Graph Layout by Path-Guided Stochastic Gradient Descent

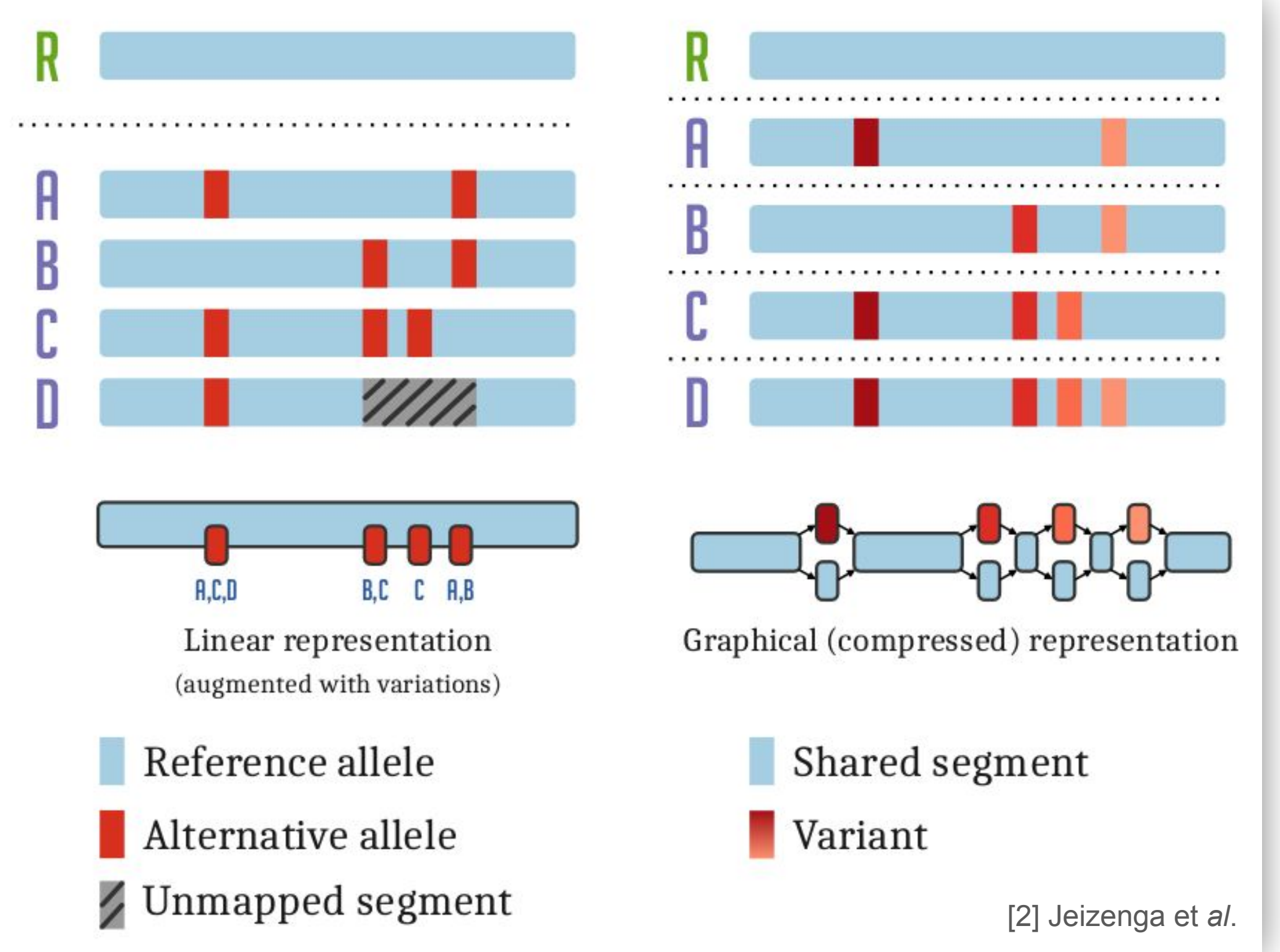
Simon Heumos[1*], Andrea Guarracino[2*], Sven Nahnsen[1,3], and Erik Garrison[4]

[1]*Quantitative Biology Center (QBiC) Tübingen, University of Tübingen, Tübingen, Germany,* [2]*Genomics Research Centre, Human Technopole, Milan, Italy,* [3]*Biomedical Data Science, Dept. of Computer Science, University of Tübingen, Tübingen, Germany,* [4]*University of Tennessee Health Science Center, Memphis, TN, USA*

**Contributed equally.*

Pangenome graphs built from raw sets of alignments may have complex structures which can introduce difficulty in downstream analyses, visualization, mapping, and interpretation. Graph sorting aims to find the best node order for a 1D and 2D layout to simplify these complex regions. Pangenome graphs embed linear pangenomic sequences as paths in the graph, but to our knowledge, no algorithm takes into account this biological information in the sorting. Moreover, existing 2D layout methods struggle to deal with large graphs. We present a new layout algorithm to simplify a pangenome graph, by using path-guided stochastic gradient descent (SGD[3]) to move a single pair of nodes at a time. We exemplify how the 1D path-guided SGD implementation is a key step in general pangenome analyses such as pangenome graph linearization and simplification.

VARIATION GRAPHS ENCODE PANGENOMES



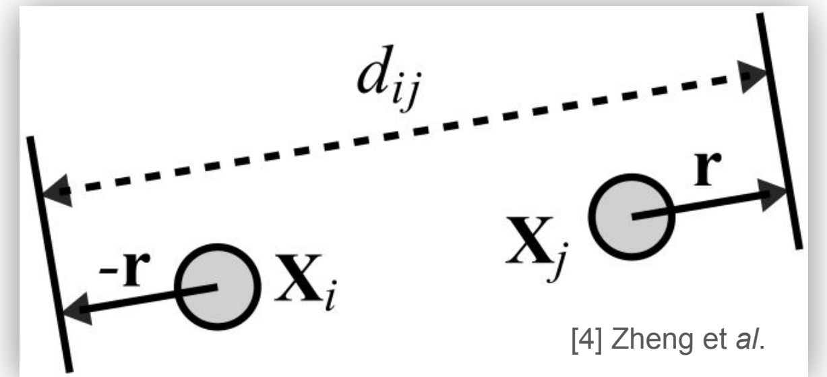
[2] Jeizenga et *al.*

A pangenome[1] models the full set of genomic elements in a given species or clade. It can efficiently be encoded[2] in the form of a variation graph, which embeds the linear sequences of the pangenome as paths in the graphs themselves.

https://bit.ly/PangenomeGraph
https://bit.ly/OptimizedDynamicGraphImplementation

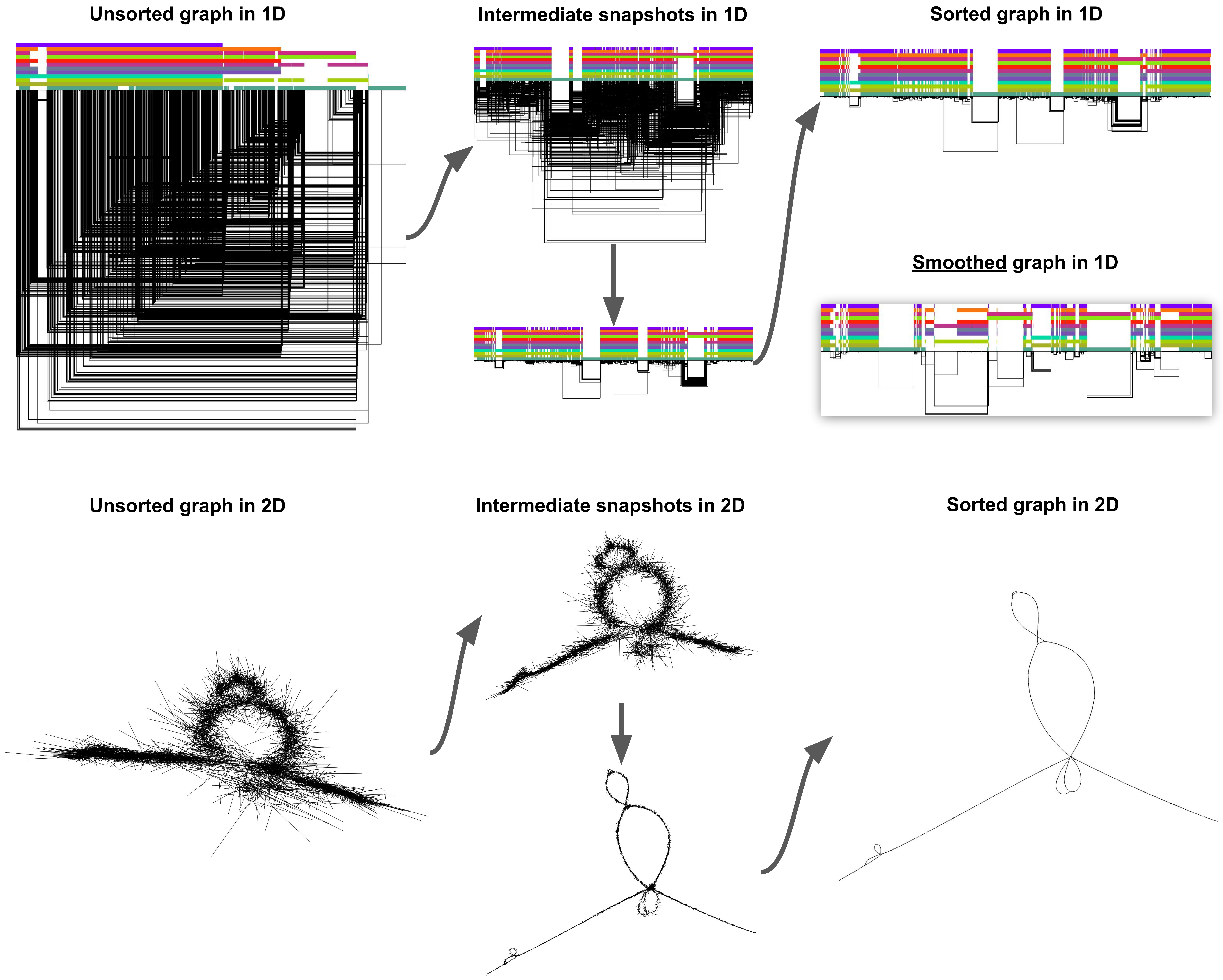
PATH-GUIDED STOCHASTIC GRADIENT DESCENT

Our algorithm moves a single pair of nodes at a time, optimizing the disparity between the layout distance of a node pair and the actual nucleotide distance of a path traversing these nodes.



[4] Zheng et *al.*

- The first node X_i of a pair is a uniform path step pick from all nodes.
- The second node X_j of a pair is sampled from the same path following a Zipfian distribution.
- The path nucleotide distance of the nodes in the pair guides the actual layout distance d_{ij} update of these nodes. The magnitude r of the update depends on the current learning rate of the SGD.



Unsorted graph in 1D

Intermediate snapshots in 1D

Sorted graph in 1D

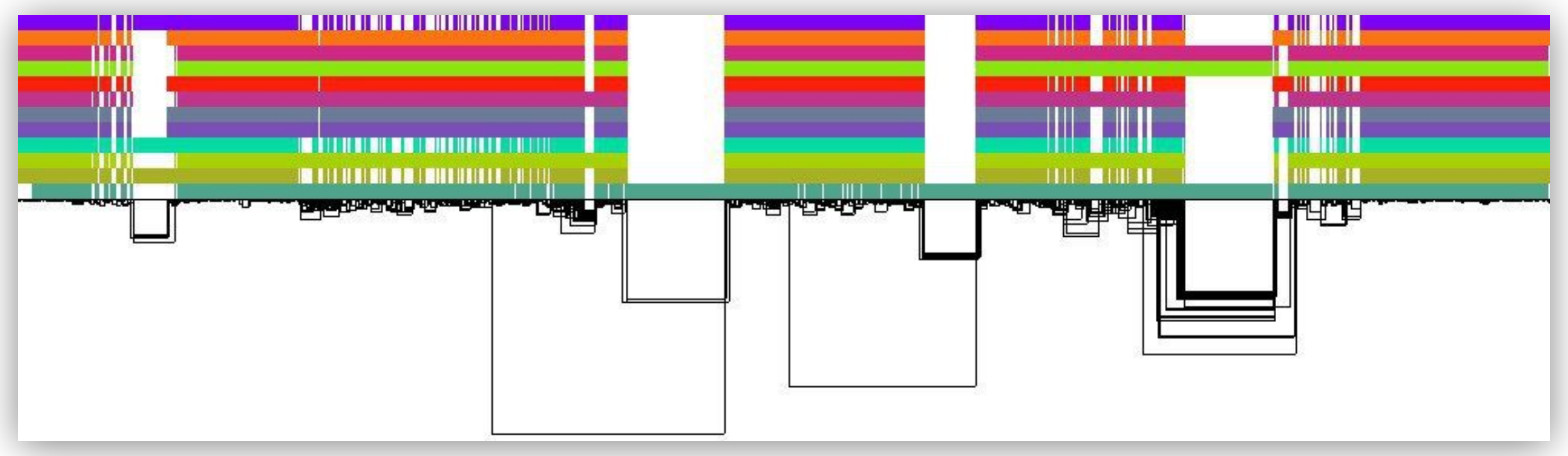
Smoothed graph in 1D

Unsorted graph in 2D

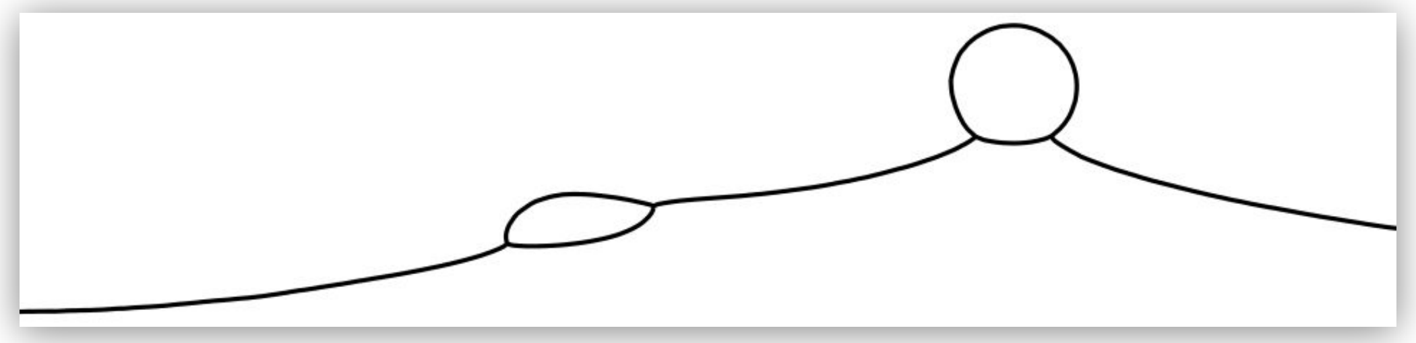
Intermediate snapshots in 2D

Sorted graph in 2D

GRAPH VISUALIZATIONS EXPLAINED



- The graph nodes' are arranged from left to right forming the pangenome's sequence.
- The colored bars represent paths versus the pangenome sequence in a binary matrix.
- The black lines under the paths, so called links, represent the topology of the graph.



- Graph topology becomes visible in 2 dimensions.
- Regions of variation appear as bubbles.

GRAPH SIMPLIFICATION PIPELINE

- Smoothxg runs abPOA for each block of paths that are collinear within a seqwish induced variation graph. A prerequisite is that the graph nodes are sorted according to their occurrence in the graph's embedded paths. Our 1D path-guided SGD algorithm is designed to provide this kind of sort.

FUTURE WORK

- Explore the path-guided SGD parameter space
- Compare our proposed 2D graph layouting algorithm with existing pangenome graph visualization tools
- Tight integration with the Vulkan-accelerated interactive GFA visualization tool gfaestus.

References

1. Eizenga et al. (2020). Pangenome Graphs. *Annual Reviews of Genomics and Human Genetics*, 21, 1.
2. Eizenga et al. (2020). Efficient dynamic variation graphs. *Bioinformatics*, btaa640.
3. Zheng et al. (2019). Graph Drawing by Stochastic Gradient Descent. *IEEE Transactions on Visualization and Computer Graphics.* 25, 2738-2748.

Acknowledgements

We thank Vincenza Colonna for organizing the Crusco Summer Hackathon and the Forentum Ritrovato museum for hosting it.
We thank the deNBI cloud for providing computational resources.
S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany.