

Data mining for new materials chemistries for immobilizing environmentally toxic elements: a systems engineering approach

Prasanna V Balachandran
Dept of Materials Science & Engg

Chitra Rajan
Dept of Economics
rajan@iastate.edu

Krishna Rajan*
Dept of Materials Science & Engg
*krajan@iastate.edu

Iowa State University, Ames, IA, USA

Abstract

This paper addresses the critical challenge of applying data mining methodologies to fuse diverse data sets relating materials science data to economics data, for creating a balance in the underlying physical system. By applying a combination of data mining techniques, the crystal chemical systematics of the complex apatite crystal structures as a potential material for immobilizing toxic materials is explored from energy-cost perspective. Potential apatite chemistries are identified as host lattices for sequestering toxic elements and a quantitative model was developed to determine the formability of the substituted lattice. The developed data mining framework is a novel approach in the chemical modeling of new chemistries even in the field of materials science for regulating hazardous materials from environment.

1. Introduction

The challenge of designing new materials to address environmental and energy issues and still provide economically an affordable solution is a daunting task. The link between developing technical solutions in parallel with an understanding of economic impact is difficult as no *a priori* theories really exist. Even for addressing the purely chemistry and physics issues in designing materials is a very high dimensional problem and adding economic constraints makes it even more complex. This paper is an introduction to the work in our group in applying a data driven discovery approach to this problem. The template described in this paper addresses the important issue of environmental remediation by finding ways to contain toxic elements. The strategy is to use data mining to discover new and yet untested materials chemistries. By applying a combination of data mining techniques, such as data dimensionality reduction techniques, cluster analysis and regression analysis we identify such new materials. Our methods can be generalized to incorporate both physics-based as well as economics-based descriptors and hence lay the foundation of a systems based engineering solution. In this paper, we present our preliminary results in the on-going research where we apply a combination of linear and non-linear data mining techniques to not only identify new

chemistries but with the developed quantitative-structure-property-relationship (QSPR) model, identify which of the chosen chemistries may be the most suited from an energy-cost perspective. We use heat of formation (enthalpy change- ΔH_f°) as our metric of energy-cost.

2. Apatite crystal structure

Apatites are a class of complex crystal structures containing 42 atoms per unit cell represented with the general chemical formula $A_{10}(BO_4)_6X_2$, where A-is the larger cation, B-is the smaller cation and X is an anion (see fig. 1). Apatites are used in a number of applications such as fuel cell electrolyte, dentistry, catalysis, remediation of hazardous materials etc. This versatility in the crystal structure can be mainly attributed to the nature of its crystal structure in imbibing a number of varied cations and anions and still maintain the crystal structure from collapsing [1-3]. The application of apatites for

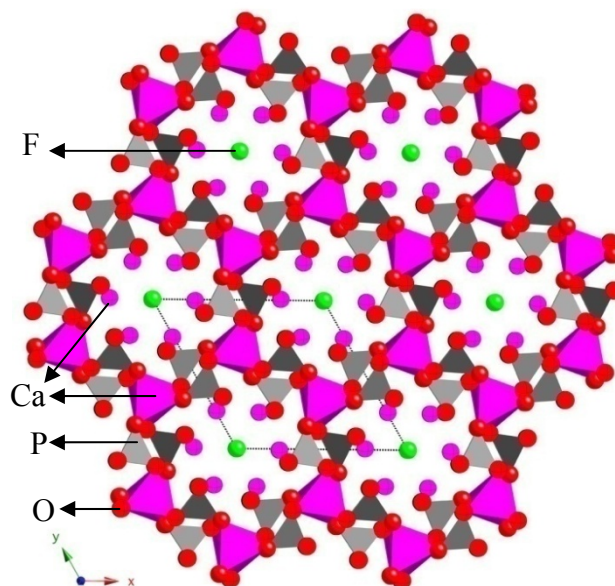


Figure 1 Crystal structure of $Ca_{10}(PO_4)_6F_2$. The crystal structure that is shown here belongs to hexagonal $P6_3/m$ space group. The various atomic sites are labeled in the figure. It is the ability to substitute for ions of varied chemistry makes this crystal structure unique for immobilizing toxic materials

immobilizing toxic materials is not a completely new topic of interest. However, very few attempts have been made to systematically model the crystal chemical aspects to design chemistries for sequestering hazardous materials [4, 5].

The most common challenge associated with analyzing any complex crystal structures such as apatites are that the nature of crystal structural data is high dimensional in nature. Parameterization of the crystal structural information needs modeling a large number of descriptors to comprehensively quantify the diverse crystal structural aspects. Therefore, in order to understand the complex crystal chemical interactions, there is a necessity to reduce the complexity of the data and visualize the interactions in the low dimensional space. This defines the application of data dimensionality reduction techniques for studying these complex chemistries. The geometric parameterization of the complex apatite crystal structures into a number of distinct bond lengths and bond angles was modeled by Mercier et al [2]. For the initial data mining exercise, a high dimensional database was constructed comprising of 25 apatite chemistries and each apatite was in turn represented by 29 descriptors. Figure 2 schematically shows the nature of database constructed for data dimensionality reduction. The nature of 25 apatite chemistries are such that, the chemical variations are made along A-site, B-site and X-site. Therefore, all the crystal structural changes can be attributed to the changes in A-site, B-site or X-site chemistries.

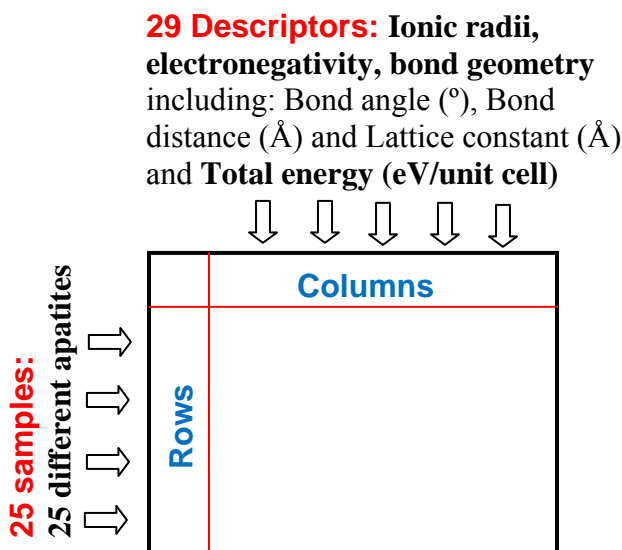


Figure 2 Apatite crystal structure data for data dimensionality reduction. The different apatites are taken along rows and the descriptors are built along the columns. In matrix nomenclature, the final dataset is a 25x29 matrix. Ionic radii, electronegativity, geometric parameters and total energy term are the descriptors used to quantify the apatite chemistry.

Linear and non-linear data dimensionality reduction techniques are employed to study the similarity between the apatite crystal chemistries. By visualizing the chemistries in the reduced manifold, critical crystal structural patterns are discovered.

3. Data dimensionality reduction

The two data mining algorithms used in this analysis are the linear principal component analysis (PCA) and non-linear isometric mapping (IsoMap). The data is first pre-processed by auto-scaling. The auto-scaled data becomes the input for the linear and non-linear manifold learning. The constructed database includes chemistries that contain toxic elements such as mercury, lead, cadmium, arsenic etc. The main purpose of the analysis is to search for chemistries that can substitute for these toxic elements favorably by discovering interesting crystal structural patterns in the apatites chemistries.

The algorithmic details are not presented here and the readers can refer to the articles for PCA and IsoMap in the reference [6-8]. For IsoMap, the analysis was done using the Euclidean metric for calculating the geodesic distances and a value of $K=4$ were used to construct the nearest neighborhood graph.

3.1. Low dimensional embedding using IsoMap

Figure 3 compares the residual variance obtained using PCA and IsoMap. The residual variances were normalized so that a direct comparison can be made between the two. Clearly, the residual variance decreases as the number of dimensions increases. The easiest way to obtain the intrinsic dimensionality will be to locate the “elbow” in the curve. Addition of an extra dimension beyond the “elbow” in the curve will not provide any significant information and hence can be ignored. From fig 3, as with PCA the true dimensionality is five, whereas with IsoMap, the “elbow” in the curve is attained for three dimensions.

The intrinsic dimensionality of the dataset is three, since the major source of crystal structural variations are brought about by changing A-site, B-site and X-site chemistries. Therefore, IsoMap reproduces the intrinsic dimensionality of the dataset whereas PCA over-estimates it. As a result, only the low dimensional embedding of the IsoMap will be considered for further analysis. By visualizing the apatite chemistries in the reduced dimensional space, distinct features can be discovered which would further aid in giving critical insights for designing new materials for sequestering toxic materials. Figure 4 shows the 3D plot containing the low dimensional embedding of all the apatite chemistries obtained from IsoMap computation. The purpose of these 3D plots is to visualize the apatite chemistries and identify compounds that show similar crystal structural

characteristics. Therefore, by locating chemistries that are close to those containing toxic elements such as Hg, Pb, Cd etc, interesting patterns can be discovered and this knowledge can be used to design new chemistries for sequestering environmentally toxic elements.

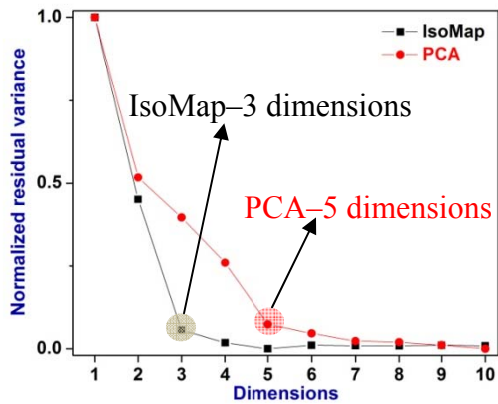


Figure 3 Residual variance of PCA (filled circle) and IsoMap (filled square). Linear PCA captures five dimensions and non-linear IsoMap captures three dimensions as the approximate intrinsic dimensionality of the dataset. Since the apatite database was constructed by varying the chemistries along A-site, B-site and X-site (three crystallographic sites) it is concluded that, IsoMap reproduces the original intrinsic dimensionality of the apatite dataset.

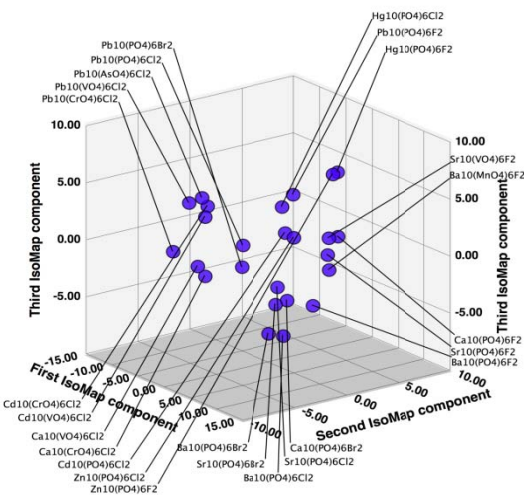


Figure 4 Low dimensional embedding of high dimensional complex apatite data using IsoMap. 3D map showing the complex crystal chemical interactions of apatites are visualized. A 29 dimensional dataset is reduced to only three dimensions. The labels in the figure are the apatite chemistries included in the database for analysis. Compounds that are located closer to one another can be considered to show a significant level of similarity in the crystal structure.

4. K-means clustering

A number of interesting features can be identified using the 3D plot shown in fig 4. There are primarily two broad ways to detect patterns in spatially distributed array of points. The first is a crude approach where we visually cluster the data points and the other approach is to formally detect clusters by applying well-defined cluster analysis methodologies. In this paper we have applied K-means clustering algorithm [9] to detect compact clusters of apatites in the low dimensional embedding obtained from IsoMap. That is, the input to the K-means clustering algorithm is the output from the IsoMap analysis. Cluster analysis is performed in the L2-norm Euclidean space. The optimum number of clusters is found by exploring the evolution of silhouette values for each value of chosen “K”. Based on this procedure, the optimal “K” was reached for K=7 clusters.

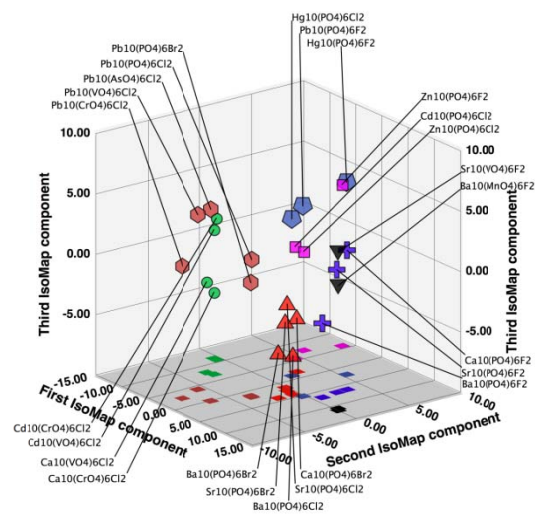


Figure 5 Evolution of patterns containing crystal structurally similar apatites based on k-means clustering. 3D IsoMap plot showing different clusters of apatites are visually presented based on the result derived from K-means clustering algorithm with K=7. Apatite chemistries belonging to the same cluster have similar color or similar symbol representation.

Figure 5 shows the result of k-means cluster analysis in the form of a 3D plot. Different apatite chemistries are grouped based on the crystal structural similarity. For the ease of visualization, apatites belonging to the same cluster are either represented with similar color or similar symbol notation. The nature of clustered chemistries is given below, where every bullet is a separate cluster of apatites:

1. $\text{Ca}_{10}(\text{PO}_4)_6\text{Br}_2$, $\text{Sr}_{10}(\text{PO}_4)_6\text{Cl}_2$, $\text{Ba}_{10}(\text{PO}_4)_6\text{Br}_2$,
 $\text{Ba}_{10}(\text{PO}_4)_6\text{Cl}_2$, $\text{Sr}_{10}(\text{PO}_4)_6\text{Br}_2$
2. $\text{Ba}_{10}(\text{PO}_4)_6\text{F}_2$, $\text{Ca}_{10}(\text{PO}_4)_6\text{F}_2$, $\text{Sr}_{10}(\text{PO}_4)_6\text{F}_2$
3. $\text{Ba}_{10}(\text{MnO}_4)_6\text{F}_2$, $\text{Sr}_{10}(\text{VO}_4)_6\text{F}_2$
4. $\text{Zn}_{10}(\text{PO}_4)_6\text{F}_2$, $\text{Zn}_{10}(\text{PO}_4)_6\text{Cl}_2$, $\text{Cd}_{10}(\text{PO}_4)_6\text{Cl}_2$
5. $\text{Hg}_{10}(\text{PO}_4)_6\text{F}_2$, $\text{Pb}_{10}(\text{PO}_4)_6\text{F}_2$, $\text{Hg}_{10}(\text{PO}_4)_6\text{Cl}_2$
6. $\text{Pb}_{10}(\text{PO}_4)_6\text{Br}_2$, $\text{Pb}_{10}(\text{PO}_4)_6\text{Cl}_2$, $\text{Pb}_{10}(\text{AsO}_4)_6\text{Cl}_2$,
 $\text{Pb}_{10}(\text{VO}_4)_6\text{Cl}_2$, $\text{Pb}_{10}(\text{CrO}_4)_6\text{Cl}_2$
7. $\text{Cd}_{10}(\text{CrO}_4)_6\text{Cl}_2$, $\text{Cd}_{10}(\text{VO}_4)_6\text{Cl}_2$, $\text{Ca}_{10}(\text{VO}_4)_6\text{Cl}_2$,
 $\text{Ca}_{10}(\text{CrO}_4)_6\text{Cl}_2$

The most common feature associated with cluster 1 is the PO_4 tetrahedron with Br and Cl occupying the X-site. Chemically, Br and Cl are very similar in characteristics. Both have a comparable ionic radii ($r_{\text{Cl}} = 1.81\text{\AA}$ and $r_{\text{Br}} = 1.96\text{\AA}$) and similar electronegativity (Pauling's electronegativity of Cl = 3.16 and Br = 2.96). Ca, Sr and Ba belong to the group 2 of the periodic table. In cluster 2, the A-site (Ca, Sr and Ba) and B-site (P) chemistries are similar but the change is with respect to the X-site. With respect to cluster 3, the A-site has group 2 elements and X-site has F, whereas the B-site is Mn and V. Mn and V belong to period 4 of the transition series elements in the periodic table with very distinct chemical characteristic differing from P. Cluster 4 contains Zn and Cd apatites with P in the B-site. Electronically Zn and Cd can be considered to be very similar with both Zn^{2+} and Cd^{2+} cations containing d10 outer-shell valence electron configuration. Cluster 5 contains Hg and Pb apatites with P in the B-site. Chemically, Hg and Pb show a greater covalent character (electronegativity of Pb=2.33 and Hg=2.0) and hence, they cluster together. Cluster 6 captures all Pb-based apatites with various B-site and X-site chemistries and cluster 7 captures the crystal structural similarities associated with Ca and Cd containing apatites with Cd and V in the B-site and Cl in the X-site. The ionic sizes of Ca^{2+} and Cd^{2+} are very similar but chemically they are distinct with Ca showing more ionic character and Cd showing more covalent character. However, it is the similarity in the B-site and X-site chemistry that brings about this clustering behavior. Clearly the different clusters capture different characteristic behavior of apatites with respect to the site chemistries.

The combination of IsoMap and K-means clustering have potentially lead to a formalized approach to discover crystal chemically similar compounds in a highly complicated crystal structures such as apatites. The impact of the derived knowledge is vital in identifying new chemistries for sequestering toxic elements. For example, As (arsenic) is a known environmentally toxic element and from the data analysis, it is discovered that As-apatites share a degree of crystal structural similarity with P, V and Cr containing apatites (cluster 6) with Pb as the A-site element. Similarly, cluster 7 identifies the similarity between Ca and toxic Cd apatites. The non-

toxic chemistries can act as a potential host site for immobilizing toxic chemistries.

5. Predictive model using PLS

The framework developed so far addresses the key issue of systematically identifying new chemistries as potentially host lattices for toxic elements. However to definitively suggest new chemistries for practice, it is important to test the new chemistries based on a figure of merit such as the thermochemical data (eg., Heat of formation, solubility product etc). By computing the thermochemical data, say heat of formation, of the participating chemistries, the “energy cost” of the final structure can be quantitatively established. In this paper, heat of formation data is used as the metric to determine the “energy cost” and a first-order quantitative-structure-property-relationship (QSPR) model (analogous to drug discovery research) is developed using partial least squares methods (PLS) to predict the heat of formation of new apatite chemistries. The heat of formation data of the apatites is taken from the published literature [10]. Ionic radii, electronegativity and lattice constants were taken as the independent variables. The choice of the variables was based on our previous research experience in modeling chemical crystallography of complex inorganic solids [11]. A database containing the heat of formation data of fourteen apatites with eight independent descriptors is constructed. A PLS model was developed to predict heat of formation as a function of eight descriptors (Ionic radii of AI-site- r_{AI} and AII-site- r_{AII} , ionic radii of X-site- r_{X} , average A-site electronegativity- A_{EN} , average X-site electronegativity- X_{EN} , lattice constants- a , c and c/a). Twelve apatites were used for training the model and two were used for testing the resulting model. The model was cross-validated by leave-one-out approach. The algorithmic details of PLS is not presented here and the readers can refer to the articles in the references [7].

Figure 6 compares the experiment vs. predicted heat of formation data of apatites. The R^2 value of 0.92 was obtained and the predictions for the test chemistries were reasonable. Based on the QSPR model, new heat of formation data for some of the apatites are predicted thereby quantifying the “energy cost” of some of the apatite chemistries for which the heat of formation data is not available in the literature. The accuracy of the developed model is high for simple apatite chemistries while the model can be extended to predict the energy-cost of complex chemistries. Comparing the weights of the descriptors derived from PLS (fig. 6), the c/a ratio dominates the model with the highest relative magnitude followed by average A-site electronegativity A_{EN} . The lattice constants data for the apatites are taken from the published work of White et al [1].

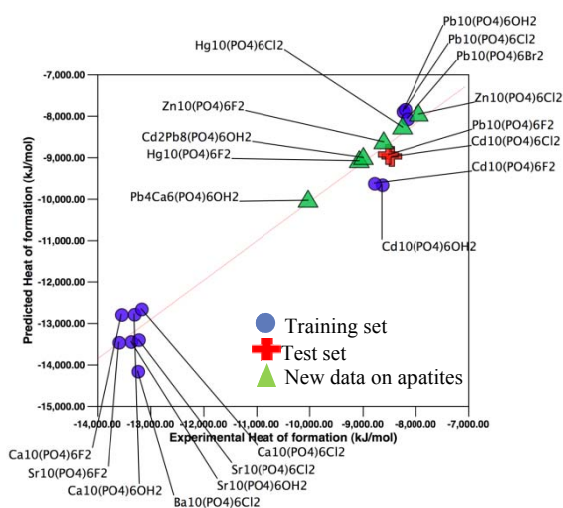


Figure 6 Predicted vs. Experiment heat of formation of apatites. The developed QSPR model is, heat of formation (energy cost) ΔH°_f (kJ/mol) = $-1599.01r_{AI} - 1651.31r_{AII} + 442.1066r_X + 3837.086A_{EN} - 835.273X_{EN} - 158.31a - 257.941c - 100053.7c/a - 11.868$ where, r_{AI} -ionic radii of Al-site, r_{AII} -ionic radii of AII-site, r_X -ionic radii of X-site, average A-site electronegativity- A_{EN} , X-site electronegativity- X_{EN} , lattice constants-a, c and c/a.

6. Conclusion

By applying a combination of data mining techniques the complexity of interactions of elements in industrially relevant crystal structures such as apatites is captured. The non-linear topology of the data structure was discovered from applying IsoMap which captures the underlying degree of chemical freedom thereby reducing the 29 dimensional dataset to only three dimensions. Interesting clusters are detected in the low dimensional embedding to identify new apatite chemistries for sequestering toxic elements (for example, the similarity observed between Cd and Ca-apatites). A quantitative model was developed to not only predict suitable host chemistries but also to identify the best chemistry from the energy cost perspective. While more data is needed, there are a number of data mining challenges to be addressed such as dealing with sparse data, simulating data in cases where getting access to experimental information is not trivial, integrating linear and non-linear data mining techniques to reliably explore more materials and suggest new chemistries for immobilizing environmentally toxic elements.

7. Acknowledgements

The authors acknowledges support from National Science Foundation: International Materials Institute Program for the Combinatorial Sciences and Materials Informatics

Collaboratory (CoSMIC-IMI), grant number DMR--08-33853, NSF-ARI Program: grant # CMMI 09-389018; NSF-CDI Type II program: grant # PHY 09-41576; the Air Force Office of Scientific Research, grant # FA95500810316 and FA95500610501; the; and DARPA grant # DARPA Centre for Interfacial Engineering for MEMS (CIEMS) grant no. 1891874036790B.

8. References

- [1] T. White, C. Ferraris, J. Kim and S. Madhavi, "Apatite – an adaptive framework structure", *Reviews in Mineralogy & Geochemistry* **57**, 2005, pp. 307-40.
- [2] P.H.J. Mercier et al, "Geometrical parameterization of the crystal chemistry of $P6_3/m$ apatites: comparison with experimental data and ab initio results", *Acta Crystallographica B* **61**, 2005, pp. 635-655.
- [3] T.J. White and D. ZhiLi, "Structural derivation and crystal chemistry of apatites", *Acta Crystallographica B* **59**, 2003, pp. 1-16.
- [4] J.Y. Kim, Zhili Dong and T.J. White, "Model apatite systems for the stabilization of toxic metals: II, cation and metalloid substitutions in chlorapatite", *Journal of American Ceramic Society*, **88**, 2005, pp. 1253-1260.
- [5] A.V. Shevade, L. Erickson, G. Pierzynski and S. Jiang, "Formation and stability of substituted pyromorphite: a molecular modeling study", *Journal of Hazardous Substance Research* **3**, 2001, pp. 1-12.
- [6] M. Ringnér. "What is Principal Component Analysis?", *Nature Biotechnology* **26**. 2008, pp. 303-304.
- [7] L. Ericksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Multi- and megavariate data analysis: principles, applications*, Umea, Sweden: Umetrics, 2001.
- [8] J.B. Tenenbaum, V. de Silva and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, **290**, 2000, pp. 2319-2323.
- [9] P-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, US, 2005.
- [10] N.J. Flora and C.H. Yoder, "Lattice energies of apatites and the estimation of ΔH°_f (PO_4^{3-} , g)", *Inorganic Chemistry* **43**, 2004, pp. 2340-2345.
- [11] K. Rajan, *Data Mining and Inorganic Crystallography* in *Data Mining in Crystallography-Structure and Bonding Series* - D.W.M. Kuleshova, N. Liudmila, Eds. vol. 134. in press (Springer-Verlag, 2010).