrix data, as well as the genomic annotations, were filtered in individual experiments. When filtering the interactions with a threshold, every value within the Hi-C matrix below the threshold is set to 0. When comparing different thresholds, it was observed that with an increasing threshold the performance of the model dropped compared to the baseline performance. With a threshold of 10, 13% percent of the entries within the matrix retain values above 0 and the performance of the model declines to 0.82 (Fig. 3B). When increasing the threshold to 100, merely 2% of the data entries remain above 0 in the matrix and the performance drops further to 0.72 (Fig. 3B). Subsequently, the genomic annotations were filtered and compared to the model's baseline on the 25kb resolution matrix. For this experiment, the occurrence of the genomic marker per bin was split into quantiles and instead of using the exact value per genomic annotation, the genomic annotation values were turned into binary values. Three experiments were conducted, with the thresholds set to the 25^{th}, 50^{th} and 75^{th} quantile, respectively. The values below the threshold were set to 0 while values above were set to 1. All three thresholds displayed a decline in performance compared to the baseline of 0.84 at 0.76, 0.64 and 0.72, respectively.

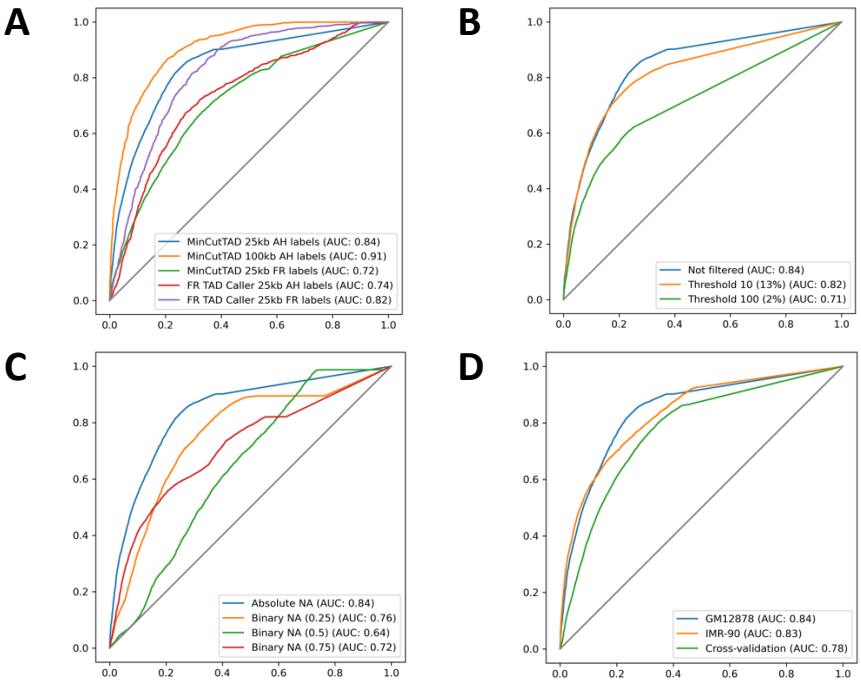

Figure 3: ROC performances of various supervised models on chromosome 15. *Subfigure A displays the ROC for the MinCutTAD model trained on the 25kb and 100kb resolution of the GM12878 cell line and the model of our partner team. (FR TAD Caller) trained on 25kb resolution. Different algorithms used for labeling were compared; Arrowhead (AH labels) and the combined algorithm preferred by our partner team (FR labels). Subfigure B exhibits the performance of the model trained on the 25kb GM12878 cell line, along with the training on the same data with two thresholds applied: 10 and 100. Subfigure C displays various trainings on the 25kb resolution GM12878 cell line data with a binary filter on the genomic annotations, with a multitude of thresholds: the 25th, 50th and 75th percentile. Subfigure D displays the performance of the model trained on the 25kb resolution of both the GM12878 and IMR-90 cell line, as well as the performance for the model trained exclusively on GM12878 cell line data and evaluated on the IMR-90 cell line data.*

Following the threshold experiments, an experiment was set up to evaluate the model trained on one cell line and tested on another cell line. This experiment was executed to investigate previously reported findings by existing models that TADs are conserved throughout evolution (Spiro *et al.,* 2022). We trained and tested our supervised model individually on both the GM12878 and IMR-90 cell lines and then cross-validated them. This cross validation was performed using the 25kb resolution Hi-C matrix and the performance of IMR-90 was evaluated to have a AUROC value at 0.83, displaying a minor decrease in performance. The model trained on the GM12878 cell line and tested on the IMR-90 cell line was evaluated to have a AUROC value of 0.78 (Fig. 3D).

In addition to the performance of the model, the underlying importance of each node feature for the supervised model trained on the 25kb GM12878 cell line was evaluated. The GNNExplainer showed that the importance of each feature varies. Within both classes, TAD and no-TAD, CTCF and RAD21 display a greater mean importance to the classification of the node than SMC3 and the housekeeping genes. Additionally, the mean importance of CTCF and RAD21 for the TAD detection was more vital to the classification than for the no-TAD classification (Fig. 4). The importance scores for each node feature showed a general tendency to be higher when calling TAD regions, and lower when calling no-TAD regions. Although this was the general tendency, a slight bimodal distribution, and thus an overlap in importance scores for TAD and no-TAD regions could be seen, especially for CTCF and RAD21. This overlap shows that some regions, although called as no-TAD, place a high importance on the genomic features, even though this importance is generally higher for TAD regions (Fig. 4).

A closer look also showed that the genomic bins that are called as TADs by Arrowhead showed a higher signal strength for the genomic annotations than the genomic bins not classified as a TAD (Fig. 5A). This change in genomic signal effect could not be seen for genomic bins classified as TADs or no-TADs by the tool our partner team uses for labeling. Additionally, the variability of the data for SMC3 was much lower than for CTCF and RAD21 corresponding to the lower importance score of SMC3 (Fig. 4).

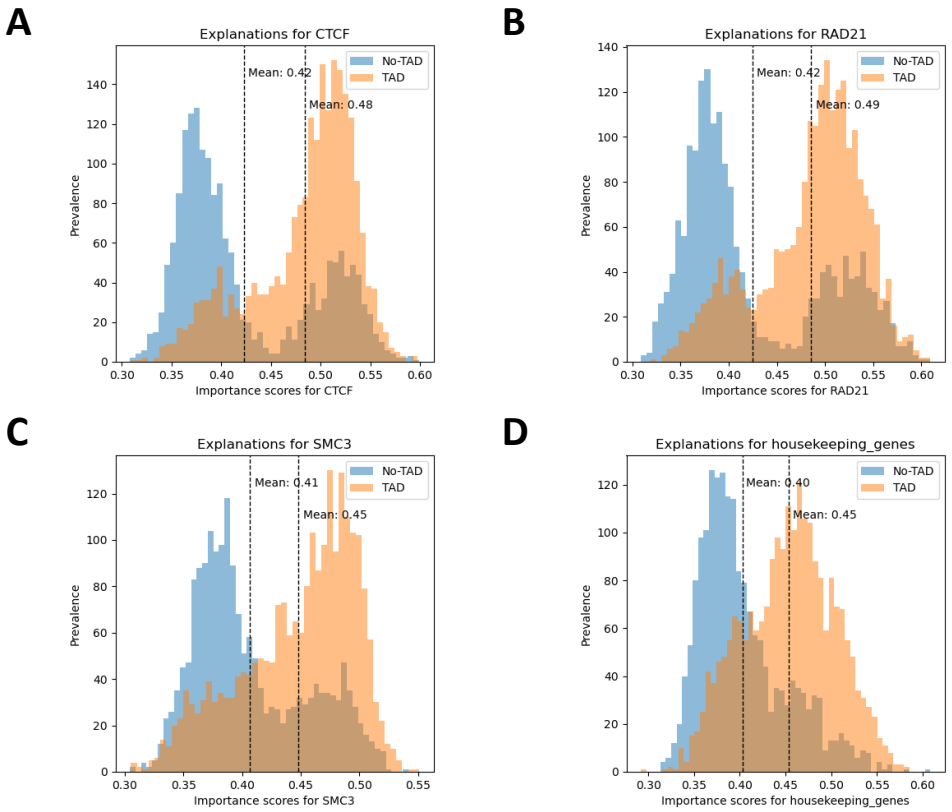

Figure 4: The importance of the node features to the GNN node label classification on chromosome 15. *Subfigures A through D represent histograms which plot the occurrence of each importance score for all four node features, CTCT (A), RAD21 (B), SMC3 (C) and housekeeping genes (D), respectively.*

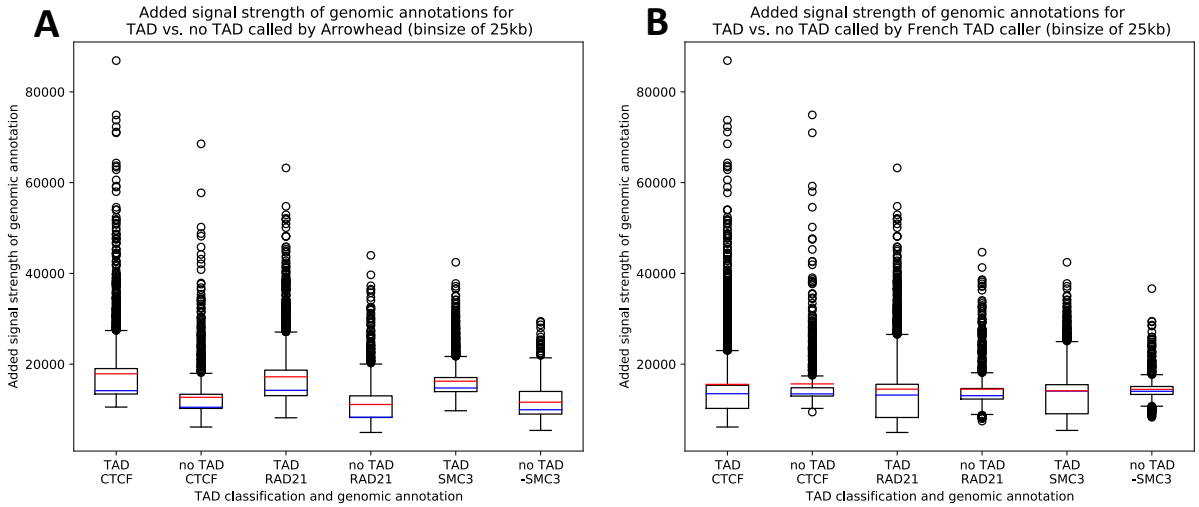

Figure 5: Boxplots of genomic annotations for TADs called by Arrowhead (A) or the TAD caller used by our French partner team (B) for chromosome 15: *The signal for the three different genomic annotations CTCF, RAD21 and SMC3 is added up for each genomic bin, depending on whether it is classified as a TAD or not. The median value is depicted as a blue line and the mean values as a red line. In subfigure A the genomic bins were classified by Arrowhead, in subfigure B the genomic bins were classified by the TAD caller used by our French partner team.*

Our partner team used a supervised CNN approach to predict the borders of a TAD region, whereas our approach was to use a more complex GNN to predict TAD regions as a whole. For this, they first used the Arrowhead algorithm for labeling and received an AUROC of 0.74 for a resolution of 25kb. Their other concept was to combine several algorithms and if 5 or more of them classify a genomic bin as a TAD, it is called as a TAD. When they used this algorithm for the labeling, they receive an

AUROC of 0.82. When our model was trained using these labels, an AUROC of 0.72 was obtained (Fig. 3A).

With the supervised model evaluated, we conclude that the model is capable of learning TAD regions. This conclusion enables us to train our model in unsupervised manner. The model was initialized with the GraphConv layer and includes the MinCut pooling layer. The model was trained on the 100kb GM12878 Hi-C data due to memory issues when using 25kb data. The goal for this model is to determine the number of TAD regions that can be found in a chromosome. After an initial improvement of performance with an increase in the number of clusters from 0 to 100, the performance flattened out when further increasing the number of clusters (Fig. 5). The MinCutTAD was able to distinguish 15 to 20 TAD regions on average per chromosome. Although the number of clusters was furthered increased, this 15-20 cluster average remained constant. Remarkably, the silhouette scores, independent of the number of clusters, are negative.

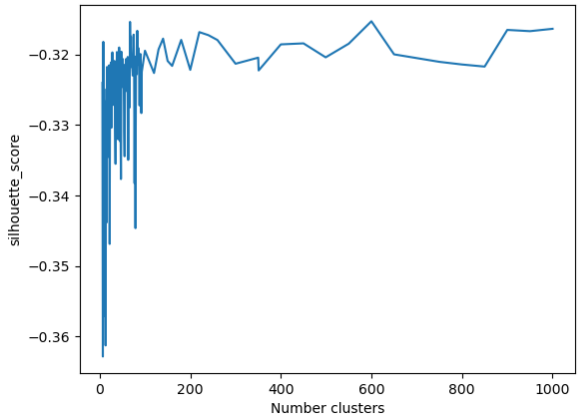

Figure 6: Silhouette score of the unsupervised model. *Plot of the silhouette score, with the number of clusters the GNN is asked to divide the data into on the x-axis, and the silhouette score on the y axis.*

Discussion

With the model architecture optimized, the developed supervised MinCutTAD was benchmarked on both the GM12878 and IMR-90 cell line. Initially, the two Hi-C resolutions of the GM12878 were tested for 25kb and 100kb. In a direct performance comparison, the models' performance varies marginally, with a performance of 0.84 (25kb) and 0.91 (100kb) (Fig. 3D). This difference in performance based on the different cell lines is coherent with previous models, indicating that some areas in different cell lines remain constant throughout evolution (Spiro *et al.*, 2022; Dixon *et al.*, 2012).
Applying different thresholds to the Hi-C matrix data, as well as to the genomic annotation data revealed that the reduction of information within these features leads to a decrease in performance of the model (Fig. 3B and C). Because of these findings, we decided not to apply any thresholds on the data for further training and testing, instead maintaining the data as provided in the Hi-C matrix and from the ENCODE genomic annotations.
The performance decline of the cross-validation is only minor on the 25kb resolution supervised model. Compared to the 0.84 and 0.83 AUROC scores of the independent training and testing on one cell line, the performance of the cross-validation model with an AUROC of 0.78 proves that our model can generalize. Compared with the model trained and tested on chromosomes of the same cell line, the training in cross-validation is performed with all chromosomes from one cell line, while the validation and testing is performed with the chromosomes of the other cell line.

While evaluating the importance scores, relevant differences are found between the 4 node features. First, the mean values for the genomic annotations are higher for regions classified as TADs by Arrowhead than for regions not classified as a TAD (Fig. 4). This finding further supports the approach to use genomic annotations as node features in our GNN. The importance score of SMC3 is probably

lower than for CTCF and RAD21, because the signal value for SMC3 has a lower variability (Fig. 4). The score is lower for the housekeeping genes as well, since, for example, chromosome 15 has one to three housekeeping genes in only 51 of 4101 genomic bins.
The distribution of the individual scores display distinct peaks for unique importance scores (Fig. 4). When looking at the distribution for CTCF, the majority of importance scores that contribute to the classification of a TAD lie between 0.5 and 0.55 (Fig. 4).
Despite these clear peaks, within the peak boundaries are importance scores that contribute to classifications resulting in the opposing label. These smaller importance score peaks, which attributed to the opposite label, may arise from either wrong classification of the MinCutTAD model or Arrowhead solution. Although the Arrowhead solutions are recognized as good predictions, they do not represent the ground truth and might include wrong labels to some extent for TAD regions within the genome.

Our algorithms optimal result on a 25kb resolution using Arrowhead for labeling (AUROC of 0.84) is slightly higher than the optimal result of our partner team using their combined approach for labeling (AUROC of 0.82)(Fig. 3A). Nevertheless, if their model is trained on the Arrowhead labels, the AUROC is much lower than for our model (0.74 vs. 0.84). Accordingly, if our model is trained on their labeling approach the AUROC is much lower as well (0.72 vs 0.82). This score might be explained by the association of the genomic annotations with the TADs used for labeling. The TADs predicted by the tool our French partner team uses for labeling do not correlate as strongly with the occurrence of CTCF, RAD21 and SMC3 as the TADs predicted by Arrowhead (Fig. 6). Since the performance of our MinCutTAD algorithm is based on genomic annotations, the predictions using the labels of our partner team cannot be as good as using Arrowhead for labeling. Also, this finding could suggest that the TADs called by their labeling tool have a lower biological accuracy.

In the unsupervised approach, the genomic bins are clustered into 15-20 TAD regions on average. Even when the number of clusters is set higher, a similar number of clusters of TAD regions are determined; the model cannot find a higher abstraction level in the data. This clustering might be the reason for the negative silhouette score (Fig. 6). Since the clustering is performed based on the similarity between the node annotations, each cluster includes similar nodes that do not have to correspond to neighboring bins. This clustering method results in a lower number of detected TADs with a higher mean length on each chromosome. This method could be improved by integrating a sliding window into the clustering, which was implemented by (Cresswell *et al.*, 2020). In the Cresswell approach, the TADs at the beginning of the Hi-C matrix that lie inside the sliding window are identified and the window is then moved to the beginning of the last identified TAD (Cresswell *et al.*, 2020). This sliding window would add another hyperparameter, but could potentially help to cluster smaller TAD regions and measure positive silhouette scores in unsupervised training. In this case, the background is subtracted, and genomic bins are not assigned in many clusters. Additionally, our model could be refined in terms of detecting subgraph structures – the TAD region nodes, which are strongly interconnected between themselves, but less interconnected with the surrounding background – to detect more concise and smaller TAD regions as described by (Alsentzer *et al.*, 2020). Moreover, this approach could help in the detection of hierarchical TADs.
For the unsupervised training, the role of the loss terms needs to be evaluated. The loss term L_o has shown to be useful in creating approximately equally sized clusters but can be problematic when detecting hierarchical TADs of different size. Furthermore, a third loss term could be added, which forces the model to cluster direct neighbors together in regions, which is not the case yet, as the model is clustering the genomic bins based on their node annotations and graph neighborhood similarity.

The previously discussed results have been achieved in a not fully optimized model. The layers within the GNN have been tested in a multitude of arrangements and set ups, however, not covering all possible scenarios, and the values of the other hyperparameters of the model were not optimized.

One parameter is the value thresholding on the Hi-C matrix. Only the two thresholds 10 and 100 were tested. With decline in performance in both cases no further experiments were conducted, and as the computational time was not reduced with this altered data, the model only converged faster with the applied thresholds. Nonetheless, further thresholding experiments, such as adjusting the threshold value to the chromosome, could be relevant.

By training our GNN for supervised and unsupervised tasks we have proven the versatility of our tool. Also, a fully trained model can easily be used by any external scientist even if no HPC or GPU is available. With the trained weights of the model, the predictions for all chromosomes require five minutes of computational time (Tab. 1). Our model is faster than Arrowhead and TopDom, which require up to five minutes computational time per chromosome (Rao *et al.*, 2014; Shin *et al.*, 2016).

The TAD regions of our model were compared to the models Arrowhead and TopDom. These Venn diagrams display the similarity of the detected TAD regions, as a large proportion of the overall detected TAD regions is the same in TopDom, Arrowhead and our supervised MinCutTAD algorithm (Fig. 7A). In contrast, the MinCutTAD unsupervised method did show a smaller overlap with the Arrowhead and the MinCutTAD supervised method (Fig. 7B). This lack of overlap can be explained by the fact that in the unsupervised method even distant bins are clustered together as TAD regions. Detected TAD regions are spread across the chromosome. Additionally, there is a large overlap in bins detected as TADs between Arrowhead, our MinCutTAD model and our partner teams' model (Fig. 7C).

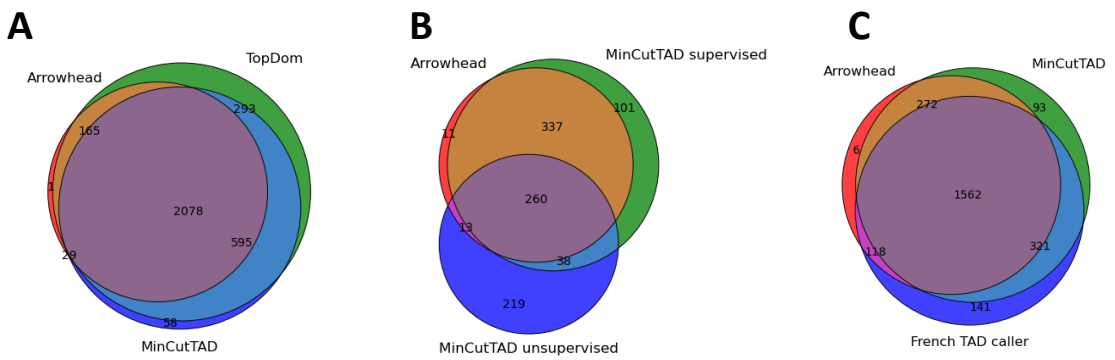

Figure 7: Venn Diagrams comparing different TAD detection methods*. In subfigure A the Venn diagram shows the overlapping tad bins detected by the methods Arrowhead, TopDom, and MinCutTAD supervised for chr15 at a 25kb resolution. In subfigure B the Venn diagram shows the overlapping tad bins detected by the methods Arrowhead, MinCutTAD supervised and MinCutTAD unsupervised for chr15 at a 100kb resolution. In subfigure C the Venn diagram shows the overlapping tad bins detected by the methods Arrowhead, MinCutTAD supervised, and our partner teams' model for chr20 at a 25kb resolution.*

In this project, we have demonstrated the application power of the MinCutTAD on various metrics and have shown it to be a feasible model to investigate and detect TAD regions. However, the current implementation still has drawbacks, such as high density in the interaction tensors leading to memory issues. This issue prevents the applications of alternative convolutional layers like the GATConv. However, these layers could enhance the prediction as the attention mechanism enables to assign different weights to the nodes in the message passing operation. This could be achieved by using sparse tensors and the development of more memory-efficient implementations. Furthermore, the validation on more cell lines may offer further insights into the generalizability of model, especially in the context of cross-validation testing with various cell lines. Lastly, it would be of interest in further experiments, to evaluate the performance impact and node feature importance for a multitude of further genomic annotations. One especially interesting genomic marker is the H3K9me3 of the histones, as this has been linked to the evolution of the 3D structure, as well as a driver for the development of the 3D structure within the DNA (Yu *et al.*, 2017).

Sources

Ashoor H, Chen X, Rosikiewicz W, Wang J, Cheng A, Wang P, Ruan Y, Li S. Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. Nat Commun. 2020 Mar 3;11(1):1173. doi: 10.1038/s41467-020-14974-x. PMID: 32127534; PMCID: PMC7054322.

S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

Alsentzer, Emily & Finlayson, Samuel & Li, Michelle & Zitnik, Marinka. (2020). Subgraph Neural Networks.

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012 Nov;58(3):268-76. doi: 10.1016/j.ymeth.2012.05.001. Epub 2012 May 29. PMID: 22652625; PMCID: PMC3874846.

Bianchi, Filippo Maria, Daniele Grattarola, and Cesare Alippi. "Spectral clustering with graph neural networks for graph pooling." *International Conference on Machine Learning*. PMLR, 2020.

Brody, S., Alon, U., & Yahav, E. (2021). How Attentive are Graph Attention Networks? *ArXiv, abs/2105.14491*.

Bronstein, M.M., Bruna, J., Cohen, T., & Velivckovi'c, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. ArXiv, abs/2104.13478.

Cresswell, K.G., Stansfield, J.C. & Dozmorov, M.G. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* **21,** 319 (2020). https://doi.org/10.1186/s12859-020-03652-w

Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res*. 2017;45(6):2994-3005. doi:10.1093/nar/gkx145

Dixon, J., Selvaraj, S., Yue, F. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012). https://doi.org/10.1038/nature11082

Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. 2021 Jan 8;49(D1):D947-D955. doi: 10.1093/nar/gkaa609

Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2020 Apr;21(4):207-226. doi: 10.1038/s41576-019-0195-2. Epub 2019 Dec 17. PMID: 31848476.

Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., & Grohe, M. (2019). Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks. *ArXiv, abs/1810.02244*.

Parker, C. (2011). An Analysis of Performance Measures for Binary Classifiers. In 2011 IEEE 11th International Conference on Data Mining pp. 517–526,.

Pati D, Zhang N, Plon SE. Linking sister chromatid cohesion and apoptosis: role of Rad21. Mol Cell Biol. 2002 Dec;22(23):8267-77. doi: 10.1128/MCB.22.23.8267-8277.2002. PMID: 12417729; PMCID: PMC134054.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Epub 2014 Dec 11. Erratum in: Cell. 2015 Jul 30;162(3):687-8. PMID: 25497547; PMCID: PMC5635824.

Schmidt, J., Marques, M. R. G., Botti, S. and Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials *5*, 83.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016 Apr 20;44(7):e70. doi: 10.1093/nar/gkv1505

Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*. 2021; 23(6):759. https://doi.org/10.3390/e23060759

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe G, Podduturi NR, Tanaka F, Hong EL, Cherry JM. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016 Jan 4;44(D1):D726-32. doi: 10.1093/nar/gkv1160

Spiro C Stilianoudakis, Maggie A Marshall, Mikhail G Dozmorov, preciseTAD: a transfer learning framework for 3D domain boundary prediction at base-pair resolution, *Bioinformatics*, Volume 38, Issue 3, 1 February 2022, Pages 621–630, https://doi.org/10.1093/bioinformatics/btab743

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio', P., & Bengio, Y. (2018). Graph Attention Networks. *ArXiv, abs/1710.10903*.

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in neural information processing systems, 32*, 9240-9251 .

Yu, C., H. Gan, A. Serra-Cardona, L. Zhang, S. Gan, S. Sharma, E. Johansson, A. Chabes, R. M. Xu and Z. Zhang (2018). A mechanism for preventing asymmetric histone segregation onto replicating DNA strands. Science 361: 1386-1389.

Zheng Zhao and Huan Liu. 2007. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (ICML '07). Association for Computing Machinery, New York, NY, USA, 1151–1157. DOI: https://doi.org/10.1145/1273496.1273641*

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., & Sun, M. (2020). Graph Neural Networks: A Review of Methods and Applications. *ArXiv, abs/1812.08434*.