# Exploiting the Structures of the U-Matrix

Jörn Lötsch[1,2] and Alfred Ultsch[3]

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany

[2] Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany

[3] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, D-35032 Marburg, Germany

## Abstract

The U-matrix has become a standard visualization of self-organizing feature maps (SOM). Here we present the abstract U-matrix, which formalizes the structures on a U-matrix such that distance calculations between best-matching units w.r.t. the height structures of a U-matrix are precisely defined (U-cell distance). This enables the assessment of the topological correctness of the SOM and the implementation of clustering algorithms that take the structures seen on the U-matrix into account. A weighted Delaunay graph of the U-cell distances allows the calculation of a dendrogram corresponding to the structures of the U-matrix. The method is shown to detect and visualize meaningful cluster structures on difficult artificial and real-life data.

## 1. Introduction

Self-organizing feature maps (SOM) [2] are often visualized by using the U-matrix [3]. A trained SOM represents a topology pre-serving mapping of n high-dimensional data points $x_i \in R^D$ onto a two dimensional grid of neurons. A neuron n and the neurons in its Moore neighborhood N(n) on the output grid of the SOM represent points in the data space. The sum of distances between n and the neurons in N(n) in the high-dimensional space is shown on a U-matrix as a height value (U-height) at neuron n. Large U-heights mean that there is a large gap in the data space. Low U-heights mean that the points in $\{n \cup N(n)\}$ are close to each other within the data space. On a 3D-display of U-matrix valleys, ridges and basins can be seen (Figure 1).

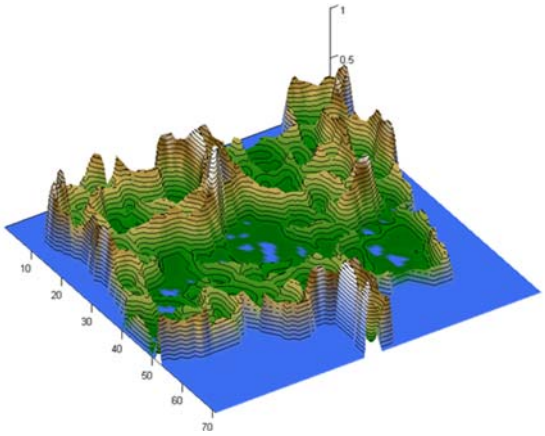


Figure 1: U-matrix of the pain data described below. [1]

If the best matching units (BMUs) of data points are located in a valley surrounded by large walls (water-basin), then these data points are within a distance-induced cluster structure in the data space. Water-sheds, respectively water-basins, on a U-matrix allow for emergence in SOM-based algorithms [3]. Emergent algorithms have the property that novel, formerly unseen structures on a macroscopic level (e.g., valley ridges, clusters) become visible on top of the only locally defined U-heights. The described usage of a SOM and its U-matrix can be used to visualize the distance structures in the high dimensional data space. If clustering of the data space is sought, additional clustering methods need to be applied such as a second SOM layer [4], Fuzzy clustering [5] or spectral clustering [6]. An alternative to these is the usage of visual observation to identify coherent valleys on the U-matrix, i.e. clusters in the data.

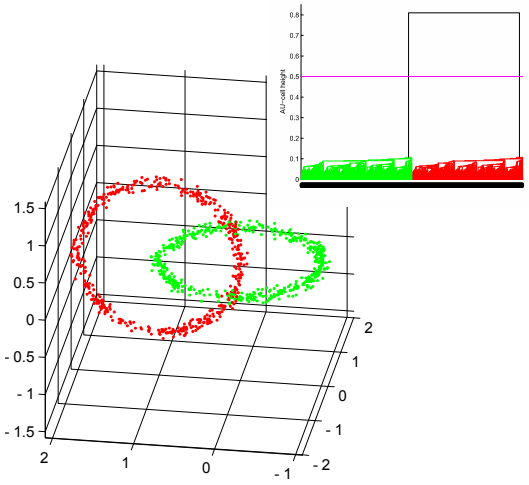


*Figure 2: The chain link data set and a dendrogram induced by the AU-cell distances, from which two classes, colored as either red or blue, clearly emerge (top right).*

In this work, we present the abstract U-matrix (AU-matrix), which formalizes the structures on a U-matrix such that distance calculations between BMUs become

meaningful. This enables assessing the topological correctness of the SOM and the implementation of clustering algorithms that take structures on the U-matrix into account.

## 2. Definitions

We assume that n high-dimensional data points $x_i \in R^D$ are projected (topology preserving) onto a two-dimensional grid of neurons trough a sufficiently trained SOM. The output grid of neurons (units) is embedded in $O \subset R^2$ (output space). The images (projections) of the points are the corresponding best-matching units (BMU). We assume that the size of the grid is large enough to map sufficiently distinct points of the data space to distinct BMU coordinates on the grid. For this type of SOM, called emergent SOM (ESOM), the size of the output grid is such that the Voronoi cells of a Voronoi tessellation [11] of the BMUs are sufficiently large. If a Voronoi-cell $V_i$ has a cell $V_j$ as neighbor, then there is an edge in the corresponding Delaunay graph $D$ [7].

Let $b_i$ and $b_j$ be BMUs of data points $x_i$ and $x_j$, and $b_i$ and $b_j$ are connected by an edge in D. Define a U-cell as follows: a U-cell has a floor shaped by the border lines of the Voronoi cell of the BMU. On each borderline there is a vertical plane. If the borderline is between $b_i$ and $b_j$, the height of the U-cell on this borderline (AU-height) is the distance $d(x_i,x_j) > 0$ of the data points in the data space. The abstract U-matrix, (AU-matrix) is then the set of all U-cells on a SOM grid. This gives a geometric structure on top of the output space O which is analog to the U-matrix. Height values on the AU-matrix are displayed on top of the Voronoi cell lines and have a clear meaning: the distance in data space of the corresponding BMUs. The usual U-matrix can be regarded as a quantized visualization of the AU-matrix (see below). A U-matrix corresponds to its AU-matrix, if for all pairs of BMUs having an edge in the Delaunay graph the sums of the U-heights on suitable paths between the pairs of BMU correlate to the AU-heights.

Let AUH denote the Delaunay graph D induced by the BMUs and weighted by the AU-cell distances. If the edges with weights above a threshold $\min(AUH) < t < \max(AUH)$ are removed from AUH the graph may be separated into different connected components. Using all possible values of t results in an ordered set of critical threshold $t_0,t_1,\ldots,t_c$ such that for $t_i < t < t_{i+1}$ the clustering is the same. Following the method proposed by Carlsson et al. [8], a dendrogram can be constructed that shows the threshold along with the number and the sizes of the resulting clusters. T Using the dendrogram the data can be clustered by providing either the number of clusters or the threshold for the maximal AU-cell distance. This is called AU-cell clustering. The adjustment of a suitable threshold respectively the number of clusters is the same problem as in hierarchical clustering. A political map of a U-matrix, resp. AU-matrix, is a top view of the AU-matrix where Voronoi cells of BMU $b_i$ and $b_j$

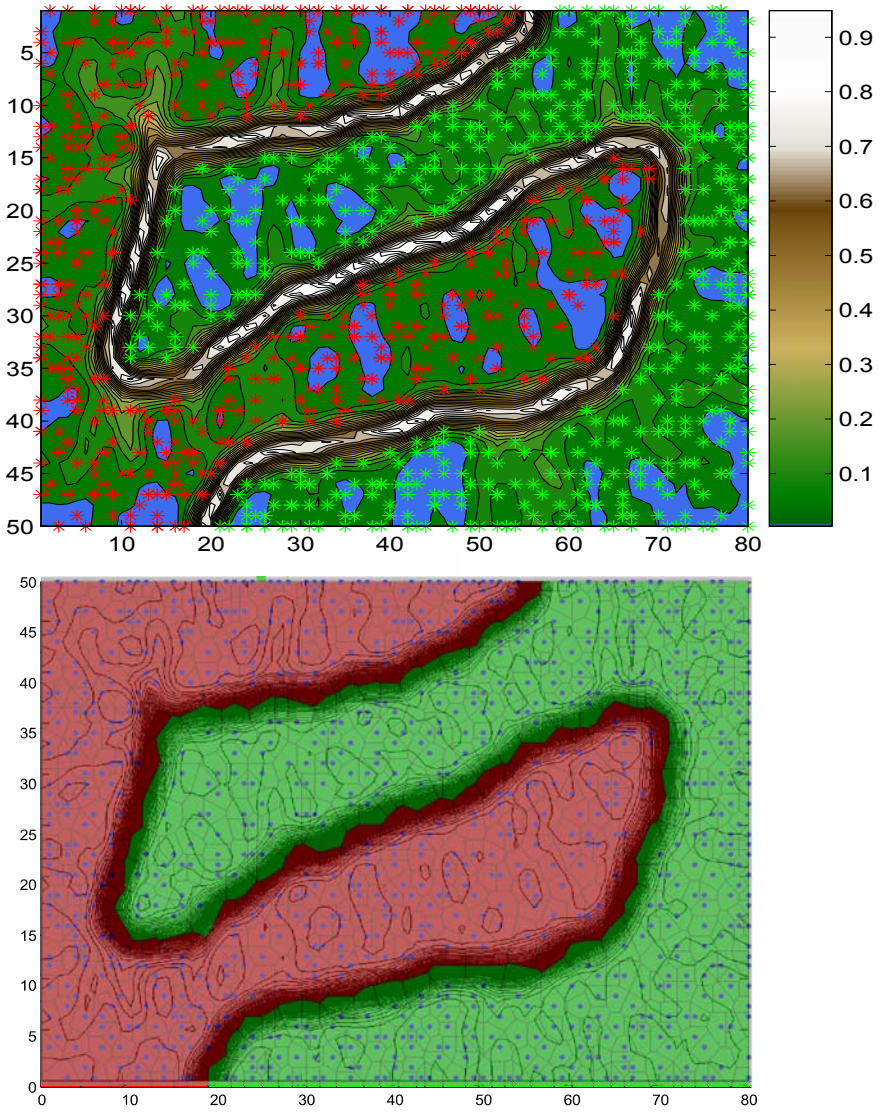


*Figure 3: Top: U-matrix of the Chainlink data, where a visual separation emerges from the ridge in the physical map. Bottom: Political map superimposed on the U-matrix. The two classes are colored in either red or green and their clear visual separation follows the edges of Voronoi cells.*

have the same color if the corresponding data points of $b_i$ and $b_j$ are assigned to the same cluster.

## 3. A first example

Data sets from the Fundamental Clustering Problems Dataset (FCPS,) are used to demonstrate the application of the AU matrix. For the three-dimensional Chainlink data set of 1000 data points Figure 2) an ESOM of grid size 80 x 50 was trained using the Databionic ESOM software [9].

A top view of the U-matrix using physical-map analogy for color-coding of the distances separates the two classes visually by a ridge between two valleys (Figure 3).

An overlay of the U-matrix with a top view of the AU-matrix, for each BMU its Voronoi cell can be seen (Figure 3). The ridge on the U-matrix coincides with the Voronoi-cell's borders having large AU-heights. The dendrogram for AU-cell clustering clearly indicates a definition of two classes. These two clusters are the two separate rings in the data.

*Table 1: Comparative of performance (accuracy [%] of data point assignment to the correct cluster) the AU based and other (Ward, k-means) clustering methods for identifying the cluster structure of data sets with different degrees of difficulty selected form the Fundamental Clustering Problems Dataset (FCPS http://www.uni-marburg.de/fb12/datenbionik/data).*

| Data Set | Main problem | Accuracy [%] of cluster membership assignment | | |
|---|---|---|---|---|
| | | **AU clustering** | **Ward** | **k-means** |
| **Hepta** | Easy | 100 % | 100 % | 100 % |
| **Lsun** | Standard | 100 % | 50 % | 50 % |
| **Tetra** | Small inter distances | 100 % | 90 % | 100 % |
| **Chainlink** | Linearly not separable | 100 % | 50 % | 50 % |
| **Atom** | Variance differences | 100 % | 50 % | 50 % |
| **Target** | Outlier | 100 % | 25 % | 25 % |
| **Golf ball** | Equidistant points | 100 % | 50 % | 0 % |

On the Chainlink data set, AU-cell clustering provides complete accuracy (100%). On other data sets from FCPS, the AU matrix method outperforms common cluster algorithms such as k-means and Ward clustering by obtaining always the correct cluster membership of a data point (100% accuracy), whereas the classical methods often provides lower accuracies with more difficult data, up to occasional complete failure (Table 1).

## 4. Application of the abstract U-matrix to real-life data

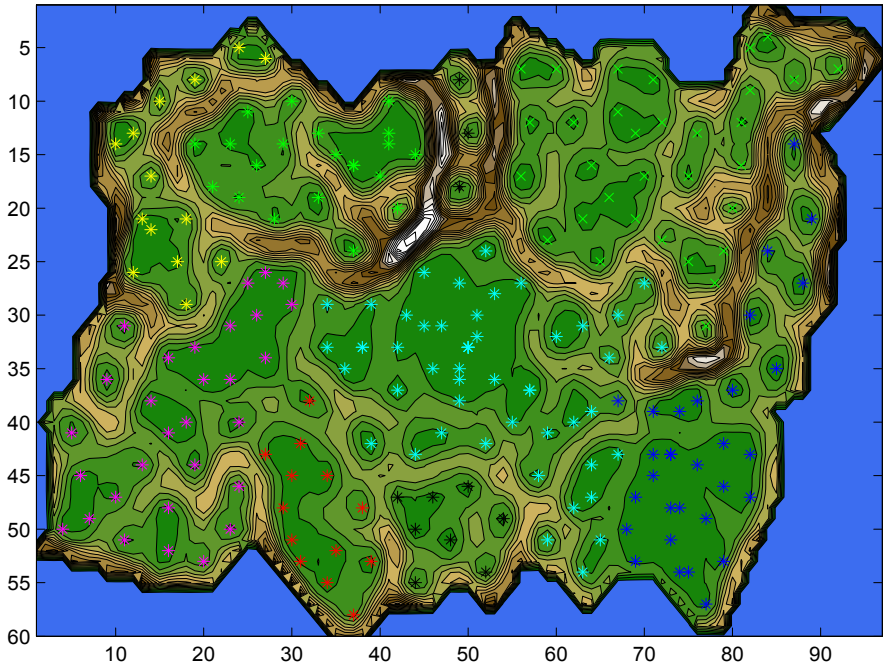


*Figure 4: U-matrix of the Pain data set*

Pain and its genetic background is a complex problem in biology. Pain is a trait defined as an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damages. Its sensory, affective, motor, vegetative and emotional components [10] are associated with a complex pathophysiology [11] reflected in the large network of molecular nociceptive pathways [12]. A genetic basis of pain and analgesia has been well established. Today, more than 410 genes have been recognized to contribute to the individual sensitivity of pain [13]. For example, red-haired women displayed greater pain relief following administration of a kappa-opioid receptor specific analgesic (pentazocine) than women without this phenotype [14]. Another example is the hereditary insensitivity to pain due to a loss-of-function genetic mutation, which was found in a single family whose members work as fakirs using the absence of pain professionally [15]. Such mutations are today a valuable source of targets of new analgesic drugs.

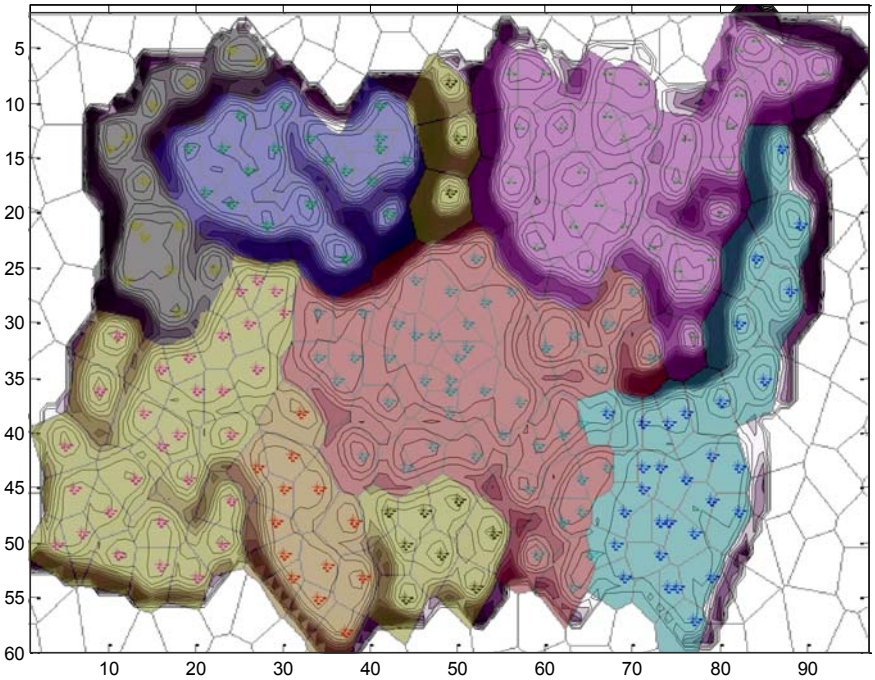


*Figure 5: Political map of the Pain data set following clustering into eight classes of subjects with similar pain sensitivity patterns and overlaid onto the U-matrix (Figure 4)*

However, the utility of genetic markers to predict pain sensitivity in the average population and to guide personalized analgesic therapy has remained modest [16] due to the complexity in both, the phenotype and the genotype of pain [17]. Initial approaches using clustering of patients with similar sensitivities to particular pain stimuli have so far not provided reproducible predictions of pain phenotypes and associations of underlying pain-relevant genotypes. These approaches used mainly k-means clustering [18]. However, as shown above, k-means clustering may provide poor cluster associations depending on the distribution of the data.

A data set [1] was obtained following administration of defined pain stimuli to

214 (105 men) healthy volunteers (approval of the Ethics Committee of the Medical Faculty of the Goethe – University and informed written consent from each participant obtained). The pain phenotype was assessed by means of measuring pain thresholds to four different pain stimuli (heat, cold, blunt pressure, electricity 0 – 20 mA).

After appropriate preprocessing (for details, see [1]) the data were projected onto a ESOM of 50 x 82 = 4200 neurons and an U-Matrix was generated (Figure 4).

The dendrogram of the AU-cell distances suggested eight clusters in the pain data resulting in a political map that can be overlaid on the U-matrix (Figure 5).

The cluster identification using the U-matrix provided a suitable basis for the desired genotype-phenotype association. That is, on the basis of a combined genotype, consisting of 10 variants in four genes (plus gender), subjects with a high pain sensitivity phenotype were predicted with an accuracy of 78 % [1].

For comparison, among single genetic markers and gender, only the latter provided a prediction better than guessing. Similarly, for a pain phenotype called "stoics with a selective high sensitivity to heat", a genetic association with a genotype composed of seven variants in three genes provided a mean cross-validated classification accuracy of 88 ± 12%. These examples clearly demonstrate the utility of AU-cell clustering for real-life pain phenotype genotype associations, suggesting that this method indeed may be essential to advance personalized therapy approaches.

## 5. Discussion

In this contribution we shed insights onto what can be seen on a U-matrix and how this can be used to identify structures and or clusters in high dimensional data. The AU-matrix can be seen as theoretical model to explain a given U-matrix. It can be used for the assessment of the topological correctness of the underlying SOM (see Figure 6) and the implementation of clustering algorithms which take the structures seen on the U-matrix into account. A dendrogram as known from hierarchical clustering algorithms which closely correspond to the structures seen on a U-matrix can be constructed. The political map of a U-matrix is a very flexible tool to visualize the result of possible clusterings. It allows to easily identify outliers and critical distance structures where the membership of data points to the same or different clusters is debatable.

Sometimes cluster structures on high dimensional data are not defined by distance structures (alone) [19]. Local densities of the data space must be taken into account. DBSCAN is an example of a distance and density based clustering algorithm [20]. The CONNvis approach recently proposed [21] integrates density information into a Delaunay graph on the high dimensional data points. In our approaches density information is regarded separately using the P- and/or U* matrix methods[22]. A corresponding technique for AU-matrices is subject to further research. So far, the here presented method provides accurate clustering in model data sets and seems to provide the necessary clustering of real-life data sets, with promising results to provide the necessary methods to identify, for example, subpopulations for individualized treatments and drug discovery and development.

## 6. References


1. Lötsch, J., Ultsch, A.: A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. Application to pain. Journal of biomedical informatics 46, 921-928 (2013)
2. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol Cybernet 43, 59-69 (1982)
3. Ultsch, A.: Emergence in Self-Organizing Feature Maps. In: International Workshop on Self-Organizing Maps (WSOM '07). Neuroinformatics Group, (2007)
4. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council 11, 586-600 (2000)
5. Sarlin, P., Eklund, T.: Fuzzy Clustering of the Self-Organizing Map: Some Applications on Financial Time Series. In: Laaksonen, J., Honkela, T. (eds.) Advances in Self-Organizing Maps, vol. 6731, pp. 40-50. Springer Berlin Heidelberg (2011)
6. Taşdemir, K.: Spectral Clustering as an Automated SOM Segmentation Tool. In: Laaksonen, J., Honkela, T. (eds.) Advances in Self-Organizing Maps, vol. 6731, pp. 71-78. Springer Berlin Heidelberg (2011)
7. Delaunay, B.: Sur la sphère vide. Izvestia Akademii Nauk SSSR, vol. 7, pp. 793-800. Otdelenie Matematicheskikh i Estestvennykh Nauk (1934)

8. Carlsson, G., M, F., #233, moli: Characterization, Stability and Convergence of Hierarchical Clustering Methods. J. Mach. Learn. Res. 11, 1425-1470 (2010)
9. Ultsch, A., Moerchen, F.: ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. (2005)
10. Tracey, I., Mantyh, P.W.: The cerebral signature for pain perception and its modulation. Neuron 55, 377-391 (2007)
11. Cross, S.A.: Pathophysiology of pain. Mayo Clin Proc 69, 375-383 (1994)
12. Julius, D., Basbaum, A.I.: Molecular mechanisms of nociception. Nature 413, 203-210 (2001)
13. Lötsch, J., Doehring, A., Mogil, J.S., Arndt, T., Geisslinger, G., Ultsch, A.: Functional genomics of pain in analgesic drug development and therapy. Pharmacology & therapeutics 139, 60-70 (2013)
14. Mogil, J.S., Wilson, S.G., Chesler, E.J., Rankin, A.L., Nemmani, K.V., Lariviere, W.R., Groce, M.K., Wallace, M.R., Kaplan, L., Staud, R., Ness, T.J., Glover, T.L., Stankova, M., Mayorov, A., Hruby, V.J., Grisel, J.E., Fillingim, R.B.: The melanocortin-1 receptor gene mediates female-specific mechanisms of analgesia in mice and humans. Proc Natl Acad Sci U S A 100, 4867-4872 (2003)
15. Cox, J.J., Reimann, F., Nicholas, A.K., Thornton, G., Roberts, E., Springell, K., Karbani, G., Jafri, H., Mannan, J., Raashid, Y., Al-Gazali, L., Hamamy, H., Valente, E.M., Gorman, S., Williams, R., McHale, D.P., Wood, J.N., Gribble, F.M., Woods, C.G.: An SCN9A channelopathy causes congenital inability to experience pain. Nature 444, 894-898 (2006)
16. Mogil, J.S.: Are we getting anywhere in human pain genetics? Pain 146, 231-232 (2009)
17. Lötsch, J., Flühr, K., Neddermayer, T., Doehring, A., Geisslinger, G.: The consequence of concomitantly present functional genetic variants for the identification of functional genotype-phenotype associations in pain. Clin Pharmacol Ther 85, 25-30 (2009)
18. Baron, R., Binder, A., Wasner, G.: Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. Lancet Neurol 9, 807-819 (2010)
19. Ultsch, A., Moutarde, F.: U*F Clustering: a new performant Cluster-mining method based on segmentation of Self-Organizing Maps. In: International Workshop on Self-Organizing Maps (WSOM '05). (2005)
20. Ester, M., Kriegel, H.-P., Sander, S., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. 226-231 (1996)
21. Tasdemir, K., Merényi, E.: SOM-based topology visualization for interactive analysis of high-dimensional large datasets. University of Bielefeld, Germany (2012)
22. Ultsch, A.: The U-Matrix as Visualization for Projections of high-dimensional data. In: Proc. 11th IFCS Biennial Conference. (2003)