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

Lucas Arnoldt, Paul Kittner, Stefanie Mantz and Charlotte Westhoven
Supervised by: Carl Herrmann

Research questions:

1. Is it possible to predict whether a genomic bin is a TAD region using a Graph neural network (GNN)?
2. To what extent can genomic annotations that are associated with TADs be used for the prediction? And how can their importance for the predictions be interpreted?

Used datasets: Hi-C matrices & genomic annotations (CTCF, RAD21, SMC3, number of housekeeping genes) for the provided genomic loci of chromosomes

Two approaches for the GNN:

- **Supervised** uses Arrowhead solutions as labels for the genomic bins and optimizes towards classifying the graph nodes accordingly.
- **Unsupervised**: no labels are provided to the model. It determines whether regions belong to a TAD or not and aggregates them. Therefore, its main goal is to cluster single TAD regions together.

The MinCutTAD model: GNN algorithm driven by spectral clustering to detect TADs. Constructed with GraphConv, a message passing layer, and if unsupervised, with a MinCut pooling layer (Bianchi et al., 2020).

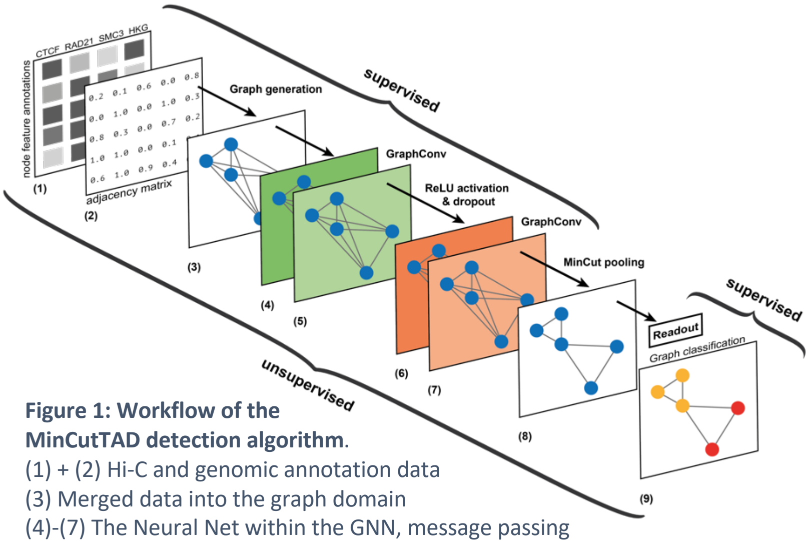

Figure 1: Workflow of the MinCutTAD detection algorithm.
(1) + (2) Hi-C and genomic annotation data
(3) Merged data into the graph domain
(4)-(7) The Neural Net within the GNN, message passing
(8) Applied MinCut pooling guided by (Bianchi 2020)
(9) Final output of the model

- **Message passing** refers to the smoothening of the information among the directly surrounding node features.
- **Pooling** refers to the aggregation of strongly similar nodes, thereby reducing the graph domain and forming sub clusters.

Conducted experiments:

The evaluation of the supervised training approach, without any pooling operation, provides insight into the performance of the model and if an unsupervised model is feasible.

- Evaluated on its performance of TAD bin detection on the test chromosomes
- The GNNExplainer provides measurements for the importance scores for each genomic annotation, determining their contribution to the classification (Ying et al., 2019).
- Generalizability of the model is verified by training on the GM12878 cell line and testing its performance on the chromosomes of the IMR-90 cell line.

If the MinCutTAD model is able to detect the underlying features of TADs and distinguish them from no TAD regions, the next step is to utilize the unsupervised model to investigate its power to determine TADs without the comparison to the true labels and aggregating these TADs into subclusters in the graph domain.

- The performance of the unsupervised model is examined by comparing the amount of overlapping TAD regions predicted by MinCutTAD and alternative models, such as Arrowhead and TopDom.

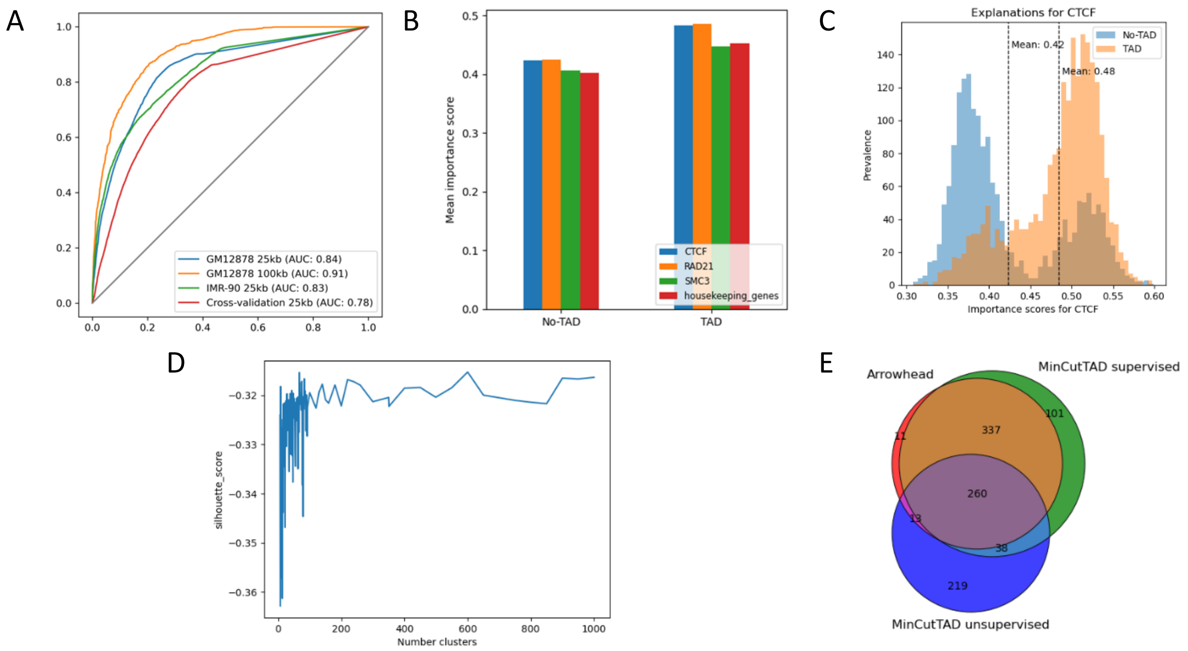

Figure 2: Performance measures and interpretations for the supervised model on chromosome 15 and the silhouette score of the unsupervised model.

Subfigure A displays the performance on the test set for the 25kb and 100kb resolution of the Hi-C matrix for the GM12878 cell line, the performance for the 25kb resolution for the cell line IMR-90, as well as the performance of the model trained on all GM12878 chromosomes and tested on the IMR-90. Subfigure B displays the mean importance score for all node annotations for each label while subfigure C exhibits the distribution of the importance scores for CTCF. The silhouette score of the unsupervised model is shown in subfigure D. In subfigure E a Venn diagram with the overlapping TAD regions between Arrowhead, MinCutTAD supervised and MinCutTAD unsupervised is displayed.

Results:

- The supervised model can replicate the performance of the Arrowhead TAD caller. It achieves an AUROC score of 0.84 and 0.91 on the 25kb and 100kb resolution of the GM12878 cell line (Fig. 2A).
- The model is able to learn the general underlying features of TADs and apply them to other cell lines. Trained on all chromosomes of GM12878, except Y, the model achieves an AUROC of 0.78 on IMR-90 (Fig. 2A). This result is in line with previous findings of various models, which state that TADs are conserved throughout evolution.
- Looking at the importance scores and their distribution, two things arise. Firstly, the distribution and mean importance scores are different for each node feature, and it becomes apparent that CTCF and RAD21 are the most meaningful for the classification for both labels, with a slightly stronger impact on the TAD label (Fig. 2B, C). Secondly, high importance scores can also be found for the no-TAD labeled bins. This score may arise from wrongly labeled nodes by the MinCutTAD model or incorrect true labels provided by the Arrowhead TAD caller.
- The results above provide evidence of the functionality of the supervised MinCutTAD method; therefore, the unsupervised approach can be tested. The unsupervised model yields an average of 15 to 20 TAD regions detected on the tested chromosomes for the maximum silhouette score (Fig. 2D). The model is not able to detect more concise and smaller TAD regions, which could be interpreted as hierarchical TADs.
- The Venn diagram shows that the supervised MinCutTAD model has a large overlap in the number of bins classified as TADs as compared to Arrowhead. The unsupervised model has a much smaller overlap, possibly due to the fact that the unsupervised model aggregates similar graph nodes, despite the fact that these nodes are not necessarily neighboring TAD regions (Fig. 2E).

Sources:

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

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.