Amirreza Rouhi– Master student
Thesis submission: April 2020
Adv: Prof. Stefano Ceri

Ensemble Feature Selection for Single Cell Hi-c data

POLITECNICO
MILANO 1863

Introduction to single-cell Hi-C data

1. Introduction to single-cell Hi-C data

2-What is the Goal?

3. Proposed Method

4. Experimental Results

5- Future works

Hi-C → Chromatin conformation capture method

Hi-C data *is often used to analyze genome-wide chromatin organization*

Single-cell Hi-C is a modification of the original Hi-C protocol

Allows us to determine **proximity of different regions** of the genome in a single cell

Single-cell assays introduce a new axis of **variation— cell-to-cell variability**—that is not directly observable in data derived from a bulk sequencing .

1. Introduction to single-cell Hi-C data

2-What is the Goal?

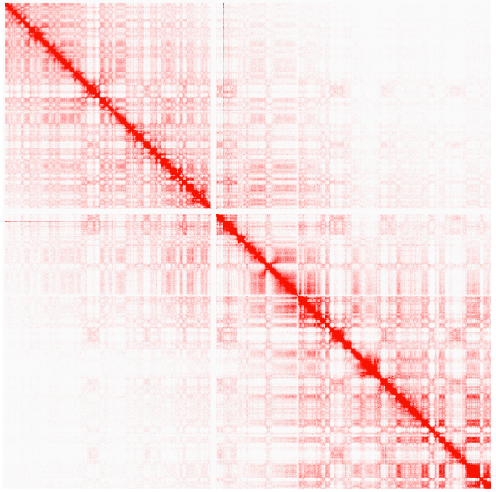
3. Proposed Method

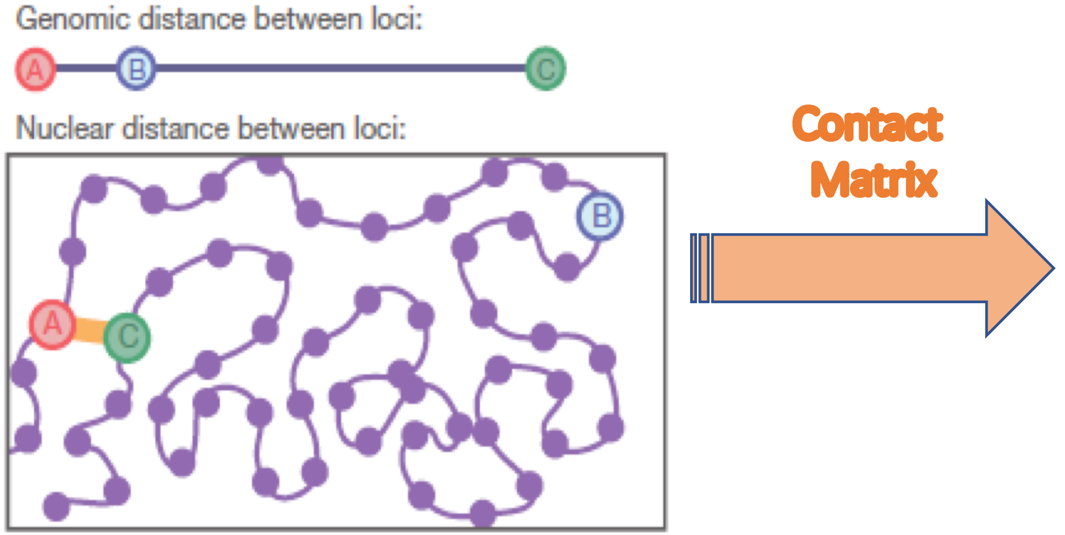
4. Experimental Results

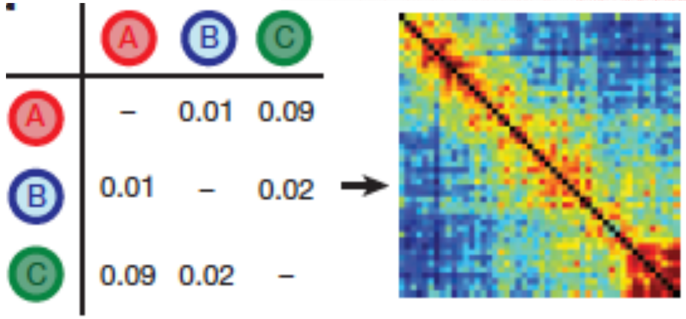
5- Future works

Single Cell Hi-C Data

Single Cell Hi-C data is represented as two-dimensional contact matrices




Genomic distance between loci:
A
B
C
Nuclear distance between loci:
A
C
B
Contact Matrix



	A	B	C
A	–	0.01	0.09
B	0.01	–	0.02
C	0.09	0.02	–

The sparsity of single-cell Hi-C data is higher than most other types of single-cell data.

Main Problem : The Sparsity of Single Cell Hi-C data

1. Introduction to single-cell Hi-C data

2-What is the Goal?

3. Proposed Method

4. Experimental Results

5- Future works

What is the Goal?

Design a Unique method to use in two different problems:

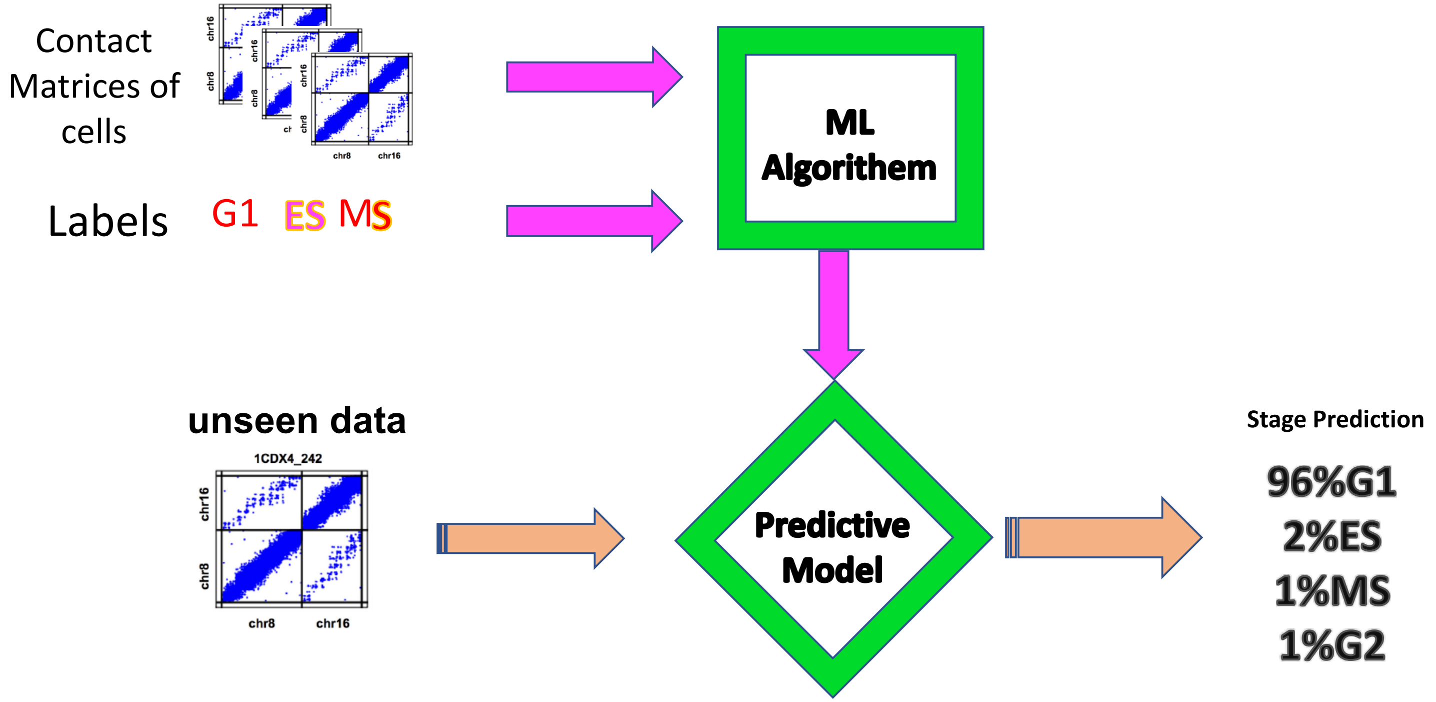
Determine the 1-Cell Cycle and 2-Cell line of each cell

Cell cycles:

- G1:Growth
- ES: Starting DNA Synthesis
- MS: Mid- DNA Synthesis
- Late S-G2 : Growth and prepreation for Mitosis & S to G2 transmission

Cell lines (Human) :

- HAP1
- HeLa
- K562
- GM12878



Contact Matrices of cells

Labels G1 ES MS

ML Algorithem

unseen data

Predictive Model

Stage Prediction

96%G1
2%ES
1%MS
1%G2

1. Introduction to single-cell Hi-C data

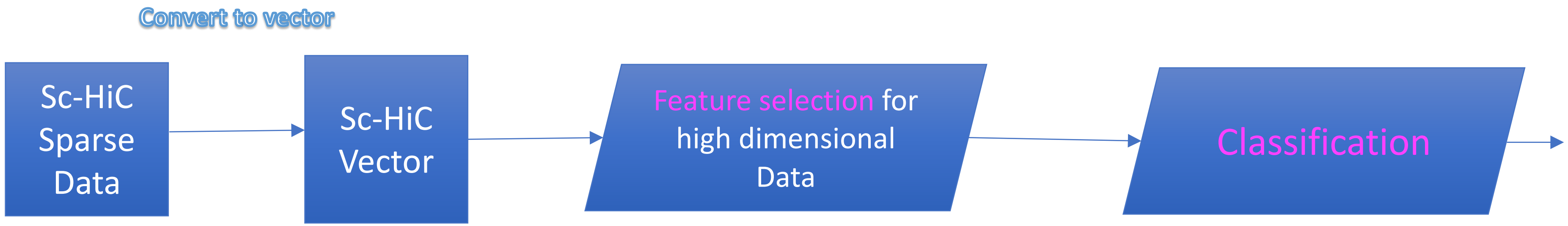
2-What is the Goal?

3. Proposed Method

4. Experimental Results

5- Future works

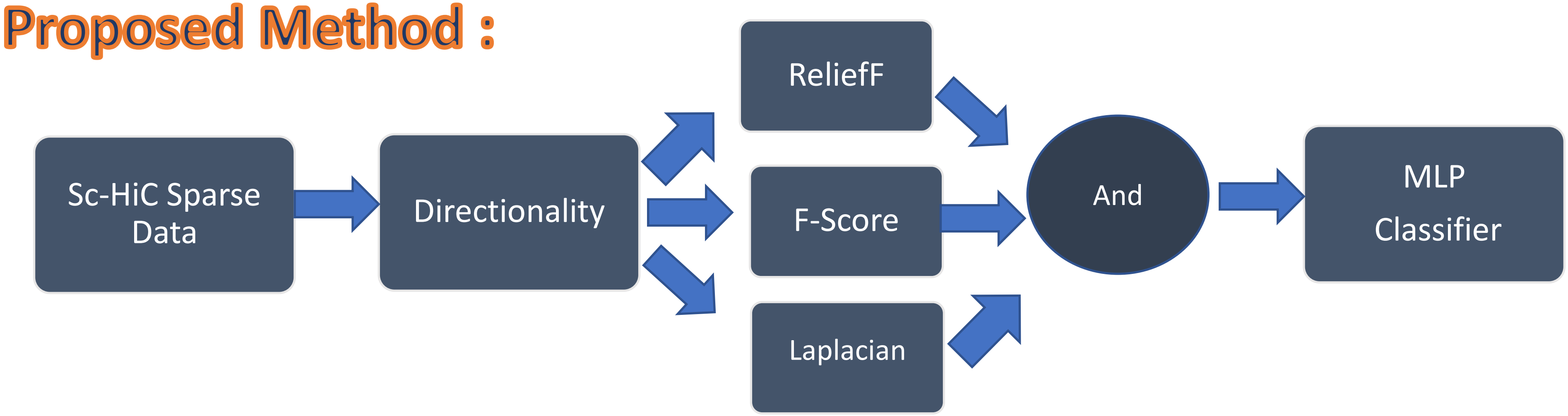
Proposed framework



Convert to vector

Sc-HiC Sparse Data → Sc-HiC Vector → Feature selection for high dimensional Data → Classification →

Proposed Method :



Sc-HiC Sparse Data → Directionality → ReliefF / F-Score / Laplacian → And → MLP Classifier

Directionality :

1. Introduction to single-cell Hi-C data

2-What is the Goal?

3. Proposed Method

4. Experimental Results

5- Future works

If in a sparse matrix X for diagonal cell (i,i) we have:

$$X(i,j) = \begin{array}{c} \\ \\ \\ X4 \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1\,\xleftarrow{A4} & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2\,\xrightarrow{B4} & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

Ai=$\sum_{j=n}^{i} X(i,j)$ *Summation of contacts between bin i and previous bins*

Bi=$\sum_{k=i}^{i+n} X(i,k)$ *Summation of contacts between bin i and next bins*

Where n is a boundaries variable. Then we can assign a score (Directionality Score) to each bin as:

Directionality_Score=$\frac{B-A}{|B-A|}$.

Now we were able to convert the matrix into vectors (for each column we have a score) .

1. Introduction to single-cell Hi-C data

2-What is the Goal?

3. Proposed Method

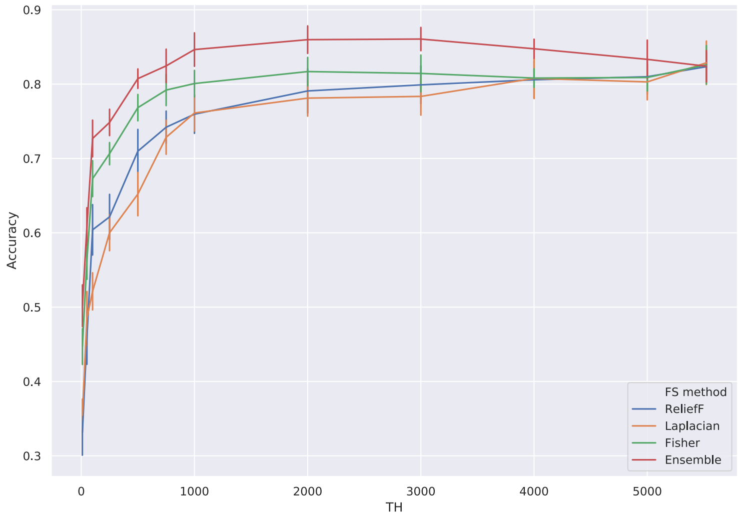
4. Experimental Results

5- Future works

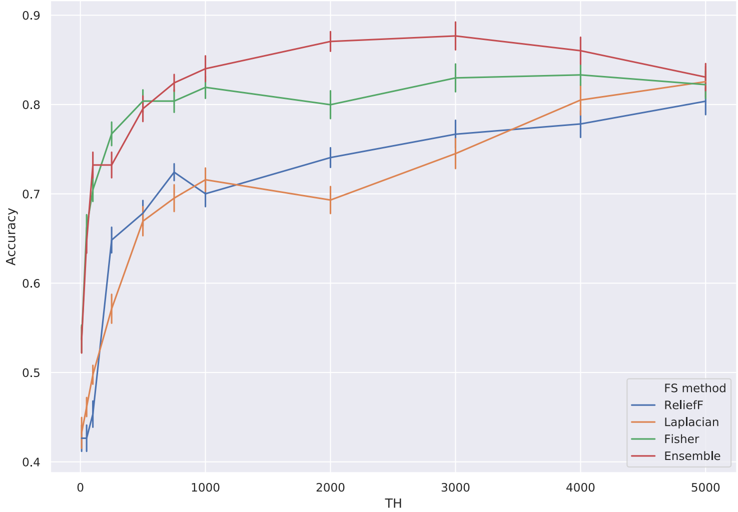
Experimental Results- Accuracy

Comparison between the accuracy rate of the proposed Ensemble method and single methods using MLP classifier (10 CV)

Cell Cycle



Cell Line



1. Introduction to single-cell Hi-C data

2-What is the Goal?

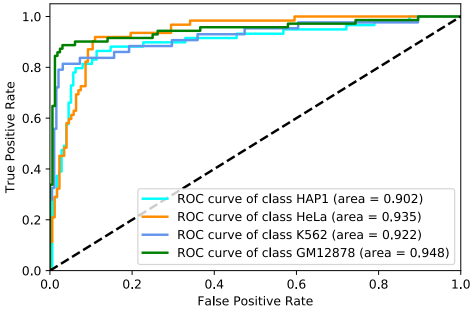
3. Proposed Method

4. Experimental Results

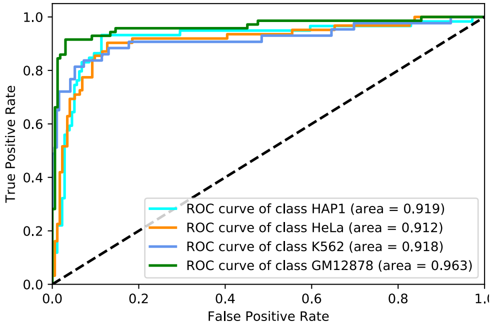
5- Future works

Experimental Results- ROC

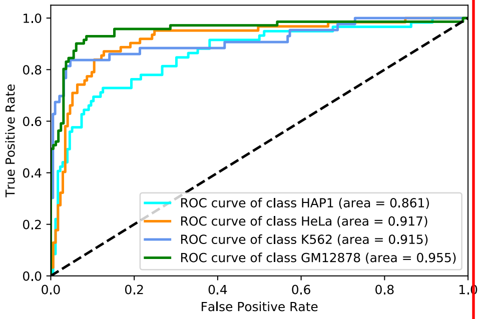
Cell Cycle:



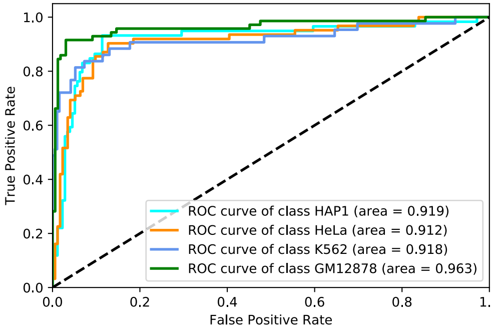
ROC Curve for applying Fscore method



ROC Curve for applying Laplacian score method

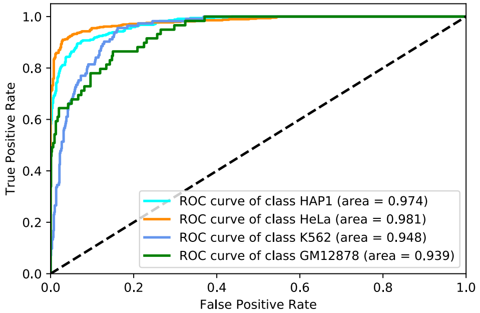


ROC Curve for applying ReliefF method

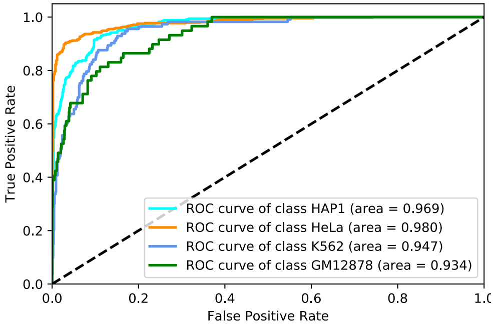


ROC Curve for applying Ensemble method

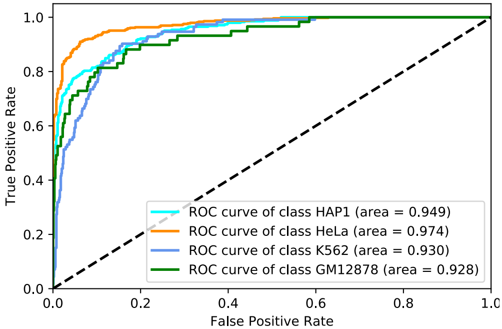
Cell Line:



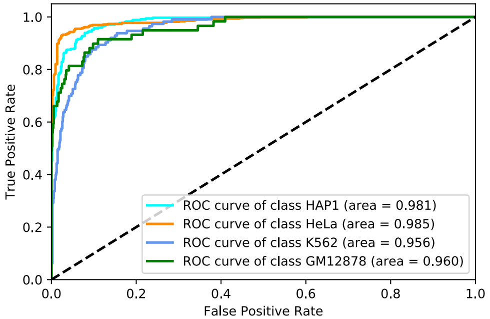
ROC Curve for applying Fscore method



ROC Curve for applying Laplacian score method



ROC Curve for applying ReliefF method



ROC Curve for applying Ensemble method

Experimental Results- Cell Cycle

1. Introduction to single-cell Hi-C data

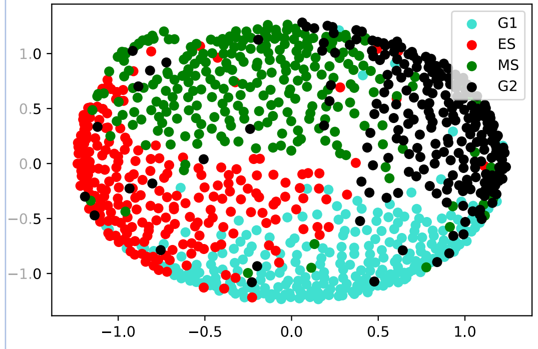
2-What is the Goal?

3. Proposed Method

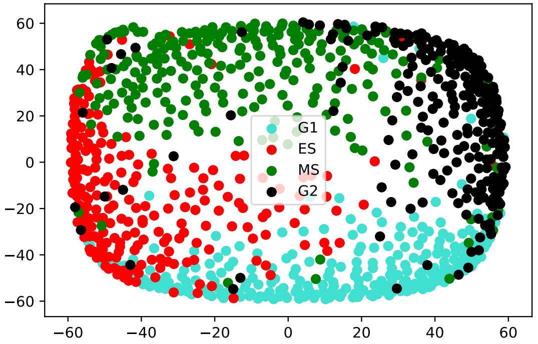
4. Experimental Results

5- Future works

MDS projections from the four cell-cycle phases when the Distance measure is calculated using cosine

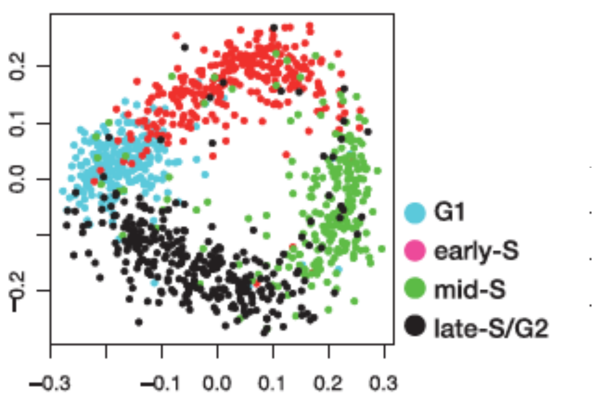


using Sigmoid

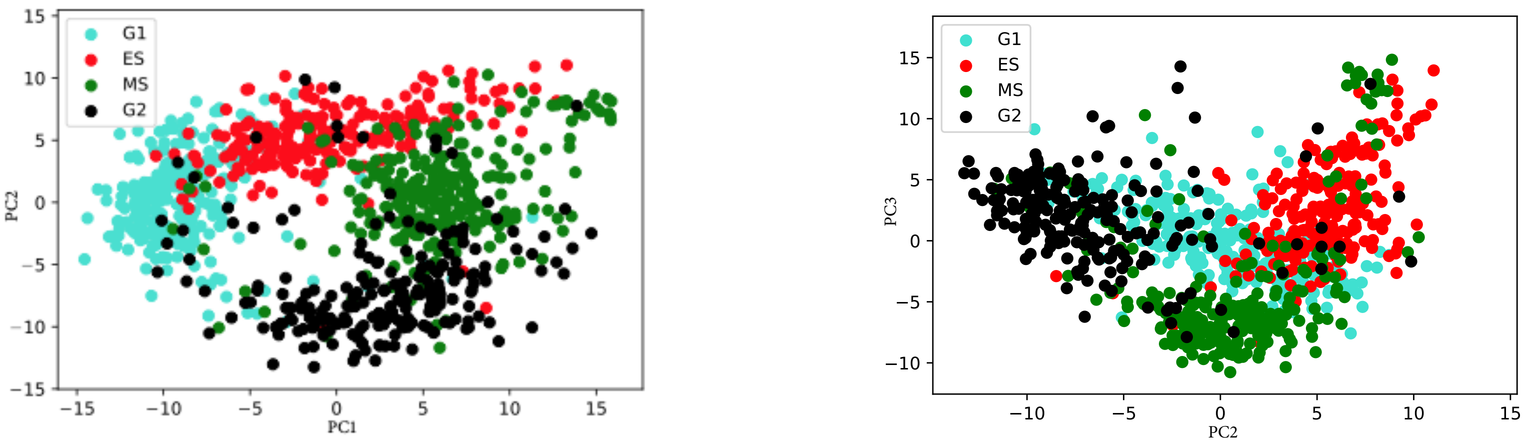


Compare with result from Li et al. :

Combined/MDS



PCA projections from the four cell-cycle phases :



1. Introduction to single-cell Hi-C data

2-What is the Goal?

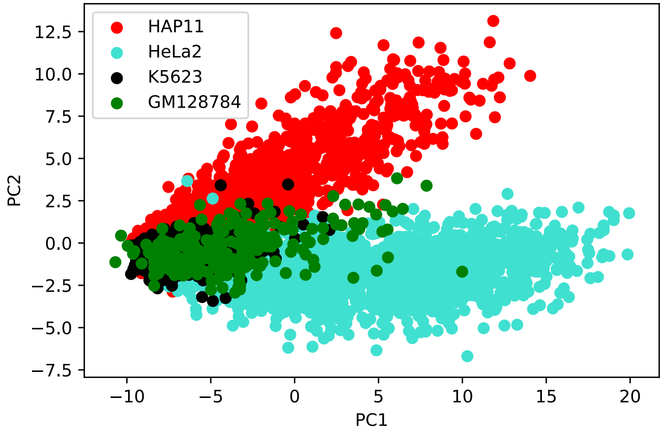
3. Proposed Method

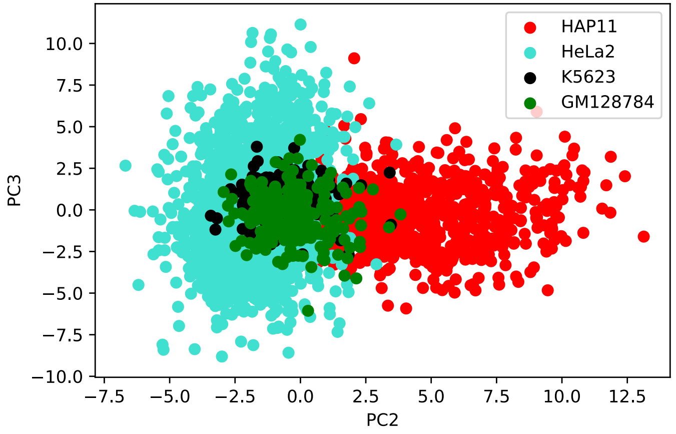
4. Experimental Results

5- Future works

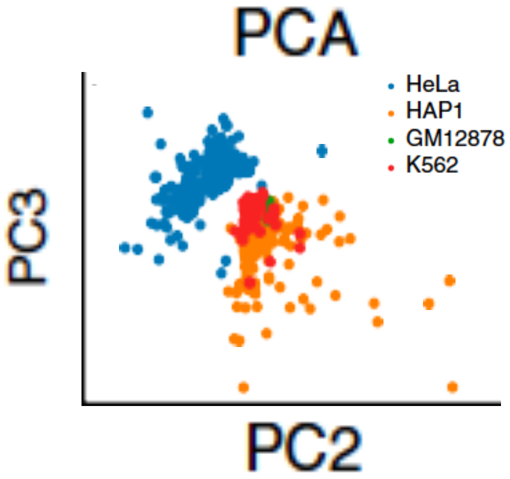
Experimental Results- Cell Line

PCA projections from the four lines





Compare with result from Zhou et al. :



1. Introduction to single-cell Hi-C data

2-What is the Goal?

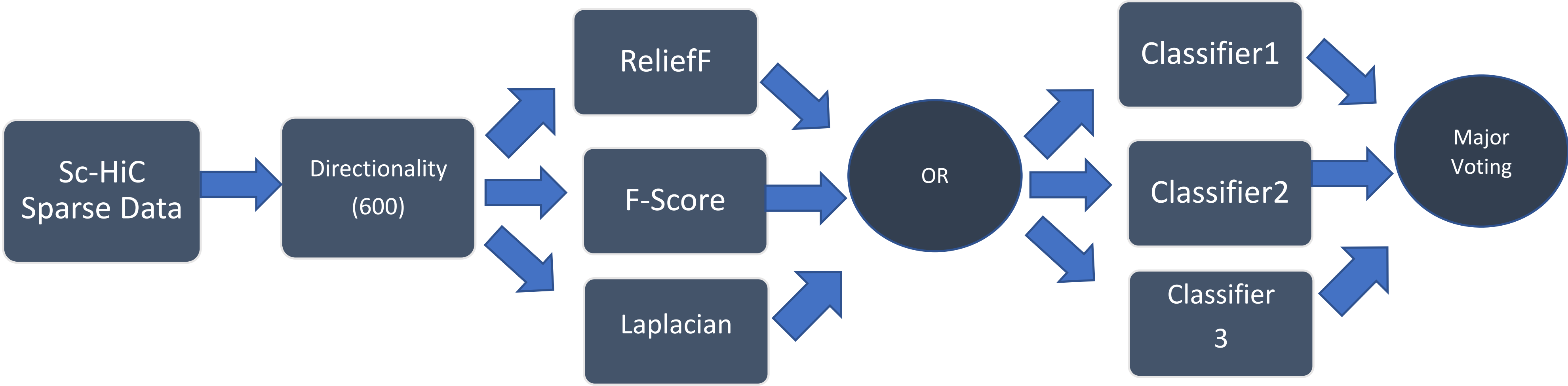
3. Proposed Method

4. Experimental Results

➡ 5- Future works

Future works

Applying Ensemble Classification



Sc-HiC Sparse Data

Directionality (600)

ReliefF

F-Score

Laplacian

OR

Classifier1

Classifier2

Classifier 3

Major Voting

Benefit:

*Increase the classification accuracy by using ensemble technique and aggregate the results

1. Introduction to single-cell Hi-C data

2-What is the Goal?

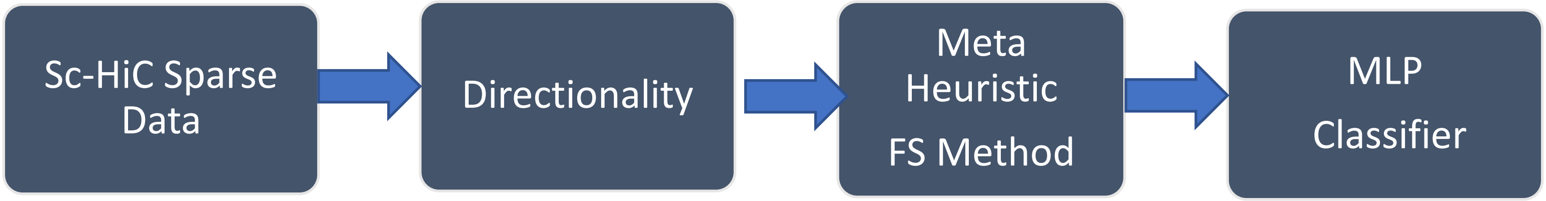
3. Proposed Method

4. Experimental Results

5- Future works

Future works

Applying meta-Heuristic Feature Selection methods



*Increase the performance of selecting the effective bins

Thank you for your attention