Powder Diffraction Indexing as a Pattern Recognition Problem: A New Approach for Unit Cell Determination Based on an Artificial Neural Network

Scott Habershon, Eugene Y. Cheung, Kenneth D. M. Harris,* and Roy L. Johnston*

School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

Received: September 9, 2003; In Final Form: November 10, 2003

An artificial neural network, in combination with local optimization, is shown to be an effective approach for determining unit cell parameters directly from powder diffraction data. The viability of this new approach is initially demonstrated using simulated powder diffraction data. Subsequently, the successful application of the method to determine unit cell parameters is illustrated for two materials using experimental powder X-ray diffraction data recorded on a standard laboratory diffractometer.

1. Introduction

The determination of the unit cell parameters of a crystalline sample via analysis of the positions of Bragg reflections in a powder X-ray (or neutron) diffraction pattern, a process commonly referred to as "indexing", is an important first step in the structural analysis of many materials. However, the process of indexing can present significant difficulties, and failure to successfully index a powder diffraction pattern usually prevents further analysis of the data—for example, complete structure determination can be carried out only if the powder diffraction data have been indexed correctly. Indeed, in our recent work concerning crystal structure determination from powder X-ray diffraction data,[1,2] we have found that indexing can be the limiting step in the structure elucidation process.

The position, $2\theta_{hkl}$, of the Bragg reflection with Miller indices (hkl) in a powder diffraction pattern is related to the unit cell parameters, $\{a, b, c, \alpha, \beta, \gamma\}$, by

$$2\theta_{hkl} = 2 \sin^{-1}\left(\frac{\lambda}{2d_{hkl}}\right) \quad (1)$$

where

$$d_{hkl} = V\,[h^2b^2c^2 \sin^2 \alpha + k^2a^2c^2 \sin^2 \beta + l^2a^2b^2 \sin^2 \gamma + 2hlab^2c(\cos \alpha \cos \gamma - \cos \beta) + 2hkabc^2(\cos \alpha \cos \beta - \cos \gamma) + 2kla^2bc(\cos \beta \cos \gamma - \cos \alpha)]^{-1/2} \quad (2)$$

In these equations, d_{hkl} is the interplanar spacing for the (hkl) set of lattice planes, V is the unit cell volume, and λ is the X-ray or neutron wavelength. The aim of indexing is to determine the unit cell parameters that correctly reproduce the set of reflection positions observed in the experimental powder diffraction pattern.

Although the indexing process requires the determination of, at most, only six parameters, there are several features of powder diffraction data that can significantly limit the chances of success. Problems of particular importance include the presence of crystalline impurity phases, significant zero-point error, systematic errors introduced by peak broadening and peak overlap and the fact that certain reflections may be weak or unobserved due to being systematically absent (as a consequence of symmetry), having intrinsically low structure factor moduli or being suppressed by the effects of preferred orientation.

In general, current approaches for indexing fall into two main categories: *search* (exhaustive or optimization) methods and *deductive* methods. Search methods tackle the indexing problem by attempting to locate the optimal set of unit cell parameters on a hypersurface $F(a, b, c, \alpha, \beta, \gamma)$ defined by a suitable figure-of-merit F (such as the M_{20} function[3]) that describes the agreement between the experimentally observed and calculated peak positions. This basic philosophy is embodied in the popular *DICVOL* program,[4] as well as in several more recent programs.[5−8] Deductive methods adopt a different approach, in which possible unit cell parameters are proposed by recognizing well-defined relationships between sets of observed peak positions. This approach forms the basis of programs such as *ITO*[9] and *TREOR*.[10]

An alternative approach, proposed in this Letter, is to view indexing as an exercise in *pattern recognition*; thus, given a set of observed peak positions in an experimental powder diffraction pattern, we wish to predict the likely unit cell parameters by utilizing knowledge derived from analysis of the peak positions in powder diffraction patterns corresponding to known unit cell parameters. Artificial neural networks (ANNs) are ideally suited to solve such problems, and this Letter reports the first application of an ANN to tackle the problem of indexing powder diffraction data.

2. Methodology

ANNs are a class of computer algorithm that, in an elementary sense, attempt to mimic the logical operation of the human brain. The potential of ANNs as tools for image recognition, process control, and feature extraction is considerable, and many applications in a wide variety of disciplines have been reported.[11,12] Previous work has included applications in areas of powder diffraction[13,14] that are unrelated to the present work.

In general, the application of an ANN to solve a given problem involves the following stages: (i) the definition of input and output variables and the selection of a suitable network topology linking the input and output, (ii) training the ANN using data for which both the input variables and the corresponding output variables are known, and (iii) application of the trained ANN to determine the (unknown) set of output

* To whom correspondence should be addressed. E-mail: K.D.M.Harris@bham.ac.uk, roy@tc.bham.ac.uk.

10.1021/jp0310596 CCC: $27.50
Published on Web 01/07/2004

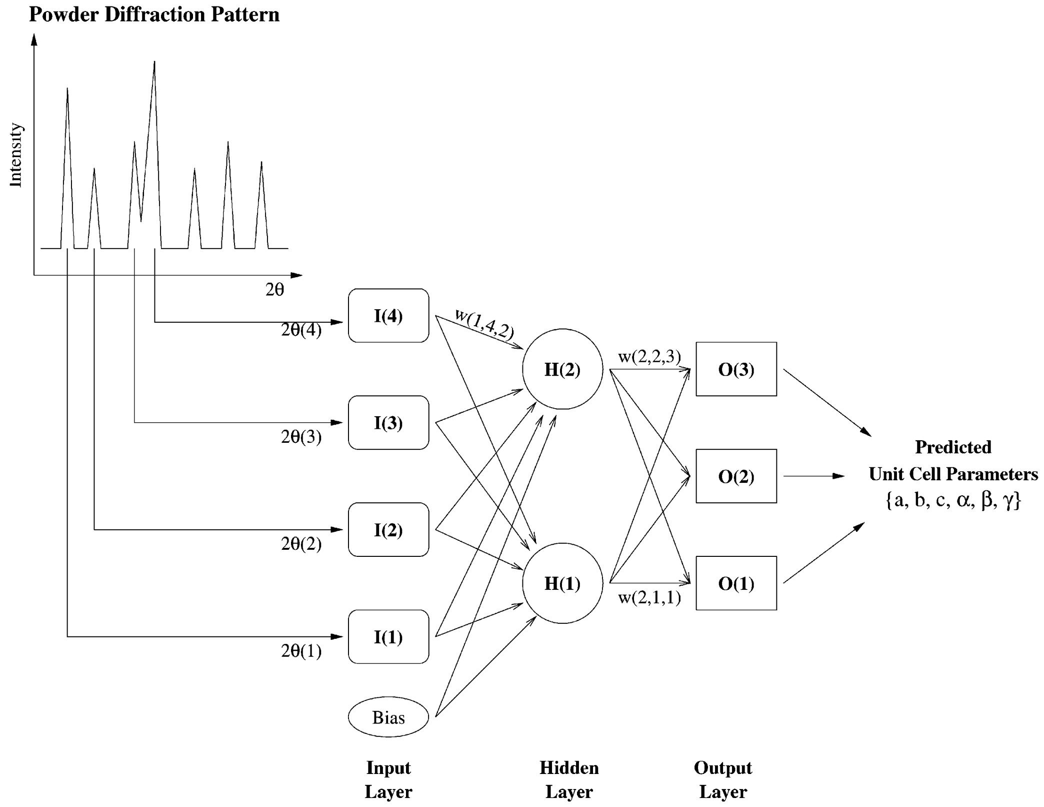

Figure 1. Schematic illustration of our ANN approach for indexing powder diffraction data. Peak positions $2\theta(l)$ are extracted from the powder diffraction pattern and are subsequently passed into the input layer of the ANN (after appropriate scaling; see text). The input signals are then transferred to the hidden layer, which processes the signals to generate the output signals at the output layer. The output signals are then interpreted to produce a set of unit cell parameters. Note that, for clarity, only a few connection weights are shown, though it should be remembered that each connection has an associated weight. In the above illustration, the weight values are specified by an extra index, indicating the connection layer, in comparison to those in eq 3.

variables corresponding to a particular set of input variables (e.g., experimental data). In the present context, the input variables are peak positions in a powder diffraction pattern and the output variables are the unit cell parameters. We now elaborate details of each of these stages in the context of the ANN that we have developed for indexing powder diffraction data.

2.1. Network Topology and Parameter Definitions. The work presented here is focused upon a particular type of ANN known as a "feed-forward multilayer perceptron network".[11,12] In general, such a network comprises three major components: (i) a layer of nodes that provide input signals to the network, referred to as the *input layer*, (ii) a number of *hidden layers* containing the *hidden units* of the network, and (iii) a layer of nodes that interpret the output of the final hidden layer, referred to as the *output layer*. A schematic illustration of a network of this type is shown in Figure 1. In our terminology, a *node* serves to provide input or collect output from the network. The key components of the ANN are the hidden units, which are the artificial analogues of biological neurons and act in a similar (though vastly simplified) manner. Thus, a hidden unit operates by summing the input signals that it receives from other units (either nodes or other hidden units) to which it is connected, analogous to the way in which a neuron sums nerve impulses. Each network connection has an associated weight, $w_{i,j}$, which dictates the effective magnitude of the received signal, $x_{i,j}$. The *activation* value, a_j, produced at unit j by the signals (labeled $i = 1, ..., N$) that it receives from the units to which it is connected is calculated as follows:

$$a_j = \sum_{i=1}^{N} w_{i,j} x_{i,j} \tag{3}$$

To simulate the continuous signal values produced by biological neurons, the output signal y_j produced by activation at unit j is calculated from a_j via a sigmoid function,

$$y_j = \frac{1}{1 + \mathrm{e}^{-a_j}} \tag{4}$$

The output signal may then act as an input signal for further units within the network.

In the present case, each node or hidden unit in a given layer is connected to every node or hidden unit in the preceding layer, with each connection being characterized by a weight value. The network is thus described as being fully connected. We note that in the present work, the ANN contains only a single hidden layer, and the input layer includes an additional node referred to as the *bias*. This node passes a constant value of 1 to all units in the hidden layer. From eqs 3 and 4, it may be seen that the effect of this bias signal is to shift the activation values into a more appropriate range, such that saturation of the sigmoid function is avoided.

Besides the actual topology of the network, another important factor is the definition of the input signals, as well as the manner in which the signals produced at the output layer of the network are interpreted. In the current work, the input to the ANN comprises the 2θ positions of N_i peaks in a powder diffraction pattern (typically the 20−30 peaks at lowest 2θ). The output signals ρ_k ($0 \leq \rho_k \leq 1$) produced by the ANN are interpreted as normalized unit cell parameters, which may be rescaled to recover each unit cell parameter x_k ($k = 1, 2, ..., 6$) via,

$$x_k = x_k^{\min} + \rho_k(x_k^{\max} - x_k^{\min}) \qquad (5)$$

where $x_k^{\min}$ and $x_k^{\max}$ are user-defined minimum and maximum allowed values of the unit cell parameter x_k.

From the above definitions, it should be clear that, if the ANN has "knowledge" of the relationship connecting the input and output variables (i.e., eqs 1 and 2 in the present work), the network will be suitable for predicting unit cell parameters when presented with N_i peak positions from a powder diffraction pattern. The manner in which such "knowledge" may be acquired by the ANN is the focus of the *training* process, which is now outlined.

2.2. Training an ANN. For an ANN to perform a useful function, the network must be trained. In computational terms, training represents an exercise in error minimization; when the ANN is presented with an input data set, which is related to a *known* target output, the error between the output calculated by the ANN and the target output should be minimal. This is achieved by adjusting the connection weights in the network, which may thus be viewed as containing the "knowledge" of the network. In practice, the training process is carried out not for a single set of input and output data, but for a large number of such data sets, which are collectively referred to as the *training set*. This approach encourages the ANN to learn the fundamental relationships that connect input and output data for the type of problem under investigation, rather than learning specific features of an individual input/output pair. If these relationships are learned correctly and to a sufficient degree of accuracy, the trained ANN may then be used to predict the correct output data when presented with a set of previously unseen input data. This ability of a trained ANN to successfully *generalize* is obviously a highly desirable quality, and underpins the predictive capabilities of an ANN.[11,12]

In the current work, the weights are adjusted during a number of *training epochs* to minimize a least-squares error function,

$$E_t = \frac{1}{N_t}\sum_{m=1}^{N_t}\sum_{k=1}^{N_o}(x_k(m) - t_k(m))^2 \qquad (6)$$

where the training error, E_t, is calculated over all N_t members of the training set, $x_k(m)$ is the scaled ANN output (eq 5) at the kth output node (of which there are N_o) when input data set m is presented at the input layer of the ANN and $t_k(m)$ is the target output corresponding to training set m. In the current work, the error minimization is achieved using a modified version of the *Rprop* algorithm,[15] namely *iRprop*$^+$.[16] In all calculations detailed below, the training process is allowed to proceed for a fixed number N_e of training epochs, though we note that other criteria, such as minimum error values or gradient norms, may be used to judge when training should be stopped.

2.3. Specification of the ANN for Indexing. Our ANN approach for indexing powder diffraction data can now be fully defined. Given the requirement to carry out training using a large number of data sets for which the unit cell parameters are known, it is not viable to use experimental powder diffraction data for this purpose. Instead, the training set is created by randomly generating N_t sets of unit cell parameters (labeled $m = 1, ..., N_t$) and the input data comprise the 2θ positions of the first N_i lines (labeled $l = 1, ..., N_i$) in the powder diffraction pattern generated for each of these sets of unit cell parameters. Clearly, the target outputs during the training period are the sets of unit cell parameters used to generate the input data. At each input node l, it is beneficial to use scaled input signals, $s_{l,m}$, which are obtained from the peak positions $2\theta_{l,m}$ according to

$$s_{l,m} = \frac{2\theta_{l,m} - \mu_l}{\sigma_l} \qquad (7)$$

where the mean input signal μ_l at each input node is given by

$$\mu_l = \frac{\sum_{m=1}^{N_t} 2\theta_{l,m}}{N_t} \qquad (8)$$

and the standard deviation σ_l is given by

$$\sigma_l = \sqrt{\frac{\sum_{m=1}^{N_t}(2\theta_{l,m} - \mu_l)^2}{N_t - 1}} \qquad (9)$$

After construction of the training set, the weights are adjusted during each of the N_e training epochs using the *iRprop*$^+$ algorithm.

Subsequently, the trained ANN can be applied to index a previously unseen set of input data (e.g., from an experimental powder diffraction pattern). The input data comprises the 2θ positions of N_i peaks in the powder diffraction pattern, scaled according to eq 7. The unit cell parameters predicted by the trained ANN are recovered from the network output signals according to eq 5. A schematic of the overall process is shown in Figure 1.

2.4. Further Crystallographic Considerations. The ability of the ANN to deal with previously unseen data may be enhanced by the introduction of several problem-specific features. First, although any powder diffraction pattern may be indexed by a primitive triclinic unit cell, it is beneficial to train the ANN separately for each different crystal metric symmetry. This strategy obviously reduces the network complexity and, consequently, the training difficulty. In the present work, only ANNs with three output nodes (i.e., $N_o = 3$ in eq 6), suitable for indexing orthorhombic unit cells, are considered, though current developments include the extension to other metric symmetries. Second, the chances of successfully indexing a powder diffraction pattern may be increased by explicit consideration of systematic absences. In our ANN approach, this has been implemented by training a separate ANN (i.e., a different set of weights is generated) for each of the possible classes of systematic absences (*powder extinction classes*) in a given crystal system.[17] When applying the method to index a previously unseen set of input data, the output from each powder extinction class-specific ANN may be assessed, providing an opportunity to establish both the correct unit cell and the correct powder extinction class. In this way, the ANN can also directly assist the process of space group assignment. Finally, a small amount of random "noise" can be introduced to the input data during training to simulate the effect of random errors in the peak positions, which are inevitably present in an experimental powder diffraction pattern. An ANN trained in this way may thus be more appropriate for subsequent application to real experimental data. The introduction of noise into the training set may also improve the generalization abilities of the ANN by "blurring out" features of the error landscape which could, in certain cases, lead to the error minimization process becoming trapped in a local minimum.[11,12] We note that the random noise is added to the input signals $s_{l,m}$ (eq 7) and not to the input 2θ values. Because σ_l generally increases with 2θ, this strategy implicitly includes the fact that errors in peak positions extracted from a powder diffraction pattern tend to increase with 2θ, due primarily to increased peak overlap.

The final aspect of our methodology concerns the application of local optimization. Because unit cell parameters predicted by the ANN will inevitably contain some errors, it is desirable to refine the unit cell parameters predicted by the ANN to achieve the best possible agreement with the experimental data. In the current work, this is achieved by optimizing (maximizing) the value of M_N, the de Wolff figure-of-merit for the N peaks in the input data, using the simplex algorithm of Nelder and Mead.[18]

3. Application, Results, and Discussion

All calculations were performed on a 1 GHz AMD Athlon processor (running RedHat Linux 7.3) using our custom-written ANN program *PIANNO*.[19] An ANN comprising 15 hidden units (in a single hidden layer), 20 input units, and 3 output units was trained using a training set of size $N_t = 1000$. The target outputs (i.e., unit cell parameters with orthorhombic metric symmetry) of the training set were randomly generated with minimum and maximum unit cell lengths of 5 and 25 Å, respectively, and with minimum and maximum allowed unit cell volumes of 100 and 4000 Å^3, respectively. The target unit cell parameters were ordered $a > b > c$. No peaks were removed from the input data due to systematic absences, and thus the set of optimized weights is specific to the $P222$ extinction class. The training was allowed to proceed for 5×10^3 training epochs, and artificial noise was not introduced during training. After training, the training error (eq 6) was 0.016 Å^2, corresponding to an average error in each unit cell parameter of just 0.042 Å.[20]

Following training, the ability of the ANN to predict the unit cell parameters for a specific set of input data was assessed. The first 20 peak positions in the simulated powder diffraction pattern of an orthorhombic $P222$ unit cell with $a = 15.00$ Å, $b = 12.00$ Å, and $c = 9.00$ Å were presented to the trained network. The initial unit cell parameters predicted by the ANN were $a = 14.99$ Å, $b = 12.02$ Å, and $c = 9.02$ Å, in excellent agreement with the correct parameters. The exact unit cell parameters were obtained following local optimization.

With the successful application of the ANN thus confirmed for simulated powder diffraction data, the potentially more difficult case of experimentally recorded powder diffraction data was considered. The powder diffraction patterns considered were those of the β phase of L-glutamic acid (β-LGA) and the peptide Piv-$^{\text{L}}$Pro-γ-Abu-NHMe (PPAN).[21] The crystal structures of both materials had been solved previously using single-crystal neutron diffraction (β-LGA[22]) and powder X-ray diffraction (PPAN[23]) respectively, and the unit cell parameters and space group ($P2_12_12_1$ in both cases) were already known.[24,25]

For these calculations, a network comprising 22 hidden units, 25 input values, and 3 output values was trained. The training set comprised 1000 randomly generated orthorhombic unit cells, with minimum and maximum unit cell lengths of 5 and 25 Å, respectively, and minimum and maximum unit cell volumes of 100 and 2000 Å^3, respectively. The input data corresponding to each member of the training set had the systematic absences of the $P2_12_12_1$ extinction class explicitly removed, such that the trained set of weights was specific to this extinction class. As described above, a random noise value $\kappa \in [-0.1, +0.1]$ was introduced into the input signals $s_{l,m}$ at the beginning of each training epoch. After 5×10^3 training epochs, the average error for the members of the training set was 1.052 Å^2, corresponding to an average parameter error of 0.342 Å.[26] The errors reported in this case are larger than those for the $P222$ case, as a result of the introduction of random noise into the training data, as well as the explicit removal of systematic absences, which can affect the input signals $s_{l,m}$ via μ_l and σ_l in eq 7.

TABLE 1: Results of the Indexing Calculation for the β Phase of L-Glutamic Acid[a]

	unit cell parameters			
	a/Å	b/Å	c/Å	M_{25}
ANN	17.282	7.339	5.128	1.53
optimized	17.369	6.979	5.180	85.16
correct	17.368	6.980	5.180	

[a] The unit cell parameters initially predicted by the ANN are given, as are the final parameters after local optimization using a simplex algorithm. The correct unit cell parameters (obtained from Le Bail fitting of the powder X-ray diffraction pattern) are given for comparison.

TABLE 2: Results of the Indexing Calculation for Piv-$^{\text{L}}$Pro-γ-Abu-NHMe[a]

	unit cell parameters			
	a/Å	b/Å	c/Å	M_{25}
ANN	17.445	10.901	8.593	1.85
optimized	16.996	10.773	9.175	22.84
correct	16.930	10.718	9.142	

[a] The unit cell parameters initially predicted by the ANN are given, as are the final parameters after local optimization using a simplex algorithm. The correct unit cell parameters (obtained after Rietveld refinement of the crystal structure using powder X-ray diffraction data[23]) are given for comparison.

The unit cell parameters determined by the ANN when presented with 25 extracted peak positions from the experimental powder X-ray diffraction patterns of β-LGA and PPAN are shown in Tables 1 and 2, respectively. In each case, the unit cell predicted by the ANN is relatively close to the correct unit cell, although with a maximum discrepancy of up to ca. 0.6 Å. However, in both cases, local optimization proceeds in a straightforward manner to give the correct unit cell parameters, indicating that the unit cell parameters predicted by the ANN are in the close vicinity of the global minimum. We note that, when the β-LGA input data are presented to a network trained with $P222$ unit cells, the predicted unit cell parameters are in error by up to ca. 7 Å (with an M_{25} value after local optimization of 1.97), indicating that the ANN approach has the ability to discriminate against incorrect extinction classes within a given crystal system.

Not surprisingly, the results indicate that the accuracy of the unit cell parameters predicted by the ANN is greater when simulated data, rather than experimental data, are used. In the present examples, this situation most likely arises as a result of two key features of the experimental input data, as illustrated in Figure 2. First, and most importantly, a number of peaks predicted by the optimized (and correct) unit cell were not included in the peak list selected as input to the ANN. These reflections were excluded either because their intensity is too weak (but not systematically absent) or because they overlap significantly with other, more pronounced peaks (we note that no preferred orientation effects are present in either experimental powder diffraction pattern). The absence of these peaks inevitably affects the performance of the ANN, because it is

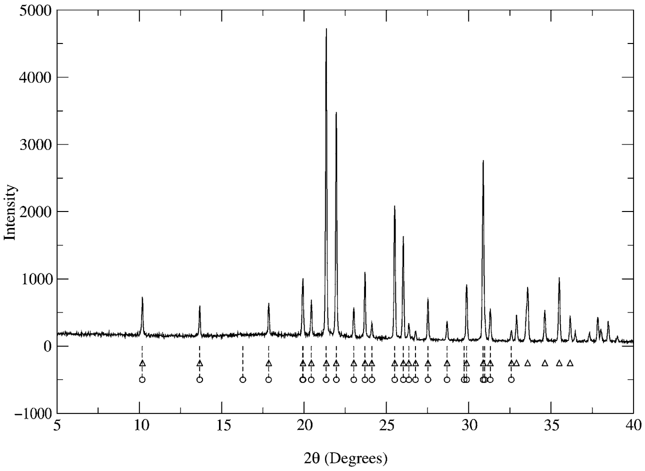

Figure 2. Experimental powder diffraction pattern of the β phase of L-glutamic acid. The peak positions selected as input data for the ANN are shown as triangles and the peak positions corresponding to the final optimized unit cell are shown as circles with dashed indicator lines. The peak predicted by the optimized unit cell at approximately 17° is absent from the input experimental peak list simply due to its weak intensity. Note the almost perfect agreement between the sets of experimental and predicted peak positions.

trained with data for which *all* of the first N_i peaks of the simulated powder diffraction pattern are included. However, the ability of the ANN to successfully determine the correct unit cell parameters indicates that the information contained in the input peak list from the experimental data is sufficient to allow the ANN to predict a unit cell at least within the basin of attraction of the global minimum on the figure-of-merit hypersurface. The second important point concerns the zero-point error in the powder diffraction pattern. Again, no explicit information concerning the zero-point error was introduced during training, though our results suggest that the ANN is not significantly perturbed by zero-point errors as large as 0.04°, as in the powder diffraction pattern of PPAN.

4. Conclusions

This Letter has highlighted a new approach for indexing powder diffraction data using an ANN in combination with local optimization. This approach, in which the process of indexing is viewed as a pattern recognition exercise, represents a new strategy in comparison with current indexing methodologies. Results obtained using both simulated and experimental powder diffraction patterns (the latter recorded on a standard laboratory powder X-ray diffractometer) demonstrate the successful application of our new methodology. The calculation times required are minimal and are of the order of a few minutes for ANN training and a few seconds for prediction of unit cell parameters from previously unseen data once the ANN has been trained.[27] Obviously, once a reliable set of weights for a given metric symmetry and extinction class has been generated, the ANN requires no further training to index powder diffraction data for the specified metric symmetry and extinction class. This strategy therefore offers the possibility of significant gains in calculation speed compared to many standard indexing algorithms.

Current work is focused upon extending our methodology to other metric symmetries, as well as developing a more detailed understanding of the factors affecting the accuracy of the unit cell parameters predicted by the ANN. Importantly, we note that, with continued tuning of the training process, as well as the selection of suitable training data, the ANN approach described here has the potential to successfully deal with many of the common problems encountered in indexing, and may therefore prove to be a reliable, robust and widely applicable method for tackling the indexing problem.

Acknowledgment. We gratefully acknowledge financial support from The University of Birmingham and EPSRC (studentship to S.H., postdoctoral fellowship to E.Y.C., and the award of computer equipment under the Strategic Equipment Initiative). We also acknowledge helpful discussions with Mr. R. J. J. Gilham and Mr. T. V. Mortimer-Jones.

References and Notes

(1) Harris, K. D. M.; Tremayne, M.; Kariuki, B. M. *Angew. Chem., Int. Ed.* **2001**, *40*, 1626.

(2) Harris, K. D. M.; Johnston, R. L.; Cheung, E. Y.; Turner, G. W.; Habershon, S.; Albesa-Jové, D.; Tedesco, E.; Kariuki, B. M. *CrystEngCommun.* **2002**, *4*, 1.

(3) de Wolff, P. M. *J. Appl. Crystallogr.* **1968**, *1*, 108.

(4) Boultif, A.; Louër, D. *J. Appl. Crystallogr.* **1991**, *24*, 987.

(5) Kariuki, B. M.; Belmonte, S. A.; McMahon, M. I.; Johnston, R. L.; Harris, K. D. M.; Nelmes, R. J. *J. Synchrotron Rad.* **1999**, *6*, 87.

(6) Neumann, M. A. *J. Appl. Crystallogr.* **2003**, *36*, 356.

(7) Hageman, J. A.; Wehrens, R.; de Gelder, R.; Buydens, L. M. *J. Comput. Chem.* **2003**, *24*, 1043.

(8) Coelho, A. A. *J. Appl. Crystallogr.* **2003**, *36*, 86.

(9) Visser, J. W. *J. Appl. Crystallogr.* **1969**, *2*, 89.

(10) Eriksson, L.; Westdahl, M. *J. Appl. Crystallogr.* **1985**, *18*, 367.

(11) Gurney, K. *An Introduction to Neural Networks*; UCL Press: London, 2002.

(12) Swingler, K. *Applying Neural Networks: A Practical Guide*; Academic Press: London, 1996.

(13) Glorieux, C.; Zolotoyabko, E. *J. Appl. Crystallogr.* **2001**, *34*, 336.

(14) Agatonovic-Kustrin, S.; Wu, V.; Rades, T.; Saville, D.; Tucker, I. G. *J. Pharm. Biomed. Anal.* **2000**, *22*, 985.

(15) Riedmiller, M.; Braun, H. A Direct Adaptive Method for Faster Back-propagation Learning: The RPROP Algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*; IEEE Press: 1993.

(16) Igel, C.; Hüsken, M. Improving the Rprop Learning Algorithm. In *Proceedings of the Second International Symposium on Neural Computation, NC′2000*; ICSC Academic Press: 2000.

(17) Hahn, T., Ed. *International Tables for X-ray Crystallography Volume A (Space Group Symmetry)*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989.

(18) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN77: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 1992.

(19) Habershon, S.; Cheung, E. Y.; Johnston, R. L.; Harris, K. D. M. PIANNO − Powder Indexing by an Artificial Neural Network with Optimization. University of Birmingham, Birmingham, United Kingdom, 2003.

(20) As a further assessment of the performance of the trained ANN, the average error (calculated using eq 6) for a set of 500 randomly generated input/output pairs, referred to as the *validation set*, was found to be 0.020 Å^2, corresponding to an average parameter error of 0.047 Å. Because the members of the validation set were not used during training, the validation set error represents a more reliable assessment of the generalization abilities of the ANN than the training error.

(21) Both powder X-ray diffraction patterns were recorded in transmission mode on a Siemens D5000 diffractometer using Ge-monochromated Cu Kα_1 radiation with a linear position-sensitive detector covering 8° in 2θ and a step size of 0.019°.

(22) Lehmann, M. S.; Koetzle, T. F.; Hamilton, W. C. *J. Cryst. Mol. Struct.* **1972**, *2*, 225.

(23) Cheung, E. Y.; McCabe, E. E.; Harris, K. D. M.; Johnston, R. L.; Tedesco, E.; Raja, K. M. P.; Balaram, P. *Angew. Chem. Int. Ed.* **2002**, *41*, 494.

(24) In the work presented here, the "correct" unit cell of the β phase of L-glutamic acid is defined as that obtained after Le Bail fitting[28] of the powder X-ray diffraction data, rather than that obtained from the reported[22] single-crystal neutron diffraction results, to avoid potential discrepanices caused by differing experimental conditions (e.g., temperature).

(25) For both β-LGA and PPAN, indexing was also attempted using the three most commonly used indexing algorithms (*ITO*,[9] *TREOR*,[10] and *DICVOL*[4]). Only one of these programs successfully determined the correct unit cell parameters in both cases. One program succeeded for β-LGA but failed for PPAN, whereas the other program succeeded for PPAN but failed for β-LGA.

(26) The average validation set error, calculated from 500 randomly generated input/output pairs, was 0.712 Å^2, corresponding to an average parameter error of 0.281 Å.

(27) The calculation times for indexing by our ANN methodology are comparable to those required for *ITO*,[9] *TREOR*,[10] and *DICVOL*[4] (when the search is constrained to the orthorhombic system).

(28) Le Bail, A.; Duroy, H.; Fourquet, J. L. *Mater. Res. Bull.* **1988**, *23*, 447.