CSCS

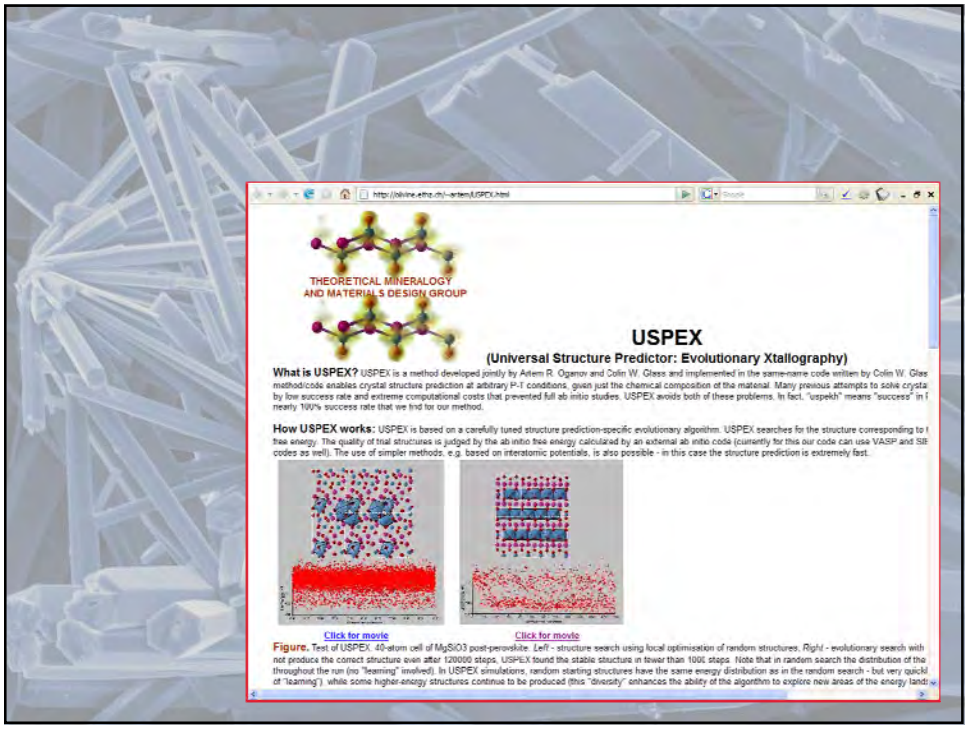
Swiss National Supercomputing Centre

Mario Valle

CECAM Workshop – 10/7/2009

Crystal fingerprints space

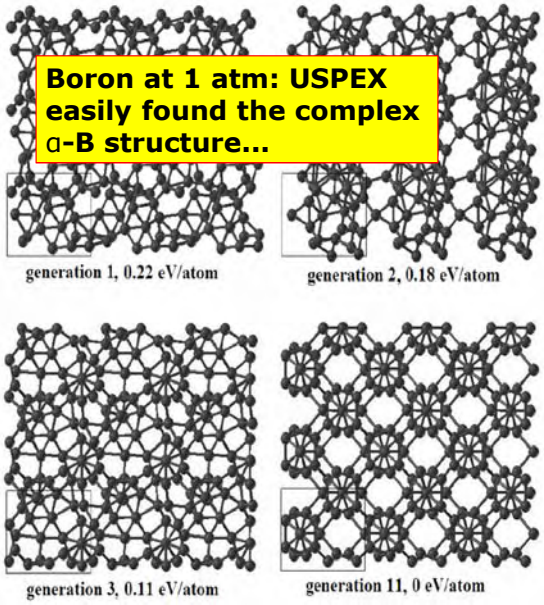
a novel paradigm to study crystal structures sets

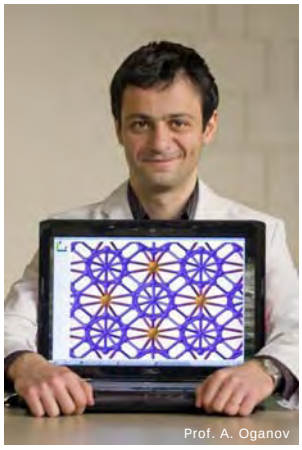


Use the genetic evolution idea!

USPEX discovered new materials

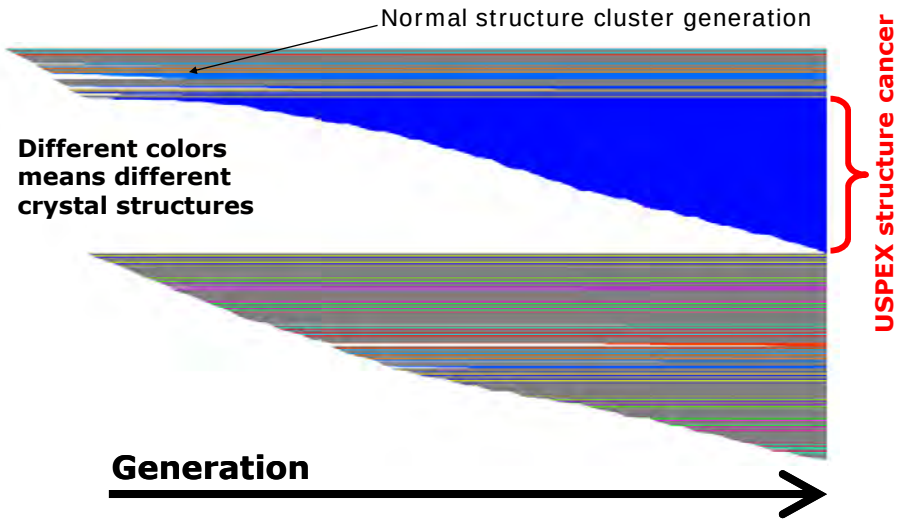
Boron at 1 atm: USPEX easily found the complex α-B structure…





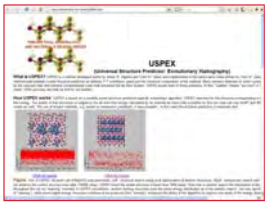
…but discovered also the superhard boron γ-B_{28} phase (*Nature*)

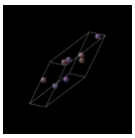
USPEX structure cancer problem



CrystalFp: the problem to solve

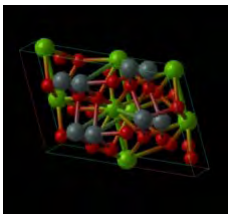
USPEX is a crystal structure predictor based on an evolutionary algorithm





Each run produces hundred of putative crystal structures…

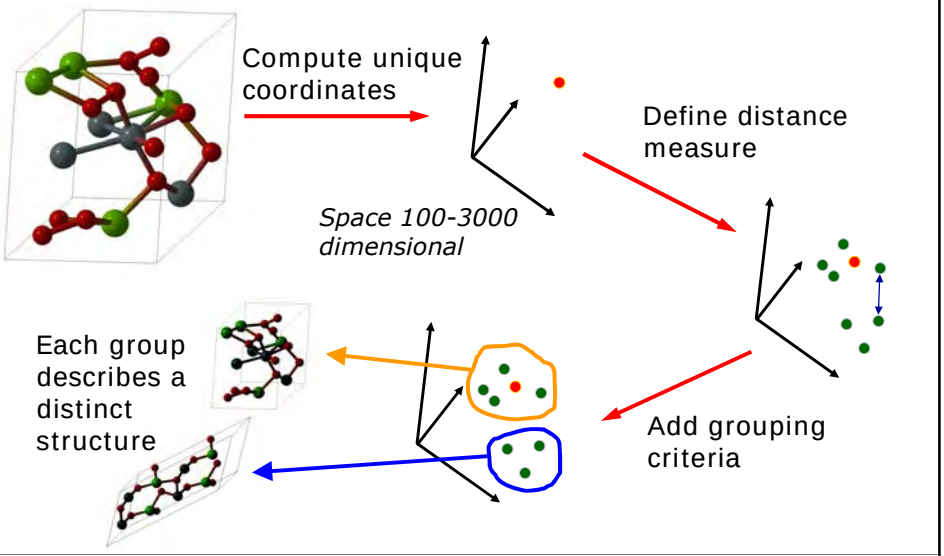
…but many of them are equal



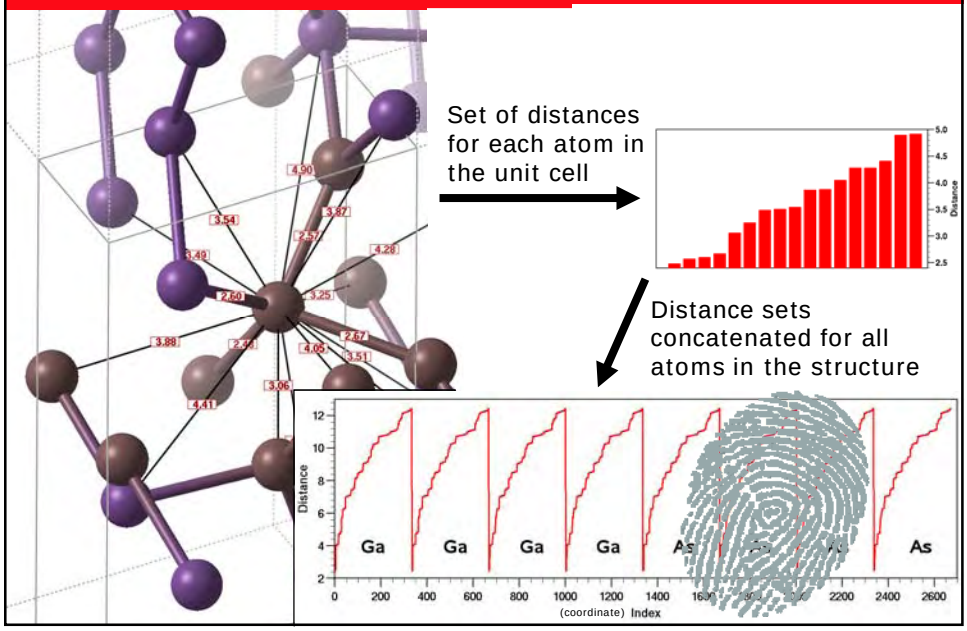
So an intensive manual labor is needed to prune duplicated structures

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

Proposed solution from High-Dim



Structure "coordinates" (fingerprint)

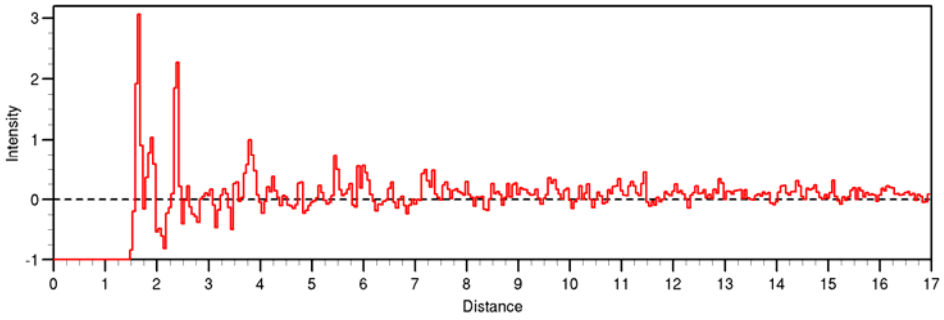


A pseudo-diffraction like method

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

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

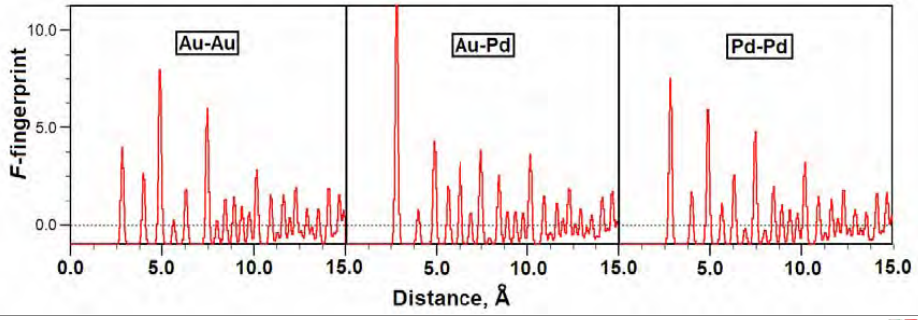
The fingerprint is cut at a user defined distance to provide 100-400 coordinate values



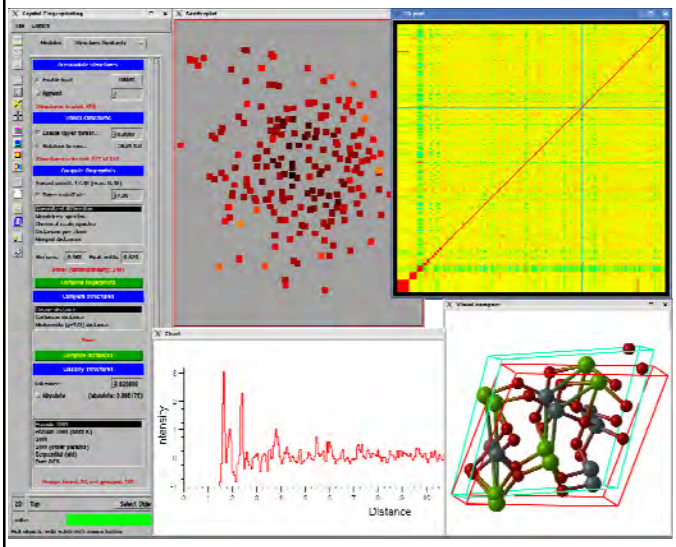
Final fingerprint (per atom type pair)

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

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



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

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

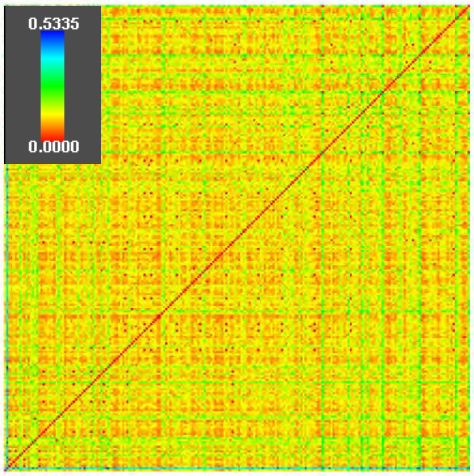
1. Load structures
2. Filter on energy
3. Compute fingerprints
4. Compute distances
5. Group structures

Visual diagnostics tools

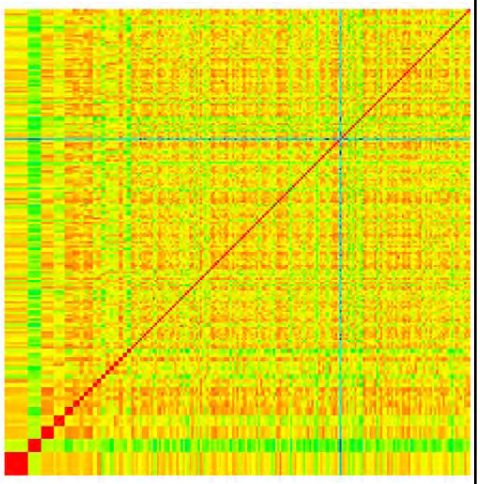
Various visualization and analysis tools to check and **validate** CrystalFp algorithms behavior

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

Visual diagnostics: distance matrix



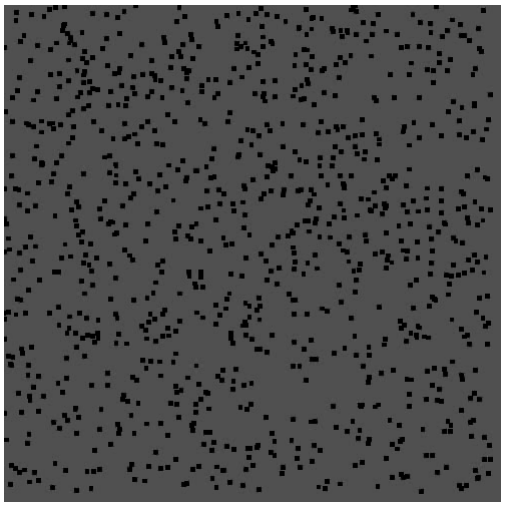
Distances between structures

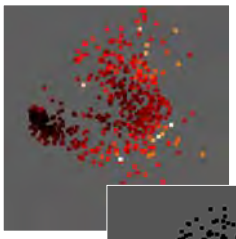


Distances ordered by group

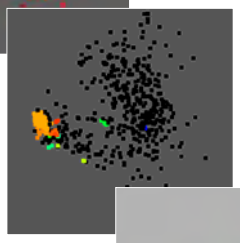
Visual diagnostics: scatterplot

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

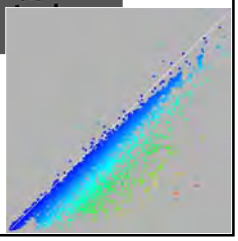




Colored by "stress" to detect local minima traps

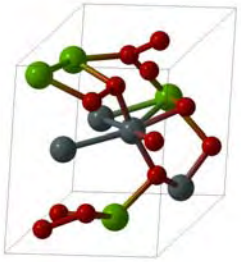


Colored by group

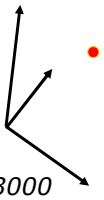


Diagnostic chart: distances in 2D vs. distances in High-D space

So, where is the problem?

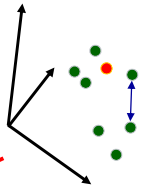


Compute unique coordinates

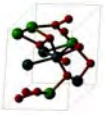


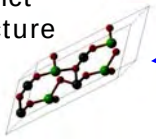
Space 100-3000 dimensional

Define distance measure



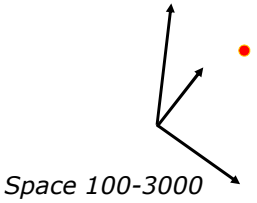
Add grouping criteria





Each group describes a distinct structure

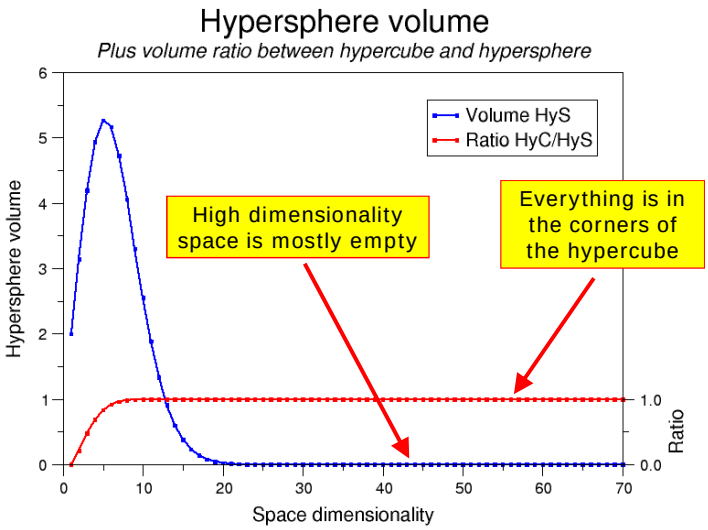
So, where is the problem?



Space 100-3000 dimensional

Space 100-3000 dimensional

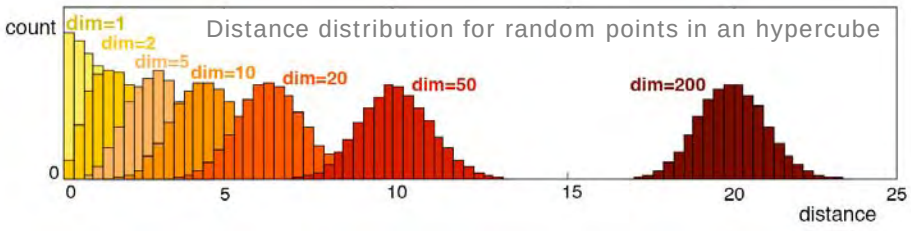
High-dimensionality is not intuitive



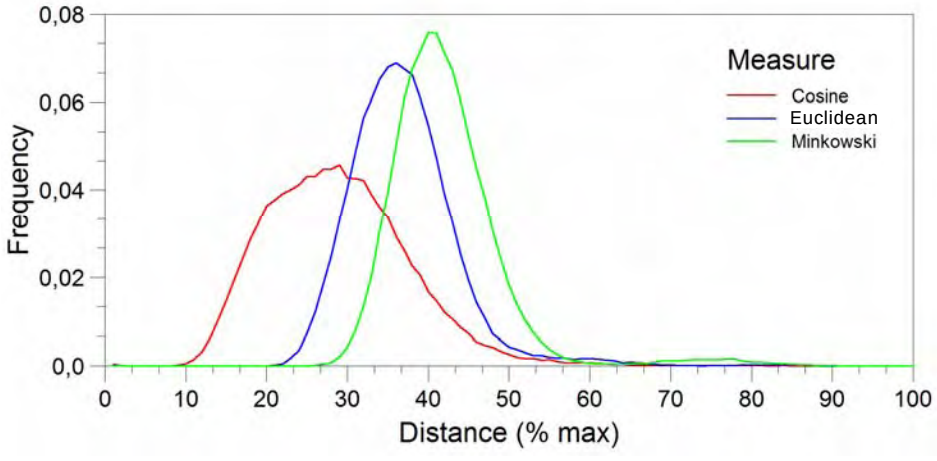
The curse of dimensionality

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…



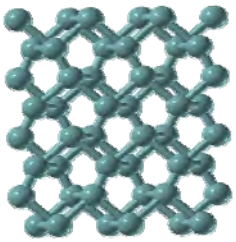
Tried various distance measures

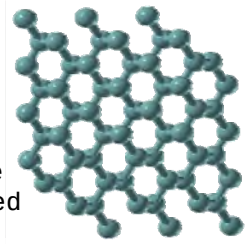


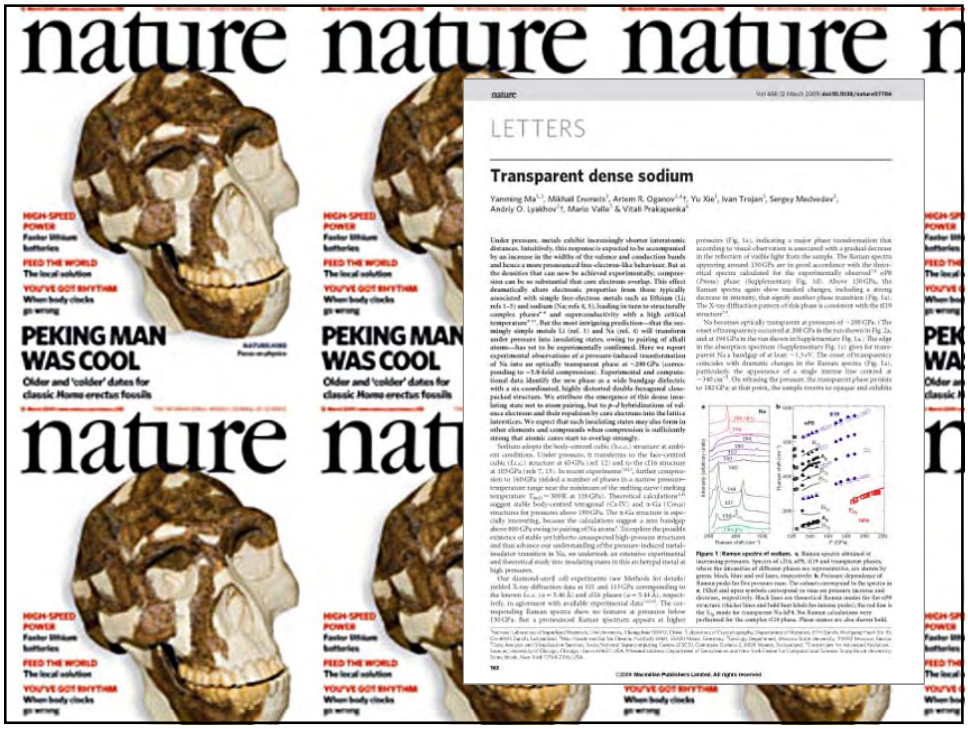
USPEX problem solved: an example

Hydrogen at 600 GPa (16 atoms)

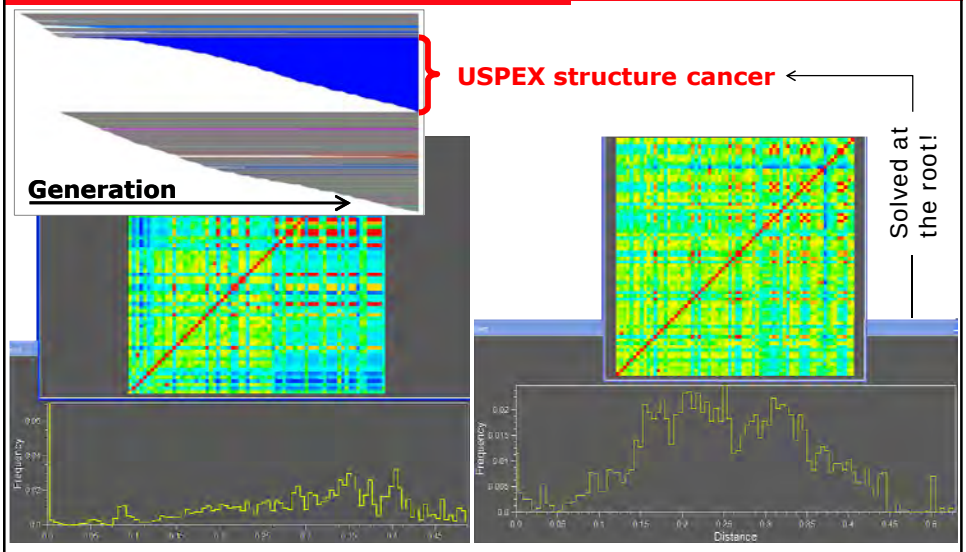
- The USPEX run produced **1274** structures
- From these the **794** within 0.5 eV from the lowest energy value found are selected
- Manual analysis to remove duplicated structures from this set: **~20h** of work
- Using the CrystalFp classifier: **~10min**
- At the end found only **4** unique structures: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







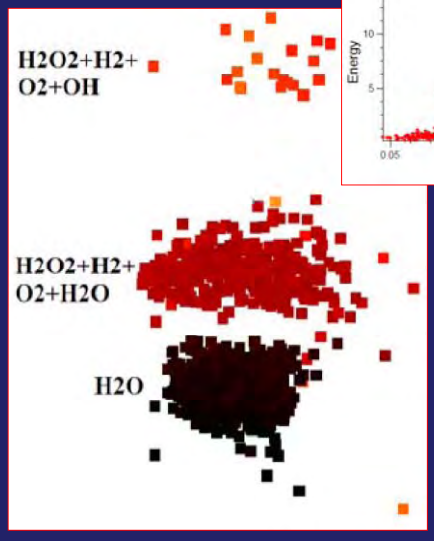
Classifier integration in USPEX

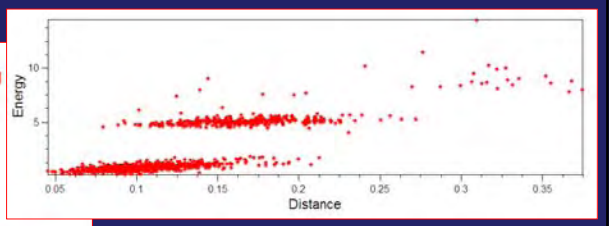


Original USPEX.
A lot of identical structures.

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

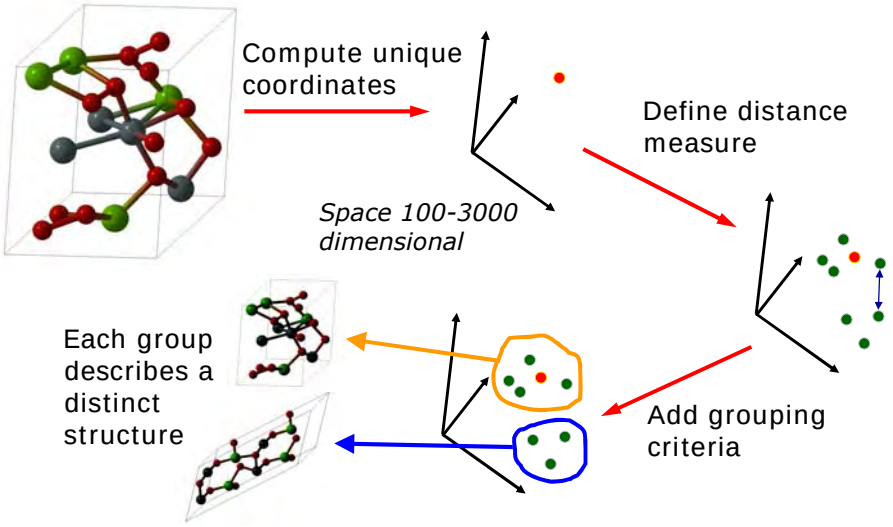
(Totally) unexpected correlations



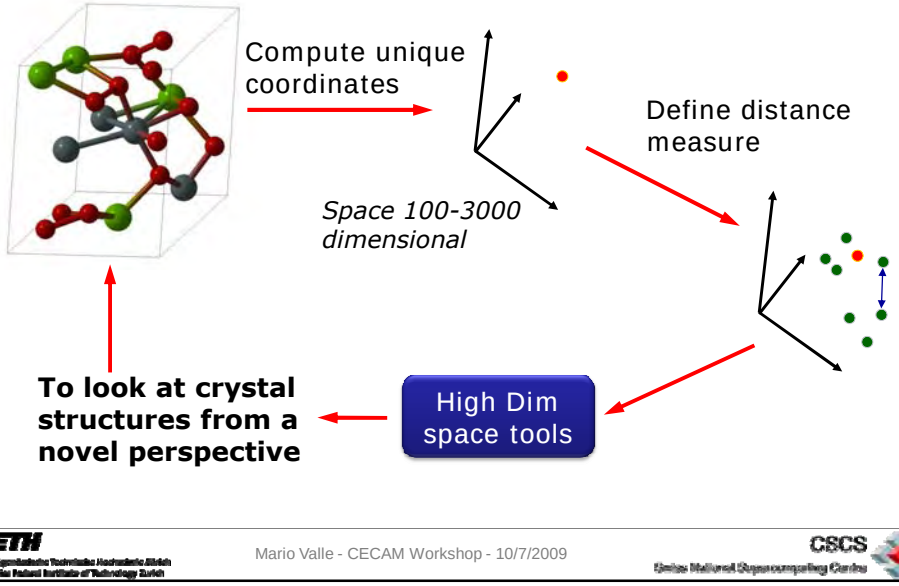


- We found unexpected correlations between distance and other physical variables
- For example the deceptively simple H_2O shows clear correlations and grouping
- This and other datasets motivated us to continue the exploration of the crystal fingerprints’ space…

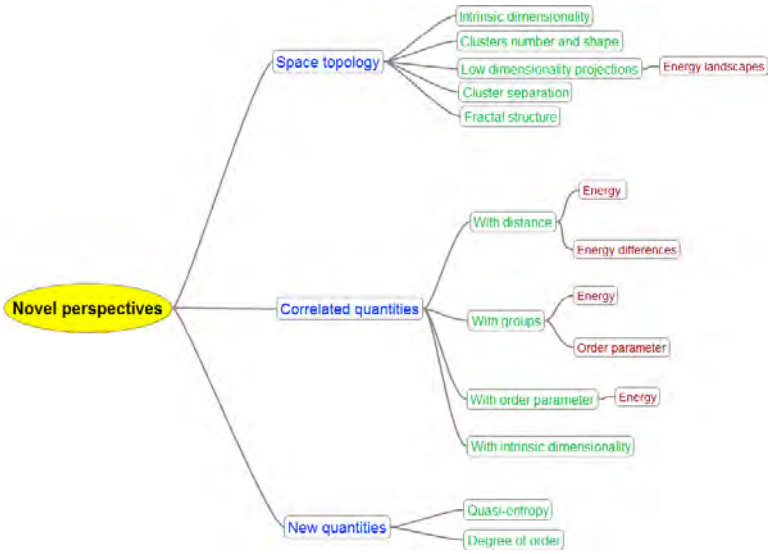
From the problem solution …



… to a new paradigm

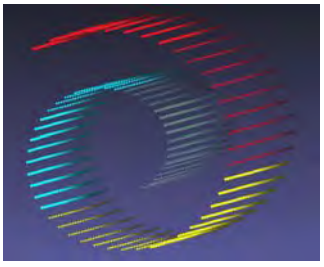


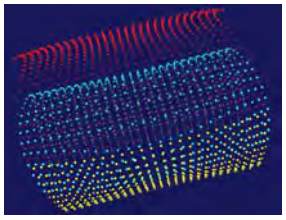
What we want to look at



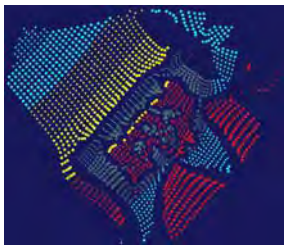
Unfold data to lower dimensions

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

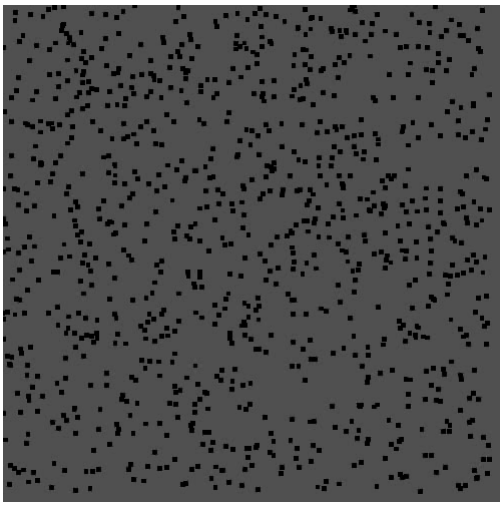


CCA mapping

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)

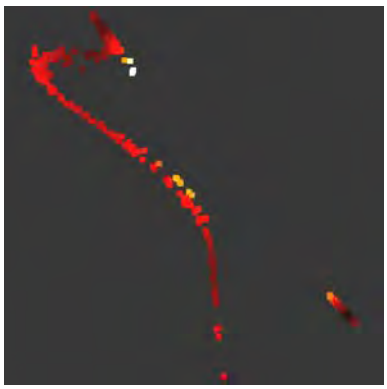
CrystalFp multi dim. scaling

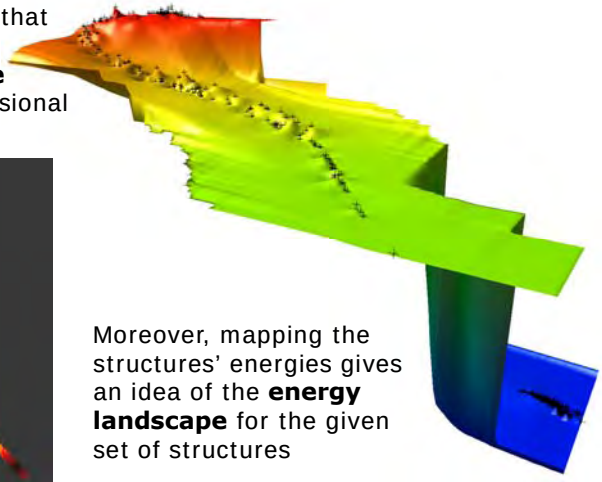
The scatterplot tool in CrystalFp is the implementation of a Force Directed Placement multidimensional scaling algorithm



Multi dim. scaling to understand

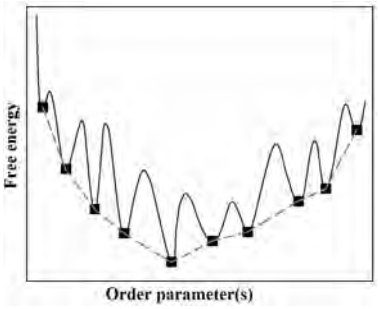
Multi dimensional scaling (that is, CrystalFp scatterplot) gives an idea of the **shape** of the abstract high dimensional fingerprint space

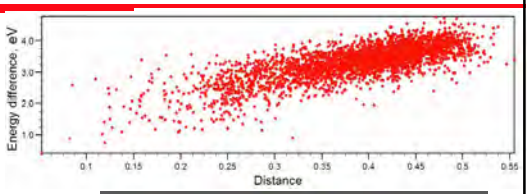


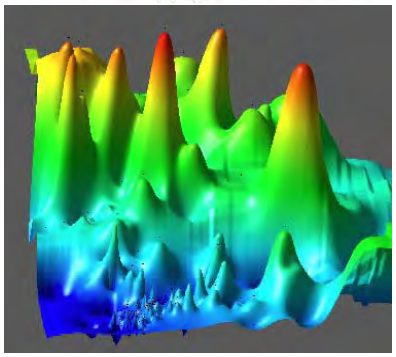


Moreover, mapping the structures' energies gives an idea of the **energy landscape** for the given set of structures

Study of energy landscapes



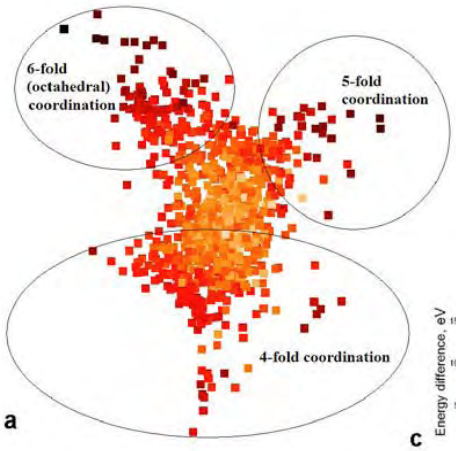


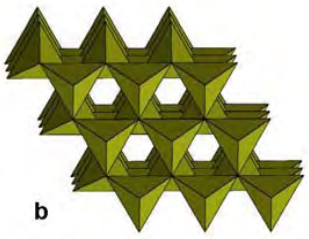


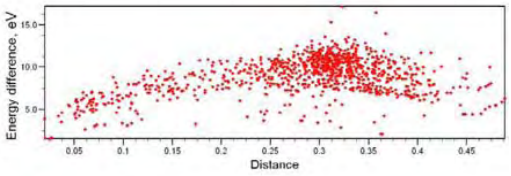
A. R. Oganov and M. Valle,
How to quantify energy landscapes of solids,
The Journal of Chemical Physics, vol. 130,
p. 104504, 2009.

Energy landscape of Au_8Pd_4 system

More complex landscapes







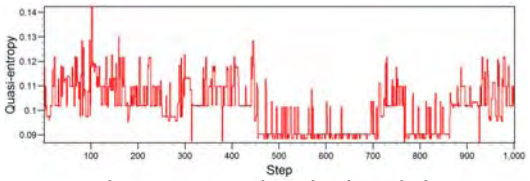
Energy landscape for MgO with 32 atoms/cell

New quantities: quasi-entropy

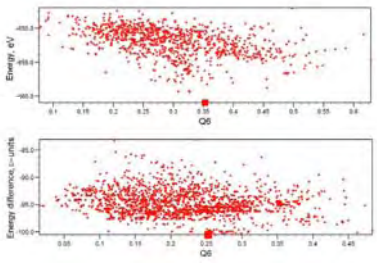
For each given structure, **quasi-entropy** is a measure of disorder and complexity of that structure.

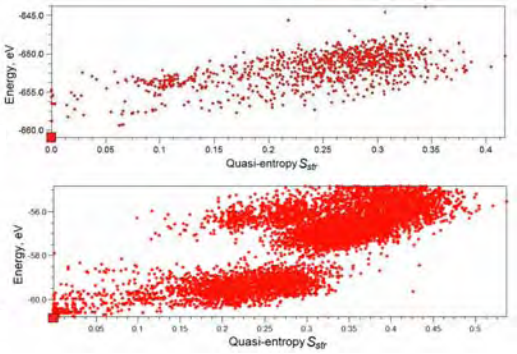
$$S_{\mathrm{str}} = -\sum_{A} \frac{N_A}{N_{\mathrm{cell}}} \langle \ln(1 - D_{A_i A_j}) \rangle$$

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

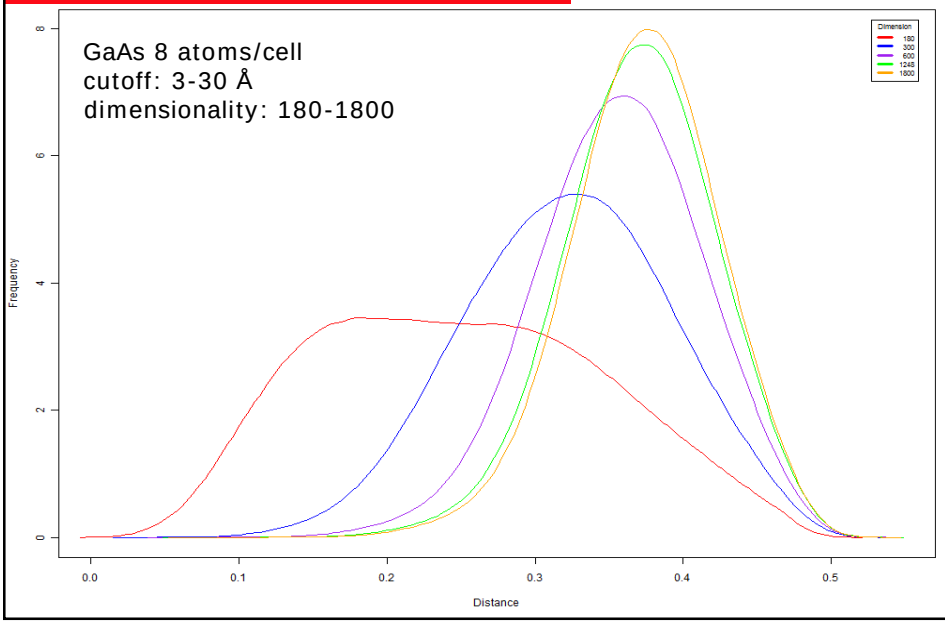


Si structures developing defects

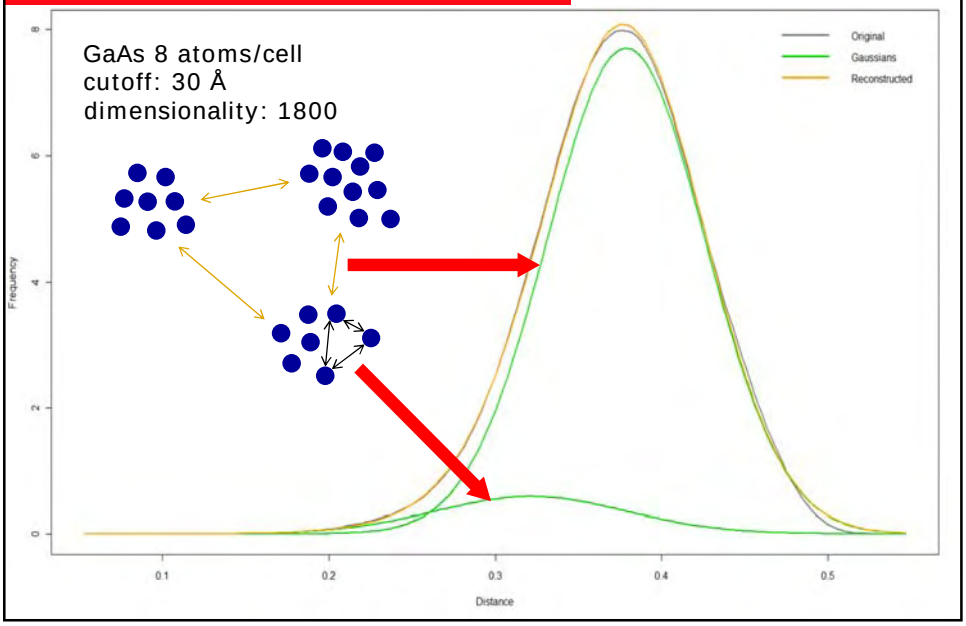




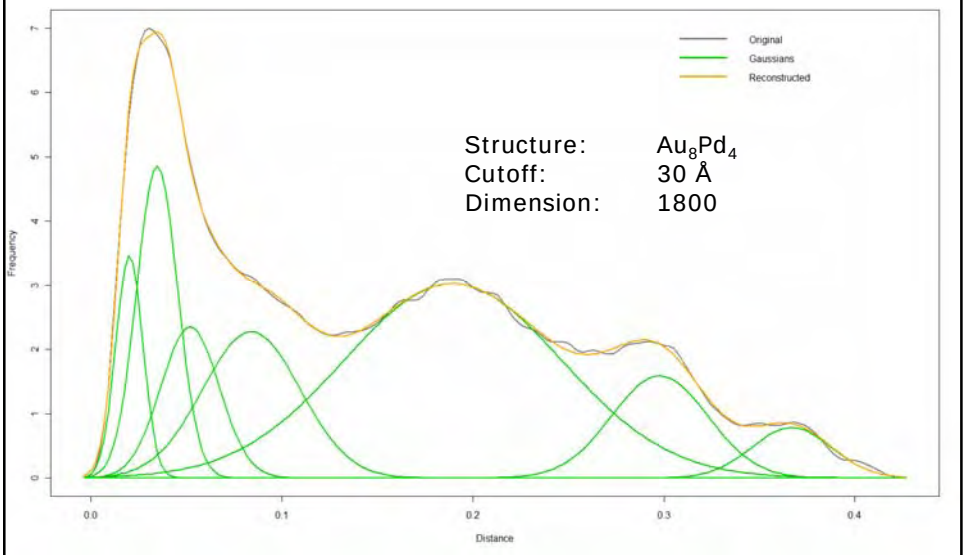
Distance vs. dimensionality



Distance decomposition



But this one?



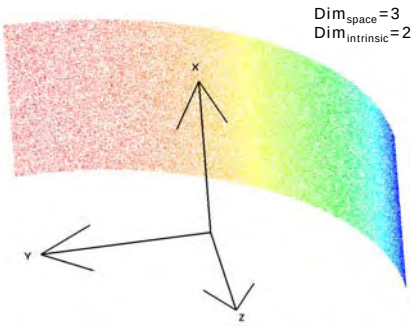
Intrinsic dimensionality

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

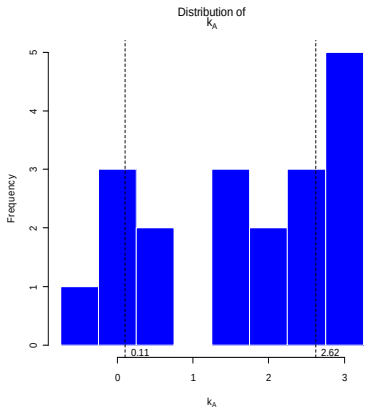
More realistic theory:

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

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

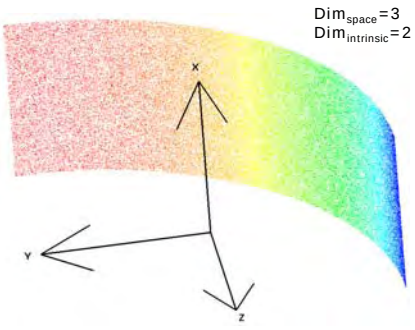
Vastly redundant!

Theory:

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

More realistic theory:

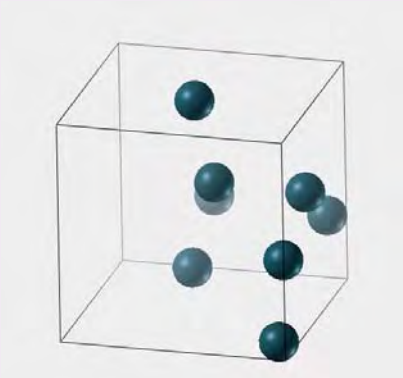
$Dim_{intrinsic} = 3*N_{atoms}+3-K$



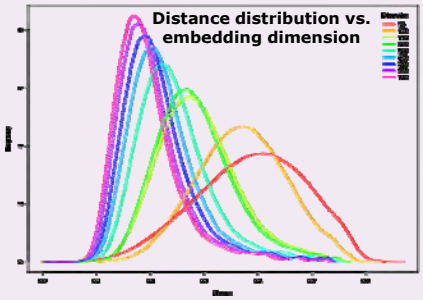
Au_8Pd_4	39	10.85
MgNH	39	32.47
MgO	99	11.62
H_2O	**39**	**80.50**

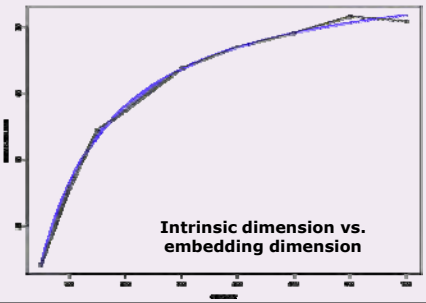
But how do you explain this?

Synthetic datasets

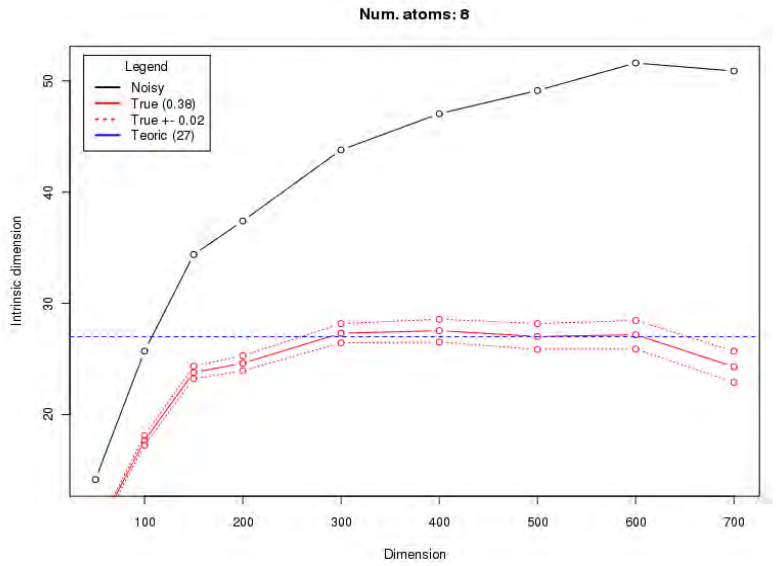


8 atoms with uniformly distributed random fractional coordinates in a cubic unit cell with 5 Å side

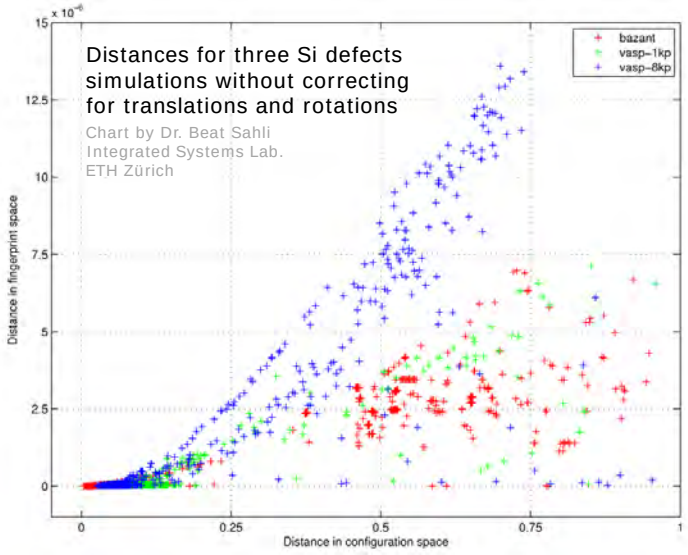




White noise removed

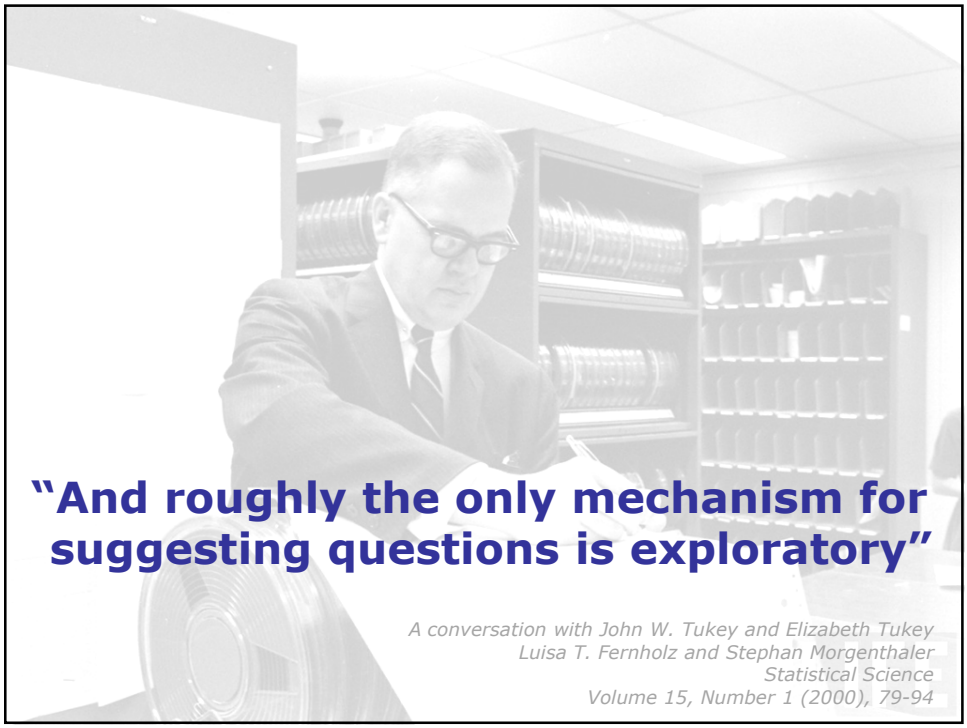


Faithful representation?



CrystalFp publications so far

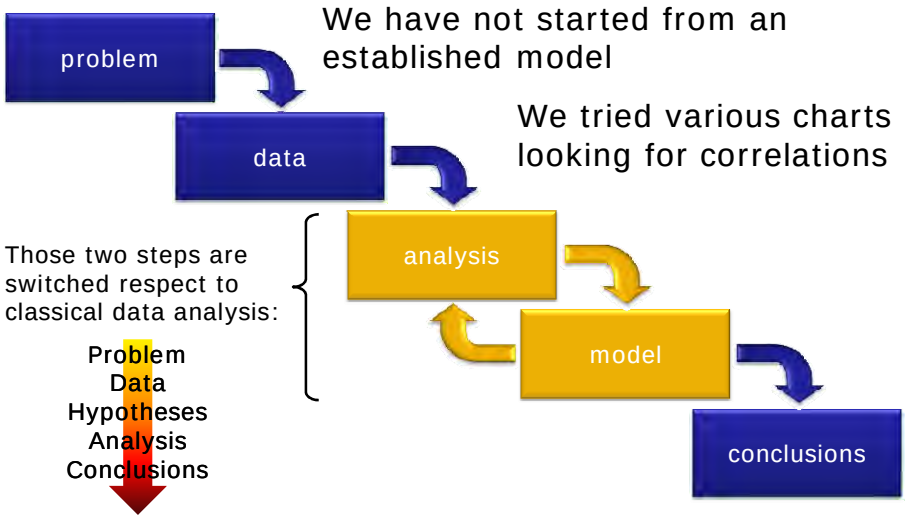
1. Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, and V. Prakapenka, **Transparent dense sodium,** *Nature*, vol. 458, pp. 182-185, Mar. 14 2009.
2. A. R. Oganov and M. Valle, **How to quantify energy landscapes of solids,** *The Journal of Chemical Physics*, 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.

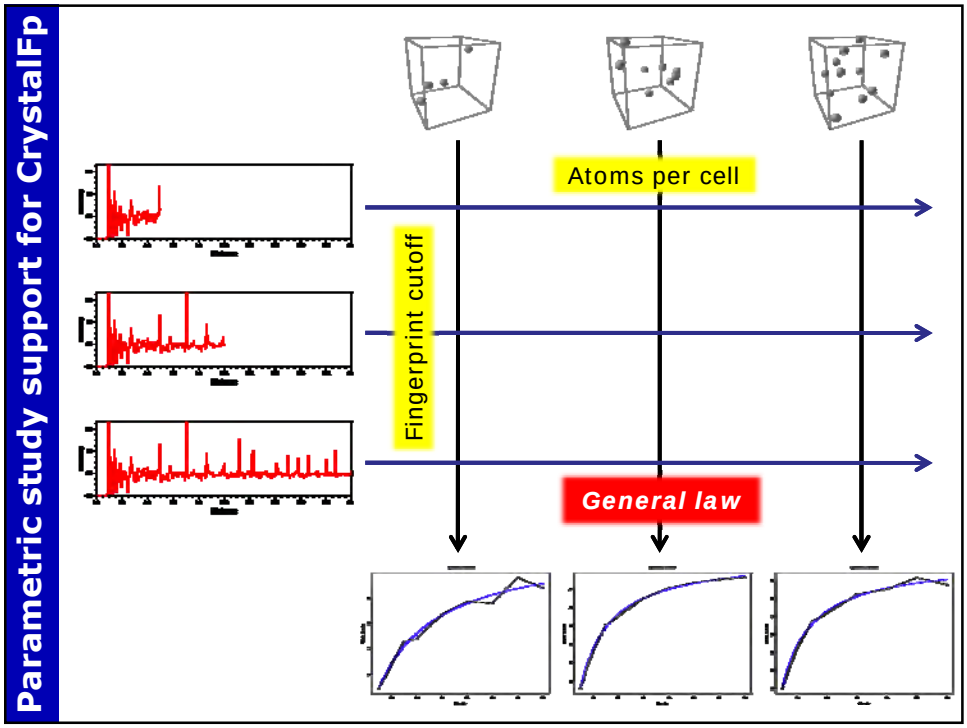


"And roughly the only mechanism for suggesting questions is exploratory"

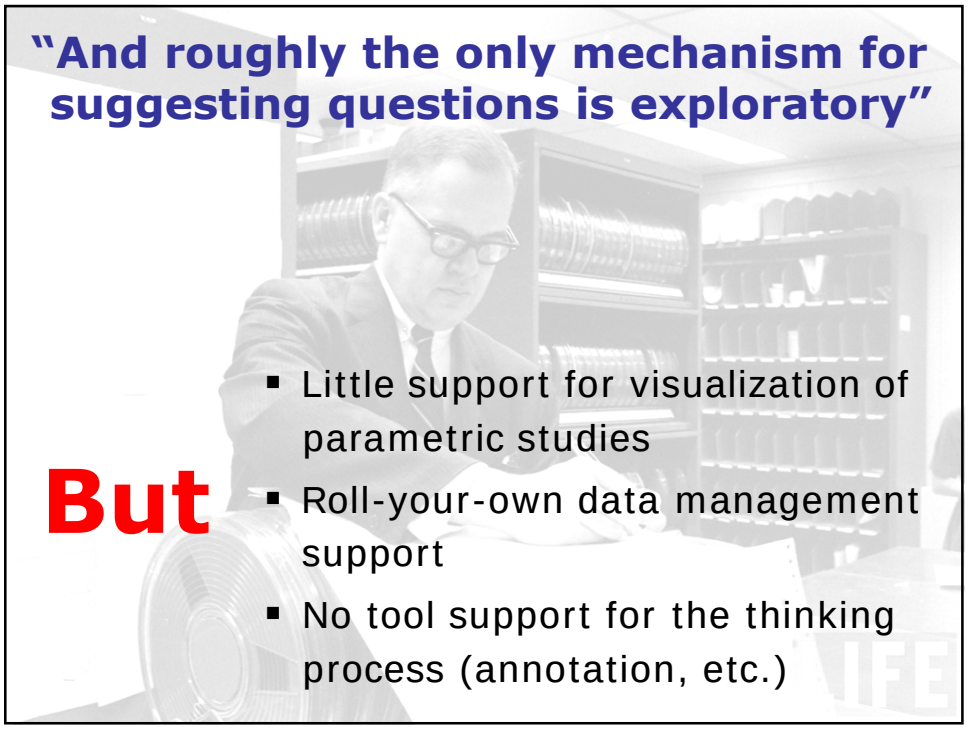
A conversation with John W. Tukey and Elizabeth Tukey
Luisa T. Fernholz and Stephan Morgenthaler
Statistical Science
Volume 15, Number 1 (2000), 79-94

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