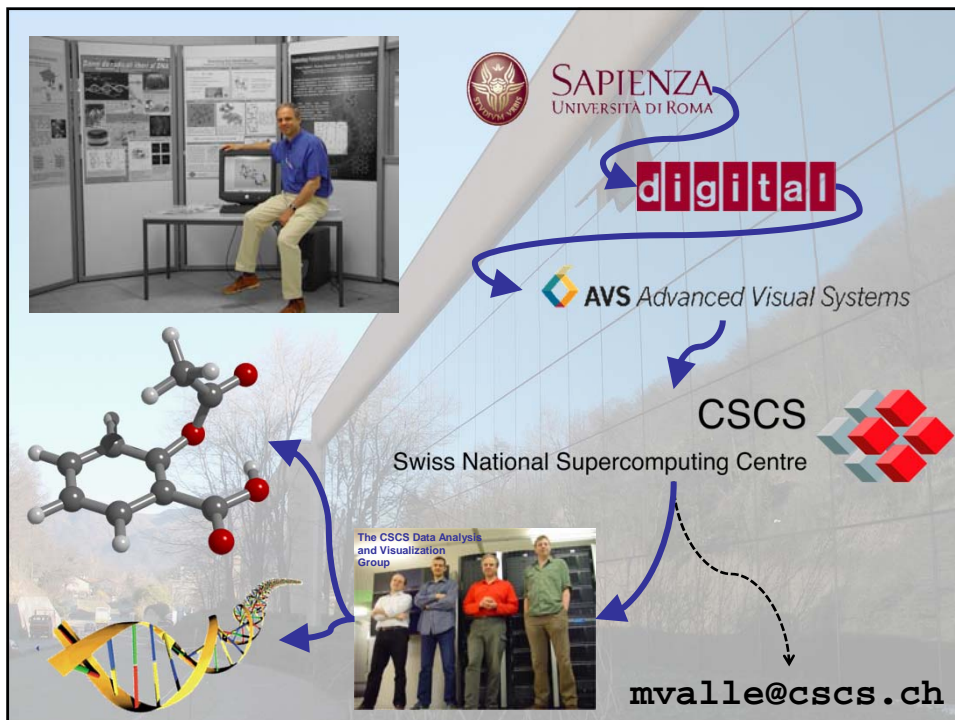
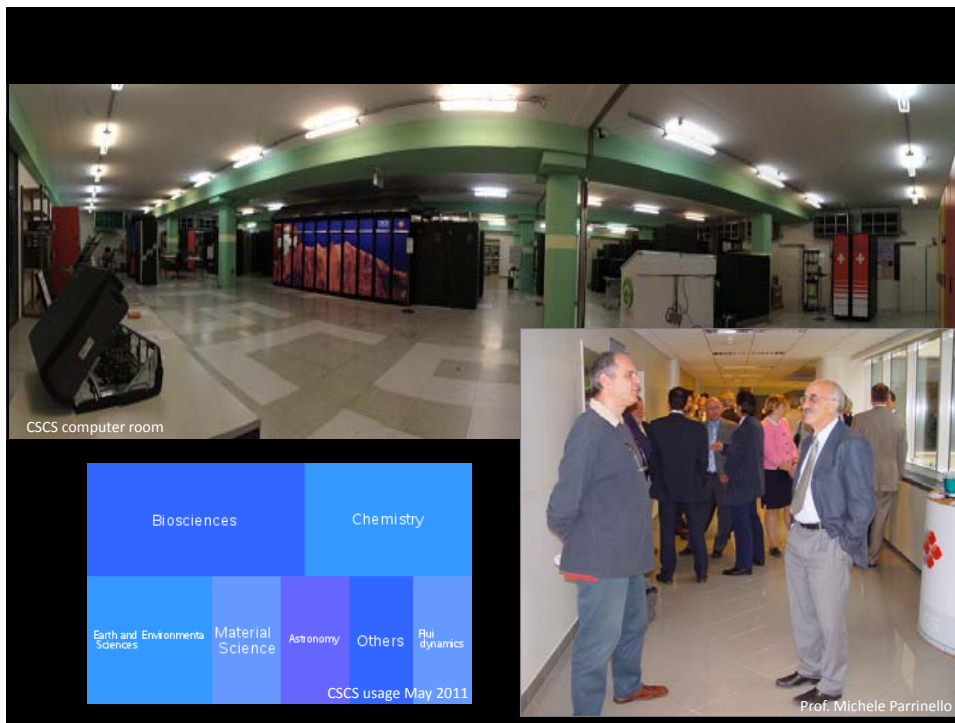


CSCS
Swiss National Supercomputing Centre

Crystal Fingerprinting and STM4 for USPEX output analysis

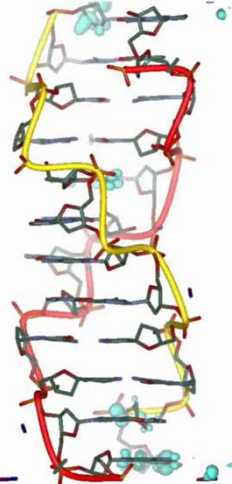
Mario Valle
USPEX 2011 Xi'an workshop – 04/08/2011



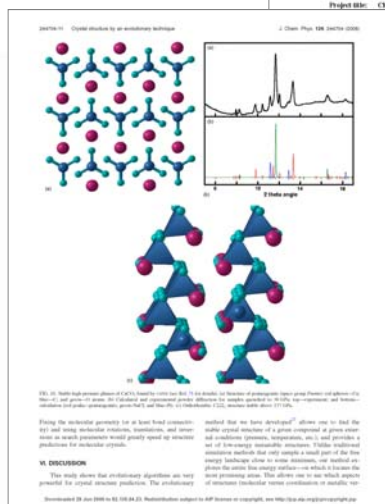


Started with postprocessing

Movies for conference talks



Francesco Gervasio – ETH Zürich



Project title: Charge Transfer and Oxidative Damage in DNA Fibers

Fields: Computational Science, Department of Chemistry and Applied Biosciences, Zürich-USD Campus, Lugano

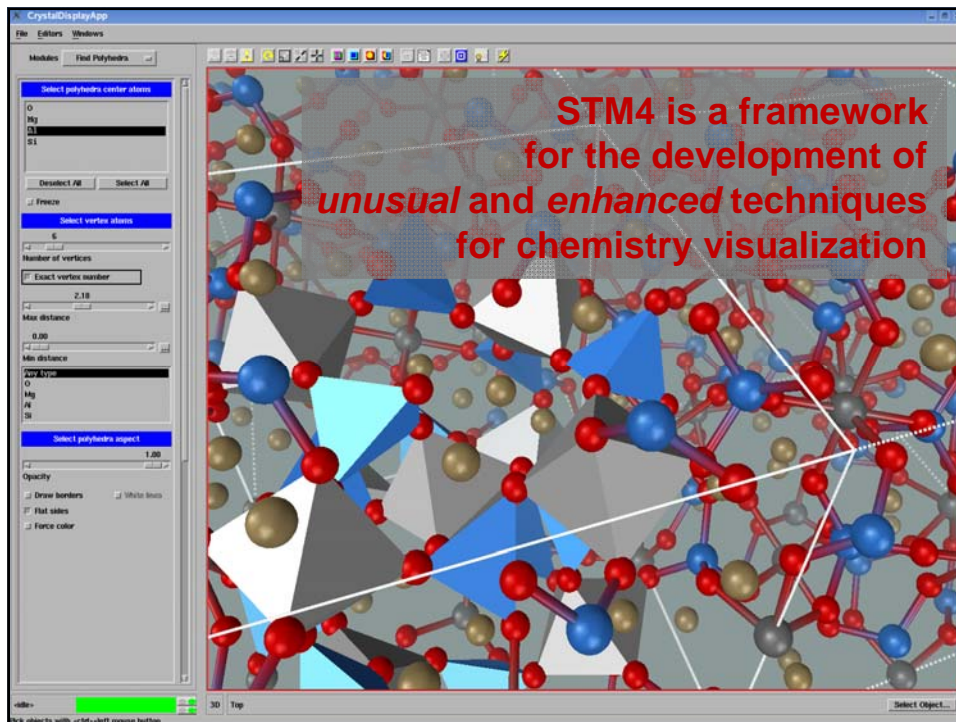
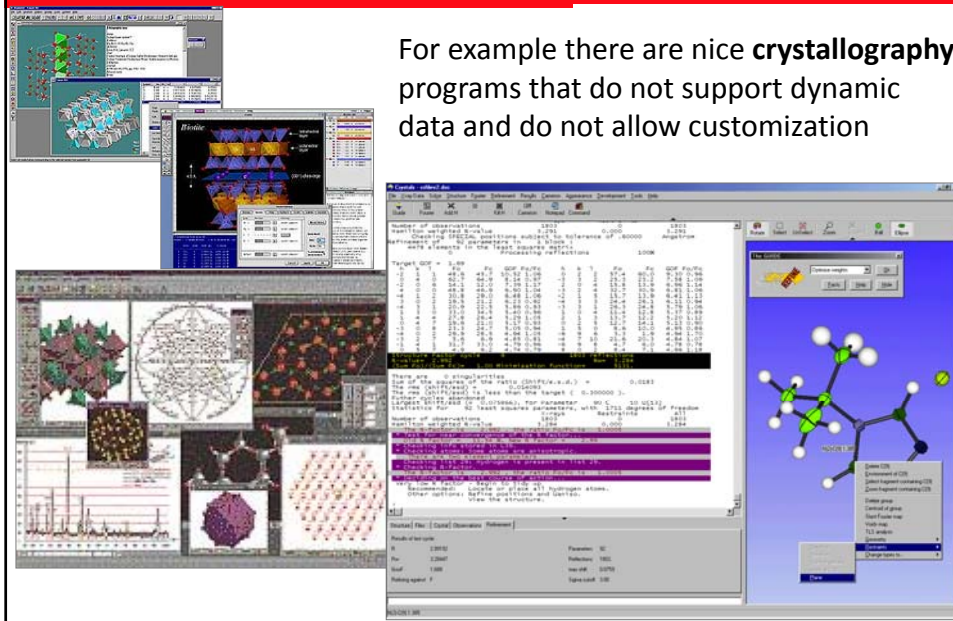
Atomic structure of DNA has recently attracted great interest due to its photochemistry. Indeed DNA has several properties that make it attractive. It is a stable polymer and can easily be handled and modified almost in one-dimensional character and regular stacking of its bases have made it easy to use as a molecular conductor. Unfortunately, experiments have very little theory could be of great help in understanding these phenomena. Important constraints of this complex molecule are taken into account (sugar backbone, solvating water and counter ions). Using state-of-the-art computational chemistry techniques we investigate the radical of the localization of the hole, possible mechanisms of charge transfer and the rate is not only important to a particular event but also charge transfer point of view so it represents the first step in DNA oxidation, a common cause of repeating mechanism eventually leads to fatal consequences such as

three-dimensional structure of the GC dimer end of the spin density localized in the radical cation state. Base molecules, counter-ions and solvent for clarity. The superphosphate backbone is represented as spheres and counter-ions as sticks and H₂O molecules as blue sticks. The molecule is C₂h, C_{2v} or C_{2h}. The ionization represented has a value of 10⁻⁷.

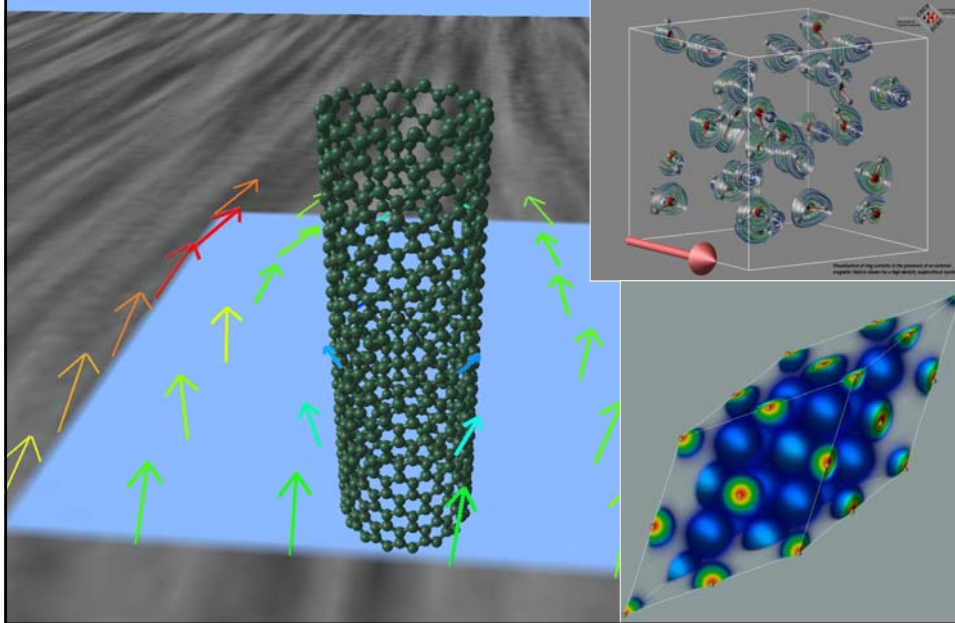
Nice images for publications

Help overcome tools inflexibility

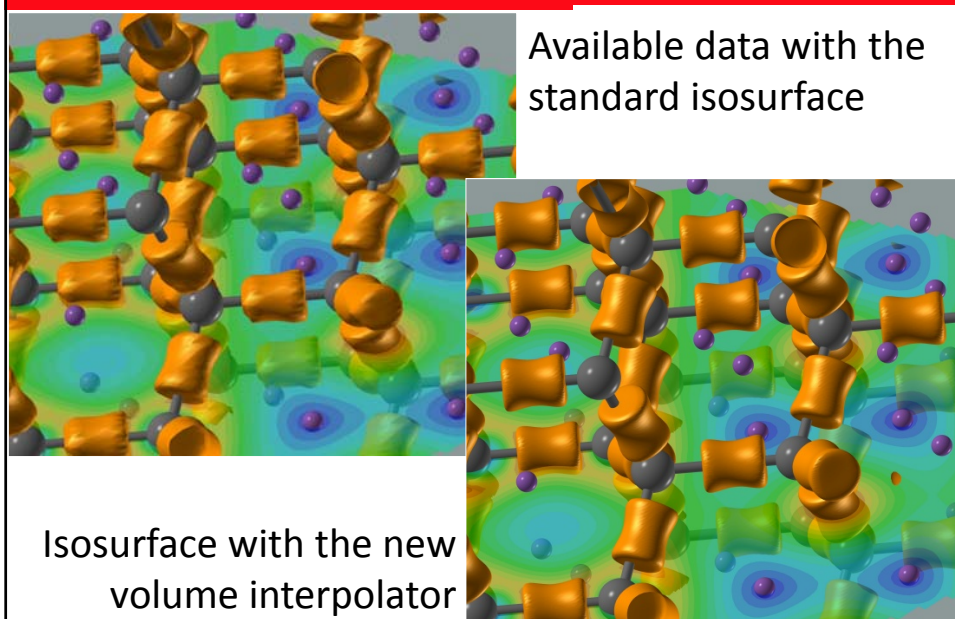
For example there are nice **crystallography** programs that do not support dynamic data and do not allow customization

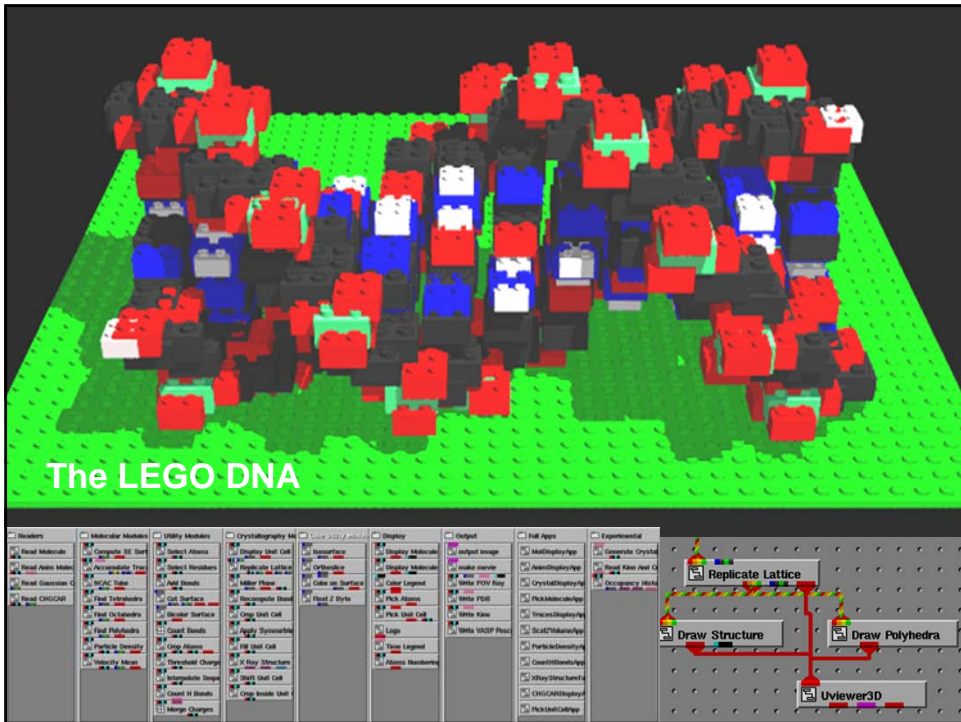
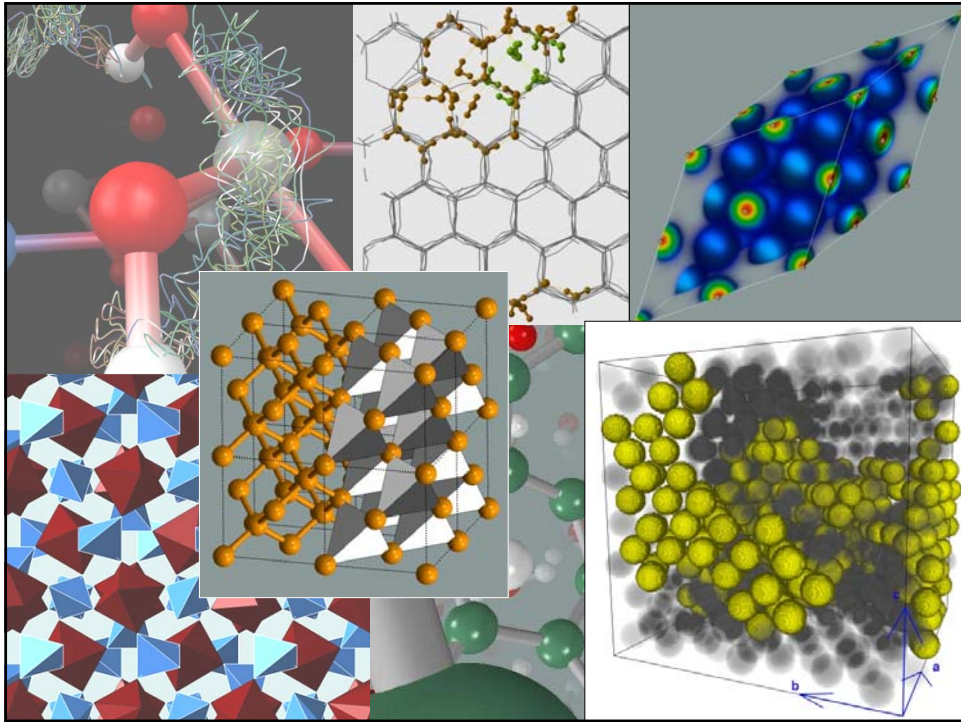


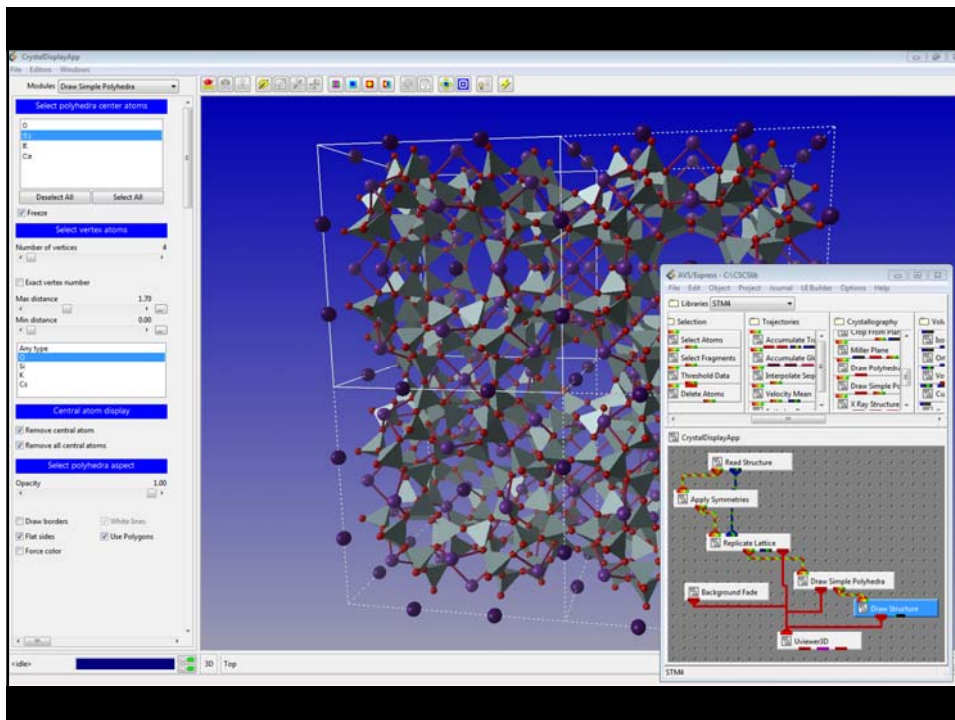
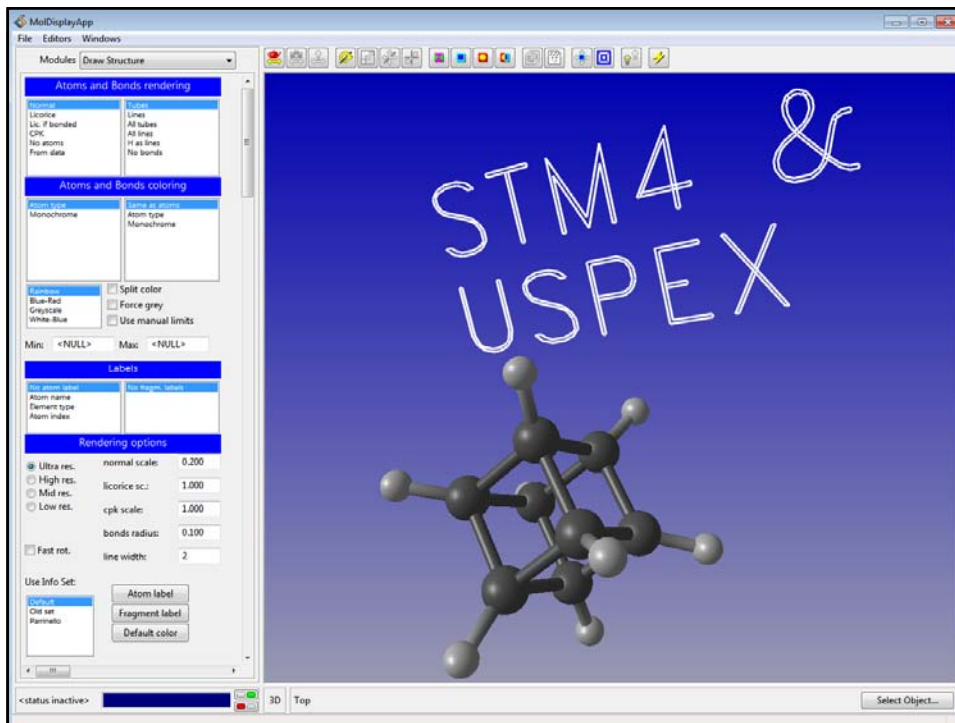
Offer broader set of techniques



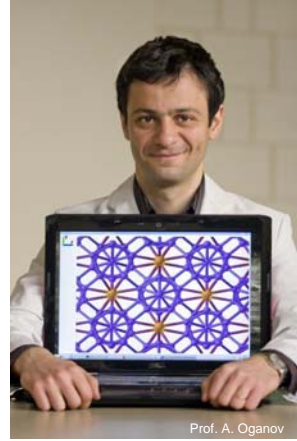
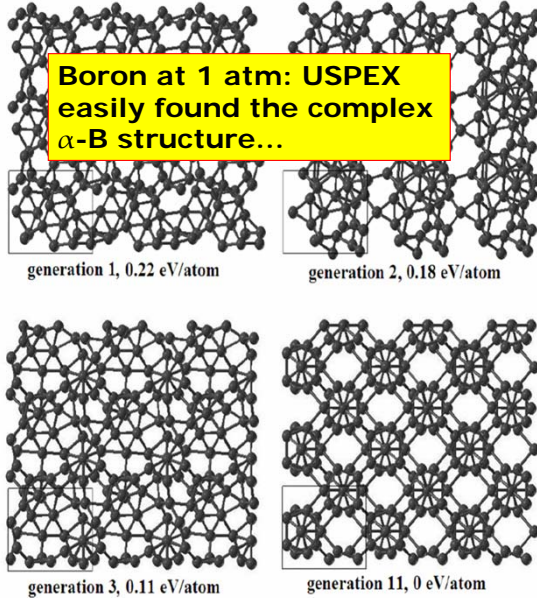
Provide enhanced techniques





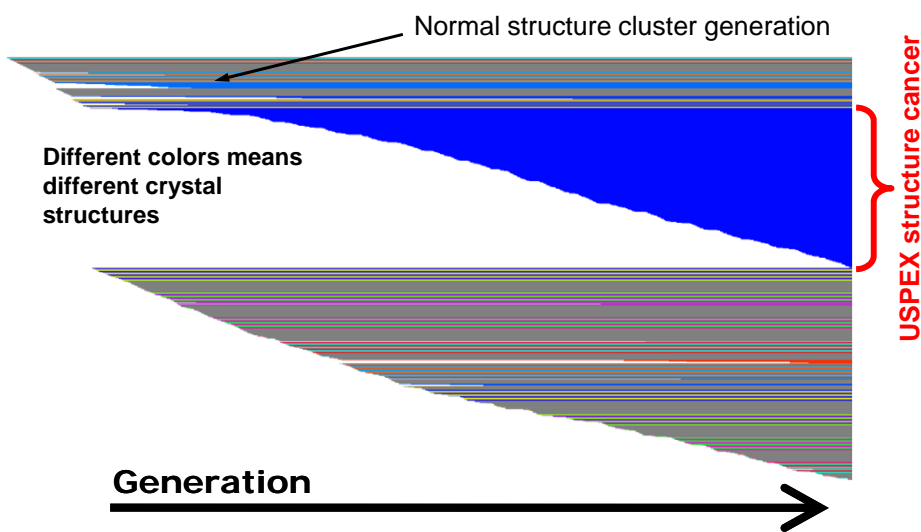


USPEX discovered new materials



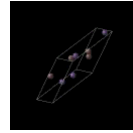
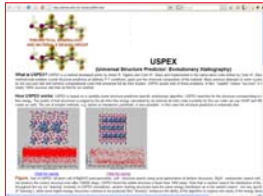
...and discovered also the superhard boron γ -B₂₈ phase (*Nature*)

USPEX structure cancer problem

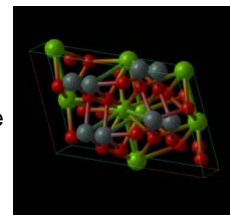


CrystalFp: the problem to solve

USPEX is a crystal structure predictor based on an evolutionary algorithm



Each run produces hundred of putative crystal structures...
...but many of them are equal



Project: develop a (semi)automatic way to extract unique structures from USPEX outputs



So an intensive manual labor is needed to prune duplicated structures

Proposed solution from High-Dim

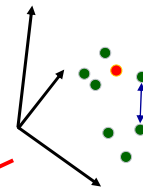


Compute unique coordinates

Space 100-3000 dimensional

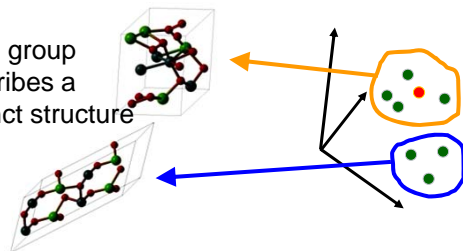


Define distance measure

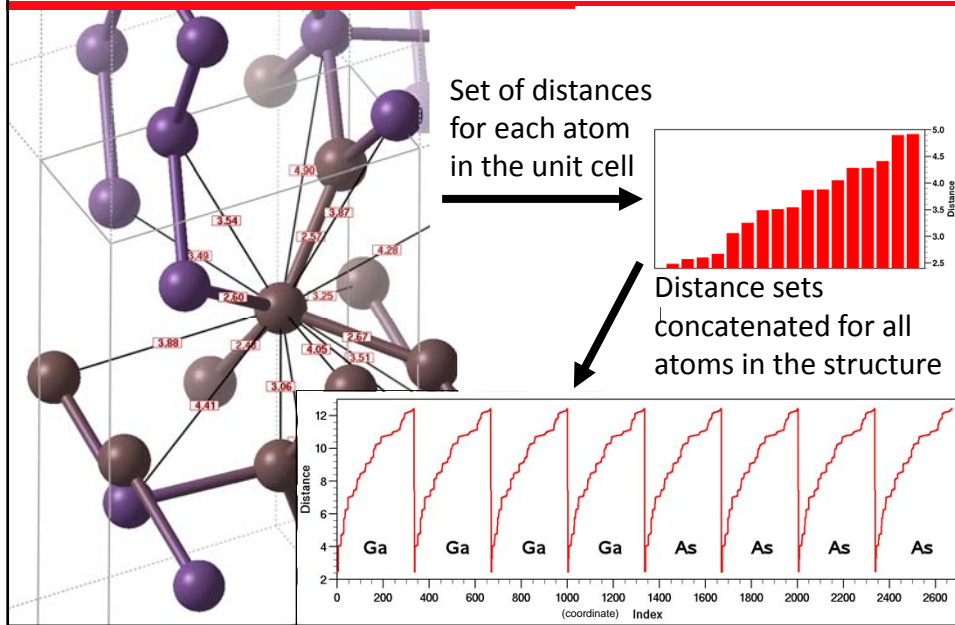


Add grouping criteria

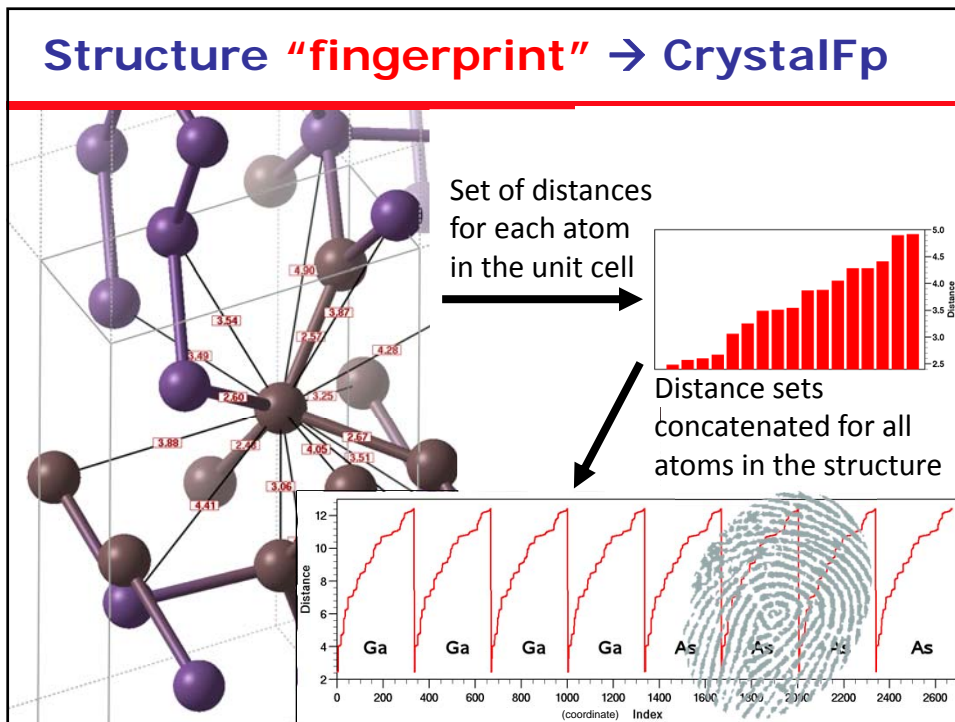
Each group describes a distinct structure

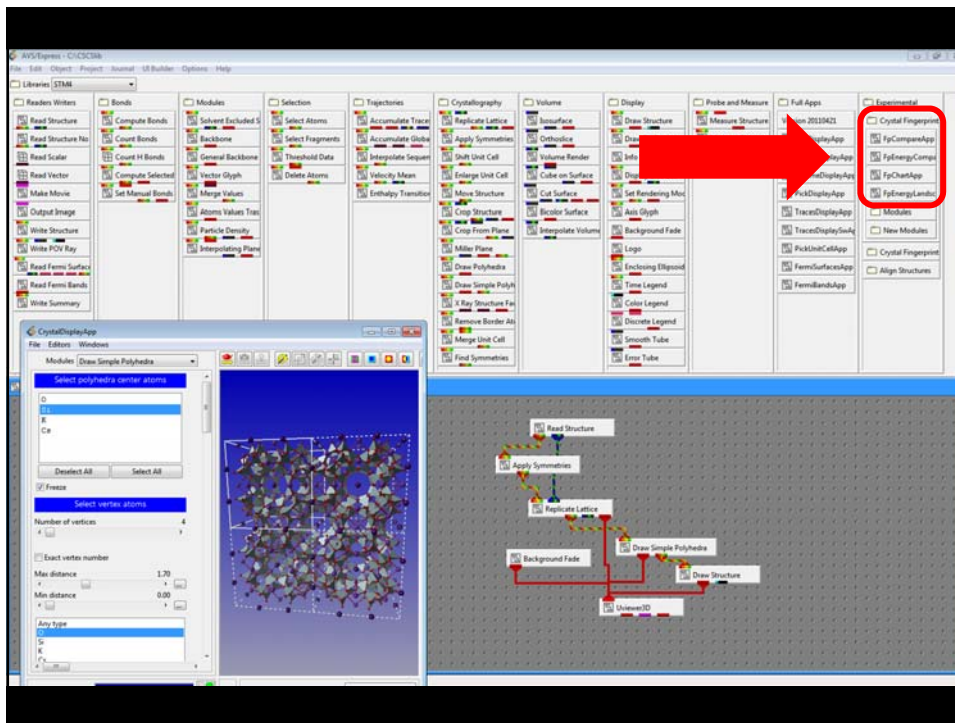


Structure "coordinates"

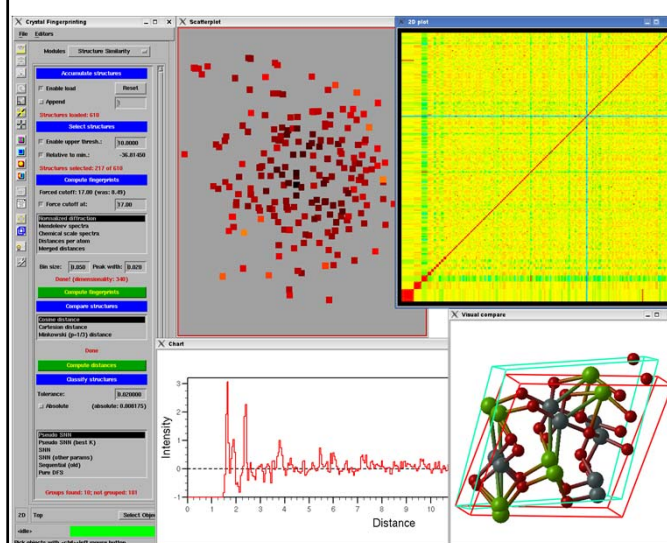


Structure "fingerprint" → CrystalFp





Visual design and validation



Built a tool to **explore** algorithm choices and parameters settings

This tool wraps the classifier library, called **CrystalFp**, and provides various **interactive visual diagnostics** to check classifier behavior

It is built inside **STM4**, the molecular visualization toolkit developed at CSCS

Workflow support

The application interface gives access to all CrystalFp algorithms and their parameters in a **clear process workflow** STM4 provided an environment that accelerated the implementation

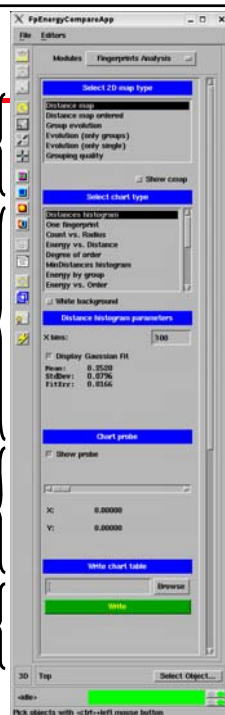
1. Load structures
2. Filter on energy
3. Compute fingerprints
4. Compute distances
5. Group structures



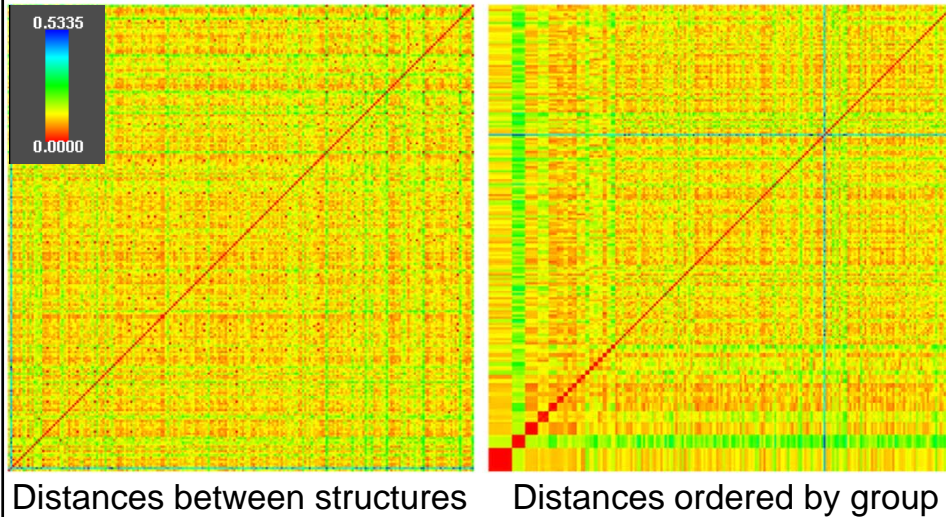
Visual diagnostics tools

Various visualization and analysis tools to check and **validate** CrystalFp algorithms behavior

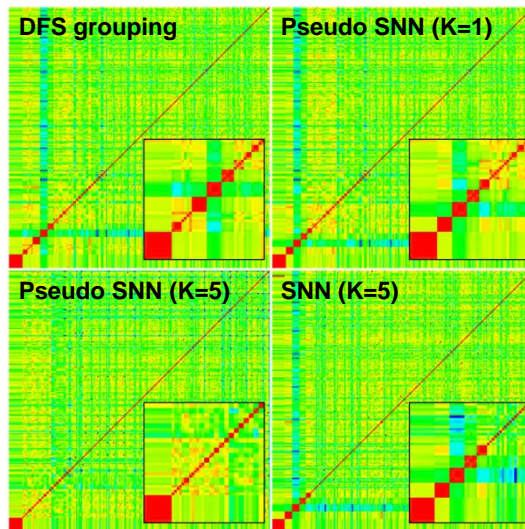
1. 2D maps
2. Charts
3. Picking for details
4. 2D data export



Visual diagnostics: distance matrix



Clustering visual diagnostic



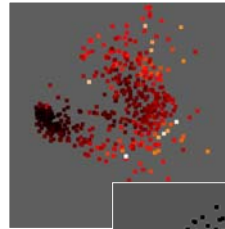
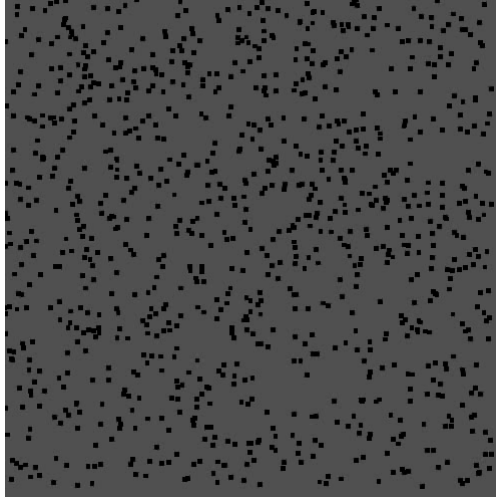
DFS: Deep first search of the neighbors nodes

Pseudo SNN: Maintain connection between nodes only if they share at least K neighbors

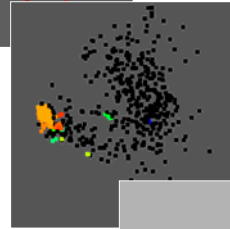
SNN: As above plus a DBSCAN pass

Visual diagnostics: scatterplot

The scatterplot tool in CrystalFp tries to map High-Dim space points to 2D preserving their relative distances

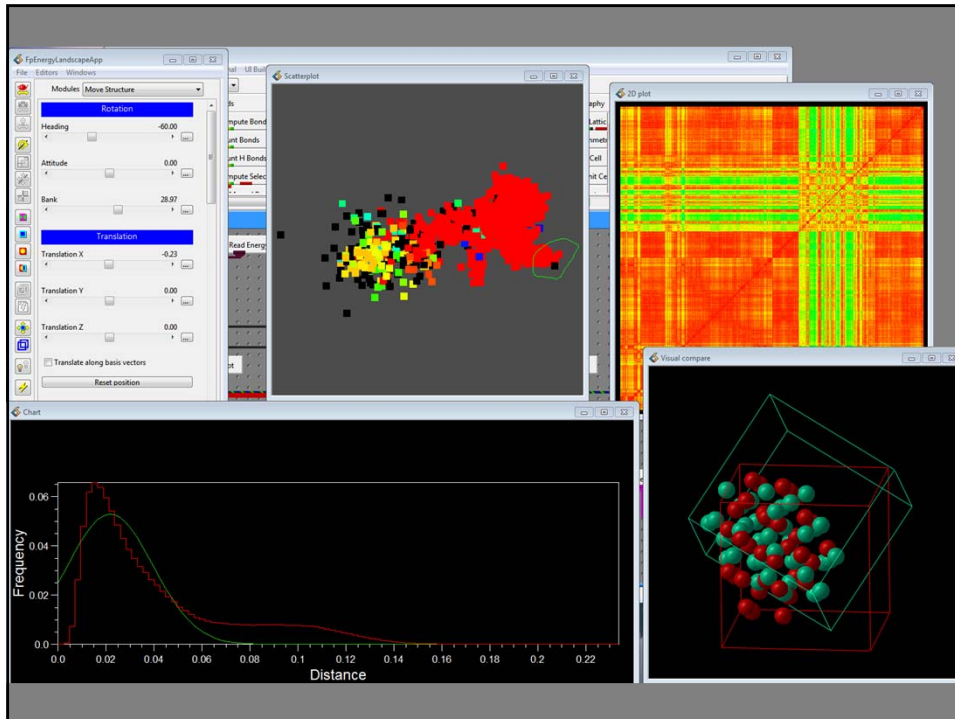
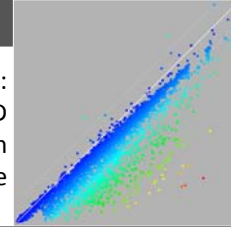


Colored by "stress" to detect local minima traps



Colored by group

Diagnostic chart: distances in 2D vs. distances in High-Dim space

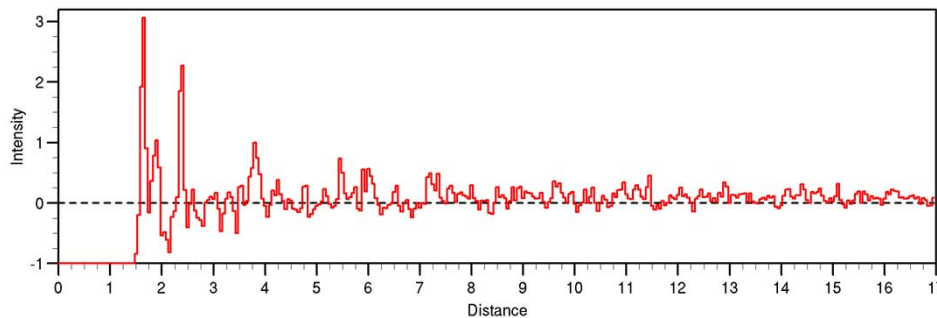


A pseudo-diffraction like method

$$\text{Fing}(R) = \sum_i Z_i \sum_j \delta(R - R_{ij}) \frac{Z_j}{4\pi R_{ij}^2 \frac{N_{uc}}{V_{uc}}}$$

This structure fingerprint is sampled on X to provide the coordinate values.

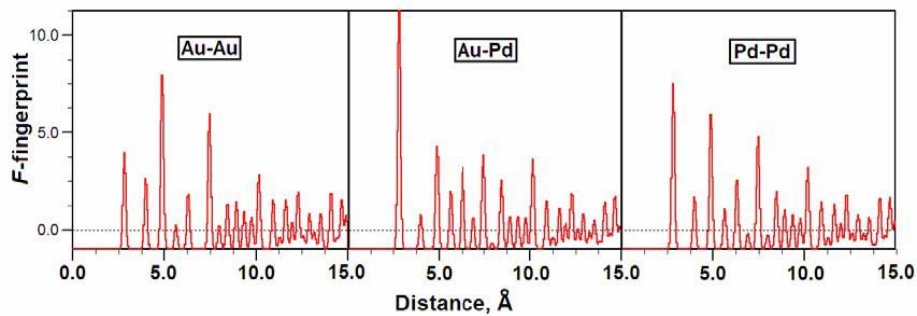
The fingerprint is cut at a user defined distance to provide 100-400 coordinate values



Final fingerprint (per atom type pair)

$$F_{AB}(R) = \sum_{A_i \text{ cell}} \sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 \frac{N_A N_B}{V}} - 1$$

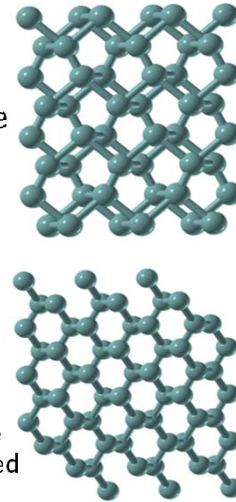
Compared to the previous fingerprinting method, this one is sensitive to the ordering of atoms in the structure and does not depend on the specific atomic species involved



USPEX problem solved: an example

Hydrogen at 600 GPa (16 atoms)

- The USPEX run produced **1274** structures
- From these the **794** within 0.5 eV from the lowest energy value found are selected
- Manual analysis to remove duplicated structures from this set: **~20h** of work
- Using the CrystalFp classifier: **~10min**
- At the end found only **4** unique structures:
 - One α -Ga type (top)
 - One Cs-IV (bottom), the ground state (i.e. the lower energy structure), and two closely related structures



LETTERS
Vol 458/12 March 2009 doi:10.1038/nature07786

Transparent dense sodium

Yanning Ma¹, Mikhail Eremin², Artem R. Ogilvie^{3,4}, Yu Xia⁵, Ivan Trejcek⁶, Sergey Medvedev⁷, Andriy O. Lyakhov⁸, Mario Valle⁹ & Vitalii Trukhanovskii⁹

Under pressure, metals exhibit increasingly shorter interatomic distances. Intuitively, this response is expected to be accompanied by an increase in the width of the valence and conduction bands and hence a more pronounced free-electron-like behaviour. But at the transition that can now be achieved experimentally, compression can be so substantial that core electrons overlap. This effect dramatically alters electronic properties from those typically associated with simple free-electron metals such as lithium (Li; refs 1, 2) and sodium (Na; refs 3, 4), leading to new structurally complex phases^{5,6} and superconductivity with a high critical temperature^{7,8}. But the most intriguing prediction—that the so-called simple metals Li (ref. 1) and Na (ref. 4) will transform under pressure into insulating states, owing to a pairing of alkali atoms—has yet to be experimentally confirmed. Here we report experimental observation of a pressure-induced transformation of Na into an optically transparent phase at 208 GPa (corresponding to ~3.6-fold compression). Experimental and computational data identify the new phase as a wide-bandgap diamond-like structure. We attribute the emergence of this dense insulating state not to atom pairing, but to the hybridization of valence electrons and their repulsion by core electrons into the lattice interstices. We expect that such insulating states may also form in other elements and compounds when compression is sufficiently strong that atomic cores start to overlap strongly.

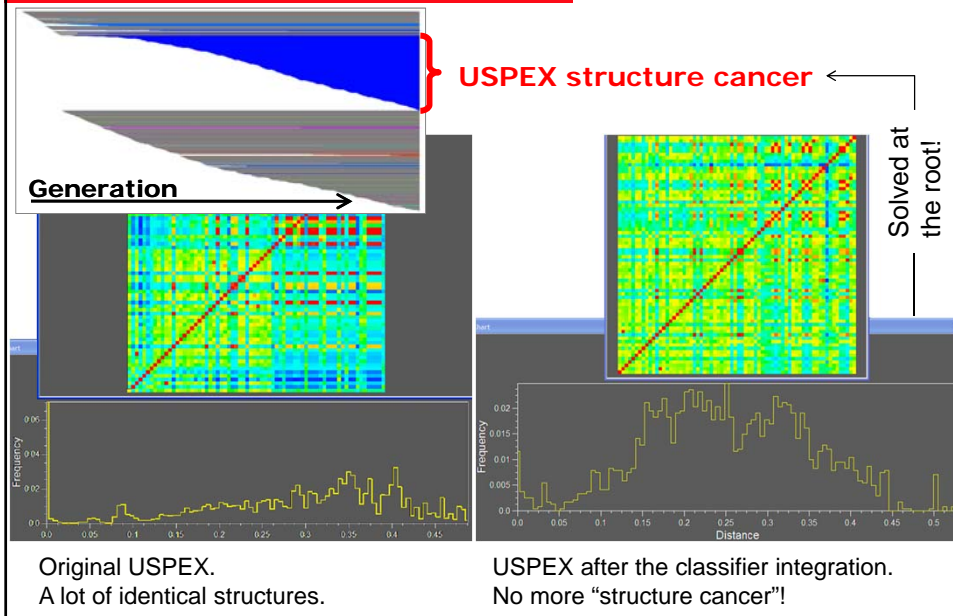
Sodium adopts the body-centred cubic (b.c.c.) structure at ambient conditions. Under pressure, it transforms to the face-centred cubic (f.c.c.) structure at 43 GPa (ref. 12) and to the c16 structure at 102 GPa (ref. 11). In recent experiments^{9,10}, further compression to 100 GPa yielded a number of phases in a narrow pressure-temperature range that the simulation of the existing core-valence interaction theory for the estimation of the energy curve (ref. 11) and theoretical calculations¹¹ suggest include body-centred tetragonal (b.c.t.) and c16 (CsCl) structures for pressures above 100 GPa. The c16 structure is especially interesting, because the calculations suggest a zero bandgap about 80 GPa owing to pairing of Na atoms¹¹. To explore the possible existence of stable yet lattice-compressed high-pressure structures and thus our own understanding of the pressure-induced metal-insulator transition in Na, we undertook an extensive experimental and theoretical study into insulating states in this archetypal metal at high pressures.

Our diamond- and cell experiments (see Methods for details) yielded X-ray diffraction data at 101 and 113 GPa corresponding to the known c.c.c. (a = 3.6 Å) and c16 phases (a = 3.6 Å, respectively), in agreement with available experimental data^{11,12}. The corresponding Raman spectra show no features at pressures below 130 GPa. But a pronounced Raman spectrum appears at higher pressures (Fig. 1a), indicating a major phase transformation that according to visual observation is associated with a gradual decrease in the reflectance of white light from the sample. The Raman spectra appearing around 130 GPa are in good accordance with the theoretical spectra calculated for the experimentally observed¹¹ c16 (Pnma) phase (Supplementary Fig. 1d). Above 150 GPa, the Raman spectra again show marked changes, including a strong decrease in intensity, that signify another phase transition (Fig. 1a). The X-ray diffraction pattern of this phase is consistent with the c14 structure¹¹.

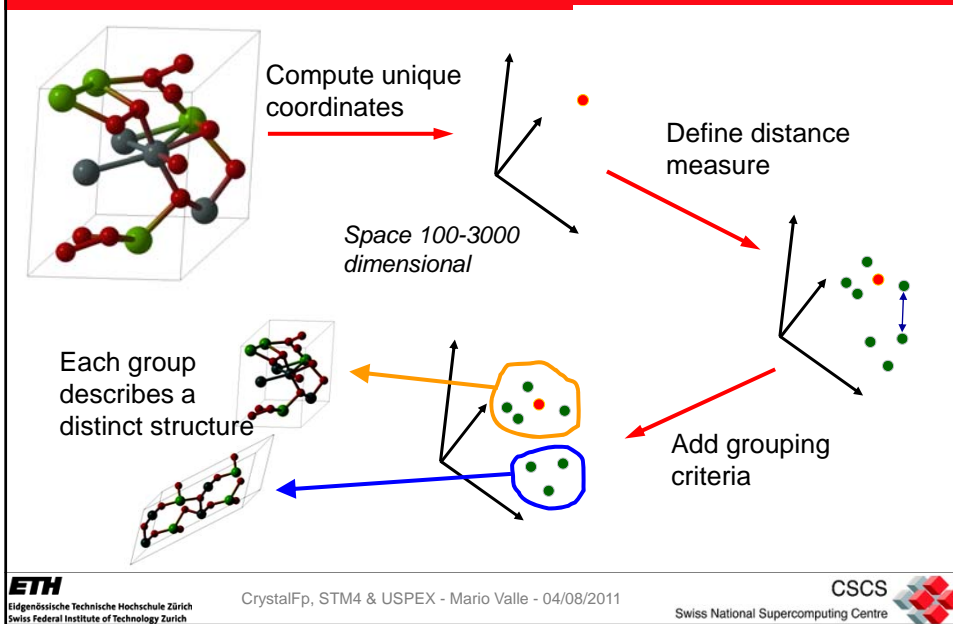
No features optically transparent at pressures of ~200 GPa (the onset of transparency occurred at 208 GPa in the run shown in Fig. 2a, and at 194 GPa in the run shown in Supplementary Fig. 1a). The edge in the absorption spectrum (Supplementary Fig. 1c) gives for transparent Na a bandgap of at least ~1.2 eV. The onset of transparency coincides with dramatic changes in the Raman spectra (Fig. 1a), particularly the appearance of a single intense line centred at ~160 cm⁻¹ (the following the pressure, the transparent phase persists at 162 GPa, at that point, the sample ceases to appear and exhibit

Figure 1 Raman spectra of sodium. a, Raman spectra obtained at increasing pressure. Spectra are shown for pressures of 101, 113, 130, 150, 162, 194, 208 and 220 GPa. The corresponding theoretical spectra are shown in black, blue and red lines, respectively. b, Pressure dependence of the Raman intensity for the transparent phase. The color scale corresponds to the spectra in a. Filled and open symbols correspond to raise pressure decrease and decrease pressure, respectively. Black lines are theoretical calculations for the c16 structure (thick lines) and b.c.c. (thin lines) for various polarizations (the red line is the fit, while the transparent Na b.c.c. Raman calculation was performed for the complex c16 phase. These curves are also shown here).

Classifier integration in USPEX



So, where is the problem?



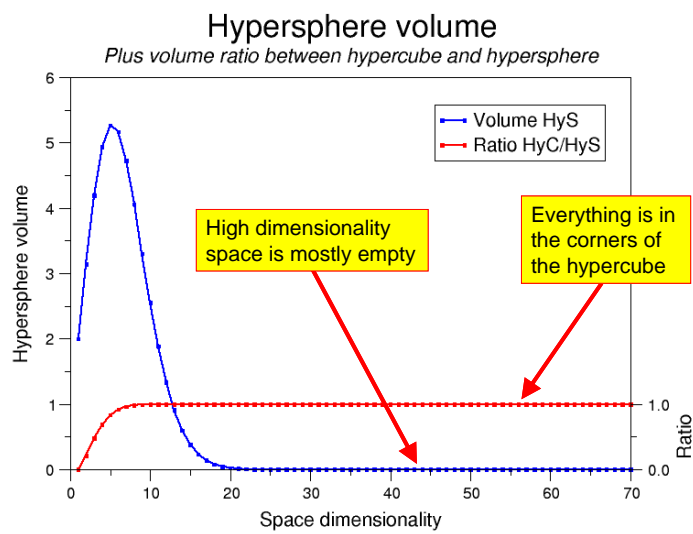
So, where is the problem?



Space 100-3000
dimensional

**Space 100-3000
dimensional**

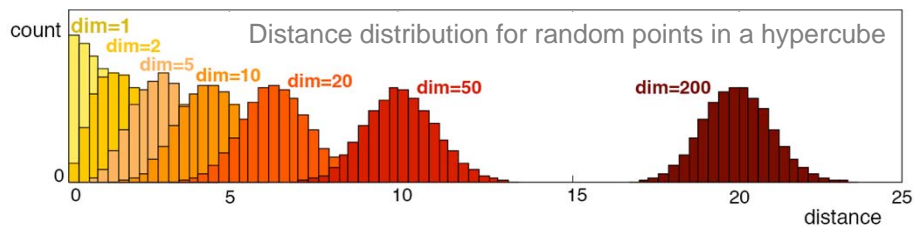
High-dimensionality is not intuitive



The curse of dimensionality

Roughly speaking, the higher the dimensionality,
the lower the power of recognizing similar objects

Because everything is at the same distance from
every other point...



But distances measures ...

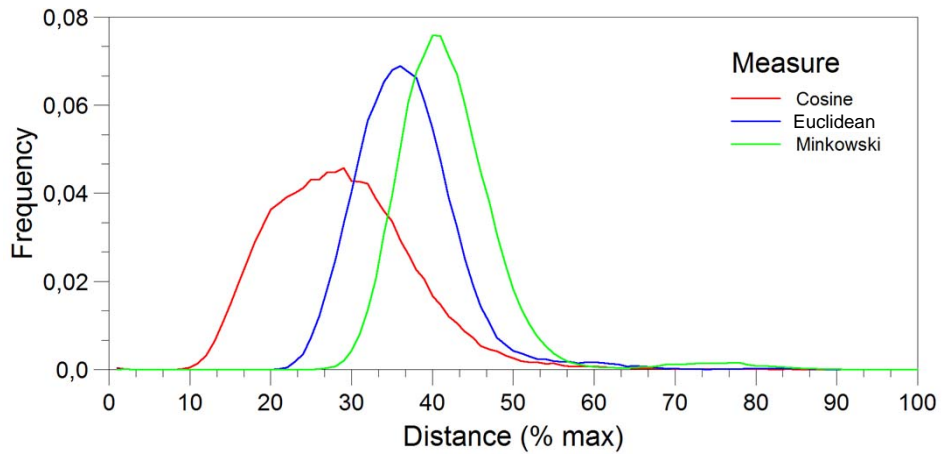
...could help contrast this curse of dimensionality

Euclidean distance:
$$D(a, b) = \sqrt{\sum_{i=1}^N (b_i - a_i)^2}$$

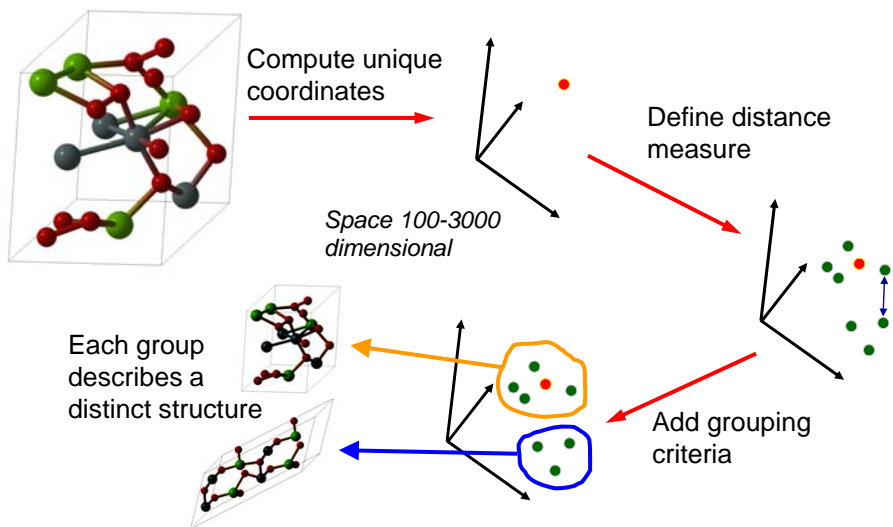
Minkowski distance:
$$D(a, b) = \sqrt[p]{\sum_{i=1}^N |b_i - a_i|^p}$$

Cosine distance:
$$D(a, b) = \frac{1}{2}(1 + \cos\theta)$$
$$= \frac{1}{2}\left(1 + \frac{A \cdot B}{\|A\| \|B\|}\right)$$

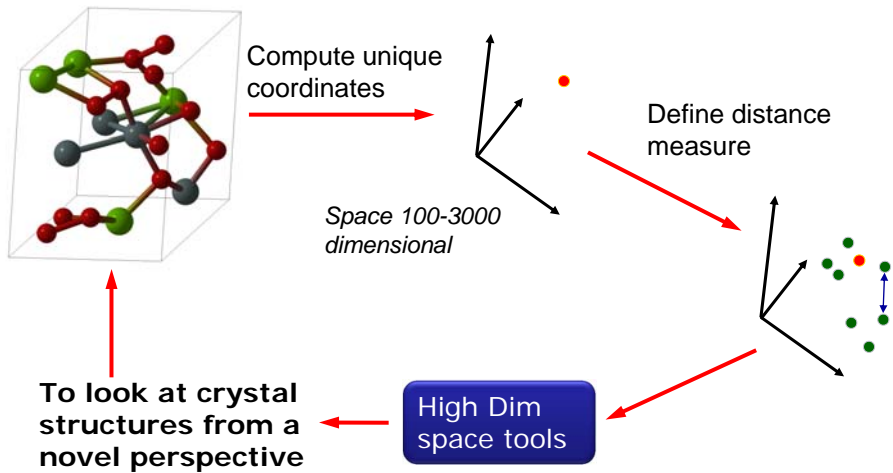
Tried various distance measures



From the problem solution ...

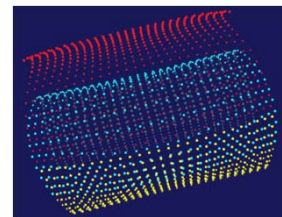
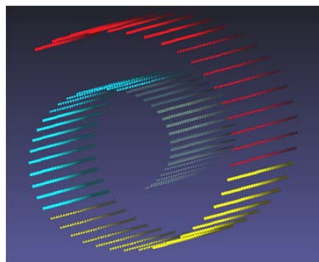


... to a new paradigm

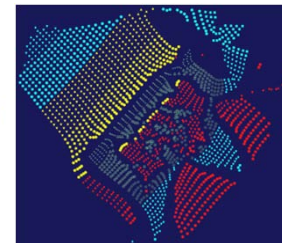


Unfold data to lower dimensions

Multidimensional scaling projects points from high dimensional space to a lower dimensional one **preserving distances between points** as faithfully as possible



Sammon mapping to 2D

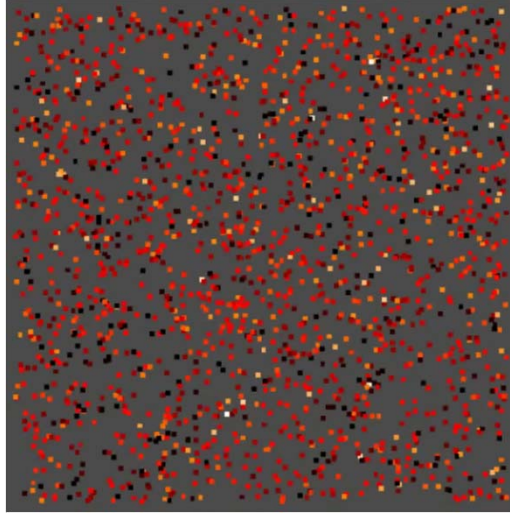


CCA mapping

One famous test dataset (right I said right!) contains points on a rolled sheet that forms a 3D shape called the "Swiss roll" (a superb example on the left)

CrystalFp multi dim. scaling

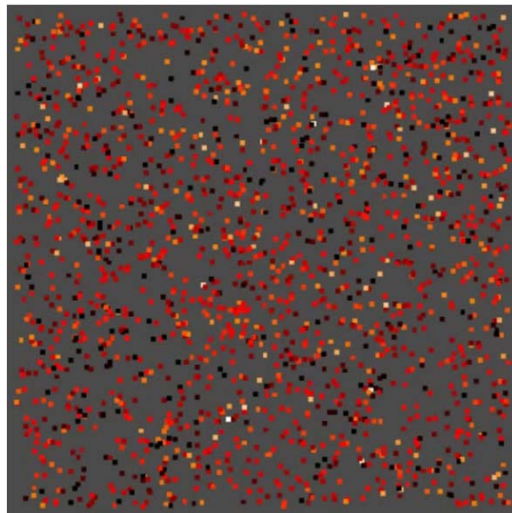
The scatterplot tool in CrystalFp implements a **Force Directed Placement** multidimensional scaling algorithm (here the points are colored by energy)



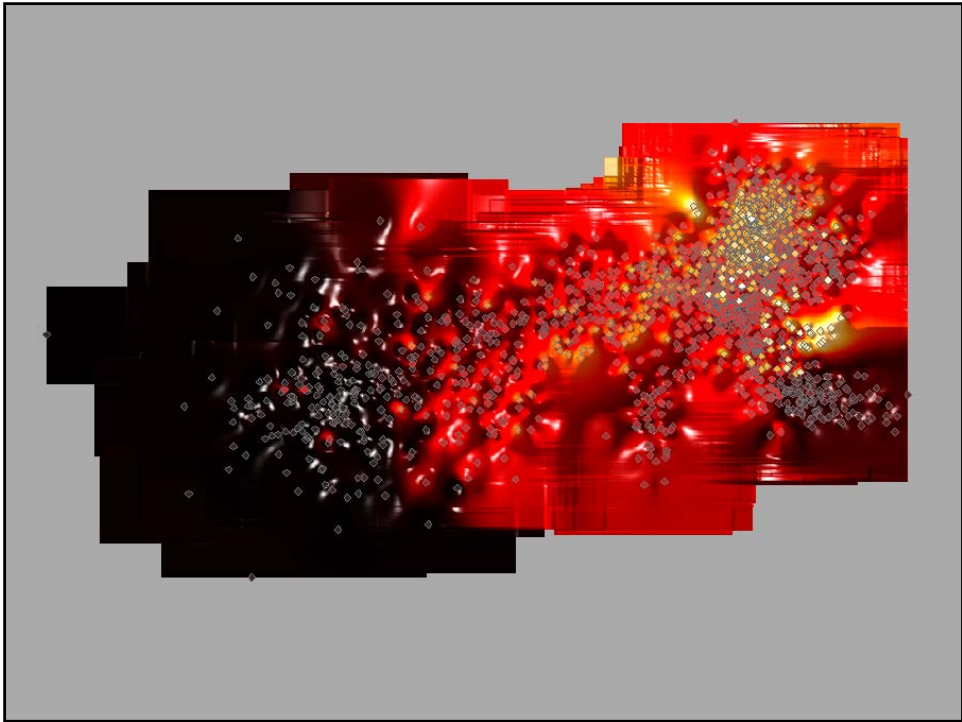
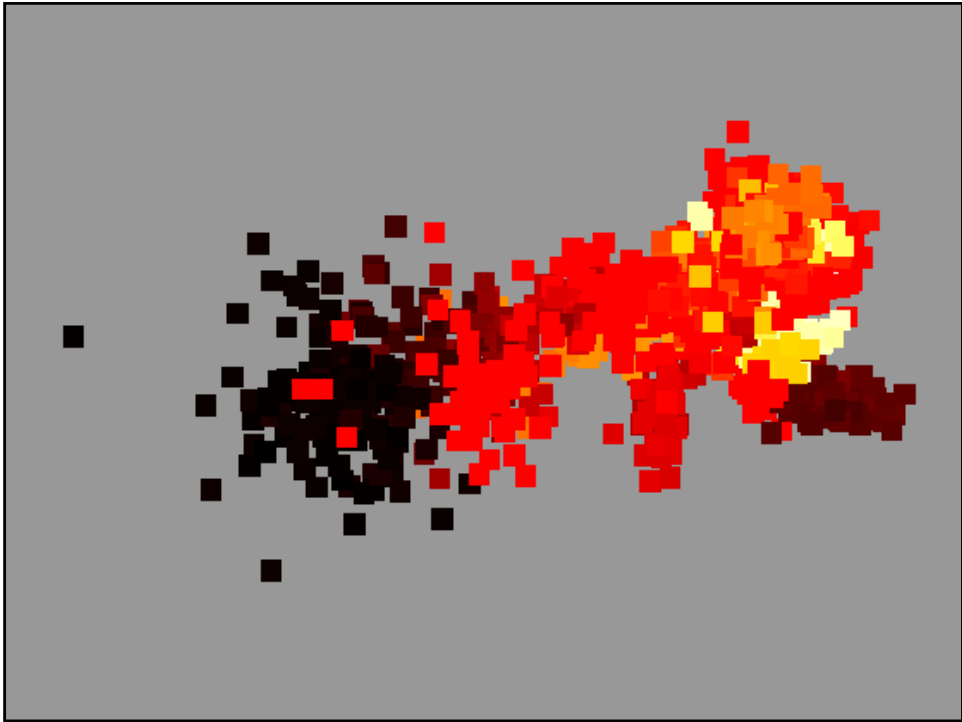
Nanocluster data from Dr. Gareth Tribello (USI)

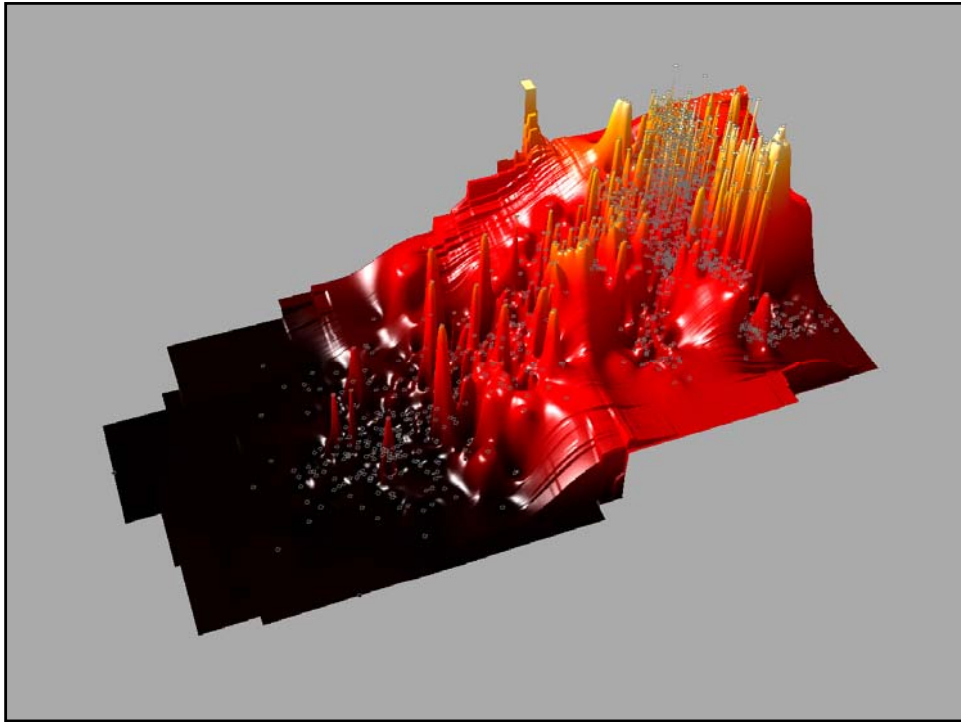
CrystalFp multi dim. scaling

The scatterplot tool in CrystalFp implements a **Force Directed Placement** multidimensional scaling algorithm (here the points are colored by energy)

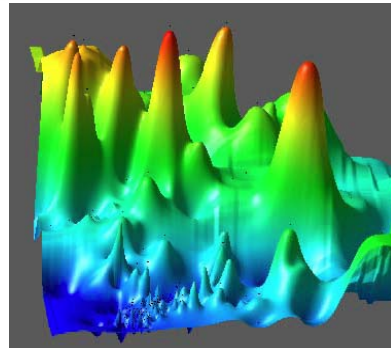
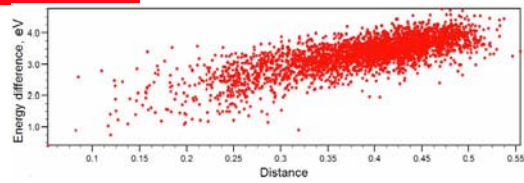
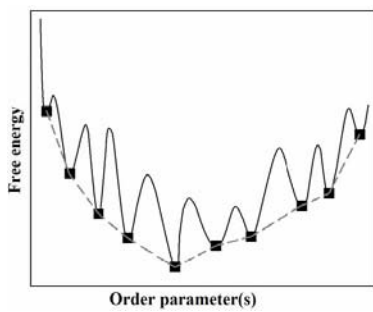


Nanocluster data from Dr. Gareth Tribello (USI)





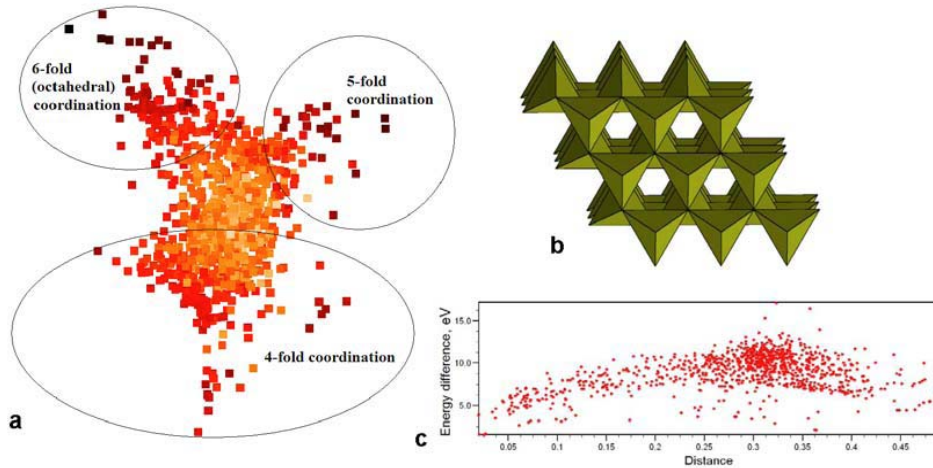
Study of energy landscapes



Energy landscape of Au_8Pd_4 system

A. R. Oganov and M. Valle,
How to quantify energy landscapes of solids,
The Journal of Chemical Physics, vol. 130,
 p. 104504, 2009.

More complex landscapes



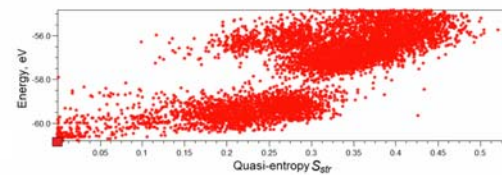
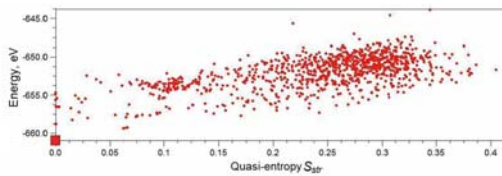
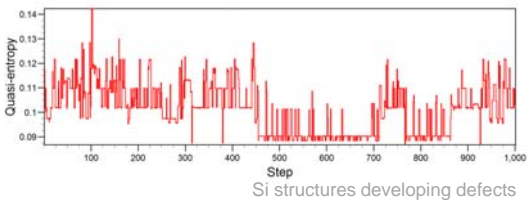
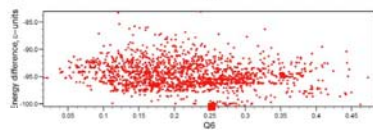
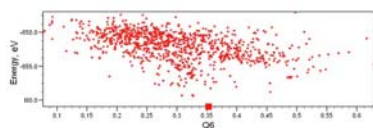
Energy landscape for MgO with 32 atoms/cell

New quantities: quasi-entropy

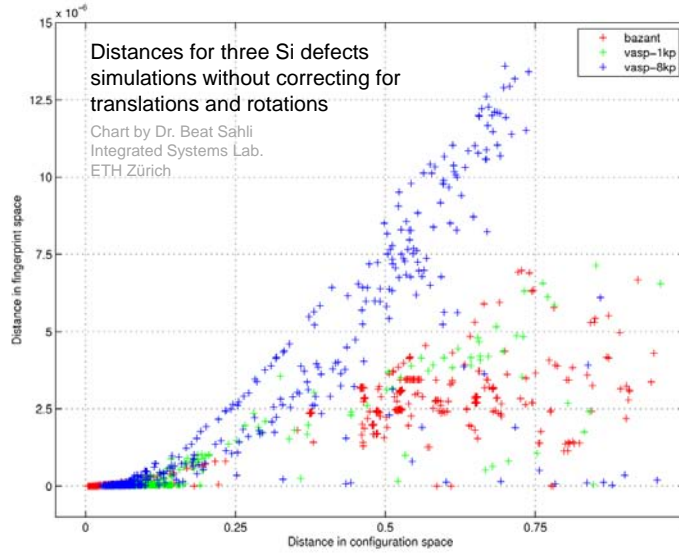
For each given structure, **quasi-entropy** is a measure of disorder and complexity of that structure.

$$S_{str} = - \sum_A \frac{N_A}{N_{cell}} \langle \ln(1 - D_{A,A_j}) \rangle$$

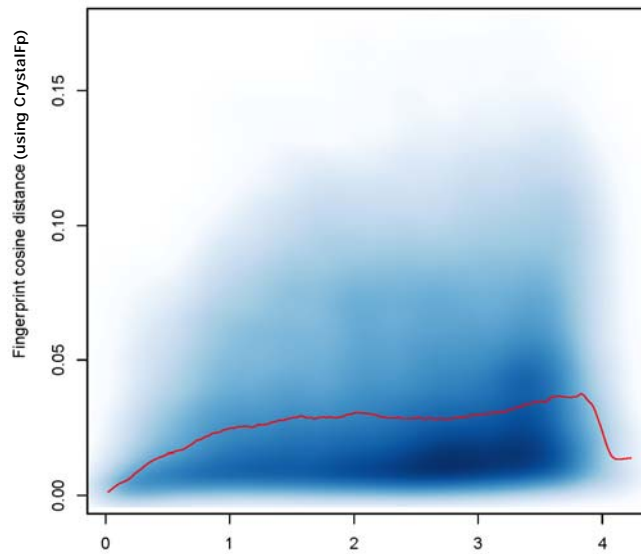
S_{str} is better correlated to energy than Steinhardt's Q_6



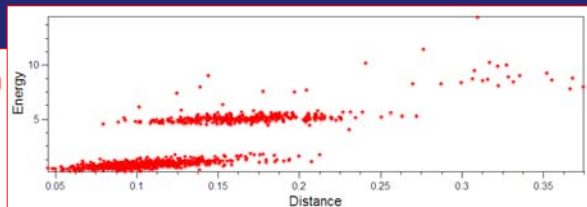
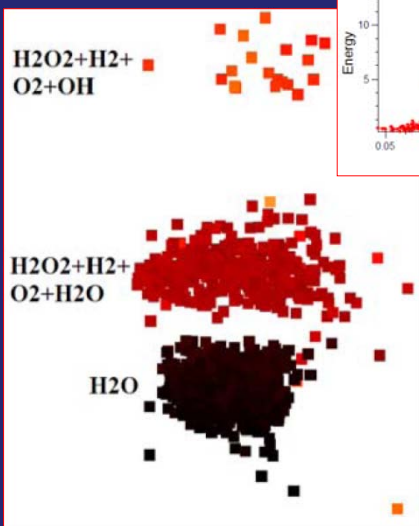
Faithful representation?



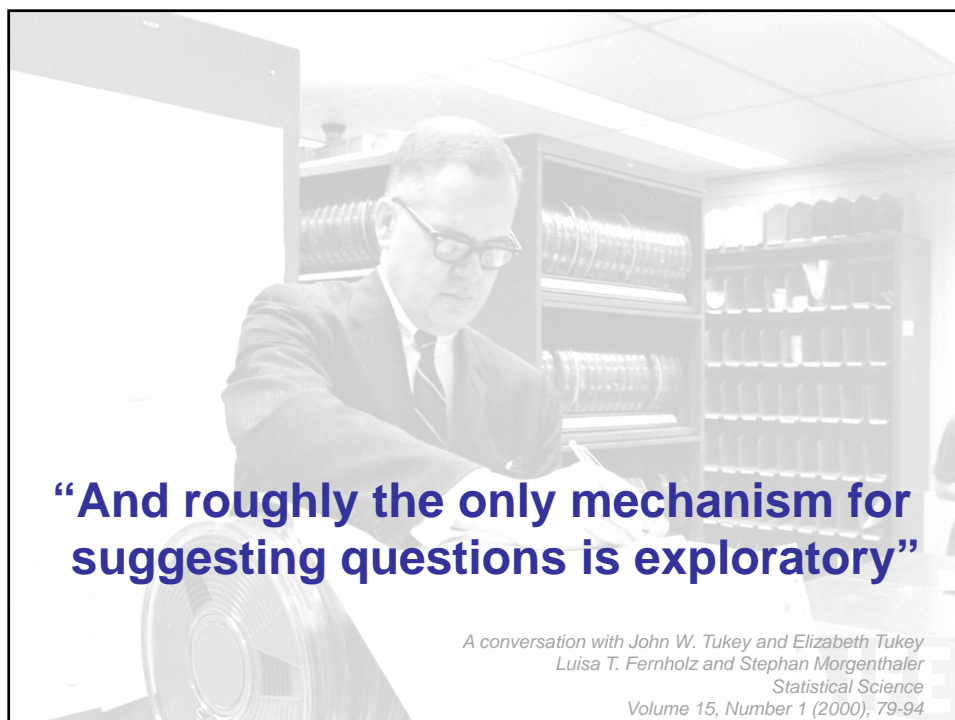
On average (on an interval), yes



(Totally) unexpected correlations



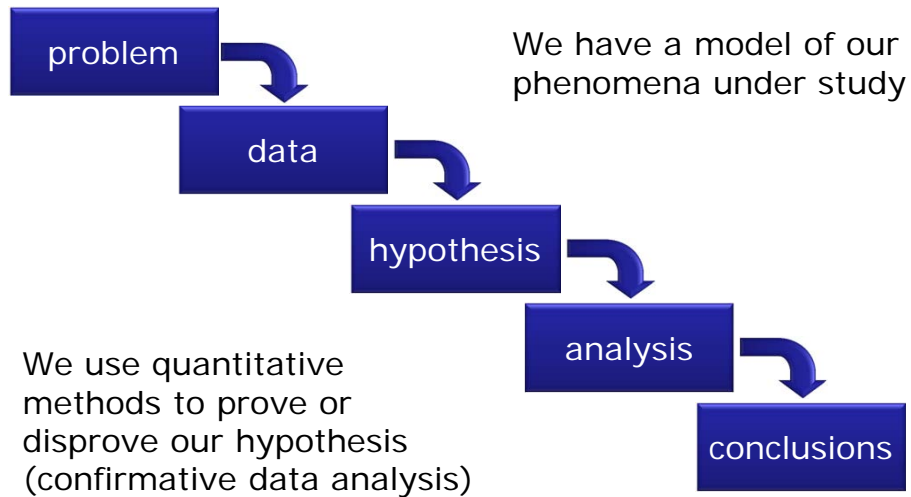
- We found unexpected correlations between distance and other physical variables
- For example the deceptively simple H₂O shows clear correlations and grouping
- This and other datasets motivated us to continue the exploration of the crystal fingerprints' space...



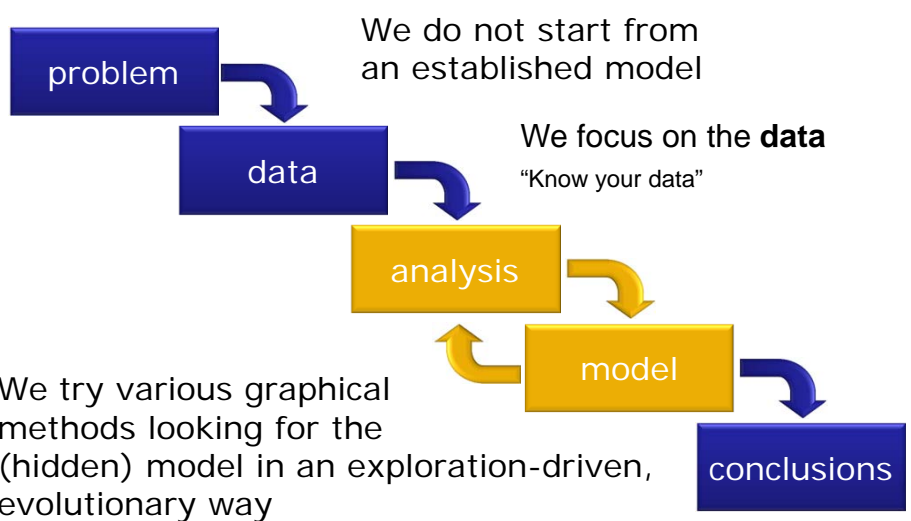
“And roughly the only mechanism for suggesting questions is exploratory”

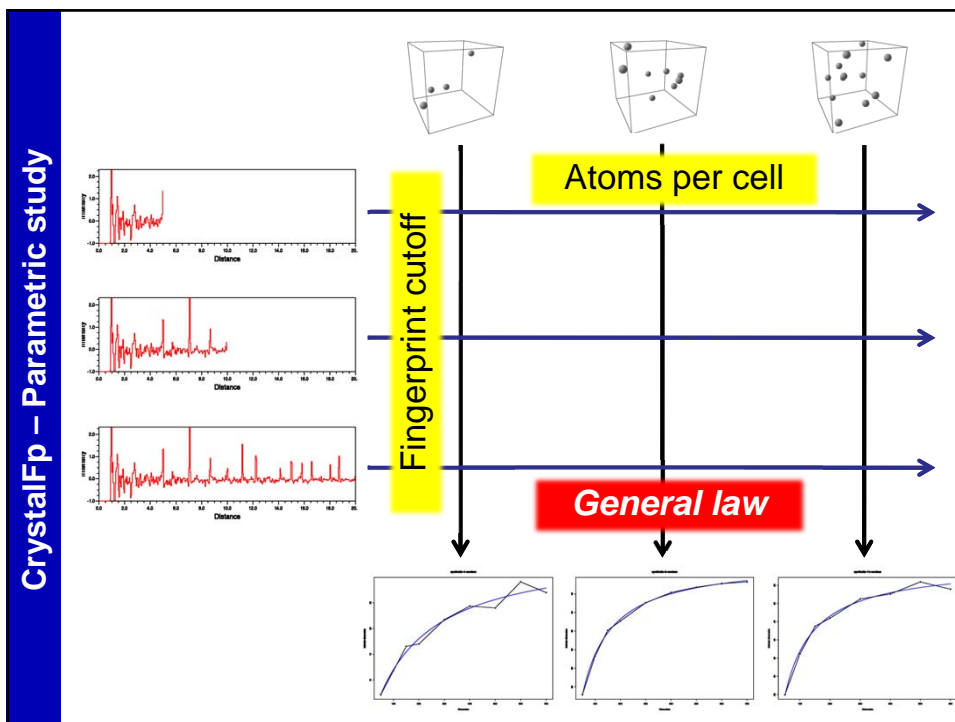
*A conversation with John W. Tukey and Elizabeth Tukey
Luisa T. Fernholz and Stephan Morgenthaler
Statistical Science
Volume 15, Number 1 (2000), 79-94*

Classical data analysis



Exploratory data analysis





CrystalFp Wiki
The Crystal Fingerprinting method on-going research

Navigation: [fingerprinting method](#) [library](#) [documentation](#) [current research](#) [pmwiki](#) [back to site](#)

CrystalFp: the crystal fingerprinting project

CrystalFp started as a way to solve a problem with the USPEX crystal structure predictor. 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 a method to find and remove duplicated structures has to be found.

The approach adopted was to apply usual high-dimensional classification concepts to the unusual field of crystallography.

We adopted a visual design and validation method to develop a classifier library (*CrystalFp*) and an end-user application to select and validate method choices, to gain users' acceptance and to tap into their domain expertise.

Using the end-user application with real datasets, 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.

The use of the classifier has already accelerated the analysis of USPEX output by at least one

SEARCH

ABOUT

To study ensembles of crystal structures, the **fingerprinting method** converts them to points in a high dimensional space. Analyzing this space, we are able to infer properties of the original set of data, like the shape of their energy landscape.

SECTIONS

- Home Page
- Fingerprinting method
- CrystalFp library**
- documentation
- Current Research


```

Usage:
CrystalFp [options] POSCARfile [ENERGIESfile]

-v --verbose (optional argument)
    Verbose level (if no argument, defaults to 1)

-? -h --help (no argument)
    This help

-t --elements (required argument)
    List of chemical elements

-es --max-step --end-step (required argument)
    Last step to load (default: all)

-ss --start-step (required argument)
    First step to load (default: first)

-et --energy-per-structure (no argument)
    Energy from file is per structure, not per atom

-e --energy-threshold (required argument)
    Energy threshold

-r --threshold-from-min (required argument)
    Threshold from minimum energy

-c --cutoff-distance (required argument)
    Fingerprint forced cutoff distance

-n --nano-clusters --nanoclusters (no argument)
    The structures are nanoclusters, not crystals

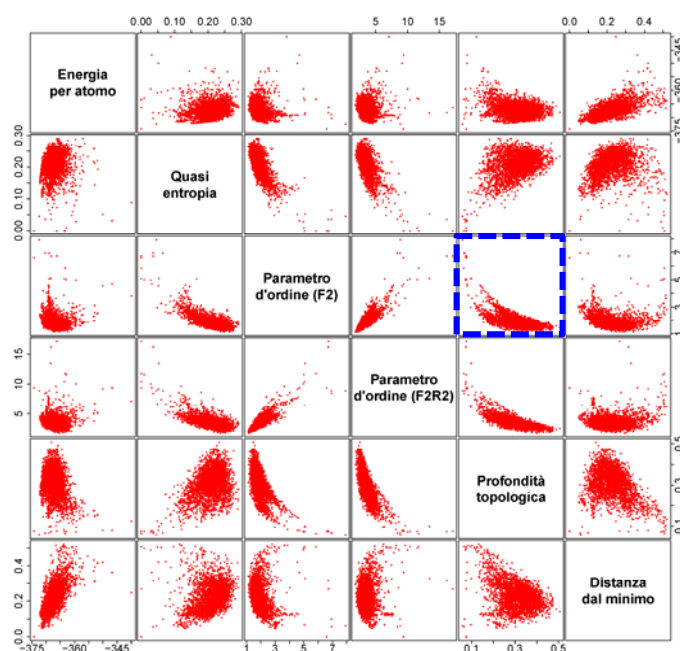
-b --bin-size (required argument)
    Bin size for the pseudo-diffraction methods

-p --peak-size (required argument)
    Peak smearing size

...

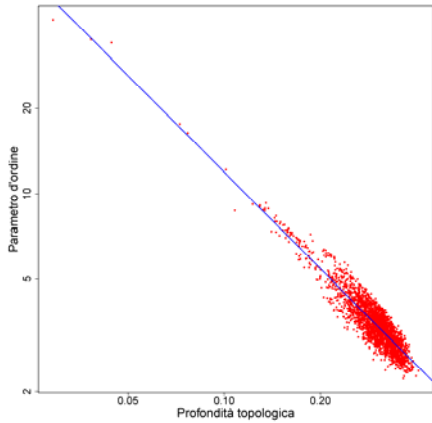
```

Dataset: SiO2-09atoms_random - Dimensionalità dello spazio: 240

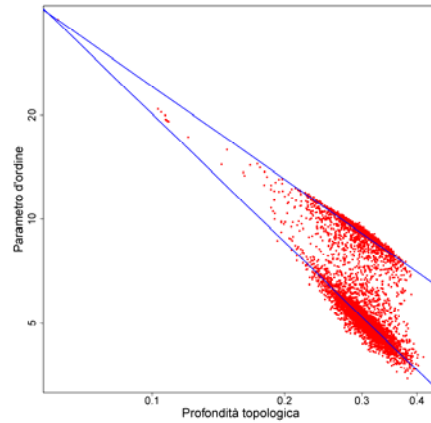


Interesting correlations

Dataset: SiO2-06atoms_random - Dimensionalità dello spazio: 1800

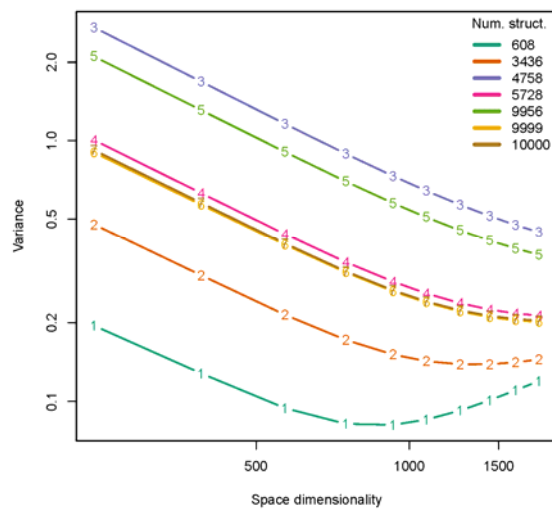


Dataset: SiO2-03atoms_random - Dimensionalità dello spazio: 1080

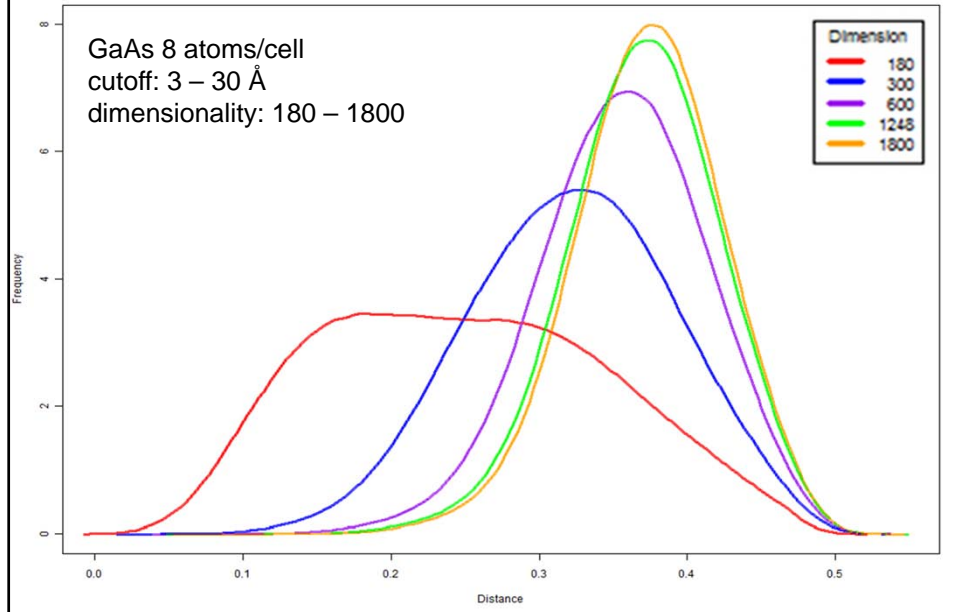


Searching an explanation

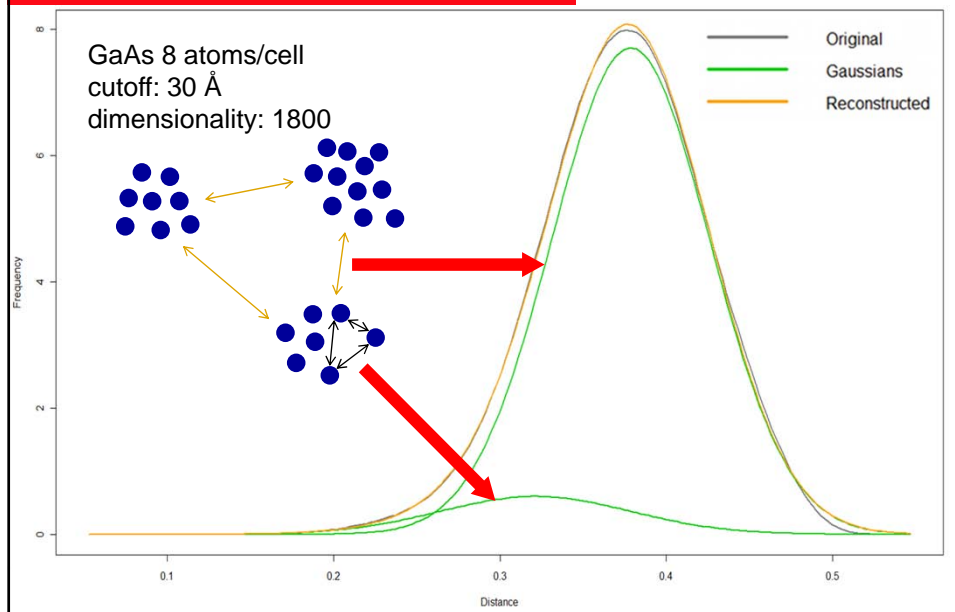
Phantom 48 atoms



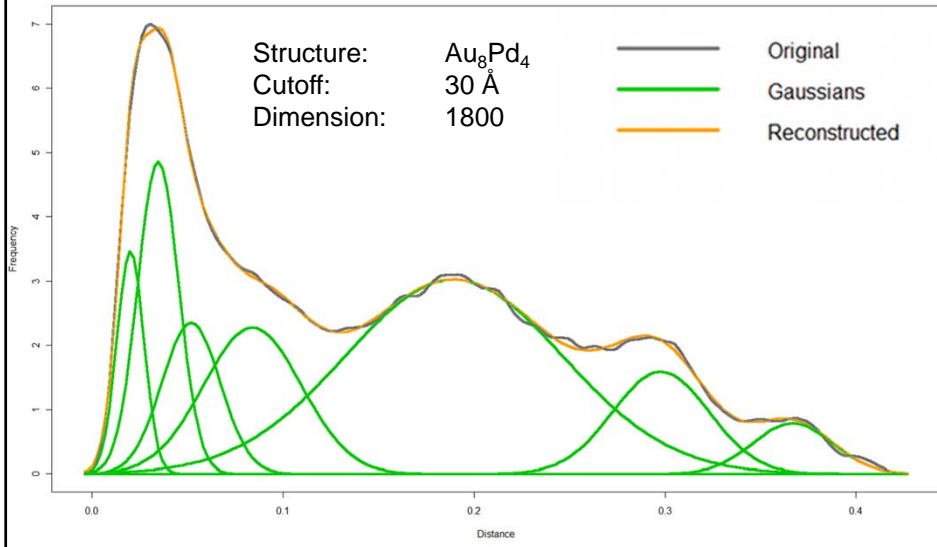
Distance vs. dimensionality



Distance decomposition



But this one?



Intrinsic dimensionality

Fingerprint space
 dimensionality:
 $100 \div 3000$

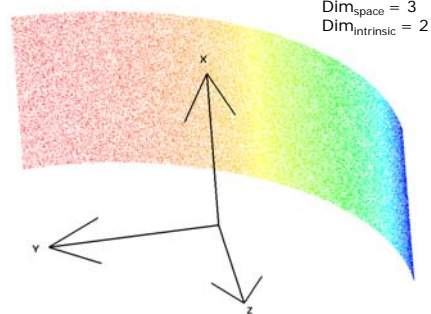
Vastly redundant!

Theory:

$$\text{Dim}_{\text{intrinsic}} = 3 * N_{\text{atoms}} + 3$$

More realistic theory:

$$\text{Dim}_{\text{intrinsic}} = 3 * N_{\text{atoms}} + 3 - K$$

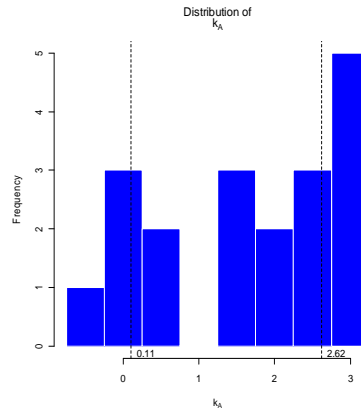


Au_8Pd_4	39	10.85
MgNH	39	32.47
MgO	99	11.62

Constraints per atom

$$Dim_{intrinsic} = 3 * N_{atoms} + 3 - K \rightarrow (3 - K_A) * N_{atoms} + 3$$

Dataset	Theor. Dim.	Intr. Dim.	k_A
au8pd4	39	11.94	2.25
Ca-16at-160GPa	51	27.37	1.48
ca-16at-300GPa	51	29.19	1.36
carbon-0GPa-8atoms-final	27	25.88	0.14
ch2-800GPa-18at	57	61.94	-0.27
gaas-8at_new	27	26.47	0.07
h2o	39	90.78	-4.32
H-300GPa-12at	39	4.48	2.88
H-500GPa-16at	51	25.78	1.58
H-500GPa-8at	27	4.29	2.84
l4j8a	39	2.30	3.06
l4j8	39	2.87	3.01
mgnh-2.5eV-threshold1	39	9.67	2.44
mgnh-total4	39	53.03	-1.17
mgo32a	99	10.58	2.76
mgofull	99	15.05	2.62
Na-140GPa-8at	27	27.22	-0.03
urea-0GPa	51	18.79	2.01
GaAs-old	27	5.64	2.67
MgSiO3_Postperovskite_120GPa	63	56.34	0.33
GaAs_random	27	23.45	0.44
MgNH-random	39	63.69	-2.06



Intrinsic dimensionality

Fingerprint space
dimensionality:
100 ÷ 3000

Vastly redundant!

Theory:

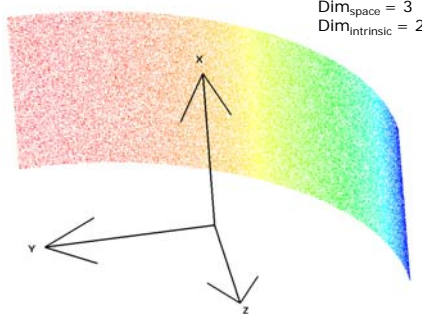
$$Dim_{intrinsic} = 3 * N_{atoms} + 3$$

More realistic theory:

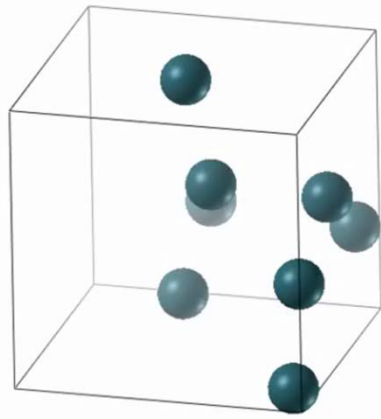
$$Dim_{intrinsic} = 3 * N_{atoms} + 3 - K$$

Au ₈ Pd ₄	39	10.85
MgNH	39	32.47
MgO	99	11.62
H₂O	39	80.50

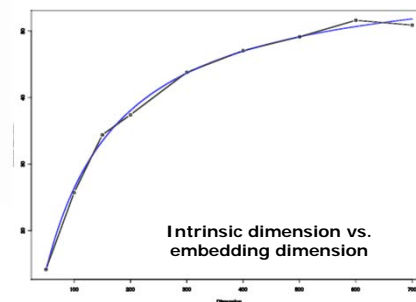
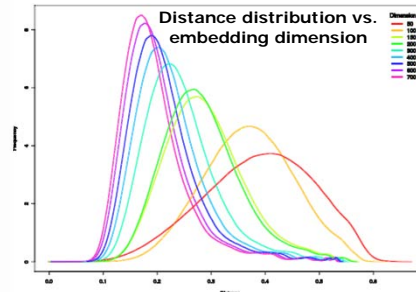
But how do you explain this?



Synthetic datasets



8 atoms with uniformly distributed random fractional coordinates in a cubic unit cell with 5 Å side



Lessons learned

From the Modeling side

- Using known concepts in unusual contexts is a source of unexpected insights
- Discoveries happen on the boundaries of disciplines
- “Seeing is believing” and convincing. Then the domain experts become a source of ideas

From the Visual Analysis side

- Quick prototyping and experimentation capabilities are critical (that is, STM4 is a big help)
- No need of fancy visualizations. What are needed are visualizations tuned to the problem at hand
- Data management is critical to keep order in the data exploration

<http://www.cscs.ch/~mvalle/STM4>

CrystalFp: the crystal fingerprinting project

CrystalFp started as a way to solve a problem with the USPEX crystal structure predictor. 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 a method to find and remove duplicated structures has to be found.

The approach adopted was to apply usual high-dimensional classification concepts to the unusual field of crystallography.

We adopted a visual design and validation method to develop a classifier 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

<http://www.cscs.ch/~mvalle/CrystalFp>

**Thank you
for your attention!**

And don't forget: mvalle@cscs.ch