# AAGI-AU-RD-HAP: Is Deep Learning Nearing Optimal Haplotype MEC?

Richard Pinter[1], Julian Taylor[2], Stefan Podgorski[1], Mario Fruzangohar[2]

[1] Australian Institute for Machine Learning, Adelaide University

[2] Biometry Hub, Adelaide University

AAGI Analytics for the Australian Grains Industry

## Background and research problem

Haplotype phasing identifies which genetic variants occur together on each chromosome copy, supporting downstream genetic and trait analysis in Australian grains research. Deep learning methods reconstruct haplotypes from sequencing reads and are commonly compared using minimum error correction (MEC). These comparisons show relative performance, but do not establish how close their MEC scores are to the optimum. The research problem is to determine how much scope remains for new architectures to improve MEC.

### Einkorn wheat



Diploid, 2n = 14 [1]. Illustrative wheat example.
Photo: Sten Porse. CC BY-SA 3.0.

In a cell

### Reference genome

A C G T A C

**One chromosome pair** Two copies from the same plant

Copy 1 → T A A C T C A C T G A C T T A C T A A C T C A C

Copy 2 → T G A C T T A C T A A C T C A C T G A C T T A C

s1 s2 s3 s4 s5 s6

Haplotype 1

Haplotype 2

Sequence

### Mixed sequencing reads

| | s1 | s2 | s3 | s4 | s5 | s6 |
|---|---|---|---|---|---|---|
| r1 | A | C | | | | |
| r2 | | | | C | G | |
| r3 | | T | A | | | |
| r4 | | | | | A | C |
| r5 | G | T | | | | |
| r6 | | | | T | A | |
| r7 | | C | G | | | |
| r8 | | | | | G | T |

SNP calls

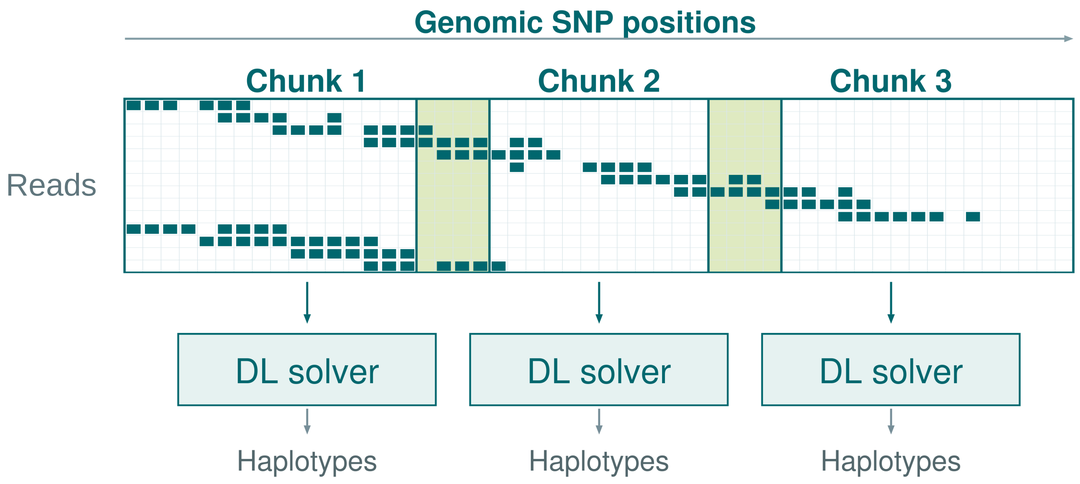
### Read-SNP matrix

| | s1 | s2 | s3 | s4 | s5 | s6 |
|---|---|---|---|---|---|---|
| r1 | A | C | – | – | – | – |
| r2 | – | – | – | C | G | – |
| r3 | – | T | A | – | – | – |
| r4 | – | – | – | – | A | C |
| r5 | G | T | – | – | – | – |
| r6 | – | – | – | T | A | – |
| r7 | – | C | G | – | – | – |
| r8 | – | – | – | – | G | T |

– = no observation

## Overlapping SNP chunks

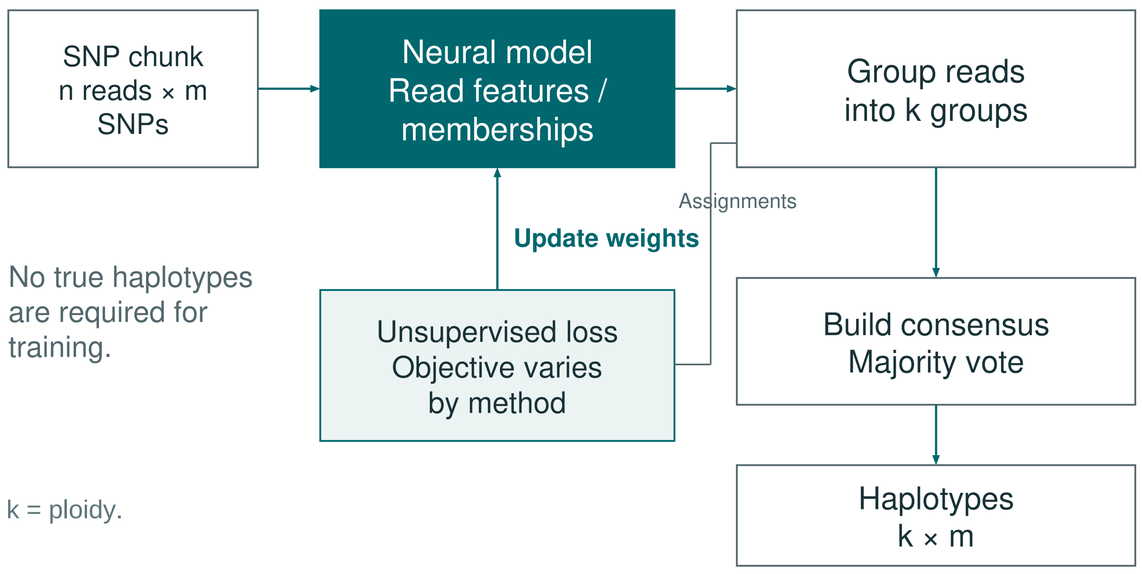
250-SNP windows, 50-SNP overlap.
Used in XHap [2] and DeepHapNet [3].




Shared SNPs support stitching adjacent haplotypes.

## General deep learning solver architecture

Shared unsupervised workflow [2-6].




## Minimum error correction (MEC)

**Edits: Corrections against H1 (orange) and H2 (purple) are highlighted for each read.**

| | SNP 1 | SNP 2 | SNP 3 | SNP 4 | Edits vs H1 | Edits vs H2 |
|---|---|---|---|---|---|---|
| **Haplotype 1** | A | C | G | T | | |
| **Haplotype 2** | G | T | A | C | | |
| Read 1 | **A** | **C** | – | **T** | **0** | 3 |
| Read 2 | **A** | **T** | **G** | – | **1** | 2 |
| Read 3 | **G** | **T** | **A** | – | 3 | **0** |
| Read 4 | **G** | – | **G** | **C** | 2 | **1** |

**MEC = 2**

**0 + 1 + 0 + 1**

"–" = no observation.

## Benchmarking against optimal MEC

Model comparisons alone do not show how close MEC scores are to optimal.

### 1 Formulate MEC

MEC is NP-hard [7]. Prior work found optima for NA12878 fosmid data [8].

### 2 Run CP-SAT

OR-Tools provides a lower bound and certifies the optimal MEC [9].

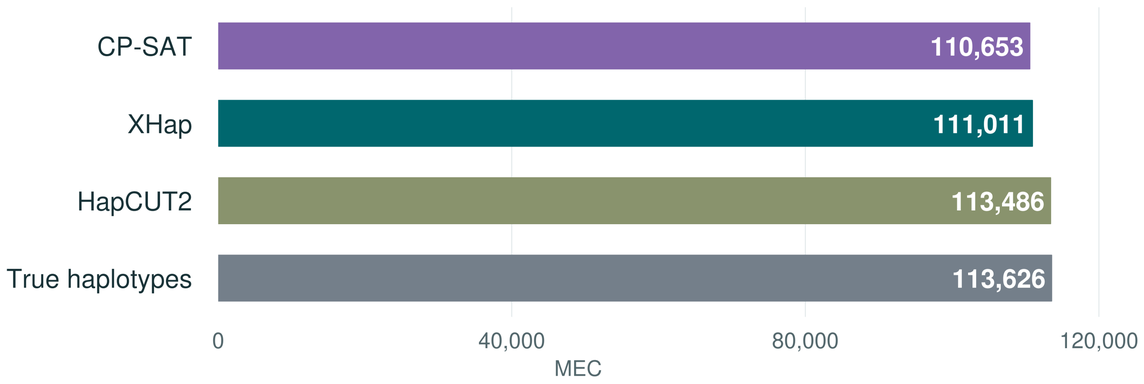
### 3 Compare per chunk

Compare the deep learning solver's MEC with the optimum for each chunk.

## NA12878 chromosome 22: stitched MEC and per-chunk optimality

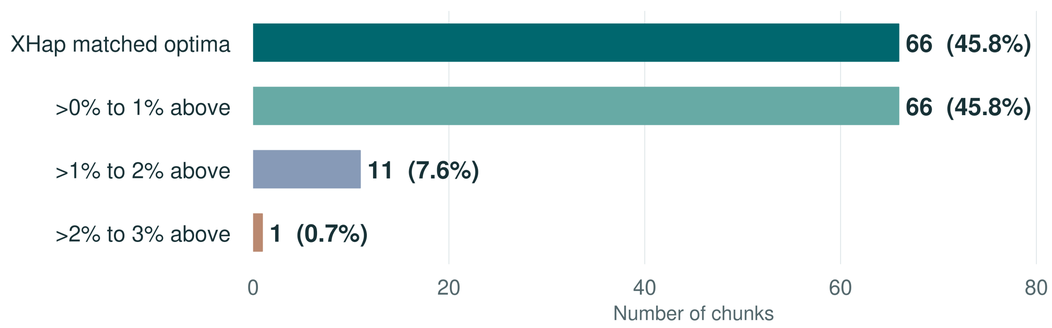
### Stitched MEC

CP-SAT reduces MEC by 358 versus XHap (0.322%)




### Gap between XHap and optimal MEC

**All 144 chunks are within 3% of optimal MEC.**




## Industry impact

By showing where current methods are already near the MEC optimum, this benchmark can help direct research effort towards more challenging phasing problems. The project is ongoing. Further work will extend evaluation to crop genomes, higher ploidy and insertions and deletions, and refine the solver architecture to support downstream analytics in Australian grains and plant breeding research.

## References

[1] Ahmed et al. (2023). Einkorn genomics. Nature. doi:10.1038/s41586-023-06389-7.
[2] Consul et al. (2023). XHap. Bioinform Adv. doi:10.1093/bioadv/vbad169.
[3] Luo et al. (2025). DeepHapNet. Brief Bioinform. doi:10.1093/bib/bbae656.
[4] Ke & Vikalo (2020). Graph auto-encoder for haplotype assembly. AAAI. doi:10.1609/aaai.v34i01.5414.
[5] Ke & Vikalo (2020). Convolutional auto-encoder for haplotype assembly. NeurIPS 33:13493-13503.
[6] Ke & Vikalo (2022). Deep learning from short and long sequencing reads. BCB. doi:10.1145/3535508.3545524.
[7] Cilibrasi et al. (2007). Complexity of SNP haplotyping. Algorithmica. doi:10.1007/s00453-007-0029-z.
[8] Chen et al. (2013). Exact algorithms for haplotype assembly. Bioinformatics. doi:10.1093/bioinformatics/btt349.
[9] Google OR-Tools. CP-SAT solver. developers.google.com/optimization/cp/cp_solver.

GRDC Grains Research & Development Corporation | Adelaide University | Curtin University | The University of Queensland Australia

The support of the Grains Research and Development Corporation (GRDC) provided through the Analytics for the Australian Grains Industry (AAGI) project is acknowledged.