VARIATIONAL GRAPH AUTO-ENCODER FOR DENOISING SINGLE-CELL HI-C DATA

Neda Shokraneh Kenari [1,2] **and Maxwell W Libbrecht** [1]
[1]Computing Science department, Simon Fraser University
[2]Human Genetics department, McGill University

ABSTRACT

Single-cell Hi-C (scHi-C) enables the study of 3D genome organization at the resolution of individual cells and cell types that cannot be isolated for bulk profiling. However, the extreme sparsity of scHi-C data presents major challenges, particularly in recovering cell-type-specific 3D structures when only a small number of cells are available. We introduce contactVI, a method that combines the strengths of graph-based models and variational autoencoders (VAEs) to account for spatial dependencies in noisy chromatin interaction data and effectively denoise them. On simulated data, contactVI outperforms existing imputation methods in recovering Hi-C contact maps at both the single-cell and cell-type levels. On real datasets, contactVI performs comparably to or better than other graph-based methods across different resolutions. When applied to jointly profiled single-cell Hi-C and RNA-seq data, contactVI successfully recovers the expected association between genome compartmentalization and gene expression. contactVI is available on GitHub: https://github.com/nedashokraneh/contactVI

Keywords scHi-C · sparsity · denoising · VGAE

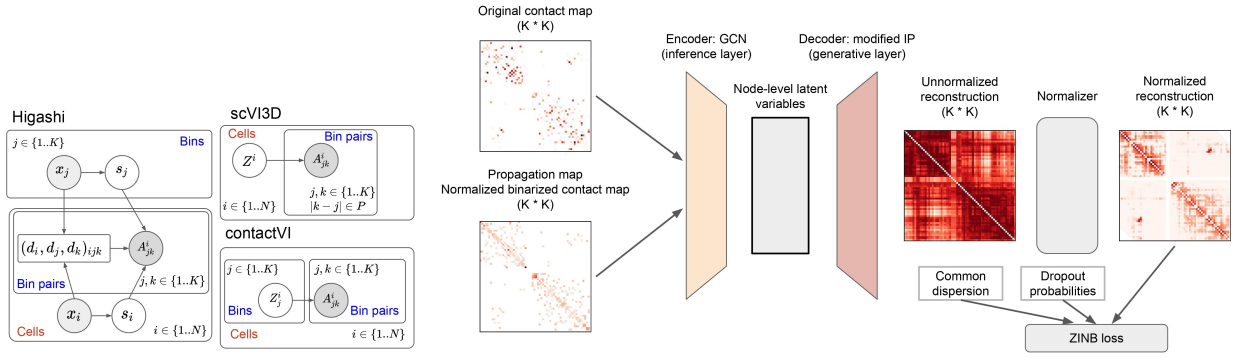
1 Introduction

Genome 3D organization plays important roles in cellular functions like replication timing and gene expression [1]. By studying genome 3D organization, researchers have discovered non-local gene regulation mechanisms [2] and unknown pathological mechanisms of gene expression [3]. Hi-C is a high-throughput technology capturing genome-wide chromatin interactions [4], widely used to study genome 3D organization. While traditional bulk Hi-C technologies [4, 5] measure average genome 3D organization across many cells, single-cell technologies are now combined with Hi-C measuring genome 3D organization at single-cell resolution, known as single-cell Hi-C (scHi-C) technology [6, 7, 8, 9, 10].

Single-cell Hi-C data allows studying genome 3D organization of rare and unknown cell types from complex tissues, its variability across cells from the same cell type, and its dynamic along development. However, the analysis of scHi-C data is challenging. For example, many zero observed interaction frequencies are not biologically zero, or all observed interaction frequencies are highly noisy, due to the limited sequencing material per cell, the problem known as sparsity. The sparsity of scHi-C data poses challenges for identifying cell types and characterizing cell- and cell-type-specific genome 3D structures. scHi-C imputation methods [11, 12, 13] share information between similar cells or neighbor genomic bins to denoise or reduce the sparsity of scHi-C data.

Traditional imputation methods are heuristic smoothing methods by applying 2D convolutional filters [14] or Random Walk with Restart (RWR) [11] on Hi-C contact maps that take into account the dependency between linear and spatial neighbor genomic bins respectively. While they have shown improvement in many downstream analyses of Hi-C and scHi-C data, they are heuristic and their performance depends on data properties which are not necessarily true.

More recently, two categories of learning-based approaches have emerged for denoising single-cell Hi-C data. First, deep latent variable models like the variational auto-encoder (VAE) have recently revolutionized the analysis of single-cell data [15, 16]. However, these methods are not directly applicable to scHi-C data due to its irregular data structure, where each cell's 3D organization is represented as a matrix or a graph, compared to a vector representation of other modalities like gene expression. scVI3D [13] adapts scVI [13] for scHi-C analysis by dividing contact maps into several pools, each containing bin pairs with similar genomic distances, and modeling each pool separately with the scVI model



(a) (b)

Figure 1: (a) Graphical model of learning-based single-cell Hi-C imputation methods. K: number of genomic bins. N: number of cells. A^i_{jk}: observed interaction frequency between genomic bins j and k in cell i. Higashi notations: x_i (cell features), x_j and x_k (genomic bin features), s_i (cell static representation), s_j (genomic bin statis representation), $(d_i, d_j, d_k)_{ijk}$ (dynamic representations of genomic bins and a cell given their features). scVI3D notation: Z^i (cell representation). contactVI notation: Z^i_j (genomic bin representations of a cell). (b) contactVI consists of three main modules: encoder, decoder, and normalizer. The encoder, a graph convolutional network, takes the original contact map as input. It takes original contact maps as node features representing genomic bins' neighborhood adjacency vectors, and a normalized version of the original contact map as a propagation matrix to produce node-level latent variables. The decoder reconstructs the graph structure from these latent variables. The normalizer converts the decoder's outputs (real values) into probabilities and scales them to represent expected interaction frequencies. When modeling observed interaction frequencies using a zero-inflated negative binomial (ZINB) distribution, a dropout module is included to predict the dropout probability between bin pairs. The likelihood is then calculated based on three learned parameters: expected interaction frequency (without dropout), dropout probability, and a dataset-wide dispersion parameter. Each contact map corresponds to one chromosome, and one model is trained per chromosome.

(Figure 1a). Although this approach leverages useful properties of scVI—such as removing library size and assuming appropriate distributions for observed counts—it overlooks both the spatial correlation between genomic bin pairs and the interaction frequency information between bin pairs in different pools.

The second category, graph-based methods such as RWR and Higashi, exploit the long-range dependencies revealed by Hi-C. These methods utilize a graph representation of the Hi-C modality, which is well-suited for capturing long-range dependencies arising from spatial positioning. For instance, if there is evidence that the genomic bin pairs (i, j) and (j, k) are spatially close, we would expect the pair (i, k) to also be spatially close, even if they are distant on the linear genome. Such dependencies can only be effectively captured with a graph representation. To do that, Higashi learns cell- and genomic-bin-level representations and decodes interaction frequencies from those representations, which allows sharing information between neighbor cells and spatially neighbor genomic bins (Figure 1a).

Here, we propose a method that combines the benefits of both VAE-based and graph-based approaches, called contactVI. We model the generative process of a single-cell contact map by defining genomic-bin-level latent variables and a decoder function generating the contact strength between two genomic bins given their embeddings and the genomic distance between them. We use a variational graph auto-encoder to learn the encoder for inferring the genomic-bin-level latent variables and the decoder for generating the contact maps. The latent variables are less noisy compared to the input contact map due to their lower dimension. Furthermore, the generative model is a probabilistic graph decoder and induces the dependency between genomic bin pairs with a common genomic bin anchor. Consequently, the generated interaction frequencies are less noisy because of information sharing across spatially neighbor genomic bins.

We show that contactVI reliably denoises a simulated scHi-C dataset and preserves the expected 3D structures at both single-cell and pseudo-bulk levels. We also demonstrate its performance on real scHi-C datasets by recovering expected 3D structures at the bulk level, identifying differential chromatin contacts between cell lines, and preserving the expected association between genome compartmentalization and gene expression. Our method models the generative process of scHi-C data, enabling effective denoising and imputation.

2 Materials and methods

2.1 scHi-C data representation

We represent a scHi-C dataset per chromosome as a set of graphs $G^1, \ldots, G^N$, where N is the number of cells. Each graph $G^i = (V, A^i)$ consists of a node set V of size K, corresponding to genomic bins in the chromosome, and a weighted adjacency matrix $A^i \in \mathbb{R}^{K \times K}$, corresponding to observed interaction frequencies in the single-cell contact map. For example, the entry A^i_{jk} denotes the observed interaction frequency between the jth and kth genomic bins in cell i. Here, we assume a chromosome C of size K, with all chromosomes modeled similarly. For simplicity, we omit the cell index i from the edge weight matrix A in subsequent discussions. From now on, we will use the terms nodes and genomic bins interchangeably, as well as adjacency matrix and contact map.

2.2 Formal problem definition

Given a set of graphs $G^1, \ldots, G^N$ with corresponding adjacency matrices $A^1, \ldots, A^N$ for chromosome C, our goal is to predict graphs $G^{*1}, \ldots, G^{*N}$ with denoised adjacency matrices $A^{*1}, \ldots, A^{*N}$. Our assumption is that A^{*i} is generated from genomic bin-level latent representations $Z^i \in R^{K \times d}$, where K and d are the number of genomic bins and the dimension of a latent representation, respectively. To achieve this, we extend the variational graph auto-encoder (VGAE) framework [17] to infer node- or genomic bin-level latent representations from the input graph structure and decode the denoised graph structures from these learned latent representations.

2.3 contactVI

We build on the idea of learning genomic bin-level representations from Higashi and extend the VAE-based framework—commonly used for modeling single-cell data—to a version tailored for scHi-C data with an explicit graph representation (Figure 1a). To denoise the observed scHi-C graphs, contactVI learns genomic bin latent representations for each cell using its adjacency matrix. The encoder (or inference layer) models this process by capturing information from linearly and spatially neighboring genomic bins. The decoder (or generative layer) then reconstructs the cell's contact map from these low-dimensional latent representations. Because the genomic bin latent representations are lower-dimensional and incorporate neighborhood information during inference, the reconstructed contact map is expected to be less noisy than the original. In addition to the encoder and decoder modules, contactVI includes a normalizer module that accounts for variable sequencing depths across cells. To handle the uncertainties inherent in single-cell measurements—often modeled with a zero-inflated negative binomial (ZINB) distribution—we train the model using either a Poisson or ZINB loss function, depending on the dataset. Figure 1b shows the overall architecture. Details of each module are explained in the following sections.

2.3.1 Encoder

The encoder in contactVI is a graph neural network (GNN) designed to learn denoised genomic bin latent representations by sharing information between spatially close genomic bins. Formally, given genomic bin features, X, and an adjacency matrix, A, the GNN module, $f(X, A)$, encodes these into genomic bin latent representations. We use a graph convolutional network (GCN) [18], which is a type of GNN, consisting of multiple propagation layers following the propagation rule:

$$H^{(l+1)} = \sigma(PH^{(l)}W^{(l)}), \tag{1}$$

where $H^{(l)}$ is the matrix of genomic bin representations at the l^{th} layer, with $H^{(0)} = X$ and $H^{(L)} = Z$ (a latent representation), and $W^{(l)}$ is the layer-specific trainable weight matrix. The matrix P is a propagation matrix, a normalized and modified version of the adjacency matrix A, which is defined as follows in GCN:

$$P = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}},\ \tilde{A} = A + I_K. \tag{2}$$

P is a normalized matrix, with edge weights normalized by the degree of the corresponding nodes ($\tilde{D}_{jj} = \sum_k \tilde{A}_{jk}$). contactVI has two propagation layers. The function σ in equation 1 is a non-linear activation function, ReLU, and identity function in the first and the second layers, respectively.

In the scHi-C dataset, we only have a graph structure (adjacency matrix) and aim to encode it into denoised representations of genomic bins. For node features, we use either an identity matrix (I_K), which is analogous to positional one-hot encoding for genomic bins or the original contact map (A) to have cell-specific node features analogous to interaction pattern with the rest of the chromosome for genomic bins. Through our ablation study, we show that the second version, using the original contact map as cell-specific node features, results in better reconstruction of the contact maps (Appendix C).

2.3.2 Decoder

The decoder in contactVI is a probabilistic graph decoder that operates on node representations. In the original VGAE [17], an inner product (IP) similarity function is used to predict edge weights between two nodes based on their embeddings. However, this function lacks parameterization and is not expressive enough to capture the variability of chromatin structures across different cells, specifically in datasets with a larger number of cells. To address this limitation, we propose alternative decoder variations: weighted IP, joint and combined.

Weighted IP: The weighted inner product (IP) decoder is similar to the standard IP decoder; however, instead of performing an unweighted average, the pairwise multiplications are weighted according to learned parameters. Formally, given the embeddings of two nodes j and k, $\overrightarrow{Z_j} = (Z_{j1}, Z_{j2}, ..., Z_{jd})$ and $\overrightarrow{Z_k} = (Z_{k1}, Z_{k2}, ..., Z_{kd})$, the unnormalized interaction frequency between the nodes, w_{jk}, is computed as: $w_{jk} = \text{MLP}(\sigma(\text{MLP}([\overrightarrow{t}, d|j,k|])))$, $\overrightarrow{t} = Z_j \odot Z_k$, where $\overrightarrow{t}$ is a pairwise multiplication vector between two node embeddings, $d|j,k|$ is a linear genomic distance between two nodes, and $[,]$ indicates the concatenation of the pairwise multiplication vector and a genomic distance. The concatenation of the size of $n+1$ is then passed into a two-layer MLP with the output layer of size one to predict the interaction frequency between these two genomic bins.

Joint: The joint decoder takes the concatenation of two node embeddings and a genomic distance as input to capture first-order relationships between node embedding dimensions. Then w_{jk} is calculated as: $w_{jk} = \text{MLP}(\sigma(\text{MLP}([\overrightarrow{Z_j}, \overrightarrow{Z_k}, d|j,k|])))$. The size of the joint decoder's input layer is $2n+1$.

Combined: The combined decoder takes the concatenation of two node embeddings in addition to the pairwise multiplication vector and a genomic distance as input to capture both first-order and second-order relationships between node embedding dimensions. The vector $\overrightarrow{t}$ is calculated similar to the weighted IP, and then w_{jk} is calculated as: $w_{jk} = \text{MLP}(\sigma(\text{MLP}([\overrightarrow{Z_j}, \overrightarrow{Z_k}, \overrightarrow{t}, d|j,k|])))$. The size of the combined decoder's input layer is $3n+1$.

Our ablation study shows that the combined decoder is the most expressive one being able to reconstruct cell-type-specific contact maps (Appendix C).

2.3.3 Normalizer

The normalizer module converts the unnormalized predicted interaction frequencies into normalized interaction frequencies using a softmax layer. We employ two variations of the normalizer in contactVI. The first, the whole normalizer, applies the softmax layer to all unnormalized interaction frequencies, $w_{jk}, \forall j,k \in 1..N,\ j<k$, ensuring that the sum of all normalized interaction frequencies across all bin pairs is one. The second variation, the pool normalizer, applies the softmax layer to each pool, separately. Pooling [19, 13] is used to split the Hi-C matrix by distance. For example, distance 0 is its own pool, distances 1 and 2 are pooled, distances 3, 4, and 5 are pooled, and so on, until all unit distances are grouped into one of p pools. This pooling approach aims to maintain a similar number of bin pairs in each pool. In the pool normalization, the unnormalized interaction frequencies corresponding to pool p, $w_{jk}, \forall (j,k) \in$ pool p, are normalized together using a softmax layer such that their sum equals one.

Next, the normalized interaction frequencies are scaled by either the total sequencing depth or the depth of the corresponding pool, depending on the normalization type. We denote the scaled normalized interaction frequencies as W_{jk}. The scaled normalized interaction frequencies are then used as the expected value of the non-zero component in the distribution of choice for interaction frequencies. In our ablation study, we show that there is a minor difference between the performance of the two normalizers, however, a pool normalizer slightly improves the performance due to better information sharing between linearly distant genomic bins (Appendix C). Therefore, we use the pool normalizer in our experiments.

2.3.4 Dropout module

In the case of using zero-inflated negative binomial (ZINB) distribution to model the measurement error of observed interaction frequencies, contactVI has an additional module to learn the dropout probability of the ZINB distributions. The dropout probabilities are cell- and genomic distance-specific. Following Higashi [12], the dropout module takes the total sequencing depth of the cell and the linear genomic distance between two genomic bins as input and predicts the dropout probability.

2.3.5 Training

We use either the Poisson (for sparser datasets) or the ZINB observational (for the less sparse dataset) model to train contactVI (Appendix C). If training by the ZINB model, a shared dispersion parameter, θ, is learned for the entire dataset. More formally, the conditional ZINB log-likelihood for the cell n is calculated as:

$$\mathrm{LL}_{\text{zinb}} = \sum_{\substack{j,k \in \{1..K\} \\ j < k}} \log\ P_{\text{zinb}}(A^i_{jk} | W^i_{jk}, \theta, \pi_{i|j-k|}), \tag{3}$$

where A^i_{jk} and W^i_{jk} are the observed and reconstructed interaction frequencies between genomic bins j and k in a cell i, respectively. $\pi_{i|j-k|}$ is a predicted dropout probability based on the total sequencing depth of cell i and a linear genomic distance between two genomic bins.

The variational inference procedure provides us with the ELBO function:

$$\mathrm{ELBO} = \mathrm{LL}_{\text{zinb}} - D_{KL}(q_\phi(Z|A^i)||p(Z)), \tag{4}$$

where Z is a genomic bin latent representation matrix, q_ϕ is an encoder module, and p is a prior for genomic bin latent representations, which is Gaussian. Note that the KL divergence is averaged across Z dimensions.

We consider each cell as a sample, pass it through the model to calculate ELBO, and use the negative ELBO to backpropagate and train a model. The encoder is a 2-layer GCN with hidden sizes of 32 and 16, respectively. We use distribution modules from scVI tools [16] to calculate the ZINB log-likelihood. We use Adam optimizer with a learning rate of 0.001 and train a model for 60 epochs.

2.4 Supplementary materials and methods

Datasets and evaluation metrics are explained in A and B, respectively.

3 Results

3.1 contactVI outperforms other imputation methods in denoising Hi-C contact maps at multiple scales in the simulated dataset

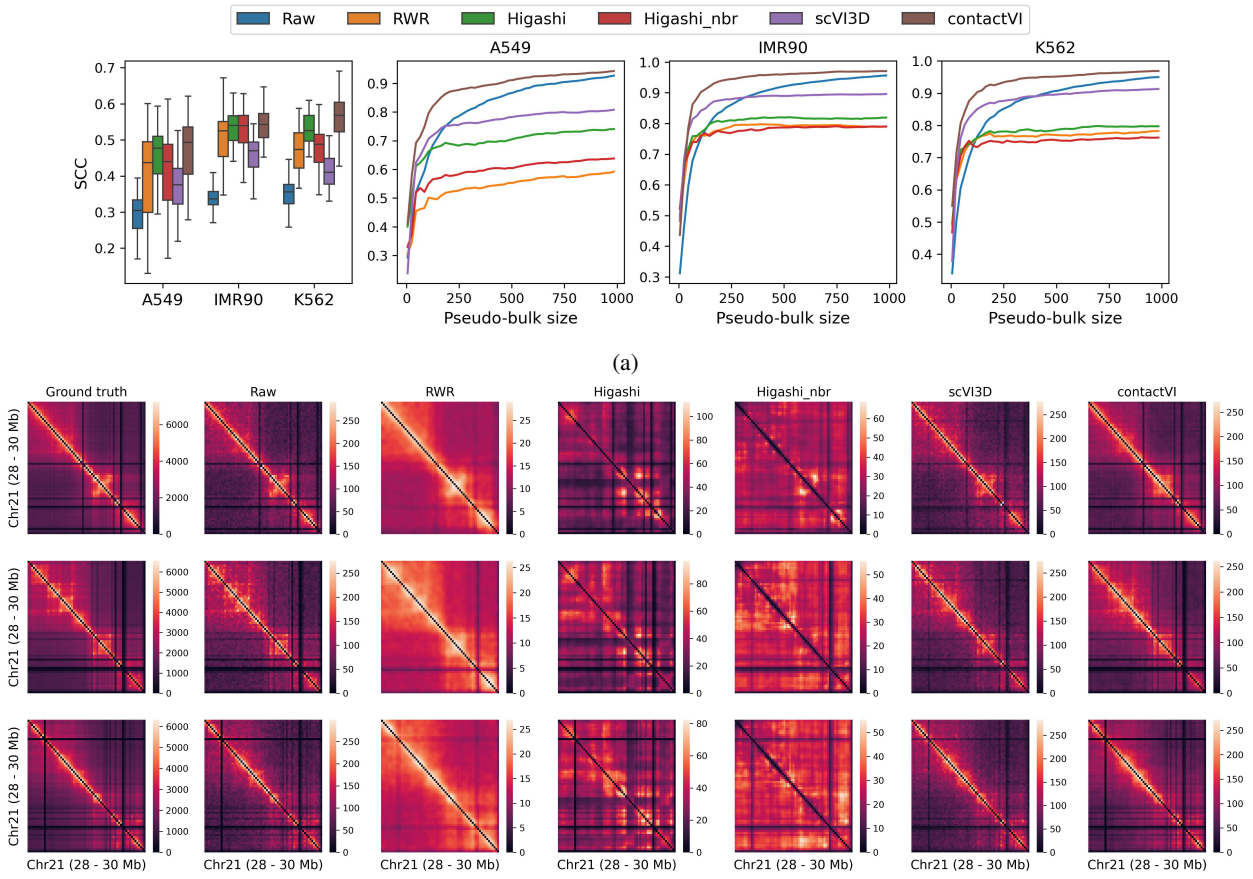
We simulated a realistic sparse scHi-C dataset profiling 1000 cells per cell line for three cell lines at 30 Kb resolution spanning chromosome 21 (28-30 Mb) based on single-cell imaging dataset (Appendix A, simulated dataset). We ran different imputation methods on this dataset, and visualized pseudo-bulked contact maps constructed from the sparse down-sampled data (Raw), the original data (Original), and the imputed versions of the sparse data (Figure 2b). The contact maps imputed with contactVI closely resemble the ground truth and appear less noisy than the raw data, especially when fewer cells are aggregated (Figure S3). In contrast, contact maps imputed with RWR are over-smoothed and lose finer structural details. Higashi's imputed contact maps appear invalid in this simulated setting—likely due to its complex, data-driven model struggling to train effectively on a sub-genomic dataset rather than a genome-wide one. The pseudo-bulked contact map imputed by scVI-3D also resembles the original data but appears less smoothed, likely due to ignoring the graph structure of the data in its model.

Second, we quantified the stratum-adjusted correlation coefficient (SCC) similarity between the pseudo-bulked imputed contact maps and the original data across different pseudo-bulk sizes. Consistent with the visualizations, contactVI achieves the highest similarity with the original data across various pseudo-bulk sizes (Figure 2a). Taking advantage of the available ground-truth single-cell contact maps, we also assessed the similarity between imputed and ground-truth contact maps at the single-cell level. Because Pearson correlation of contact map diagonals is highly sensitive to outliers, especially given the sparsity of the data, we instead created pseudo-bulked contact maps of size 5 and compared each to the corresponding ground-truth pseudo-bulk (constructed from the same 5 cells). This analysis confirms that contactVI denoises contact maps effectively across different pseudo-bulk sizes (Figure 2a, left).

3.2 contactVI-denoised scHi-C data closely matches pseudo-bulked high-depth scHi-C and bulk Hi-C

Furthermore, we simulated a genome-wide sparse scHi-C dataset by down-sampling from a scHi-C dataset with a relatively higher sequencing depth (*Wu2024*) at 1 Mb resolution. More specifically, we generated a simulated scHi-C contact map per cell from the original dataset by down-sampling it to $\frac{1}{8}$ of the sequencing depth of the original cell. Then, we applied different imputation methods to the simulated dataset and assessed the similarity of the single-cell and pseudo-bulked (for GM12878 cell line) down-sampled and imputed contact maps with the single-cell and pseudo-bulked original contact maps, respectively. Unfortunately, we observed that all imputation methods decrease the similarity of the down-sampled data to the original data at single cells (Figure S4b). This likely indicates that, although imputation methods reduce noise, they also introduce subtle biases on a single cell level. All existing imputation methods, including contactVI, exhibit a similar extent of such bias (Figure S4c).

Due to the limited sequencing depth of the original data, we alternatively used the bulk Hi-C data for the GM12878 as a ground truth. We also considered the pseudo-bulk of all GM12878 single-cells from the original data (n = 220), which



(a)

(b)

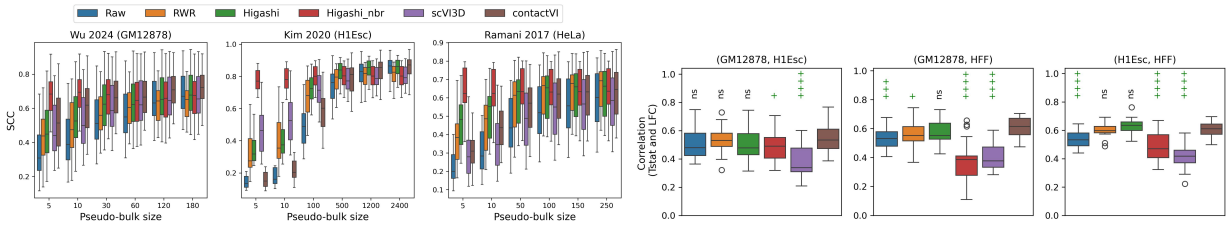
Figure 2: (a) Comparison of imputation methods based on their similarity to the ground-truth pseudo-bulk contact map, measured using the stratum-adjusted correlation coefficient (SCC). Panel 1 shows SCC values for pseudo-bulks of size 5, with each point representing one group of 5 cells. Panels 2–4 show SCC values across varying pseudo-bulk sizes for each of the three cell lines. Pseudo-bulk size refers to the number of aggregated cells. (b) Heatmap visualization of pseudo-bulked contact maps generated from the down-sampled raw data, the original data, and the imputed data, by aggregating all 1,000 cells from each cell line.

can be considered a mini-bulk dataset. We observe a different pattern based on these metrics, such that imputation methods improve the similarity with bulk or mini-bulk data (Figure S4a). Excluding Higashi with neighbor imputation that removes cell-specific structures (pseudo-bulking more cells does not improve the similarity with the bulk data), contactVI and Higashi outperform RWR and scVI3D based on these metrics, although the gap between the raw down-sampled and imputed data becomes smaller by increasing the pseudo-bulk size.

3.3 contactVI performs comparably or outperforms other imputation methods in preserving cell-type-specificity and differential signals in real datasets

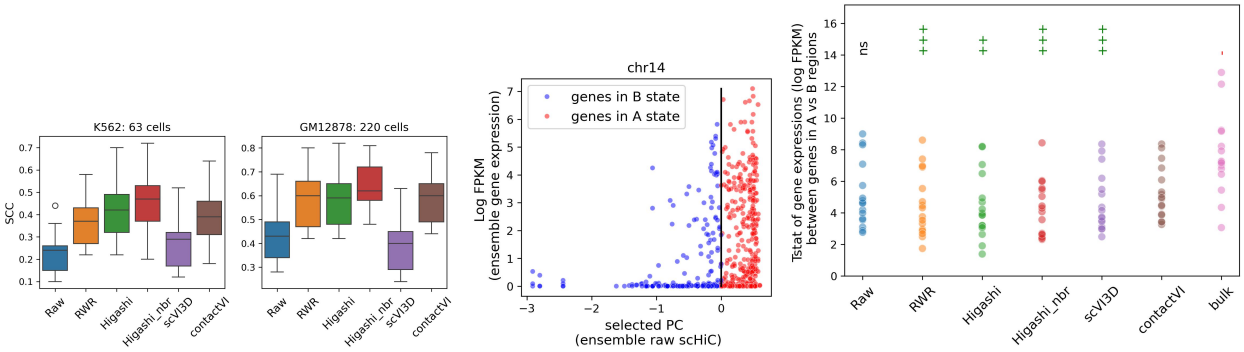
Next, we compared different imputation methods based on resembling the deeply sequenced bulk data after pseudo-bulking in real datasets at 1 Mb resolution. We imputed three real scHi-C datasets profiling human cell lines, *Ramani2017* [7], *Kim2020* [8], and *Wu2024* [10] with existing bulk Hi-C data available. Then, we calculated the similarity between the bulk contact map of a given cell line and a pseudo-bulked contact maps after the imputation for different number of aggregated cells.

Wu2024 [10] profiled 431 cells from four human cell lines. Although this dataset has higher coverage compared to other single-cell Hi-C datasets, the small number of profiled cells leads to challenges, and pseudo-bulking 63 K562 cells results in noisy long-range interactions (Figure S5b, Raw). However, after contactVI imputation, the long-range interaction patterns closely resemble the ground-truth bulk data (Figure S5b, contactVI). Additionally, the overall similarity (SCC) between the pseudo-bulked imputed single-cell contact maps and the bulk map across chromosomes is highest with contactVI compared to other imputation methods (Figure 3a, left).



(a) (b)

Figure 3: (a) Comparison of imputation methods based on the similarity of pseudo-bulked imputed contact maps with the ground-truth bulk contact map, measured using the stratum-adjusted correlation coefficient (SCC). (b) Comparison of imputation methods based on their ability to recover known differential chromatin contacts identified in bulk analyses. For each chromosome, the correlation between t-statistics (Tstat, from a single-cell DCC caller) and log fold changes (LFC, from a bulk DCC caller) among candidate bin pairs is computed. Each point in the boxplot represents one chromosome. Results are shown for three cell line pairs from Kim 2020.



(a) (b) (c)

Figure 4: (a) Comparison of imputation methods based on the similarity between pseudo-bulked imputed contact maps and the ground-truth bulk contact map, measured using the stratum-adjusted correlation coefficient (SCC) for two cell populations in the *Wu2024* dataset at 100 Kb resolution. (b) Normalized gene expression (log FPKM) as a function of compartment scores, derived from the selected principal component (PC) of PCA on ensemble (pseudo-bulked) scHi-C data. Both gene expression and compartment scores are computed from 220 GM12878 cells profiled with scRNA-seq and scHi-C. (c) T-statistics comparing normalized gene expression between genes in A vs. B compartments, with compartments inferred from pseudo-bulked raw, pseudo-bulked imputed, or bulk GM12878 data. Each point in the boxplot represents one chromosome (chromosomes 6–22).

Quantitative evaluations on other datasets with more cells show that contactVI performs comparably to RWR and Higashi in reconstructing ground-truth bulk data when a sufficient number of cells are available (Figure 3a, middle and right). However, the pseudo-bulked contact map is not necessarily cell-type-specific because of its high similarity with that of a bulk map. This is due to the existence of similar structures between different cell types, therefore, reconstructing the shared patterns between different cell types and removing the cell-type-specific structures does not severely decrease the similarity performance. Therefore, we analyzed differential chromatin contacts after the imputation and compared it with known differential contacts from the bulk data in three pairs of cell types from *kim2020* dataset (Appendix B.2). Figure 3b shows that contactVI performs comparably to RWR and Higashi in terms of recovering differential chromatin contacts and performs better than scVI-3D. The worse performance of scVI-3D is due to splitting up the dataset into the pools, therefore, the latent space of the further pools and consequently their reconstructions would not be cell-type-specific. This observation confirms the importance of the extension of existing deep latent variable models to Hi-C modality without data splitting or ignoring the data structure.

3.4 contactVI denoises scHi-C data across resolutions and preserves biological signals

To assess the performance of contactVI at finer resolutions, we imputed the less sparse scHi-C dataset, *Wu2024*, at 100 Kb resolution. First, we calculated the similarity between the pseudo-bulked imputed contact maps for two cell populations—GM12878 and K562—and their corresponding bulk data. Our results show that all three graph-based imputation methods—RWR, Higashi, and contactVI—increase the similarity with the bulk data (Figures 4a, S6),

highlighting the benefit of inductive bias in enhancing global similarity. However, this inductive bias may introduce false positive signals, as seen in the heatmap visualization (Figure S6). While these false positives have little effect on global similarity, they may obscure true biological signals.

To further evaluate the imputation methods in terms of preserving biological signal, we developed a metric based on the well-established association between genome compartmentalization and gene expression. The genome is divided into A and B compartments, representing transcriptionally active and inactive regions, respectively. Compartment states can be inferred from principal component analysis (PCA) of Hi-C contact maps [4]. Because gene expression data for the same cells are available in the *Wu2024* dataset, we could directly assess the relationship between inferred compartment states and gene expression.

We computed pseudo-bulked raw and imputed contact maps for the GM12878 cells (n = 220) and inferred their compartment annotations. We similarly annotated compartments from bulk GM12878 data to serve as a reference. A compartment score was then assigned to each gene based on its genomic position. As shown in Figure 4b, genes located in the A compartment (score > 0) are significantly more expressed than those in the B compartment (score < 0). We used a t-statistic comparing the normalized expression levels of genes in the A vs. B compartments as a metric to evaluate the association between compartment calls and gene expression (Appendix B.3). Figure 4c shows that all imputation methods, except contactVI, slightly reduce this association compared to the raw data, whereas contactVI preserves it (Figure S7).

This discrepancy between improved contact map similarity and reduced biological signal may arise from the features being compared. Similarity metrics assess each observed or unobserved interaction between locus pairs, making them sensitive to noise. In contrast, compartment annotations summarize the entire contact map into a single score per genomic bin, smoothing out noise. As a result, false positives introduced by imputation can outweigh the benefits of correctly imputed interactions and degrade biological relevance. These findings highlight the need for caution when applying imputation methods in downstream analyses.

4 Discussion

We proposed contactVI, a deep latent variable model designed specifically for single-cell Hi-C (scHi-C) data. We demonstrated that contactVI produces accurate imputations on both simulated and real scHi-C datasets. It outperforms or performs comparably to existing imputation methods in reconstructing chromatin contact maps at both the single-cell and pseudo-bulk levels.

In particular, contactVI performs well in recovering long-range interaction patterns, even when only a small number of cells are available. These long-range interactions are critical for inferring higher-order 3D genome structures such as compartments and subcompartments, which play important roles in gene regulation. We also showed that contactVI-imputed data preserves the expected association between genome compartmentalization and gene expression.

While contactVI improves data quality for downstream tasks, it may be less advantageous for characterizing cell types when a sufficiently large number of cells from each type is available. This limitation may stem from its emphasis on denoising the data rather than selectively imputing dropout values. In contrast, statistical methods developed for single-cell RNA-seq, like scImpute [20], model uncertainty in the observed values and perform more conservative imputation. These methods estimate imputed values while preserving confidence in observed data. HiCImpute [21] follows a similar principle by distinguishing between structural and dropout zeros and imputing only the latter. However, it has not yet been benchmarked against other single-cell Hi-C imputation methods.

A promising direction for future work is to compare such conservative, uncertainty-aware approaches to contactVI and other methods benchmarked in this study. Given that contactVI is a likelihood-based framework, it could be extended to model uncertainty by estimating residuals based on the observed values, their expectations, and standard deviations. These residuals could be used to assess the confidence in each observation and enable selective shrinkage of observed values toward their expected values depending on the magnitude of the residuals.

References

[1] Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nature structural & molecular biology*, 20(3):290–299, 2013.

[2] David U Gorkin, Danny Leung, and Bing Ren. The 3d genome in transcriptional regulation and pluripotency. *Cell stem cell*, 14(6):762–775, 2014.

[3] William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suvà, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016.

[4] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[5] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[6] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[7] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263–266, 2017.

[8] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS computational biology*, 16(9):e1008173, 2020.

[9] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou, Sahaana Chandran, Angeline Rivkin, Anna Bartlett, Joseph R Nery, Conor Fitzpatrick, Carolyn O'Connor, Jesse R Dixon, et al. Simultaneous profiling of 3d genome structure and dna methylation in single human cells. *Nature methods*, 16(10):999–1006, 2019.

[10] Honggui Wu, Jiankun Zhang, Fanchong Jian, Jinxin Phaedo Chen, Yinghui Zheng, Longzhi Tan, and X Sunney Xie. Simultaneous single-cell three-dimensional genome and gene expression profiling uncovers dynamic enhancer connectivity underlying olfactory receptor choice. *Nature Methods*, pages 1–9, 2024.

[11] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J Sejnowski, Jesse R Dixon, and Joseph R Ecker. Robust single-cell hi-c clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, 116(28):14011–14018, 2019.

[12] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature biotechnology*, 40(2):254–261, 2022.

[13] Ye Zheng, Siqi Shen, and Sündüz Keleş. Normalization and de-noising of single-cell hi-c data with bandnorm and scvi-3d. *Genome biology*, 23(1):1–34, 2022.

[14] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.

[15] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[16] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022.

[17] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[19] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multihiccompare: joint normalization and comparative analysis of complex hi-c experiments. *Bioinformatics*, 35(17):2916–2923, 2019.

[20] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.

[21] Qing Xie, Chenggong Han, Victor Jin, and Shili Lin. Hicimpute: A bayesian hierarchical model for identifying structural zeros and enhancing single cell hi-c data. *PLoS computational biology*, 18(6):e1010129, 2022.

[22] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Biophysical Journal*, 112 (3):217a, 2017.

[23] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, 2018.

[24] Aaron TL Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):1–11, 2015.

Supplementary notes

A Datasets

Real datasets: We used two types of real scHi-C datasets: *single-cell combinatorial indexed Hi-C* (sciHi-C) [7, 8] and *Linking mRNA to Chromatin Architecture* (LiMCA) [10], which jointly profiles the 3D genome and the transcriptome.

The first sciHi-C dataset is *Ramani2017* [7] [1], which profiles four human cell lines and includes 620 cells after filtering out low-quality cells, with an average of approximately 24K total contacts per cell. The second sciHi-C dataset is *Kim2020* [8] [2], profiling five human cell lines with $8,021$ cells and an average of about 11.4K total contacts per cell. The LiMCA dataset (*Wu2024*) [10] [3] profiles four human cell lines with 421 cells and an average of about 1M total contacts per cell, making it substantially less sparse than sciHi-C datasets.

Simulated dataset:

The single-cell imaging assay [22, 23] is an orthogonal, low-throughput technology for scHi-C data. It detects the 3D coordinates of a limited number of genomic positions, enabling 3D reconstruction of a genomic region, typically spanning less than 10 Mb. These reconstructions are more accurate than those from sequencing-based assays, as they are not constrained by the amount of DNA material. We use single-cell imaging data from Bintu et al. [23] to simulate a sparse scHi-C data, following the procedure used in Higashi [12]:

For each cell,

1. Calculate the pairwise distance between all genomic bins based on their 3D coordinates.
2. The inverse of these distance measurements is proportional to the contact probability between genomic bins. Normalize the inverse distances to sum to 1, representing the contact probabilities. The product of these contact probabilities and an arbitrary sequencing depth serves as the ground truth contact map.
3. To generate the simulated scHi-C dataset, use the contact probabilities from step 2 as parameters for a multinomial distribution P. Given the lower bound (lb) and upper bound (ub) of sequencing depth for a dataset to be simulated, sample a sequencing depth, s, from a uniform distribution defined over the range $[lb, ub]$. Then, sample s contacts from the distribution P.

We repeated this process for $3,000$ cells, including $1,000$ cells each from the A549, IMR90, and K562 cell lines. We set lb and ub to 100 and 250, respectively. The dataset profiles a genomic region on chromosome 21 (28–30 Mb) at 30 Kb resolution [4].

B Evaluation metrics

B.1 The similarity of imputed and ground-truth contact maps

Generally, there is no ground truth for single-cell contact maps, except in the case of simulated datasets. When analyzing simulated data, we assess imputation performance by comparing the imputed map to the original dense map from which the sparse data were generated. However, if the single-cell contact maps are highly sparse, similarity metrics can be sensitive to outliers and thus unreliable. To mitigate this, we aggregate groups of five cells (from the same cell state) and compute their similarity to the aggregation of the corresponding original (pre-sparsified) cells (referred to as single-cell similarity).

We also assess similarity at the pseudo-bulk level by aggregating a specific number of imputed cells into a pseudo-bulk contact map and computing its similarity to the aggregation of the corresponding original cells (pseudo-bulk similarity using the same set of cells).

For real sparse scHi-C datasets without ground truth, we expect that the pseudo-bulk profile—created by aggregating many single cells from the same cell state—will converge to the corresponding bulk profile. Using three scHi-C datasets profiling human cell lines with available bulk profiles, we aggregate a specific number of imputed single-cell profiles to create pseudo-bulk profiles and then calculate their similarity to the bulk profiles (pseudo-bulk similarity with the bulk profile).

[1] source: Filtered cells in data.txt

[2] source: sci-Hi-C .matrix files

[3] source: GSE240114 from GEO

[4] source: A549 and IMR90, and K562

To measure similarity, we use the stratum-adjusted correlation coefficient (SCC) metric from the HiCRep framework [14]. Given the wide range of signal intensities across different genomic distances, an overall correlation coefficient might be skewed by unwanted factors such as distance effects. SCC addresses this by computing the correlation between two samples at each genomic distance and then calculating a weighted average of these correlations, assigning greater weight to smaller genomic distances.

B.2 The similarity of dynamic 3D genome organization identified from single-cell and bulk data

Another important property of a good imputation approach is its ability to accurately reconstruct known dynamics between two populations of cells after imputation. The dynamics of 3D genome organization occur at multiple scales, including compartments, TADs (Topologically Associating Domains), and loops. In this evaluation, we focus specifically on identifying differential chromatin contacts (DCCs) rather than inferring dynamic 3D structures.

We apply a DCC calling algorithm proposed in SnapHiC-D [9] with slight modification on imputed datasets. Given the imputed and annotated scHi-C dataset and two cell states, A and B, we use n single cells per state as its samples. We apply quantile normalization to the entire dataset, rather than using the distance-centric quantile normalization method from SnapHiC-D, to maintain consistency with ground-truth DCCs. Specifically, we use diffHiC [24] to identify ground-truth DCCs from bulk data, which incorporates whole-dataset normalization. Therefore, it is essential to use the same normalization strategy for consistency. Then, we apply the two-sided two-sample t-test similar to SnapHiC-D to calculate t-statistics and p-values corresponding to each candidate bin pair and adjust for the multiple comparison problem.

B.3 Explaining gene expression

We expect a good imputation method to preserve biological signal. The 3D organization of the genome plays a regulatory role in gene expression through various mechanisms, including its compartmentalization into two states: A (active) and B (inactive). Compartment annotations can be inferred from Hi-C contact maps using principal component analysis (PCA) [4]. Using the *Wu2024* dataset, we pseudo-bulked the raw or imputed contact maps of the GM12878 population ($n = 220$) and inferred compartment annotations based on either the raw or imputed data. Similarly, we aggregated the gene expression profiles of the same 220 GM12878 cells to construct a pseudo-bulk gene expression profile, from which we computed FPKM-normalized gene expression values.

We expect compartment annotations to reflect gene expression, such that genes in the A compartment exhibit significantly higher expression than those in the B compartment. To assess this, we calculated t-statistics of FPKM-normalized gene expression between genes assigned to A and B compartments, using annotations derived from each imputation method, the raw data, and the bulk GM12878 dataset as a reference.

C VGAE with a weighted decoder and a proper loss function robustly denoises scHi-C data

We imputed three real and one simulated scHi-C datasets profiling human cell lines using different variations of contactVI to study the impact of input features (identity matrix or raw contact map), decoder type (inner product (IP), weighted IP, joint, or combined), and scaling method (whole- or pool-based). For each dataset, we generated a pseudo-bulk contact map for one or more cell lines with matching bulk Hi-C data and calculated its similarity to the ground truth. In the simulated dataset, the pseudo-bulk of the original (non-downsampled) simulated scHi-C served as the ground truth.

Across all datasets, particularly those with a larger number of cells (e.g., Kim 2020 and Bintu 2018), using the raw contact map as the input feature outperformed the identity matrix (Figure S1). Using the identity matrix implies identical node features across cells, limiting the model's expressivity. In contrast, providing the raw contact map allows the model to better capture cell-to-cell variability. However, when using the raw contact map, a more expressive decoder than the unweighted inner product is required to accurately model interaction frequencies from the more heterogeneous node embeddings. Accordingly, the combination of raw contact maps as input features with weighted decoders consistently outperformed other configurations.

Due to the relatively high similarity of Hi-C contact maps across cell types and the insensitivity of SCC to fine-scale structures, high SCC values may still be observed even when cell-type-specific or fine-scale patterns are lost after imputation. To better compare the impact of different decoder types, we visualized pseudo-bulked imputed contact maps for three cell lines in the simulated scHi-C dataset. Although the weighted IP, joint, and combined decoders yield similar SCC scores, the weighted IP decoder fails to preserve cell-type-specific patterns. This suggests that a decoder relying solely on pairwise similarities in node embedding dimensions lacks the expressivity needed to capture the

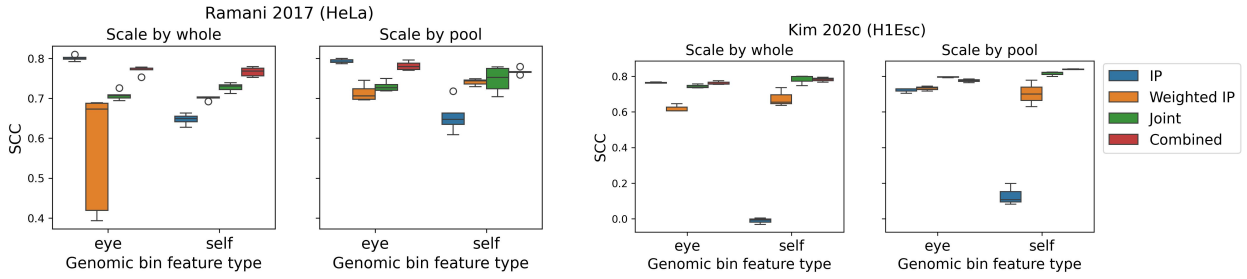
heterogeneity of interaction frequencies. While the joint decoder maintains cell-type specificity, the combined decoder further improves preservation of fine-scale structures, such as loops in IMR90 (Figure S2, black box in the second row). Based on these results, we use the combined decoder for all subsequent analyses.

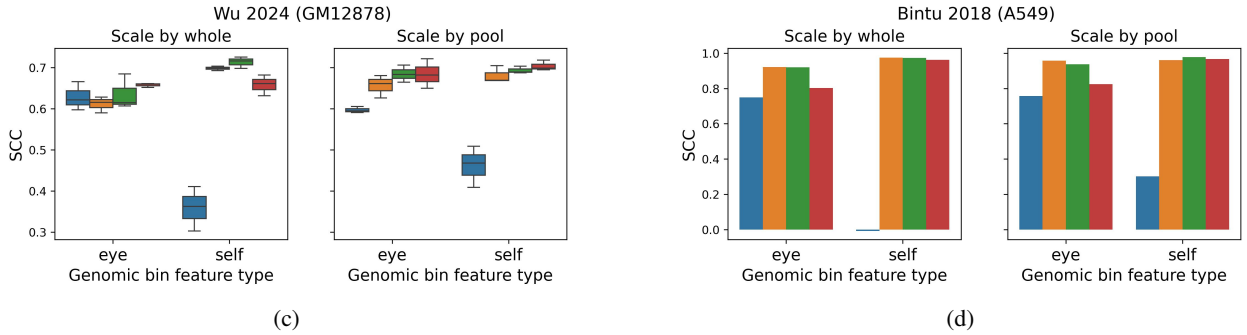
The difference between the two normalization approaches—whole-based and pool-based—is relatively minor. However, pool-based normalization performs better, particularly for larger datasets such as Kim. In the learning scheme, pool-based normalization predicts the ratio of interaction frequencies within a local pool, while whole-based normalization predicts the ratio across all interactions in a cell. As a result, bin pairs with varying genomic distances are scaled to a similar range of interaction frequency ratios. This allows long-range interactions to have greater influence during training, promoting better information sharing across all bin pairs. Therefore, we use the combined decoder followed by pool-based normalizer layer for our analyses.

Lastly, we explore the effect of the distribution assumption for observed interaction frequencies. While the Poisson assumption is usually enough for single-cell datasets with a relatively low sequencing depth, negative binomial (NB) or zero-inflated negative binomial (ZINB) usually better fit the measurements from more deeply sequenced datasets. This is due to the non-linear mean-variance relationship specifically for larger measurements. We also observed the deviation of mean-variance relationship from $y = x$ assumption of Poisson distribution, specifically for Wu 2024 dataset with the highest sequencing depth (Figure S1e). Consequently, while the validation loss (negative log likelihood (NLL)) of the Poisson model is comparable to more complex distributions in Ramani 2017, there is a larger gap in Wu 2024, indicating the necessity of non-Poisson distribution models for the datasets with larger sequencing depths.

These differences are also reflected in the similarity between the pseudo-bulked imputed contact maps and the ground truth. The ZINB distribution improves imputation performance on the Wu dataset, while it performs comparably to the Poisson model on the Ramani dataset. Notably, although the NB model shows a lower validation loss than ZINB, it does not lead to improved imputation performance. This counterintuitive result likely stems from the form of the ZINB likelihood: in zero-inflated settings, the model explicitly accounts for excess zeros via a mixture of a point mass at zero and a NB distribution. Consequently, the probability assigned to each observation may be lower than in the NB model, which assumes that all variability originates from a single NB process. This added distributional complexity can reduce the overall likelihood, even when ZINB more accurately captures the underlying data-generating process. Due to the lower computational cost of the Poisson model, we use it for all sparser datasets except Wu 2024—a newer and deeply sequenced dataset—for which we employ the ZINB model.

Supplementary figures





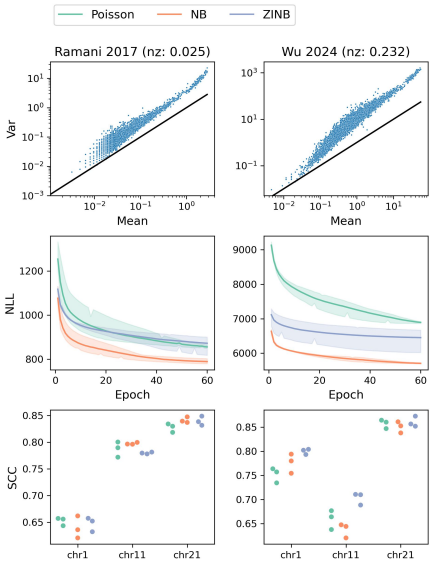


Figure S1: (a–d) Ablation analysis of contactVI modules, based on the similarity between pseudo-bulked imputed contact maps and the ground truth, measured using the stratum-adjusted correlation coefficient (SCC) metric. One population per dataset is shown, as indicated in each panel title. (e) Top: mean-variance relationship of observed interaction frequencies across cells. Middle: validation loss (negative log-likelihood, NLL) per epoch for models trained using different distributional assumptions. Bottom: similarity of pseudo-bulked contact maps after imputation, compared to ground truth, based on three distributional assumptions. Results are shown for two datasets with differing sparsity levels (left: Ramani 2017; right: Wu 2024). NLL plots correspond to training on chromosome 11. SCC (stratum-adjusted correlation coefficient) is computed for HeLa and GM12878 cell lines from the Ramani and Wu datasets, respectively.

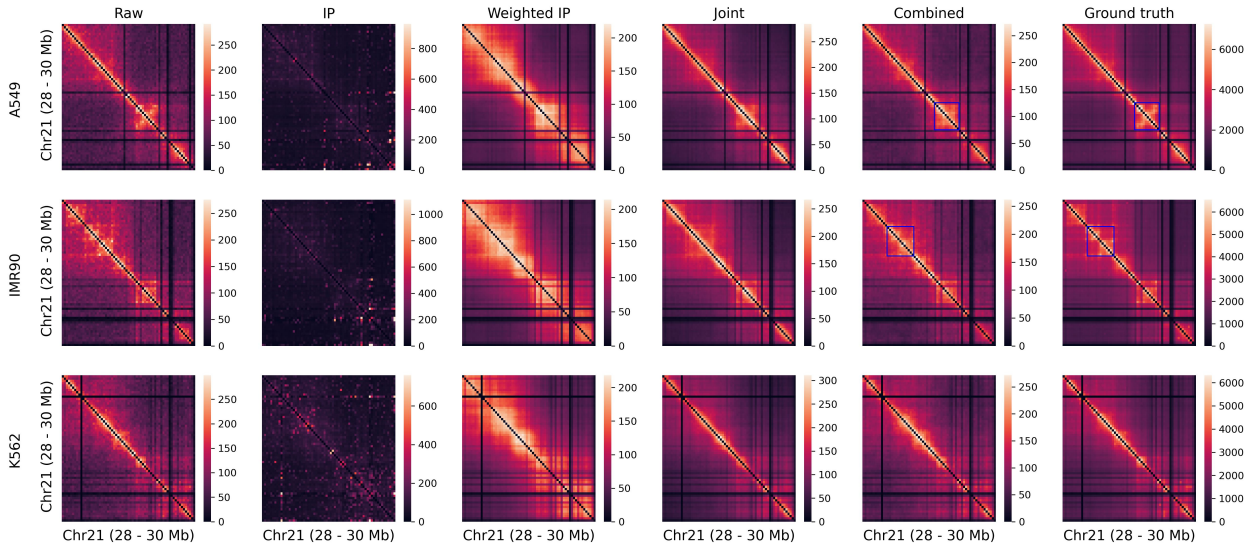


Figure S2: Heatmap visualization of pseudo-bulked contact maps generated from downsampled raw data, contactVI-imputed data using different decoder types (genomic bin feature type: self; normalization type: whole), and the original data. Each pseudo-bulk map is created by aggregating 1000 cells from each cell line in the Bintu2018 dataset.

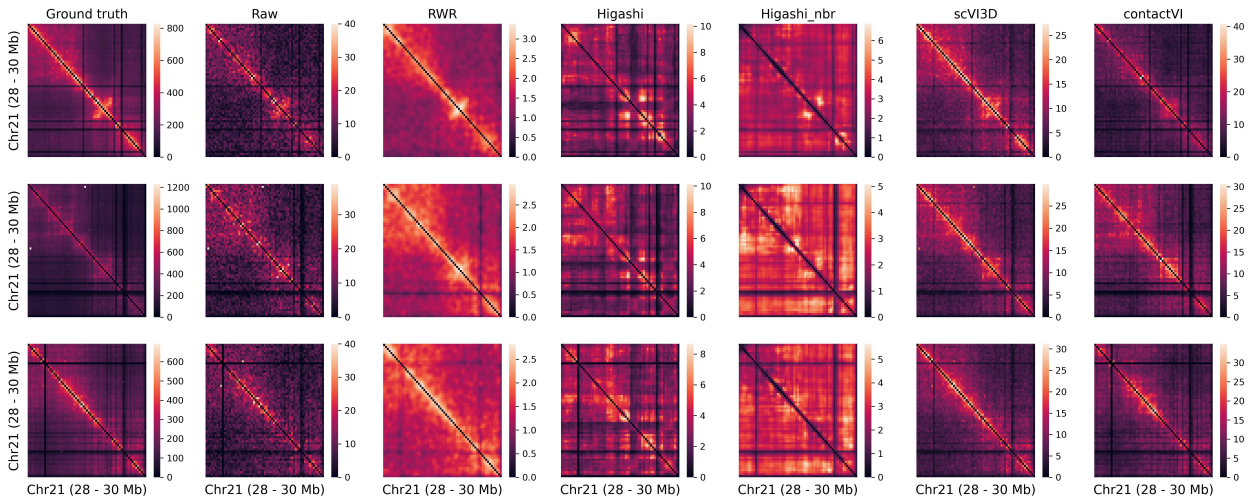


Figure S3: Heatmap visualization of pseudo-bulked contact maps generated from the down-sampled raw data, the original data, and the imputed data, by aggregating 100 cells from each cell line.

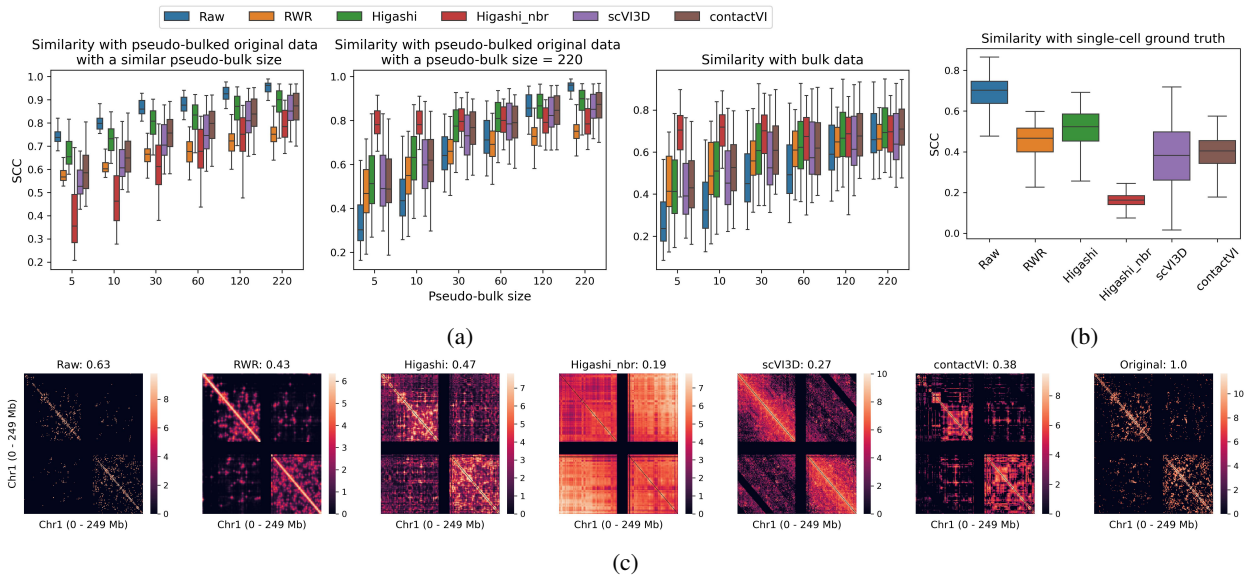


Figure S4: (a) Comparison of imputation methods based on similarity of the pseudo-bulked imputed contact maps with (left) the pseudo-bulked original contact map (with the same pseudo-bulk size), (middle) the pseudo-bulked original contact map (with pseudo-bulk size = 220), (right) bulk contact map, measured using the stratum-adjusted correlation coefficient (SCC). Pseudo-bulked and bulk contact maps are from the GM12878 cell line. (b) Comparison of imputation methods based on similarity of the single-cell imputed contact maps with the single-cell original contact map. (c) Heatmap visualization of single-cell contact maps generated from the down-sampled raw data, the original data, and the imputed data (cell: GM12878_cell_042).

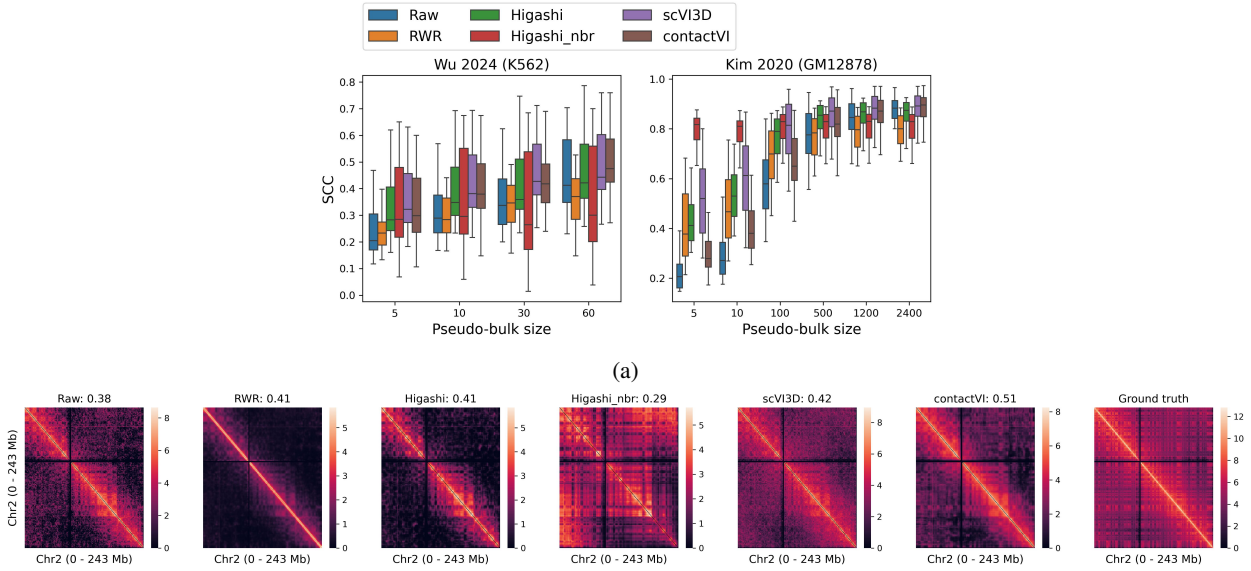


Figure S5: (a) Comparison of imputation methods based on the similarity between pseudo-bulked imputed contact maps and the ground-truth bulk contact map, measured using the stratum-adjusted correlation coefficient (SCC). (b) Heatmap visualization of pseudo-bulked contact maps generated from raw and imputed data by aggregating 63 K562 cells, alongside the bulk K562 contact map (ground truth).

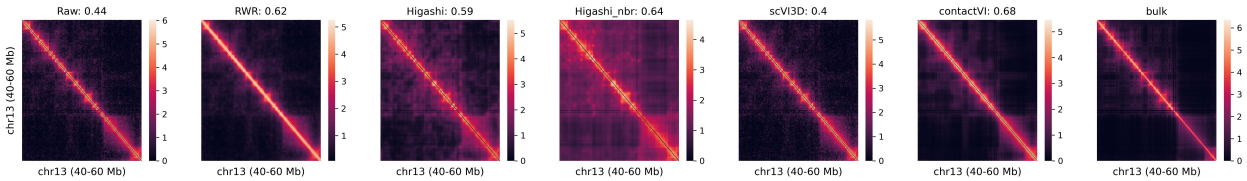


Figure S6: Heatmap visualization of pseudo-bulked contact maps generated from the raw and the imputed data by aggregating 220 GM12878 cells from $Wu2024$ dataset and the corresponding bulk data at 100 Kb resolution.

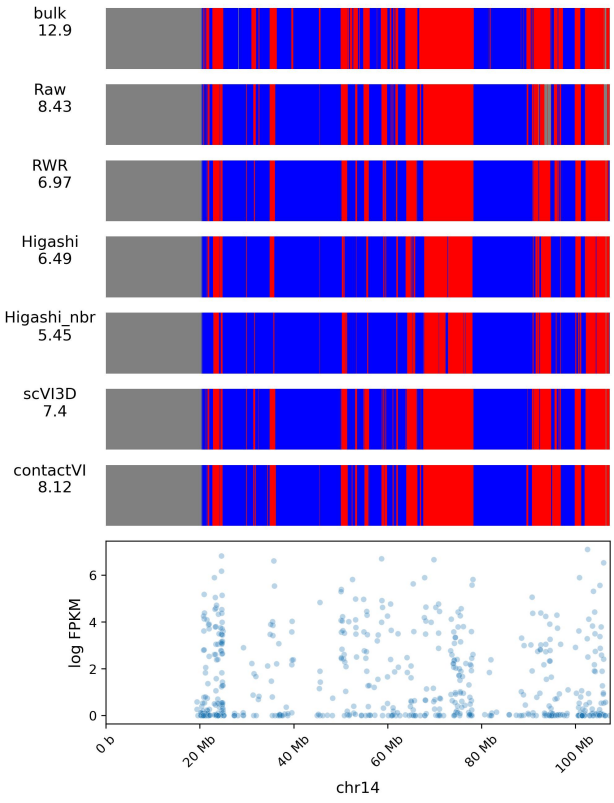


Figure S7: Example of compartment annotations based on pseudo-bulked raw, pseudo-bulked imputed, and bulk data. The bottom panel shows the normalized gene expression across the chromosome. The y-axis label of each compartment annotation panel indicates the data type and the t-statistic from comparing the normalized expression of genes in A vs. B compartments, as defined by that annotation.