A Graph Auto-Encoder for Haplotype Assembly and Viral Quasispecies Reconstruction

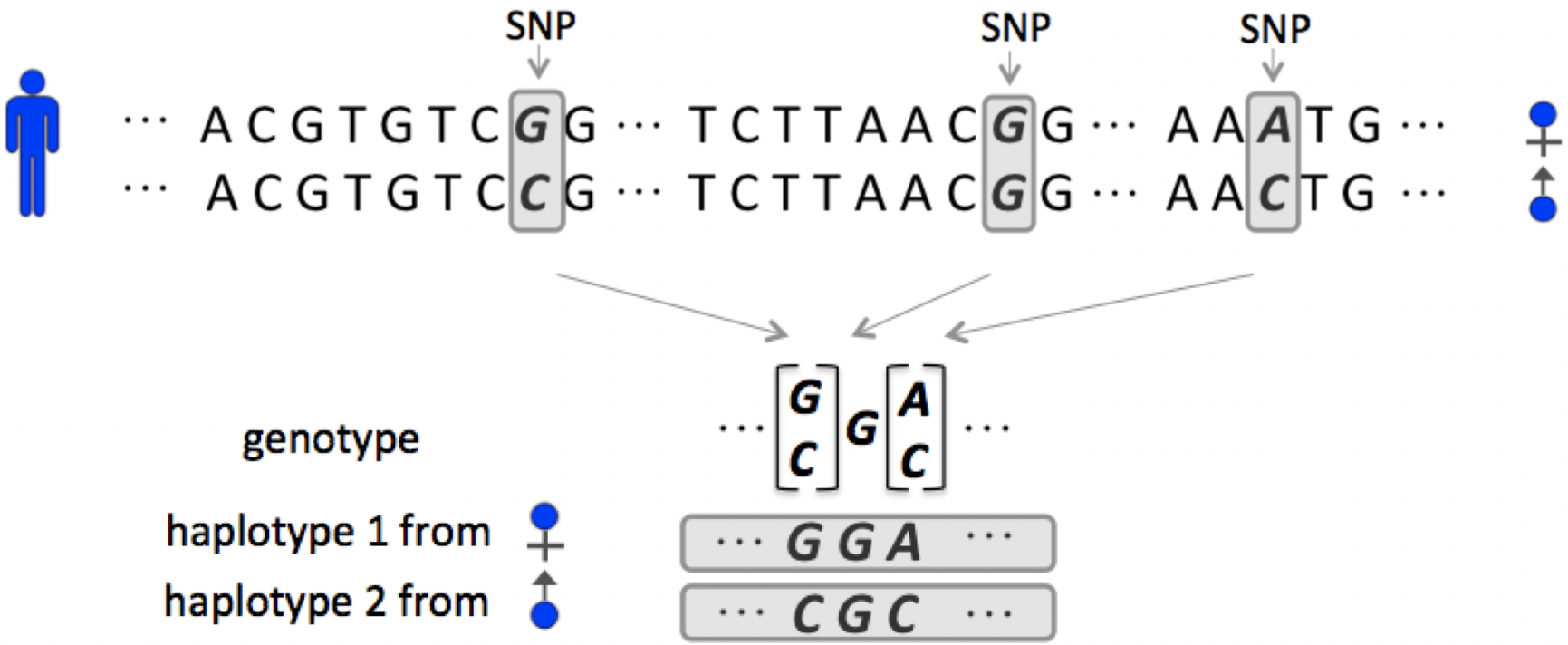
Ziqi Ke and Haris Vikalo

The University of Texas at Austin

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)
New York February 7-12, 2020

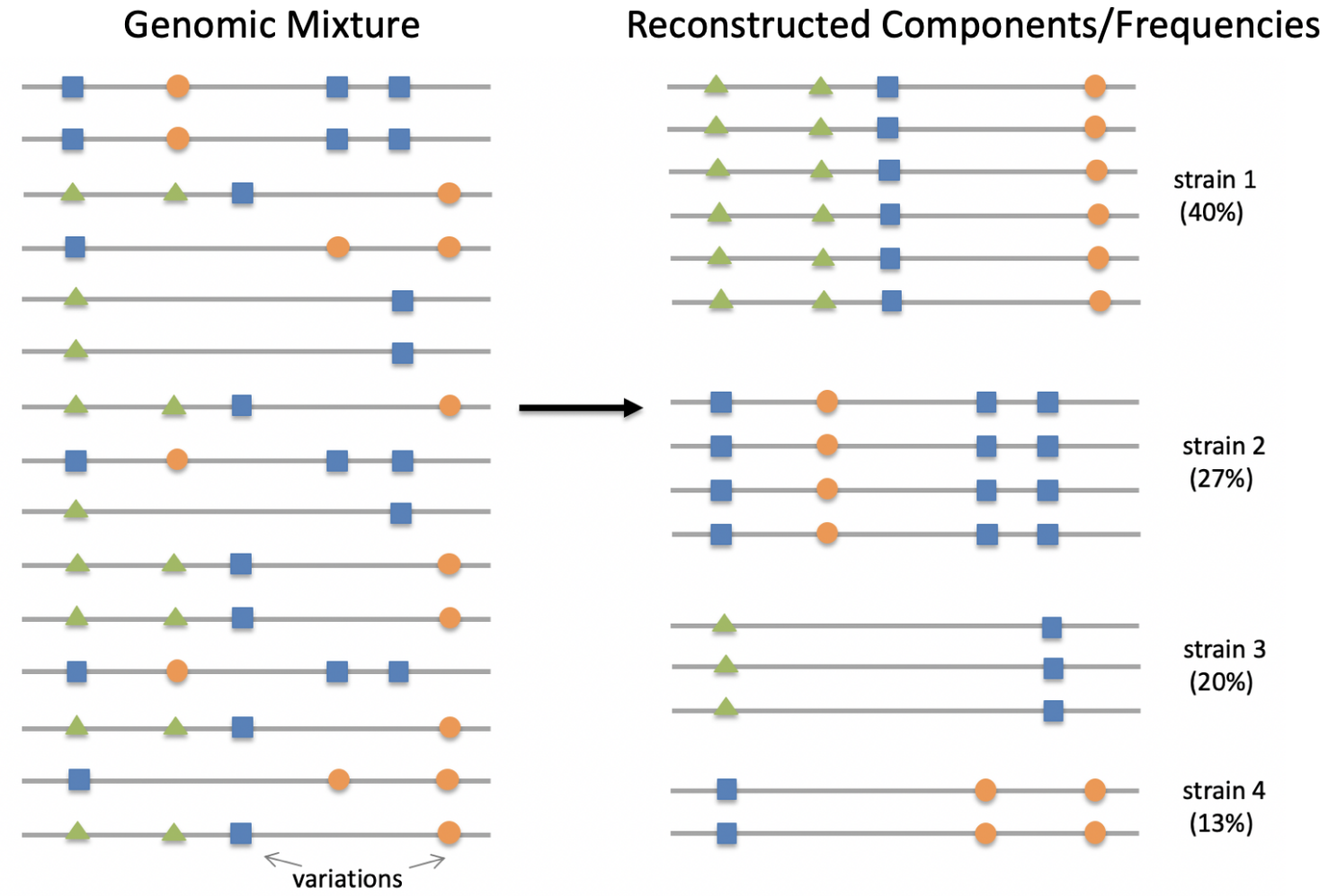
Motivation: Analysis of Haplotype Assembly

- Haplotypes
  - diploid (Human), polyploid (Potato et al.)

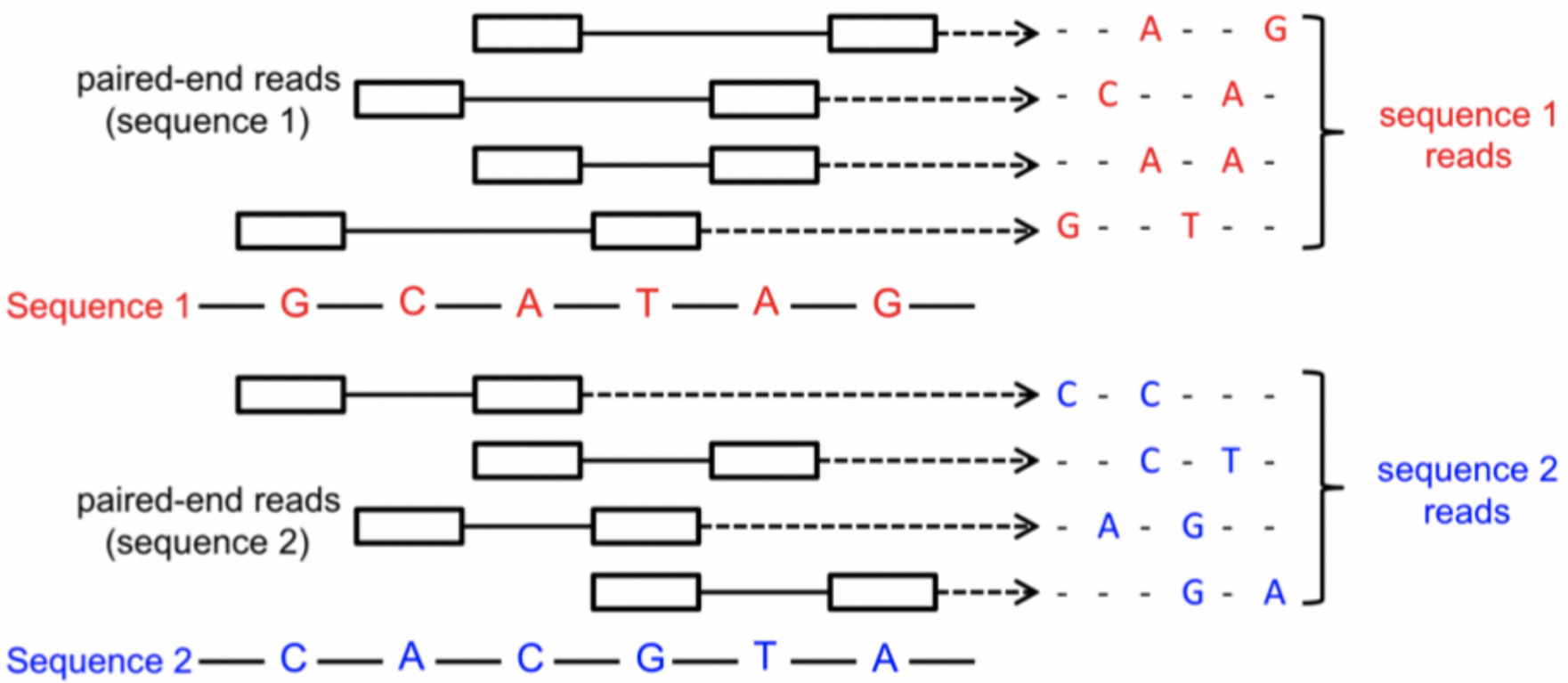


Motivation: Analysis of Viral Quasispecies

- RNA Viruses
  - HIV, HCV, Ebola, Zika



High-Throughput Sequencing Data



Reconstructing Genomic Components

- Challenging for several reasons
  - Haplotype assembly
    - limited read length
    - sequencing errors

- Existing methods
  - Haplotype assembly: HapCompass [Aguiar et al., 2012], H-PoP [Xie et al., 2016], HapCUT2 [Edge, et al., 2017], AltHap (Hashemi, et al. 2018)

Reconstructing Genomic Components

- Challenging for several reasons
  - Haplotype assembly
    - limited read length
    - sequencing errors
  - Viral quasispecies reconstruction
    - unknown population size
    - uneven frequencies of strains
- Existing methods
  - Haplotype assembly: HapCompass [Aguiar et al., 2012], H-PoP [Xie et al., 2016], HapCUT2 [Edge, et al., 2017], AltHap (Hashemi, et al. 2018)
  - Viral reconstruction: PredictHaplo [Prabhakaran et al., 2014], aBayesQR [Ahn, et al. 2017], TenSQR [Ahn, et al. 2018]

Preliminaries: Organizing Data in a Read Matrix

- Organize data into a matrix R, rows correspond to reads
  - utilize only the heterozygous sites

SNPs (Mutations)

Reads

1	2	0	4
1	0	3	4
0	3	2	1
4	3	2	0
2	4	0	3
0	4	1	3

SNP Fragment Matrix

Preliminaries: Organizing Data in a Read Matrix

- Organize data into a matrix R, rows correspond to reads
  - utilize only the heterozygous sites

SNPs (Mutations)

Reads				
	1	2	0	4
	1	0	3	4
	0	3	2	1
	4	3	2	0
	2	4	0	3
	0	4	1	3

SNP Fragment Matrix

- The goal: analyze R to jointly identify origins of the reads and assemble haplotypes or viral quasispecies/haplotypes

Preliminaries: Organizing Data in a Read Matrix

- Organize data into a matrix R, rows correspond to reads
  - utilize only the heterozygous sites

SNPs (Mutations)

Reads				
	1	2	0	4
	1	0	3	4
	0	3	2	1
	4	3	2	0
	2	4	0	3
	0	4	1	3

SNP Fragment Matrix

- The goal: analyze R to jointly identify origins of the reads and assemble haplotypes or viral quasispecies/haplotypes
- Note: R is obtained by sampling, with errors, an underlying ground truth matrix M; each row of M is one of the haplotypes

The Genomic Mixture Reconstruction Problem

- The read matrix can be represented as

$$R = \mathcal{P}_\Omega(UH + N)$$

U is the read origin indicator matrix, H is the component matrix, $\mathcal{P}_\Omega(.)$ is the sampling operator, N models errors

The Genomic Mixture Reconstruction Problem

- The read matrix can be represented as

$$R = \mathcal{P}_\Omega(UH + N)$$

  U is the read origin indicator matrix, H is the component matrix, $\mathcal{P}_\Omega(.)$ is the sampling operator, N models errors

- The GMR problem: Given R, find U and H

The Genomic Mixture Reconstruction Problem

- The read matrix can be represented as

$$R = \mathcal{P}_\Omega(UH + N)$$

  U is the read origin indicator matrix, H is the component matrix, $\mathcal{P}_\Omega(.)$ is the sampling operator, N models errors

- The GMR problem: Given R, find U and H

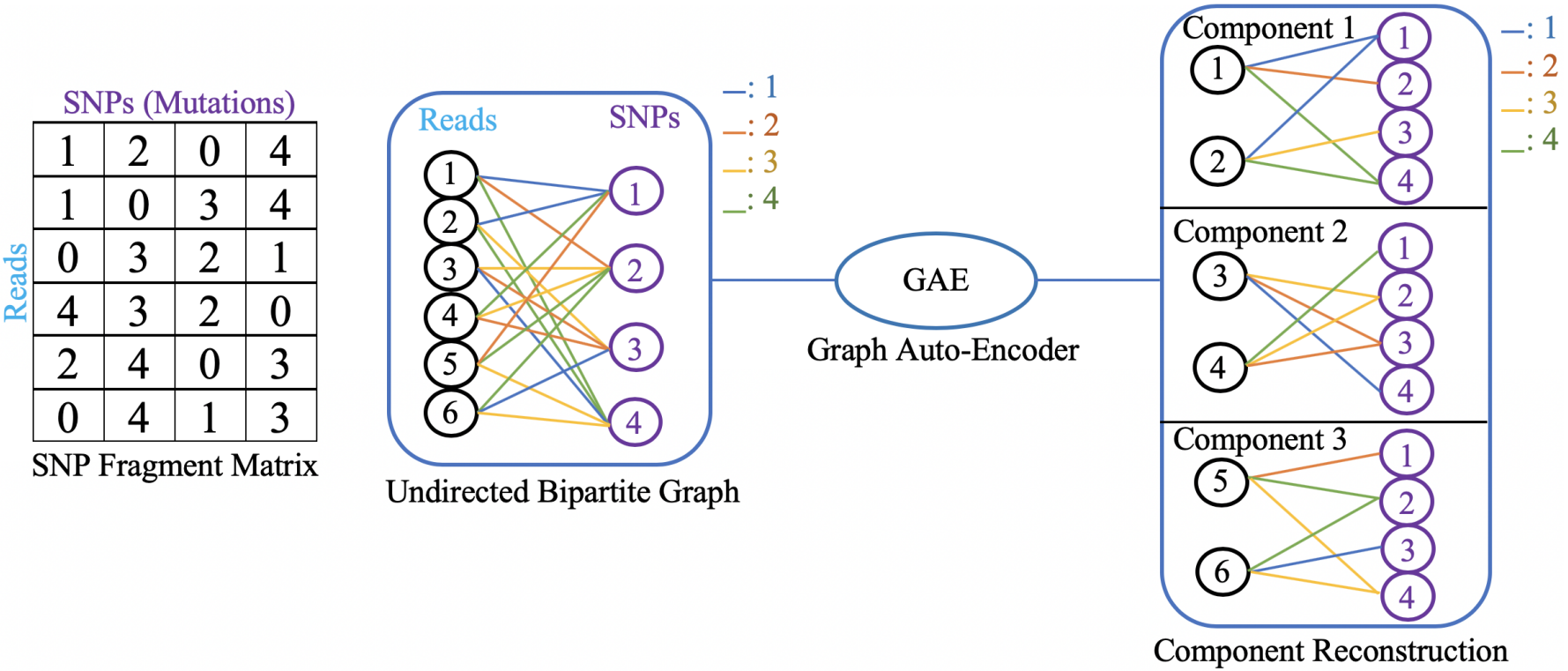
- Performance: correct phasing rate (CPR) and MEC score

$$\text{CPR} = 1 - \frac{1}{kn}(\min \sum_{i=1}^{k} \text{HD}(H_{i:}, \mathcal{M}(H_{i:})))$$

$$\text{MEC} = \sum_{i=1}^{m} \min_{j=1,2,\ldots,k} \text{HD}(R_{i:}, H_{j:})$$

Solving the GMR Problem via a Graph Auto-Encoder

- An undirected bipartite graph $G = (V, E, \mathcal{W})$

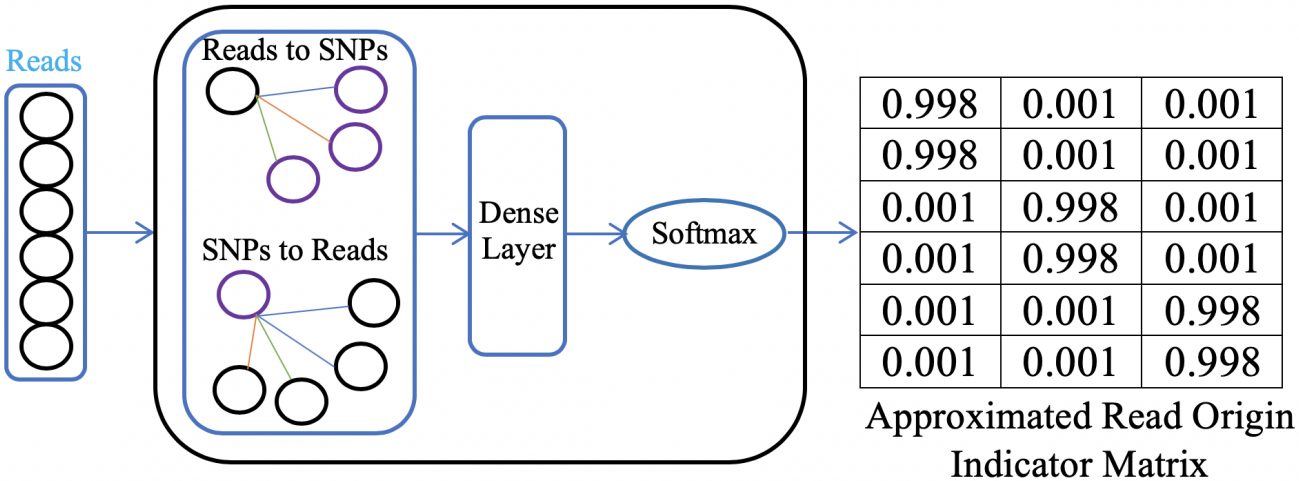


Graph Encoder

- The messages from read nodes to SNP nodes

$$M_{(1)} = \sigma(\sum_{w=1}^{4} D_s^{-1} A_w^T R W_w^{(1)} + B_w^{(1)})$$

D is the diagonal degree matrix, A is the graph adjacency matrix, W and B denote the weights and biases

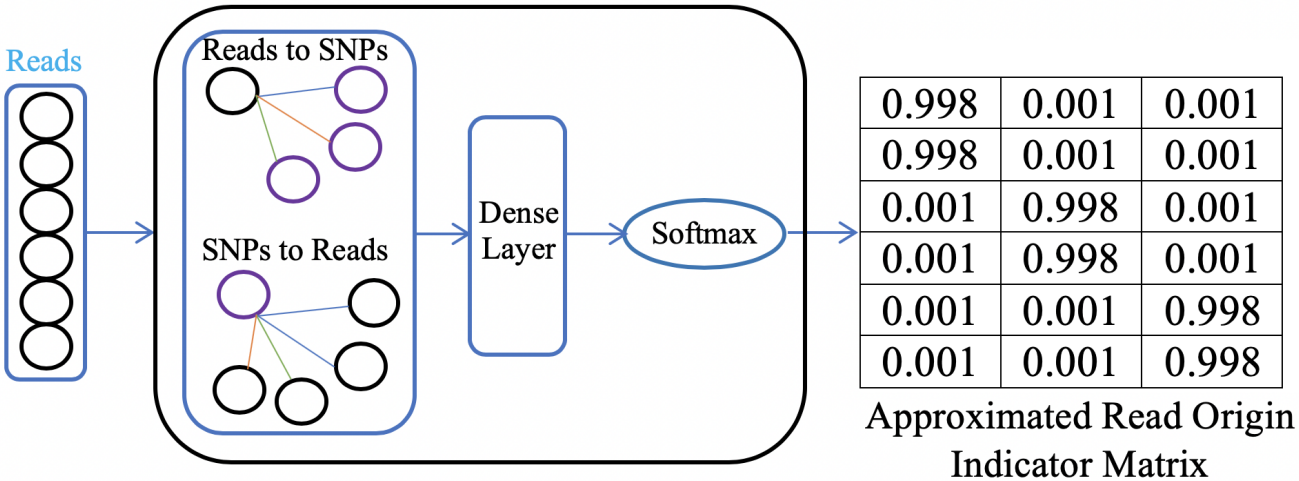


Graph Encoder

- The messages from SNP nodes to read nodes

$$M_{(2)} = \sigma\left(\sum_{w=1}^{4} D_r^{-1} A_w M_{(1)} W_w^{(2)} + B_w^{(2)}\right)$$

D is the diagonal degree matrix, A is the graph adjacency matrix, W and B denote the weights and biases



0.998	0.001	0.001
0.998	0.001	0.001
0.001	0.998	0.001
0.001	0.998	0.001
0.001	0.001	0.998
0.001	0.001	0.998

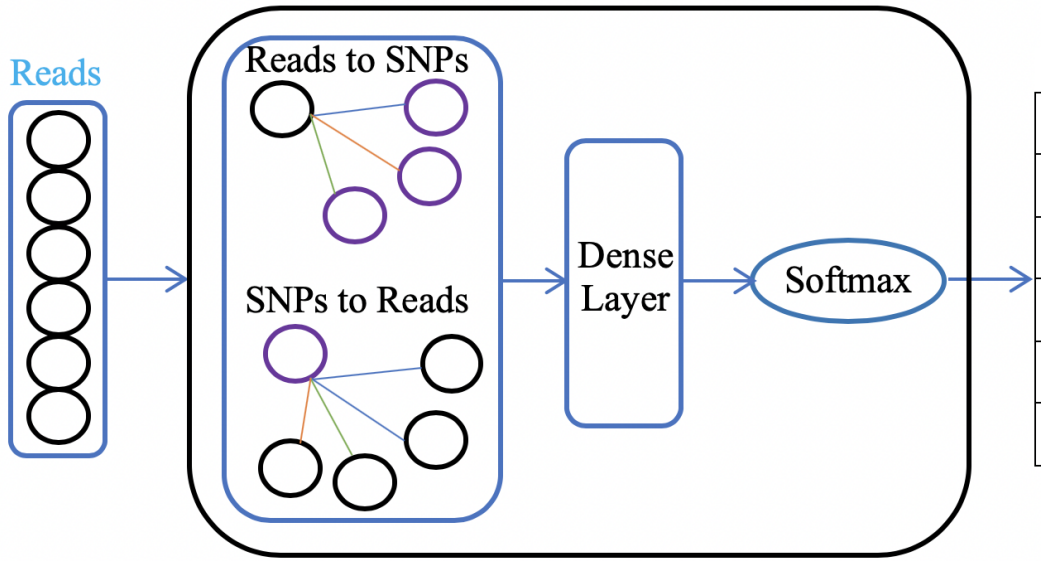
Approximated Read Origin Indicator Matrix

Graph Encoder

- The dense layer with softmax funcation

$$O = \sigma(M_{(2)} W_d + B_d)$$

$$Z_{ij} = \frac{e^{\beta O_{ij}}}{\sum_{j=1}^{k} e^{\beta O_{ij}}}$$



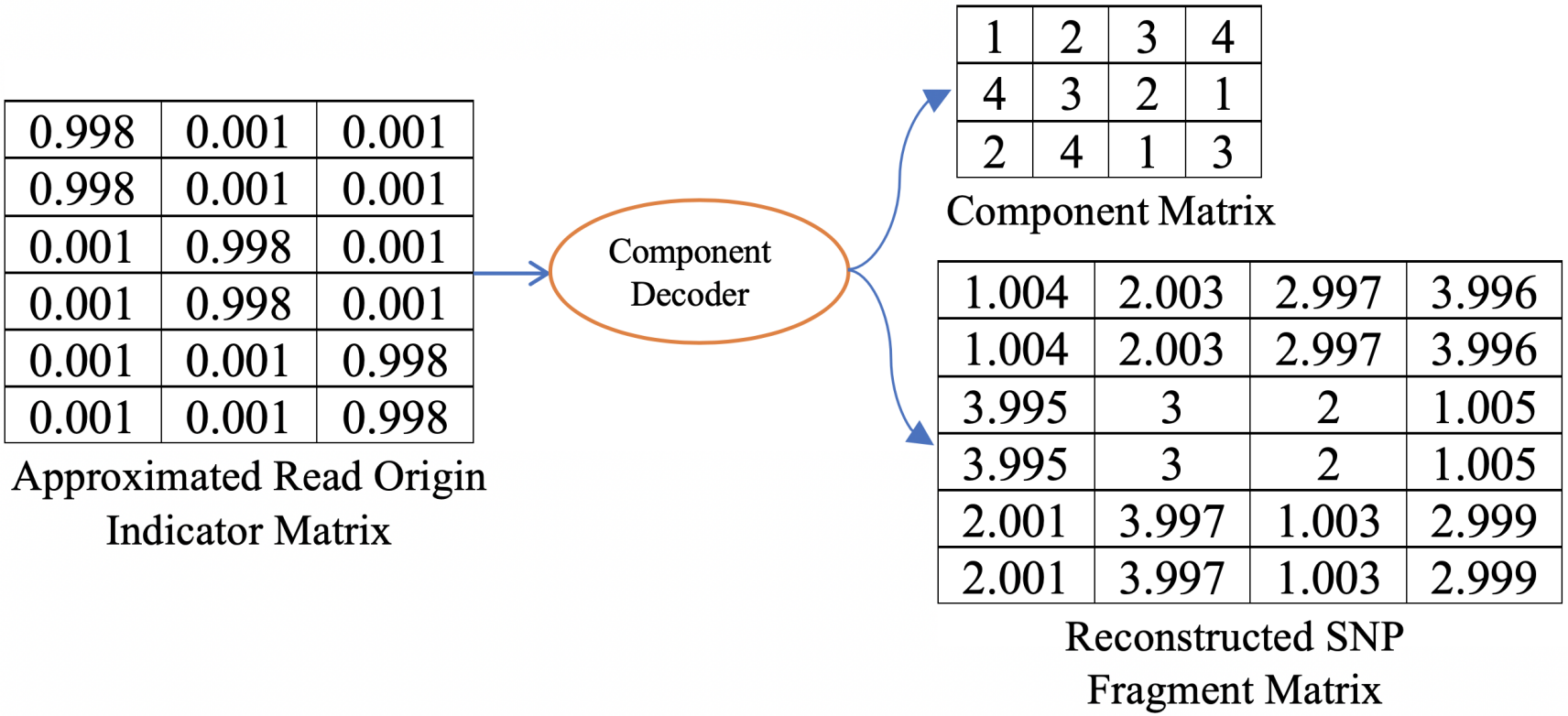
Stacked Graph Encoder

0.998	0.001	0.001
0.998	0.001	0.001
0.001	0.998	0.001
0.001	0.998	0.001
0.001	0.001	0.998
0.001	0.001	0.998

Approximated Read Origin Indicator Matrix

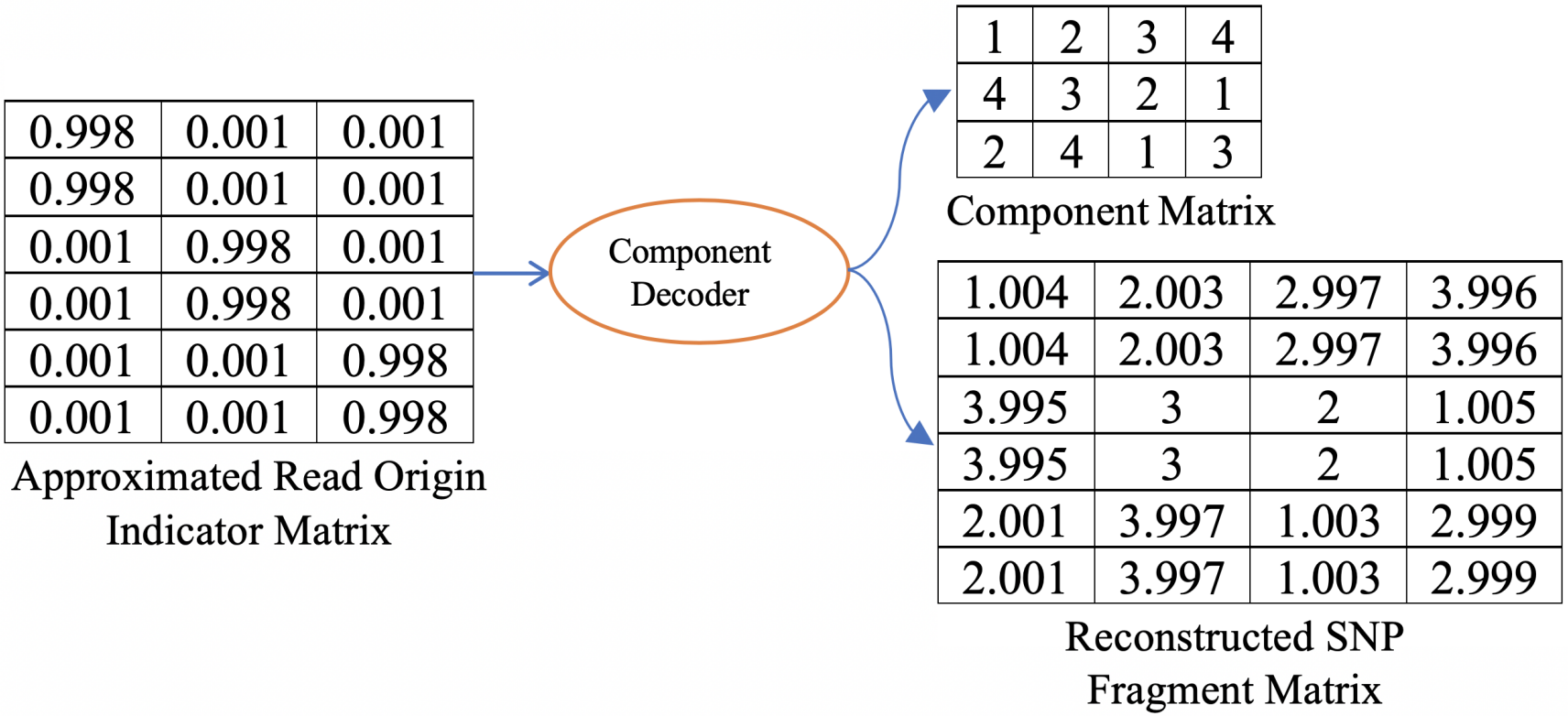
Component Decoder

- Reconstruct component matrix via majority voting



Component Decoder

- Reconstruct component matrix via majority voting
- $\mathcal{L} = \frac{1}{2}||\mathcal{P}_\Omega(\mathcal{R} - Z\mathcal{H})||_F^2$



Determining the Number of Strains, k

- Improvement rate of MEC [Ahn, et al. 2017]

$$MECimpr(k) = \frac{\text{MEC}(k) - \text{MEC}(k+1)}{\text{MEC}(k)}$$

$$\ldots \text{MEC}(k\text{-}2) \gg \text{MEC}(k\text{-}1) \gg \text{MEC}(k) \geq \text{MEC}(k{+}1) \ldots$$

Determining the Number of Strains, k

- Improvement rate of MEC [Ahn, et al. 2017]

$$MECimpr(k) = \frac{\mathrm{MEC}(k) - \mathrm{MEC}(k+1)}{\mathrm{MEC}(k)}$$

$$\ldots \mathrm{MEC}(k\text{-}2) \gg \mathrm{MEC}(k\text{-}1) \gg \mathrm{MEC}(k) \geq \mathrm{MEC}(k{+}1) \ldots$$

- Estimate k based on binary search
  - Starting from k_0

    $k_{\tau+1} \leftarrow 2k_\tau$ or $\lfloor (k_\tau + \min_{i=1,\ldots,\tau-1} k_i)/2 \rfloor$ if $MECimpr(k_\tau) > \eta$ $(k_i > k_\tau)$

    $k_{\tau+1} \leftarrow \lfloor (k_\tau + \max_{i=1,\ldots,\tau-1} \{1, k_i\})/2 \rfloor$ if $MECimpr(k_\tau) \leq \eta$

  - $\hat{k} \leftarrow k_\tau + 1$ when $k_{\tau+1} = k_\tau$

Results on Semi-Experimental Data

- Solanum Tuberosum semi-experimental data
  - 2×250bp-long Illumina's MiSeq
  - 5kbp genome (*Solanum Tuberosum* chromosome 5)
  - 10 instances
- Performance
  - MEC Score
  - Correct Phasing Rate: fraction of accuracy of each reconstructed strain

$$\text{CPR} = 1 - \frac{1}{kn}(\min \sum_{i=1}^{k} \text{HD}(H_{i:}, \mathcal{M}(H_{i:})))$$

Results on Semi-experimental Data

- Performance comparison on biallelic Solanum Tuberosum semi-experimental data

Coverage		MEC Mean	SD	CPR Mean	SD
15	GAEseq	**8.200**	**4.686**	**0.822**	0.048
	HapCompass	100.700	66.150	0.763	**0.046**
	H-PoP	28.700	32.667	0.783	0.066
	AltHap	59.100	28.125	0.709	0.054
25	GAEseq	**8.400**	**4.719**	**0.831**	0.081
	HapCompass	124.800	132.156	0.810	0.063
	H-PoP	33.800	47.434	0.798	**0.046**
	AltHap	92.600	83.649	0.756	0.068
35	GAEseq	**10.700**	**3.234**	**0.857**	0.087
	HapCompass	217.400	174.135	0.775	**0.072**
	H-PoP	41.700	53.971	0.823	0.094
	AltHap	164.000	101.583	0.754	0.093

Results on HIV-1 Data

- 5 HIV-1 strains, frequency 10-27%, pairwise distances 2.61-8.45%

		p17	p24	p2-p6	PR	RT	RNase	int	vif	vpr	vpu	gp120	gp41	nef
GAEseq	PredProp	**1**	1	**1**	**1**	**1.2**	**1**	**1**	**1**	**1**	1.2	1	1	1
	CPR_{HXB2}	**100**	99.4	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	96.2	96.7	100
	$CPR_{89.6}$	**100**	99.4	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.2	99.4	100	98.2
	CPR_{JR-SCF}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	99.9	100	99.3
	CPR_{NL4-3}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	100	100	99.8
	CPR_{YU2}	**100**	100	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100	99.6	100	98.1
PredictHap	PredProp	**1**	0.6	**1**	**1**	**1**	0.8	0.8	0.8	**1**	0.8	0.8	0.8	0.8
	CPR_{HXB2}	**100**	0	**100**	**100**	**100**	98.9	100	100	**100**	93.2	0	0	0
	$CPR_{89.6}$	**100**	100	**100**	**100**	**100**	100	99.8	100	**100**	0	97.8	100	98.8
	CPR_{JR-SCF}	**100**	100	**100**	**100**	**100**	100	100	100	**100**	100	99.7	100	100
	CPR_{NL4-3}	**100**	99.1	**100**	**100**	**100**	100	100	100	**100**	100	100	100	100
	CPR_{YU2}	**100**	0	**100**	**100**	**100**	0	0	0	**100**	100	98.6	100	100
TenSQR	PredProp	**1**	1.6	**1**	1	1.4	**1**	**1**	**1**	**1**	1.6	2.2	1.2	0.8
	CPR_{HXB2}	**100**	98.9	**100**	100	99.2	**100**	**100**	**100**	**100**	92.8	96.0	99.0	0
	$CPR_{89.6}$	**100**	100	**100**	100	98.0	**100**	**100**	**100**	**100**	94.0	97.2	100	95.7
	CPR_{JR-SCF}	**100**	100	**100**	100	100	**100**	**100**	**100**	**100**	100	98.3	97.7	99.8
	CPR_{NL4-3}	**100**	99.3	**100**	100	99.5	**100**	**100**	**100**	**100**	100	99.8	99.5	99.7
	CPR_{YU2}	**100**	99.3	**100**	99.7	99.7	**100**	**100**	**100**	**100**	100	94.9	100	98.6
aBayesQR	PredProp	1	1	**1**	**1**	1	**1**	1	1	**1.2**	1	0.8	0.8	1.2
	CPR_{HXB2}	**100**	99.4	**100**	**100**	98.5	**100**	99.9	100	**100**	99.6	98	0	95.8
	$CPR_{89.6}$	**100**	98.7	**100**	**100**	98.6	**100**	100	100	**100**	92	96.5	98.9	95.5
	CPR_{JR-SCF}	**100**	99.6	**100**	**100**	99	**100**	100	100	**100**	98.8	97.7	99.1	98.2
	CPR_{NL4-3}	**100**	100	**100**	**100**	98.9	**100**	100	99.8	**100**	100	96.3	98.8	100
	CPR_{YU2}	**100**	99.7	**100**	**100**	99.2	**100**	99.5	99.7	**100**	100	0	98.6	99.2

Summary and Acknowledgements

- GAEseq: The first learning framework for haplotype assembly and viral quasispecies reconstruction
  - an undirected bipartite graph
  - a graph auto-encoder
  - not discussed: computational setting and more results – details in the paper and supplementary material
- Software: https://github.com/WuLoli/GAEseq
- Acknowledgements
  - support from NSF CCF 1618427