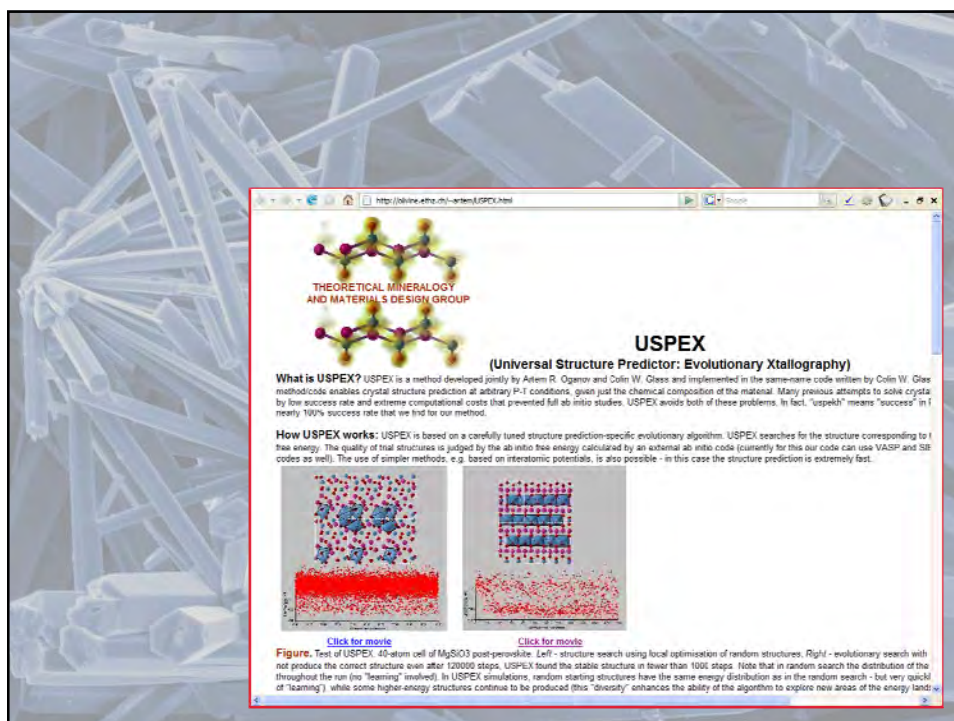





Mario Valle

CECAM Workshop – 10/7/2009

Crystal fingerprints space *a novel paradigm to study crystal structures sets*



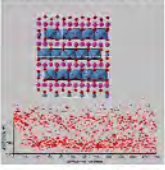
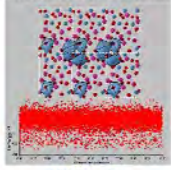

THEORETICAL MINERALOGY
AND MATERIALS DESIGN GROUP

USPEX

(Universal Structure Predictor: Evolutionary Xtallography)

What is USPEX? USPEX is a method developed jointly by Adem R. Öganov and Colin W. Glass and implemented in the same-name code written by Colin W. Glass. The code enables crystal structure prediction at arbitrary P-T conditions, given just the chemical composition of the material. Many previous attempts to solve crystals by low success rate and extreme computational costs that prevented full ab initio studies. USPEX avoids both of these problems. In fact, "uspekh" means "success" in Russian, nearly 100% success rate that we find for our method.

How USPEX works: USPEX is based on a carefully tuned structure prediction-specific evolutionary algorithm. USPEX searches for the structure corresponding to the lowest free energy. The quality of trial structures is judged by the ab initio free energy calculated by an external ab initio code (currently for this our code can use VASP and SIESTA codes as well). The use of simpler methods, e.g. based on interatomic potentials, is also possible – in this case the structure prediction is extremely fast.



[Click for movie](#) [Click for movie](#)

Figure. Test of USPEX: 40-atom cell of MgSiO_3 post-perovskite. Left – structure search using local optimization of random structures. Right – evolutionary search with no "learning" involved. In USPEX simulations, random starting structures have the same energy distribution as in the random search – but very quickly of "learning" while some higher-energy structures continue to be produced (this "diversity" enhances the ability of the algorithm to explore new areas of the energy landscape).

Use the genetic evolution idea!

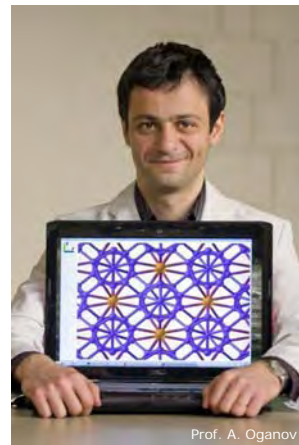
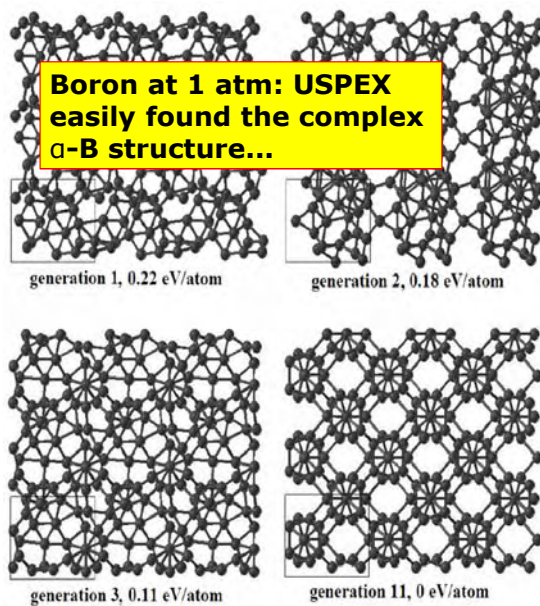


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

Mario Valle - CECAM Workshop - 10/7/2009

CSCS
Swiss National Supercomputing Centre

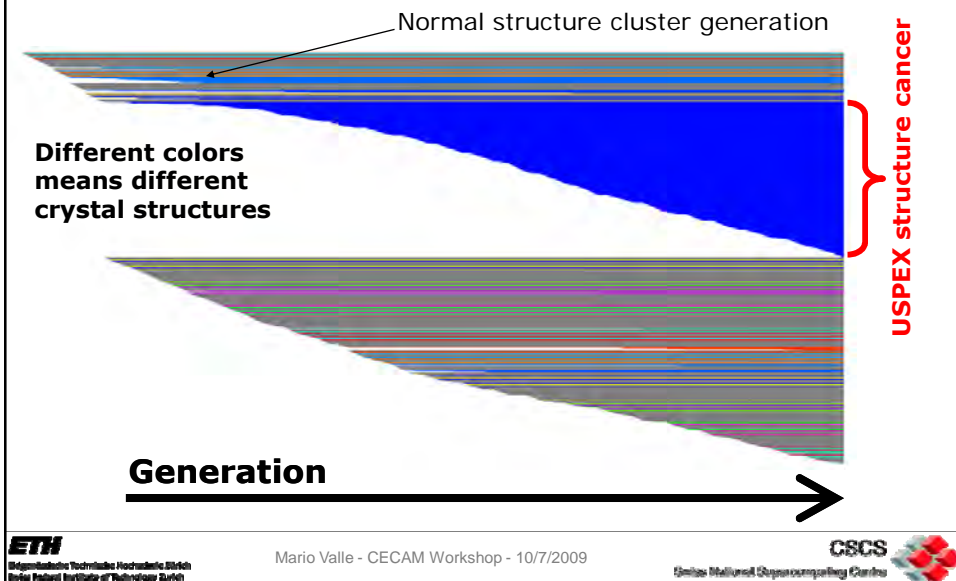
USPEX discovered new materials



Prof. A. Oganov

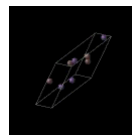
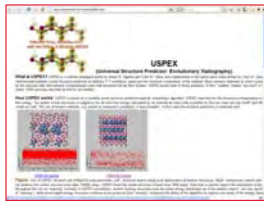
...but discovered also the superhard boron γ -B₂₈ phase (*Nature*)

USPEX structure cancer problem

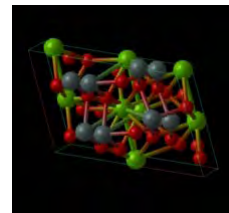


CrystalFp: the problem to solve

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

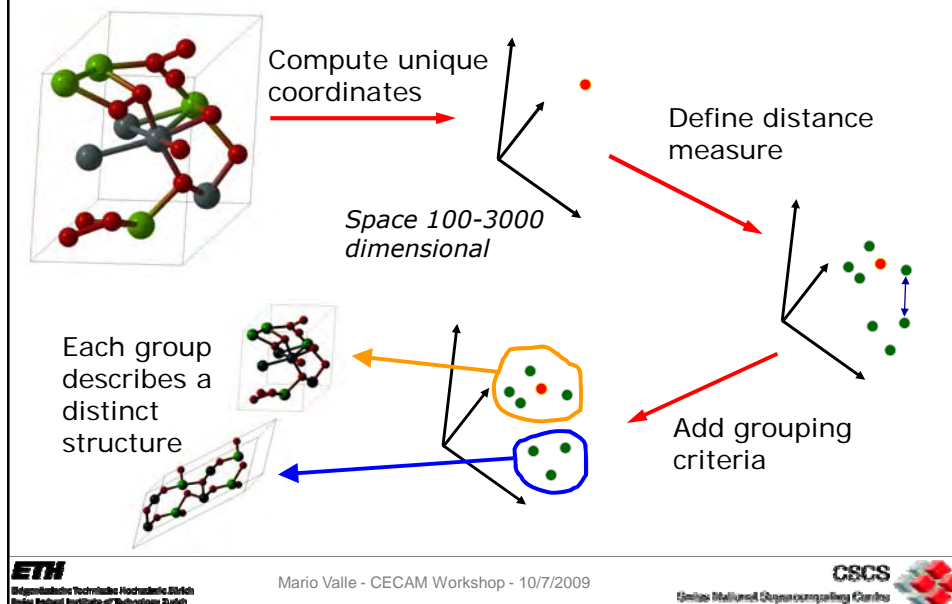


Project: develop a (semi)automatic way to extract unique structures from USPEX outputs

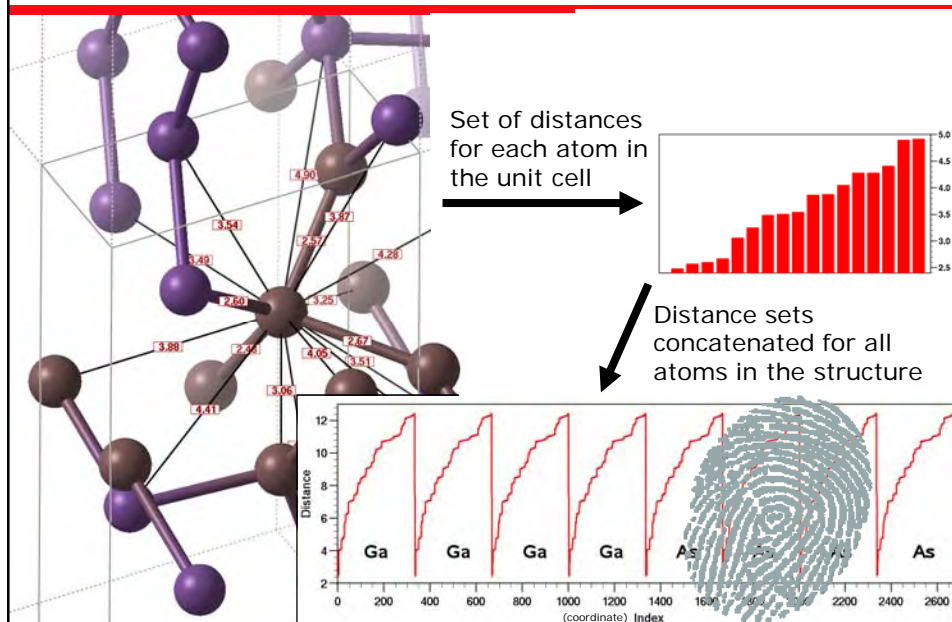


So an intensive manual labor is needed to prune duplicated structures

Proposed solution from High-Dim



Structure "coordinates" (fingerprint)

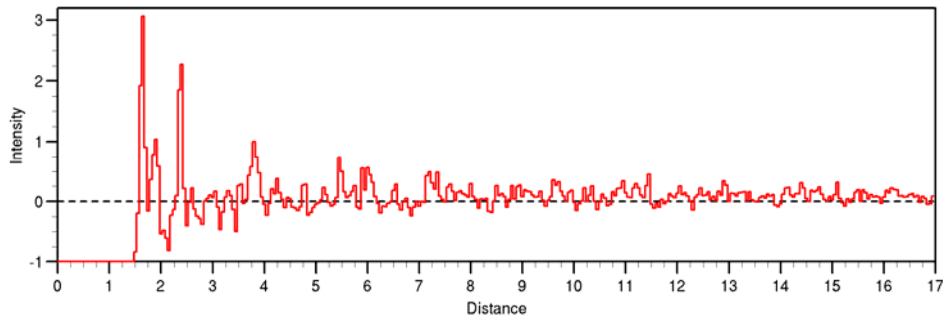


A pseudo-diffraction like method

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

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



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

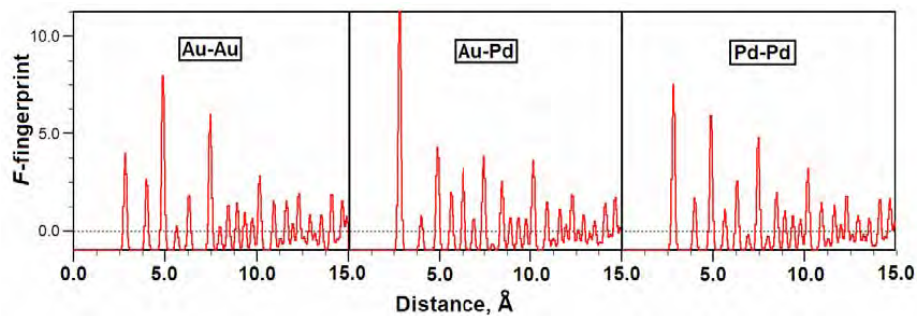
Mario Valle - CECAM Workshop - 10/7/2009

CSCS
Eidgenössische Technische Hochschule Zürich
Swiss National Supercomputing Centre

Final fingerprint (per atom type pair)

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

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

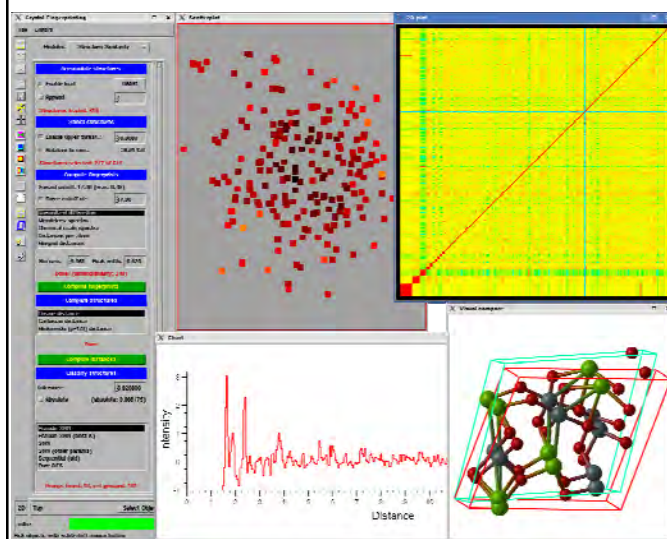


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

Mario Valle - CECAM Workshop - 10/7/2009

CSCS
Eidgenössische Technische Hochschule Zürich
Swiss National Supercomputing Centre

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

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

Mario Valle - CECAM Workshop - 10/7/2009

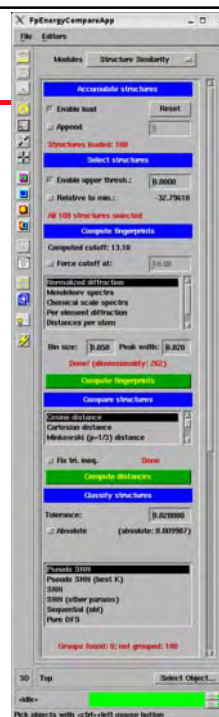
CSCS
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

STM4 interface

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

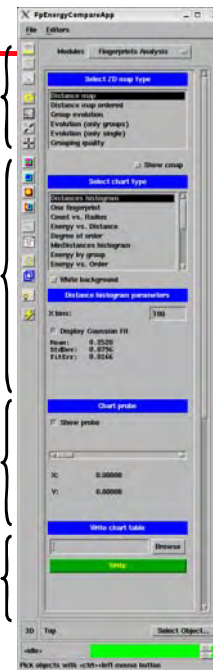
1. Load structures
2. Filter on energy
3. Compute fingerprints
4. Compute distances
5. Group structures



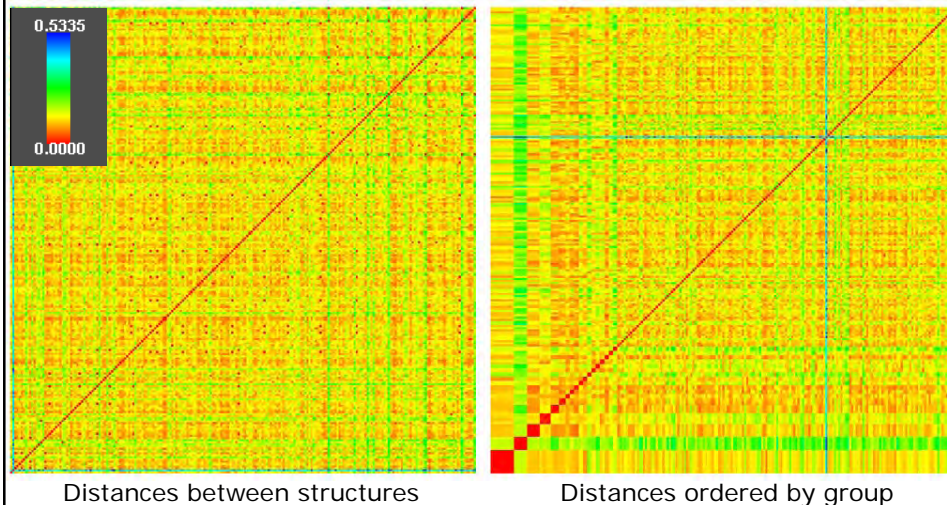
Visual diagnostics tools

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

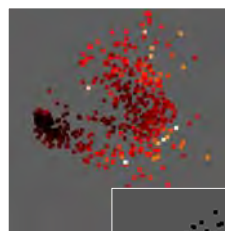
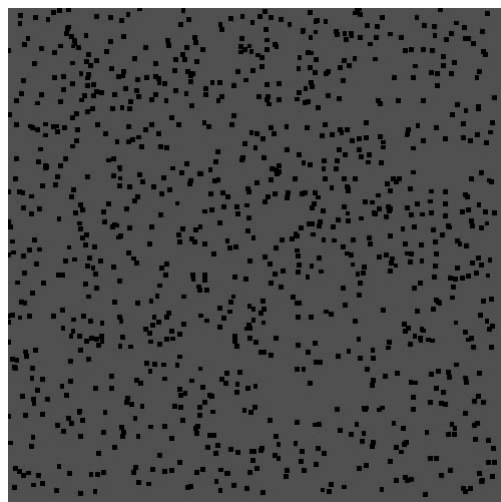


Visual diagnostics: distance matrix

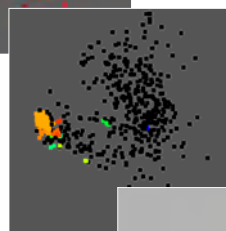


Visual diagnostics: scatterplot

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

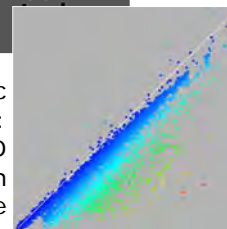


Colored by "stress" to detect local minima traps

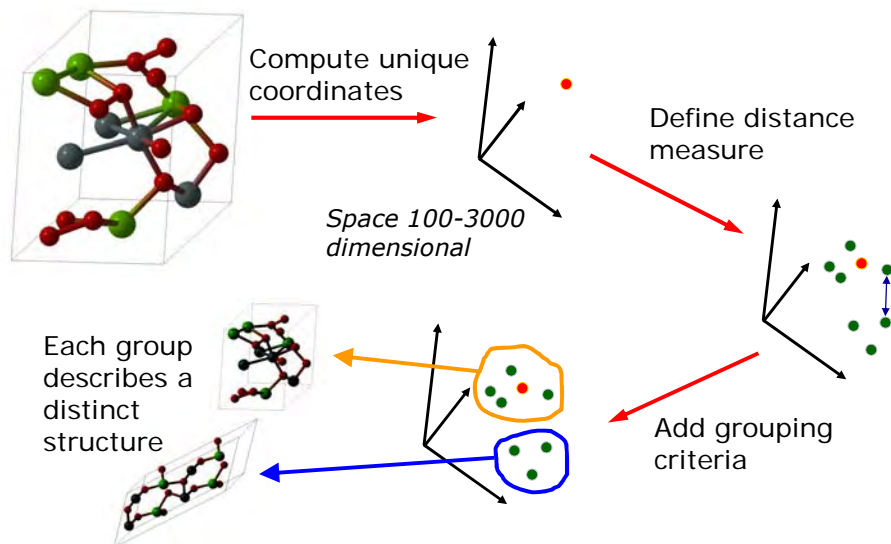


Colored by group

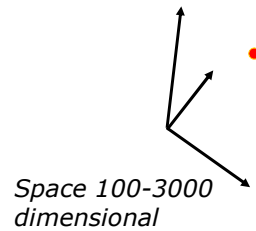
Diagnostic chart: distances in 2D vs. distances in High-D space



So, where is the problem?

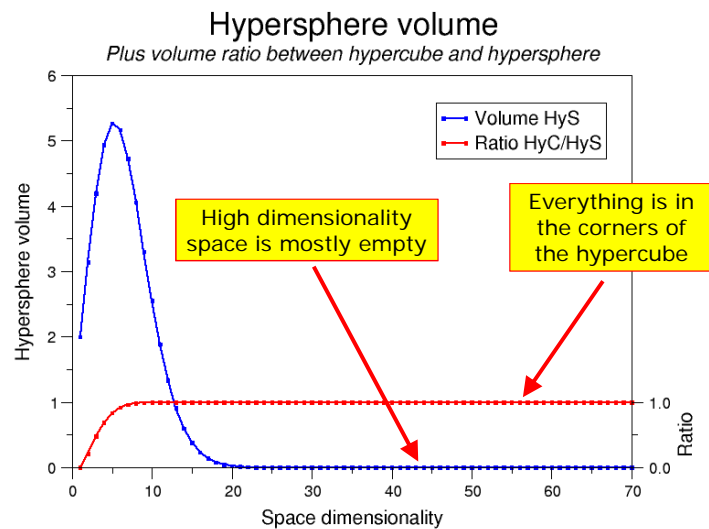


So, where is the problem?



Space 100-3000 dimensional

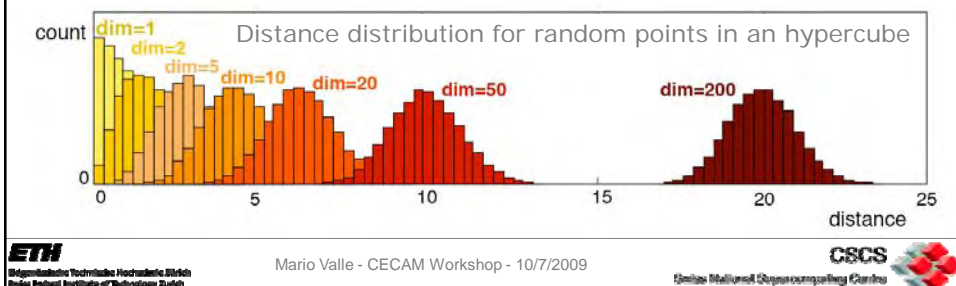
High-dimensionality is not intuitive



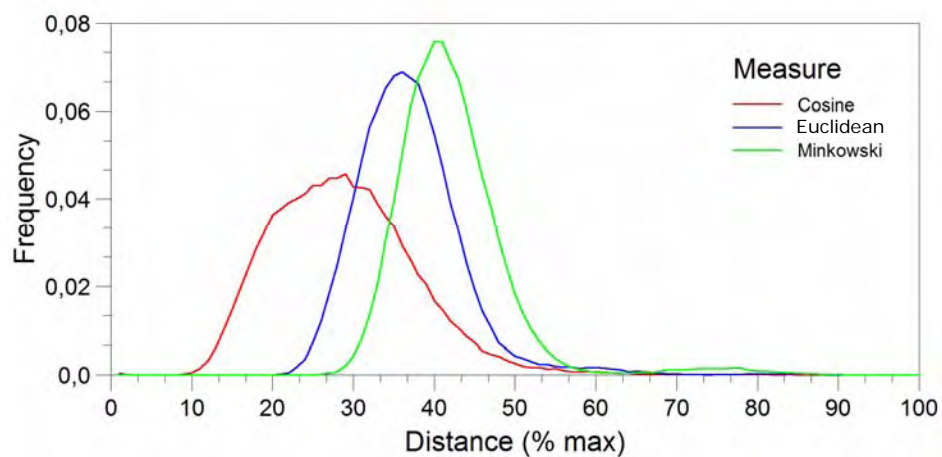
The curse of dimensionality

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...



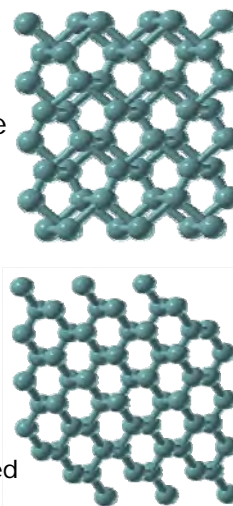
Tried various distance measures



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



nature

PEKING MAN WAS COOL
Older and 'colder' dates for classic Homo erectus fossils

PEKING MAN WAS COOL
Older and 'colder' dates for classic Homo erectus fossils

nature

PEKING MAN WAS COOL
Older and 'colder' dates for classic Homo erectus fossils

PEKING MAN WAS COOL
Older and 'colder' dates for classic Homo erectus fossils

LETTERS

Transparent dense sodium

Yanning Ma^{1,2}, Mikhail Eremin³, Artem R. Oganov^{1,4}, Yu Xie¹, Ivan Trojan⁵, Sergey Medvedev⁶, Andriy O. Lyakhov⁷, Mario Valle¹ & Vitalii Prakapenka¹

Under pressure, metals exhibit increasingly shorter interatomic distances. Intriguingly, this response is expected to be accompanied by an increase in the widths of the valence and conduction bands and hence a more pronounced free-electron-like behaviour. But at the densities that can now be achieved experimentally, compression can be so substantial that core electron overlap, this effect dramatically alters electronic properties: from those typically associated with simple transition metals such as lithium (Li), rubidium (Rb) and cesium (Cs), leading to more or less strongly complex phases^{1–4} and superconductivity with a high critical temperature^{5,6}. But the most surprising prediction—due to the seemingly simple metals Li (ref. 7) and Na (ref. 8)—will transform under pressure into insulating states, owing to pairing of d/f electrons that are not to be experimentally confirmed. Here we report experimental observation of a pressure-induced insulator-to-metal transition in an optically transparent phase at 200 GPa (corresponding to a 3.4-fold compression). Experimental and computational data identify the new phase as a wide bandgap electronic insulator with a well-defined, highly distorted double-hexagonal close-packed structure. We attribute the emergence of this dense insulating state not to core pairing, but to a redistribution of valence electrons and their repulsion by core electrons into the lattice interstices. We expect that such insulating states may also exist in other elements and compounds when compression is sufficiently strong that atomic cores start to overlap strongly.

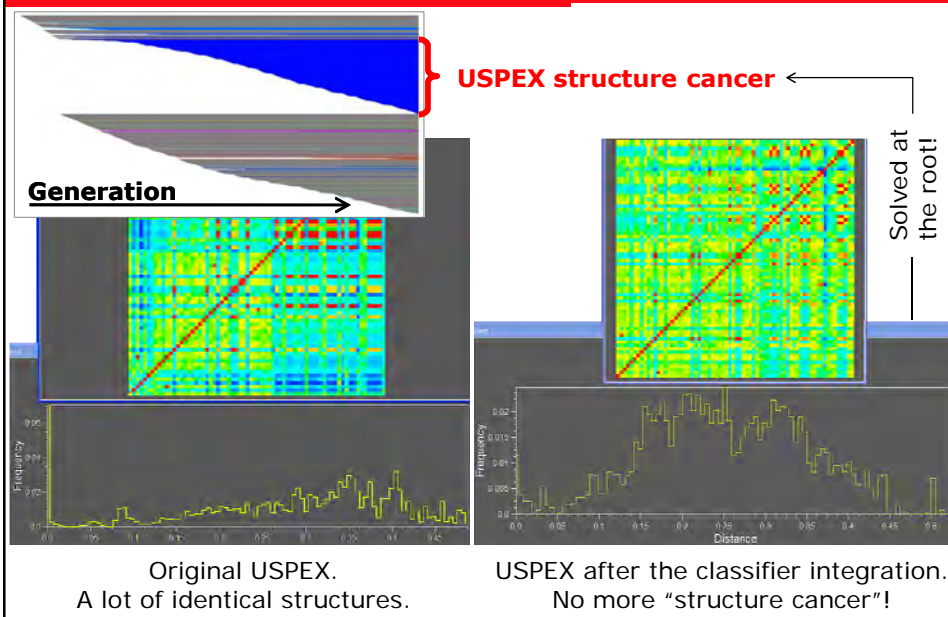
Sodium adopts the body-centred cubic (bcc) structure at ambient conditions. Under pressure, it transforms to the face-centred cubic (fcc) structure at 60 GPa (ref. 12) and to the Cs structure at 100 GPa (ref. 7, 13). In our experiments^{14,15} under compression to 180 GPa yielded a number of phases in a narrow pressure-temperature range near the minimum of the melting curve (melting temperature $T_m \approx 300$ K at 100 GPa). Theoretical calculations¹⁶ suggest stable body-centred tetragonal (bct) and fcc structures for pressures above 180 GPa. The bct structure is especially interesting, because the calculated negative α thermal expansivity above 180 GPa owing to pairing of d electrons¹⁷. Compared the possible structure of stable insulating phases under high pressure, we expect that when core overlap is sufficiently strong, the pressure-induced metal-to-insulator transition in Na, we undertake an extensive experimental and theoretical study, involving access to this as yet unexplored field at high pressures.

Our diamond-cell experiments (see Methods) for details) yielded X-ray diffraction data at 100 and 110 GPa corresponding to the bcc (a.c.a. = 0.36 Å) and fcc phases (a.c.a. = 0.34 Å), respectively, in agreement with available experimental data^{18,19}. The corresponding Raman spectra show no features at pressure below 110 GPa, but a pronounced Raman spectrum appears at higher pressures. Interestingly, because the calculated negative α thermal expansivity above 180 GPa owing to pairing of d electrons¹⁷, we expect that when core overlap is sufficiently strong, the pressure-induced metal-to-insulator transition in Na, we undertake an extensive experimental and theoretical study, involving access to this as yet unexplored field at high pressures.

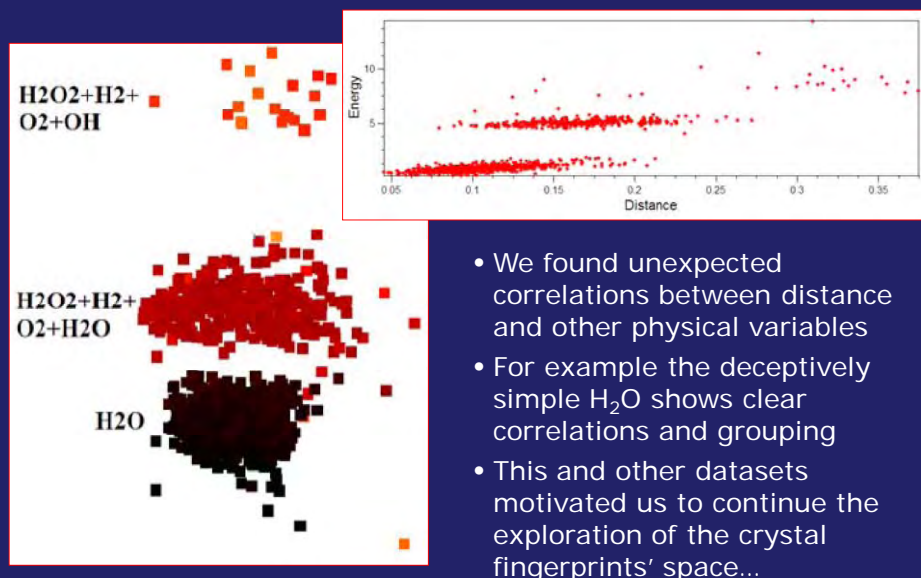
Our diamond-cell experiments (see Methods) for details) yielded X-ray diffraction data at 100 and 110 GPa corresponding to the bcc (a.c.a. = 0.36 Å) and fcc phases (a.c.a. = 0.34 Å), respectively, in agreement with available experimental data^{18,19}. The corresponding Raman spectra show no features at pressure below 110 GPa, but a pronounced Raman spectrum appears at higher pressures. Interestingly, because the calculated negative α thermal expansivity above 180 GPa owing to pairing of d electrons¹⁷, we expect that when core overlap is sufficiently strong, the pressure-induced metal-to-insulator transition in Na, we undertake an extensive experimental and theoretical study, involving access to this as yet unexplored field at high pressures.

Figure 1 Raman spectra of sodium. A series of spectra obtained at increasing pressures. Spectra of 0, 20, 40, 60 and 200 GPa are shown. The spectra show a transition from a broad peak at low pressures to a sharp peak at high pressures, indicating a phase transition.

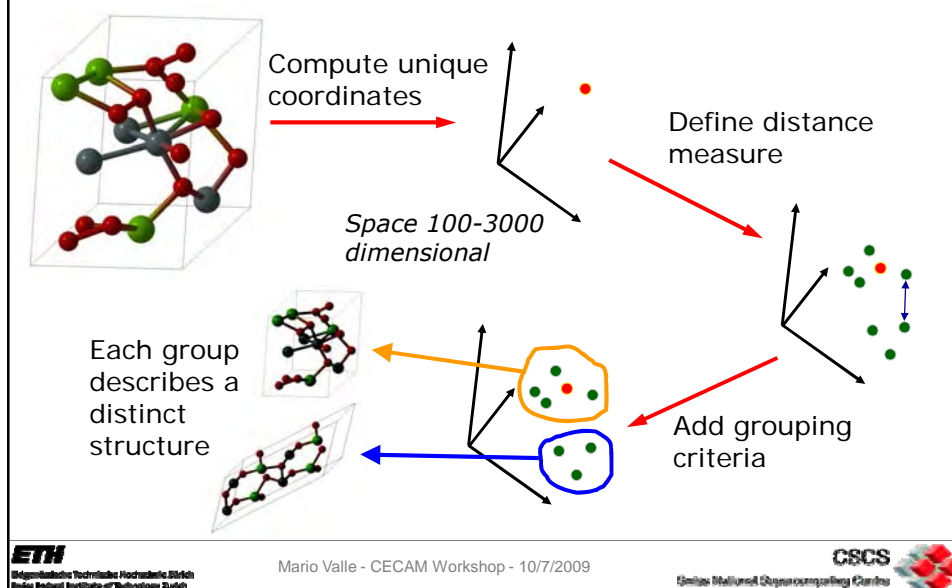
Classifier integration in USPEX



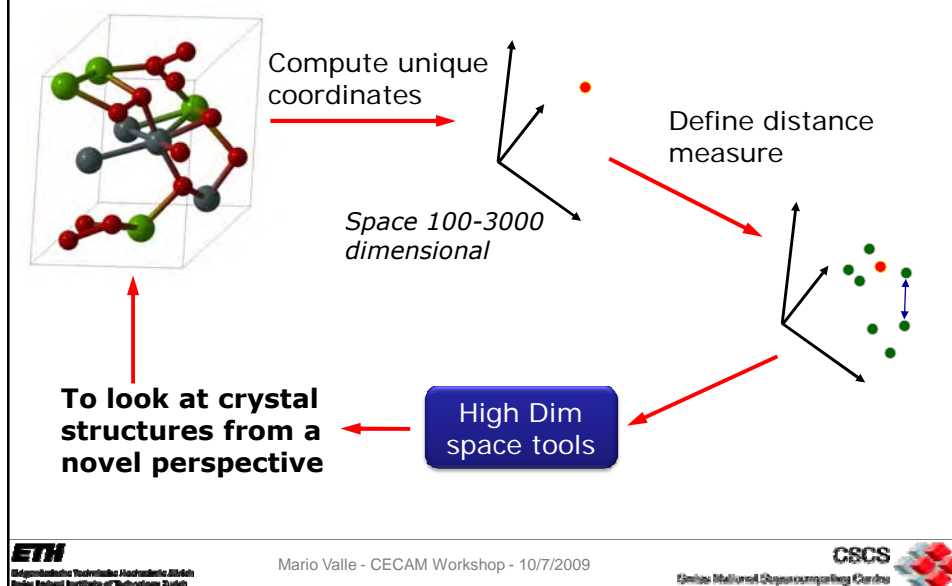
(Totally) unexpected correlations



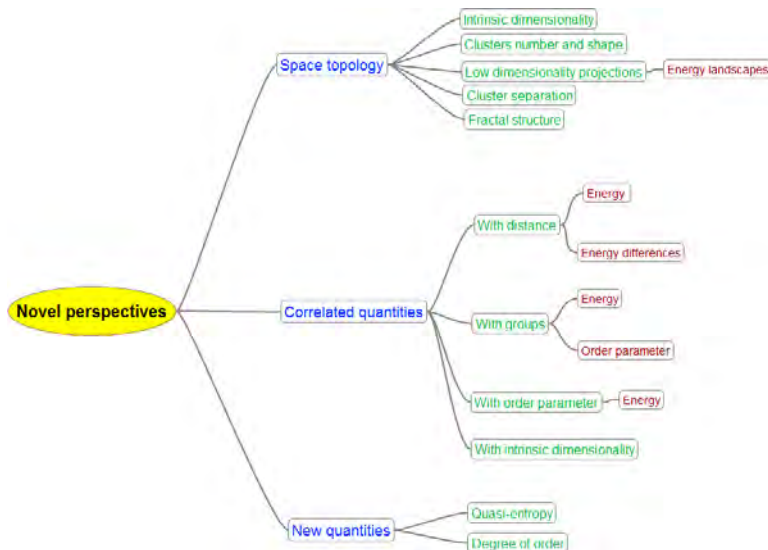
From the problem solution ...



... to a new paradigm

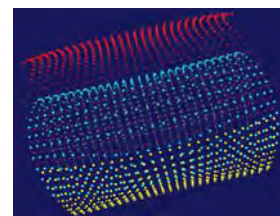
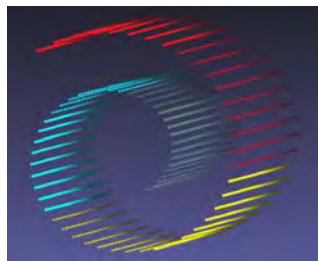


What we want to look at

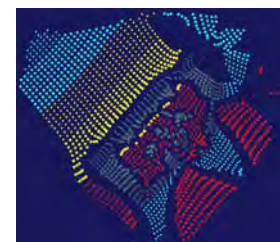


Unfold data to lower dimensions

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

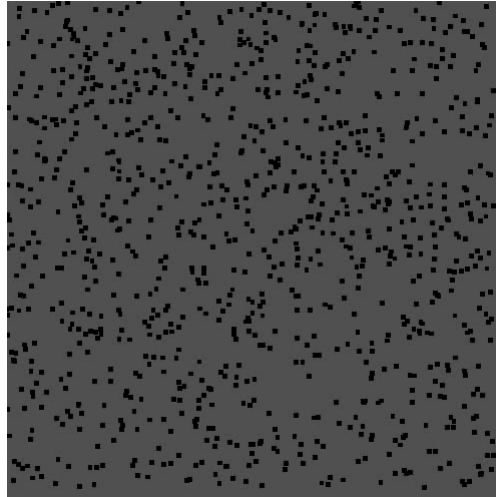


CCA mapping

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)

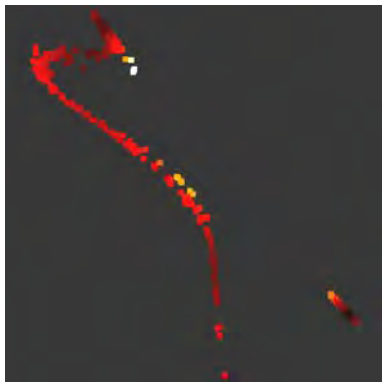
CrystalFp multi dim. scaling

The scatterplot tool in CrystalFp is the implementation of a Force Directed Placement multidimensional scaling algorithm

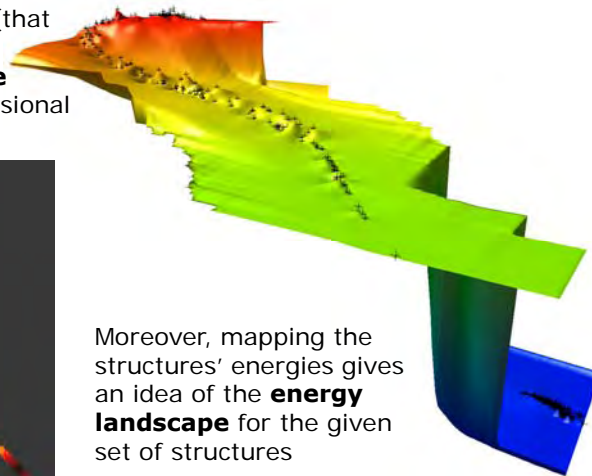


Multi dim. scaling to understand

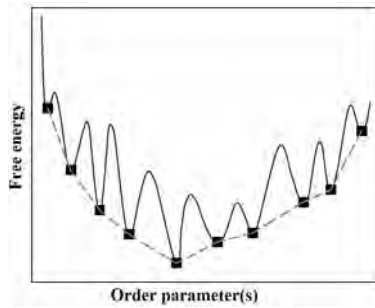
Multi dimensional scaling (that is, CrystalFp scatterplot) gives an idea of the **shape** of the abstract high dimensional fingerprint space



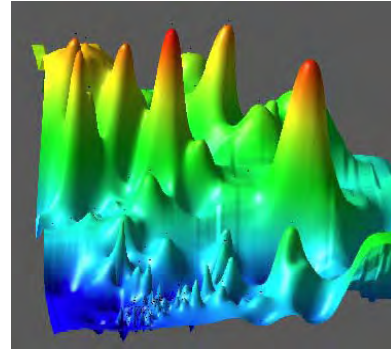
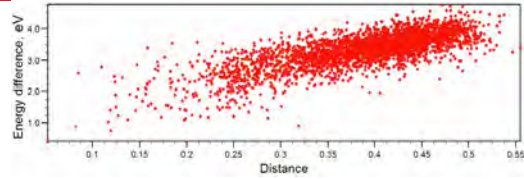
Moreover, mapping the structures' energies gives an idea of the **energy landscape** for the given set of structures



Study of energy landscapes

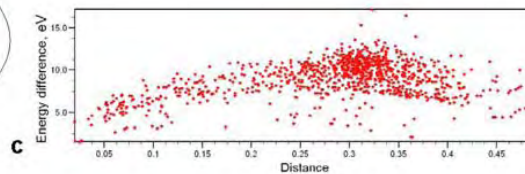
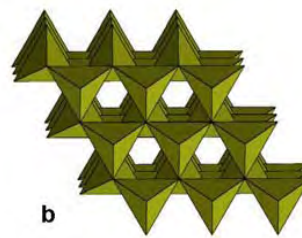
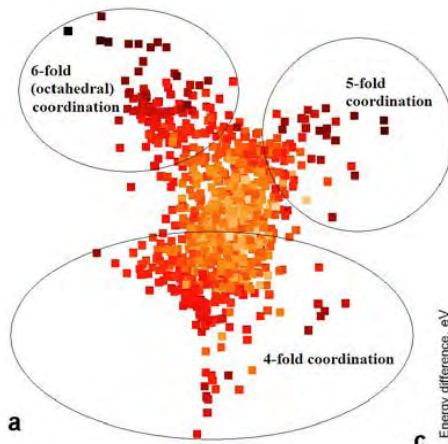


A. R. Oganov and M. Valle,
How to quantify energy landscapes of solids,
The Journal of Chemical Physics, vol. 130,
p. 104504, 2009.



Energy landscape of Au_8Pd_4 system

More complex landscapes



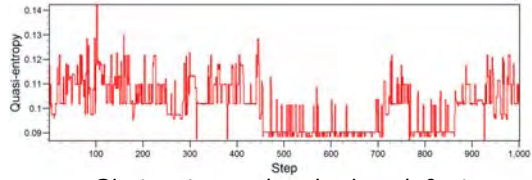
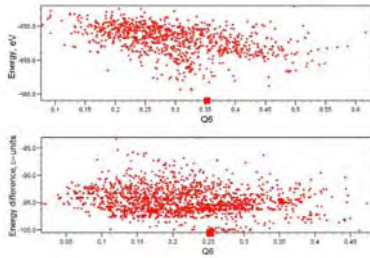
Energy landscape for MgO with 32 atoms/cell

New quantities: quasi-entropy

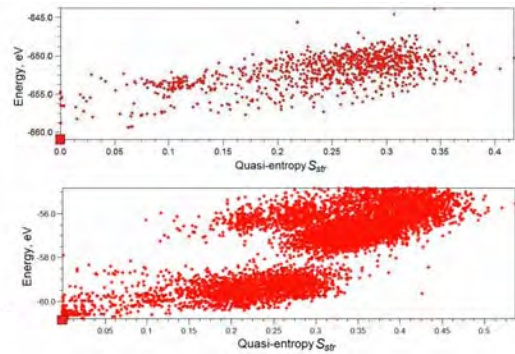
For each given structure, **quasi-entropy** is a measure of disorder and complexity of that structure.

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

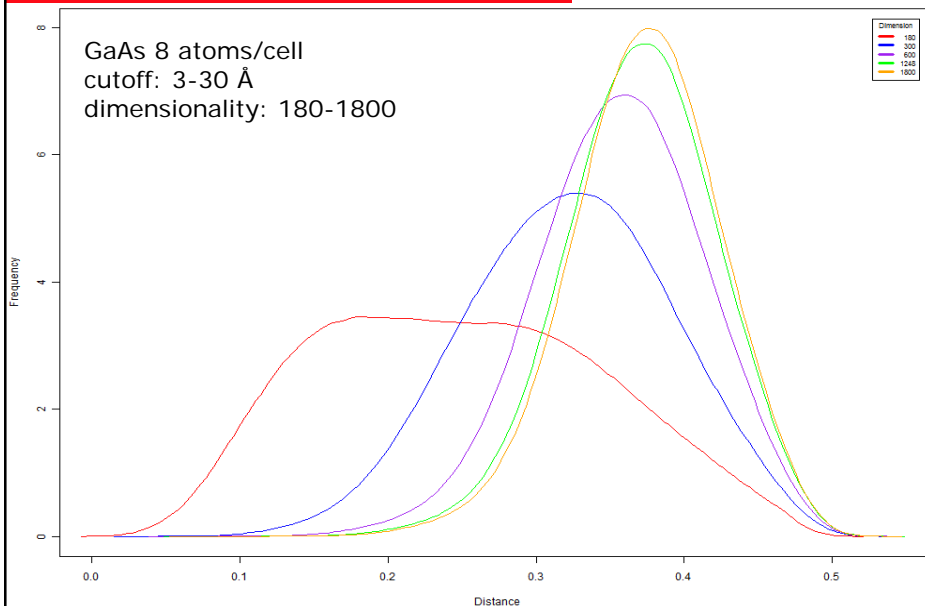
S_{str} is better correlated to energy than Steinhardt's Q_6



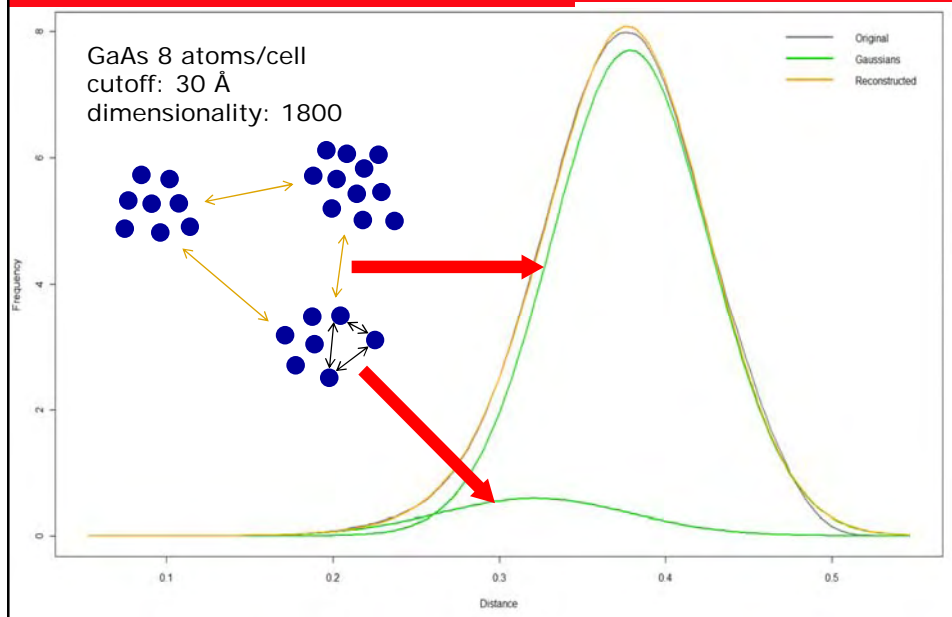
Si structures developing defects



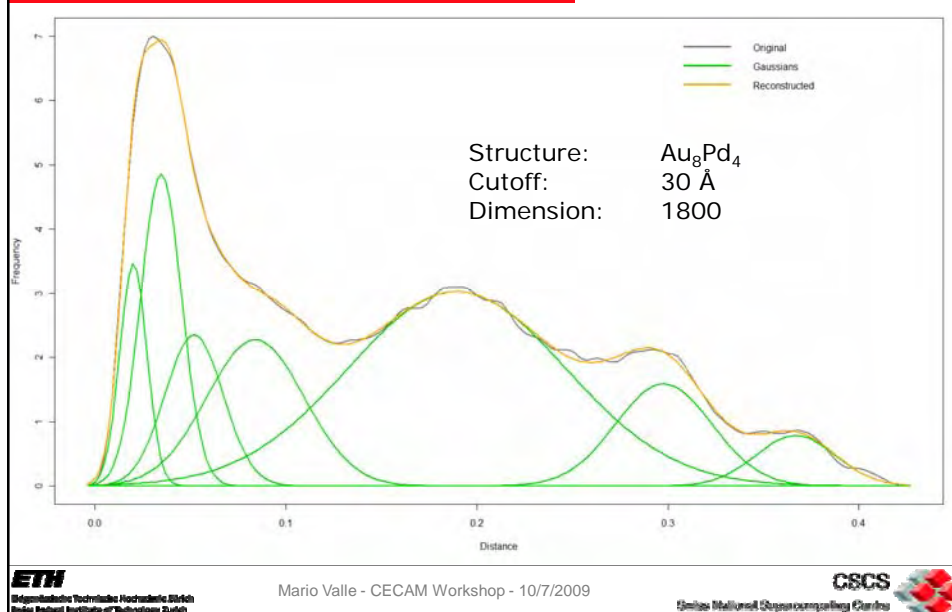
Distance vs. dimensionality



Distance decomposition



But this one?



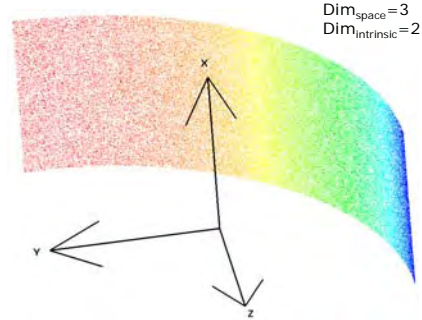
Intrinsic dimensionality

Fingerprint space
dimensionality:
 $100 \div 3000$

Vastly redundant!

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

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

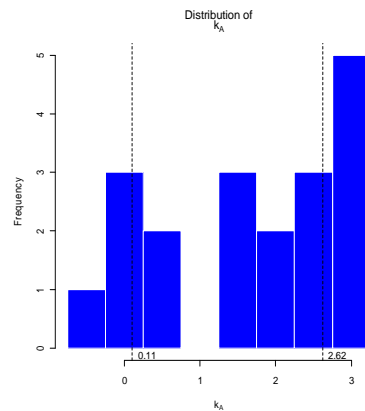


Au ₈ Pd ₄	39	10.85
MgNH	39	32.47
MgO	99	11.62

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:
 $100 \div 3000$

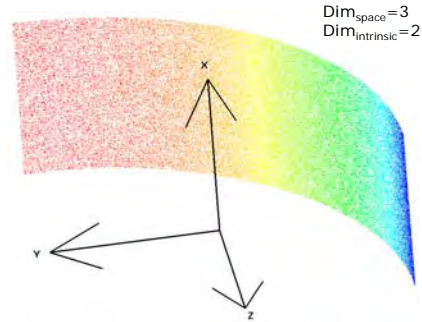
Vastly redundant!

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

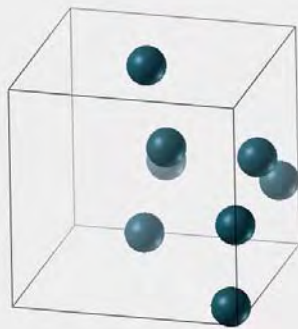
More realistic theory:
 $Dim_{intrinsic} = 3 * N_{atoms} + 3 - K$

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

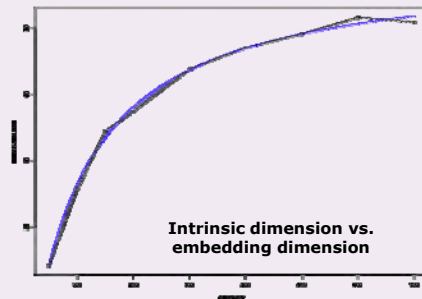
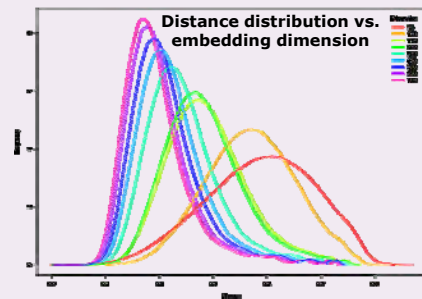
But how do you explain this?



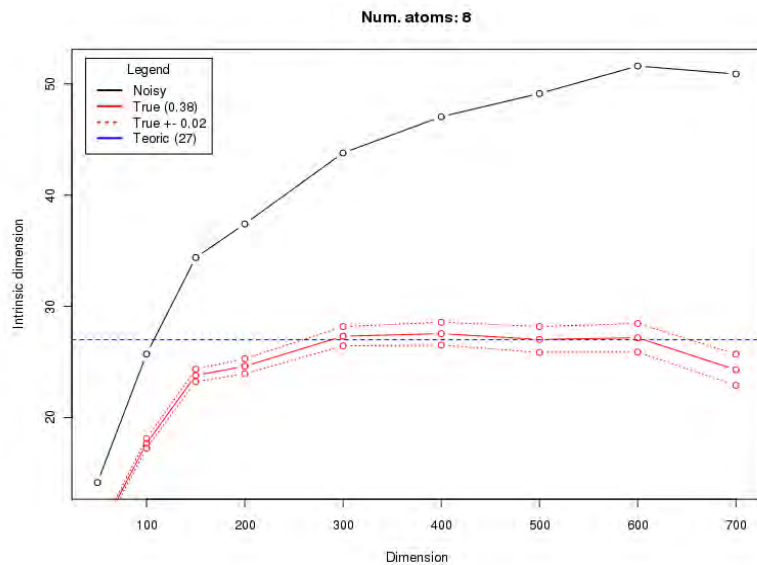
Synthetic datasets



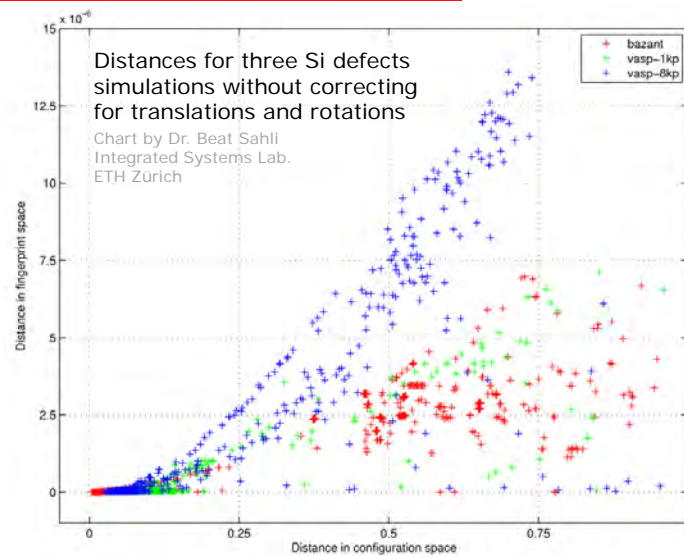
8 atoms with uniformly
distributed random fractional
coordinates in a cubic unit
cell with 5 Å side



White noise removed



Faithful representation?



CrystalFp publications so far

1. Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, and V. Prakapenka, **Transparent dense sodium**, *Nature*, vol. 458, pp. 182-185, Mar. 14 2009.
2. A. R. Oganov and M. Valle, **How to quantify energy landscapes of solids**, *The Journal of Chemical Physics*, vol. 130, p. 104504, Mar. 14 2009.
3. M. Valle and A. R. Oganov, **Crystal Structures Classifier for an Evolutionary Algorithm Structure Predictor**, in *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08.*, pp. 11-18, Oct. 19 - 24 2008.
4. A. R. Oganov, M. Valle, A. O. 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.
5. 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.



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

Mario Valle - CECAM Workshop - 10/7/2009

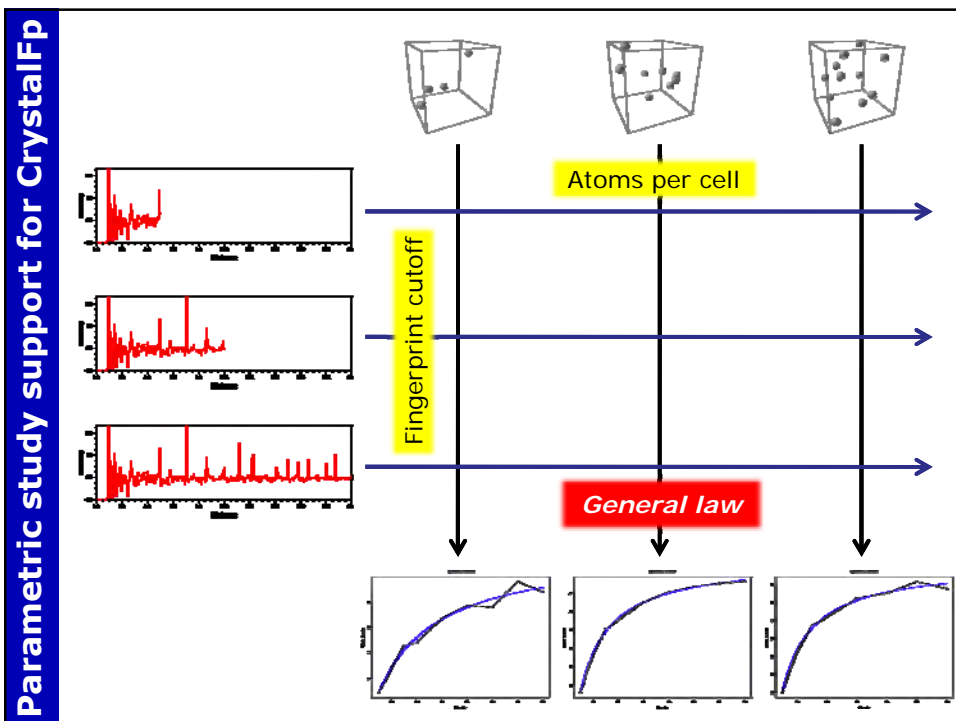
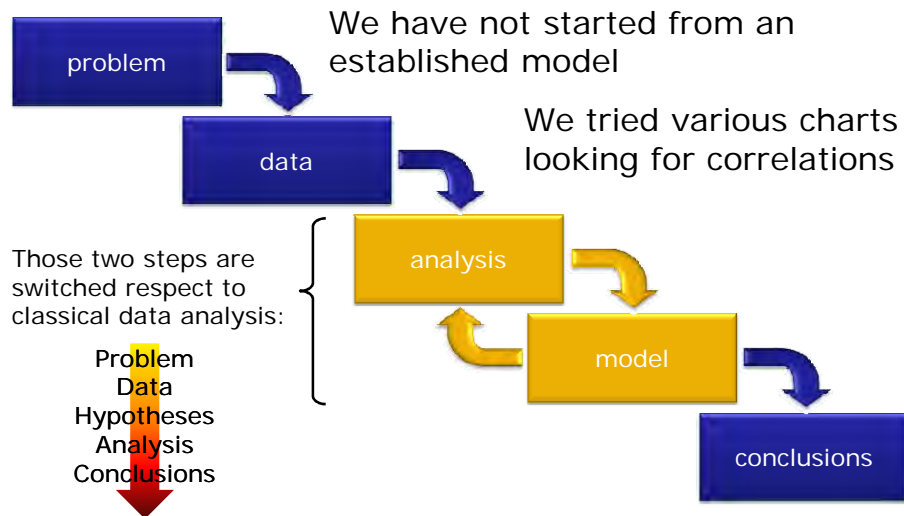


CSCS
Swiss National Supercomputing Centre

“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*

Role of exploration



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

But

- Little support for visualization of parametric studies
- Roll-your-own data management support
- No tool support for the thinking process (annotation, etc.)

The lessons learned ...

- 1. Looking outside** our own discipline produced unexpected outcomes
- 2. These unexpected outcomes** transformed the project from a point solution to a more general tool for the field
- 3. Visual exploration** plays a non-marginal role, but currently has very shallow support from tools



**Thank you
for your attention!**

And don't forget: www.cscs.ch/~mvalle/CrystalFp