Machine learning approaches for comparative genome structure analysis

Carlos Rojas[1]*, Minh N. Tran[1]*, Linh Huynh[1], Fereydoun Hormozdiari[1,2,3]

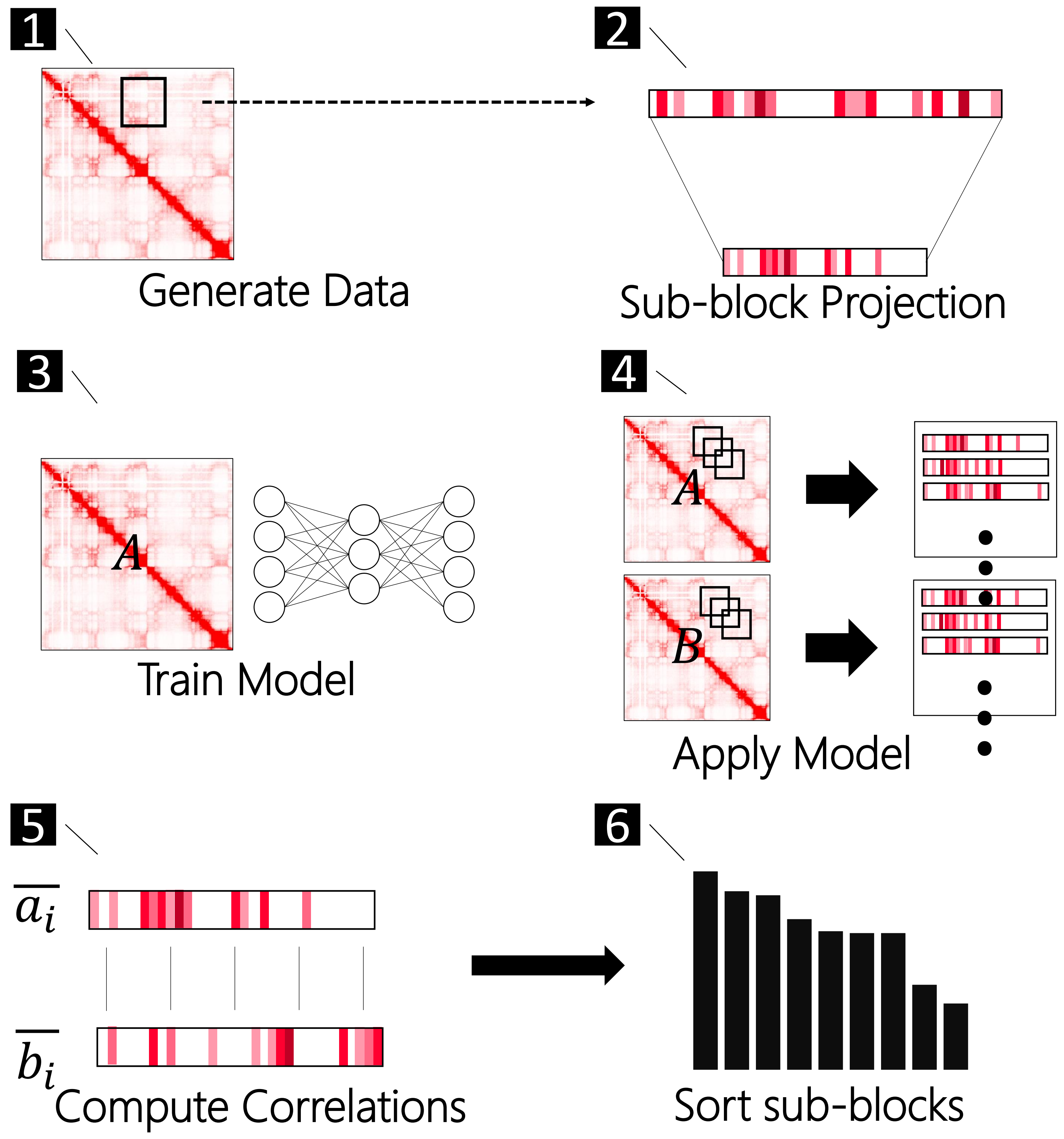
Problem: Comparing Hi-C matrices

High-throughput chromosome conformation capture techniques (Hi-C) has produced wealth data to study 3D genome. We investigated methods to find conserved or specific genomic interactions between two Hi-C contact matrices.

We compared Hi-C matrices with

- Principal Components Analysis (PCA)
- Root-mean-square error (RMS)
- Pearson Correlation
- Convolutional Autoencoder

Method: Convolutional Autoencoder



Generate Data

Sub-block Projection

Train Model

Apply Model

Compute Correlations

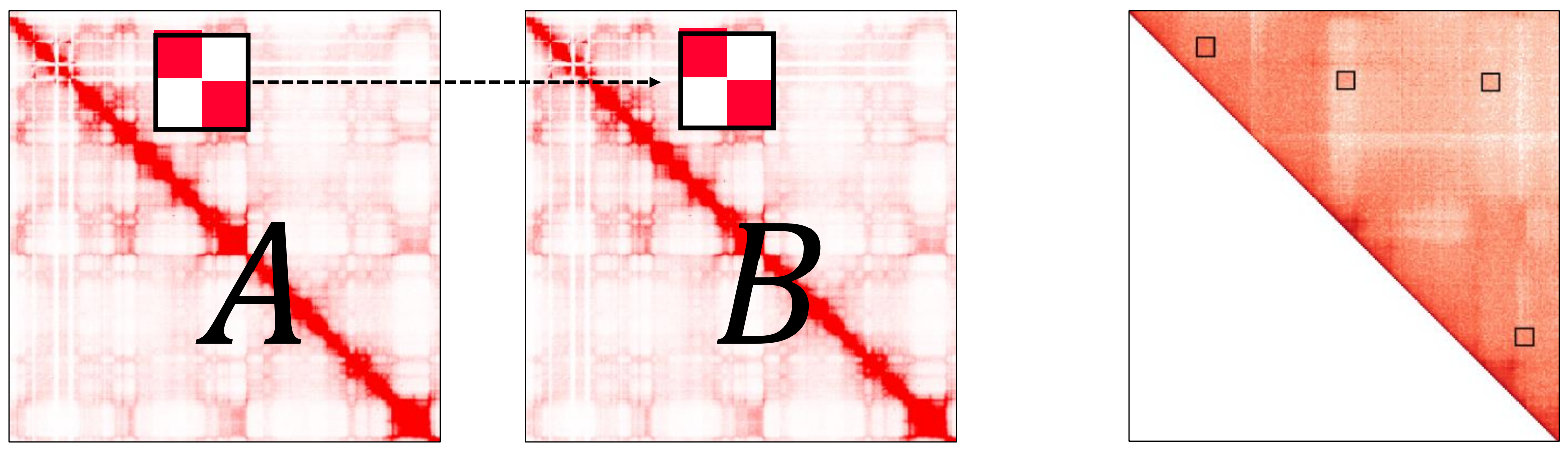
Sort sub-blocks

Algorithm

1. Generate $n \times n$ dimensional feature sub-blocks from a Hi-C matrix.
2. Project the $n \times n$ dimensional sub-blocks to a k dimensional vector $\overline{v}_i$.
3. Train the autoencoder model on the Hi-C matrix A.
4. Apply model on matrix A and B to produce projections.
5. Compute correlation between corresponding projections.
6. Sort the correlations and visualize the top ranked differences.

Data: Simulations

To evaluate the comparison between Hi-C matrices we simulated changes on a base matrix B. We simulated changes by copying areas from a matrix A. Next, we applied Gaussian noise to B. We show an actual example on the right figure. We varied the size of the changes between $10 \times 10, 24 \times 24, 32 \times 32, 48 \times 48$.

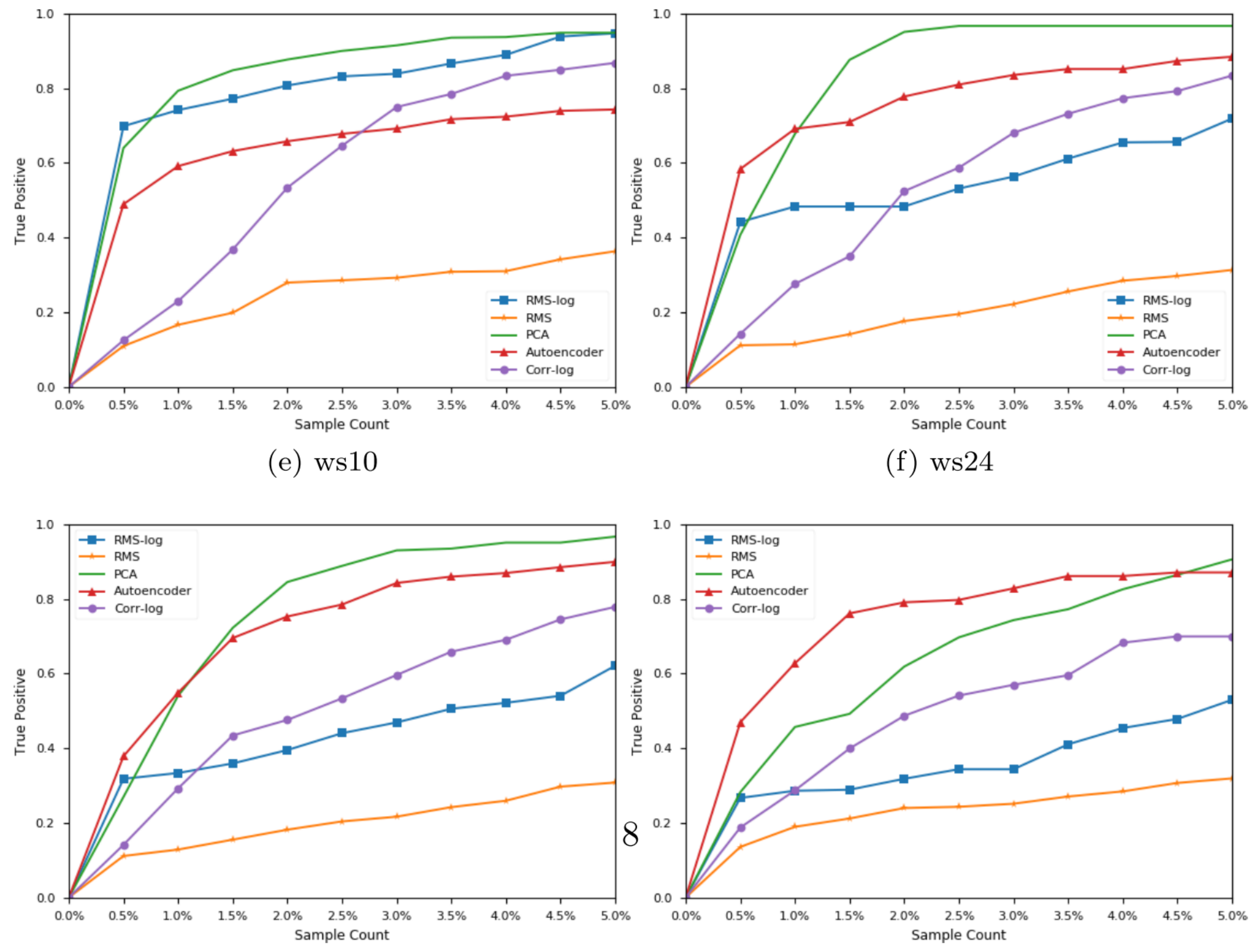


Simulated Changes

Example

Here we show a figure where the x-axis represents the total number of feature blocks and the y-axis represents the total number of changes correctly identified. We placed the simulated changes uniformly and off-diagonal of the matrices. We used the GM12878 and K562 datasets from Rao et al. [RHD 14].

Uniform Locations



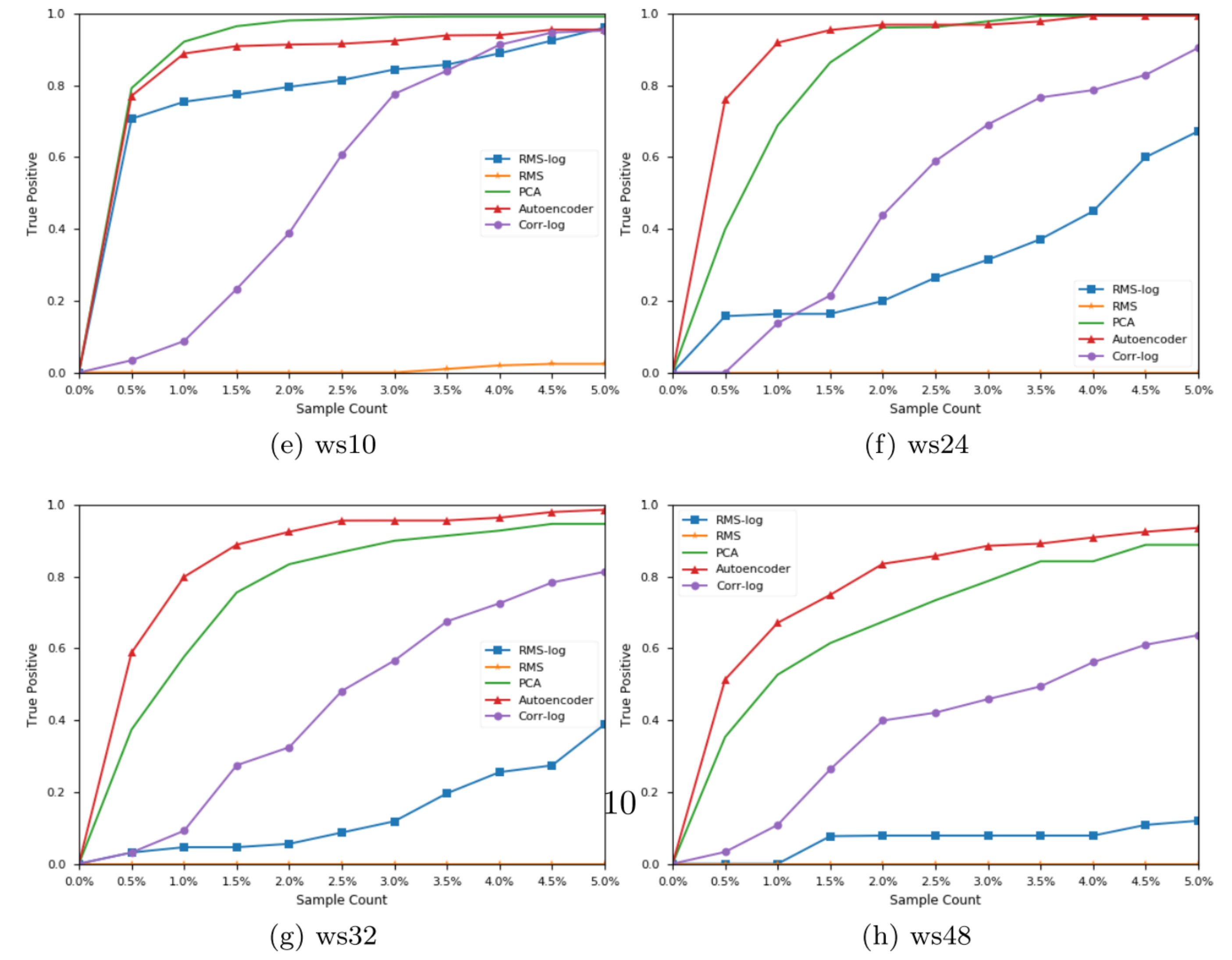
(e) ws10

(f) ws24

(g) ws32

(h) ws48

Off-diagonal Locations



(e) ws10

(f) ws24

(g) ws32

(h) ws48

Conclusion

An autoencoder can quickly locate areas of interest, especially in areas off the diagonal. Areas near the diagonal have large changes in frequencies, which can easily be picked up by RMS and PCA. Nonetheless, as we increase the size of the sub-block the results of the autoencoder improve.

References

[RHD 14] Suhas S.P. 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, and Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell, 159(7):1665–1680, December 2014.

[1] UC-Davis Genome Center, University of California, Davis, USA
[2] MIND Institute, University of California, Davis, USA
[3] Biochemistry and Molecular Medicine, University of California, Davis, USA