Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Representing structural databases in a self-organizing map

Ron Wehrens, Willem Melssen, Lutgarde Buydens* and René de Gelder

Institute for Molecules and Materials, Radboud University Nijmegen, Toernooiveld 1, Nijmegen, 6525 ED, The Netherlands

Correspondence e-mail: l.buydens@science.ru.nl

Received 9 March 2005
Accepted 27 June 2005

This paper presents a way to accomodate large numbers of crystal structures, as present in *e.g.* the Cambridge Structural Database (CSD), in a self-organizing map. The structures are represented by their calculated powder diffraction patterns. The use of a recently introduced similarity criterion is essential: the weighted cross-correlation. This accurately reflects the similarities of the powder patterns and therefore, indirectly measures the resemblance of crystal packings. It will be shown that good results are obtained, even if the network is trained with a small subset of a complete database. This makes it possible to construct the map on common hardware in a few hours. Such a map provides several possibilities for two-dimensional visualization, but additionally has a number of important applications. Two such applications are fast and easy screening of a database, and providing an overview of the contents of a database in terms of structural diversity of specific chemical classes of compounds, *e.g.* steroids or peptides. A third is the selection of archetypical structures, covering the complete structural space.

1. Introduction

A unique and practical representation of a crystal structure is not easy to define and, therefore, sophisticated methods are needed to assess structural similarity. A powder diffraction pattern is a global descriptor of a crystal structure or, more precisely, of the periodic electron density in a crystal. Such a descriptor can be calculated from database entries and, moreover, can be obtained experimentally. In general, the similarity between powder diffraction patterns reflects the packing similarity between underlying crystal structures. This way of expressing the relations between crystal structures is different from, and in a way more general than, the traditional criteria of (reduced) cell dimensions and space group.

The positions of the peaks in powder diffraction patterns are very sensitive to small deviations in the unit-cell parameters, whereas the variations in the scattering power of atoms influence the intensities of the peaks. As a result, strongly related structures may give powder patterns that look similar from an overall point of view but may differ significantly on a more local scale. For this reason comparison of powder diffraction patterns, for the purpose of analysing structural similarity, is not straightforward.

The recently introduced weighted cross-correlation (WCC; de Gelder *et al.*, 2001) is a similarity measure for patterns in which the primary information is in both the positions and the amplitudes of the features. It is based on cross-correlation and therefore uses a neighbourhood in the calculation of similarity.

This neighbourhood is taken into account using a triangular weighting function, with a user-defined width. The validity of WCC-based similarities has been shown in several applications, such as the clustering of powder patterns (de Gelder *et al.*, 2001; Willighagen *et al.*, 2005), finding unit-cell parameters from powder patterns (also known as indexing; Hageman *et al.*, 2003) and identifying rotational constants from high-resolution fluorescence spectra (Hageman *et al.*, 2000; Meerts *et al.*, 2004).

The WCC can also be used to investigate the relations between compounds in a large database. However, direct pairwise comparisons are not computationally feasible given the current database sizes and since new data are being generated at an ever faster rate, this problem will become even more difficult in the future. One way to avoid this is to provide a mapping of all the database compounds to two dimensions. This kind of visualization is an appealing way to obtain a complete overview of a large database. Several general methods exist, but not all are suitable for powder diffraction patterns. As an example, principal component analysis (PCA; Jackson, 1991) in essence compares data on a point-by-point basis and as a result distorts similarity relations when applied to powder patterns. In fact, one should start from similarities, such as calculated by a measure like the WCC. Performing PCA on the similarity matrix, however, is not easy, if possible at all, because of the size of the matrix. The same problem exists for multi-dimensional scaling (MDS; Mardia *et al.*, 1979), a collection of non-linear extensions of PCA, based on the similarity matrix. In MDS, the objects are positioned in a two-dimensional space in such a way that distances in the plane are a direct measure of dissimilarities in the original space. While this is clearly a desirable feature, the method does not provide a direct mapping to two dimensions; for new data, the (dis)similarity matrix should be recalculated and the MDS repeated.

An alternative is formed by a class of artificial neural networks called Self-Organizing Maps (SOMs) or Kohonen maps (Kohonen, 1982, 2001). Objects are mapped to a two-dimensional grid of units, rather than a continuous space, in such a way that very similar objects are mapped to the same unit or to units which are close together in the map. Thus, it is topology rather than dissimilarity that is graphically represented in the map. In other words, the dissimilarity of the neighbouring units is not constant throughout the map. During the training phase, the optimal unit weights are adapted (see §3). Although this procedure can be time-consuming, the amount of memory needed is limited. Moreover, training has to be performed only once: after training, new objects can be mapped quickly to the network by determining which unit in the map is most similar. The net can also easily be updated when new data become available. In this paper, we combine the WCC function with Kohonen maps to visualize large structural databases, such as the Cambridge Structure Database (Allen, 2002).

Such a visualization may serve a variety of purposes. Firstly, the visualization itself is of scientific interest as it may show groupings in the data or relationships that might otherwise have gone unnoticed. The structure of the map reflects the structure of the database, but in a much more accessible format. We will show several illustrative examples below (see §5); other examples, using standard difference measures, can be found in the literature, *e.g.* the mapping of the IR spectra of organic compounds (Melssen *et al.*, 1993), dihedral angles of DNA dinucleotides (Beckers *et al.*, 1997) and lipids (Hyvonen *et al.*, 2001), and the complexation properties of metal ions (Pletnev & Zernov, 2002).

The second application that we mention here is the rapid screening of the similarities of new compounds. Given that a database itself may often contain hundreds of thousands of compounds, comparing a new compound with the whole database may take quite some time. Comparison with the units in a SOM is much quicker. One can then concentrate on all the compounds which are mapped to the units that are more similar than a certain cutoff. A similar example in the field of proteomics can be found in Vracko & Basak (2004).

The paper is organized as follows: first we review the background of the WCC measure. Next, we show how the self-organizing maps are created and in which way the WCC is employed in both training and classification phases. Next, the data that are used to illustrate our approach, and the software (available *via* the web), are briefly described. The results show that meaningful maps can be obtained with modest resources. Several applications are shown. The paper ends with a discussion of the possibilities for further enhancements.

2. Comparison of powder diffraction patterns

As shown in earlier publications (de Gelder *et al.*, 2001; Hageman *et al.*, 2003), a meaningful comparison of powder patterns to assess structural similarity is not possible unless the fact that peaks may shift with respect to each other is taken into account. A (dis)similarity criterion that does this is the weighted cross-correlation (WCC; de Gelder *et al.*, 2001), which basically is the area under the cross-correlation curve, weighted by the shift, and normalized so that identical patterns give a similarity value of 1. This can be written as

$$\mathrm{WCC} = \frac{f' W g}{(f' W f)^{1/2} (g' W g)^{1/2}}, \qquad (1)$$

where f and g are powder pattern profiles (column vectors, the prime symbol indicates the transpose), and W is a weight matrix (see also Stephenson & Binsch, 1980). The latter is a banded matrix containing values of one on the diagonal. In our application, values decrease linearly with the distance from the diagonal; further away than a specific threshold (the triangle width) values are zero. A WCC with a triangle width of zero, corresponding to a diagonal weight matrix, leads (for mean-centred patterns) to the well known Pearson product-moment correlation.

3. Self-organizing feature maps

A Kohonen neural network or self-organizing feature map (Kohonen, 2001) consists of a set of non-interconnected units which are spatially ordered according to some topology; typically a two-dimensional hexagonal or rectangular grid is chosen. Each unit is equipped with a weight vector, of which the number of elements is equal to the number of variables per input object (in this case, a powder diffraction pattern). Intuitively, the operation of a Kohonen neural network can be compared with the well known non-linear Mercator projection of the three-dimensional Earth onto a flat, two-dimensional, topographical map.

Before the training of the network, the elements of all weight vectors are initialized by random values in a data-set specific range. Then, all the objects of the pre-selected training set are presented to all units in the network, in random order. The unit in the map possessing the weight vector most similar to the presented object is assigned to be the winner. Subsequently, the weight vectors of this unit and its closest neighbours in the map are updated in such a way as to become more similar to the presented input object. The amount of change is governed by a parameter, the learning rate. This iterative process of weight updating is repeated until all objects belonging to the training set are presented a sufficient number of times to the network.

The size of the neighbourhood is of vital importance to guarantee that relevant features of the entire input space are embedded in the weight vectors. Initially, the size of the neighbourhood is equal to that of the size of the map itself. In this phase of network training, global characteristics of the database are captured into the weights. During training, the size of the neighbourhood is gradually decreased. This neighbourhood shrinkage forces local clusters of units to represent specific combinations of features, which are present in the data set. During the last phase, which takes most of the learning iterations, only the weight of the winning unit itself is adapted; as a consequence, such a unit becomes specialized to those objects which are frequently mapped onto it. It should be noted that initially the leaning rate is relatively high (forcing global adaptation of the units in the map). During the training process, this rate is decreased gradually to a small value (allowing individual units to diversify).

Various similarity measures can be applied to determine the winning unit. Rather than using the common Euclidean distance measure, in this paper we adopt the WCC as the similarity criterion. Obviously, the width of the WCC function interferes with the granularity of the final mapping: a very narrow WCC triangle will result in a speckled 'high-definition' but noisy map, whereas a triangle which is too wide yields a smoothing effect which probably blurs the desired specificity of the units.

After training, mapping a new object is carried out by simply calculating the similarity of the new object with all the unit vectors and assigning the object to the unit with maximal similarity. This is a very fast operation since the number of similarity calculations is equal to the number of units in the map. Usually, this is orders of magnitudes smaller than the size of the training set.

4. Experimental

4.1. Data

To illustrate the method, we used several data sets. The first is a small data set of 205 powder patterns, calculated from structures in the CSD (November 2003 release, plus January and April 2004 updates) by searching on structures similar to 12 quite different 'seed' structures. For this, the *IsoQuest* package (de Gelder & Smits, 2004, 2005) was used; the cutoff value for including structures in a class was set to a value that leads to a consistent set of related compounds. Each of these 12 seed structures led to a specific class, as shown in Fig. 1. Not all classes are equally 'tight': for instance, in class 3 (ECARAB) there is much more structural diversity than in class 12 (ELAMIN). This difference is reflected in the powder patterns.

The second data set is a random selection of 11 165 widely varying crystal structures from the CSD (up to April 2004). Only five of the patterns from the small set occur in the larger set: there is one pattern from each of the ECARAB, DOSKEB and CUXQAN classes, and two patterns are from the ALACAC01 class.

Two further data sets are selected from the CSD (November 2004 release): a set of 1262 peptides and a set of 2303 steroids. The latter are selected by performing a search on the typical arrangement of three six-membered rings and one five-membered ring, such as is visible in the structure of ECARAB in Fig. 1. It should be noted that this selection procedure yields a few structures that do not exactly conform to the definition of a steroid. The peptides are selected by searching for structures containing at least one amino acid, using the program *Conquest* (Bruno *et al.*, 2002). One structure, SIQVIX, appears in both the peptide and steroid sets. Finally, all 5789 structures in the July 2004 update of the CSD are also mapped to the trained network; of course, none of these is present in the April 2004 subset that was used for training.

All data sets include 2θ values up to 25°, with a sampling rate of 0.05°. Values below 1° are not taken into account since no features were present. A pattern therefore consists of 481 intensity values (counts). The Cu $K\alpha_1$ wavelength is used to calculate the powder diffraction patterns. These settings lead to a crystal structure description with a resolution of approximately 3.6 Å. Other choices are possible, of course. Intensity counts are scaled by taking square roots, analogous to the *IsoQuest* program (de Gelder & Smits, 2004, 2005). The largest intensity is then set to 100 units.

4.2. Similarity calculations

Comparing crystal structures on the basis of *e.g.* cell parameters is very easy and fast, but only considers periodicity and ignores the electron density distribution within the unit cell. Additionally, there is a risk of missing similarities between compounds, either because of the ambiguity in choosing the unit cell or because of symmetry-related issues. Directly comparing crystal packings, as described by powder diffraction

patterns, offers a complementary approach where these difficulties are not present.

Triangle widths of 0.5, 1.0, 1.5 and 2.0° are used to determine similarities by the WCC criterion. Triangles which are too narrow ignore the neighbourhood of features in calculating similarities; triangles which are too broad take too much of the neighbourhood into account, leading to uniformly high similarity values without much discriminatory power. Calcu-

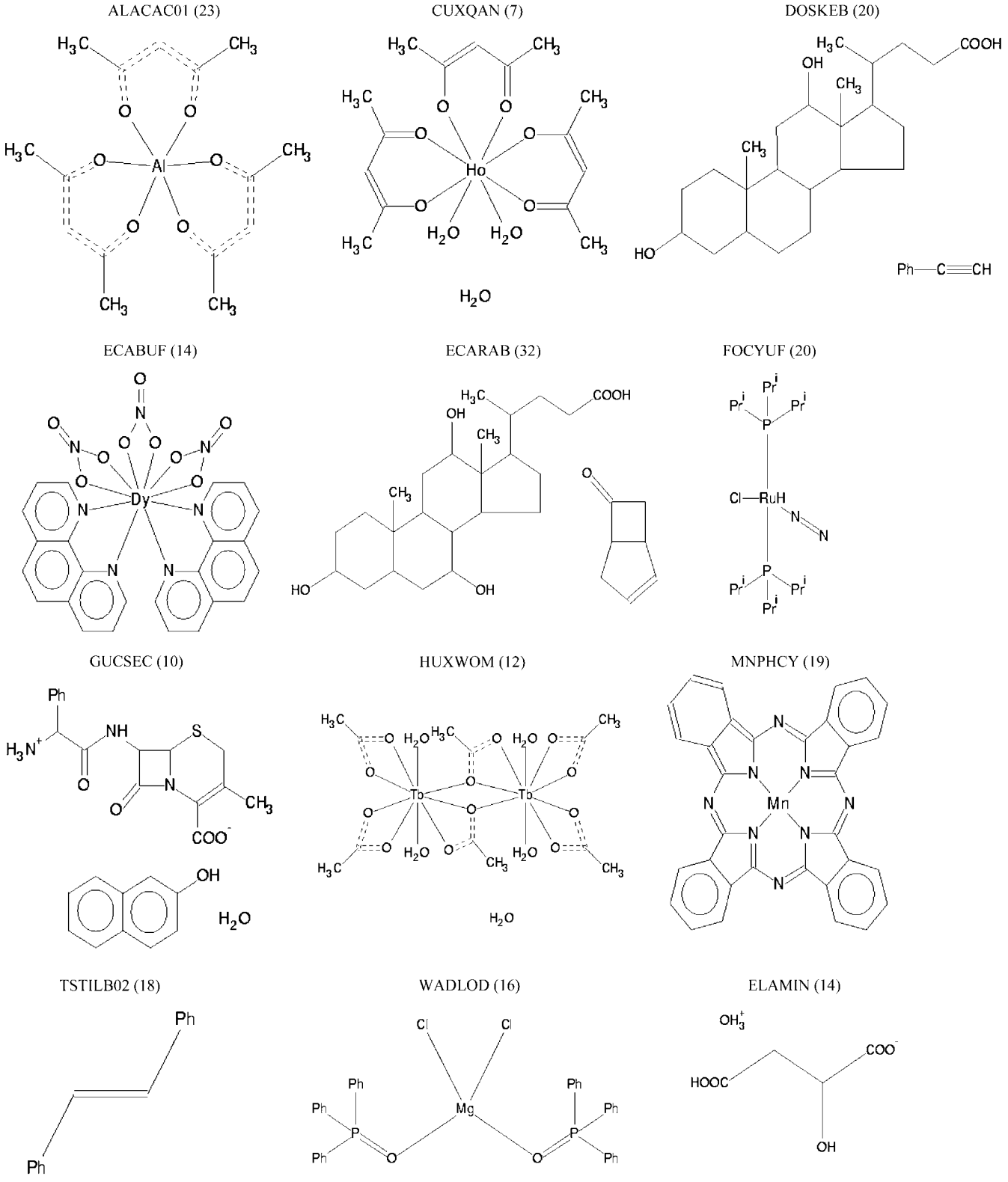

Figure 1
The 12 seed structures. The number of structures in each class is indicated in parentheses.

lation times for broad triangles are also longer in the current implementation. The best triangle is the narrowest one that still gives consistent and true similarities.

4.3. Network training

Several parameters must be set for training self-organizing maps: map size and topology, the learning parameter α and the neighbourhood. The size of the map depends on the amount of detail visualized: the more units, the more distinct features can be distinguished. For the set of 11 165 structures, which will primarily be used for visualization and to quickly find similar compounds, we use a 40×40 map. Thus, the number of units in the map is approximately 15% of the number of objects in the database.

The network topology in this paper is a hexagonal network, in which the distances to all six neighbours of a unit are equal. Moreover, all edges of the network are joined so that there are no edge units and all units have six neighbours. One can also imagine replicates of the map placed in all eight horizontal, vertical and diagonal neighbouring positions. In the plots, of course, this cannot be visualized. As a consequence, any unit in the map may be used as the central unit without changing the properties of the map. In this paper, we show the maps without this kind of transposition.

The number of training events for the maps shown in this paper is 200 times the number of patterns that are presented. An update consists of a weighted average of the unit weight and the new pattern; the weight of the new pattern is the learning parameter α. During training, α decreases linearly from 0.05 to 0.01. The neighbourhood decreases exponentially. At the start of training the whole map is part of the neighbourhood. The neighbourhood decreases in such a way that after one third of the training phase only the winning unit is updated.

4.4. Software

All procedures are implemented in *R* (Ihaka & Gentleman, 1996), with time-critical elements, such as the calculation of cross- and autocorrelations and the training of the network, in *C*. It is available as an *R* package '`wccsom`' from the web at http://www.cac.science.ru.nl/software. The core of the package is modelled after the SOM functions in the recommended *R* package '`class`', written by Brian Ripley. Many additional functions, most notably plotting features, have been added.

5. Results

As an example, we used the 205 test patterns to train a six-by-six Kohonen map. The resulting mapping is shown in the left plot of Fig. 2. In principle, the mapping depends on the (random) initial starting point, but the results shown here are representative of repeated mappings. All classes are mapped into different units. Class 3, showing the highest diversity in the patterns, is mapped to five units; bear in mind that the edges of the map are folded onto each other so that all the units containing patterns of class 3 are actually neighbours. Class 5 is mapped to three units and all other classes are mapped to only one unit. In this case, a triangle width of 1.5° is used for the WCC function; similar results were obtained for other widths.

Fig. 3 shows the patterns mapped to units 1, 7 and 18 of the six-by-six map, respectively, and the weights associated with these units. Unit 1 contains all patterns from class 4, and unit 18 contains a large portion of the class 3 patterns. Clearly, the weight vectors are very similar to the mapped patterns. Unit 7 serves as a transition between units 1 and 18 (since the edges of the map are joined, unit 7 is also a neighbour of unit 18). This is a general feature of Kohonen maps: even 'empty' units are important. The width of the features in the unit weights is directly related to the triangle width employed, in this case 1.5°. Narrower triangles lead to narrower features.

In the right plot of Fig. 2, the effect of doubling the number of units in both directions is shown. Again, class 3 is the most dispersed, although now all classes occupy more than one unit. In almost all cases, though, patterns from one class are in one contiguous cluster of units. As there is more freedom, repeated mappings may look quite different. Again, in all cases a similar class separation is achieved. This shows that good, interpretable maps may be achieved as long as the map has at least a certain minimal size.

Obviously, the data set of 11 165 patterns needs to be mapped to a larger grid than the 205 test set patterns. Moreover, since this is a random data set, there is no specific class structure and it is to be expected that all units will contain patterns. Indeed, this can be seen in Fig. 4, where the 1600 units in the 40×40 map are coloured according to the number of patterns mapped to them: all units

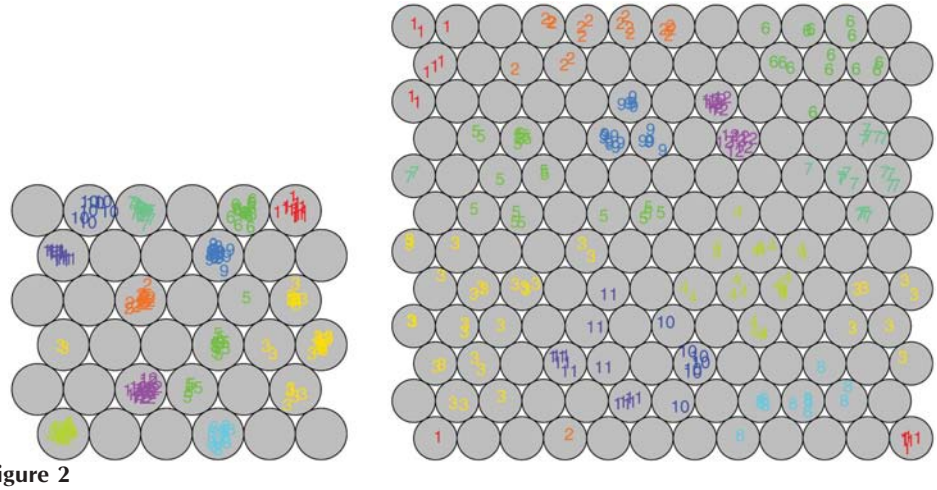

Figure 2
Mapping of the 205 patterns onto a 6×6 grid (left plot) and a 12×12 grid (right plot). Objects are indicated by their class number; for each class a different colour is used.

contain at least one pattern. The unit weights of the map show similar transitions as in Fig. 3. Interestingly, not all weight vectors show the same number of features. Some appear quite empty and seem to function mainly as buffers between more explicitly defined weight vectors.

In Fig. 5 the convergence of the network during training is visualized. The x axis shows the number of iterations, where one iteration corresponds to one presentation of the whole data set to the network. The y axis shows the median absolute deviation of the patterns with the weight vectors of the network. The vertical grey lines indicate when the network neighbourhood has shrunk to such an extent that fewer units will be marked as 'neighbours'; the rightmost grey line (after one-third of all iterations) indicates that, from that moment on, only the winning unit is updated. Clearly, at the end of the training no further changes occur.

To further assess the quality of the trained network, we project the small data set of 205 patterns on the large map, using the WCC value with a triangle of 1.5° – the same width that was used during the training of the map. Remember that only five of the patterns in the small set were also present in

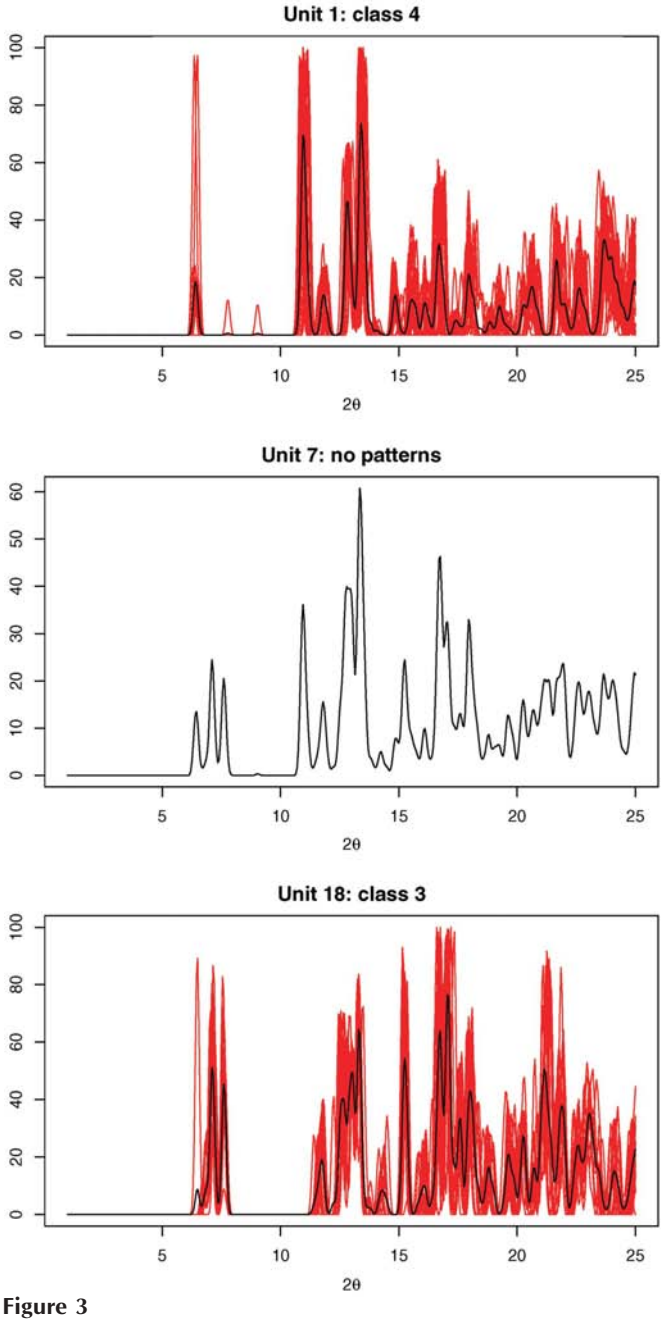

Figure 3
Weight vectors of units 1 (top plot) and 18 (bottom plot), with the patterns mapped to these units superimposed in red. Unit 1 is the unit at the bottom left, containing all patterns of class 4; unit 18 is the rightmost unit at the third row from the bottom and contains patterns from class 3.

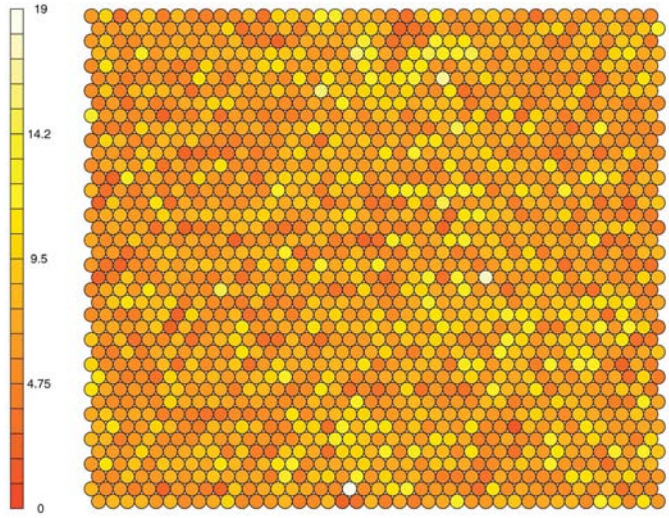

Figure 4
Number of patterns mapped on each of the units in the 40 × 40 grid: the total number is 11 165. Note that the distribution is fairly even: at most 19 patterns are mapped to a unit. The triangle width for the WCC value is 1.5°.

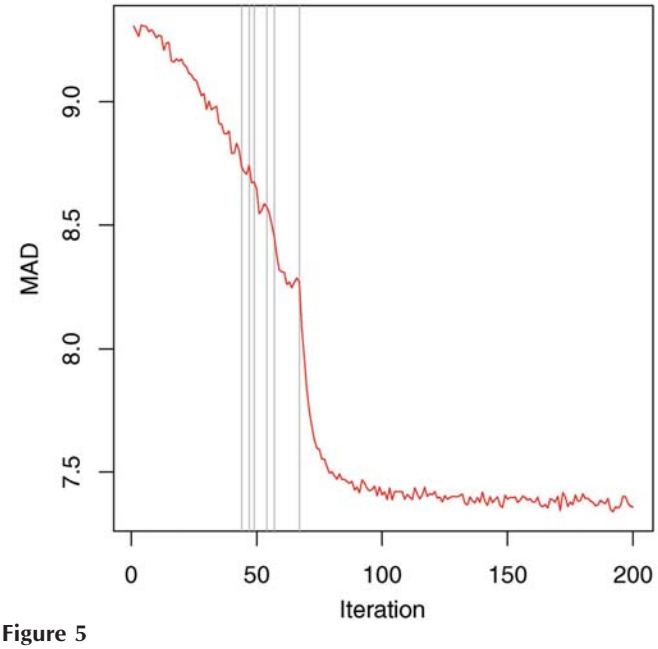

Figure 5
Progress during training: the median absolute deviation between a presented pattern and the weight vector of the winning unit. After one-third of the iterations (rightmost grey line), only the winning unit is updated.

the large set. If the network really has captured the structure of the powder patterns, the classes again should be visible in the map. This is indeed the case, as shown in Fig. 6. In most cases, there is a strong concentration of compounds of one class into one or two units, and some class members that are mapped elsewhere. Class 3 is essentially mapped to two separate units; as is class 12. The split of class 12, one of the least diverse classes, shows that it is possible that two or more units in the network end up with approximately the same unit vectors. Members of such a class will be mapped on any of these units. Some other classes, such as 1 and 11, are more spread out.

As an illustration, Fig. 7 shows the similarity with unit-code vectors for the 12 'seed' compounds. Each compound is mapped to the unit with the highest similarity (indicated with the blue crosshairs). Obviously, the similarity pattern is markedly different for different classes, but patterns from one class in general show very similar images.

When mapping new compounds to a trained map, a logical objective is to quickly find similar compounds. One could concentrate on the units that show the highest similarity to the new compound, and investigate all patterns that are mapped to these units. We can use our small test set of 205 compounds to investigate the influence of the triangle width for such a purpose. Suppose we map each of the 'seed' patterns, as in Fig. 7, one could ask: how many units should be selected to include all members of the respective classes? The results are given in Table 1 for four different triangle widths. In many cases, only a few units have to be considered to find all members of a class; in some cases, this number seems relatively high. However, this is usually because of one or a few patterns that are less similar to the seed pattern. As an example, for a triangle width of 1.5°, all in all, 11 of the 205 structures are not within the best 20 units, and 22 of the 205 structures are not in the top five. There is obviously a trade-off between speed and completeness. Overall, triangle widths of 1.0–1.5° are optimal, in agreement with earlier results (de Gelder *et al.*, 2001).

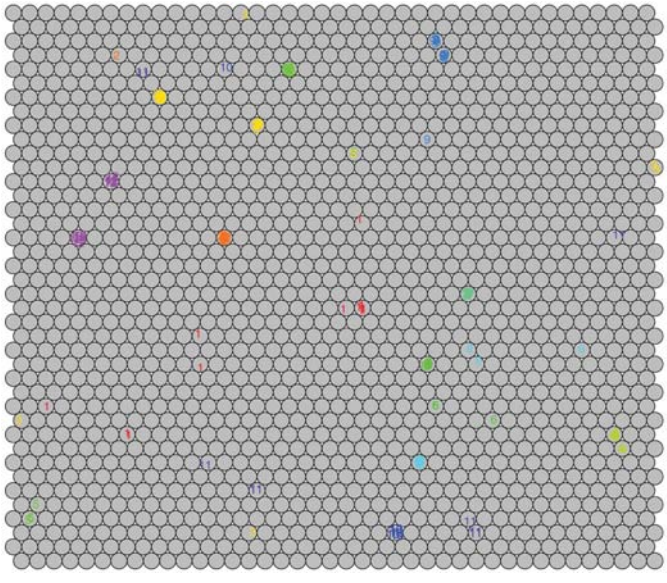

Figure 6
Projection of the test set of 205 compounds into the map trained with 11 165 compounds. Again, the classes are clearly localized.

Table 1
The number of units used to consider, in order to capture all the class members, upon projection of the 12 'seed' compounds to the trained map.

The optimal width is somewhere around 10–15°.

Class	Seed compound	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$
1	WADLOD	257	205	73	14
2	TSTILB02	136	129	138	106
3	ECARAB	511	577	217	371
4	ALACAC01	8	22	11	120
5	DOSKEB	55	9	107	206
6	ECABUF	22	6	3	5
7	FOQYUF	10	2	1	8
8	HUXWOM	35	11	13	4
9	MNPHCY	307	71	4	209
10	CUXQAN	108	18	5	12
11	GUCSEC	180	15	196	210
12	ELAMIN	1	1	2	1
	average	135.8	88.8	64.2	105.5

6. Applications

Several applications of Kohonen maps, as detailed here, can be envisaged (see also Zupan & Gasteiger, 1999). First, the map can be used to visualize the occupancy of structural space for various classes of compounds. Mapping both the peptide and the steroid data sets to the trained network leads to the images shown in Fig. 8. The 2303 steroids are mapped to 654 units; unit 1097 has most hits, with 32 steroids mapped to it. The more diverse set of 1262 peptides is mapped to 638 units, with no unit containing more than 10 peptide structures. There are 297 units containing representatives of both classes. Clearly, the two data sets show a different mapping. This is remarkable and unexpected, since it is known that small changes in chemical structure in general may lead to major changes in crystal packing. In contrast, the July 2004 update of the CSD covers almost the whole map (1472 units).

Furthermore, one may use the map to identify compounds with similar crystal packings. Not only is it no longer necessary to perform a pair-wise comparison between all objects in the database, the map also presents an appealing visualization of groupings in the data and as such is more informative than lists of compounds with a high similarity. Many publications have appeared that address the issue of isostructurality. In one of the papers of Kálmán *et al.* (1993), a small set of steroid structures is described that show isostructural relationships: digitoxigenin (DIGTOX), (21*S*)-methyldigitoxigenin (JIDNOZ), (21*R*)-methyldigitoxigenin (BOKTIE), digirezigenin (CUXYAV) and 3-epidigitoxigenin (DHCENO). If we take digitoxigenin as a search compound and map its corresponding pattern in the trained network, we can select the five units with the highest similarity and analyse their contents. This leads to a set of 59 other steroids. These include the other compounds mentioned by Kálmán *et al.* (1993) showing that this is a quick and elegant

way to search for isostructural compounds. Obviously, this also holds for new compounds. One can of course investigate all units that show a similarity above a certain threshold (*e.g.* 0.9) rather than of a predefined number of the most similar units.

This procedure is much quicker than a comparison with all objects in the database. Not only is the map much smaller than the database, but with only a very low number of matching units one captures a high percentage of similar compounds. Mapping DIGTOX, selecting the five best units and with that the 59 most similar steroid patterns, and sorting the results, takes just over half a second on not very modern hardware (Athlon 1800+ MHz).

A particular advantage of using powder patterns as a representation of crystal structure is that experimental patterns may also be used, *i.e.* patterns for which cell parameters and space group are not known. The map may be used to infer what packing patterns are most likely for a new compound. Suppose a new experimental steroid powder pattern is mapped to a unit which contains several other steroids. Then one could expect that the new compound would share some global characteristics with them. In principle, this information could reduce the number of possible unit cells and space groups and makes structure solution easier. This needs further research.

Furthermore, the map may serve as a tool for stratified sampling. The goal is then to identify a small set of 'representative' or 'archetypical' compounds. This is often used in classification or regression applications to divide a data set into training and test sets (*e.g.* Guha *et al.*, 2004). In the current setting of mapping databases of structures, it is possible to extract a small set of structures that is representative for the whole database. Or, put differently, if we would take one random structure from each of the 1600 units in the map, we would cover the same chemical space as the 11 165 structures that were used to train the map. Random sampling would be less efficient, because small, specific groups can easily be missed. In polymorph prediction, this feature can be used to reduce the number of structures before energy minimization

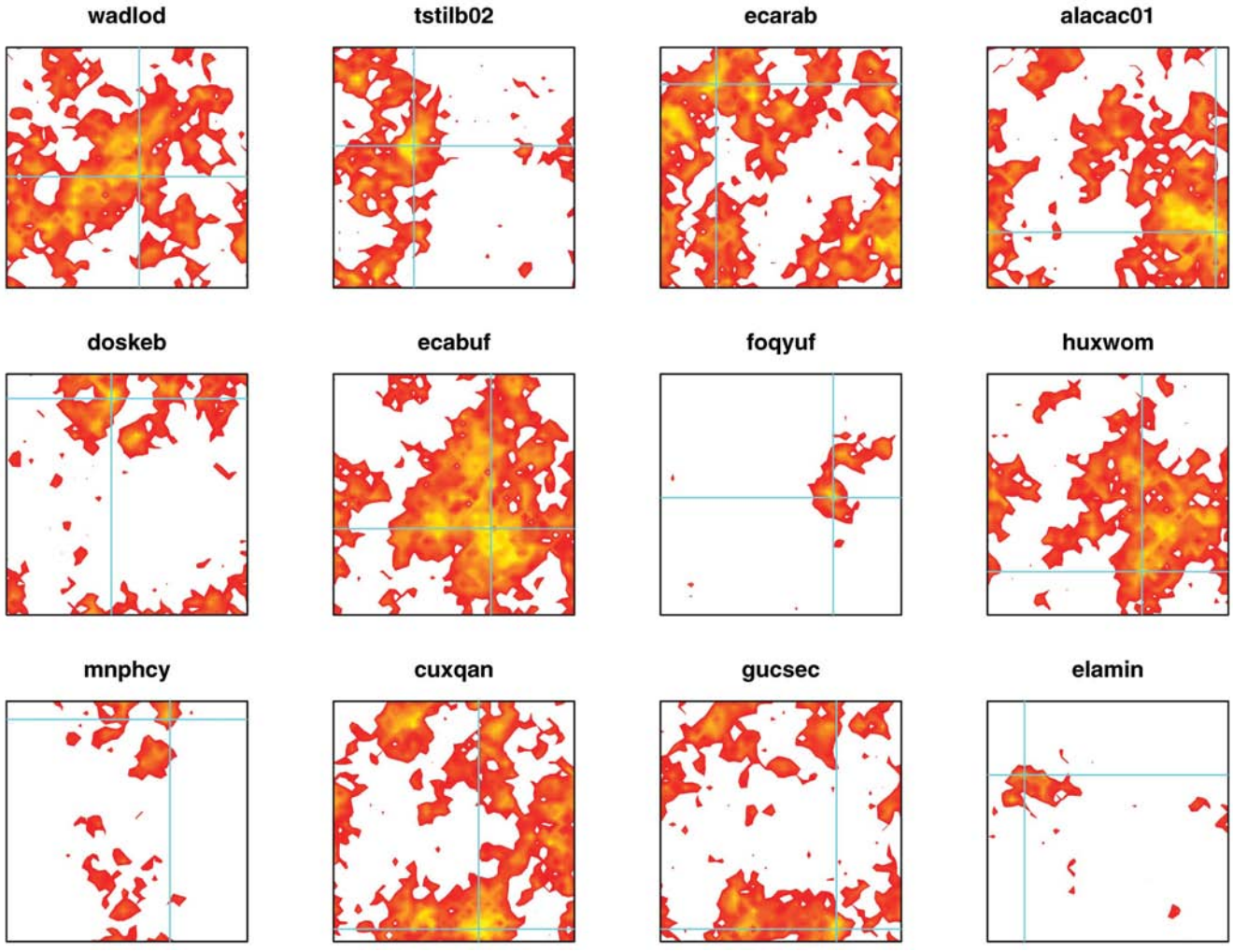

Figure 7
Smoothed representation of similarities between the 12 calculated powder diffraction patterns of the 'seed' compounds and the trained map (the triangle is 1.5°). The compound is projected onto the point indicated by the crosslines. The colour scale runs from 0.75 (red) to 1.0 (white); a grey colour indicates similarities lower than 0.75.

without losing potentially interesting candidates. Also the final ensemble of polymorphs can be assessed in this way.

7. Discussion and future work

Mapping databases of crystal structures to a two-dimensional grid of neural network units has obvious advantages: it provides, at a glance, a visualization of the similarities of all the structures and may show groupings that are not easily found otherwise. We have argued that the combination of similarities of powder patterns, as measured by the WCC, and Kohonen maps is a useful tool for this mapping. Of course, other choices can be made: it is very well possible to construct a Kohonen map that shows similarities between *e.g.* cell parameters. We think that focusing on electron distributions at low resolution has several advantages; moreover, it provides the opportunity to map experimental patterns which may lead to an easier structure elucidation.

The mapping of the 11 165 structures onto 1600 units shows that the proposed approach is feasible. The map can be interpreted and mapping new compounds is straightforward. Compounds from different classes are mapped to specific areas in the network. This is not only the case for the small set of 12 test classes, but also for the less well defined peptide and steroid sets. The mapping of *e.g.* the steroid class shows that related structures are not randomly scattered in the map. One therefore can speak of a 'steroid' region. Of course, this region overlaps with regions occupied by other chemical families. Obviously, we have used only a small subset of the CSD in training the Kohonen map. Nevertheless, the mapping of the test set of 205 structures, and the peptide and steroid classes shows that there is enough diversity in the map to accomodate differences in powder patterns.

One advantage of Kohonen maps is that they can be updated fairly easily. One can see a situation where the complete CSD is used for training and where CSD updates are used to re-train the map. This re-training should strike a balance between representing the information already in the map and describing the new structures. The size of such a map would probably be somewhat larger than the map we have presented here.

Finally, one can enhance the Kohonen maps as detailed here for prediction purposes (Kohonen, 2001) in several ways. These basically consist of a Kohonen map where the units are arranged in such a way that, in addition to the topological structure of the powder diffraction data, the properties of interest, such as a crystal system or unit-cell volume, are used to colour specific parts of the map. A projection in a certain part of the map may then be directly translated to a value for the property of interest. The essential part is that *e.g.* the unit-cell volume, during training, partly determines what units are updated and how they are updated. Therefore, two groups with similar powder patterns but different cell volumes will be better separated better than in the unsupervised Kohonen maps, as seen in the current paper. We will report on these extensions in the near future.

In conclusion, the visualization of large powder pattern databases such as those derived from the CSD is made possible by the application of self-organizing networks utilizing the WCC similarity measure. This opens up many new

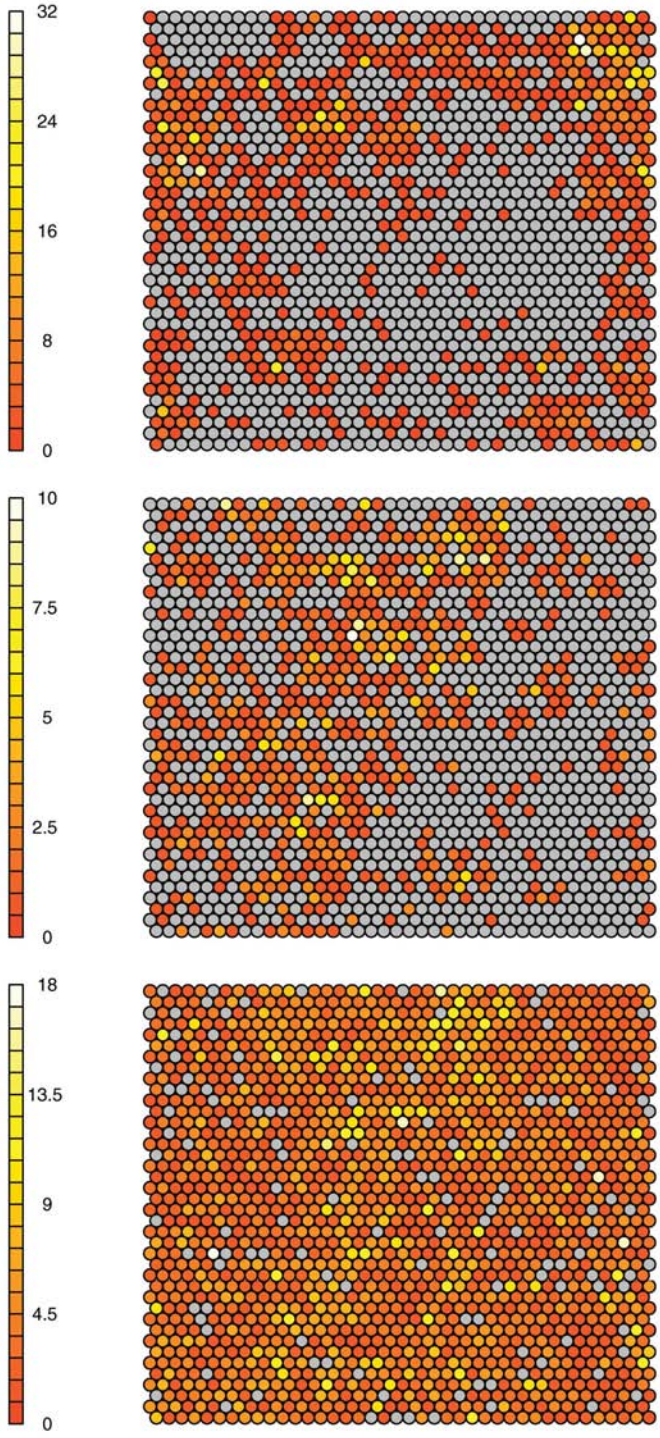

Figure 8
Mapping of steroids (top), peptides (middle) and July 2004 update (bottom) on the trained map. The colour indicates the number of compounds mapped to that unit; no compounds are mapped to units depicted in grey. Whereas the peptide and steroid mappings concentrate in specific regions in the map, the July 2004 update is more or less uniformly spread out over the whole map, like the original training set.

possibilities to explore and utilize the full potential of a combination of many different types of compound.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Beckers, M. L. M., Buydens, L., Pikkemaat, J. & Altona, C. (1997). *J. Biomol. NMR*, **9**, 25–34.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Gelder, R. de & Smits, J. M. M. (2004). *Acta Cryst.* A**60**, s78.

Gelder, R. de & Smits, J. M. M. (2005). Submitted for publication.

Gelder, R. de, Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**, 273–289.

Guha, R., Serra, J. R. & Jurs, P. C. (2004). *J. Mol. Graphics Mod.* **23**, 1–14.

Hageman, J. A., Wehrens, R., de Gelder, R. & Buydens, L. M. C. (2003). *J. Comput. Chem.* **24**, 1043–1051.

Hageman, J. A., Wehrens, R., de Gelder, R., Meerts, W. L. & Buydens, L. M. C. (2000). *J. Phys. Chem.* **113**, 7955–7962.

Hyvonen, M. T., Hiltunen, Y., El-Deredy, W., Ojala, T., Vaara, J., Kovanen, P. T. & Ala-Korpela, M. (2001). *J. Am. Chem. Soc.* **123**, 810–816.

Ihaka, R. & Gentleman, R. (1996). *J. Comput. Graphic. Statist.* **5**, 299–314.

Jackson, J. E. (1991). *A User's Guide to Principal Components.* New York: Wiley.

Kálmán, A., Párkányi, L. & Argay, G. (1993). *Acta Cryst.* B**49**, 1039–1049.

Kohonen, T. (1982). *Biol. Cyb.* **43**, 59–69.

Kohonen, T. (2001). *Self-Organizing Maps*, No. 30 in Springer Series in Information Sciences, 3rd ed. Berlin: Springer.

Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate Analysis.* New York: Academic Press.

Meerts, W. L., Schmitt, M. & Groenenboom, G. C. (2004). *Can. J. Chem.* **82**, 804–819.

Melssen, W. J., Smits, J. R. M., Rolf, G. H. & Kateman, G. (1993). *Chemom. Intell. Lab. Syst.* **18**, 195–204.

Pletnev, I. V. & Zernov, V. V. (2002). *Anal. Chim. Acta*, **455**, 131–142.

Stephenson, D. S. & Binsch, G. (1980). *J. Magn. Reson.* **37**, 409–430.

Vracko, M. & Basak, S. C. (2004). *Chemom. Intell. Lab. Syst.* **70**, 33–38.

Willighagen, E. L., Wehrens, R., Verwer, P., de Gelder, R. & Buydens, L. M. C. (2005). *Acta Cryst.* B**61**, 29–36.

Zupan, J. & Gasteiger, J. (1999). *Neural Networks in Chemistry and Drug Design: An Introduction*, 2nd ed. New York: Wiley.