# HiC2Self: SELF-SUPERVISED HI-C CONTACT MAP DENOISING

Rui Yang[1,2], Alireza Karbalayghareh[1], Christina Leslie[1]

[1] Memorial Sloan-Kettering Cancer Center
[2] Tri-Institutional Computational Biology Program

Memorial Sloan Kettering Cancer Center

## Overview

**Background** Hi-C is a genome-wide chromatin conformation capture assay that is used to study 3D genomic organization. High-resolution Hi-C conact maps require generation of multiple replicate libraries and extremely high sequencing coverage (1-2B reads), incurring considerable costs.

**Aim** Develop a self-supervised Hi-C contact map denoising method

**Method** HiC2Self – a self-supervised neural network for denoising Hi-C contact maps that only needs low-coverage data, and could directly process low coverage raw count matrices without additional normalization steps.

**Application** We implemented HiC2Self on a single low-coverage library of GM12878 (GSE63625), and recovered Chromosome 18 for down-stream analysis.

## Self-Supervision Framework

**Noise2Self** (Batson & Royer, 2019) is a self-supervised denoising framework that uses $\mathcal{J}$-invariant functions $f$, where $\mathcal{J}$ represents a partition of the input data dimensions $m$ into subsets, and we consider a subset $J \in \mathcal{J}$ and its complement $J^C$.

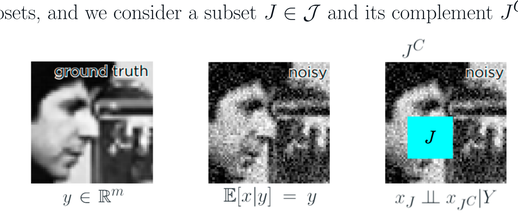


Fig. 1: Input dimension partition and assumptions

Two assumptions:

- Zero-mean noise: $\mathbb{E}[x|y] = y$
- Conditional independence: $x_J \perp\!\!\!\perp x_{J^C}|Y$

$\mathcal{J}$-invariant function $f : \mathbb{R}^m \to \mathbb{R}^m$, $f(x)_J \perp\!\!\!\perp x_J$ for every $J \in \mathcal{J}$.
With this $\mathcal{J}$-invariant function $f$, the ordinary denoising loss function defined as

$$\mathcal{L}_f = \mathbb{E}_{x,y}||f(x) - y||^2 = \mathbb{E}_x||f(x) - x||^2 + \mathbb{E}||x - y||^2 - 2\mathbb{E}\langle\mathbb{E}(f(x) - x), \mathbb{E}(x - y)\rangle$$

could be simplified to

$$\mathcal{L}(f) = \sum_{J\in\mathcal{J}} \mathbb{E}||f_J(x_{J^C}) - x_J||^2$$

**Mask design** The upper and lower tri-angle of Hi-C contact maps are symmetric. Masks are used for $\mathcal{J}$-invariant partition. To satisfy the independence assumption, we designed diagonal-symmetric masks.

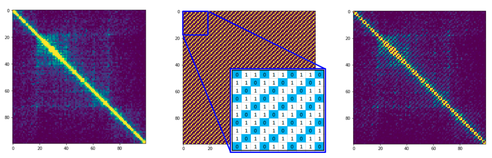

Fig. 2: Mask designed for Hi-C contact map

## Proposed Method

**HiC2Self**

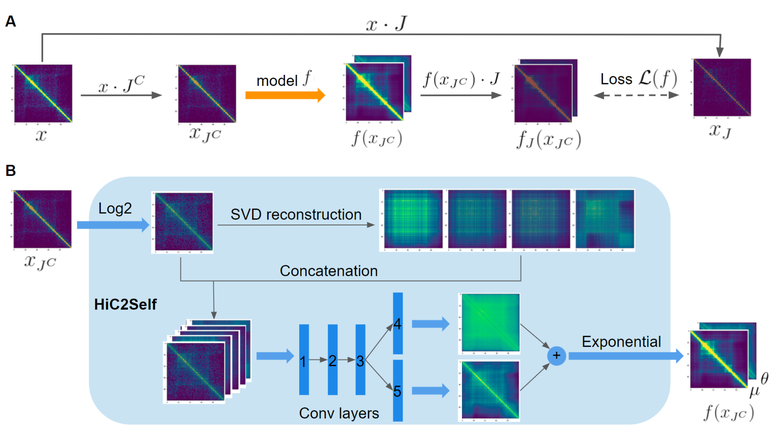


Fig. 3: HiC2Self Framework

**Key points:**

- **Framework**: Self-supervision framework (Fig 3 A) with designed masks
- **Model**: Convolutional network with top-4 SVD reconstructed channels (Fig 3 B)
- **Loss function**: Negative-binomial loss to predict values in raw counts scale

**Loss function** Inspired by DCA model (Eraslan et al. 2019), who uses a negative binomial loss for single cell data analysis. We assume counts from each bin ($x_{ij}$) of the contact map $X$ also follow a negative binomial distribution, $x_{ij} \sim NB(\mu_{ij}, \theta_{ij})$, and define log-likelihood loss function as

$$\mathcal{L}(f) = -logL_{NB} = \sum(log\Gamma(x+1) + log\Gamma(\theta) - log\Gamma(x+\theta) + \theta log(\frac{\mu+\theta}{\theta}) + xlog(\frac{\mu+\theta}{\mu}))$$

## HiC2Self denoising results

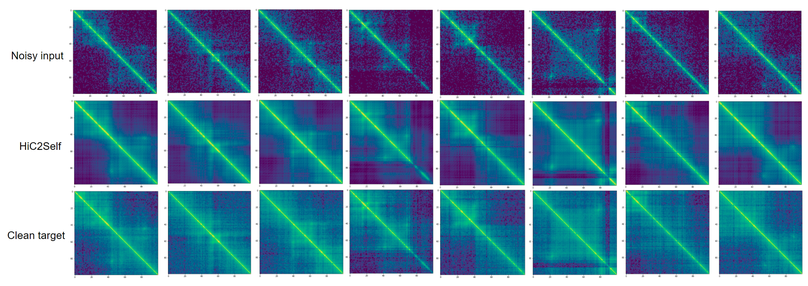


Fig. 4: HiC2Self denoising results

## Method comparison

**Compared with hicGAN** hicGAN (Liu et al. 2019) is a supervised Hi-C denoising method. Training and testing data are normalized as hicGAN paper suggested. HiC2Self was trained using MSE loss. Chromosome 3, 8, 12 were used as testing data, and all others as training.

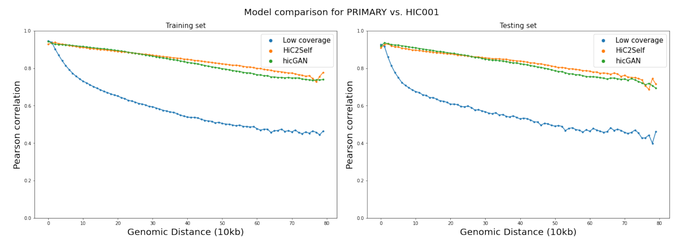


Fig. 5: Comparison with existing supervised method

## Whole Chromosome Prediction

**Genome-wide prediction assembly** HiC2Self produces denoised results as raw counts, which can easily be assembled into a whole-chromosome prediction (Fig 6). Whole genome predictons were assembled by averaging each sub-matrix along the diagonal, and further saved in .hic format for visualization using Juicebox (Robinson et al, 2018).

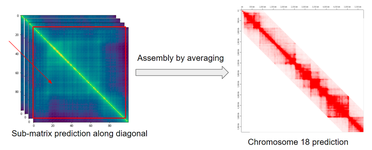


Fig. 6: Prediction assembly for an entire Chromosome

**Downstream analysis** We compared the assembled prediction for Chromosome 18 with high coverage library, along with another independent high-coverage library with 3B reads. Number pf significant 3D interactions called by HiC-DC+ (Sahin et al. in prep) is shown.

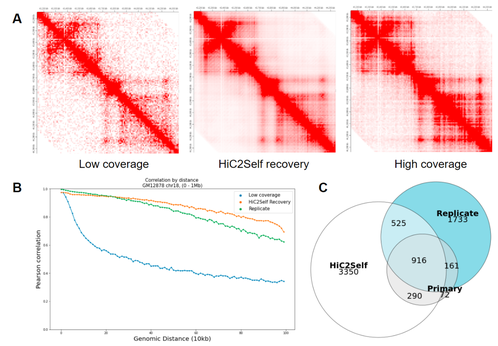


Fig. 7: Prediction assembly for an entire Chromosome