Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Structure maps for $A^{I}_{4}A^{II}_{6}(BO_4)_6X_2$ apatite compounds *via* data mining

Prasanna V. Balachandran and Krishna Rajan*

Department of Materials Science and Engineering and Institute of Combinatorial Discovery, Iowa State University, Ames, IA 50011, USA

Correspondence e-mail: krajan@iastate.edu

Received 27 September 2011
Accepted 15 December 2011

This paper describes a method to identify key crystallographic parameters that can serve as strong classifiers of crystal chemistries and hence define new structure maps. The selection of this pair of key parameters from a large set of potential classifiers is accomplished through a linear data-dimensionality reduction method. A multivariate data set of known $A^{I}_{4}A^{II}_{6}(BO_4)_6X_2$ apatites is used as the basis for the study where each $A^{I}_{4}A^{II}_{6}(BO_4)_6X_2$ compound is represented as a 29-dimensional vector, where the vector components are discrete scalar descriptors of electronic and crystal structure attributes. A new structure map, defined using the two distortion angles α_{AII} (rotation angle of $A^{II}-A^{II}-A^{II}$ triangular units) and $\psi^{AIz\,=\,0}_{AI-O1}$ (angle the $A^{I}-O1$ bond makes with the c axis when $z = 0$ for the A^{I} site), is shown to classify apatite crystal chemistries based on site occupancy on the A, B and X sites. The classification is accomplished using a K-means clustering analysis.

1. Structure maps and structure classification

Structure maps have played an important role as a useful *a priori* tool for establishing structure–chemistry relationships in a two-dimensional way (Mooser & Pearson, 1959; Zunger, 1980; Villars, 1984; Pettifor, 1986; Rabe *et al.*, 1992; Kleinke & Harbrecht, 2000; Hauck & Mika, 2002; Zhang *et al.*, 2007; Li *et al.*, 2008). Basically, the structure map approach involves visualizing the data of known compounds with known crystal structures in a two-dimensional space using two scalar descriptors (normally heuristically chosen) that are associated with physical/chemical properties, crystal chemistry or electronic structure. The objective is to map out the relative geometric position of each structure type from which one tries to discern qualitatively if there are strong associations of certain structure types to certain bivariate combinations of parameters.

From the data-mining perspective, a structure map is a data classification tool. Data classification can be accomplished in two ways (Han & Kamber, 2006): supervised and unsupervised learning. In the case of *supervised learning* we have a tuple $X = (x_1, x_2, \ldots x_m)$ depicting m independent measurements (chemical compositions) that is represented by an n-dimensional attribute vector $A = (A_1, A_2, \ldots, A_n)$. Each attribute (discrete scalar descriptors of crystal and electronic structure) in A represents a feature of X. Each tuple, X, is assumed to belong to a predefined class as determined by another attribute called the class label attribute (crystal structure type). The objective of supervised learning is to map a function $y = f(X)$ that can predict the associated class label y of a given tuple in X. Typically this mapping is represented in the form of classification rules, decision trees or mathematical

Table 1
List of 29 discrete descriptors of electronic and crystal structure parameters.

Descriptor	Brief description
a (Å)	Lattice constant of the hexagonal unit cell
c (Å)	Lattice constant of the hexagonal unit cell
c/a	Variable axial ratio (no unit)
r_{AI} (Å)	Shannon's ionic radii of A^{I}-site ion (nine-coordination)
r_{B} (Å)	Shannon's ionic radii of B-site ion
r_{AII} (Å)	Shannon's ionic radii of A^{II}-site ion (seven-coordination for F^- and eight-coordination for Cl^- and Br^-; Đordević *et al.*, 2008)
r_{X} (Å)	Shannon's ionic radii of X-site ion
Av CR (Å)	Average crystal radius = $[(r_{AI}x4) + (r_{AII}x6) + (r_{B}x6) + (r_{O}x24) + r_{X}x2)]/42$
$A_{EN} - O_{EN}$	Electronegativity difference A atom and O atom
$B_{EN} - O_{EN}$	Electronegativity difference B atom and O atom
$A_{EN} - X_{EN}$	Electronegativity difference A atom at A^{II} site and X atom
$A_{EN} - B_{EN}$	Electronegativity difference A atom at A^{I} site and B atom
A^{I}—O1 (Å)	Distance between A^{I} and O1 atom
A^{I}—O1$^{AIz=0}$ (Å)	Distance between A^{I} and O1 atom with the constraint $z = 0$ at A^{I}
Δ_{AI-O} (Å)	Difference in the lengths A^{I}—O1 and A^{I}—O2
$\Delta_{AI-O}^{AIz=0}$ (Å)	Difference in the lengths A^{I}—O1 and A^{I}—O2 with the constraint $z = 0$ at A^{I}
ψ_{AI-O} (°)	The angle that the A^{I}—O1 bond makes with respect to c
$\psi_{AI-O}^{AIz=0}$ (°)	The angle that the A^{I}—O1 bond makes with respect to c with the constraint $z = 0$ at A^{I}
δ_{AI} (°)	Counter-rotation angle of $A^{I}O_6$ structural unit
φ_{AI} (°)	Metaprism twist angle ($\pi/3 - 2\delta_{AI}$)
α_{AI} (°)	Orientation of $A^{I}O_6$ unit with respect to a
$\langle B$—O$\rangle$ (Å)	Average B—O bond length
$\langle\tau_{O-B-O}\rangle$ (°)	Average O—B—O bond-bending angle
ρ_{AII} (Å)	A^{II}—A^{II} triangular side length
A^{II}—X (Å)	Distance between A^{II} and X atom
α_{AII} (°)	Orientation of A^{II}—A^{II}—A^{II} triangles with respect to a
A^{II}—O3 (Å)	Distance between A^{II} and O3 atom
$\Phi_{O3-AII-O3}$ (°)	O3—A^{II}—O3 angle
E_{total} (eV)	Total energy calculated from *ab initio* calculations

formulae. The key difference between a supervised learning and *unsupervised learning* is that we have no *a priori* information on the predefined class of each tuple. The objective of unsupervised learning is to assign a class to tuple in X without having a known target class. This approach provides value in finding correlations and similarities in data sets and the mapping is typically represented in the form of data clustering. Traditional structure maps belong to the supervised learning scheme where the crystal structure information (predefined class label) for every chemical composition (X) is ascertained beforehand. In this paper we develop a data-mining approach using unsupervised learning to discover the best classifiers for constructing structure maps for $A^{I}_4A^{II}_6(BO_4)_6X_2$ apatite-type compounds without any *a priori* assumptions. One of the fundamental challenges in chemical crystallography is to discover the key structural descriptors such as specific interatomic distances and angles that reflect the structural systematics in complex crystal chemistries and this work establishes a methodology for extracting such descriptors in a statistically robust yet physically meaningful manner. In our group we have applied data mining to explore a variety of questions in crystal chemistry (Gadzuric *et al.*, 2006; Rajagopalan & Rajan, 2007; George *et al.*, 2009; Suh & Rajan, 2005, 2009; Aourag *et al.*, 2010; Broderick *et al.*, 2010; Rajan, 2010; Balachandran *et al.*, 2011). Apatite-type compounds were chosen because of their fundamental and technological significance and we are taking advantage of a wealth of experimentally and computational based studies on the crystal chemistry of this class of compounds (Elliott, 1994; White & ZhiLi, 2003; White *et al.*, 2005; Mercier *et al.*, 2005, 2007; Kim *et al.*, 2005; Sugiyama, 2007; Pramana *et al.*, 2008; Baikie *et al.*, 2007, 2010).

The rest of the article is organized as follows: in §2 we describe the crystal structure of apatites and the data set used in this study, in §3 we briefly give an account of the mathematics of principal component analysis (PCA) as an unsupervised learning technique, in §4 the results are discussed, in §5 the new structure map is defined and in §6 we conclude this paper.

2. Apatite crystal chemistry

Chemically, apatites are described by the general formula $A^{I}_4A^{II}_6(BO_4)_6X_2$, where A^{I} and A^{II} are distinct crystallographic sites that usually accommodate larger monovalent (Na^+, K^+ *etc.*), divalent (Ca^{2+}, Sr^{2+}, Ba^{2+}, Pb^{2+} *etc.*) or trivalent (Y^{3+}, La^{3+}, Ce^{3+}, Sm^{3+} *etc.*) cations; B sites are filled by smaller metals and metalloids (P^{5+}, As^{5+}, V^{5+}, Si^{4+} *etc.*) and the X site is filled by halide, hydroxide or oxide anions (F^-, Cl^-, Br^-, I^-, OH^-, O^{2-}). The crystal structure of a typical $Ca_{10}(PO_4)_6F_2$ fluorapatite belonging to the space group $P6_3/m$ is shown in Fig. 1. The unit cell consists of 42 atoms and the complex structure can be decomposed into three basic building (or structural) units based on the principles of coordination polyhedra (Hughes & Rakovan, 2002): $A^{I}O_6$ metaprism, BO_4 tetrahedra and $A^{II}O_6X_{1,2}$ polyhedron. This process aids in the development of an enumeration scheme where a crystal structure is described using numerous discrete descriptors of electronic and crystal structure parameters such as ionic radii, electronegativity differences, bond length, bond angles, lattice constants and total energy.

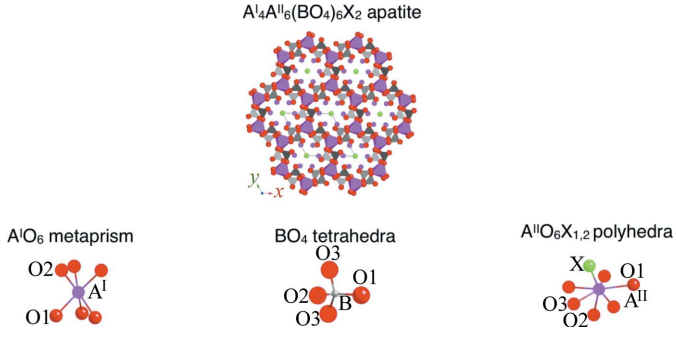


Figure 1
Crystal structure of a typical $P6_3/m$ hexagonal $Ca^{I}_4Ca^{II}_6(PO_4)_6F_2$ apatite with the atoms projected along the (001) plane shown. The complex crystal structure is decomposed into three basic structural units: $A^{I}O_6$ metaprism, BO_4 tetrahedra and $A^{II}O_6X_{1,2}$ polyhedra.

Table 2
List of compounds used in this work.

All compound chemistries are taken from the work of Mercier *et al.* (2005).

$Ba_{10}(PO_4)_6Br_2$	$Pb_{10}(AsO_4)_6Cl_2$
$Ba_{10}(PO_4)_6Cl_2$	$Pb_{10}(PO_4)_6F_2$
$Sr_{10}(PO_4)_6Br_2$	$Sr_{10}(PO_4)_6F_2$
$Ca_{10}(PO_4)_6Br_2$	$Sr_{10}(VO_4)_6F_2$
$Ca_{10}(PO_4)_6F_2$	$Cd_{10}(PO_4)_6Cl_2$
$Ca_{10}(CrO_4)_6Cl_2$	$Cd_{10}(CrO_4)_6Cl_2$
$Ca_{10}(VO_4)_6Cl_2$	$Cd_{10}(VO_4)_6Cl_2$
$Ba_{10}(PO_4)_6F_2$	$Pb_{10}(PO_4)_6Cl_2$
$Ba_{10}(MnO_4)_6F_2$	$Hg_5(PO_4)_3F$
$Sr_{10}(VO_4)_6F_2$	$Hg_5(PO_4)_3Cl$
$Pb_{10}(PO_4)_6Br_2$	$Zn_5(PO_4)_3F$
$Pb_{10}(CrO_4)_6Cl_2$	$Zn_5(PO_4)_3Cl$
$Pb_{10}(VO_4)_6Cl_2$	

The raw data for this study comes from the detailed geometrical parameterization scheme developed by Mercier *et al.* (2005) where they quantify the bond distortions in the $P6_3/m$ apatite compounds using 15 algebraically independent bond lengths and angles. Our objective is to screen these discrete bond distortion descriptors to identify a subset of dominant descriptors that could be chosen as the coordinates for constructing new structure maps. The geometrical parameters carry structural information and understanding their trends allows for the systematic tracking of structural modifications (Mercier *et al.*, 2005, 2007). In addition, we also included the ionic radii and electronegativity data taken from the work of Shannon (1976) and Pauling (1960), respectively. The rationale behind choosing ionic radii and electronegativity differences to describe the chemistry and electronic structure was dictated by past theoretical and experimental work in the literature (Suzuki *et al.*, 1984; Flora *et al.*, 2004; Matsunaga *et al.*, 2008). In Table 1 all 29 attributes of the apatite crystal structures used in this study are briefly defined.

A data set of 25 compounds with the stoichiometry $A_4^{\rm I}A_6^{\rm II}(BO_4)_6X_2$ was developed. In the data set the following chemical elements were included: Ca, Ba, Sr, Pb, Hg, Zn and Cd in the *A* site, P, As, V, Cr and Mn in the *B* site and F, Cl and Br in the *X* site. A table containing the list of compound chemistries considered in this work is given in Table 2.

3. Data dimensionality reduction with principal component analysis

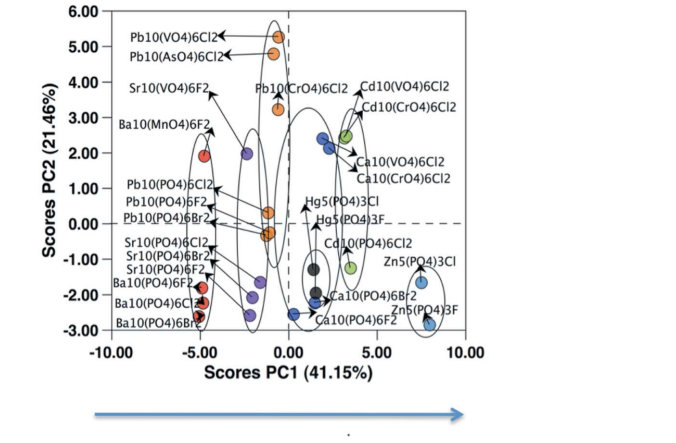
Principal component analysis (PCA) is one of the well known unsupervised linear manifold learning methods. The mathematical background that is reviewed here follows the treatment of several published literatures (Jolliffe, 2002; Morris, 2004; Ringnér, 2008; Rajan *et al.*, 2009). Manifold learning or dimensionality reduction is concerned with projecting the high-dimensional multivariate data into a new low-dimensional subspace with the loss of minimal information (Rajan, 2010). The central idea of PCA was developed by relying on the fact that the dataset consists of a large number of intercorrelated descriptors. The purpose here is to reduce the dimensionality of a data set, while retaining maximum variability in the data. This is achieved by transforming the original set of variables to a new set of derived variables, called the principal components (PCs), which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002). The first PC accounts for the maximum variance (highest eigenvalue) in the dataset. The second PC is orthogonal to the first and accounts for most of the remaining variance. Thus, the *m*th PC is orthogonal to all others and has the *m*th largest variance in the set of PCs. Once all the PCs have been calculated, only those with eigenvalues above a critical level are retained. Each PC is a linear combination of the weighted contribution of all attributes and the magnitude of the weight determines the relative impact of each descriptor in affecting the PC. From knowledge of the calculated PCs one can easily determine the relative importance of each descriptor and the correlation between any two descriptors. Information pertaining to the relative importance of descriptors will be helpful in identifying the dominant descriptors and the correlation information will be helpful in screening the dominant descriptors to avoid choosing redundant descriptors. Thus, from the evaluation of the sole variance–covariance information or correlation information (without including any information about the predefined class label – crystal structure type), we can screen a large library of descriptors and select only a dominant few as the potential coordinates for defining the structure map.

The main computational steps of PCA are summarized here. We have a large library of discrete attributes $A = (A_1, A_2, ..., A_k)$ that represent the features of apatite crystal structure. The data is assumed to follow a multivariate normal distribution. The first step in our analysis involves preprocessing A by mean centering and standardization, and denoting the preprocessed vector as $\hat{A}$. This preprocessing step transforms each attribute in A to have a zero mean and unit variance. The standardized data matrix becomes our data set and it is mathematically described in the Euclidean space as $\hat{\mathbf{A}} = \left(\hat{A}_1, \hat{A}_2, ..., \hat{A}_k\right) \in \Re^{25\times 29}$, where the integers 25 and 29 indicate the number of apatite compounds and the number of electronic and crystal structure parameters considered in this study. Let $\hat{A}_j = \left\{a_{ij}\right\}$, where $\hat{A}_j$ corresponds to the column vector and a_{ij} represent the discrete features of each chemical composition (index i) of the attribute $\hat{A}_j$. Let **S** be the sample covariance matrix of $\hat{\mathbf{A}}$. The next step involves performing eigenanalysis on the sample covariance matrix **S**. The eigenanalysis results in the generation of eigenvectors and eigenvalues of the matrix **S**. The eigenvalues and the corresponding eigenvectors are ordered in descending order. Only those eigenvectors are retained whose eigenvalues are greater than 1. The eigenvectors (or principal components, PCs) form a new set of derived variables that captures several important characteristics of the features (A) of apatite crystal structure that are helpful in developing structure maps.

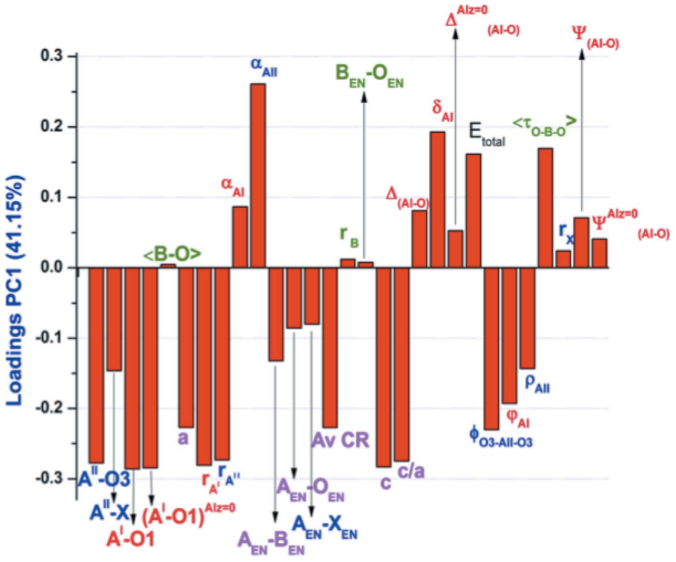
In short, PCA decomposes the data matrix $\hat{\mathbf{A}} = \left(\hat{A}_1, \hat{A}_2, ..., \hat{A}_k\right) \in \Re^{25\times 29}$ into two matrices **P** and **U**, where **P** is called the loadings matrix (PCs or eigenvectors of sample covariance matrix **S**) and **U** is called the scores matrix (obtained by matrix multiplication of $\hat{\mathbf{A}}$ and **P**). Therefore, the final model can be mathematically described as

$$\hat{\mathbf{A}} = \mathbf{P}^T\mathbf{U} + \mathbf{R}, \quad (1)$$

where **R** is the residual matrix. In the past PCA has been applied to analyze bond geometries of crystal and molecular structures (Murray-Rust & Motherwell, 1978; Pawlak *et al.*, 1999; Bürgi, 1998) and in this work we have applied it for the first time to identify the key crystallographic parameters of apatite crystal chemistries.



(*a*)

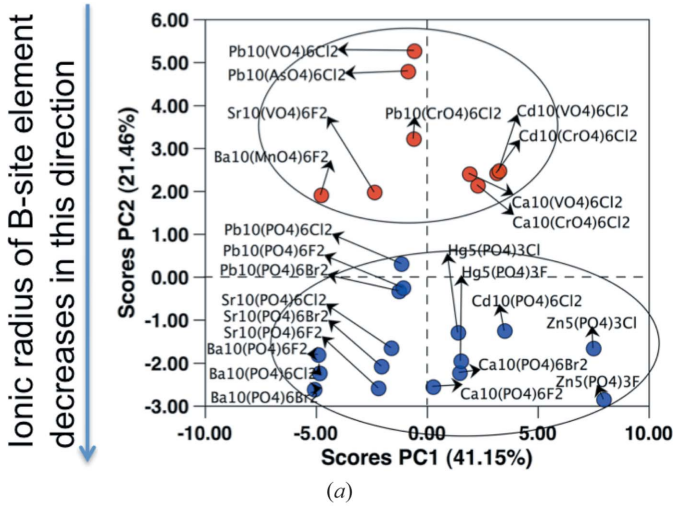


(*b*)

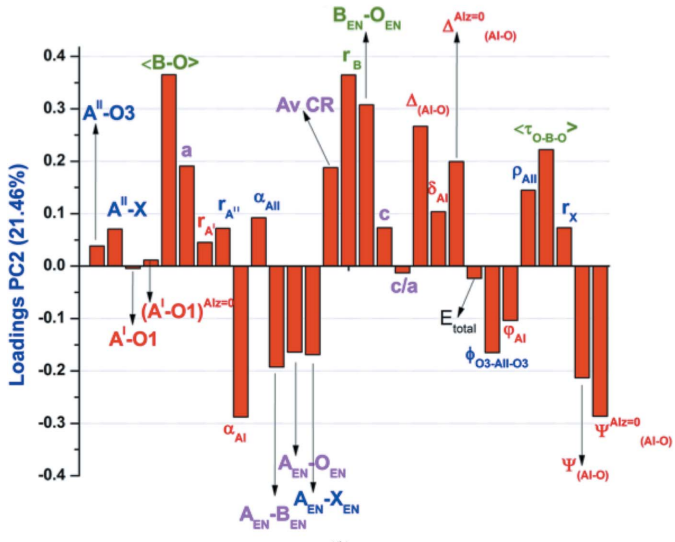
Figure 2
(*a*) This is a two-dimensional plot, commonly referred to as a scores map, between latent variables PC1 and PC2, and each data point is a correlation position representing an apatite compound as influenced by 29 descriptors. Along the PC1 axis, distinct clusters of apatite compounds with the same *A*-site elements are identified. The ionic radii of *A*-site elements decrease in the direction shown in the figure and clearly Zn stoichiometries could be seen to be well separated from the rest of the compounds in this classification map, which correlates well with its uncertain existence. (*b*) Having identified the structural correlations along the PC1 axis, the relative influence of the dominant descriptors that are responsible for the observed pattern is defined in this plot (commonly referred to as a loadings map). The descriptors that dominate the PC1 axis are r_{AI}, r_{AII}, Av CR, $A^{I}-O1$, $(A^{I}-O1)^{AIz=0}$, $A^{II}-O3$, α_{AII}, $\Phi_{O3-AII-O3}$, and lattice constants (*a*, *c* and *c*/*a*) since these descriptors carry the largest weight. The abbreviations of the 29 descriptors used in the loadings map are given in Table 1.

4. Results and discussion

From applying PCA to the multivariate apatite data, it was identified that the first three PCs together explain more than 80% of the variation in the data. We will use the three PCs and demonstrate new strategies for constructing structure maps for apatite-type compounds. Before we show how to construct a structure map from knowledge of the computed PCs we will discuss the results obtained from the loadings and scores matrix. The loadings matrix yields insight into the relationship



(*a*)



(*b*)

Figure 3
(*a*) The scores map shown here captures the pattern associated with the same *B*-site element along the PC2 axis. Along the PC2 axis, apatites with P (phosphorus) in the *B*-site are clustered together and are seen to be well separated from the other cluster containing V, As, Mn and Cr in the *B* site. (*b*) The loadings map defines the dominant descriptors that are responsible for the typical pattern observed in the scores map. The significant electronic structure and bond-distortion variables are r_B, $B_{EN}-O_{EN}$, $\langle B-O\rangle$, $\langle\tau_{O-B-O}\rangle$, α_{AI}, $\psi_{AI-O}^{AIz=0}$ and Δ_{AI-O}. The abbreviations of the 29 descriptors used in the loadings map are given in Table 1.

between site chemistry-geometrical parameters and the relationship between algebraically independent geometrical parameters. The scores matrix captures the correlation between the apatite chemistries.

4.1. Mapping the structural patterns associated with *A*-site chemistry

Each data point in Fig. 2(*a*) (referred to as a scores map) represents a correlation position of an apatite compound as influenced by 29 descriptors. The percentage labelled on the axes corresponds to the amount of variance of the total dataset captured by the respective axes. Along the PC1 axis, clusters of apatite compounds based on the differences in the ionic radii of *A*-site chemical elements are recognized. As a generic trend we find that the ionic size of the *A*-site element decreases as we move from left to right along the PC1 axis, as shown in Fig. 2(*a*). Accordingly, the compounds belonging to the same *A*-site chemical element (Ba, Sr, Pb, Ca, Hg, Cd and Zn) are grouped together. Clearly, the two Zn-based stoichiometries (located farthest right with a relatively large PC1 value) could be seen well separated from the rest of the compounds. This correlates well with the experimental results and thermodynamic calculations on the uncertain existence of fully stoichiometric Zn-based apatite compounds (Grisafe & Hummel, 1970; Flora *et al.*, 2004).

The variables that are responsible for the observed pattern can be understood by visualizing the PC1 axis, as shown in Fig. 2(*b*) (referred to as a loadings map). The weights in the histogram give the relative contribution of each descriptor on the PC1 axis. The descriptors that dominate the PC1 axis are r_{AI}, r_{AII}, Av CR (average crystal radius), A^{I}—O1, A^{II}—O3, α_{AII}, $\Phi_{O3-AII-O3}$ and lattice constants (*a*, *c* and *c*/*a*) since these descriptors carry the largest weight. Besides identifying the dominant descriptors, we find that r_{AI}, r_{AII}, Av CR, A^{I}—O1, A^{II}—O3, $\Phi_{O3-AII-O3}$ and lattice constants are directly correlated to one another (all have negative PC1 values) and are inversely correlated to the distortion angle α_{AII} (has positive PC1 value). While the inverse relationship between α_{AII} and $\Phi_{O3-AII-O3}$ is already known (Mercier *et al.*, 2005), the impact of ionic radii on α_{AII} and $\Phi_{O3-AII-O3}$ is identified for the first time. The dominant influence of ionic radii (r_{AI} and r_{AII}) compared with the electronegativity differences

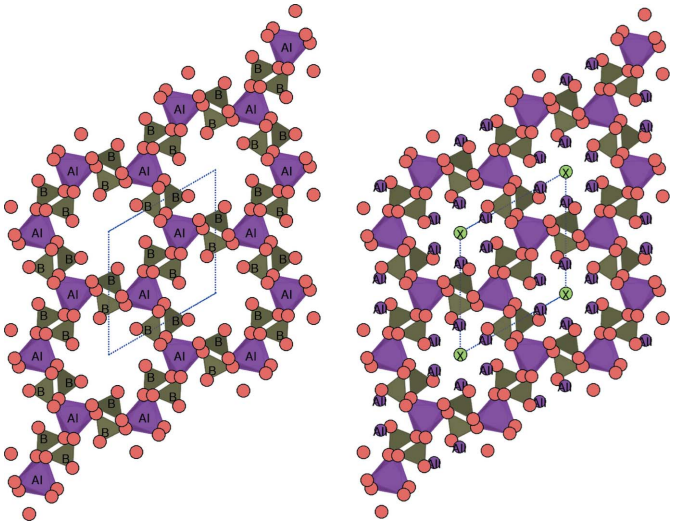


Figure 4
The figure on the left is the apatite framework $[A^{I}_{4}(BO_4)_6]^{10-}$ composed of $A^{I}O_6$ metaprisms linked together with BO_4 tetrahedra with channels extending right through the structure. In these channels are inserted the remaining A^{II} and *X* ions forming an $[A^{II}_{6}X_2]^{10+}$ complex to neutralize the framework. The final crystal structure is typically of the form as shown on the right. The dotted blue cell represents the basic unit cell of the apatite crystal structure.

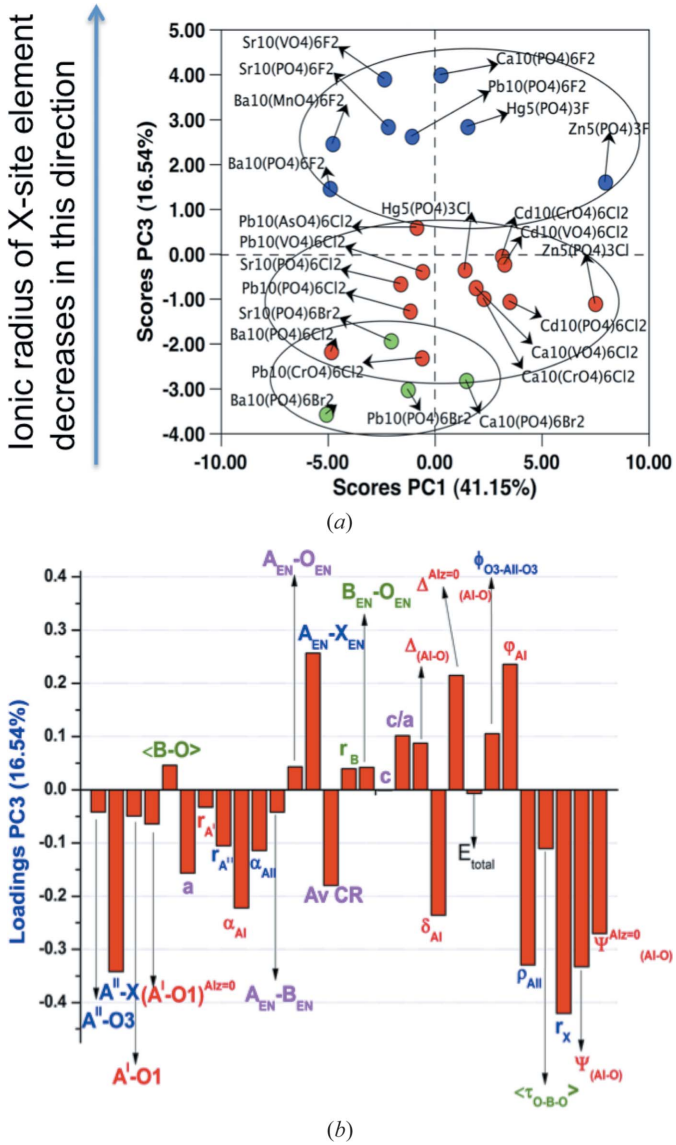


Figure 5
(*a*) The scores map shown here captures the pattern associated with the same *X*-site element along the PC3 axis. Along the PC3 axis three clusters of apatites are identified separating F, Cl and Br in the *X* site. (*b*) The loadings map identifies the significant electronic structure and bond-distortion variables that are responsible for the typical pattern observed in the scores map. The descriptors that dominate the PC3 axis are r_X, A^{II}—*X*, A_{EN}—X_{EN}, ψ_{AI-O}, ρ_{AII}, $\Delta^{AI_{z}=0}_{AI-O}$ and φ_{AI}. The abbreviations of the 29 descriptors used in the loadings map are given in Table 1.

($A_{EN}-O_{EN}$, $A_{EN}-B_{EN}$, $A_{EN}-X_{EN}$) suggest that the complex bond distortions due to *A*-site substitution are controlled predominantly by the steric constraints resulting from the packing of atoms.

4.2. Mapping the structural patterns associated with *B*-site chemistry

In Fig. 3(*a*) it is clearly shown that the PC2 axis recognizes the pattern associated with *B*-site elements. Apatites with P (phosphorus) in the *B* site are clustered together and are seen to be well separated from V, As, Mn and Cr containing apatites. To understand the dominant variables behind the classification pattern, the PC2 axis is visualized (Fig. 3*b*). The descriptors that dominate the PC2 axis are r_B, $B_{EN}-O_{EN}$, $\langle B-O\rangle$, $\langle\tau_{O-B-O}\rangle$, α_{AI}, $\psi_{AI-O}^{AIz=0}$ and Δ_{AI-O}. Among these descriptors, r_B, $B_{EN}-O_{EN}$, $\langle B-O\rangle$, $\langle\tau_{O-B-O}\rangle$ and Δ_{AI-O} are directly correlated to one another and are inversely correlated to α_{AI} and $\psi_{AI-O}^{AIz=0}$ (which are in-turn directly correlated to one another). It should be noted here that the three bond distortion descriptors, Δ_{AI-O}, α_{AI} and $\psi_{AI-O}^{AIz=0}$ belong to the A^IO_6 structural unit. This indicates that any site substitution along the *B* site not only affects the BO_4 tetrahedra but also significantly affects the geometry of the A^IO_6 structural unit. The relationship between *B*-site cations and the variability in Δ_{AI-O}, α_{AI} and $\psi_{AI-O}^{AIz=0}$ distortion parameters could be explained by considering the apatites as framework structures (White *et al.*, 2005), where the A^IO_6 metaprism and BO_4 tetrahedra together form the framework $[A^I_4(BO_4)_6]^{10-}$ with channels extending right through the structure and in the channels are located $[A^{II}{}_6X_2]^{10+}$ complex ions neutralizing the framework (Fig. 4). Therefore, any lattice substitution in the *B* site affects the geometry of the A^IO_6 metaprism unit, which is in turn expressed in the distortion parameters Δ_{AI-O}, α_{AI} and $\psi_{AI-O}^{AIz=0}$. The ionic size of the *B*-site atom (r_B) is found directly correlated to lattice constants (*a* and *c*), however, we find that variation in *a* is significantly higher compared with *c* (Fig. 3*b*).

4.3. Mapping the structural patterns associated with X-site chemistry

In Fig. 5(*a*) three distinct clusters of apatites are recognized along the PC3 axis, differentiating various apatite compounds with respect to the *X*-site elements (F, Cl and Br). To further understand the crystal chemical meaning behind the classification pattern, the PC3 axis is visualized (Fig. 5*b*). The descriptors that dominate the PC3 axis (Fig. 5*b*) are r_X, $A^{II}-X$, $A_{EN}-X_{EN}$, ψ_{AI-O}, $\psi_{AI-O}^{AIz=0}$, ρ_{AII}, $\Delta_{AI-O}^{AIz=0}$ and φ_{AI}. Among these, r_X, $A^{II}-X$, ψ_{AI-O}, $\psi_{AI-O}^{AIz=0}$ and ρ_{AII} are directly correlated to one another and are inversely correlated to $A_{EN}-X_{EN}$, $\Delta_{AI-O}^{AIz=0}$ and φ_{AI} (which are directly correlated to one another). It should be noted here that the three bond distortion variables $\Delta_{AI-O}^{AIz=0}$, φ_{AI} (or δ_{AI}) ψ_{AI-O} and $\psi_{AI-O}^{AIz=0}$ belong to the A^IO_6 structural unit. Hence, any site-substitution along the *X* site significantly affects the geometry of the A^IO_6 structural unit (ψ_{AI-O}, $\psi_{AI-O}^{AIz=0}$, $\Delta_{AI-O}^{AIz=0}$ and φ_{AI}). The role of ionic size (r_X) and chemical bonding ($A_{EN}-X_{EN}$) in impacting the bond distortions is very evident from the interpretation of the PC3 axis. The lattice constant *a* is found to be directly correlated to r_X, whereas the lattice constant *c* appears to be essentially unaffected (Fig. 5*b*), which suggests that any site substitution along the *X* site affects the lattice constant *a* far more than compared with the lattice constant *c*.

Hughes *et al.* (1989) reported that in $Ca_5(PO_4)_3X$ (*X* = F, Cl, OH) apatites, changes induced in the crystal structure as a result of differences in the *X* anions propagate throughout the structure with the $Ca^{II}O_6X_{1,2}$ polyhedron being greatly affected, but have a minor impact on the average P—O and

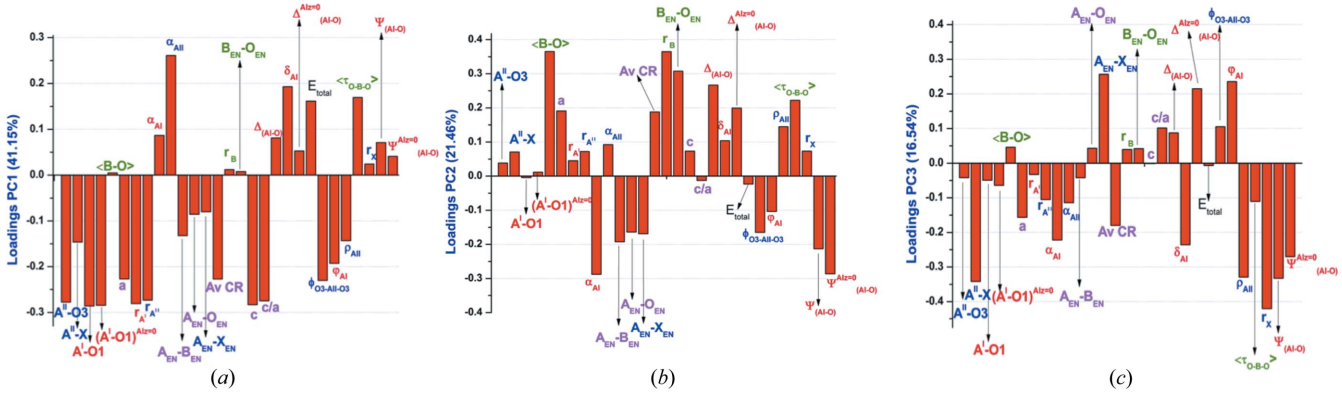


Figure 6
The coordinates for constructing new structure maps are obtained from the loadings maps. The loadings maps are used for two purposes: (i) to identify the dominant bond-geometrical descriptors and (ii) to screen the dominant descriptors for structure map construction. Measuring the absolute distance from the origin identifies dominant descriptors. The impact of the descriptors is increased as its distance from the origin is increased. The purpose of screening is to remove intercorrelated descriptors. Within a PC, all descriptors are intercorrelated. However, the PCs are orthogonal and uncorrelated to each other. Therefore, we can construct a structure map by picking any two dominant descriptors from the (*a*) PC1, (*b*) PC2 and (*c*) PC3 axes. Here we have demonstrated in detail the construction of a structure map by choosing α_{AII} and $\psi_{AI-O}^{AIz=0}$ bond distortion angles as the dominant descriptors. This logic can also be extended to other dominant descriptors. It should be noted that in this figure we have repeated the same loadings maps shown in Figs. 2(*b*), 3(*b*) and 5(*b*), however, we have interpreted them in a totally different manner. The role of the loadings maps shown here is to identify key bond-distortion parameters for constructing new structure maps for apatites. In the previous case, the loadings maps were used to rationalize the structural patterns observed in Figs. 2(*a*), 3(*a*) and 5(*a*) related to *A*, *B* and *X* site occupancy.

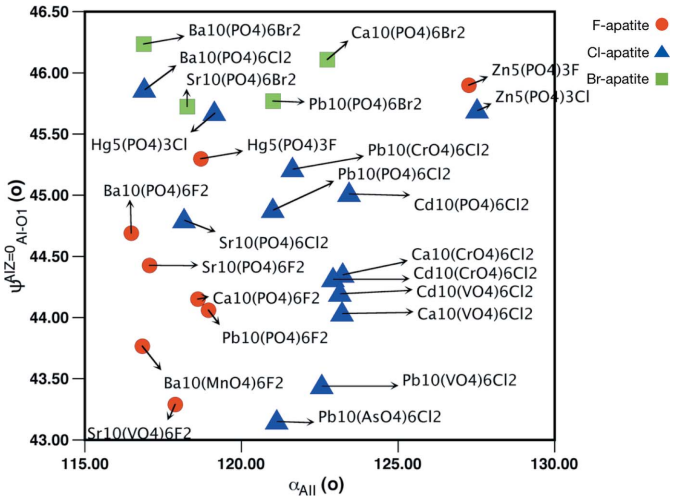


Figure 7
The new structure map for apatite compounds defined using the two orthogonal coordinates α_{AII} and $\psi_{AI-O1}^{AIz=0}$ identified through PCA is shown. Filled circles, triangles and squares represent apatites that have fluoride (F), chloride (Cl) and bromide (Br) anions in the X site, respectively.

$Ca^{I}-O$ bond lengths. However, our data-mining work suggests that any site substitution in the X site, not only affects the $A^{II}O_6X_{1,2}$ structural unit but also has a significant impact on the orientation of the $A^{I}O_6$ structural unit. For example, if we consider the relative magnitude (in absolute scale) of each descriptor in the loadings map (shown in Fig. 5*b*), it is clear that $\langle B-O\rangle$ and $A^{I}-O1$ bond-length descriptors, representing the average bond lengths in BO_4 and $A^{I}O_6$ structural units, have a relatively minor impact on the PC3 axis. On the other hand, bond-length descriptors associated with $A^{II}O_6X_{1,2}$ structural unit such as $A^{II}-X$ and ρ_{AII} have a significant impact on the PC3 axis. This pattern, corresponding to the variations in the average bond lengths as a result of X-site substitution, is in agreement with the observation of Hughes *et al.* (1989). However, our loadings map also identifies that φ_{AI} (or δ_{AI}) ψ_{AI-O} and $\psi_{AI-O}^{AIz=0}$ bond-angle descriptors undergo relatively significant variation due to X-site substitution, an effect that was undetected by Hughes *et al.* (1989). This observation leads to the conclusion that structural adjustments caused due to the differences in column anions can be more clearly understood by critically examining the local variations caused due to the changes in the following bond distortion angles: φ_{AI} (or δ_{AI}) ψ_{AI-O} and $\psi_{AI-O}^{AIz=0}$.

5. New structure map for apatites

In the results discussed in §4 we employed the scores map. In this section we will use the loadings map (Fig. 6) to screen the bond geometrical descriptors so that we can choose two dominant yet uncorrelated descriptors that can be used as the coordinates for a new structure map. The process involves two steps:

Step 1 – Identification of dominant bond geometrical descriptors: The relative impact of each descriptor in a loadings map is identified by measuring the absolute distance from the origin. The impact of the descriptors is increased as its distance from the origin is increased. Therefore, from the loadings maps shown in Fig. 6 we identify the following dominant bond geometrical descriptors:

(i) With respect to the PC1 axis (Fig. 6*a*), $A^{I}-O1$, $(A^{I}-O1)^{AIz=0}$, $A^{II}-O3$, $\Phi_{O3-AII-O3}$ and α_{AII} are dominant.

(ii) With respect to the PC2 axis (Fig. 6*b*), $\langle B-O\rangle$, α_{AI}, $\psi_{AI-O}^{AIz=0}$, ψ_{AI-O} and $\langle\tau_{O-B-O}\rangle$ are dominant.

(iii) With respect to the PC3 axis (Fig. 6*c*), $A^{II}-X$, ψ_{AI-O}, ρ_{AII}, $\psi_{AI-O}^{AIz=0}$, φ_{AI} (or δ_{AI}) and α_{AI} are dominant.

Step 2 – Screening the dominant descriptors for structure map construction: The purpose of screening is to select two strong classifiers that are dominant yet uncorrelated. Within a PC axis, all descriptors are intercorrelated. For example, although $A^{I}-O1$, $(A^{I}-O1)^{AIz=0}$, $A^{II}-O3$, $\Phi_{O3-AII-O3}$ and α_{AII} are dominant with respect to the PC1 axis, they are intercorrelated. This implies that from knowledge of any one of these descriptors the variation in the other can be estimated. On the other hand, these PCs are orthogonal and uncorrelated to each other. As a result we can choose any two variables, one from each set $\{A^{I}-O1, (A^{I}-O1)^{AIz=0}, A^{II}-O3, \Phi_{O3-AII-O3}$ and $\alpha_{AII}\}$, $\{\langle B-O\rangle$, α_{AI}, $\psi_{AI-O}^{AIz=0}$, ψ_{AI-O} and

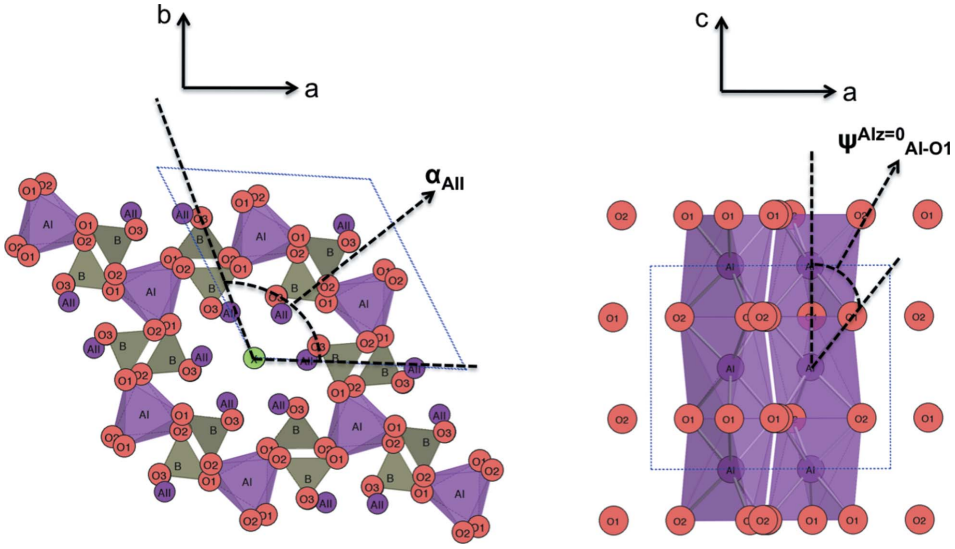


Figure 8
The key bond-distortion angles, α_{AII} and $\psi_{AI-O1}^{AIz=0}$ identified from PCA, are schematically shown here. α_{AII} is defined as the rotation angle of $A^{II}-A^{II}-A^{II}$ triangular units and $\psi_{AI-O1}^{AIz=0}$ is defined as the angle the $A^{I}-O1$ bond makes with respect to the c axis with the additional constraint $z = 0$ at the A^{I} site (Mercier *et al.*, 2005). It should be noted that the two algebraically independent distortion angles are manifested in different orientations of the crystal structure as shown in the figure.

Table 3
Interpretation of data clusters obtained from the K-means clustering method.

The relationship linking various clusters shown in Fig. 9 with the site occupancy is described.

Clusters	Site occupancy
k = 1 and k = 2	*A* site: Ba, Pb, Sr, Ca
	B site: P, V, Mn
	X site: F
k = 3	*A* site: Ba
	B site: P
	X site: Cl and Br
k = 4	*A* site: Sr, Hg
	B site: P
	X site: Cl and Br
k = 5	*A* site: Ca, Cd, Pb
	B site: V, Cr, As
	X site: Cl
k = 6	*A* site: Ca, Pb
	B site: P
	X site: Cl and Br
k = 7	*A* site: Zn
	B site: P

$\langle\tau_{O-B-O}\rangle\}$ and $\{A^{II}-X$, ψ_{AI-O}, ρ_{AII}, $\psi^{AIz=0}_{AI-O}$, φ_{AI} (or δ_{AI}) and $\alpha_{AI}\}$ to construct new structure maps. This strategy ensures robustness in the chosen variables because such pairs of variables are likely to have only a low correlation between them.

Following this two-step approach, a library of new structure maps can be developed for apatites. Here we demonstrate the logic in detail for constructing a new structure map defined using the two distortion angles α_{AII} (rotation angle of A^{II}–A^{II}–A^{II} triangular units) and $\psi^{AIz=0}_{AI-O1}$ (the angle the $A^{I}-O1$ bond makes with respect to the c axis with the constraint $z = 0$ at the A^{I} site). This process can also be extended to other dominant descriptors.

The new structure map for apatites defined between α_{AII} and $\psi^{AIz=0}_{AI-O1}$ distortion angles is shown in Fig. 7. While α_{AII} is the key bond-distortion angle of the apatite crystal structure carrying the largest PC1 coefficient (among bond angles), the significance of $\psi^{AIz=0}_{AI-O1}$ is evident from the loadings map of the PC2 and PC3 axis. With respect to the PC2 axis, $\psi^{AIz=0}_{AI-O1}$ and α_{AI} carry similar weights indicating their similarity. However, with respect to the PC3 axis, the relative weight of $\psi^{AIz=0}_{AI-O1}$ is significantly higher than α_{AI} and is comparable to ψ_{AI-O1} (which in turn has a relatively lower significance with respect to the PC2 axis). Since the variance captured by the PC2 axis is higher compared with the PC3 axis, we selected α_{AII} and $\psi^{AIz=0}_{AI-O1}$ as the two orthogonal coordinates of the new structure map. The schematic of the geometry of distortion angles α_{AII} and $\psi^{AIz=0}_{AI-O1}$, as described by Mercier *et al.* (2005), is shown in Fig. 8.

The structure map picks up site occupancy information that demarcates broad regimes in apatite crystal chemistry. This is shown in Fig. 9 where we have quantitatively mapped out various clusters through the K-means clustering method. Unlike most structure maps where the proximity between data points is explored through visual inspection, we have employed the K-means clustering method to detect natural groupings in the data. The K-means clustering approach provides a quantitative metric for proximity estimation (see *Appendix A*).

The classification of apatite crystal chemistries after K-means clustering is shown in Fig. 9. The clusters are numbered from 1 to 7 in an arbitrary order for ease of interpretation. The relationship linking seven data clusters with the site-occupancy information is summarized in Table 3. The structure map identifies new and unexplored patterns of behavior of apatite compound chemistries, reinforcing the fact that the two distortion angles, α_{AII} and $\psi^{AIz=0}_{AI-O1}$ are strong classifiers. The validity of our methods is aided by the fact that we can recover known information, and at the same time add significant new information that would not have been easily discernible. For example, clusters 1 and 2 (k = 1 and k = 2) correspond to F-apatites. They are well localized in the structure map and are characterized by relatively low α_{AII} and $\psi^{AIz=0}_{AI-O1}$. The two F-apatites that do not belong to the clusters k = 1 and k = 2 are $Hg_5(PO_4)_3F$ (in k = 4) and $Zn_5(PO_4)_3F$ (in k = 7). While the existence of a fully stoichiometric $Zn_5(PO_4)_3F$ apatite compound is uncertain due to the relatively smaller ionic size of Zn^{2+} cations (Grisafe & Hummel, 1970; Flora *et al.*, 2004), the relative position of $Hg_5(PO_4)_3F$ suggests some peculiar characteristics. Even though Ca^{2+} and Hg^{2+} cations have roughly the same ionic size (1.18 and 1.23 Å in the A^{I} site), their electronegativity data indicates that Hg atoms (electronegativity value of 2 in Pauling scale) are relatively highly covalent compared with the Ca atoms (electronegativity value of 1 in Pauling scale). In the structure map this covalent character is predicted to be manifested in the bond distortion angle $\psi^{AIz=0}_{AI-O1}$. This observation of unusual behaviour in $Hg_5(PO_4)_3F$ stoichiometry, discovered based solely on data-mining methods, identifies a new topic for

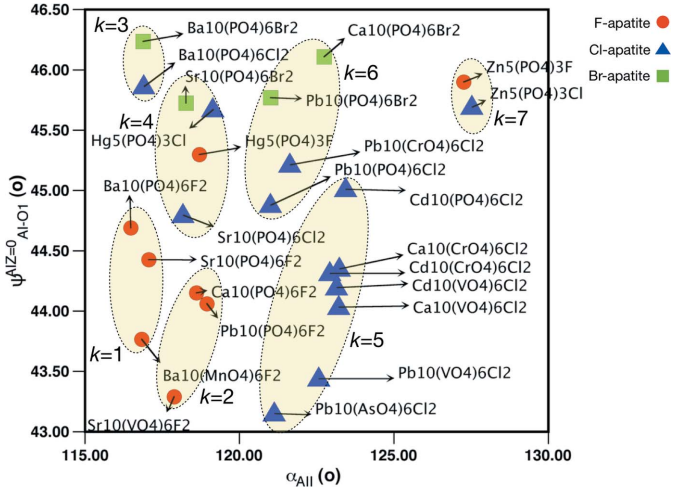


Figure 9
The structure map with the data classification is shown. The classification is accomplished through a quantitative K-means clustering method. The clusters are numbered from 1 to 7 in an arbitrary order for ease of interpretation. We find that the clustering effectively classifies apatite chemistries based on site occupancy on the *A*, *B* and *X* sites. The relationship linking data clusters with site occupancy information is summarized in Table 3.

detailed computational investigations and has initiated another study of further exploration, to be reported in another paper. It should be noted that there is no study reported in the literature on the synthesis and determination of the phase stability of Hg-containing apatites.

6. Conclusion

We have demonstrated a data-mining approach based on an unsupervised learning scheme using principal component analysis (PCA) and K-means clustering for developing structure maps without any *a priori* choice of key factors governing the complex apatite crystal structure. We have identified a new structure map that serves as a strong classifier and captures several regularities in the bond distortions of apatite compounds that were associated with A, B and X site occupancy.

APPENDIX *A*

K-means clustering is a well known methodology used in pattern recognition studies for discovering natural grouping(s) in complex data sets (Jain, 2010). Given a representation of n objects or data points, the K-means clustering attempts to find k groups (or clusters) based on a measure of similarity such that the similarities between data points in the same cluster are high while the similarities between data points in different cluster are low. The algorithm proceeds as follows (Han & Kamber, 2006). First, it randomly selects k of the objects (the value of k must be provided at the beginning), each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. While there are different similarity metrics, we have chosen the conventional Euclidean distance (L2-norm). It then computes the new mean for each cluster. This process iterates until the squared-error criterion function converges

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2, \qquad (2)$$

where E is the sum of the square error for all objects in the data set; p is the point in the space representing a given object (chemical composition); m_i is the mean of cluster C_i. In other words, for each object in each cluster, the distance from the object to the cluster center is squared and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible. The above process was repeated over several trials for different values of k (3–9) to determine the optimal k clusters in the two-dimensional structure map shown in Fig. 6. For k = 7, we obtained an optimal partition in the data. We followed the criterion suggested by Tibshirani *et al.* (2001) to determine the optimal number of clusters. At first, K-means is run independently for different values of k. The sum of the squared error distance decreases as k increases, but from k = 7 onwards we found that the decrease in the sum of the squared error distance markedly reduces thereby forming an elbow, which indicates the appropriate number of clusters.

The authors acknowledge support from the Air Force Office of Scientific Research, grants #FA9550-06-10501 and #FA9550-08-1-0316; NSF-ARI Program: CMMI 09-389018; NSF-CDI Type II program: grant #PHY 09-41576, NSF CRI-IAD grant #07-51157 and NSF-AF grant #CCF09-17202 and Army Research Office grant #W911NF-10-0397. KR also acknowledges support from the Wilkinson Professorship of Interdisciplinary Engineering.

References

Aourag, H., Broderick, S. R. & Rajan, K. (2010). *Phys. Status Solidi B*, **247**, 115–121.
Baikie, T., Mercier, P. H. J., Elcombe, M. M., Kim, J. Y., Le Page, Y., Mitchell, L. D., White, T. J. & Whitfield, P. S. (2007). *Acta Cryst.* B**63**, 251–256.
Baikie, T., Pramana, S. S., Ferraris, C., Huang, Y., Kendrick, E., Knight, K., Ahmad, Z. & White, T. J. (2010). *Acta Cryst.* B**66**, 1–16.
Balachandran, P. V., Broderick, S. R. & Rajan, K. (2011). *Proc. R. Soc. A*, **467**, 2271–2290.
Broderick, S. R., Nowers, J. R., Narasimhan, B. & Rajan, K. (2010). *J. Comb. Chem.* **12**, 270–277.
Bürgi, H. B. (1998). *Acta Cryst.* A**54**, 873–885.
Đordević, T., Šutović, S., Stojanović, J. & Karanović, Lj. (2008). *Acta Cryst.* C**64**, i82–i86.
Elliott, J. C. (1994). *Structure and Chemistry of the Apatites and Other Calcium Orthophosphates.* New York: Elsevier Science.
Flora, N. J., Yoder, C. H. & Jenkins, H. D. (2004). *Inorg. Chem.* **43**, 2340–2345.
Gadzuric, S., Suh, C., Gaune-Escard, M. & Rajan, K. (2006). *Metall. Trans. A*, **37**, 3411–3414.
George, L., Hrubiak, R., Rajan, K. & Saxena, S. (2009). *J. Alloys Compd*, **478**, 731–735.
Grisafe, D. A. & Hummel, F. A. (1970). *Am. Mineral.* **55**, 1131–1145.
Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*, 2nd ed. California: Morgan Kauffman Publishers.
Hauck, J. & Mika, K. (2002). *Cryst. Eng.* **5**, 105–121.
Hughes, J. M., Cameron, M. & Crowley, K. D. (1989). *Am. Mineral.* **74**, 870–876.
Hughes, J. M. & Rakovan, J. (2002). *Rev. Mineral. Geochem.* **48**, 1–12.
Jain, A. K. (2010). *Pattern Recognit. Lett.* **31**, 651–666.
Jolliffe, I. T. (2002). *Principal Component Analysis.* New York: Springer-Verlag.
Kim, J. Y., Dong, Z.-L. & White, T. J. (2005). *J. Am. Ceram. Soc.* **88**, 1253–1260.
Kleinke, H. & Harbrecht, B. (2000). *Z. Anorg. Allg. Chem.* **626**, 1851–1853.
Li, C., Lu, X., Ding, W., Feng, L., Gao, Y. & Guo, Z. (2008). *Acta Cryst.* B**64**, 702–707.
Matsunaga, K., Inamori, H. & Murata, H. (2008). *Phys. Rev. B*, **78**, 094101.
Mercier, P. H. J., Dong, Z., Baikie, T., Le Page, Y., White, T. J., Whitfield, P. S. & Mitchel, L. D. (2007). *Acta Cryst.* B**63**, 37–48.
Mercier, P. H. J., Le Page, Y., Whitfield, P. S., Mitchell, L. D., Davidson, I. J. & White, T. J. (2005). *Acta Cryst.* B**61**, 635–655.
Mooser, E. & Pearson, W. B. (1959). *Acta Cryst.* **12**, 1015–1022.
Morris, R. J. (2004). *Acta Cryst.* D**60**, 2133–2143.
Murray-Rust, P. & Motherwell, S. (1978). *Acta Cryst.* B**34**, 2534–2546.
Pauling, L. (1960). *The Nature of the Chemical Bond.* New York: Cornell University Press.

Pawlak, D. A., Woźniak, K. & Frukacz, Z. (1999). *Acta Cryst.* B**55**, 736–744.
Pettifor, D. G. (1986). *J. Phys. Solid State Phys.* **19**, 285–313.
Pramana, S. S., Klooster, W. T. & White, T. J. (2008). *J. Solid State Chem.* **181**, 1717–1722.
Rabe, K. M., Phillips, J. C., Villiars, P. & Brown, I. D. (1992). *Phys. Rev. B*, **45**, 7650–7676.
Rajagopalan, A. & Rajan, K. (2007). *Combinatorial and High-Throughput Discovery and Optimization of Catalysts and Materials*, edited by W. Maier & R. A. Potyrailo. Boca Raton: CRC Press.
Rajan, K. (2010). *Data Mining in Crystallography-Structure and Bonding Series*, edited by D. W. M. Kuleshova & N. Liudmila, Vol. 134, pp. 59–87. Berlin: Springer-Verlag.
Rajan, K., Suh, C. & Mendez, P. F. (2009). *Data Mining*, **1**, 361–371.
Ringnér, M. (2008). *Nat. Biotechnol.* **26**, 303–304.
Shannon, R. D. (1976). *Acta Cryst.* A**32**, 751–767.
Sugiyama, S. (2007). *Phosphorus Res. Bull.* **21**, 1–8.
Suh, C. & Rajan, K. (2005). *QSAR Comb. Sci.* **24**, 114–119.
Suh, C. & Rajan, K. (2009). *Mater. Sci. Technol.* **25**, 466–471.
Suzuki, T., Ishigaki, K. & Miyake, M. (1984). *J. Chem. Soc. Faraday Trans. 1*, **80**, 3157–3165.
Tibshirani, R., Walther, G. & Hastie, T. (2001). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411–423.
Villars, P. (1984). *J. Less-Common Met.* **102**, 199–211.
White, T. J. & ZhiLi, D. (2003). *Acta Cryst.* B**59**, 1–16.
White, T. J., Ferraris, C., Kim, J. & Madhavi, S. (2005). *Rev. Mineral. Geochem.* **57**, 307–401.
Zhang, H., Li, N., Li, K. & Xue, D. (2007). *Acta Cryst.* B**63**, 812–818.
Zunger, A. (1980). *Phys. Rev. B*, **22**, 5839–5872.