UC Riverside

UC Riverside Previously Published Works

Title

SHICEDO: single-cell Hi-C data enhancement with reduced over-smoothing

Permalink

https://escholarship.org/uc/item/6qg0w6sw

Journal

Bioinformatics, 41(12)

ISSN

1367-4803

Authors

Huang, Jingong

Ma, Rui

Strobel, Michael

et al.

Publication Date

2025-12-01

DOI

10.1093/bioinformatics/btaf575

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

J_ID: Customer A_ID: **BTAF575** Copyedited by: Manuscript Category: **Original Paper** Cadmus Art: **OP-CBIO250610**

NOT FOR PUBLIC RELEASE

Author Query Form

Journal: *Bioinformatics*
Article Doi: 10.1093/bioinformatics/btaf575
Article Title: SHICEDO: single-cell Hi-C data enhancement with reduced over-smoothing
First Author: Jingong Huang
Corr. Author: Tao Jiang and Wenxiu Ma

AUTHOR QUERIES – TO BE ANSWERED BY THE CORRESPONDING AUTHOR

These proofs are for checking purposes only. They are not in final publication format. Please do not distribute them in print or online. Do not publish this article, or any excerpts from it, anywhere else until the final version has been published with OUP. For further information, see https://academic.oup.com/journals/pages/authors

Figure resolution may be reduced in PDF proofs and in the online PDF, to manage the size of the file. Full-resolution figures will be used for print publication.

Select each question and describe any changes we should make on the proof. Changes against journal style will not be made and proofs will not be sent back for further editing.

AQ1: Please note that we have received multiple versions of the article title. Check the version used and provide correct wording, if necessary, directly in the text. Note that significant changes to the title may be sent to the journal editor for approval.

AQ2: Please check the spelling and accuracy of all author names and affiliations, particularly for any co-authors. Also check that author surnames are correctly highlighted. This is to ensure that forenames and surnames are tagged correctly for online indexing. Incorrect names and affiliations may lead to an author not being credited for their work by funders, institutions, or other third parties. Make any changes directly in the text. Note that changes to your authorship list (adding or removing authors, changing designation of corresponding author) require additional approvals. Email journals.corrections@oup.com if this is required.

AQ3: Please provide the complete details (department/unit, university/institution, and postal address with road/street name) for the corresponding author directly in the text.

AQ4: Please note that the Abstract should be formatted as following. Format 1: Only the subheading “Summary, Availability and implementation”; Format 2: The subheadings “Motivation, Results, and Availability and implementation”, with no other subheadings. Please check and make any amends directly in the text.

AQ5: As per the journal style, the section heading ‘Methods’ has been changed to ‘Materials and methods’. Please check and confirm by making changes directly in the text.

AQ6: If your manuscript has figures or text from other sources, please ensure you have permission from the copyright holder. Also ensure that the copyright owner is properly credited, for instance in the figure legends. Make any changes directly in the text. For any questions about permissions, contact jnls.author.support@oup.com.

AQ7: This journal requires authors to list author contributions using CRediT taxonomies. Please check the author contribution section carefully to ensure CRediT information is accurate. Any changes will be sent to the editorial team for review. For more information on CRediT, see https://credit.niso.org/

AQ8: We have received the following files for publication as supplementary material. • supplementary_clean_without_red_color_changes. These will be published in the format provided, without edits. If any files need to be removed from this list, specify by responding to this query, stating the filename. If any files need to be added, upload them directly or email them with your proof corrections. If any files need to be updated, upload them directly or email them, without altering the filename. We will seek editor approval for any changes to supplementary material, which may delay publication of your article.

AQ9: Please ensure all "conflicts of interest" (Or "disclosures") have been included for you and your co-authors, and that this section is correct. Edit the text directly if changes are required.

NOT FOR PUBLIC RELEASE

AQ10: Journal policy requires authors to provide a data availability statement in their manuscript. Please confirm that this statement is included in your manuscript and that any required links or identifiers for your data are present in the manuscript as described or provide edits with the required information directly in the text. See the sample data availability statements at https://academic.oup.com/journals/pages/authors/preparing_your_manuscript/research-data-policy#data2

AQ11: Please check that funding is recorded in a separate funding section, if applicable. Use the full official names of any funding bodies and include any grant numbers. Insert any changes to the funding section directly in the text.

AQ12: Please provide the publisher location (city/state/country), publisher name and location where the proceedings were held for reference Hu et al. (2018) directly in the reference and if reference has more than 3 authors, please list the names of the first 3 authors followed by et al.; otherwise, provide the names of all the authors. Make changes directly in the reference.

AQ13: Please provide the page range for reference Wang et al. (2023) directly in the reference.

AQ14: The caption for Figure 3 currently contains reference to colour used within the figure. For accessibility purposes reference to colour should be avoided unless essential to the understanding of the content. If the reference to colour is not essential to the understanding of the content please provide updated wording for the figure caption to remove reference to colour.

AQ15: The caption for Figure 4 currently contains reference to colour used within the figure. For accessibility purposes reference to colour should be avoided unless essential to the understanding of the content. If the reference to colour is not essential to the understanding of the content please provide updated wording for the figure caption to remove reference to colour.

AQ16: Please note that the part labels in figure 2 is given in the artwork but not mentioned in the caption of Figure 2. Please provide an updated caption or confirm that this part of the figure should be deleted.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

Bioinformatics, 2025, btaf575
https://doi.org/10.1093/bioinformatics/btaf575
Advance Access Publication Date: 23 October 2025
Original Paper

OXFORD

Genome analysis

SHICEDO: single-cell Hi-C data enhancement with reduced over-smoothing

Jingong Huang[1], Rui Ma[2], Michael Strobel[1], Yangyang Hu[1], Tiantian Ye[2], Tao Jiang[1,3,*], Wenxiu Ma[2,3,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, United States
[2]Department of Statistics, University of California, Riverside, CA 92521, United States
[3]Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, United States

*Corresponding authors. Tao Jiang. E-mail: jiang@cs.ucr.edu; Wenxiu Ma. E-mail: wenxiu.ma@ucr.edu.

Associate Editor: Anthony Mathelier

Abstract

Single-cell Hi-C (scHi-C) technologies have significantly advanced our understanding of the 3D genome organization. However, scHi-C data are often sparse and noisy, leading to substantial computational challenges in downstream analyses. In this study, we introduce SHICEDO, a novel deep-learning model specifically designed to enhance scHi-C contact matrices by imputing missing or sparsely captured chromatin contacts through a generative adversarial framework. SHICEDO leverages the unique structural characteristics of scHi-C matrices to derive customized features that enable effective data enhancement. Additionally, the model incorporates a channel-wise attention mechanism to mitigate the over-smoothing issue commonly associated with scHi-C enhancement methods. Through simulations and real-data applications, we demonstrate that SHICEDO outperforms the state-of-the-art methods, achieving superior quantitative and qualitative results. Moreover, SHICEDO enhances key structural features in scHi-C data, thus enabling more precise delineation of chromatin structures such as A/B compartments, TAD-like domains, and chromatin loops.

Availability and implementation: SHICEDO is publicly available at https://github.com/wmalab/SHICEDO.

1 Introduction

The 3D genome architecture is fundamental to the regulation of key biological processes such as gene transcription, DNA replication, and cell division (Misteli 2020). Hi-C techniques (Lieberman-Aiden *et al.* 2009, Duan *et al.* 2010, Rao *et al.* 2014, Ma *et al.* 2015) provide genome-wide mapping of chromatin contacts, shedding light on the principles of 3D genome organization. Analyses of Hi-C contact frequency matrices have unveiled multiple levels of chromatin organization, including active and inactive (A/B) compartments (Lieberman-Aiden *et al.* 2009), topologically associated domains (TADs) (Dixon *et al.* 2012), and chromatin loops (Rao *et al.* 2014).

Recently, the advent of single-cell Hi-C (scHi-C) technologies (Nagano *et al.* 2013, 2017, Ramani *et al.* 2017) has further revolutionized the field by enabling the investigation of 3D genome architecture at the single-cell level, yielding invaluable insights into the variability and dynamics of spatial genome organization of individual cells. However, scHi-C data remain limited to a few cell lines or tissues due to experimental constraints and the high sequencing cost. Furthermore, the presently available scHi-C datasets often suffer from low sequencing depth, substantial sparsity, experimental biases, and noise, all of which pose significant computational challenges for downstream data analyses. Among these challenges, low sequencing depth and sparsity are particularly significant, as they substantially hinder comprehensive analyses of scHi-C data.

Several computational methods, commonly referred to as imputation or data enhancement approaches, have been developed to address this challenge. These methods aim to infer missing or sparsely captured chromatin contacts and computationally increase the effective sequencing depth, thereby reducing sparsity and improving the overall data quality and interpretability of scHi-C data. Due to the inherent sparse nature of scHi-C data, such methods typically do not increase the resolution of scHi-C matrices in the traditional sense, i.e. they do not reduce the size of genomic bins or increase the dimensionality of the contact matrices.

For example, scHiCluster (Zhou *et al.* 2019) used convolution and random walk with restart (RWR) imputation to mitigate data sparsity before downstream clustering. Similarly, SnapHiC (Yu *et al.* 2021) used RWR-based imputation to improve loop detection. However, imputation strategies based on convolution and random walk often rely on local information, which can introduce false positive contacts and lead to over-smoothing, a phenomenon where fine structural details in the imputed matrices are obscured. scHiCcompare (Nguyen *et al.* 2025) proposed a different approach: it incorporated genomic distance decay into its design by using a distance-aware random forest model, grouping chromatin contacts of similar genomic distance for imputation.

Received: 18 March 2025; **Revised:** 10 September 2025; **Accepted:** 9 October 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

Although this helps reduce bias from unrelated contacts, it may still contribute to over-smoothing by borrowing information across similar distances without preserving fine-grained structures. Zhang *et al.* (2022) introduced Higashi, a hypergraph-based deep learning method designed to tackle the sparsity issue in scHi-C data. By conceptualizing scHi-C data as a hypergraph, Higashi effectively harnesses global information across cells by aggregating signals from similar cells to enhance shared chromatin features. While Higashi improves scHi-C data quality, its aggregation strategy may also introduce over-smoothing (as depicted in Fig. 1, available as supplementary data at *Bioinformatics* online). Over-smoothed scHi-C matrices often lose fine structural details, which can lead to false identification of chromatin features. Moreover, over-smoothing can inadvertently reduce cell-to-cell variability.

Alternatively, deep learning-based methods have been developed to enhance Hi-C data. These approaches typically adopt a supervised learning framework, in which a downsampled matrix is enhanced to approximate the original high-coverage Hi-C matrix. Several convolutional neural network (CNN) and generative adversarial network (GAN) models have been developed for enhancing bulk Hi-C data, including HiCPlus (Zhang *et al.* 2018), hicGAN (Liu *et al.* 2019), DeepHiC (Hong *et al.* 2020), and EnHiC (Hu and Ma 2021). While originally designed for bulk Hi-C data, these methods can be adapted for sparse scHi-C data. However, such adaptation often results in over-smoothed scHi-C matrices (see Fig. 1, available as supplementary data at *Bioinformatics* online). More recently, ScHiCEDRN (Wang *et al.* 2023) has been developed to enhance scHi-C data by treating it as a one-channel image and using a GAN framework inspired by super-resolution imaging techniques. While promising, this approach occasionally produces image artifacts and tends to predict an excessive number of false positive contacts (see Fig. 1, available as supplementary data at *Bioinformatics* online).

Traditional CNN-based models, though effective in image applications, face unique challenges when applied to inherently sparse scHi-C data. Unlike natural images, scHi-C data are represented as one-channel symmetric matrices, where structural patterns are strongly influenced by genomic distance rather than spatial continuity. This contrasts with the smooth intensity gradients typically observed in images. In this context, standard CNN parameter sharing may fail to capture key features. Moreover, designing convolutional kernels that effectively extract both fine-grained signals and large-scale structural patterns is particularly challenging in the presence of high data sparsity.

To overcome these limitations, we developed SHICEDO, a novel GAN-based deep-learning approach specifically designed for scHi-C data enhancement. SHICEDO incorporates specialized scHi-C feature extraction and channel-wise attention mechanisms, building on our previous work EnHiC (Hu and Ma 2021). It substantially improves the effective coverage and quality of scHi-C data while preserving its unique structural characteristics. We conducted a comprehensive evaluation to assess the accuracy and quality of SHICEDO-enhanced scHi-C data. Our analyses include both pixel-wise and Hi-C-specific similarity measures, as well as performance in key downstream applications. Results consistently show that SHICEDO effectively enhances sparse scHi-C data while maintaining its structural integrity. The capability and effectiveness of SHICEDO are further demonstrated across downstream analyses, including A/B compartment identification, TAD-like domain detection, and single-cell loop detection. Thus, SHICEDO provides an effective and robust solution to scHi-C data sparsity, enabling in-depth exploration of chromatin organization at the single-cell level.

2 Materials and methods

2.1 Overview of the SHICEDO model architecture

In this work, we introduce SHICEDO, a novel deep-learning AQ5 model specifically designed to enhance the quality of scHi-C data while mitigating the over-smoothing issue. Built on a GAN framework, SHICEDO's generator processes sparse or downsampled scHi-C input of different sizes or scales, producing enhanced contact maps as output. Leveraging our previous work on bulk Hi-C enhancement, EnHiC (Hu and Ma 2021), we have incorporated and refined its rank-one feature extraction and reconstruction techniques, along with our newly developed feature refinement modules, into the SHICEDO framework.

As illustrated in Fig. 1, the GAN model in SHICEDO con- AQ6 sists of two key components: the generator (Enhancement Network) and the discriminator (Discriminator Network). The generator transforms sparse, low-coverage scHi-C matrices into enhanced matrices with denser and more complete chromatin contact patterns, while the discriminator distinguishes between enhanced and authentic scHi-C matrices. This adversarial interplay drives the model to continuously improve, yielding high-quality, biologically meaningful enhancements of scHi-C data. Further details on the SHICEDO model are provided in Notes 2.1–2.7, available as supplementary data at *Bioinformatics* online.

SHICEDO's generator is designed with flexibility, supporting multiple input branches that can process submatrices of different sizes or resolution scales (see Note 2.6, available as supplementary data at *Bioinformatics* online). This design allows users to select multi-size input for broader contextual information or multi-scale input for capturing both fine-grained and large-scale structural features. Guided by our ablation study (see Note 2.8, available as supplementary data at *Bioinformatics* online), we adopted a dual-branch configuration for the datasets analyzed in this study. In this setup, the two branches process sparse, low-coverage submatrices of different sizes, enabling the extraction of both central and boundary features. This dual-branch strategy also effectively reduces edge artifacts when merging submatrices back into complete scHi-C matrices.

2.2 Key design strategies of SHICEDO

The SHICEDO model architecture incorporates two specialized design components, each tailored to address the unique challenges of scHi-C data enhancement. The first, rank-one feature extraction, improves the capture of structural information from sparse scHi-C matrices. The second, channel-wise feature selection, mitigates the over-smoothing problem that commonly arises in existing Hi-C enhancement methods.

2.2.1 Rank-one feature extraction

An integral part of SHICEDO's feature selection strategy is the adaptation of the rank-one feature extraction concept, first introduced in our earlier bulk Hi-C enhancement work, EnHiC (Hu and Ma 2021). In SHICEDO, this module has

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610

NOT FOR PUBLIC RELEASE

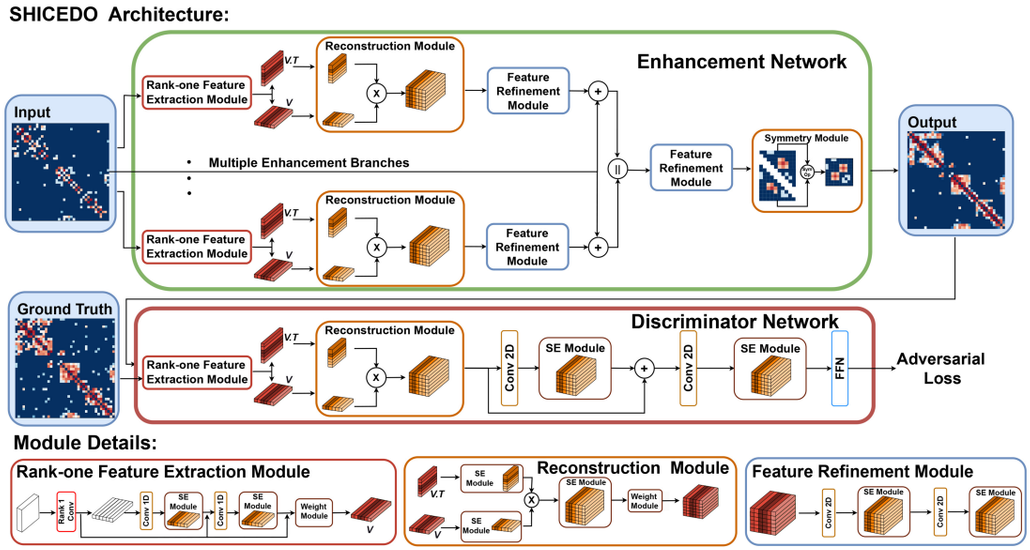


Figure 1. Schematic illustration of the SHICEDO model for scHi-C data enhancement. The model architecture consists of a generator (Enhancement Network) and a discriminator (Discriminator Network) built on a GAN framework. Symbols and abbreviations are explained in Note 2.1, available as supplementary data at *Bioinformatics* online.

been refined to better capture key structural features in sparse scHi-C data.

For an input matrix of size $n \times n$, the rank-one module applies a convolution kernel of size $n \times 1$. This design is well-suited to sparse, symmetric scHi-C matrices, where conventional square kernels may fail to capture informative patterns. The rank-one kernel promotes parameter sharing across genomic bins, achieving more efficient feature extraction. Because sparse scHi-C matrices exhibit heterogeneous patterns within local regions, this approach helps preserve meaningful structural details that might otherwise be overlooked by traditional 2D convolutional layers.

Compared to our previous bulk Hi-C implementation, this refined rank-one module integrates two additional 1D convolutional layers for more precise feature extraction. Additionally, we incorporated a squeeze-and-excitation (SE) module (Hu *et al.* 2018) and skip connections to improve information flow and mitigate information loss. Furthermore, the reconstruction module derives feature tensors through the outer product of rank-one tensors, ensuring that the enhanced matrix preserves the intrinsic symmetry characteristic of Hi-C matrices.

Further details on the design and implementation of rank-one feature extraction are provided in Notes 2.3 and 2.4, available as supplementary data at *Bioinformatics* online.

2.2.2 Channel-wise feature selection

Another key component of SHICEDO is the channel-wise attention mechanism (Hu *et al.* 2018), specifically designed to enhance feature selection across multiple channels of the feature tensor. Although each scHi-C submatrix enters the network as a single-channel contact map, subsequent convolutions expand this into many feature channels. Not all channels are equally informative: some capture biologically meaningful structures, while others encode noise.

Channel-wise attention adaptively weighs these channels, amplifying useful features while suppressing uninformative ones. This is especially important for sparse scHi-C data, where conventional CNN channel summation often produces over-smoothed results that obscure fine structural details. Over-smoothing arises not from failure to recognize structures but from indiscriminate mixing of pixels, which blurs local interactions and reduces cell-to-cell variability.

To address this, SHICEDO integrates SE modules throughout the architecture. Each SE module applies pooling and a small two-layer perceptron to generate learnable channel weights, dynamically adjusting feature contributions. By focusing on biologically relevant patterns and filtering noise, SHICEDO preserves fine structural details, improves enhancement quality, and mitigates over-smoothing.

2.3 scHi-C datasets and preprocessing

We applied SHICEDO to four publicly available scHi-C datasets: the human brain prefrontal cortex dataset (Lee *et al.* 2019), the developing mouse embryo dataset (Liu *et al.* 2023), the mouse embryonic stem (ES) cell dataset (Nagano *et al.* 2017), and the developing mouse brain Dip-C dataset (Tan *et al.* 2021) (Table 1, available as supplementary data at *Bioinformatics* online). After preprocessing, we generated scHi-C matrices at three resolutions (1 Mb, 100 kb, and 50 kb) with five downsampling ratios (2×, 4×, 9×, 16×, and 36×).

Matrices at suitable resolutions and chromatin contacts within predefined genomic-distance thresholds were used for downstream analyses, including A/B compartment identification, TAD-like domain detection, and chromatin loop calling (Table 2, available as supplementary data at *Bioinformatics* online). Additional details on data preprocessing steps are described in Note 2.9, available as supplementary data at *Bioinformatics* online, and training and prediction configurations are provided in Note 2.10, available as supplementary data at *Bioinformatics* online.

3 Results

3.1 SHICEDO faithfully enhances scHi-C matrices

To evaluate the performance of SHICEDO in scHi-C data enhancement, we conducted a comprehensive assessment using

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610

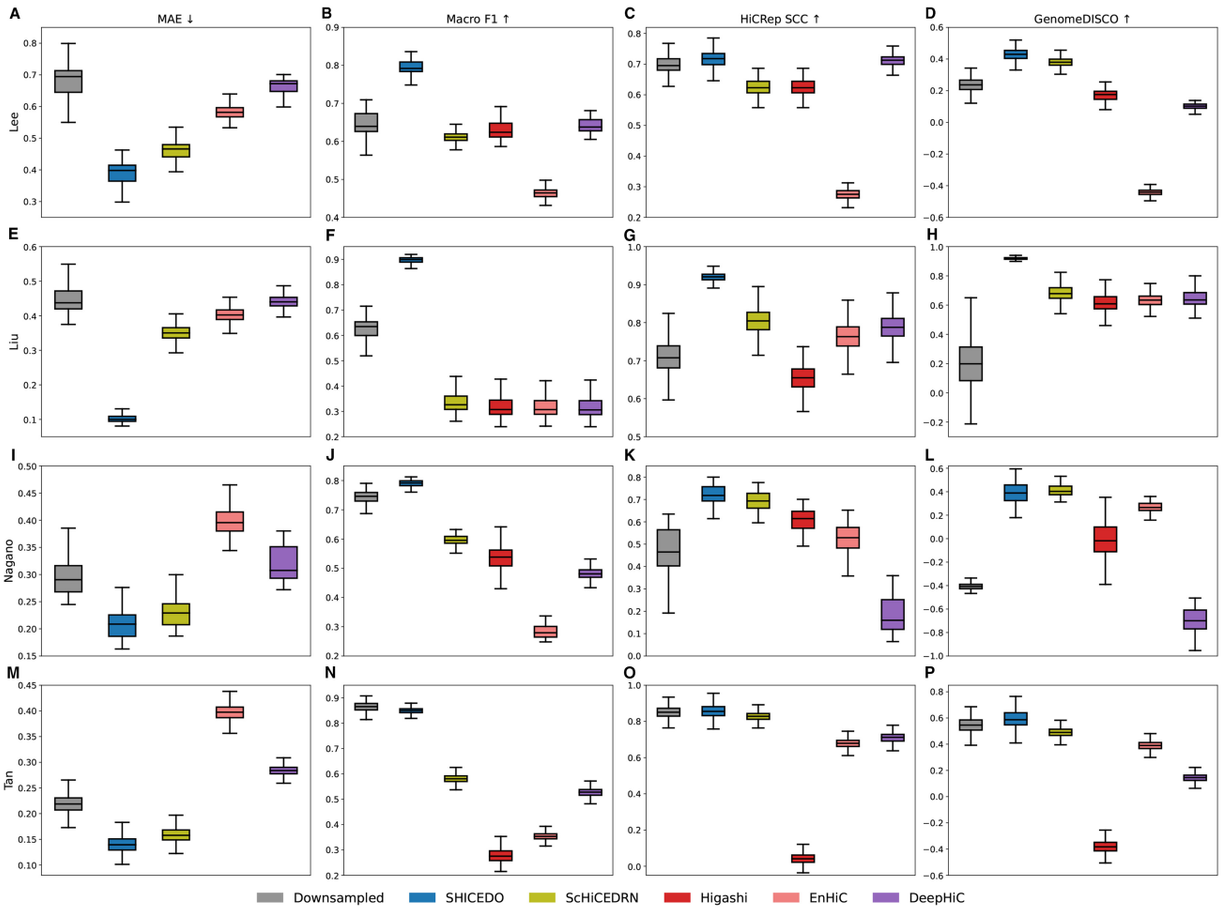
NOT FOR PUBLIC RELEASE

both pixel-wise metrics and Hi-C-specific similarity measures. We compared SHICEDO against two state-of-the-art scHi-C enhancement methods, ScHiCEDRN (Wang *et al.* 2023), and Higashi (Zhang *et al.* 2022), as well as two bulk Hi-C enhancement models, EnHiC (Hu and Ma 2021), and DeepHiC (Hong *et al.* 2020). This evaluation aimed to assess SHICEDO's ability to accurately enhance sparse scHi-C matrices while preserving their structural properties.

For pixel-wise evaluations, we used two key metrics: mean absolute error (MAE) and macro F1 score. MAE quantifies pixel-level accuracy, with lower values indicating closer agreement with the ground truth, the original high-coverage matrices. The macro F1 score provides a balanced measure of precision and recall, effectively identifying potential over-smoothing effects. To compute it, sparse scHi-C matrices were binarized, where the smallest non-zero value in the ground truth served as the threshold, and all values above this threshold in the enhanced matrix were treated as positives. The macro F1 score ranges from 0 to 1, with higher values indicating better performance.

We first evaluated the scHi-C dataset from human prefrontal cortex cells by Lee *et al.* at 1-Mb resolution using 36× downsampled sparse input matrices. The enhanced matrices were compared against the ground truth (i.e. the original high-coverage matrices). As shown in Fig. 2, SHICEDO consistently outperformed baseline methods in both MAE and macro F1 score (Fig. 2A and B). For MAE, SHICEDO achieved the lowest score of 0.3908, surpassing the best-performing baseline method, ScHiCEDRN (P-value $= 2.5 \times 10^{-100}$, t-test). For macro F1 score, SHICEDO obtained the highest score of 0.7951, significantly outperforming the best-performing baseline, DeepHiC (P-value $= 2.1 \times 10^{-247}$, t-test). Note that the high macro F1 score in downsampled sparse input is largely due to the inherent sparse nature of scHi-C matrices, where most pixels are zeros. Downsampling reduces the values of high-count pixels but rarely sets them to zero, which increases precision and leads to an inflated macro F1 score. In contrast, other baseline methods tended to over-smooth the matrices, introducing false-positive pixels that lowered precision and, consequently, macro F1 scores. Therefore, the macro F1 score serves as a reliable indicator of over-smoothing, and our results confirm that SHICEDO effectively mitigates this issue.

In addition to pixel-wise evaluations, we assessed performance using two Hi-C-specific similarity measures: HiCRep (Yang *et al.* 2017, Lin *et al.* 2021) and GenomeDISCO (Ursu



AQ16 **Figure 2.** Pixel-wise and Hi-C-specific similarity evaluation of enhanced scHi-C data. Evaluation results are presented for four scHi-C datasets at 1-Mb resolution: Lee *et al.* (36× downsampling), Liu *et al.* (16× downsampling), Nagano *et al.* (9× downsampling), and Tan *et al.* (4× downsampling). For each dataset, two pixel-wise metrics, mean absolute error (MAE) and macro F1 score, were computed by comparing downsampled sparse input and enhanced scHi-C submatrices against the ground truth. Note that MAE values for Higashi are not plotted, as the method does not specifically optimize for pixel-wise predictions. Additionally, two Hi-C-specific similarity measures, stratum-adjusted correlation coefficients (SCC) from HiCRep and GenomeDISCO scores, were calculated using chromosome-wide matrices.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

et al. 2018). These metrics further validated SHICEDO's effectiveness in preserving structural features in scHi-C data. HiCRep accounts for Hi-C-specific biases and noise by applying smoothing and stratifying contact counts based on genomic distance. The resulting stratum-adjusted correlation coefficient (SCC) ranges from −1 to 1, with 1 indicating identical Hi-C matrices. On the other hand, GenomeDISCO evaluates Hi-C matrix similarity by modeling it as a graph, producing a similarity score also ranging from −1 to 1, with 1 reflecting perfect agreement.

Figure 2C and D presents the average HiCRep SCC and GenomeDISCO values of merged intra-chromosomal matrices across individual cells in the Lee *et al.* dataset. Notably, SHICEDO achieved the highest HiCRep SCC value of 0.7175, outperforming DeepHiC (P-value $= 3.5 \times 10^{-3}$, t-test). Additionally, SHICEDO obtained the highest GenomeDISCO scores of 0.4245, significantly surpassing ScHiCEDRN (P-value $= 5.9 \times 10^{-42}$, t-test).

Furthermore, we evaluated SHICEDO on three additional scHi-C datasets: developing mouse embryonic cells from Liu *et al.* (1-Mb resolution, 16× downsampling), mouse ES cells by Nagano *et al.* (1-Mb resolution, 9× downsampling), and developing mouse brain cells by Tan *et al.* (1-Mb resolution, 4× downsampling). These datasets span different scHi-C protocols, cell numbers, and sequencing depths. Our results demonstrated that SHICEDO consistently outperformed baseline methods across all datasets in terms of MAE, macro F1 scores, and HiCRep SCC values. Additionally, SHICEDO achieved the highest GenomeDISCO scores on the Liu *et al.* and Tan *et al.* datasets, and performed comparably to ScHiCEDRN in GenomeDISCO score on the Nagano *et al.* dataset (Fig. 2). Collectively, these results confirm that SHICEDO robustly enhances scHi-C data quality while mitigating over-smoothing.

To contextualize SHICEDO's performance, we further compared it against several lightweight imputation methods, which are computationally efficient but less sophisticated. Unsurprisingly, SHICEDO consistently achieved the best performance across all evaluation metrics, whereas the lightweight baselines showed mixed or degraded performance depending on sparsity (see Note 2.11, available as supplementary data at *Bioinformatics* online). Taken together, our results demonstrate that SHICEDO not only restores fine-scale chromatin contacts but also preserves global chromatin organization more effectively than both state-of-the-art deep learning models and classical lightweight methods.

3.2 SHICEDO effectively enhances scHi-C data across different cell types and species

The Lee *et al.* dataset comprises 14 distinct cell types from the human prefrontal cortex. To evaluate SHICEDO's adaptability to new data, we trained the model on six selected cell types: ODC, Astro, OPC, Sst, Endo, and Vip, and validated on MP and L6 cells. Testing was subsequently conducted on the remaining six cell types: MG, Ndnf, Pvalb, L4, L5, and L23. The training, validation, and test set consisted of approximately 75%, 5%, and 20% of the cells, respectively.

Following enhancement with SHICEDO, the model achieved an MAE value of 0.3074, a macro F1 score of 0.8357, a HiCRep SCC of 0.7829, and a GenomeDISCO score of 0.4856 on the test data. For comparison, when trained and tested on the same Lee *et al.* dataset without distinguishing cells by their types, under identical conditions (i.e. 1-Mb resolution and a 16× downsampling ratio.), SHICEDO yielded an MAE value of 0.3164, a macro F1 score of 0.8304, a HiCRep SCC of 0.8194, and a GenomeDISCO score of 0.5305. While MAE and macro F1 scores showed slight improvement, HiCRep SCC and GenomeDISCO scores exhibit minor decreases. Overall, SHICEDO's performance remained within the same range across these two experimental settings, indicating its robustness in handling diverse cell types without requiring specialized retraining.

To further evaluate the generalizability of SHICEDO across species, we trained the model on scHi-C data from human brain prefrontal cortex cells by Lee *et al.* at 1-Mb resolution with a 16× downsampling ratio, and tested it on mouse ES cells from Nagano *et al.* at 1-Mb resolution with 9× downsampling ratio.

On the mouse ES cell dataset, SHICEDO achieved an MAE of 0.2265, representing a 25.02% reduction compared to downsampled sparse input. It also yielded a macro F1 score of 0.8004, reflecting an 8.85% improvement, a HiCRep SCC of 0.6632, marking a 26.28% increase, and a GenomeDISCO score of 0.2686, showing a 166.45% improvement over the downsampled sparse input data. In comparison, when trained and tested on the same Nagano *et al.* dataset under identical conditions (i.e. 1-Mb resolution, 9× downsampling), SHICEDO achieved an MAE of 0.2110, a macro F1 score of 0.7854, a HiCRep SCC score of 0.7137, and a GenomeDISCO score of 0.3953. Notably, when trained and tested on different species, SHICEDO exhibited minor decreases in MAE, HiCRep SCC, and GenomeDISCO scores but showed an increase in macro F1 score.

These results demonstrate that SHICEDO maintains strong performance across different cell types and exhibits only a slight performance drop across species. This indicates that the model effectively learns transferable chromatin structure features. Such generalizability is crucial, as it allows SHICEDO to enhance scHi-C data quality across diverse cell types and species without requiring extensive retraining.

3.3 SHICEDO facilitates robust identification of single-cell A/B compartments

In this section, we evaluate SHICEDO's ability to enhance scHi-C data and improve the detection of single-cell A/B compartments at 1-Mb resolution. We applied SHICEDO and baseline methods to enhance the 16× downsampled scHi-C matrices from the Lee *et al.* dataset, followed by A/B compartment analysis on each single cell, adopting Higashi's algorithm (Zhang *et al.* 2022) for identifying single-cell A/B compartments.

To assess the effectiveness of the enhancement methods, we calculated the precision, recall, and macro F1 scores of single-cell A/B compartment assignments obtained from downsampled sparse input and enhanced scHi-C matrices, using the ground truth as reference. A prediction was considered correct if the assigned compartment (A or B) matched the true label at a genomic bin. Given the inherent sparsity of scHi-C data, we incorporated a flexible bin-shift threshold, allowing small positional offsets to still be treated as correct matches.

As shown in Fig. 3, Fig. 2, available as supplementary data at *Bioinformatics* online, and Table 3, available as supplementary data at *Bioinformatics* online, SHICEDO-enhanced scHi-C data consistently demonstrated highly accurate single-

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

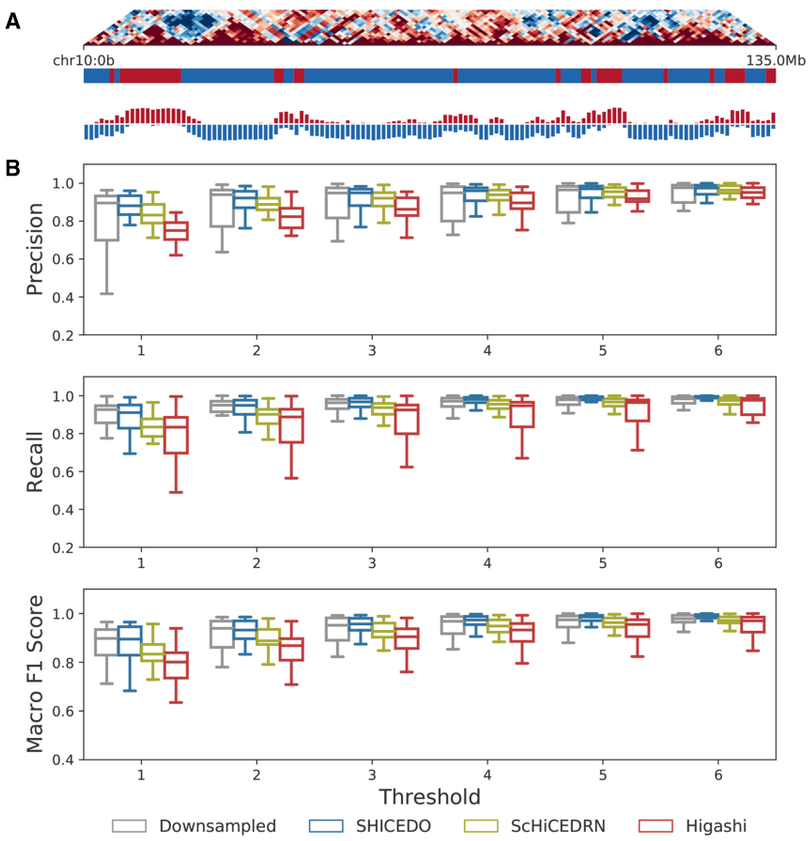


Figure 3. Identification of A/B compartments in enhanced scHi-C data using the Lee *et al.* dataset. (A) Pseudo-bulk Hi-C heatmap, as well as pseudo-bulk and single-cell A/B compartment profiles, are displayed for SHICEDO-enhanced scHi-C matrices on chromosome 10. The pseudo-bulk Hi-C heatmap illustrates the correlation of interaction profiles between pairs of 1-Mb genomic bins; the horizontal barcode represents A/B compartments determined by pseudo-bulk Hi-C, colored in blue and red; the vertical barplot displays the number of cells assigned to A or B compartments at each corresponding genomic bin. (B) Precision, recall, and macro F1 score evaluations for single-cell A/B compartment calls on downsampled sparse input and enhanced scHi-C matrices, compared to the ground truth. Each genomic bin is treated as an independent observation. Predictions and true labels are considered matching if their genomic distance difference falls within a specified threshold (ranging from 1 to 6 bins). Results are based on 1-Mb resolution data with AQ14 36× downsampling.

cell A/B compartment identification, evident in both visual representation and numerical evaluations. Figure 3A highlights the strong concordance between single-cell compartment labels and those derived from pseudo-bulk data aggregated across all cells, confirming the reliability of the single-cell A/B compartment calling. When comparing A/B compartments identified from downsampled sparse input with the ground truth, we observed that the identification of higher-order genome structures, such as A/B compartments, remains relatively robust despite data sparsity (Fig. 2, available as supplementary data at *Bioinformatics* online). Yet, SHICEDO-enhanced scHi-C matrices outperformed all tested methods, particularly at relaxed thresholds, as indicated by comparable or higher precision, recall and macro F1 scores, along with reduced cell-to-cell variations (Fig. 3B and Table 3, available as supplementary data at *Bioinformatics* online). In contrast, ScHiCEDRN and Higashi exhibited a tendency to over-smooth the data, resulting in decreased performance.

These results confirm SHICEDO's ability to accurately preserve active and inactive states of genomic bin states across various cells while effectively enhancing sparse scHi-C data into more detailed and biologically meaningful representations.

3.4 SHICEDO enhances accurate detection of TAD-like domain boundaries

To further assess SHICEDO's capability to promote accurate detection of TAD-like domain boundaries in scHi-C data, we analyzed scHi-C matrices from the Lee *et al.* dataset at 100-kb resolution with 9× downsampling. After enhancement, we identified TAD-like domains using an insulation score based procedure used by Higashi (Zhang *et al.* 2022). Consistent with previous approaches such as EnHiC and ScHiCEDRN, our TAD-like domain detection focused on chromatin contacts within a 2-Mb genomic distance.

Figure 4A and Fig. 3, available as supplementary data at *Bioinformatics* online illustrate single-cell TAD-like boundaries detected from the ground truth, downsampled sparse input, and enhanced scHi-C matrices. The substantial variability of single-cell domain boundaries across cells highlights the high heterogeneity of single-cell TAD structures. Given this variability, we applied a flexible threshold allowing small bin shifts to determine matching boundaries between enhanced matrices and the ground truth. Figure 4B and Table 4, available as supplementary data at *Bioinformatics* online present the precision, recall, and macro F1 score under this relaxed criterion.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610

NOT FOR PUBLIC RELEASE

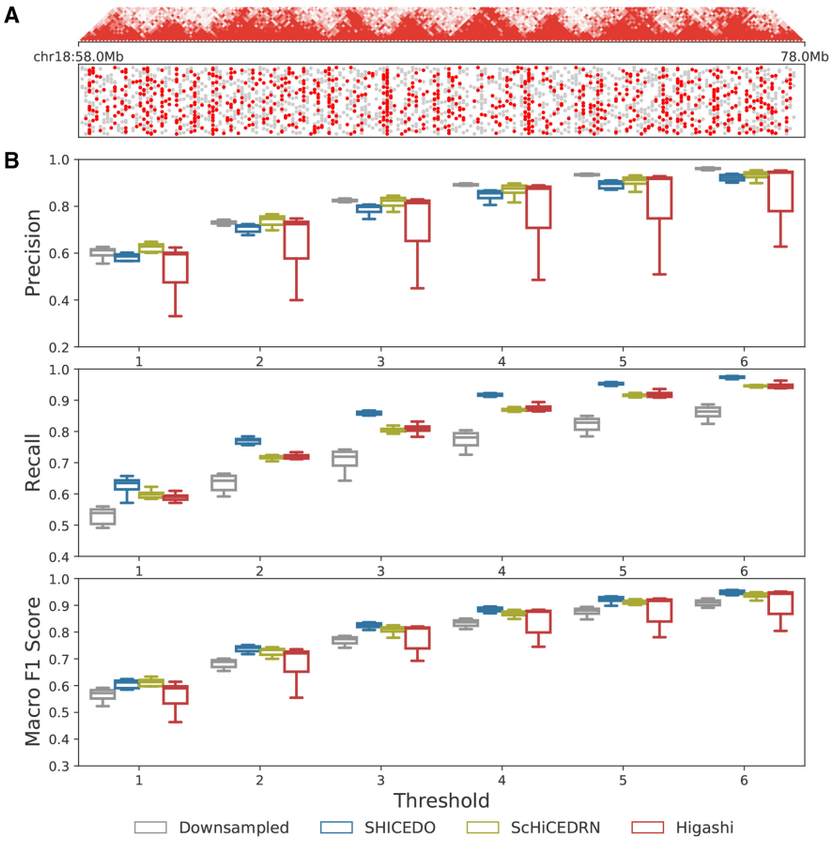


Figure 4. Detection of TAD-like domains in enhanced scHi-C data from the Lee *et al.* dataset. (A) Snapshot of single-cell TAD-like domain boundaries detected based on SHICEDO-enhanced scHi-C matrices within chr18:58–78 Mb is shown in dot plots, one line per cell. TAD structures on the pseudo-bulk heatmap are displayed at the top as a reference. Red dots indicate the boundaries that overlap with the ground truth; while non-overlapping boundaries are shown in gray. (B) Precision, recall, and macro F1 scores of TAD-like domain boundaries detected from downsampled sparse input and enhanced scHi-C data, compared to the ground truth. Two boundaries are considered overlapping if their bin distance falls within the specified threshold, ranging from 1 to 6 bins. Results are based on 100-kb resolution data with 9× downsampling.

Compared to downsampled sparse input and baseline methods, SHICEDO achieved a distinct improvement in recall. Since domain boundary bins are significantly outnumbered by non-boundary bins, the dataset is highly imbalanced, explaining the higher precision observed in downsampled data. However, even against baseline methods, SHICEDO maintained comparable precision while consistently achieving the highest macro F1 scores across all thresholds, demonstrating its superior overall performance. Furthermore, comparing results across multiple thresholds allows us to assess how well SHICEDO preserves structural details as noisy data points are removed. As the threshold value increased, evaluation scores of baseline methods approached those of SHICEDO, highlighting the impact of over-smoothing in these baseline methods.

In summary, these results provide compelling evidence that SHICEDO promotes reliable and robust detection of TAD-like domains in single cells.

3.5 SHICEDO-enhanced scHi-C matrices facilitate fine-scale loop detection

To evaluate SHICEDO's ability to preserve fine-scale structural features, we assessed its performance in chromatin loop detection using the Lee *et al.* dataset. The 50-kb resolution scHi-C matrices were downsampled by a factor of 2, and chromatin loops were identified in the ground truth, downsampled sparse input, and SHICEDO-enhanced matrices using the state-of-the-art loop caller SnapHiC (Yu *et al.* 2021). Following SnapHiC's recommendations, we restricted the analysis to chromatin contacts within a genomic distance of 1 Mb.

We conducted four sets of experiments to detect chromatin loops within genomic distances of 250 kb, 500 kb, 750 kb, and 1 Mb using SnapHiC. SHICEDO-enhanced scHi-C matrices produced fewer detected chromatin loops overall (Fig. 5A), but the majority of them (e.g. 2750 out of 4744 loops within a 250-kb genomic distance) perfectly overlapped with the ground truth, without a single pixel shift. In contrast, baseline methods ScHiCEDRN and Higashi detected substantially more chromatin loops than the number of true loops across all genomic distance ranges, primarily due to over-smoothing (Fig. 5A). This discrepancy was especially pronounced at the 1-Mb genomic distance range, where the number of loops detected from Higashi-enhanced scHi-C data exceeded the true number by more than 4-fold. Such over-smoothing generated an excessive number of false-positive pixels, which were incorrectly identified as chromatin loops, reducing the reliability of single-cell loop detection.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

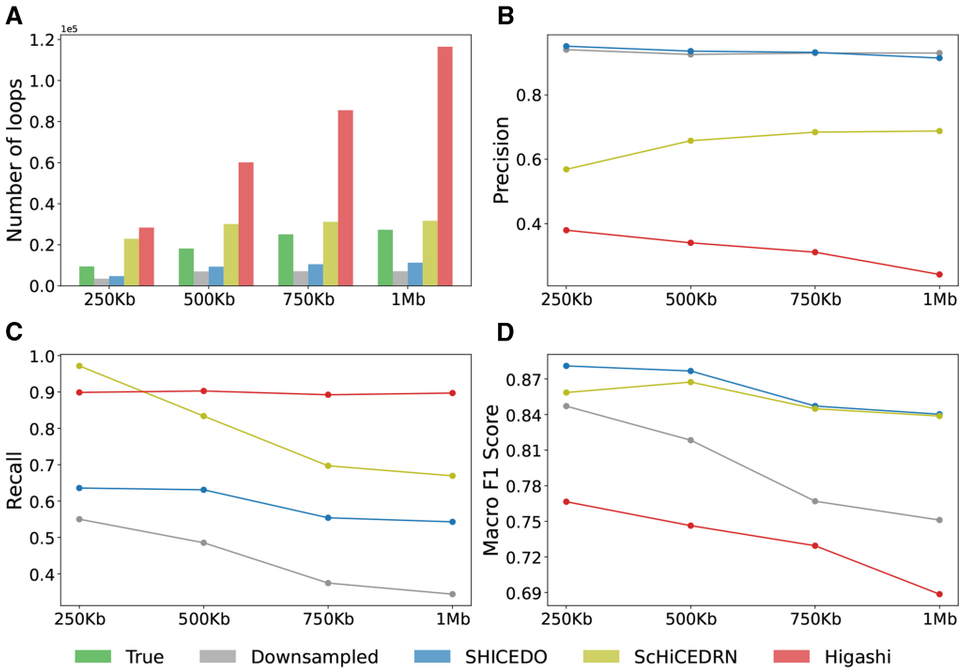


Figure 5. Chromatin loops detected in enhanced scHi-C data from the Lee *et al.* dataset. (A) Total number of detected loops across different genomic distance ranges. (B) Precision, (C) recall, and (D) macro F1 score of loops detected from downsampled sparse input and enhanced scHi-C matrices, compared to the ground truth data. The *X*-axis represents the maximum genomic distance considered for loop detection. A single-pixel shift was allowed while categorizing identical loops. Results are based on 50-kb resolution data with 2× downsampling.

Visual examples of detected chromatin loops in Fig. 4, available as supplementary data at *Bioinformatics* online illustrate the impact of over-smoothed scHi-C data on loop detection.

To quantify loop detection performance, we computed precision, recall, and macro F1 scores. As shown in Fig. 5B, SHICEDO achieved consistently high precision (∼0.93) across all four genomic distance ranges, while capturing more loops as the genomic distance threshold increases. Figure 5C shows that ScHiCEDRN and Higashi exhibited high recall due to excessive loop predictions. Lastly, as reflected in Fig. 5D, SHICEDO attained the highest macro F1 scores across all thresholds, particularly for loops detected at shorter genomic distances, demonstrating its balanced performance of precision and recall.

In summary, SHICEDO effectively mitigates over-smoothing, leading to more reliable loop detection and facilitating the identification of fine-scale chromatin features in scHi-C data.

4 Conclusions and discussion

In this study, we introduced SHICEDO, a GAN-based deep learning approach specifically designed to enhance scHi-C data. SHICEDO improves the effective quality and coverage of scHi-C matrices while preserving structural features and mitigating over-smoothing. Building upon our previous work EnHiC (Hu and Ma 2021), the model integrated refined rank-one feature extraction and reconstruction modules, and further optimizes channel-wise information extraction through Squeeze-and-Excitation Networks (Hu *et al.* 2018).

We rigorously evaluated SHICEDO's performance on publicly available scHi-C datasets generated from multiple experimental protocols and technologies, spanning different resolutions and sparsity levels to demonstrate its robustness. To comprehensively assess its ability to preserve structural information, we performed a series of multi-scale downstream analyses, including A/B compartment analysis, TAD-like domain detection, and loop detection. Compared with state-of-the-art methods, SHICEDO consistently outperformed existing models, demonstrating its capability to enhance scHi-C matrices while preserving cell-to-cell structural variability and mitigating over-smoothing.

Potential directions for future research include leveraging multi-resolution scHi-C data to achieve a more comprehensive understanding of both local and global chromatin structures, thereby improving prediction accuracy. The main challenge lies in designing model architectures that can effectively handle such complexity. Our current model framework already supports multi-scale input, laying the groundwork for future explorations. Another promising strategy is to directly incorporate global Hi-C information rather than relying solely on local submatrices, which would require innovative computational strategies for efficiently integrating diverse data sources. Furthermore, as scHi-C technologies continue to evolve and datasets expand rapidly in scale and quality, the scalability of current enhancement methods remains an open question. Exploring or improving the scalability of SHICEDO to handle much larger and higher-quality datasets will be critical to ensure its long-term applicability. Together, these future directions hold promise for advancing scHi-C data enhancement methods, enabling further improvements in data quality, recovery of structural details, and broader applicability in 3D genome research.

J_ID: Customer A_ID: BTAF575 Copyedited by: Manuscript Category: Original Paper Cadmus Art: OP-CBIO250610 D

NOT FOR PUBLIC RELEASE

Acknowledgements

The authors would like to thank Guoyao Hao, Huiling Liu, Jinli Zhang, and Yusen Zhang for their valuable discussions.

Author contributions

Jingong Huang (Conceptualization [supporting], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [supporting], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [supporting], Writing—review & editing [supporting]), Rui Ma (Data curation [supporting], Investigation [supporting], Methodology [supporting], Software [supporting], Validation [supporting], Visualization [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Michael Strobel (Formal analysis [supporting], Methodology [supporting], Validation [supporting], Writing—original draft [supporting]), Yangyang Hu (Conceptualization [supporting], Methodology [supporting], Writing—original draft [supporting]), Tiantian Ye (Formal analysis [supporting], Methodology [supporting]), Tao Jiang (Conceptualization [equal], Formal analysis [supporting], Funding acquisition [supporting], Investigation [supporting], Methodology [supporting], Project administration [supporting], Resources [equal], Supervision [equal], Validation [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), and Wenxiu Ma (Conceptualization [equal], Data curation [supporting], Formal analysis [supporting], Funding acquisition [lead], Investigation [supporting], Methodology [supporting], Project administration [supporting], Resources [equal], Supervision [equal], Validation [supporting], Writing—original draft [supporting], Writing—review & editing [lead])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the National Institute of Health [R35GM133678, 1R01NS125018] and the National Science Foundation [DBI-1751317]. Computational experiments were performed using resources provided by the High-Performance Computing Center (HPCC) at UC Riverside, which was funded by grants from the National Science Foundation [MRI-2215705, MRI-1429826] and the National Institute of Health [1S10OD016290-01A1].

Data availability

The SHICEDO source code, together with detailed documentation, is publicly available at https://github.com/wmalab/SHICEDO under the MIT license (DOI: 10.5281/zenodo.17069263).

References

Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.

Duan Z, Andronescu M, Schutz K *et al.* A three-dimensional model of the yeast genome. *Nature* 2010;**465**:363–7.

Hong H, Jiang S, Li H *et al.* DeepHiC: a generative adversarial network for enhancing Hi-C data resolution. *PLoS Comput Biol* 2020;**16**:e1007287.

Hu J *et al.* Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 7132–41.

Hu Y, Ma W. EnHiC: learning fine-resolution Hi-C contact maps using a generative adversarial framework. *Bioinformatics* 2021;**37**:i272–9.

Lee D-S, Luo C, Zhou J *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;**16**:999–1006.

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.

Lin D, Sanders J, Noble WS *et al.* HiCRep.py: fast comparison of Hi-C contact matrices in python. *Bioinformatics* 2021;**37**:2996–7.

Liu Q, Lv H, Jiang R *et al.* hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* 2019;**35**:i99–i107.

Liu Z, Chen Y, Xia Q *et al.* Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. *Science* 2023;**380**:1070–6.

Ma W, Ay F, Lee C *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincrna genes. *Nat Methods* 2015;**12**:71–8.

Misteli T. The self-organizing genome: principles of genome architecture and function. *Cell* 2020;**183**:28–45.

Nagano T, Lubling Y, Stevens TJ *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**:59–64.

Nagano T, Lubling Y, Várnai C *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 2017;**547**:61–7.

Nguyen M, Wall BPG, Harrell JC *et al.* scHiCcompare: an R package for differential analysis of single-cell Hi-C data. *J Mol Biol* 2025;**437**:169155.

Ramani V, Deng X, Qiu R *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* 2017;**14**:263–6.

Rao SSP, Huntley MH, Durand NC *et al.* A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.

Tan L, Ma W, Wu H *et al.* Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* 2021;**184**:741–58.e17.

Ursu O, Boley N, Taranova M *et al.* Genomedisco: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* 2018;**34**:2701–7.

Wang Y, Guo Z, Cheng J *et al.* Single-cell Hi-C data enhancement with deep residual and generative adversarial networks. *Bioinformatics* 2023;**39**.

Yang T, Zhang F, Yardımcı GG *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* 2017;**27**:1939–49.

Yu M, Abnousi A, Zhang Y *et al.* SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data. *Nat Methods* 2021;**18**:1056–9.

Zhang R, Zhou T, Ma J *et al.* Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat Biotechnol* 2022;**40**:254–61.

Zhang Y, An L, Xu J *et al.* Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun* 2018;**9**:750.

Zhou J *et al.* Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. *Proc Natl Acad Sci USA* 2019;**116**:14011–8.