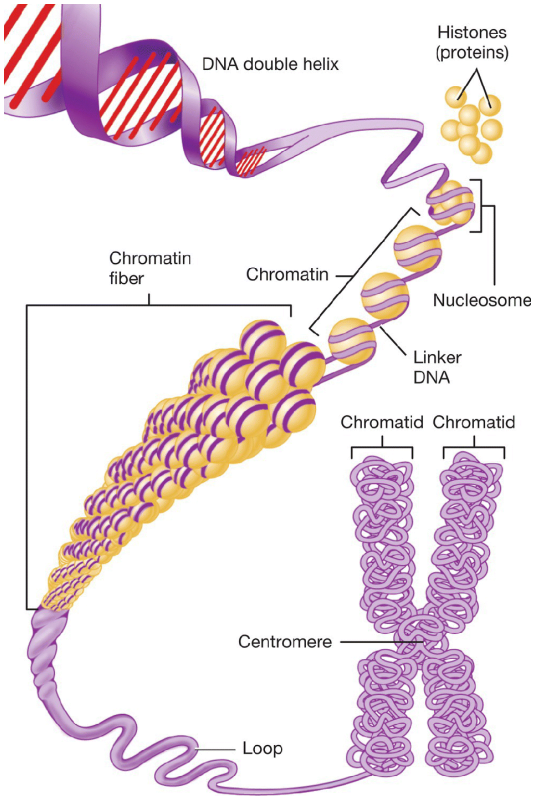
Smooth Zero-Inflated Modeling on Counting Tensors

Elena Tuzhilina
Department of Statistical Sciences, University of Toronto
elena.tuzhilina@utoronto.ca

JSM 2026, Boston, USA — August 2026

Conformation Reconstruction

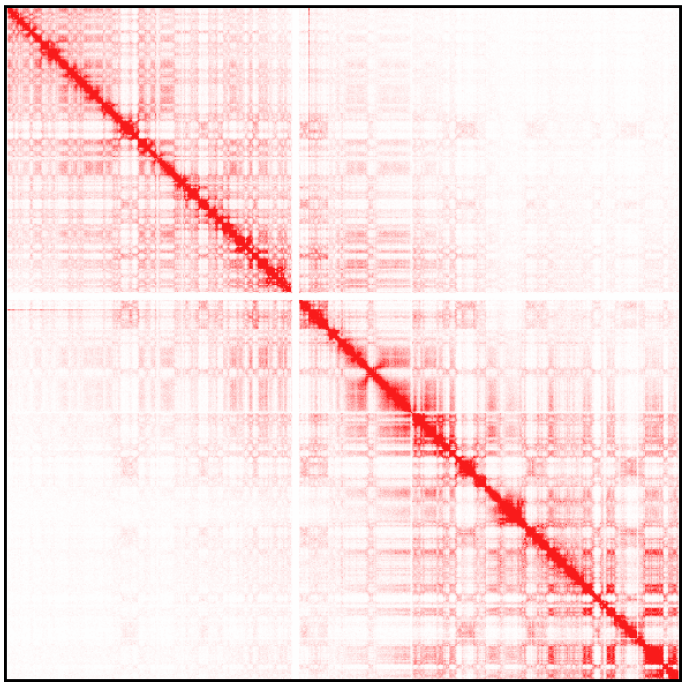
3D Conformation



source

Hi-C $\longrightarrow$

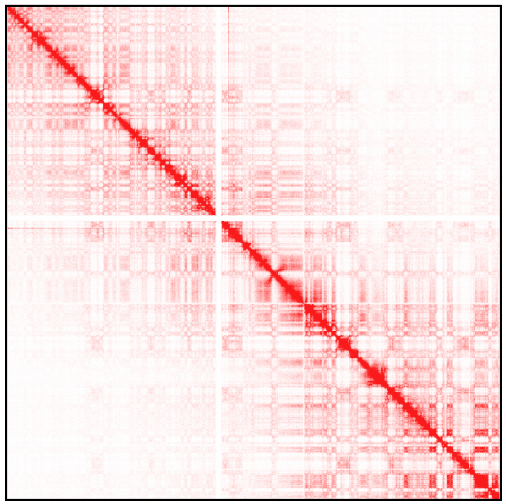
Contact Matrix



Let N be the number of genomic loci. The **contact matrix** $C \in \mathbb{Z}_+^{N \times N}$ has entries C_{ij} representing interaction frequencies. High values indicate spatial proximity.

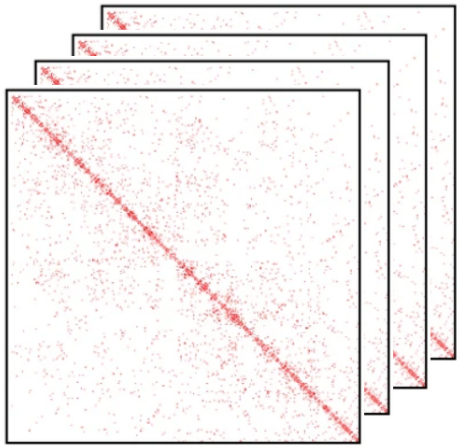
Bulk vs Single-Cell Hi-C

Bulk Hi-C



- Aggregates many cells into one matrix
- Loses cell-to-cell variability

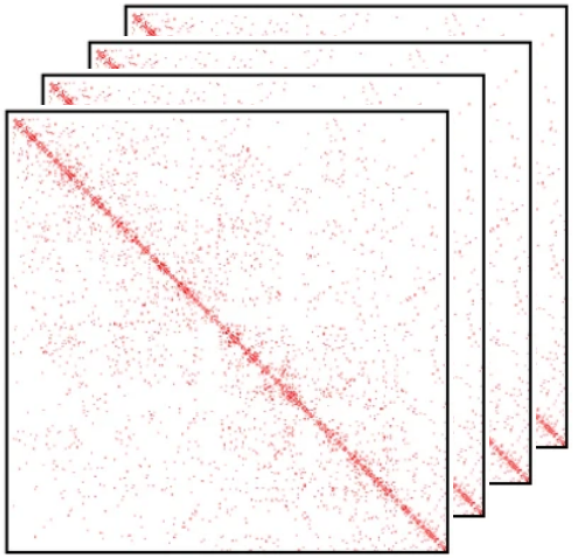
Single-Cell Hi-C



- Measures contacts in individual cells
- Captures heterogeneity
- Extremely sparse and noisy

Single-Cell Hi-C Structure

Single-cell Hi-C data form a tensor $\mathcal{C} \in \mathbb{Z}_{+}^{N \times N \times K}$, where C_{ijk} denotes the contact frequency between loci i and j in cell k.



- **True zeros**: no physical contact
- **Dropouts**: missed contacts

Which observed zeros are false?

N = number of genomic loci

K = number of cells

Model: Sparsity

Entries of $\mathcal{C}$ follow a zero-inflated Poisson model:

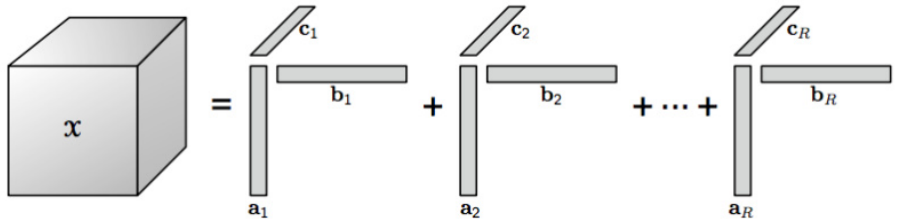
$$C_{ijk} \sim \mathrm{ZIP}(p_{ijk}, \lambda_{ijk})$$

After proper transformations,

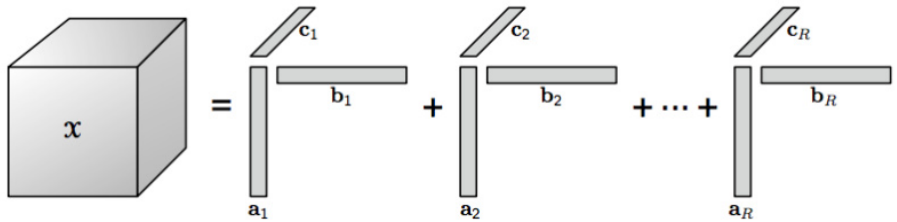
$$\eta_{ijk} = \log \lambda_{ijk}, \qquad \theta_{ijk} = \mathrm{logit}(1 - p_{ijk}),$$

the parameter tensors admit low-rank CP decompositions:

$$\eta = \mathcal{I} \times_1 \alpha \times_2 \alpha \times_3 \beta$$



$$\Theta = \mathcal{I} \times_1 \alpha \times_2 \alpha \times_3 \xi$$



Latent Embeddings

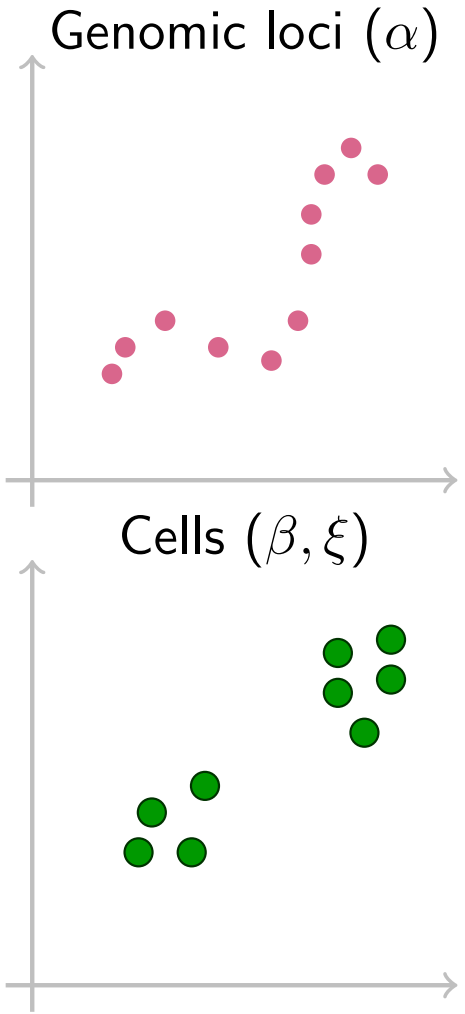
Parameter Tensors

$$\eta = \mathcal{I} \times_1 \alpha \times_2 \alpha \times_3 \beta$$

$$\Theta = \mathcal{I} \times_1 \alpha \times_2 \alpha \times_3 \xi \quad \Longrightarrow$$

- α: genomic locus embeddings
- β, ξ: cell embeddings

Latent Embedding Space



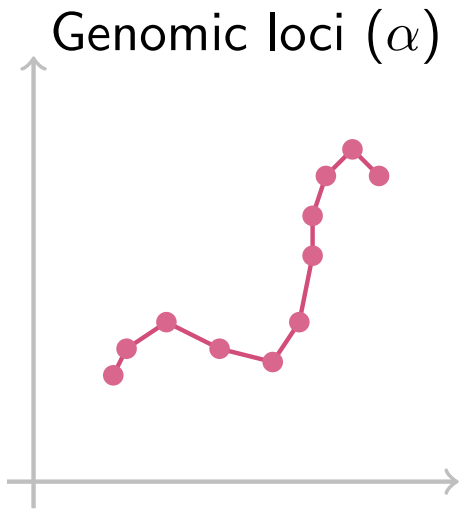
Model: Smoothness

Nearby genomic loci have similar latent representations, so the embedding varies smoothly along genomic position.

Curve representation

$$\alpha = \begin{pmatrix} \gamma(1) \\ \dots \\ \gamma(N) \end{pmatrix}, \quad \gamma(t) \in \mathbb{R}^L$$

Each genomic position corresponds to a point on a smooth latent curve.



Model: Smoothness

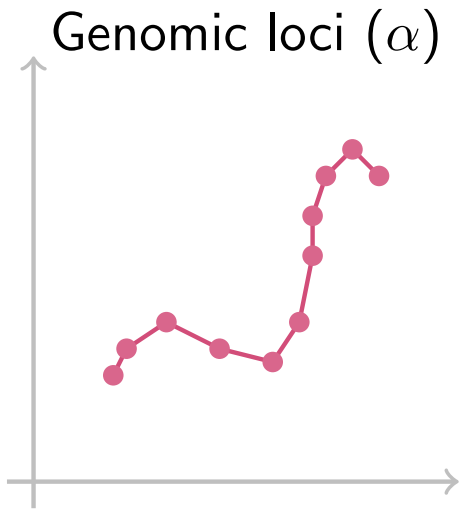
Nearby genomic loci have similar latent representations, so the embedding varies smoothly along genomic position.

Spline parameterization

$$\text{curve } \gamma(t) = (\gamma_1(t), \ldots, \gamma_L(t))$$

$$\text{spline basis } h_1(t), \ldots, h_Q(t)$$

$$\gamma_j(t) = \sum_{q=1}^{Q} \delta_{qj} h_q(t), \quad j = 1, \ldots, L$$



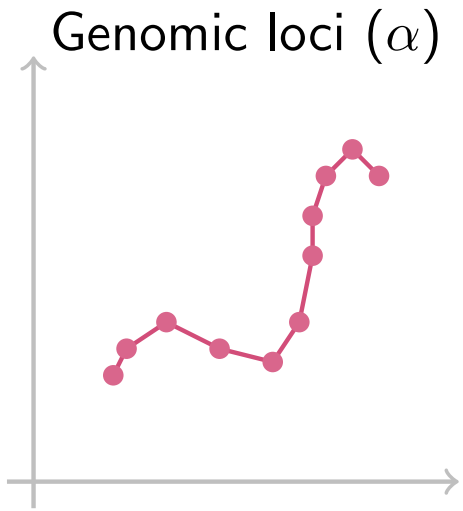
Model: Smoothness

Nearby genomic loci have similar latent representations, so the embedding varies smoothly along genomic position.

Matrix form

$$\alpha = H\Delta$$

- $H \in \mathbb{R}^{N \times Q}$: spline basis matrix
- $\Delta \in \mathbb{R}^{Q \times L}$: coefficient matrix



Zero-Inflated Tensor model with Smoothing (ZITS)

Likelihood model (zero-inflated counts)

$$C_{ijk} \sim \mathrm{ZIP}(p_{ijk}, \lambda_{ijk})$$

Link function

$$\eta_{ijk} = \log \lambda_{ijk}, \qquad \theta_{ijk} = \mathrm{logit}(1 - p_{ijk})$$

Low-rank decomposition + smoothness

$$\eta = \mathcal{I} \times_1 (H\Delta) \times_2 (H\Delta) \times_3 \beta$$

$$\Theta = \mathcal{I} \times_1 (H\Delta) \times_2 (H\Delta) \times_3 \xi$$

Parameters

spline coefficients $\Delta \in \mathbb{R}^{Q \times L}$, cell embeddings $\beta, \xi \in \mathbb{R}^{K \times L}$

Hyperparameters

embedding dimension L, spline degrees-of-freedom Q

False-Zero Detection and Imputation

Detection: an observed zero is classified as false if

$$\widehat{p}_{ijk} > \frac{1}{e^{\widehat{\lambda}_{ijk}} - 1}.$$

Imputation: false zeros are replaced with

$$C_{ijk} \leftarrow \widehat{\lambda}_{ijk}.$$

Zeros with high dropout probability but non-negligible intensity are treated as missing.

Real Data Analysis

Dataset Single-cell Hi-C data from Ramani et al. (2017, *Nature*) at 10 Mb resolution on chromosomes 1–4.

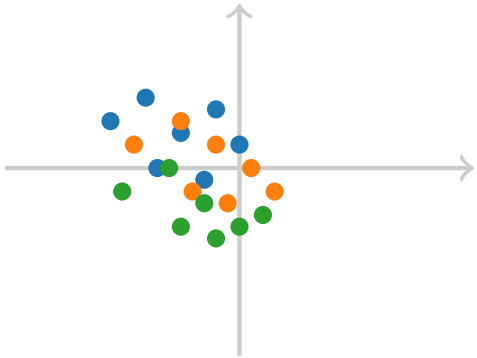
Cell types *HAP1*, *HeLa*, and *K562*

Data characteristics

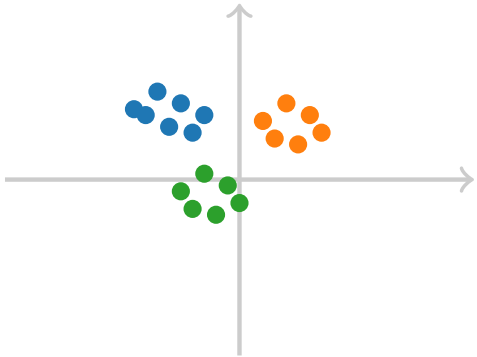
- Cells: $K = 144$ (48 per cell type)
- Genomic bins: $N \approx 20$–25
- Sparsity: $\approx 80\%$ zeros

Does imputation improve the separability of the three cell types?

Before imputation

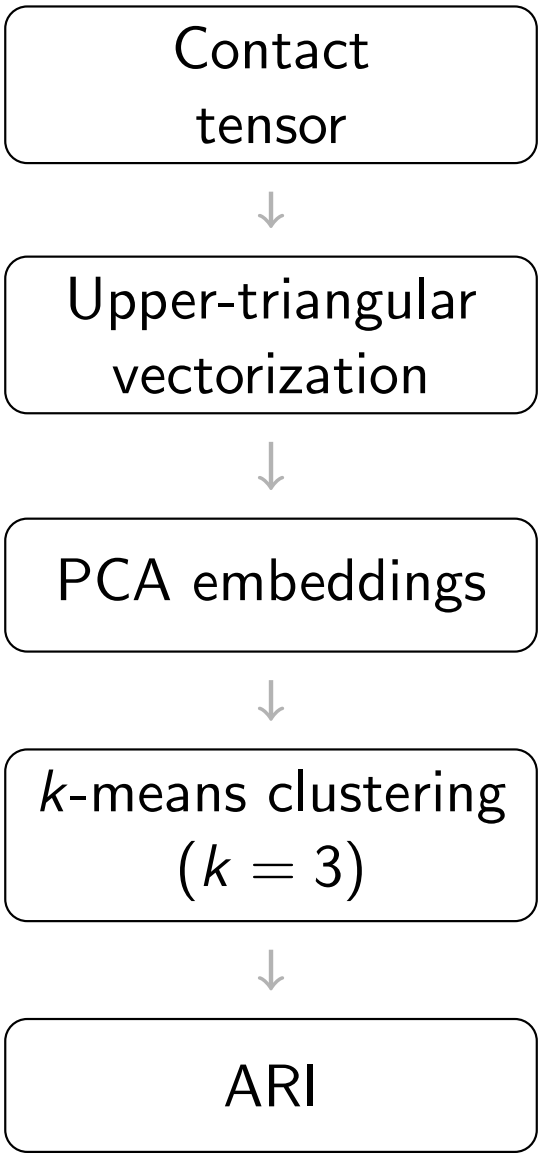


After imputation

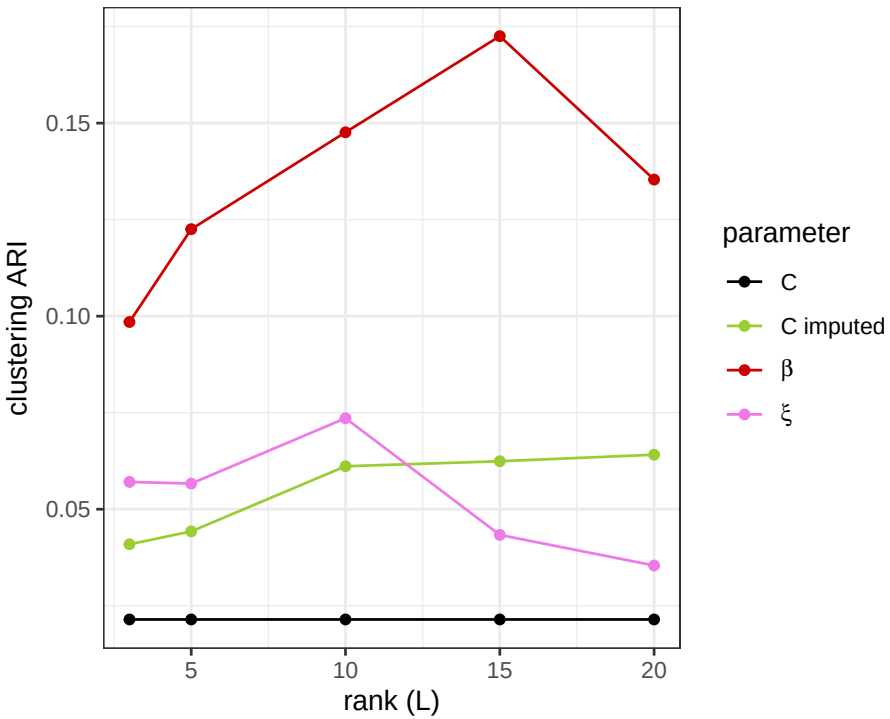


HAP1 HeLa K562

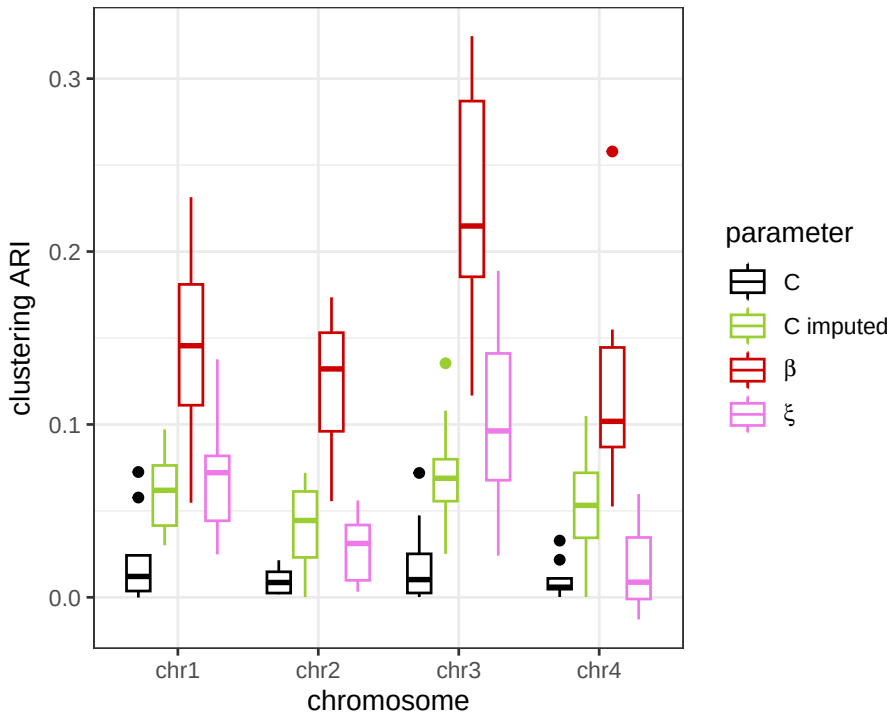
Real Data Analysis: Model Evaluation



Real Data Analysis: Imputation Evaluation



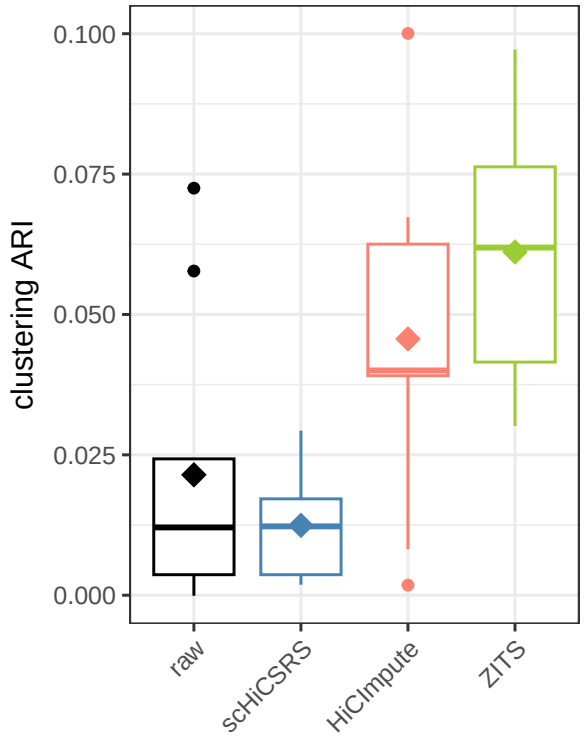
ARI vs embedding dimension L (Chr 1)



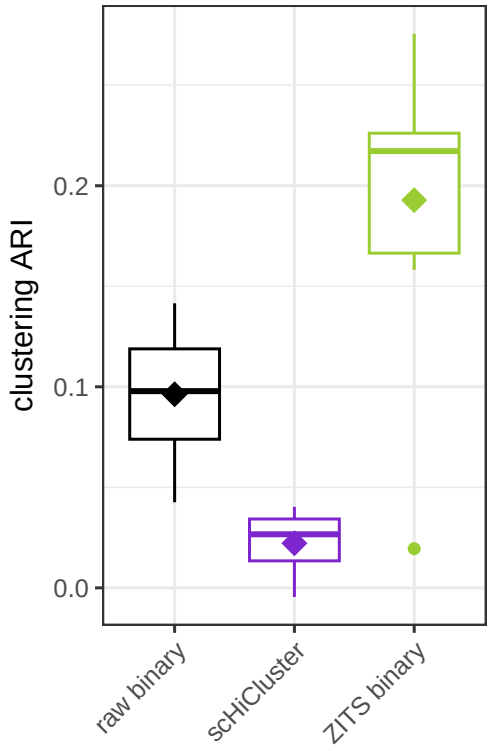
ARI across chromosomes ($L = 10$)

- Imputation improves separability of cell-types
- CP embeddings β capture cell-type structure effectively
- Performance is stable across chromosomes

Real Data Analysis: Competitors



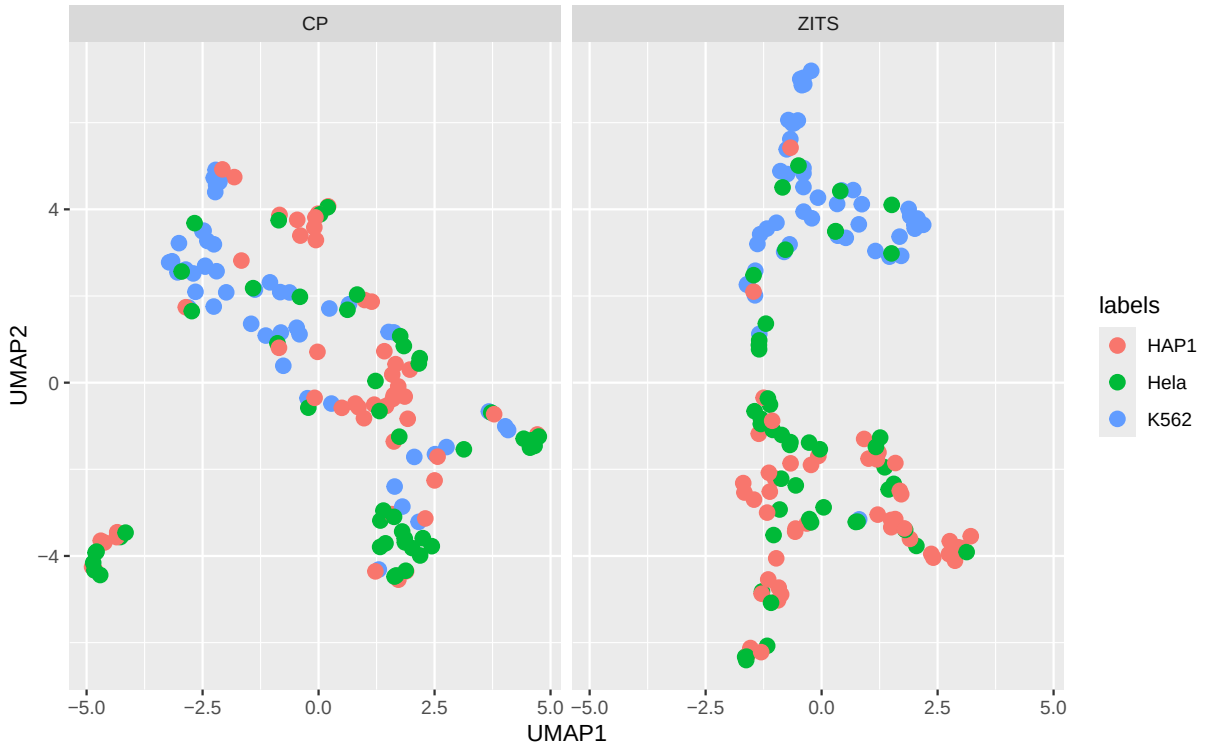
Count-based imputation methods



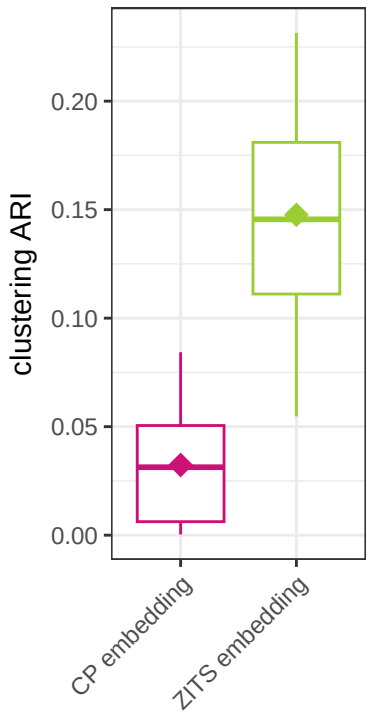
Binary contact imputation methods

ZITS achieves consistently higher ARI across settings with significant improvement in binary case.

Real Data Analysis: Embeddings Comparison



CP on observed tensor vs. ZITS embedding β



Cell-type separation

Joint modeling of counts and zero inflation yields more informative cell embeddings.

Conclusions

We proposed `ZITS`, a smooth zero-inflated tensor model for single-cell Hi-C data.

- Models dropout events using a zero-inflated Poisson likelihood.
- Learns cell embeddings via CP tensor decomposition.
- Captures genomic locus structure through smooth spline-based embeddings.

Imputation is performed using a Bayes-optimal decision rule under the fitted model.

Joint modeling of counts, dropout events, and genomic smoothness improves imputation quality.

Acknowledgements

NSERC CRSNG

McLaughlin Centre
UNIVERSITY OF TORONTO

UNIVERSITY OF TORONTO | DATA SCIENCES INSTITUTE

Yaoming Zhen
School of Data Science
The Chinese University of Hong Kong, Shenzhen

Thank You

Questions?

Elena Tuzhilina

Department of Statistical Sciences

University of Toronto

elena.tuzhilina@utoronto.ca