Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 2 May 2001
Accepted 27 November 2001

Analysis of polynuclear aromatic hydrocarbons *via* the Voronoi tessellation approach: classification of atom types using artificial neural networks

Stefan W. Christensen

Materials Research Group, School of Engineering Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, England. Correspondence e-mail: swc@soton.ac.uk

When identifying the correct atom types to occupy the specific atomic locations within newly observed structures or when assessing the plausibility of new suggested structures with specific locations for specific types of atoms, any information quantifying geometrically the local environments around those locations is valuable, provided known characteristic differences exist, with respect to this geometric information, between the different atom types. A powerful tool for quantifying such geometries is the Voronoi tessellation; this has been used in a pilot study of polynuclear aromatic hydrocarbons. It has been found that perfect identification of all C and H atoms may be achieved through the examination of polyhedral volumes and surface areas. The use of a weighted face-area average is also found to be a useful measure of local structure. Simple neural network models that may be used for atom-type prediction are given in the paper. It is expected that the present approach will be useful in distinguishing between atoms that have close scattering curves whilst displaying similar crystallographic behaviour.

1. Introduction

With the aid of X-ray, neutron and electron diffraction, it is possible to analyse a crystalline structure and deduce, limited by the resolution of the experiment, where the individual atoms are situated relative to the unit cell, the inherent dimensions and angles of which may also be determined. Each atomic position does not, however, have its own beacon openly signalling its type to the experimenter and the diffraction experiment therefore begs the question: 'Which atom type goes where?'.

Of a somewhat similar nature is another question that constantly appears in the modelling of new, as yet unrealized, structures: 'Given a new suggested structure, with given atom types in specific positions relative to a specific unit cell, could it physically exist as a stable state?'.

A very obvious way, for both questions, to start seeking an answer is to try to quantify the local environments in which the atom types involved are known to occur and subsequently compare the local environments found in, or proposed for, the new structure with those known to be typical of the various atom types.

If a particular local environment in a new structure is significantly akin to one known to apply to a certain atom type, while being significantly unlike those known for all other atom types, then it may be inferred that the corresponding atom in the new structure is of that same type. Moreover, if the new local environment is unlike all those seen before then, if the new structure is real, it may be concluded that an entirely new situation has been found[1] whereas, if the new structure has been suggested by atomistic modelling, it may, at least initially, be assumed that the suggested structure cannot exist as a stable state in the physical world.

This procedure, then, involves two distinct problems:

(i) how to quantify an atom's local environment;

(ii) how to compare a newly observed local environment with those already known.

1.1. Quantifying the local environment around an atom

The immediate suggestion arising in the mind of any person with a knowledge of chemistry will probably be to make use of bond lengths and bond angles. Atoms believed to be bonded are invariably in each other's proximity and maintain a distance between them that is strongly influenced by their respective atom types and valences. Hence, bond lengths represent a good way of quantifying an atom's surroundings. Clearly, however, these interatomic distances cannot exhaustively characterize the spatial distribution of other atoms surrounding a central one; partly because not all atoms in the central atom's neighbourhood will be bonded to it, and so will not be constrained by a bond length, partly because bonded atoms, whose distance from the central atom is constrained, still have two degrees of freedom left to position themselves

[1] Or that an error has occurred, either in the diffraction experiment or in the subsequent analysis.

around the central atom. The existence, then, of bond angles, which measure the angle between neighbouring bonds, and which also show a strong regularity for specific pairs of bond types (as given by the atom types and valences involved), certainly aids in the quantification of local environments but, again, applies only to atoms bonded in a chemical sense.

An alternative much more comprehensive and certainly more objective[2] way of characterizing the local geometry in which an atom is embedded is to first calculate the Voronoi polyhedron of that atom and then calculate the various features of the polyhedron.

1.2. The Voronoi tessellation

The Voronoi tessellation is a construction defined for a group of generating points existing in a distinct space: each generating point is enclosed in exactly one polyhedron and these polyhedra taken together completely fill the space, without overlapping, and without leaving any gaps between them. Moreover, every location within the distinct space that is closer to one of the generating points than to any of the others will be enclosed in that point's Voronoi polyhedron. Locations equidistant from two generating points, and nearer to those than to any others, will lie in the Voronoi face that separates the polyhedra of those two points (which, therefore, are Voronoi neighbours). Locations equidistant from three generating points, and nearer those than any others, lie on the edge where the polyhedra of these three meet. Locations equidistant from four, or more, generating points, and nearer those than any others, are Voronoi vertices – the corners of the polyhedra of the generating points involved.

The Voronoi tessellation is thus a purely mathematical concept but, given that the equilibrium positions of atomic nuclei in solids are determined, to a good accuracy, as points, it is not unreasonable to apply this tessellation to crystal structures. If this is in fact performed, a specific interpretation of the geometric structures of the tessellation suggests itself: the polyhedra may be considered to represent the atoms, the faces represent the interactions between neighbouring atoms, the edges represent interactions between three mutually adjacent atoms with the vertices, finally, representing either the interactions between all atoms meeting there or, alternatively, the voids between the atoms.

If accepting this interpretation, a number of ways exist of quantifying the local environment in which an atom is embedded. Representative measures of atomic size would naturally be polyhedral volume and surface area, as well as the sum of the lengths of all edges on the polyhedron (giving three-, two- and one-dimensional measures of atomic size). Interactions between two neighbours could, for example, be represented *via* interatomic distance (nucleus to nucleus),[3] area of face representing interaction, solid angle under which face is seen from either of the atoms' nucleus, volume subtended at the nuclei under the face, or length of circumference of face. Interactions between three mutually adjacent atoms could, for example, be quantified *via* length of the common edge, angle under which this edge is seen from either of the nuclei, or area or length of circumference of triangle given by the edge and any one of the nuclei. Voids could be quantified *via* distances to surrounding nuclei or distances to neighbouring vertices (each vertex being the neighbour of another if they are joined *via* an edge).

Evidently, there are numerous alternative ways of quantifying specific physical/chemical concepts and, ideally, they would all be considered.[4] This may not always be feasible, if ever, but clearly the more unique views of a specific feature are available, the more truthfully can it be characterized and understood.

In summary, the Voronoi tessellation may be used to quantify the arrangement of nucleonic equilibrium positions by virtue of their point-like nature and, on accepting the representational scheme outlined, be used to quantify the local environments around individual atoms in terms of atomic interactions. It should be borne in mind, however, that there will be grievances as to the validity of this scheme if approaching it from a traditional chemical point of view owing to its overt neglect of any information concerning atomic radii. This was addressed in an earlier paper (Christensen & Thomas, 1999). For the task at hand, learning the characteristics of local environments in crystals, we need not be concerned, however, as long as we approach it consistently; we do not *need* to interpret the Voronoi geometric structures in any particular way, so far as we are considering arrangements of nucleonic equilibrium positions. On the other hand, we *may* readily adopt any one, in order to guide our understanding of our calculations, if we so desire.

1.3. Comparing local environments

Once a set of variables quantifying local geometry has been decided upon, the problem becomes one of comparison between these geometric characteristics of the atoms in the new structure and those of atoms in crystal structures observed and analysed earlier. This is by no means a trivial task. First, to get the most complete understanding of a particular atom type, it is necessary to take all known instances of this into consideration; a daunting if not insurmountable project, with the overall number of observed structurally unique atoms running into the tens of millions. However, a great deal of similarity must be expected to exist between structurally unique atoms of the same type, and so it may be hoped that, taking only a smaller subset of all known structurally distinct instances of an atom type into consideration, it may still be possible to get a good understanding of that atom type's local geometric characteristics. A reasonable approach to selecting such a subset may be to include crystal structures of a similar nature to that of the new structure in point; *i.e.* if

[2] The bond model of molecules and solids is, as is any model, necessarily partly subjective.

[3] *Cf.* bond length, but note that this would also apply for non-bonded interactions.

[4] As long as they all contribute unique information; if a specific variable can be derived from a combination of the others, it is logically redundant and should be omitted.

the new structure is inorganic, include only inorganic compounds, if a protein, include only proteins *etc.*

With the subsets of atoms from the literature selected, and their local geometric features quantified, the actual comparisons may proceed. This may be accomplished on an instance-by-instance basis; if the issue is to ascertain the type of a specific atom in a new structure, this may be compared, feature for feature, against all the atoms amongst the established structures, with the type eventually given by the one known atom whose features most closely resemble those of the new one. This nearest-neighbour approach may be prohibitively slow, however, and a more prudent way to address the problem may be to summarize the information contained in the data of the atoms in the known structures *via* data-driven modelling, establishing a model of local geometric characteristics for each atom type and, subsequently, compare the new atoms with the models.

2. Data-driven modelling

To extract knowledge about the nature of the properties of an underlying data-generating system, given a set of data, and, in turn, utilize this knowledge to predict the properties, given other data, is the problem addressed by modelling. Arguably, this is the most basic of tasks since it, in essence, is the very process known as learning. Not surprisingly, it has been of the highest concern to all areas of science and, equally unsurprisingly, it has been approached in very many different ways – more often than not in complete or near-complete ignorance of the understanding of modelling previously obtained in other areas of science and mathematics.

Three fundamental questions must be answered before a mathematical model may be derived:

(i) which kinds of mathematical functions may be employed?

(ii) how is the quality of a model assessed?

(iii) which guiding principle should be used to search for the best model among those that may be considered [*cf.* (i) above]?

The various techniques, and all their subvariants, differ, often profoundly, in their answers to these questions, reflecting the fact that they have arisen in very disparate scientific communities with widely different fundamental assumptions for the modelling process.[5] No method has established itself as the indisputably best in every regard, and all methods have their specific strengths, weaknesses, protagonists and antagonists. Moreover, as mathematical modelling is very much still an active research area, all methods are continually improved while new approaches are introduced at irregular intervals.

For the present project, which represents a first attempt at quantifying structural characteristics of crystals *via* adaptive numeric modelling, it was decided to employ one of the most basic techniques available: the feed-forward neural network, also known as the multilayer perceptron, MLP.

[5] Notable disciplines where modelling plays the focal role include frequentist statistics, Bayesian statistics, data mining and machine learning.

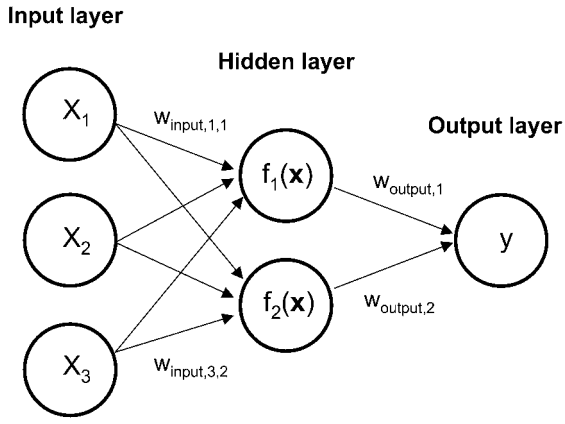

Figure 1
Schematic illustration of the structure of a MLP with one hidden layer.

Fig. 1 shows the schematic of a MLP; it is a complex mathematical function, realized through a network of interconnected artificial neurons, most commonly distributed in three layers: input, hidden and output, with each connection between two neurons given a unique weight. Mathematically, the function is given as

$$y = \sum_{i=1}^{m} w_{\mathrm{output},i} f_i(\mathbf{x}), \tag{1}$$

where y is the output, m is the number of hidden layer neurons, $\mathbf{x}$ is the input vector,[6] $w_{\mathrm{output},i}$ are the output weights (from hidden to output layer) and

$$f_i(\mathbf{x}) = a_i\left(\sum_{j=0}^{n} w_{\mathrm{input},j,i}\, x_j\right) \tag{2}$$

are subfunctions, performed at the hidden neurons, where n is the number of inputs, $w_{\mathrm{input},j,i}$ are the weights (from input to hidden layer), x_j are the individual inputs (x_0 is a constant set permanently to 1) and a, the so-called activation function, is typically, although by no means necessarily, given by a sigmoid function, *e.g.* the tanh function.[7]

In other words, the MLP is a large equation of $\mathbf{x}$, with numerous additive terms, each of which includes a coefficient (termed weight in the machine-learning community).

The MLP is therefore capable of calculating arbitrary values, based on the values of the input vectors and the weights; the trick is to tune the weights so that the network always predicts the right output, given a specific input. This is achieved during the training phase: the network weights are given random initial values, the MLP is presented with a set of observed (input, output) pairs and gives, based on the input values and the weights, a prediction for each training point. These initial predictions are very likely very different from the observed output values, and the training therefore proceeds

[6] The input vector is simply a vector containing the variables, based on which the output is sought calculated/predicted. In our case, for example, a relevant input vector could be (Voronoi volume of atom, Voronoi surface area of atom)^T. The output, then, is that property which is sought predicted.

[7] Crucially, this activation function must be non-linear to allow the MLP to adapt to non-linear system behaviour.

with successive alterations of the weights and subsequent re-evaluation of the prediction accuracies. This continues until the predictions are adequately accurate.[8]

In the present case, where the objective is prediction of atom type, we are concerned with classification, rather than regression, and the outputs are therefore not real numbers; we have two possible outcomes: H or C, and we want the MLP to predict one or the other. This is achieved by assigning a single specific numeric output value to all data points of type C and a different single output value to all data points of type H while adopting the convention that, if the predicted output value is closer to the former, the prediction is C, while it is H if closer to the latter. The specific values used here are governed by the fact that the tanh function converges to ± 1 as its argument approaches $\pm\infty$ and useful values therefore must lie entirely within that range (chosen here were 0.9 for C and 0.1 for H, with the discriminating border lying at 0.5).

The optimization of the MLP weights is a highly complex problem, in that many different solutions may exist that are all fairly good though not optimal. In reality, a search for the perfect set of weights is futile and one must make do with a suboptimal one, one that may have relative strengths in certain ways and relative weaknesses in others. If, now, the optimization process is repeated numerous times, with a different solution found every time (a highly realistic situation), it may well be that some models are strong where the others are weak and so by combining the models into a committee it may be hoped that all inherent model weaknesses are offset by equally inherent strengths, and, as long as the individual models are at least fairly good on their own, the committee will be better, in terms of prediction accuracy, than the individual models are on average. This has been established in the modelling communities, both experimentally and on theoretical grounds (Krogh & Vedelsby, 1995). For this reason, it was decided for the present work to generate several individual models and combine them into a committee.

3. Data and procedure

Two individual problems pertaining to crystallography were stated in the *Introduction*: first, obtaining the ability to ascertain the atom type for any atom, based only on knowledge of its coordinates and those of the surrounding atoms and, secondly, obtaining the ability to predict, given only proposed atomic coordinates, whether these are physically feasible for atoms of a proposed type. It is the first of these that has been addressed in the present work, while the second is the subject of ongoing research by the author.

[8] This is somewhat simplified in that also the network complexity will be subject to alteration, primarily through the number of hidden layer neurons; the more of these, the more complex a functionality the MLP may accommodate. If, however, the MLP is given too much flexibility, it is highly prone to introducing quite wild fluctuations, while being 'right on target' when the training data are concerned. In this case, the MLP is said to be overfitting the training data; these have been learned extremely well, at the cost of poor generalization ability; the wild fluctuations mean the MLP function deviates strongly from the physical system in between the training points.

Table 1
Normalization constants.

Input			
Name	Symbol	$x_{\min}$	$x_{\max}$
Surface area	A_t	22.14 Å^2	38.69 Å^2
Volume	V	5.90 Å^3	18.13 Å^3
Face area average	$\langle A \rangle$	2.70 Å^2	7.41 Å^2
Face area standard deviation	σ	1.66 Å^2	3.25 Å^2

The raw data, on which the current modelling problem is based, are the accurate locations of all atoms in a range of polynuclear aromatic hydrocarbons (PAH), in crystalline form, along with information on all unit-cell dimensions and angles. The specific PAHs studied are listed in Appendix *A*; they were obtained from the Chemical Database Service at Daresbury (Fletcher *et al.*, 1996). The reason for this particular choice is entirely pedagogical; clearly, distinguishing between C and H in aromatic hydrocarbons is not an unresolved problem, difficulties only arise when attempting to distinguish between atoms with similar scattering curves when they display similar crystallographic behaviour, *e.g.* between N and O, O and F, Al and Si. The choice of C and H will, however, serve well to illustrate the approach since the procedure is the same in any case, and whilst the resulting models discriminating between C and H may be of low dimensionality (C and H being really very different) and thus possibly rendered graphically, models for the more challenging cases are entirely unlikely to be equally good-natured, and therefore less satisfactory to demonstrate the approach.

Each of the PAHs was subjected to a Voronoi tessellation; each atom enveloped in its own Voronoi polyhedron and various characteristics of these polyhedra calculated. Four variables quantifying the local environment around the atoms were chosen for modelling: (i) volume and (ii) total surface area of the Voronoi polyhedron, (iii) a weighted average of the areas of the faces on the polyhedron, and (iv) the similarly weighted standard deviation from this weighted average. Polyhedral volume and overall surface area have been studied many times in various settings in the past, see *e.g.* Mackay (1972), Richards (1977), Koch & Fischer (1980), Blatov *et al.* (1995), Andersson & Hovmöller (1998); the weighted averages need some clarification, however. The typical Voronoi polyhedron obtained in crystal structures has many faces, most of which are very small. If a straight average of all face areas were to be calculated, it would be dominated by these small faces. Paradoxically, it is the largest faces that should be expected to correspond to the most structurally significant interactions, while the small faces should be expected to be of little consequence. Subsequently, a better measure of average face area should be obtainable by weighting the contribution of each face in accordance with its expected importance. One way of doing this is by weighting the contribution of a face with its fraction of the entire polyhedral surface area:

$$\langle A\rangle = \sum_{i=1}^{n} (A_i/A_t)\, A_i, \tag{3}$$

where $\langle A\rangle$ is the weighted face area average, n the number of faces, A_i is the area of face i and A_t is the total surface area of the polyhedron.[9] Similarly, a weighted standard deviation from this average can be defined as:

$$\sigma = \left[\sum_{i=1}^{n} (A_i/A_t)(A_i - \langle A\rangle)^2\right]^{1/2}. \tag{4}$$

These measures were adopted for the subsequent modelling. The values were all normalized to lie between 0 and 1:

$$x_{\text{normalized}} = \frac{x_{\text{non-normalized}} - x_{\min}}{x_{\max} - x_{\min}}, \tag{5}$$

where $x_{\min}$ and $x_{\max}$ are the smallest and largest values found in the data set; *i.e.* the largest volume in the normalized data is 1 while the smallest is 0. Likewise for the other inputs. The normalization constants are given in Table 1.

Two distinct modelling line-ups were chosen; one in which all four properties were included and one where only polyhedral volume and surface area were employed. The reason for conducting the smaller modelling problem, which is of course entirely included in the larger problem, is mostly pedagogical; models that try to describe a system of four different independent variables may be four-dimensional, models of systems with only two independent variables may never be more than two-dimensional and therefore readily be illustrated graphically. Moreover, a two-dimensional modelling problem is much smaller than a four-dimensional one[10] and, with a given amount of effort, a better solution[11] may be expected for the smaller problem. The four-dimensional problem was considered as well, just in case perfect discrimination could not be achieved with the two-dimensional model.

The number of data points is limited by the numbers of symmetrically inequivalent atoms in the structures studied and was, for this study, 1568 atoms, 960 of which were C and the remaining 608 were H.

These data points are shown in six bivariate plots (Figs. 2–7) for the six possible bivariate views (V–A_t; V–$\langle A\rangle$; $\langle A\rangle$–σ; V–σ; A_t–$\langle A\rangle$; A_t–σ); C given by solid markers, H by open circles. It is readily apparent that the C and H atoms tend to group separately, and for that reason it was expected that a MLP with only a single neuron in the hidden layer would be able to separate the two atom types well, while being in no danger of overfitting the data. Accordingly, that MLP structure was chosen. As a consequence hereof, there would be only one output weight for each model, hence this weight was set to 1 and not allowed to vary (effectively removing it from the model).

[9] A different approach is to disregard all faces smaller than a certain set fraction of the total surface area [suggested in Fischer & Koch (1979), though not used there in conjunction with average face areas]; this, however, gives the somewhat arbitrary cut-off value great influence on the result, and the smoothly varying influence of faces with face area, given here, was preferred.
[10] This is a result of what, in modelling terminology, generally is referred to as the 'curse of dimensionality'; the size of the modelling space grows exponentially with increased dimensionality.
[11] Relative to the theoretically best possible for the particular problem.

For the purpose of obtaining a realistic estimate of the model's ability to correctly classify atoms from new, hitherto unseen, systems, three systems were excluded from the training process and used only for testing: tribenzopyrene, tetrabenzoperylene and anthrabenzonaphthopentacene. These three systems were among the most complex with large numbers of atoms in the asymmetric unit. Hence, the numbers of atoms available for training were 822 C and 538 H, the test set comprising 138 C and 70 H atoms.

In all, five individual models were trained on the training data, and a committee known as a basic ensemble (Perrone & Cooper, 1993) of these was formed. The predictive capability of this ensemble, which is simply an equally weighted linear

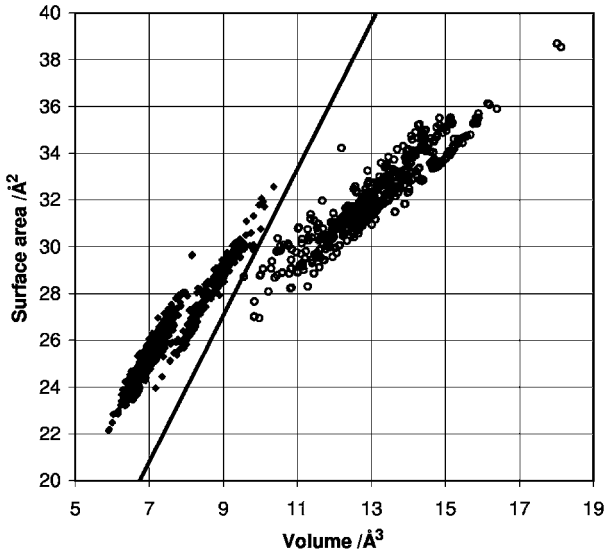

Figure 2
Data-distribution plot showing volume and total surface area. C given by solid markers, H by open circles. The discriminating border due to the two-variable committee is indicated.

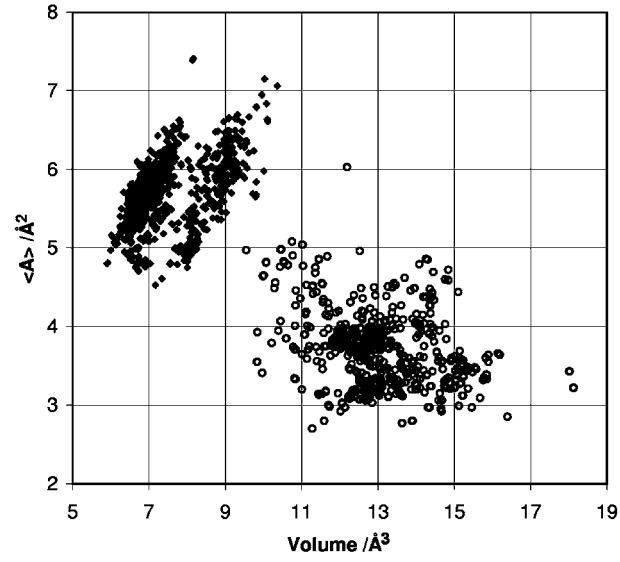

Figure 3
Data-distribution plot showing volume and average face area. C given by solid markers, H by open circles.

Table 2
Performance of models: two-variable problem.

	Misclassified C		Misclassified H	
Model	No.	%	No.	%
1	0	0	0	0
2	0	0	1	1.43
3	0	0	0	0
4	0	0	1	1.43
5	0	0	0	0
Average	0	0	0.4	0.57
Committee	0	0	0	0

Table 3
Performance of models: four-variable problem.

	Misclassified C		Misclassified H	
Model	No.	%	No.	%
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
Average	0	0	0	0
Committee	0	0	0	0

combination of the constituent models, was then assessed on the test data in terms of the misclassification risk (in %), calculated separately for C and H atoms. The individual risks of misclassification for each of the constituent models were also calculated.

4. Results

The misclassification risks for the two problems, two-variable and four-variable, are shown in Tables 2 and 3, respectively. For the two-variable problem, three of the constituent models correctly classified all the atoms in the test set, while the remaining two each classified 1 H atom wrongly as a C atom, corresponding to 1.43% of all H atoms. On average, the models then misclassified 0.57% of H atoms. The committee, however, classified all atoms correctly. The expectation is, therefore, that this committee will correctly classify all atoms in polynuclear aromatic hydrocarbons, based solely on their Voronoi volumes and surface areas. The discriminating border between the C and H regions, due to the committee, has been drawn in Fig. 2.

The weights for the individual models are given in Table 4; these apply to equation (2) where a is the tanh function, while $m = 1$ and $w_{\text{output},i} = 1$ in equation (1). Thus, for example, model 1 is

$$y_{\text{model 1}} = \tanh(1.718 + 3.746A_t - 8.909V), \qquad (6)$$

where the inputs, A_t and V have been normalized [*cf.* Table 1 and equation (5)]. The committee prediction is

$$y_{\text{committee}} = \tfrac{1}{5}\sum_{i=1}^{5} y_{\text{model } i}. \qquad (7)$$

Table 4
Model weights: two-variable problem.

Model	$w_{\text{input 0,1}}$ constant	$w_{\text{input 1,1}}$ area	$w_{\text{input 2,1}}$ volume	$w_{\text{output 1}}$ (clamped)
1	1.718	3.746	8.909	1
2	1.956	3.110	8.728	1
3	1.467	3.769	8.199	1
4	1.239	3.251	6.606	1
5	1.631	4.322	9.616	1

Table 5
Model weights: four-variable problem.

Model	$w_{\text{input 0,1}}$ constant	$w_{\text{input 1,1}}$ area	$w_{\text{input 2,1}}$ volume	$w_{\text{input 3,1}}$ $\langle A\rangle$	$w_{\text{input 4,1}}$ σ	$w_{\text{output 1}}$ (clamped)
1	0.871	0.029	9.686	4.199	0.855	1
2	0.173	8.170	8.150	9.882	1.014	1
3	0.378	2.962	1.488	5.376	0.217	1
4	0.192	1.249	9.462	8.511	1.006	1
5	0.309	6.208	5.570	6.238	3.135	1

For the four-variable problem, all individual models correctly classified all atoms in the test set and the committee, consequently, did the same. The weights for these models are given in Table 5.

5. Discussion

The main objectives of this study have been to illustrate an approach to using Voronoi tessellation variables to quantify local structural characteristics of atom types with carbon and hydrogen atoms in polynuclear aromatic hydrocarbons providing an example. The bivariate plots in Figs. 2–7 clearly illustrate that with such an approach there does exist a possibility to distinguish between these two atom types; two main clusters of points, one for C and one for H, stand out in several of the plots.[12] As undeniable as this tendency towards separate clustering of C and H data points is, it is equally undeniable that some combinations of variables give better separation than others. The combination of volume and total surface area, which possibly will be the one most easily arrived at with the majority of Voronoi tessellation software in existence, is fairly good, with H atoms generally having both larger volumes and surface areas. Note that taking only one of these variables into consideration could not possibly lead to perfect separation; both are required. And yet, even with these two variables, the separation is not large. The two two-variable models that misclassified one H atom both faltered on the same atom; the one with the smallest volume, and it is clear that this atom does not give much leeway for a separating border.

The weighted face area average, $\langle A\rangle$, does seem to be a very useful means of separating C and H, though, again, it is not

[12] The C cluster will be seen to consist of two smaller clusters; this corresponds to a difference in bond structure and is subject to on-going research by the author.

adequate on its own. Combined with either volume or total surface area, however, it does provide a very good opportunity for separation. The fact that the four-variable models on the whole fared better than the two-variable models is almost certainly, in part, due to the use of $\langle A \rangle$. More generally, though, the use of more variables, each of which is somewhat useful as a means of separation, will make such separation of clusters easier; if the distance along a variable between the centres of the clusters is d_i, where i runs over the number of variables, then the distance in the high-dimensional space is

$$d = \left(\sum_i d_i^2 \right)^{1/2}. \qquad (8)$$

In other words, the more variables, the more the centres of the clusters are pulled apart; *i.e.* better separation is a consequence.

Among the four, the weighted standard deviation from $\langle A \rangle$, σ, stands out as not being very useful, although there is a tendency for carbon atoms to have the higher values. σ could probably be dropped as an input without making the separation task more difficult and, in fact, the model optimization would be faster, with one weight less to determine. One thing this would seem to suggest is that, if possible, as many variables as can be thought of should be calculated and their use for separation, one at a time, be assessed. Finally, those most useful should be included in the modelling process. However, variables that are highly correlated with others might be better not included as optimization time might otherwise become too long.

Correlations clearly exist, as one would of course expect. Volume and surface area, for example, are both measures of size and should be correlated, as indeed they are. The nature of the correlation differs, however, between C and H, and so gives rise to discrimination. More surprising, correlations exist between volume (or surface area) and average face area (for C only), between volume (or surface area) and σ (for C only) and, vaguely, between average face area and σ (both C and H). The fact that these correlations are more pronounced for C indicate that there is a larger degree of randomness to the H positions than to their C counterparts; *i.e.* the C atoms are more constrained in terms of position.

The MLP models generally performed well; a quick glance at the bivariate distribution plots will suggest, however, that nearest-neighbour models would probably also perform well. The considerable drawback with those models is, of course, that they are much more cumbersome (*i.e.* slower) to use and also are not suited for printing in tabular form, particularly when there are many atoms in the database from which the training data are obtained. This does not preclude the possibility that they would be more accurate, with a lower risk of misclassification and for such classification tasks they should not be readily dismissed. Nor should it be supposed that the MLP, in general, is particularly well suited for these tasks; it is one of the most basic of modelling approaches and, whilst the ease of its implementation is attractive, there are many other, potentially more powerful, techniques that may prove more useful, particularly for more complex classification problems where several atom types are involved and/or more polyhedral features are required for separation.

In this ability to choose more features with which to characterize the local environments, to which the atoms are confined, lies a prominent strength of the Voronoi approach, as opposed to the bond length/angle approach. Whilst the latter can clearly be used for perfect classification in moderately simple cases, the much larger number of possible Voronoi variables, giving a more detailed view of the local structure, may prove especially valuable when the complexity of the problem is very high. It must be remembered, however, that mere quantity is not necessarily implying quality; many

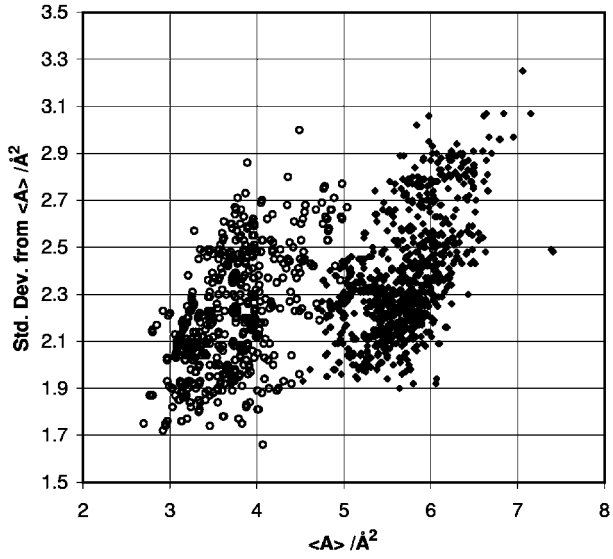

Figure 4
Data-distribution plot showing average face area and standard deviation from this area, σ. C given by solid markers, H by open circles.

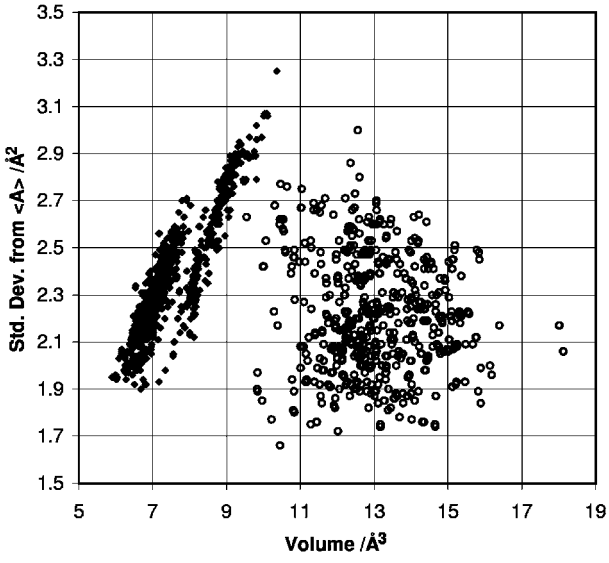

Figure 5
Data-distribution plot showing volume and σ. C given by solid markers, H by open circles.

Voronoi variables may indeed be found to contain little, if any, information. The weighted standard deviation from the weighted face-area average, σ, being an example thereof although, admittedly, a variable that on its own does not seem to provide any separation at all may still, in conjunction with another, or even several others, enable such separation in a higher-dimensional feature space.

Finally, the expectation that the two-variable models might end up being better than the four-variable models, because the optimization problem was smaller, turned out to be a moot point; the optimization of both systems required only little CPU time and so even the larger problem was quite tractable. Nevertheless, if considerably more complex problems are to be investigated, with greater similarity between the atom types involved and possibly tens of thousands of atoms to be included, then this issue will probably come to the fore. In future work by the author, concerned with elucidating the feasibility of using this technique on proteins, this will be investigated.

6. Conclusions

In this study, it has been found that it is possible to distinguish between C and H atoms in crystalline polynuclear aromatic hydrocarbons through calculating the Voronoi polyhedra of the atoms and, in turn, examining their volumes, surface areas, weighted face-area averages and a similarly weighted standard deviation from these.

On test data, not used for training the models, a simple committee of five multilayer perceptrons predicted, without error, the correct atom type from information only on polyhedral volume and surface area, although two of the individual models incorrectly identified one H atom as C. Using also the two remaining variables, all five models trained predicted the correct atom type for all atoms.

The success of the models stems from the fact that, in a variable space spanned by these variables, C and H Voronoi polyhedra tend strongly to cluster separately. This clustering is particularly pronounced with respect to volume, total surface area and weighted face-area average, whereas the weighted standard deviation from the latter average is of lesser importance.

It is expected that similar success will be possible on systems of much greater complexity, containing atoms with similar scattering curves, although it is conceivable that more Voronoi variables will be required in those cases. Future work will be devoted to elucidating this point fully, with specific focus initially lying on proteins and inorganic compounds.

APPENDIX *A*

The crystallographic data for this study were obtained with the Crystal Structure Search and Retrieval facility (CSSR) at Daresbury and are taken from the Cambridge Structural Database. Numbers quoted are CSSR reference numbers.

The following compounds were included in the study.

Benzene: (2334, 2335, 2337, 2338). Naphthalene: (15282, 15283, 39183, 39184, 39185, 39186, 39187, 41288, 149593). Anthracene: (1334, 1335, 1336, 44319, 44320, 92498, 92499, 92500, 92501, 92502, 92503). Phenanthrene: (17326, 17327, 17328, 95753). Biphenyl: (2526, 2529, 2530, 148595, 148596, 148597). Chrycene: (6089). Triphenylene: (21749, 21750). Quaterphenyl: (18814). Benzo[*c*]phenanthrene: (3647). Picene: (54600). Dibenzanthracene: (6988, 6989). Pyrene: (18657, 18658, 18659, 18660). Perylene: (17163). Dinaphthoanthracene: (7949). Quaterrylene: (18773). Benzopyrene: (2704). Annulene: (1387, 148567). Dibenzoperylene: (6998). Coronene: (5445). Ovalene: (16356). Tribenzopyrene: (19923).

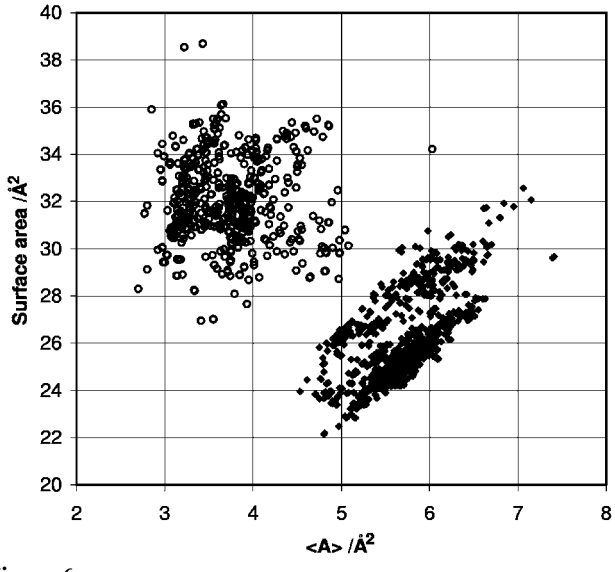

Figure 6
Data-distribution plot showing average face area and total surface area. C given by solid markers, H by open circles.

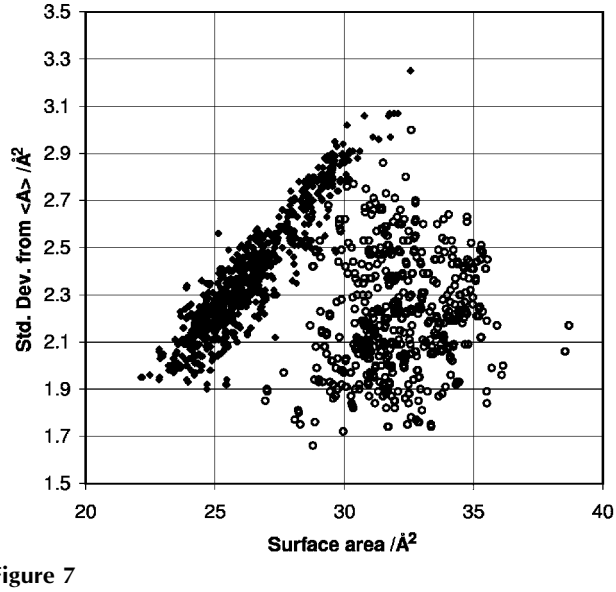

Figure 7
Data-distribution plot showing total surface area and σ. C given by solid markers, H by open circles.

Tetrabenzoperylene: (19920). Anthrabenzonaphthopentacene: (41591). Diperinaphthyleneanthracene: (108608).

The author wishes to acknowledge the use of the EPSRC's Chemical Database Service at Daresbury.

References

Andersson, K. M. & Hovmöller, S. (1998). *Z. Kristallogr.* **213**, 369–373.
Blatov, V. A., Shevchenko, A. P. & Serezhkin, V. N. (1995). *Acta Cryst.* A**51**, 909–916.
Christensen, S. W. & Thomas, N. W. (1999). *Acta Cryst.* A**55**, 811–820.
Fischer, W. & Koch, E. (1979). *Z. Kristallogr.* **150**, 245–260.
Fletcher, D. A., McMeeking, R. F. & Parkin, D. (1996). *J. Chem. Inf. Comput. Sci.* **36**, 746–749.
Koch, E. & Fischer, W. (1980). *Z. Kristallogr.* **153**, 255–263.
Krogh, A. & Vedelsby, J. (1995). *Advances in Neural Information Processing Systems 7*, edited by G. Tesauro, D. S. Touretzky & T. K. Leen, pp. 231–238. Cambridge, MA: MIT Press.
Mackay, A. L. (1972). *J. Microsc.* **95**, 217–227.
Perrone, M. P. & Cooper, L. N. (1993). In *Neural Networks for Speech and Image Processing*, edited by R. J. Mammone. London: Chapman & Hall.
Richards, F. M. (1977). *Ann. Rev. Biophys. Bioeng.* **6**, 151–176.