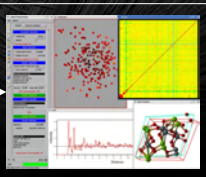
# Crystal Structures Classifier for an Evolutionary Algorithm Structure Predictor

Mario Valle – Swiss National Supercomputing Centre (CSCS)
Artem Oganov – ETH Zürich

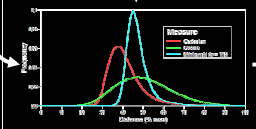
## Two parallel stories

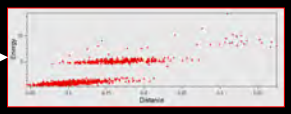
One talk, two stories

The original problem



The visual analytics story



The modeling story

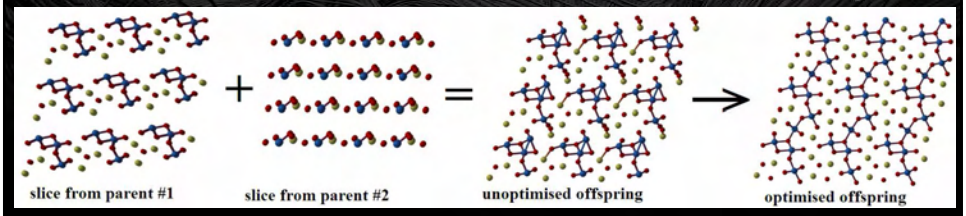
*Thanks to:*
- *ETH Zürich*
- *Swiss National Supercomputing Centre (CSCS)*
- *Joint Russian Supercomputer Centre (Russian Academy of Sciences)*

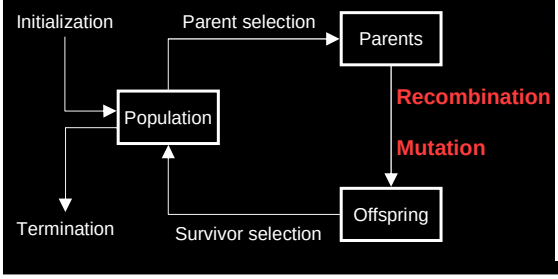
## Crystal structure prediction: major unsolved problem

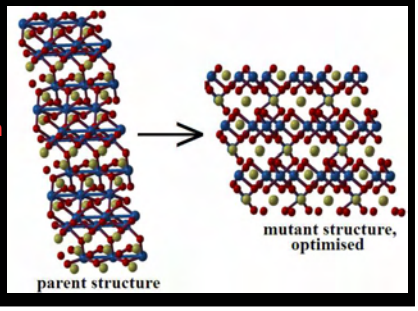
> One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition. Who, for example, would guess that graphite, not diamond, is the thermodynamically stable allotrope of carbon at ordinary temperature and pressure? Solids such as crystalline water (ice) are still thought to lie beyond mortals' ken.
>
> **J. Maddox, editorial in *Nature* (1988)**

- **Prediction** of the **stable crystal structure** on the basis of only the chemical composition is one of the central problems of condensed matter physics, which for a long time remained **unsolved**.
- The ability to solve this problem would open new ways also for the understanding of the behaviour of materials.

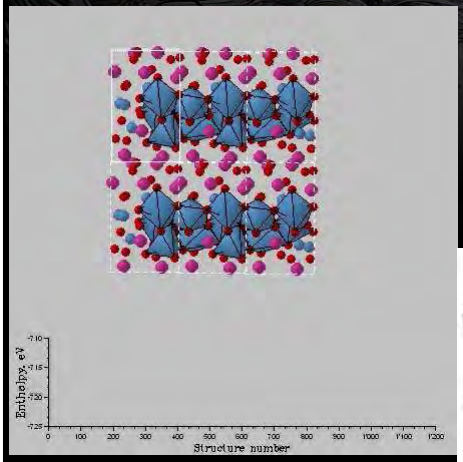
## USPEX an evolutionary algorithm and system for crystal structure prediction



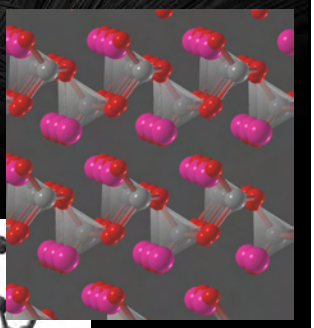
## Examples of USPEX predictions

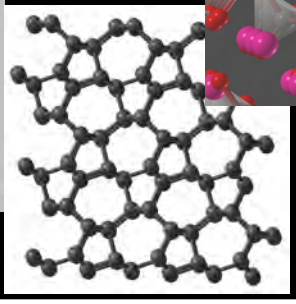



40-atom cell of $MgSiO_3$ post-perovskite

From: http://olivine.ethz.ch/~artem/USPEX.html

Novel high pressure phases of $CaCO_3$

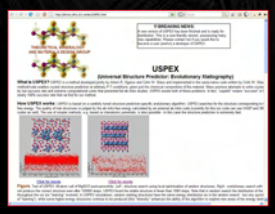




Low-energy 3D carbon structure

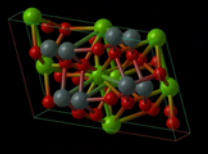
## The problem to solve

USPEX is a crystal structure predictor based on an evolutionary algorithm




Each run produces hundred of putative crystal structures…

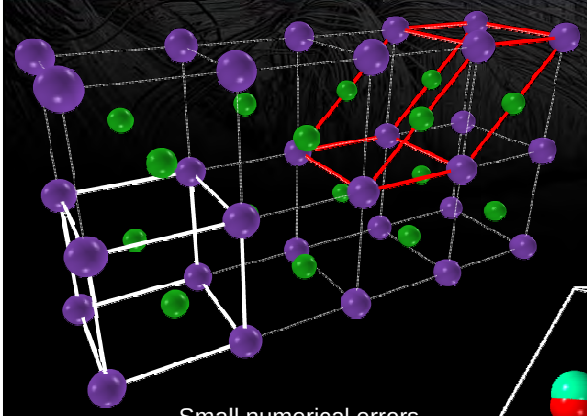
…but many of them are equal



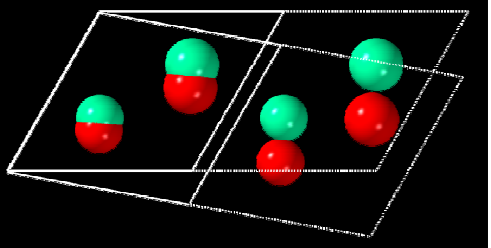
So an intensive manual labor is needed to prune duplicated structures

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

## Comparison problems for crystal structures

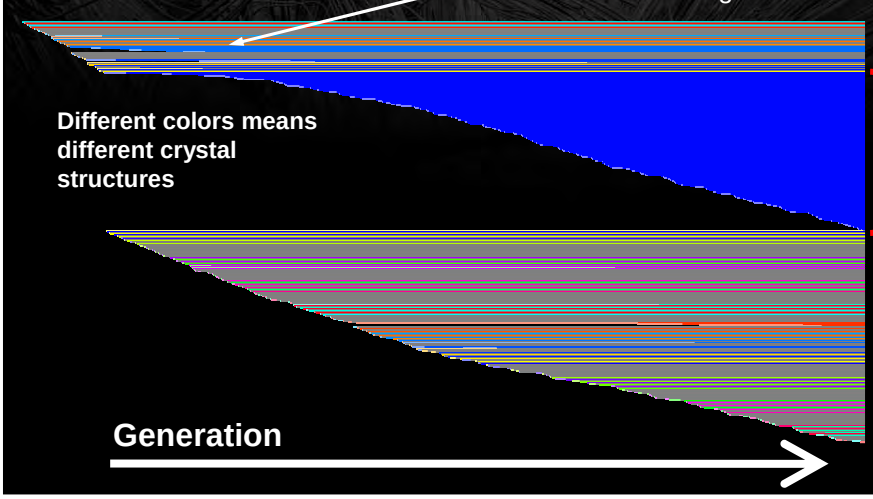


More than one unit cell could describe the same crystal structure

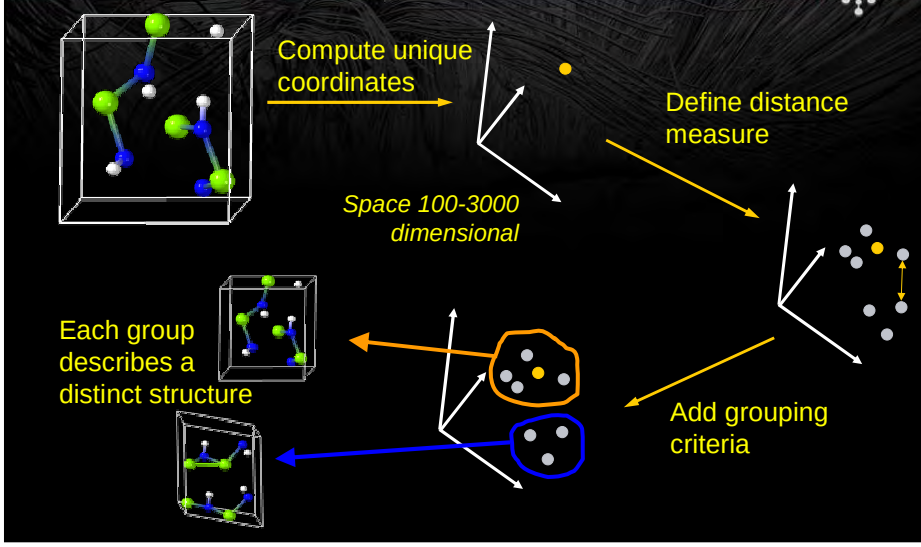


Small numerical errors make structures diverge when move away from base unit cell

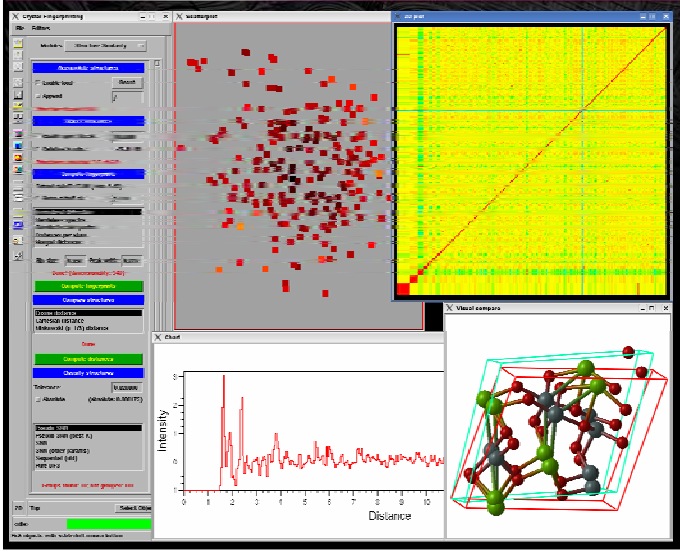
## The USPEX problem (but common to all evolutionary algorithms)



## Proposed solution: use methods and ideas from multidimensional spaces




## Visual design and validation support

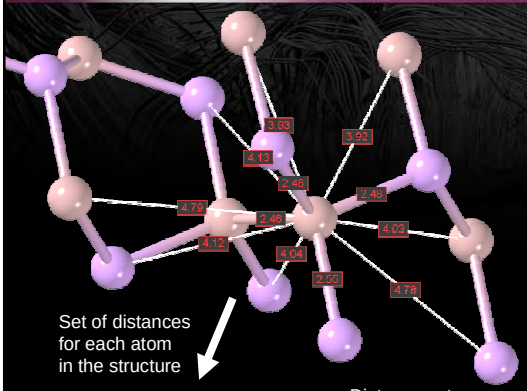



- Built a tool to explore algorithm choices and parameters settings
- This tool wraps the classifier library and provides various interactive visual diagnostics to check classifier behavior
- It is built inside STM4, the molecular visualization toolkit developed at CSCS

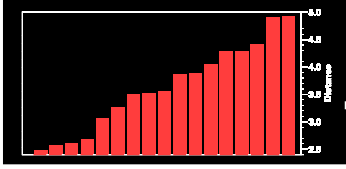
**Why this approach?**

- We had to win user support and confidence
- It supports experimentation for library design
- It provides at no cost the tool to select and remove identical structures

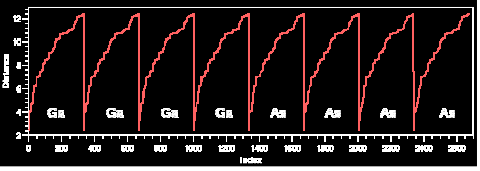
## Structure coordinates (fingerprint) from interatomic distances




Set of distances for each atom in the structure



Distance sets concatenated for all atoms in the structure




Coordinates based on interatomic distances are independent from:

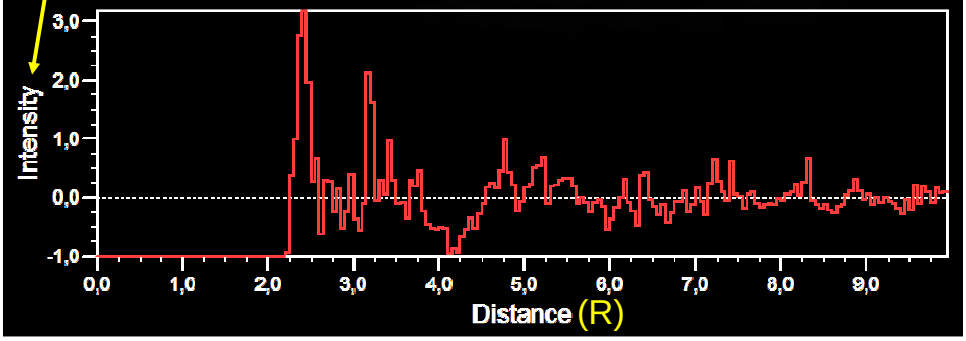
1. Translation and rotation of the structure
2. Choice of unit cell among equivalent unit cells
3. Ordering of cell axis and atoms in the cell
4. Inversion and mirroring of the structure.

## A better domain based choice: the pseudo-diffraction fingerprint

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

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



## Experimented with various types of distance measure

- Classical Euclidean distance

$$dist_{euclidean}(i,j) = \sqrt{\left(\sum_k |fp_{i_k} - fp_{j_k}|^2\right)}$$

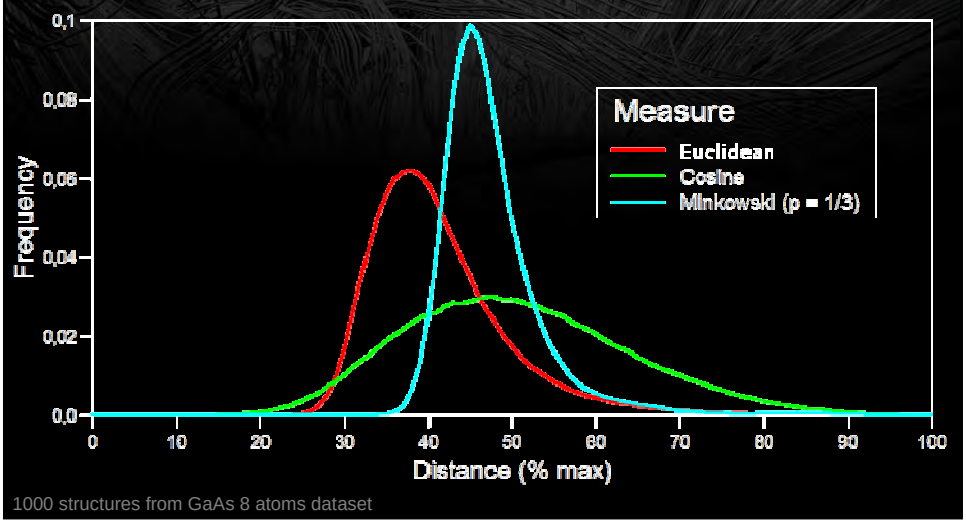
- Minkowski distance (with p = $^1/_3$)

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

- Cosine distance

$$dist_{cosine}(i,j) = \frac{1}{2}\left(1 - \frac{\overrightarrow{FP_i} \cdot \overrightarrow{FP_j}}{\|\overrightarrow{FP_i}\| \, \|\overrightarrow{FP_j}\|}\right)$$

## Goal: to have better relative contrast (spread) for distances



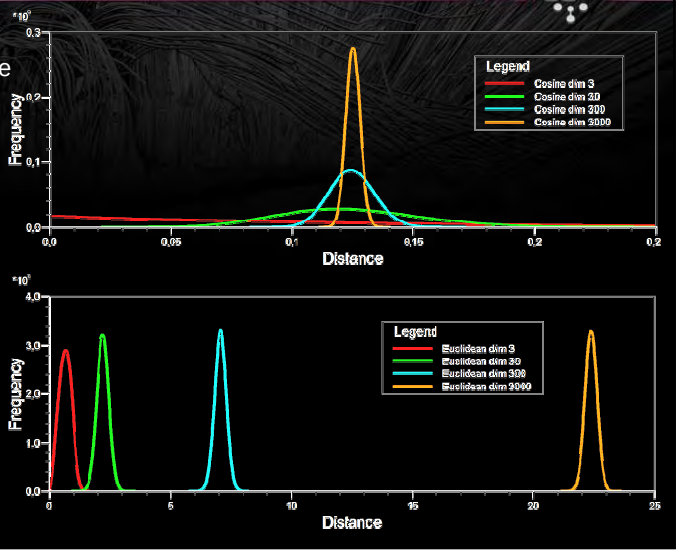

1000 structures from GaAs 8 atoms dataset

## Cosine and Euclidean distances give different relative contrast

Relative contrast is higher for cosine distance (here from a synthetic dataset of uniformly distributed points in the unit hypercube)

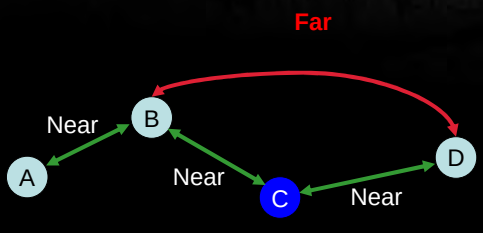
Relative contrast is estimated from Gaussian fit of the peaks by: `mean/FWHM`

| Dim. | Cos. | Eucl. |
|---|---|---|
| 30 | 0.520 | 0.259 |
| 300 | 0.172 | 0.080 |
| 3000 | 0.055 | 0.025 |

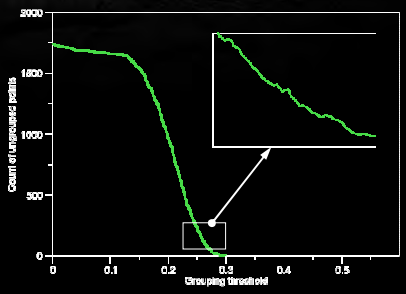



## Grouping challenges

For the cosine distance the triangle inequality $D_{bd} \leq D_{bc}+D_{cd}$ does not hold

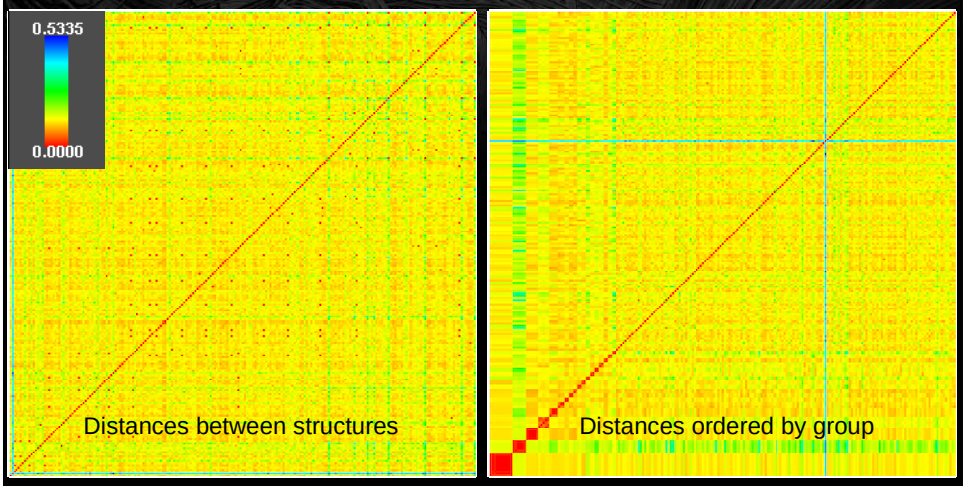



Which is the correct grouping?
ABC + D or AB + CD or ABCD

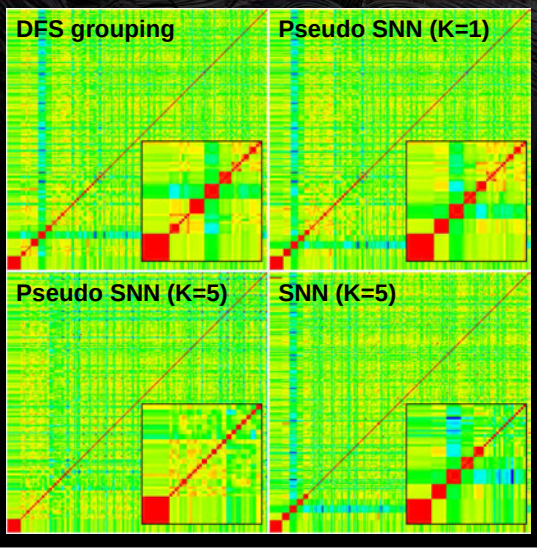


The problem is made visible with one of the diagnostic charts

## Visual diagnostics: distance matrix and clustering




## Visual diagnostic of the clustering algorithms




**DFS**: Deep first search of the neighbors nodes

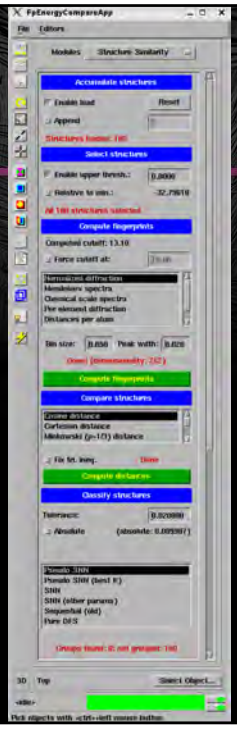
**Pseudo SNN**: Maintain connection between nodes only if they share at least K neighbors

**SNN**: As above plus a DBSCAN pass

## Access to all CrystalFp parameters

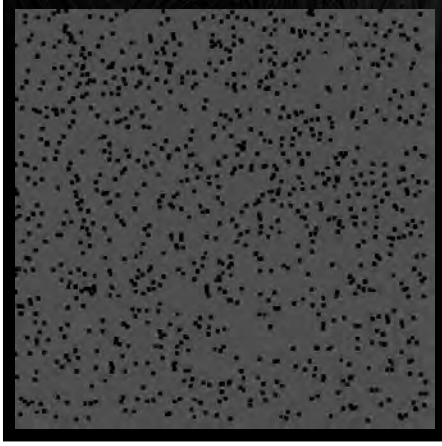
- The End User application makes possible the choice of algorithms and their parameters manipulation in a clear process workflow

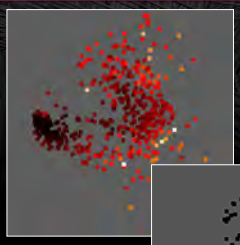
1. Load structures
2. Filter on energy
3. Compute fingerprints
4. Compute distances
5. Group structures



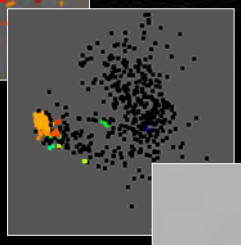
## Visual diagnostics: scatterplot

The scatterplot tries to map High-D space points to 2D preserving their relative distances

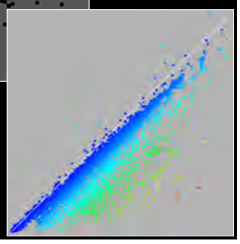




Colored by "stress" to detect local minima traps

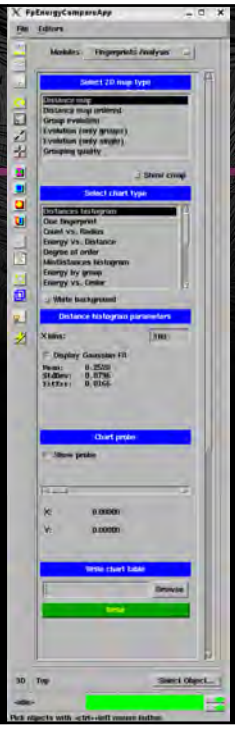


Colored by group



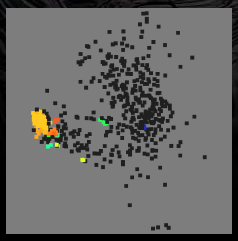
Diagnostic chart: distances in 2D vs. distances in High-D space

## Various visual diagnostics tools

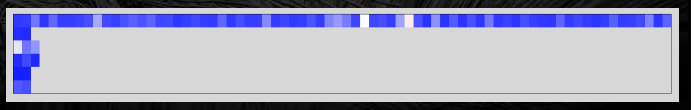


1. 2D maps
2. Charts
3. Picking for details
4. 2D data export

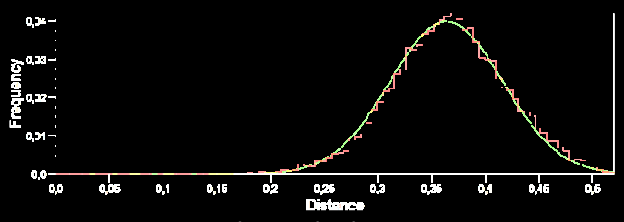
## Visual diagnostic tools



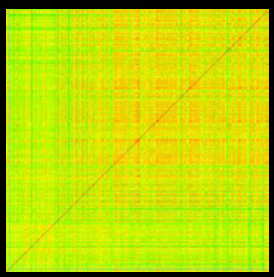
Scatterplot



Grouping quality: silhouette coefficients




Diagnostic charts

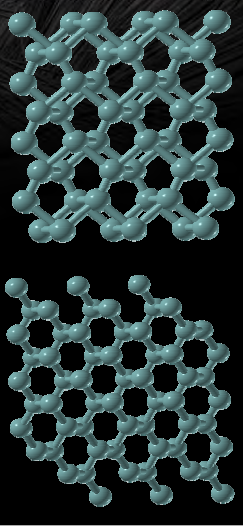


Distance matrix

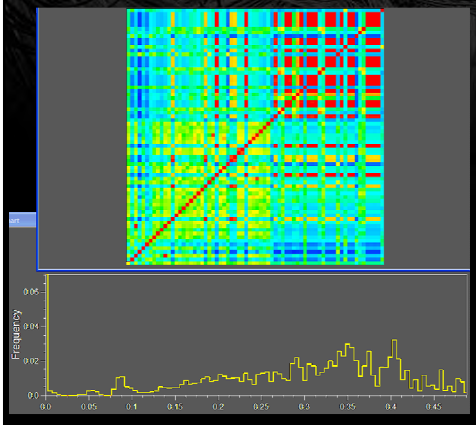
## USPEX problem solved: An example

**Hydrogen at 600 GPa (16 atoms)**

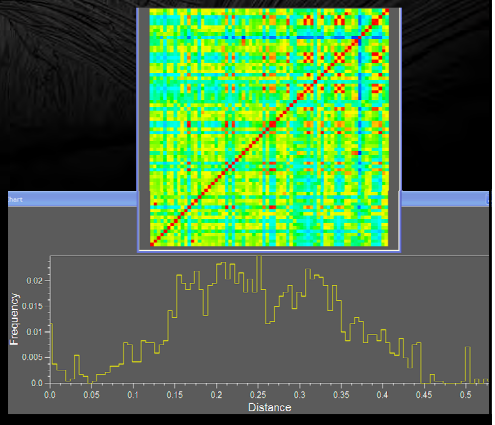
- The USPEX run produced **1274** structures
- From these the **794** within 0.5 eV from the lowest energy value found are selected
- Manual analysis to remove duplicated structures from this set: **2-20h** of work
- Using the CrystalFp classifier: **~10min**
- At the end found only **4** unique structures:
  - One α-Ga type (top)
  - One Cs-IV (bottom), the ground state (i.e. the lower energy structure), and two closely related structures



## The visual analytics story has an happy end…

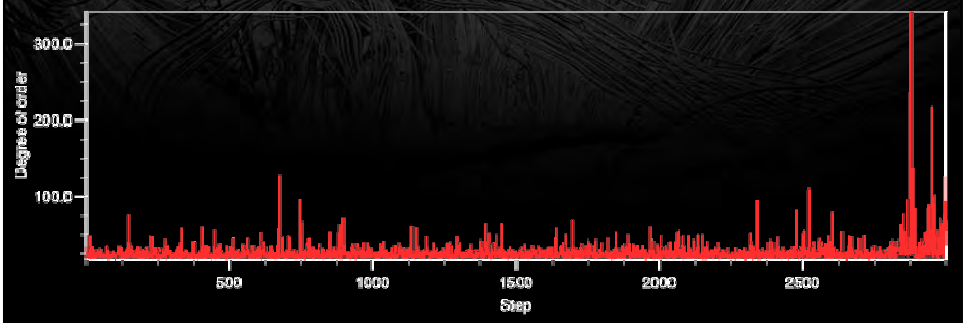



Original USPEX.
A lot of identical structures



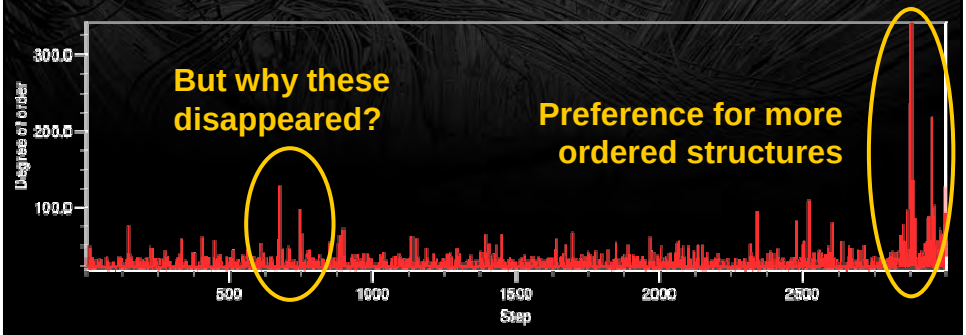

USPEX after the classifier integration.
No more “structure cancer”

## New visual analysis tools



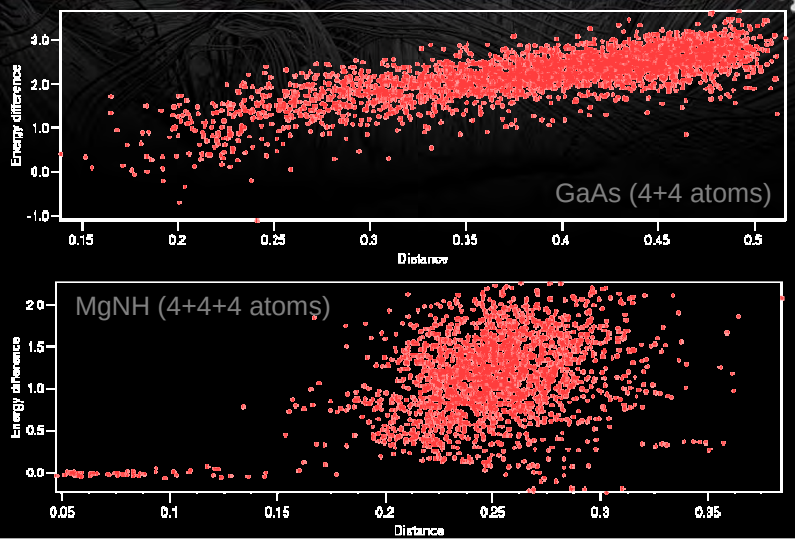

Other derived quantities, that are not strictly needed for validation, but provided useful insight on USPEX behavior, are obtained almost for free from our multidimensional approach

## (Somehow) unexpected phenomena

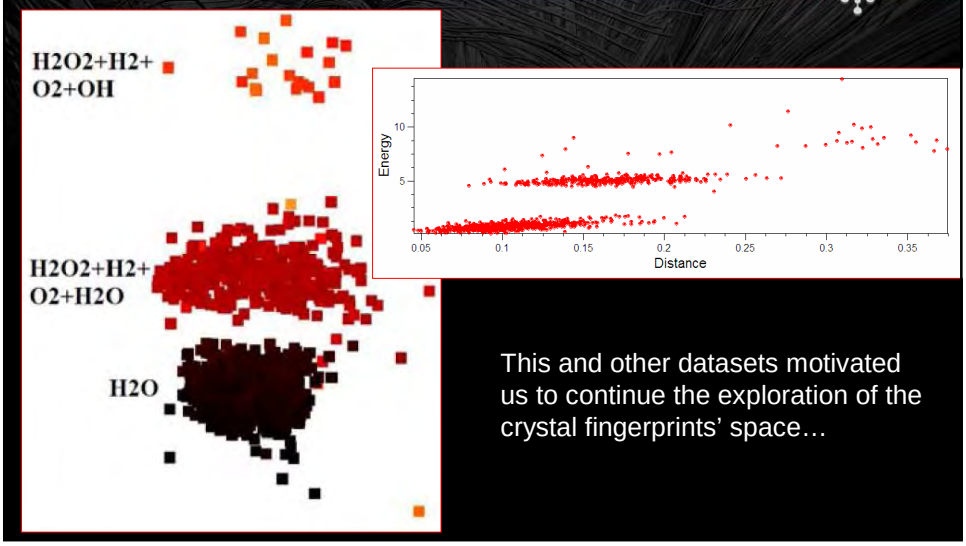



Latest generation has lower energy than previous ones. Normally low energy implies more ordered structure.

## (Totally) unexpected correlations




## 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…

## Lessons learned

**From the Visual Analytics story**

- Quick prototyping and experimentation capabilities are critical
- No need of fancy visualizations. What are needed are visualizations tuned to the problem at hand
- Credibility and user support are critical. When gained, the user becomes a source of ideas

**From the Modeling story**

- Using known concepts in unusual contexts is a source of unexpected insights
- Discoveries happen on the boundaries of disciplines
- "Seeing is believing" and convincing

## Project pages

Source code, testing results and related material:

- **http://www.cscs.ch/~mvalle/CrystalFp**

Publications:

- A. R. Oganov, M. Valle, A. Lyakhov, Y. Ma, and Y. Xie, **Evolutionary crystal structure prediction and its applications to materials at extreme conditions,** in *Proceedings IUCr2008*, Aug. 23 - 31 2008.
- 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*, vol. 84, pp. 1-10, Dec. 2007.
- Other already submitted…


Thank you
for your attention!
BTW, I'm mvalle@cscs.ch