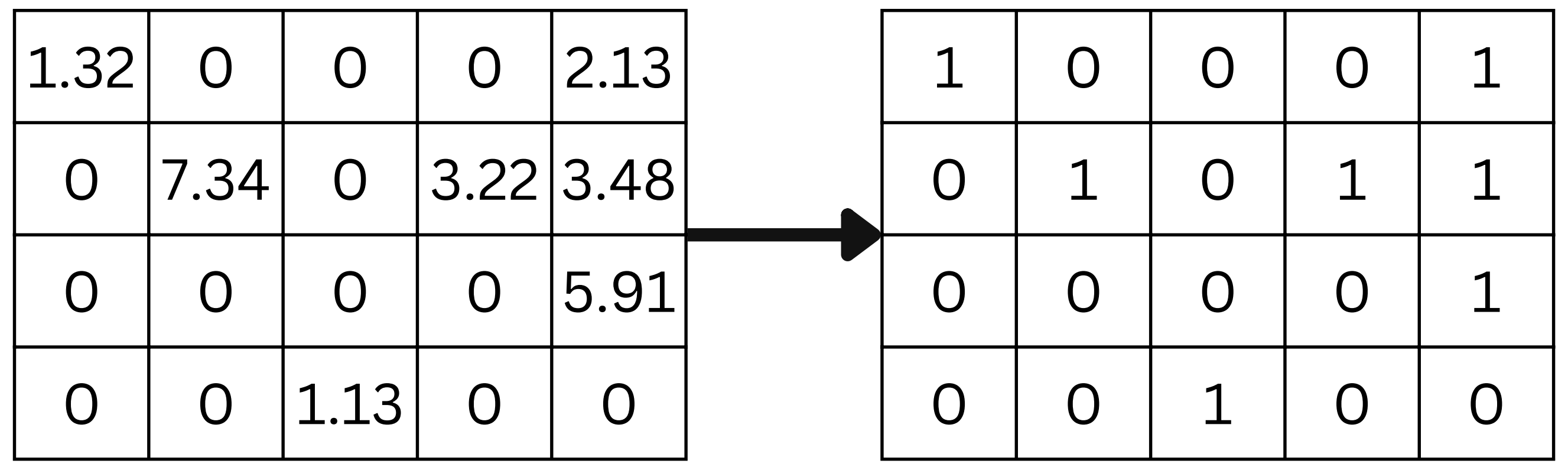
Similarity metrics for binary cell clustering

How close can we get to state-of-the-art ?

Author: Bartosz Golik (b.p.golik@student.tudelft.nl)
Supervisors: Prof.dr.ir. M.J.T. Reinders, Gerard Bouland

01 INTRODUCTION

- Single-cell RNA sequencing:
  - tool for studying heterogeneity of cell populations
  - research for cancer treatment
  - early embryo development
- Single-cell data represented as **expression matrix**
  - rows → cells, columns → genes
  - fields show how much a gene is expressed in a cell
- **Problem:** single-cell data requires too much memory
- **Solution:** expression matrix can be **binarized** [1]



1.32	0	0	0	2.13
0	7.34	0	3.22	3.48
0	0	0	0	5.91
0	0	1.13	0	0

1	0	0	0	1
0	1	0	1	1
0	0	0	0	1
0	0	1	0	0

Figure 1: Binarization of expression matrix

02 AIM

In order to analyze single-cell data, a common next step is **cell clustering,** where we choose a **similarity metric** to compare two cells and compute **similarity matrix** and **kNN graph.**

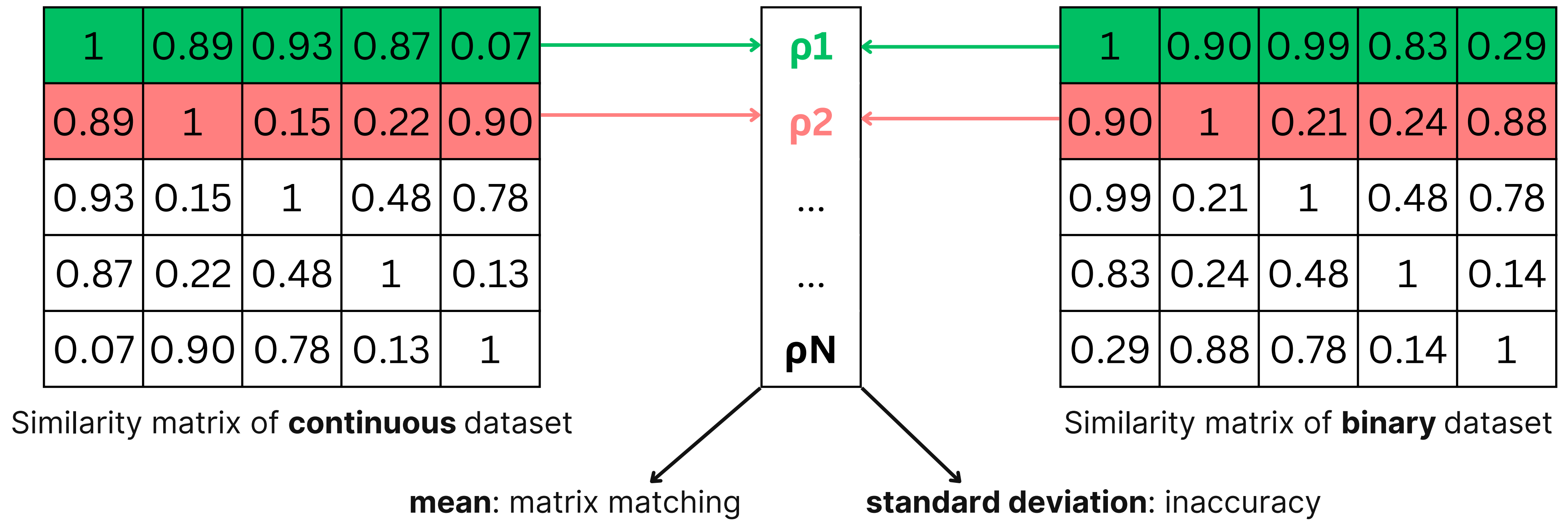
How close are the results of binary similarity metrics to the ones produced by continuous metrics when applied on single-cell data ?

05 CONCLUSION

Through experimental analysis it was shown that the output of specific **continuous similarity metrics can,** to some extent, **be consistently and accurately reproduced with binary metrics** when applied on single-cell data. The quality of kNN graph reproduction is debatable and **further evaluation on larger and sparser datasets** is required. This, however, **requires immense computational power**, which limited our evaluation possibilities. We advise future researchers to carefully consider their code implementation to decrease the impact of memory limitations.

03 METHODOLOGY

- Experimental studies on 8 real single-cell datasets
- 9 continuous metrics chosen based on existing evaluations
- 18 binary metrics with low correlation for the greatest variety of results
- Comparing similarity matrices: cell-wise **Spearman's rank correlation ρ** (Fig. 2)
- Comparing kNN graphs: cell-wise **Jaccard index J** (Fig. 3) in adjacency list form
- Implementation: C++ binary cell clustering pipeline with R used as evaluation framework, integrated together with *Rcpp* [2]



1	0.89	0.93	0.87	0.07
0.89	1	0.15	0.22	0.90
0.93	0.15	1	0.48	0.78
0.87	0.22	0.48	1	0.13
0.07	0.90	0.78	0.13	1

Similarity matrix of **continuous** dataset

ρ1
ρ2
...
...
ρN

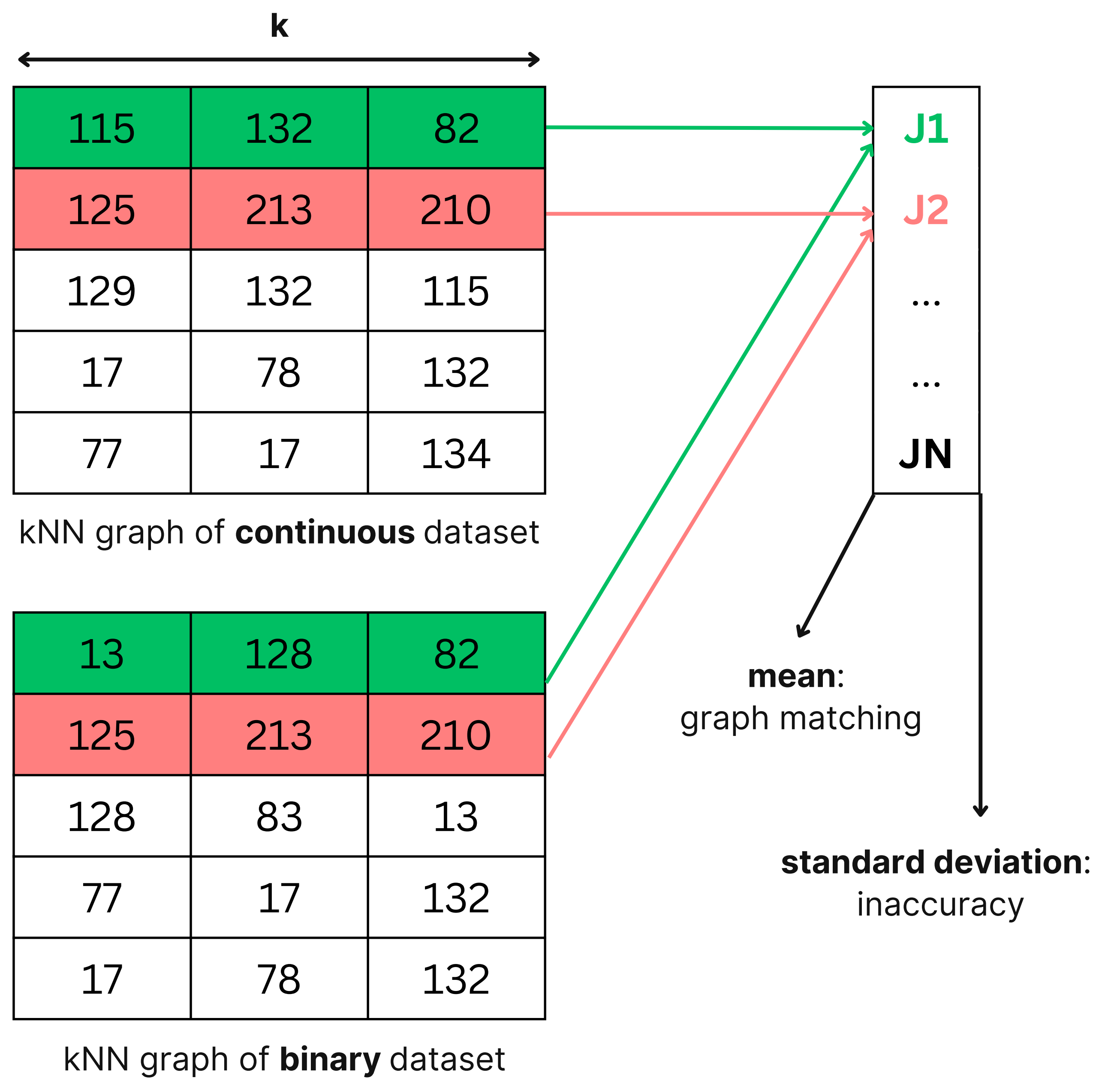
1	0.90	0.99	0.83	0.29
0.90	1	0.21	0.24	0.88
0.99	0.21	1	0.48	0.78
0.83	0.24	0.48	1	0.14
0.29	0.88	0.78	0.14	1

Similarity matrix of **binary** dataset

mean: matrix matching

standard deviation: inaccuracy

Figure 2: Computation flow for comparing similarity matrices



k

115	132	82
125	213	210
129	132	115
17	78	132
77	17	134

kNN graph of **continuous** dataset

J1
J2
...
...
JN

13	128	82
125	213	210
128	83	13
77	17	132
17	78	132

kNN graph of **binary** dataset

mean: graph matching

standard deviation: inaccuracy

Figure 3: Computation flow for comparing kNN graphs

04 RESULTS

- **Not all continuous metrics have a matching binary counterpart** (Fig. 4). Highest scores noted for correlation-based metrics (Spearman, Weighted rank, Kendall, ZI Kendall). Lowest scores for true distance metrics (Euclidean, Manhattan), excluding Canberra.
- **Optimal matchings stay consistent across datasets**. Exceptions to this rule suggest that matching consistency may be sparsity dependant.
- **kNN graphs are not reproducible**. Further evaluation is required on larger & sparser datasets. Relation between matrix matching and graph matching is not linear, but exponential (Fig. 5). Higher graph similarity for higher k (Fig. 5) suggests more order mismatch in closest neighbours.
- **Better matrix matchings are more accurate within cells** (Fig. 6). Good matching quality is reflected across the entire dataset.
- **No correlation with data charecteristics was found**. This could change for bigger datasets, as metric performance was shown to be input dependant [3].
- For each of the continuous metrics, a set of most fitting binary metrics was identified. Not all binary metrics proved to match with continuous metrics.

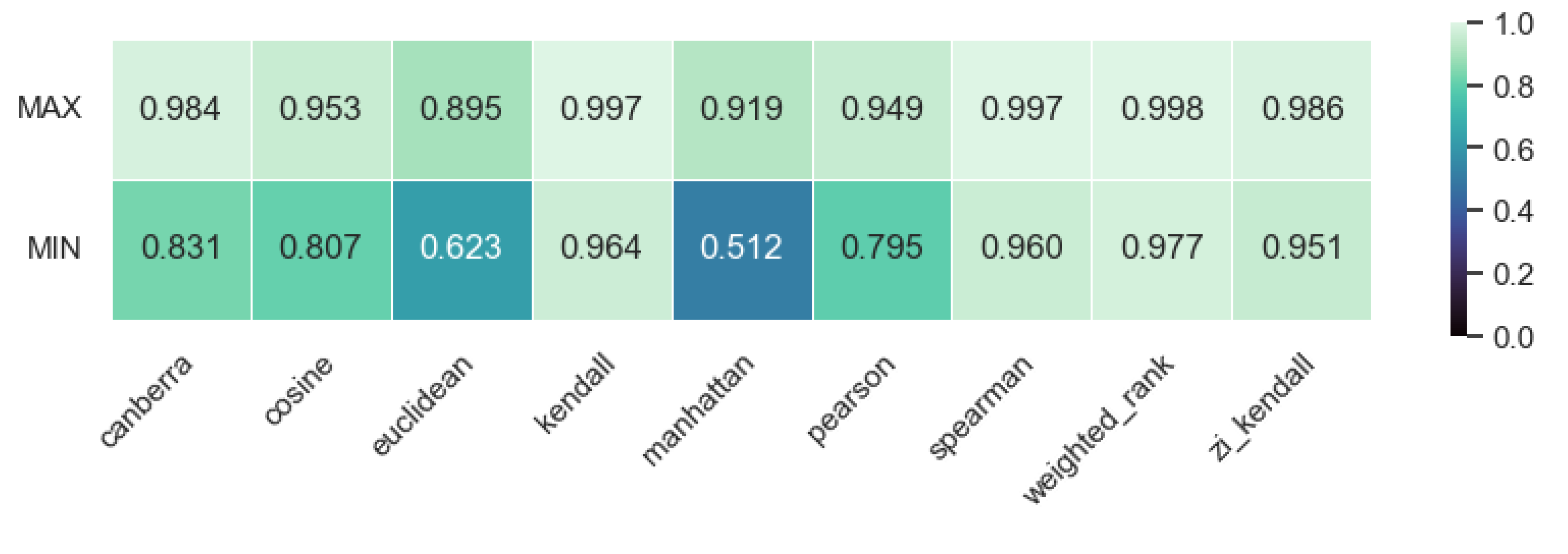


Figure 4: Lower and upper bounds of optimal matrix matchings for each of the continuous metrics

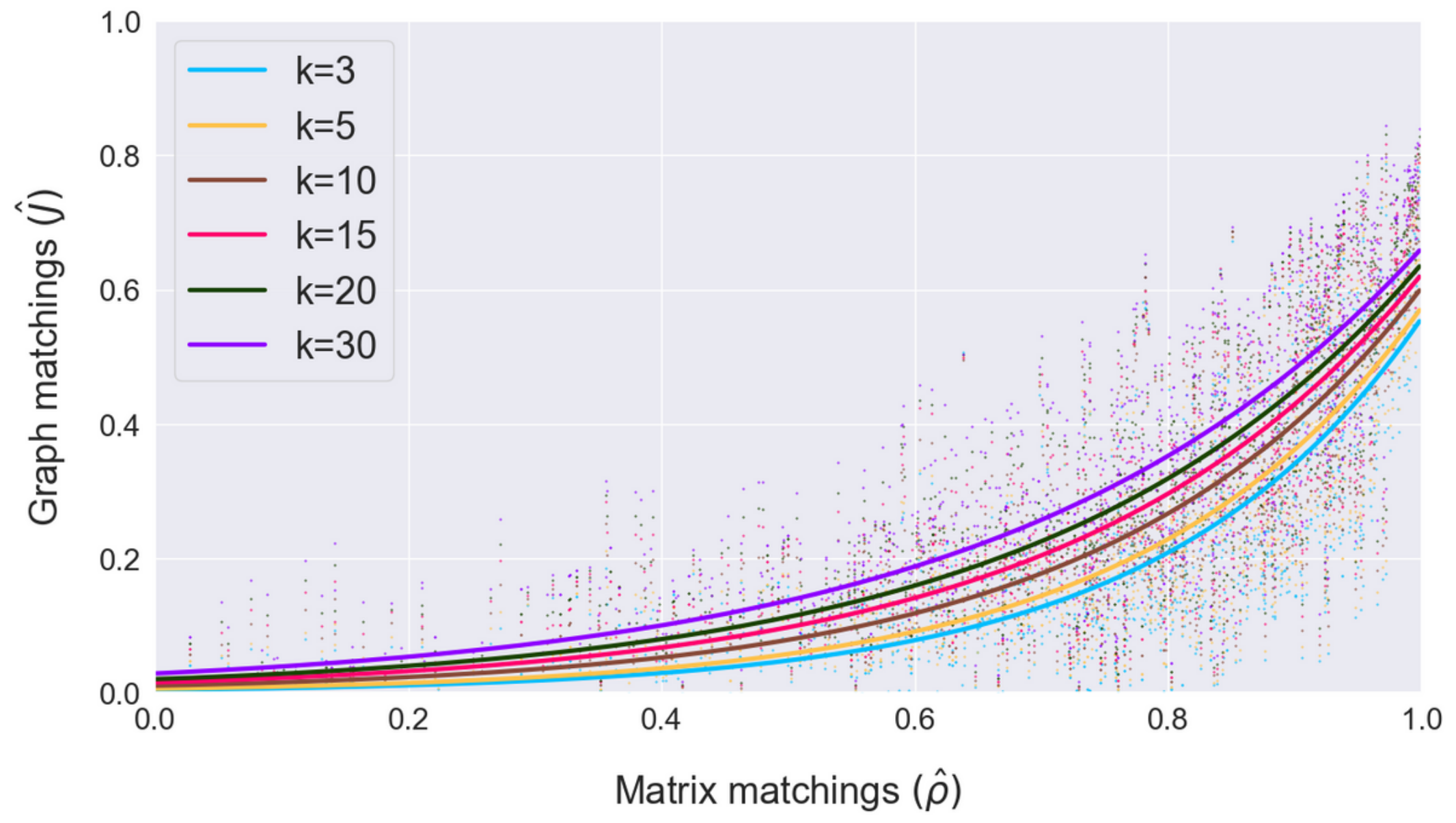


Figure 5: Relation between matrix matching and graph matching across all datasets

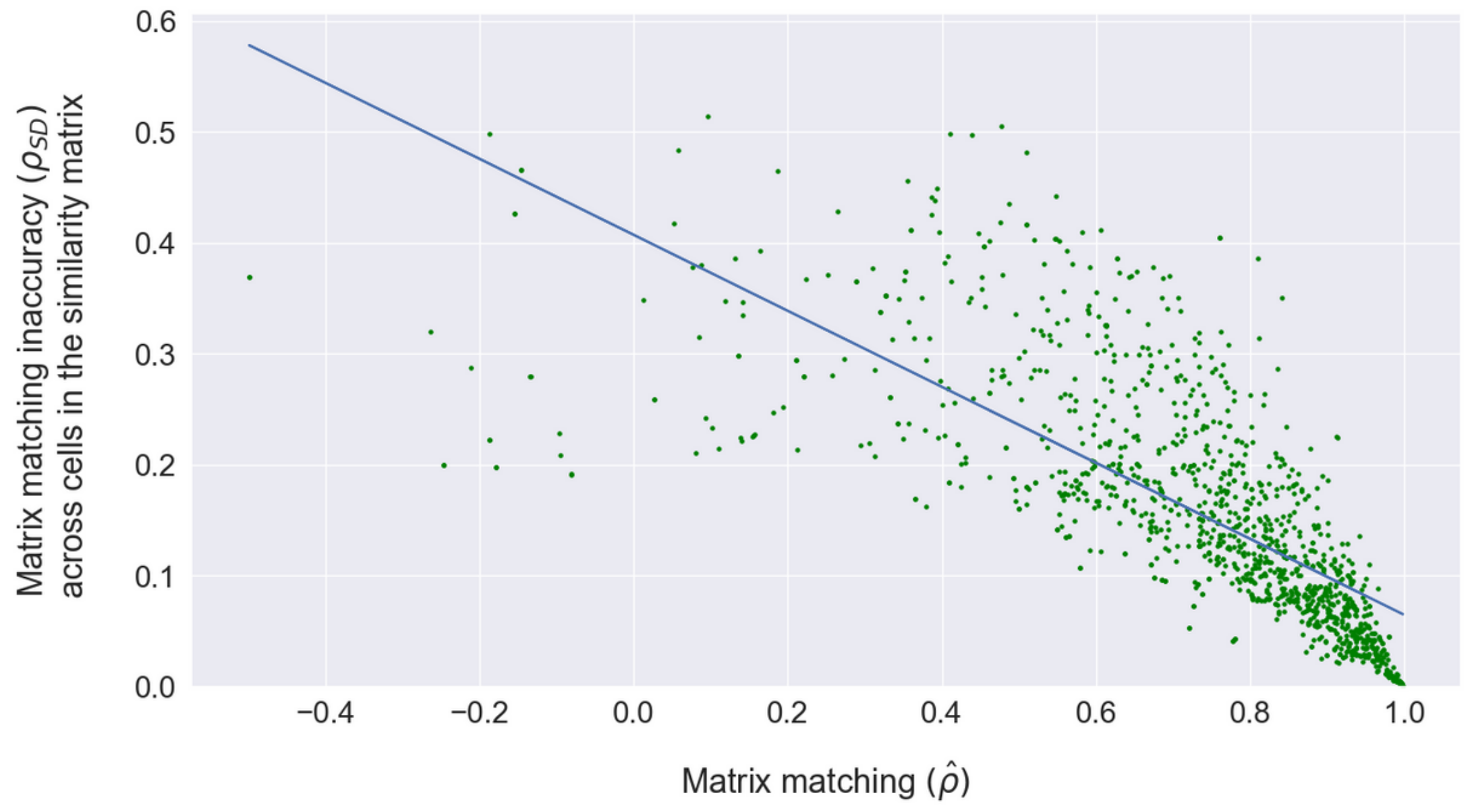


Figure 6: Relation between matrix matching inaccuracy and its overall quality

REFERENCES

[1] Bouland, G. A., Mahfouz, A., & Reinders, M. J. T. (2022). The rise of sparser single-cell RNAseq datasets; consequences and opportunities. https://doi.org/10.1101/2022.05.20.492823

[2] D. Eddelbuettel, Seamless R and C++ Integration with Rcpp. New York: Springer, 2013. ISBN 978-1-4614- 6867-7

[3] E. R. Watson, A. Mora, A. T. Fard, and J. C. Mar, "How does the structure of data impact cell-cell similarity? evaluating how structural properties influence the performance of proximity metrics in single cell rna-seq data," Briefings in bioinformatics, vol. 23, 11 2022.

TU Delft Delft University of Technology