Crystal Structures Classifier for an Evolutionary Algorithm Structure Predictor

Mario Valle*
Data Analysis and Visualization Services
Swiss National Supercomputing Centre (CSCS)

Artem R. Oganov†
Laboratory of Crystallography, Department of Materials
ETH Zürich

ABSTRACT

USPEX is a crystal structure predictor based on an evolutionary algorithm. Every USPEX run produces hundreds or thousands of crystal structures, some of which may be identical. To ease the extraction of unique and potentially interesting structures we applied usual high-dimensional classification concepts to the unusual field of crystallography. We experimented with various crystal structure descriptors, distinct distance measures and tried different clustering methods to identify groups of similar structures. These methods are already applied in combinatorial chemistry to organic molecules for a different goal and in somewhat different forms, but are not widely used for crystal structures classification. We adopted a *visual design and validation* method in the development of a library (CrystalFp) and an end-user application to select and validate method choices, to gain users' acceptance and to tap into their domain expertise. The use of the classifier has already accelerated the analysis of USPEX output by at least one order of magnitude, promoting some new crystallographic insight and discovery. Furthermore the visual display of key algorithm indicators has led to diverse, unexpected discoveries that will improve the USPEX algorithms.

Index Terms: J.2 [Physical Sciences and Engineering]: Chemistry; I.5.2 [Pattern Recognition]: Design Methodology—Classifier design and evaluation

1 INTRODUCTION

USPEX [20] is a computational method and application based on an evolutionary algorithm that enables crystal structure prediction at arbitrary P-T conditions, given just the chemical composition of the material.

Due 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. Before analysis by the crystallographers it is therefore necessary to reduce the results to a set of unique structures to concentrate their analysis on the configurations that could give insight on new phenomena. This is indeed an intensive manual labor, consisting mainly in judging equality from side-by-side visualization of pairs of structures.

This reduction step, once automated, could be exploited also in another context. It could be integrated inside USPEX to improve the effectiveness of its evolutionary algorithm by avoiding the dilution of diversity at each generation caused by the presence of identical structures.

We decided therefore to design an automatic structure comparison and clustering method, initially to support the post-run classification task, but with the final goal of incorporating it inside USPEX.

*e-mail: mvalle@cscs.ch
†e-mail: a.oganov@mat.ethz.ch

IEEE Symposium on Visual Analytics Science and Technology
October 21 - 23, Columbus, Ohio, USA
978-1-4244-2935-6/08/$25.00 ©2008 IEEE

The approach adopted applies to the classification problem methods common to the visual analytics and data mining communities, but not widely employed in the crystallography field. Crystal structures are thus described as points in a multidimensional space, each identified by a multidimensional coordinate set (here called *fingerprint*). This space has a similarity metric defined so we can measure structure "closeness" and then use clustering methods to group equivalent structures. This model is targeted to the project specific usage and not necessarily intended to be a solution to the general problem of finding equivalent structures for generic molecules or judging their degree of similarity as is required, for example, in combinatorial chemistry (a survey can be found in [18]).

This specificity means that the method should be tailored to the comparison of structures as used by crystallographers (see sect. 1.1) and that in general it could be limited to a binary same/different crystal structure answer. Nonetheless, during the domain experts' exploration of the classifier capabilities, they found that a suitable distance definition could show interesting correlations with other structure's properties (see sect. 5.1), hence we adopted more usual, but crystal structure-specific, distance measures.

Standard multidimensional techniques cannot be blindly applied to crystal structure data. The difficulties stem mainly from the very same structure of crystallographic data: a crystal is an infinite repetition of a basic cell (*unit cell*, see fig. 1) and different unit cells could describe the same crystal structure. Not to mention that small numerical errors in atom positions and unit cell parameters could make automated comparison difficult when progressing from a single unit cell to a whole crystal (fig. 2).

For this project we adopted a *visual supported design* approach for the classifier design. Therefore we implemented an end-user application to access the fingerprinting and grouping algorithms making them immediately usable on real problems and providing *interactive diagnostics* on their behavior. We thus had the opportunity to refine the library structure and algorithms starting from usage results on real crystal classification problems.

The visual design approach provided also other benefits: first, it greatly facilitated gaining the users support for the project by convincing them that the approach was feasible and at least as good as the manual way of work. They indeed feared to miss important results using an automated method for which they see only the final results. Second, a visual exploratory approach simplified the access to the domain expert experience, to explore different alternatives, to validate design decisions and to propose and test unanticipated ideas.

The use of the resulting classifier library (CrystalFp) and the related end-user application has already accelerated the analysis of USPEX output by at least one order of magnitude, promoting some new crystallographic insight and discovery, and has increased the user confidence in the viability of the classification methods on real problems.

1.1 Problem Context

To put the problem in context we should consider the input data origin and the kind of crystallographic structures processed.

A crystal structure is defined by its *unit cell*, the smallest group

of atoms or molecules whose repetition at regular intervals in three dimensions produces the whole infinite structure of a crystal. For any crystal the unit cell is not unambiguously defined because an infinite crystal structure could be "cut" in different ways to define the elementary cell (see fig. 1). Each cell thus not necessarily contains the same number of atoms.

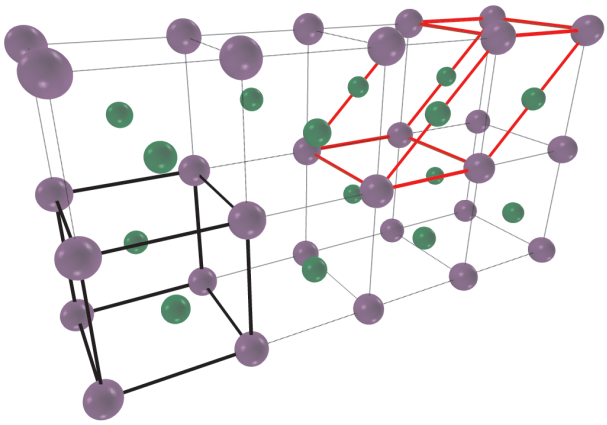

Figure 1: Crystal unit cell (lower left) and its repetitions that build the whole crystal structure. An alternative unit cell for the same crystal is show on the top right.

Another (incomplete) structure descriptor is the space group, i.e. the set of all symmetry operators in the structure. If the symmetry for a crystal is known, then it provides a strong constraint on structure similarity.

A value of internal energy or enthalpy could be associated to a structure. Lower values mean more stable structures. Equal structures, also if represented in different ways, should have equal energy values. But the energy equality alone is not sufficient as a criterion for grouping due to numerical imprecision or because different structures could have energies so similar that discrimination is made impossible.

The crystal structures considered in this work differ from structures of interest in other fields, like biochemistry for example, mainly in the number of atoms composing the unit cell and the kind of chemical elements involved. Another important difference is that biomolecules have bonds topologies that could be used to constrain the classification process; instead crystal structures generally do not carry uniquely defined bonds information.

A typical USPEX run produces 300–3000 structures usually containing 6 to 40 atoms of one to four element types. The USPEX output data are in concatenated VASP [16] POSCAR file format. This format does not carry the exact atoms' types and the optional structure energy; these information should be provided externally to the application. The structure symmetries are not available too, and thus are not considered as a possible input to the clustering process.

1.2 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 structure descriptors. Some of these few use symmetry criteria or some form of structure standardization in the comparison phase [1–3, 6, 7, 9, 14, 22]. For our application these methods seem not sufficiently robust with respect to numerical errors. Others use related structure data relevant to the problem they try to solve, like partial charges [31] or powder diffraction patterns [5]. Furthermore methods based on crystal symmetries are not applicable because we do not have symmetry data available and because symmetry is extremely sensitive to small numerical errors. The work of Hundt et al. [14] gives a comprehensive survey of existing methods with a focus on calculating some form of distance metric between structures.

Chisholm and Motherwell [4] use interatomic distances to investigate molecular packing and inspired one of the methods we tried (see sect. 3.1). Interatomic distances are a good choice for a structure identifier because they are independent from coordinate reference frame and cell choices.

Radial Distribution Function (RDF) is another possible structure descriptor method based only on local characteristics. Beside interatomic distances, it considers also other data associated to the structure. The method of Willighagen et al. [31] computes a RDF using distances from a central atom weighted by the atoms partial charge to include electrostatic interactions that play a major role in crystal packing. This method then calculates dissimilarities on the basis of powder diffraction patterns as in [5]. In our work we use a rapidly convergent function based on distance distributions and related to RDF and diffraction spectra, but we focus more on standard multidimensional methods for distance computation. Another interesting application of RDF is the work of Hemmer et al. [13] that uses RDF to match structures to IR spectra using a counterpropagation neural network. Their goal is indeed different from ours, but they also found that RDF's could be good crystal structures identifiers.

2 Visual Approach

We approached the design of the classifier from two sides. The first is the application of multidimensional methods to crystallographic data (see sect. 3). The second was in retrospect the most important one: we decided to use a *visual supported design and validation* approach for the classifier design. Visual analytics is not only fancy graphics, it is the use of visualizations, also simple ones, that could foster scientific insight in the domain expert user.

The visual approach could lead to better scientific ideas and results because the scientists are involved in the design on a level they are familiar with, the system could visualize quickly the result of their suggestions and, last, the visual imagery has the power to suggest unexpected ideas and correlations not planned in the design phase.

3 Crystal Fingerprinting

The proposed method associates a descriptor, called a *fingerprint*, to each structure. This descriptor is a vector of N real values; each structure becomes thus a point in a N-dimensional space. A distance measure between these vectors is then used to cluster them into groups of "near" fingerprints, that is, groups of similar structures. This is indeed an application to the crystallography field of concepts already taken for granted in visual analytics or data mining.

To support the classification phase the fingerprint should, as much as possible, uniquely identify every structure, tolerate numerical errors and enhance contrast between different structures. The chosen distance between fingerprints measure could help classification too by increasing contrast between neighbor points and by using as much as possible the information available. In our high dimensional fingerprints space peculiar phenomena, like distance concentration [11], work against this goal. The distance measurement method should therefore counteract this effect too. In the last step, to increase the classification quality, the classifier should create well separated, but highly internally connected clusters.

3.1 Fingerprint Definition

To support the specific kind of data we consider, the fingerprint associated to a structure should be independent from: 1) translation

and rotation of the structure; 2) the choice of unit cell among equivalent unit cells; 3) the ordering of cell axis and atoms in the cell; 4) inversions and mirroring of the structure.

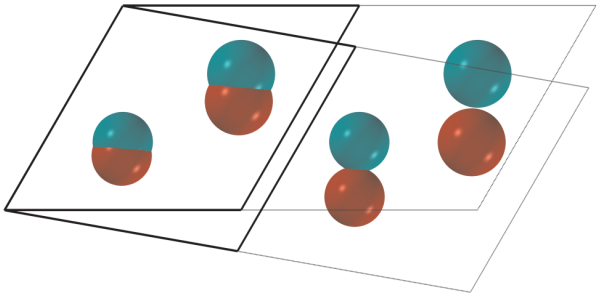

Figure 2: The effect of numerical errors on the unit cell is visible moving away from the origin along the crystal structure.

Whichever definition we chose for the structure fingerprint, it should be computed over the "infinite" crystal structure. In practice after a distance of $D_{uc}/2$, where D_{uc} is the longest unit cell diagonal, everything start repeating in every direction. Therefore in place of the whole crystal structure, a set of unit cell repetitions that cover the maximum distance over all structures in all directions around the base unit cell is used. We call this the *extended unit cell*.

We experimented with two fingerprint definitions: 1) per-atom distance sets and 2) a rapidly convergent function based on distance distributions and related to RDF and diffraction spectra. The idea behind these fingerprint definitions is to reduce a global property, like the crystal structure, to a local one, like interatomic distances or RDF. Both definitions satisfy the criteria stated above for a good fingerprint.

The per-atom distance sets fingerprint is composed by a section for each atom in the unit cell. Each section is an ascending ordered set of distances from the corresponding atom to all the other atoms of the extended unit cell. All these sets contain the same number of distances. Furthermore each section is labeled with the element type of the corresponding atom. An example is shown in fig. 3. The idea behind this labeling method is this: if two structures are

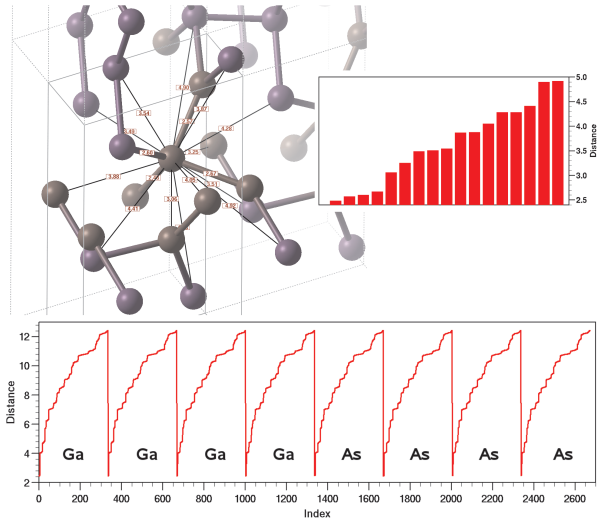

Figure 3: Local atom distances for a GaAs crystal (top left) are concatenated to form a fingerprint section (top right). Sections are then assembled to form the structure fingerprint (bottom).

the same, then at least one atom of the same type in each should have the same set of distances from its neighbors. This definition of fingerprint then shifts the burden of corresponding atoms matching to the distance computation phase (see sect. 3.2).

The second fingerprint definition we have experimented with starts from the following function based on atoms identities and distances distribution:

$$FP(R) = \frac{V_{uc}}{N_{uc}B} \sum_{i \in uc} \sum_{j \in euc} \frac{Z_i Z_j}{4\pi {R_{ij}}^2} \delta(R - R_{ij}) \text{ where } i \neq j \qquad (1)$$

Here i runs over the atoms of the unit cell, j over the atoms of the extended unit cell; Z is the atomic number; R_{ij} is the distance between atoms i and j; V_{uc} is the unit cell volume; N_{uc} is the number of atoms in the unit cell and B is the bin size used to compute the value of $FP(R)$ from the discrete peaks. Each peak is then smoothed using a Gaussian kernel with σ set by the user (usually 0.02 Å) and accumulated into a histogram with bin size B (usually 0.05 Å). The resulting function is closely related to the diffraction spectra of the crystal. To remove fingerprint dependency from bin size and cutoff distance, the histogram is normalized:

$$FP_{norm}(R) = \frac{FP(R)}{\sum_i \sum_j Z_i Z_j N_i N_j} - 1 \qquad (2)$$

Here N_i is the number of atoms in the unit cell with atomic number Z_i and the two sums go over all distinct Z values. One example of normalized diffraction-like fingerprint is given in fig. 4.

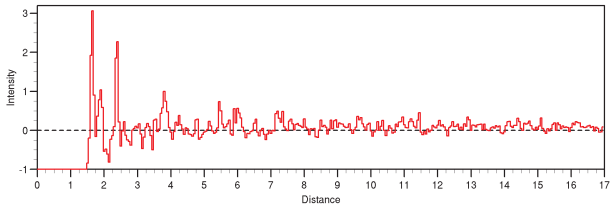

Figure 4: Diffraction like fingerprint.

When atomic numbers Z are used as weights, the fingerprint is related to X-ray powder diffraction spectra. However, atoms with very different properties may have similar atomic numbers (e.g. $Z(I) = 53$ and $Z(Cs) = 55$) and will be hard to distinguish them by X-rays or by the above fingerprint. For these cases, instead of Z, we can use the Chemical Scale χ value or Mendeleev Number m [23] which correctly maps chemical differences between elements. This substitution could increase the discriminating power of the fingerprint in realistic situations: for example the relative variation of one fingerprint term due to the exchange of two atoms is proportional to $|Z_1 - Z_2|/Z_1 Z_2$ and thus for a $Si \leftrightarrow O$ exchange it varies by 5.4% using Z, by 0.2% using m, but using χ it changes by 23.0%.

The decision to use a certain algorithm or a certain set of parameters (e.g. cutoff distance, bin size, Gaussian smoothing width) is made by the domain expert after interacting with the CrystalFp application to classify sets of real data that have been previously manually analyzed. So the algorithm performance is determined in a rather informal way. The evaluation result is nevertheless compatible with the expert's crystallographic intuition. In the case of the fingerprint definition this experimentation selected the fingerprint defined by eq. (1) and (2).

3.2 Distance Measurement

To make possible the classification of structures we should define a distance or pseudo-distance (i.e. one distance measure for which

triangular inequality does not hold) between fingerprints. We experimented with three distance measures: 1) Cartesian distance; 2) Minkowski norm with fractional exponent and 3) cosine distance.

The Cartesian distance between same type atoms is the more physically-based measure, and the most obvious one. For each atom of the first structure we select from the second structure the atom of the same type with the minimum Cartesian distance between the two atoms fingerprint sections. Remember that an atom fingerprint section is composed by the ordered set of distances between this atom and its neighbors up to a certain cutoff distance. After this pairing has been done for all atoms, the sections of the two fingerprints are reordered to have corresponding atoms sections aligned. Then the usual Cartesian distance between the reordered fingerprints is computed.

As seen before, we discontinued the use of Cartesian distance measure in favor of more crystal-specific measures after experimentation. Cartesian distance is thus a good illustration of how blindly applying multidimensional techniques is not the best strategy to use them in the crystallographic domain.

The second proposed distance measure replaces the Cartesian distance with a more general Minkowski norm with a fractional exponent $0 < p < 1$ as suggested by [11, sect. 2.6]. The distance between fingerprints based on the Minkowski norm with exponent p is defined as:

$$dist(i,j) = \left(\sum_k |fp_{i_k} - fp_{j_k}|^p \right)^{\frac{1}{p}} \tag{3}$$

Where $FP_i = (fp_{i_1}, fp_{i_2}, \ldots)$ is the fingerprint associated to structure i. The Minkowski norm is really a pseudo-distance, but could alleviate the distance concentration phenomena visible in high dimensional spaces [11]: in these spaces all pair-wise distances seem to be equal or at least very similar. The reduction of concentration is application dependent, so we experimented with various p values; for our data better results have been obtained with $p = 1/3$.

The cosine distance is a popular norm in the text mining community [25, 26]. Here every text has associated a 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 FP_i associated to structure i:

$$dist(i,j) = \frac{1}{2} \left(1 - \frac{FP_i \cdot FP_j}{\|FP_i\| \|FP_j\|} \right) \tag{4}$$

To summarize, the choice of metric depends on the kind of data set under analysis, and the search of the best metric is exactly one of the goals of our visual design approach. As we will see, we have obtained better results using the cosine distance metric (eq. 4), not last for its ability to counteract the distance concentration phenomena, as seen in fig. 5, because it spreads distances much more than the other methods.

3.3 Structure Clustering

After building the matrix of distances between all pairs of fingerprints, we assign two structures to the same cluster if their distance is less than a user-defined threshold. This operation transforms the distance matrix into a binary connection matrix. The graph described by this matrix is then separated into connected components that are our clusters of similar structures. We explored the following methods to extract the connected components (groups) and unconnected entries (singles): 1) Depth First Search (DFS) [15]; 2) Shared Nearest Neighbor (SNN) [8] and 3) Pseudo SNN.

The first method adds one by one connected structures to a cluster doing a Depth First Search on the connection graph. The SNN

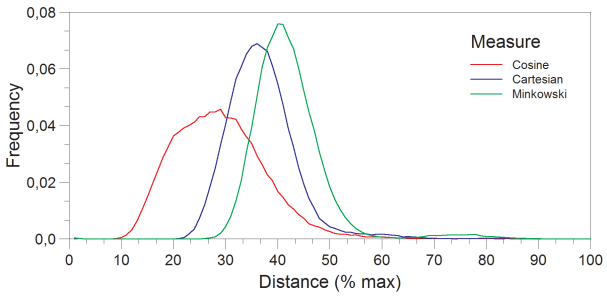

Figure 5: Distance distribution for the three distance definitions over the $MgSiO_3$ Post-perovskite 120GPa dataset. The curves are generated by one of the interactive diagnostics tools (see sect. 4.3) and presented together after normalizing the distance values.

algorithms are instead density based: they define density as the number of nearest neighbors points shared between pairs of connected points and confirm this last connection only if the number is at least K. The Pseudo SNN clustering method stops at this point, instead the full SNN algorithm adds a DBSCAN [27] step to refine the cluster points' membership.

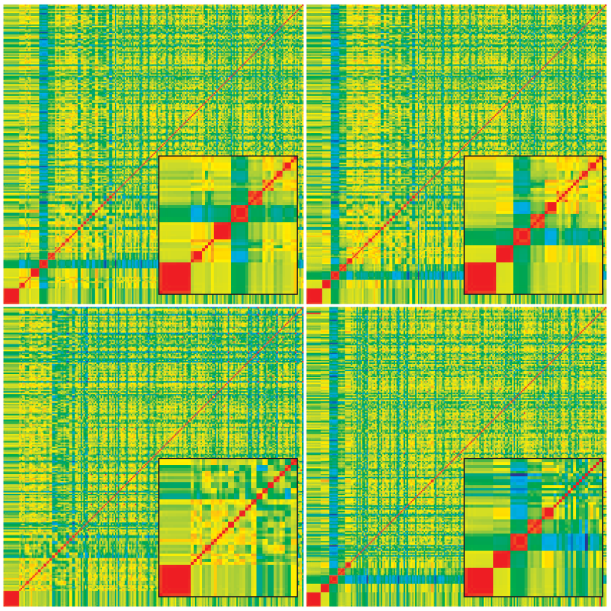

Figure 6: Ordered distance matrix for: (top left) DFS grouping; (top right) Pseudo SNN with $K = 1$; (bottom left) Pseudo SNN with $K = 5$; (bottom right) SNN with $K = 5$. Inserts show the first 48 structures grouped.

Fig. 6 shows the effect of algorithm choice on the grouping structure. Every cell of the matrix is colored from red to blue by the distance between the corresponding row and column structures and the structures are then ordered to keep the ones pertaining to the same group together in decreasing order of group size (see sect. 4.3). Beside the first group, that has uniformly low distances between structures inside the group for all methods, the differences are in the number of groups and in the uniformity of distances inside groups. The the patch color uniformity is visually estimated. DFS groups structures that have distances that are not uniformly low. Pseudo SNN with $K = 5$ creates smaller groups, but with the same non uniformity of distances and it does not consider various similar pairs

for grouping as testified by the regularly spaced red points far from the diagonal. SNN creates few groups containing similarly low distances. Pseudo SNN with $K = 1$ seems a good compromise between SNN and DFS. There is another reason for not choosing the full SNN method even if it is resistant to noise and can create more connected clusters: it is well known that its DBSCAN phase does not work well with varying cluster densities and high-dimensional data. We therefore opted to use the pseudo SNN clustering method with $K = 1$.

4 Visual Design and Validation

To support the *visual design* of the classifier library we must provide ways to explore algorithm choices and parameters setting and we must make available visual tools to verify and validate these settings.

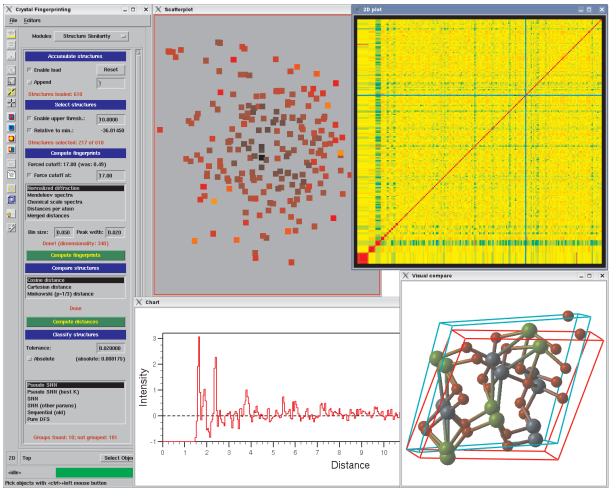

Figure 7: The CrystalFp end-user application. The control panel (left), the scatterplot and ordered distance matrix (top), one diagnostic chart and the visual pair-wise structure comparison (bottom).

To make this exploration possible we built an end-user application around the classifier library to support the USPEX results analysis workflow and to provide *interactive visual diagnostics* on the behavior of the algorithms. As we have seen in the previous section, the choices made for the fingerprint definition, the distance metric and the classifier algorithm are direct consequence of the domain expert exploration and validation of the library during analysis runs made on real data.

4.1 Analysis Workflow

The analysis starts from the USPEX output that is composed by one or more files containing crystal data and, optionally, files with the corresponding energies or enthalpies. These files are then loaded inside the CrystalFp end-user application. Multiple results files are needed, for example, when investigating the relationship between distances and energy differences from a ground-state structure (see sect. 5.1). During load the user should input the element type of the atoms, because this information is not contained in the USPEX output files.

To focus the analysis over the most stable structures a filtering step is applied after loading to retain only the lowest energy ones. Then the three phases of the classification –fingerprint computation, distance computation and grouping– are run in sequence. This splitting into three independent phases facilitates methods exploration and parameters adjustment driven by the visual diagnostic tools as we see below.

After classification, to support further analysis, the system could generate two kinds of outputs: 1) a report of the groups' content together with structure energies and 2) the reduced structure and energy files, where each group is substituted by its most characteristic structure. This structure is the lowest energy one or, if energies have not been loaded, the structure with the highest silhouette coefficient in the group (see sect. 4.3).

4.2 Previous Analysis Workflow

To appreciate the streamlined workflow made possible by the CrystalFp end-user application, we take a look at the previous manual structure classification method.

The USPEX output is loaded inside a visualization tool and then the crystallographer has to retrieve on screen all the pairs of structures and visually compare them. Besides being a time consuming method, it is worth mentioning its main difficulty: the comparison in general cannot be limited to a single pair of unit cells, but should be done on the "infinite" crystal structure, and this makes comparison methods based on naïve superposition strategies difficult to use.

4.3 Interactive Diagnostics

The classification application provides visual representation of key algorithm's quantities so the domain expert could judge the algorithm behavior. The visual validation and analysis, plus the user's algorithm selection and parameters modification, supports a very effective exploratory design approach. The visual tools provided are: 1) distance matrix; 2) grouping quality; 3) group evolution display; 4) scatterplot; 5) diagnostic charts and 6) visual structure pair comparison.

Distance matrix The full distance matrix between structures visually reveals distance trends and overall distance distribution. When the distance map is sorted to put together structures of the same group, it becomes the primary method to judge grouping quality. Good grouping produces uniformly red squares on the bottom left with the rest of the map tending toward green or blue (fig. 8).

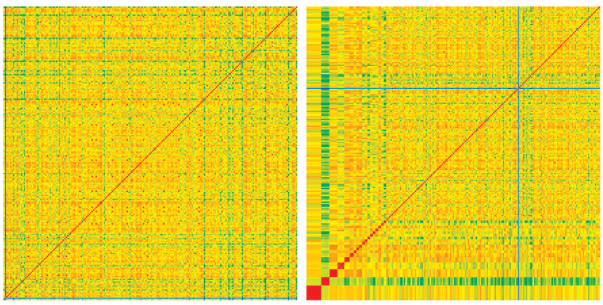

Figure 8: Distance matrix (left) and sorted distance matrix (right).

Grouping quality To judge grouping quality we use the popular method of silhouette coefficients [28, p. 541]. This quality

Figure 9: Silhouette coefficients for all grouped structures. Along Y are the groups, along X the elements of the group. Blue: +1, white: 0, red: -1.

measure combines cohesion inside a group with separation between groups. A value of -1 means the element is probably in the wrong group, because it has low cohesion with the rest of its group and is

too close to another group, whereas a value of $+1$ means the element is closer to the other elements of its group than to the elements of all other ones. The coefficients are visualized in a chart where every horizontal slab represents a group with the elements colored by the respective silhouette coefficient.

Group evolution This visualization shows the various groups forming and evolving during an USPEX run. Here every structure

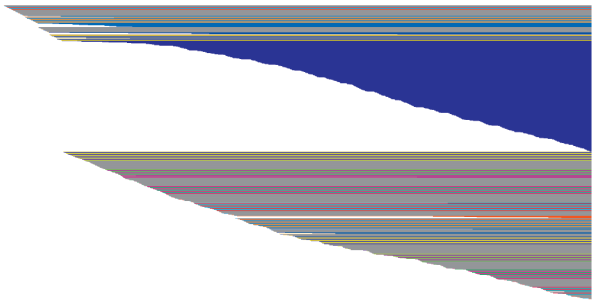

Figure 10: Group evolution display.

starts a horizontal line at a position from the left side proportional to its sequential index. If the structure belongs to a group its line is added below the ones representing the structures already pertaining to the group, otherwise it is added below all the other lines. Fig. 10 shows an example where a big group appears and continues steadily to grow.

Scatterplot To provide an intuitive view of the fingerprints' multidimensional space, we map it to a 2D space. We use a simple implementation of force-directed 2D point's placement algorithm [12] plus a random perturbation step to let the point configuration escape from local minima. The scatterplot points could be colored by group with black representing non-grouped structures (fig. 11), this way the user has an intuitive view of the goodness of the clustering algorithm. If the points are colored by a measure of

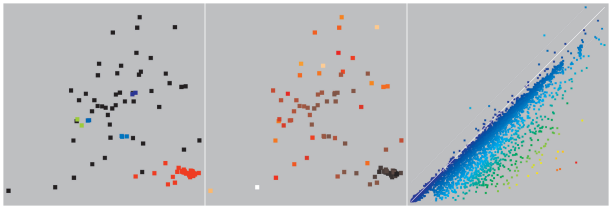

Figure 11: Scatterplot with points colored by group (left) or by temperature (center) and the distance mapping fidelity chart (right).

the total forces acting on the point at equilibrium (here called temperature), the scatterplot shows if the point configuration has been trapped in a local minima. This is one of the scatterplot's own diagnostic tools; another one is the mapping fidelity chart that shows how multidimensional distances are mapped to 2D distances: the more points lay close to the diagonal, the better the mapping is.

Other diagnostic charts Various charts have been developed to support visual diagnostics: 1) distances distribution charts, to grasp the overall structure of the multidimensional space; 2) fingerprint display, to validate the fingerprint computation; 3) energy distribution inside groups, to check that structures with similar energies have been grouped together.

Visual structures comparison The definitive test for the validity of the classification algorithm is the visual comparison and superposition of structures assigned to the same group. The user can select structures in two ways: manually, from the groups' content listings and graphically, circling a group of points in the scatterplot. Both structures could be replicated and color coded. One of them can be translated and rotated to manually superimpose it over the other one.

4.4 New Visual Analysis Tools

We added to the above list of interactive diagnostic tools few visualizations that are not strictly needed for validation of the classification schema, but provided new analysis methods of the USPEX results. We added them because existing algorithms made their implementation almost effortless and because in the visual design phase we discovered their usefulness. As we see in sect. 5.1, these analysis and visualization tools provided unexpected insight to the researchers. The first of them is a measure of the *degree of order* of the structures. It is computed from the simulated diffraction spectra and it is defined as:

$$DegreeOfOrder = \frac{1}{R_0}\int F(x)^2\,dx \qquad (5)$$

Where R_0 is the distance at which the fingerprint first crosses zero (see fig. 4). Order seems to increase and saturate or remain almost constant during an USPEX run, but exhibits increasing number of isolated high order peaks at the end of the run (fig. 12). Visualizing

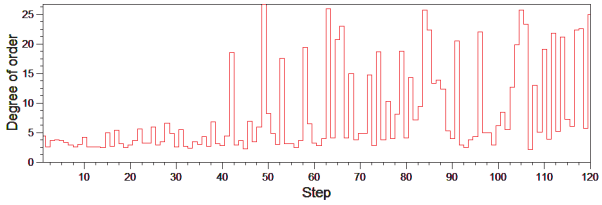

Figure 12: Order evolution in the GaAs (8 atoms/cell) dataset.

energy vs. order and energy differences vs. distances visually revealed unexpected correlations also (see sect. 5.1).

5 Visual Analytics Outcomes

A typical *ab-initio* evolutionary run samples ≈ 1000 structures (which takes several thousand CPU hours on today's clusters). Manual analysis of such a dataset, aimed at finding a handful of the most promising distinct structures, requires between 2 to 20 hours of work. With the CrystalFp end-user application, this can be achieved within ≈ 10 minutes.

To illustrate this, let us look at a rather pathological case, hydrogen under pressure. The pathology is easy to recognize using the CrystalFp visual analysis tools: the number of distinct enthalpy minima is rather small (so that even such a poor global optimization strategy as random sampling can find the global minimum [24]), but their enthalpies are extremely close. In such cases, one cannot use enthalpy as the sole grouping criterion and there will be a large number of identical structures found in evolutionary runs. Here we analyzed a dataset obtained with USPEX for hydrogen at 600 GPa, with 16 atoms in the (super)cell. Our dataset contains 1274 structures, 794 of which had enthalpies within 0.5 eV of the lowest value found. The large number of energetically reasonable structures makes this dataset difficult to analyze manually. Analyzing these structures with CrystalFp, we recognized an unusually high degeneracy: these 794 structures could be grouped into only 4 unique structures. This analysis took only ≈ 5 minutes, but doing the same work manually would take many hours. Among the four structures one belongs to the Cs-IV structure type (fig. 13a), two are closely related to it, and one belongs to the alpha-Ga type (fig. 13b). The

Cs-IV structure is the ground state, while the alpha-Ga type phase has a 20-30 meV/atom higher enthalpy. Both are atomic phases (i.e. non-molecular), and we confirm the conclusion of Pickard and Needs [24] that hydrogen adopts non-molecular structures at pressures above 500 GPa. The case of compressed hydrogen is rather rare in the extent of degeneracies, but even in more normal cases the use of the analysis methods presented here makes data analysis much easier compared to manual analysis.

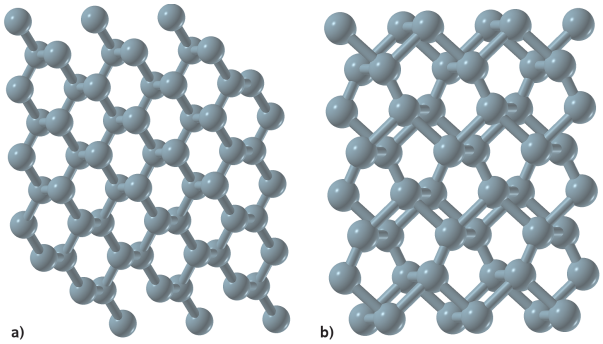

Figure 13: Low-enthalpy structures of hydrogen at 600 GPa found with USPEX: (a) Cs-IV structure, (b) alpha-Ga structure.

Various other examples of USPEX results with the role played by CrystalFp in their analysis are collected in the work of Oganov et al. [21]. But more interesting are the unexpected outcomes made possible by the CrystalFp visual approach. These outcomes where not searched for, but the availability of the CrystalFp visualizations helped pose the right scientific questions. Here below is collected a quick survey of these outcomes, instead their analysis will appear elsewhere.

5.1 Unexpected Outcomes

Genetic algorithm structure cancer phenomena Standard evolutionary algorithms, when allowed to run indefinitely, tend to converge to one solution, which tends to create its own replicas. Sometimes this solution may be suboptimal, and such premature convergence precludes efficient exploration of other possible solutions. CrystalFp visual diagnostics can visualize this phenomenon (see the large blue wedge in fig. 10). A procedure based on fingerprinting has be incorporated in the USPEX code, and proved to be very effective in precluding such "cancer growth" phenomena.

Energy vs. Distance correlation One of the basic assumptions made in the construction of the structure prediction method USPEX is that the energy landscape has an overall shape, where low-energy structures are clustered together. Energy-distance correlation enables checking this assumption for real systems. In most cases this assumption is confirmed, but we have found cases where there are several "clusters" of low-energy minima separated by large distances (see fig. 14). In such cases evolutionary algorithms are less efficient. Again, it is possible to maximize their efficiency using fingerprinting inside the evolutionary algorithm.

Random structures distance distribution Looking at statistics of distances between structures in a random set of structures, we discovered a striking Gaussian-type shape of distributions, with a clear peak (fig. 5). In the multidimensional space, fingerprint vectors describing crystal structures form a diffuse spherical shell (lengths of the vectors are similar within an order of magnitude, but directions differ) and the distance distribution corresponding to this geometry is similar to a Gaussian. In some cases we find more than one peak in the distance distribution, and this is a signal of complex chemistry involving different coordination numbers. Interestingly, as the number of atoms increases, fingerprints of randomly

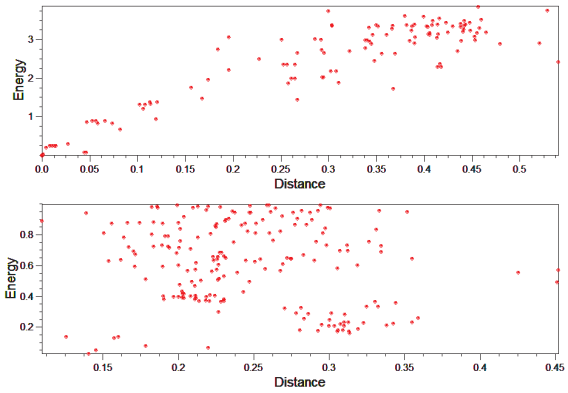

Figure 14: GaAs structures exhibits clear energy–distance correlation (top). Instead MgNH shows a more complex landscape (bottom).

generated structures (and structures themselves, and their energies) become increasingly close to random (where the fingerprint is -1 at small distances and 0 otherwise) and progressively indistinguishable. This can be viewed as a decrease of the radius of the hyperdimensional sphere, until it collapses to a point. This picture clarifies why Cartesian distances (measuring the distance between two points on a sphere) suffer more from the "curse of dimensionality" than cosine distances (which measure angular differences) (see sect. 3.2).

Energy vs. Order correlation In many runs we see order increasing during the run, and it is correlated with energies: high order usually means low energies. This is natural, disordered structures are expected to have high energies. Energy-order correlations can be used to uncover cases of geometric frustration, where less ordered or more complex structures are made energetically favorable by competition of oppositely directed factors.

6 System Implementation

The CrystalFp library has been implemented in C++ with only one dependency on other software[1] to make its integration inside applications easier. Its API various areas –fingerprint calculation, distances, grouping and analysis– have been implemented as separate classes as orthogonal as possible to simplify the addition of other algorithms for testing. The library has been tested on Linux, Windows and Mac OSX, but potentially can run on any platform.

The end-user application, which supports the library visual design and validation, has been built inside the molecular visualization toolkit STM4 [29, 30] based on AVS/Express [10]. Each step in the analysis workflow has been implemented as an AVS/Express module. The complete application has been visually assembled using these modules together with the ones provided by STM4 and AVS/Express itself.

This choice provided us a threefold benefit: first, we don't have to implement from scratch functionalities already available, like data file readers and graphical crystal structure rendering, but we can concentrate on our problem delegating the application control to the AVS/Express "data flow" architecture [10]. The second benefit, which perfectly matches with the project goal of user involvement in the visual design, lays in the immediate support of rapid prototyping provided by the component architecture and visual programming paradigm of AVS/Express. Last, the packaging of the CrystalFp functionalities as separate components make them reusable inside other crystallographic applications built using STM4.

[1] The exception is the use of ANN [19], a library for approximate nearest neighbor searching.

7 Discussion

The main contribution of this work to the crystallographic field is the reformulation of the crystal structures classification problem using multidimensional analysis and visualization methods. This reformulation provided flexibility for the exploration of new solutions and a concrete use case for others to consider visual analytics methods in the crystallographic and chemistry fields.

The visual exploration and diagnosis of key algorithm methods and parameters has facilitated the involvement of the domain experts in the classifier design. The rapid testing and modification of the algorithms implementation has been made possible by coupling visual design to the fast prototyping capabilities offered by the STM4 environment. This visual approach made a real difference in the project, also without introducing sophisticated visualizations, but simply relaying on the grounding ideas of visual analytics. Moreover this approach is not constrained to the crystallographic field, but can be reused also in other contexts.

The algorithms selected for the various library classification phases have been collected in sections 3.1, 3.2 and 3.3 together with an informal evaluation in section 5. Fortunately nearly-ground-truth datasets were abundantly present and used for this informal testing and validation of the algorithms.

There are still two open points that we plan to address in the future. The distance concentration [11] phenomena overshadow every attempt to define a better fingerprint or a better distance measure. We started experimenting with a weighting of distances that takes into account this phenomenon that has been introduced by the data-driven high-dimensional scaling (DD-HDS) [17, sect. IV.A] visualization method. The second open point concerns too many free parameters taken by the algorithm. Some of these parameters have been fixed based on crystallographic expertise, but the parameters related to clustering remains arbitrary.

8 Conclusion and Future Work

After the visual design and validation of the CrystalFp library, we started its integration inside the USPEX evolutionary crystal predictor algorithm. This integration will surely raise other interesting problems beside the open points already listed.

The work on the CrystalFp library and end-user application will continue, driven by the domain expert requests elicited by the visual exploration tools put in place during the design phase. For example we already have to experiment with a new fingerprinting method and a request for new functionalities in the end-user application to ease the superposition and comparison of crystal structures.

But so far the most important lesson learned in this project is the importance to have on board domain experts deeply interested in the project success and to have an end-user application, built using language and concepts from the crystallographic domain, they can use and experiment with.

Acknowledgements

Calculations were performed at the Joint Russian Supercomputer Centre (Russian Academy of Sciences, Moscow), ETH Zürich and Swiss National Supercomputing Centre (Manno) that the authors gratefully acknowledge.

References

[1] L. C. Andrews and H. J. Bernstein. Bravais lattice invariants. *Acta Crystallogr. Sect. A*, 51:413–416, May 1995.

[2] L. C. Andrews, H. J. Bernstein, and G. A. Pelletier. A perturbation stable cell comparison technique. *Acta Crystallogr. Sect. A*, 36:248–252, Mar. 1980.

[3] J. Apostolakis, D. W. M. Hofmann, and T. Lengauer. Derivation of a scoring function for crystal structure prediction. *Acta Crystallogr. Sect. A*, 57:442–450, July 2001.

[4] J. A. Chisholm and S. Motherwell. COMPACK: a program for identifying crystal structure similarity using distances. *J. Appl. Crystallogr.*, 38:228–231, Feb. 2005.

[5] R. de Gelder. Quantifying the similarity of crystal structures. *IUCr CompComm Newsletter*, 7:59–69, Nov. 2006.

[6] A. V. Dzyabchenko. Method of crystal-structure similarity searching. *Acta Crystallogr. Sect. B*, 50:414–425, Aug. 1994.

[7] B. P. V. Eijck and J. Kroon. Fast clustering of equivalent structures in crystal structure prediction. *J. Comput. Chem.*, 18:1036–1042, 1997.

[8] L. Ertoz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM '03)*, Philadelphia, 2003.

[9] L. Fábián and A. Kálmán. Volumetric measure of isostructurality. *Acta Crystallogr. Sect. B*, 55:1099–1108, Dec. 1999.

[10] J. Favre and M. Valle. AVS and AVS/Express. In C. Hansen and C. Johnson, editors, *The Visualization Handbook*, pages 655–672. Academic Press, Dec. 2004.

[11] D. François, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE T. Knowl. Data. En.*, 17(7):873–886, July 2007.

[12] T. M. J. Fruchterman and E. M. Reingold. Graph Drawing by Force-directed Placement. *Software Pract. Exper.*, 21(11):1129–1164, 1991.

[13] M. C. Hemmer, V. Steinhauer, and J. Gasteiger. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.*, 19:151–164, Feb. 1999.

[14] R. Hundt, J. C. Schön, and M. Jansen. CMPZ — an algorithm for the efficient comparison of periodic structures. *J. Appl. Crystallogr.*, 39:6–16, Feb. 2006.

[15] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[16] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169–11186, Oct. 1996.

[17] S. Lespinats, M. Verleysen, A. Giron, and B. Fertil. DD-HDS: A Method for Visualization and Exploration of High-Dimensional Data. *IEEE T. Neural Networ.*, 18:1265–1279, 2007.

[18] D. J. Livingstone. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.*, 40(2):195–209, 2000.

[19] D. M. Mount and S. Arya. ANN — A Library for Approximate Nearest Neighbor Searching. `http://www.cs.umd.edu/~mount/ANN`. Online, last checked: Jun. 2008.

[20] A. R. Oganov and C. W. Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.*, 124(24):244704, 2006.

[21] A. R. Oganov, Y. Ma, C. W. Glass, and M. Valle. Evolutionary crystal structure prediction: overview of the USPEX method and some of its applications. *Psi-k Newsletter*, 84:1–10, Dec. 2007.

[22] E. Parthé and L. M. Gelato. The standardization of inorganic crystal-structure data. *Acta Crystallogr. Sect. A*, 40:169–183, May 1984.

[23] D. G. Pettifor. The structures of binary compounds. I. Phenomenological structure maps. *J. Phys. C*, 19(3):285–313, 1986.

[24] C. J. Pickard and R. J. Needs. Structure of phase III of solid hydrogen. *Nat. Phys.*, 3:473–476, July 2007.

[25] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.*, 24(5):513–523, 1988.

[26] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[27] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Min. Knowl. Discov.*, 2(2):169–194, June 1998.

[28] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[29] M. Valle. STM4 — the molecular visualization toolkit. `http://www.cscs.ch/~mvalle/STM4`. Online, last checked: Jun. 2008.

[30] M. Valle. STM3: a chemistry visualization platform. *Z. Kristallogr.*, 220:585–588, 2005.

[31] E. L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L. M. C. Buydens. Method for the computational comparison of crystal structures. *Acta Crystallogr. Sect. B*, 61:29–36, Feb. 2005.