The unassigned distance geometry problem applied to find atoms in nanoclusters for sustainable energy

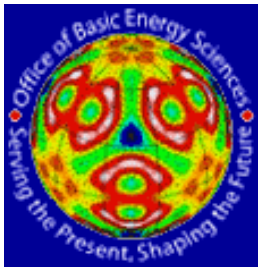
S.J.L. Billinge[1,2] Pavol Juhas[3], Phil Duxbury[3]

[1]*Dept. of Applied Physics and Applied Mathematics, Columbia University*

[2]*CMPMS, Brookhaven National Laboratory (BNL)*

[3]*Center for Data-Driven Discovery, BNL*

[4]*Department of Physics, Michigan State University*



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Columbia University in the City of New York
Brookhaven National Laboratory

National Synchrotron Light Souce-II (NSLS-II) =>

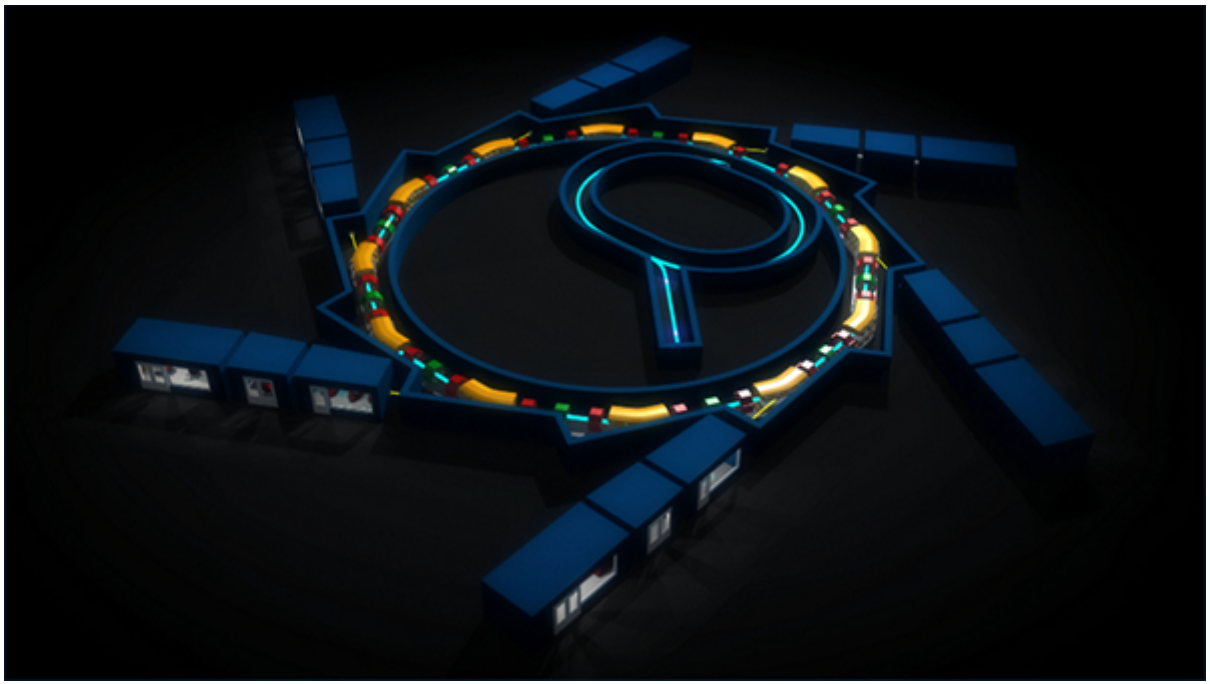
- XPD beamline
- Coherence
- Small beams
- High energy resolution
- Resonant scattering

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

A short side-trip

- Synchrotron: a very intense source of x-rays
- Relativistic electrons (>0.99 c) are “wiggled” and radiate x-rays
- Relativistic squeezing produces a pencil-narrow beam in the direction of travel of the electrons
- We put our samples in that beam



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Why bother?

Materials are the bottleneck to technological solutions to some of mankinds most pressing problems

- Photovoltaics with improved efficiency
  - Nanoparticles in the light collecting layer
- High energy density batteries
  - Electrodes
  - Electrolytes
- Fuel cells for transportation applications
  - Electrodes
  - Electrolytes
  - Catalysts
  - Hydrogen storage
- Sequestration
  - Functionalized mesoporous materials

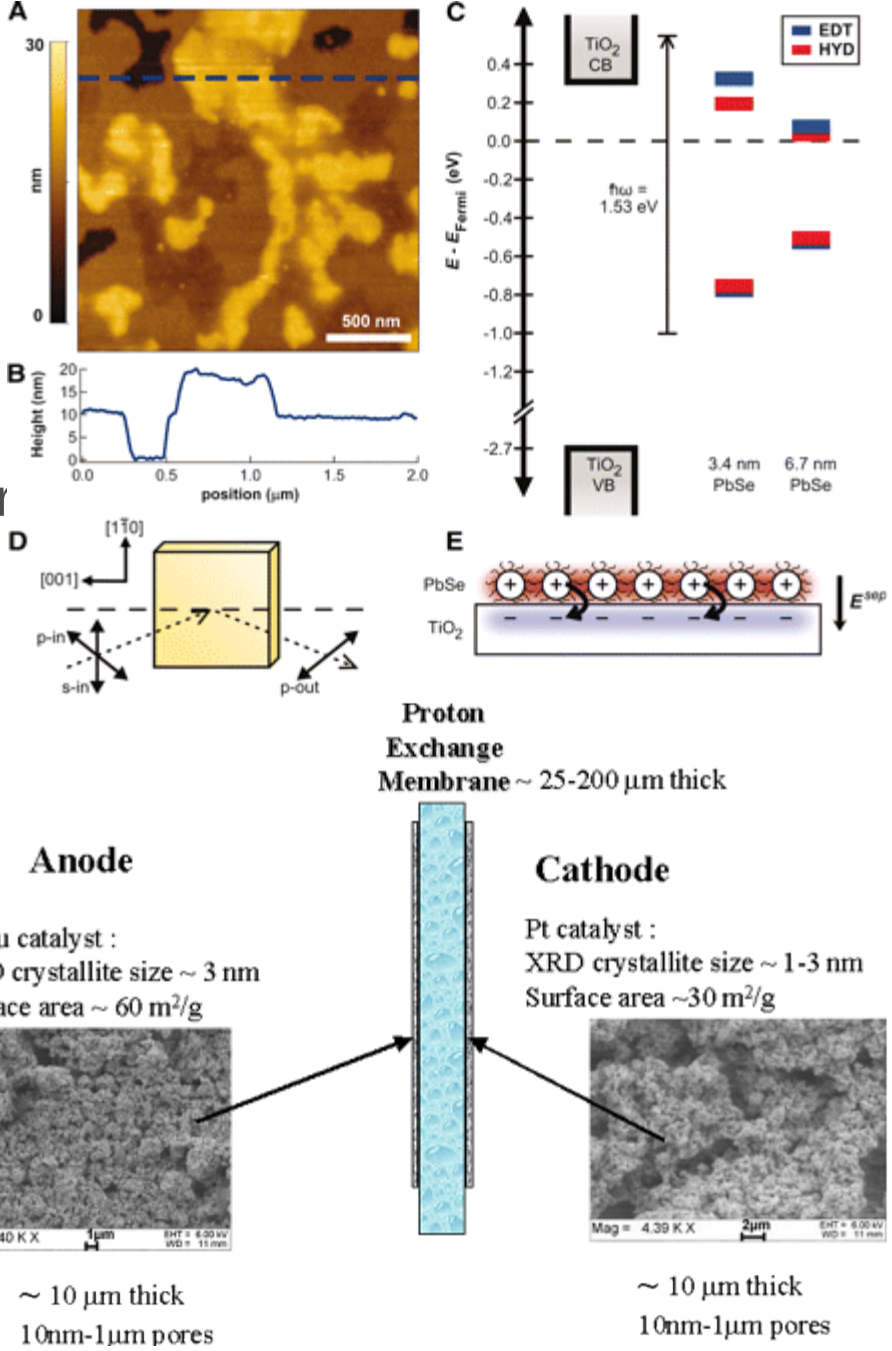


Image credits: 10.1126/science.1185509
U. Uppsala

Structure and properties

Structure has a profound affect on properties

Take pure carbon for example:

- Diamond:
  - hard
  - transparent
  - insulating
  - expensive
- Graphite:
  - soft
  - black
  - metallic (semimetallic anyway)
  - cheap

A big diamond:
Portugese, 31.93 carats

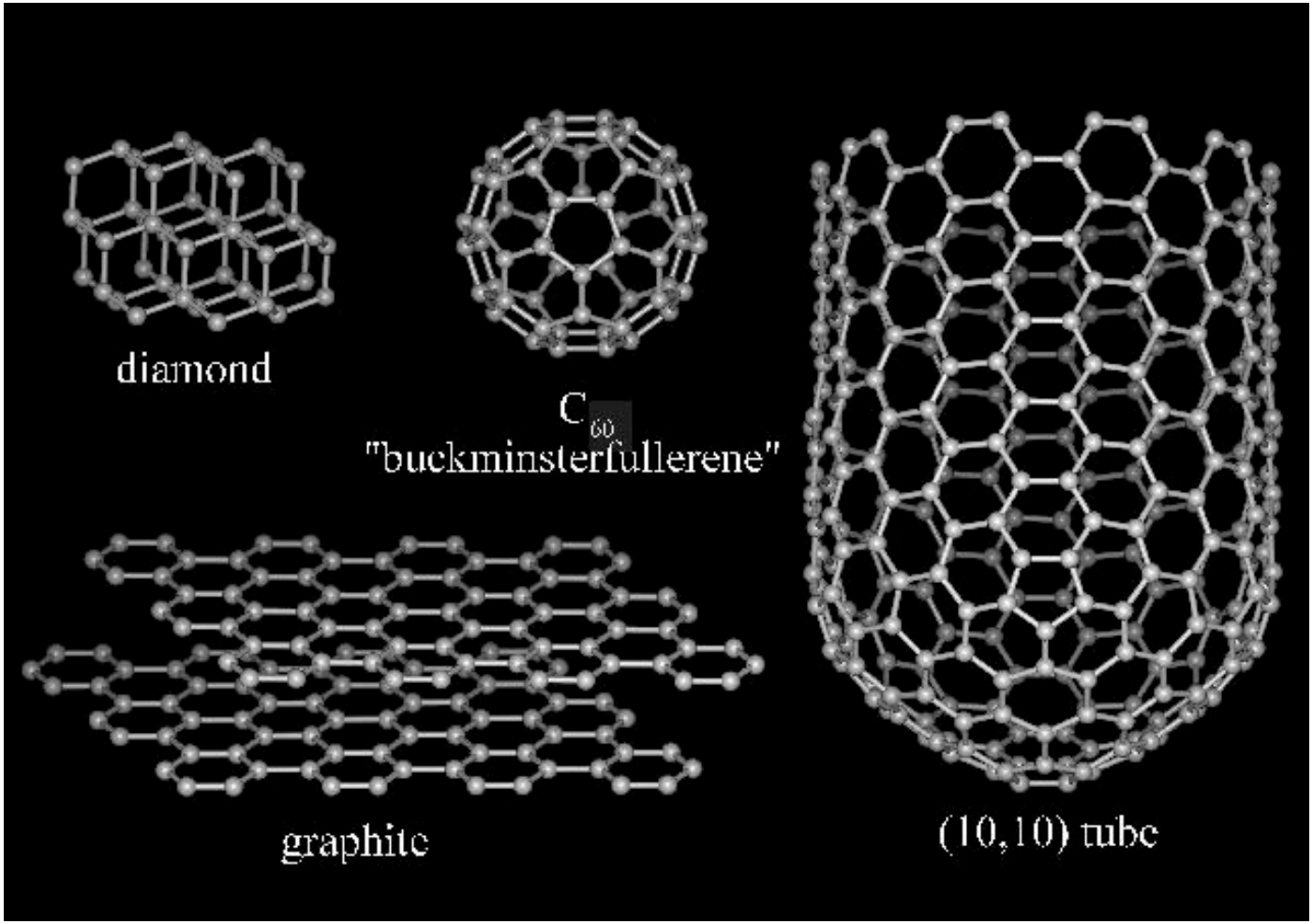
Something big being built out of graphite (don't try this with diamond)

It's all just pure carbon…The difference?

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

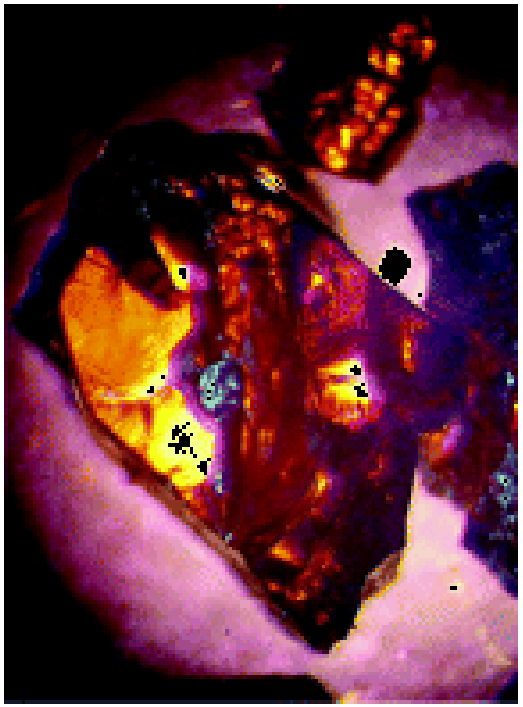
Structure!



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

The Crystal Structure Problem



- Problem:
  - Here is a crystal, what is its structure?
- Solution:
  1. Give it to your grad student
  2. She puts it on the x-ray machine
  3. …Pushes the button
     1. Machine tells you the structure
     2. Or Machine gets stuck
        1. Throw away the crystal
        2. Make it the subject of her thesis

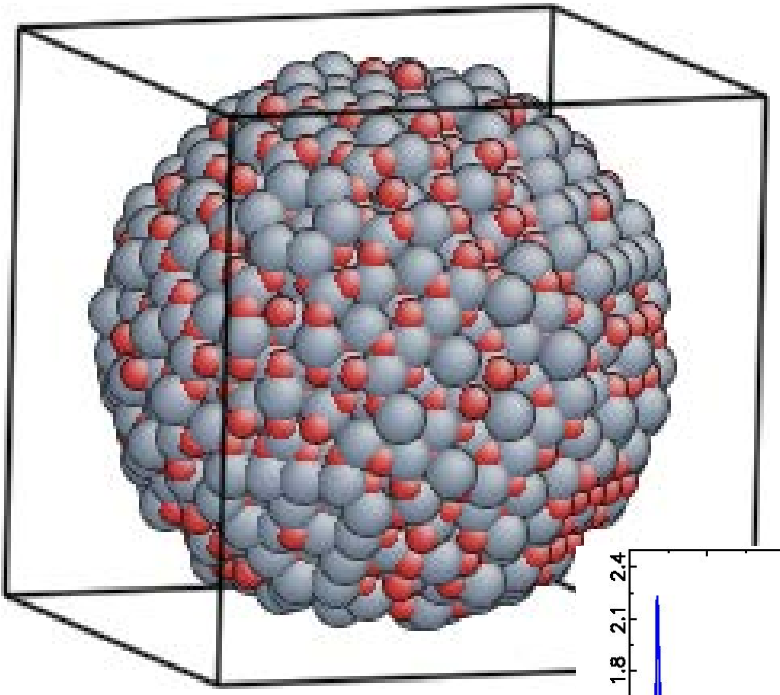
Crystallography is largely a solved problem

From LiGaTe2: A New Highly Nonlinear Chalcopyrite Optical Crystal for the Mid-IR L. Isaenko, et al., J. Crystal Growth, 5, 1325 – 1329 (2005)

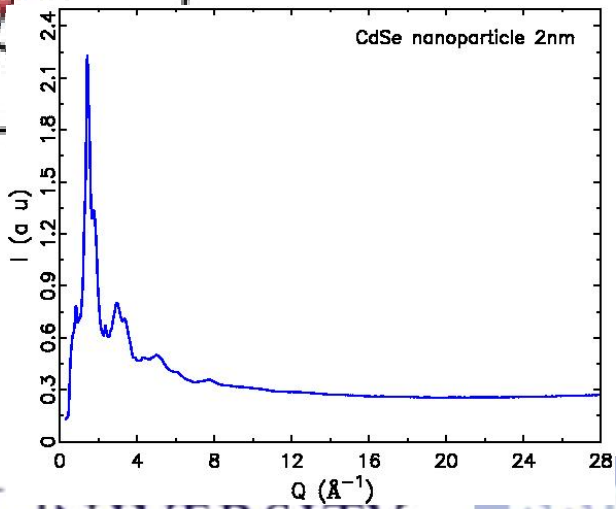
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

The Nanostructure Problem



- Problem:
  - Here is a nanoparticle, what is its structure?
- Solution:
  1. Give it to your grad student
  2. She puts it on the x-ray machine
  3. …Pushes the button



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

Crystal Structure Problem

The crystal structure problem reduces to a “phase retrieval” problem in most cases.

- The phase retrieval problem
  - Recover a general signal, an image, for example, from the magnitude of its Fourier transform
- In crystallography
  - The signal is the amplitudes of a large set of discrete Fourier coefficients in a discrete Fourer series over a periodic basis (the reciprocal lattice)
  - Solution is the contents of a periodically averaged unit cell

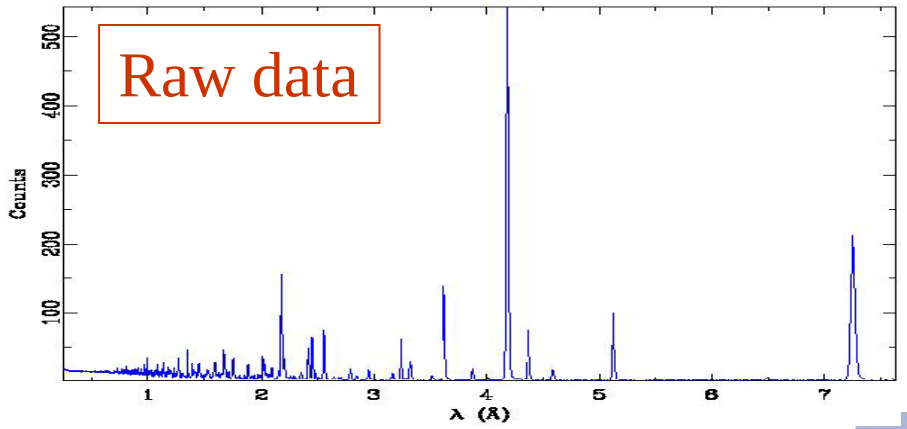
The Nanostructure Problem

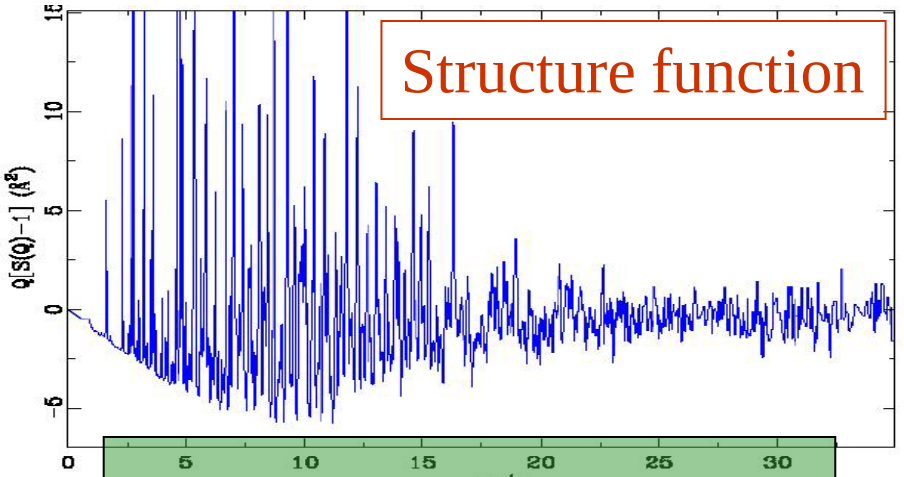
- At the nanoscale crystalloraphy breaks down, the structure is not translationally invariant, but discovery of nanostructure is nonetheless very important
- Potential approaches
- “Single” nanocrystallography
  - Isolate a single (or a few) nanoparticle(s) and take diffraction patterns or atomically resolved images at all different angles
  - Reconstruct the structure using phase retrieval (continuous now) or tomographic reconstruction
- “Powder” nanocrystallography
  - Get a signal from a large number of similar nanoparticles with undetermined orientations
  - Reconstruct the atomic structure from the degraded signal
  - Atomic PDF

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

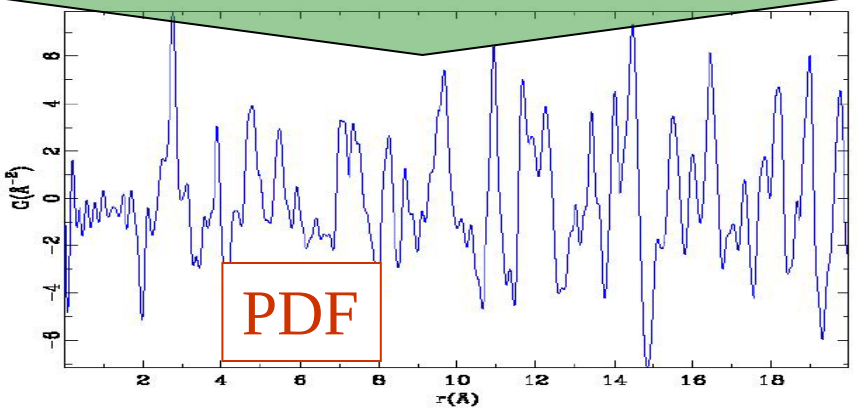
BROOKHAVEN
NATIONAL LABORATORY

The atomic Pair Distribution Function





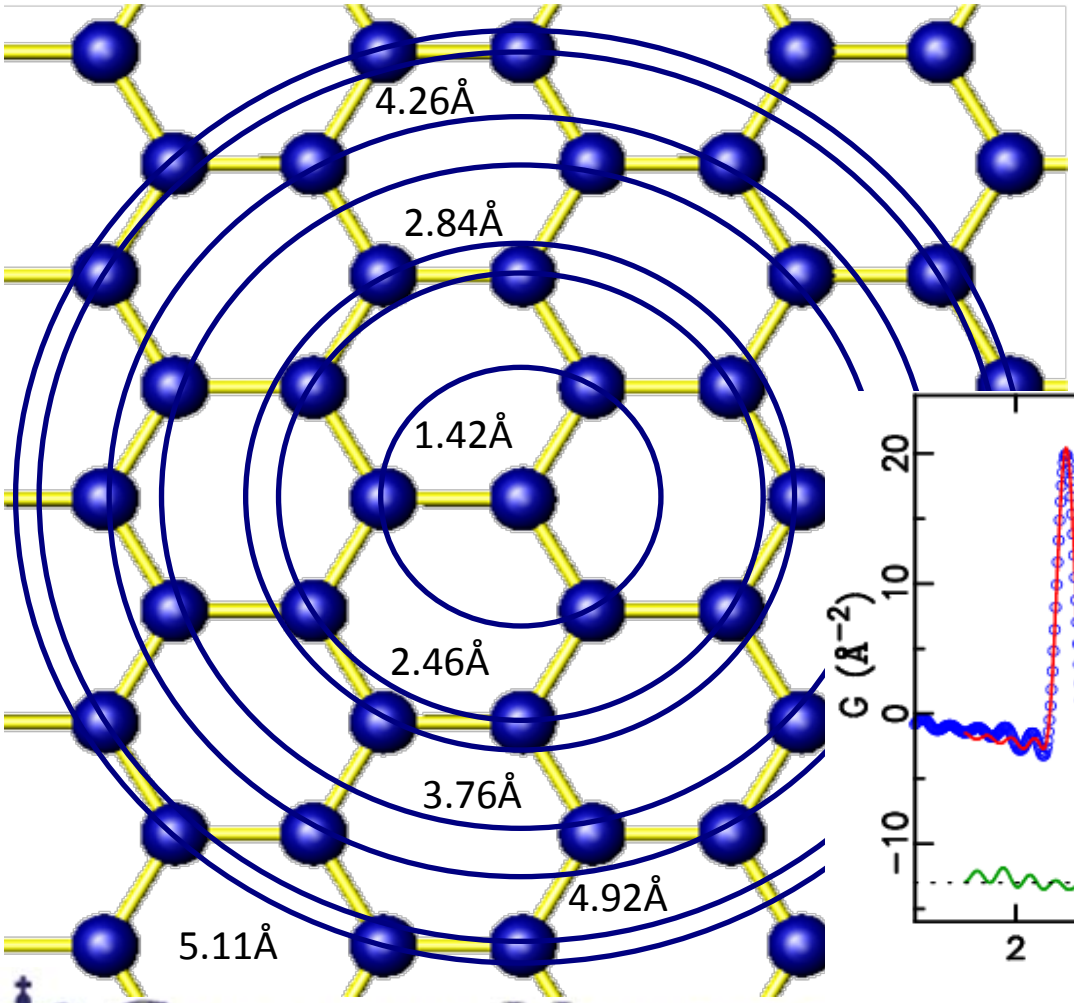
$$G(r) = \frac{2}{\pi}\int_0^\infty Q[S(Q)-1]\sin Qr dQ$$



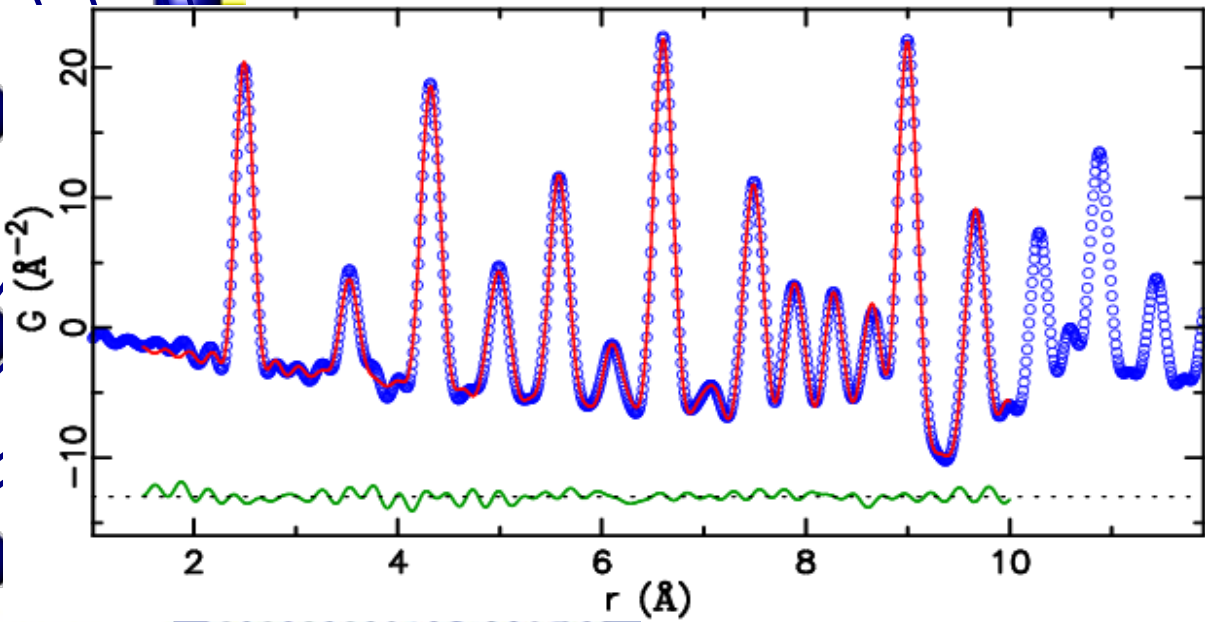
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Nanostructure refinement



Pair distribution function (PDF) gives the probability of finding an atom at a distance “r” from a given atom.



COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

Where does Distance Geometry Come in?

- From Leo Liberti, Carlile Lavor, Nelson Maculan, Antonio Mucherino SIAM review 2014

DISTANCE GEOMETRY PROBLEM (DGP). Given an integer $K > 0$ and a simple undirected graph $G = (V, E)$ whose edges are weighted by a nonnegative function $d : E \to \mathbb{R}_+$, determine whether there is a function $x : V \to \mathbb{R}^K$ such that

$$\forall \{u, v\} \in E, \quad \|x(u) - x(v)\| = d(\{u, v\}). \tag{1.1}$$

This is fine, but it assumes that we know the whole graph, (V,E).

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

Unassigned Distance Geometry Problem

- To be explicit we rename that as the assigned DGP, *aDGP*
- We define a new problem, the unassigned DGP or *uDGP where there is no assignment of vertices to distances*
- This is a much harder problem because the graph structure itself has to be discovered as well as the embedding
- Problem formalized by my collaborator Phil Duxbury (Michigan State University) based on work going back to mid oughties

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

V **204**, pp 117 (2016)

The unassigned distance geometry problem

P.M. Duxbury [a,*], L. Granlund [a], S.R. Gujarathi [a], P. Juhas [b], S.J.L. Billinge [c]

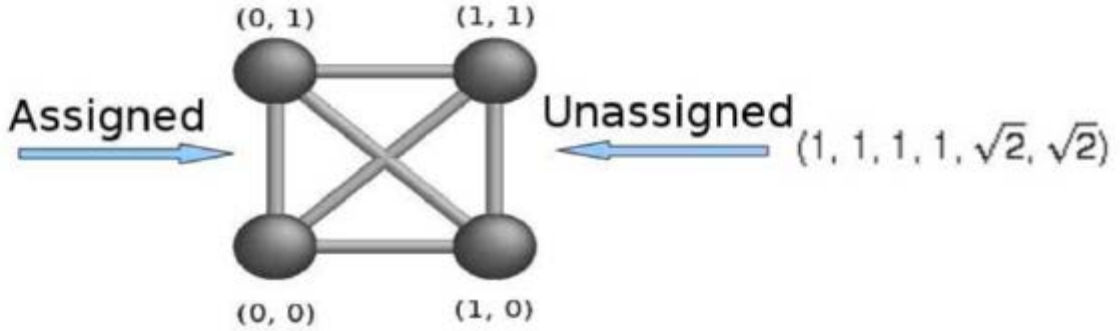
Unassigned DGP

Definition 2 (*Unassigned Distance Geometry Problem (uDGP)*). Given an integer $K > 0$ and a set D of strictly positive distances, determine whether there is an assignment $\alpha : L \to E$, and an embedding $x : V \to \Re^K$ such that,

$$\forall l \in L,\ u, v \in V : \{u, v\} = \alpha(l) \text{ and } \|x_u - x_v\| = d_l. \tag{3}$$

- L is the set of m indices that enumerate the distances, d, in our distance list, D

$$\begin{pmatrix} 0 & 1 & \sqrt{2} & 1 \\ 1 & 0 & 1 & \sqrt{2} \\ \sqrt{2} & 1 & 0 & 1 \\ 1 & \sqrt{2} & 1 & 0 \end{pmatrix}$$



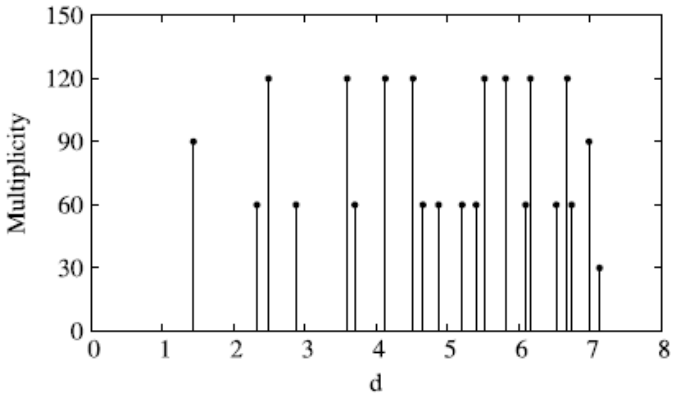
Expression of the uDGP as an optimization problem

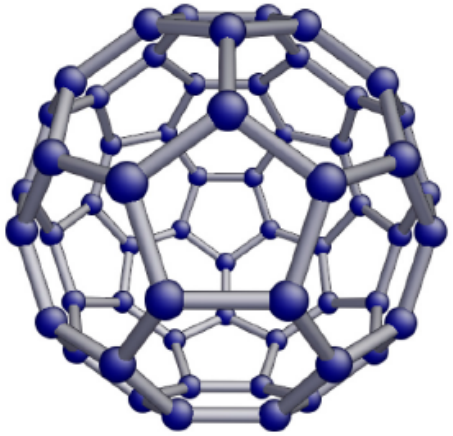
$$\min_{\alpha: L \to E} \left[\min_{X} \sum_{\{u,v\} \in E} f\left(\|x(u) - x(v)\| - d_{uv} \right) \right]$$

- The minimization is over all possible assignments of d_l to $d_{u,v}$ as well as over the placement of vertices, x(u) in $\Re^K$
- $f(y)$ is a convex penalty function
- The α mapping is bijective in the ideal case, but in many real cases, D is not complete, there are missing distances resulting in an injective and non-surjective mapping

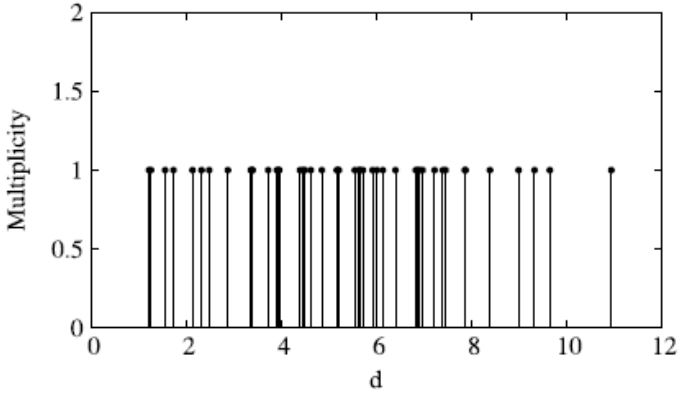
Examples

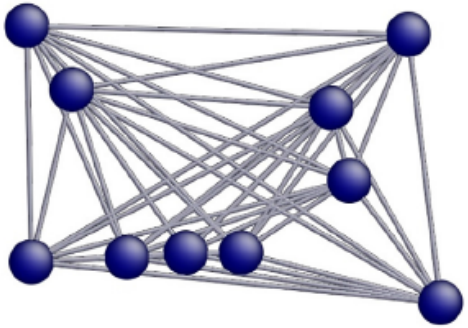
- C_{60} (Buckminsterfullerene)





- Random points in the plane



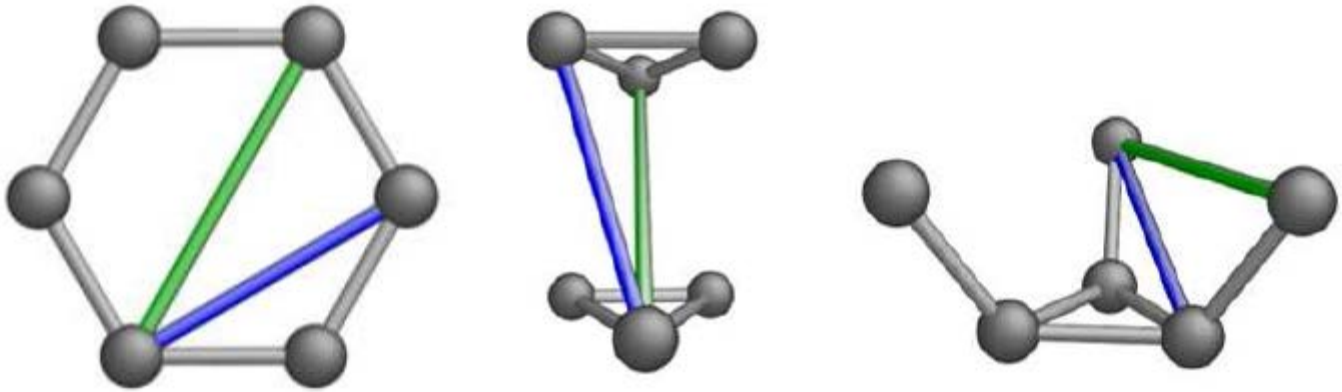


- Note, in the upper plot the edges shown are bonds
- Exactly the info we get from the PDF

Is this problem unique? Can we solve it?

Unique

- For small systems we can find multiple solutions, so it is not unique by inspection



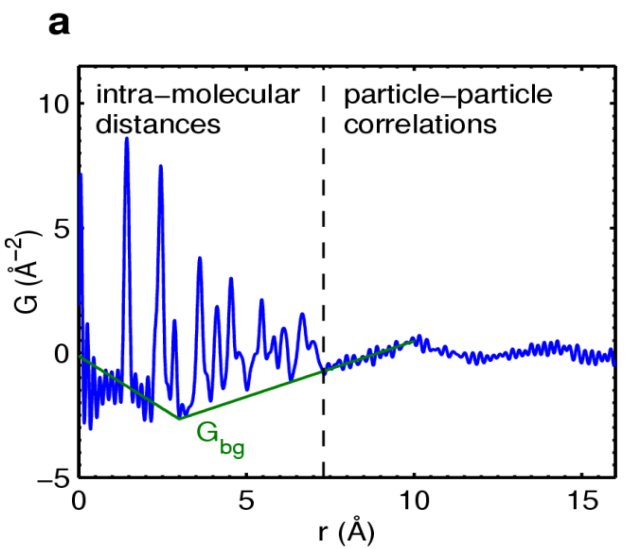
- At least 3 3d embeddings are weakly homometric for the distance list of a planar hexagon
- But these are trivial problems, what about when the problem gets larger?

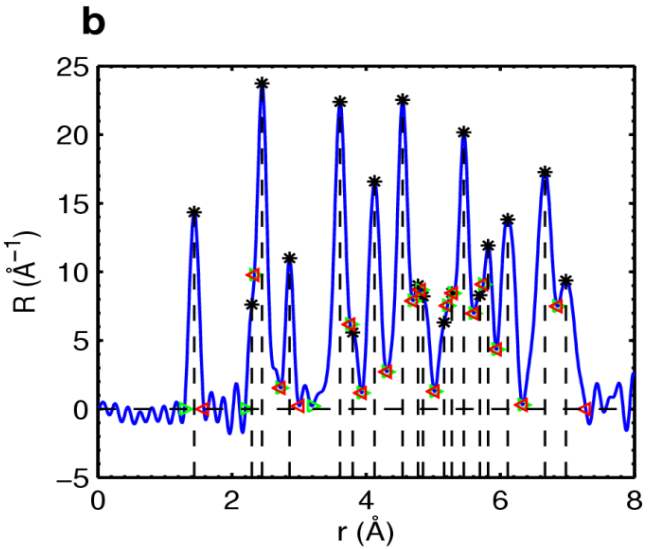
Large systems

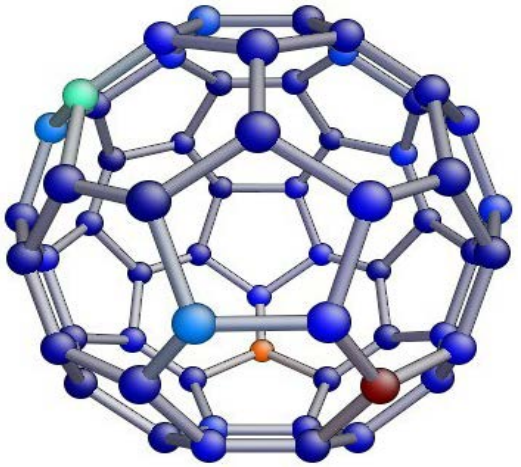
- For large systems we can argue that the probability of multiple solutions is vanishingly small
- In *K* dimensions, there are *nK* translational degrees of freedom for *n* vertices
- A rigid body in *K* dimensions has *K(K+1)/2* translational and rotational degrees of freedom
- If we have a generic graph and we know all the distances exactly we have *n(n-1)/2* distances in our list
- Since $n(n-1)/2 >> nK - K(K+1)/2$ it is highly likely that we will have a unique solution
- OK, let's push on and try and solve it.

Structure determination from PDF

- neutron diffraction PDF data from C_{60}
- 60 atoms, => n(n-1)/2 = 1770 distances
- extracted 18 out of 21 unique distance values
- structure determination still successful







[Juhás et. al, Nature 440, 655-658 (2006)]
[Juhás et. al, Acta Cryst. A 64, 631-640 (2008)]

- algorithm extended for multiple atom-types and periodic boundary conditions

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

SrMise: model independent PDF peak extraction

Community Publications Products Search

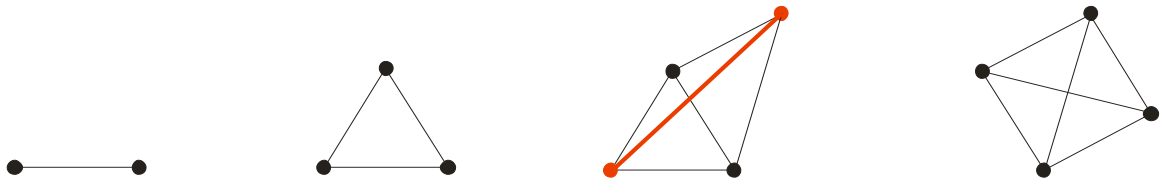
```
from diffpy.srmise.base
from diffpy.srmise.appl

import matplotlib.pyplot

# Load PDF
ppe = PDFPeakExtraction
ppe.loadpdf("si325_mesh_

# Define linear baseline
blf = Polynomial(degree=

# Initialize peak extrac
kwds = {}
kwds["rng"] = [1.5, 15.
kwds["baseline"] = blf.
kwds["dg"] = 0.5
ppe.setvars(**kwds)

# Perform extraction
ppe.extract()

# Plot results
makeplot(ppe)
plt.show()
```

- The PDF equivalent of LeBail/Pawley refinement
- Uses Information Theory to decide what is a peak and what isn't
- Alpha version available
- Granlund, Billinge and Duxbury, Acta A, submitted

SrMise
WWW.DIFFPY.ORG

Illustration of cluster buildup

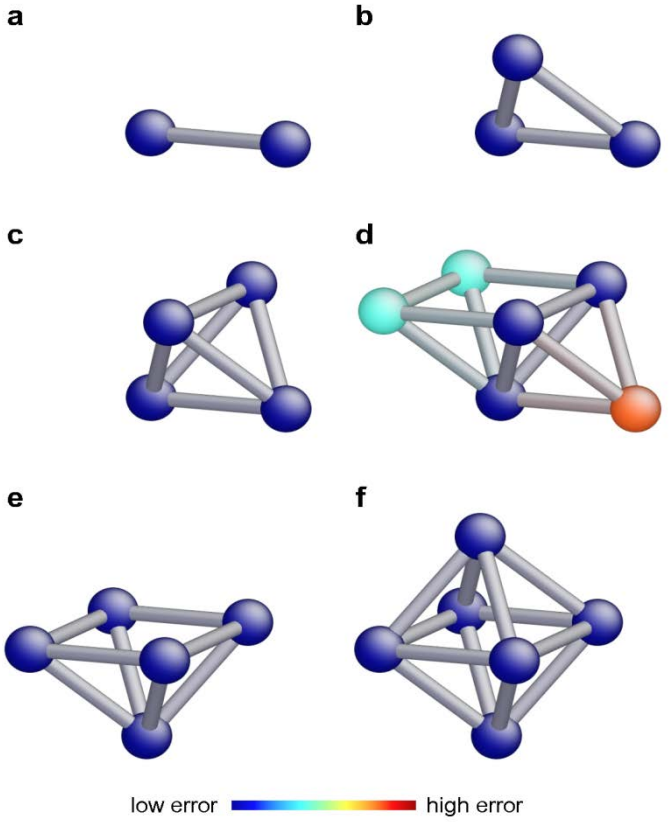
- square-distances = [4×1, 2×sqrt(2)]



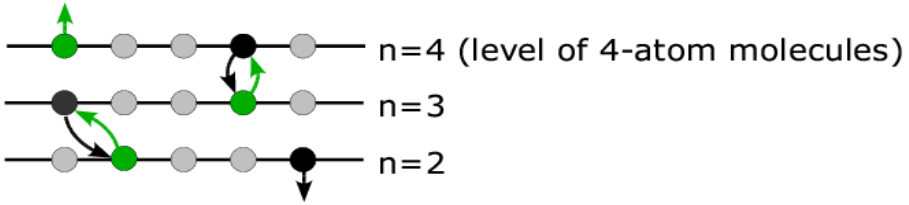
- octahedron-distances = [12×1, 3×sqrt(2)]

- minimized cost function:

$$\mathrm{var}(d) = \frac{1}{P}\sum_{k=1}^{P}[d_k - t_{l(k)}]^2$$



Juhas, SJB et al., Nature 2006



Liga algorithm

Division 1

Division 2

Division 3

Liga algorithm

Division 1




Division 2





Division 3




Division 1






Division 2





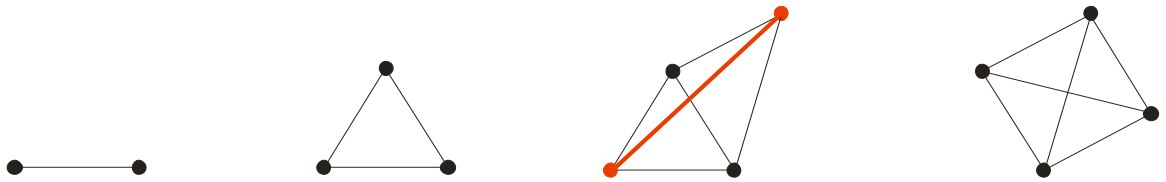
Division 3





Illustration of cluster buildup

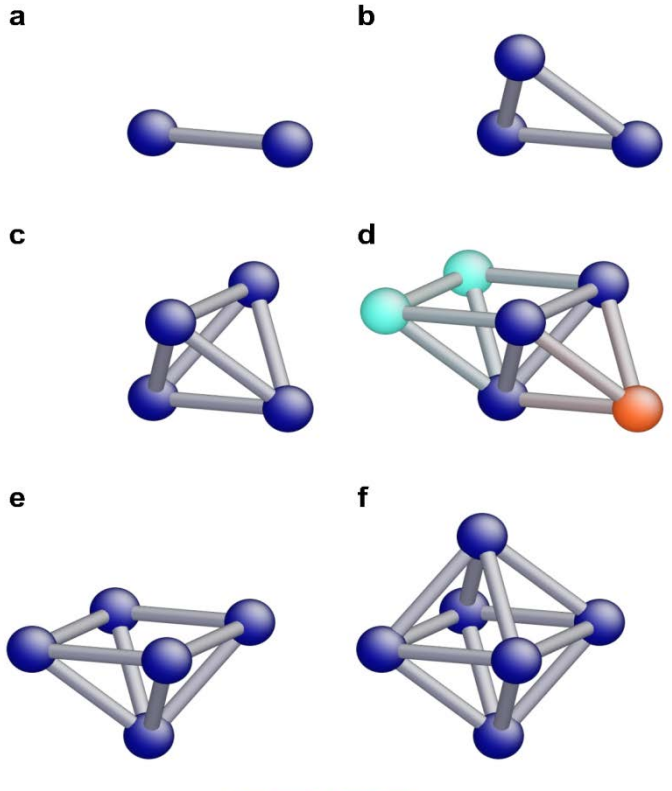
- square-distances = [4×1, 2×sqrt(2)]



- octahedron-distances = [12×1, 3×sqrt(2)]

- minimized cost function:

$$\mathrm{var}(d) = \frac{1}{P}\sum_{k=1}^{P}[d_k - t_{l(k)}]^2$$

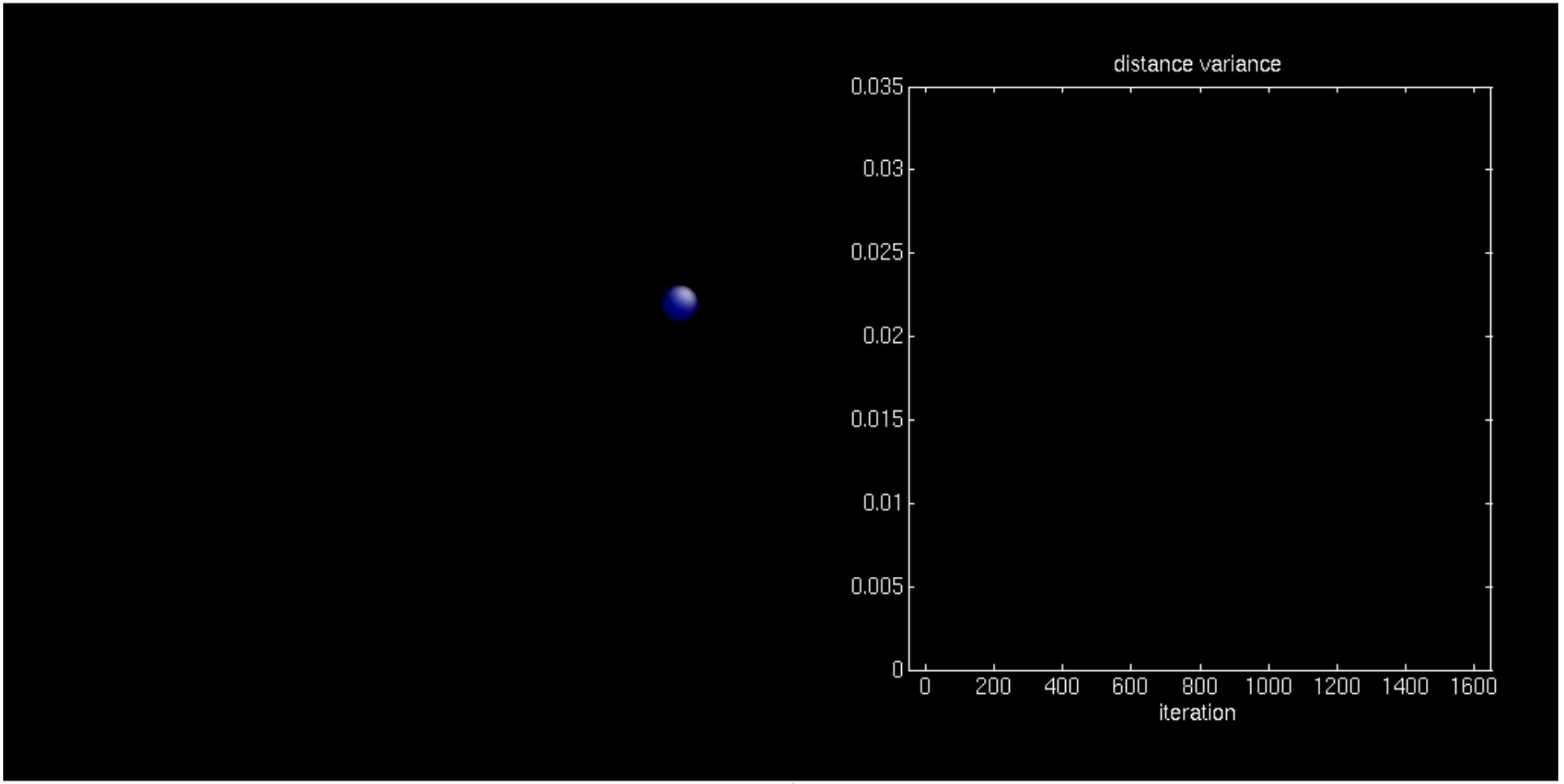


Juhas, SJB et al., Nature 2006

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

ab-initio structure solution directly from PDF data

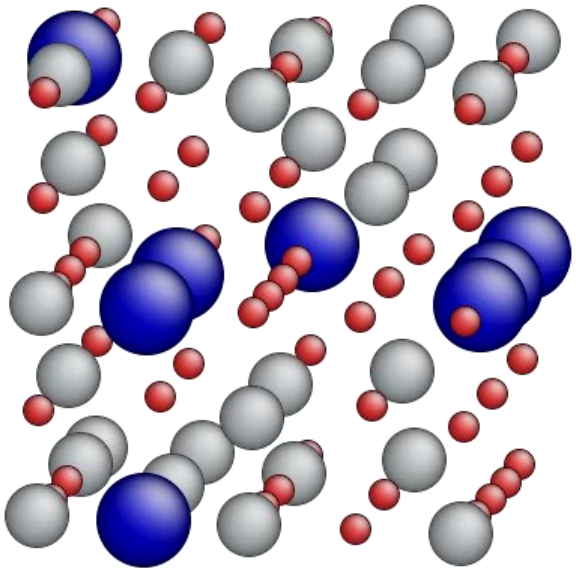


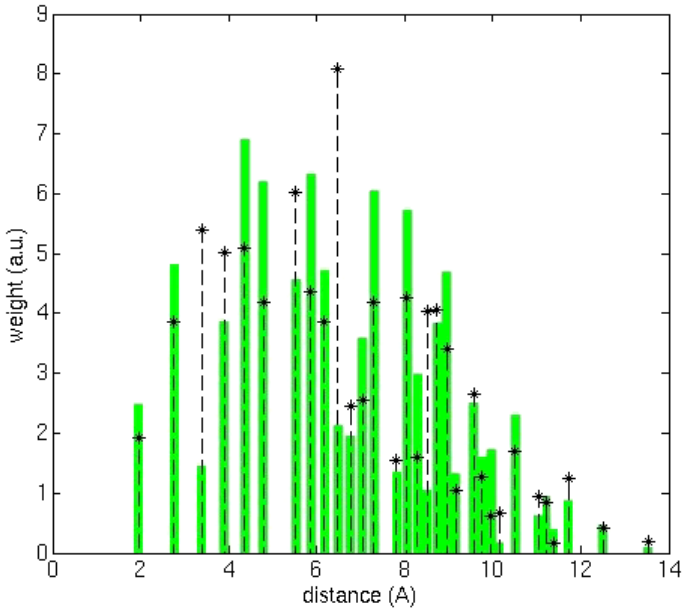
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Solving the coloring problem

- cutout from $SrTiO_3$ - 89 sites containing 8 Sr, 27 Ti, 54 O
- site assignment can be solved from ideal PDF
- downhill search:
  - start with random site arrangement
  - flip sites which improve match between model and ideal peak weights





Journal of
Applied
Crystallography

ISSN 0021-8898

Received 20 October 2009
Accepted 16 March 2010

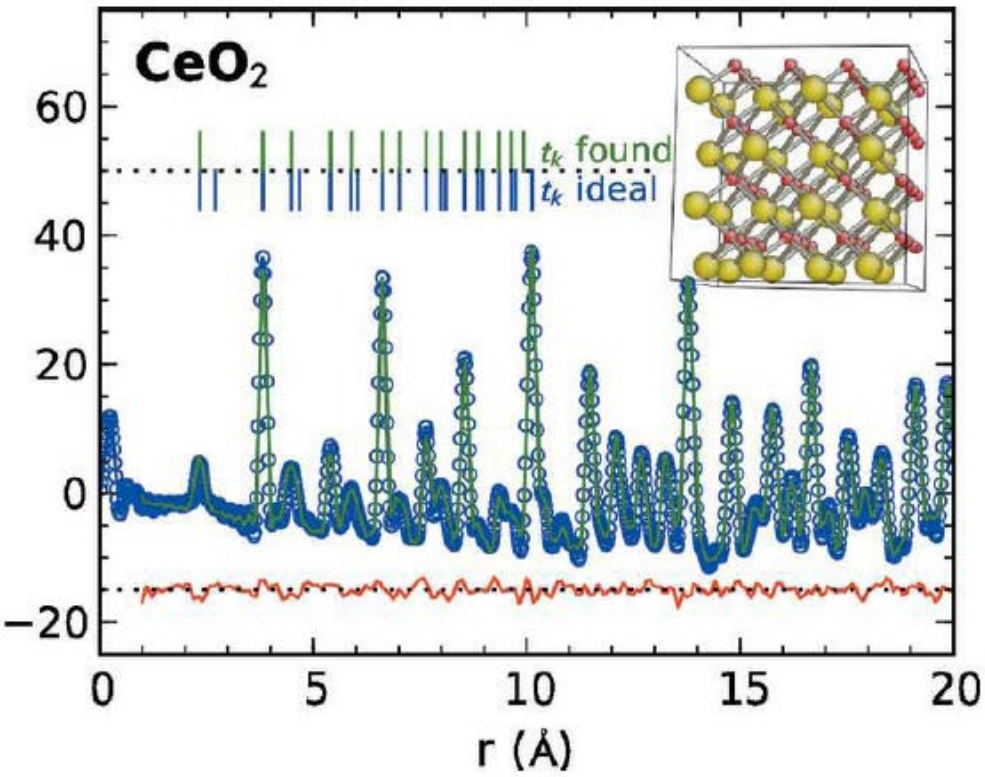
Crystal structure solution from experimentally determined atomic pair distribution functions

P. Juhás,[a]* L. Granlund,[b] S. R. Gujarathi,[b] P. M. Duxbury[b] and S. J. L. Billinge[a,c]

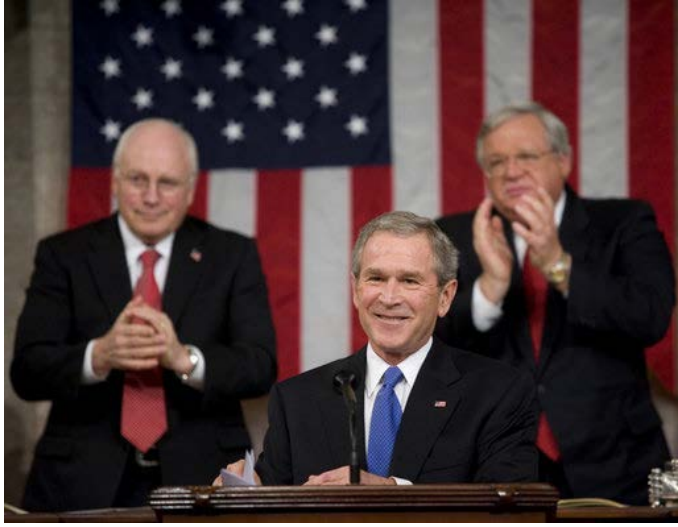
[a]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA, [b]Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA, and [c]Condensed Matter Physics and Materials Science Department, Brookhaven

C_d and C_c are the distance and atom-overlap costs, as defined in equations (3) and (4). s_x, s_y and s_z are the standard deviations in the f
normalized to a simple [111] cell. s_r (Å) is the root mean-square displacement of the solved sites from the reference CIF positions.

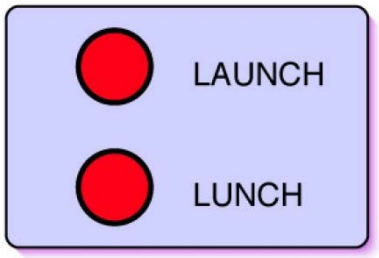
Sample (supercell)	Atoms	Cost C_d (0.01 Å^2) Liga	CIF	Cost C_c (Å^2) Liga	CIF	Deviation of coordinates s_x	s_y	
Successful solutions								
Ag [111]	4	0.0232	0.136	0	0.001	0	0	
Ag [222]	32	0.0097	0.136	0	0.001	0.00025	0.00024	
$BaTiO_3$ [111]	5	0.370	0.394	0.040	0.042	0.0057	0.0066	
$BaTiO_3$ [112]	10	0.392	0.394	0.058	0.042	0.00023	0.039	
C graphite [111]	4	0.396	0.574	0.010	0.016	0.0029	0.0029	
C graphite [221]	16	0.420	0.574	0.010	0.016	0.0086	0.0065	
CdSe [111]	4	0.107	0.138	0	0.001	0	0	
CdSe [221]	16	0.0856	0.138	0	0.001	0.00010	0.00013	
CeO_2 [111]	12	0.515	0.554	0	0	0	0	
NaCl [111]	8	1.75	1.71	0	0	0	0	
NaCl [222]	64	1.20	1.71	0	0	0.00031	0.00031	
Ni [111]	4	0.0024	0.0024	0	0	0	0	
Ni [222]	32	0.0025	0.0024	0	0	0.00015	0.00013	
PbS [111]	8	0.0125	0.0104	0.010	0.011	0	0	
PbS [222]	64	0.0140	0.0104	0.010	0.011	0.00005	0.00004	
PbTe [111]	8	0.0024	0.0127	0.097	0.090	0	0	
PbTe [222]	64	0.0022	0.0127	0.097	0.090	0.00011	0.00011	
Si [111]	8	0.0045	0.0045	0	0	0	0	
Si [222]	64	0.0048	0.0045	0	0	0.00010	0.00009	
$SrTiO_3$ [111]	5	0.437	0.437	0.002	0.002	0	0	
Zn [111]	2	0.495	0.470	0	0	0	0	
Zn [222]	16	0.564	0.470	0	0	0.00010	0.00006	
ZnS sphalerite [111]	8	0.150	0.0647	0	0	0	0	
ZnS sphalerite [222]	64	0.160	0.0647	0	0	0.00029	0.00033	
ZnS wurtzite [111]	4	0.141	0.152	0	0	0	0	
ZnS wurtzite [221]	16	0.165	0.152	0	0	0.00003	0.00002	
Failed solutions								
$CaTiO_3$ [111]	20	0.4967	0.902	0.52	0.072	0.16	0.14	
TiO_2 rutile [111]	6	0.5358	0.758	0.40	0.009	0.081	0.24	



Software Projects that go wrong



- Who: US department of homeland security
- What: Develop a GUI for the US president to undertake his most important functions
- Functional Requirements: Must be very simple and easy to use
- Solution:



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Can graph theory help?

A survey on assigned and unassigned Distance Geometry: applications to Nanostructures and Biological Molecules

Simon J.L. Billinge[1] · Phillip M. Duxbury[2] · Douglas S. Gonçalves[3] · Carlile Lavor[4] · Antonio Mucherino[5]

- 4OR, to appear(?)

- Worth noting we are not the first to work on this:

S. Skiena, W. Smith, and P. Lemke. Reconstructing sets from interpoint distances. *in Sixth ACM Symposium on Computational Geometry*, pages 332–339, 1990.

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

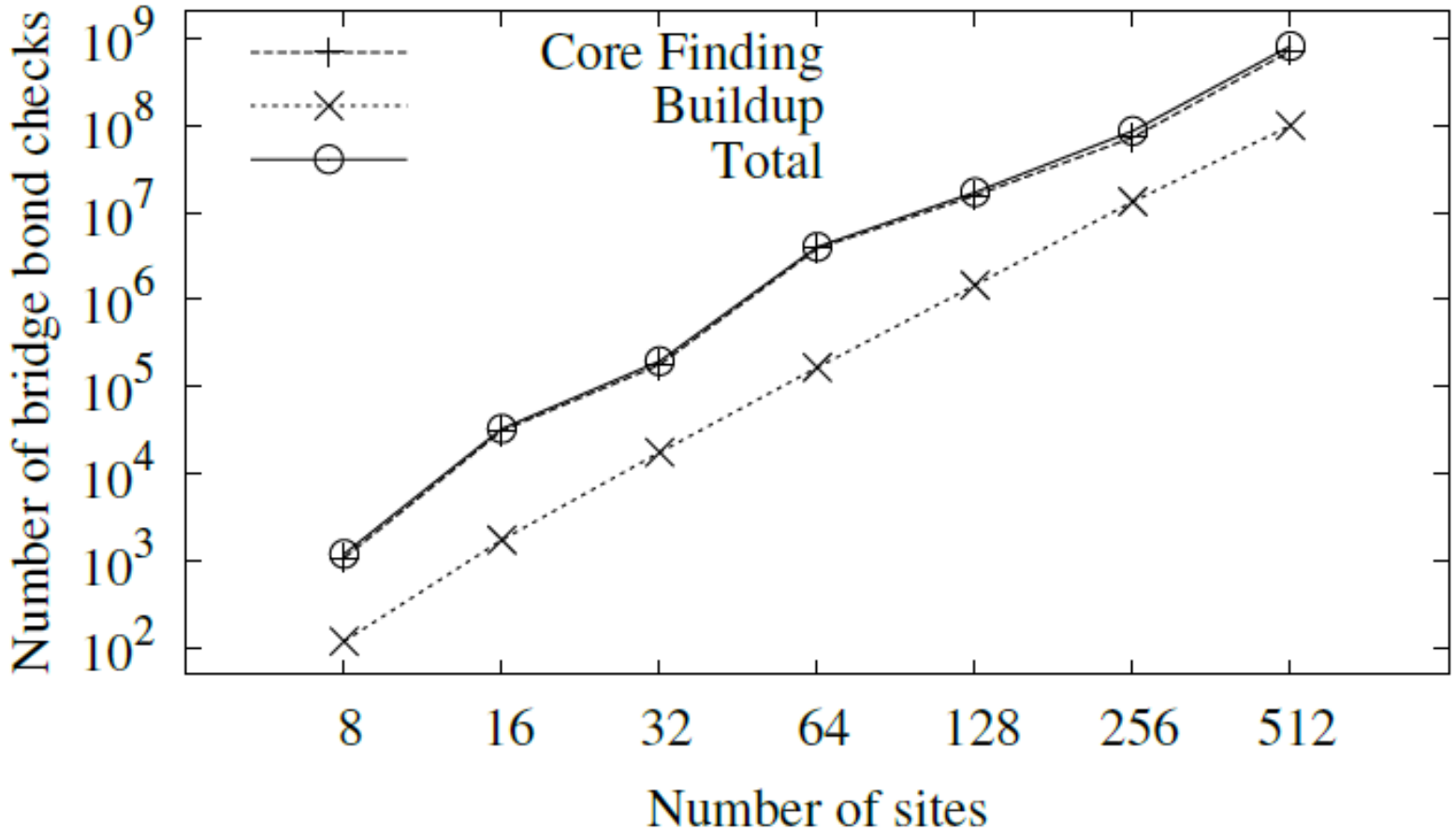
BROOKHAVEN NATIONAL LABORATORY

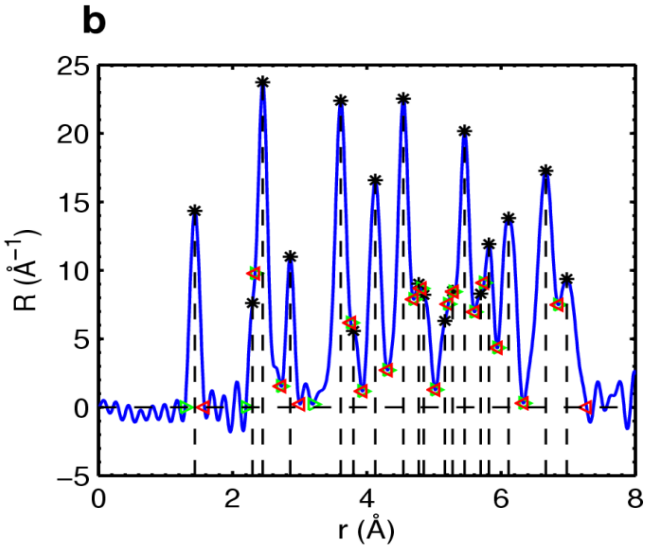
Can Graph theory help?

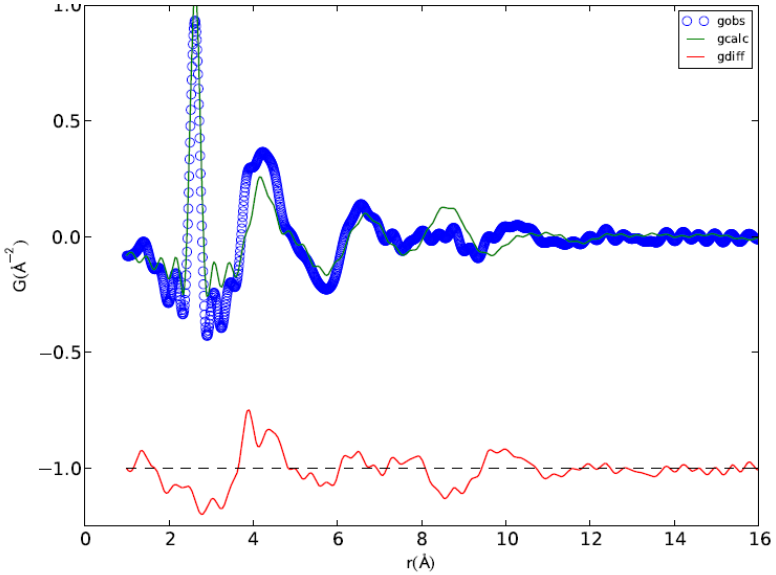
- Use Generic Graph Rigidity
- The condition for a unique graph representation is that the graph is globally rigid, each edge is redundant and the kernel of the stress matrix has dimension *K+1*
- Stress matrix: $\mathcal{S}(\mathbf{r}) = \sum_{edges} \omega_{ij}(\mathbf{r}_i - \mathbf{r}_j)^2,$ satisfying equilibrium $\sum_j \omega_{ij}(\mathbf{r}_j - \mathbf{r}_i) = 0.$
- Use combinatorial algorithms based on Laman's theorem that test for global rigidity
- => develop graph build-up methods, testing for global rigidity at each step of the buildup, globally rigid buildup (GRB) methods, can greatly reduce the search-space of viable solutions.
- LIGA is an example of a GRB and is highly efficient
- For an ideal case of generic graph with exact distances: TRIBOND (Phil Duxbury) is a deterministic GRB and polynomial

Tribond

- Core-finding is slow, buildup is fast
- Works in 2D. Extension to 3D looks promising







60 atoms

~64 atoms

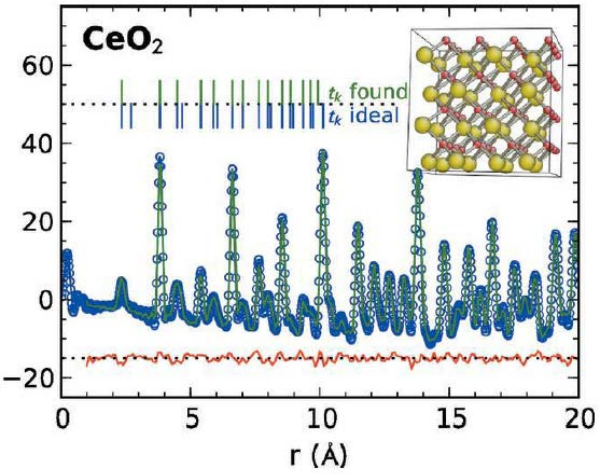
C_{60}

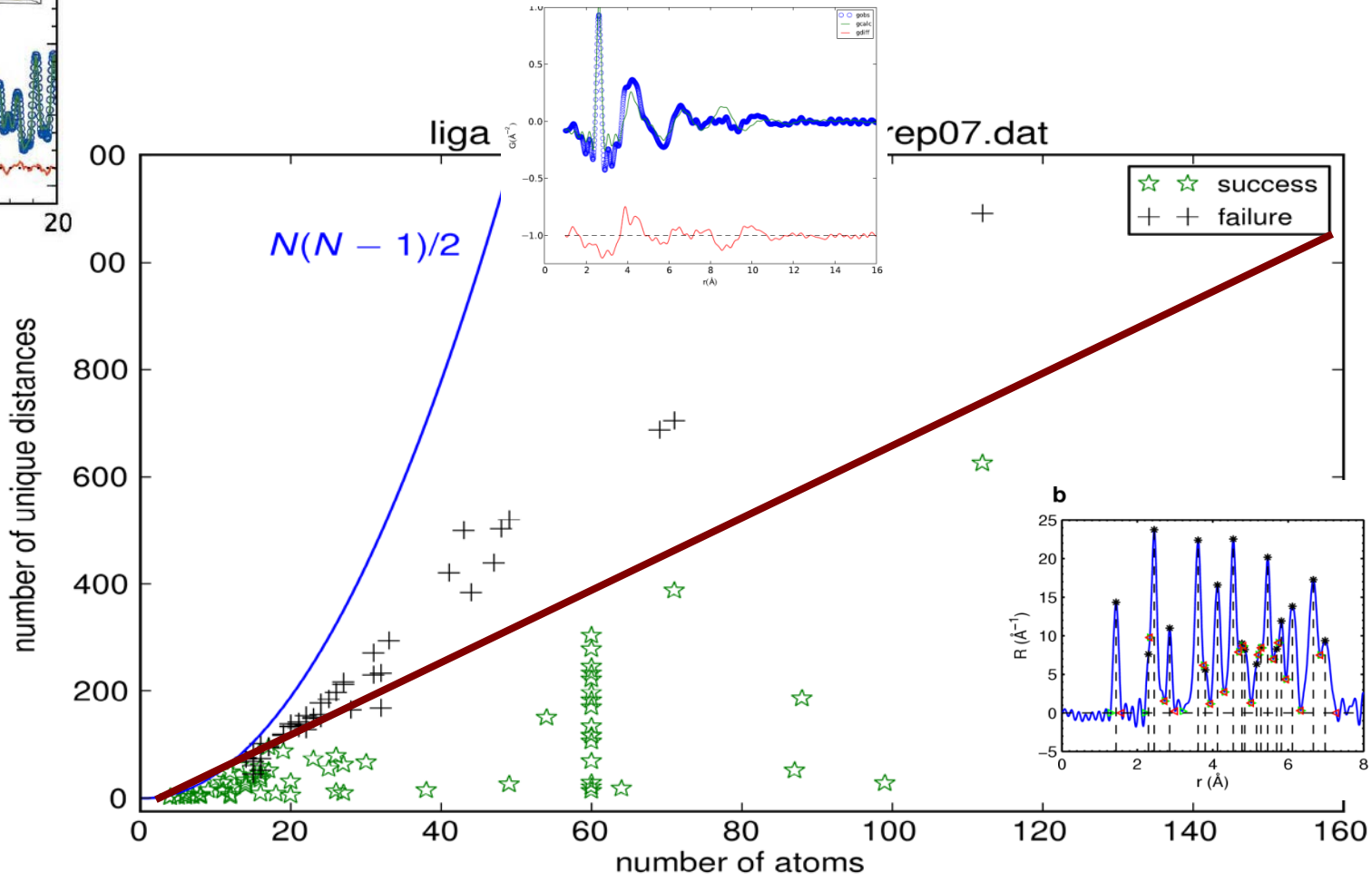
Ultra-small CdSe NPs

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

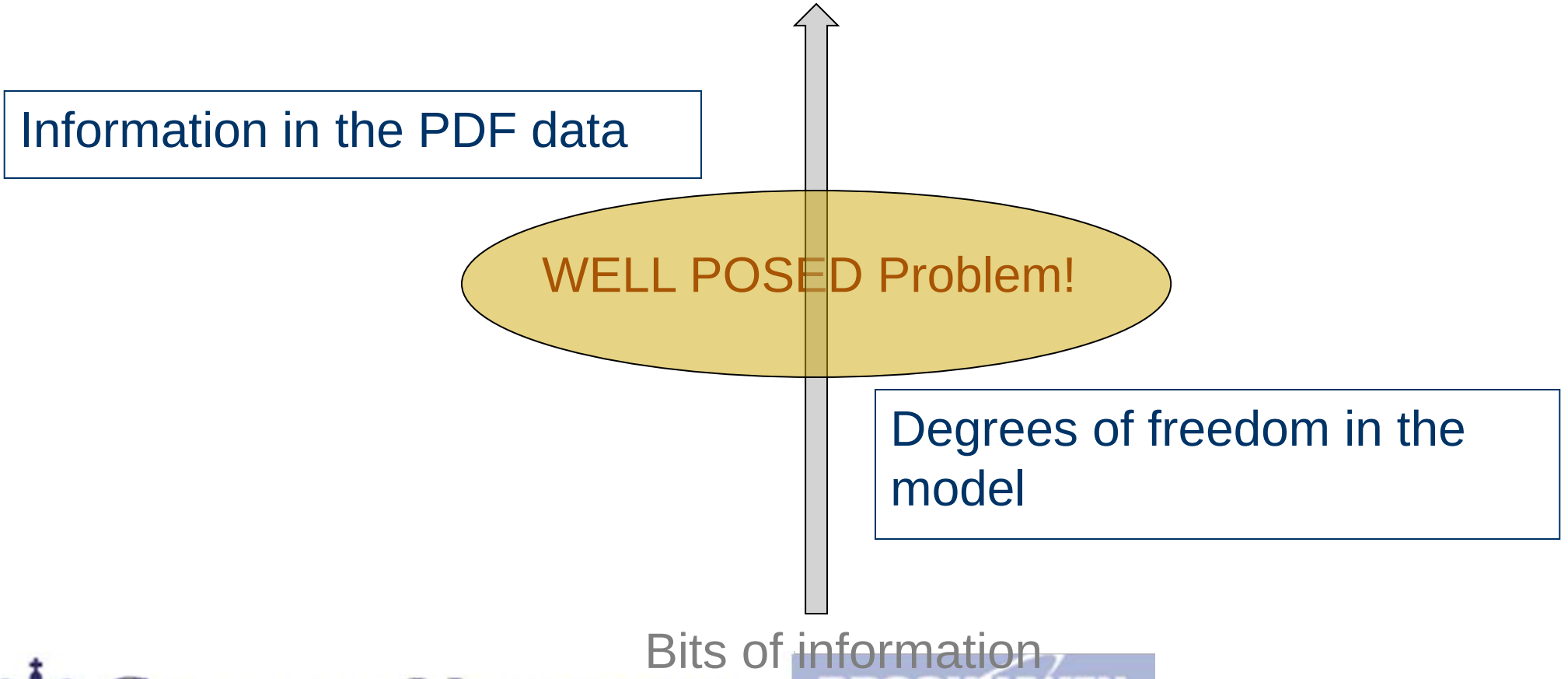
Successology





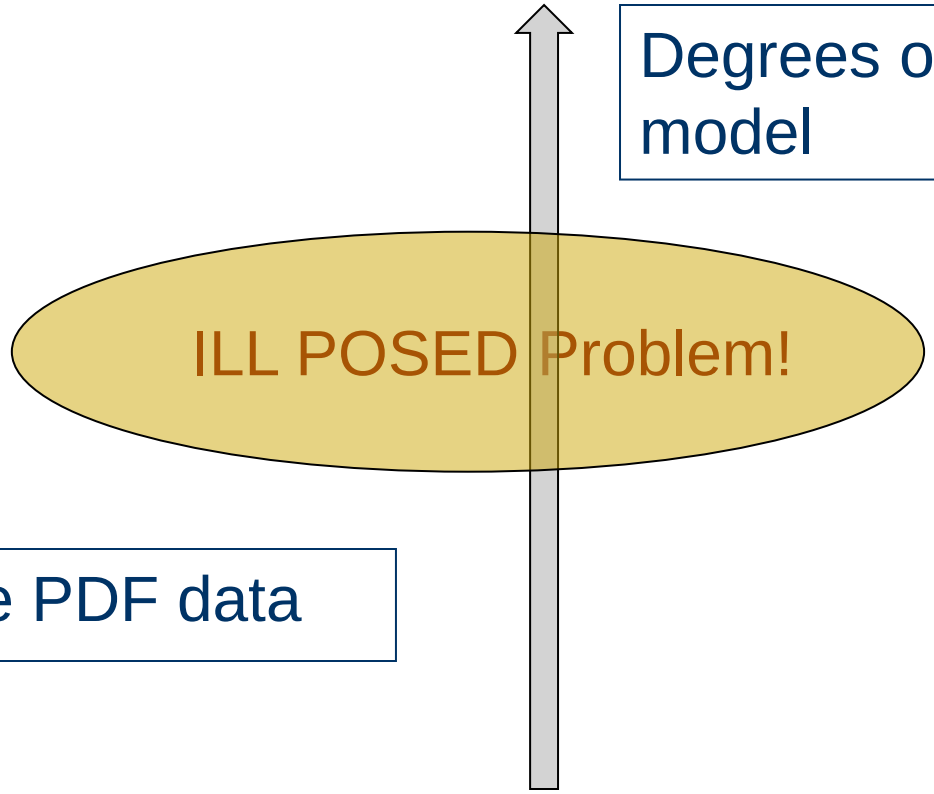
Problem

Well posed problem:



Problem

Ill posed problem:



Structure Solution

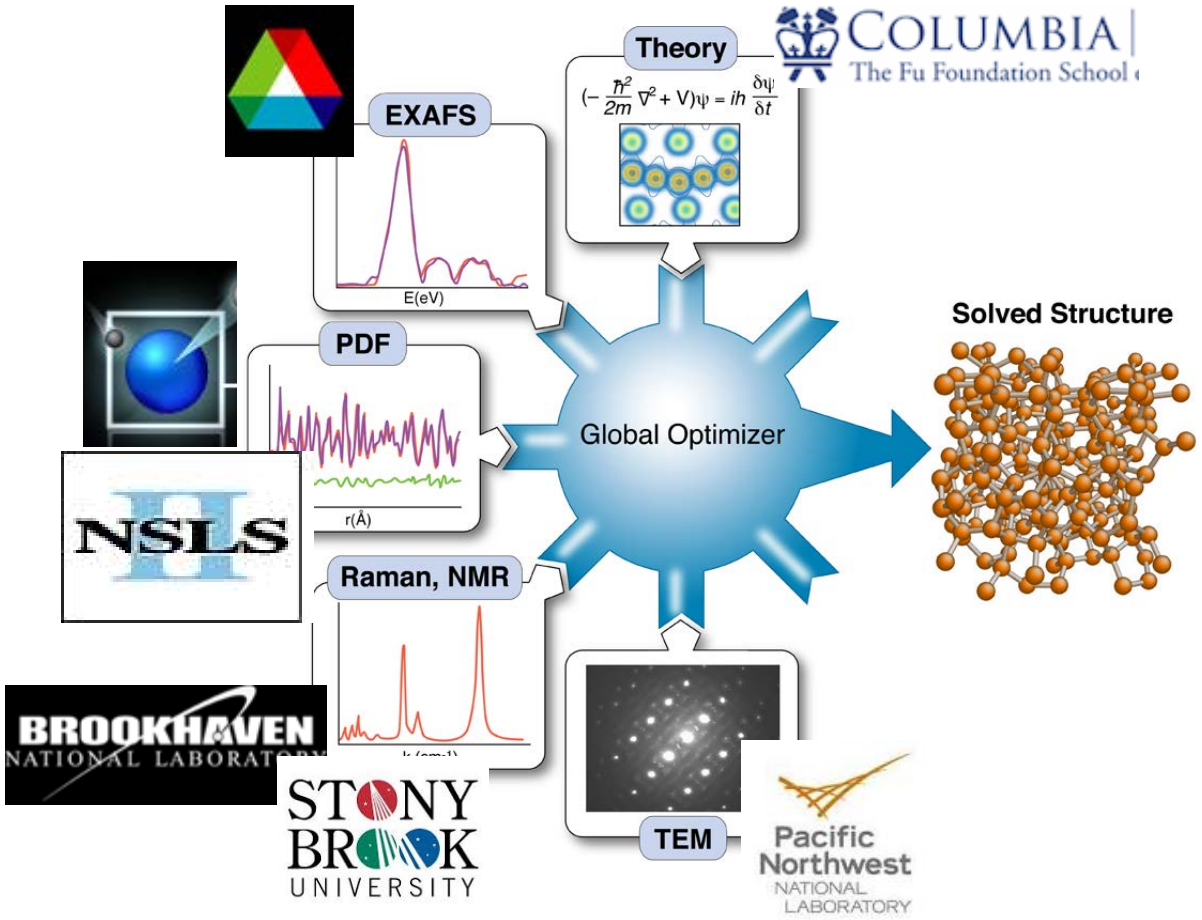
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Complex Modeling Solution

- c = a + ib – complex number mixes real and imaginary parts
- m = e + it – complex modeling mixes experiment and theory in a coherent computational framework

- Billinge and Levin, Science 2007



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Diffpy project (BNL LDRD)
Complex Modeling infrastructure: Diffpy-CMI

Official release of Diffpy-CMI v0.1

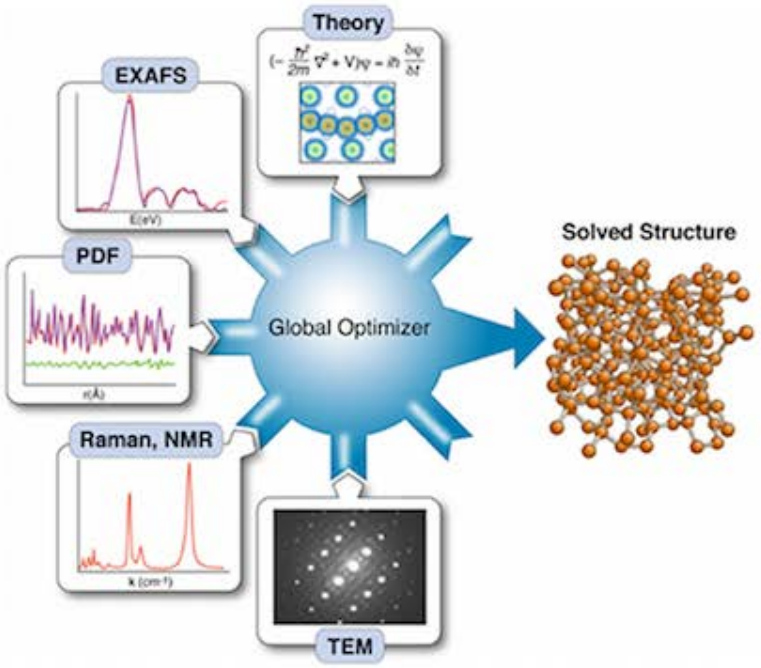
www.diffpy.org

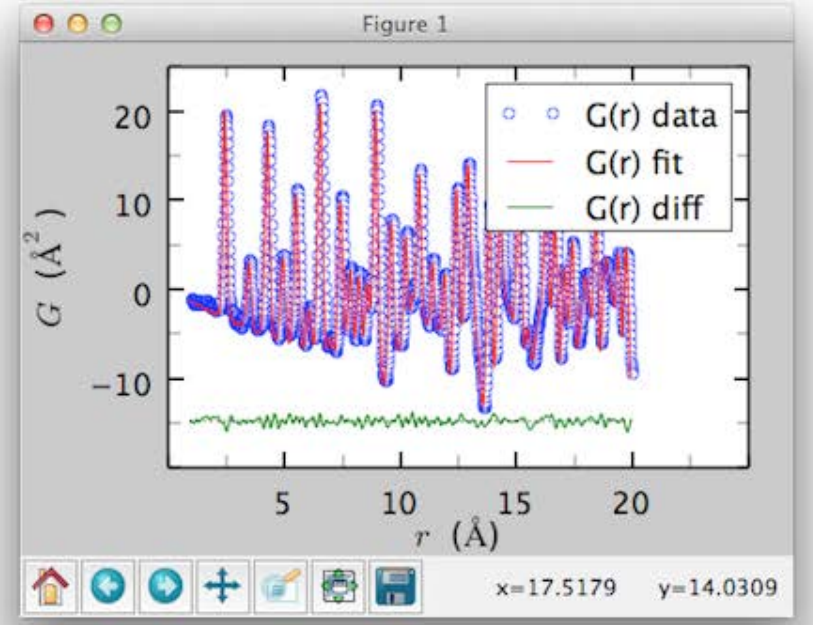
DiffPy Community Publications Products Search

DiffPy-CMI is now available! Get DiffPy-CMI Credits

DiffPy - Atomic Structure Analysis in Python

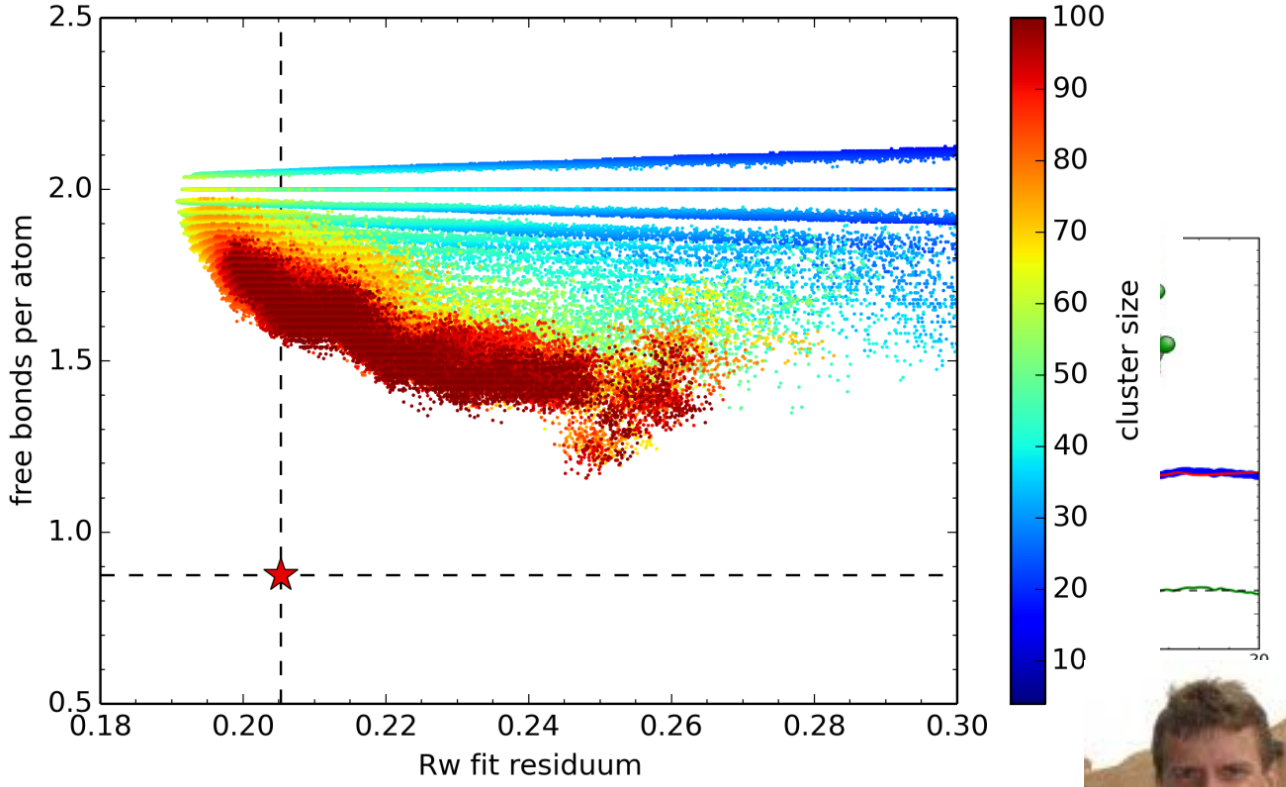
A free and open source software project to provide python software for diffraction analysis and the study of the atomic structure of materials.





To test this idea let's consider a rather well-defined problem

- Example: CdSe quantized growth nanoparticles.
- X-ray PDF analysis alone does not produce a unique solution.
- Complex Modeling approach is required.



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

Plea

We need some Applied Math help!

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY