Integrating Materials and Manufacturing Innovation
a SpringerOpen Journal

RESEARCH Open Access

A software framework for data dimensionality reduction: application to chemical crystallography

Sai Kiranmayee Samudrala[1*], Prasanna Venkataraman Balachandran[2], Jaroslaw Zola[3], Krishna Rajan[4*] and Baskar Ganapathysubramanian[5*]

*Correspondence:
ssamudrala@me.gatech.edu;
krajan@iastate.edu;
baskarg@iastate.edu
[1]School of Mechanical Engineering, Georgia Tech, Atlanta, GA 30332-0405, USA
[4]Department of Materials Science and Engineering, Iowa State University, Ames, IA 50011, USA
[5]Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA
Full list of author information is available at the end of the article

Abstract

Materials science research has witnessed an increasing use of data mining techniques in establishing process-structure-property relationships. Significant advances in high-throughput experiments and computational capability have resulted in the generation of huge amounts of data. Various statistical methods are currently employed to reduce the noise, redundancy, and the dimensionality of the data to make analysis more tractable. Popular methods for reduction (like principal component analysis) assume a linear relationship between the input and output variables. Recent developments in non-linear reduction (neural networks, self-organizing maps), though successful, have computational issues associated with convergence and scalability. Another significant barrier to use dimensionality reduction techniques in materials science is the lack of ease of use owing to their complex mathematical formulations. This paper reviews various spectral-based techniques that efficiently unravel linear and non-linear structures in the data which can subsequently be used to tractably investigate process-structure-property relationships. In addition, we describe techniques (based on graph-theoretic analysis) to estimate the optimal dimensionality of the low-dimensional parametric representation. We show how these techniques can be packaged into a modular, computationally scalable software framework with a graphical user interface - Scalable Extensible Toolkit for Dimensionality Reduction (SETDiR). This interface helps to separate out the mathematics and computational aspects from the materials science applications, thus significantly enhancing utility to the materials science community. The applicability of this framework in constructing reduced order models of complicated materials dataset is illustrated with an example dataset of apatites described in structural descriptor space. Cluster analysis of the low-dimensional plots yielded interesting insights into the correlation between several structural descriptors like ionic radius and covalence with characteristic properties like apatite stability. This information is crucial as it can promote the use of apatite materials as a potential host system for immobilizing toxic elements.

Keywords: Non-linear dimensionality reduction; Process-structure-property; Apatites; Materials science; High-throughput analysis

Background

Using data mining techniques to probe and establish process-structure-property relationships has witnessed a growing interest owing to its ability to accelerate the process of tailoring materials by design. Before the advent of data mining techniques, scientists used a variety of empirical and diagrammatic techniques [1], like pettifor maps [2], to establish relationships between structure and mechanical properties. Pettifor maps, one of the earliest graphical representation techniques, is exceedingly efficient except that it requires a thorough understanding and intuition about the materials. Recent progress in computational capabilities has seen the advent of more complicated paradigms - so-called virtual interrogation techniques - which span from first-principles calculations to multi-scale models [3-7]. These complex multi-physics and/or statistical techniques and simulations [8,9] result in an integrated set of tools which can predict the relationships between chemical, microstructural, and mechanical properties producing an exponentially large collection of data. Simultaneously, experimental methods - combinatorial materials synthesis [10,11], high-throughput experimentation, atom probe tomography - allow synthesis and screening of a large number of materials while generating huge amounts of multivariate data.

A key challenge is then to efficiently probe this large data to extract correlations between structure and property. This data explosion has motivated the use of data mining techniques in materials science to explore, design, and tailor materials and structures. A key stage in this process is to *reduce the size of the data, while minimizing the loss of information during this data reduction*. This process is called data dimensionality reduction. By definition, dimensionality reduction (DR) is the process of reducing the dimensionality of the given set of (usually unordered) data points and extracting the low-dimensional (or parameter space) embedding with a desired property (for example, distance, topology, etc.) being preserved throughout the process. Examples for DR methods are principal component analysis (PCA) [12], Isomap [13], Hessian locally linear embedding (hLLE) [14], etc. Applying DR methods enables visualization of the high-dimensional data and also estimates the optimal number of dimensions required to represent the data without considerable loss of information. Additionally, burgeoning cyberinfrastructure-based tools and collaborations sustained by the government's recent Materials Genome Initiative (MGI) provides a great platform to leverage the data dimensionality reduction tools. This will enable integration of information obtained from the individual high-throughput simulations and experimentation efforts in various domains (e.g., mechanical, electrical, electro-magnetic, etc.) and at multiple length-scales (macro-meso-micro-nano) in a fashion as never seen before [15].

Data dimensionality reduction is not a novel concept. Page [16] describes different techniques of data reduction and their applicability for establishing process-structure-property relationships. Statistical methods like PCA [17] and factor analysis (FA) [18] have been used on materials data generated by first-principles calculations or by experimental methods. However, dimensionality reduction techniques like PCA or factor analysis to establish process-structure-property relationships traditionally assume a linear relationship among the variables. This is often not strictly valid; the data usually lies on a non-linear manifold (or surface) [13,19]. Non-linear dimensionality reduction (NLDR) techniques can be applied to unravel the non-linear structure from unordered data. An example of such application for constructing a low-dimensional stochastic representation

of property variations in random heterogenous media is [19]. Another exciting application of data dimensionality reduction is in combination with quantum mechanics-based calculations to predict the structure [20-22]. For a more mathematical list of linear and non-linear DR techniques, the interested reader can consult [23,24].

In this paper, the theory and mathematics behind various linear and non-linear dimensionality reduction methods is explained. The mathematical aspects of dimensionality reduction are packaged into an easy-to-use software framework called Scalable Extensible Toolkit for Dimensionality Reduction (SETDiR) which (a) provides a user-friendly interface that successfully abstracts user from the mathematical intricacies, (b) allows for easy post-processing of the data, and (c) represents the data in a visual format and allows the user to store the output in standard digital image format (eg: JPEG), thus making data more tractable and providing an intuitive understanding of the data. We conclude by applying the techniques discussed on a dataset of apatites [25-29] described using several structural descriptors. This paper is seen as an extension of our recent work [30]. Apatites ($A_4^{\mathrm{I}}A_6^{\mathrm{II}}(BO_4)_6X_2$) have the ability to accommodate numerous chemical substitutions and hence represent a unique family of crystal chemistries with properties catering many technological applications, such as toxic element immobilization, luminescence, and electrolytes for intermediate temperature solid oxide fuel cells, to name a few [25-29].

The outline of the paper is as follows: The section 'Methods: dimensionality reduction' briefly describes the concepts of DR, algorithms, and the dimensionality estimators that can be used to estimate the dimensionality. The software framework, SETDiR, developed to apply DR techniques is described in the section 'Software: SETDiR'. The section 'Results and discussion' discusses the interpretation of low-dimensional results obtained by applying SETDiR to the apatite dataset.

Methods: dimensionality reduction

The problem of dimensionality reduction can be formulated as follows. Consider a set of data, X. This set consists of n data points, x_i. Each of the data points x_i is vectorized to form a 'column' vector of size D. Usually, D is large. Thus, $X = \{x_0, x_1, \ldots, x_{n-1}\}$ of n points, where $x_i \in \mathbb{R}^D$ and $D \gg 1$. Visualizing and analyzing correlations, patterns, and connections within high-dimensional dataset is difficult. Hence, we are interested in finding a set of *equivalent low-dimensional points*, $Y = \{y_0, y_1, \ldots, y_{n-1}\}$, that exhibit the same correlations, patterns, and connections as the high-dimensional data. This is mathematically posed as

Find $Y = \{y_0, y_1, \ldots, y_{n-1}\}$, such that $y_i \in \mathbb{R}^d$, $d \ll D$ and $\forall_{i,j}\ |x_i - x_j|_h = |y_i - y_j|_h$. Here, $|a-b|_h$ denotes a specific norm that captures properties, connections, or correlations we want to preserve during dimensionality reduction [23].

For instance, by defining h as Euclidean norm, we preserve Euclidean distance, thus obtaining a reduction equivalent to the standard technique of PCA [12]. Similarly, defining h to be the angular distance (or conformal distance [31]) results in locally linear embedding (LLE) [32] that preserves local angles between points. In a typical application [33,34], x_i represents a state of the analyzed system, e.g., temperature field, concentration distribution, or characteristic properties of a system. Such state description can be derived from experimental sensor data or can be the result of a numerical simulation. However, irrespective of the source, it is characterized by high dimensionality, that is D is typically of the order of 10^2 to 10^6 [35,36]. While x_i represents just a single state of

the system, contemporary data acquisition setups deliver large collections of such observations, which correspond to the temporal or parametric evolution of the system [33]. Thus, the cardinality n of the resulting set X is usually large ($n \sim 10^2$ to 10^5). Intuitively, information obfuscation increases with the data dimensionality. Therefore, in the process of DR, we seek as small a dimension d as possible, given the constraints induced by the norm $|a-b|_h$ [23]. Routinely, $d < 4$ as it permits, for instance, visualization of the set Y.

The key mathematical idea underpinning DR can be explained as follows: We encode the desired information about X, i.e., topology or distance, in its entirety by considering all pairs of points in X. This encoding is represented as a matrix $A_{n\times n}$. Next, we subject matrix A to unitary transformation V, i.e., transformation that preserves the norm of A (thus, preserving connectivities and correlations in the data), to obtain its sparsest form Λ, where $A = V\Lambda V^T$. Here, $\Lambda_{n\times n}$ is a diagonal matrix with rapidly diminishing entries. As a result, it is sufficient to consider only a small, d, number of entries of Λ to capture all the information encoded in A. These d entries constitute the set Y. The above procedure hinges on the fact that unitary transformations preserve original properties of A [37]. Note also, that it requires a method to construct matrix A in the first place. Indeed, what differentiates various spectral data dimensionality methods is the way information is encoded in A.

We focus on four different DR methods: (a) PCA, a linear DR method; (b) Isomap, a non-linear isometry-preserving DR method; (c) LLE, a non-linear conformal-preserving DR method; and (d) Hessian LLE, a topology-preserving DR method.

Principal component analysis

PCA is a powerful and a popular DR strategy due to its simplicity and ease in implementation. It is based on the premise that the high-dimensional data is a linear combination of a set of hidden low-dimensional axes. PCA then extracts the latent parameters or low-dimensional axes by reorienting the axes of the high-dimensional space in such a way that the variance of the variables is maximized [23].

PCA algorithm

1. Compute the pair-wise Euclidean distance for all points in the input data X. Store it as a matrix $[E]$.
2. Construct a matrix $[W^*]$ such that the elements of $[W^*]$ are -0.5 times the square of the elements of the euclidean distance matrix $[E]$.
3. Find the dissimilarity matrix $[A]$ by double centering $[W^*]$:

$$[A] = \left[H^T\right]\left[W^*\right][H] \tag{1}$$

$$H_{ij} = \begin{cases} (1-1/n)\ \forall\ \ \mathbf{i} = \mathbf{j}, \\ (-1/n)\ \forall\ \ \mathbf{i} \neq \mathbf{j}. \end{cases} \tag{2}$$

4. Solve for the largest d eigenpairs of $[A]$:

$$[A] = [U][\Lambda]\left[U^T\right]. \tag{3}$$

5. Construct the low-dimensional representation in $\mathbb{R}^d$ from the eigenpairs:

$$[Y] = [I][\Lambda]^{1/2}\left[U^T\right]. \tag{4}$$

The functionality of the identity matrix is to extract the most important d-dimensions from the eigenpairs of $[A]$.

The limitation of PCA is that it assumes the data lies on a linear space and hence performs poorly on the data that are inherently non-linear. In these cases, PCA also tends to over-estimate the dimensionality of the data.

Isomap

Isomap relaxes the assumption of PCA that the data lies on a linear space. A classic example of a non-linear manifold is the Swiss roll. Figure 1 shows how PCA tries to fit the best linear plane while Isomap unravels the low-dimensional surface. Isomap essentially smooths out the non-linear manifold into a corresponding linear space and subsequently applies PCA. This smoothing out can intuitively be understood in the context of the spiral, where the ends of the spiral are pulled out to straighten the spiral into a straight line. Isomap accomplishes this objective mathematically by ensuring that the geodesic distance between data points are preserved under transformations. The geodesic distance is the distance measured along the curved surface on which the points rest [23]. Since it preserves (geodesic) distances, Isomap is an isometry (distance-preserving) transformation. The underlying mathematics of the Isomap algorithm assumes that the data lies on a manifold which is convex (but not necessarily linear). Note that both PCA and Isomap are isometric mappings; PCA preserves pair-wise Euclidean distances of the points while Isomap preserves the geodesic distance.

Isomap algorithm

1. Compute the pair-wise Euclidean distance matrix $[E]$ from the input data X.
2. Compute the k-nearest neighbors of each point from the distance matrix $[E]$.
3. Compute the pair-wise geodesic distance matrix $[G]$ from $[E]$. This is done using Floyd's algorithm [38].
4. Construct a matrix $[W^*]$ such that the elements of $[W^*]$ are -0.5 times the square of the elements of the geodesic distance matrix $[G]$.
5. Find the dissimilarity matrix $[A]$ by double centering $[W^*]$:

$$[A] = \left[H^T\right]\left[W^*\right][H] \tag{5}$$

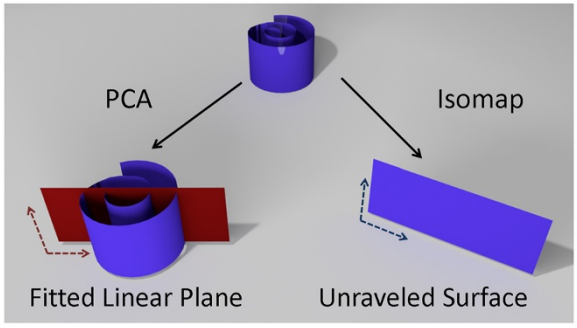

Figure 1 Comparison of performance of PCA and Isomap on a dataset lying on non-linear manifold.

$$H_{ij} = \begin{cases} (1 - 1/n) \ \forall \ \ \mathbf{i} = \mathbf{j}, \\ (-1/n) \ \forall \ \ \mathbf{i} \neq \mathbf{j}. \end{cases} \tag{6}$$

6. Solve for the largest d eigenpairs of A:

$$[A] = [U]\,[\Lambda]\left[U^T\right]. \tag{7}$$

7. Construct the low-dimensional representation in $\mathbb{R}^d$ from the eigenpairs:

$$[Y] = [I]\,[\Lambda]^{1/2}\left[U^T\right]. \tag{8}$$

The non-linearity in the data is accounted for by using geodesic distance metric. The graph distance is used to approximate the geodesic distance [39]. Graph distance between a pair of points in a graph (V, E) is the shortest path connecting the two given points. The graph distances are calculated using the well-known Floyd's algorithm [38].

Locally linear embedding

In contrast to PCA and Isomap methods which preserve distances, LLE preserves the local topology (or local orientation, or angles between data points). LLE uses the notion that locally a non-linear manifold (or curve) is well-approximated by a linear curve. In other words, the manifold is locally linear and hence can be represented as a patchwork of linear curves. The algorithm first divides the manifold into patches and reconstructs each point in the patch based on the information (or weights) obtained from its neighbors (i.e., infer how a specific point is located with respect to its neighbors). This process extracts the local topology of the data. Finally, the algorithm reconstructs the global structure by combining individual patches and finding an optimized, low-dimensional representation. Numerically, local topology information is constructed by finding the k-nearest neighbors of each data point and reconstructing each point from the information about the weights of the neighbors. The global reconstruction from the local patches is accomplished by assimilating the individual weight matrices to form a global weight matrix $[W]$ and evaluating the smallest eigenvalues of normalized global weight matrix $[A]$.

LLE algorithm

1. For each of the n input vectors from $X = \{x_0, x_1, \ldots, x_{n-1}\}$:
   (a) Find the k-nearest neighbors of the data point x_i.
   (b) Construct the local covariance or Gram matrix $\mathbf{G_i}$

$$g_{r,s}(i) = (x_i - x_r)^T (x_i - x_s) \tag{9}$$

   where x_r and x_s are neighbors of x_i.
   (c) Weight vector, w_i is computed by solving the linear system:

$$\mathbf{G_i} w_i = \mathbf{1} \tag{10}$$

   where $\mathbf{1}$ is a $k \times 1$ vector of ones.
2. Using the vectors w_i, build the sparse matrix W. The (i, j) of W is zero if x_i and x_j are not neighbors. If x_i and x_j are neighbors, then $W(i, j)$ takes the values of the corresponding with vector, $w_i(j)$.
3. From W, build A:

$$[A] = (I - W)^T (I - W). \tag{11}$$

4. Compute the eigenpairs (corresponding to the smallest eigenvalues) for A:
$$[A] = [U]\,[\Lambda]\left[U^T\right]. \tag{12}$$
5. Compute the low-dimensional points in $\mathbb{R}^d$ from the smallest eigenpairs.

Hessian LLE

Hessian LLE [14] (hLLE) is a modification of LLE and Laplacian Eigenmaps [40]. Mathematically, hLLE replaces the Laplacian (first derivative) operator with a Hessian (second derivative) operator over the graph. hLLE constructs patches, performs a local PCA on each patch, constructs a global Hessian from the eigenvectors thus obtained, and finally finds the low-dimensional representation from the eigenpairs of the Hessian. hLLE is a topology preservation method and assumes that the manifold is locally linear.

hLLE algorithm

1. At each given point x_i, construct a $k \times n$ neighborhood matrix M_i such that each row, j, of the matrix represents a point
$$x_j = x_j - \bar{x_i}, \tag{13}$$
where $\bar{x_i}$ is the mean of k neighboring points.
2. Perform singular value decomposition (SVD) of the M_i to obtain the SVD matrices, U, V, D.
3. Construct the $(N * d(d+1)/2)$ local Hessian matrix $[H]^i$ such that the first column is a vector of all ones and the next d columns are the columns of U followed by the products of all the d columns of $[U]$.
4. Compute Gram-Schmidt orthogonalization [37] on the local Hessians $[H]^i$ and assimilate the last $d(d+1)/2$ orthonormal vectors of each to construct the global Hessian matrix $[A]$ [14].
5. Compute the eigenpairs (corresponding to the smallest eigenvalues) of the Hessian matrix:
$$[A] = [W]\,[\Lambda]\,[W]^T. \tag{14}$$
6. Compute the low-dimensional points $[Y]$ in $\mathbb{R}^d$ from the eigenpairs:
$$[Y] = [W]\left([W]^T\,[W]\right)^{-1/2}. \tag{15}$$

An important point to note here is that, as discussed in the section 'Methods: dimensionality reduction', matrix $[A]$ encodes the required information for each of the DR techniques, and the construction of this matrix is what differentiates a spectral DR method from the rest. Matrix $[A]$ is a normalized Euclidean matrix in the case of PCA, a normalized geodesic matrix in the case of Isomap, a normalized Hessian matrix for hLLE, and so on.

Dimensionality estimators

A key step in constructing the low-dimensional points from the data is the choice of the low dimensionality or optimal dimensionality d. Methods like PCA and Isomap have an implicit technique to estimate the low dimensionality (approximately) using scree plots. We introduce a graph-based technique that rigorously estimates the latent dimensionality of the data, which can be used in conjunction with the scree plot.

Dimensionality from the scree plot

Scree plot is a plot of the eigenvalues with the eigenvalues arranged in decreasing order of their magnitude. Scree plots obtained from PCA and Isomap (distance-preserving methods) give an estimate of the dimensionality. A heuristic method of identifying the dimensionality is by identifying the elbow in the scree plot. A more quantitative estimate of dimensionality is estimated by choosing a value for $p_{\text{var}}(d)$ that ensures a threshold of the minimum percentage variability. If $\lambda_1 >= \lambda_2 >= \ldots \lambda_n$ are the individual eigenvalues arranged in descending order, the percentage variability ($p_{\text{var}}(d)$) covered by considering first d eigenvalues is given by

$$p_{\text{var}}(d) = 100 \times \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{16}$$

A usual approach is to choose a d that takes 95% of the variability into account.

Geodesic minimal spanning tree estimator

We have recently utilized a dimensionality estimator based on the BHH theorem (Breadwood-Halton-Hammersley Theorem) [41]. This theorem states that the rate of convergence of the length of minimal spanning tree[a] gives a measure of the latent dimensionality. This theorem allows one to express the dimensionality (d) of an unordered dataset as a function of the length of geodesic minimal spanning tree (GMST) of the graph of the dataset. Specifically, the slope of a $\log(n)$ vs. $\log(L_n)$ plot constructed by calculating the GMST length (L_n) with respect to increasing size of randomly chosen data points (n) provides an estimate of the dimensionality: $d = \frac{1}{(1-m)}$, where m is the slope of the log-log plot [19].

Correlation dimension

Correlation dimension is a space-filling dimension which is derived from a more generic fractal dimension by assigning a value of $q = 2$ in

$$C(\mu, \epsilon) = \int \left[\mu \bar{B_e}(z)\right]^{q-1} d\mu(z) \tag{17}$$

where μ is a Borel probability measure on a metric space $\mathbb{Z}$. $\bar{B_e}(z)$ is a closed ball of radius ϵ centered on z.

Numerical definition of correlation dimension is given by

$$d_{\text{cor}}(\epsilon_1, \epsilon_2) = \frac{\log(\hat{C}_2(\epsilon_2)) - \log(\hat{C}_2(\epsilon_1))}{\log(\epsilon_2) - \log(\epsilon_1)} \tag{18}$$

where $\hat{C}_2(\epsilon_2)$ is a measure of proportion of distances less than ϵ [23,42]. Intuitively, these ϵ values are like window ranges through which one zooms through the data. Too small ϵ will render the data as individual points, while too huge ϵ will make the entire dataset look like a single fuzzy spot. Hence, correlation dimension is sensitive to the epsilon values. One important point to note, however, is that the correlation dimension provides the user with a lower bound of the optimal dimensionality.

Software: SETDiR

These DR techniques are packaged into a modular, scalable framework for ease of use by the materials science community. We call this package, *SETDiR*. This framework contains two major components:

1. Core functionality: developed using C++
2. User interface: developed based on Java (Swing)

Figure 2 describes the scope of the functionality of both modules in SETDiR.

Core functionality

Functionality is developed using object-oriented C++ programming language. It implements the following methods: PCA, Isomap, LLE, and dimensionality estimators like GMST and correlation dimension estimators [23].

User interface

A graphical user interface (shown in Figure 3) is developed using Java™ Swings Components with the following features which make it user-friendly:

1. Abstracts the user from the mathematical and programming details.
2. Displays the results graphically and enhances the visualization of low-dimensional points.
3. Easy post-processing of results: in-built cluster analysis, ability to save plots as image files.
4. Organized settings tabs: Based on the niche of the user, the solver settings are organized as Basic User and Advanced User tabs which abstract a new or a naive user from, otherwise overwhelming, details.

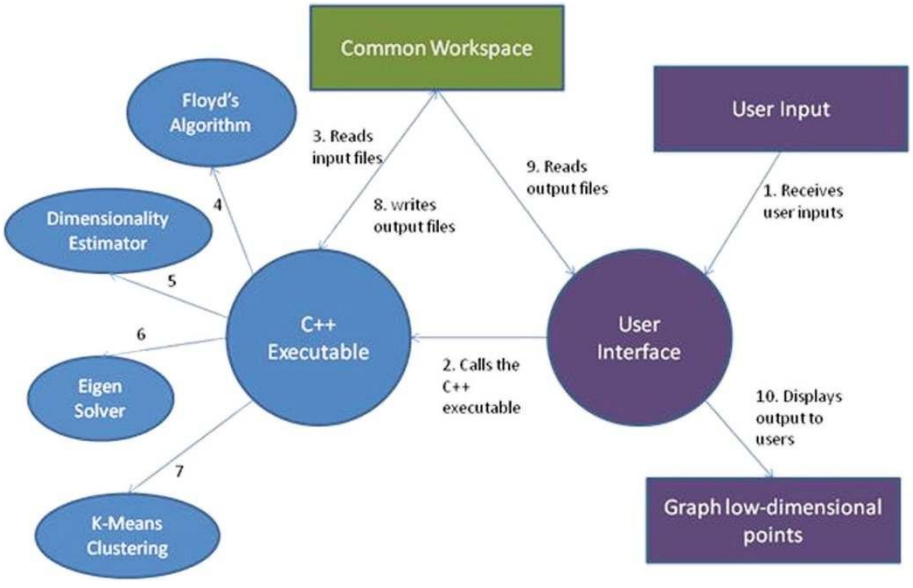

Figure 2 Description of the software.

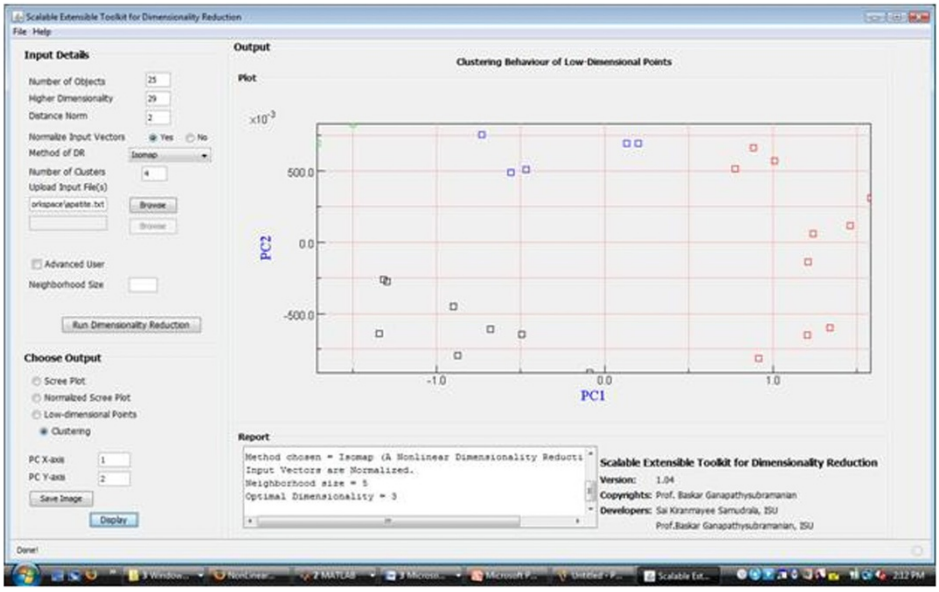

Figure 3 Snapshot of clustering pattern displayed using SETDiR for apatite dataset.

This framework can be downloaded from SETDiR (http://setdir.engineering.iastate.edu/doku.php?id=download). A more detailed discussion of the parallel features of the code is deferred to another publication. We next showcase the framework and the mathematical strategies on the apatite dataset.

Results and discussion

In this section of the paper, we compare and contrast the algorithms on an interesting dataset of apatites with immense technological and scientific significance. Apatites have the ability to accommodate numerous chemical substitutions and exhibit a broad range of multifunctional properties. The rich chemical and structural diversity provides a fertile ground for the synthesis of technologically relevant compounds [25-29]. Chemically, apatites are conveniently described by the general formula $A^{\mathrm{I}}_4 A^{\mathrm{II}}_6 (BO_4)_6 X_2$, where A^{I} and A^{II} are distinct crystallographic sites that usually accommodate larger monovalent (Na^+, Li^+, etc.), divalent (Ca^{2+}, Sr^{2+}, Ba^{2+}, Pb^{2+}, etc.), and trivalent (Y^{3+}, Ce^{3+}, La^{3+}, etc.), *B*-site is occupied by smaller tetrahedrally coordinated cations (Si^{4+}, P^{5+}, V^{5+}, Cr^{5+}, etc.), and the *X*-site is occupied by halides (F^-, Cl^-, Br^-), oxides, and hydroxides. Establishing the relationship between the microscopic properties of apatite complexes with those of the macroscopic properties can help us in gaining an understanding and promote the use of apatites in various technological applications. For example, information about the relative stability of the apatite complexes can promote the utilization of apatites as a suitable host material for immobilizing toxic elements such as lead, cadmium, and mercury (i.e., by identifying an apatite chemical composition that contain at least one of the aforementioned toxic elements and yet remaining thermodynamically stable). DR techniques offer unique insights into the originally intractable high-dimensional datasets by enabling visual clustering and pattern association, thereby establishing process-structure-property relationship for chemically complex solids such as apatites.

Apatite data description

The crystal structure of the aristotype $P6_3/m$ $Ca^{I}_4Ca^{II}_6(PO_4)_6F_2$ apatite with hexagonal unit cell is shown in the Figure 4 with the atoms projected along the (001) axis. The polyhedral representation of $A^{I}O_6$ and BO_4 structural units are clearly shown with the Ca^{II}-site (pink atoms) and F-site (green atoms) occupying the tunnel. Thin black line represents the unit-cell of the hexagonal lattice.

The sample apatite dataset considered consists of 25 different compositions described using 29 structural descriptors. These structural descriptors, when modified, affect the crystal structure [44]. Therefore, by establishing the relationship between the crystal structure and these structural descriptors and analyzing the clustering of different compositions, conclusions can be drawn about how the changes in these structural descriptors (defining the atomic features) could affect the macroscopic properties (such as elastic modulus, band gap, and conductivity). The bond length, bond angle, lattice constants, and total energy data are taken from the work of Mercier et al. [26]; the ionic radii data are taken from the work of Shannon [45] and the electronegativity data is based on the Pauling's scale [46]. The ionic radii of A^{I}-site ($r_{A^{I}}$) has a coordination number nine and A^{II}-site ($r_{A^{II}}$) has a coordination number seven (when the X-site is F^-) or eight (when the X-site is Cl^- or Br^-). Our database describes Ca, Ba, Sr, Pb, Hg, Zn, and Cd in the A-site; P, As, Cr, V, and Mn in the B-site; and F, Cl, and Br in the X-site. The 25 compounds considered in this study belong to the aristotype $P6_3/m$ hexagonal space group. We utilize SETDiR on the apatite data and present some of the results below. More information regarding the source of the apatite data can be found in [44]. A preliminary analysis (focusing only on PCA) can be found in [30].

Figure 4 Crystal structure of a typical $P6_3/mCa^{I}_4Ca^{II}_6(PO_4)_6F_2$ apatite with hexagonal unit cell [43,44].

Dimensionality estimation

SETDiR first estimates the dimensionality using the scree plot. A scree plot is a plot of eigenvalue indices vs. eigenvalues. The occurrence of an elbow (or a sharp drop in eigenvalues) in a scree plot gives the estimate of the dimensionality of the data. Figure 5 displays the scree plots when the input vectors $\{x_0, x_1, \ldots, x_{n-1}\}$ were normalized with respect to that when they were not normalized. We plot for comparison the eigenvalues that are obtained from both PCA and Isomap. This plot shows how the second eigenvalue collapses to zero when the input vectors are not normalized and hence emphasizes the importance of normalization of input vectors[b]. It is also interesting to compare the eigenvalues of PCA and Isomap for normalized input: PCA being a linear method overestimates the dimensionality as 5, while Isomap estimates it to be 3. SETDiR subsequently uses the geodesic minimal spanning tree method to estimate the dimensionality of the apatite data. This method gives a rigorous estimate of 3 (Figure 6), which matches the outcome of the more heuristic scree plot estimate.

Low-dimensional plots

In this section, we discuss the visual interpretation of the low-dimensional plots obtained by applying the dimensionality reduction techniques - PCA, Isomap, LLE, and hLLE - to a set of apatites described using structural descriptors. Figure 7 (left) shows the 2D plot between principal components 2 and 3. The reason for showing this plot is that PC2-PC3 map captures pattern that is similar to Isomap components 1 and 2. While we find associations among compounds that are similar to those as shown in Figure 7 (right), the nature of information is manifested differently. This is mainly attributed to the differences in the underlying mathematics of the two techniques, where PCA is essentially a linear technique and, on the other hand, Isomap is a non-linear technique. To further interpret the hidden information captured by Isomap classification map (Figure 7), we have focused on the three regions separately.

Figure 7 (right) shows a two-dimensional classification map with isomap components 1 and 2 in the orthogonal axes. The two-dimensional classification map groups various apatite compounds into three distinct regions that capture various interactions between

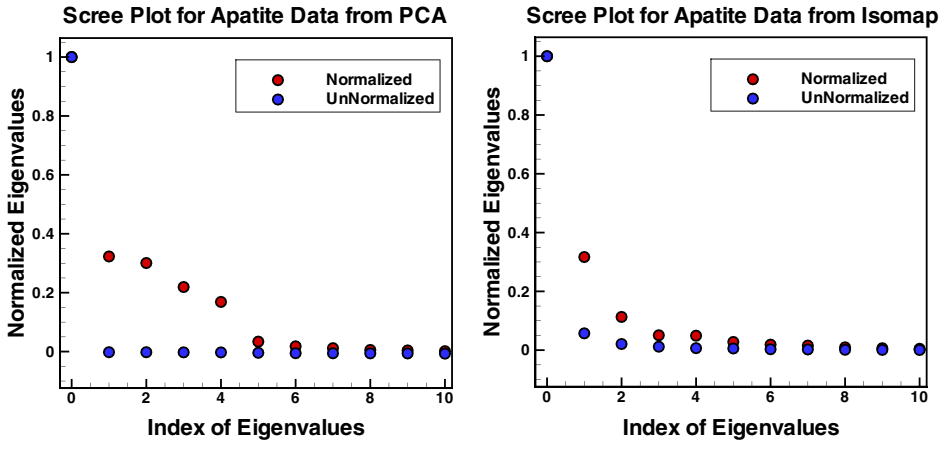

Figure 5 Scree plots for PCA and Isomap - normalized vs. unnormalized input.

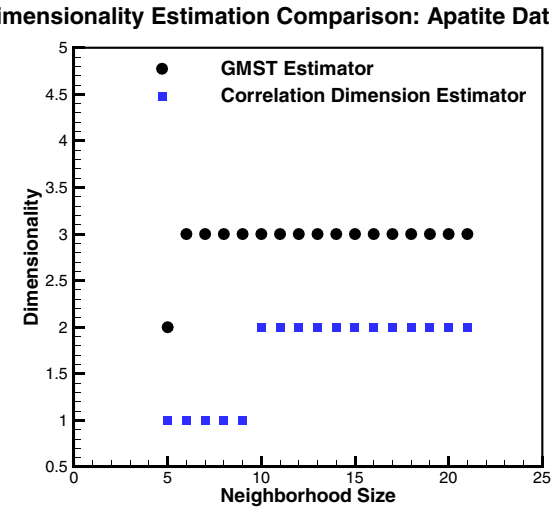

Figure 6 Sensitivity analysis of dimensionality estimators with change in neighborhood size.

A-, *B*-, and *X*-site ions in complex apatite crystal structure. Region 1 corresponds to apatite compounds with fluoride (F) ion in the *X*-site. All apatite compounds in this region contain only F in the *X*-site but has different *A*-site (Ca, Sr, Pb, Ba, Cd, Zn) and *B*-site elements (P, Mn, V). Therefore, this unique region classifies F-apatites from Cl and Br-apatites. Region 2 belongs to apatite compounds with phosphorus (P) ion in the *B*-site and contains Cl and Br ions in the *X*-site. The uniqueness of this region is manifested mainly due to the presence of only smaller P ions in the *B*-site. Similarly, region 3 belongs to apatite compounds with Cl ions in the *X*-site and contains larger *B*-site Cr, V, and As cations.

Figure 8 (right) presents the results from hLLE. It can be observed that the compounds that have highly covalent *A*-site cation (e.g., Hg^{2+} and Pb^{2+}) and highly covalent *B*-site cation (P^{5+}) clearly separate out from the rest. An exception to this rule is $Pb_{10}(CrO_4)_6Cl_2$. Our PCA-derived structure map also revealed similar pattern - i.e.,

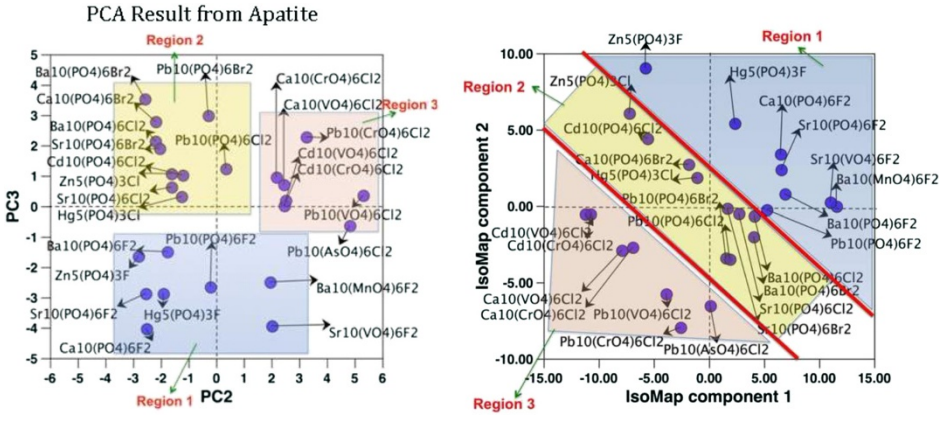

Figure 7 Apatite PCA (left) and Isomap (right) result interpretation [43].

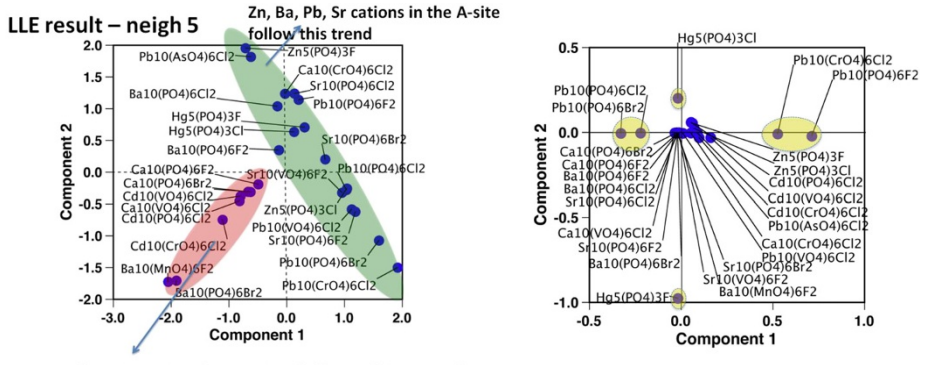

Figure 8 Apatite LLE (left) and hLLE (right) result interpretation [43].

$Pb_{10}(CrO_4)_6Cl_2$ was found to not obey the general trend [44]. Note that the presence of Cr cation in the *B*-site has been known to cause structural distortions in apatites.

For example, $Sr_{10}(PO_4)_6Cl_2$ has a $P6_3/m$ symmetry, whereas $Sr_{10}(CrO_4)_6Cl_2$ has a distorted $P6_3$ symmetry [28]. Based on the previous PCA work [44], we attribute the cause for this exception to two bond distortion angles: (i) rotation angle of A^{II}-A^{II}-A^{II} triangular units and the angle that bond A^{I}-O_1 makes with the *c*-axis. Compared to Hessian LLE, we cannot find any clear pattern with respect to chemical bonding in the LLE result Figure 8 (left).

Figure 9 shows a zoomed-in plot of the Hessian LLE result.[c] Around the origin, we can find two clusters of compounds: (i) one on the left with negative component 1 value corresponding to compounds that have ionic alkaline earth metal cations in the *A*-site and (ii) one on the right with positive component 1 value corresponding to compounds that have covalent *A*-site cations. An exception here is $Ca_{10}(CrO_4)_6Cl_2$, which is found among the covalent *A*-site cluster indicating that $Ca_{10}(CrO_4)_6Cl_2$ may have a distorted symmetry. It is important to recognize that neither PCA nor Isomap identifies $Ca_{10}(CrO_4)_6Cl_2$ as an exception. Compared to hLLE, we do not find any intriguing insights from the LLE analysis and therefore, we do not discuss LLE results.

One needs to explore different manifold methods to fully understand high dimensional correlations and mappings. Hence, in the following section, we shall explore the impact of the Isomap analysis.

In Figure 10 region 1, the ionic radii of *A*-site elements increases along the direction shown, with Zn^{2+} cation being the smallest and Ba^{2+} being the largest. Note that this *A*-site ionic radii trend is not clearly seen in the PC2-PC3 classification map (Figure 7). One of the key outcomes from Figure 10 is the identification that $Pb_{10}(PO_4)_6F_2$ compound is an outlier. In terms of Shannon's ionic radii scale, Pb^{2+} is larger than Ca^{2+} but smaller than Sr^{2+} cation. Ideally (assuming apatites as ionic crystals), the relative position of $Pb_{10}(PO_4)_6F_2$ should have been between $Ca_{10}(PO_4)_6F_2$ and $Sr_{10}(PO_4)_6F_2$ compounds in the map. However, this was not the case. The physical reason behind this observation could be attributed to the electronic structure of Pb^{2+} ions [47]. The theoretical electronic structure calculations indicate that in the atom-projected density of states curves, the Pb^{2+} ions have active $6s^2$ lone-pair electrons that hybridize with oxygen $2p$ electrons resulting in a strong covalent bond formation. Indeed, recent density functional theory

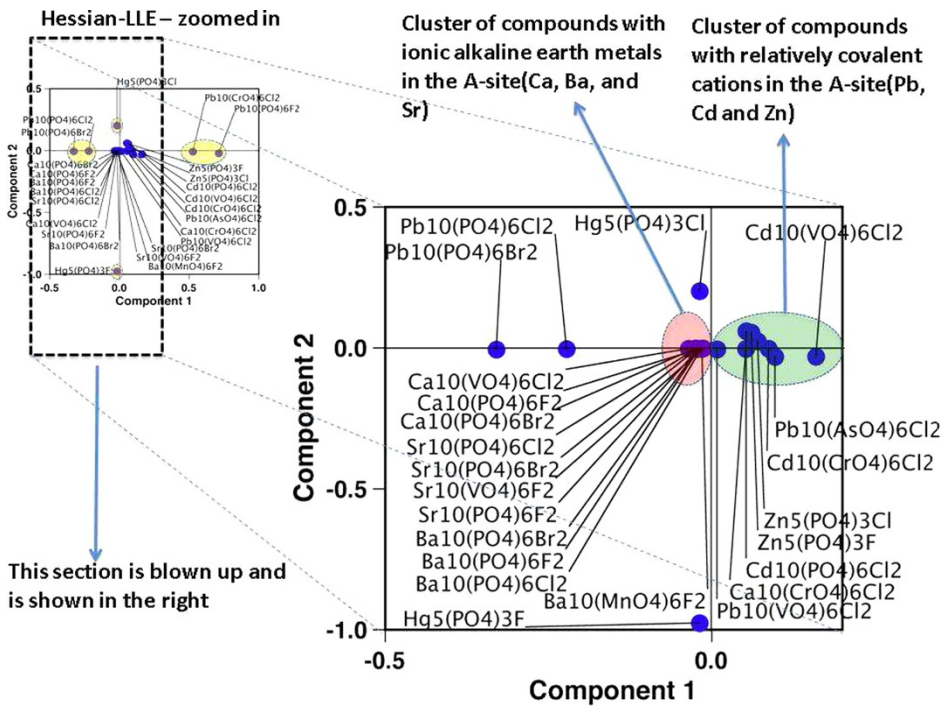

Figure 9 Apatite hLLE result interpretation [43].

(DFT) calculations [48] show that the electronic band gap (at the generalized gradient approximation (GGA) level) for $Pb_{10}(PO_4)_6F_2$ is 3.7 eV, which is approximately 2 eV smaller compared to $Ca_{10}(PO_4)_6F_2$ (5.67 eV) and $Sr_{10}(PO_4)_6F_2$ compounds (5.35 eV).

In our dataset, the electronic structure information of *A*-site elements was quantified using Pauling's electronegativity data. While PCA captures this behavior, the dominating effect of the electronic structure of Pb^{2+} ions is more transparent within the mathematical framework of non-linear Isomap analysis.

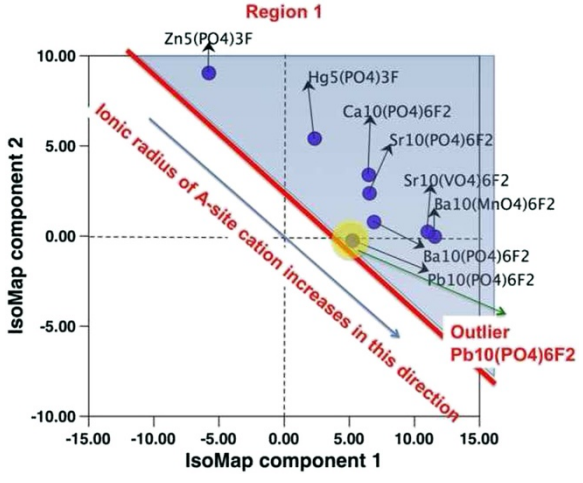

Figure 10 Apatite Isomap result interpretation (region 1) [43].

Besides, from Figure 10, it can also be inferred that the bond distortions of Zn apatite is different from other compounds. This trend correlates well with the non-existence of $Zn_{10}(PO_4)_6F_2$ compounds due to the difficulty in experimental synthesis [49]. On the other hand, the relative correlation position of $Hg_{10}(PO_4)_6F_2$ compound offer intriguing insights. In fact, the uniqueness of $Hg_{10}(PO_4)_6F_2$ chemistry was previously detected in a PCA-derived structure map [44], which clearly identified the composition as an outlier among other isostructural compounds. Guided by this original insight from PCA, recently, Balachandran *et al.* [48] showed using DFT calculations that the ground state structure of $Hg_{10}(PO_4)_6F_2$ is triclinic (space group $P\bar{1}$). Although the ionic size of Hg^{2+} is very close to that of Ca^{2+}, the aristotype $P6_3/m$ symmetry distorts to $P\bar{1}$ symmetry in $Hg_{10}(PO_4)_6F_2$ due to the mixing of fully occupied Hg-$5d^{10}$ orbitals with the empty Hg-$6s^0$ orbitals. This mixing is unavailable to the $Ca_{10}(PO_4)_6F_2$ compound, because it does not have orbitals of appropriate symmetry.

In Figure 11, region 2 is highlighted where we find a clear trend of apatite compounds with respect to the ionic radii of *A*-site elements. Similar to region 1, Pb apatites manifest themselves as outliers in region 2. The unique electronic structure of Pb^{2+} cations in forming a covalent bond with oxygen 2*p*-states is identified as the reason for the deviation of Pb apatites from the expected trend. The covalent bonding among Pb compounds appear to be independent of *X*-site anion, when the *B*-site is occupied by phosphorus cations. In Figure 11, $Hg_{10}(PO_4)_6Cl_2$ compound is found to be closely associated with $Ca_{10}(PO_4)_6Br_2$ indicating some similarity in the bond distortions of the two compounds. In comparing the relative correlation position of all Cl-containing apatites (except Pb-based compounds) in region 2, we predict $Hg_{10}(PO_4)_6Cl_2$ to have a stable apatite structure type (in sharp contrast to $Hg_{10}(PO_4)_6F_2$).

Figure 12 describes region 3 where we find clusters of apatite compounds with Cl ions in the *X*-site and contain larger V, Cr, and As cations in the *B*-site. The ionic radius of *A*-site element increases in the direction as shown in the figure, and in this case, the Pb apatites are not outliers. The presence of large V, Cr, and As cations (compared to smaller P cations in regions 1 and 2) in the *B*-site were identified as the reason for this behavior.

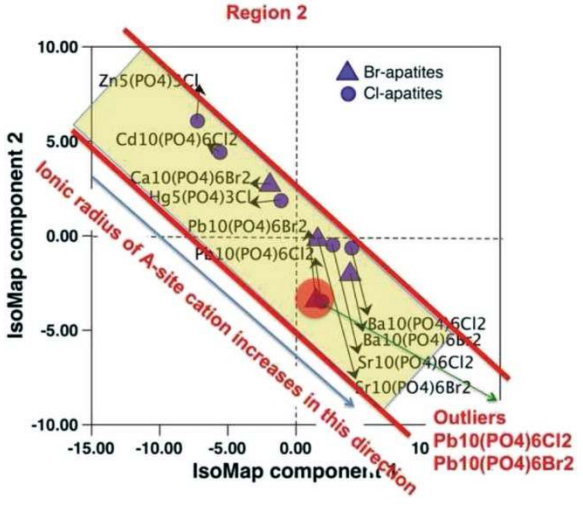

Figure 11 Apatite Isomap result interpretation (region 2) [43].

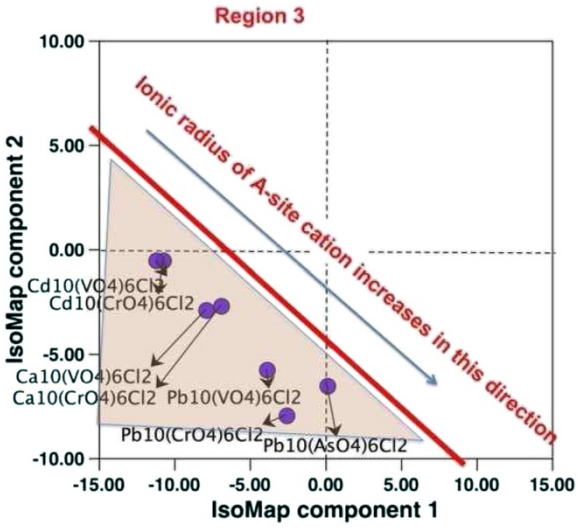

Figure 12 Apatite Isomap result interpretation (region 3) [43].

Region 3 also identifies the existence of complex relationship between *A*-site and *B*-site chemistries in Cl apatites.

Several topological observations can be made on the data. Firstly, since low-dimensional points obtained are different for both Isomap and PCA, it could be interpreted that the apatite data lie on a non-linear manifold in the embedding space. However, a counter argument can be made based on the fact that PC2-PC3 plot shows similar trends and clustering as that in Isomap1-Isomap2. One possible reason for this happening could be due to the existence of outliers dominating and deviating the first PCA component (PC1) while Isomap being unaffected by this outlier; in which case, the data could actually be lying on a linear manifold. Secondly, the different clustering phenomena observed along different dimensionality reduction techniques might imply that the pattern/features seen in PCA and Isomap clusters are a function of the distance preserved, while those in hLLE and LLE is a function of the topology preserved. Hence, these chosen features represented by these clusters happen to be preserved all along the dimensionality reduction process from the embedded space to the lower-dimensional space.

Conclusions

In this paper, we have detailed a mathematical framework of various data dimensionality reduction techniques for constructing reduced order models of complicated datasets and discussed the key questions involved in data selection. We introduced the basic principles behind data dimensionality reduction[d]. The techniques are packaged into a modular, computational scalable software framework with a graphical user interface - SETDiR. This interface helps to separate out the mathematics and computational aspects from the scientific applications, thus significantly enhancing utility of DR techniques to the scientific community. The applicability of this framework in constructing reduced order models of complicated materials dataset is illustrated with an example dataset of apatites. SETDiR was applied to a dataset of 25 apatites being described by

29 of its structural descriptors. The corresponding low-dimensional plots revealed previously unappreciated insights into the correlation between structural descriptors like ionic radius, bond covalence, etc., with properties such as apatite compound formability and crystal symmetry. The plots also uncovered that the shape of the surface on which the data lies could be non-linear. This information is crucial as it can promote the use of apatite materials as a potential host lattice for immobilizing toxic elements.

Availability of supporting data

Information regarding the source of the apatite data can be found in [44].

Endnotes

[a]A tree is a graph where each pair of vertices is connected exactly with one path. A spanning tree of a graph $G(V, E)$ is a sub-graph that traces all the vertices in the graph. A minimal spanning tree (MST) of a weighted graph $G(V, E, W)$ is a spanning tree with a minimal sum of the edge weights (length of the MST) along the tree. A geodesic minimal spanning tree (GMST) is an MST with edge weight representing geodesic distance. Computationally, GMST is computed using Prim's (greedy) algorithm [50].

[b]Normalization of a variable is forcing a limit of $[-1, 1]$ or $[0, 1]$ to an existing limit of $[a, b]$ of a variable by dividing the sequence of numbers with the maximum absolute value of the sequence.

[c]Hessian LLE is highly sensitive to neighborhood size and is much more sensitive to the input estimated dimensionality. Incorrect input of estimated dimensionality implies construction of tangent planes of incorrect dimensions which, in turn, implies sub-optimal low-dimensional representation.

[d]A comprehensive catalogue of non-linear dimensionality reduction techniques along with the mathematical prerequisites for understanding dimensionality reduction could be found in [23].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SKS, BG, and JZ formulated the mathematical framework. SKS and JZ implemented the mathematical framework. SKS and PVB performed the model reduction on the apatite data to extract the low-dimensional representation. PVB and KR interpreted the results. SKS, PVB, BG, JZ, and KR discussed the results and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge the support from the National Science Foundation (NSF) grant CDI- NSF-CDI -PHY 09-41576. KR acknowledges the support from NSF: DMR- 13-07811 and DMS-11-25909, Department of Homeland Security/NSF-ARI Program: CMMI 09-389018; Army Research Office grant W911NF-10-0397, Air Force Office of Scientific Research SFA9550-12-1-0456, and the Wilkinson Professorship of Interdisciplinary Engineering. BG also acknowledges the support from NSF CAREER CMMI-11-49365.

Author details

[1]School of Mechanical Engineering, Georgia Tech, Atlanta, GA 30332-0405, USA. [2]Department of Materials Science, Drexel University, Philadelphia, PA 19104, USA. [3]Rutgers Discovery Informatics Institute, Rutgers University, Piscataway, NJ 08854, USA. [4]Department of Materials Science and Engineering, Iowa State University, Ames, IA 50011, USA. [5]Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA.

Received: 2 December 2013 Accepted: 29 April 2014
Published: 29 June 2014

References

1. Rabe KM, Phillips JC, Villars P, Brown ID (1992) Global multinary structural chemistry of stable quasicrystals, high-t_c ferroelectrics, and high-t_c superconductors. Phys Rev B 45:7650–7676
2. Morgan D, Rodgers J, Ceder G (2003) Automatic construction, implementation and assessment of pettifor maps. J Phys: Condens Matter 15(25):4361

3. Chawla N, Ganesh VV, Wunsch B (2004) Three-dimensional (3d) microstructure visualization and finite element modeling of the mechanical behavior of SiC particle reinforced aluminum composites. Scripta Materialia 51(2):161–165
4. Langer SA, Jr. Fuller ER, Carter WC (2001) OOF: an image-based finite-element analysis of material microstructures. Comput Sci Eng 3(3):15–23
5. Liu ZK, Chen LQ, Raghavan P, Du Q, Sofo JO, Langer SA, Wolverton C (2004) An integrated framework for multi-scale materials simulation and design. J Comput Aided Mater Des 11:183–199
6. van Rietbergen B, Weinans H, Huiskes R, Odgaard A (1995) A new method to determine trabecular bone elastic properties and loading using micromechanical finite-element models. J Biomech 28(1):69–81
7. Yue ZQ, Chen S, Tham LG (2003) Finite element modeling of geomaterials using digital image processing. Comput Geotechnics 30(5):375–397
8. McVeigh C, Liu WK (2008) Linking microstructure and properties through a predictive multiresolution continuum. Comput Methods Appl Mech Eng 197(4142):3268–3290
9. Zabaras N, Sundararaghavan V, Sankaran S (2006) An information-theoretic approach for obtaining property PDFs from macro specifications of microstructural variability. TMS Lett 3:1–2
10. Meredith JC, Smith AP, Karim A, Amis EJ (2000) Combinatorial materials science for polymer thin-film dewetting. Macromolecules 33(26):9747–9756
11. Takeuchi I, Lauterbach J, Fasolka MJ (2005) Combinatorial materials synthesis. Mater Today 8(10):18–26
12. Lumley JL (1967) The structure of inhomogeneous turbulent flows. Atmospheric turbulence and radio wave propagation 166–178
13. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
14. Donoho DL, Grimes C (2003) Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci 100:5591–5596
15. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson K (2013) The materials project: a materials genome approach to accelerating materials innovation. APL Mater 1(1):011002
16. Page YL (2006) Data mining in and around crystal structure databases. MRS Bulletin 31:991–994
17. Rajan K, Suh C, Mendez PF (2009) Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering. Stat Anal Data Mining 1(6):361–371
18. Brasca R, Vergara LI, Passeggi MCG, Ferrona J (2007) Chemical changes of titanium and titanium dioxide under electron bombardment. Mat Res 10:283–288
19. Ganapathysubramanian B, Zabaras N (2008) A non-linear dimension reduction methodology for generating data-driven stochastic input models. J Comput Phys 227(13):6612–6637
20. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G (2003) Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 91:135503
21. Fischer CC, Tibbetts KJ, Morgan D, Ceder G (2006) Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 5(8):641–646
22. Morgan D, Ceder G, Curtarolo S (2005) High-throughput and data mining with ab initio methods. Meas Sci Technol 16(1):296
23. Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Springer
24. Van der Maaten LJP, Postma EO, Van Den Herik HJ (2009) Dimensionality reduction: a comparative review
25. Elliott JC (1994) Structure and chemistry of the apatites and other calcium orthophosphates, volume 4. Elsevier, Amsterdam
26. Mercier PHJ, Le Page Y, Whitfield PS, Mitchell LD, Davidson IJ, White TJ (2005) Geometrical parameterization of the crystal chemistry of P63/m apatites: comparison with experimental data and ab initio results. Acta Crystallogr Sect B: Structural Sci 61(6):635–655
27. Pramana SS, Klooster WT, White TJ (2008) A taxonomy of apatite frameworks for the crystal chemical design of fuel cell electrolytes. J Solid State Chem 181(8):1717–1722
28. White T, Ferraris C, Kim J, Madhavi S (2005) Apatite–an adaptive framework structure. Rev Mineralogy Geochem 57(1):307–401
29. White TJ, Dong ZL (2003) Structural derivation and crystal chemistry of apatites. Acta Crystallogr Sect B: Structural Sci 59(1):1–16
30. Samudrala S, Rajan K, Ganapathysubramanian B (2013) Data dimensionality reduction in materials science In: Informatics for materials science and engineering: data-driven discovery for accelerated experimentation and application. Elsevier Science
31. Bergman S (1950) The kernel function and conformal mapping. Am Math Soc
32. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
33. Fontanini A, Olsen M, Ganapathysubramanian B (2011) Thermal comparison between ceiling diffusers and fabric ductwork diffusers for green buildings, Energy and Buildings 43(11):2973–2987. ISSN 0378–7788. http://dx.doi.org/10.1016/j.enbuild.2011.07.005
34. Amini H, Sollier E, Masaeli M, Xie Y, Ganapathysubramanian B, Stone HA, Di Carlo D (2013) Engineering fluid flow using sequenced microstructures. Nature Communications 4:2013
35. Guo Q (2013) Incorporating stochastic analysis in wind turbine design: data-driven random temporal-spatial parameterization and uncertainty quantication. Graduate Theses and Dissertations. Paper 13206. http://lib.dr.iastate.edu/etd/13206
36. Wodo O, Tirthapura S, Chaudhary S, Ganapathysubramanian B (2012) A novel graph based formulation for characterizing morphology with application to organic solar cells. Org Electron:1105–1113
37. Golub GH, Van Loan CF (1996) Matrix computations. The John Hopkins University Press
38. Floyd RW (1962) Algorithm 97: shortest path. Commun ACM 5(6):345

39. Bernstein M, De Silva V, Langford JC, Tenenbaum JB (2000) Graph approximations to geodesics on embedded manifolds. Technical report, Department of Psychology, Stanford University
40. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
41. Beardwood J, Halton JH, Hammersley JM (1959) The shortest path through many points. Math Proc Camb Philos Soc 55:299–327
42. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Phys D: Nonlinear Phenomena 9(12):189–208
43. Balachandran PV (2011) Statistical learning for chemical crystallography. PhD thesis, Iowa State University
44. Balachandran PV, Rajan K (2012) Structure maps for $A^{I}{}_4A^{II}{}_6(BO_4)_6X_2$ apatite compounds via data mining. Acta Crystallogr Sect B 68(1):24–33
45. Shannon RD (1976) Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallographic Sect A: Crystal Phys Diffraction Theor Gen Crystallography 32(5):751–767
46. Pauling L (1960) The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry, vol 18. Cornell University Press
47. Matsunaga K, Inamori H, Murata H (2008) Theoretical trend of ion exchange ability with divalent cations in hydroxyapatite. Phys Rev B 78:094101
48. Balachandran PV, Rajan K, Rondinelli JM (2014) Electronically driven structural transitions in $A_{10}(PO_4)_6F_2$ apatites (A = Ca, Sr, Pb, Cd and Hg). Acta Crystallogr Sect B 70: 612–615
49. Flora NJ, Hamilton KW, Schaeffer RW, Yoder CH (2004) A comparative study of the synthesis of calcium, strontium, barium, cadmium, and lead apatites in aqueous solution. Synthesis Reactivity Inorganic Metal-organic Chem 34(3):503–521
50. Prim RC (1957) Shortest connection networks and some generalizations. Bell Syst Tech J 36(6):1389–1401

doi:10.1186/s40192-014-0017-5
Cite this article as: Samudrala *et al.*: A software framework for data dimensionality reduction: application to chemical crystallography. *Integrating Materials and Manufacturing Innovation* 2014 **3**:17.