Lötsch et al. Big Data Analytics (2018) 3:5
https://doi.org/10.1186/s41044-018-0032-1

Big Data Analytics

SOFTWARE Open Access

Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix)

Jörn Lötsch[1,2*], Florian Lerch[2,3], Ruth Djaldetti[4], Irmgard Tegder[1] and Alfred Ultsch[3]

* Correspondence:
j.loetsch@em.uni-frankfurt.de
[1]Institute of Clinical Pharmacology, Goethe – University, Theodor Stern Kai 7, 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, 60590 Frankfurt am Main, Germany
Full list of author information is available at the end of the article

Abstract

Background: Unsupervised machine-learned analysis of cluster structures, applied using the emergent self-organizing feature maps (ESOM) combined with the unified distance matrix (U-matrix) has been shown to provide an unbiased method to identify true clusters. It outperforms classical hierarchical clustering algorithms that carry a considerable tendency to produce erroneous results. To facilitate the application of the ESOM/U-matrix method in biomedical research, we introduce the interactive R-based bioinformatics tool "Umatrix", which enables valid identification of a biologically meaningful cluster structure in the data by training a Kohonen-type self-organizing map followed by interface-guided interactive clustering on the emergent U-matrix map.

Results: The ability to detect clinical relevant subgroups was applied to a data set comprising plasma concentrations of d = 25 lipid markers including endocannabinoids, lysophosphatidic acids, ceramides and sphingolipids acquired from $n = 100$ patients with Parkinson's disease and n = 100 controls. Following ESOM training, clear data structures in the high-dimensional data space were observed on the U-matrix, allowing separation of patients from controls almost perfectly. When the data structure was destroyed by Monte-Carlo random resampling, the U-matrix became unstructured and patients and controls were mixed. Obtained results are biologically plausible and supported by empirical evidence of a regulation of several classes of lipids in Parkinson's disease.

Conclusions: Sophisticated analysis of structures in biomedical data provides a basis for the mechanistic interpretation of the observations and facilitates subsequent analyses focusing on hypothesis testing. The freely available R library "Umatrix" provides an interactive tool for broader application of unsupervised machine learning on complex biomedical data.

Background

Biomedical research generates increasingly complex data [1, 2] that are challenging for information processing and knowledge discovery. However, novel methods employing data driven approaches to complex clinical information are increasingly being esteemed [3, 4]. This is currently facilitated by developments in data science considered as a rapidly growing interdisciplinary research area that deals with the problem-

oriented processing of large amounts of complex data with the aim to discover and process knowledge [5–7].

Among key technologies for complex data evaluation figures machine learning [8]. Its major forms are supervised or unsupervised learning. For the first, the data space, $D = \{(x_i, y_i) | x_i \in X, y_i \in Y, i = 1 \ldots n\}$ consists of an input space X comprising vectors $x_i = <x_{i,1}, \ldots x_{i,d}> \in P_\delta$ with $d > 0$ different parameters (features) acquired from $n > 0$ cases, and the output space Y comprising $y_i \in C = \{1, \ldots, c\}$ of c possible classes for the cases. Classes are, for example, treatment groups or diagnoses. The task is to learn a suitable algorithm that maps the input cases to the output classes and can be used for assignment of unlabeled cases to the right class. By contrast, in the case of unsupervised machine learning the data space $D_u = \{x_i \mid x_i \in X, i = 1 \ldots n\}$ lacks the class labeling and the task is to find interesting structures in the d-dimensional feature space $D_u \subset R^d$ that is accessible to biomedical interpretation.

Identification of structures in high-dimensional data by means of unsupervised machine learning can be used to explore whether a complex data set contains information that reflects the experimental setting. For example, the possibility to separate patients from healthy subjects via structures in a data set encourages further statistical exploration as this observation strongly supports, that the acquired data carries problem-relevant information. Structures in high-dimensional data are also employed for subgroup stratification known as clustering. While several different clustering methods are commonly used, most of them occasionally produce erroneous results by assigning data points to the wrong groups or by imposing cluster structures on cluster-free data sets [9, 10]. As it is imperative that structures identified in biomedical data correctly reflects the cluster structure, we have recently proposed emergent self-organizing feature maps (ESOM), combined with the U-matrix, as a viable, unbiased alternative method to identify true clusters in the high-dimensional data space [11, 12]. To facilitate data driven research approaches, the present work introduces a bioinformatics toolbox for unsupervised machine learning implemented as emergent self-organizing feature maps (ESOM [13]) combined with the U-matrix [14].

Methods

Biomedical data set

A data set suitable for the present assessment was available as plasma concentrations of $d = 25$ lipid markers (Fig. 1) assayed in probes drawn from Parkinson patients and healthy controls. The assessment of lipid markers in the context of Parkinson's disease is based on evidence of an involvement of lipid regulation [15–21].

Subjects and study design

The study followed the Declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty of the Goethe – University Frankfurt am Main, Germany (reference number 197/13). Informed written consent into study participation and publication of the results in an anonymized form was obtained from all subjects. Employing a parallel group design, patients with Parkinson's disease ($n = 128$, age = 69 ± 8.2 years (mean ± standard division, body mass index, BMI = 25.3 ± 4.3, 85 men) and healthy controls ($n = 350$, age = 26.9 ± 6.6 years, BMI = 22.7 ± 3.4, 117 men) were consecutively recruited from outpatients and inpatients of the Department of Neurology

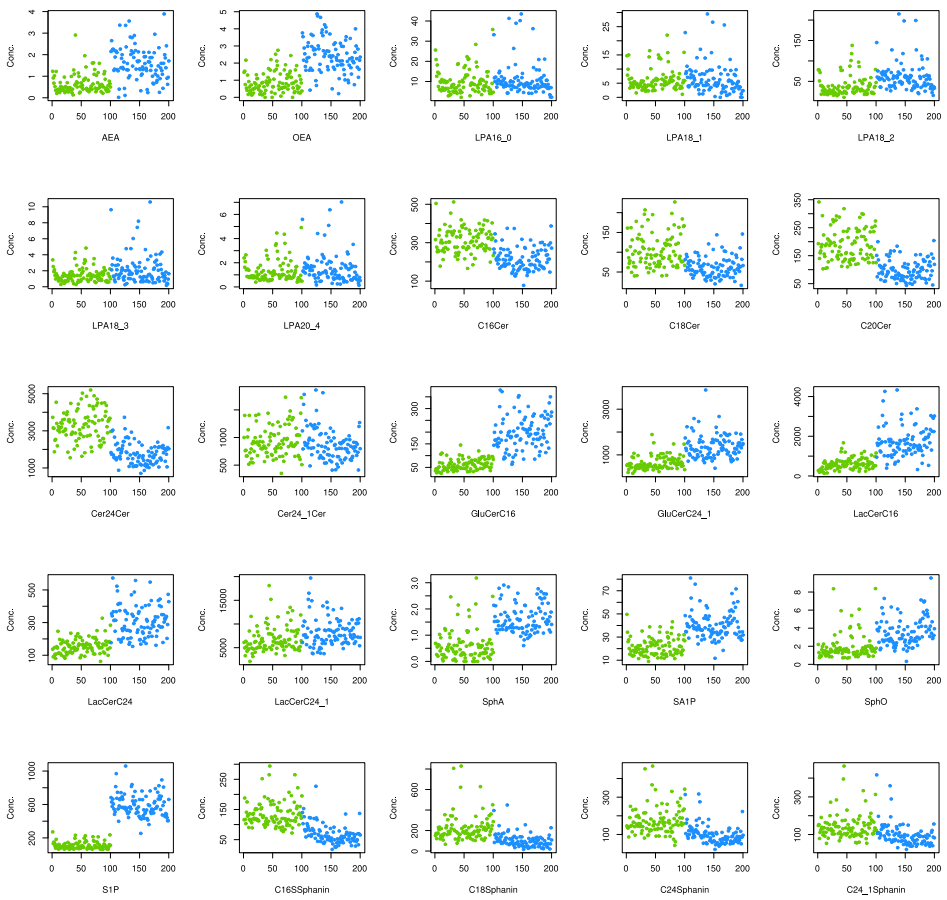

Fig. 1 Dotplot of plasma concentrations of various lipid markers. The dots display the single data, sorted in order of consecutive study number, starting with the Parkinson patients ($n = 100$, green dots) and followed by the healthy subjects (n = 100, blue dots). The figure has been created using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24])

(patients) and from students and staff members of the hospital (controls) who routinely reported at the institutional occupational health service. From this, BMI and sex matched samples of $n = 100$ subjects per group were drawn whereas the age difference was addressed during data preprocessing (see respective section below). Inclusion criteria were age ≥ 18 years, for patients a neurologically verified diagnosis of Parkinson's disease and for controls no current medical condition queried by medial interview, and no drug intake for at least 1 week except contraceptives, vitamins and L-thyroxin.

Lipid mediator plasma concentration analysis

From each subject, a venous blood sample (9 ml) was collected into a serum tube and centrifuged at 3000 rpm for 10 min. Plasma was separated and frozen at − 80 °C until assay. A total of $n = 43$ different lipid mediators (Fig. 1) was analyzed from the plasma samples. Plasma concentration analyses were performed using liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) essentially as described previously [22, 23]. The selection included endocannabinoids (AEA, OEA), lysophosphatidic acids (LPA16:0, LPA18:1, LPA18:2, LPA18:3, LPA20:4), ceramides

(Cer16:0, Cer18:0, Cer20:0, Cer24:0, Cer24:1, GluCerC16:0, GluCerC24:1, LacCerC16:0, LacCerC24:0, LacCerC24:1; Cer = ceramide, GluCer = glucosylceramide, LacCer = lactosylceramide), and sphingolipids (sphinganine, sphingosine, S1P, SA1P C16Sphinganine, C18Sphinganine, C24Sphinganine, C24:1Sphinganine). For all analytes, the concentrations of the calibration standards, quality controls and samples were evaluated by the Analyst software 1.6 and MultiQuant Software 3.0 (Sciex) using the internal standard method (isotope-dilution mass spectrometry). Calibration curves were calculated by linear regression with 1/x weighting for ceramides and LPA.

Data analysis

Data were analyzed using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24]) on an Intel Xeon® computer running on Ubuntu Linux 16.04.2 64-bit. The data space consisted of d = 25 lipid markers measured in the plasma of Parkinson patients or healthy subjects. Specifically, from the original cohort of $n = 478$ subjects, BMI and sex matched samples of $n = 100$ Parkinson patients and n = 100 healthy subjects were drawn using a propensity score matching (PSM [25]). This obtained samples of subjects who were comparable, i.e., did not statistically significantly differ, on the covariates BMI (Parkinson patients: BMI 24.9 ± 3.8, healthy subjects: BMI 24.7 ± 3.5, t-test: $t = -0.36307$, df = 197.13, $p = 0.7169$) and sex (Parkinson patients: 52/48 men/women, healthy subjects: 64/36men/women, χ^2-test: $\chi^2 = 2.4836$, df = 1, $p = 0.115$).

The data space was examined with the task to find interesting structure, for which unsupervised machine learning provides the adequate methodology. Specifically, for unsupervised machine learning the data space $D_u = \{x_i \mid i = 1 \ldots n\}$ does not include the class information, i.e., diagnose of Parkinson disease, and the task is to find distance and density based structures in the d-dimensional feature space $D_u \subset R^d$ that can subsequently be interpreted in a biomedical context such as a clinical diagnosis. The analyses were performed using an interactive R toolbox ("Umatrix"); the analytical steps included (i) data preprocessing, (ii) identification of distance and density-based structures in the data space, and (iii) interpretation of these structures with respect to group respectively cluster structures and their relation to the clinical diagnosis, which will be described in detail as follows.

Data prepossessing

The exploration of the data space was preceded by data preprocessing as an important step in the knowledge discovery process, which is considered as a fundamental building block of data mining. Data preprocessing comprised of (i) log transformation, (ii) age correction, (iii) uniform scaling and (iv) imputation of missing data. Specifically, (i) as quantile-quantile plots pointed at log-normal distribution of the data, which is in line with general observations in blood-derived concentration data [26], data was zero invariant log-transformed, except for the two endocannabinoids (AE and OEA) for which the plots suggested to prefer the original linear scaling. Subsequently, (ii) the influences of age on the lipid marker plasma concentrations were reduced by applying corrections based on robust linear regression using the Levenberg-Marquardt nonlinear least-squares algorithm implemented in the R library "minpack.lm" (https://cran.r-project.org/package=minpack.lm [27]). To obtain (iii) a uniform scaling of all lipid marker

plasma concentrations suitable to be assessed for Euclidean distances, data were transformed into percentages [28], i.e., into the interval [0,100]. Finally, a single missing data point was imputed using a k nearest neighbor algorithm with $k = 3$ [29] and applying the weighted average method and Euclidean distance as implemented in the "DMwR" R library (https://cran.r-project.org/package=DMwR [30]).

Identification of distance and density based structures in the data space

The data respectively feature space $D = \{x_i, i = 1, \ldots, 200\} \subset \mathbb{R}^d$, d = 25 was explored for emergent structures using unsupervised machine learning [8]. Of note, the information about the presence or absence of the disease was not included in this analysis. The aim of this analysis was to find interesting structures in the data space accessible to subsequent biomedical interpretation. This was obtained by the application of emergent self-organizing feature maps (ESOM) [14] that are based on a topology-preserving artificial neuronal network (Kohonen SOM [13, 31]). The ESOM was used to project high-dimensional data points $x_i \in R^d$ onto a two dimensional grid of neurons (map space). The analysis followed the workflow specified in Fig. 2.

The identification of emergent data structures in the 25-marker plasma lipidomics data set required four major steps comprising (i) the preparation of the neuronal grid for data projection from the high-dimensional onto a two-dimensional space, (ii) the learning phase of the self-organizing map, (iii) the detection of distance based data structures on this map and (iv) the detection of density based data structures on this map, which will be described as follows. In addition, the same analyses were applied after destruction of the possible high-dimensional structure in this data set was destroyed by Monte-Carlo [32] random resampling of $n = 200$ data among each lipid marker concentration vector using the R library "sampling" (https://cran.r-project.org/package=sampling [33]).

Preparation of the neuronal grid

A characteristic of ESOM as a tool for the characterization of the high dimensional data space is the large number of neurons. The grid size is chosen on the criteria that (i) it should contain a sufficiently large number of neurons to avoid that SOMs degenerate to a k-means like clustering algorithm with no potential to show emergent structures [34], (ii) it should, however, not be too large to avoid that each input data point can be represented on the map on a separate neuron with a surrounding area of other neurons interpolating the data space, and (iii) the grid should not be quadratic, but observe edge ratios between 1.2 and the golden ratio of 1.6. This has been shown empirically to improve representation performance [35]. Combining the requirements for size and form, as a starting point for the exploration of structures in a data set a SOM with 4000 (80×50) neurons has been successful in many applications and was also applied to the present data set. High dimensional datasets are usually projected by ESOM onto a finite but borderless output space. The borderless space is obtained by the embedding of the finite grid of neurons on the surface of a torus (toroid). This avoids the problems of borderline neurons [14] and subsequent boundary effects [14].

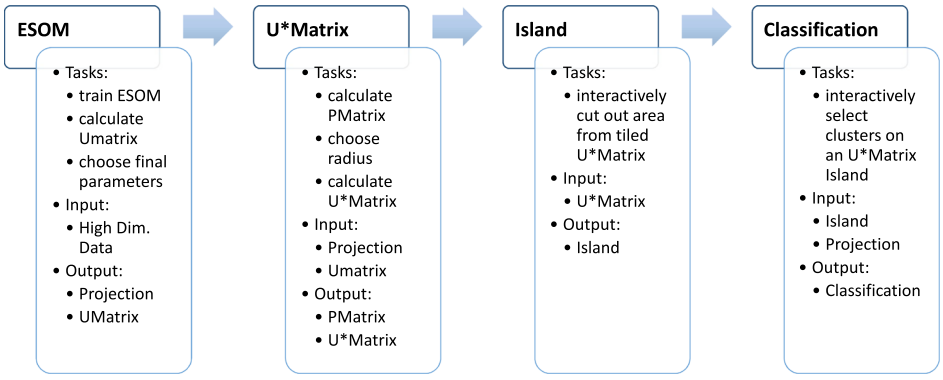

Fig. 2 Scheme of the workflow of generating a U-matrix with subsequent identification of clusters in the data using the R package "Umatrix" (https://cran.r-project.org/package=Umatrix). The process starts with a topology-preserving projection of high-dimensional data points onto a two-dimensional self-organizing network consisting of a grid of neurons of the Kohonen type [13] obtained via training of an emergent self-organizing map (ESOM). Subsequently, the distances between data points are projected on top of the grid as so-called U-matrix. Since the ESOM is toroid, i.e., opposite edges are connected, a so-called island is cut off the U-matrix, which contains each part of the U-matrix only once. Finally, on the finalized U-matrix clusters can be obtained interactively

Learning of the self-organizing map

On the SOMs each neuron represented a vector located in the d dimensional feature space, traditionally called "weights" (w_i). Each component's weight was initially randomly drawn from the range of the corresponding feature of the data and subsequently adapted to the data during the learning phase that used 25 epochs, i.e., presentations of the case data D_u. The weights were adapted according to the SOM learning rule: $\Delta w_i = \eta(t)\ h(bmu_i, r, t)(x_i - w_i)$ where x_i is a presented data point, bmu_i the closest neuron for x_i in the SOM (best matching unit, BMU), w_i the weight vector of neuron n_i, $h(...) \in [0,1]$ the neighborhood which depended in particular on the distance r from the bmu_i to the neuron n_i on the grid of neurons and $\eta(t) \in [0, 1]$ the learning rate. Learning rate and neighborhood radius are decreased during learning [13]. The result of this procedure is a data-driven self-organized and topology (neighborhood) preserving projection of the d dimensional feature space onto the two dimensional grid of neurons. The cases, i.e., data points of the input space, are represented in form of localizations of their respective BMU.

Detection of distance based data structures

Basically, any projection of high dimensional data points onto two dimensions is unable to preserve all the distances between the points [36]. This implies the distances between the best matching units on the grid of neurons are not always proportional to the data's distances in feature space. To solve this problem an U-matrix is constructed on top of the grid of neurons. This illustrates the local distance structures of the input space on top of layout of the data. The U-matrix is the canonical tool for the display of the distance structures of the input data on ESOM [14]. Specifically, the U-matrix is based on the local topology of the neuron space. If U_i denotes the set of neurons in the immediate neighborhood of a neuron n_i in the map space, the *U-height* of a neuron $uh(n_i)$ is given by the sum of all distances $d(...)$ from the weight vector of n_i to the

weight vectors of the neurons in U_i: $uh(n_i) = \sum_{n \in U_i} d(w(n_i), w(n))$. U-heigts represent distance structures within the data space in the vicinity of the weight vector *w(n)* of a given neuron *n*. A visualization of all U-heights at the neuron's coordinates is called the U-matrix [14] on top of the toroid SOM grid. For a two-dimensional display of a toroid map space, the U-matrix is presented as four adjacent copies in a tiled display, which has the disadvantage that each input data point is repeated on four locations. Therefore, the tiled U-matrix is trimmed at suitable borders in a way that each data point is represented only once. This results in a 3-dimensional U-matrix landscape with curved boundaries (U-map or "Island") where large "heights" represent large distances in the high-dimensional feature space while low "valleys" represent similar data. "Mountain ranges" visually separate clusters, and the view is enhanced by applying isohyptic coloring scheme derived from earth orbiting satellite measurements [37].

Detection of density based data structures

In addition to distance-based clusters, groups of data can also emerge from differences in data density. Therefore, the U-matrix was further enhanced by calculating a P-matrix [14] displaying the point density in data space. This density $p(n_i)$ was estimated as the number of data points in a sphere with radius *r* around the weight vector $w(n_i)$ for each neuron n_i on the ESOM's output *grid* $p(n_i) = |\{data\ points\ x |\ d(w(n_i),x) <= r\}|$. The U*-matrix combines distance structures (U-matrix) and density structures (P-matrix) into a single matrix [14]. This may provide better distinguishable cluster borders and can be applied when improvement of a standard U-matrix is desired.

Implementation

Implementation of a visualization guided ESOM/U-matrix R tool

An interactive tool ESOM/U-matrix unsupervised machine learning was realized using the freely available R software package (version 3.4.0 for Windows; http://CRAN.R-project.org/ [24]) and our newly introduced R package "Umatrix" (https://cran.r-project.org/package=Umatrix). For the graphical interface the open source web application framework Shiny for R was used (https://CRAN.R-project.org/package=shiny [38]).

An R library ("Umatrix"; Figs 3 and 4) was programmed that performs the tasks in separate modules dedicated to (i) data projection comprising the learning respectively training of an emergent SOM (subroutine "iTrainEsom"), (ii) distance-based data structure visualization (U-matrix, subroutine "iUstarmatrix"), (iii) density-based data structure visualization (P-matrix), which can be combined with the U-matrix into a U*-matrix (subroutine "iUstarmatrix"), and (iv) clustering and data classification (subroutine "iClassification"). In addition to the interactive modules, data analysis can also be performed via direct call of R functions. For example, the first algorithm comprising the learning respectively training of an emergent SOM (ESOM) can be also accessed via the R-script (esomTrain(…)). Similarly, visualizations can be obtained via the R-script (ustarmatrixCalc(…)). Since most of the projections map the data onto a borderless toroid map space, an interactive tool ("iUmapIsland") is provided to create a planar structure, called "U-map" or "island" (Fig. 4a). Finally, for flexible visualization purposes the package contains routines for two dimensional (top view, "plotMatrix(…)") and three dimensional plots of U-, P- and U*matrices.

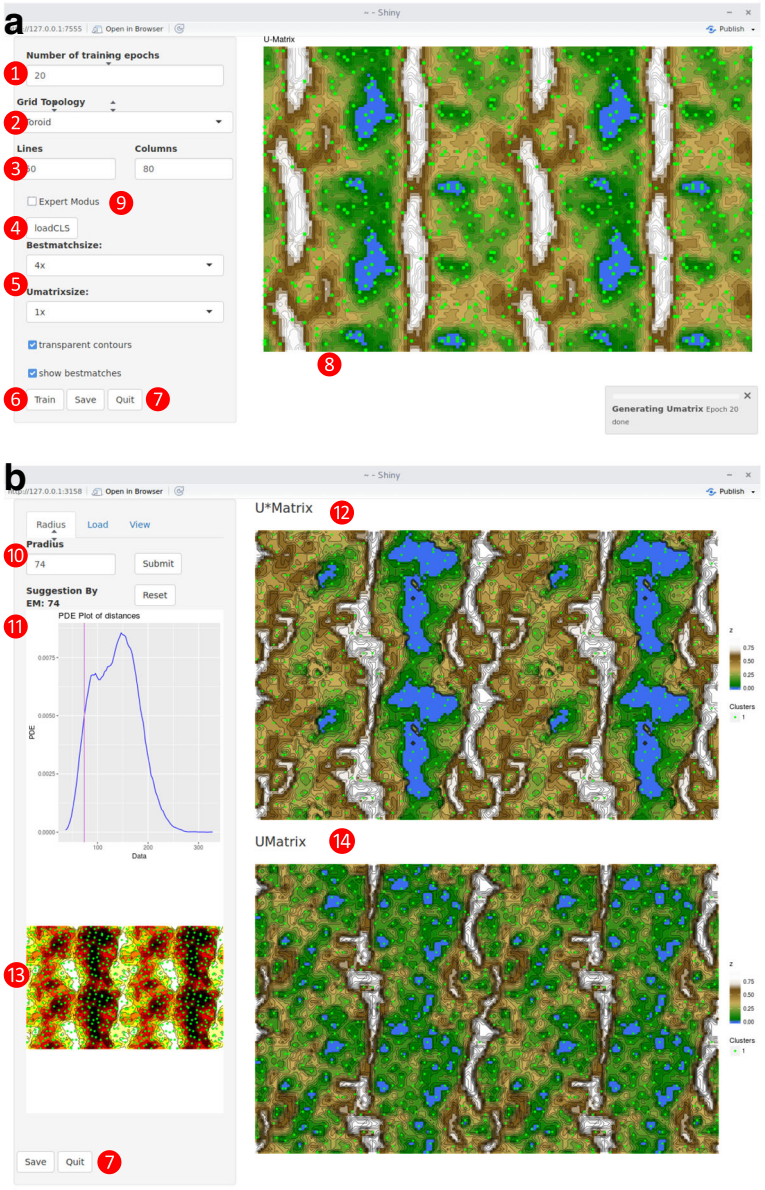

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Screenshots of interface components "iEsomTrain" and "iUstarmatrix" of the interactive "Umatrix" R library. **a**: The "iEsomTrain" component performs the training of the emergent self-organizing map (ESOM) and displays the resulting U-matrix. Default values are shown. The following user interactions are implemented: ❶ Selection of the number of training cycles of ESOM. ❷ Selection of the projection grid as either toroid, where opposite edges are connected, or planar. ❸ Selection of the size of the ESOM. The sizes should meet the following three criteria: Firstly, it should not be too small as it has been shown that in that case SOMs degenerate to a k-means like clustering [34]. Secondly, it should not be too large to avoid that each input data point can be represented on the map on a separate neuron with a surrounding area of other neurons interpolating the data space. Thirdly, edge ratios between 1.2 and the golden ratio of 1.6 should be applied as it has been observed that SOMs perform better if the edge lengths of the map are not equal [35]. However, a SOM with the default sizes of 4000 (80 × 50) neurons has been successful in many applications. ❹ A prior classification can be loaded from a structured text file. ❺ The display of the U-matrix can be visually modified such as changing the size, the diameter of the best matching units or making the colors slightly transparent which enhanced data structure visibility. ❻ After all parameters have been set, the training of the ESOM is started by pressing the "Train" button, the numerical results can be saved to a file and ❼ the interface is finally closed. ❽ The trained U-matrix is shown at the right of the interface panel. ❾ Further parameters such as learning rate can be set in a special expert mode, for details, see the description delivered within the R-package. **b:** The "iUstarmatrix" component calculated the data density based P-matrix and displays the U- and the resulting U*-matrix. The following user interactions are implemented: ❿ the radius of the hyperspheres for density estimation can be selected based on a suggestion obtained from the probability density distribution of the distances between the data points. This distribution is displayed below as a Pareto density estimation (PDE) [72] with the suggested radius indicated as a magenta line adjustable by the user. . At the top right part of the interface, the U*-matrix is displayed, which results from superposition of the data-density based P-matrix with the original data-distance based U-matrix . The figure has been created using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24]) using the R package "Umatrix" (https://cran.r-project.org/package=Umatrix)

Results

A matrix of 25 lipid biomarkers assayed in the plasma of 200 subjects provided the data space $D = \{x_i, i = 1, \ldots, 200\} \subset \mathbb{R}^d$ for the present analysis. A SOM was trained to obtain a topology pre-serving mapping of n high-dimensional data points $x_i \in R^d$ onto a two dimensional grid of neurons 50 × 80 neurons. The output grid of neurons (units) was embedded in O, a toroid output space where the projections of the points are the corresponding best-matching units (BMU). These BMUs were visualized on the SOM as dots (Fig. 5). On this SOM, each neuron n and the neurons in its neighborhood $N(n)$ represented 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 while 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, this can be visualized as valleys, ridges and basins (Fig. 6).

As the grid size was chosen large enough to map sufficiently distinct points of the data space to distinct BMU coordinates on the grid, data points within a distance-induced cluster structure in the data space can be separated by water-sheds allowing for emergence in the SOM-based algorithm [11]. 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. If b_i and b_j denote best matching units (BMUs) of data points x_i and x_j, and b_i and b_j are connected by an edge in D, a U-cell is defined as follows: a U-cell has a floor shaped by the border lines of the Voronoi cell of the BMU. On each borderline was 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 largest U-heights along

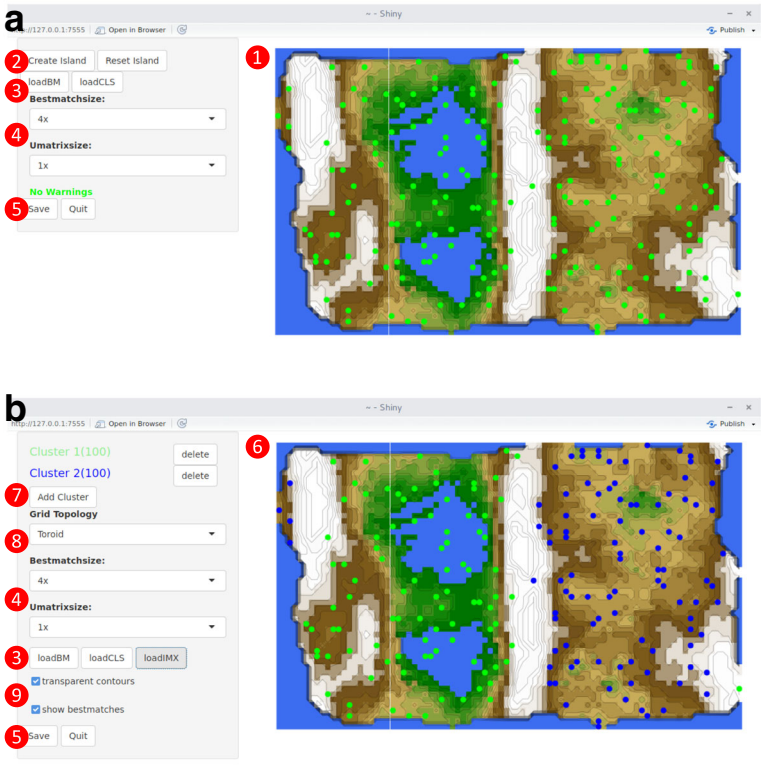

Fig. 4 Screenshots of interface components "iUmapIsland" and "iClassification" of the interactive "Umatrix" R library. **a**: The "iUmapIsland" component provides an interface to manually cut a so-called "island" out of the toroid Umatrix. The island should contain each part of the U-matrix only once. It is usually cut along cluster borders enhancing the visual data structure of emerging in the U-matrix. The following user interactions are implemented: ❶ the trained U-matrix is shown at the right of the interface panel. Using the mouse, on the toroid U-matrix (see Fig. 3) a unique region can be marked that contains every best matching unit (BMU) only once. ❷ The marked region can be cut out of the toroid U-matrix providing the so-called "island" as the standard representation of the U-matrix. ❸ A prior classification can be loaded from a structured text file. ❹ The display of the U-matrix can be visually modified such as changing the size, the diameter of the best matching units. ❺ The numerical results can be saved to a file and the interface is finally closed. **b**: The "iClassification" component is an interactive shiny tool that visualizes a given U-matrix and allows the user to select areas and mark them as clusters. The following user interactions are implemented: On the island, ❻ areas located in the same regions can be marked with the mouse as clusters, ❼ that can be added to the available clauses, or, clusters can be deleted manually. ❹ The display of the U-matrix can be visually modified such as changing the size, the diameter of the best matching units or ❾ making the colors slightly transparent which enhanced data structure visibility. ❽ Selection of the projection grid as either toroid where opposite edges are connected or planar. The figure has been created using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24]) using the R package "Umatrix" (https://cran.r-project.org/package=Umatrix)

the cell borders emerged as mountain ridges pointing at coherent valleys on the U-matrix, i.e. clusters in the data, which could be identified visually (Figs. 5 and 6).

The cluster structure seen on the U*-matrix indicated two main clusters (Figs. 5 and 6), clearly separated by a high "snow-covered" mountain ridge in the middle of the U*-matrix. A projection of the original classification into Parkinson patients versus healthy subjects onto the distance and density based data structures showed that the cohort was completely separated by lipid-marker plasma concentration data structure. Hence, plasma lipid markers carry sufficient information to distinguish Parkinson disease as a separate lipid marker pattern. This was further emphasized by the absence of such structures when the data was permutated using Monte-Carlo random resampling. The

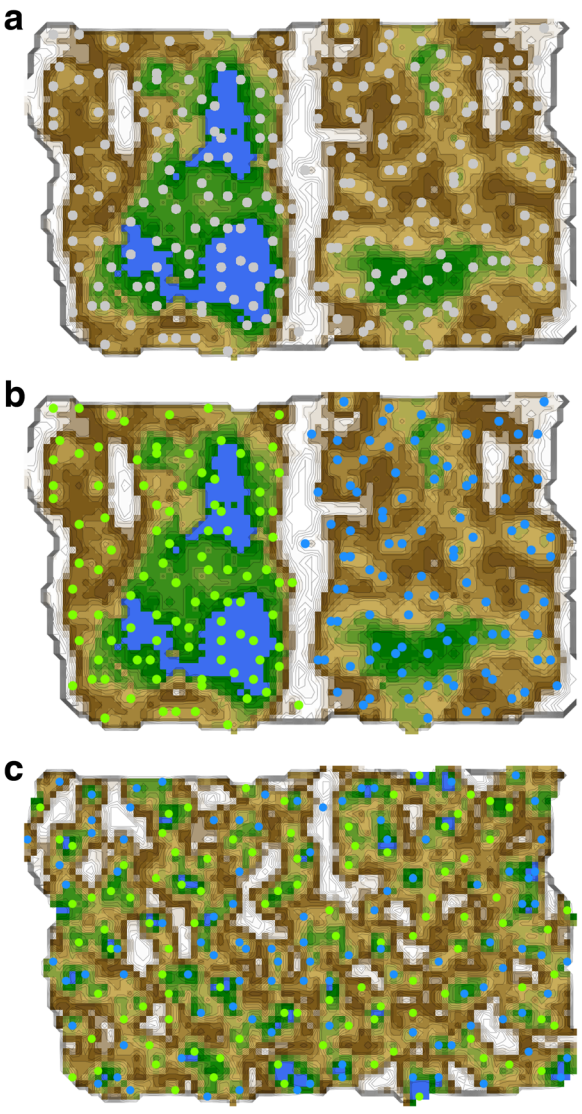

Fig. 5 U*-matrix representations of lipid marker plasma concentration patterns (d = 25 markers, Fig. 1) observed in $n = 200$ blood samples. The figure has been obtained using a projection of the data points onto a toroid grid of $50 \times 80 = 4000$ neurons where opposite edges are connected. The U*-Matrix was colored as a geographical map with brown (up to snow-covered) heights and green valleys with blue lakes. Valleys indicate clusters and watersheds indicate borderlines between different clusters. The dots indicate the so-called "best matching units" (BMUs) of the self-organizing map (SOM), which are those neurons whose weight vector is most similar to the input. A single neuron can be the BMU for more than one data pint or subject, hence, the number of BMUs may not be equal to the number of subjects as in the present case. **a:** top view on the U matrix showing two distinct regions (clusters) on the left and right of the white "mountain range" in the middle. The BMUs are colored neutrally (grey). **b:** The BMUs were differently colored to analyze the distribution of subjects across the cluster structure of the data space. When the group membership to either the Parkinson patients (green dots) or the healthy subjects (blue dots) is projected onto the U*-matrix, it becomes clear that the separate clusters perfectly coincide with the diagnostic classification of the subjects. **c:** The cluster structure was destroyed by permutation of the data, resulting on the U-matrix display in a clearly absent cluster structure with data from Parkinson patients mixed with data from healthy subjects and no clear "mountain ranges". The figure has been created using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24]). Specifically, the figures displaying geographical map analogies have been created using our R package "Umatrix" (https://cran.r-project.org/package=Umatrix)

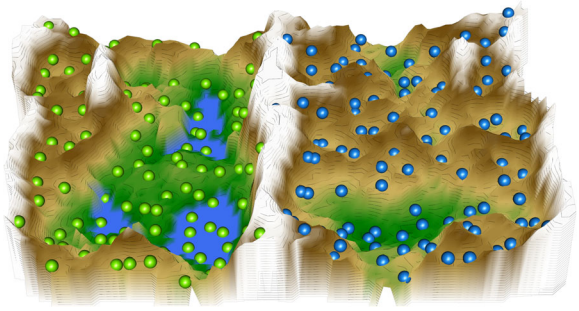

Fig. 6 3D-display of the U-matrix representation of the lipid marker (Fig. 1) plasma concentration pattern. It shows the Parkinson patients (green dots) are completely separated, i.e., located in a different cluster, from the healthy subjects (blue dots). A clear separation is indicated by a mountain range in the middle of the U*-matrix. The figure has been created using the R software package (version 3.4.0 for Linux; http://CRAN.R-project.org/ [24]). Specifically, the figures displaying geographical map analogies have been created using the R package "Umatrix" (https://cran.r-project.org/package=Umatrix)

resulting U-matrix was devoid of any recognizable cluster structure (Fig. 6.) and resembled earlier demonstrations of U-matrices obtained in structure-less data (see Figs. 3 and 6 in [12]).

Discussion

In this paper, unsupervised machine learning has been shown to identify biologically plausible structures in high-dimensional biomedical data. Following projection of the data onto a two-dimensional neuronal grid, training of a self-organizing map and visualization of the distances between the data points as a color-coded third dimension, two distinct clusters emerged. This structure coincided almost perfectly with the diagnosis of either Parkinson's disease or healthy controls. Figure 3 shows the prior clinical classification on top of the ESOM/U-matrix. By contrast, when destroying the lipid-marker plasma concentration pattern by random permutation of the data, the ESOM/U-matrix method ceased to show any emergent structure (Fig. 5).

Considering the repeated comparative demonstrations of the ability of this method to identify true structure in complex data, both artificial and biomedical [12, 14], comparisons with alternative clustering methods were not reiterated. Present results reproduce earlier demonstrations that ESOM/U-matrix correctly detects clusters in artificial [14] and biomedical data sets [39], while overcoming the imposing of spurious clusters [12, 14]. When the structure in biomedical data is destroyed by permutation, classical clustering algorithms may still suggest a structure. Hence, it is crucial to select an adequate clustering algorithm. The presently introduced R-library for self-organized structure finding using an unsupervised machine learning method enhances the availability of such a method. Previous applications of the ESOM/U-matrix method support the utility of this approach in biomedical research [40–45]. Moreover, ESOM/U-matrix rate is an intuitive representation with a sound basis in bioinformatics [46]. Its inherent 3D structure allows intuitive cluster recognition and can also provide a haptic access to the data space by means of 3D-printing of the structures of the complex feature space [47].

The present analysis used the ESOM/U-matrix method with the task to identify interesting structures in a high dimensional data set composed of lipid mediator plasma concentrations acquired in paints with a neurodegenerative disease or healthy subjects.

This suggests that the acquired data contains information that reflects the clinical setting, or any other experimental setting. In the present data set consisting of Parkinson patients and healthy subjects, this was the expected outcome. Specifically, Parkinson's disease is a neurodegenerative disease characterized by a progressive loss of dopaminergic neurons predominantly in the *Substantia nigra*. Its pathophysiology involves, e.g., accumulation and spreading of misfolded proteins [48], neuro-inflammation including persistent microglial activity [49], neuro-inflammation including persistent microglial activity [49] and dysfunctions of mitochondrial biogenesis and quality control systems [50]. The latter processes are essentially regulated by bioactive lipids including ceramides, sphingolipids, lysophospholipids, eicosanoids, endocannabinoids, HETEs, omega-3 and omega-6 lipids and EETs [17, 51–59]. The relevance of subtle changes in complex regulatory circuits has been taken into account by modern omics-based analysis tools using "gene set enrichment analyses" for the detection of pattern changes [60, 61]. Hence, the observation that a set of serum lipid markers contain information relevant for the separation of Parkinson patients from healthy subjects is biologically plausible. Following data structure detection performed with the present R library, subsequent analyses may focus on hypothesis testing or may apply supervised machine learning methods for further data exploration or for the creation of a diagnostic biomarker.

Among methods for the detection of distance and/or density based structures in high-dimensional, complex data, ESOM/U-matrix is accompanied by a number of alternatives, for example projection methods such as principal component analysis (PCA) [62], multidimensional scaling [63], isomap [64], t-SNE [65] and many more. PCA and MDS are continuous projection methods and fail when the cluster structures are not separable using hyperplanes. Isomap is a projection method that models high dimensional structures using a graph connecting k-nearest neighbors in the data. Whether or not the graphs show meaningful structures depends very much on the correct choice of k. This is a clear disadvantage to the ESOM/U-matrix method where the number of clusters does not need to be specified. t-SNE uses ideas derived from SOM, i.e., it employs a definition of neighborhood that is large at the beginning of the training and shrinks during the construction (learning) of the projection. It uses t- distributions to model neighborhoods of data points in the high-dimensional space as well as in the low dimensional projection space, which are parameterized by a neighborhood variance s^2. For the calculation of s^2 the (dis-)similarity between the neighborhood distributions in form of the Kullback-Leibler-Divergence [66] is used. The neighborhood size parameter s^2 is critical for a correct representation of the structures of the input space. The t-SNE method is not recommended for data with a high intrinsic dimensionality. Thus, the ESOM/U-matrix method well competes with alternatives partly aimed at the same goal, i.e., the detection of structures in high dimensional data.

Technical implementations of self-organizing maps including options to generate a U-matrix have been proposed previously such as, for example, the "yasomi" R package (https://r-forge.r-project.org/R/?group_id=1021) or the Matlab toolbox "somtoolbox" (http://www.cis.hut.fi/projects/somtoolbox/). These alternatives lack the essential feature of projecting the data on a toroid grid, which is essential to avoid border effects. These packages also lack tools for interactive cluster selection and data classification, such as cutting the islands. For comparison with the present R library, an example of

the results of an alternative implementation can be viewed in Fig. 3(a) of [67]. In addition, the present package provides further options not fully employed in the present data analysis example. That is, also all fundamental tools were applied in the present report, in a so-called "expert mode", which can be selected via a checkbox in the graphical user interface (item 9 in Fig. 3), the common selections for SOM calculations are available including different neighborhood functions, initialization parameters and parameters to tune the cooling, i.e. speed, of the ESOM algorithm. In addition, all functions offered in the graphical user interfaces are also available through scripting, which enables automated calculations. Finally, it should be mentioned that the present toolbox is satisfactorily scalable. Specifically, there is no reason to assume that the strategy behind the ESOM will fail with a more data points although the present implementation has not particularly been optimized for processing really big data sets. For most analytical steps the scaling of the Umatrix R library tends to be linear depending on the distribution of data. The present biomedical example data set comprising $200 \times 25 = 5000$ data points the ESOM was generated on an Intel® Xeon® Prozessor E5–2690 v2 with 64 GByte memory within less than 1 minute. A dataset containing a million data points and three parameters was processed on a Macbook Pro with 8 GByte memory and a 2.9 GHz Intel Core i5 processor within approximately 15 min. A one million data set with 9000 attributes was learned in less than 1 hour.

In the present report we focused on a single example biomedical data set. Comparisons of the ESOM/U-matrix method with alternative cluster algorithms have been extensively shown previously and were therefore not repeated [11, 12]. Such comparisons consistently supported the ESOM/U-matrix method as a viable, unbiased method to identify true clusters outperforming classical clustering methods such as Ward or k-means, which by imposing predefined cluster shapes occasionally produce cluster structures that are non-existent in the data [9, 10]. As cluster identification is among central targets of high-dimensional biomedical data analyses for discovery and prediction of classes independently of previous biological knowledge [68], the choice of a suitable analytical method is crucial to avoid non-reproducible pattern recognition. The present implementation of the ESOM/U-matrix method as an R library may provide a broader access to such a method. Previous successful applications of this method in biomedical informatics include data sets related to pain research [43, 69], computational drug discovery [40, 70] or lipidomics research in neurodegenerative diseases [71].

Conclusions

In the present report, unsupervised machine-learned analysis [8] was applied to a data set comprising lipid marker concentrations assayed in blood plasma 200 subjects. It was demonstrated that the combination of contemporary data science with analytical techniques for biomarkers allows recognizing Parkinson patients from plasma patterns of lipid markers. This unsupervised machine-learned analysis provides biologically plausible candidates compatible with prior knowledge from molecular research that finally may be complied into a complex biomarker. The implementation of the machine-learned analysis [8] applied using the emergent self-organizing feature maps (ESOM) [13] and the so-called U-matrix [14] into the freely available R environment facilitates broad accessibility of the method.

Acknowledgements
Michael C. Thrun re-implemented some of the ESMOM/U-matrix functions originally written in Matlab into the R programming language.

Funding
This work has been funded by the Landesoffensive zur Entwicklung wissenschaftlich - ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL, IT) and in addition, by the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 602919 (JL, GLORIA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
Data are available from the author IT on request.

Authors' contributions
Conceived and designed the analysis: AU, JL. Analyzed the data: JL. Wrote the paper: JL, AU, Performed the programming: FL, Provided data and probes: IT, RD. All authors read and approved the final manuscript.

Ethics approval and consent to participate
The study followed the Declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty of the Goethe – University Frankfurt am Main, Germany (reference number 197/13).

Consent for publication
Informed written consent into study participation and publication of the results in an anonymized form was obtained from all subjects.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Institute of Clinical Pharmacology, Goethe – University, Theodor Stern Kai 7, 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, 60590 Frankfurt am Main, Germany. [3]DataBionics Research Group, University of Marburg, Hans - Meerwein - Straße, 35032 Marburg, Germany. [4]Department of Neurology, Movement Disorder Clinic, Rabin Medical Center, Beilinson Hospital, 49100 Petach Tikva, Israel.

Received: 15 January 2018 Accepted: 6 April 2018
Published online: 20 April 2018

References

1. McDermott JE, Wang J, Mitchell H, Webb-Robertson B-J, Hafen R, Ramey J, Rodland KD. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert opinion on medical diagnostics. 2013;7(1):37–51.
2. Rinaldi A. Teaming up for biomarker future: many problems still hinder the use of biomarkers in clinical practice, but new public–private partnerships could improve the situation. EMBO Rep. 2011;12(6):500–4.
3. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author); 2001. p. 199–231.
4. Lötsch J, Geisslinger G. Bedside-to-bench pharmacology: a complementary concept to translational pharmacology. Clin Pharmacol Ther. 2010;87(6):647–9.
5. President's Information Technology Advisory C: Report to the president: computational science: ensuring America's competitiveness. 2005.
6. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.
7. Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64–73.
8. Murphy KP: Machine learning: a probabilistic perspective: the MIT press; 2012.
9. Kleinberg JM: An Impossibility Theorem for Clustering. In: Advances in Neural Information Processing Systems 15. Edited by Becker S, Thrun S, Obermayer K. Cambridge: MIT Press. 2003:463–470.
10. Jardine N, Sibson R. The construction of hierarchic and non-hierarchic classifications. Comput J. 1968;11(2):177–84.
11. Ultsch A: Emergence in self-organizing feature maps. In: International workshop on self-organizing maps (WSOM '07): 2007; Bielefeld, Germany. Neuroinformatics Group.
12. Ultsch A, Lötsch J. Machine-learned cluster identification in high-dimensional data. J Biomed Inform. 2017;66:95–104.
13. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybernet. 1982;43:59–69.
14. Ultsch A: Maps for visualization of high-dimensional data spaces. In: WSOM: 2003 2003; Kyushu, Japan. 2003: 225–230.
15. Pisani A, Fezza F, Galati S, Battista N, Napolitano S, Finazzi-Agro A, Bernardi G, Brusa L, Pierantozzi M, Stanzione P, et al. High endogenous cannabinoid levels in the cerebrospinal fluid of untreated Parkinson's disease patients. Ann Neurol. 2005;57(5):777–9.
16. Pyszko J, Strosznajder JB. Sphingosine kinase 1 and sphingosine-1-phosphate in oxidative stress evoked by 1-methyl-4-phenylpyridinium (MPP+) in human dopaminergic neuronal cells. Mol Neurobiol. 2014;50(1):38–48.

17. Mielke MM, Maetzler W, Haughey NJ, Bandaru VV, Savica R, Deuschle C, Gasser T, Hauser AK, Graber-Sultan S, Schleicher E, et al. Plasma ceramide and glucosylceramide metabolism is altered in sporadic Parkinson's disease and associated with cognitive impairment: a pilot study. PLoS One. 2013;8(9):e73094.
18. Li Z, Zhang J, Sun H. Increased plasma levels of phospholipid in Parkinson's disease with mild cognitive impairment. J Clin Neurosci. 2015;22(8):1268–71.
19. Xing Y, Tang Y, Zhao L, Wang Q, Qin W, Ji X, Zhang J, Jia J. Associations between plasma ceramides and cognitive and neuropsychiatric manifestations in Parkinson's disease dementia. J Neurol Sci. 2016;370:82–7.
20. France-Lanord V, Brugg B, Michel PP, Agid Y, Ruberg M. Mitochondrial free radical signal in ceramide-dependent apoptosis: a putative mechanism for neuronal death in Parkinson's disease. J Neurochem. 1997;69(4):1612–21.
21. Boutin M, Sun Y, Shacka JJ, Auray-Blais C. Tandem mass spectrometry multiplex analysis of glucosylceramide and Galactosylceramide isoforms in brain tissues at different stages of Parkinson disease. Anal Chem. 2016;88(3):1856–63.
22. Zschiebsch K, Fischer C, Pickert G, Haeussler A, Radeke H, Grosch S, Ferreiros N, Geisslinger G, Werner ER, Tegeder I. Tetrahydrobiopterin attenuates DSS-evoked colitis in mice by rebalancing redox and lipid signaling. J Crohns Colitis. 2016;10(8):965–78.
23. Sisignano M, Angioni C, Ferreiros N, Schuh CD, Suo J, Schreiber Y, Dawes JM, Antunes-Martins A, Bennett DL, McMahon SB, et al. Synthesis of lipid mediators during UVB-induced inflammatory hyperalgesia in rats and mice. PLoS One. 2013;8(12):e81228.
24. R Development Core Team: R: a language and environment for statistical computing. 2008.
25. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
26. Lacey LF, Keene ON, Pritchard JF, Bye A. Common noncompartmental pharmacokinetic variables: are they normally or log-normally distributed? J Biopharm Stat. 1997;7(1):171–8.
27. Elzhov TV, Mullen KM, Spiess A-N, Bolker B: Minpack.Lm: R Interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds. In.; 2016.
28. Milligan GW, Cooper MC. A study of standardization of variables in cluster analysis. J Classif. 1988;5(2):181–204.
29. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
30. Torgo L: Data mining with R: learning with case studies: chapman \& hall/CRC; 2010.
31. Kohonen T. Self-Organizing Maps. Berlin: Springer; 1995.
32. Good PI. Resampling methods : a practical guide to data analysis. Boston: Birkhäuser; 2006.
33. Tillé Y, Matei A: Sampling: survey sampling. In.; 2016.
34. Murtagh F, Hernández-Pajares M. The Kohonen self-organizing map method: an assessment. J Classif. 1995;12(2):165–90.
35. Ultsch A, Herrmann L. The architecture of emergent self-organizing maps to reduce projection errors. In: Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005): 2005 2005; Bruges, Belgium; 2003. p. 1–6.
36. Koch I. Analysis of multivariate and high-dimensional data. Cambridge: Cambridge University Press; 2013.
37. Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L *et al*: The shuttle radar topography mission. Rev Geophys 2007, 45(2):n/a-n/a.
38. Chang W, Cheng J, Allaire J, Xie Y, McPherson J: Shiny: web application framework for R. In.; 2016.
39. Ultsch A, Behnisch M, Lötsch J. ESOM Visualizations for Quality Assessment in Clustering. In: Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016. Edited by Merenyi E, Mendenhall MJ, O'Driscoll P, vol. 428. New York: Springer; 2016.
40. Lötsch J, Ultsch A. Process pharmacology: a pharmacological data science approach to drug development and therapy. CPT Pharmacometrics Syst Pharmacol. 2016;5(4):192–200.
41. Lötsch J, Hummel T, Ultsch A. Machine-learned pattern identification in olfactory subtest results. Sci Rep. 2016;6:35688.
42. Lötsch J, Dimova V, Hermens H, Zimmermann M, Geisslinger G, Oertel BG, Ultsch A. Pattern of neuropathic pain induced by topical capsaicin application in healthy subjects. Pain. 2015;156(3):405–14.
43. Lötsch J, Ultsch A. A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. Application to pain. J Biomed Inform. 2013;46(5):921–8.
44. Knothe C, Oertel BG, Ultsch A, Kettner M, Schmidt PH, Wunder C, Toennes SW, Geisslinger G, Lötsch J. Pharmacoepigenetics of the role of DNA methylation in mu-opioid receptor expression in different human brain regions. Epigenomics. 2016;8(12):1583–99.
45. Lötsch J, Thrun MC, Lerch F, Brunkhorst R, Schiffmann S, Thomas D, Tegeder I, Geisslinger G, Ultsch A. Machine-learned data structures of lipid marker serum concentrations in multiple sclerosis patients differ from those in healthy subjects. Int J Mol Sci. 2017; **in press**
46. Lötsch J, Ultsch A. Exploiting the structures of the U-matrix. In: Villmann T, Schleif F-M, Kaden M, Lange M, editors. Advances in intelligent systems and computing, vol. 295. Heidelberg: Springer; 2014. p. 248–57.
47. Ultsch A, Weingart M, Lötsch J. 3-D printing as a tool for knowledge discovery in high dimensional data spaces. In: Fürstberger A, Lausser L, Kraus JM, Schmid M, Kestler HA, editors. Statistical computing, vol. 2015-04. Schloss Reisensburg (Günzburg): Universität Ulm, Fakultät für Ingenieurwissenschaften und Informatik; 2015. p. 12–3.
48. Danzer KM, Kranich LR, Ruf WP, Cagsal-Getkin O, Winslow AR, Zhu L, Vanderburg CR, PJ ML. Exosomal cell-to-cell transmission of alpha synuclein oligomers. Mol Neurodegener. 2012;7:–42.
49. Crane PK, Gibbons LE, Dams-O'Connor K, Trittschuh E, Leverenz JB, Keene CD, Sonnen J, Montine TJ, Bennett DA, Leurgans S, et al. Association of Traumatic Brain Injury with Late-Life Neurodegenerative Conditions and Neuropathologic Findings. JAMA neurology. 2016;73(9):1062–9.
50. Celardo I, Martins LM, Gandhi S. Unravelling mitochondrial pathways to Parkinson's disease. Br J Pharmacol. 2014; 171(8):1943–57.
51. Sentelle RD, Senkal CE, Jiang W, Ponnusamy S, Gencer S, Selvam SP, Ramshesh VK, Peterson YK, Lemasters JJ, Szulc ZM, et al. Ceramide targets autophagosomes to mitochondria and induces lethal mitophagy. Nat Chem Biol. 2012;8(10):831–8.
52. Mayo L, Trauger SA, Blain M, Nadeau M, Patel B, Alvarez JI, Mascanfroni ID, Yeste A, Kivisakk P, Kallas K, et al. Regulation of astrocyte activation by glycolipids drives chronic CNS inflammation. Nat Med. 2014;20(10):1147–56.

53. Bras J, Singleton A, Cookson MR, Hardy J. Emerging pathways in genetic Parkinson's disease: potential role of ceramide metabolism in Lewy body disease. FEBS J. 2008;275(23):5767–73.
54. Lovinger DM, Mathur BN. Endocannabinoids in striatal plasticity. Parkinsonism Relat Disord. 2012;18(Suppl 1):S132–4.
55. Yang XY, Zhao EY, Zhuang WX, Sun FX, Han HL, Han HR, Lin ZJ, Pan ZF, Qu MH, Zeng XW, et al. LPA signaling is required for dopaminergic neuron development and is reduced through low expression of the LPA1 receptor in a 6-OHDA lesion model of Parkinson's disease. Neurol Sci. 2015;36(11):2027–33.
56. Pyszko JA, Strosznajder JB. The key role of sphingosine kinases in the molecular mechanism of neuronal cell survival and death in an experimental model of Parkinson's disease. Folia Neuropathol. 2014;52(3):260–9.
57. Gregoire L, Smith T, Senanayake V, Mochizuki A, Miville-Godbout E, Goodenowe D, Di Paolo T. Plasmalogen precursor analog treatment reduces levodopa-induced dyskinesias in parkinsonian monkeys. Behav Brain Res. 2015;286:328–37.
58. Hacioglu G, Seval-Celik Y, Tanriover G, Ozsoy O, Saka-Topcuoglu E, Balkan S, Agar A. Docosahexaenoic acid provides protective mechanism in bilaterally MPTP-lesioned rat model of Parkinson's disease. Folia Histochem Cytobiol. 2012;50(2):228–38.
59. Meng Q, Luchtman DW, El Bahh B, Zidichouski JA, Yang J, Song C. Ethyl-eicosapentaenoate modulates changes in neurochemistry and brain lipids induced by parkinsonian neurotoxin 1-methyl-4-phenylpyridinium in mouse brain slices. Eur J Pharmacol. 2010;649(1–3):127–34.
60. Martins M, Rosa A, Guedes LC, Fonseca BV, Gotovac K, Violante S, Mestre T, Coelho M, Rosa MM, Martin ER *et al*: Convergence of miRNA expression profiling, alpha-synuclein interacton and GWAS in Parkinson's disease. PLoS One 2011, 6(10):e25443. doi: https://doi.org/10.1371/journal.pone.0025443. Epub 0022011 Oct 0025447.
61. Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML, Eklund AC, Zhang-James Y, Kim PD, Hauser MA, et al. PGC-1alpha, a potential therapeutic target for early intervention in Parkinson's disease. Sci Transl Med. 2010; 2(52):52ra73.
62. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2(4):433–59.
63. Borg I, Groenen P. Modern multidimensional scaling: theory and applications. New York: Springer; 2005.
64. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319–23.
65. van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. J Mach Learn Res. 2008;9:2579–605.
66. Kullback S, Leibler RA. On information and sufficiency. Ann Math Statist. 1951;22(1):79–86.
67. Friedel M. Data-driven modeling of background and mine-related acidity and metals in river basins. Environ Pollut. 2013;184:530–9.
68. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.
69. Lötsch J, Geisslinger G, Heinemann S, Lerch F, Oertel BG, Ultsch A. Quantitative sensory testing response patterns to capsaicin- and ultraviolet-B-induced local skinhypersensitization in healthy subjects: a machine-learned analysis. Pain; 2017. https://doi.org/10.1097/j.pain.0000000000001008. [Epub ahead of print]
70. Lötsch J, Ultsch A. A machine-learned computational functional genomics-based approach to drug classification. Eur J Clin Pharmacol. 2016;72(12):1449–61.
71. Lötsch J, Thrun M, Lerch F, Brunkhorst R, Schiffmann S, Thomas D, Tegder I, Geisslinger G, Ultsch A. Machine-learned data structures of lipid marker serum concentrations in multiple sclerosis patients differ from those in healthy subjects. Int J Mol Sci. 2017;18:6.
72. Ultsch A. Pareto density estimation: a density estimation for knowledge discovery. In: Innovations in classification, data science, and information systems - proceedings 27th annual conference of the German classification society (GfKL). Berlin: Springer; 2003.