Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 7 July 2009
Accepted 4 July 2010

Crystal fingerprint space – a novel paradigm for studying crystal-structure sets

Mario Valle[a]* and Artem R. Oganov[b,c]

[a]Data Analysis and Visualization Group, Swiss National Supercomputing Centre (CSCS), via Cantonale Galleria 2, 6928 Manno, Switzerland, [b]Department of Geosciences, Department of Physics and Astronomy, and New York Center for Computational Sciences, State University of New York, Stony Brook, NY 11794-2100, USA, and [c]Geology Department, Moscow State University, 119992 Moscow, Russia. Correspondence e-mail: mvalle@cscs.ch

The initial aim of the crystal fingerprint project was to solve a very specific problem: to classify and remove duplicate crystal structures from the results generated by the evolutionary crystal-structure predictor *USPEX*. These duplications decrease the genetic diversity of the population used by the evolutionary algorithm, potentially leading to stagnation and, after a certain time, reducing the likelihood of predicting essentially new structures. After solving the initial problem, the approach led to unexpected discoveries: unforeseen correlations, useful derived quantities and insight into the structure of the overall set of results. All of these were facilitated by the project's underlying idea: to transform the structure sets from the physical configuration space to an abstract, high-dimensional space called the fingerprint space. Here every structure is represented as a point whose coordinates (fingerprint) are computed from the crystal structure. Then the space's distance measure, interpreted as structure 'closeness', enables grouping of structures into similarity classes. This model provides much flexibility and facilitates access to knowledge and algorithms from fields outside crystallography, *e.g.* pattern recognition and data mining. The current usage of the fingerprint-space model is revealing interesting properties that relate to chemical and crystallographic attributes of a structure set. For this reason, the mapping of structure sets to fingerprint space could become a new paradigm for studying crystal-structure ensembles and global chemical features of the energy landscape.

1. From a problem to a new paradigm

USPEX (Oganov & Glass, 2006) is a computational method and code based on an evolutionary algorithm which enables crystal-structure prediction at arbitrary conditions of pressure and temperature given just the chemical composition of the material. Its latest versions (Lyakhov *et al.*, 2009) enable even the prediction of the chemical compositions of all stable phases given just the names of the constituent chemical atoms.

Owing to the algorithm's evolutionary nature, every *USPEX* run produces hundreds or thousands of putative crystal structures, but in practice many of them are the same structure, perhaps described in a different, but equivalent, way or based on a different coordinate reference frame or made different by small numerical errors. It is therefore necessary to reduce the results to a set of unique structures to concentrate analysis on the configurations that could give insight on new phenomena. This involves intensive manual labor, consisting mainly of judging similarity from side-by-side visualization of pairs of structures. We decided therefore to design an automated structure comparison and clustering method.

The approach adopted describes crystal structures as points in a multidimensional space, each identified by a multidimensional coordinate set (here called the fingerprint). This space has a distance measure defined on it so we can quantify structure 'closeness' and then use clustering methods to find equivalent structures. Each equivalence group is then reduced to a single representative structure.

The resulting *CrystalFp* classifier (Valle & Oganov, 2008) has already substantially improved the quality of *USPEX* structure predictions, leading to some new crystallographic insight and discoveries (Ma *et al.*, 2009; Oganov & Valle, 2009). Now it has been incorporated inside *USPEX*, thus solving the duplicate-structures generation problem at its roots.

We have solved the initial problem, but have also foreseen a new line of research: gathering insight on crystal-structure sets by analyzing them after transformation to the fingerprint space. This shift was motivated by the unexpected correlations we found (Oganov & Valle, 2009), for example, between

structures' energy differences and their distance in fingerprint space, by the insight gained using new derived quantities and, finally, by the opportunities offered by the rich semantics of high-dimensional spaces. Our thesis is that fingerprint spaces have interesting properties that relate to the chemical properties of the whole set of structures, *i.e.* its 'global' chemistry. Thus this model could become a new analysis tool for crystal-structure ensembles, like the ones created in structure-prediction simulations or in simulations of structural transitions.

The present work starts with a review of the original design and underlying ideas that give rise to the crystal-fingerprint model as presented by Valle & Oganov (2008). Then we analyze how the solution adopted changed from being a point solution to become the foundation of a more general paradigm for studying ensembles of crystal structures. The paper then covers the current work and future directions of fingerprint-space research.

2. Previous work

In the literature there are plenty of works proposing suitable structure descriptors for organic molecules and distance metrics based on them, but very few focused on crystal structures. The work of Hundt *et al.* (2006) gives a comprehensive survey of existing methods and is focused on calculating some form of distance measure between structures.

Interatomic distances are a good choice as structure identifiers because they are independent of the coordinate reference frame and unit-cell choices. Chisholm & Motherwell (2005) use them to investigate molecular packing and inspired one of the methods we tried (see §3.1). Radial distribution functions (RDFs) are other possible structure descriptors based only on local characteristics. Willighagen *et al.* (2005) use an RDF computed using distances from a central atom weighted by the involved atomic partial charges to include electrostatic interactions, which play a major role in crystal packing. This method then calculates dissimilarities on the basis of powder diffraction patterns as proposed by de Gelder (2006). In our work we use a rapidly convergent function based on distance distributions and related to RDFs and diffraction spectra [equation (1)], but we focus on standard multidimensional methods for distance computation. Another interesting application of RDFs is the work of Hemmer *et al.* (1999), which uses them to match structures to IR spectra using a counterpropagation neural network. Their goal is indeed different from ours, but they also found that RDFs could be good crystal-structure identifiers.

The ideas and results of high-dimensional spaces, the basis of the methods we employ in this work, will be covered in §6.

3. Crystal fingerprint spaces

Each structure's fingerprint is a vector of N real values computed from its structural parameters; each structure thus becomes a point in an N-dimensional fingerprint space. A distance measure between these vectors is defined and then used to cluster them into groups of 'near' fingerprints, that is, groups of similar structures.

3.1. Fingerprint definition

To be useful as a structure identifier, the associated fingerprint should be independent of: (1) translation and rotation of the structure; (2) the choice of unit cell among equivalent unit cells (modular invariance); (3) the ordering of cell axis and atoms in the cell; and (4) inversions and mirroring of the structure.

Whichever definition we choose for the structure fingerprint, it should be computed over the 'infinite' crystal structure. In practice, after a distance of $D_{\rm uc}/2$, where $D_{\rm uc}$ is the longest unit-cell diagonal, everything starts repeating in every direction. Therefore in place of the unlimited crystal structure, a set of unit-cell repetitions that cover the maximum distance $R_{\rm max} = \max_i(D_{{\rm uc}_i}/2)$ over all structures in all directions around the base unit cell is used. We call the union of these replicas and the original unit cell the extended unit cell.

Our choice of fingerprint started from a simple concatenation of per-atom distances (Fig. 1) and then moved to a rapidly convergent function based on distance distributions and related to RDFs and diffraction spectra (Fig. 2) that was modified to take account of the different atomic species present in the structure (Fig. 3). This evolution was driven by the visual design approach (see §4) that guided us toward fingerprinting methods that better identify structures and their spatial relationships.

To make the selected fingerprint sensitive to the ordering of atoms in a given structure and independent of their types, we separate the components of the fingerprint function coming from different pairs of atom types A–B, making the total fingerprint a matrix, each element of which is a function

$$F_{AB}(R) = \sum_{A_i}\sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 (N_A N_B / V_{\rm uc})\Delta} - 1, \qquad (1)$$

where i runs over all N_A atoms of type A within the unit cell and j runs over all N_B atoms of type B in the extended unit cell, R_{ij} is the distance between these atoms, $V_{\rm uc}$ is the unit-cell volume and δ is the Dirac distribution. Each peak is smoothed before calculating the sum using a Gaussian kernel with σ set by the user (usually 0.02 Å) and accumulated into a histogram with bin size Δ (usually 0.05 Å). Note that each F_{AB} starts with the value of -1 (at $R = 0$) and converges to zero. One example of this fingerprint is given in Fig. 3.

We also note that our fingerprint definition is based on a two-body correlation function, and in specific situations there may be advantages in using three-, four- and higher-order many-body correlation functions (the formal extension of our formalism to such cases is trivial, but the computation is much heavier).

This definition of fingerprint provided the best discriminative power and became the foundation for the definition of new quantities, such as the quasi-entropy (see §7.2), but it also

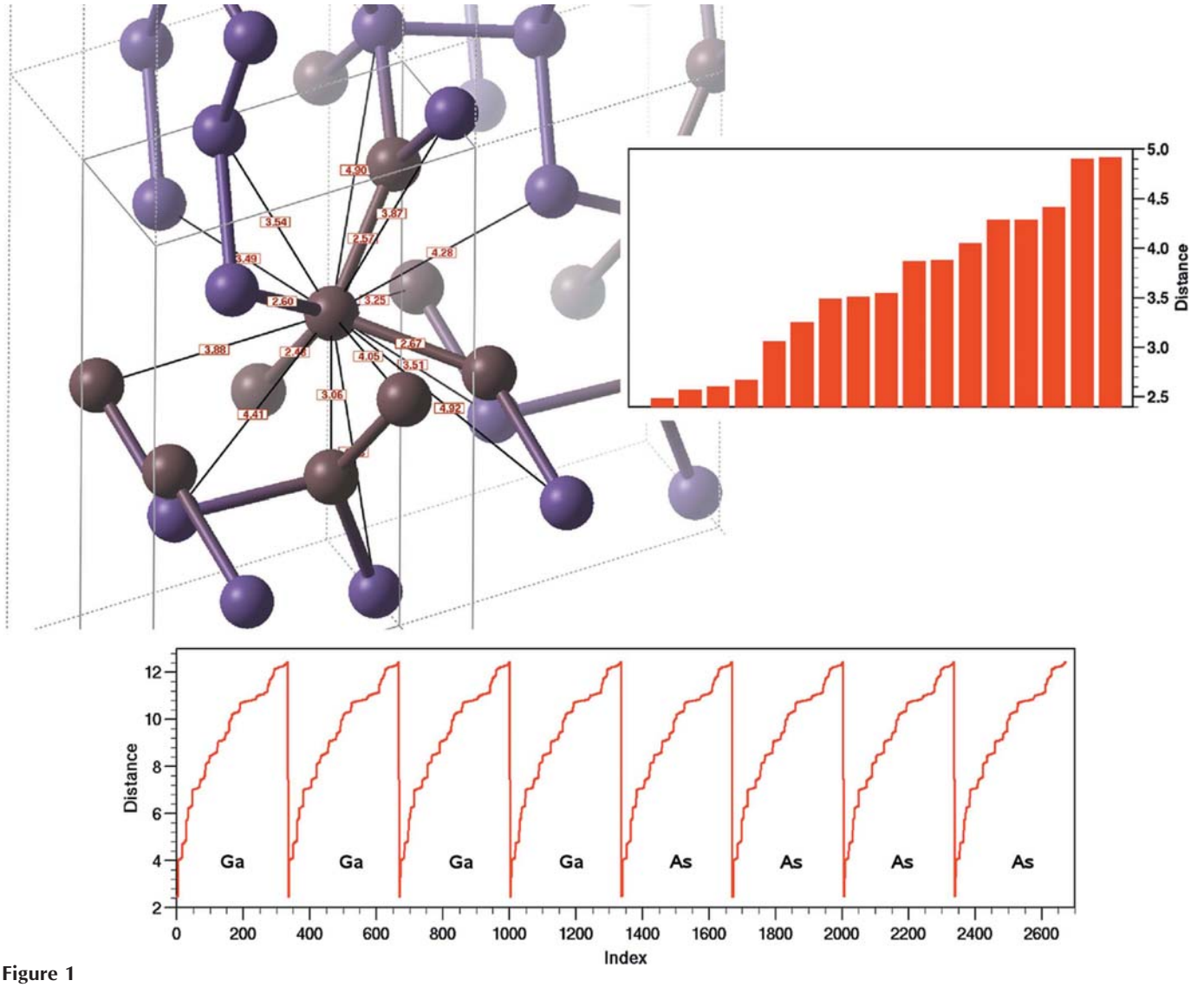

Figure 1
Per-atom distances fingerprint. Local atom distances for a GaAs crystal (top left) are concatenated to form a fingerprint section (top right). Sections are then assembled to form the structure fingerprint (bottom).

has some unpleasant characteristics: (1) some coordinates are in a way redundant, for example, those having a value of -1 before the first peak; (2) the coordinates are not independent, due to the Gaussian kernel smoothing; and (3) not all coordinates have the same importance: the further one moves to the right, the less different are the coordinate values and the less dependent on the structure they are.

3.2. Distances measure

To enable the classification of structures, we should define a distance or pseudo-distance[1] measure between fingerprints. We tried three distance measures (Valle & Oganov, 2008): (1) Euclidean distance; (2) Minkowski norm with a fractional exponent; and (3) cosine distance.

[1] Pseudo-distance is a distance measure for which the triangular inequality $\text{dist}_{AB} \leq \text{dist}_{AC} + \text{dist}_{CB}$ does not hold.

The cosine distance, the one finally adopted, is a popular norm in the text-mining community (Salton & McGill, 1983; Salton & Buckley, 1988). Here every text has an associated vector of word frequencies and the similarity between texts is based on the dot product between these vectors. We use a slightly modified definition of similarity that produces a distance in the $[0 \ldots 1]$ interval using, in place of the word-frequency vectors, the fingerprint F_i associated with structure i,

$$\text{dist}(i, j) = \frac{1}{2}\left(1 - \frac{F_i \cdot F_j}{\|F_i\|\|F_j\|}\right). \qquad (2)$$

We chose the cosine-distance measure for its ability to counteract the distance-concentration phenomena (defined in §6) by spreading distances much more than the other methods, as seen in Fig. 4. The only unpleasant characteristics of the cosine distance are that it is not translation-invariant and it does not satisfy the triangular inequality.

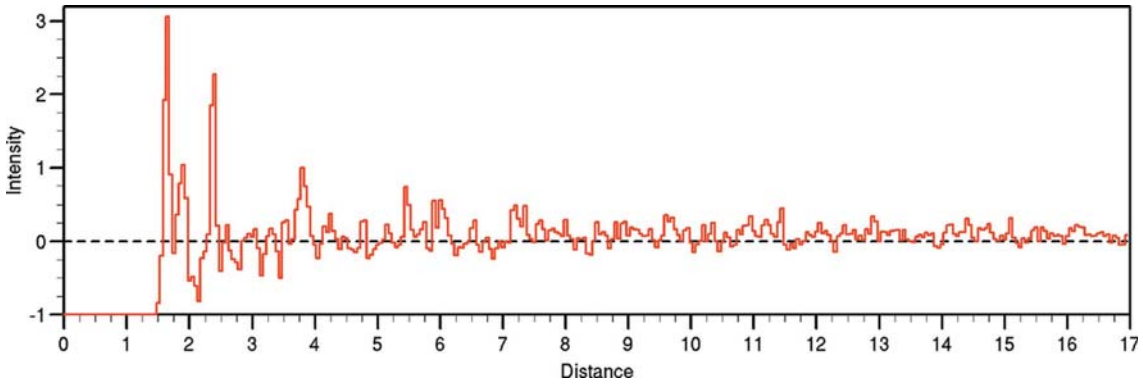

Figure 2
A fingerprint related to RDFs and diffraction spectra. Note that for this fingerprint values start at −1 and converge to 0.

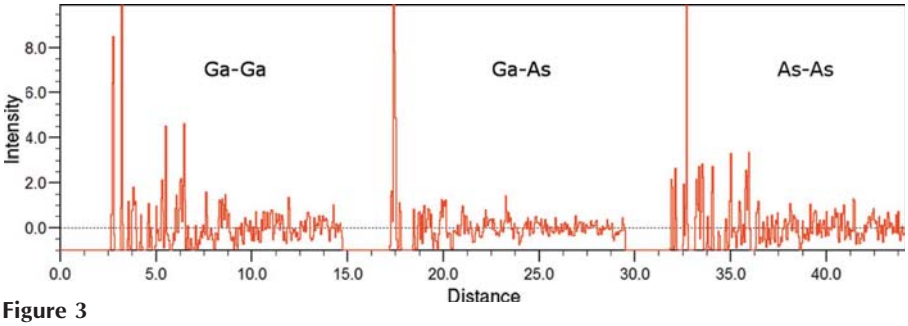

Figure 3
A multicomponent fingerprint. A fingerprint like the one in Fig. 2 is computed for each pair of element types. Then all of them are concatenated to form the structure fingerprint (this example is for a GaAs structure).

3.3. Structure clustering

The basic idea of clustering is to assign two structures to the same group if their distance, *i.e.* the distance between their fingerprints, is less than a user-defined threshold. Then to move from pairwise grouping to cluster building, we can proceed by aggregation or by graph walking (Jain *et al.*, 1999).

The clustering by aggregation is a 'bottom-up' approach: each point starts in its own cluster, and pairs of clusters are merged if the distance between their closest or furthest elements falls below the threshold (Hastie *et al.*, 2001).

The other approach starts from the graph generated by the binary connection matrix obtained by thresholding the distance matrix. The graph described by this matrix is then separated into connected components that are our clusters of similar structures. To refine these components we resort to a density-based criterion: a connection is confirmed only if the two connected vertices share at least K nearest neighbors. We call this clustering approach the pseudo-shared nearest neighbor (pseudoSNN) method to distinguish it from the full SNN method (Ertoz *et al.*, 2003), which adds a refining step based on the cluster's density. After experimentation we selected the pseudoSNN clustering method with at least one shared nearest neighbor.

4. Visualization role in design and analysis

How did the various definitions of a fingerprint, distance metric and classification algorithm evolve within the common framework? They changed as direct consequence of the domain-expert exploration and validation of the classifier during analysis runs made on real data. To make this exploration possible we built an end-user application around the classifier to support the *USPEX* results analysis workflow and to provide interactive visual diagnostics on the behavior of the algorithms (Fig. 5). These provide visual representations of key algorithm quantities so the domain expert could judge the algorithm behavior. The visual validation and analysis, plus the user's algorithm selection and parameter modifications, support a very effective exploratory design approach.

The end-user application was built inside the molecular visualization toolkit *STM4* (Valle, 2005, 2009) based on *AVS/Express* (Lever *et al.*, 2000; Advanced Visual Systems, 2009). Complete coverage of the tool and the visualization is provided by Valle & Oganov (2008).

From the beginning of this work we embraced an exploratory approach to finding the correct scientific questions to pose. In this approach, visualization continues to play a prominent role. As we see in §5, visualization made possible the identification of interesting correlations between data, derived quantities and surrogate data sets. As a consequence,

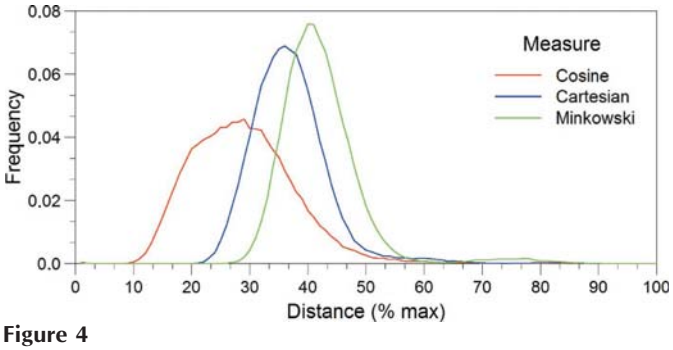

Figure 4
Distance distributions for the $MgSiO_3$ post-perovskite 120 GPa data set using the three distance measures analyzed. The distributions are presented together after normalizing the distance values.

we plan to rely on visualization help in the next steps of this research as well. To this end, the initial end-user tool has been enhanced with a more flexible way to correlate data across data sets and across parameter changes.

5. Transition to the new paradigm

The initial problem of removing duplicate crystal structures from the results generated by the evolutionary crystal-structure predictor *USPEX* was solved at its roots by incorporating the *CrystalFp* classifier inside it. But the story does not end here. During the visual design of the algorithm, the ease of addition of new visualizations and new analysis to the library, coupled with the simplicity in accessing them from the end-user application, uncovered interesting and unanticipated insights, such as the following.

Energy versus distance correlation. For some structures the chart depicting the energy difference *versus* distance shows a strong linear correlation; for others the relationship is more complex, but not random (see Fig. 6). We were able to relate this behavior to the shape of the energy landscape (Oganov & Valle, 2009).

Random structures distance distribution. Looking at the statistics for distances between structures in a random set of structures, we often discovered a striking Gaussian-like shape of distributions with a clear peak (Fig. 7*a*). In some cases we find more than one peak (Fig. 7*b*); this could be a signal of complex chemistry involving different coordination numbers, or could be caused by non-random sampling of the configuration space.

Energy versus order correlation. In many tests of *USPEX* we see order increasing during the run (see Fig. 10), and observe a clear correlation with energies: high order usually means low energies. This is natural; disordered structures are expected to have high energies. But energy–order correlations also revealed cases of geometric frustration, where less ordered or more complex structures are made energetically favorable by competition of opposing factors.

These discoveries prompted us to consider the fingerprint-space choice not just as a point solution, but also as a useful tool for analyzing ensembles of crystal structures. Our hope is that the fingerprint space could have attributes correlated to the chemical and crystallographic properties of the starting structure set. We want to analyze, for example, fingerprint-

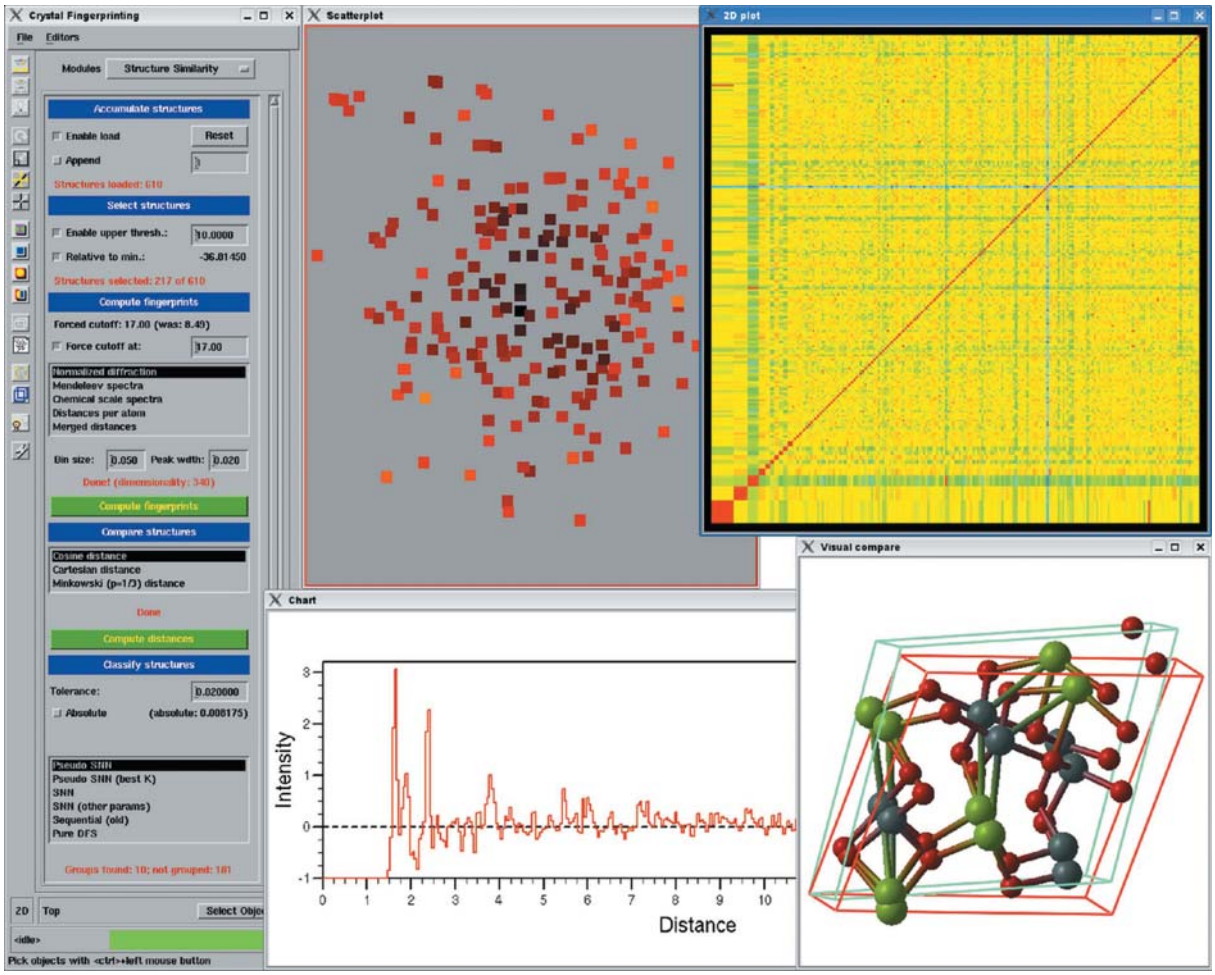

Figure 5
The *CrystalFp* end-user application. The control panel (left), the scatter plot and ordered distance matrix (top), one diagnostic chart and the visual pairwise structure comparison (bottom) are shown.

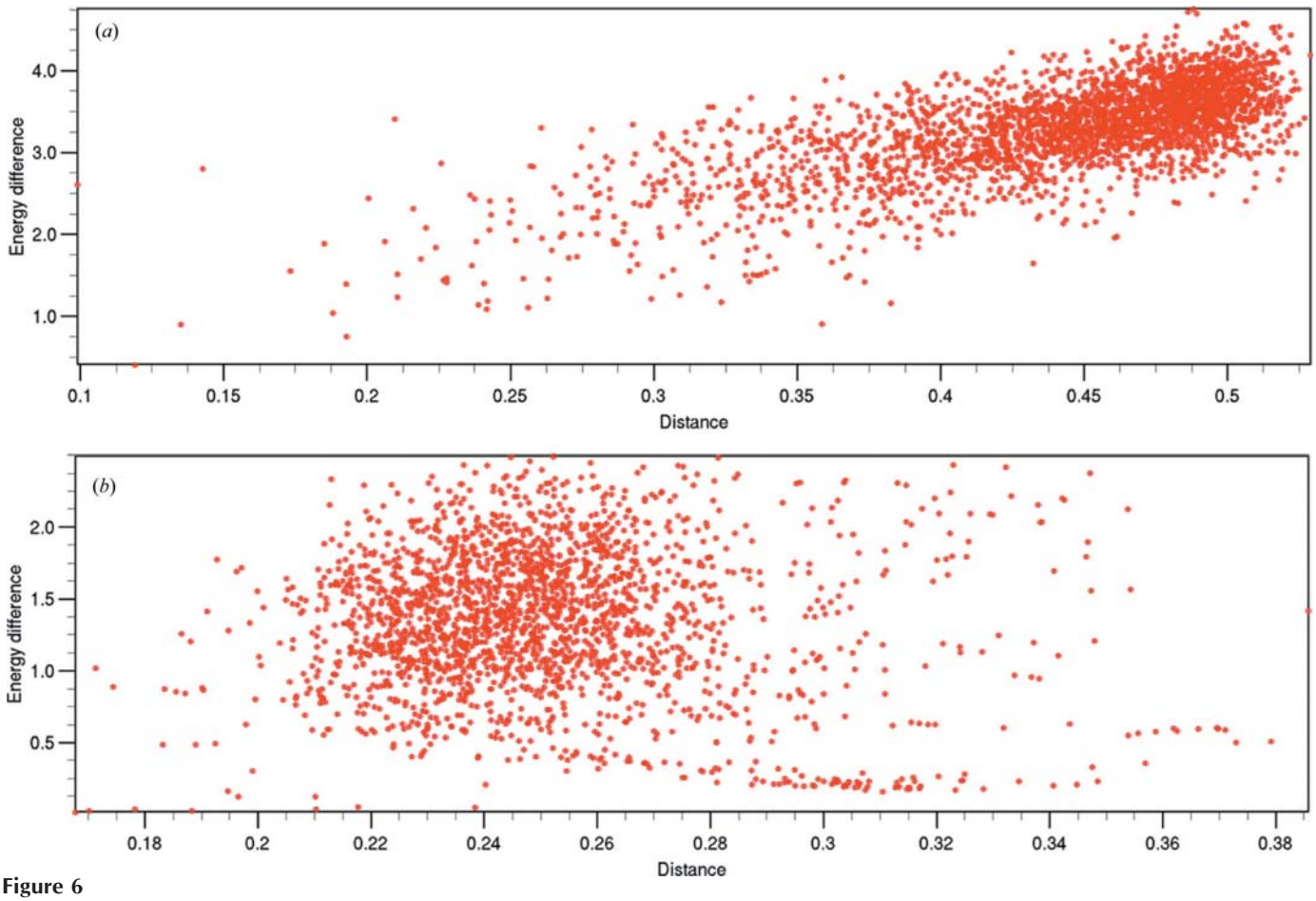

Figure 6
(*a*) GaAs structures exhibit a clear energy–distance correlation. (*b*) In contrast, MgNH shows a more complex landscape.

space topology, how it relates to chemical composition and how it changes with experimental conditions. We also hope to find interesting new correlations between physical quantities and distances in fingerprint space. As a side result, we expect to shed more light on the behavior of the *USPEX* algorithm.

Why is this mapping interesting? Because high-dimensional spaces are semantically richer than physical spaces; because there are available tools, methods and ideas from entirely different fields, like data mining, and because looking at known problems from outside the discipline increases the chances of finding unplanned 'cross-fertilization' between fields.

However, there are problems to surmount. The first one derives from the peculiar behavior of high-dimensional spaces (see §6) that forces us always to check whether findings made in fingerprint space originate from chemical and crystallographic reasons or whether they are artifacts of the mapping. Another problem is how to validate the modeling approach itself: that is, how to verify that the mapping of structures to fingerprints is a faithful one (see §9).

6. High-dimensional spaces

We usually limit our analysis to data that readily fit into our physical space – data that we can see in our mind and visualize on a computer screen. But the really interesting data, like crystal fingerprints, are almost always multivariate. These data normally have no trivial connection with a physical space, and have a sizable number of attributes associated with each data point. We can say that these points live in a high-dimensional space.

The most striking differences to physical spaces are that *high-dimensional spaces are almost empty* and that *their volume is concentrated in strange places*.[2]

In Fig. 8 the volume of a hypersphere of radius 1 is plotted (blue line). At space dimensionality $D = 20$ its volume is already almost zero and thus cannot contain any points. Now the hypersphere is put inside a hypercube of side 2. The ratio between the volume of the circumscribed hypercube that lies outside the hypersphere and the volume of the hypercube itself (red line in Fig. 8) shows that the space is concentrated outside the sphere, in the 'corners' of the hypercube. To grasp an idea of this monstrosity, consider that the diagonals of the hypercube grow as $D^{1/2}$ so the hypercube is a highly anisotropic body with 'spikes' around the contained hypersphere (Köppen, 2000).

[2] Introductions to these and other phenomena are given in Weber *et al.* (1998) and Köppen (2000).

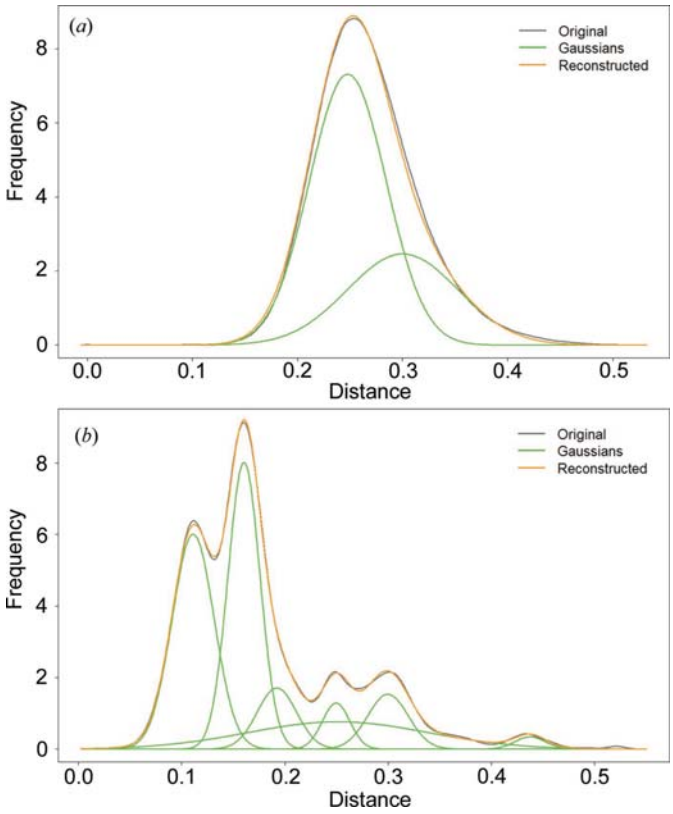

Figure 7
(*a*) Some data sets, like SiO_2 (nine atoms per cell), have a distance distribution decomposable into two Gaussians. (*b*) Others, like H_2O (12 atoms per cell), have a complex distance distribution that needs many Gaussians to be acceptably approximated.

In high-dimensional spaces the points are all at about the same distance from each other (Beyer *et al.*, 1999). This distance grows steadily with dimensionality and decreases only marginally as the number of points increases (Weber *et al.*, 1998). Instead, the variance of their pairwise distance depends strongly on the distance measure adopted and the shape of the space. This phenomenon is very simple to observe experimentally (Köppen, 2000; François *et al.*, 2007). For example, a numerical derivation for Euclidean distances gives for the expectation $E(\text{dist}) \rightarrow (D/6)^{1/2}$ and for the variance $\sigma^2(\text{dist}) \rightarrow 0.49$ (Schmitt, 2001). Instead, for cosine distances we have $E(\text{dist}) \rightarrow 0.5$ and $\sigma^2(\text{dist}) \rightarrow 1/4D$. The concept of relative contrast, $C_R = (\text{dist}_{max} - \text{dist}_{min})/\text{dist}_{min}$, formalizes this phenomenon as $\lim_{D\rightarrow\infty} C_R = 0$.

The above phenomenon, which causes pairwise distances to seem the same for all points, is called 'concentration of distances' (François *et al.*, 2007) and makes some of the concepts that we take for granted in low-dimensional spaces meaningless. One of them is the concept of nearest neighbor. It is no longer meaningful, not only because all points are almost at the same distance, but also because a small perturbation can change the nearest point into the farthest one.

Space partitioning, needed for example by search algorithms, *becomes intractable* as the space dimension grows. Simply cutting in half the space along each dimension generates 2^D partitions, each containing zero or a small number of points. That is, the space is almost empty (Weber *et al.*, 1998). For example, in a 100-dimensional space the first partition contains about 10^{30} blocks. If the space contains 10^6 points, on average only one partition in 10^{24} contains a single point!

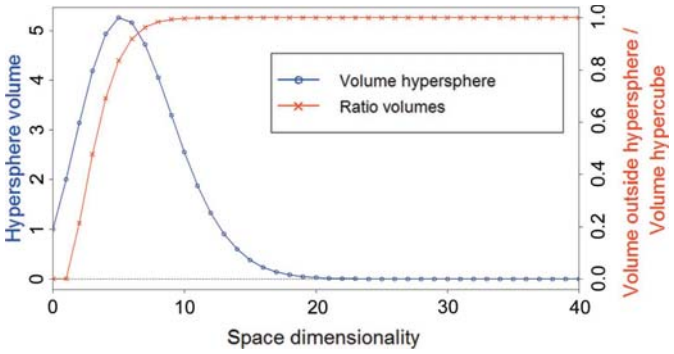

Figure 8
Unit-radius hypersphere volume and the ratio between the circumscribed hypercube volume outside the hypersphere and the hypercube volume itself as function of space dimensionality.

This empty-space phenomenon is the reason why in high-dimensional spaces *there are never enough data points*. For example, histograms are often very different from the underlying theoretical probability density function, because there are too few points to construct them. The same happens for interpolation: having too few points means we cannot be sure of the quality of any of the results.

All these phenomena are usually called 'the curse of dimensionality', an expression introduced by Bellman (1961). Luckily, this curse is not inevitable. Knowing why it arises helps us design strategies for overcoming or at least mitigating its effects. Strategies include selecting a distance measure different from the usual Euclidean one or checking whether the high-dimensional data under analysis are instead a low-dimensional data set embedded into a high-dimensional space.

6.1. High-dimensional analysis tools

Moving into high-dimensional spaces does not change the basic operations we want to do on data: understand them, discover hidden knowledge, search for similar points, group and partition them and, finally, use the data to predict missing points. Visualization is a useful tool for understanding high-dimensional data and for discovering information and knowledge hidden in them. Unfortunately, the techniques used for high-dimensional data are generally unintuitive and far more abstract than scientific visualization techniques (Keim *et al.*, 2004). The use of parallel coordinates (Inselberg, 1985) tries to overcome these difficulties through its sound mathematical foundation.

Another approach to understanding high-dimensional data is to reduce their dimensionality so usual visualization techniques could be applied. To this effect the first step is checking whether the data are truly high-dimensional. Often the data live in a low-dimensional subspace of the high-dimensional one: that is, we need fewer than D free variables to uniquely identify the points. In this case, we can ignore the extra

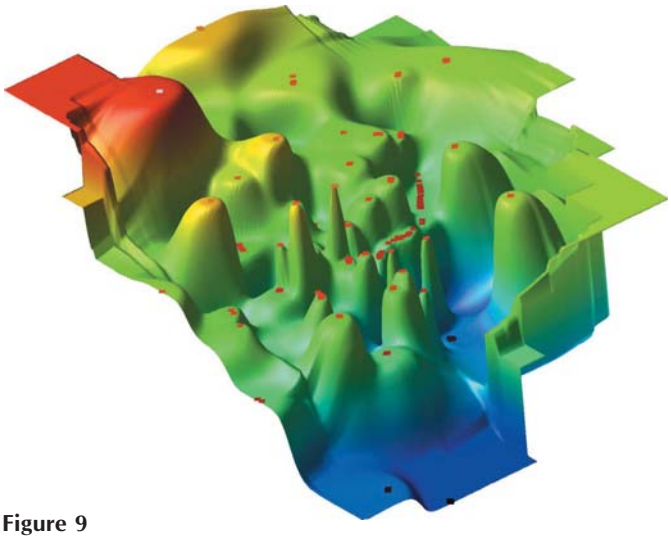

Figure 9
An example of an energy landscape. The points are the two-dimensional projection of the structures in fingerprint space with height values proportional to the structures' energies. The surface interpolates height values between points.

dimensions and project the data into the low-dimensional space without loss of information. Even if this is not possible, we can attempt the same projection while trying to minimize the loss of information and the inevitable distortions. In both cases the first step is to compute the effective dimensionality of the data (Camastra, 2003; Levina & Bickel, 2005; Pettis *et al.*, 1979). Then the extra dimensions can be removed using methods such as principal component analysis (PCA) (Jolliffe, 2002) and multidimensional scaling (MDS) (Borg & Groenen, 2005; Lee & Verleysen, 2007), or we can try to find, and remove, them using our knowledge of the data.

7. Current results

The exploration of the paradigm is still at its beginning, but some results have already been obtained, as seen in §5 and here below, and some promising areas are actively being worked on. The areas on which we plan to work in the near future will be covered in the next section.

7.1. Energy landscapes

To understand the relationship between structures, we project the fingerprint-space points to a two-dimensional or three-dimensional physical space. The 'scatter plot' tool in the end-user application does exactly this, projecting the fingerprint space into the plane. Interpolating the energies associated with each structure (point) creates an energy surface (Fig. 9) which gives an idea of the true energy landscape (Oganov & Valle, 2009). This result adds a highly complementary point of view to the traditional concepts in studies of energy landscapes, such as those reviewed by Wales & Bogdan (2006). Moreover, energy *versus* order and energy differences *versus* distance charts visually revealed unexpected correlations (see §5) that can be interpreted in the framework of these energy landscapes.

7.2. Definition of new quantities

For structures we specified several degree-of-order measures derived from their fingerprints. The simplest one is defined as

$$\Pi = \frac{\Delta}{(V_{\rm uc})^{1/3}}|F|^2, \qquad (3)$$

where $V_{\rm uc}$ is the unit-cell volume and Δ is defined in equation (1). While the angles between fingerprint vectors F measure structural differences, the lengths of these vectors show the degree of order of each structure (Oganov & Valle, 2009). The order seems to increase and saturate or remain almost constant during an *USPEX* run, but exhibits an increasing number of isolated high-order peaks at the end (Fig. 10).

Based on our distance metric, we have also introduced a novel measure of disorder and complexity of structures, called 'quasi-entropy' (Oganov & Valle, 2009) and defined as

$$S_{\rm str} = -\sum_A (N_A/N_{\rm tot})\langle (1 - D_{A_iA_j})\ln(1 - D_{A_iA_j})\rangle, \qquad (4)$$

where distances $D_{A_iA_j}$ are measured between fingerprints of all ith and jth sites occupied by chemical species A, and the total quasi-entropy is a weighted sum over all chemical species. Note that the definition above is slightly modified from the paper of Oganov & Valle (2009), to make it look more like a traditional formula for the entropy.

Quasi-entropy $S_{\rm str}$ has a much better correlation with energy than the set of orientational bond-order parameters Q_n proposed by Steinhardt *et al.* (1983). These parameters successfully differentiate between liquid-like and crystal-like configurations and could be expected to be good predictors of the energy. However, for all systems examined, even as simple as MgO, the correlation of Q_n parameters with energy is either very weak or nonexistent.

7.3. Intrinsic dimensionality

Typical dimensionalities of fingerprint spaces are between 10^2 and 10^3. These values are clearly redundant compared to the theoretical dimensionality $D = 3N + 3$ derived from degrees-of-freedom considerations, where N is the number of atoms in the unit cell. However, even D overestimates the true dimensionality of the space, because it ignores short-range order that leads to certain constraints κ on the relative positions of the atoms. The actual intrinsic dimensionality is thus $D_{\rm intrinsic} = 3N + 3 - \kappa$ or, better, $D_{\rm intrinsic} = (3 - \kappa_A)N + 3$, where $\kappa_A = \kappa/N$ is the mean constraint per atom.

In our tests (see Table 1) we saw this rule almost always obeyed, with κ_A distributed in the range $[0 \ldots 3]$. However, we encountered a few data sets that violated this rule, where the computed intrinsic dimensionality is much higher than the theoretical one. There are various possible explanations for this anomaly: (1) the quality of the statistics; (2) non-randomness in the data sets; (3) approximate relaxation that introduces noise; and (4) the thresholds used for removing identical structures (before analyzing the dimensionality of the data set, we remove identical structures, but the result depends on the distance threshold used). The analysis of these

Table 1
Intrinsic dimensionalities.

Data set	No. of atoms	Theoretical dimensionality	Intrinsic dimensionality	κ_A
Structures from random sampling				
SiO_2	3	12	8.71	1.10
SiO_2	6	21	32.49	−1.92
SiO_2	9	30	29.69	0.03
SiO_2	12	39	29.85	0.76
SiO_2	24	75	26.17	2.03
SiO_2	36	111	35.98	2.08
SiO_2	48	147	50.04	2.02
H_2O	12	39	17.91	1.76
GaAs	8	27	23.45	0.44
MgNH	12	39	63.69	−2.06
Structures from *USPEX* runs and random sampling				
Au_8Pd_4†	12	39	11.94	2.25
Binary Lennard-Jones crystal A_4B_8‡	12	39	2.87	3.01
$Mg_{16}O_{16}$‡	32	99	15.05	2.62

† Structures from an *USPEX* run. ‡ Random sampling and *USPEX* data.

and other effects that may distort the dimensionality estimates will be part of our next steps.

We computed the intrinsic dimensionality using the Grassberger–Procaccia algorithm (GPA) (Grassberger & Procaccia, 1983) plus Camastra's correction (Camastra & Vinciarelli, 2001). This correction is needed because the GPA, to give correct estimations of the intrinsic dimensionality, requires an unrealistically large number of points. Camastra's method computes a function that computes the real intrinsic dimensionality given the measured one, and utilizes it to correct the results of the GPA applied to limited-size data sets. The function above is obtained from the measurement of the intrinsic dimensionality for sets of random points. For them the intrinsic dimension is equal to the embedding one, so the function can correlate the measured intrinsic dimension to the real, known one.

We then applied a second correction to remove noise. In fact, noise increases subspace dimensionality when the embedding space dimension grows. The result is an intrinsic dimensionality measure that grows with embedding dimension. After correction we have an estimate of the intrinsic dimensionality that reaches a constant value above a minimum embedding dimensionality.

8. Future

To continue this line of research we plan to deal with various scientific questions related to the structure of fingerprint spaces and to study the relationship between physical and crystallographic quantities with abstract quantities extracted from the space structure. Below are collected the main areas we plan to approach together with a collection of questions and references. There are also issues related to the definition of more robust crystal fingerprints, to clustering algorithms and to implementation choices that we should approach, but these points will not be covered here.

Analyze the intrinsic dimensionality of fingerprint spaces. First we must gain better understanding of the Grassberger–Procaccia algorithm (GPA). This algorithm started as a tool for the study of chaotic dynamic systems, but those systems are temporally correlated and have small embedding dimension, so they are quite different from ours. Are these differences significant for our usage of the GPA? We should also explore the GPA validity, derive statistical confidence intervals for computed intrinsic dimensions and look at alternatives to the GPA, like the Judd (1994) and Takens (1985) estimators. On the scientific side, we must understand how κ_A depends on the experimental conditions and we must find out what causes a few data sets to violate the $D_{\text{intrinsic}} \leq 3N + 3$ rule.

Study local space behavior. Intrinsic data dimensionality is a local feature, but we compute the global intrinsic dimension, which is an average of the intrinsic dimensionalities of the subspaces. We should try the local intrinsic dimension (LID) method (Buzug *et al.*, 1995) to see whether, from a local analysis, we can gain new insights on single structure behavior and a rough idea of the cluster shapes. Are there local topological quantities that help us to locate a point (inside, on the surface *etc.*) with respect to these clusters?

Study cosine distance behavior. The behavior of the cosine distance measure using random sets of fingerprints should be studied. These distances have a beta distribution (similar to Gaussian, but with support limited to $[0 \ldots 1]$ and the capability to model asymmetric distributions). We have already found that the variance of the distance distributions depends on the intrinsic dimension according to $\sigma^2 \propto 1/D_{\text{intrinsic}}$. Could this fact be generalized?

Study distance distributions. Looking at the statistics of distances between structures, we often discover a striking

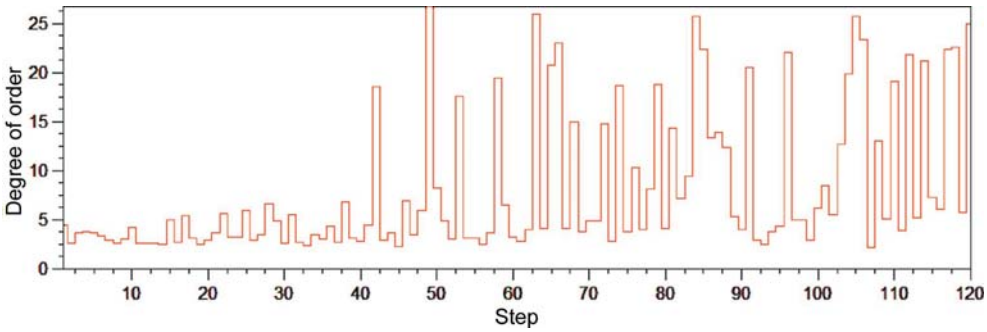

Figure 10
Emergence of order from disorder in the evolutionary structure prediction for GaAs (eight atoms in the unit cell).

Gaussian-like shape of distributions, with a clear peak (Fig. 7*a*), but not always (Fig. 7*b*). We have already started analyzing these curves by decomposing them into sums of beta distributions.

Understand correlations with other physical quantities. Initial tests show correlation between pressure and intrinsic dimensionality and between pressure and order. It should be interesting to study magnetic behavior, hardness, superconducting T_c and other quantities as they correlate with distance and space topology. For the multicomponent fingerprint we will also study the correlation of each fingerprint component $F_{AB}(R)$ with energy.

Study landscape structure. Can we classify landscape extrema in fingerprint space directly? Is the scaling behavior of the number of minima *versus* the number of atoms, the distribution of minima energy and the number of minima *versus* energy consistent with published studies?

9. Is this model valid?

In the search for confirmations of the validity of the fingerprint method, we are prepared not to be able to assess insight quality in the same manner as we do in physical spaces. Here we can only look at the analysis results and compare them with the real world using our crystallographic intuition. That said, we do not have any objective proof that distances as computed in fingerprint space really reflect true differences between crystal structures, but we are confident that the things we study are not meaningless for various reasons:

(1) Studies of sets of very similar structures (*e.g.* an Si crystal with moving vacancies) show a clear correlation between distances in fingerprint space and distances between physical configurations (Sahli, 2009).

(2) Similar results are obtained using synthetic data sets. In this case there are no chemical or crystallographic constraints that could affect the results.

(3) Pragmatically, the results obtained so far make sense both from the numerical and from the crystallographic point of view, so we conclude that the underlying model makes sense. For example, the correlations we have found between distances and other quantities were unexpected outcomes, but in some sense they validate our choice of space structure and distance measure.

(4) Last, by analyzing how results change when the computation method changes, we become confident that what we see is not an artifact of a particular computational method.

10. Lessons learned

This research is still at its beginning and there is a lot of work to do (see §8). However, the paradigm has already demonstrated its usefulness in real situations (Ma *et al.*, 2009; Oganov *et al.*, 2007; Oganov & Valle, 2009; Oganov *et al.*, 2008). We are thus confident that fingerprint spaces could become a useful tool for studying structural transitions and crystal ensembles.

Apart from the scientific results, we have learned various valuable lessons from this research:

(1) As has been said before, 'discoveries happen on the border between disciplines'. This is the most important lesson learned.

(2) An exploratory approach is very important for understanding data and, primarily, for helping to formulate the right questions.

(3) The role of visualization when coupled to analysis is fundamental for data understanding.

Calculations were performed at the Joint Russian Supercomputer Centre (Russian Academy of Sciences, Moscow), ETH Zürich, Swiss National Supercomputing Centre (Manno) and New York Center for Computational Sciences. The authors gratefully acknowledge the use of these facilities.

References

Advanced Visual Systems (2009). *AVS/Express*, http://www.avs.com/software/soft_t/avsxps.html.
Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.
Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. (1999). *ICDT '99. Proceedings of the 7th International Conference on Database Theory*. *Lect. Notes Comput. Sci.* **1540**, 217–235.
Borg, I. & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. New York: Springer-Verlag.
Buzug, T. M., von Stamm, J. & Pfister, G. (1995). *Phys. Lett. A*, **202**, 183–190.
Camastra, F. (2003). *Pattern Recognit.* **36**, 2945–2954.
Camastra, F. & Vinciarelli, A. (2001). *Neural Process. Lett.* **14**, 27–34.
Chisholm, J. A. & Motherwell, S. (2005). *J. Appl. Cryst.* **38**, 228–231.
Ertoz, L., Steinbach, M. & Kumar, V. (2003). *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM '03)*, pp. 47–58. Philadelphia: Society for Industrial & Applied Mathematics.
François, D., Wertz, V. & Verleysen, M. (2007). *IEEE Trans. Knowl. Data Eng.* **17**, 873–886.
Gelder, R. de (2006). *IUCr CompComm Newsl.* **7**, 59–69.
Grassberger, P. & Procaccia, I. (1983). *Physica D*, **9**, 189–208.
Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. USA: Springer.
Hemmer, M. C., Steinhauer, V. & Gasteiger, J. (1999). *Vib. Spectrosc.* **19**, 151–164.
Hundt, R., Schön, J. C. & Jansen, M. (2006). *J. Appl. Cryst.* **39**, 6–16.
Inselberg, A. (1985). *Vis. Comput.* **1**, 69–91.
Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). *ACM Comput. Surv*. **31**, 264–323.
Jolliffe, I. T. (2002). *Principal Component Analysis.* USA: Springer.
Judd, K. (1994). *Physica D*, **71**, 421–429.
Keim, D. A., Ankerst, M. & Sips, M. (2004). In *The Visualization Handbook*, edited by C. Hansen & C. Johnson, pp. 813–825. Academic Press.
Köppen, M. (2000). *The Curse of Dimensionality*. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), 4–8 September. IEEE Finland Section.
Lee, J. A. & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction.* USA: Springer.
Lever, P. G., Leaver, G. W., Curington, I., Perrin, J. S., Dodd, A. W., John, N. W. & Hewitt, W. T. (2000). *Design Issues in the AVS/Express Multi-Pipe Edition*, 8th IEEE Visualization Conference (Vis'00), Salt Lake City, USA. *IEEE Visualization 2000 Conference Works in Progress*, pp. 1-9. IEEE.
Levina, E. & Bickel, P. J. (2005). *Maximum Likelihood Estimation of Intrinsic Dimension*. In *Advances in NIPS*, Vol. 17, edited by L. K. Saul, Y. Weiss & L. Bottou, pp. 777–784. Cambridge: MIT Press.

Lyakhov, A. O., Oganov, A. R., Ma, Y., Wang, Y. & Valle, M. (2009). In *Modern Methods of Crystal Structure Prediction*, edited by A. R. Oganov, ch. 8. USA: Wiley-VCH.
Ma, Y., Eremets, M., Oganov, A. R., Xie, Yu., Trojan, I., Medvedev, S., Lyakhov, A. O., Valle, M. & Prakapenka, V. (2009). *Nature (London)*, **458**, 182–185.
Oganov, A. R. & Glass, C. W. (2006). *J. Chem. Phys.* **124**, 244704.
Oganov, A. R., Ma, Y., Glass, C. W. & Valle, M. (2007). *Psi-k Newsl.* **84**, 142–171.
Oganov, A. R. & Valle, M. (2009). *J. Chem. Phys.* **130**, 104504.
Oganov, A. R., Valle, M., Lyakhov, A., Ma, Y. & Xie, Y. (2008). *Evolutionary crystal structure prediction and its applications to materials at extreme conditions*. XXI IUCr Congress, 23–31 August, Osaka, Japan. *Acta Cryst.* (2008). A**64**, C41–C42
Pettis, K. W., Bailey, T. A., Jain, A. K. & Dubes, R. C. (1979). *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 25–37.
Sahli, B. (2009). Personal communication.
Salton, G. & Buckley, C. (1988). *Inf. Process. Manag.* **24**, 513–523.
Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.
Schmitt, I. (2001). *Nearest Neighbor Search in High Dimensional Space by Using Convex Hulls*, Preprint No. 6, University of Magdeburg, Germany.
Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. (1983). *Phys. Rev. B*, **28**, 784–805.
Takens, F. (1985). *Dyn. Syst. Bifurcations*, **1125/1985**, 99–106.
Valle, M. (2005). *Z. Kristallogr.* **220**, 585–588.
Valle, M. (2009). *STM4 - the molecular visualization toolkit*, http://www.cscs.ch/~mvalle/STM4.
Valle, M. & Oganov, A. R. (2008). *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST08)*, pp. 11–18. IEEE.
Wales, D. J. & Bogdan, T. V. (2006). *J. Phys. Chem. B*, **110**, 20765–20776.
Weber, R., Schek, H.-J. & Blott, S. (1998). *A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces*. *Proceedings of the 24th VLDB Conference, New York, USA*, pp. 194–205. Morgan Kaufmann.
Willighagen, E. L., Wehrens, R., Verwer, P., de Gelder, R. & Buydens, L. M. C. (2005). *Acta Cryst.* B**61**, 29–36.