

## Supplementary Material

for

# Effect of data preprocessing and machine learning hyperparameters on mass spectrometry imaging models

Wil Gardner<sup>1</sup>, David A. Winkler<sup>2,3,4</sup>, David L.J. Alexander<sup>5</sup>, Davide Ballabio<sup>6</sup>, Benjamin W. Muir<sup>7</sup> and Paul J. Pigram<sup>1,a)</sup>

<sup>1</sup>Centre for Materials and Surface Science and Department of Mathematical and Physical Sciences, La Trobe University, Melbourne, Victoria, Australia

<sup>2</sup>La Trobe Institute for Molecular Sciences, La Trobe University, Melbourne, Victoria, Australia

<sup>3</sup>Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, 3052, Australia

<sup>4</sup>Advanced Materials and Healthcare Technologies, School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, U.K

<sup>5</sup>CSIRO Data61, Clayton, VIC 3168, Australia

<sup>6</sup>Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, 20126, Milano, Italy

<sup>7</sup>CSIRO Manufacturing, Clayton, Victoria, 3168, Australia

a) Email: [p.pigram@latrobe.edu.au](mailto:p.pigram@latrobe.edu.au)

---

**ABSTRACT:** This supplemental information contains expanded mathematical descriptions, as well as the supplemental figures and tables, for the manuscript entitled: *The effect of preprocessing and hyperparameter selection on machine learning applied to mass spectrometry imaging data*. Mathematical descriptions are provided for the V-measure score. The figures and tables are: Schematic outlining the semi-synthetic data algorithm; example class membership maps generated using the semi-synthetic hyperspectral data generator; example results from the grid-search of the preprocessing-hyperparameter space, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data sets; and results from the various multiple linear regression models for the microarray ToF-SIMS data set, trained using the various preprocessing methods and hyperparameters, with and without interactions.

---

## V-measure score

V-measure<sup>1</sup> is an entropy-based measure of the overall performance of clustering algorithms. It is defined as the weighted harmonic mean of the homogeneity and completeness scores. Mathematically, this is given by

$$V_\beta = \frac{(1 + \beta) \cdot hc}{\beta h + c} \quad (7)$$

where  $h$  is the homogeneity score,  $c$  is the completeness score and  $\beta$  is a constant that controls the weighting of  $h$  and  $c$ .

The completeness score is a measure of how effectively the clustering has assigned a class (in this case, the polymer type) to a single cluster (in this case, a neuron on the SOM). This is defined as 1 if there is only one cluster, and otherwise given by

$$c = 1 - \frac{H(\tilde{K}|\tilde{C})}{H(\tilde{K})} \quad (8)$$

where

$$H(\tilde{K}|\tilde{C}) = - \sum_{c=1}^C \sum_{k=1}^K \frac{n_{c,k}}{n} \cdot \log\left(\frac{n_{c,k}}{n_c}\right), \quad (9)$$

$$H(\tilde{K}) = - \sum_{k=1}^K \frac{n_k}{n} \cdot \log\left(\frac{n_k}{n}\right). \quad (10)$$

Here,  $H(\tilde{K}|\tilde{C})$  is the conditional entropy of the set of clusters  $\tilde{K}$  given the set of classes  $\tilde{C}$ , and  $H(\tilde{K})$  is the entropy of the clusters. Further,  $n$  is the total number of samples,  $n_k$  and  $n_c$  are the numbers of samples corresponding to cluster  $k$  and class  $c$ , respectively, and  $n_{c,k}$  is the number of samples in both class  $c$  and cluster  $k$ . Note that we only sum over nonzero values for  $n_{c,k}$ .

Inversely, the homogeneity score measures how effectively the clustering has assigned a cluster to a single class. This is defined as 1 if there is only one class, and otherwise given by

$$h = 1 - \frac{H(\tilde{C}|\tilde{K})}{H(\tilde{C})} \quad (11)$$

where

$$H(\tilde{C}|\tilde{K}) = - \sum_{k=1}^K \sum_{c=1}^C \frac{n_{c,k}}{n} \cdot \log\left(\frac{n_{c,k}}{n_k}\right), \quad (12)$$

$$H(\tilde{C}) = - \sum_{c=1}^C \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right). \quad (13)$$

It can be observed that when the number of clusters approaches the number of samples in the data set (in this case, the number of pixels in the hyperspectral image), the homogeneity score approaches 1 and the completeness score approaches zero. The opposite is true as the number of clusters approaches 1. Therefore, the V-measure score attempts to avoid such extreme scenarios by using a harmonic mean of the two scores. The homogeneity score is more important than the completeness score for the SOM. This is because it is not necessarily undesirable for the SOM to assign multiple neurons to the same class, given its self-organizing and topology-preserving nature. As such, we only consider this score in our evaluations. Furthermore, to be consistent with other metrics used in our evaluation for which a smaller score is better, we convert the homogeneity to what we call heterogeneity, given simply as  $1 - h$ . In this form, a heterogeneity of zero is considered ideal.

## Figures and tables

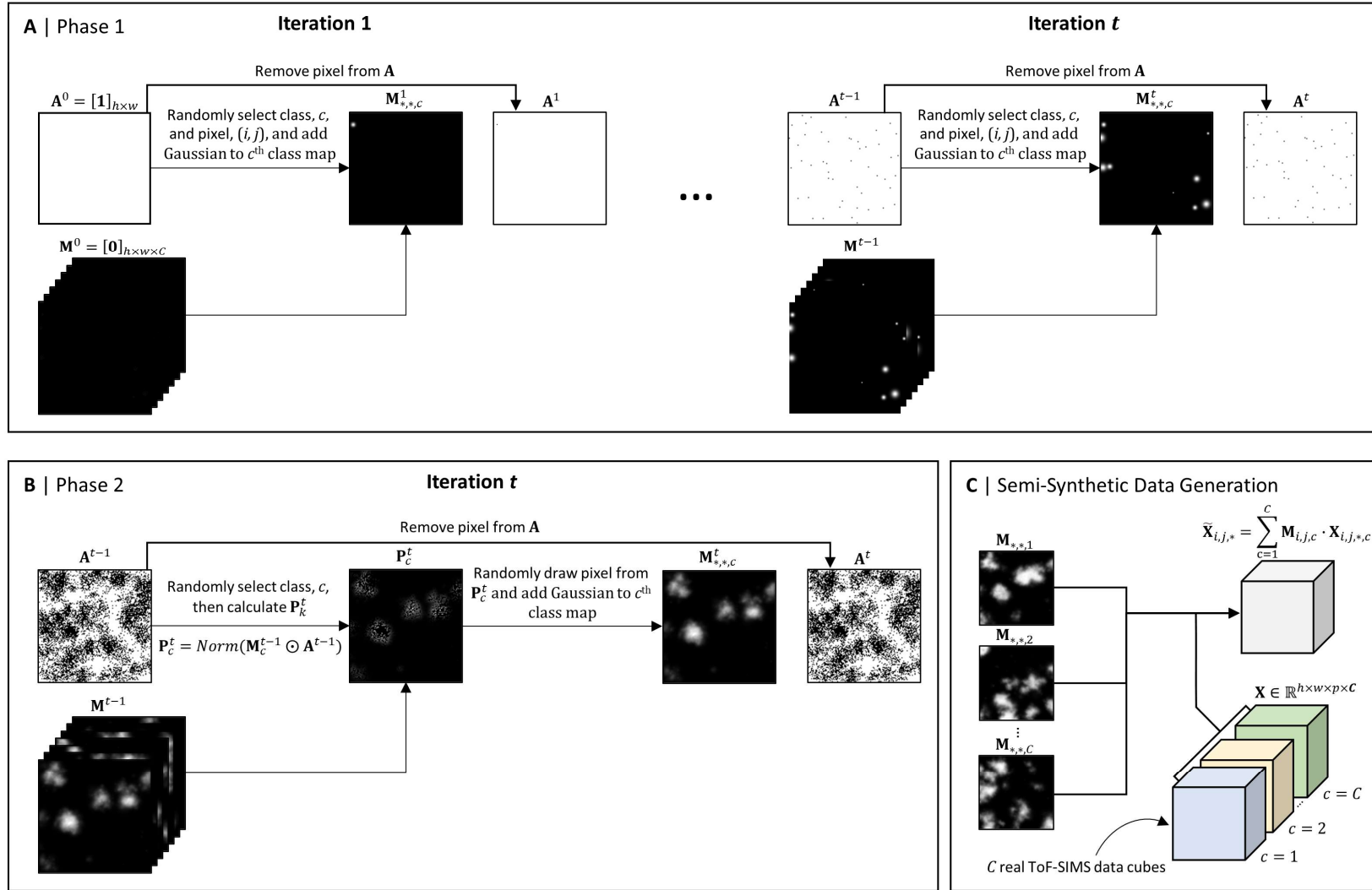


Fig S1. Schematic outlining the semi-synthetic data algorithm. In Phase 1 (A), class membership maps are constructed by randomly assigning pixels to each class. In Phase 2 (B), the maps are completed by assigning nearby pixels to the same class, thereby increasing spatial autocorrelation. At the semi-synthetic data generation step (C), the  $C$  class membership maps are used to construct a single semi-synthetic ToF-SIMS data cube from  $C$  real ToF-SIMS data cubes.

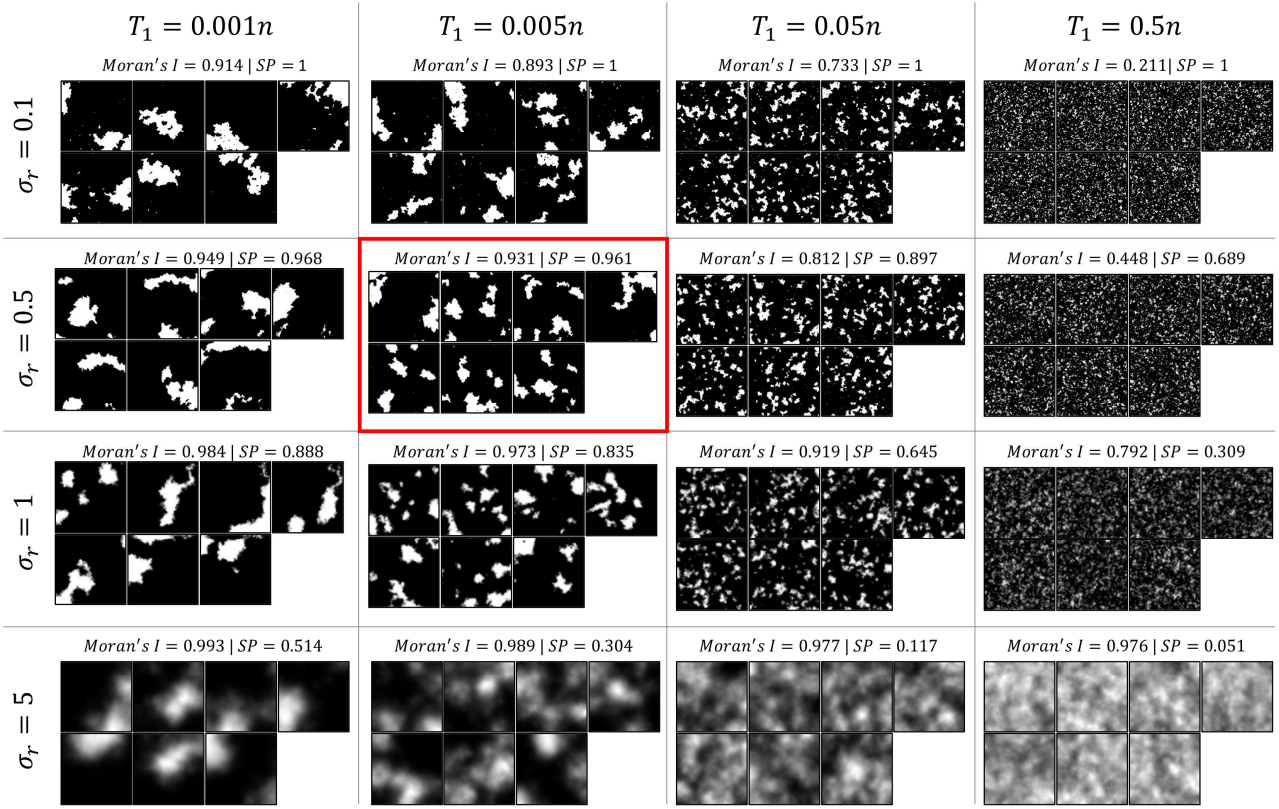


Fig S2. Example class membership maps generated using the semi-synthetic hyperspectral data (Eq. 1-6 and Fig S1). Total number of classes,  $C$ , was 7 and  $T_1$  is the number of pixels assigned during Phase 1 of the algorithm, as a fraction of the total number of pixels,  $n$ .  $\sigma_r$  represents the scale factor for the Rayleigh distribution, which is used to calculate a new standard deviation,  $\sigma^t$ , at each iteration,  $t$ . For each set of maps, we calculated the mean Moran's I value and the spectral purity,  $SP$ , defined in Eq. 2 in the main text.

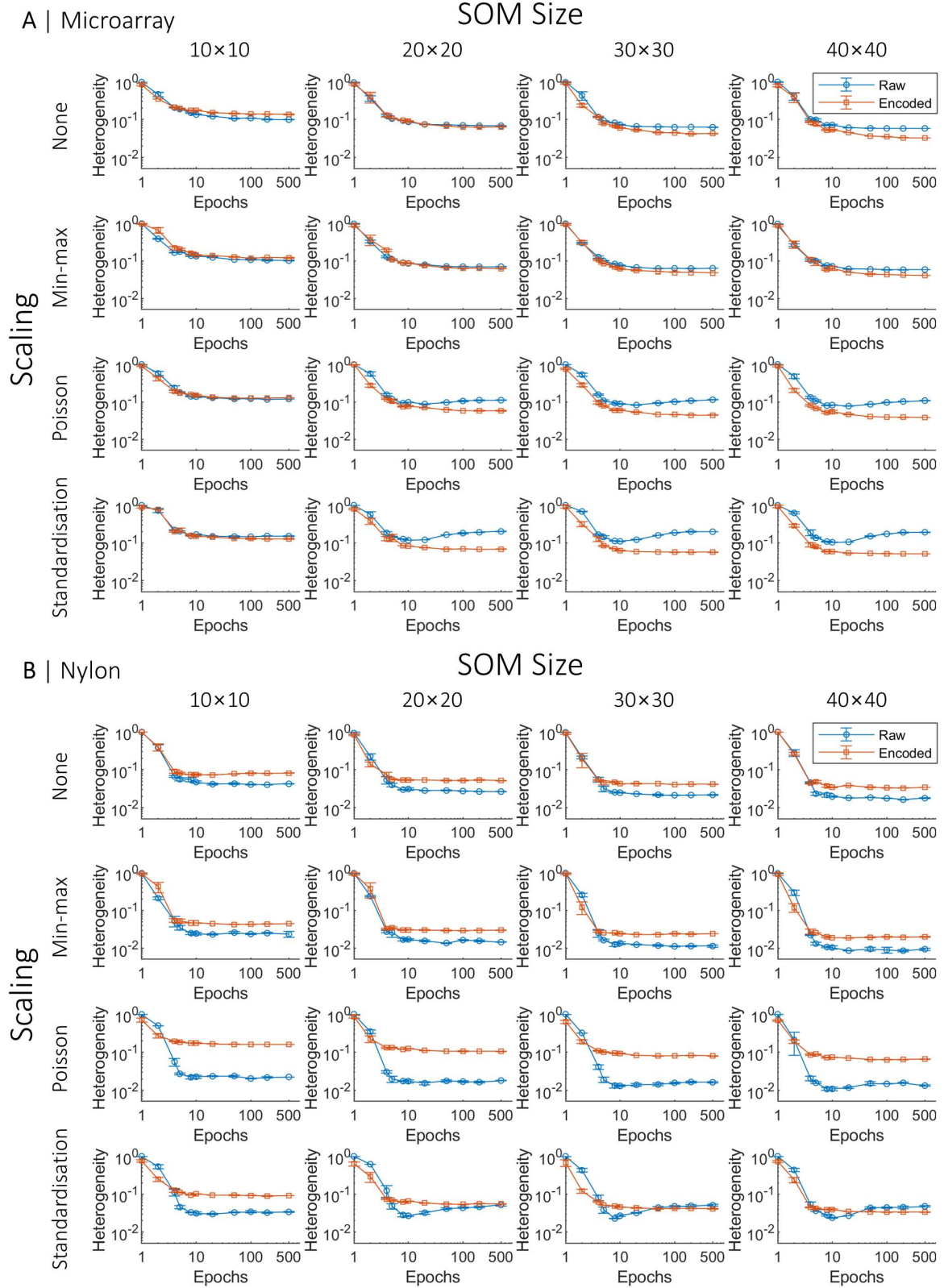


Fig S3. Heterogeneity score for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data not normalized to TIC. In each case, square neurons with a toroidal SOM were used. Each plot compares the heterogeneity score as a function of training epochs, using either raw data or data encoded to 100 features using the CNAE. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.



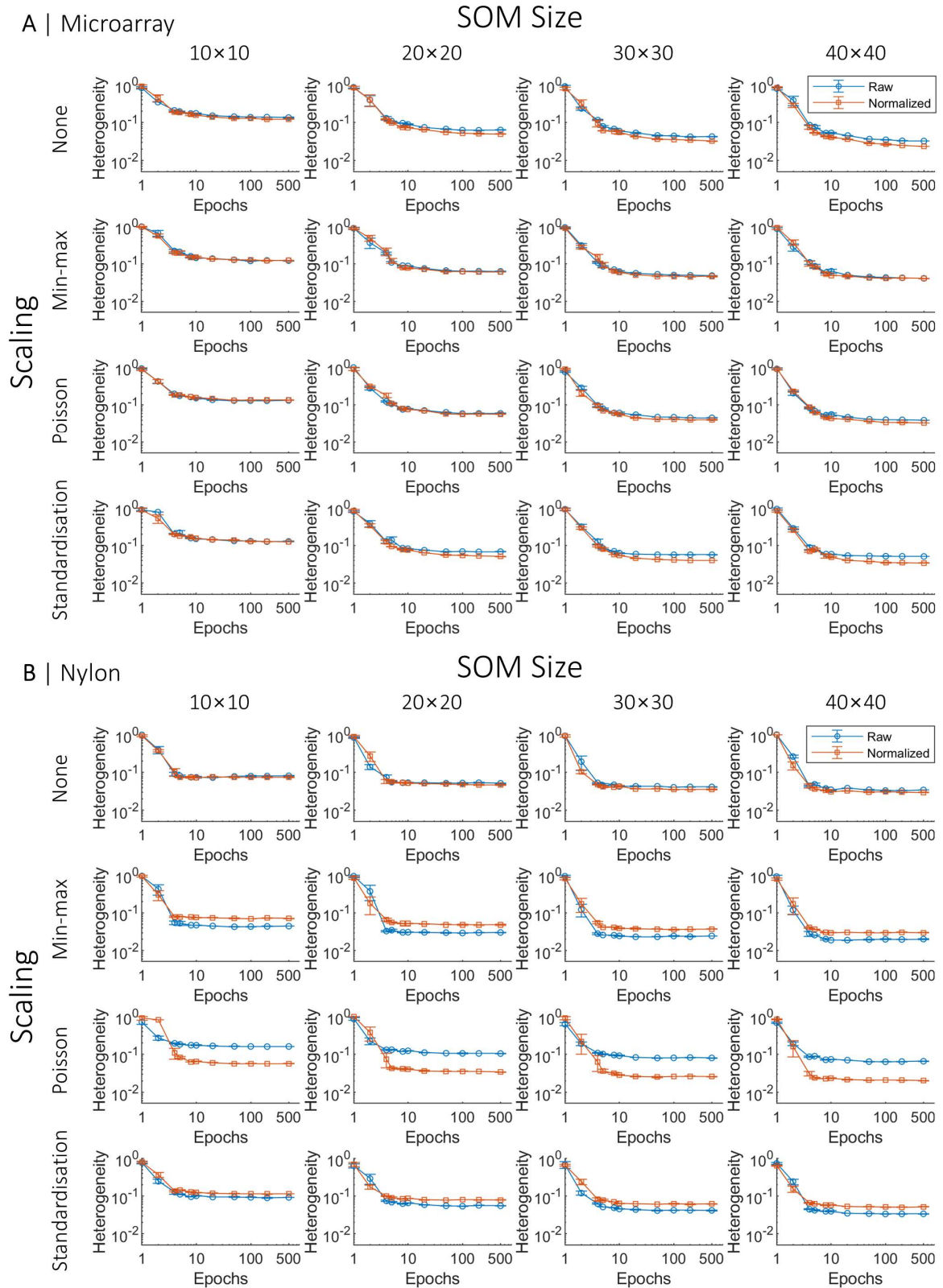


Fig S4. Heterogeneity score for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data encoded to 100 features by the CNNAE. In each case, square neurons with a toroidal SOM were used. Each plot compares the heterogeneity score as a function of training epochs, using either raw data or data normalized (per pixel) to TIC. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.



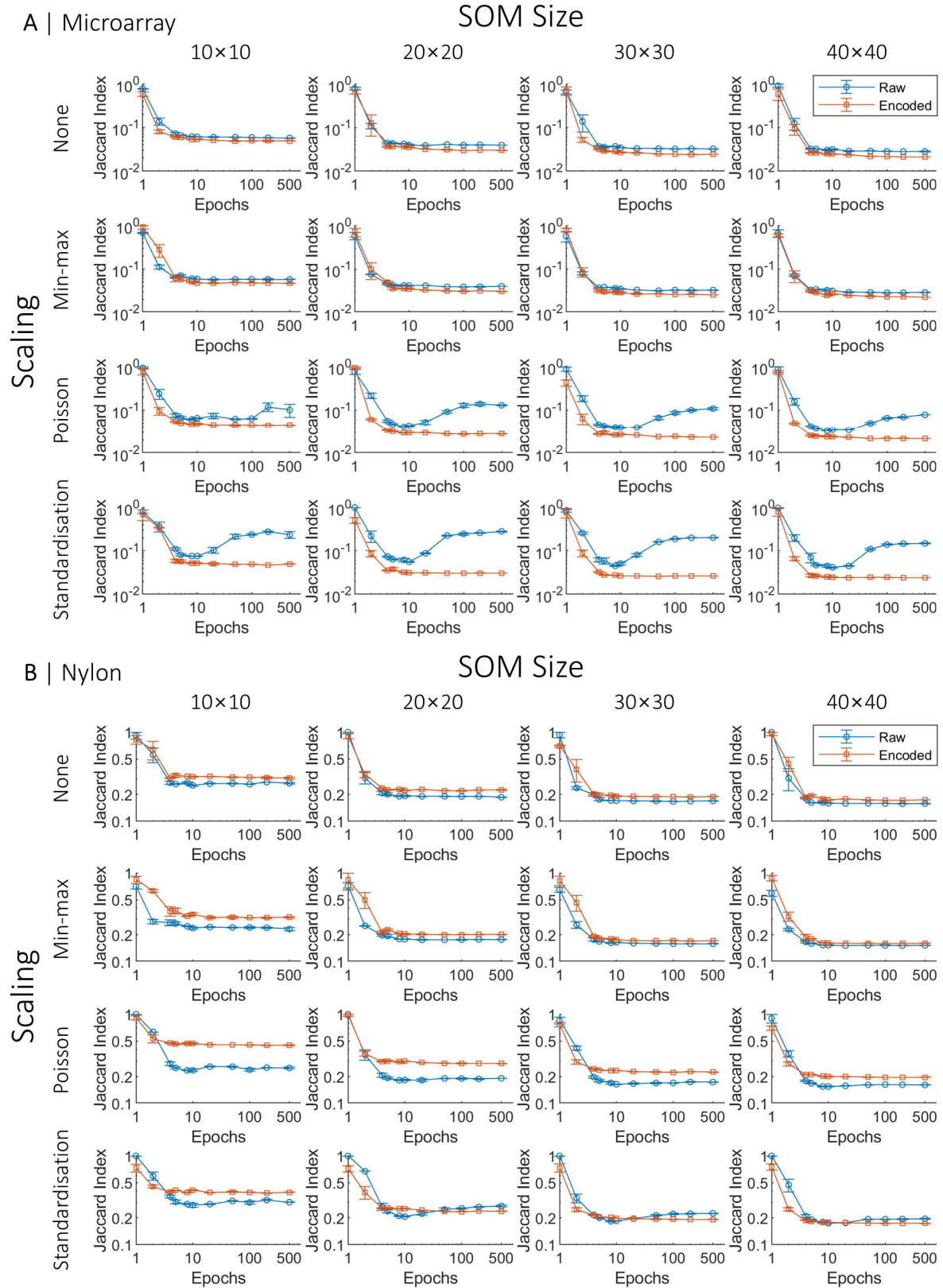


Fig S5. Jaccard index for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data not normalized to TIC. In each case, square neurons with a toroidal SOM were used. Each plot compares the Jaccard index as a function of training epochs, using either raw data or data encoded to 100 features using the CNNAE. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.

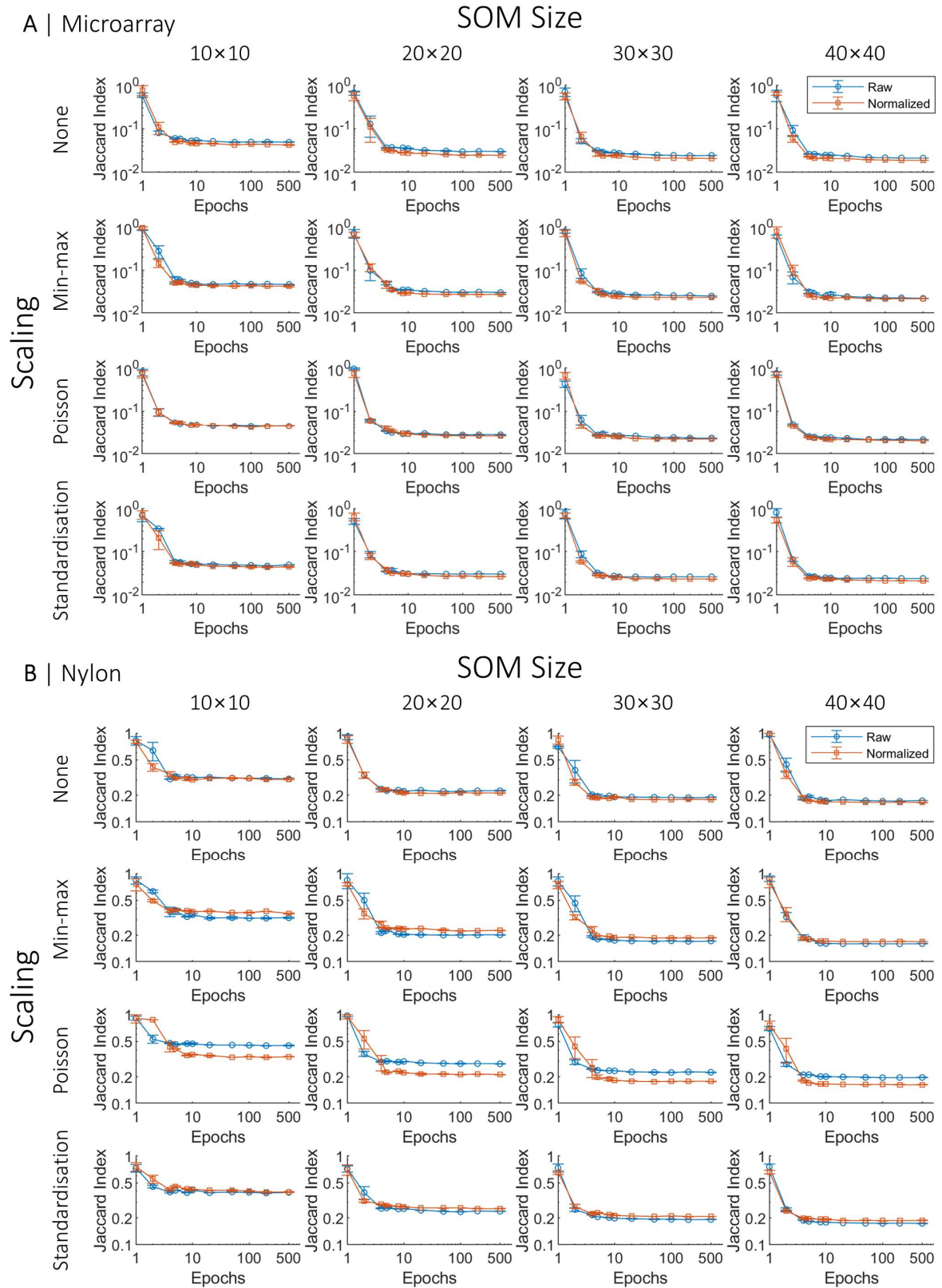


Fig S6. Jaccard index for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data encoded to 100 features by the CNNAE. In each case, square neurons with a toroidal SOM were used. Each plot compares the Jaccard index as a function of training epochs, using either raw data or data normalized (per pixel) to TIC. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.

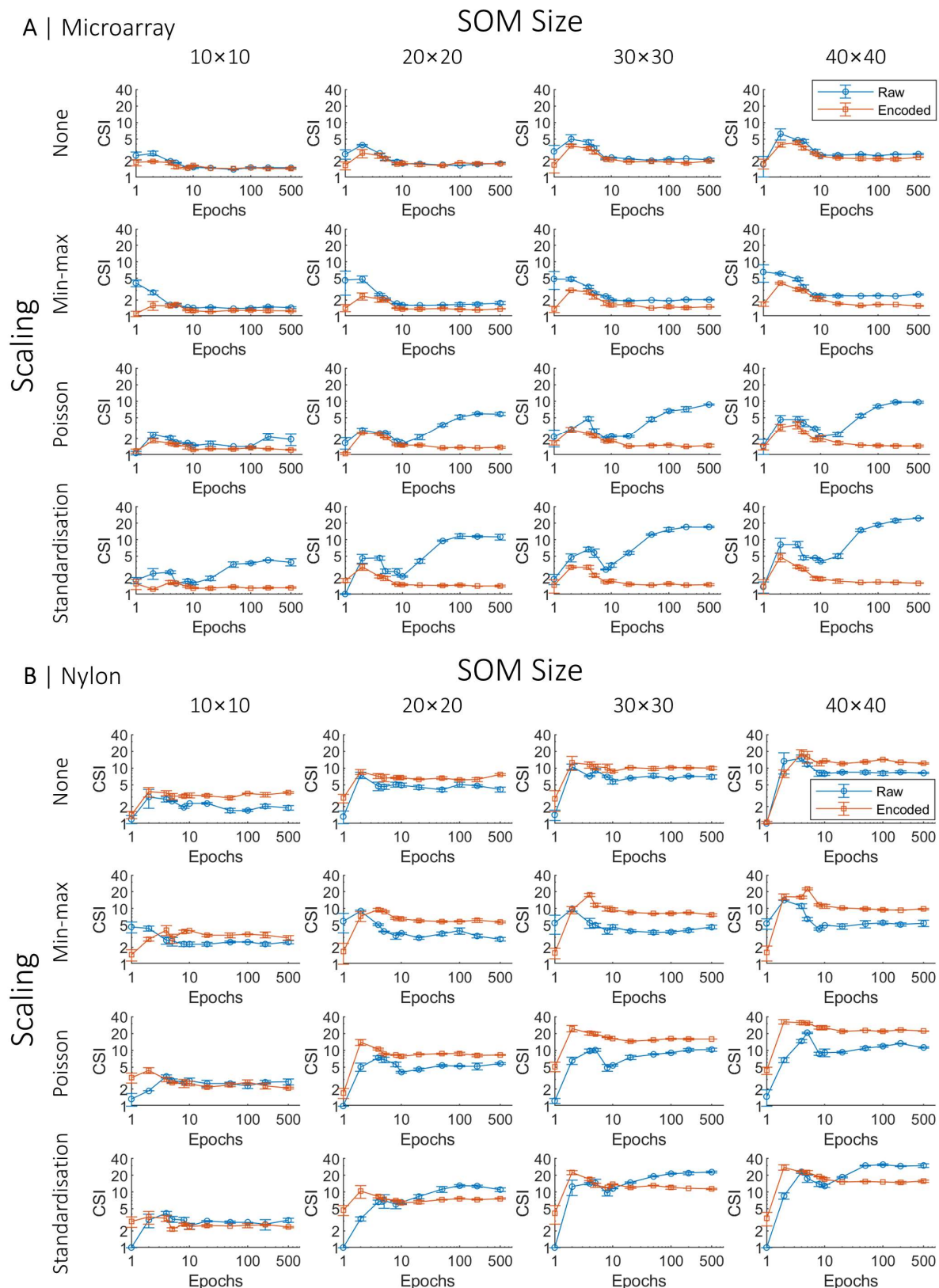


Fig S7. Class scatter index (CSI) for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data not normalized to TIC. In each case, square neurons with a toroidal SOM were used. Each plot compares the CSI as a function of training epochs, using either raw data or data encoded to 100 features using the CNNAE. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.



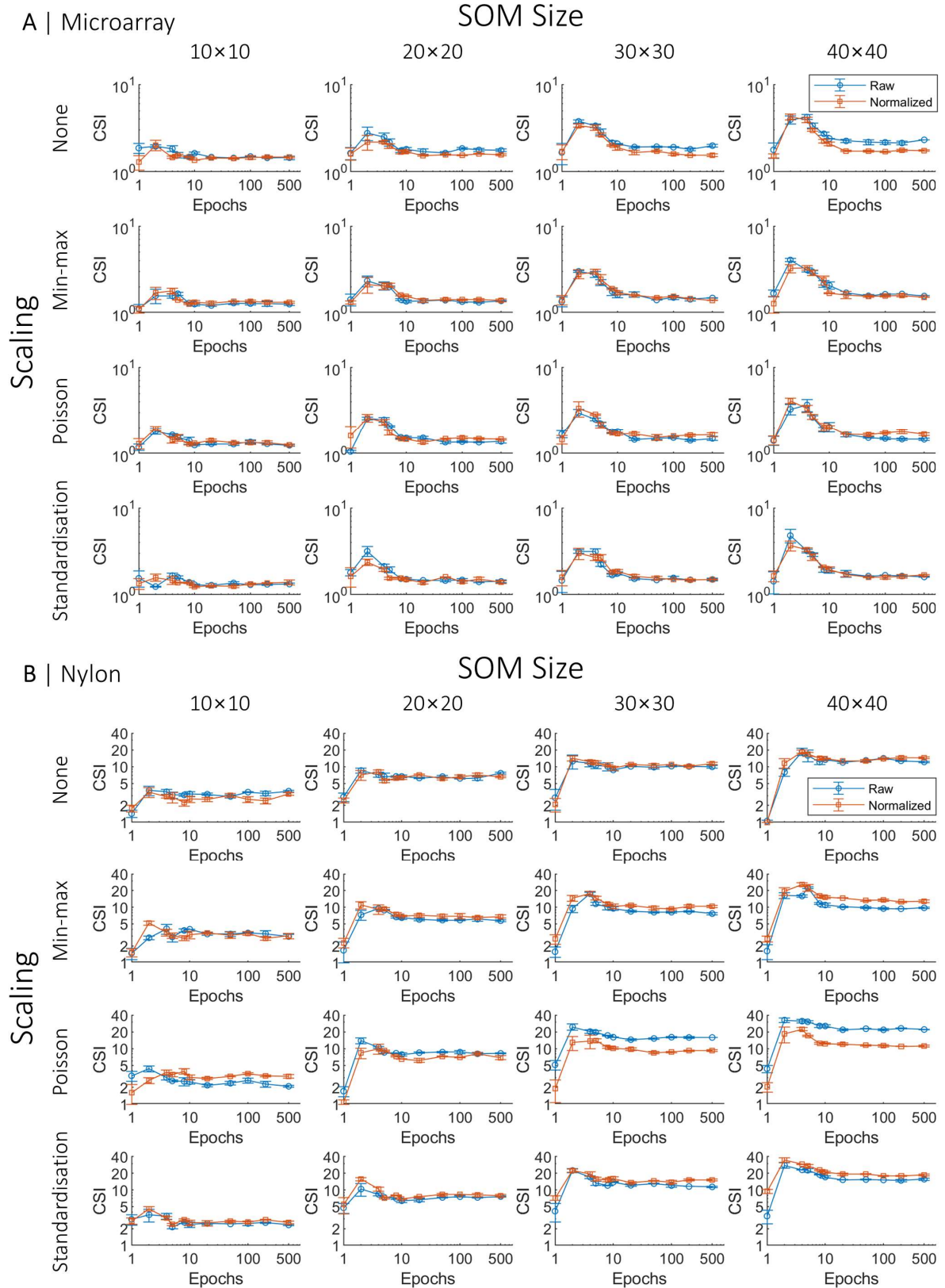


Fig S8. Class scatter index (CSI) for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data encoded to 100 features by the CNNAE. In each case, square neurons with a toroidal SOM were used. Each plot compares the CSI as a function of training epochs, using either raw data or data normalized (per pixel) to TIC. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.

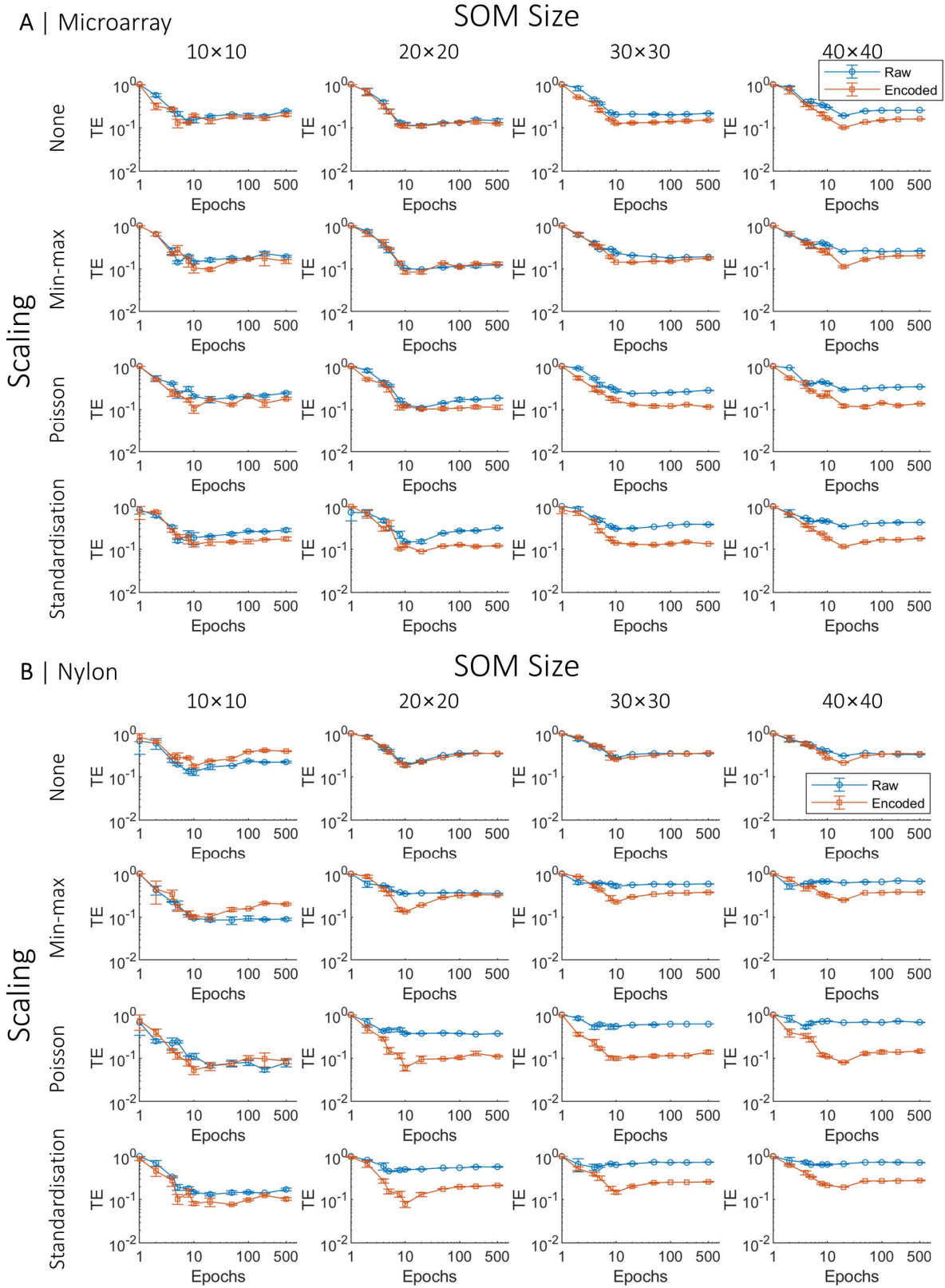


Fig S9. Topographic error (TE) for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data not normalized to TIC. In each case, square neurons with a toroidal SOM were used. Each plot compares the topographic error as a function of training epochs, using either raw data or data encoded to 100 features using the CNNAE. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.

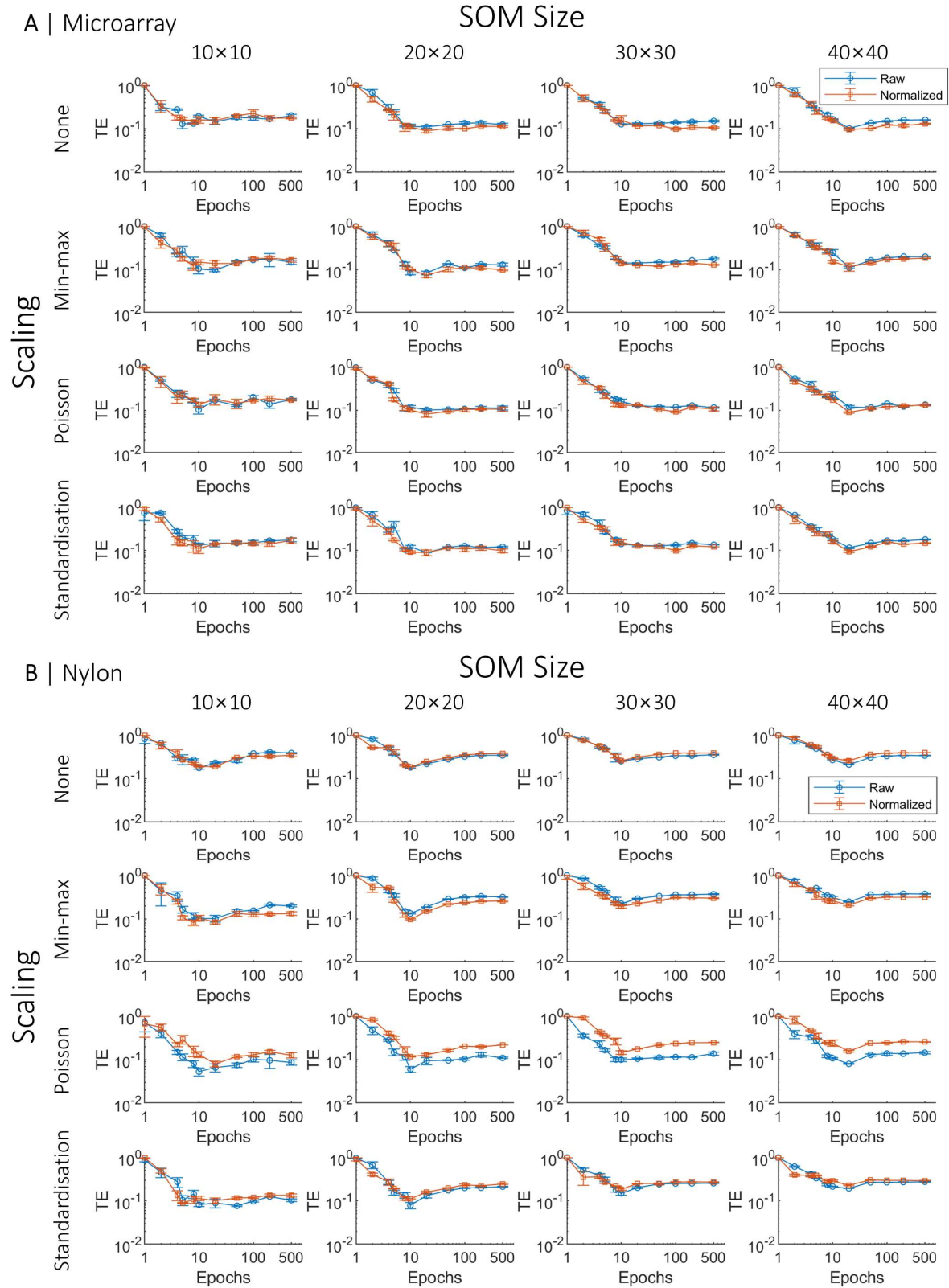


Fig S10. Topographic error (TE) for a grid-search of the preprocessing-hyperparameter space for a range of SOMs of different sizes, trained using data scaled using different methods, for the microarray (A) and semi-synthetic nylon (B) ToF-SIMS data encoded to 100 features by the CNNAE. In each case, square neurons with a toroidal SOM were used. Each plot compares the TE as a function of training epochs, using either raw data or data normalized (per pixel) to TIC. Error bars show standard deviation of 3 replicates, and the axis scales are logarithmic.





Table S3. Standardized regression coefficients from MLR models of the microarray data set, trained using the various preprocessing methods and hyperparameters, as well as their interactions. Bolded entries were statistically significant at  $p < 0.05$ . Stars represent significance levels:

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . Intercepts are shaded in grey, whereas coefficients are shaded using the color scheme shown. Note that coloring is relative, per model.

| Type                     | Metric            | Epochs | Adj R <sup>2</sup> | Intercept                   | TIC Norm         | Minmax            | Poisson          | Standard          | Encoded                                   | Toroidal       | Hexagon         | SOM Size         | TIC Norm Minmax | TIC Norm Poisson | TIC Norm Standard | TIC Norm Encoded | Minmax Encoded    | Poisson Encoded  | Standard Encoded |                  |          |
|--------------------------|-------------------|--------|--------------------|-----------------------------|------------------|-------------------|------------------|-------------------|---|----------------|-----------------|------------------|-----------------|------------------|-------------------|------------------|-------------------|------------------|------------------|------------------|----------|
| Class-Cluster Similarity | Heterogeneity     | 10     | 0.81               | 0.15***                     | -0.013***        | -0.015**          | -0.0031          | 0.015***          | 0.038***                                  | -0.0012        | -0.0038***      | -0.048***        | 0.0056          | 0.0034           | 0.00041           | NA               | -0.00020          | -0.011**         | -0.028***        | Better           |          |
|                          |                   | 100    | 0.89               | 0.10***                     | -0.0037          | -0.0073           | 0.022***         | 0.071***          | 0.053***                                  | 0.0057*        | -0.0012         | -0.024***        | 0.0083**        | 0.0053           | -0.00071          | -0.0062**        | 0.00014           | -0.031***        | -0.093***        |                  |          |
|                          |                   | 500    | 0.91               | 0.098***                    | -0.0029          | -0.0090*          | 0.025***         | 0.077***          | 0.058***                                  | 0.0061*        | -0.0018         | -0.022***        | 0.0081**        | 0.0048           | -0.0020           | -0.0072***       | 0.00029           | -0.037***        | -0.11***         |                  |          |
|                          | Jaccard Index     | 10     | 0.85               | 0.068***                    | -0.0059***       | -0.0032*          | 0.0021           | 0.018***          | -0.011***                                 | NA             | -0.00057        | -0.023***        | 0.0010          | 0.0017           | 0.002*            | -0.00083         | 0.00037           | -0.0046***       | -0.015***        |                  |          |
|                          |                   | 100    | 0.94               | 0.077***                    | -0.0026          | -0.0022           | 0.061***         | 0.20***           | -0.043***                                 | 0.0061*        | -0.0052         | -0.031***        | 0.0016          | -0.0019          | -0.010***         | 0.0061**         | 0.00063           | -0.051***        | -0.15***         |                  |          |
|                          |                   | 500    | 0.94               | 0.073***                    | 0.000070         | -0.0025           | 0.061***         | 0.20***           | -0.036***                                 | NA             | -0.0018         | -0.027***        | 0.0018          | 0.00092          | -0.011***         | 0.0053*          | 0.00024           | -0.057***        | -0.16***         |                  |          |
| SOM Topology             | CSI               | 10     | 0.84               | 0.82***                     | -0.089           | -0.078            | -0.089           | 0.21*             | 0.85***                                   | -0.043         | -0.0085         | 1.1***           | 0.14*           | 0.32***          | 0.50***           | -0.37***         | -0.12             | -0.31***         | -1.3***          | Neutral          |          |
|                          |                   | 100    | 0.92               | -1.1**                      | -0.11            | -0.018            | 1.4***           | 4.9***            | 5.2***                                    | 0.082          | -0.40           | 2.5***           | 0.098           | 0.79**           | 1.1***            | -0.88***         | -0.19             | -4.1***          | -12***           |                  |          |
|                          |                   | 500    | 0.90               | -2.1***                     | -0.21            | -0.035            | 1.3*             | 4.2***            | 7.1***                                    | -0.026         | -0.098          | 3.5***           | 0.090           | 0.86*            | 1.5***            | -1.1***          | -0.14             | -4.9***          | -14***           |                  |          |
|                          | Topographic Error | 10     | 0.84               | 0.037***                    | 0.027***         | -0.022            | 0.010            | 0.0051            | 0.096***                                  | 0.019***       | 0.13***         | 0.13***          | NA              | NA               | NA                | -0.030***        | -0.027**          | -0.056***        | -0.074***        |                  |          |
|                          |                   | 100    | 0.86               | 0.15***                     | 0.041***         | -0.042***         | 0.013            | 0.068***          | 0.010                                     | 0.025***       | 0.12***         | 0.048***         | NA              | NA               | NA                | -0.027***        | 0.024**           | -0.028***        | -0.096***        |                  |          |
|                          |                   | 500    | 0.86               | 0.17***                     | NA               | -0.060***         | 0.0042           | 0.090***          | 0.0049                                    | 0.023**        | 0.11***         | 0.042***         | NA              | NA               | NA                | NA               | 0.026***          | -0.040***        | -0.11***         |                  |          |
|                          |                   |        |                    | Hyperparameter Interactions |                  |                   |                  |                   | Preprocessing-Hyperparameter Interactions |                |                 |                  |                 |                  |                   |                  |                   |                  |                  |                  | Worse    |
| Type                     | Metric            | Epochs | Toroidal Hexagon   | Toroidal SOM Size           | Hexagon SOM Size | TIC Norm Toroidal | TIC Norm Hexagon | TIC Norm SOM Size | Minmax Toroidal                           | Minmax Hexagon | Minmax SOM Size | Poisson Toroidal | Poisson Hexagon | Poisson SOM Size | Standard Toroidal | Standard Hexagon | Standard SOM Size | Encoded Toroidal | Encoded Hexagon  | Encoded SOM Size |          |
| Clustering Performance   | Heterogeneity     | 10     | NA                 | NA                          | NA               | NA                | NA               | 0.0028            | NA  | NA             | 0.011***        | NA               | NA              | 0.0093**         | NA                | NA               | 0.012***          | NA               | NA               | -0.036***        |          |
|                          |                   | 100    | NA                 | -0.0040                     | NA               | NA                | NA               | NA                | NA  | NA             | 0.0067*         | NA               | NA              | 0.0092**         | NA                | NA               | 0.026***          | NA               | NA               | -0.051***        |          |
|                          |                   | 500    | NA                 | -0.0038                     | NA               | NA                | NA               | NA                | NA  | NA             | 0.0083**        | NA               | NA              | 0.012***         | NA                | NA               | 0.033***          | NA               | NA               | -0.056***        |          |
|                          | Jaccard Index     | 10     | NA                 | NA                          | NA               | NA                | NA               | 0.0019*           | NA  | NA             | 0.0022*         | NA               | NA              | 0.00070          | NA                | NA               | -0.0031**         | NA               | NA               | 0.0032***        |          |
|                          |                   | 100    | NA                 | -0.0031                     | 0.0021           | -0.0030           | NA               | -0.0022           | NA  | NA             | 0.0014          | NA               | NA              | -0.0076**        | NA                | NA               | -0.040***         | NA               | 0.0030           | 0.026***         |          |
|                          |                   | 500    | NA                 | NA                          | NA               | NA                | NA               | NA                | 0.005768**                                | NA             | -0.00019        | 0.0019           | NA              | -0.0056          | -0.0022           | NA               | -0.0034           | -0.032***        | NA               | 0.0041           | 0.020*** |
| SOM Topology             | CSI               | 10     | 0.30***            | NA                          | NA               | -0.047            | -0.079           | 0.23***           | NA  | NA             | -0.10           | NA               | NA              | 0.019            | NA                | NA               | 0.56***           | 0.11*            | NA               | -0.85***         |          |
|                          |                   | 100    | 0.49*              | NA                          | 0.31             | NA                | NA               | 0.43*             | NA  | NA             | -0.12           | NA               | NA              | 1.9***           | NA                | NA               | 5.4***            | -0.20            | NA               | -4.4***          |          |
|                          |                   | 500    | 0.48               | NA                          | NA               | NA                | NA               | NA                | 0.61*                                     | NA             | NA              | -0.15            | NA              | NA               | 2.7***            | NA               | NA                | 7.9***           | NA               | NA               | -6.1***  |
|                          | Topographic Error | 10     | NA                 | NA                          | -0.023***        | NA                | NA               | NA                | NA  | NA             | 0.047***        | NA               | NA              | 0.051***         | NA                | NA               | 0.068***          | 0.038***         | 0.027***         | -0.13***         |          |
|                          |                   | 100    | 0.0071             | -0.016**                    | 0.011*           | NA                | NA               | -0.019***         | NA  | 0.0034         | 0.026***        | NA               | -0.0026         | 0.015            | NA                | -0.018*          | 0.041***          | 0.0087           | 0.019***         | -0.067***        |          |
|                          |                   | 500    | 0.0098             | -0.022***                   | 0.012*           | NA                | NA               | NA                | NA  | -0.0058        | 0.040***        | NA               | -0.011          | 0.027***         | NA                | -0.024**         | 0.034***          | 0.019***         | 0.028***         | -0.074***        |          |

Table S4. Standardized regression coefficients from MLR models of the nylon data set, trained using the various preprocessing methods and hyperparameters, as well as their interactions. Bolded entries were statistically significant at  $p < 0.05$ . Stars represent significance levels: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . Intercepts are shaded in grey, whereas coefficients are shaded using the color scheme shown. Note that coloring is relative, per model.

|                          |                   |        |                             |                   | Preprocessing                             |   |                  |                   |                 | Hyperparameters |                 |                  | Preprocessing Interactions |                  |                   |                  |                   |                  |                  |                  |  |  |  |  |  |
|--------------------------|-------------------|--------|-----------------------------|-------------------|---|---|------------------|-------------------|-----------------|-----------------|-----------------|------------------|----------------------------|------------------|-------------------|------------------|-------------------|------------------|------------------|------------------|--|--|--|--|--|
| Type                     | Metric            | Epochs | Adj R <sup>2</sup>          | Intercept         | TIC Norm                                  | Minmax                                    | Poisson          | Standard          | Encoded         | Toroidal        | Hexagon         | SOM Size         | TIC Norm Minmax            | TIC Norm Poisson | TIC Norm Standard | TIC Norm Encoded | Minmax Encoded    | Poisson Encoded  | Standard Encoded |                  |  |  |  |  |  |
| Class-Cluster Similarity | Heterogeneity     | 10     | 0.85                        | 0.044**           | 0.0034                                    | -0.029***                                 | 0.0067           | -0.0075*          | 0.057***        | -0.0024**       | NA              | -0.013***        | 0.012***                   | -0.034***        | 0.016***          | -0.014***        | 0.0038            | 0.040***         | 0.022***         | Better           |  |  |  |  |  |
|                          |                   | 100    | 0.82                        | 0.040***          | 0.0032                                    | -0.028***                                 | 0.0072*          | 0.0019            | 0.059***        | -0.0018         | NA              | -0.012***        | 0.011***                   | -0.031***        | 0.016***          | -0.014***        | 0.0015            | 0.031***         | 0.0029           |                  |  |  |  |  |  |
|                          |                   | 500    | 0.82                        | 0.040***          | 0.0032                                    | -0.028***                                 | 0.0071*          | 0.0030            | 0.060***        | -0.0021         | NA              | -0.012***        | 0.011***                   | -0.032***        | 0.016***          | -0.014***        | 0.0022            | 0.031***         | 0.00038          |                  |  |  |  |  |  |
|                          | Jaccard Index     | 10     | 0.82                        | 0.27***           | NA  | -0.0020                                   | 0.014            | 0.051***          | 0.10***         | -0.014*         | NA              | -0.058***        | NA                         | NA               | NA                | NA               | 0.018**           | 0.053***         | 0.025***         |                  |  |  |  |  |  |
|                          |                   | 100    | 0.82                        | 0.27***           | NA  | -0.0056                                   | 0.017*           | 0.083***          | 0.089***        | -0.0065*        | NA              | -0.059***        | NA                         | NA               | NA                | NA               | 0.015**           | 0.042***         | -0.0086          |                  |  |  |  |  |  |
|                          |                   | 500    | 0.82                        | 0.26***           | NA  | -0.0064                                   | 0.017*           | 0.085***          | 0.087***        | -0.0062*        | NA              | -0.058***        | NA                         | NA               | NA                | NA               | 0.017**           | 0.042***         | -0.012*          |                  |  |  |  |  |  |
| SOM Topology             | CSI               | 10     | 0.88                        | -0.12             | 1.2**                                     | 0.53                                      | -0.13            | -1.4**            | -1.3**          | 0.14            | -0.42           | 4.7***           | 0.24                       | -2.5***          | 1.6***            | -2.4***          | 1.6***            | 2.2***           | -1.7***          | Neutral          |  |  |  |  |  |
|                          |                   | 100    | 0.88                        | -2.3***           | 1.1*                                      | 1.3*                                      | 0.83             | 2.0**             | 2.6***          | 0.27            | -0.51*          | 7.2***           | 0.58                       | -2.1***          | 0.93*             | -2.2***          | 0.85*             | -0.018           | -8.4***          |                  |  |  |  |  |  |
|                          |                   | 500    | 0.88                        | -2.6***           | 1.3*                                      | 1.4*                                      | 1.1              | 1.4*              | 3.1***          | 0.44            | -0.41           | 7.2***           | 0.60                       | -2.4***          | 1.1*              | -2.0***          | 0.37              | -0.50            | -9.2***          |                  |  |  |  |  |  |
|                          | Topographic Error | 10     | 0.92                        | -0.042**          | 0.028*                                    | 0.011                                     | 0.040*           | 0.085***          | 0.27***         | 0.0087          | 0.13***         | 0.26***          | -0.012                     | 0.033**          | 0.031**           | -0.038***        | -0.22***          | -0.26***         | -0.31***         |                  |  |  |  |  |  |
|                          |                   | 100    | 0.91                        | 0.13***           | 0.022                                     | -0.16***                                  | -0.17***         | -0.073***         | 0.33***         | 0.024           | 0.12***         | 0.18***          | -0.030*                    | 0.042***         | 0.0061            | -0.029***        | -0.16***          | -0.27***         | -0.31***         |                  |  |  |  |  |  |
|                          |                   | 500    | 0.90                        | 0.13***           | 0.036**                                   | -0.17***                                  | -0.18***         | -0.075***         | 0.35***         | 0.038**         | 0.12***         | 0.17***          | -0.041***                  | 0.033**          | 0.0059            | -0.036***        | -0.16***          | -0.28***         | -0.33***         |                  |  |  |  |  |  |
|                          |                   |        |                             |                   | Preprocessing-Hyperparameter Interactions |   |                  |                   |                 |                 |                 |                  |                            |                  |                   |                  |                   |                  |                  |                  |  |  |  |  |  |
| Type                     | Metric            | Epochs | Hyperparameter Interactions |                   |   | Preprocessing-Hyperparameter Interactions |                  |                   |                 |                 |                 |                  |                            |                  |                   |                  |                   |                  |                  |                  |  |  |  |  |  |
|                          |                   |        | Toroidal Hexagon            | Toroidal SOM Size | Hexagon SOM Size                          | TIC Norm Toroidal                         | TIC Norm Hexagon | TIC Norm SOM Size | Minmax Toroidal | Minmax Hexagon  | Minmax SOM Size | Poisson Toroidal | Poisson Hexagon            | Poisson SOM Size | Standard Toroidal | Standard Hexagon | Standard SOM Size | Encoded Toroidal | Encoded Hexagon  | Encoded SOM Size |  |  |  |  |  |
| Clustering Performance   | Heterogeneity     | 10     | NA                          | NA                | NA  | NA  | NA               | 0.0022            | NA              | NA              | 0.0067**        | NA               | NA                         | -0.0044          | NA                | NA               | -0.0019           | NA               | NA               | -0.027***        |  |  |  |  |  |
|                          |                   | 100    | NA                          | NA                | NA  | NA  | NA               | 0.0029            | -0.0013         | NA              | 0.0083***       | -0.0017          | NA                         | -0.0022          | -0.0046           | NA               | 0.0045            | 0.0037*          | NA               | -0.029***        |  |  |  |  |  |
|                          |                   | 500    | NA                          | NA                | NA  | NA  | NA               | 0.0028            | -0.0011         | NA              | 0.0085***       | -0.0016          | NA                         | -0.0017          | -0.0047           | NA               | 0.0062*           | 0.0037*          | NA               | -0.030***        |  |  |  |  |  |
|                          | Jaccard Index     | 10     | NA                          | 0.0056            | NA  | NA  | NA               | NA                | NA              | NA              | -0.0085         | NA               | NA                         | -0.024***        | NA                | NA               | -0.032***         | 0.0049           | NA               | -0.068***        |  |  |  |  |  |
|                          |                   | 100    | NA                          | NA                | NA  | NA  | NA               | NA                | NA              | NA              | -0.0051         | NA               | NA                         | -0.020***        | NA                | NA               | -0.035***         | 0.0074           | NA               | -0.061***        |  |  |  |  |  |
|                          |                   | 500    | NA                          | NA                | NA  | NA  | NA               | NA                | NA              | NA              | -0.0049         | NA               | NA                         | -0.020***        | NA                | NA               | -0.034***         | 0.0073           | NA               | -0.060***        |  |  |  |  |  |
| SOM Topology             | CSI               | 10     | 1.4***                      | -1.0***           | NA  | -0.27                                     | NA               | 0.94***           | -0.074          | NA              | -2.2***         | -0.49            | NA                         | 1.2***           | -0.66             | NA               | 4.4***            | 0.29             | 0.43             | 4.0***           |  |  |  |  |  |
|                          |                   | 100    | 1.6***                      | -1.5***           | NA  | NA  | NA               | 0.69*             | -0.26           | NA              | -3.0***         | -0.80            | NA                         | 1.7***           | -1.8***           | NA               | 7.3***            | 1.2***           | 0.76*            | NA               |  |  |  |  |  |
|                          |                   | 500    | 1.5***                      | -1.4***           | NA  | -0.45                                     | NA               | 0.59              | -0.31           | NA              | -3.1***         | -0.78            | NA                         | 1.7***           | -1.4**            | NA               | 8.2***            | 1.1***           | 0.34             | NA               |  |  |  |  |  |
|                          | Topographic Error | 10     | NA                          | NA                | -0.027***                                 | -0.016*                                   | NA               | 0.018*            | 0.0093          | NA              | 0.15***         | 0.036**          | NA                         | 0.11***          | 0.023*            | NA               | 0.13***           | 0.056***         | 0.031***         | -0.27***         |  |  |  |  |  |
|                          |                   | 100    | NA                          | -0.030***         | -0.019*                                   | NA  | NA               | 0.018*            | -0.024*         | 0.012           | 0.25***         | -0.0087          | 0.021                      | 0.23***          | 0.0025            | 0.0035           | 0.25***           | 0.060***         | 0.025**          | -0.29***         |  |  |  |  |  |
|                          |                   | 500    | NA                          | -0.037***         | -0.020*                                   | NA  | NA               | 0.014             | -0.029*         | 0.012           | 0.26***         | -0.0074          | 0.020                      | 0.24***          | -0.0062           | 0.0034           | 0.26***           | 0.055***         | 0.020*           | -0.29***         |  |  |  |  |  |

## References

<sup>1</sup>A. Rosenberg and J. Hirschberg. *V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure*; 2007.