Bioinformatics, 2026, **42**(5), btag284
https://doi.org/10.1093/bioinformatics/btag284
Published: 11 May 2026
Original Paper | Genome analysis

OXFORD

SCUDDO: an unsupervised clustering algorithm for single-cell Hi-C maps using diagonal diffusion operators

Luka Maisuradze[1,*], Mark D. Shattuck[2], Corey S. O'Hern[3,4]

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, United States
[2]Benjamin Levich Institute and Physics Department, The City College of New York, New York, NY 10031, United States
[3]Department of Mechanical Engineering, Yale University, New Haven, CT 06520, United States
[4]Graduate Program in Computational Biology and Biomedical Informatics, Yale University, New Haven, CT 06520, United States
*Corresponding author. Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Ave, New Haven, CT 06520, United States.
E-mail: luka.maisuradze@yale.edu
Associate Editor: Peter Robinson

Abstract

Motivation: **Advances in high-throughput chromatin conformation capture have provided insight into the three-dimensional structure and organization of chromatin. While bulk Hi-C experiments capture spatio-temporally averaged chromatin interactions across millions of cells, single-cell Hi-C experiments report on the chromatin interactions of individual cells. Supervised and unsupervised algorithms have been developed to embed single-cell Hi-C maps and identify different cell types. However, single-cell Hi-C maps are often difficult to cluster due to their high sparsity, with state-of-the-art algorithms achieving a maximum Adjusted Rand Index (ARI) of only $\lesssim$0.4 on several datasets.**

Results: **We introduce a novel unsupervised algorithm, Single-cell Clustering Using Diagonal Diffusion Operators (SCUDDO), to embed and cluster single-cell Hi-C maps. We evaluate SCUDDO on four previously difficult-to-cluster single-cell Hi-C datasets, and show that it can outperform other current algorithms in ARI by $\gtrsim$0.2. Further, SCUDDO outperforms all other tested algorithms even when we restrict the number of intrachromosomal maps for each cell type and when we use only a small fraction of contacts in each Hi-C map. Thus, SCUDDO can capture the underlying latent features of single-cell Hi-C maps and provide accurate labelling of cell types even when cell types are not known *a priori*.**

Availability and implementation: **SCUDDO is freely available at https://www.github.com/lmaisuradze/scuddo as well as https://doi.org/10.6084/m9.figshare.31759915. The tested datasets are publicly available and can be downloaded from the Gene Expression Omnibus.**

1 Introduction

Elucidating the structure and dynamics of chromatin in cell nuclei is essential for understanding numerous cellular processes such as DNA transcription and replication (Misteli 2020). Advances in whole-genome analyses, e.g. chromosome conformation capture techniques such as Hi-C, have provided important insights into long-range chromatin interactions and hierarchical chromatin organization (Lieberman-Aiden et al. 2009). Hi-C experiments provide chromatin contact maps, often represented as a symmetric matrix $\mathcal{A}$, where $\mathcal{A}_{ij}$ is the number of times that loci i and j of chromatin come into close proximity. Bulk Hi-C contact maps provide information on chromatin fragment interactions averaged over millions of cells. In contrast, *single-cell* Hi-C maps give the frequency of chromatin contacts in each individual cell.

It is now well established that chromatin structure and organization can differ significantly across cell populations (Finn et al. 2019, Misteli 2020). Transcription analyses and imaging studies have shown that gene expression profiles and cell morphology can differ even between genetically identical cells (Shah et al. 2018, Finn et al. 2019, Rodriguez et al. 2019). In addition, the size and location of chromatin loops and topologically associating domains (TADs) can vary between the Hi-C maps of individual cells for a given organism (Cattoni et al. 2017, Bintu et al. 2018, Finn and Misteli 2018, Mateo et al. 2019). As a result, the loci that possess high contact frequencies in bulk Hi-C maps can differ from those that are in close spatial proximity in fluorescence *in situ* hybridization (FISH) experiments, in part due to the heterogeneity in chromatin structure across individual cells (Bickmore and van Steensel 2013, Williamson et al. 2014, Giorgetti and Heard 2016, Fudenberg and Imakaev 2017, Shi and Thirumalai 2019). Thus, bulk Hi-C maps

Received: 29 October 2025. **Revised:** 21 April 2026. **Accepted:** 28 April 2026
© The Author(s) 2026. 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.

cannot be used to capture the structure and organization of chromatin in individual cells.

Several single-cell Hi-C technologies have been developed to capture chromatin interactions for large numbers of individual cells (Nagano et al. 2013, 2017, Flyamer et al. 2017, Stevens et al. 2017, Li et al. 2019). Single-cell Hi-C techniques enable studies of genome organization in individual cells, as well as comparisons of chromatin structure and organization across different cell types. Using data from single-cell Hi-C experiments, computational studies have focused on TAD, loop, and compartment identification for individual cells within a population (Flyamer et al. 2017, Stevens et al. 2017). However, despite the rapid advances in genome-wide assays, single-cell Hi-C maps are still sparse, only capturing a fraction of the interactions that are obtained in bulk Hi-C experiments (Galitsyna and Gelfand 2021, Zhen et al. 2022). For example, in Fig. 1, we show a pseudo-bulk Hi-C map for chromosome 5 in mouse oocyte cells (Collombet et al. 2020) and compare it to a single-cell Hi-C map for the same chromosome and cell type. The single-cell Hi-C map shows significant sparsity, with most of the off-diagonal elements having a count of 0, as well as large variability for elements near the diagonal.

While techniques like fluorescence-activated cell sorting can be used to label single cells during chromosome conformation capture methods, these techniques are more expensive, lower throughput, and not as widely available as single-cell Hi-C experiments. Thus, the development of classification algorithms for single-cell Hi-C maps may enable researchers to identify the key chromatin interactions that distinguish different cell types. Algorithms developed for bulk Hi-C analysis, including topologically associating domain callers (Maisuradze et al. 2024), often work with limited efficacy on raw single-cell Hi-C data. Thus, due to their inherent variability and sparsity, specialized algorithms must be developed to identify robust features in single-cell Hi-C maps. In this article, we focus on the specific task of classifying single-cell Hi-C maps based on the cell labels that have been provided by experimental studies. Most algorithms for clustering single-cell Hi-C maps use dimensionality reduction, treating each single-cell Hi-C map as a point in high-dimensional space and then mapping each point to a lower-dimensional space to cluster the data (Liu et al. 2018, Galitsyna and Gelfand 2021, Zhen et al. 2022). Despite the fact that there are more than a dozen algorithms to date for clustering single-cell Hi-C maps, there are many single-cell Hi-C datasets for which these methods achieve a maximum adjusted Rand index ARI$\lesssim$0.4 (Zhen et al. 2022, Zheng et al. 2022). Moreover, there are many cases where one clustering method performs well on one single-cell Hi-C dataset, but then performs poorly on another dataset (Galitsyna and Gelfand 2021, Zhen et al. 2022), suggesting that current methods have trouble identifying features that generalize across multiple single-cell Hi-C datasets for clustering.

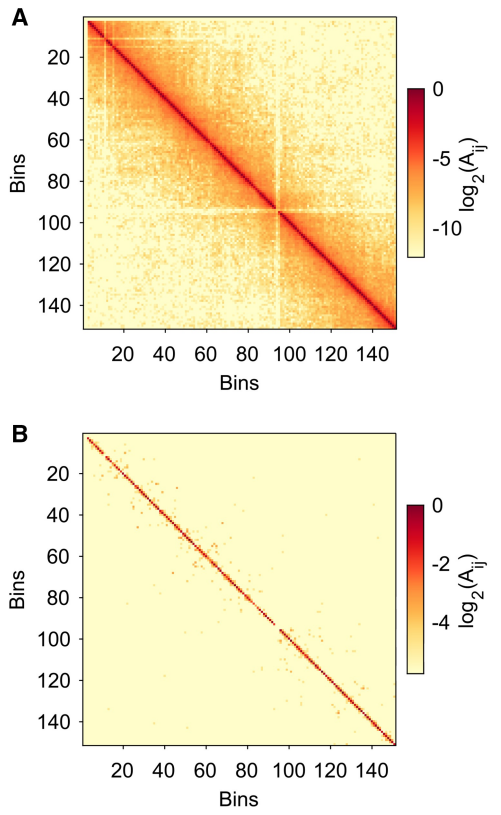

Figure 1 (A) A pseudo-bulk Hi-C map ($\log_2 \mathcal{A}_{ij}$ where zero counts are rendered with the colormap's minimum color) for chromosome 5 using pooled mouse oocyte cells before division from the Collombet dataset (Collombet et al. 2020) (normalized so that $\max(\mathcal{A}_{ij}) = 1$). (B) An example single-cell Hi-C map from the Collombet dataset for chromosome 5 of a mouse oocyte cell using the same normalization.

We develop a novel algorithm, single-cell clustering using diagonal diffusion operators (SCUDDO), which is fully unsupervised, fast, and easy to interpret to separate single-cell Hi-C maps into distinct groups. We then compare the predicted labels of the single-cell Hi-C maps to the cell types that are provided by experimental studies. To quantify the accuracy of the clustering, we calculate the ARI and normalized mutual information (NMI) using the predicted and ground truth labels. We find that SCUDDO outperforms current state-of-the-art methods on four difficult-to-cluster single-cell Hi-C datasets, achieving an ARI and NMI greater than those for all of the tested methods on each of the datasets. We also find that SCUDDO achieves higher accuracy for clustering single-cell Hi-C maps compared to other algorithms when using only a fraction of the number of intrachromosomal maps and a fraction of the superdiagonals in each map.

The remainder of the manuscript is organized as follows. In the Materials and Methods section, we describe the key elements and hyperparameters of SCUDDO for clustering single-cell Hi-C maps. We then describe the four difficult-to-cluster datasets for benchmarking SCUDDO and the two metrics (ARI and NMI) for quantifying the clustering accuracy. In the Results section, we provide the ARI and NMI scores for SCUDDO and five current methods for clustering single-cell Hi-C maps on each of the four difficult-to-cluster Hi-C datasets. We also show SCUDDO's performance across different hyperparameter regimes and when limiting the number of superdiagonals and intrachromosomal maps sampled. In the Discussion, we include some interpretations of the results, our conclusions, and promising future research directions.

2 Materials and methods

The Materials and Methods is organized into three subsections. We first define the necessary notation and summarize the steps

of SCUDDO to cluster single-cell Hi-C maps (see Figure 2 for a summary of SCUDDO algorithm). Second, we describe four difficult-to-cluster single-cell Hi-C datasets that will be used to benchmark SCUDDO alongside five other single-cell Hi-C clustering algorithms. Finally, we define the two metrics, ARI and NMI, which are used to quantify clustering accuracy.

2.1 SCUDDO algorithm

The SCUDDO algorithm takes as input a set of intrachromosomal Hi-C maps with nonnegative integer entries for a cells, each with b chromosomes, totaling $a \times b$ intrachromosomal Hi-C maps. As for bulk Hi-C maps, single-cell Hi-C maps are represented as symmetric matrices with elements $\mathcal{A}_{ij}$ that represent the number of contacts between loci i and j on chromatin. To distinguish between the cell and chromosome indices, we define $\mathcal{A}^k_{s,ij}$ as the ijth element of the $n^k \times n^k$ Hi-C map for chromosome k of cell s. n^k only depends on k since the dimensions of the Hi-C map only vary across different chromosomes. Given a set of intrachromosomal matrices, SCUDDO returns a low-dimensional embedding of the Hi-C matrices. This embedding is then used as an input into a clustering algorithm, e.g. K-means++ (Arthur and Vassilvitskii 2007), where each cell is assigned to one of l predicted labels.

SCUDDO starts by pre-processing and performing imputation on each intrachromosomal matrix $\mathcal{A}^k_s$. First, SCUDDO checks each chromosome's intrachromosomal matrices for any zero (empty) matrices and discards all intrachromosomal matrices for a chromosome if greater than ten percent are empty. Next, SCUDDO calculates a version of the regularized graph Laplacian kernel (Smola and Kondor 2003) for each $\mathcal{A}^k_s$:

$$\bar{\mathcal{A}}^k_s = (2I - D^{-1}(\mathcal{A}^k_s + I))^{-1}, \tag{1}$$

where I is the identity matrix, and D is the degree matrix of $\mathcal{A}^k_s + I$. Then, each $\bar{\mathcal{A}}^k_s$ is symmetrized and then reshaped into a $r \times r$ matrix, $\mathcal{A}'^k_s$, using a bicubic interpolation kernel, where $r = \sum_{k=1}^{b} n^k / b$. SCUDDO then convolves each intrachromosomal matrix with a Gaussian kernel:

$$\mathcal{A}''^k_{s,ij} = \sum_{\chi=1}^{3}\sum_{\omega=1}^{3} \mathcal{G}_{\chi\omega}\mathcal{A}'^k_{s,(i-1+\chi)(j-1+\omega)}, \tag{2}$$

where $\mathcal{G}$ is a two-dimensional 3×3 Gaussian kernel with standard deviation $\sigma = 0.5$ that uses replicate padding, where values outside of the bounds of the original Hi-C map are set to the values of the nearest border entry (for more details see Fig. 1, available as supplementary data at *Bioinformatics* online). $\mathcal{G}$ smooths local regions in each individual intrachromosomal matrix. The final pre-processing step is to normalize each intrachromosomal matrix and apply a matrix exponential:

$$\mathcal{B}^k_s = e^{\left(-\frac{\mathcal{A}''^k_s}{\sum_{ij}\mathcal{A}''^k_{s,ij}}\right)}, \tag{3}$$

where e^A denotes the matrix exponential, the convergent power series defined for any square matrix A as: $e^A = \sum_{k=0}^{\infty} \frac{1}{k!}A^k$. In the context of continuous-time network dynamics, e^{tA} serves as a time-evolution operator, where t dictates the duration and scale of information propagation across the network. While positive values of t model forward diffusion, which smooths the data and spreads the signal across the network, setting $t = -1$ as SCUDDO does yields a dissipative system. We suspect that this inverse operation functions as a spectral high-pass filter where rather than blurring structural features, it penalizes indirect transitive paths and sharpens the local topological boundaries of the contact map. Next, we construct high-dimensional embeddings of the intrachromosomal Hi-C maps for each cell. Let $d^w(\mathcal{B}^k_s)$ be the ordered set of Hi-C map entries on the wth superdiagonal of the $r \times r$ matrix $\mathcal{B}^k_s$:

$$d^w(\mathcal{B}^k_s) = \{\mathcal{B}^k_{s,1,1+w}, \mathcal{B}^k_{s,2,2+w}, \ldots, \mathcal{B}^k_{s,r-w,r}\}. \tag{4}$$

For a given w, $d^w(\mathcal{B}^k_s)$ for each chromosome k for cell s is concatenated to form the embedding vector:

$$\vec{e}_s{}^w = \{d^w(\mathcal{B}^1_s), d^w(\mathcal{B}^2_s), \ldots, d^w(\mathcal{B}^b_s)\}, \tag{5}$$

where $\alpha = 1, \ldots, b(r-w)$ indexes the entries in $\vec{e}_s{}^w$. Every embedding vector, $\vec{e}_s{}^w$, is z-score normalized such that

$$\vec{e}'_s{}^w = \frac{\vec{e}^w_s - \mu}{\sqrt{\frac{1}{b(r-w)-1}\sum_{\alpha=1}^{b(r-w)}\left((\vec{e}^w_s)_\alpha - \mu\right)^2}}, \tag{6}$$

where $\mu = \sum_{\alpha=1}^{b(r-w)}(\vec{e}_s{}^w)_\alpha / [b(r-w)]$. This pooling approach is similar to previous work (Zheng et al. 2022, Fletez-Brant et al. 2024) that employs band normalization for Hi-C matrices. Next, each embedding vector is transformed into a signed difference vector:

$$\vec{f}^w_s = \mathrm{sgn}(\nabla(\vec{e}'_s{}^w)), \tag{7}$$

where $\nabla(\vec{e}'_s{}^w)_\alpha = (\vec{e}'_s{}^w)_\alpha - (\vec{e}'_s{}^w)_{\alpha+1}$ and sgn is the sign function. (Note that we set the last entry of the chromosome difference vector $(\vec{f}_s{}^w)_{b(r-w)} = (\vec{e}'_s{}^w)_{b(r-w)}$.) Eq. 7 transforms $\vec{e}'_s{}^w$ into a ternary vector with values 1, 0, or -1. For a given w, each cell's difference vector, $\vec{f}_1{}^w, \vec{f}_2{}^w, \ldots, \vec{f}_a{}^w$ is used to calculate the distance matrix between cells i and j using cosine similarity:

$$D^w_{ij} = 1 - \frac{\vec{f}_i{}^w \cdot \vec{f}_j{}^w}{|\vec{f}_i{}^w||\vec{f}_j{}^w|}, \tag{8}$$

where $|\vec{X}|$ indicates the magnitude of $\vec{X}$. A separate distance matrix is calculated for $\vec{e}'_s{}^w$:

$$D'^w_{ij} = 1 - \frac{\vec{e}'_i{}^w \cdot \vec{e}'_j{}^w}{|\vec{e}'_i{}^w||\vec{e}'_j{}^w|}, \tag{9}$$

and combined to form a final exponentiated distance matrix using element-wise exponentiation:

$$\mathcal{K}_{ij}^{w} = e^{(D_{ij}^{w} + D_{ij}^{w})}. \quad (10)$$

Finally, SCUDDO uses nonmetric multidimensional scaling (MDS) (Kruskal 1964) to transform the $a \times a$ matrix $\mathcal{K}^w$ into a lower-dimensional representation, i.e. an $a \times p$ matrix where $p < a$, which preserves the distances in $\mathcal{K}^w$. The MDS is followed by principal component analysis to further reduce the dimension to an $a \times q$ matrix $\mathcal{U}^w$, where $q < p$ ($p = 30$ and $q = 5$). This procedure is performed for the diagonal ($w = 0$) and a given number of superdiagonals ($w = \zeta > 0$), and each set of dimensionality-reduced representations are concatenated, forming the $a \times (q(\zeta+1))$ matrix $\mathcal{R} = \mathcal{U}^0, \mathcal{U}^1, \ldots, \mathcal{U}^\zeta$. $\mathcal{R}$ is then normalized feature-wise using the softmax function, $\mathcal{R}'_{ij} = e^{\mathcal{R}_{ij}} / \sum_{\theta=1}^{a} e^{\mathcal{R}_{\theta j}}$, and a distance matrix is constructed using the L^1 metric:

$$\mathcal{S}_{ij} = \sum_{\lambda=1}^{q(\zeta+1)} |\mathcal{R}'_{i\lambda} - \mathcal{R}'_{j\lambda}|. \quad (11)$$

Another round of dimensionality reduction is performed using MDS to reduce the dimension of $\mathcal{S}$ to the embedding size ε, which gives the $a \times \varepsilon$ matrix, $\mathcal{V}$. Because there is no guarantee of convexity associated with each cluster when clustering single-cell Hi-C matrices, we use spectral decomposition before performing the clustering. In particular, SCUDDO transforms $\mathcal{V}$ into the similarity matrix, $\mathfrak{A}_{ij} = e^{-\mathcal{Z}_{ij}^2}$, where $\mathcal{Z}_{ij} = |\vec{\mathcal{V}}_{i*} - \vec{\mathcal{V}}_{j*}|$ and $\mathcal{V}_{i*}$ is the vector consisting of all elements in the ith row of $\mathcal{V}$. Next, we calculate the final $a \times (l-1)$ spectral embedding $\mathcal{C}$, where the columns of $\mathcal{C}$ are the smallest $l-1$ eigenvectors (ignoring the trivial eigenvector) of the random-walk Laplacian matrix constructed from $\mathfrak{A}$ using the Shi-Malik algorithm (Shi and Malik 2000) with $\log(a)$ nearest neighbors. We then input $\mathcal{C}$ and the number of labels l into a clustering algorithm, such as K-means++ (Arthur and Vassilvitskii 2007), which returns the predicted labels for each single-cell Hi-C map.

SCUDDO includes two tunable hyperparameters: ζ, the set of (super)diagonals, $w = 0, 1, \ldots, \zeta$ used to construct the embedding vectors, and the dimension ε, to which $\mathcal{V}$ is reduced. By default, SCUDDO outputs two embeddings: $\mathcal{C}$ and $\mathcal{V}$. Both ζ and ε are varied in Section 3 to study their effects on SCUDDO's performance for each dataset. For all results in this study unless otherwise noted, we use $\mathcal{C}$ as the input into K-means++ and set $\zeta = 25$ and $\varepsilon = 5$. Our results are not sensitive to the values of the dimensions p and q.

2.2 Benchmarking of single-cell Hi-C clustering algorithms

We focus our studies on four difficult-to-cluster datasets of single-cell Hi-C maps from recent benchmarking studies (Zhen et al. 2022, Ma et al. 2024, Plummer et al. 2025). In particular, we consider the Li et al. (2019) dataset (GEO ID: GSE119171) consisting of $a = 150$ mouse embryonic stem cells that are separated into $l = 3$ labels: "2i," "Serum1," and "Serum2," the Flyamer et al. (2017) dataset (GEO ID: GSE80006) consisting of $a = 169$ cells (154 after filtering) from developing mouse zygotes and oocytes with $l = 3$ cell types: "Oocyte," "ZygP," and "ZygM" as labels, the Collombet et al. (2020) dataset (GEO ID: GSE129029) consisting of $a = 468$ mouse embryo cells (466 after filtering) with labels that represent $l = 5$ different cell stages: 1-cell, 2-cell, 4-cell, 8-cell, and 64-cell stages (we also study the full Collombet dataset of $a = 648$ cells), and the Liu et al. (2023) dataset (hereafter denoted as "HiRES"; GEO ID: GSE223917) consisting of $a = 399$ mouse brain cells with $l = 7$ labels: "Astro," "Ex1," "Ex2," "Ex3," "In1," "In2," and "Oligo." In previous benchmarking studies (Zhen et al. 2022), none of the eight tested methods for single-cell Hi-C map clustering achieved ARI or NMI ≥ 0.6 on the Collombet et al. dataset when using all 648 cells, and in another study (Zheng et al. 2022), none of the eight methods tested achieved an ARI > 0.45 on the Li et al. dataset across any clustering algorithm (not just k-means). Similarly, when using a more lenient cell filter for the Flyamer et al. dataset, another benchmarking study (Plummer et al. 2025) found that none of the thirteen tested methods surpassed an average ARI (across several clustering algorithms not just k-means) of 0.6 and for the HiRES dataset none surpassed 0.32. For each dataset in the manuscript, we use 1 Mb bin sizes for the single-cell Hi-C maps, and re-bin those with higher resolution, as discussed in Zhou et al. (2019). For the Flyamer, Collombet, and HiRES datasets, we use the preprocessed single-cell Hi-C files from a previous benchmarking study (Plummer et al. 2025) (GEO ID: GSE305523) when applicable and for the Li dataset we use the dataset provided in another study (Zheng et al. 2022). We use the filtering procedure outlined in Plummer et al. (2025), which is based on the procedure outlined in Zhou et al. (2019). In particular, if the sum of all non-diagonal nonzero pairs of elements in the intrachromosomal Hi-C maps for a given cell is less than 5000, the data for this cell was not included in the analysis. Also, for each individual chromosome of size x for a cell, if the intrachromosomal Hi-C map for that chromosome has a sum of non-diagonal contacts that is less than x, all intrachromosomal Hi-C maps are not considered for that cell.

To select appropriate methods to benchmark for SCUDDO, we surveyed previous single-cell Hi-C clustering studies to identify the most consistently high-performing algorithms (Zheng et al. 2022, Ma et al. 2024, Plummer et al. 2025). In a recent comprehensive evaluation of 13 algorithms (Plummer et al. 2025), Higashi (Zhang et al. 2022), ScHiCluster (Zhou et al. 2019), and SnapATAC2 (Zhang et al. 2024) achieved the highest average ARI scores across eleven datasets at 1 Mb resolution. A separate study evaluating eight methods at 500 kb resolution (Ma et al. 2024) identified scVI-3D (Zheng et al. 2022) and scHiCluster as the top-performing methods on the Nagano et al. dataset, while Higashi, scVI-3D, and scHiCTools (Li et al. 2021) were among the top-performing methods on the Tan et al. dataset. Furthermore, a third study (Zheng et al. 2022) placed scVI-3D, Higashi, and scHiCluster within the top four algorithms (out of eight total) based on the median rank across four datasets. Drawing on these findings, we selected Higashi, scHiCluster, snapATAC2, scVI-3D, and InnerProduct (the highest-performing algorithm within the scHiCTools suite) as the methods for comparison with SCUDDO. These selections provide a well-rounded comparison that encompasses both statistically driven approaches (InnerProduct, scHiCluster) and training-intensive deep learning models (Higashi, scVI-3D). Importantly, all algorithms that we tested are unsupervised or self-supervised, and do not require labels for training. Other algorithms that require labels or significant pretraining are unable to cluster unlabeled single-cell Hi-C datasets, and thus they are not included in this manuscript. We ran each algorithm using its default hyperparameters; if none were specified, we

utilized the parameters from a previous benchmarking study (Plummer et al. 2025). Because several methods are non-deterministic, we generated 10 independent embeddings for each algorithm. These embeddings were fed directly into K-means++ clustering without further processing to calculate mean ARI and NMI scores. Following prior benchmarking studies (Zhang et al. 2024, Plummer et al. 2025), latent embeddings for each method were capped at a maximum of 30 dimensions.

2.3 Metrics for clustering accuracy

To assess the accuracy of the predicted labels, we calculate the ARI (Hubert and Arabie 1985) and NMI (Vinh et al. 2010). Let $\Omega_T(s)$ and $\Omega_G(s)$ be functions that map each cell index s (from 1 to a) to the integers l' and l'' respectively, where l' is the ground truth label for cell s and l'' is the predicted label for cell s. We then define $P_T = \{X_1, X_2, \ldots X_l\}$ as the "ground-truth" label set, where $X_{l'}$ denotes the set of cells such that $\Omega_T(s) = l'$, and $P_G = \{Y_1, Y_2, \ldots Y_l\}$ as the "predicted" label set, where $Y_{l''}$ is the set of cells such that $\Omega_G(s) = l''$.

The ARI determines the similarity between the sets of cells with given ground truth and predicted labels:

$$\text{ARI} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{l}\binom{\beta_{ij}}{2} - \left(\sum_{i=1}^{l}\binom{\Gamma_i}{2}\sum_{j=1}^{l}\binom{\Delta_j}{2}\right)\Big/\binom{a}{2}}{\frac{1}{2}\left(\sum_{i=1}^{l}\binom{\Gamma_i}{2} + \sum_{j=1}^{l}\binom{\Delta_j}{2}\right) - \left(\sum_{i=1}^{l}\binom{\Gamma_i}{2}\sum_{j=1}^{l}\binom{\Delta_j}{2}\right)\Big/\binom{a}{2}}, \quad (12)$$

where $\beta_{ij} = [X_i \cap Y_j]$, $\cap$ is the intersection between two sets, $[X]$ is the number of elements in set X, $\Gamma_k = \sum_{i=1}^{l}\beta_{ki}$, $\Delta_k = \sum_{j=1}^{l}\beta_{jk}$, and $\binom{m}{n} = \frac{m!}{n!(m-n)!}$. ARI $= 1$ indicates a perfect match between P_T and P_G, whereas ARI $= 0$ indicates the match between P_T and P_G is no better than that achieved by random assignments in P_G.

We also quantify the accuracy of the clustering of the single-cell Hi-C maps using the NMI. NMI measures how much information can be learned about a given clustering by observing a different, but related clustering. NMI is defined as:

$$\text{NMI} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{l}\mathcal{H}(i,j)\log_2\frac{\mathcal{H}(i,j)}{\mathcal{H}(i)\mathcal{H}(j)}}{\sqrt{\left(-\sum_{i=1}^{l}\mathcal{H}(i)\log_2\mathcal{H}(i)\right)\left(-\sum_{j=1}^{l}\mathcal{H}(j)\log_2\mathcal{H}(j)\right)}}, \quad (13)$$

where $\mathcal{H}(i) = \frac{[X_i]}{a}$, $\mathcal{H}(j) = \frac{[Y_j]}{a}$, and $\mathcal{H}(i,j) = \frac{[X_i \cap Y_j]}{a}$. $0 \leq \text{NMI} \leq 1$, where NMI $= 1$ indicates that $P_T = P_G$ and NMI $= 0$ indicates that there is no correlation between P_T and P_G. We calculate both ARI and NMI since they can differ for different-sized clusters: ARI is preferable when the sets in P_T are similar in size, whereas NMI is preferable when the sets in P_T are unbalanced. For all datasets and algorithms, we calculate the ARI and NMI after using the native embedding and K-means++ clustering.

3 Results

We carried out single-cell Hi-C map clustering on four difficult-to-cluster datasets from Collombet et al. (2020), Flyamer et al. (2017), and Li et al. (2019), and on the HiRES dataset (Liu et al. 2023), using five current algorithms [Higashi (Zhang et al. 2022), InnerProduct (Li et al. 2021), scHiCluster (Zhou et al. 2019), snapATAC2 (Zhang et al. 2024), and scVI-3D (Zheng et al. 2022)] and compared the results to those obtained from SCUDDO. We

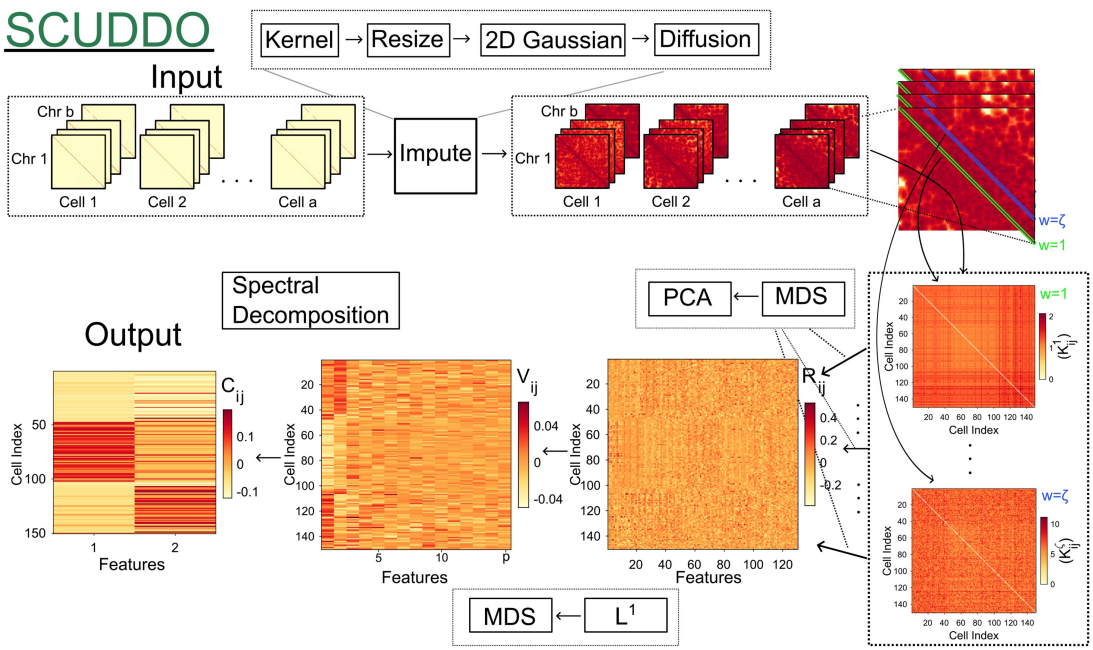

Figure 2 A schematic of the SCUDDO algorithm for clustering single-cell Hi-C maps. SCUDDO first imputes the set of intrachromosomal Hi-C matrices (indexed by $\mathcal{A}_s^k$) for each cell and then samples each superdiagonal (indexed by w) from each Hi-C matrix to form a feature matrix $\mathcal{K}^w$ for each sampled superdiagonal. Principal component analysis (PCA) and nonmetric MDS are then applied to each feature matrix to form the matrix $\mathcal{R}$, which is then embedded in a lower-dimensional latent space using the L^1 norm to form the embedding, $\mathcal{V}$, which is then finally spectrally decomposed into $\mathcal{C}$.

plot ARI versus NMI for each dataset and algorithm in Fig. 3. Overall, SCUDDO outperforms the other five methods for all datasets tested. For the Li et al. dataset, we find a significant separation in accuracy between SCUDDO and the next most accurate method (scHiCluster). SCUDDO achieves an ARI $\sim$ NMI ≈ 0.4, while scHiCluster achieves ARI ≈ 0.31 and NMI ≈ 0.33. For the Flyamer et al. dataset, we find that SCUDDO achieves an ARI $\approx$ 0.75 and an NMI ≈ 0.67, whereas the next most accurate method, snapATAC2, achieves an ARI ≈ 0.43 and NMI ≈ 0.46. Because the implementation of filtering we follow in this manuscript [from Plummer et al. (2025)] is more lenient (filtering on autosomes), we also ran benchmarks with a more restrictive Flyamer et al. dataset using only $a = 134$ cells [from Zhen et al. (2022)] and found that SCUDDO still performed the best among the tested methods, achieving an ARI ≈ 0.74 and NMI ≈ 0.70 (scHiCluster achieves a notable ARI ≈ 0.65 and NMI ≈ 0.63). Interestingly, using a larger kernel size and standard deviation for the Gaussian kernel for SCUDDO allows SCUDDO to achieve ARI/NMI scores greater than 0.9 (see Fig. 1, available as supplementary data at *Bioinformatics* online). For the Collombet et al. dataset, we find that SCUDDO has ARI ≈ 0.85 and NMI ≈ 0.83, whereas the next most accurate methods, snapATAC2 and scHiCluster, achieve ARI $\sim$ NMI < 0.65. Similar to the Flyamer et al. dataset, we also tested SCUDDO on a separate version of the Collombet et al. dataset [from Zhen et al. (2022)], consisting of more cells ($a = 648$) and found again that SCUDDO performs the best among the tested methods scoring an ARI $\sim$ NMI ≈ 0.62 whereas no other method surpassed ARI/NMI scores of 0.5. Lastly, for the HiRES dataset, we again find that SCUDDO outperforms all tested methods, scoring an ARI ≈ 0.28 and NMI ≈ 0.41 whereas the next best methods, InnerProduct and snapATAC2, score ARI ≈ 0.18 and NMI ≈ 0.34. Interestingly, when setting the ζ hyperparameter to lower values (i.e. sampling fewer superdiagonals), SCUDDO achieves ARI scores greater than 0.45, suggesting that higher superdiagonals potentially degrade clustering ability. In addition to preserving major cell-type lineages in the HiRES dataset, we additionally tested whether SCUDDO could simultaneously capture structural features during the cell cycle for the HiRES dataset. To ensure that this separation was an inherent property of SCUDDO's high-dimensional space and not an artifact of a 2D non-linear projection, we performed a k-nearest neighbors (k-NN, $k=5$) analysis directly on the raw, unmanipulated SCUDDO embeddings. We found that a cell's closest high-dimensional structural neighbors were highly predictive of its continuous mitotic

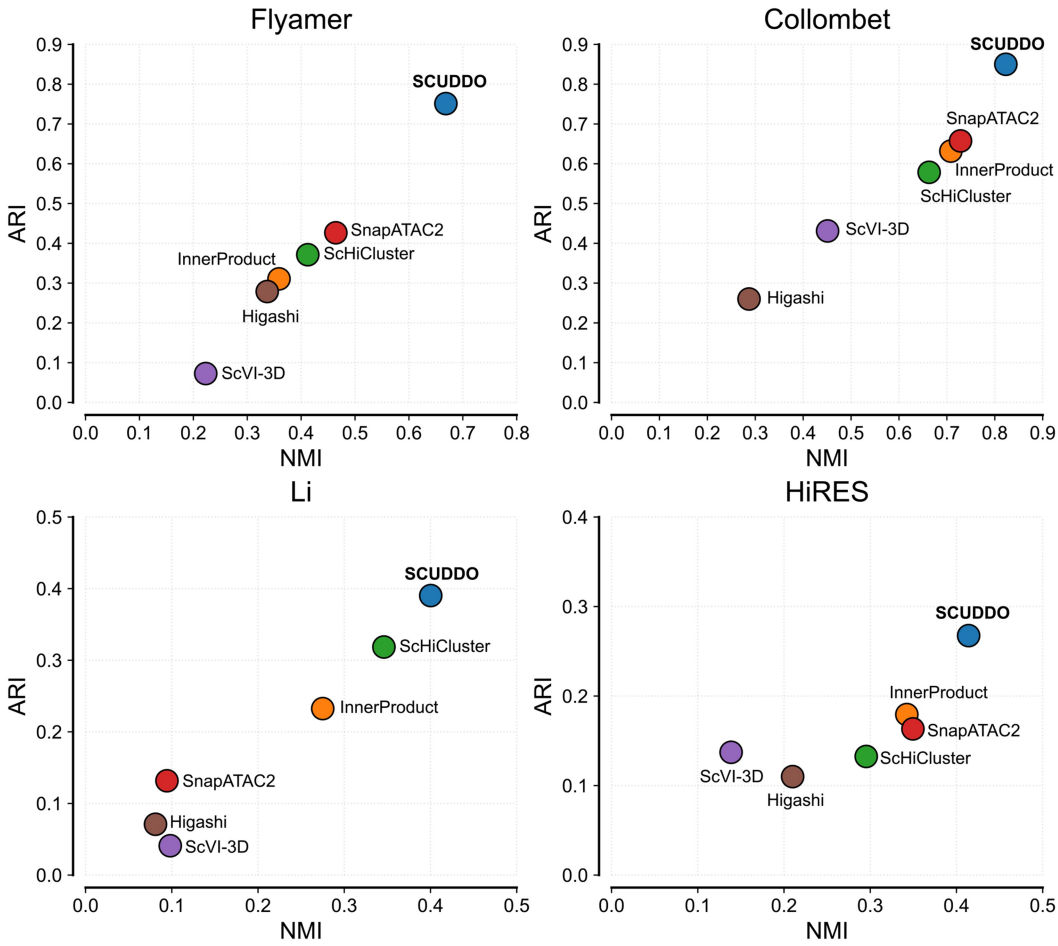

Figure 3 Accuracy of the six single-cell Hi-C map clustering algorithms [Higashi (Zhang et al. 2022) (brown), InnerProduct (Li et al. 2021) (orange), scHiCluster (Zhou et al. 2019) (green), snapATAC2 (Zhang et al. 2024) (red), scVI-3D (Zheng et al. 2022) (purple), and SCUDDO (blue)] on four difficult-to-cluster single-cell Hi-C datasets. We plot the ARI versus the NMI for each algorithm on the (top left) Flyamer et al. (2017), (bottom left) Li et al. (2019), (top right) Collombet et al. (2020), and (bottom right) HiRES (Liu et al. 2023) datasets.

percentage and short-range contact fraction. The mean mitotic score of a cell's local neighborhood strongly and significantly correlated with the cell's actual score (Pearson $R=0.8$) and the same held true for short-range contact fraction (Pearson $R=0.85$). This effect was weaker (Pearson $R=0.34$) when looking at nearby neighbors for similar replication scores, suggesting that SCUDDO is better able to capture some cell cycle features than others.

We next show that SCUDDO can accurately embed single-cell Hi-C maps using a reduced amount of information for already highly sparse single-cell Hi-C maps, surpassing the accuracy of previous algorithms using fewer intrachromosomal matrices for each cell, as well as fewer sampled superdiagonals for each matrix. We study the performance of SCUDDO after restricting the single-cell Hi-C data available to it in two ways: first by varying the hyperparameter ζ for the number of superdiagonals to sample for each intrachromosomal Hi-C map, as well as varying the hyperparameter ε for the embedding dimension of $\mathcal{V}$. In Fig. 4, we show heatmaps of the ARI and NMI for the SCUDDO algorithm, while varying $0 \leq \zeta \leq 40$ and $1 \leq \varepsilon \leq 40$. The outlined pixels in the ζ-ε plane indicate ARI or NMI values for the SCUDDO algorithm that exceed those for all other methods (when they use all of the available single-cell Hi-C data).

For the Li et al. dataset, we show in the leftmost column of Fig. 4A and B that for $\approx 93\%$ and 72% of the ζ-ε plane, SCUDDO outperforms all methods in ARI and NMI, respectively. In particular, across all ζ values, SCUDDO's mean ARI (over all ε) exceeds the next best method. Similarly, when $\varepsilon > 1$, SCUDDO gives mean ARI values over all ζ that exceed the values for all other methods. Even in regimes with low superdiagonal sampling and embedding dimension, SCUDDO can obtain ARI values that surpass the next best method (e.g. $\zeta = 0, \varepsilon = 2$). The maximum ARI and NMI for clustering the Li et al. dataset using SCUDDO in the sampled hyperparameter space were ≈ 0.47 and ≈ 0.44, respectively.

On the Flyamer et al. dataset, SCUDDO outperforms the other methods again over a large region of the hyperparameters ζ and ε, as shown in Fig. 4A and B, with $\approx 80\%$ and $\approx 95\%$ of (ζ, ε) input pairs into SCUDDO resulting in ARI and NMI scores that surpassed the next best method's ARI and NMI scores, respectively. We find that SCUDDO requires generally higher ζ values to perform ideally on the Flyamer et al. dataset, showing a sudden improvement around $\zeta = 20$ for both ARI and NMI. The exception is when $\zeta = 0$, where a strong band exists once $\varepsilon \geq 6$. Across all ζ and ε pairs, SCUDDO achieves a maximum ARI of ≈ 0.89 and NMI of ≈ 0.77, both surprisingly when $\varepsilon = 17$ and $\zeta = 0$.

For the Collombet et al. dataset, SCUDDO outperforms the next best method in ARI and NMI over $\approx 87\%$ and $\approx 90\%$ of the ζ-ε plane. For $\zeta > 5$, the mean ARI and NMI across all ε values for SCUDDO is larger than the best ARI and NMI from all other tested algorithms. Similarly, SCUDDO outperforms all the other methods in mean ARI and NMI when $\varepsilon > 1$ (across all ζ). Across the sampled hyperparameters, we find that the maximum ARI and NMI are ≈ 0.87 and ≈ 0.85, respectively.

On the HiRES dataset, SCUDDO performs ideally when $\zeta < 30$ and $\varepsilon > 20$. Despite this, SCUDDO performs very well over a large region of the hyperparameters ζ and ε, with $\approx 79\%$ and $\approx 87\%$ of (ζ, ε) input pairs into SCUDDO resulting in ARI and NMI scores that surpassed the next best method's ARI and NMI scores, respectively. For $\zeta > 2$, the mean ARI and NMI across all ε values for SCUDDO is larger than the best ARI and NMI from all other tested algorithms. Similarly, SCUDDO outperforms all the other methods in mean ARI and NMI when $\varepsilon > 2$ (across all ζ) for the HiRES dataset. Across all ζ and ε pairs, SCUDDO achieves a maximum ARI of ≈ 0.49 and NMI of ≈ 0.55, which is a significant improvement in ARI and NMI from the default hyperparameter values.

Previous single-cell clustering algorithms often require a large number of dimensions ($\varepsilon \gtrsim 100$) to achieve reasonable clustering accuracy on single-cell Hi-C maps (Zhou et al. 2019). In addition, the ARI and NMI can possess large fluctuations as a function of the embedding dimension and depend strongly on the specific dimensionality reduction technique that is implemented, for instance

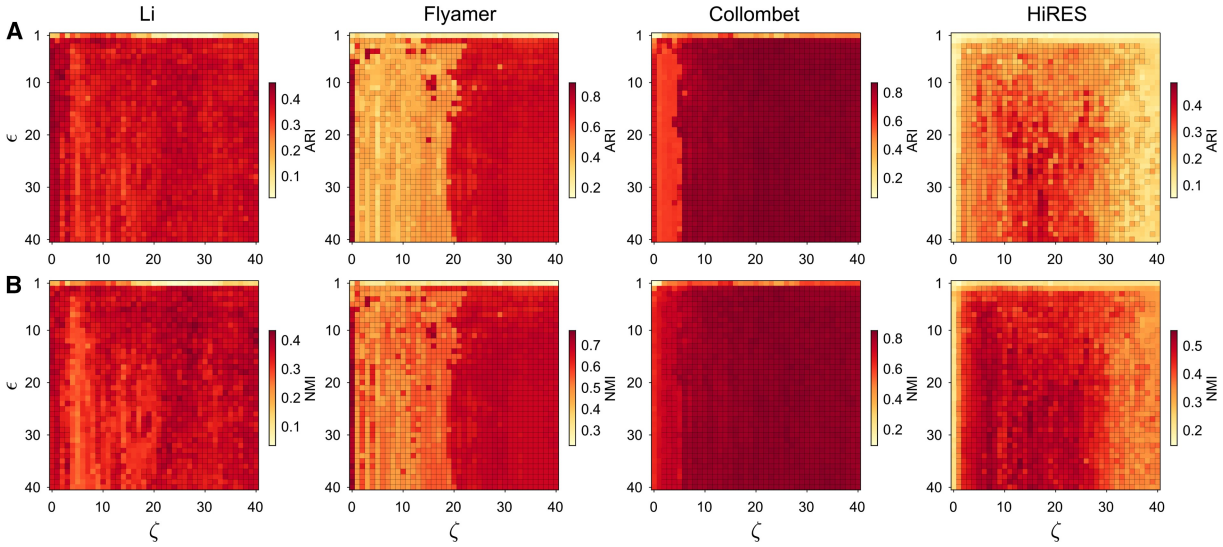

Figure 4 The clustering accuracy [ARI in (A) and NMI in (B)] for the SCUDDO algorithm for each of the four datasets, Li et al. (2019), Flyamer et al. (2017), Collombet et al. (2020), and the HiRES dataset (Liu et al. 2023), plotted as a function of the number of sampled superdiagonals ζ and the embedding dimension ε prior to spectral decomposition. The pixels in the heatmap are outlined when the ARI or NMI for SCUDDO exceed those of the next best performing method. The default values for SCUDDO are $\zeta = 25$ and $\varepsilon = 5$.

with some methods requiring specific dimensionality reduction techniques to be competitive (Zhang et al. 2022). In contrast, we have shown that the ARI and NMI scores for SCUDDO are large at both very low embedding dimensions and when sampling only a few superdiagonals. This result is true even when we treat ε as the final embedding dimension of the output for SCUDDO, despite the fact that SCUDDO always outputs a $l-1$ dimensional embedding, where $l<\varepsilon$ for all datasets except for HiRES. In addition, we find that SCUDDO does not depend sensitively on the specific dimensionality reduction technique. For instance on the Collombet et al. dataset, SCUDDO performs roughly equivalently when $\mathcal{V}$ is embedded spectrally (i.e. the default embedding) (ARI ≈ 0.85), embedded using UMAP (McInnes et al. 2018) (ARI ≈ 0.84), embedded using t-SNE (Hinton and Roweis 2002) (ARI ≈ 0.85), and when there is no further dimensionality reduction and using $\mathcal{V}$ directly (ARI ≈ 0.78). For detailed UMAP and t-SNE embeddings of SCUDDO for all datasets, see Fig. 4, available as supplementary data at *Bioinformatics* online. While the default values for ζ and ε for SCUDDO were not optimized to the four selected datasets, the default parameters give excellent results for ARI and NMI for these datasets. However, there are (ζ,ε) pairs that give superior performance across all datasets in this manuscript.

In Fig. 5, we calculate ARI and NMI for the four datasets versus the number of chromosomes used by SCUDDO. For these calculations, we sample all chromosomes with an index less than or equal to b, for instance, if we set $b=4$, SCUDDO samples only chromosomes 1, 2, 3, and 4. For the Li et al. and Collombet et al. datasets, we find that the ARI and NMI for the SCUDDO algorithm first exceed those for the next best method when $b \gtrsim 5$, respectively (with the exception of Collombet showing a dip at $b=16$). For the Flyamer et al. dataset, we find that when $b \gtrsim 10$ there is a jump in SCUDDO ARI/NMI scores, past which the ARI/NMI is mostly stable and higher than the next best ARI/NMI. The HiRES dataset requires the most chromosomes to sample, with SCUDDO only achieving state-of-the-art performance when $b \gtrsim 11$ and even then with some dips in performance at $b=14,16$.

4 Discussion

In this article, we develop a novel algorithm, SCUDDO, to determine a low-dimensional representation and then cluster single-cell Hi-C maps. We focused on four difficult-to-cluster single-cell Hi-C map datasets, where the datasets include ground-truth labels for each single-cell Hi-C map. We compared the ARI and NMI metrics for clustering accuracy from SCUDDO to those from five other clustering algorithms that were the most accurate in

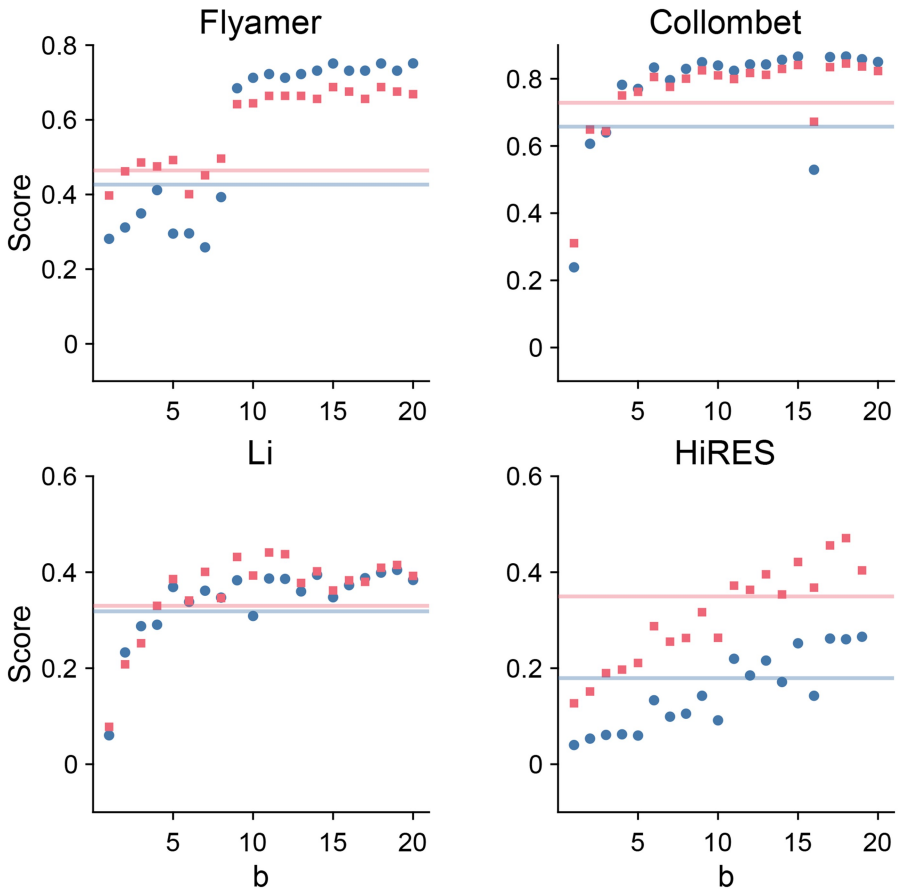

Figure 5 The clustering accuracy, ARI (red squares) and NMI (blue circles), for the SCUDDO algorithm plotted as a function of the number of sampled chromosomes b for the four datasets: (top left) Flyamer et al. (2017), (bottom left) Li et al. (2019), (top right) Collombet et al. (2020), and (bottom right) the HiRES dataset (Liu et al. 2023). The faint horizontal red and blue lines represent the maximum values of ARI and NMI achieved by the other methods. The next best ARI/NMI scores do not have to come from one method.

previous single-cell Hi-C map clustering benchmarking studies (Zhen et al. 2022, Ma et al. 2024, Plummer et al. 2025). SCUDDO is more accurate in all cases in terms of both ARI and NMI compared to the other methods for all datasets. We also find that SCUDDO can accurately cluster single-cell Hi-C maps using a fraction of the intrachromosomal Hi-C maps and their superdiagonals, as well as fewer embedding dimensions.

SCUDDO has several advantages over other existing methods for clustering single-cell Hi-C maps. First, SCUDDO uses, to our knowledge, a new imputation technique for single-cell Hi-C maps, calculating a graph kernel, then smoothing over local neighborhood features in individual intrachromosomal matrices using a Gaussian kernel, and finally using matrix exponentiation. We find that this technique readily improves the accuracy for clustering, since it allows both short-range and long-range information transfer within an intrachromosomal Hi-C map (for a detailed ablation analysis of SCUDDO, see Fig. 2, available as supplementary data at *Bioinformatics* online). SCUDDO also mainly uses PCA and MDS for dimensionality reduction, both of which are much more interpretable than dimensionality reduction techniques like UMAP and t-SNE or using complex inscrutable networks that require training to find features. SCUDDO is also fast and scalable (Fig. 3, available as supplementary data at *Bioinformatics* online) and is highly parallelizable since each (super)diagonal matrix can be processed independently in parallel. Additionally, we provide a version of SCUDDO ("fast SCUDDO") that uses several computational shortcuts including a truncated Taylor series expansion for the matrix exponential, classical MDS instead of non-classical MDS, and a Neumann series approximation for the graph kernel to achieve extremely fast runtime (≈ 5000 cells were run in <5 minutes) at the expense of some accuracy (a mean loss in ARI of ≈ 0.05).

The SCUDDO algorithm combines two key features: the diffused (normalized) superdiagonals of each single-cell Hi-C map and the trinarized differences along the diffused superdiagonals. The algorithm then calculates the angles between these features for each superdiagonal (using cosine similarity) and performs several steps of dimensionality reduction to achieve a final embedding. We find that for some datasets both features are necessary to achieve the best clustering accuracy, e.g. for the HiRES dataset using only one feature scores a maximum ARI of only ≈ 0.15. While the details of the features and SCUDDO algorithm are easy to interpret mathematically, the biophysical interpretation of these features is less clear. For example, it is unclear whether the diffusion and smoothing steps used by SCUDDO have a clear biophysical interpretation.

While our results show that SCUDDO maintains state-of-the-art clustering performance when using both high-dimensional latent vectors and 2D projections (e.g. UMAP and t-SNE; see Section 3 and Fig. 4, available as supplementary data at *Bioinformatics* online), the preservation of global higher-order structures in 2D may not hold across all single-cell Hi-C datasets due to topological distortions with further dimensionality reduction. Thus, utilizing these lower-dimensional visualizations for critical applications such as aligning modalities, defining cell-type boundaries, or aggregating cells for pseudo-bulk analysis can introduce false associations if the projection distorts the underlying manifold. We therefore recommend that users exercise caution in making inferences from lower-dimensional features like neighbors or distances and when possible, to use the higher-dimensional latent embeddings rather than relying solely on visual proximity in 2D space.

Interesting future studies involve coupling molecular dynamics simulations of polymer models of chromosomes (Sun et al. 2021) with the SCUDDO algorithm to further improve clustering accuracy and to better understand the biophysical mechanisms that support the efficacy of the methods used in the SCUDDO algorithm. In addition, the SCUDDO algorithm can be applied to synthetic datasets with ground-truth labels and tunable noise and sparsity, as well as to experimental datasets without labels to predict cell fate.

Acknowledgements

We thank Parisa A. Vaziri for insightful comments. We also thank the reviewers for their constructive comments and helpful suggestions, which significantly improved the quality of this manuscript.

Author contributions

Luka Maisuradze (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [equal]), Mark D. Shattuck (Writing—review & editing [supporting]), and Corey S. O'Hern (Funding acquisition [lead], Supervision [supporting], Writing—review & editing [equal])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

No competing interests are declared.

Funding

This work was supported by the National Science Foundation [1830904 to L.M. and C.S.O. and 2124558 to L.M. and C.S.O.].

References

Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on DiscreteAlgorithms, SODA '07, New Orleans, LA, USA. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics*, 2007, 1027–35.

Bickmore WA, van Steensel B. Genome architecture: domain organization of interphase chromosomes. *Cell* 2013;**152**: 1270–84.

Bintu B, Mateo LJ, Su J-H *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 2018;**362**:eaau1783.

Cattoni DI, Cardozo Gizzi AM, Georgieva M *et al.* Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat Commun* 2017;**8**:1753.

Collombet S, Ranisavljevic N, Nagano T *et al.* Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* 2020;**580**:142–6.

Finn EH, Misteli T. Molecular basis and biological function of variability in spatial genome organization. *Science (1979)* 2018; **365**:6457.

Finn EH, Pegoraro G, Brandão HB *et al.* Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* 2019;**176**:1502–15.e10.

Fletez-Brant K, Qiu Y, Gorkin DU *et al.* Removing unwanted variation between samples in Hi-C experiments. *Brief Bioinform* 2024;**25**:bbae217.

Flyamer IM, Gassler J, Imakaev M *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 2017;**544**:110–4.

Fudenberg G, Imakaev M. FISHing for captured contacts: towards reconciling FISH and 3C. *Nat Methods* 2017;**14**:673–8.

Galitsyna AA, Gelfand MS. Single-cell Hi-C data analysis: safety in numbers. *Brief Bioinform* 2021;**22**:bbab316.

Giorgetti L, Heard E. Closing the loop: 3C versus DNA FISH. *Genome Biol* 2016;**17**:215.

Hinton G, Roweis S. Stochastic neighbor embedding. *Adv Neural Inf Process Syst* 2002;**15**:857–64.

Hubert L, Arabie P. Comparing partitions. *J Classif* 1985; **2**:193–218.

Kruskal JB. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 1964;**29**:115–29.

Li G, Liu Y, Zhang Y *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* 2019; **16**:991–3.

Li X, Feng F, Pu H *et al.* scHiCTools: a computational toolbox for analyzing single-cell Hi-C data. *PLoS Comput Biol* 2021; **17**:e1008978.

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.

Liu J, Lin D, Yardimci GG *et al.* Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* 2018;**34**:i96–104.

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 R, Huang J, Jiang T *et al.* A mini-review of single-cell Hi-C embedding methods. *Comput Struct Biotechnol J* 2024; **23**:4027–35.

Maisuradze L, King MC, Surovtsev IV *et al.* Identifying topologically associating domains using differential kernels. *PLoS Comput Biol* 2024;**20**:e1012221.

Mateo LJ, Murphy SE, Hafner A *et al.* Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* 2019;**568**:49–54.

McInnes L, Healy J, Saul N *et al.* UMAP: uniform manifold approximation and projection. *JOSS* 2018;**3**:861.

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.

Plummer D, Lang X, Zhang S *et al.* A comprehensive benchmark of single-cell Hi-C embedding tools. *Nat Commun* 2025; **16**:9119.

Rodriguez J, Ren G, Day CR *et al.* Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell* 2019; **176**:213–26.e18.

Shah S, Takei Y, Zhou W *et al.* Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* 2018;**174**: 363–76.e16.

Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000;**22**:888–905.

Shi G, Thirumalai D. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. *Nat Commun* 2019;**10**:3894.

Smola AJ, Kondor R. Kernels and regularization on graphs. *Lect Notes Comput Sci* 2003;**2777**:144–58.

Stevens TJ, Lando D, Basu S *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 2017; **544**:59–64.

Sun Q, Perez-Rathke A, Czajkowsky DM *et al.* High-resolution single-cell 3D-models of chromatin ensembles during Drosophila embryogenesis. *Nat Commun* 2021;**12**:205.

Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J Mach Learn Res* 2010;**11**:2837–54.

Williamson I, Berlivet S, Eskeland R *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev* 2014;**28**:2778–91.

Zhang K, Zemke NR, Armand EJ *et al.* A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat Methods* 2024;**21**:217–27.

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

Zhen C, Wang Y, Geng J *et al.* A review and performance evaluation of clustering frameworks for single-cell Hi-C data. *Brief Bioinform* 2022;**23**:bbac385.

Zheng Y, Shen S, Keleş S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biol* 2022;**23**:222.

Zhou J, Ma J, Chen Y *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proc Natl Acad Sci U S A* 2019;**116**:14011–8.

© The Author(s) 2026. 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.
Bioinformatics, 2026, 42, 1–10
https://doi.org/10.1093/bioinformatics/btag284
Original Paper