ELSEVIER

Checking the projection display of multivariate data with colored graphs

Bruno Bienfait and Johann Gasteiger

Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, D-91052 Erlangen, Germany

Projection methods such as principal component analysis (PCA), nonlinear mapping (NLM), and the self-organizing map (SOM) are valuable algorithms for visualizing multidimensional data in a two-dimensional plane. Unfortunately, the reduction of the dimensionality involves distortions. In an attempt to graphically localize the distortions of the projected data, we suggest superposing colored graphs onto the 2D plots. The color of the edges of these graphs encodes the original high-dimensional distances between the connected points. The method is applied to a cluster analysis of 37 biologically active compounds and 471 molecules represented by a structural 3D descriptor. © 1998 by Elsevier Science Inc.

Keywords: principal component analysis, PCA, nonlinear mapping, NLM, self-organizing map, SOM, Kohonen's network, minimal spanning tree, MST, k*-largest distortions graph,* k*-LDG, interatomic distances, histogram, frequency distribution, autocorrelation, QSAR*

INTRODUCTION

Multivariate data often consist of sets of high-dimensional vectors. In chemical applications, a vector could be a series of physical measurements or calculated properties made on a molecule. A dataset of compounds may be a series of related molecules collected for, e.g., a structure–activity study. If the vectors are only two-dimensional, they can be plotted in a plane. This allows the visual inspection of the structure of the dataset to identify clusters and particular objects, i.e., to perform an exploratory data analysis. When dealing with vectors whose dimensions are larger than two, it is not possible to represent them graphically in a plane. One way to overcome this problem is to transform the *N*-dimensional vectors into two dimensions. Many projection methods have been developed for this task. A good projection method preserves as faithfully as possible the original structure of the high-dimensional data. Unfortunately, the true distances between the vectors in the original high-dimensional space cannot be preserved exactly in the projected two-dimensional display. The two-dimensional plot thus obtained must distort in some way the original picture. Such distortions can cause misleading plots. Among the many papers concerned with the projection of multivariate data, the checking of the projections remains mostly an exception. The goal of this article is to present techniques that graphically reveal distortions due to dimensionality reductions. The main idea is to draw colored lines between some points represented on the 2D plot, where the color of each line is proportional to the original *N*-dimensional distance.

To illustrate the usefulness of this method, three different projection methods and two datasets of molecules are studied.

The first dataset deals with a series of 37 biologically active compounds represented by four physicochemical descriptors. The second dataset is much more complex. It is composed of 417 molecules divided in three structural groups. Each compound is encoded by a 512-dimensional vector calculated from the frequency distribution of the interatomic distances. For these two examples, three projection methods are applied: principal component analysis (PCA), nonlinear mapping (NLM), and the self-organizing map (SOM) neural network.

METHODS AND ALGORITHMS

Projection methods

Projection algorithms can be either supervised or unsupervised. Because this article deals with exploratory data structure analysis, only unsupervised methods are used. The algorithms described in this section are linear (principal component analysis) and nonlinear (nonlinear mapping, self-organizing map). Comparisons of the quality of projection methods have been described.[1,2]

Principal component analysis (PCA) is probably one of the most popular projection methods. Its principal feature is to rotate the vector space using the eigenvectors (principal components) of the covariance matrix as a new basis. The principal components corresponding to the two largest eig-

Color Plates for this article are on pages 254–258.

Address reprint requests to: Dr. B. Bienfait, Laboratory of Medicinal Chemistry, National Cancer Institute, National Institutes of Health, Building 37, Room 5B20, Bethesda, Maryland 20892, USA.
Received 11 June 1996; revised 26 June 1997; accepted 30 July 1997.

© 1998 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/97/$17.00
PII S0263-7855(97)00078-7

envalues (variance) are used to produce two-dimensional plots. The quality of the projection is commonly expressed by the retained variance of the first two principal components. In addition, plots of other components, such as the first against the third, etc., might be useful.

Nonlinear mapping (NLM) was introduced by Sammon.[3] The principle of NLM is to calculate for each high-dimensional vector a 2D vector such that all interpoint distances between the 2D vectors are as close as possible to the corresponding original interpoint distances in the high-dimensional space. The quality of the projection is defined here as the mapping error, i.e., a criterion calculated from the differences between interpoint distances in the high-dimensional and the two-dimensional spaces. NLM was introduced in chemistry by Kowalski and Bender[4] and was reviewed by Domine et al.[5]

The self-organizing map (SOM) is a neural network model developed to simulate organized maps of neurons found in the brain.[6] This neural network consists of a two-dimensional grid (or map) of artificial neurons. A high-dimensional vector is associated with each neuron. During a learning phase, the neurons are ordered on the map in such a way that neighboring high-dimensional vectors are mapped onto neighboring neurons. When used as a projection algorithm, the SOM produces two-dimensional vectors composed of the *x, y* coordinates of the neurons. The first application of SOM in chemistry (QSAR) was described by Rose et al.[7] Zupan and Gasteiger reported several chemical applications.[8]

Graphs with colored edges

As described earlier, one of the main problems of projection methods is that the interpoint distances on the plot are not necessarily equal to those calculated in the high-dimensional space. The idea introduced in this article is to connect some points with lines, where the color encodes the relative interpoint distance measured in the original high-dimensional space. Using graph terminology, the lines connecting two points are edges whereas the points are the vertices. The edges and the vertices constitute a graph. In this application, each edge has an associated weight equal to the high-dimensional interpoint distance.

Because the number of interpoint distances to be checked is proportional to $n(n-1)/2$, where n is the number of vertices, a selection among the interpoint distances to be displayed must be made. In this work, two selection mechanisms are used: a *k*-largest distortions graph and a minimum spanning tree.

k-*Largest distortions graphs (*k-*LDG)* For this graph, we simply select the k edges among the $n(n-1)/2$ edges that are associated with the largest distortions, i.e., the k largest absolute differences between the corresponding interpoint distance in the high-dimensional space and in the two-dimensional space.

Minimum spanning tree (MST) A minimum spanning tree is a weighted undirected graph for which (1) all vertices are connected, (2) no closed loops occur, and (3) the subset of edges is the one with the smallest sum of weights. Algorithms to solve the MST problem are well known from the literature.[9] A simple example of an MST is displayed in Color Plate 2. One important property of MSTs for our application is that each vertex is connected to its nearest neighbors. This feature allows one to detect neighborhood relationships that are distorted by the projection (see, for example, Color Plate 1).

Verification of projections in the literature

To assert the correctness of the projection, most authors reported only a global measurement, e.g., the amount of variance retained by the first principal components (for PCA) or the mapping error (for NLM). However, this is not sufficient. As shown by Gower and Ross, a two-dimensional PCA plot that accounts for 89% of the total variance still presents some distortions that lead to false interpretations.[10] Domine et al. plotted the local mapping error of each individual point on a two-dimensional nonlinear map.[5] This procedure reveals the individual distorted points but not the distorted neighborhood relationships between the points. Other works reported the application of a minimal spanning tree (MST) calculated in the original high-dimensional space but plotted using the projected two-dimensional coordinates.[2,11,12] Miyashita et al. have combined NLM and MST for a QSAR study.[11] A plot of an MST combined with the multidimensional scaling (MDS) projection was described by Wienke et al.[12] These authors also superposed an MST onto self-organizing maps. The MST is plotted with edges of four different thicknesses to symbolize four categories of distances from the original high-dimensional space. Minimum spanning trees were also applied by Biswas et al. to quantify the global distortion of seven projection algorithms.[2]

PROGRAMS

NLM and SOM projections were performed with public domain programs.[13] The colored plots shown in this article are generated using the encapsulated PostScript (EPS) graphics file format. The original program used to produce these files is freely available on the Internet (http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne).

RESULTS AND DISCUSSIONS

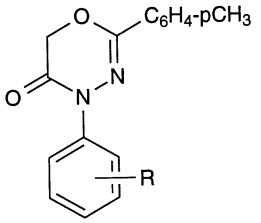
Dataset 1: Projection plots of thirty-seven 2,4-diphenyl-1,3,4-oxadiazin-5-ones

Dekeyser et al. synthesized thirty-seven 2,4-diphenyl-1,3,4-oxadiazin-5-ones and measured their acaracidal activities.[14] The chemical structures, physicochemical parameters, and the biological activities are collected in Table 1, as reported by Dekeyser et al. This dataset was also studied by Domine et al. for the evaluation of the NLM method in the fields of SAR and SPR.[5]

This example is a relatively small problem both in terms of vector dimension (only four parameters) and number of datapoints (only 37). The aim of this selection is not to repeat a complete QSAR study, but instead to take advantage of its relative simplicity to exemplify the new methods presented in this article.

The four physicochemical descriptors (π, F, R, MR) were selected from Table 1 as described by Domine et al.[5] They were autoscaled to have a mean of 0 and variance of 1. The four-dimensional data were projected using three methods: PCA, NLM, and SOM.

Table 1. Substituents, physicochemical descriptors, and efficacy data for thirty-seven 2-(4-methylphenyl)-4(*R*) phenyl-1,3,4-oxadiazin-5-one[a]



2-(4-Methylphenyl)-4(*R*) phenyl-1,3,4-oxadiazin-5-one

	R	π[b]	F[c]	R[d]	MR[e]	ED_{50}[f]
1	H	0	0	0	0	152
2	*o*-CH_3	0.84	−0.05	−0.11	4.7	>10 000
3	*m*-CH_3	0.52	−0.04	−0.04	4.7	33
4	*p*-CH_3	0.6	−0.04	−0.13	4.7	123
5	*o*-$CH_2C_6H_5$	2.01	−0.07	0	29	66
6	*o*-C_6H_5	2.39	0.1	−0.07	24.3	25
7	*p*-cyclo-C_6H_{11}	2.51	−0.17	0.04	25.7	905
8	*p*-NH_2	−1.3	0.02	−0.68	4.2	453
9	*m*-$N(CH_3)_3^+$	−5.96	0.87	−0.02	20.2	>1 000
10	*o*-NO_2	0.11	0.84	0.09	6	>10 000
11	*m*-NO_2	0.11	0.66	0.04	6	>10 000
12	*p*-NO_2	0.22	0.67	0.11	6	>823
13	*p*-OCH_3	−0.03	0.26	−0.53	6.5	29
14	*o*-$SO_2C_6H_5$	0.27	0.71	0.12	32.2	655
15	*o*-F	0	0.54	−0.32	−0.4	917
16	*p*-F	0.15	0.43	−0.37	−0.4	425
17	*o*-Br	0.84	0.55	−0.18	7.6	>1 000
18	*p*-Br	1.19	0.44	−0.21	7.6	171
19	*m*-$OCH_2C_6H_5$	1.66	0.21	−0.15	30.7	31
20	*p*-$OCH_2C_6H_5$	1.66	0.21	−0.43	30.7	26
21	*p*-C_2H_5	1.1	−0.05	−0.1	9.4	151
22	*o*-NH_2	−1.4	0.02	−0.59	4.2	>1 000
23	*m*-NH_2	−1.29	0.02	−0.24	4.2	>1 000
24	*p*-OH	−0.61	0.29	−0.66	1.5	>1 000
25	*m*-OCH_3	0.12	0.25	−0.18	6.5	728
26	*p*-SO_2CH_3	−1.2	0.54	0.18	12.5	714
27	*m*-F	0.22	0.43	−0.13	−0.4	1 000
28	*o*-Cl	0.76	0.51	−0.16	4.8	>1 000
29	*o*-C_2H_5	1.39	−0.06	−0.09	9.4	120
30	*m*-C_2H_5	0.99	−0.05	−0.04	9.4	116
31	*m*-SCH_3	0.64	0.19	−0.07	13	117
32	*p*-C_6H_5	1.74	0.08	−0.09	24.3	902
33	*p*-$CO_2C_2H_5$	0.46	0.33	0.12	16.2	>2 251
34	*p*-CO_2H	−0.32	0.33	0.12	5.9	>4 000
35	*m*-$CH_2C_6H_5$	2.01	−0.05	0	29	30
36	*m*-C_6H_5	1.92	0.08	−0.03	24.3	34
37	*p*-$OCONHCH_3$	−0.42	0.41	−0.15	15.3	>1 000

[a] Data taken from Refs. 5 and 14.
[b] Lipophilicity.
[c] Inductive effect.
[d] Resonance effect.
[e] Molar refractivity.
[f] Acaracidal efficacy data.

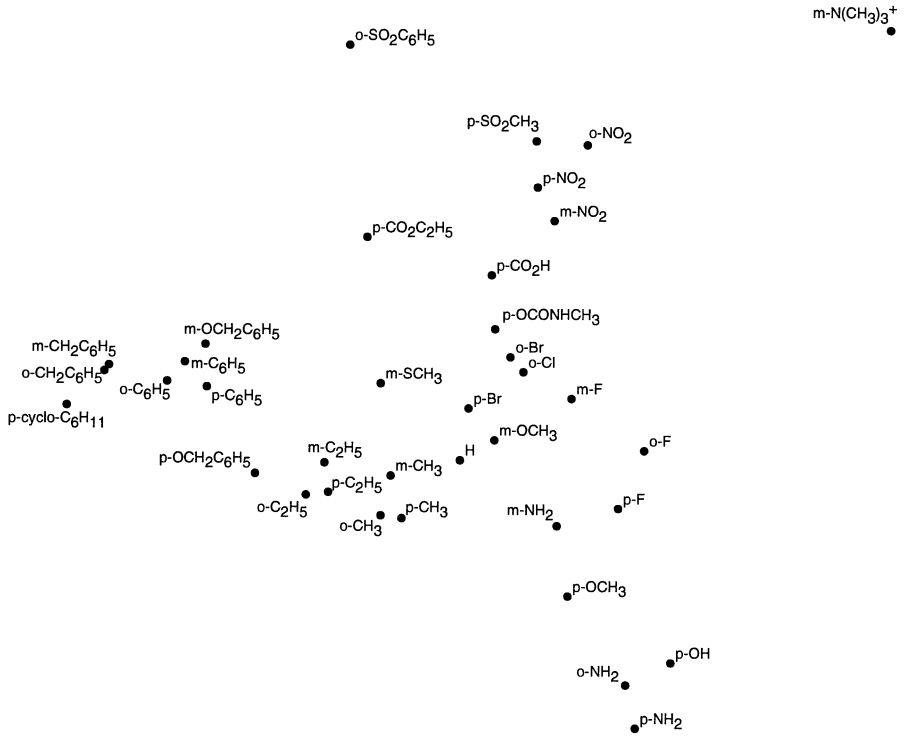


Figure 1. Dataset 1. Plot showing the projection calculated with principal component analysis (PCA) in the space defined by the two first principal components (74% of total variance). The labels show the nature of the R substituent (cf. Table 1).

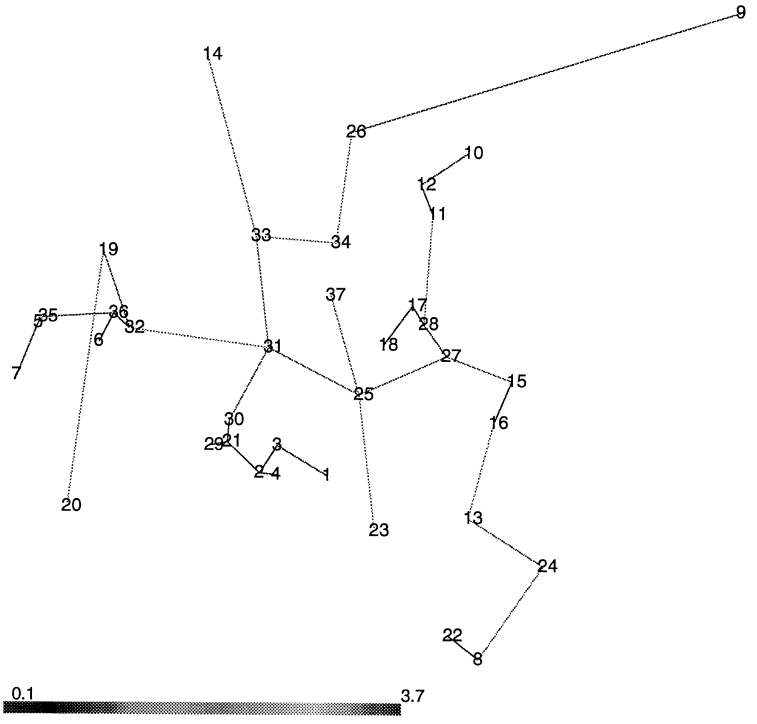


Figure 2. Dataset 1. Superposition of the minimal spanning tree (MST) calculated using the 4D data of Table 1 onto the 2D plot obtained with NLM. The colored scale at the bottom left encodes the Euclidean distance calculated in the original 4D space. The original color graphics file is available at our WWW site: http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne/Publication/.

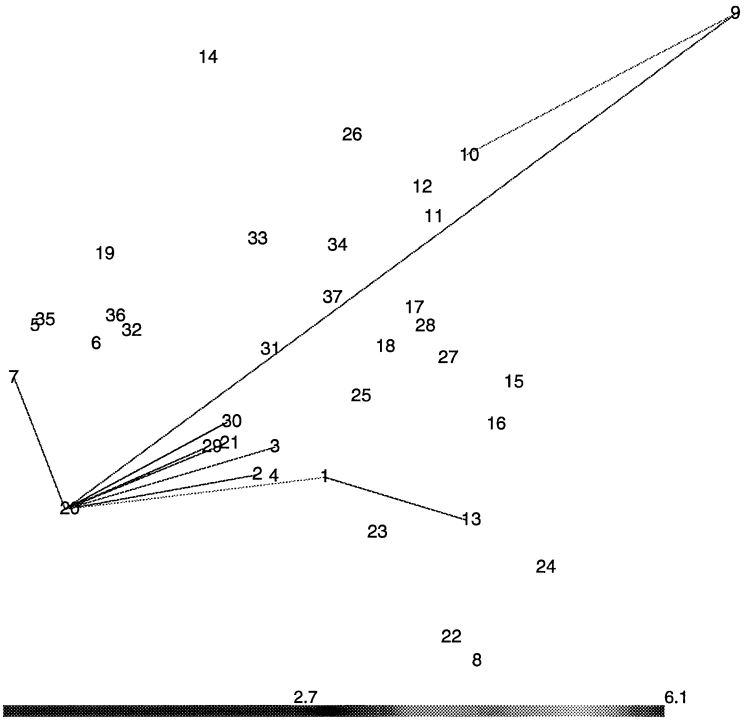


Figure 3. Dataset 1. Superposition of a graph showing the 10 largest distortions (10-LDG) onto the NLM plot. The colored scale is different from the one of the MST shown in Color Plate 3. The original color graphics file is available at our WWW site: http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne/Publication/.

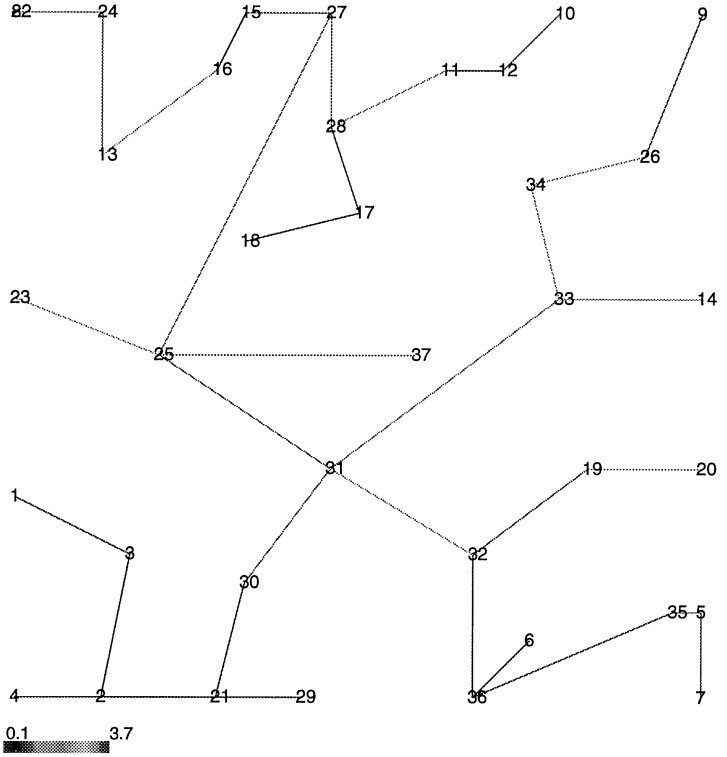


Figure 4. Dataset 1. Superposition of the minimal spanning tree (MST) calculated using the 4D data of Table 1 onto the 2D plot obtained with a 25 × 25 Kohonen SOM. The original color graphics file is available at our WWW site: http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne/Publication/.

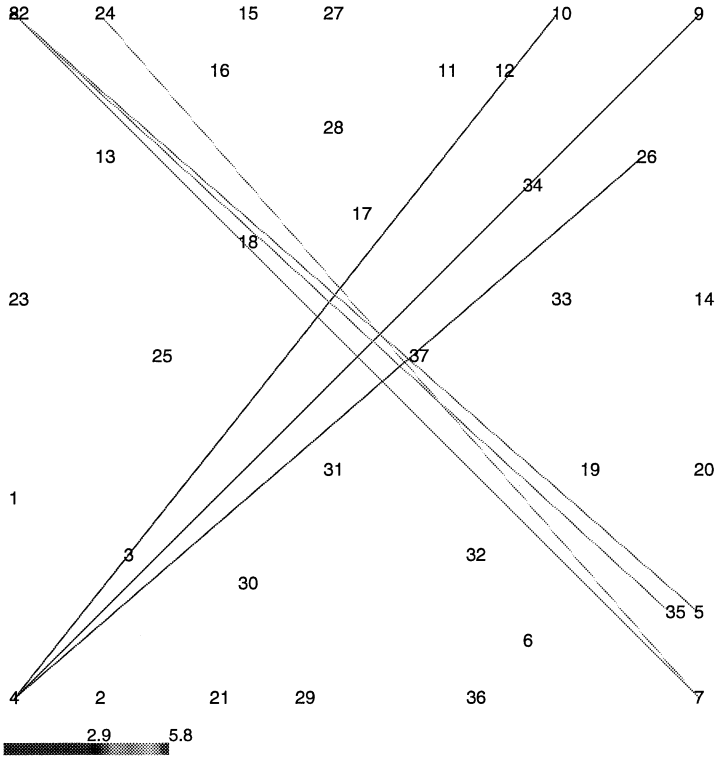


Figure 5. Dataset 1. Superposition of a graph showing the 10 largest distortions (10-LDG) onto the SOM plot. The original color graphics file is available at our WWW site: http://schiele.organik.uni-erlangen.de/BrunoBienfait/Spinne/Publication/.

1 · Projection with PCA The PCA projection obtained from the autoscaled vectors from Table 1 is displayed in Figure 1 and Color Plate 1. The two axes of the plot are those that have the largest variance. The sum of the two largest variances is 74.2%, which is identical to the value reported by Domine et al.[5] Each molecule of Table 1 is labeled by its entry number in Color Plate 1. A minimal spanning tree with colored edges is superposed onto the plot. This tree has been calculated in the original four-dimensional space and plotted using the two-dimensional coordinates obtained with PCA. In the context of this example, the MST is a weighted graph whose vertices are the four-dimensional autoscaled vectors calculated from Table 1, and whose edges are weighted by the Euclidean distances between the connected vertices. The edges are colored according to the Euclidean distances calculated in the original four-dimensional space (Color Plate 1). The colored scale at the bottom of the plot shows the range of the original Euclidean distances. The shortest distance (0.1) is coded in violet whereas the largest is coded in red (3.7). Comparison of the color of the edges and the colored scale allows one to visually detect distortions due to the projection.

To ease the visual interpretation of the superposed colored MST plots of the kind displayed in Color Plate 1, we describe and illustrate three rules.

Rule 1: Crossing lines reveal distortions.

Two pairs of crossing lines are visible (37–25, 17–18 and 19–20, 31–32). This is a first indication of a distortion due to the projection. The absence of crossing lines is a condition for achieving the shortest path. Indeed, an MST connects all points such that the total length of the path is minimal. In other words, the MST in the original 4D space has no crossing lines, but its distorted representation in 2D may show some.

Rule 2: If a vertex v_1 of an MST is a leaf, then it is connected to only one other vertex, which is its nearest neighbor v_2. Consequently, all distances between this vertex v_1 and all other vertices must be larger or equal to the distance between the vertex v_1 and its nearest neighbor v_2.

The point labeled 37, which is a leaf, is connected to point 25. Consequently, the 4D distances between point 37 and all other points must be larger or equal to distance 37–25. Indeed, the Euclidean distances calculated in the original 4D space are 1.11 for edge (25–37), 1.26 for edge (17–37), 1.37 for edge (18–37), and 1.38 for edge (28–37).* On the PCA projected plot, this is not true: Label 37 is closer to labels 17, 28, and 18 than to label 25 (Color Plate 1). Therefore, the relative positions of these points are distorted by the projection.

Rule 3: From an adjacency set $A(v)$ containing all the vertices of the MST connected to a vertex v, we can select $d(v)$, that is the shortest distance among the distances calculated between v and all the other vertices included in $A(v)$. Then, any distance between v and another vertex not included in the adjacency set $A(v)$ must be larger or equal to $d(v)$.

Note that the second rule is a particular case of this rule, where the number of vertices included in $A(v)$ is 1. An illustrative example is the relative positions of the points

* Note that these distances cannot be deduced from the plot. They are calculated from the 4D coordinates using another program.

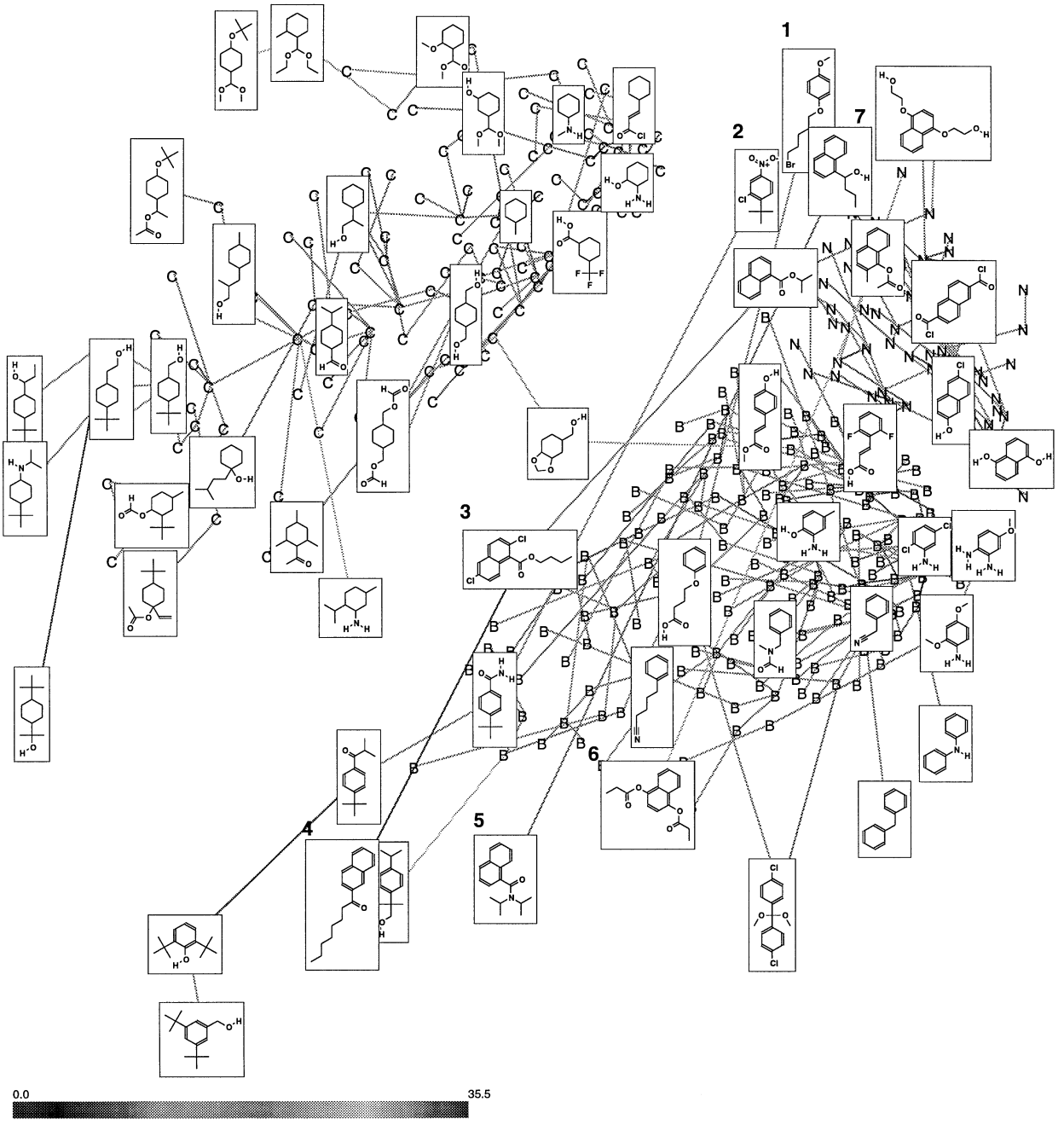


Figure 6. Dataset 2. Superposition of a minimal spanning tree (MST) onto the NLM plot. A part of the molecular structures of the dataset is displayed. The original color graphics file is available on our WWW site: http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne/Publication/.

labeled 26, 10, 12, and 11 (Color Plate 1). Label 26 is connected to labels 9 and 34. With the help of the colored scale placed at the bottom of the plot (the red color represents the largest distance), the shortest distance edge can be determined: (26–34). The distances of edges (10–26), (11–26), and (12–26) must be larger than or equal to the distance of edge (26–34).† On the projected plot, this is not true: Label 26 is closer to labels 10, 11, and 12 than to label 34. Therefore, the relative positions of these points are not correct.

Color Plate 2 also shows an MST superimposed on the 2D PCA projection. Unlike the previous plot, this MST is calculated from the 2D coordinates. The edges are colored according to the original 4D space. Note that if the projection were free of distortions, the MST calculated in the 2D space would be identical to the MST calculated in the 4D space. The interest of this representation is to view other interpoint distances. The

† The Euclidean distances calculated in the original 4D space are as follows: 1.19 for edge (26–34), 1.29 for edge (12–26), 1.35 for edge (11–26), and 1.59 for edge (10–26).

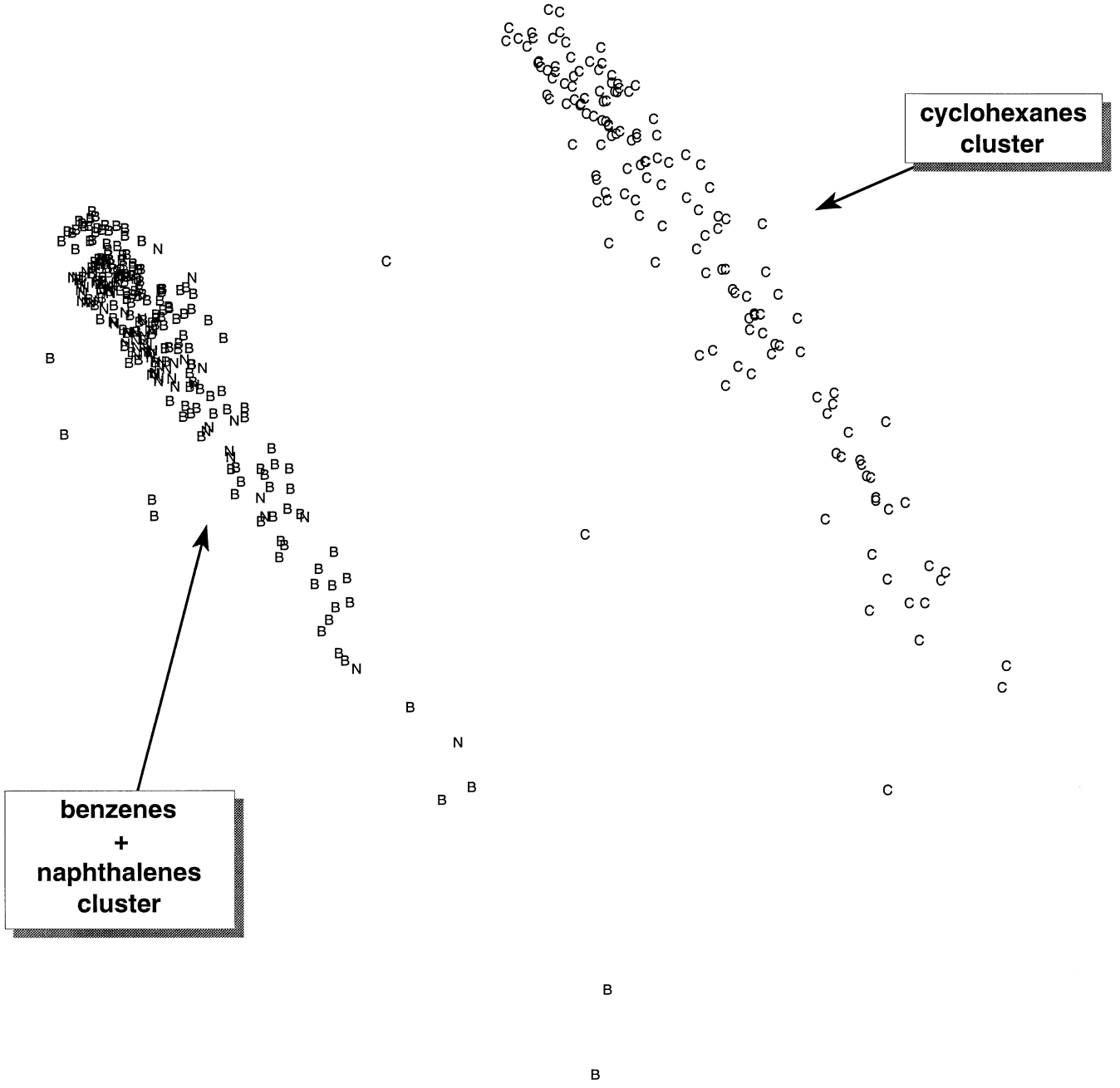


Figure 7. Dataset 2. Plot showing the projection calculated with principal component analysis (PCA) in the space defined by the two first principal components (57% of total variance). The labels of the cyclohexane, benzene, and naphthalene derivatives are, respectively, C, B, and N.

distortions are detected only by comparing the distance on the paper with the distance reveal by the color (the three rules described above cannot be used because this MST was calculated from the 2D coordinates). For instance, the points representing compounds 20 (R = *p*-$OCH_2C_6H_5$) and 29 (R = *o*-C_2H_5) are much too close to each other.

The plot of Color Plate 3 shows the 10 edges associated with the largest differences between the interpoint distances in the original 4D space and in the 2D space. Unlike the plots based on MSTs, the selected edges do not connect only close points. For the PCA-based projection, the edge associated with the largest difference is (20–30). The Euclidean distance in the 4D space is 3.0, whereas in the projected space its value is only 0.5. The points labeled 20, 9, and 1 are those that are shared among the largest number of edges. One can say that these points are the least well placed on the PCA plot.

Chemical interpretation of the plots. The aim of this article is not to present a complete interpretation of the chemical relevance of the projections. However, it is of interest to point out a few cases in which the superimposed colored graph can help to correct the interpretation.

Globally, the map is organized as follows (Figure 1). The

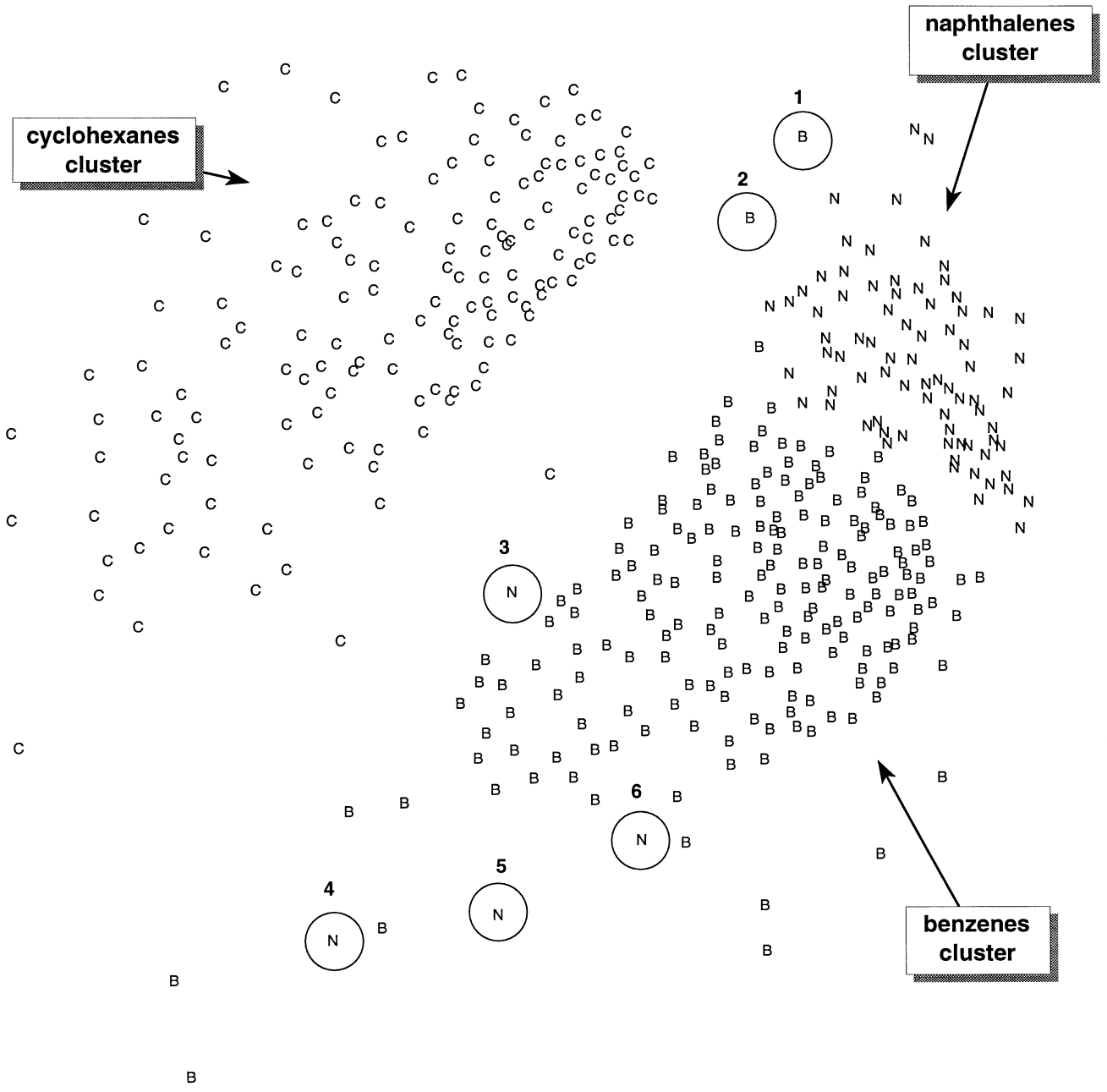


Figure 8. Dataset 2. Projection with nonlinear mapping (NLM). Points that appear to be misclassified are marked by a circle. The corresponding molecules are shown in Figure 9.

polarity of the substituents decreases from left to right whereas the electron-withdrawing strength of the substituents increases from the bottom to the top.

On the PCA plot of Color Plate 1, the points labeled 10, 26, 12, and 11 appear as a cluster. Compounds **10**, **11**, and **12** are the three derivatives of the dataset that bear a nitro substituent. Compound **26**, on the other hand, is substituted by a mesylate. When looking at the MSTs of Color Plates 1 and 2, one can deduce that labels 10, 11, and 12 are close neighbors whereas label 26 should be placed farther away. Thus, mesylate derivative **26** does not belong to the nitro cluster.

Compounds **30**, **21**, and **29** are substituted by an ethyl group. They form a cluster on the PCA plot as shown by the violet color of the connections (Color Plate 1). On the right and left side of this cluster are placed, respectively, the methyl cluster (compounds **2**, **3**, and **4**) and the *p*-phenylethoxy derivative **20**. The ethyl cluster appears slightly closer to compound **20** than to the methyl cluster. This closeness is not correct as proved by the three graphs displayed in Color Plates 1, 2, and 3:

The MST of Color Plate 1 shows that the nearest neighbor of compound **20** is not ethyl derivative **29** but compound **19**, the other phenylethoxy derivative.

The MST of Color Plate 2 reveals a 4D distance of 2.8 between labels 20 and 29. This value is much larger than the 4D distance between the two alkyl clusters (0.5).

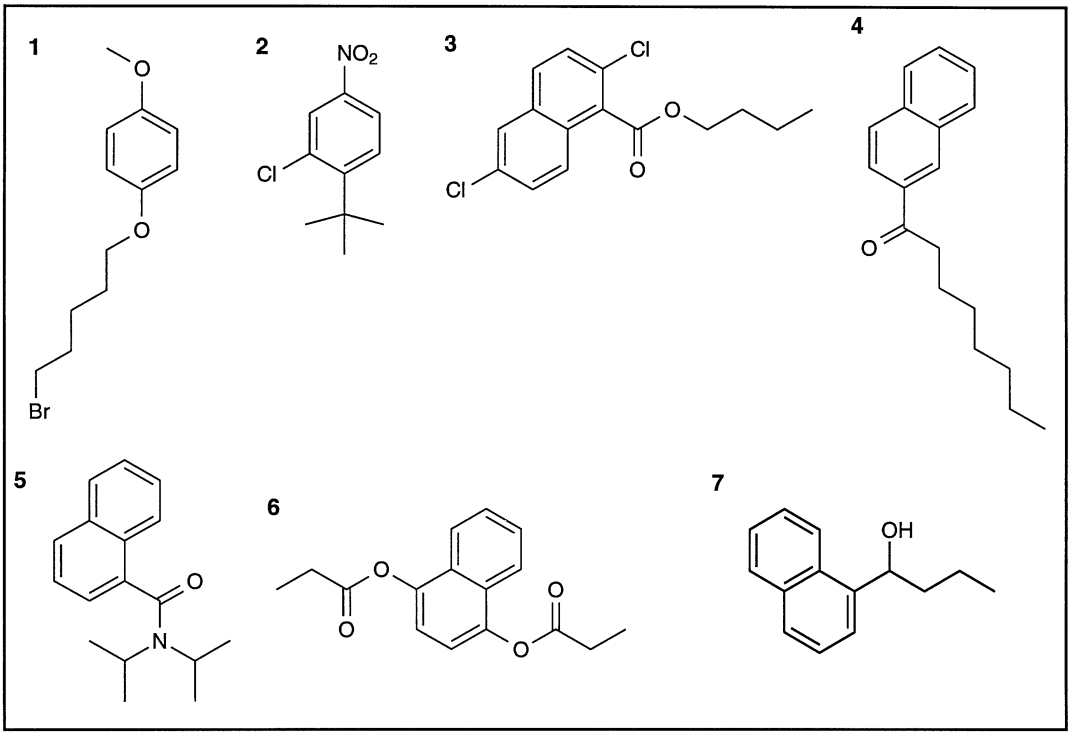


Figure 9. Selection of molecular structures from the dataset for color plates 4, 5, 6 and figures 8 and 10.

The *k*-LDG of Color Plate 3 shows that phenylethoxy **20** is much too close to alkyl derivatives **30**, **3**, and **2**.

2 · Projection with NLM The nonlinear map displayed in Figure 2 has a relatively low global mapping error of 0.021, which is an indication of a good projection. It was calculated using the 2D PCA coordinates as initial values, hence the same orientation of the two kinds of projection. The superimposed MST reveals only one crossing line. Some of the projection problems described previously for the PCA projection are now absent. Compound **26** is now farther away from the nitro cluster (**10**, **11**, and **12**). The phenylethoxy derivative **20** is not mapped as close to the alkyl derivatives. Thus, for this example, the nonlinear map provides a better projection than the PCA plot.

The plot of Figure 3 shows the 10 edges associated with the largest absolute differences between the interpoint distances in the original 4D space and the interpoint distances in the projected 2D space (10-LDG). The point labeled by 20 has the largest number of edges. This means that even if this point is better projected than on the PCA plot, it is the least well placed on the nonlinear map. Domine et al. arrived at a similar conclusion by graphically representing each individual mapping error on the map.[5] The superposition of a *k*-LDG is more informative because it displays the interpoint distances.

3 · Projection with an SOM The projection obtained with the SOM is displayed in Figure 4. This map is a grid composed of 25 × 25 neurons. Unlike the previous methods (NLM, PCA), the coordinates of the projected points are integers ranging from 0 to 24. Consequently, the SOM must not be interpreted in terms of interpoint distances but instead in terms of neighborhood relationships. The superimposed MST with colored edges helps to recover the lost interpoint distances. Because of the discrete nature of the coordinates, the superposition of largest distortions graphs cannot be used for checking the projection obtained with an SOM (Figure 5).

Dataset 2: Projection plots of 417 derivatives of three ring systems encoded with a 3D molecular descriptor

The availability of programs able to quickly generate three-dimensional structures from chemical connection tables[15] has promoted the development of new molecular descriptors based on the spatial relationships among the atoms or among points on the molecular surface. These structural descriptors are invariant to changes in atom numbering, molecular translation, and global rotation. The automatic 3D builder CORINA was used in our group to study two structural descriptors for QSAR,[16,17] for the simulation of infrared spectra,[17] and to obtain the 3D structure from infrared spectra.[18] In this work, we test a third structural descriptor, *interatomic distances histograms*, which is equivalent to the previously described spatial autocorrelation descriptor[19] in which the property on each atom is equal to 1. A histogram is a simple statistical tool used to visualize the distribution of a large number of observations with bar charts. Histograms are also called frequency distributions or frequency diagrams. Like other 3D structural descriptors, the interatomic distances histograms are *n*-dimensional vectors, which can be analyzed by various statistical or neural network methods.

The dataset includes 417 molecular structures selected from the SpecInfo datafile.[20] It is the same as the one described by Schuur et al.[17] It includes three classes of compounds: 148 cyclohexane, 188 benzene, and 81 naphthalene derivatives.

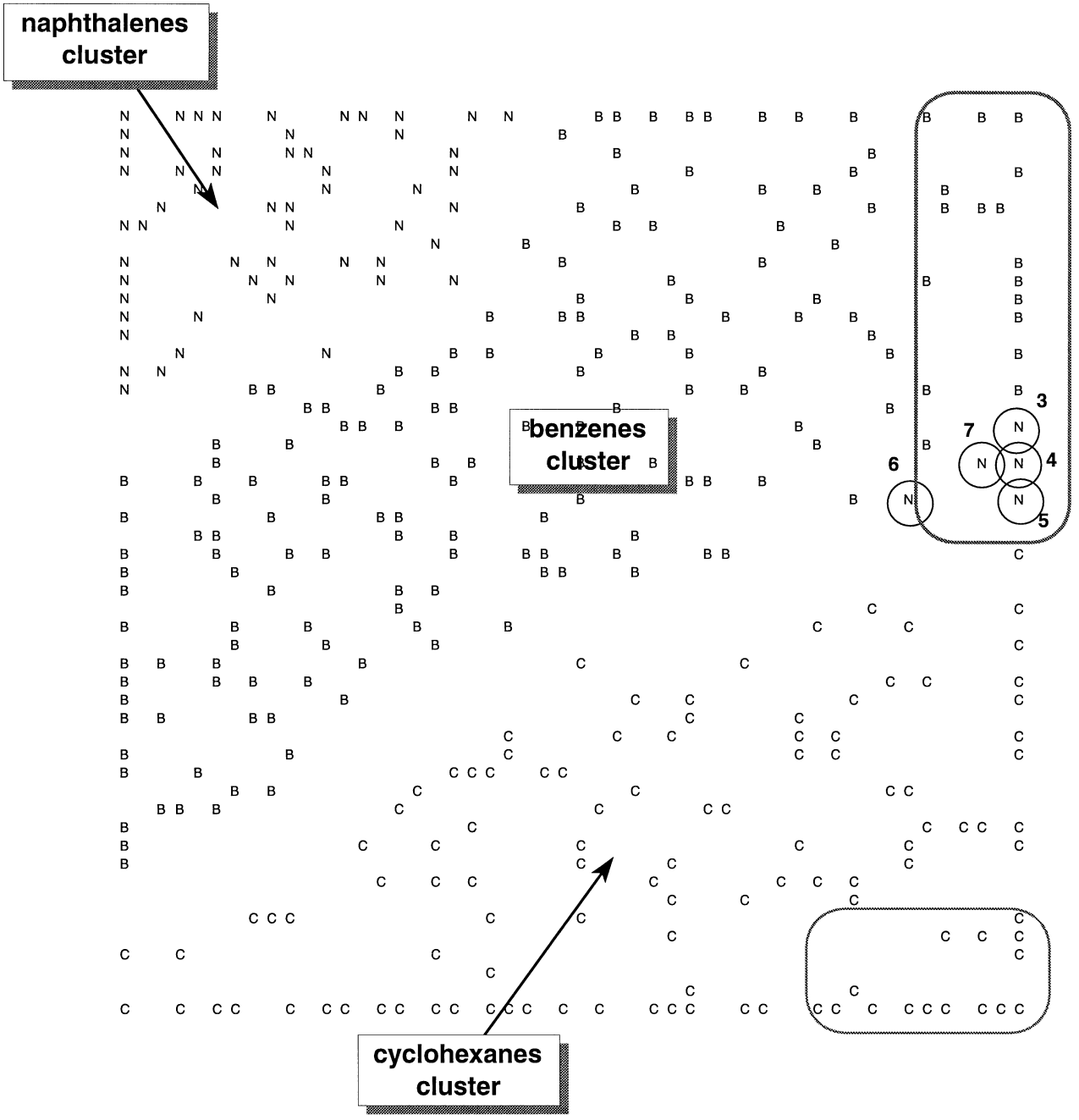


Figure 10. Dataset 2: Projection with a self-organizing map (SOM) with 50 × 50 neurons). The points that appear to be misclassified are marked by a circle. The corresponding molecules are shown in Figure 9. The two regions marked by a gray line include compounds with many alkyl substituents.

Examples of molecules belonging to each of the three classes are shown in Figure 6. The 3D atomic coordinates were calculated with CORINA, a program for generating 3D structures from connection tables and stereo information.[15] The interatomic distances histograms were calculated for 512 intervals of interatomic distances ranging from 0.9 to 7.9 Å. Thus, each of the 417 molecules of the dataset is represented by a 512-dimensional vector. Using such a large number of elements and such high-dimensional vectors makes this dataset much more difficult to project correctly into a 2D plane.

4 · Projection with PCA The three classes of compounds that are revealed by labels C, B, and N in the PCA plot shown in Figure 7 are the cyclohexanes, benzenes, and naphthalenes, respectively. Two dense clusters can be identified: cluster C and cluster B/N. The two classes of aromatic derivatives appear on the plot as one group. This representation is not correct, as we will see later from the nonlinear projection plot (cf. Figure 8). The two first principal components include only 57% of the total variance, which indicates the high-dimensional contents of the data. Because of the high density of the clusters, the

superposition of colored graphs (not shown here) onto this 2D plot does not help to distinguish the two different aromatic clusters.

5 · Projection with NLM The NLM projection performed on the dataset has a mapping error of 0.067. This value is about three times higher than the mapping error of the NLM projection of the first dataset. Unlike the PCA plot, three clusters corresponding to the three classes of compounds can be identified (Figure 8). Compounds with alkyl substituents are more numerous in the left and bottom regions of the plots (Figure 6). The two classes of aromatic derivatives labeled B and N form two separate clusters, with the exception of six points, which are not placed near their respective cluster centers. These particular points are marked on the projection plot with circles (Figure 8). Are these problems due to the molecular representation or to the projection? We answer this question in the following sections by superimposing colored graphs onto the plot.

Superposition of a minimal spanning tree (MST) onto the NLM plot. The MST reveals a great deal of distortion in this projection (Color Plate 4): many crossing lines and mismatches between the color of the edges and the scale displayed at the bottom of the plot (see Fig. 9 for molecules corresponding to apparently misclassified points). The MST connects closest neighbors. It also reveals that the nearest neighbors of the points that appear to be misclassified are connected to points of the same class.

Superposition of a *k*-largest distortions graph (*k*-LDG) onto the NLM plot. This graph connects points whose 2D interpoint distances differ the most from the original 512-dimensional interpoint distances. In other words, it reveals the largest distortions of the projection. The 30 edges with the largest distortions are displayed in Color Plate 5. The plot shows that some misclassified points, e.g., the two circles 4 and 5 at the bottom of the plot, are much too close to their immediate neighbors.

The two graphs described above show that the points marked by circles are not correctly projected by the NLM algorithm. Therefore, the problem of the misclassified points is not caused by the molecular representation but the projection.

SOM plot

The result obtained with a self-organizing map (SOM) is displayed in Figure 10. This projection must be interpreted differently from an NLM plot. An SOM does not try to preserve the original *N*-dimensional interpoint distances but, instead, the neighborhood relationships from the high-dimensional space.

Superposition of a minimal spanning tree (MST) onto the self-organizing map (SOM) The MST superimposed onto the SOM plot is displayed in Color Plate 6. Unlike in the NLM projection, no member of the benzene class appears to be misclassified. Four of the misclassified naphthalenes (circles 3–6) are the same as those on the NLM plot. One possible interpretation is that these naphthalene derivatives belong to two different classes: primarily to the naphthalene class, as shown by the lines of the MST, and secondarily to the class of compounds bearing many alkyl groups. These compounds are found predominantly in regions located at the opposite ends of the plots when compared with the naphthalenes derivatives cluster (cf. Figure 6 and Color Plate 6). For example, in Figure 6, the majority of naphthalene derivatives are located in the upper right corner of the plot, whereas compounds bearing many alkyl groups are mostly found in the lower left-hand corner.

Finally, in both projection plots, the compounds that appear to be misclassified are leaves, or are close to the leaves, of the minimal spanning tree.

CONCLUSIONS

The data structures of two datasets composed of high-dimensional vectors were visually analyzed with the help of two-dimensional plots. The reduction of the dimensionality was performed using three projection methods: principal component analysis, Sammon's nonlinear mapping, and Kohonen's self-organizing map. Each projection method necessarily distorts the original high-dimensional data somewhat. The amount of distortion is indicated by the superposition of graphs with colored edges onto the plots. Two kinds of graphs were applied: the minimum spanning trees reveal mainly misleading close neighbors whereas the *k*-largest distortions graphs detect the most strongly distorted interpoint distances. An exploratory data structure analysis was performed on two datasets of molecular structures. In both examples, the chemical interpretation of the projection plots is corrected and enhanced. However, plots with high densities of points projected from a high-dimensional space form clutters difficult to analyze with superimposed graphs.

We believe that when high-dimensional data are projected onto a two-dimensional space, the possible introduction of distortions should always be checked for.

ACKNOWLEDGMENTS

B. Bienfait expresses his thanks to the Alexander von Humboldt Foundation and the European Community for financial support. The largest distortions graph was suggested by M. Wagener.

REFERENCES

1 Kraaiveld, M.A. and Mao, J. A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Trans. Neural Networks* 1995, **6,** 548–559
2 Biswas, G., Jain, A.K., and Dubes, R.C. Evaluation of projection algorithms. *IEEE Trans. Pattern Anal. Machine Intell.* 1981, **PAMI-3,** 701–708
3 Sammon, J.W., Jr. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 1969, **C-18,** 401–409
4 Kowalski, B.R. and Bender, C.F. Pattern recognition. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* 1972, **94,** 5632–5639
5 Domine, D., Devillers, J., Chastrette, M., and Karcher, W. Non-linear mapping for structure–activity and structure–property modelling. *J. Chemometrics* 1993, **7,** 227–242
6 Kohonen, T. *Self-Organizing Maps.* Springer-Verlag, Heidelberg, Germany, 1995
7 Rose, V.S., Croall, I.F., and MacFie, H.J.H. An application of unsupervised neural network methodology (Kohonen topology-preserving mapping) to QSAR analysis. *Quant. Struct.-Activity Relat.* 1991, **10,** 6–15

8 Zupan, J. and Gasteiger, J. *Neural Networks for Chemists—an Introduction.* VCH, Weinheim, Germany, 1993
9 Cormen, T.H., Leiserson, C.E., and Rivest, R.L. *Introduction to Algorithms.* MIT Press, Cambridge, Massachusetts, 1994
10 Gower, J.C. and Ross, J.S. Minimum spanning trees and single linkage cluster analysis. *Appl. Stat.* 1969, **18,** 54–64
11 Miyashita, Y., Takahashi, Y., Yotsui, Y., Abe, H., and Sasaki, S. Application of pattern recognition to structure–activity problems. Use of minimal spanning tree. *Anal. Chim. Acta* 1981, **133,** 615–624
12 Wienke, D., Xie, Y., and Hopke, P.K. Classification of airborne particles by analytical scanning electron microscopy imaging and a modified Kohonen neural network (3MAP). *Anal. Chim. Acta* 1995, **310,** 1–14
13 SOM_PAK, the Self-Organizing Map Program Package, was prepared by the SOM Programming Team of the Helsinki University of Technology, Laboratory of Computer and Information Science, Rakentajanaukio 2 C, SF-02150, Espoo, Finland. It is available free of charge by anonymous FTP connection on the Internet (cochlea.hut.fi or 130.233.168.48). Both UNIX and MS-DOS versions are available
14 Dekeyser, M.A., Borth, D.M., Moore, R.C., and Mishra, A. Quantitative structure–activity relationships in acaracidal 4*H*-1,3,4-Oxadiazin-5(6*H*)-ones. *J. Agric. Food Chem.* 1991, **39,** 374–379
15 Sadowski, J., Gasteiger, J., and Klebe, G. Comparison of automatic 3-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 1000–1008
16 Wagener, M., Sadowski, J., and Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* 1995, **117,** 7769–7775
17 Schuur, J.H., Selzer, P., and Gasteiger, J. The coding of three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 334–344
18 Steinhauer, L., Steinhauer, V., and Gasteiger, J. Obtaining the 3D structure from infrared spectra of organic compounds using neural networks. *In: Software Development in Chemistry 10* (Gasteiger, J., ed.). Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1996
19 Broto, P., Moreau, G., and Vandycke, C. Molecular structures: Perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.-Chim. Ther.* 1984, **19 66–70**
20 Bremser, W. Structure elucidation and artificial intelligence. *Angew. Chem.* 1988, **100,** 252–265; *Angew. Chem. Int. Ed. Engl.* 1988, **27,** 247–260
21 Moreau, G. and Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau J. Chim.* 1980, **4,** 359–360
22 Jakes, S.E. and Willett, P. Pharmacoric pattern matching in files of 3-D chemical structures: Selection of interatomic distance screens. *J. Mol. Graphics* 1986, **4,** 12–20
23 Soltzberg, L.J. and Wilkins, C.L. Molecular transforms: A potential tool for structure–activity studies. *J. Am. Chem. Soc.* 1977, **99,** 439–443
24 Bemis, G.W. and Kuntz, I.D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Design* 1992, **6,** 607–628
25 Bath, P.A., Poirrette, A.R., Willett, P., and Allen, F.H. Similarity searching in files of three-dimensional chemical structures—comparison of fragment-based measures of shape similarity. *J. Chem. Int. Comput. Sci.* 1994, **34,** 141–147