PLUXter: Rapid Discovery of Metal-Organic Framework Structures Using PCA and HCA of High Throughput Synchrotron Powder Diffraction Data

Deborah Lau[*], David G. Hay, Matthew R. Hill, Benjamin W. Muir, Scott A. Furman and Danielle F. Kennedy

CSIRO Materials Science & Engineering, Clayton South, Victoria 3169, Australia

Abstract: This paper reports on a systematic method – called PLUXter – developed for screening and data mining of large numbers of potential metal-organic framework compounds that have been synthesized then subsequently analyzed with high throughput synchrotron radiation X-ray powder diffraction. The first part of the method utilizes principal components analysis (PCA), which allows materials to be ranked in order of crystallinity so that undesirable amorphous materials may be identified and eliminated. The second part allows structural grouping within and between samples to be observed using hierarchical cluster analysis (HCA). Classification using a single linkage distance produced unsatisfactory clusters however the dendrogram's structural relationships were used to establish and guide the boundaries of groups. The resultant grouping identities allowed further structure-property studies to be undertaken on representative structures from the clusters, significantly reducing time and the use of resources.

Keywords: Hierarchical cluster analysis, principal components analysis, powder X-ray diffraction, metal-organic frameworks, synchrotron radiation.

INTRODUCTION

Materials discovery is being accelerated by high throughput (HT) approaches to synthesis and characterization. HT screening approaches to materials discovery [1] rely on a cycle of rapid synthesis, characterization, data interpretation or data mining and feedback. The method of characterization utilized is very much dependent on the chemical composition of the material being developed [2]. Importantly, different characterization techniques produce different data outputs, and in some cases an individual technique can produce several data outputs that are specific to a type of detector, instrument or file format. Since there are many possible data structures that are a consequence of characterization, individual data analysis methodologies that are specific to a technique need to be developed to interpret and examine structural relationships in a sample set. For simple measurements, this can be a single parameter so that sample evaluation is a straightforward correlation with a single value. This is known as univariate data analysis [3]. Data-intensive, numerical results such as those produced by spectral or diffraction analyses are multivariate and are much more complex in the interpretation of both data structure and the relationships within the sample set. This can often be a rate limiting factor in the information feedback cycle and materials development that is dependent on it.

Due to advantages in source brightness and detector response compared with laboratory based X-ray diffraction, synchrotron radiation powder X-ray diffraction (SR-PXRD) characterization is ideally suited to and has been adapted for HT characterization. The diffraction patterns produced from

SR-PXRD experiments are data intense (about 41,000 variables per sample) and interpretation of vast numbers of patterns from HT experimentation requires the development of new approaches to multivariate data mining and data visualization. This work describes an analytical system driven by the need to examine the structural diversity of and compositional relationships between synthesized metal-organic frameworks (MOFs). MOFs are coordinated lattices that are highly crystalline. They are comprised of metal ions or clusters linked by rigid organic molecules. An example of a crystal structure is depicted in Fig. (**1**), which was generated in Diamond [4] and rendered with Povray. Despite the varied synthetic conditions, the structure was found to be isostructural with $Gd_2(BDC)_3(DEF)_2 \cdot H_2O$ reported by Poulsen *et al.* [5].

A defining character of this new class of materials (i.e. MOFs) is their high porosity and molecular capturing capacity, due to which they hold great promise for numerous applications, including renewable energy for gas storage [6] and catalysis [7]. More recently, MOFs have been identified as potential diagnostic contrast agents in biomedical imaging [8, 9]. Magnetic resonance imaging (MRI) techniques utilize high spin paramagnetic metal complexes as contrast agents due to the influence they impose on the relaxation of the protons of water. Gd-based contrast agents are preferred, because of their wider dynamic range and action as positive rather than negative contrast agents. There has been extensive research geared to developing improved compounds displaying higher relaxivities and lower toxicity. The work presented in this paper discusses techniques developed for the data interpretation step in the HT cycle for the development of Gd-MOFs as effective contrast agents.

The production of MOFs is difficult as there are a large number of variables that govern the formation of these materials and the mechanism of their formation is not understood. Add to this the need to optimize the MOF for a

*Address correspondence to this author at the CSIRO Materials Science and Engineering, Private Bag 33, Clayton South, Victoria 3169, Australia; Tel: +613-9545-2830; E-mail: Deborah.Lau@csiro.au
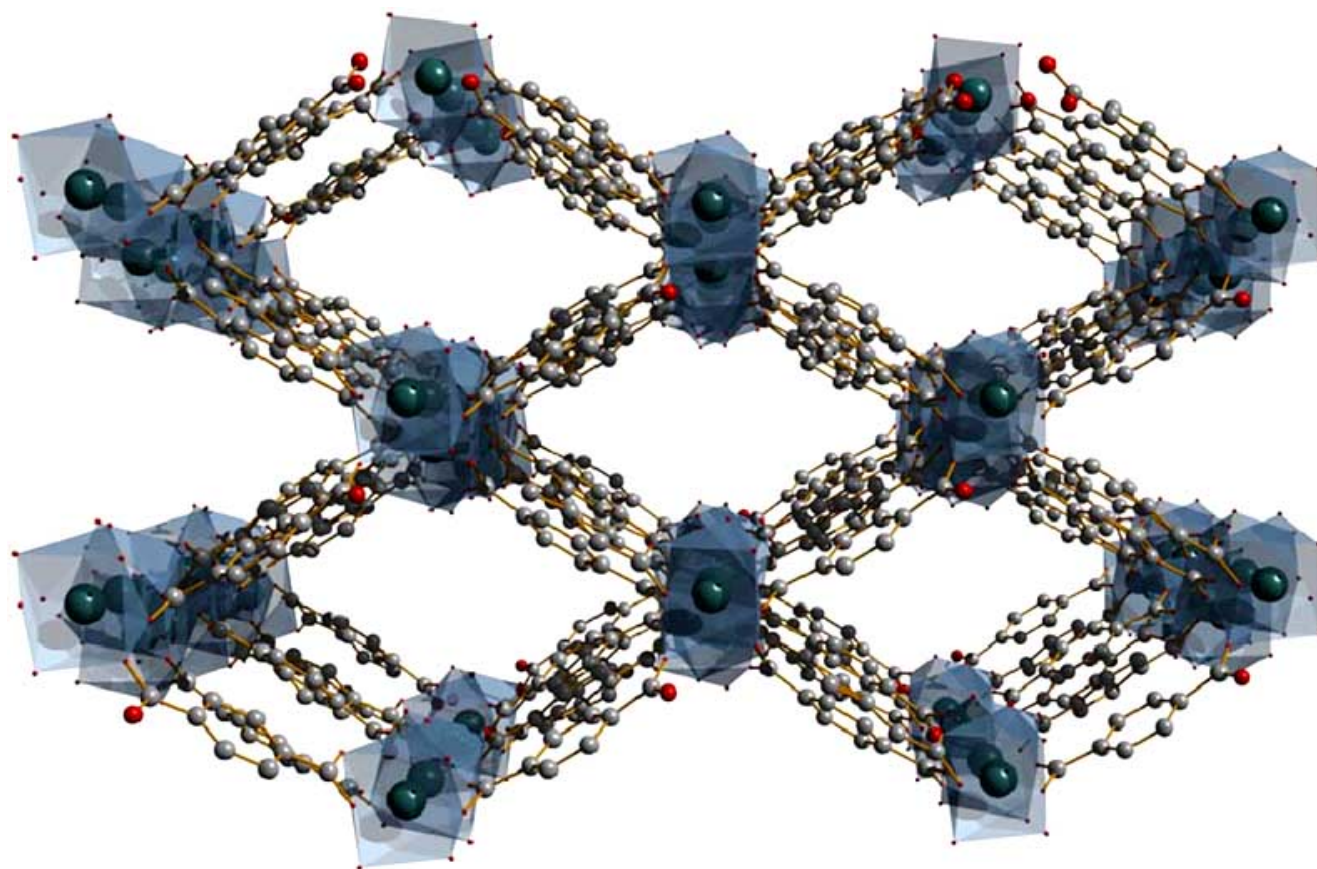
Fig. (1). Crystal structure of metal-organic framework $Gd_2BDC_3DEF_2$ isostructural with previously reported crystal structure [5]. Gd-green, O-red, C-grey, DEF molecules and hydrogen atoms have been removed for clarity and polyhedra included to highlight the metal centres. The structural regularity and large cavity dimensions have unique properties for many applications.

specific property such as MRI contrast and there are numerous combinations which need to be explored. HT synthesis is therefore an excellent approach for the discovery of new application-specific MOF structures [10, 11]. Some of the structurally defining parameters include the metal ion and the metal cluster formed, the molecular composition of the organic ligand, proportions of metal salt and ligand, solvent polarity and solvent mixtures, temperature and pressure of the reaction. In the screening of materials formed in these reactions for new MOF structures, the extent of crystallinity of the product is a key evaluation criterion, since MOFs are known to be highly crystalline regular arrays it is the primary basis for their functional efficiency. It is therefore essential to be able to screen out the amorphous samples in the search for ideal materials. The second stage is to investigate the phase structure, which may then be used to study fundamental composition and structure-performance correlations.

Crystallinity and microstructure are most effectively established using XRD. Laboratory-based XRD can be automated with sample holders for analysis of multiple samples. However, typical collection times for low signal-to-noise ratio (S/N) diffraction patterns are in the order of 30–60 minutes per sample. The development and construction of automated high-throughput instrumentation for synchrotron radiation powder X-ray diffraction (HT-SR-PXRD) has

opened up the possibility for hundreds of materials to be analysed in a relatively short period of time [12].

Faced with the need to identify potential MOFs possessing high crystallinity among the large number of materials generated, and to understand the structure and relationships between large numbers of synthesised MOFs, it became an imperative to develop a new method for the simultaneous examination of hundreds of XRD patterns. Firstly, there was a need to identify the samples that were sufficiently crystalline and separate those from amorphous products. Secondly, it was necessary to determine samples that were compositionally similar to one another, so that:

- Identical, isostructural and polymorphic materials could be quickly identified.

- Material clusters identified could be used to map out materials space in designed experiments.

- Crystallographic structure refinement could be performed on individual materials.

- Subsequent performance testing could be reduced through the elimination of isostructural materials and therefore the results related to individual materials.

In summary, the aim of this work was to develop a systematic method for the management and interpretation of hundreds of data-intense SR-PXRD diffraction patterns, each containing nearly 41,000 variables. The specific problem to

address with this data set was that the overall number of samples was too large to observe any data groupings or structures without assistance. Although there are several multivariate data analysis techniques that can perform ordination and clustering, such as k-means and c-means, there is a degree of contamination and inhomogeneity in some of the products that necessitated step-wise screening, sorting and clustering.

EXPERIMENTAL

Materials

All materials were used as supplied: gadolinium(III) chloride hexahydrate (GdCl$_3$.6H$_2$O, 99%, Sigma), gadolinium(III) nitrate hexahydrate (Gd(NO$_3$).6H$_2$O, 99.9%, Sigma), gadolinium(III) trifluoromethanesulfonate (Gd(OTf)$_3$, 98 %, Aldrich), 1,4-benzenedicarboxylate (BDC, 97%, Hopkin and Williams Ltd), 2,6-napthalenedicarboxylic acid (2,6-NDC, TCI), 1,4-napthalenedicarboxylic acid (1,4-NDC, 94%, Aldrich), 1,3,5-benzenetricarboxylic acid (1,3,5-BTC, Sigma), 1,2,4-benzenetricarboxylic acid (1,2,4-BTC, >99%, Aldrich), 4,4'-biphenyldicarboxylic acid (4,4'-BPhDC, 97%, Aldrich), 2-amino-1,4-benzenedicarboxylic acid (2NH$_2$-BDC, 98% Aldrich), 2,5-dichloro-1,4-benzenedicarboxylic acid (2,5-Cl$_2$BDC, Aldrich), 2-bromo-1,4-benzenedicarboxylic acid (2-BrBDC, 95%, Aldrich), 2,5-OH2BDC, phthalic acid (phthal, >99.9%, Aldrich), homophthalic acid (homophthal, >99%, Fluka), 1,2-benzenedinitrile (BDCN, donated from chemical library), pamoic acid (pamoic, 95%, Sigma), napthalene tetracarboxylic acid (NTC, 95%, Sigma-Aldrich), 1,2,3,4-benzenetetracarboxylic acid (1,2,4,5-BtetC, 96%, Aldrich), 4,5-imidazoledicarboxylic acid (4,5-imDC, 95%, Sigma-
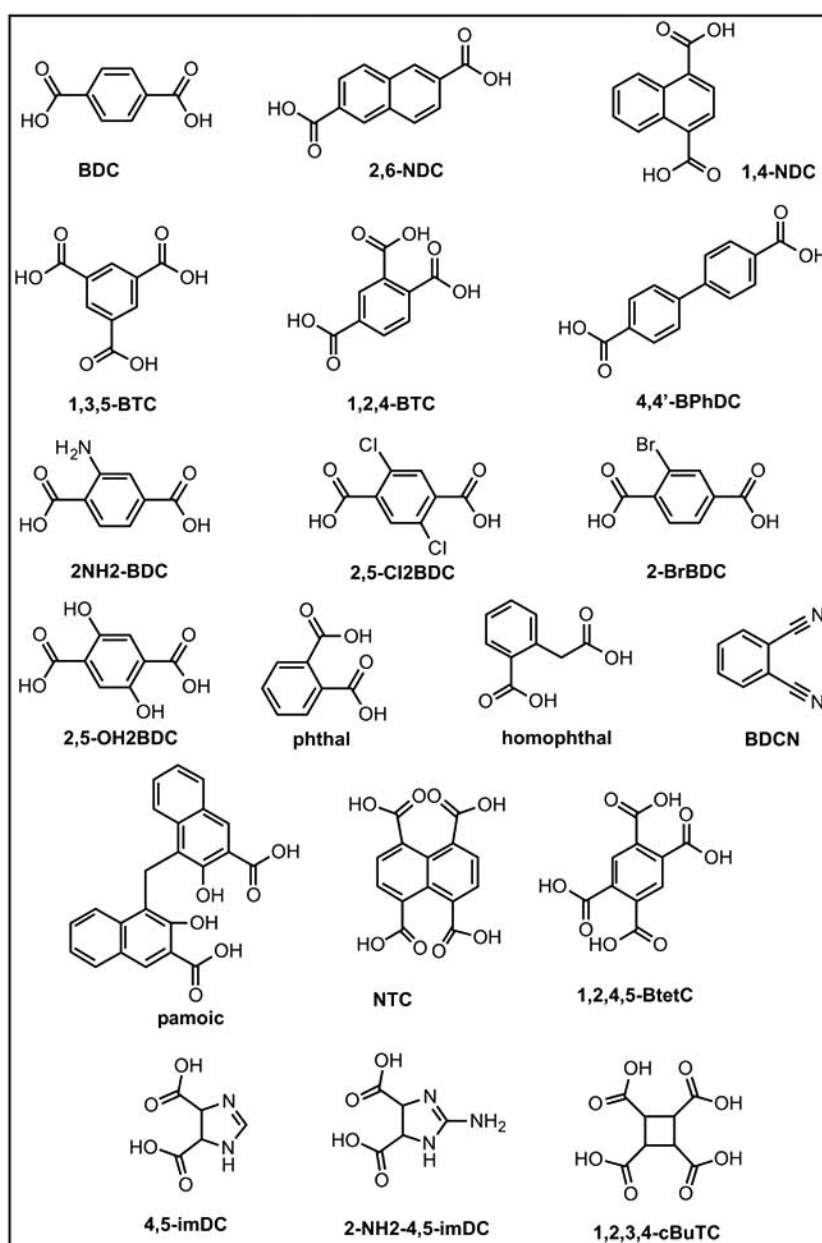

Fig. (2). Chemical structure of the ligands used in the materials investigated in this study.

Aldrich), 2-amino-4,5-imidazoledicarbonitrile (2-NH$_2$-4,5-imDC, 97%, Aldrich), 1,2,3,4-cyclobutanetetracarboxylic acid (1,2,3,4-cBuTC, 98%, Aldrich). The chemical structures of the compounds are shown in Fig. (**2**).

Generation of a Library of Gd Hybrid Materials

The method for the synthesis of the materials was broadly based on the method of Kaye *et al.* [13]. The metal precursors used were GdCl$_3$.6H$_2$O, Gd(NO$_3$)$_3$.6H$_2$O and Gd(OTf)$_3$. The ligands investigated and their abbreviations are illustrated in Fig. (**2**), and included benzenedicarboxylic acid (BDC), 1,4-napthalenedicarboxylic acid, 2,6-napthalenedicarboxylic acid and 1,3,5-benzenetricarboxylic acid, amongst others. Solvents used included dimethylform-amide (DMF) and diethylformamide (DEF), with mixed solvent systems also employed including DMF or DEF with minor amounts of methanol, water and DMSO. All materials were used as supplied. MOFs were generated by one of three methods:

- *Method 1:* Stock solutions of the metal precursors and ligands were prepared in DEF. Varying amounts of these stock solutions were pipetted into 4 mL glass vials, fitted with a Teflon-lined screw cap and heated in an aluminium dry block at a fixed temperature.

- *Method 2:* The required amount of ligand was weighed and placed into a 4 mL glass vial with a Teflon-lined screw cap. Stock solutions of the metal precursor in DMF or DEF were added, followed by the solvent additive if required. The vials were sealed and heated in an aluminium dry block at a fixed temperature.

- *Method 3:* The required amount of ligand and metal precursor was weighed and placed into a 4 mL glass vial with a Teflon-lined screw cap. DMF or DEF were added, followed by the solvent additive if required. The vials were sealed and heated in an aluminium dry block at a fixed temperature.

All reactions that were noted to form crystalline or insoluble solids were sampled. The solid was dabbed dry on filter paper and loaded into the powder diffraction sample cassette without further preparation, i.e. the samples were either powders or a mixture of crystals of varying size and shape from sub-micron to as large as 1 mm in length [12]. Numerous batches were prepared and examined, but the resultant sample set presented in this article comprised 129 samples.

High Throughput Characterisation Using SR-PXRD

Diffraction data were collected under ambient conditions (29°C) using the procedure described by Kennedy *et al.* [12]. Samples were mounted in a 114 multi-well sample cassette and examined in transmission mode at the PXRD beamline, 10-BM, of the Australian Synchrotron in Victoria using the Mythen microstrip detector. The beam was operated at 12.39837 keV or 0.9998 Å wavelength [14]. Data were collected over the 2-theta range 2–82°. Diffraction data were collected in two frames, shifted by 0.5° in order to cover the small gaps in the detector (~0.2°), which occur every 5°. Each collection took approximately 1 minute. Data were

merged using a custom written macro and converted to equally spaced data using locally developed in-house software.

Data Analysis Part 1: Interpreting and Examining Samples for Crystallinity

PCA analysis was performed using "The Unscrambler" V9.8, (CAMO ASA, Oslo, Norway). Spectra were pre-treated by the selection of a data subset of 2–40° in 2-theta. Normalization over the range 0–1, baseline correction and 4-point data averaging were performed.

Data Analysis Part 2: Determining Structural Relationships

HCA was performed using the freeware package "KNIME" V2.0.3 using a Euclidian distance function and average linkage type and an agglomerative clustering algorithm. Low order branching structure groups were identified and the constituent diffraction patterns overlaid. Adjacent individual patterns were compared to the group and included or excluded from the group on the basis of comparative reflections. Contaminated samples were observed to contain all the reflections expected in a similar sample, but additional reflections from other components. Contaminated samples were considered part of the group, but crystallographically different materials were identified as belonging to a different group.

RESULTS AND DISCUSSION

Evaluation of Ordination and Clustering Approaches

The interpretation and mining of spectroscopic data is not uncommon, however PXRD data structure is unlike most spectroscopic data, having an extremely large number of variables with relatively few, precisely positioned, extremely high intensity peaks (narrow full width-half max (FWHM)), measured in counts, and shown in Fig. (**3**). This is in many ways similar to mass spectroscopy spectra. In diffraction data obtained from MOFs, the signal intensity is not consistent from sample to sample due to the availability (crystalline size) of material and differences in packing the sample holder, leading to preferred orientation effects. In addition, there may be additional peaks due to varying degrees of different impurities (unreacted ligand, trapped solvents and minor by-products), therefore simply nominating a selected subset of variables (2-theta angles) provides inconsistent information for a multivariate data analysis (MVA). The raw data contained 40,926 data points and processing the entire pattern was considered computationally inefficient and too large to be managed by many proprietary MVA software programs and desktop computers. Examination of the individual peak widths in the PXRD patterns showed that no data information would be lost by averaging the data over 4 points. In addition, a negligible number of peaks existed above 40° in 2-theta due to large unit cell size, and high angle peaks are less important from a structural refinement perspective. Therefore, truncating the data range did not result in any loss of information from the samples. The combined treatment of averaging and data truncation reduced the number of

variables to approximately 5000 with no observable loss in the peak resolution or S/N.
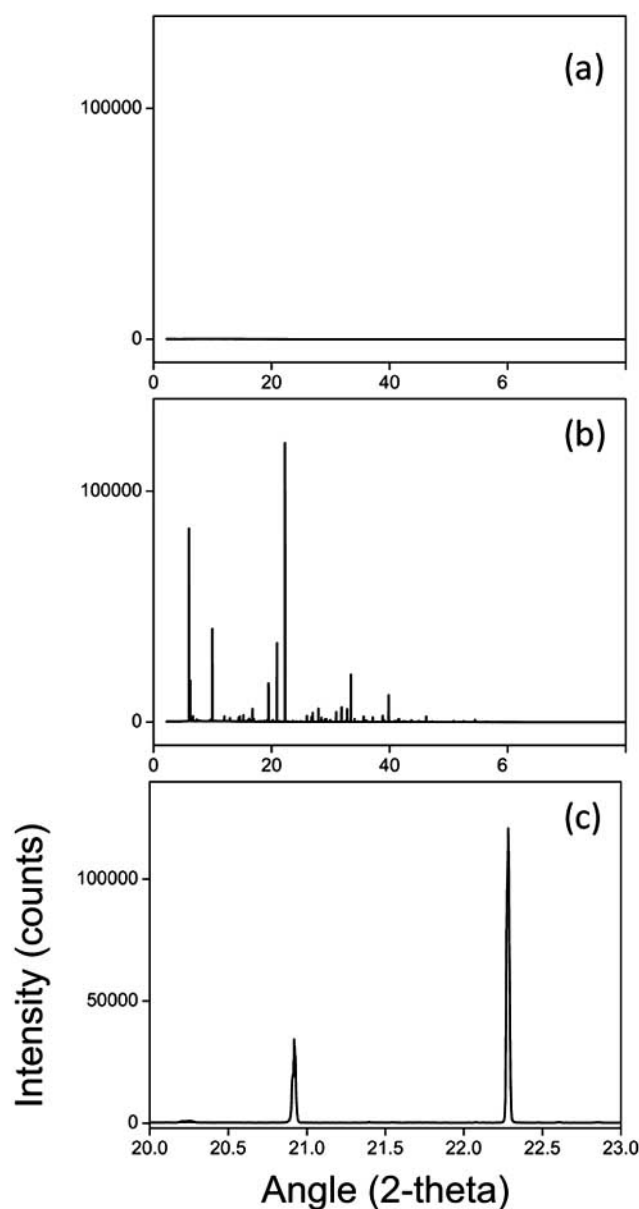

Fig. (3). Example X-ray diffraction patterns of (**a**) amorphous, (**b**) crystalline and (**c**) magnified view of a crystalline MOF synchrotron radiation X-ray diffraction pattern illustrating the flat baseline and narrow FWHM of the reflection peaks.

Generation and Interpretation of PCA Models of MOFs

The first and primary question regarding the quality of MOF products is whether they are crystalline or amorphous. MOFs are defined as ordered 3D arrays of metal clusters bridged by rigid organic ligands. Any amorphous material generated is therefore not a MOF due to a lack of organized structure, and it is necessary to identify the reaction components and conditions that are unsuitable. It also provides an opportunity to eliminate them from further analysis or performance characterization which would otherwise be a waste of resources.

Most of the crystalline MOF materials produced through the HT synthesis approach were completely novel. As the structures were previously unknown or unpredictable, the selection of individual peak positions (diffraction angles) to generate a smaller sample matrix for PCA was not possible. Truncation and averaging were found to provide suitable data reduction.

In the visual examination of SR-XRD patterns, amorphous materials do not exhibit sharp, well-defined peaks. They are comprised of a general, near-zero intensity continuum as shown in Fig. (**3a**). Crystalline samples contain relatively few non-zero (or near-zero) variables, which can be seen in Fig. (**3b, c**). A standard PCA ordination that relies on a covariance matrix was found to be inefficient in sufficiently separating an amorphous material from a highly crystalline one because of the data structure described in the previous section and shown in Fig. (**3**). The presence or absence of peaks in the SR-PXRD pattern means too few variables were solely attributable to "amorphousness" as a property represented in the diffraction pattern.

Rescaling by normalizing the sample intensity data for each sample over the range 0-1 meant amorphous samples were transformed to have relatively high values over the entire collection angle range. This manifests as a significant variance, which was captured in PC1 using a PCA transformation and ordination of the first two principal components (Fig. **4**). The first PC accounted for 81% of the sample variance while the second PC accounted for 5% and each subsequent one less than 1%. The first PC was considered attributable to the inherent degree of crystallinity of the samples. By ranking the samples sequentially in order of the first PC, amorphous samples were then able to be differentiated from crystalline ones. This sample set of 129 samples fell into three generalized groupings: amorphous, semi-crystalline and highly crystalline. The majority (114) were crystalline, 11 semi-crystalline and 4 amorphous. The 114 crystalline samples were selected for further examination.

Generation and Interpretation of HCA Models for MOFs

The second key question was to understand the relationship within the samples and the observation of any groups or clusters that may have formed – which products were similar and which were dissimilar and how close were they to each other in structure? Many clustering algorithms such as k-means and c-means provide a numerical output in which it is difficult to observe the overall data structure. They may be plotted in three dimensions, but for large numbers of groups with a small number of members, the visualization of associations becomes very difficult.

HCA provides a visual overview of the sample relationships and their connections simultaneously and was considered suitable for this data set. The second advantage of HCA is that distance measures may be examined at the same time as the grouping structure. Since the diffraction patterns sometimes differed only on the basis of containing a small amount of unreacted material (excess ligand or metal salt) or another component, a subjective grouping could be manually overlaid guided by the observable grouping structure, a feature not possible in the commercial program PolySNAP
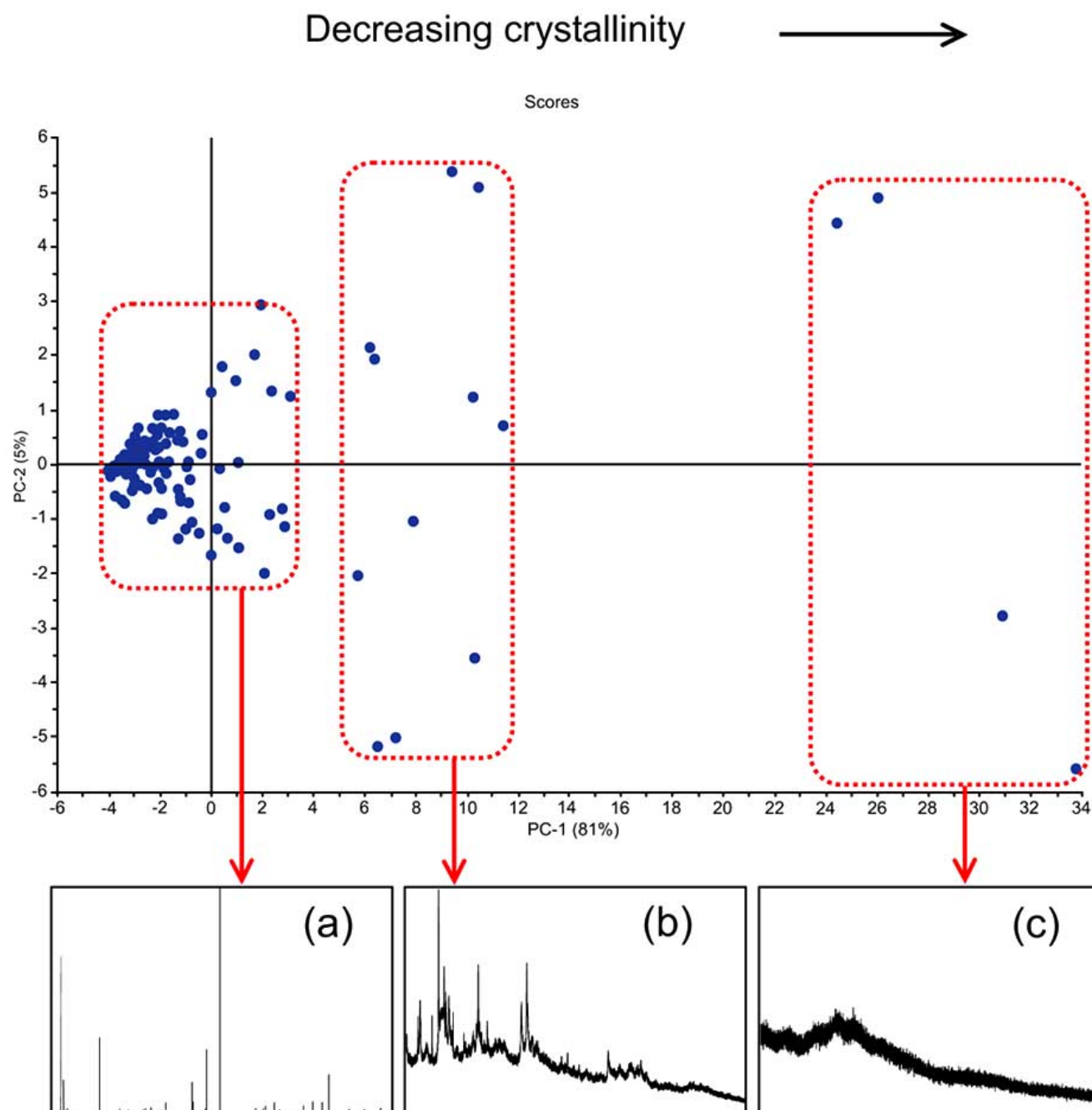
Fig. (4). The PCA scores plot of PC1 versus PC2. PC1 describes the gradient of crystallinity amongst the MOF samples. Insets show examples of normalised XRD patterns from 2–40° in 2-theta of (**a**) crystalline, (**b**) semi-crystalline and (**c**) amorphous phases.

distributed by Bruker AXS [15]. The procedure was performed by examining, selecting and overlaying individual diffraction patterns within each dendrogram agglomeration to ascertain their grouping placement. Suitably identified groups were then manually represented using coloured boxes around the groups, as shown in the dendrogram in Fig. (**5**). Unique crystallographic compositions were ungrouped and existed as clusters containing one member.

It can be seen that there was no fixed or single distance measure that satisfied all the groupings, but the dendrogram allowed the grouping of similar samples into "suggested" groups which were confirmed or denied manually by examining overlays of individual XRD patterns that were clustered close by. There were 25 individual MOF types (a cluster with 1 member) and 25 clusters with between 2 and 4 structurally similar members. Fig. (**6**) illustrates the success of the methodology for the identification of identical materials, with the successful aggregation of similar patterns as in cluster 1. These materials were all made from various combinations of Gd precursors and the ligand 4,5-imDC. Their common structure is observable as similarity in the main and low angle reflections in their diffraction patterns, but importantly the similarity in structure was not able to be observed otherwise amongst the vast number of scans.
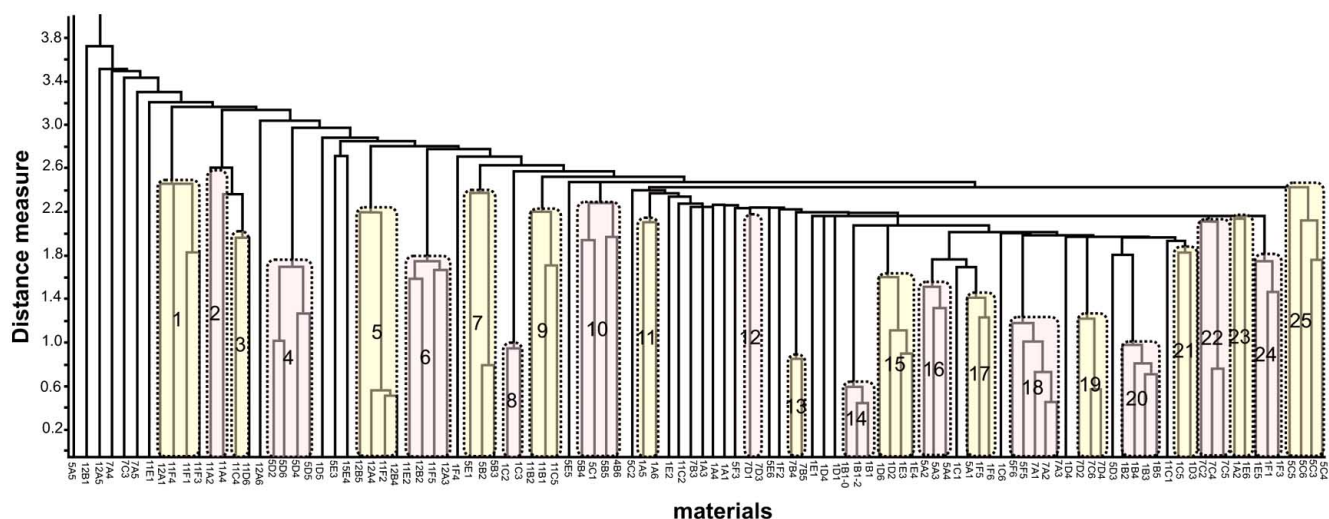
Fig. (5). HCA dendrogram of crystalline Gd-MOF samples. Identified groupings are indicated by the orange and pink boxes around sample clusters. Unique Gd-MOF structures are ungrouped. The clusters are numbered to assist recognition.

An overview of the PLUXter method is shown in Fig. (**7**), which summarizes the sequence of steps described here. The final process of MOF discovery, after the identification of individual and unique crystalline phases using PLUXter, was to perform crystallographic structure determination and refinement on representative isolated large, white cubic crystals using the macromolecular beamline at the Australian Synchrotron.
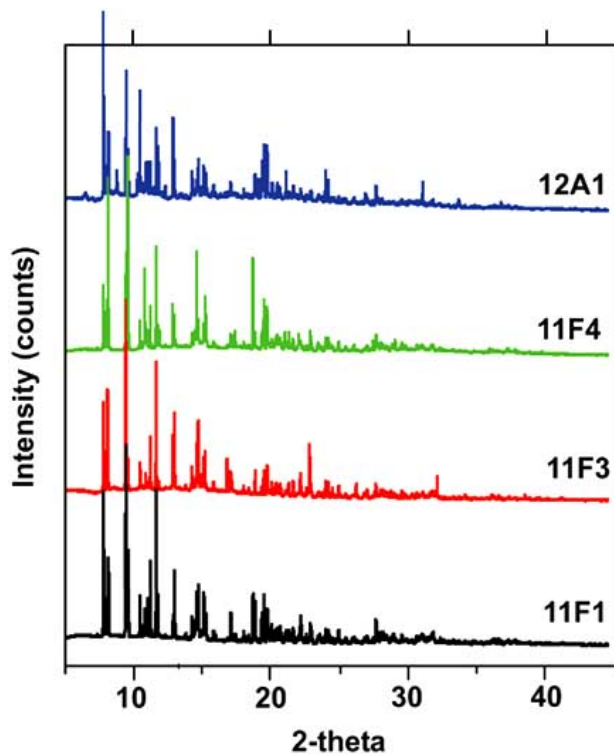
property studies from 129 to 50 samples. A 62% reduction in sample processing has obvious implications for the rate of materials development and discovery, thereby accelerating the HT cycle in general, and more specifically the development of Gd-MOFs for effective MRI contrast agents.
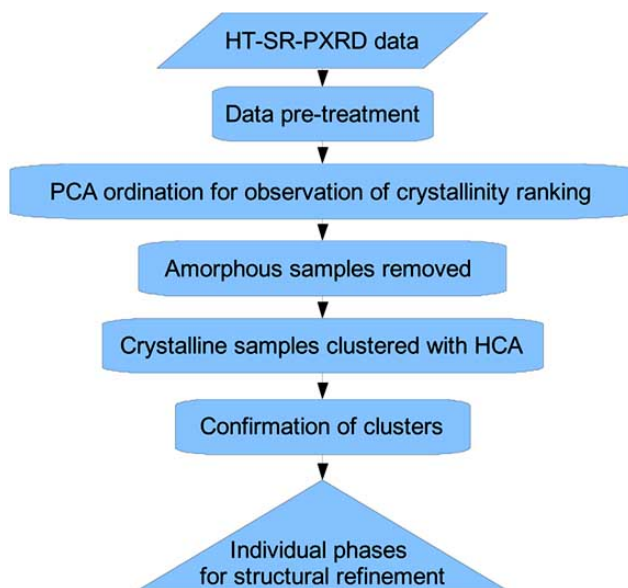

Fig. (6). Comparison of diffraction patterns in Cluster 1. The diffraction patterns in this plot are not binned or normalized, however they have been offset for clarity.

This method has drastically reduced the time and resources required for structural refinement and further

Fig. (7). Outline of the PLUXter methodological sequence for sorting and clustering HT-SR-PXRD patterns.

CONCLUSIONS

The development of HT techniques for the preparation and characterisation of materials has necessitated the development of associated data handling and interpretation strategies. For HT-SR-PXRD data, the number of samples is often in the hundreds, and the number of variables (diffraction angles) can be in the order of tens of thousands. For such high volume data, standard methods of examination

are inadequate and a manually guided approach to cluster analysis is required.

This work has developed a systematic method for screening and data mining. We have shown that after pre-treatment for data reduction, PCA allows materials to be ranked in order of crystallinity, so that undesirable amorphous materials may be identified and separated. We have then shown that structural grouping within and between samples can be observed using HCA. Using a single distance measure was found to be inadequate in separating the clusters because it was not applicable to all samples, and an interactive grouping approach was found to be necessary for the assignment of acceptable clusters.

Although the PLUXter method was developed to interpret high volume data for the synthesis and screening of materials for MOFs, it is generally applicable to other high volume XRD applications or similarly structured spectra where compositional structural relationships are of interest.

ACKNOWLEDGEMENTS

REFERENCES

[1] Maier, W.F.; Stöwe, K.; Sieg, S. Combinatorial and high-throughput materials science. *Angewandte Chemie*, **2007**, *46*, 6016-6067.

[2] Yang, Y.; Wang, X. Information flow modeling and data mining in high-throughput discovery of functional nanomaterials. *Comput. Aided Chem. Engin.*, **2009**, *26*, 135-140.

[3] Potyrailo, R.A. Analytical spectroscopic tools for high-throughput screening of combinatorial materials libraries. *Trends Analyt. Chem.*, **2003**, *22*, 374-384.

[4] Brandenburg, K.; Putz, H.; Berndt, M. *Diamond*; Crystal Impact GbR: Bonn, Germany, **1999**.

[5] Poulsen, R.D.; Overgaard, J.; Chevallier, M.A.; Clausen, H.F.; Iversen, B.B. A gadolinium-based metal-organic framework, poly[[tris(mu(4)-benzene-1,4-dicarboxylato)bis(mu(2)-N,N-diethylformamid e)digadolinium(III)] monohydrate]. *Acta. Crystallogr. Sect. E Struct. Rep. Online,* **2005**, *61*, m1337-m1339.

[6] Murray, L.J.; Dinca, M.; Long, J.R. Hydrogen storage in metal-organic frameworks. *Chem. Soc. Rev.,* **2009**, *38*(5), 1294-1314.

[7] Lee, J.; Farha, O.K.; Roberts, J.; Scheidt, K.A.; Nguyen, S.T.; Hupp, J.T. Metal-organic framework materials as catalysts. *Chem. Soc. Rev.,* **2009**, *38*(5), 1450-1459.

[8] Rieter, W.J.; Taylor, K.M.L.; Lin, W.B. Surface modification and functionalization of nanoscale metal-organic frameworks for controlled release and luminescence sensing. *J. Am. Chem. Soc.*, **2007**, *129*(32), 9852-9853.

[9] Taylor, K.M.L.; Rieter, W.J.; Lin, W.B. Manganese-based nanoscale metal-organic frameworks for magnetic resonance imaging. *J. Am. Chem. Soc.*, **2008**, *130*(44), 14358-14359.

[10] Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O.M. High-throughput synthesis of zeolitic imidazolate frameworks and application to CO_2 capture. *Science*, **2008**, *319*(5865), 939-943.

[11] Bauer, S.; Serre, C.; Devic, T.; Horcajada, P.; Marrot, J.; Ferey, G.; Stock, N. High-throughput assisted rationalization of the formation of metal organic frameworks in the iron(III) aminoterephthalate solvothermal system. *Inorg. Chem.,* **2008**, *47*(17), 7568-7576.

[12] Kennedy, D.F.; Hill, M.R.; Wright, N.M.; Wallwork, K.S.; Kimpton, J.A.; Gu, Q.F.; Hay, D.G. A two-dimensional transmission XRD stage for laboratory and synchrotron high throughput analysis. *J. Appl. Crystallogr.,* **2010**, in review.

[13] Kaye, S.S.; Dailly, A.; Yaghi, O.M.; Long, J.R. Impact of preparation and handling on the hydrogen storage properties of $Zn_4O(1,4-benzenedicarboxylate)_3$ (MOF-5), *J. Am. Chem. Soc.*, **2007**, *129*(46), 14176-14177.

[14] Wallwork, K.S.; Kennedy, B.J.; Wang, D. In: Proceedings of the 9th International Conference on Synchrotron Radiation Instrumentation (SRI 2006), Daegu, South Korea, May 28 to June 2, 2006; Choi, J.-Y.; Rah, S., Eds.; American Institute of Physics, Maryland, USA, **2006**; pp. 879-882.

[15] Barr, G.; Dong, W.; Gilmore, C.J. High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis. *J. Appl. Crystallogr.,* **2004**, *37*(2), 243-252.