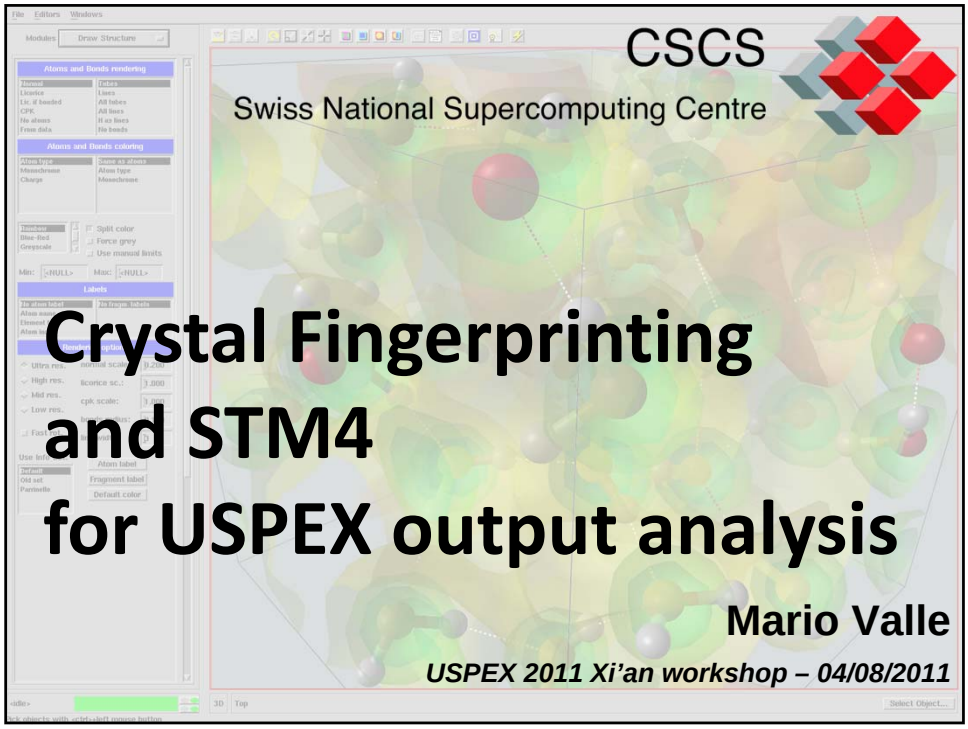
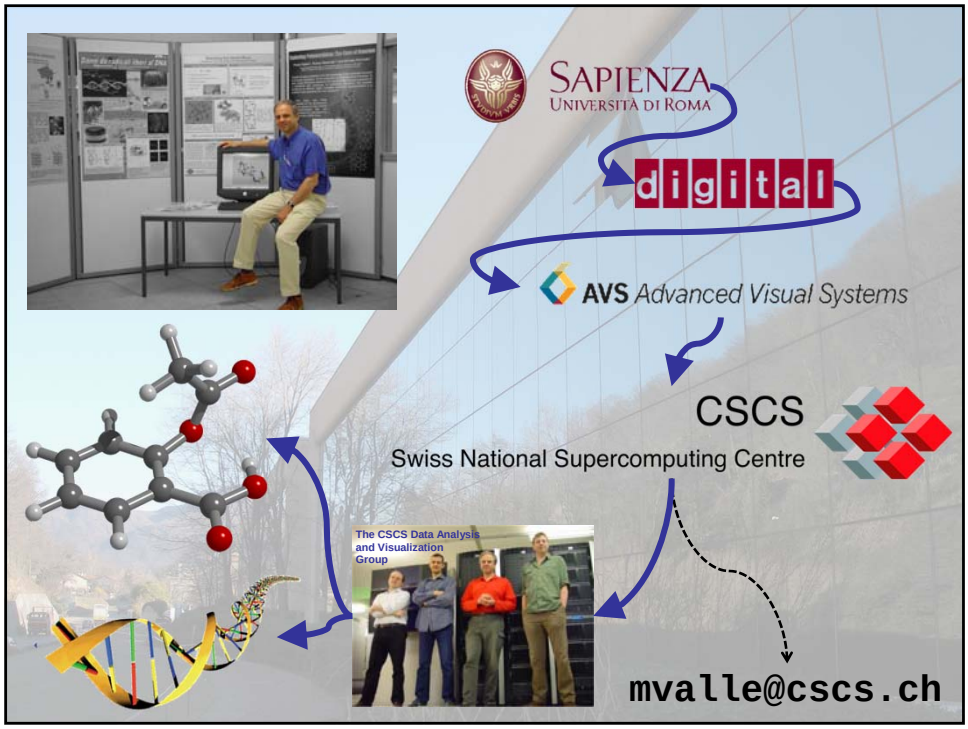
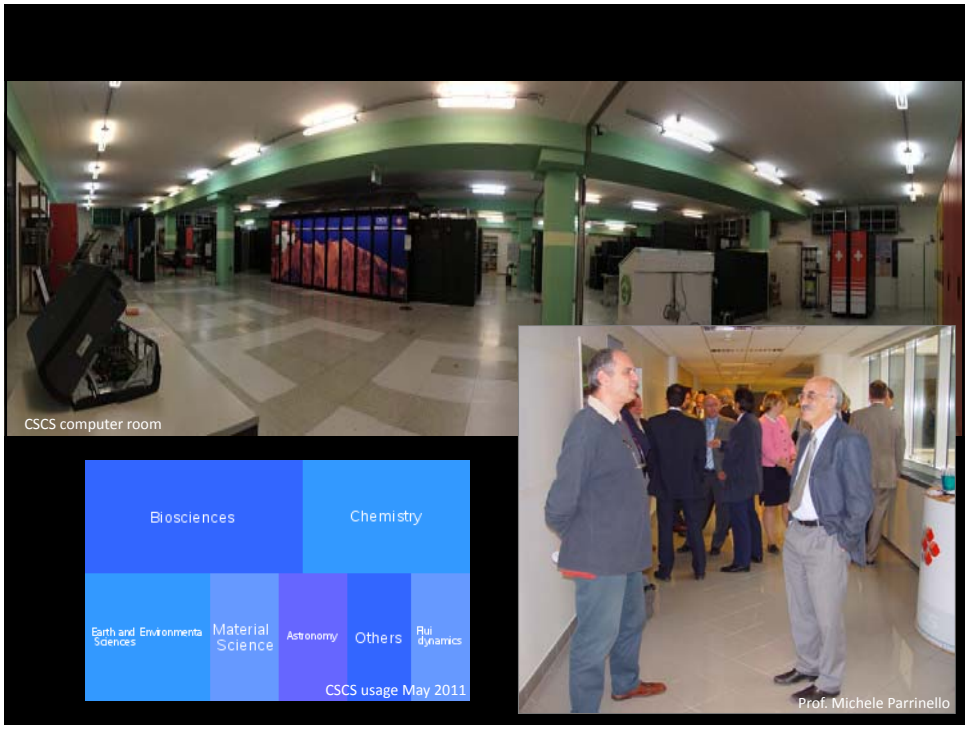
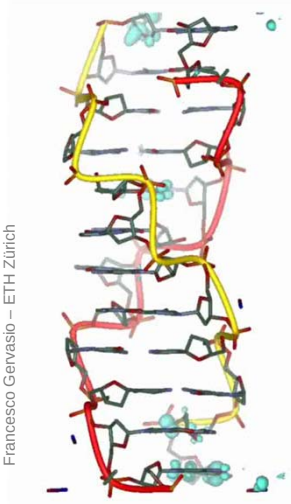

CSCS
Swiss National Supercomputing Centre
Crystal Fingerprinting
and STM4
for USPEX output analysis
Mario Valle
USPEX 2011 Xi'an workshop – 04/08/2011

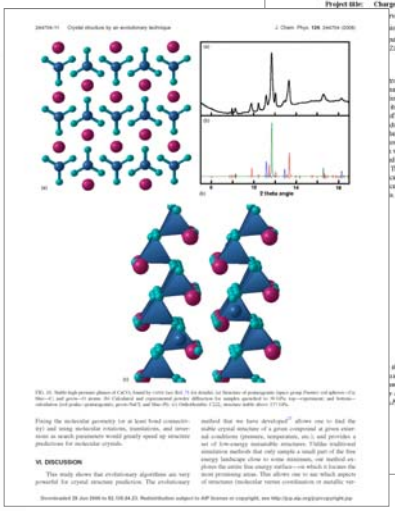

SAPIENZA
UNIVERSITÀ DI ROMA
digital
AVS Advanced Visual Systems
CSCS
Swiss National Supercomputing Centre
The CSCS Data Analysis and Visualization Group
mvalle@cscs.ch

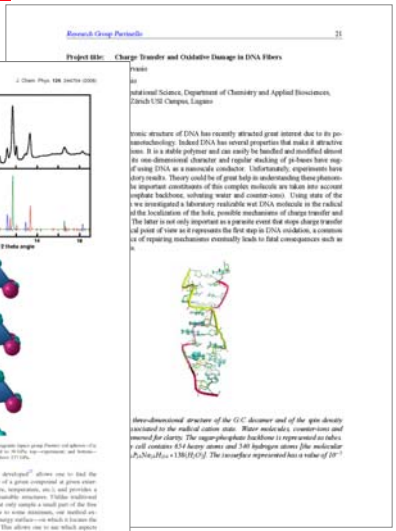


Started with postprocessing

Movies for conference talks





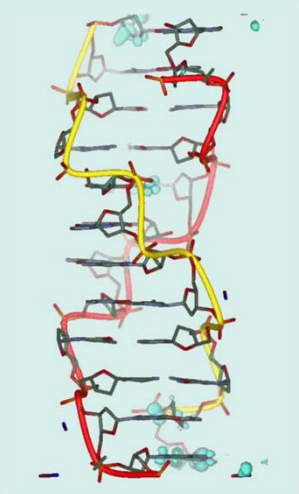


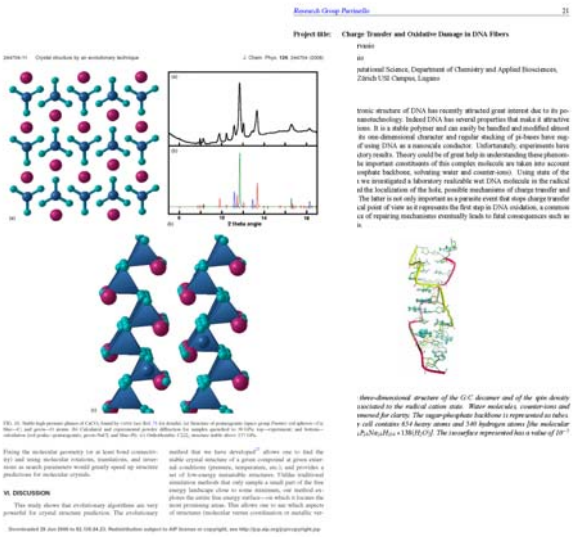
Nice images for publications

Francesco Gervasio – ETH Zürich

Started with postprocessing

Movies for conference talks

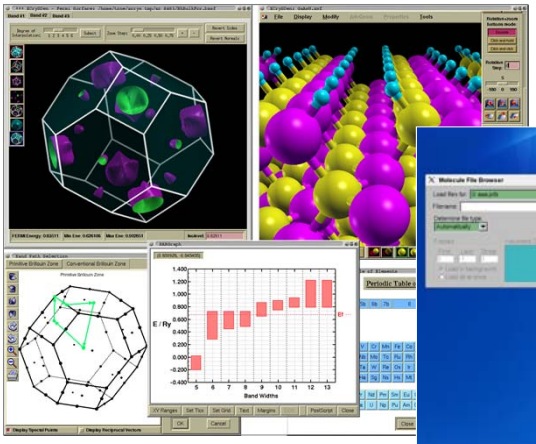




Nice images for publications

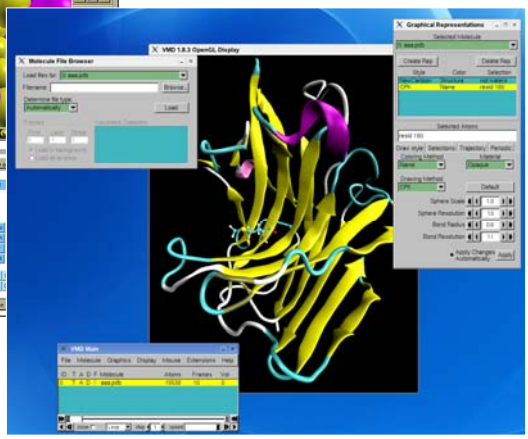
Francesco Gervasio – ETH Zürich

Data from disjoint subfields

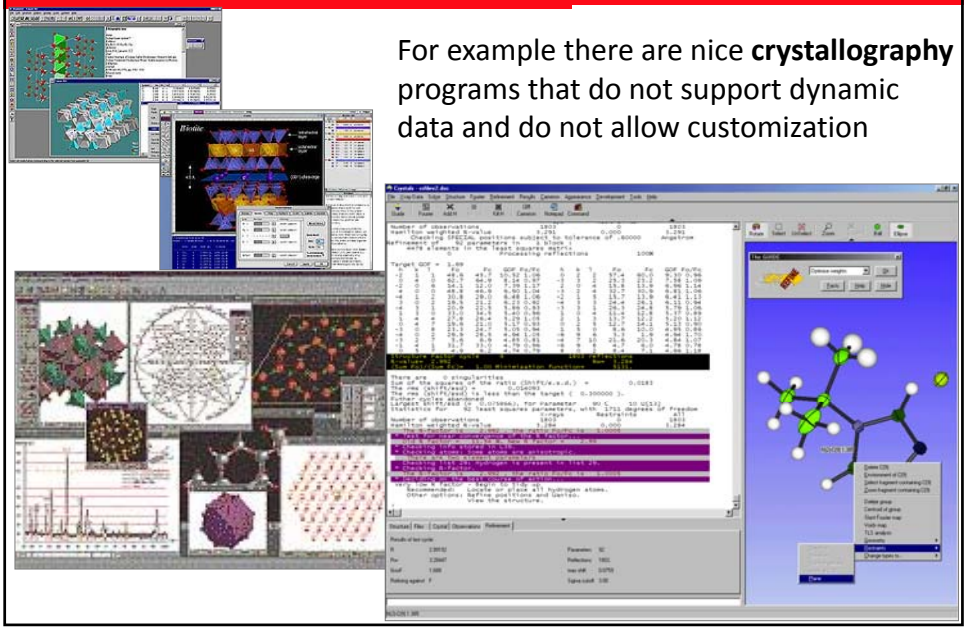


Crystallography

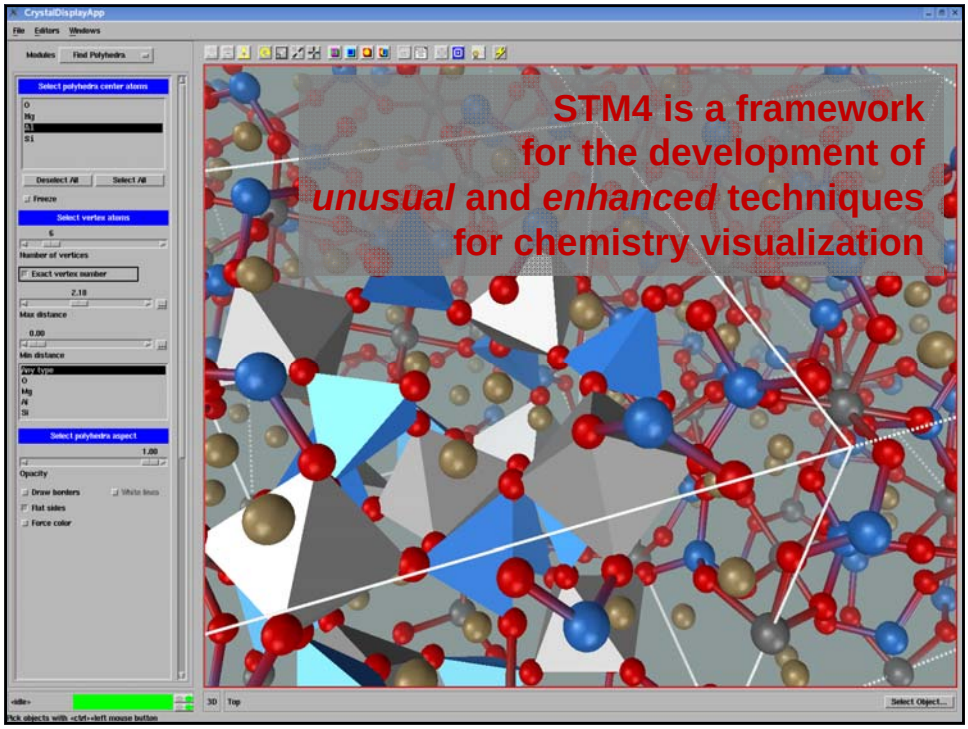
Macromolecules



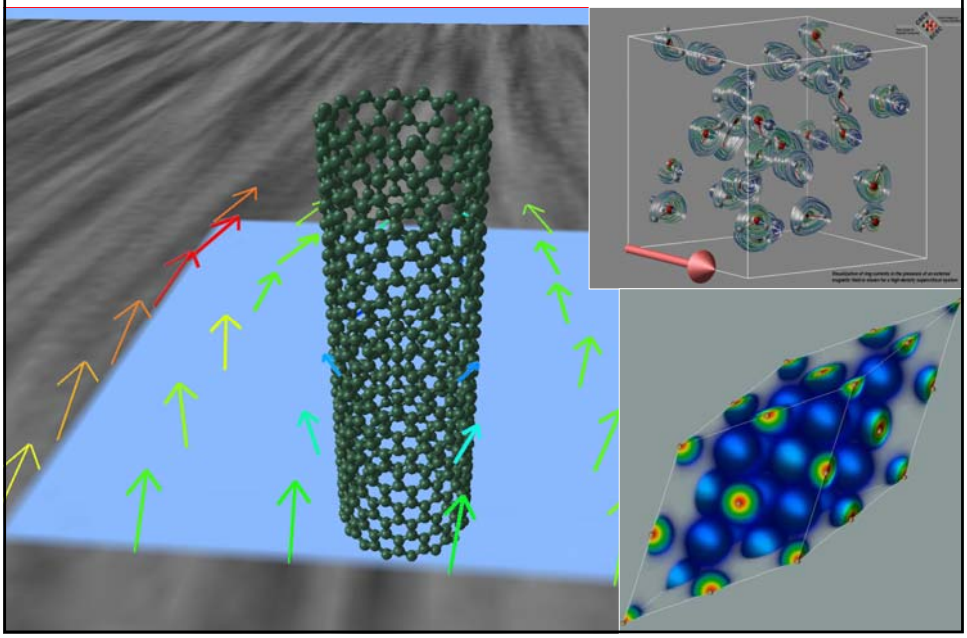
Help overcome tools inflexibility



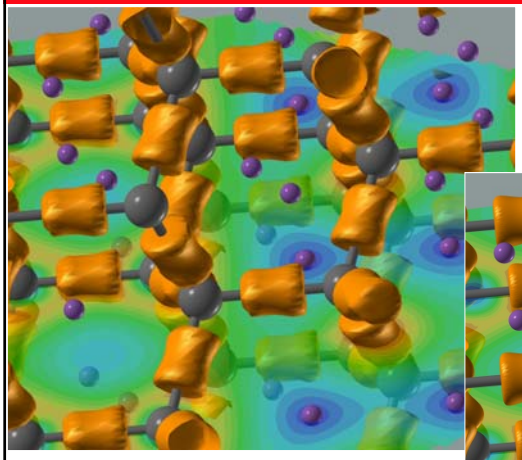
For example there are nice **crystallography** programs that do not support dynamic data and do not allow customization



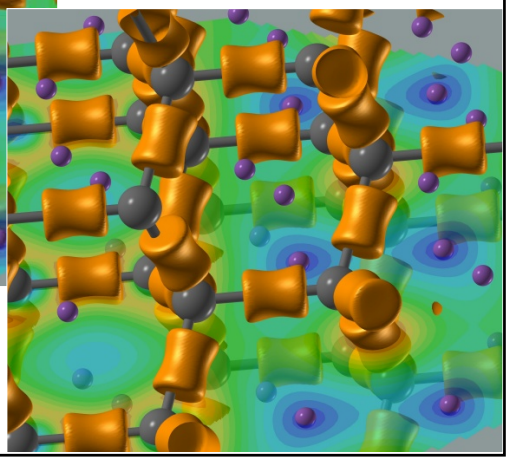
Offer broader set of techniques



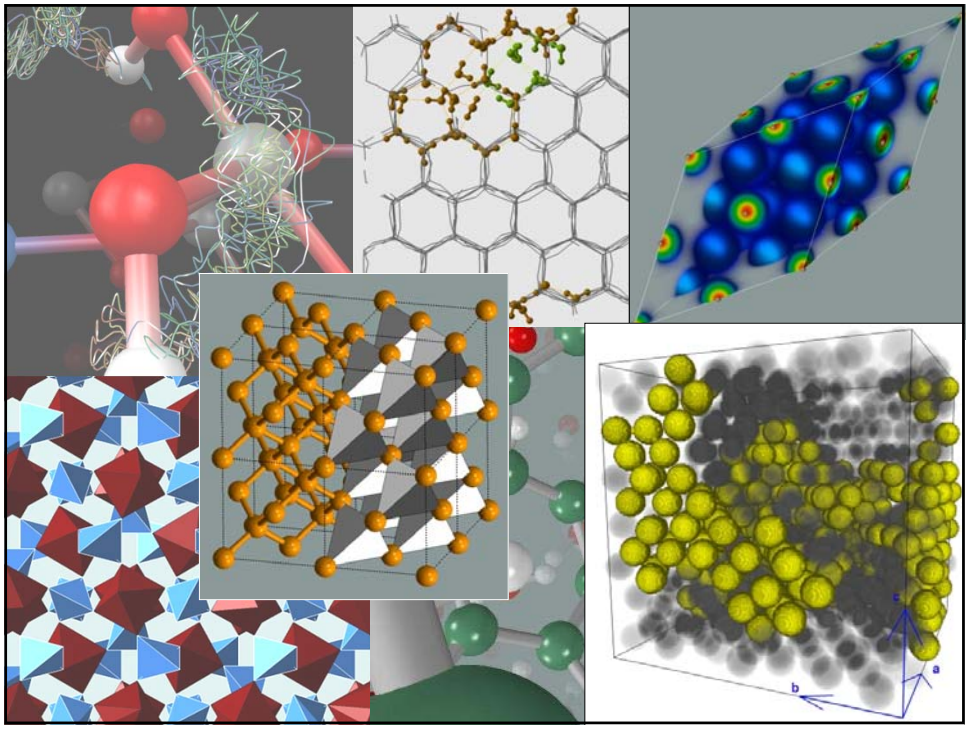
Provide enhanced techniques

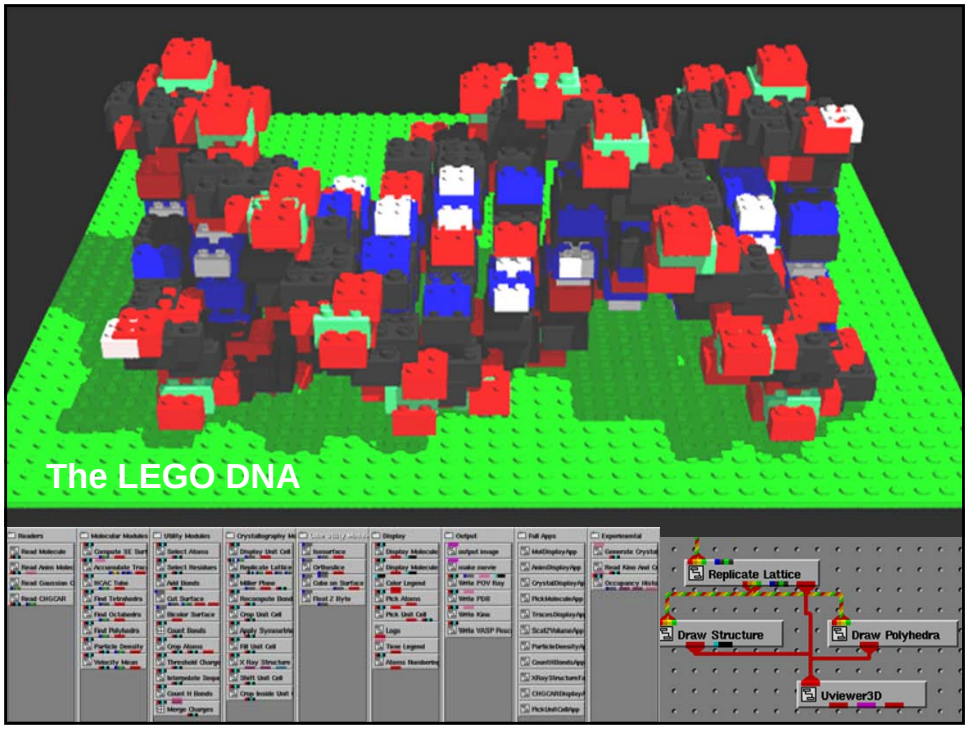


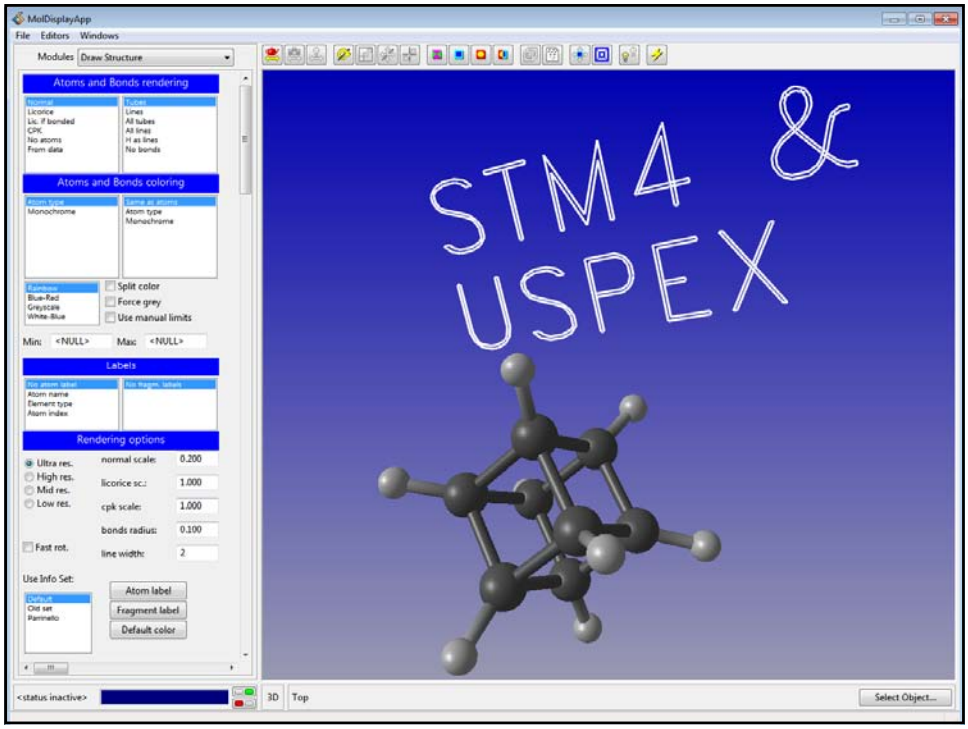
Available data with the standard isosurface

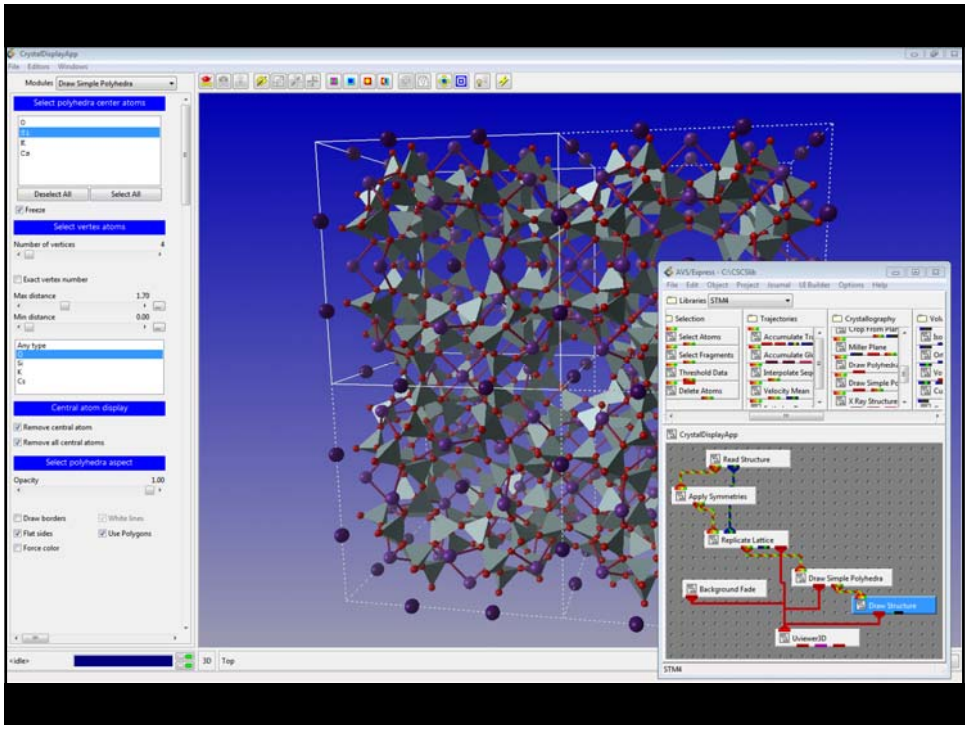


Isosurface with the new volume interpolator

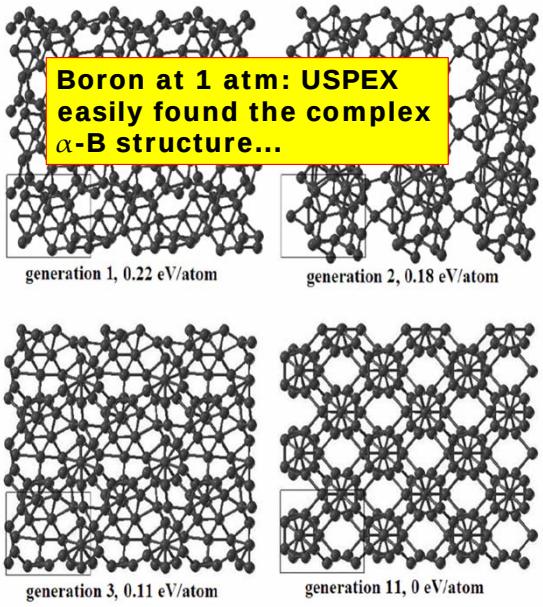



The LEGO DNA
Replicate Lattice
Draw Structure
Draw Polyhedra
Uviewer3D

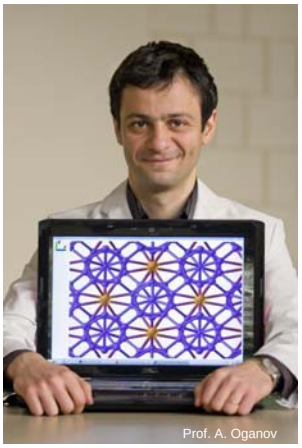

MolDisplayApp
File Editors Windows
Modules Draw Structure
Atoms and Bonds rendering
Atoms and Bonds coloring
Split color
Force grey
Use manual limits
Min: <NULL> Max: <NULL>
Labels
Rendering options
Ultra res.
High res.
Mid res.
Low res.
normal scale: 0.200
licorice sc.: 1.000
cpk scale: 1.000
bonds radius: 0.100
line width: 2
Fast rot.
Use Info Set:
Atom label
Fragment label
Default color
STM4 &
USPEX
<status inactive>
3D Top
Select Object...


CrystalDisplayApp
Modules Draw Simple Polyhedra
Select polyhedra center atoms
Deselect All
Select All
Freeze
Select vertex atoms
Number of vertices
Exact vertex number
Max distance
Min distance
Any type
Central atom display
Remove central atom
Remove all central atoms
Select polyhedra aspect
Opacity
Draw borders
Flat sides
Force color
Use Polygons
<idle>
3D Top
AVS/Express
Libraries STM
Selection
Trajectories
Crystallography
Select Atoms
Select Fragments
Threshold Data
Delete Atoms
Accumulate Tr
Interpolate Segs
Velocity Mean
Miller Plane
Draw Polyhedra
Draw Simple Pol
X Ray Structure
CrystalDisplayApp
Read Structure
Apply Symmetries
Replicate Lattice
Background Fade
Draw Simple Polyhedra
Draw Structure
UviewerJD
STMM

USPEX discovered new materials

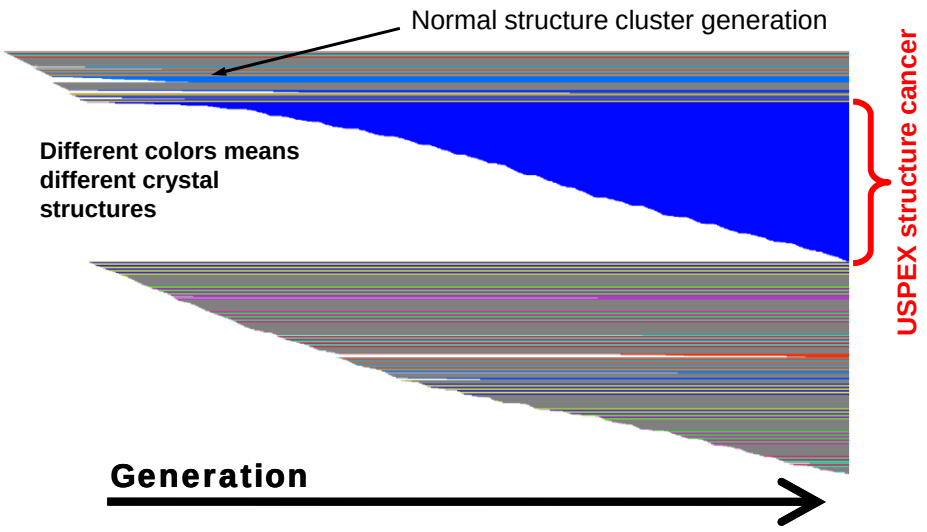


Boron at 1 atm: USPEX easily found the complex α-B structure...



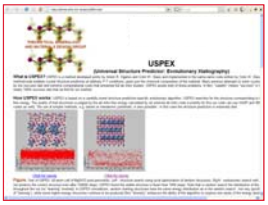
...and discovered also the superhard boron γ-B_{28} phase (*Nature*)

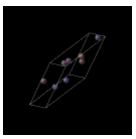
USPEX structure cancer problem



CrystalFp: the problem to solve

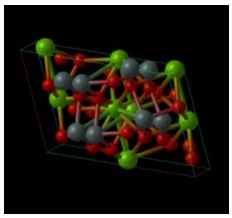
USPEX is a crystal structure predictor based on an evolutionary algorithm





Each run produces hundred of putative crystal structures…

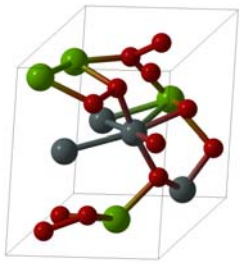
…but many of them are equal



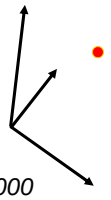
So an intensive manual labor is needed to prune duplicated structures

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

Proposed solution from High-Dim

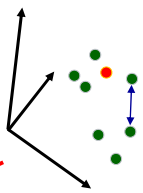


Compute unique coordinates

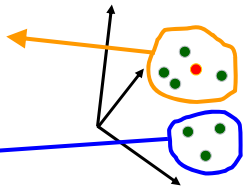


Space 100-3000 dimensional

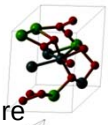
Define distance measure

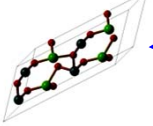


Add grouping criteria

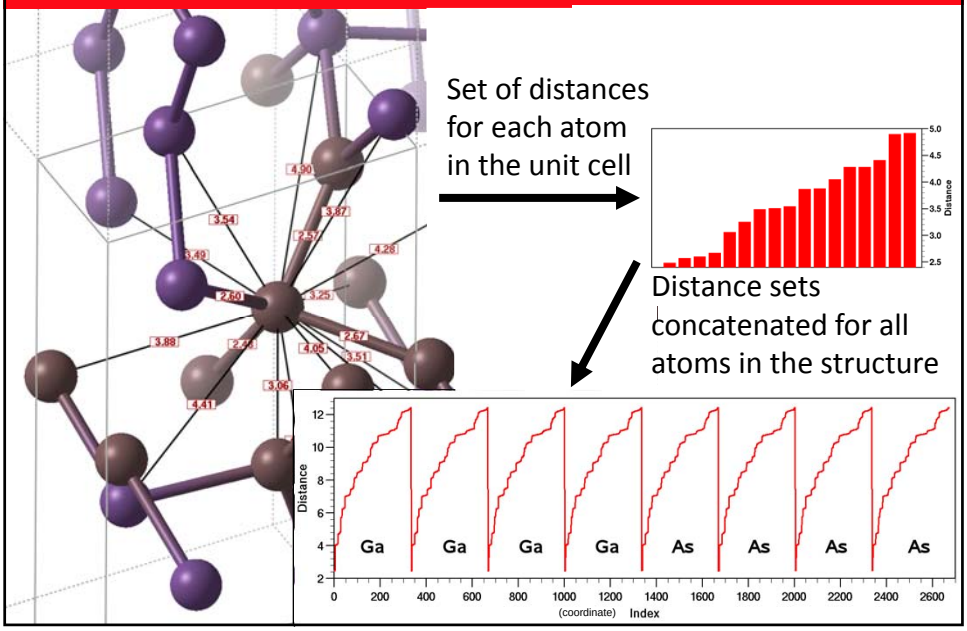


Each group describes a distinct structure

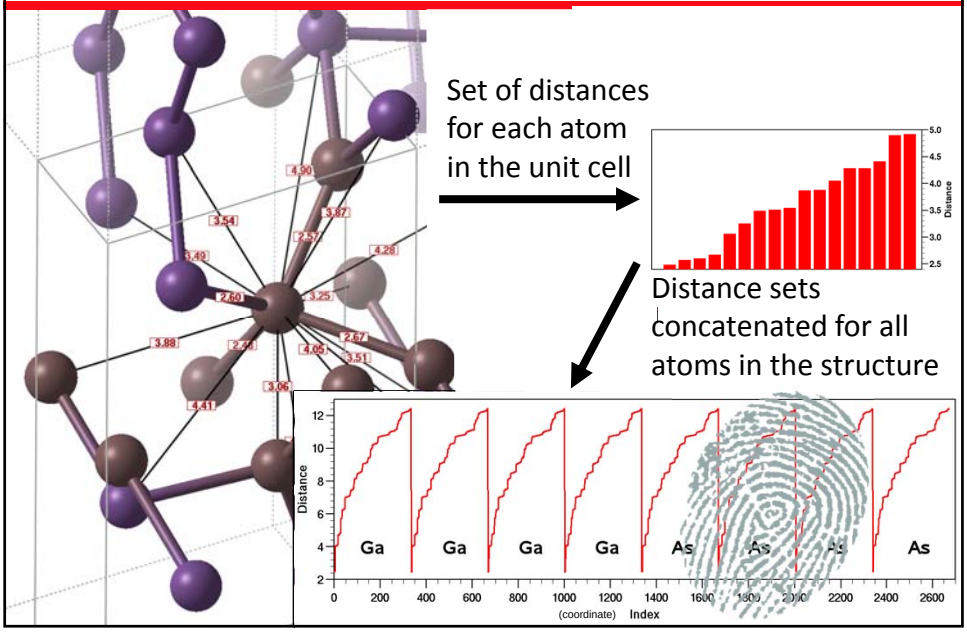


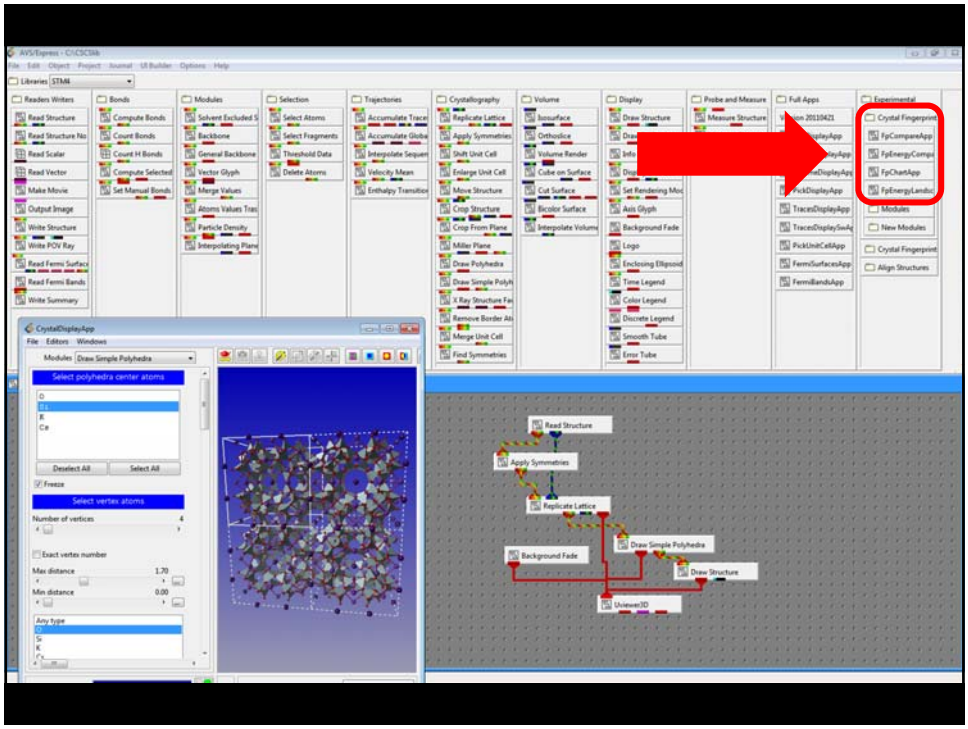


Structure “coordinates”

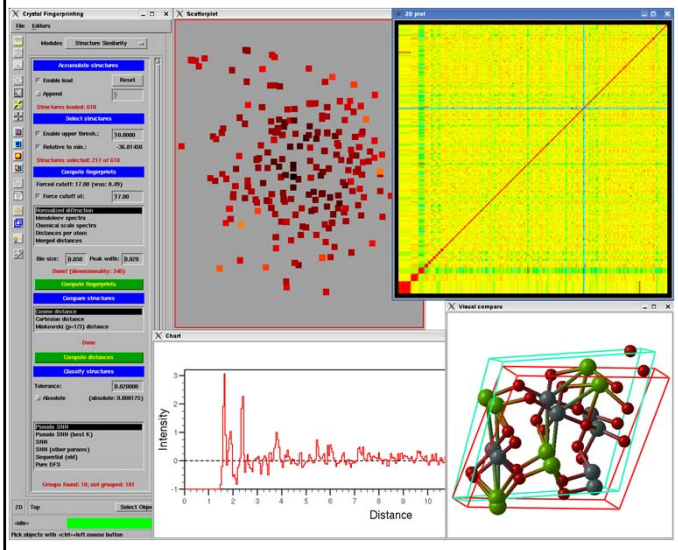


Structure “fingerprint” → CrystalFp





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

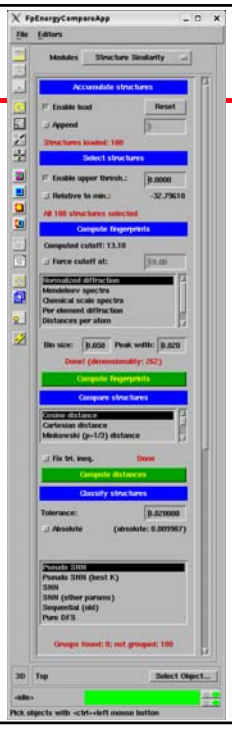
It is built inside STM4, the molecular visualization toolkit developed at CSCS

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

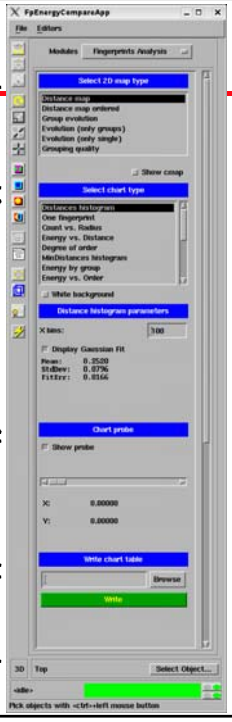
1. Load structures
2. Filter on energy
3. Compute fingerprints
4. Compute distances
5. Group structures



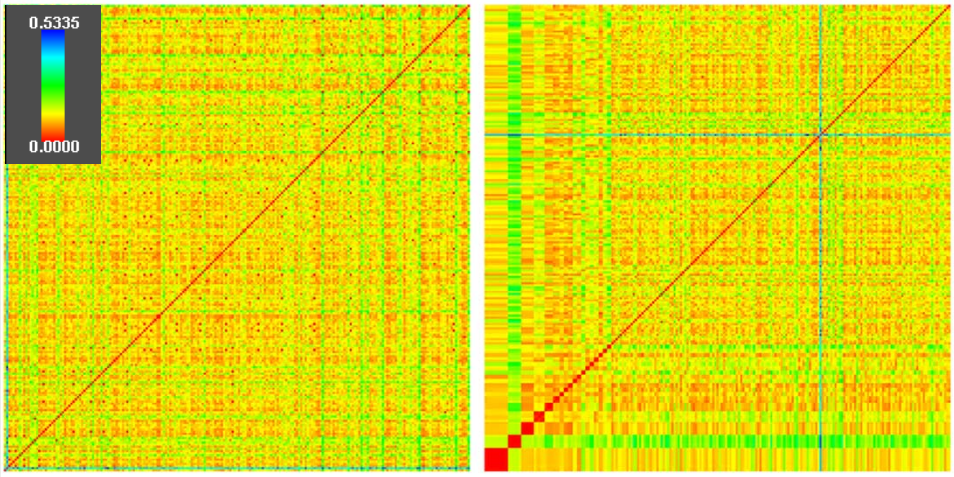
Visual diagnostics tools

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

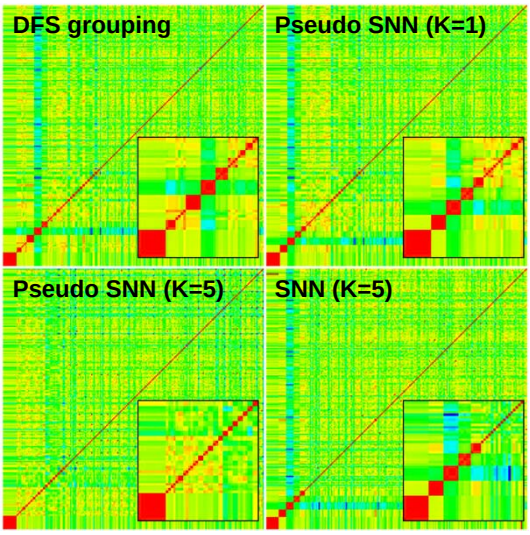


Visual diagnostics: distance matrix



Distances between structures Distances ordered by group

Clustering visual diagnostic



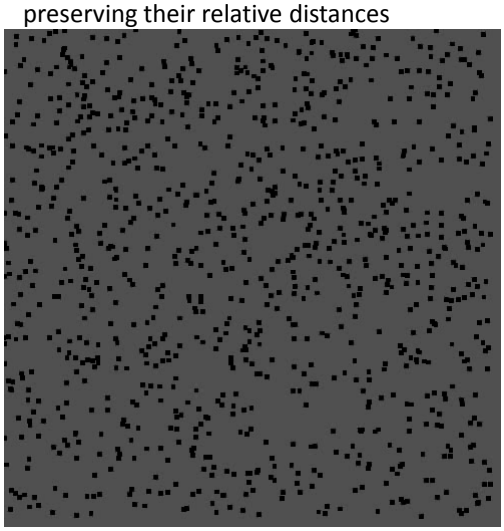
DFS: Deep first search of the neighbors nodes

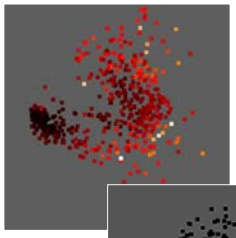
Pseudo SNN: Maintain connection between nodes only if they share at least K neighbors

SNN: As above plus a DBSCAN pass

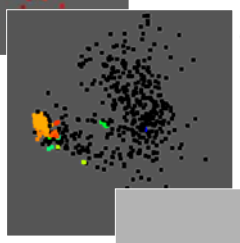
Visual diagnostics: scatterplot

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

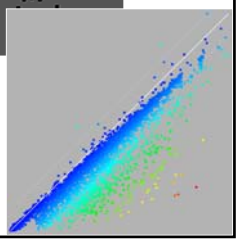




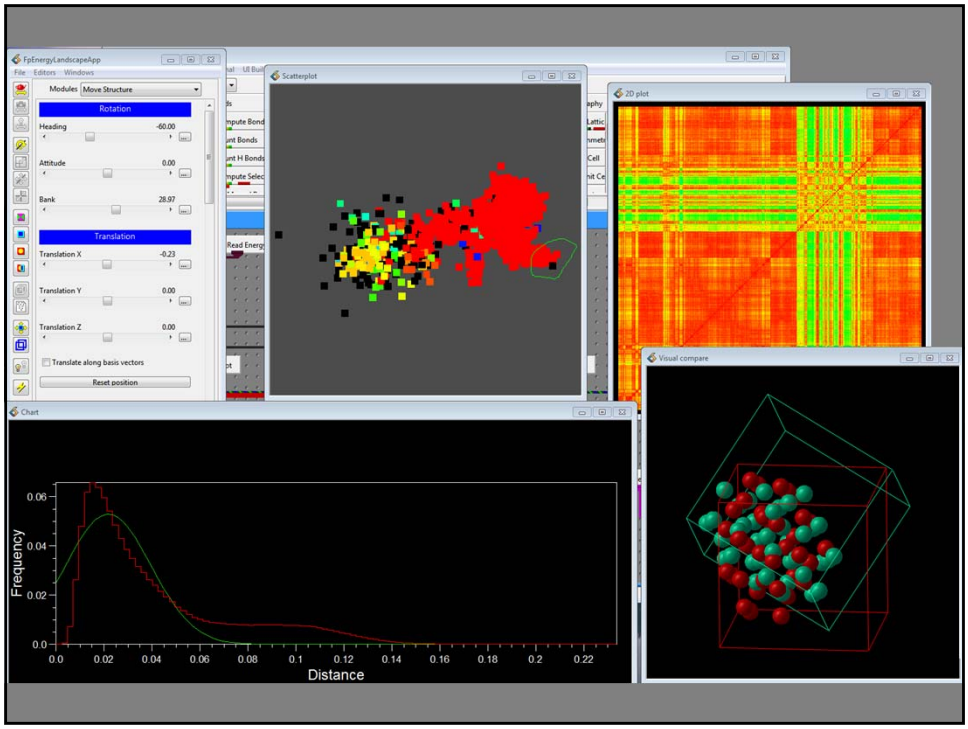
Colored by "stress" to detect local minima traps



Colored by group



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

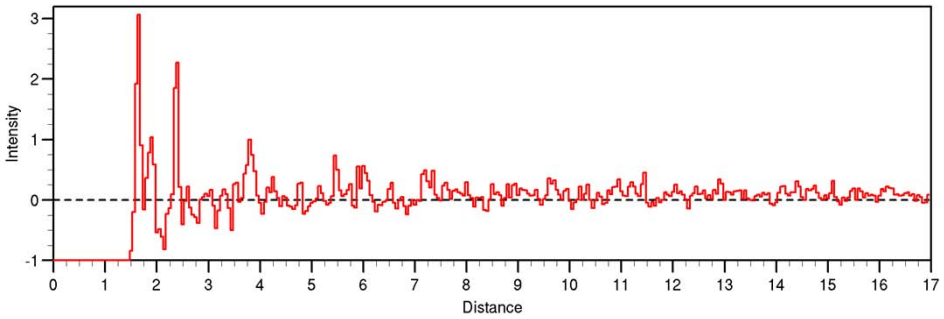


A pseudo-diffraction like method

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

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

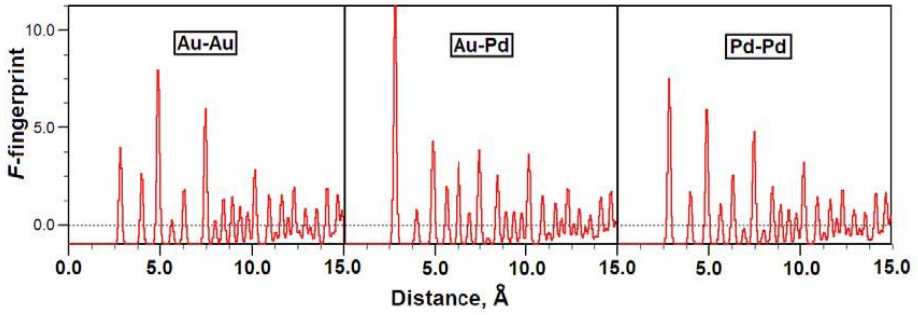
The fingerprint is cut at a user defined distance to provide 100-400 coordinate values



Final fingerprint (per atom type pair)

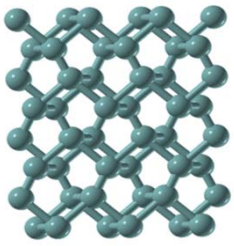
$$F_{AB}(R) = \sum_{A_i cell} \sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 \frac{N_A N_B}{V} \Delta} - 1$$

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

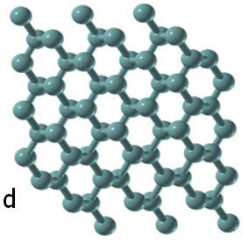


USPEX problem solved: an example

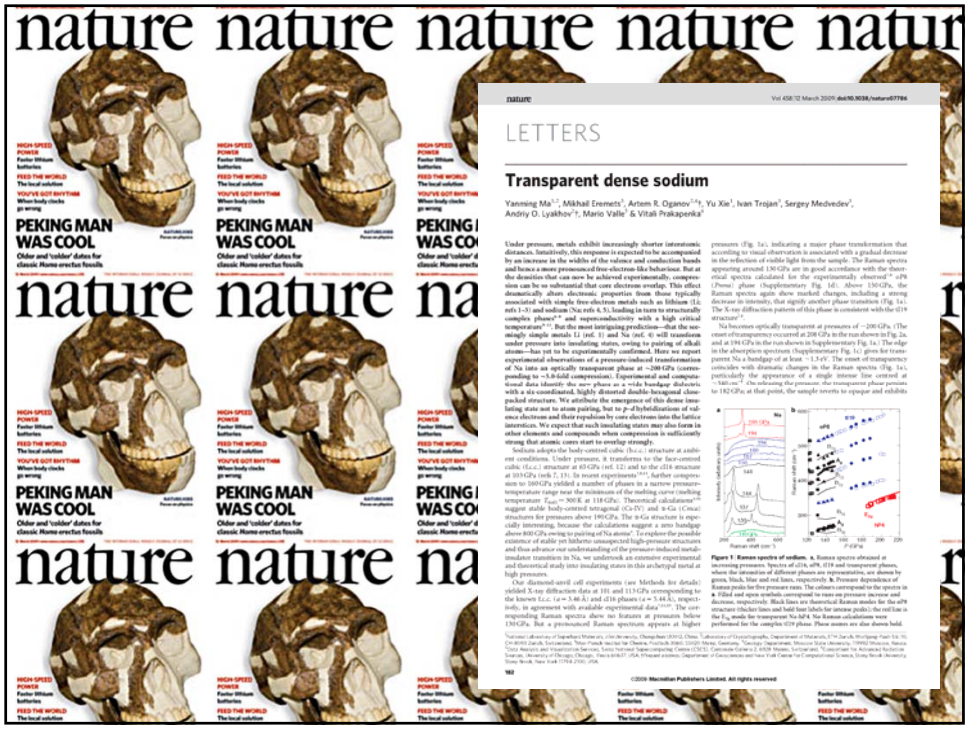
Hydrogen at 600 GPa (16 atoms)



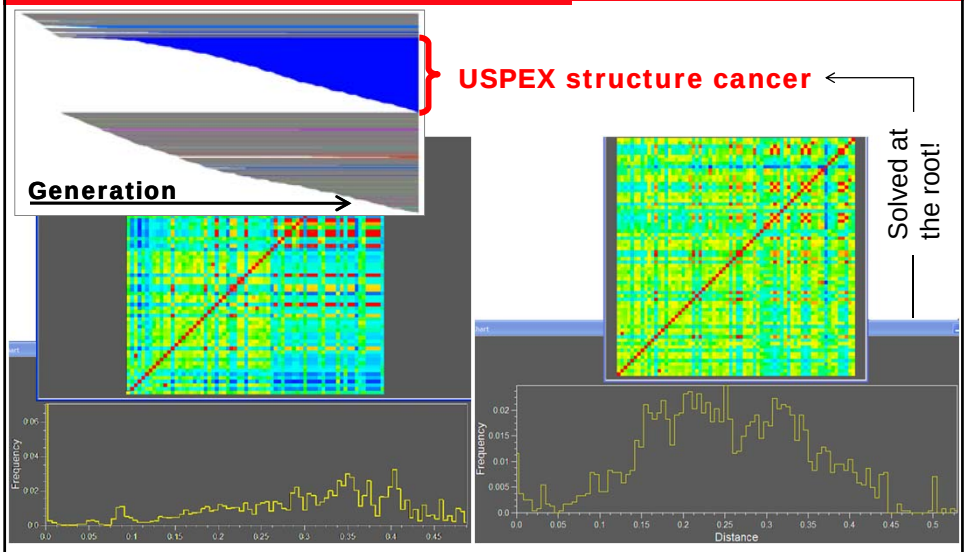
- The USPEX run produced **1274** structures
- From these the **794** within 0.5 eV from the lowest energy value found are selected
- Manual analysis to remove duplicated structures from this set: **~20h** of work
- Using the CrystalFp classifier: **~10min**



- 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



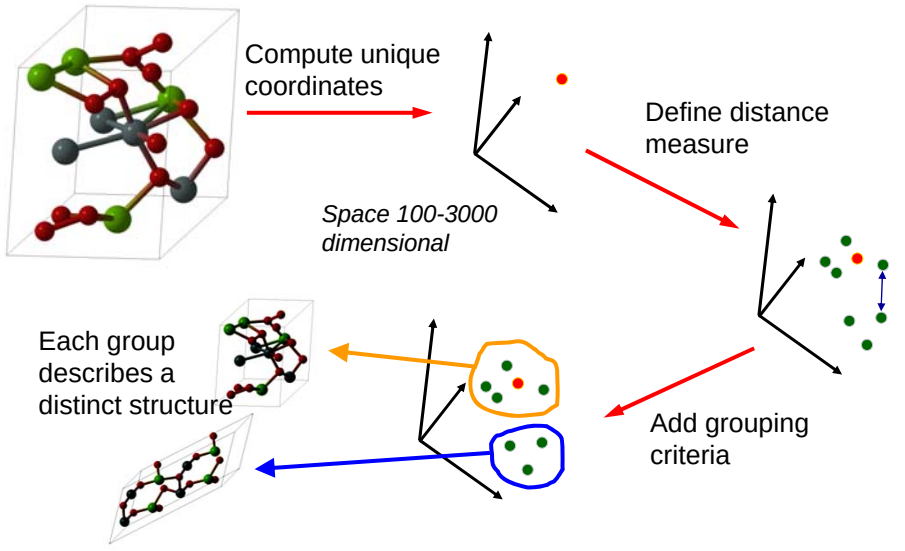
Classifier integration in USPEX



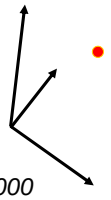
Original USPEX.
A lot of identical structures.

USPEX after the classifier integration.
No more “structure cancer”!

So, where is the problem?



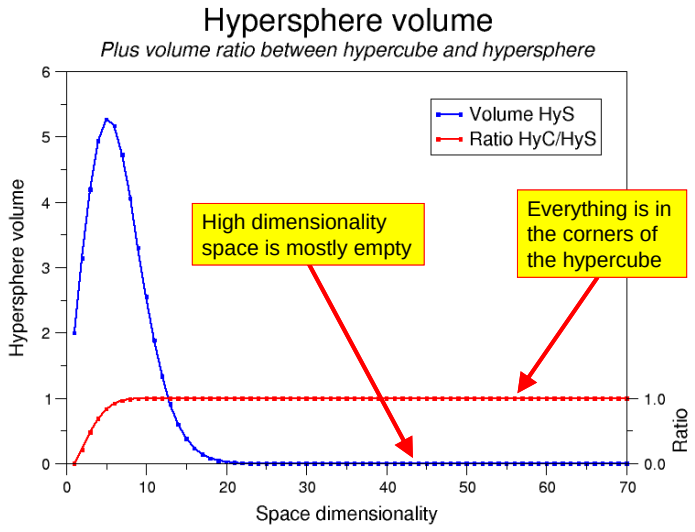
So, where is the problem?



Space 100-3000 dimensional

Space 100-3000 dimensional

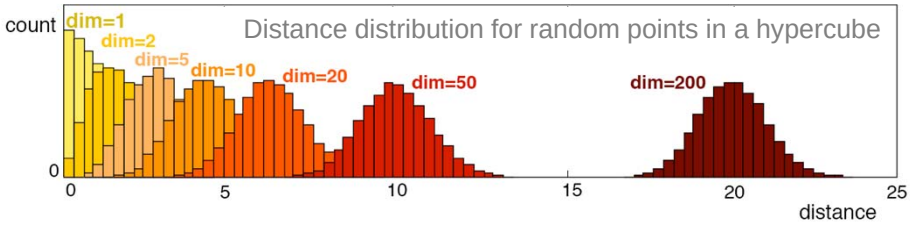
High-dimensionality is not intuitive



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 …

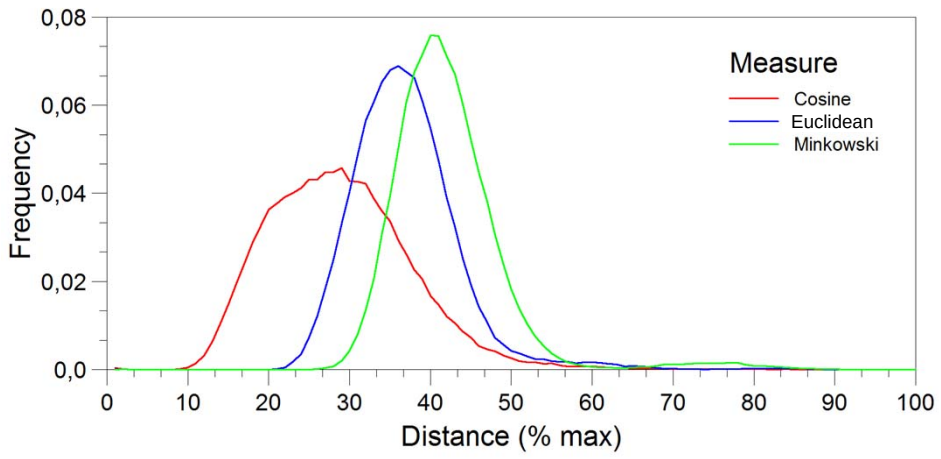
…could help contrast this curse of dimensionality

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

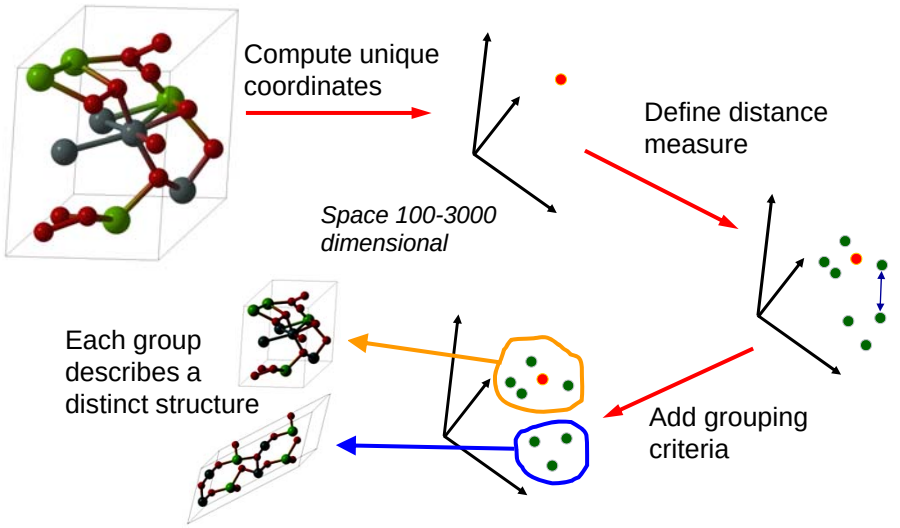
Minkowski distance: $D(a,b) = \sqrt[p]{\sum_{i=1}^{N}|b_i - a_i|^p}$

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

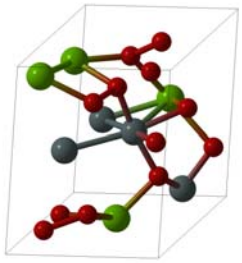
Tried various distance measures



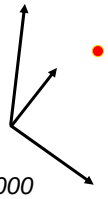
From the problem solution …



… to a new paradigm

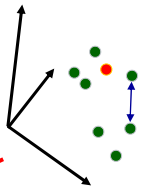


Compute unique coordinates



Space 100-3000 dimensional

Define distance measure

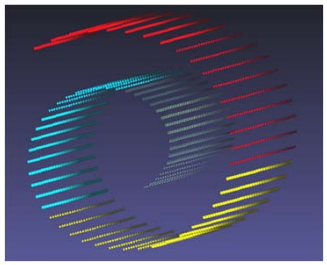


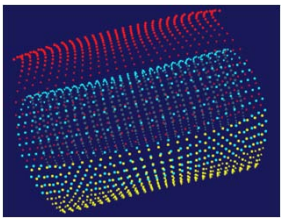
To look at crystal structures from a novel perspective

High Dim space tools

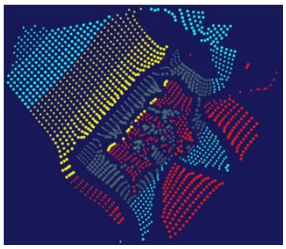
Unfold data to lower dimensions

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

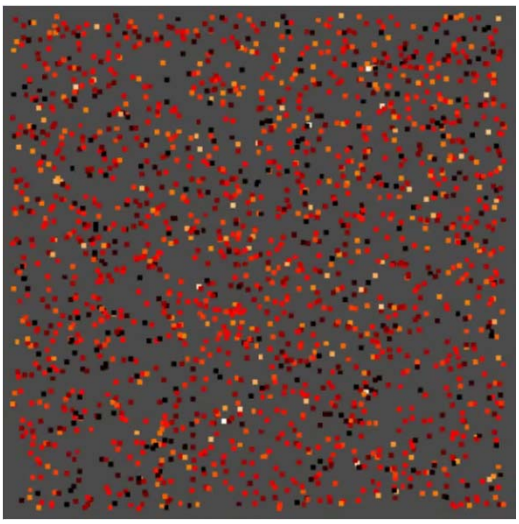


CCA mapping

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

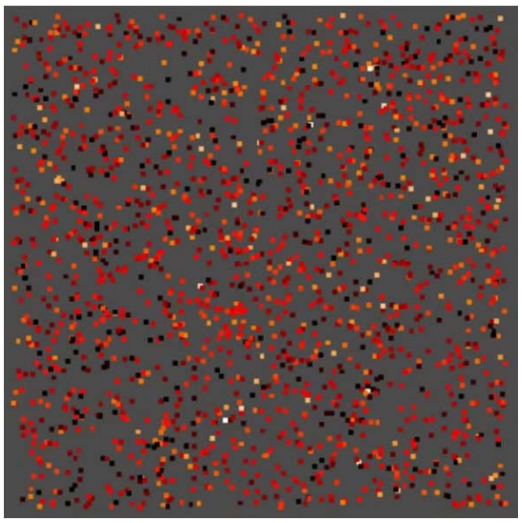
The scatterplot tool in CrystalFp implements a **Force Directed Placement** multidimensional scaling algorithm (here the points are colored by energy)



Nanocluster data from Dr. Gareth Tribello (USI)

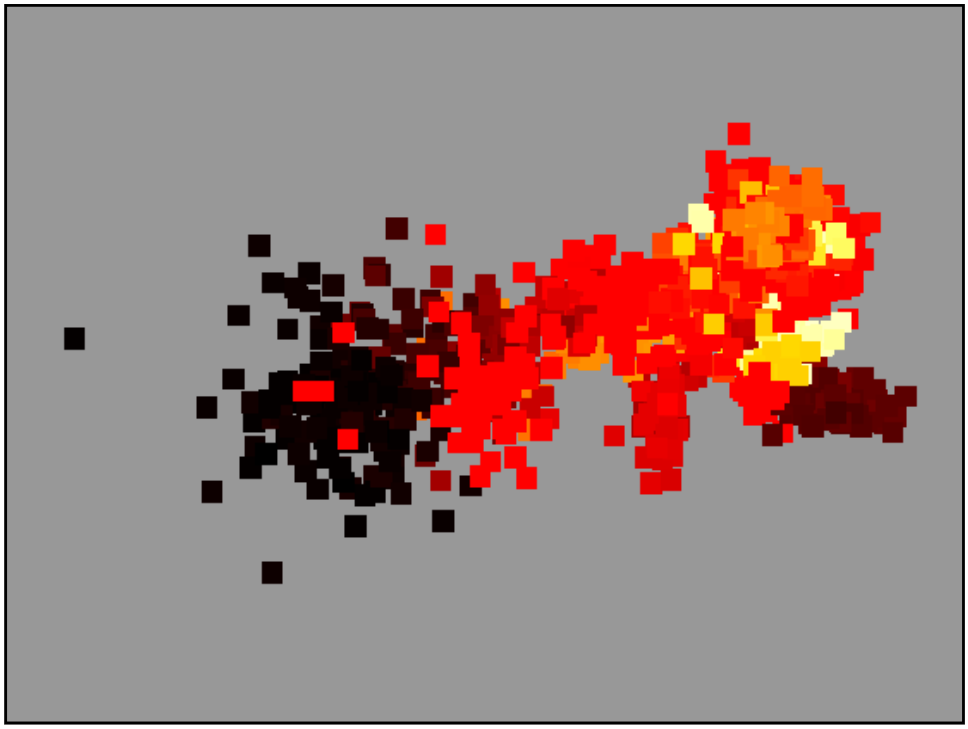
CrystalFp multi dim. scaling

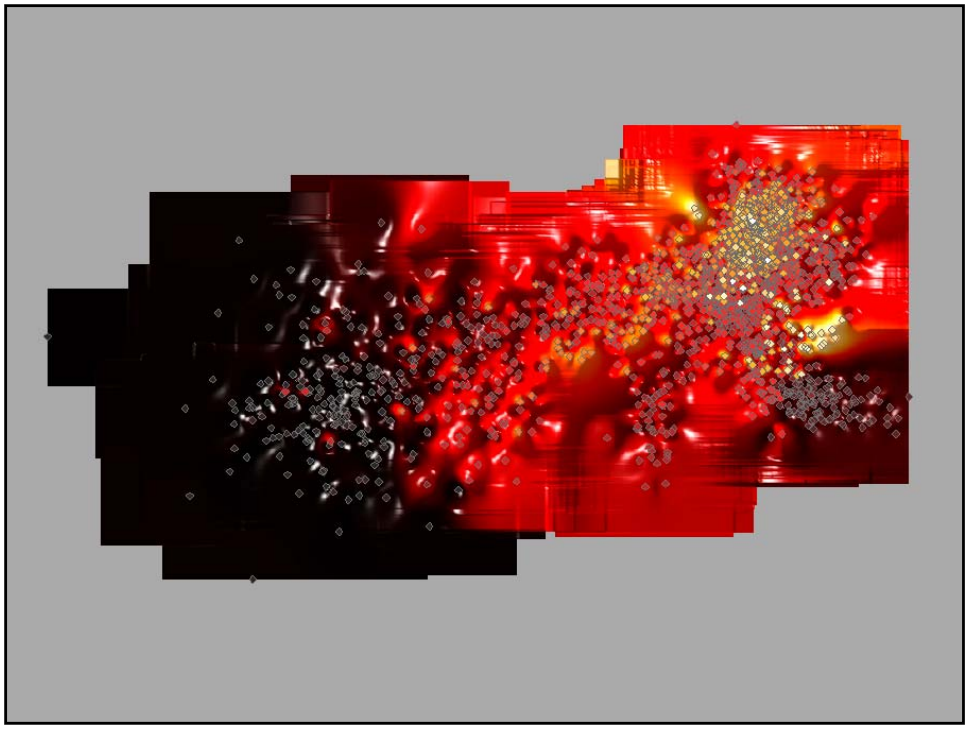
The scatterplot tool in CrystalFp implements a **Force Directed Placement** multidimensional scaling algorithm (here the points are colored by energy)

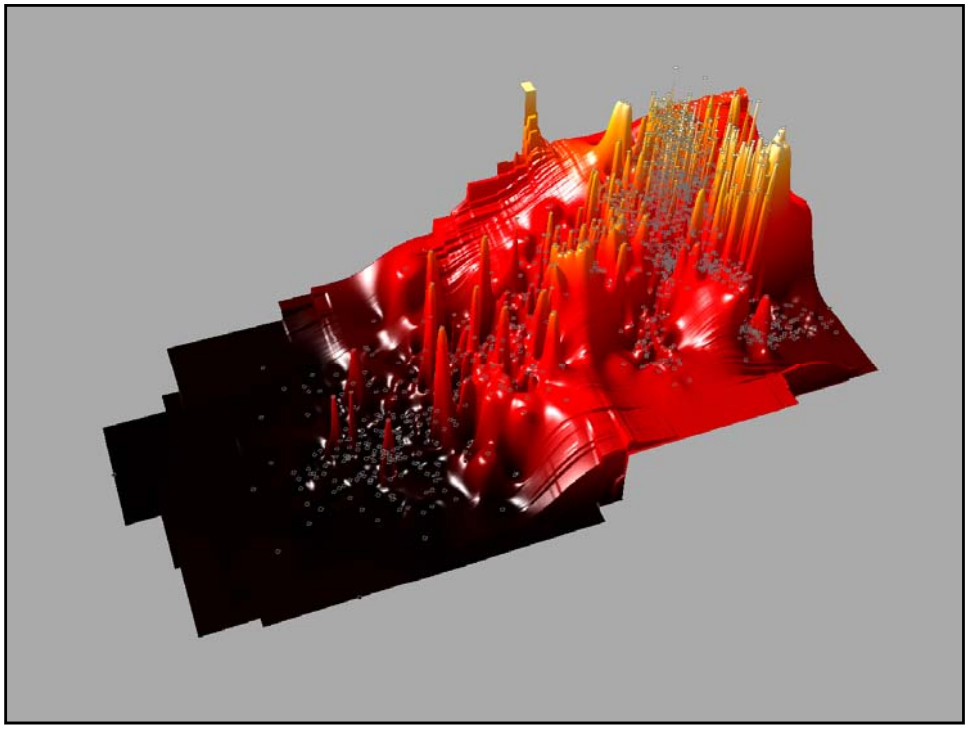


Nanocluster data from Dr. Gareth Tribello (USI)

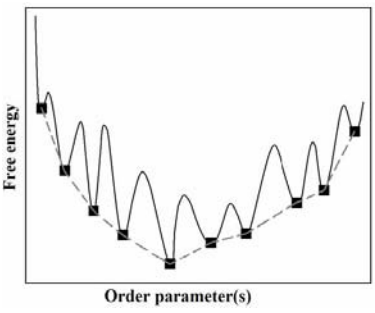
ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich
CrystalFp, STM4 & USPEX - Mario Valle - 04/08/2011
CSCS Swiss National Supercomputing Centre

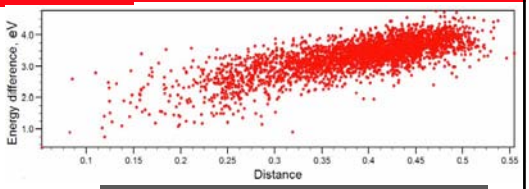




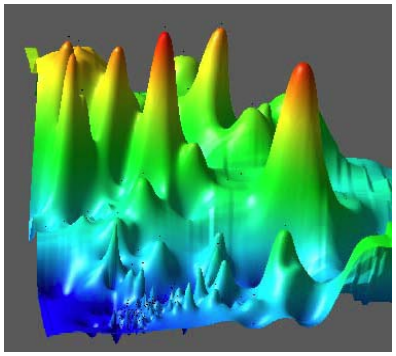


Study of energy landscapes



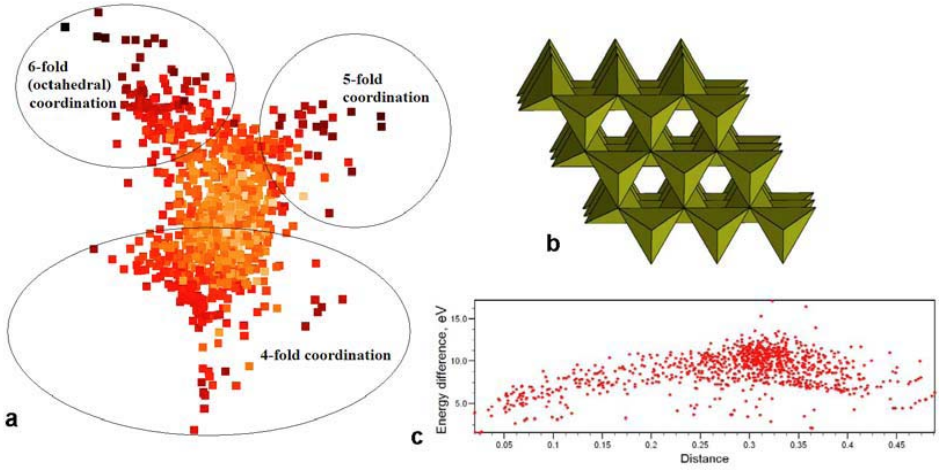


A. R. Oganov and M. Valle,
How to quantify energy landscapes of solids,
The Journal of Chemical Physics, vol. 130,
p. 104504, 2009.



Energy landscape of Au_8Pd_4 system

More complex landscapes

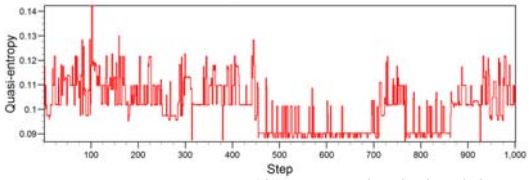


Energy landscape for MgO with 32 atoms/cell

New quantities: quasi-entropy

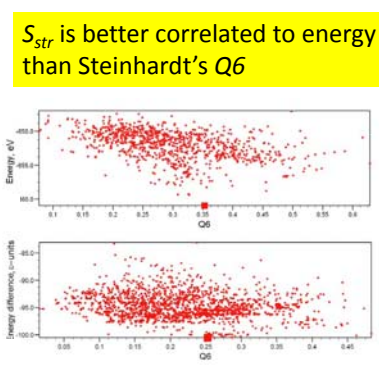
For each given structure, **quasi-entropy** is a measure of disorder and complexity of that structure.

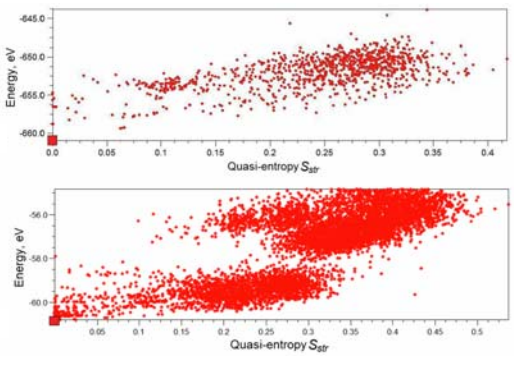
$$S_{\mathrm{str}} = -\sum_{A} \frac{N_A}{N_{\mathrm{cell}}} \langle \ln(1 - D_{A_i A_j}) \rangle$$



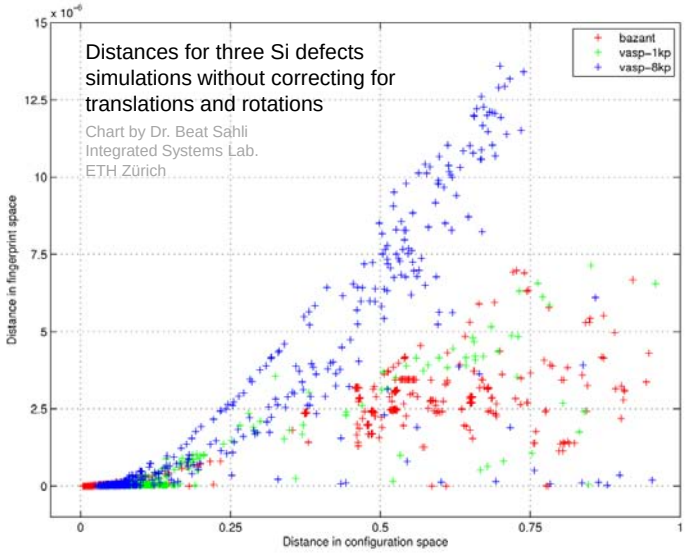
Si structures developing defects

S_{str} is better correlated to energy than Steinhardt's *Q6*





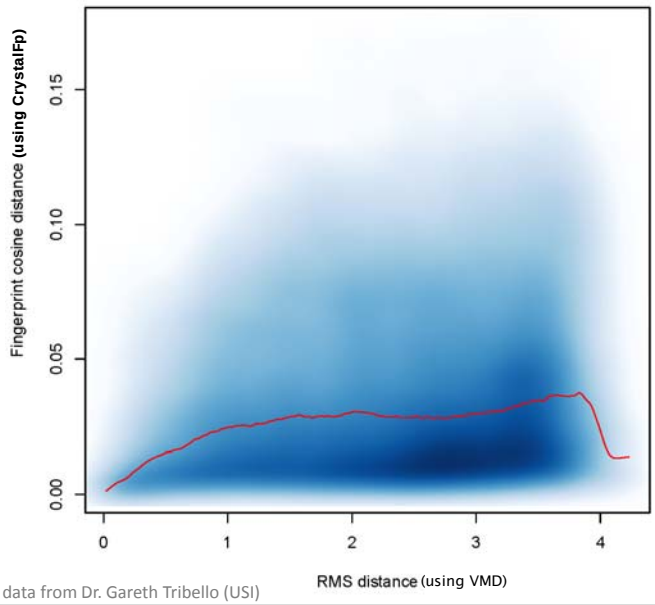
Faithful representation?



Distances for three Si defects simulations without correcting for translations and rotations

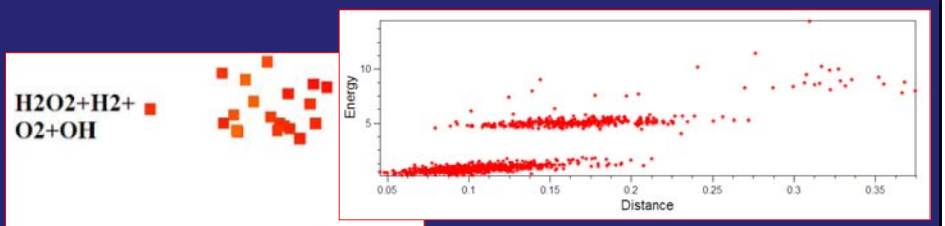
Chart by Dr. Beat Sahli
Integrated Systems Lab.
ETH Zürich

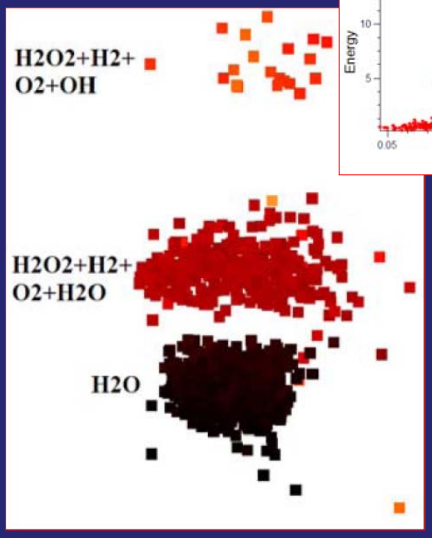
On average (on an interval), yes



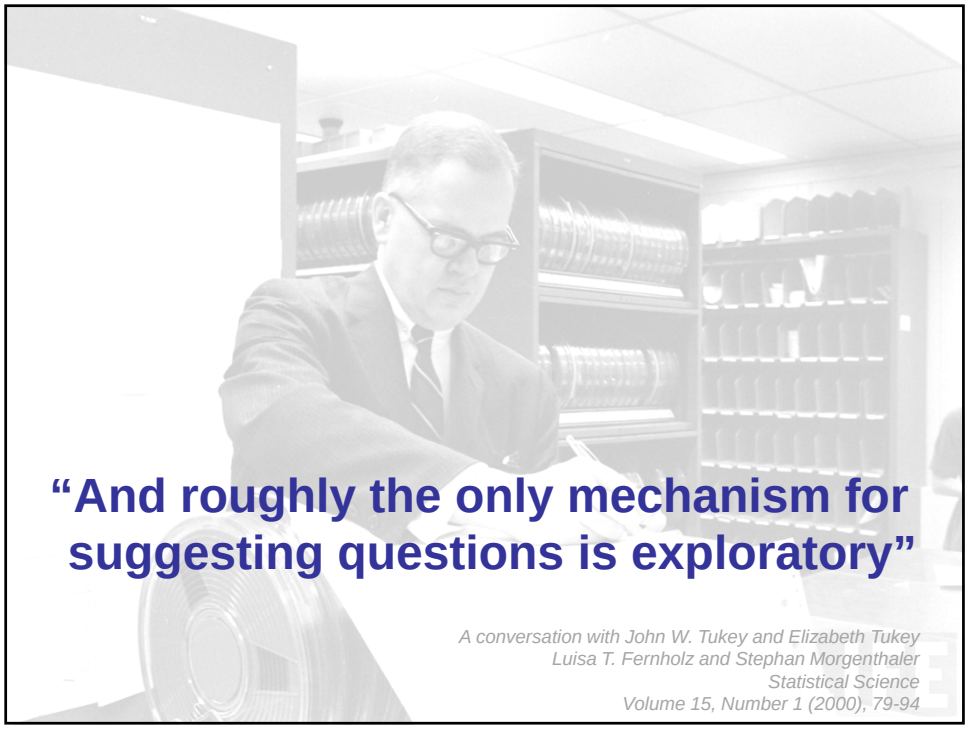
Nanocluster data from Dr. Gareth Tribello (USI)

(Totally) unexpected correlations





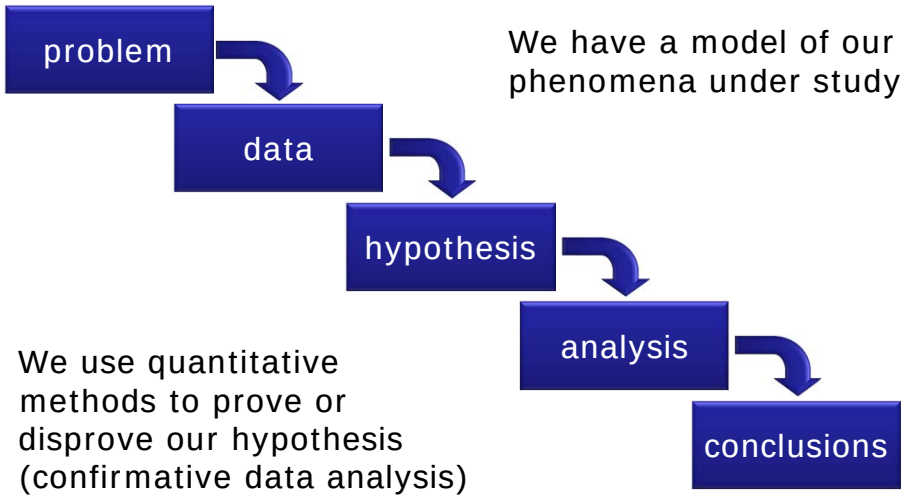
- We found unexpected correlations between distance and other physical variables
- For example the deceptively simple H_2O shows clear correlations and grouping
- This and other datasets motivated us to continue the exploration of the crystal fingerprints' space…



"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

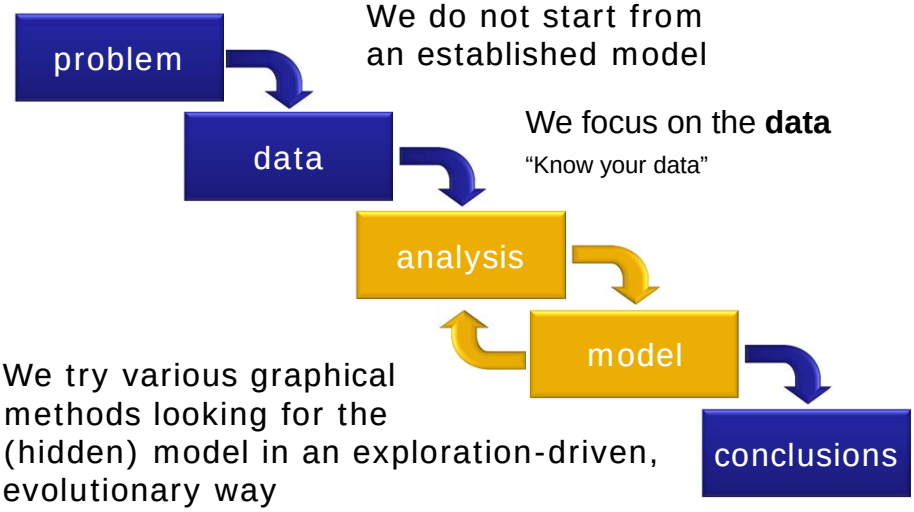
Classical data analysis



We have a model of our phenomena under study

We use quantitative methods to prove or disprove our hypothesis (confirmative data analysis)

Exploratory data analysis

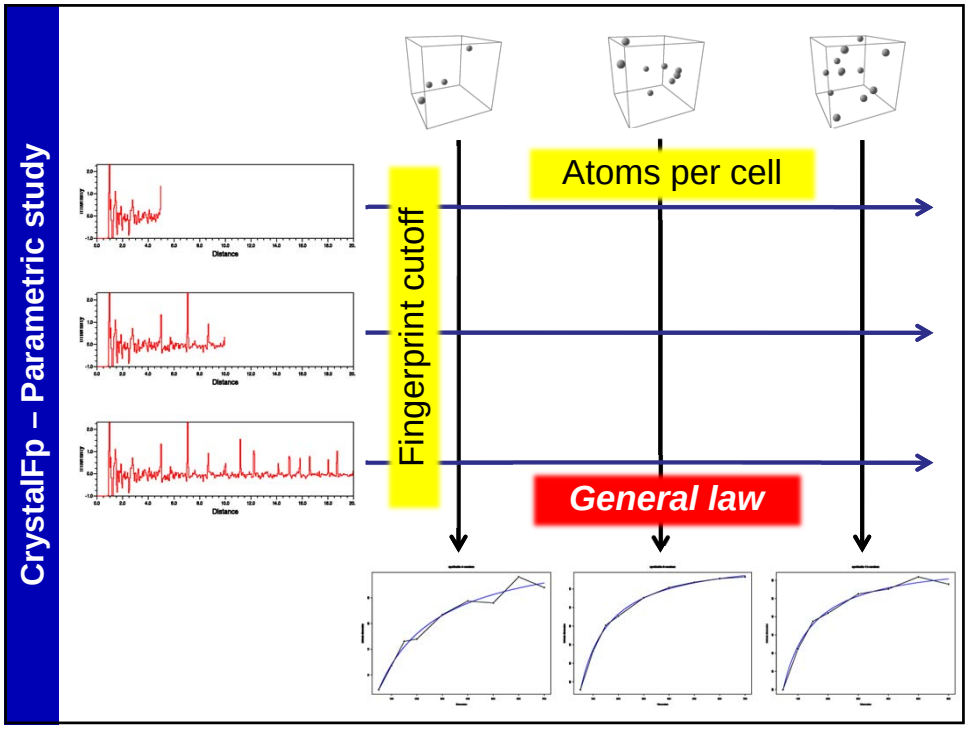


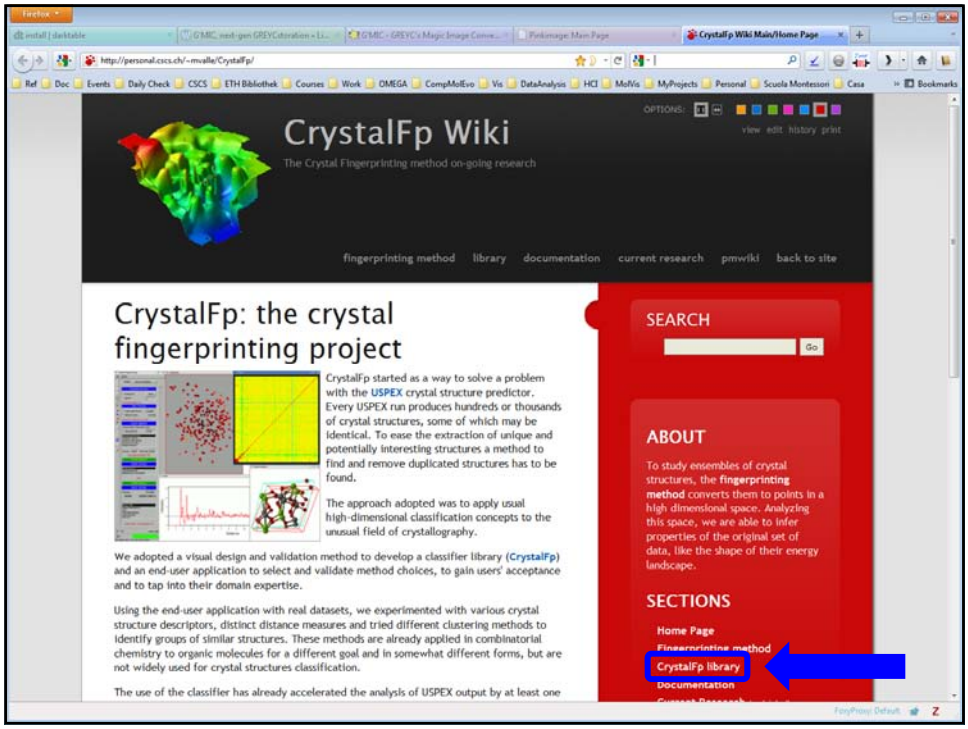
We do not start from an established model

We focus on the **data**

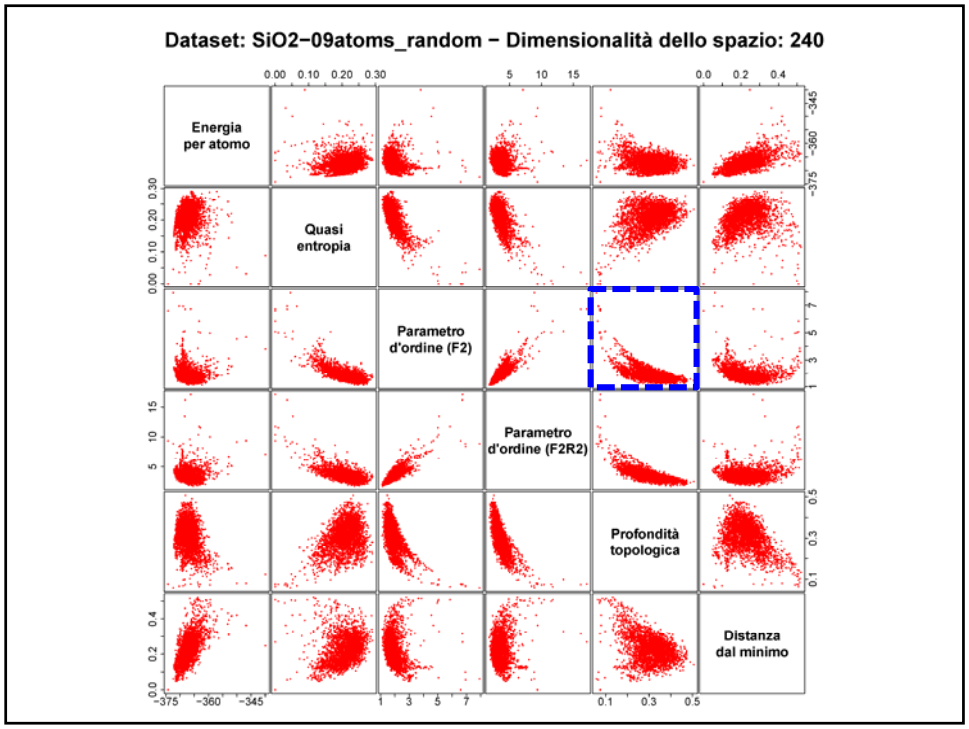
"Know your data"

We try various graphical methods looking for the (hidden) model in an exploration-driven, evolutionary way

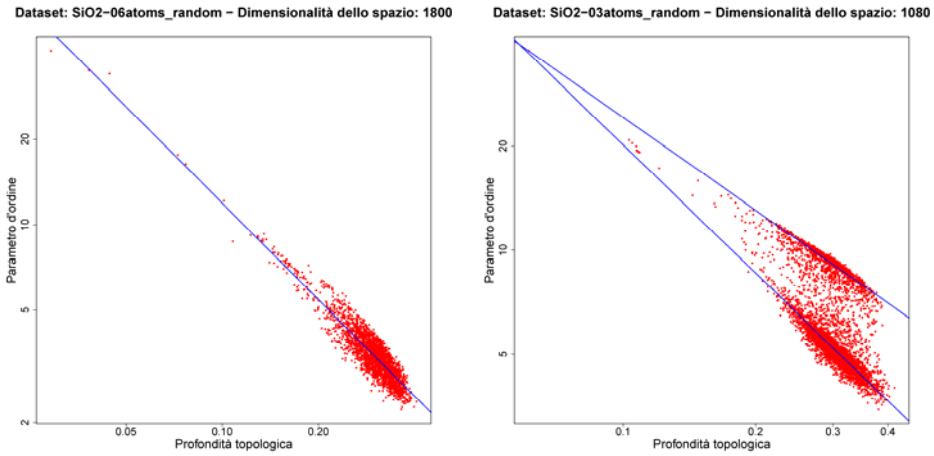

CrystalFp – Parametric study
Atoms per cell
Fingerprint cutoff
General law


CrystalFp Wiki
The Crystal Fingerprinting method on-going research
fingerprinting method
library
documentation
current research
pmwiki
back to site
CrystalFp: the crystal fingerprinting project
SEARCH
Go
ABOUT
SECTIONS
Home Page
CrystalFp library
Documentation

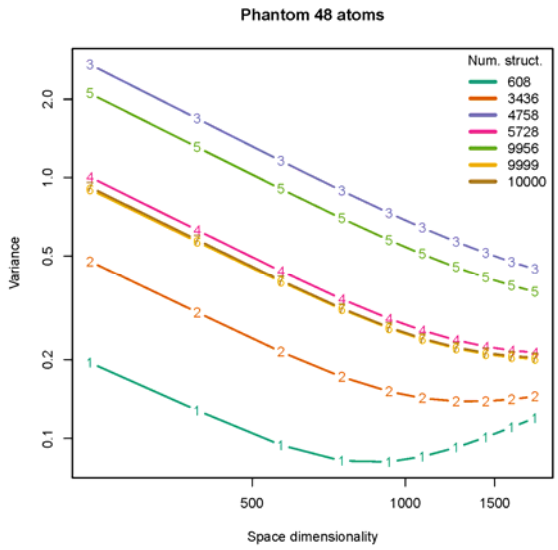
```
Usage:
    CrystalFp [options] POSCARfile [ENERGIESfile]

-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
 ...
```



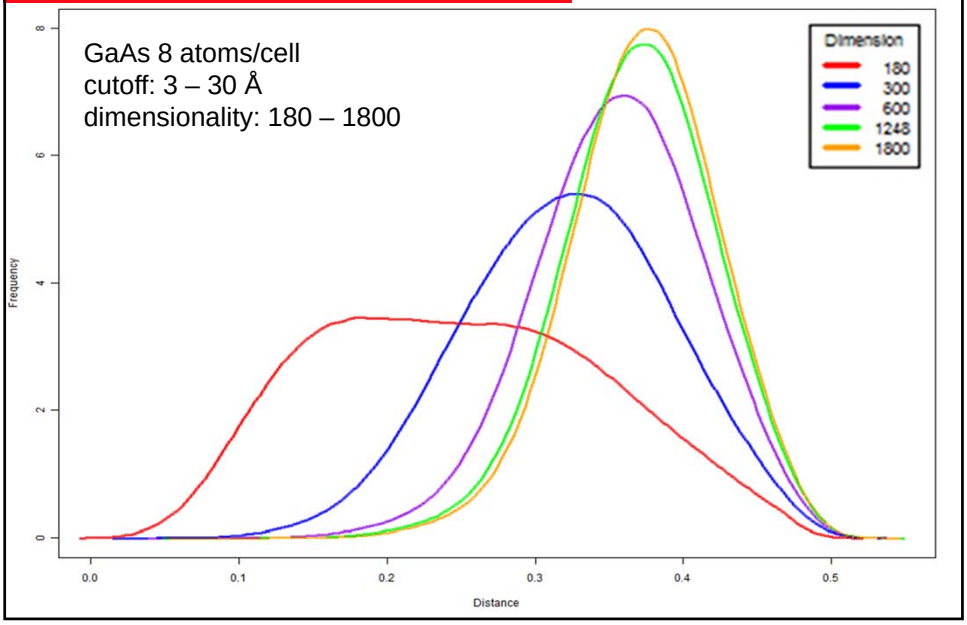
Interesting correlations



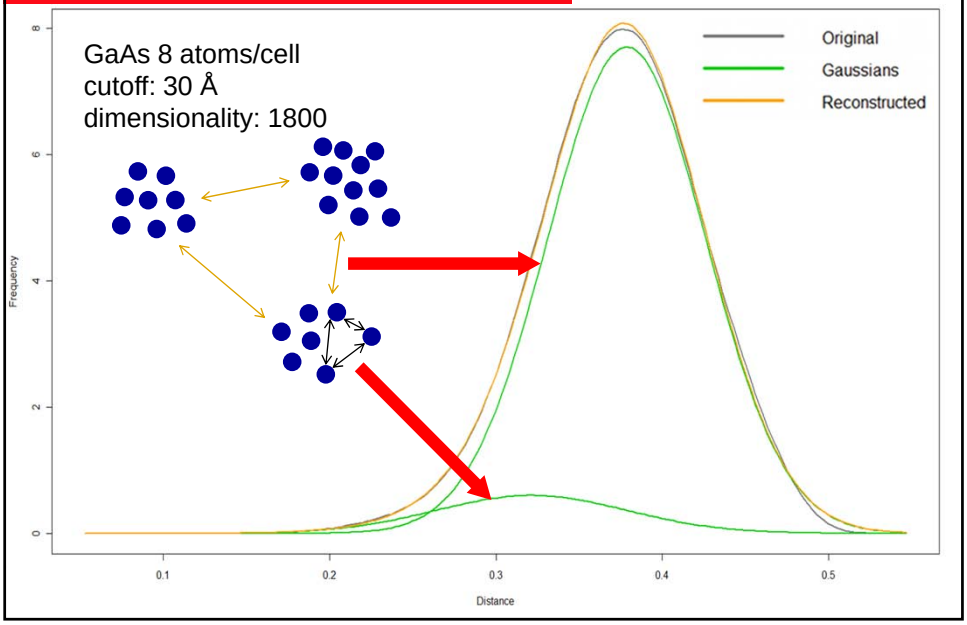
Searching an explanation



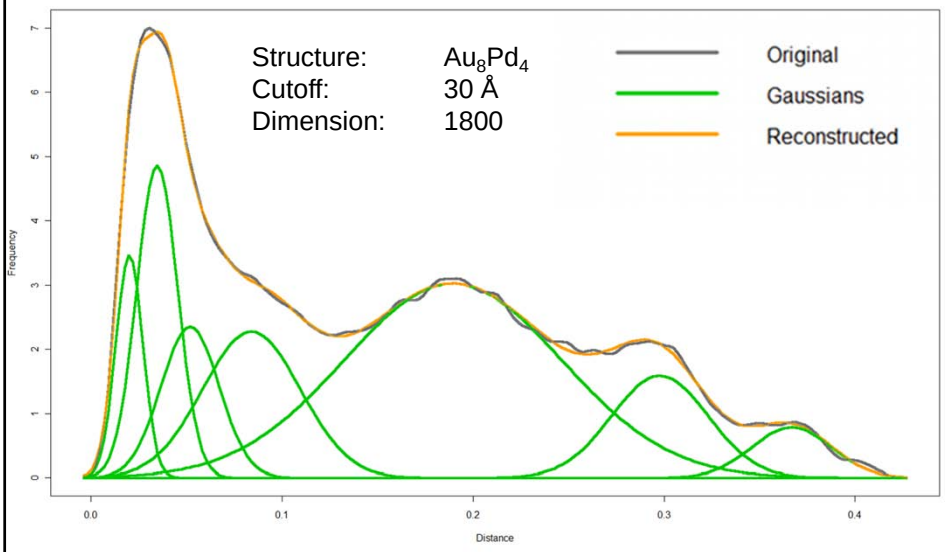
Distance vs. dimensionality



Distance decomposition



But this one?

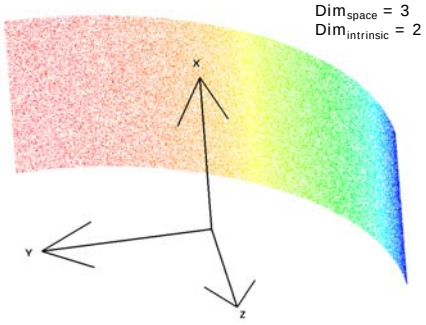


Intrinsic dimensionality

Fingerprint space dimensionality: 100 ÷ 3000

Vastly redundant!

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



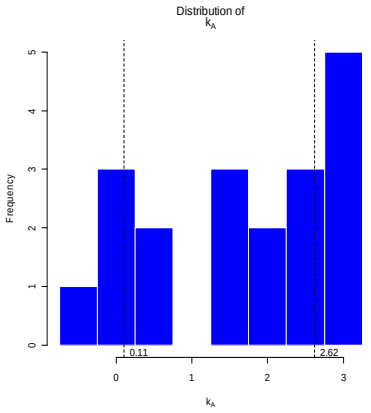
Au_8Pd_4	39	10.85
MgNH	39	32.47
MgO	99	11.62

More realistic theory:
$Dim_{intrinsic} = 3*N_{atoms}+3-K$

Constraints per atom

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

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

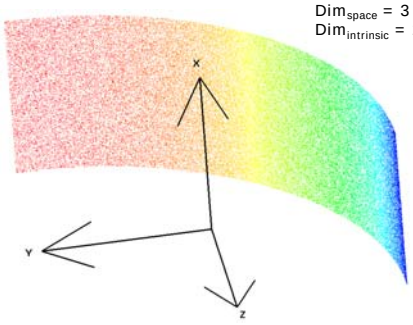
Fingerprint space dimensionality: 100 ÷ 3000

Vastly redundant!

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

More realistic theory: $Dim_{intrinsic} = 3*N_{atoms}+3-K$

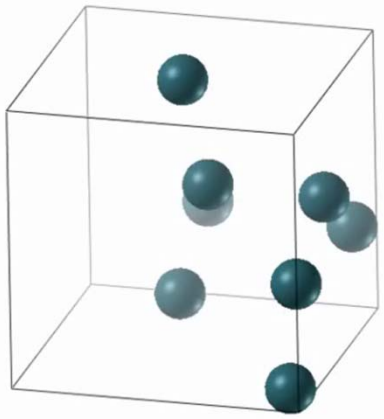
$Dim_{space} = 3$
$Dim_{intrinsic} = 2$



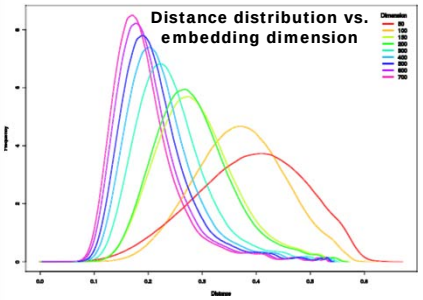
Au_8Pd_4	39	10.85
MgNH	39	32.47
MgO	99	11.62
H_2O	**39**	**80.50**

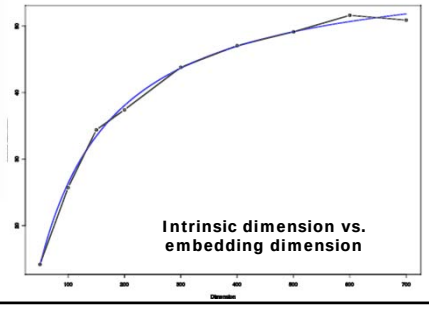
But how do you explain this?

Synthetic datasets



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





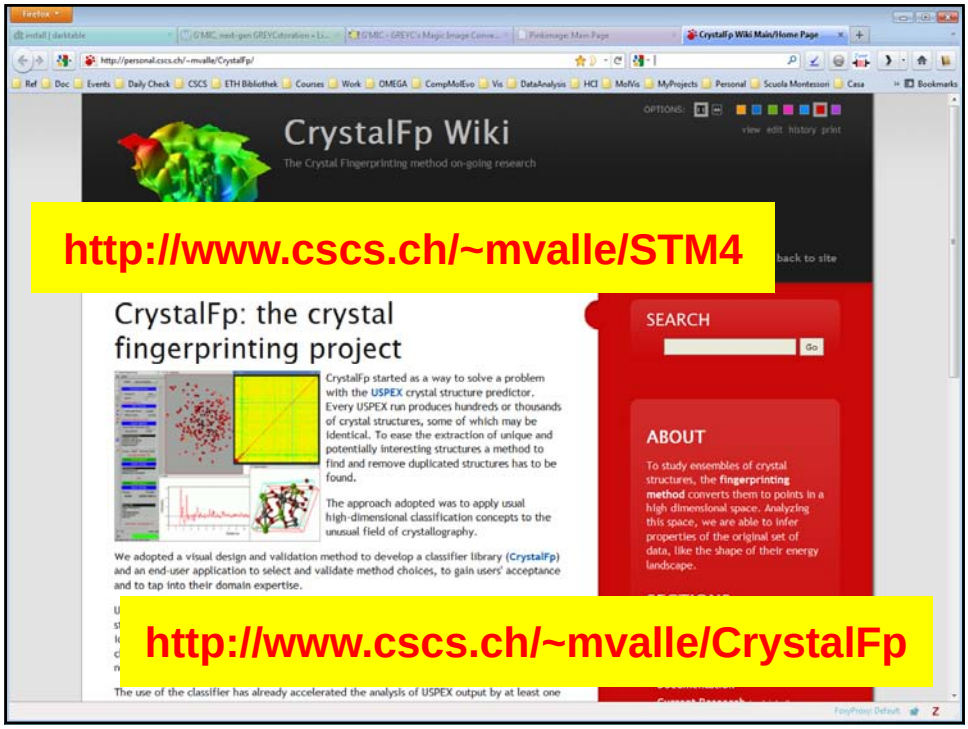
Lessons learned

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


CrystalFp Wiki
The Crystal Fingerprinting method on-going research
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.
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.
http://www.cscs.ch/~mvalle/CrystalFp
The use of the classifier has already accelerated the analysis of USPEX output by at least one


Thank you
for your attention!
And don't forget: mvalle@cscs.ch