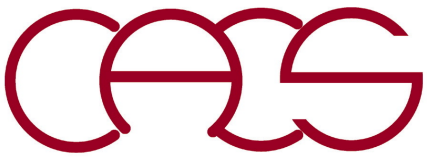
Scientific Data Mining & Machine Learning

Aiichiro Nakano

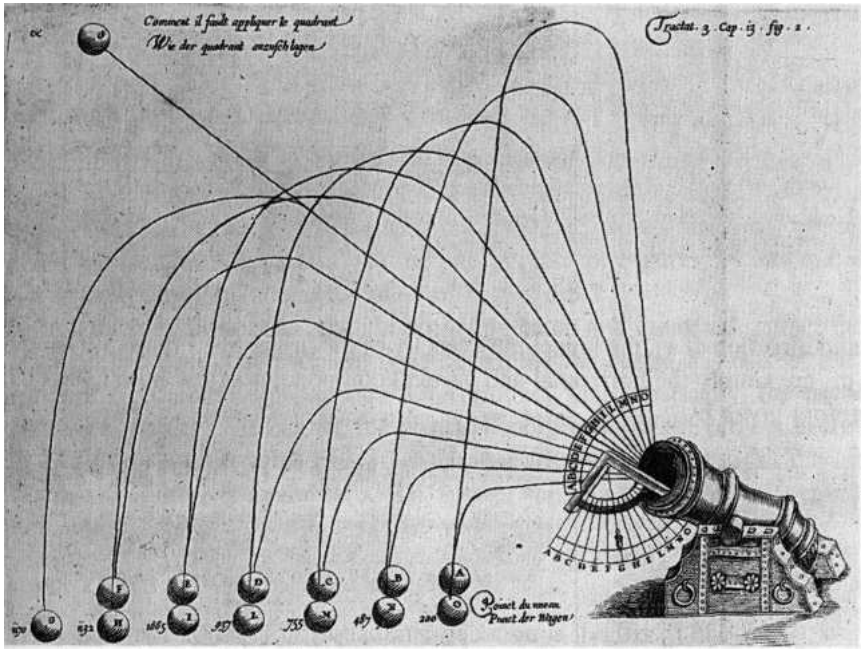
Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy
Department of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu



Scientific Data Mining

- **Scientific data mining:** Automated detection of knowledge hidden in large & often noisy scientific (experimental, simulation, *etc*.) datasets
- **Knowledge:** Simplest (*i.e*., minimal description length) explanation to replace exhaustive enumeration of the original data



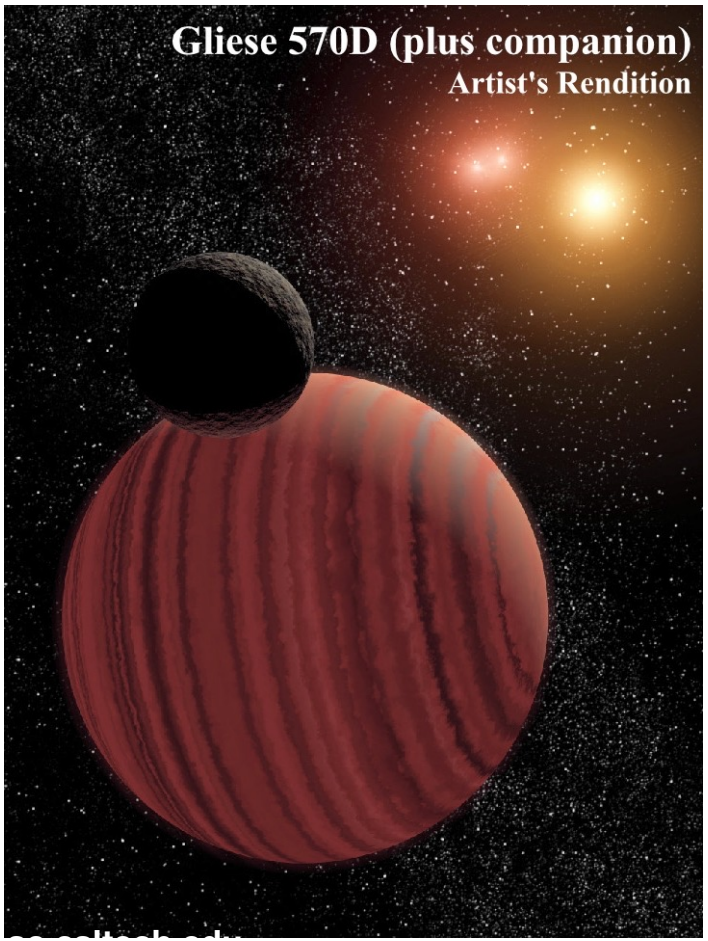
Data

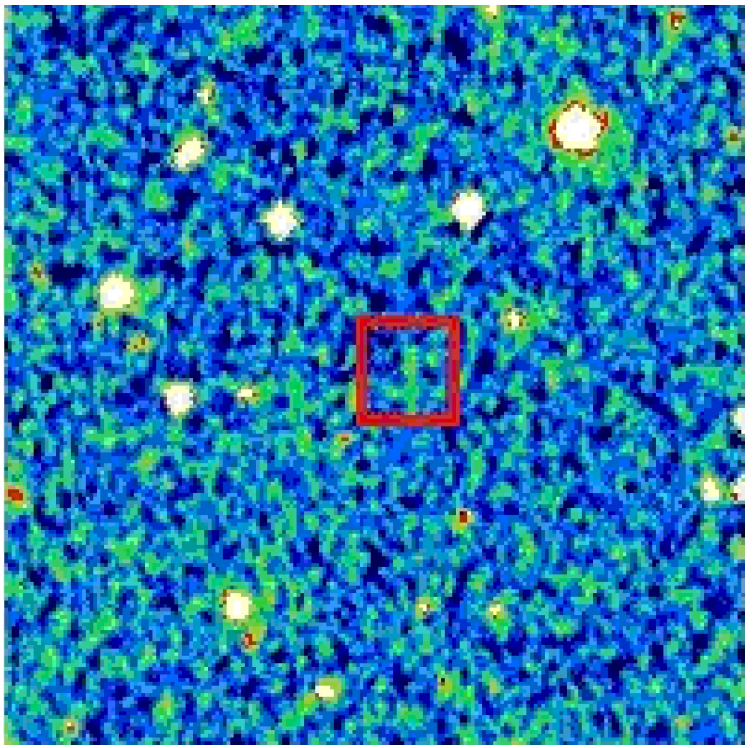
$$m\frac{d^2}{dt^2}\vec{r}(t) = \vec{F}$$

Knowledge

Google Science in the Flat World

Parallel computing on globally distributed supercomputers & visualization platforms will revolutionize & democratize science & engineering (*e.g.*, Google astronomy in the flat world)





SDSS image of brown dwarf, 2MASSI J0104075-005328

Atomistic Data as a Graph

- **Molecular dynamics data**
  - —Atomic data: species, positions, velocities, stresses,…

$$\{\lambda_i, \vec{r}_i, \vec{v}_i, \overleftrightarrow{\sigma}_i, \ldots | i = 1, \ldots, N\}$$

  - —Atomic-pair data: bond order, pair distance,…

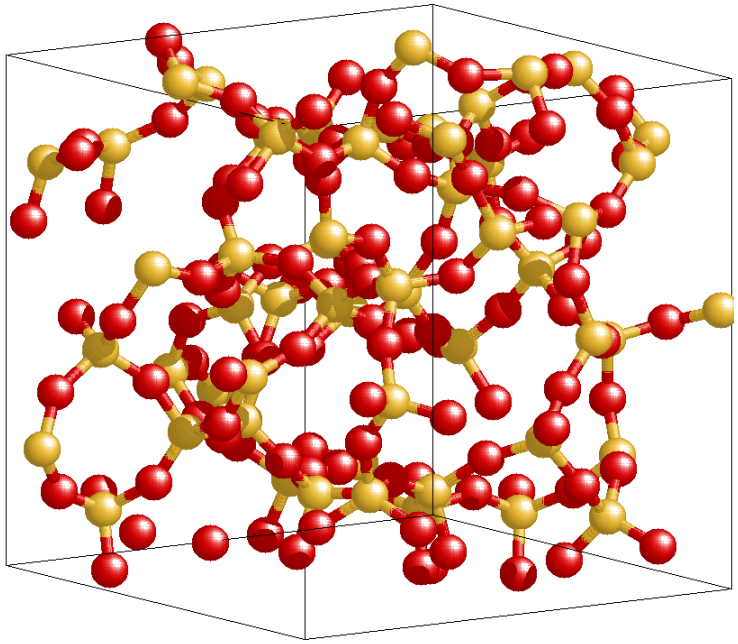
$$\{B_{ij}, \vec{r}_{ij}, \ldots | i, j = 1, \ldots, N; i \neq j\}$$

- **Chemical bond network**
  - —Node degrees
  - —Paths
  - —Rings
  - —Frequently occurring subgraphs

$G = (V, E)$

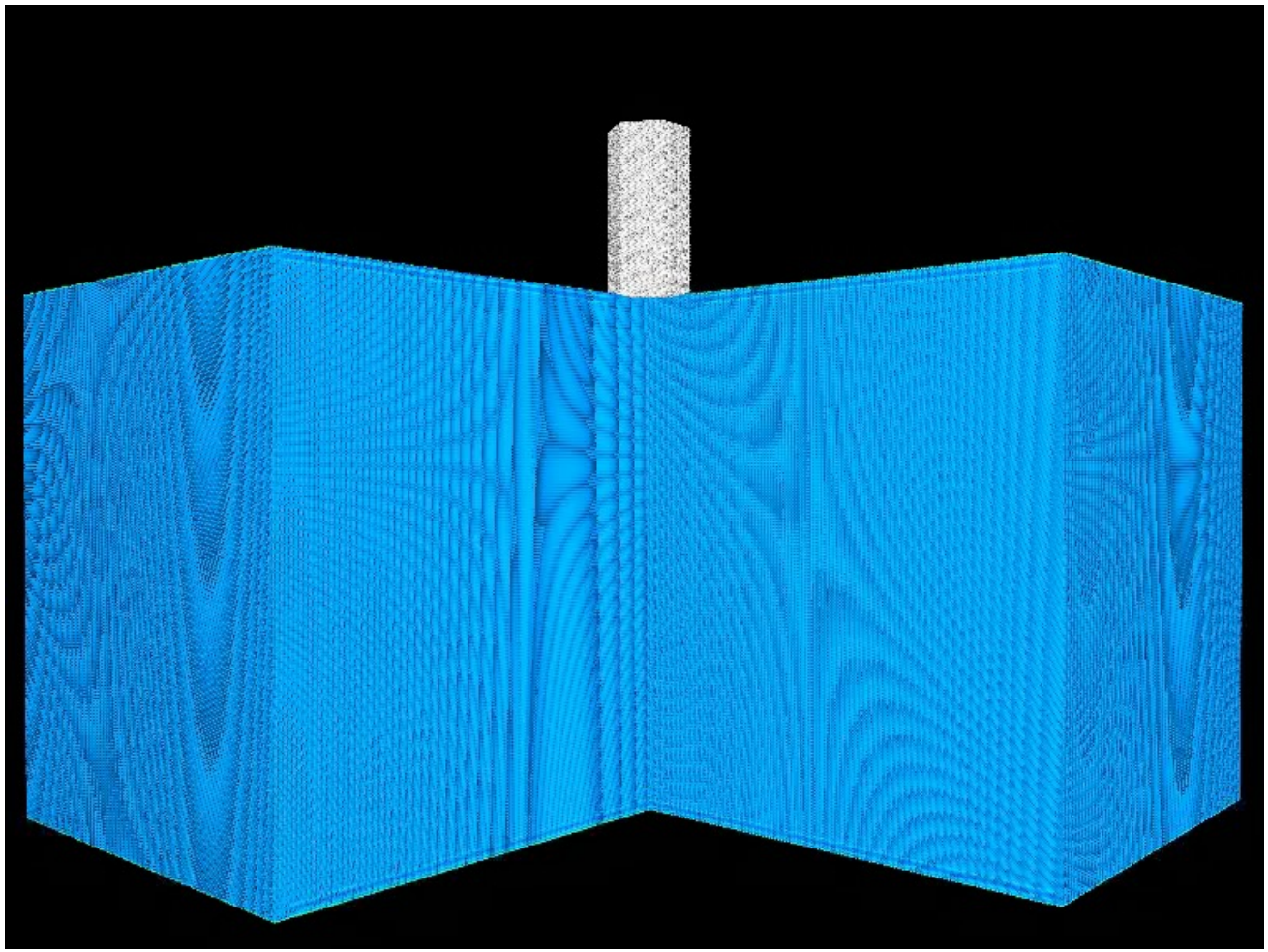
V: Set of atoms

E: Set of bonds

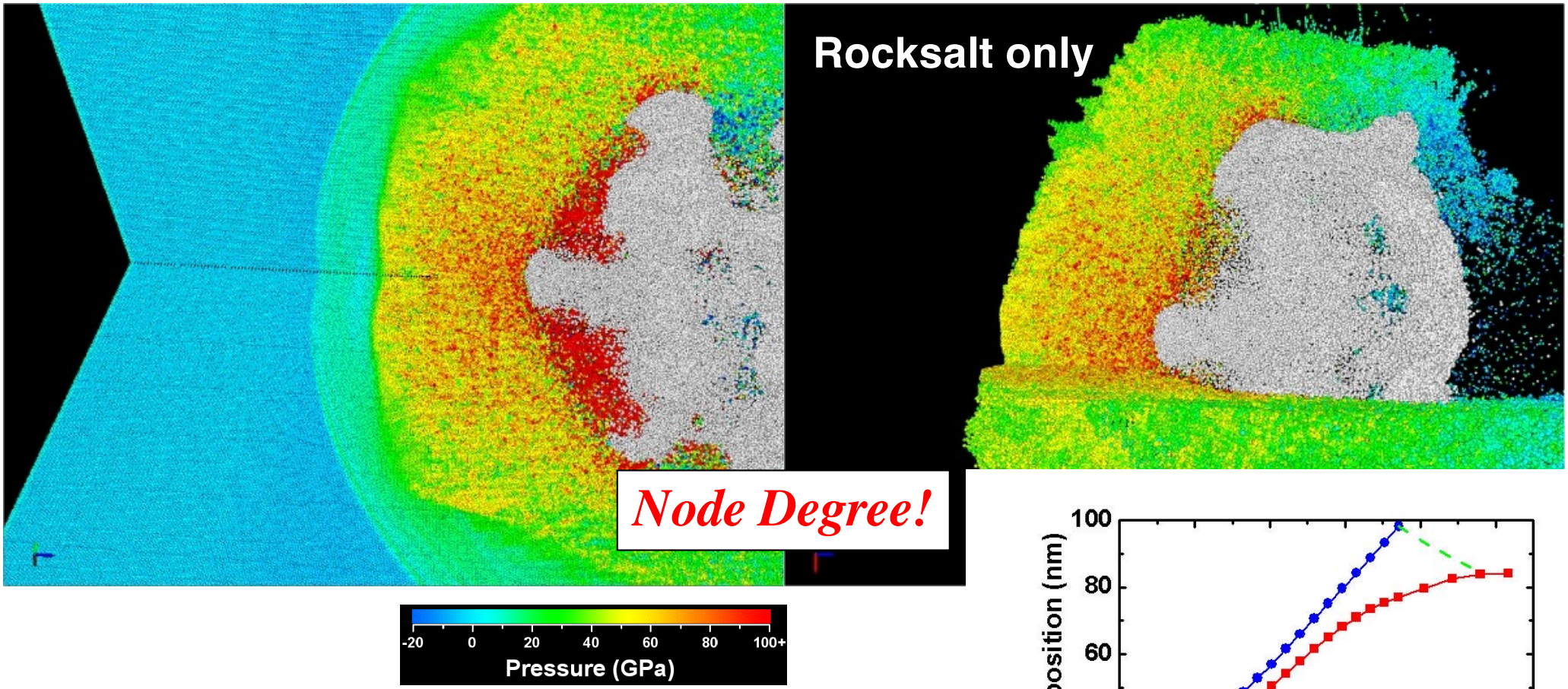


Hypervelocity Impact on Ceramics

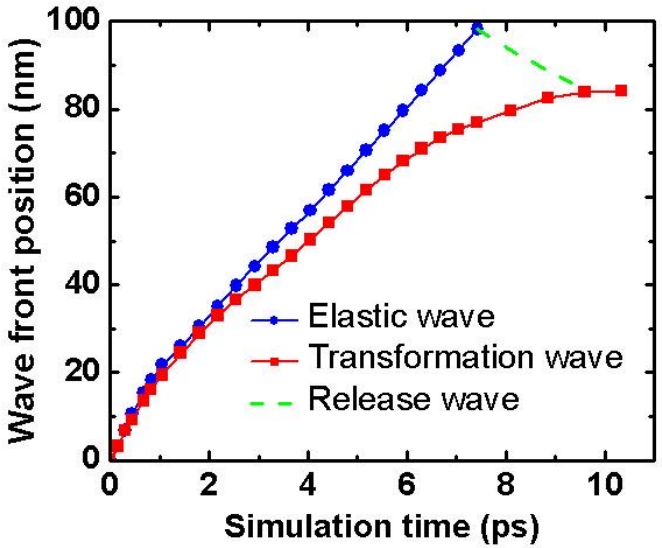
- 209M-atom MD of AlN
- 300M-atom MD of SiC
- 540M-atom MD of Al_2O_3



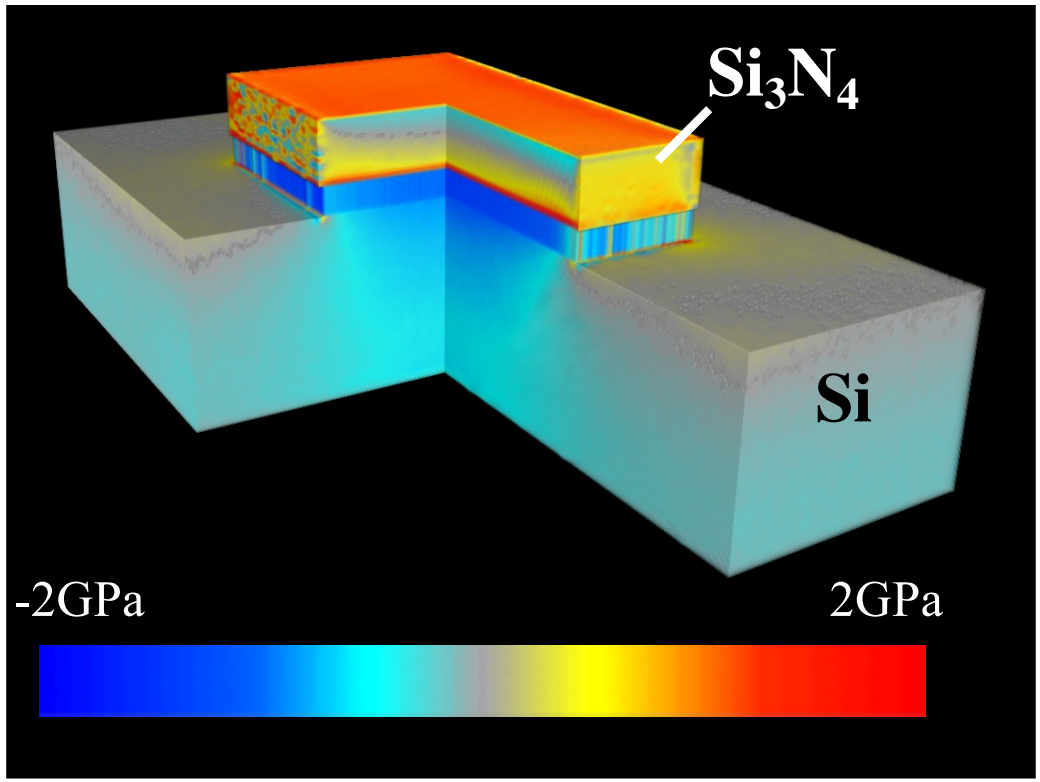
Shock-Induced Structural Phase Transformation in AlN

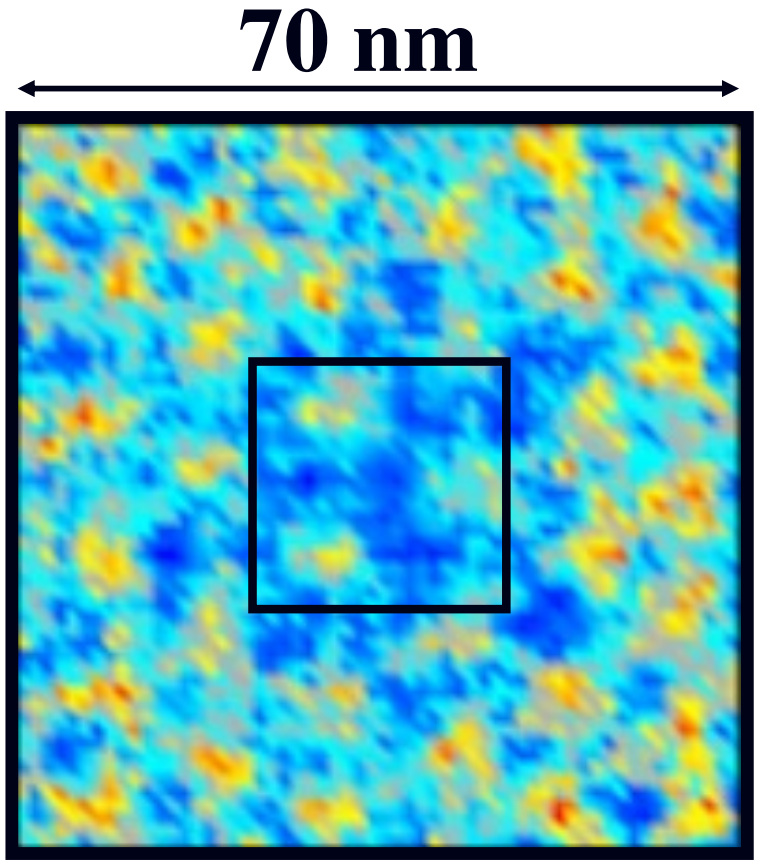


- **Wurtzite (4-coordinated) to rocksalt (6-coordinated) phase transformation at 20 GPa**



Stress Domains in Si_3N_4/Si Nanopixels

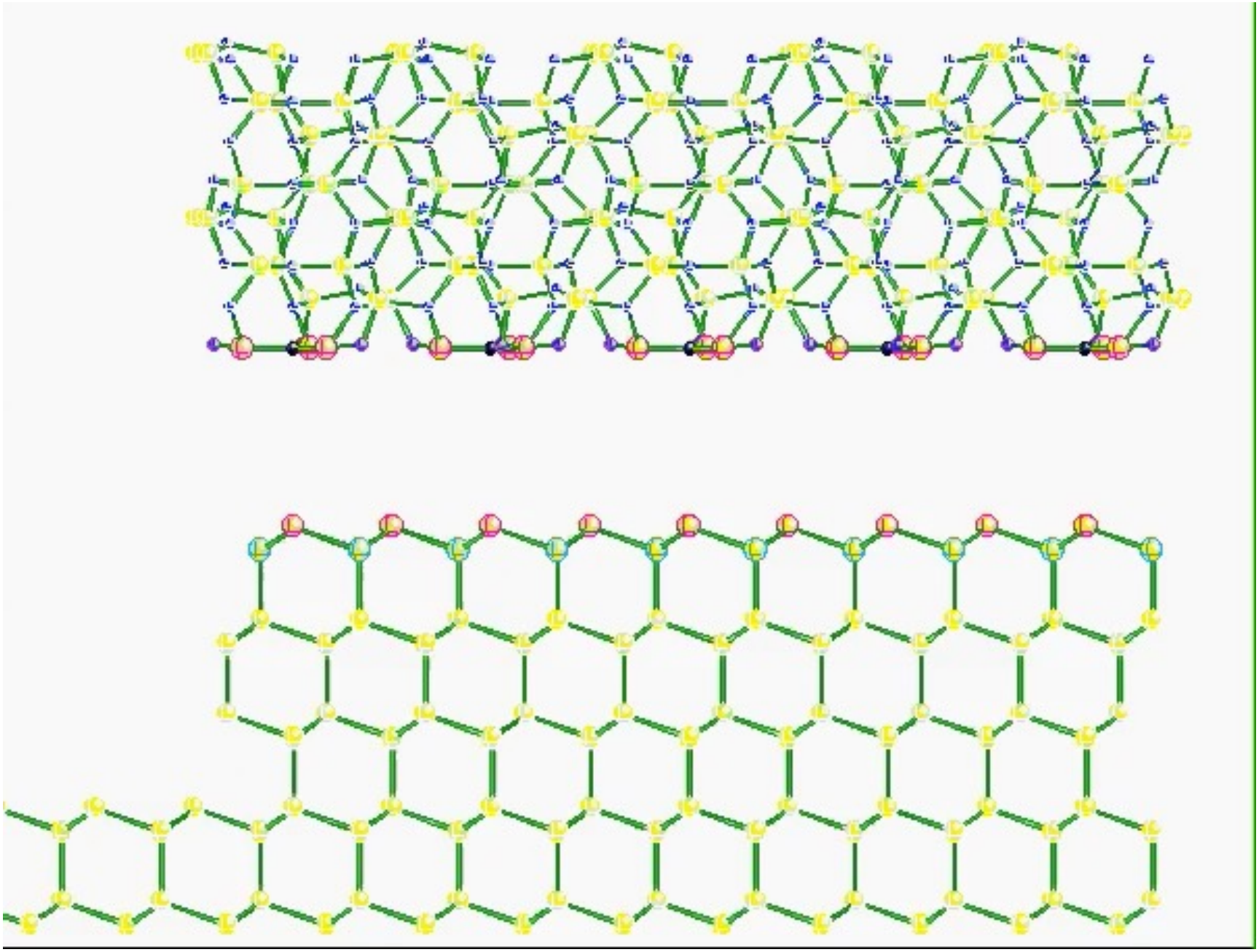




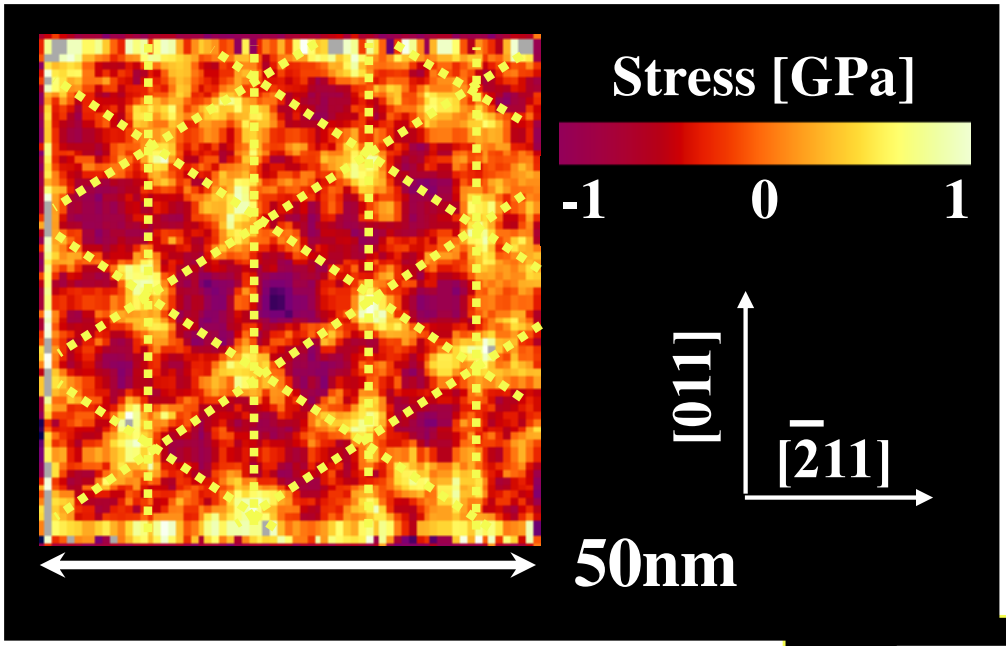
Stress well in Si with a crystalline Si_3N_4 film due to lattice mismatch

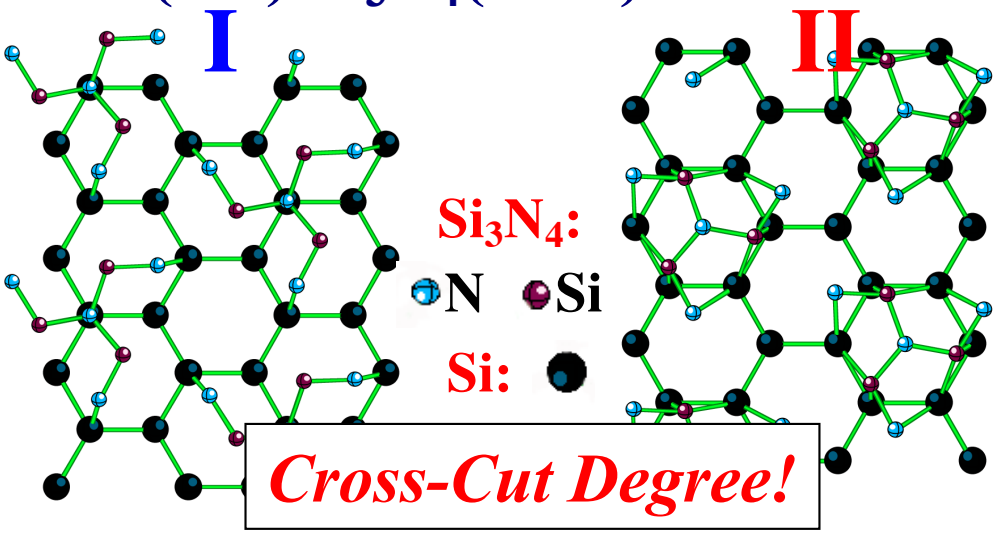
Stress domains in Si due to an amorphous Si_3N_4 film

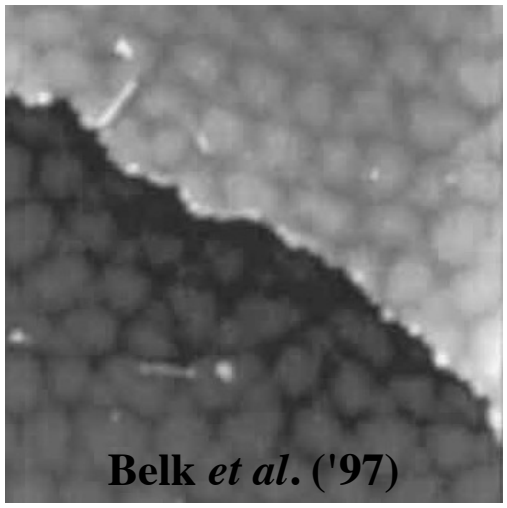
$Si(111)/Si_3N_4(0001)$ Interface



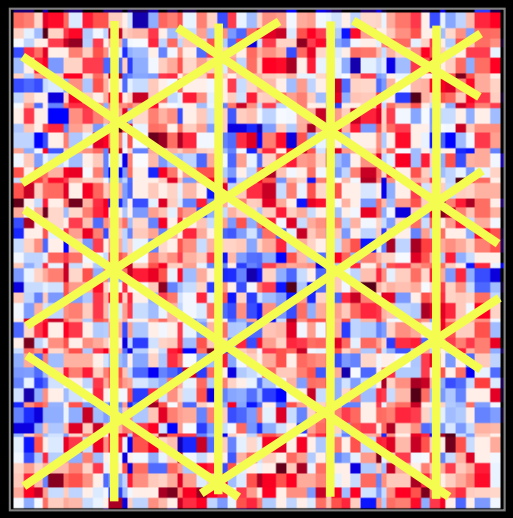
Stress Domains in Si/Si_3N_4 Nanopixel

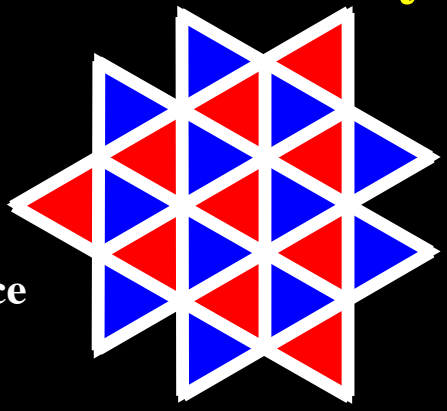






Misfit dislocation network in InAs/GaAs(111)





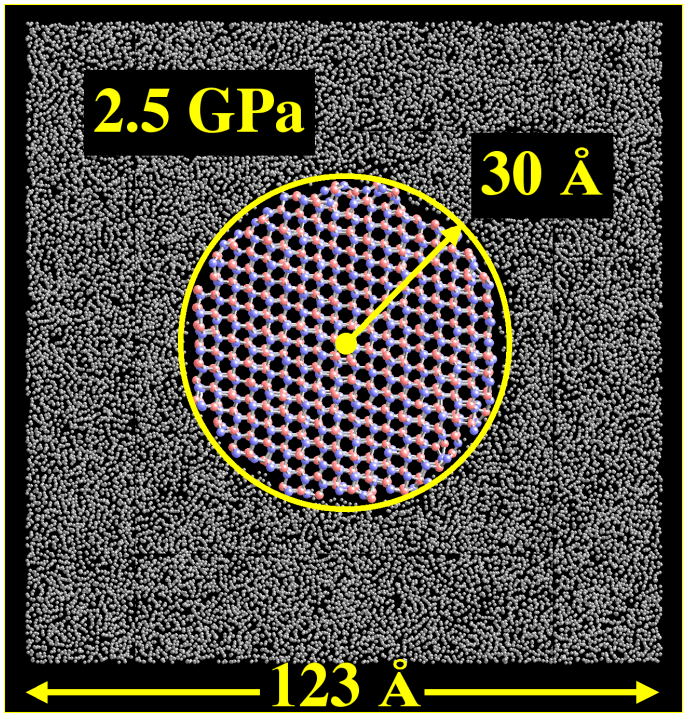
Lattice mismatch (1%) induced interfacial domain array

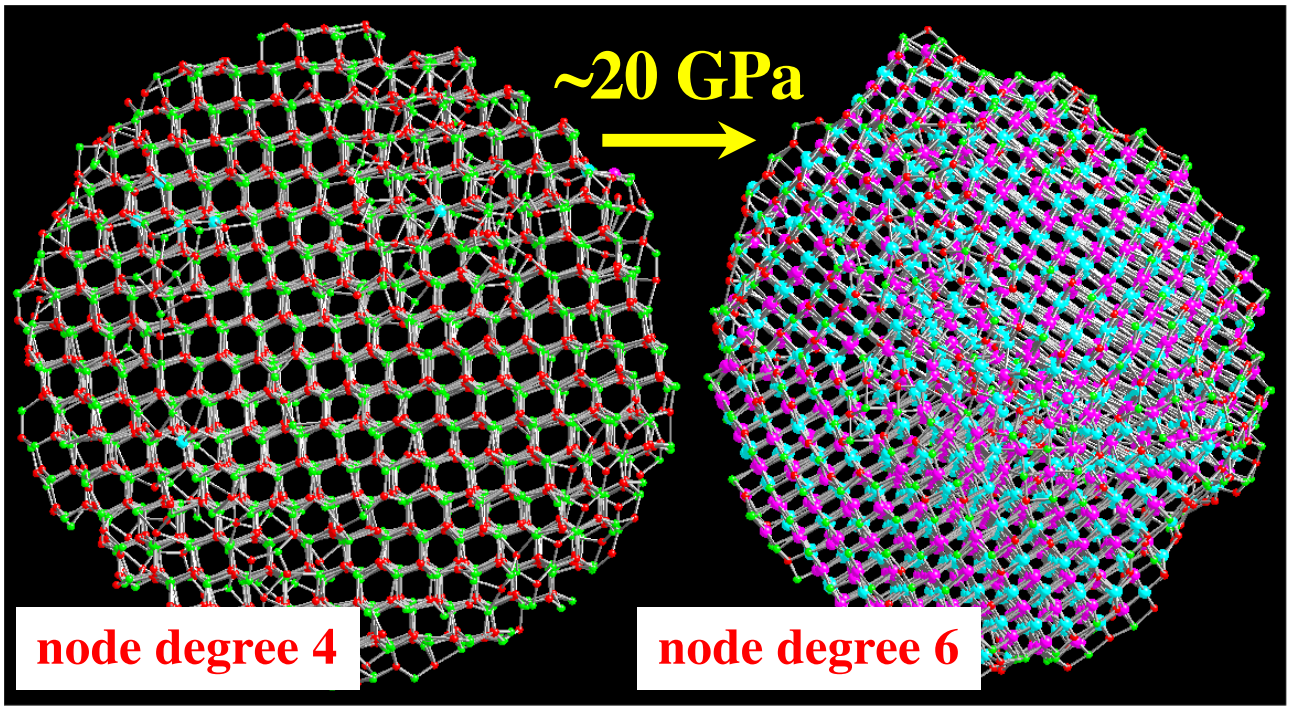
N-Si coordination across interface

1 2 3

High-Pressure Structural Transformation

- Wurzite (node degree 4) to rocksalt (node degree 6) structural transformation of a GaAs nanoparticle under high pressure

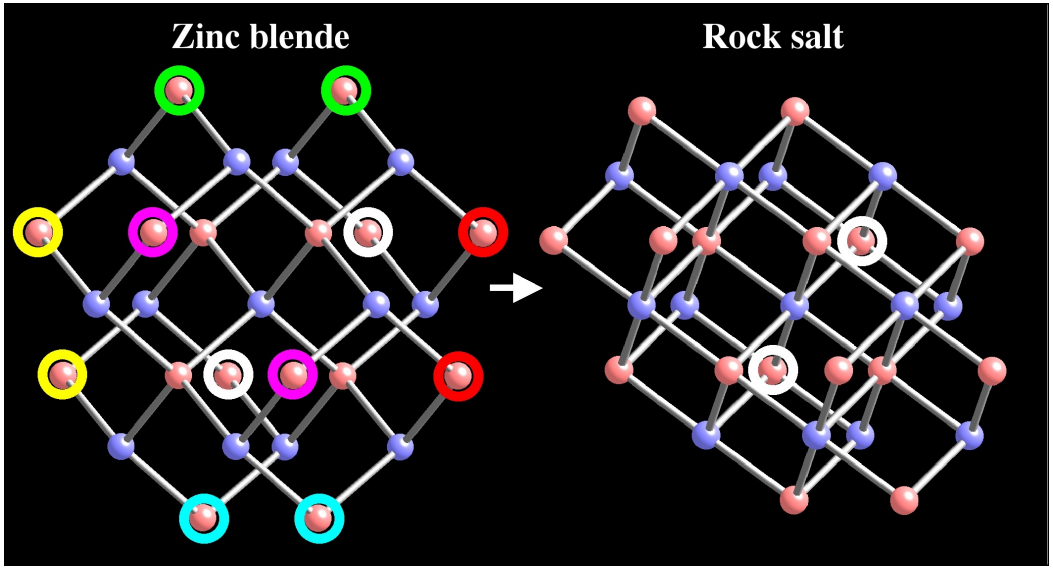


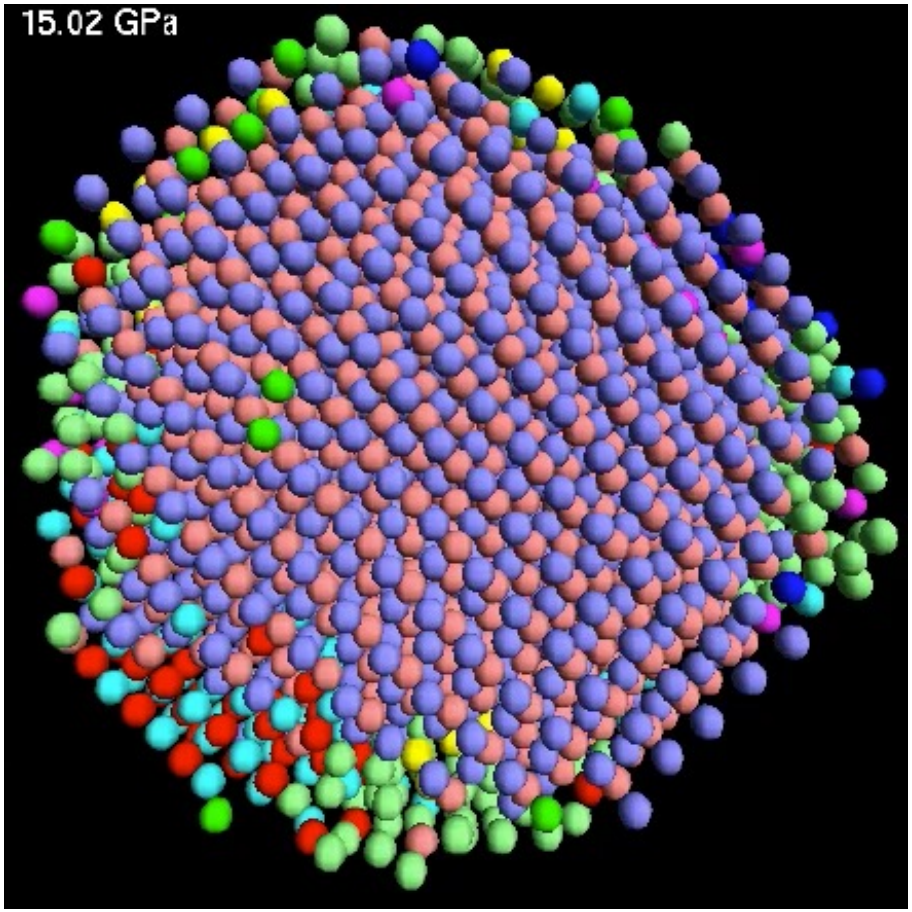


- Existence of multiple domains?

Graph-Transition Tracking

- **Finite set of graph transitions as a classifier**



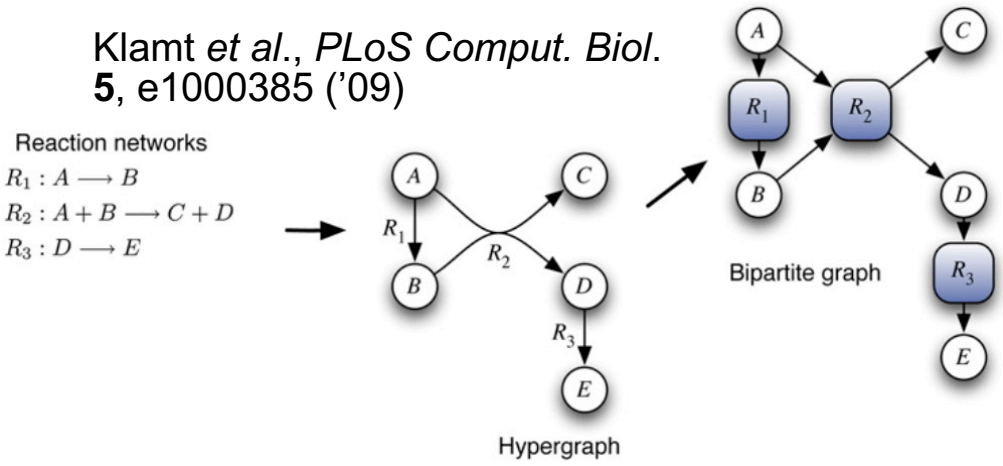


$$G = (V,E)$$
$$\downarrow$$
$$G' = (V,E')$$
$$E \subset E'$$

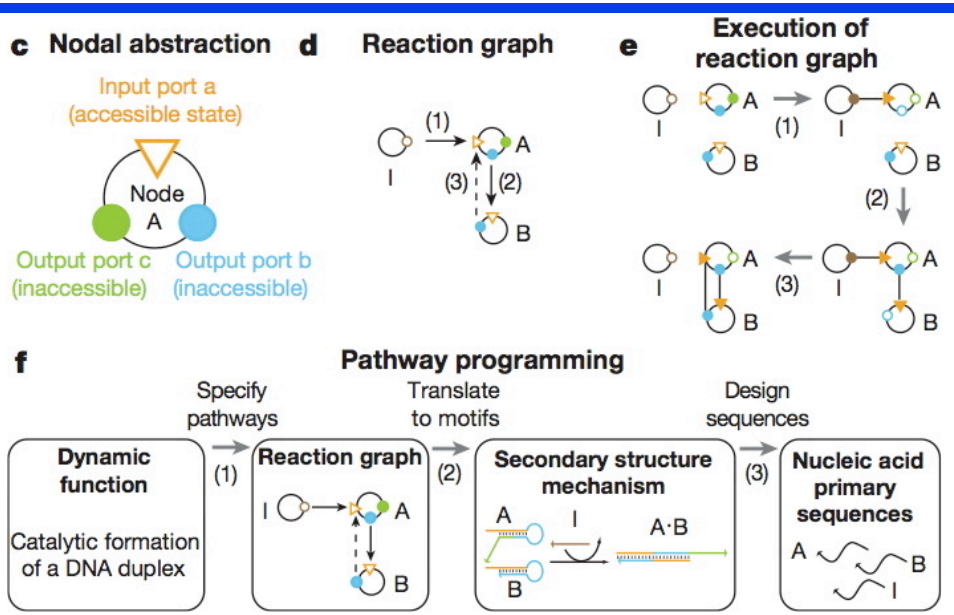
Graph Transition!

Chemical Reaction Network

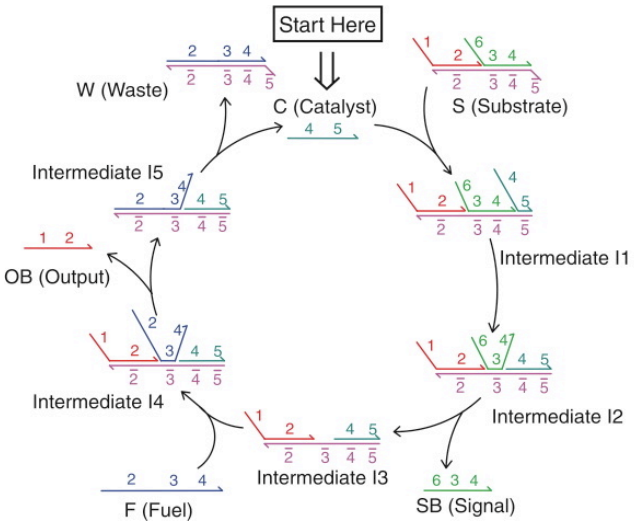
Klamt *et al.*, *PLoS Comput. Biol.* **5**, e1000385 ('09)

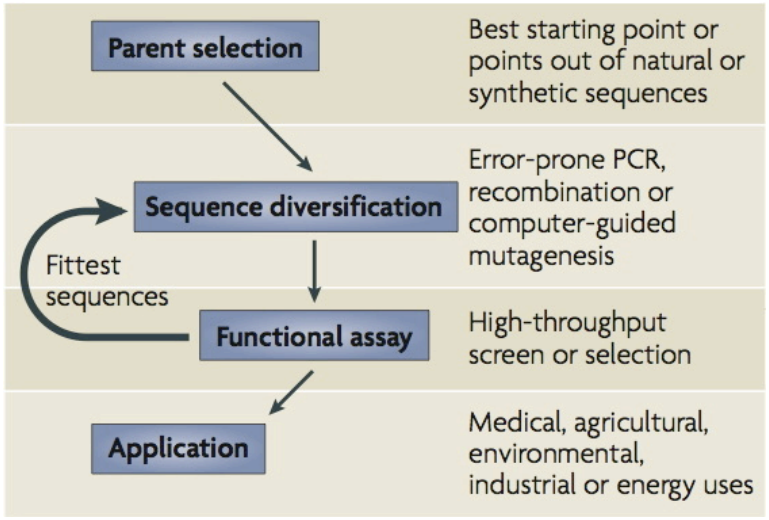


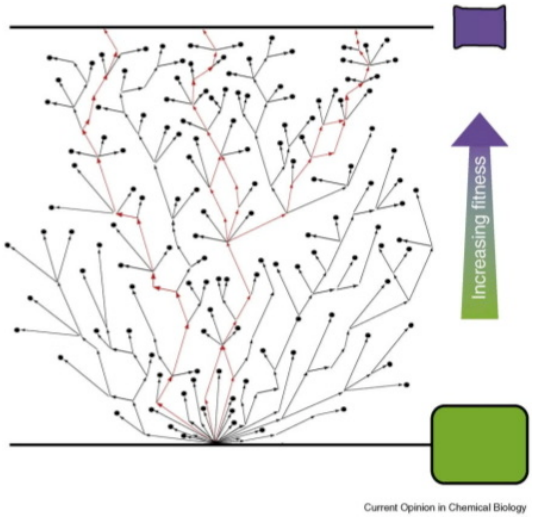
Yin *et al.*, *Nature* **451**, 318 ('08)



Zhang *et al.*, *Science* **318**, 1121 ('07)







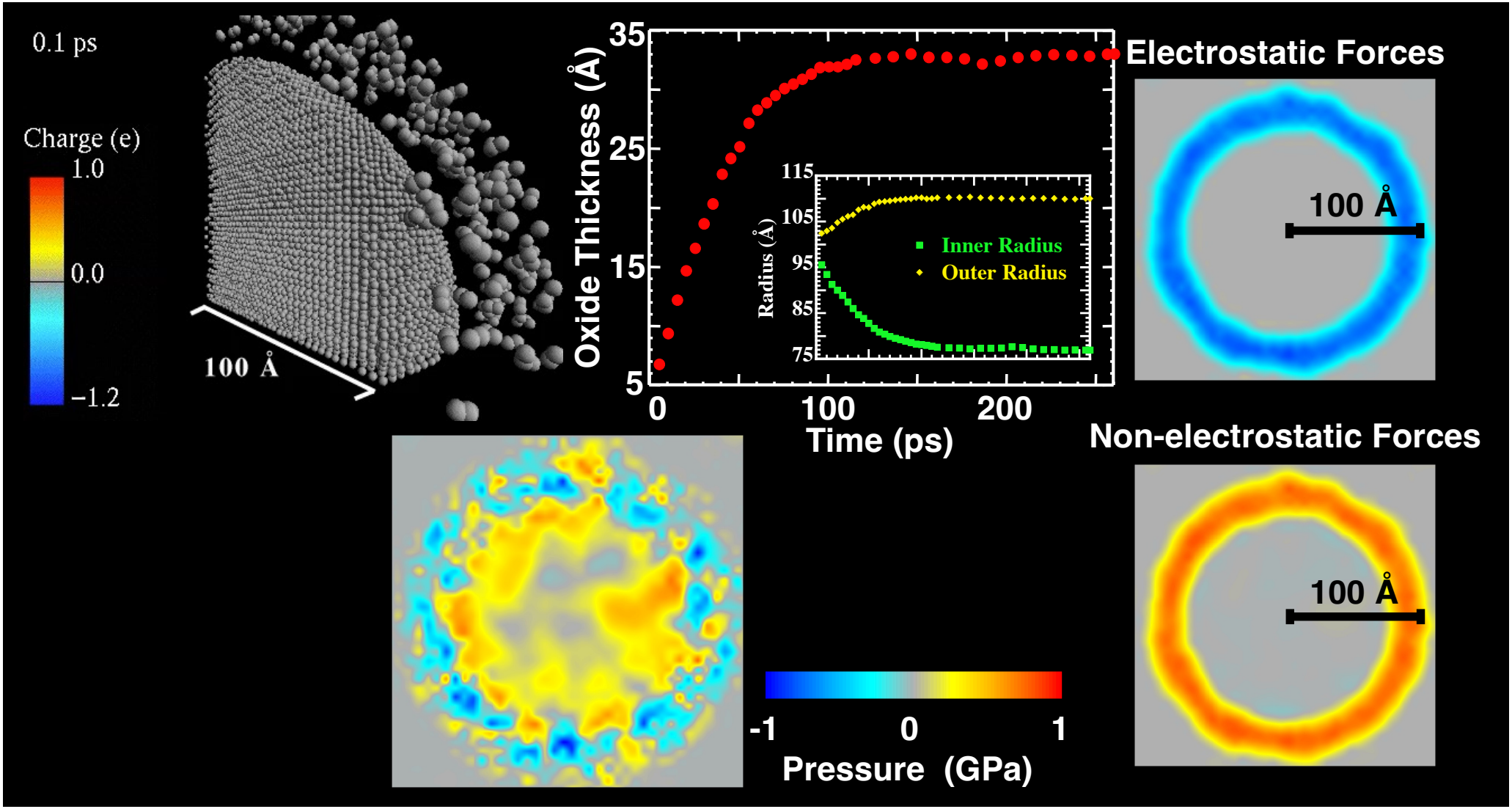
Arnold group, *Nature Rev. MCB* **10**, 867('09); *COCB* **13**, 3 ('09)

Reaction graph = language for self-assembly & catalytic cycle design

Directed & accelerated evolution

Chen *et al.*, *Nature Nanotechnol.* **8**, 755 ('13)

Oxidation of an Al Nanoparticle (n-Al)

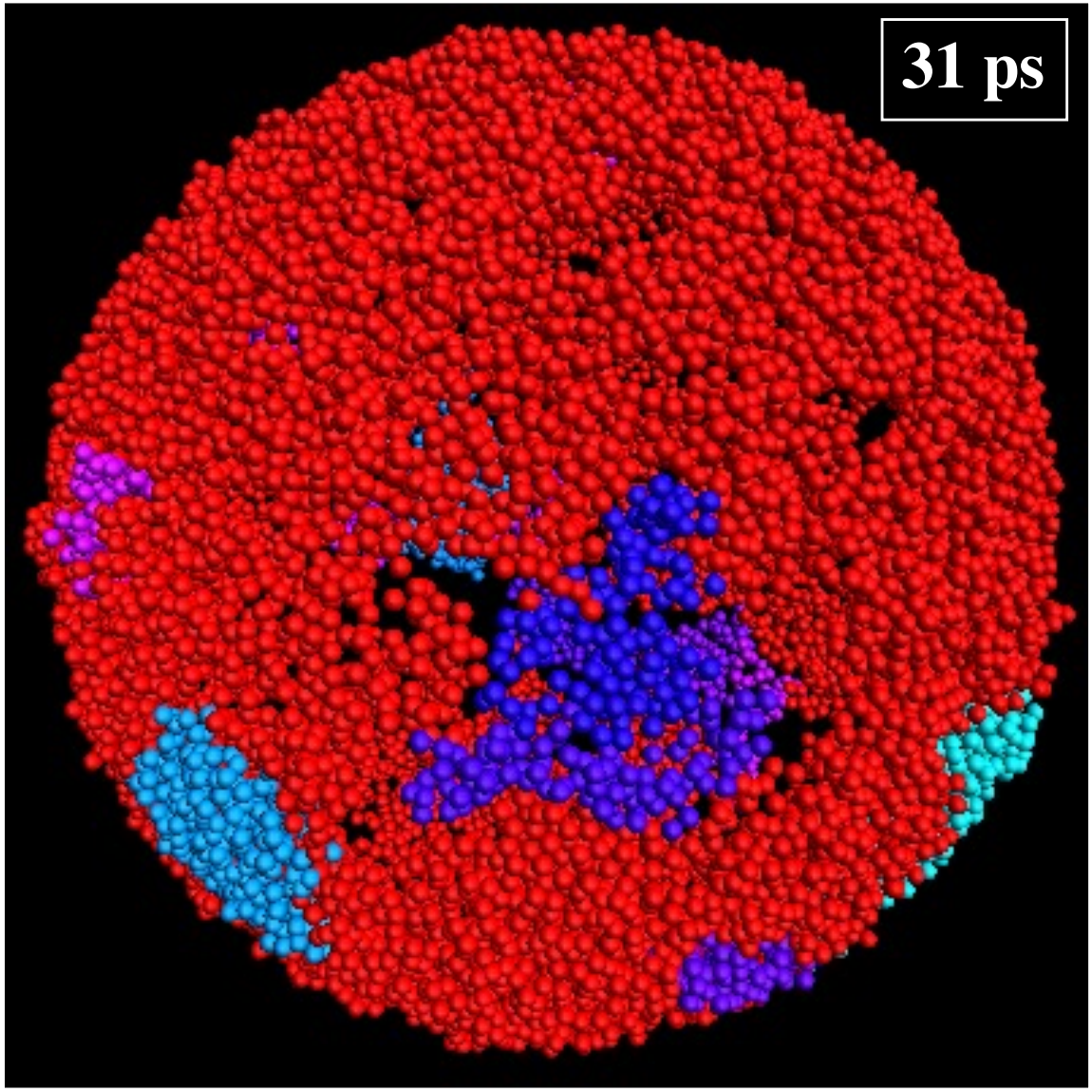


- Oxide thickness saturates at 40 Å after 0.5 ns, in agreement with experiments
- Oxide region/metal core is under negative/positive pressure
- Attractive Al-O Coulomb forces contribute large negative pressure in the oxide

Oxidative Percolation

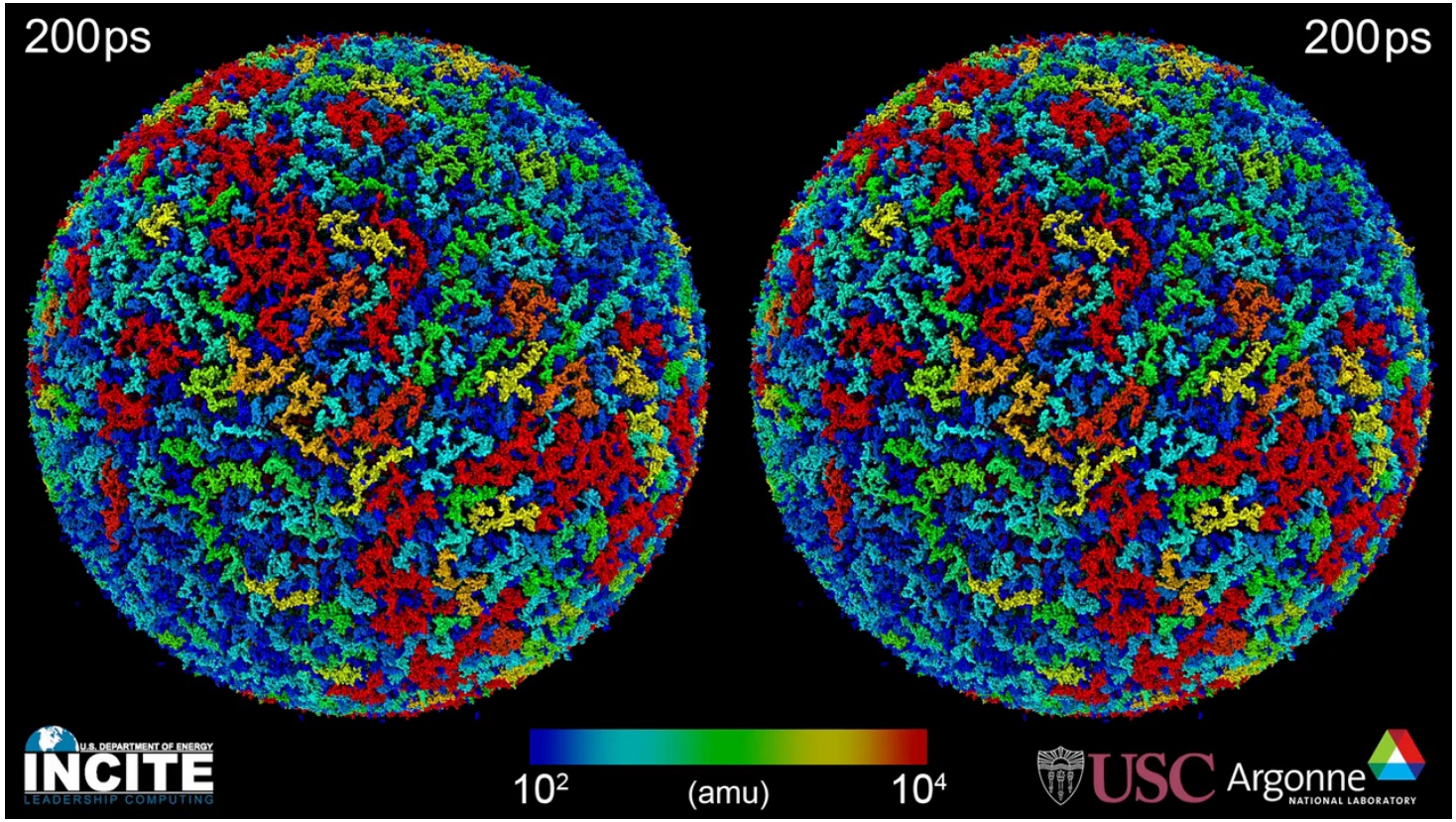
Clusters of OAl_4 coalesce to form a neutral, percolating tetrahedral network that impedes further growth of the oxide

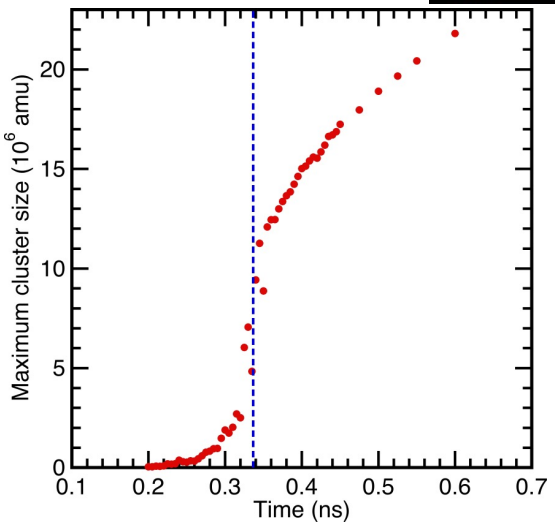
Percorative Connected Components!



Fractal Nanocarbon Product

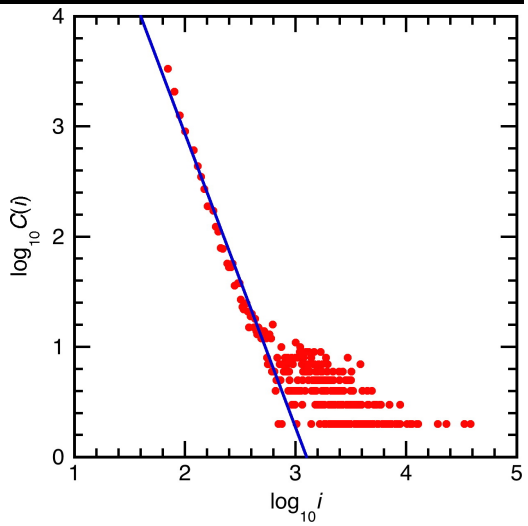
- **Percolation transition causes carbon clusters to exhibit power-law distribution of sizes: $C(i) \sim i^{-\tau}$**





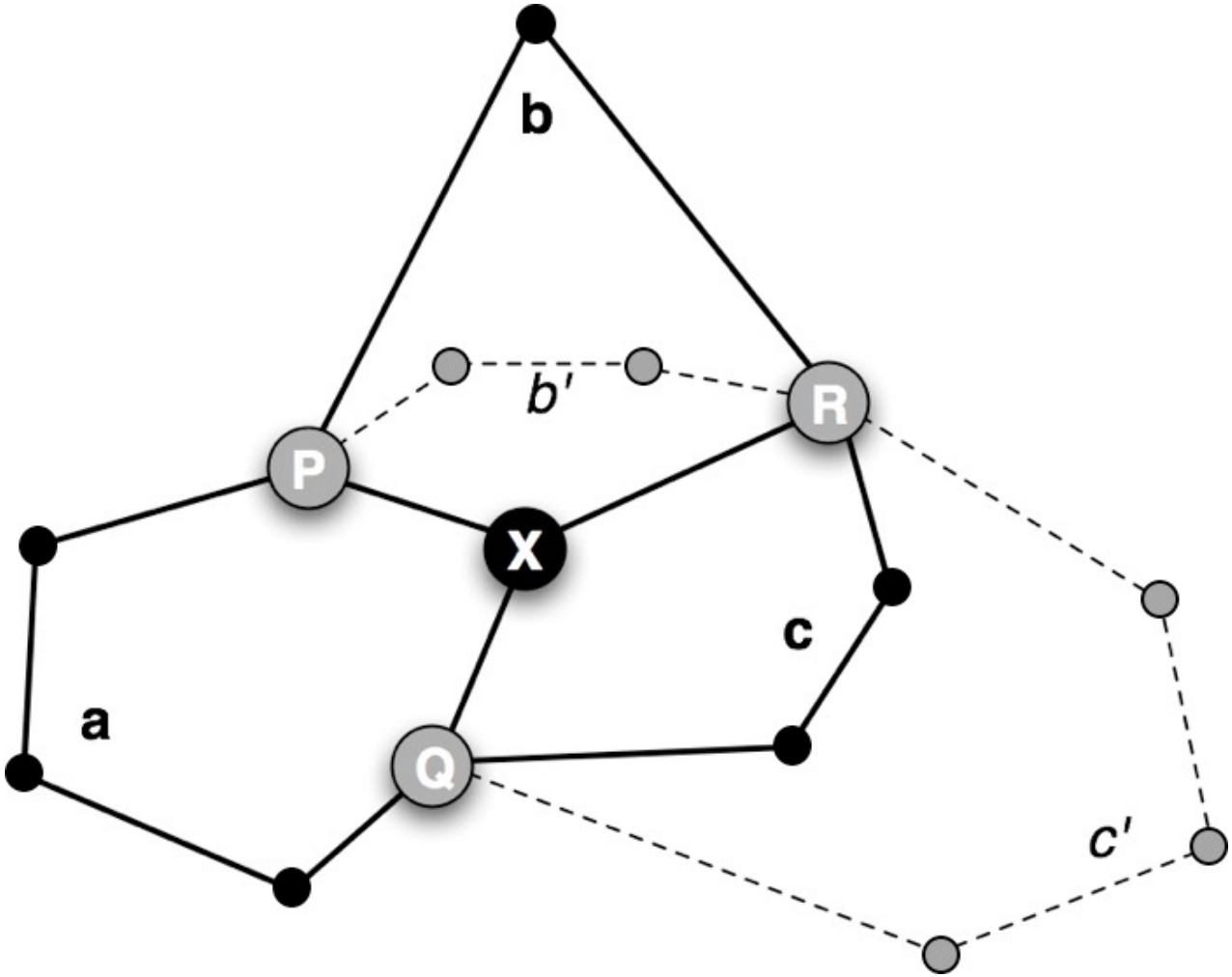
- **Fractal nanocarbon product with large surface areas may find supercapacitor, battery-electrode & mechanical metamaterial applications: $d_f = d/(\tau - 1) \sim 1.85$**

K. Nomura *et al.*, *Sci. Rep.* **6**, 24109 ('16)
J. Insley *et al.*, *IEEE/ACM SC16*



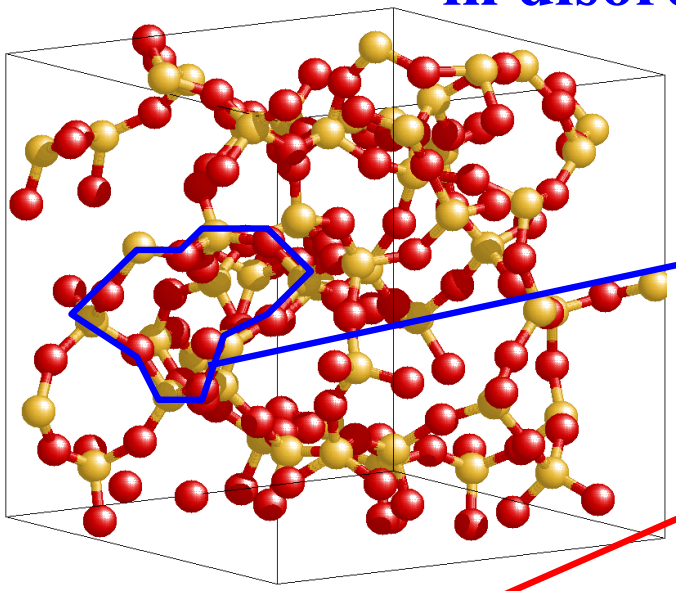
Shortest-Path Rings

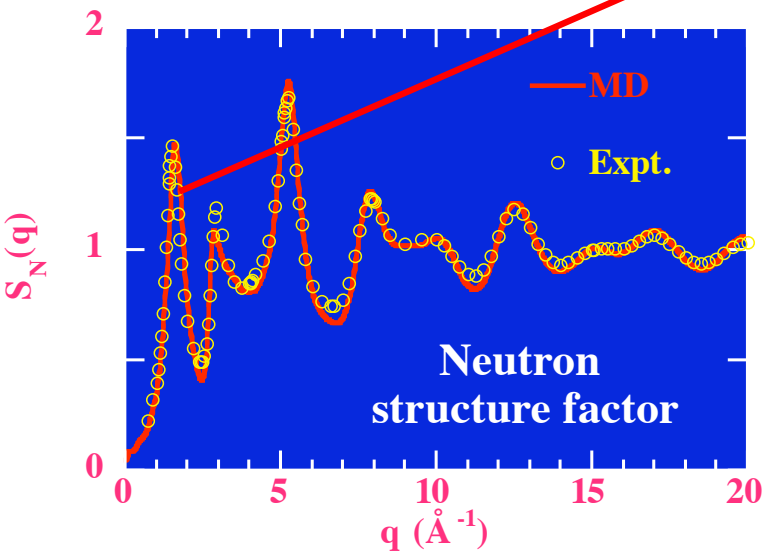
- **K-ring:** Given a vertex *x* & two of its neighbors *w* & *y*, a K-ring generated by the triplet *w*–*x*–*y* is any ring containing the edges [*w*–*x*], [*x*–*y*] and a shortest path *w*–*y* path in *G*–*x*

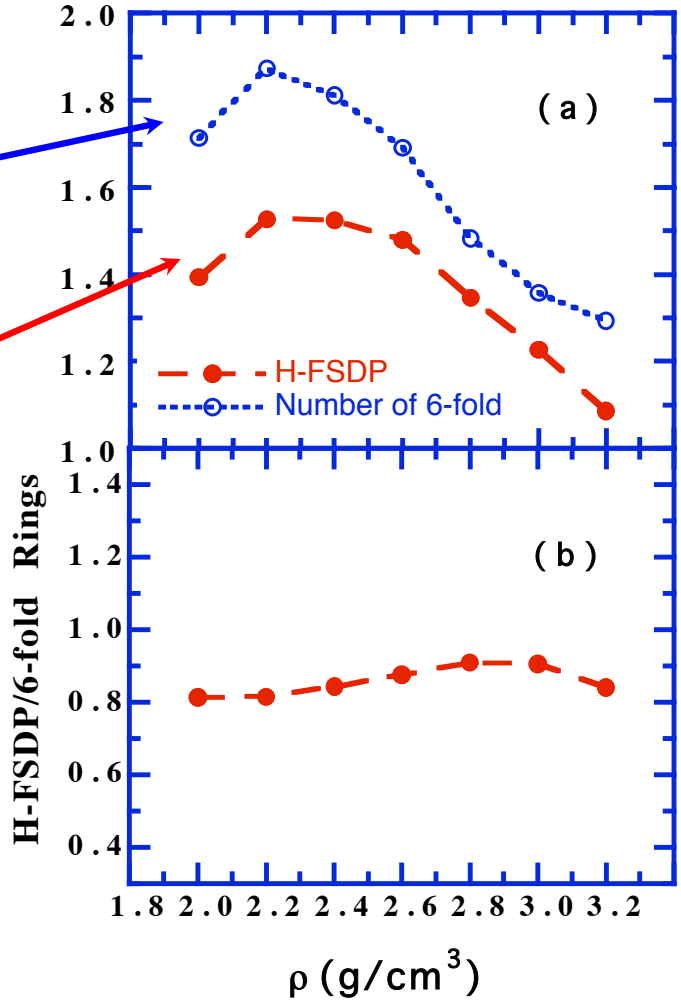


Ring-based Data Mining

Shortest-path ring analysis of intermediate-range order (IRO) in disordered materials

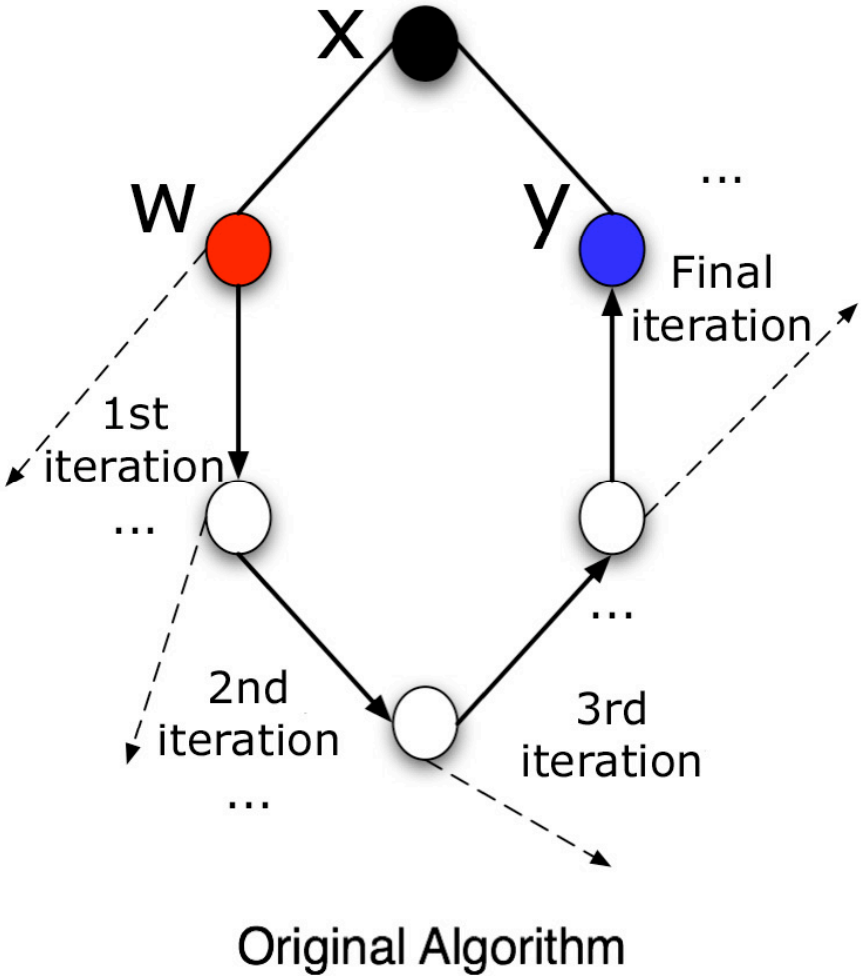




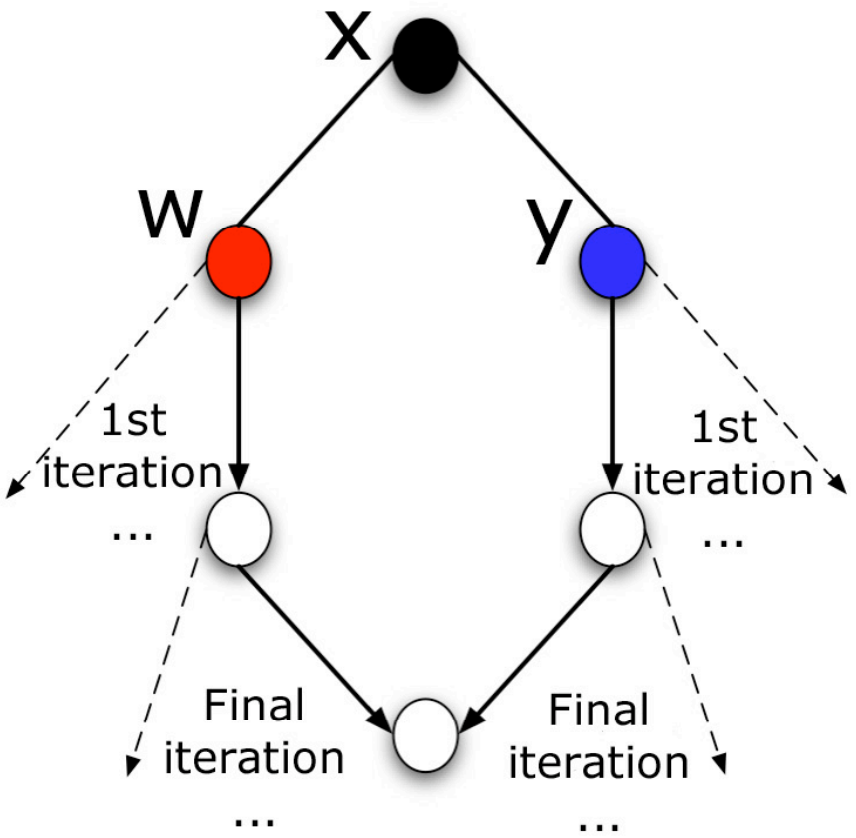


Correlation between IRO in neutron scattering & ring distribution

Fast Ring Analysis: Dual-Tree Expansion



Original Algorithm



Proposed Algorithm

C. Zhang *et al.*, *Computer Physics Communications* **175**, 339 ('06)

DTE Algorithm

Algorithm dual_tree_expansion()

Input:

V = Set of all vertices (i.e., atoms)
R_c = Ring cutoff range (Euclidean)
R_{bc} = Bond cutoff distance (Euclidean)
L_{MAX} = Maximum length of ring (integer)
P = Number of compute nodes

Output:

The K-ring statistics for all vertices in the network
List of atoms with abnormal ring profile

Variables:

Neighbors(V) = Set of vertices that share an edge with vertex V
$K_v(p)$ = Number of p-member rings that go through vertex V
L_b = Length of the ring formed with path (V_i, V, V_j)

Steps:

0 coarse grained spatial decomposition of atoms on P compute nodes with a thin boundary extension of R_c distance (This step is for the parallel version only)
1 create adjacency list G for all node in V_p using R_{bc} as cutoff distance
2 for every vertex $V \in V_p$
for each vertex pair V_i and V_j in Neighbors(V) do
$A_1 = \{V_i\}$
$A_2 = \{V_j\}$
$L_b = 0$
while ($A_1 \cap A_2 = \varnothing$ AND $L_b < L_{MAX}$) do
$L_b = L_b + 2$
if ($A_1 \cap$ Neighbors(A_2) $\neq \varnothing$ OR $A_2 \cap$ Neighbors(A_1) $\neq \varnothing$)
$L_b = L_b + 1$
break
else if (Neighbors(A_1) $\cap$ Neighbors(A_2) $\neq \varnothing$)
$L_b = L_b + 2$
A_1 = Neighbors(A_1)
A_2 = Neighbors(A_2)
if ($L_b < L_{MAX}$) ++ $K_v(L_b)$

Spatial Hash-Function Tagging

Algorithm spatial hash function tagging (SHAFT)

Input:

$C(V)$ = 3D coordinates of all vertices (i.e., atoms)
R_c = Ring cutoff range (Euclidean)
R_{bc} = Bond cutoff distance (Euclidean)
L_{MAX} = Maximum length of ring (integer)

Output:

The integer index that is unique for all vertices in the maximum ring span

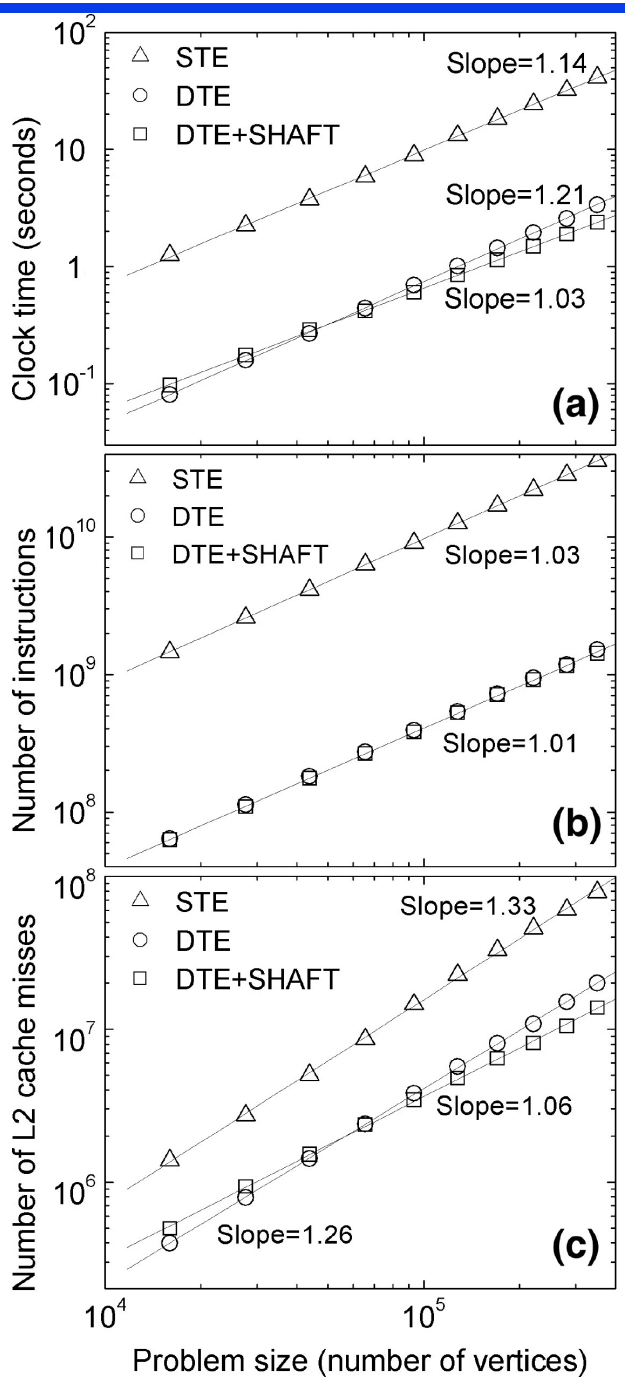
$b = R_{lower} / \sqrt{3}$
$c = R_{upper} L_{max}$
$m = \lceil c/b \rceil$

Step:

for each vertex
for each spatial dimension i from 1 to 3
$q_i = \lfloor C_i/b \rfloor$
$q_i \% = m$
return $q = q_3 \times m^2 + q_2 \times \mathrm{m} + q_1$

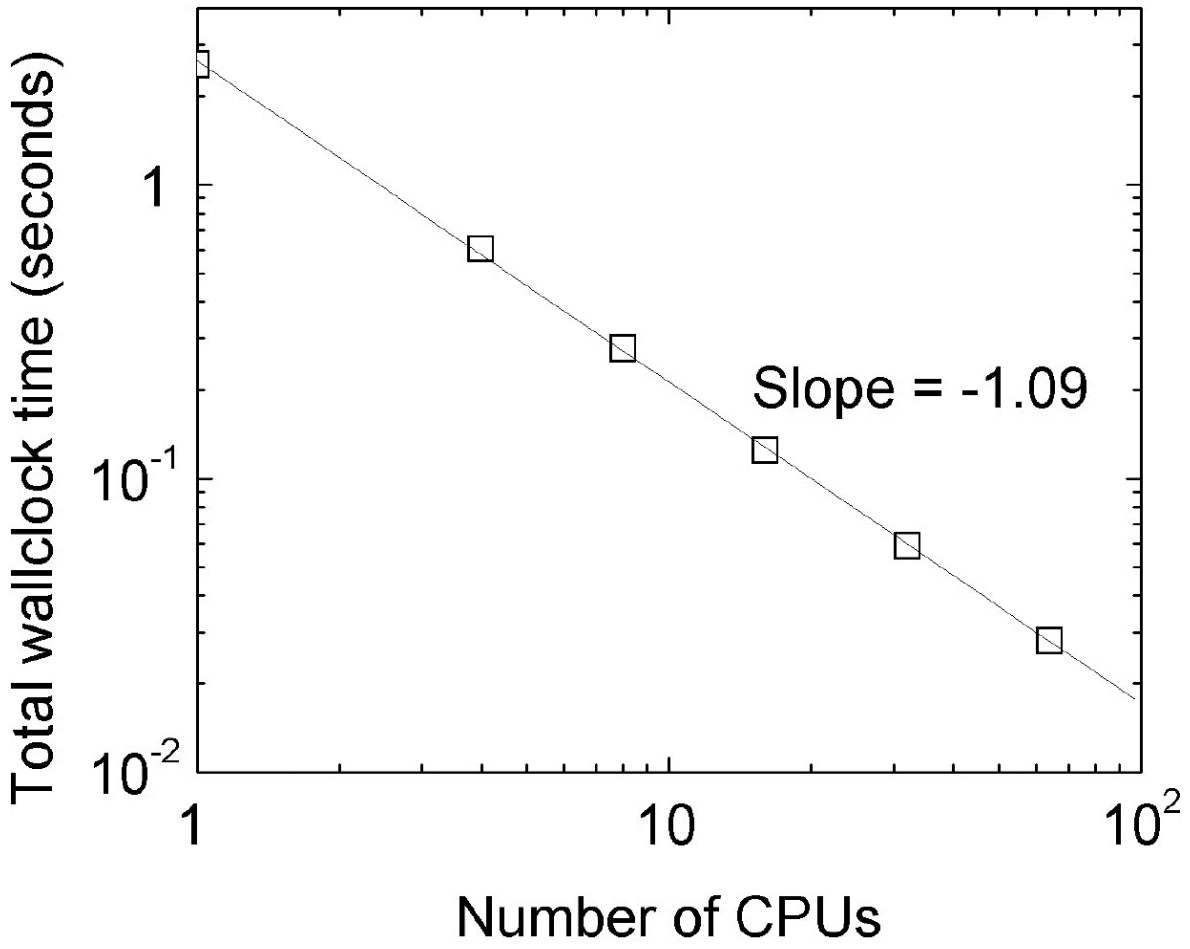
0	1	2	3	4	0	1	2	3	4
5	6	7	8	9	5	6	7	8	9
10	11	12	13	14	10	11	12	13	14
15	16	17	18	19	15	16	17	18	19
20	21	22	23	24	20	21	22	23	24
0	1	2	3	4	0	1	2	3	4
5	6	7	8	9	5	6	7	8	9
10	11	12	13	14	10	11	12	13	14
15	16	17	18	19	15	16	17	18	19
20	21	22	23	24	20	21	22	23	24

Numerical Tests



Linear scaling on the problem size

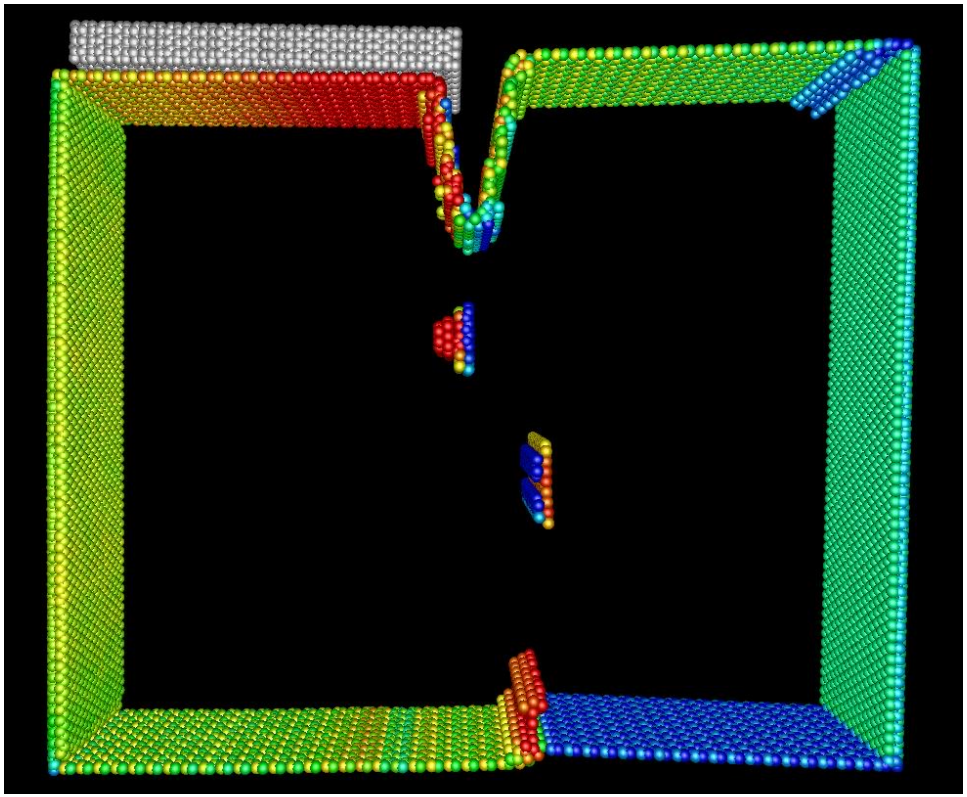
Superlinear (strong) scaling on the number of CPUs



Dislocation Mining

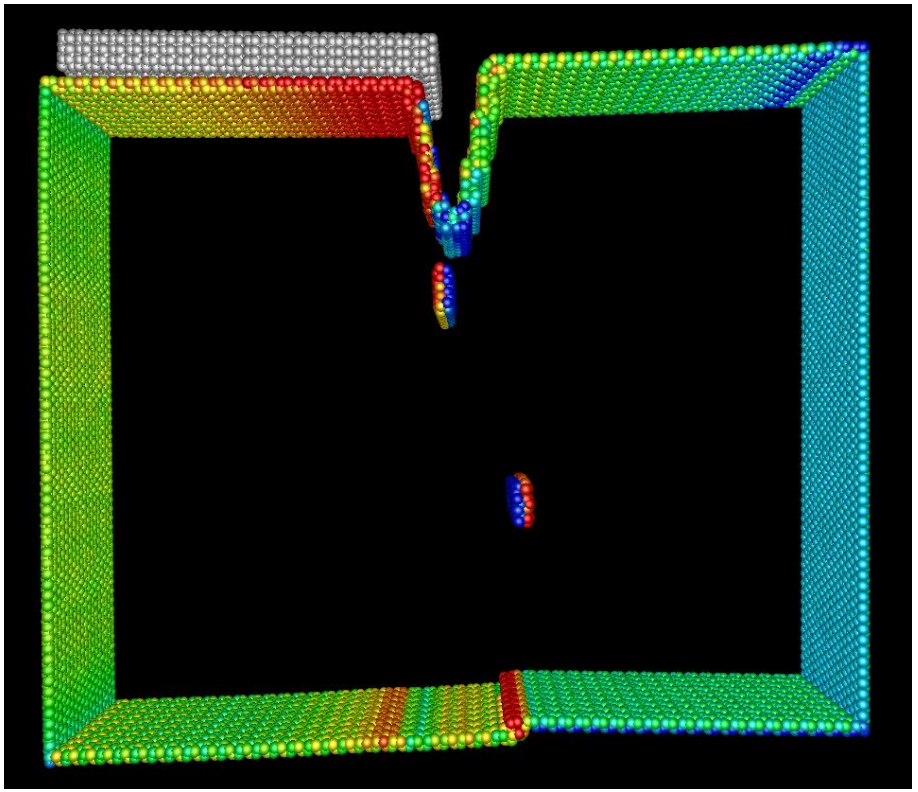
Based on potential energy

Shown atoms with high energy compared to bulk



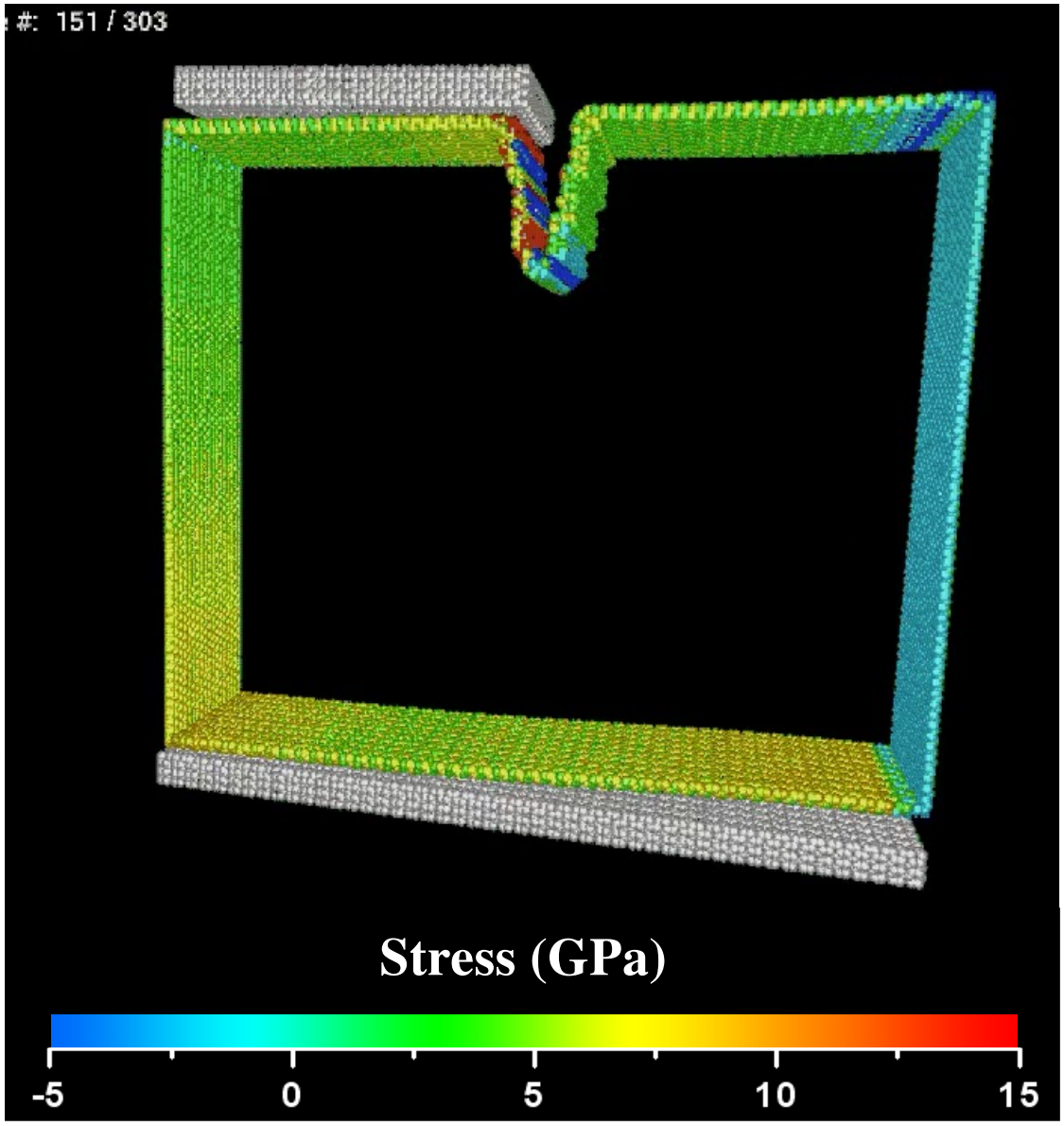
Based on shortest-path ring statistics

Shown atoms with less than 12 6-membered rings



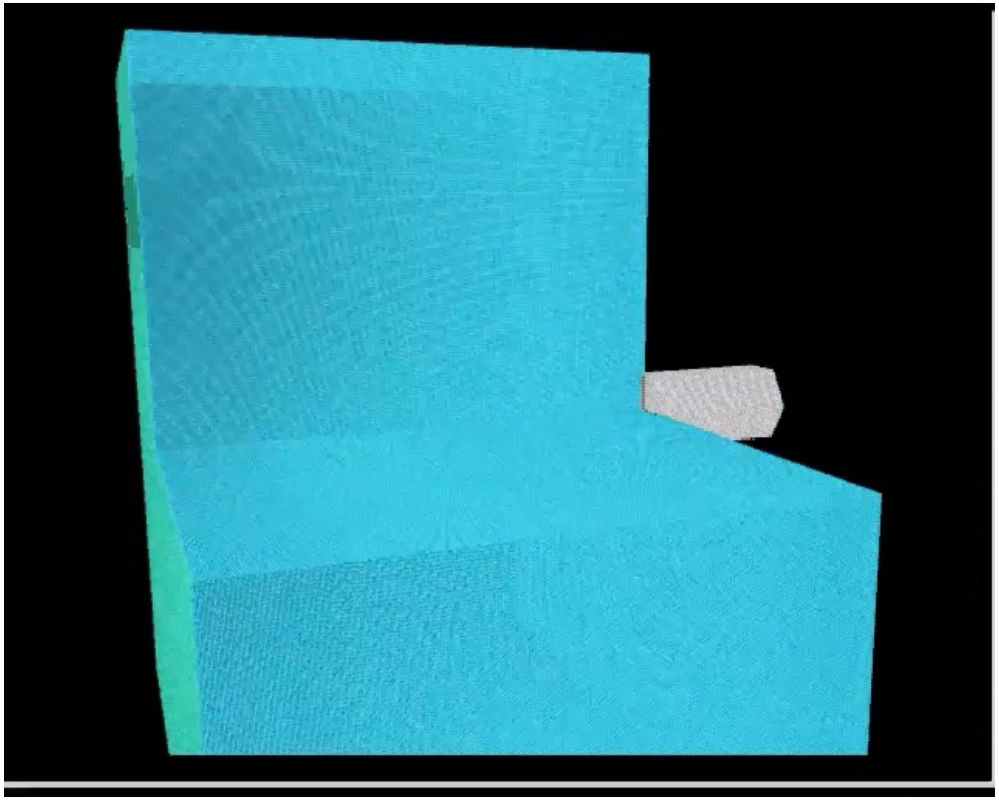
100 km/s Impact on Notched AlN

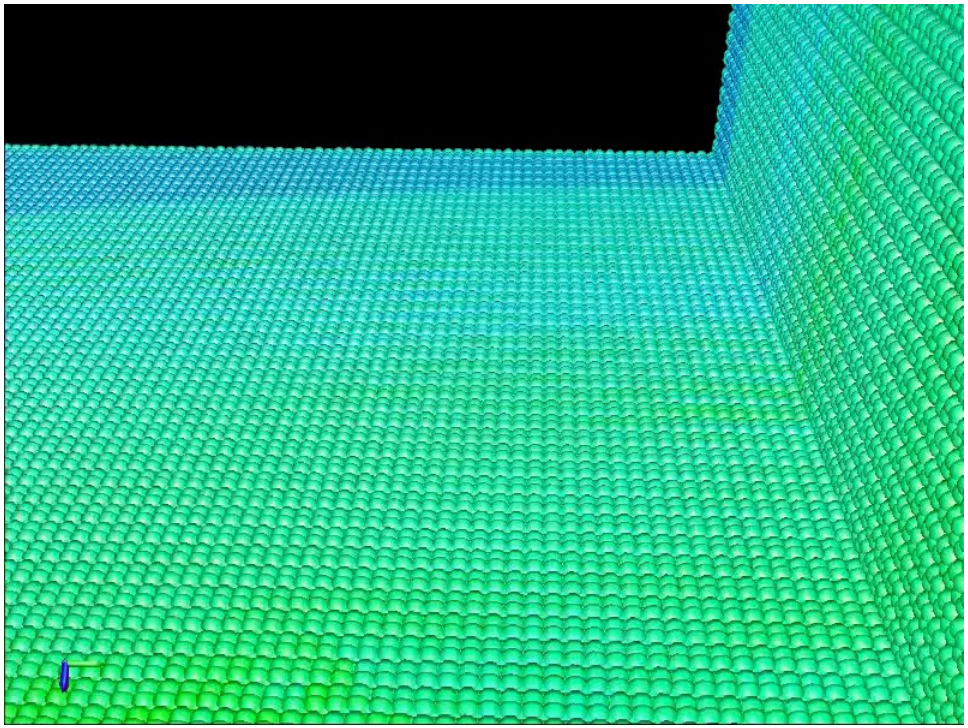
- Dislocation nucleation & emission from notch during impact
- Dislocations & surface atoms mined by ring statistics

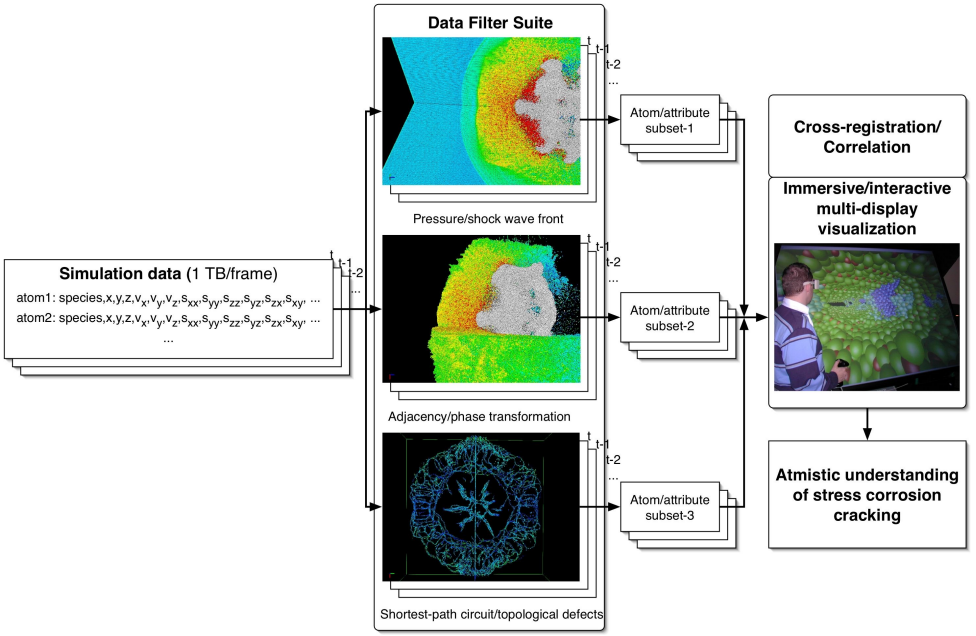


Impact-Damage Tolerant Ceramics?

Inverse problem: design materials with desired properties

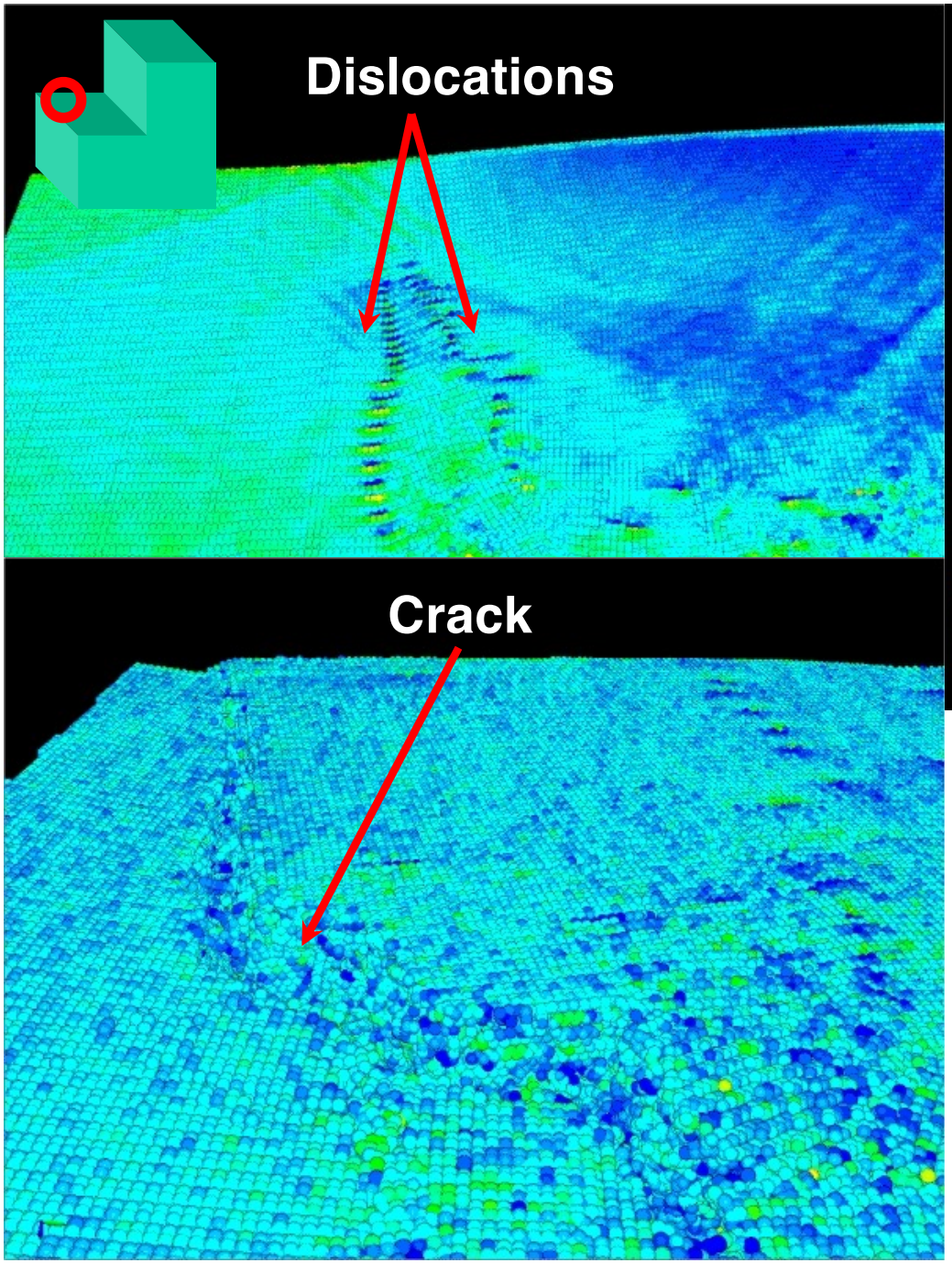


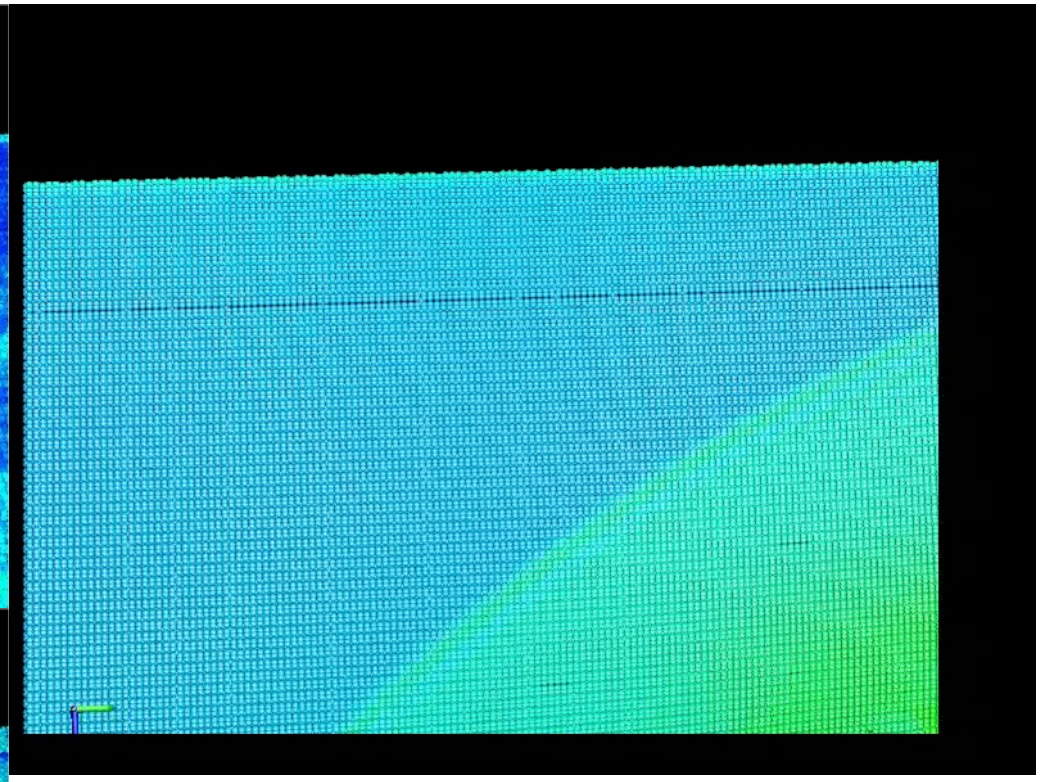




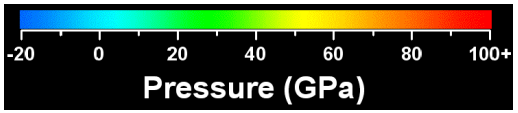
209 million atom MD of hypervelocity impact in AlN for the design of light-weight ceramic armors

Crack Nucleation at Kink Bands



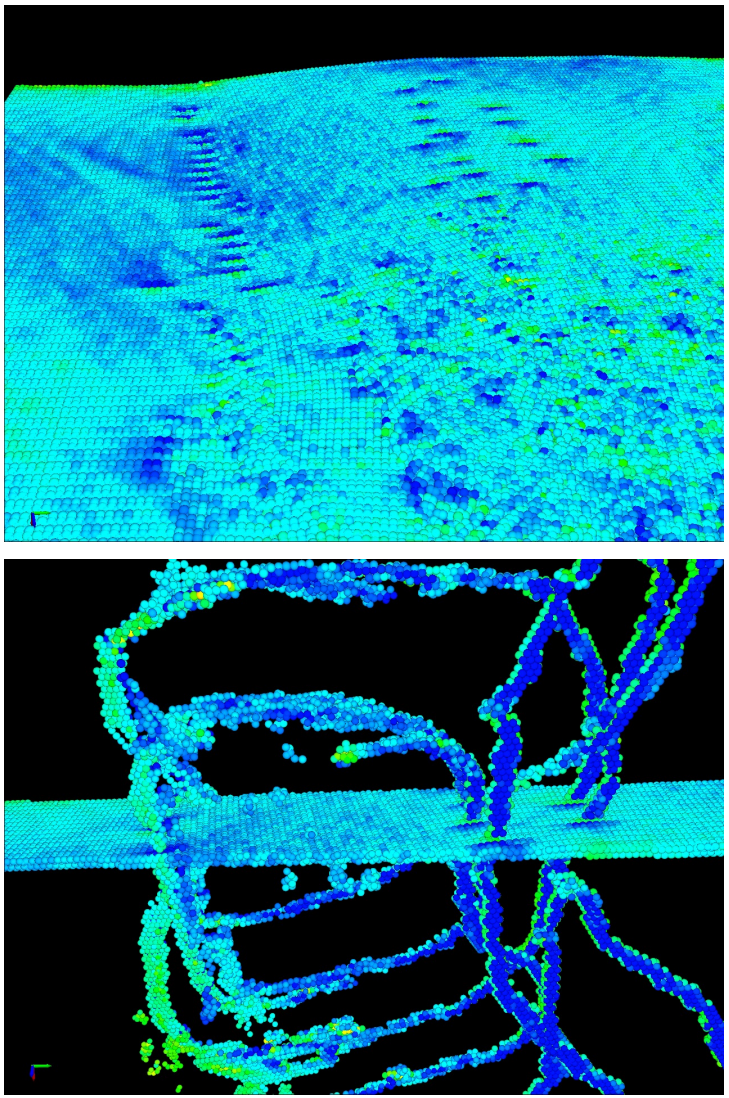


- Series of dislocation dipoles with opposite Burgers vectors form a kink band to releases stress
- Tilt grain boundaries of the kink bands act as sources of mode-II (shear) crack nucleation

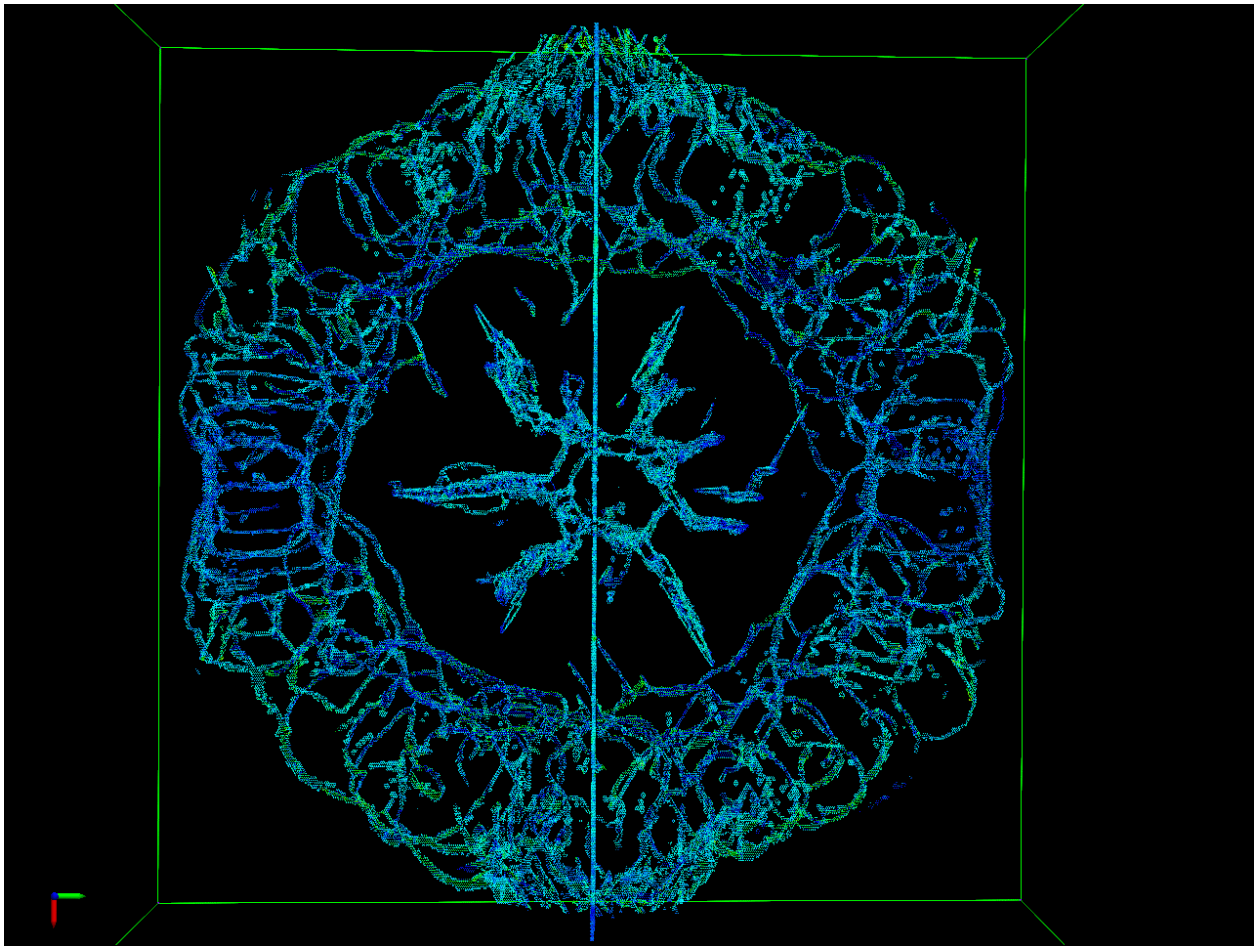


Dislocation Loops at Kink Bands

Graph (shortest-path circuit) based mining of topological defects



Atoms participating in non-6-member circuits



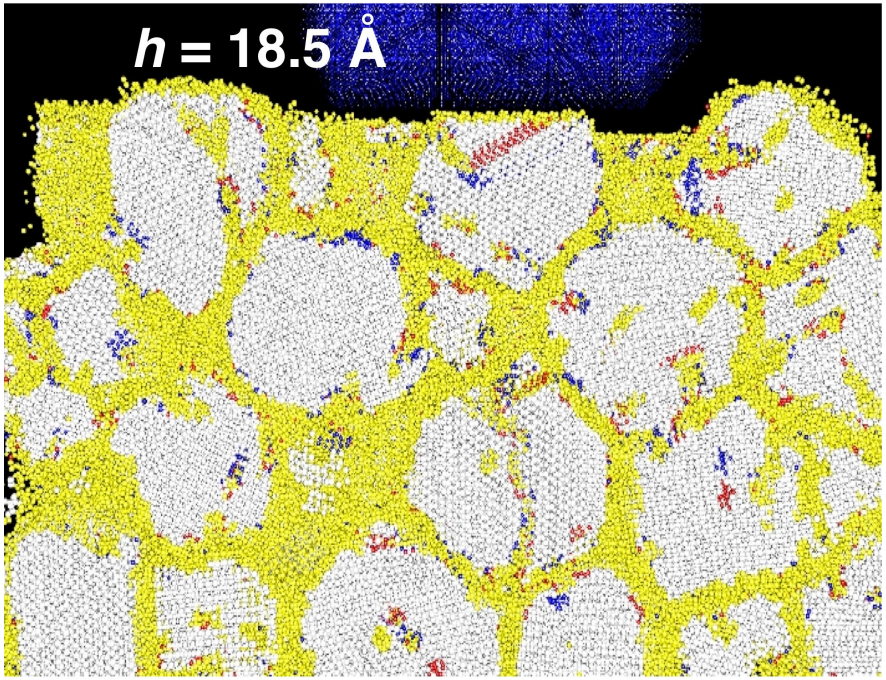
Dislocation network

Nanoindentation on Nanophase SiC

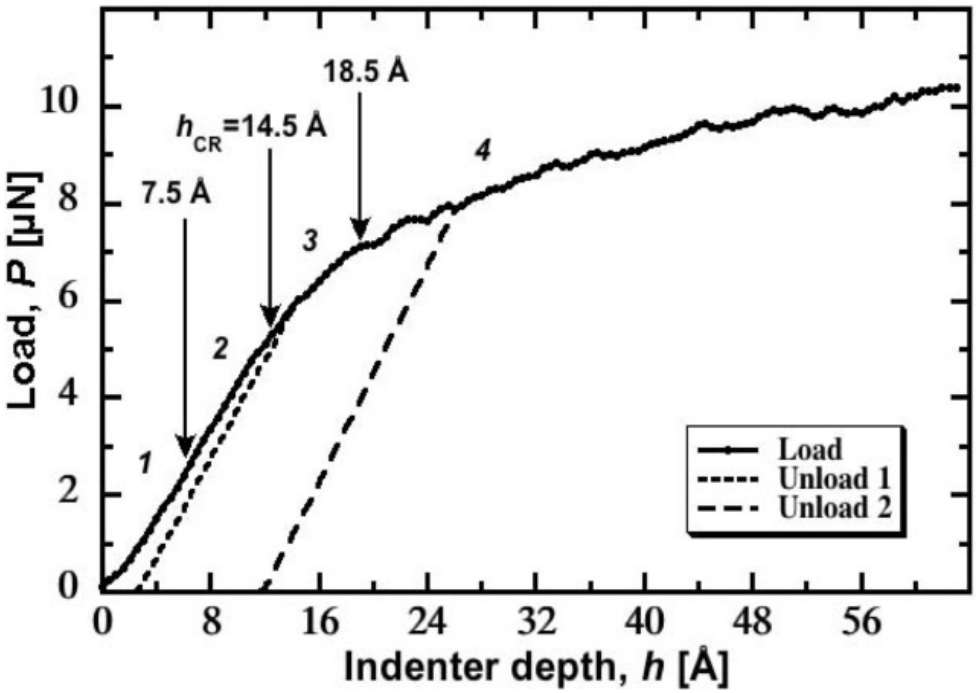
Superhardness

MD: 39 GPa
(grain size, d = 8 nm)

Expt.: 30-50 GPa
(d = 5-20 nm)
[Liao *et al.*, *APL*, '05]



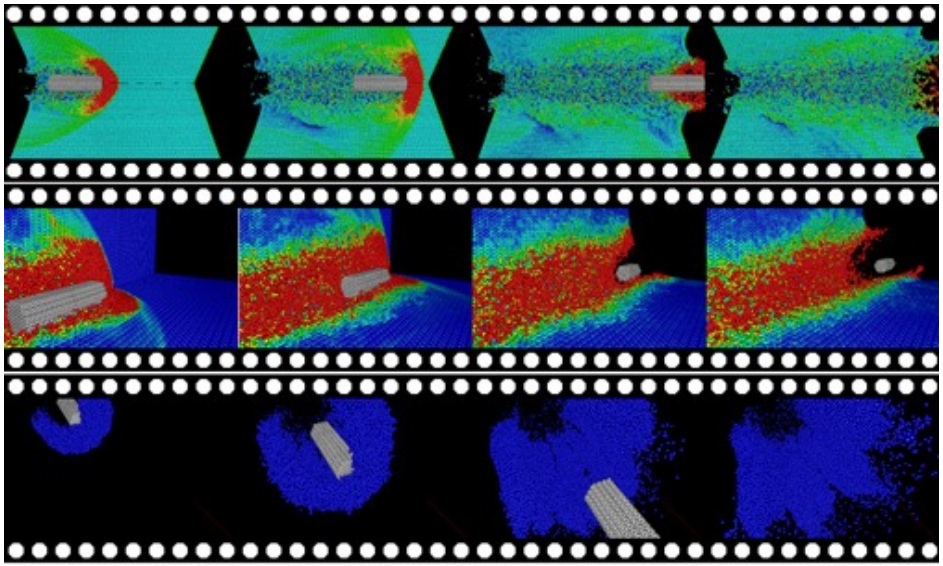
Load-displacement curve

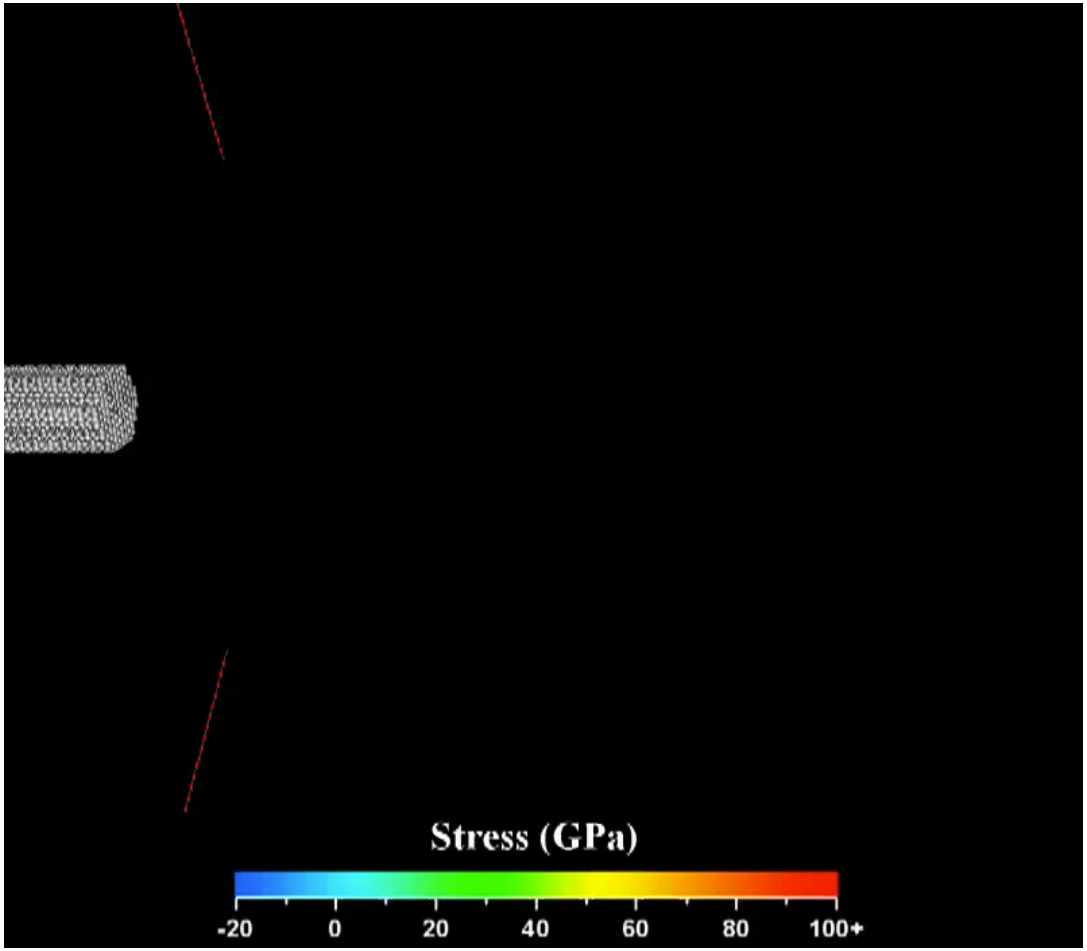


Crossover from intergrain continuous response to intragranular discrete response

Szlufarska, Nakano & Vashishta, *Science* **309**, 911 ('05)

Multimodal Multidisplay Visualization





Hypervelocity penetration through an AlN plate

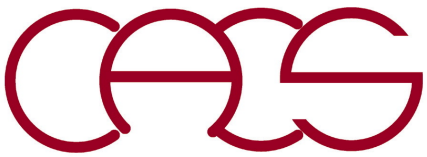
Singular Value Decomposition & Data Mining

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy,
Dept. of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu

Data mining ≅ data compression



Rank of a Matrix

- ***N*×*M* matrix *A* as a mapping: $R^M \rightarrow R^N$**

$$M \begin{bmatrix} x \end{bmatrix}^{1} \quad x\left(\in R^M\right) \xrightarrow[A]{} b = Ax\left(\in R^N\right) \quad \begin{bmatrix} b \end{bmatrix}^{1} \quad N$$

- **Range of *A*: Vector space $\{b = Ax | \forall x\}$**
- **Rank of *A*: Number, *m*, of linearly-independent vectors in the range, *i.e.*, how many linearly-independent *N*-element vectors are there in the range, such that**

$$b = A^{\forall} x = \sum_{\nu=1}^{m} c_\nu \left| \nu \right\rangle$$

Low Rank Approximations of a Matrix

- **Rank-1 approximation: $NM \rightarrow N + M$**

$$\underset{N}{\left[\begin{array}{c}\overset{M}{\psi}\end{array}\right]} \cong \left[\begin{array}{c} u \end{array}\right]\left[\begin{array}{c} v \end{array}\right] \qquad |u\rangle\langle v|\forall x\rangle \propto |u\rangle$$

- **Rank-2 approximation: $NM \rightarrow 2(N + M)$**

$$\left[\begin{array}{c}\psi\end{array}\right] \cong \left[\begin{array}{c} u_1 \end{array}\right] w_1 \left[\begin{array}{c} v_1 \end{array}\right] + \left[\begin{array}{c} u_2 \end{array}\right] w_2 \left[\begin{array}{c} v_2 \end{array}\right]$$

- **Rank-m ($m << N, M$) approximation: $NM \rightarrow m(N + M)$**

$$\left[\begin{array}{c}\psi\end{array}\right] \cong \sum_{\nu=1}^{m} \left[\begin{array}{c} u_\nu \end{array}\right] w_\nu \left[\begin{array}{c} v_\nu \end{array}\right]$$

Singular Value Decomposition

- **Problem:** Optimal approximation of an $N\times M$ matrix ψ of rank-m ($m << N$)?
- **Theorem:** An $N\times M$ matrix ψ (assume $N \geq M$) can be decomposed as

$$\psi = UDV^T = \sum_{\nu=1}^{M} U_{i\nu} d_\nu V_{j\nu} = \sum_{\nu=1}^{M} u_i^{(\nu)} d_\nu v_j^{(\nu)}$$

where $U \in R^N \times R^M$ & $V \in R^M \times R^M$ are column orthogonal & D is diagonal

$$U^T U = V^T V = I_M$$

$$\underset{N}{\left[\begin{array}{c}\\ \psi \\ \\ \end{array}\right]}_{M} = \underset{N\times M}{\left[\begin{array}{c}\\ U \\ \\ \end{array}\right]} \underset{M\times M}{\left[\begin{array}{ccc} d_1 & & \\ & \ddots & \\ & & d_M \end{array}\right]} \underset{M\times M}{\left[\begin{array}{c}\\ V^T \\ \\ \end{array}\right]}$$

See appendix on polar & singular decompositions

- **Theorem:** Sort the SVD diagonal elements in descending order, $d_1 \geq d_2 \geq \ldots \geq d_M \geq 0$, & retain the first m terms

$$\psi^{(m)} \equiv \sum_{\nu=1}^{m} u^{(\nu)} d_\nu v^{(\nu)T}$$

which is optimal among $\forall$rank-m matrices in the 2-norm sense with the error

$$\min_{rank(A)=m} \|A - \psi\|_2 = \left\|\psi^{(m)} - \psi\right\|_2 = d_{m+1}$$

cf. singular.c & svdcmp.c

Use the program!

SVD for Image Compression

Original Image

5 Iterations

10 Iterations

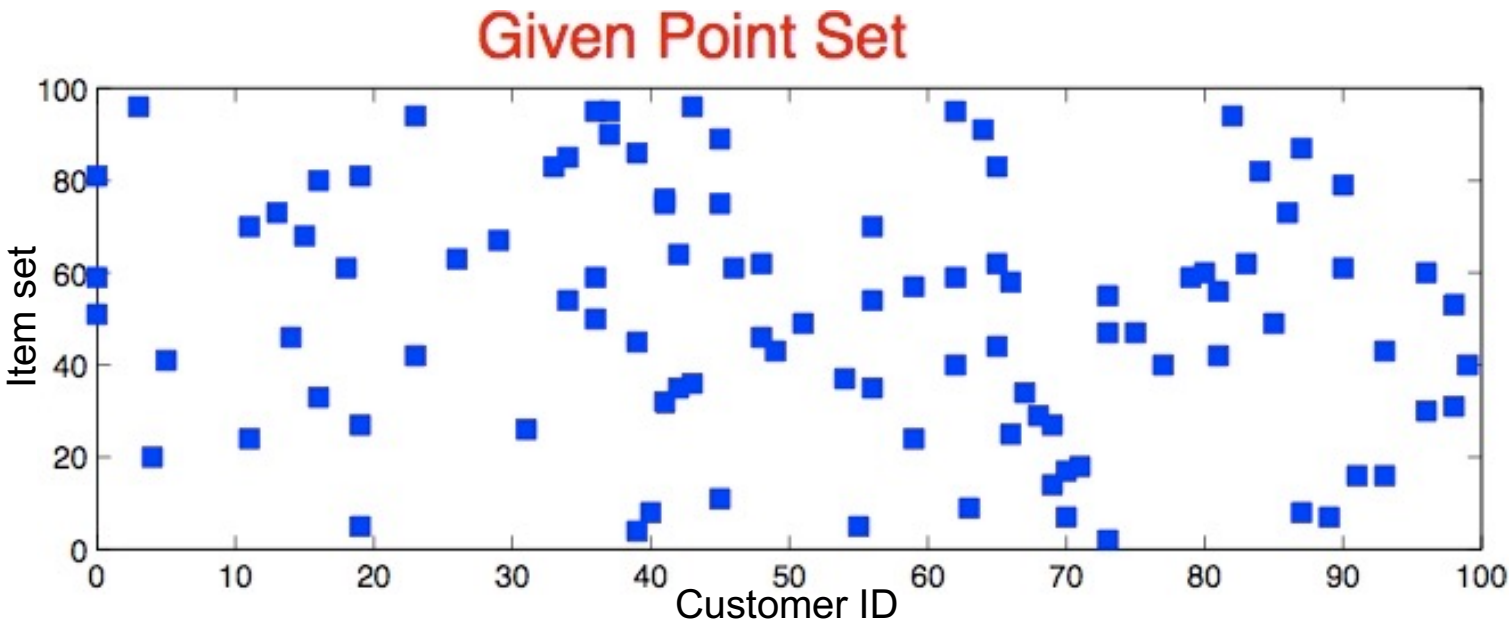
D. Richards & A. Abrahamsen

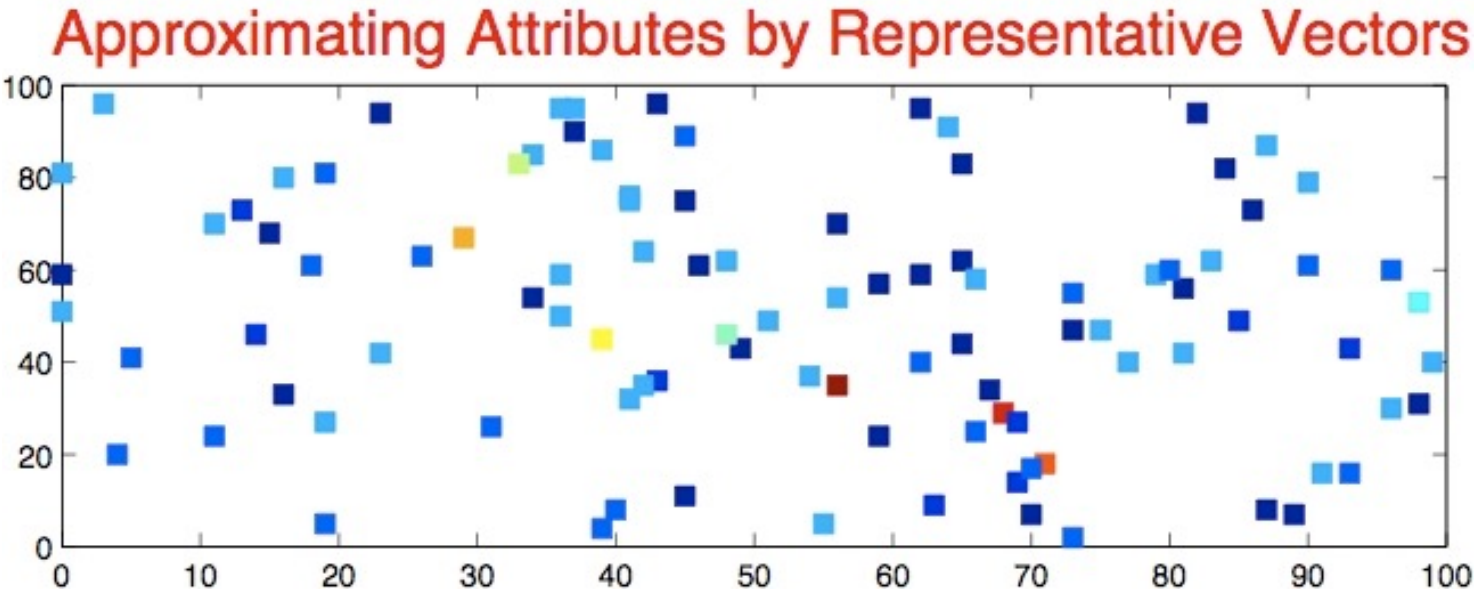
20 Iterations

60 Iterations

100 Iterations

SVD in Data Mining





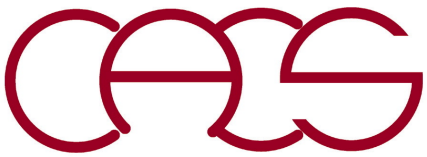
N. Ramakrishnan & A. Y. Grama

Machine Learning in Simulation

Aiichiro Nakano

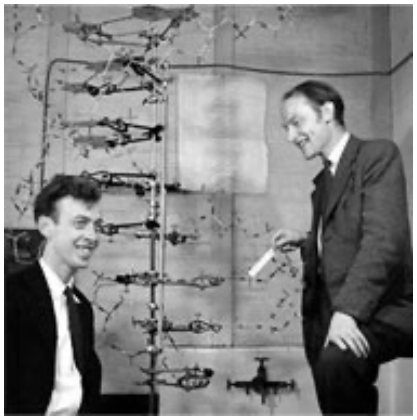
Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy,
Dept. of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu

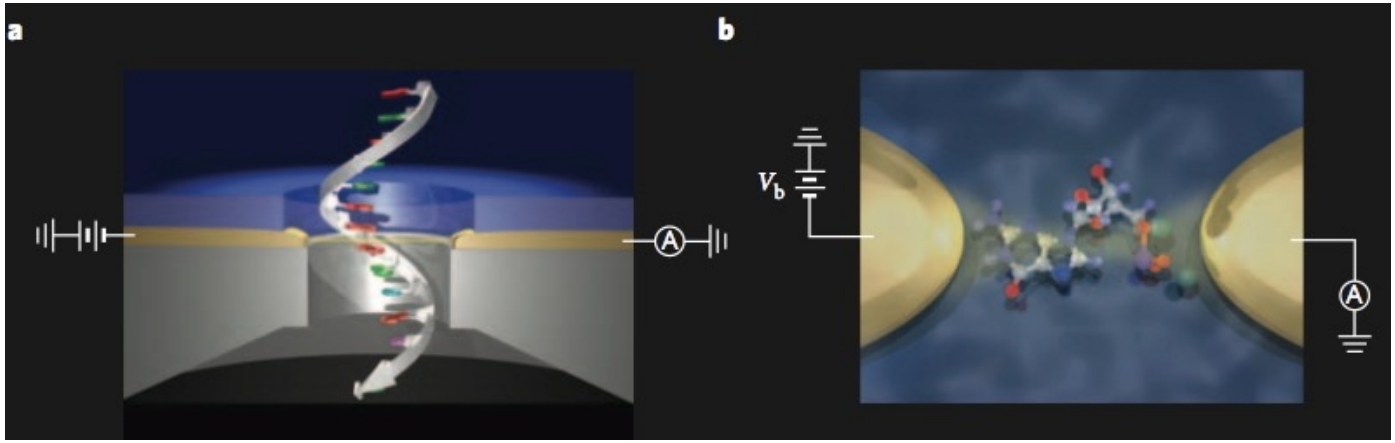


SVD for Rapid Genome Sequencing

- **$10M Archon X prize for decoding 100 human genomes in 10 days & $10K per genome (**http://genomics.xprize.org**): Preemptive attack on diseases**

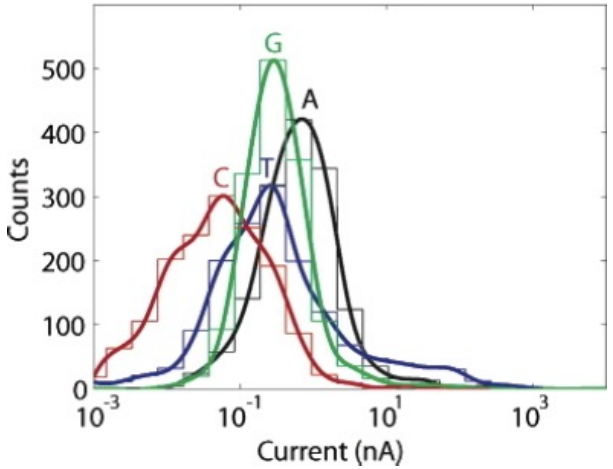


- **Quantum tunneling current for rapid DNA sequencing?**



Tsutsui *et al*., *Nature Nanotechnology* ('10)

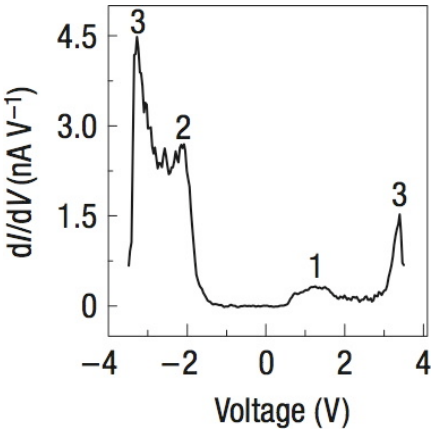
Lagerqvist *et al*., *Nano Letters* ('06)



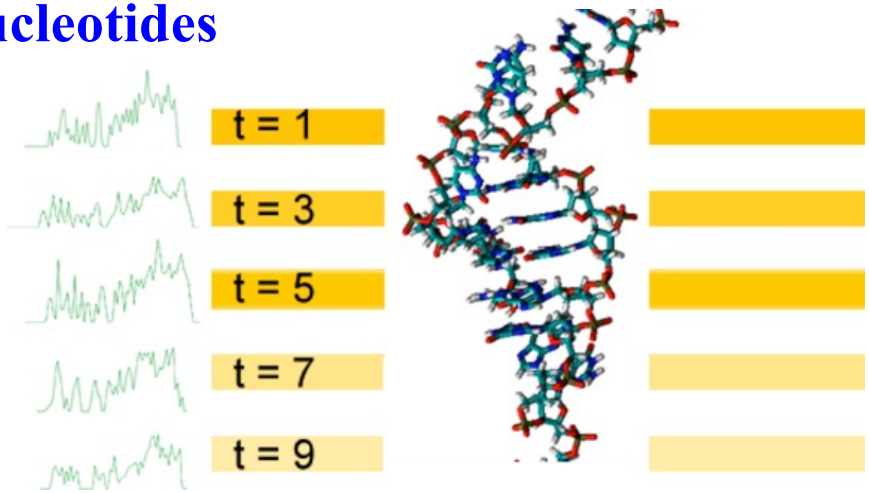
- **Tunneling current alone cannot distinguish the 4 nucleotides (A, C, G, T)**

Rapid DNA Sequencing *via* Data Mining

- **Use tunneling current (I)-voltage (V) characteristic (or electronic density-of-states) as the 'fingerprints' of the 4 nucleotides**

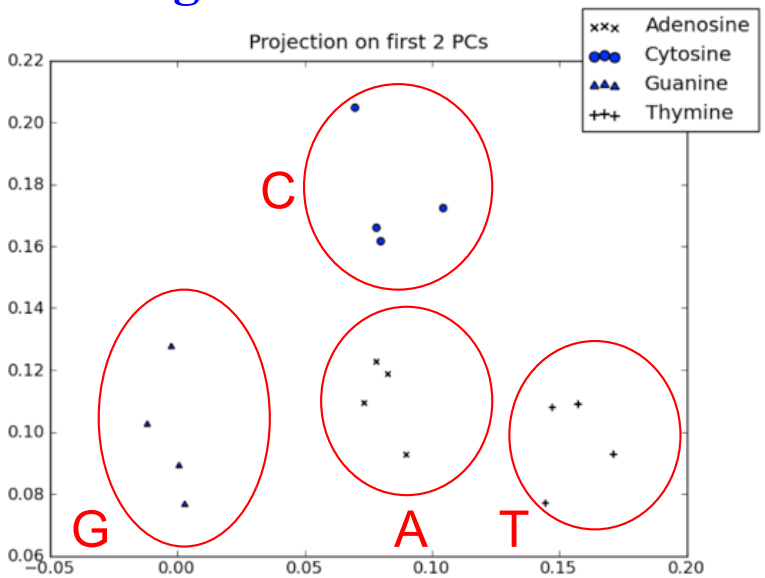


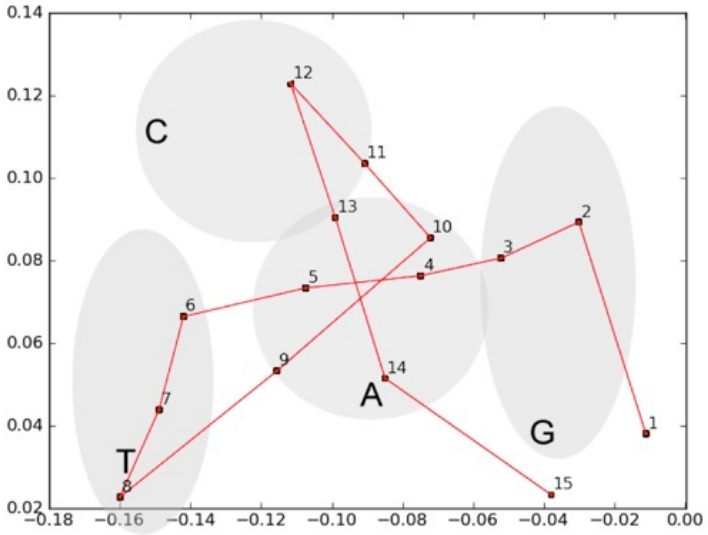
Shapir *et al.*,
Nature Materials ('08)



- **Principal component analysis (PCA) & fuzzy c-means clustering clearly distinguish the 4 nucleotides**

H. Yuen *et al.*, *IJCS* **4**, 352 ('10)





http://www.henryyuen.net/

- **Viterbi algorithm for even higher-accuracy sequencing**

SVD *vs.* PCA (in Economics)

- **SVD of N (number of companies) × T (number of time points) of stock-price time series**

$$\underset{T\times N}{\Xi}^{T} = \underset{T\times N}{\mathbf{U}}\ \underset{N\times N}{\Sigma}\ \underset{N\times N}{\mathbf{V}}^{T}$$

- **Stock correlation matrix**

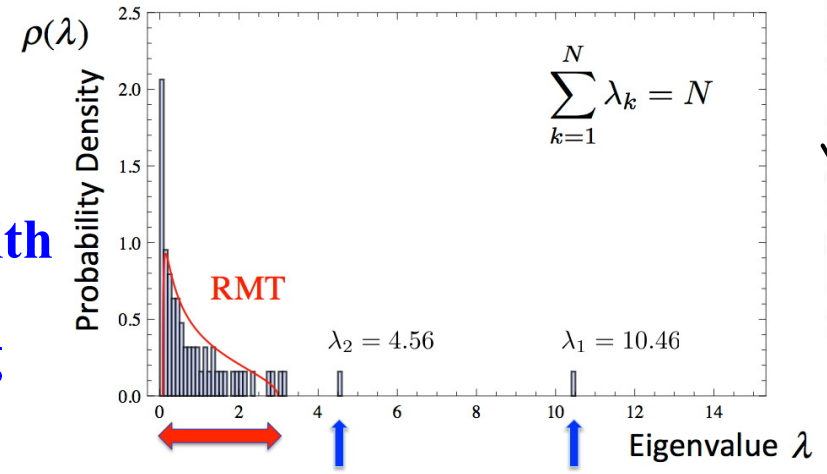
$$\underset{N\times N}{\mathbf{C}} = \underset{N\times T}{\Xi}\ \underset{T\times N}{\Xi}^{T}$$

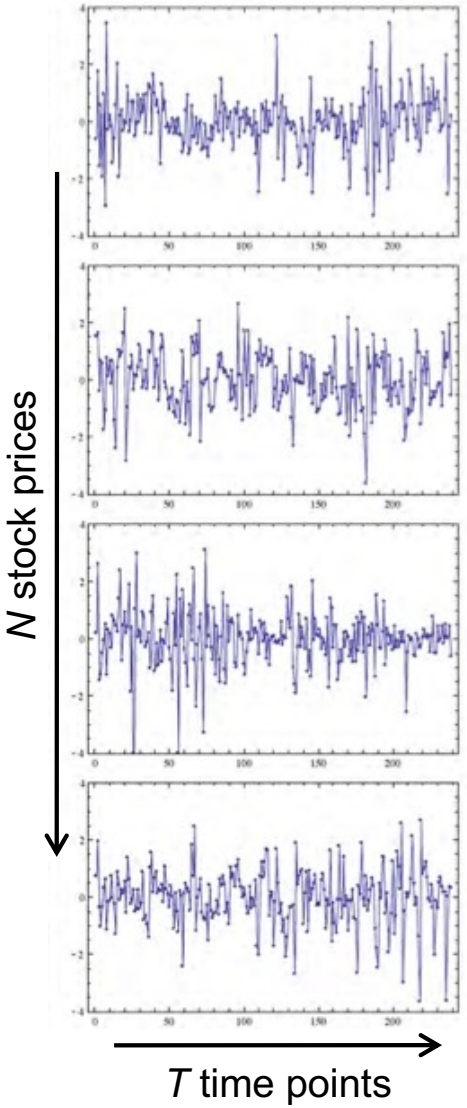
- **Principal component analysis (PCA): Eigen decomposition of the correlation matrix**

$$\mathrm{C} = \Xi\Xi^{T}$$
$$= \mathrm{V}\Sigma\overbrace{\mathrm{U}^{T}\mathrm{U}}^{\mathrm{I}}\Sigma\mathrm{V}^{T}$$
$$= \mathrm{V}\Sigma^{2}\mathrm{V}^{T}$$

- **Compare the spectrum with that of random matrix theory (RMT) for judging statistical significance**

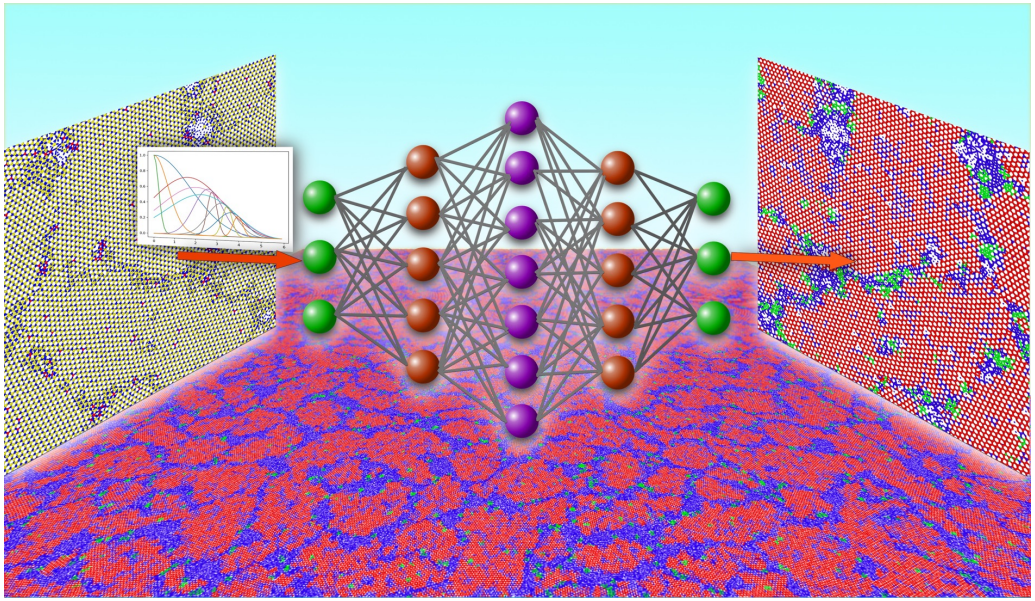
$$\rho_{\mathrm{RMT}} = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_{+}-\lambda)(\lambda-\lambda_{-})}}{\lambda}$$
$$\lambda_{\pm} = \left(1+\frac{1}{\sqrt{Q}}\right)^{2}$$
$$Q = T/(2N) \quad N,T\to\infty$$





Y. Kichikawa *et al.*, *Proc. Comp. Sci.* **60**, 1836 ('15)

Learning Materials Phases & Defects

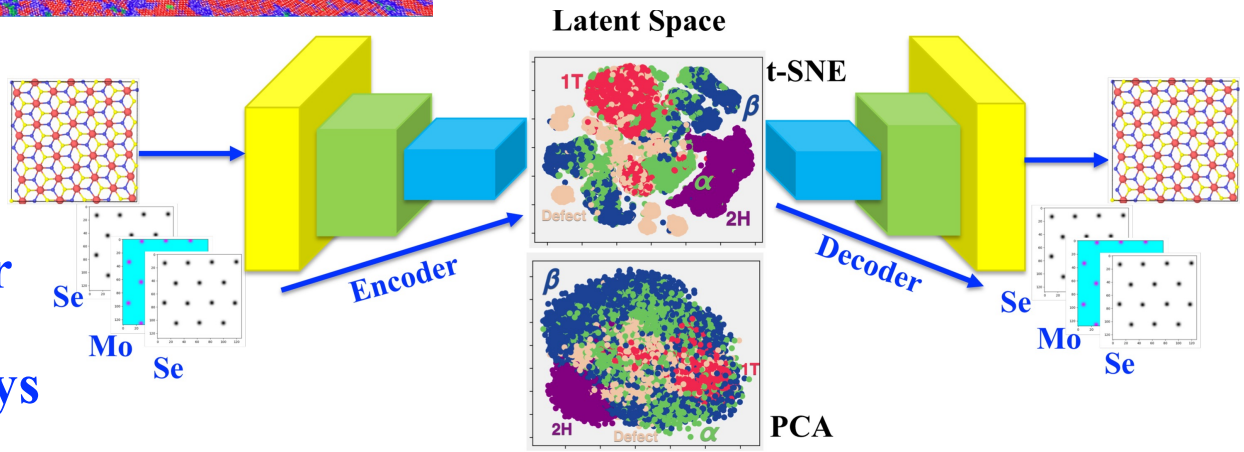


- **Feedforward neural network to learn phases from local symmetry functions**

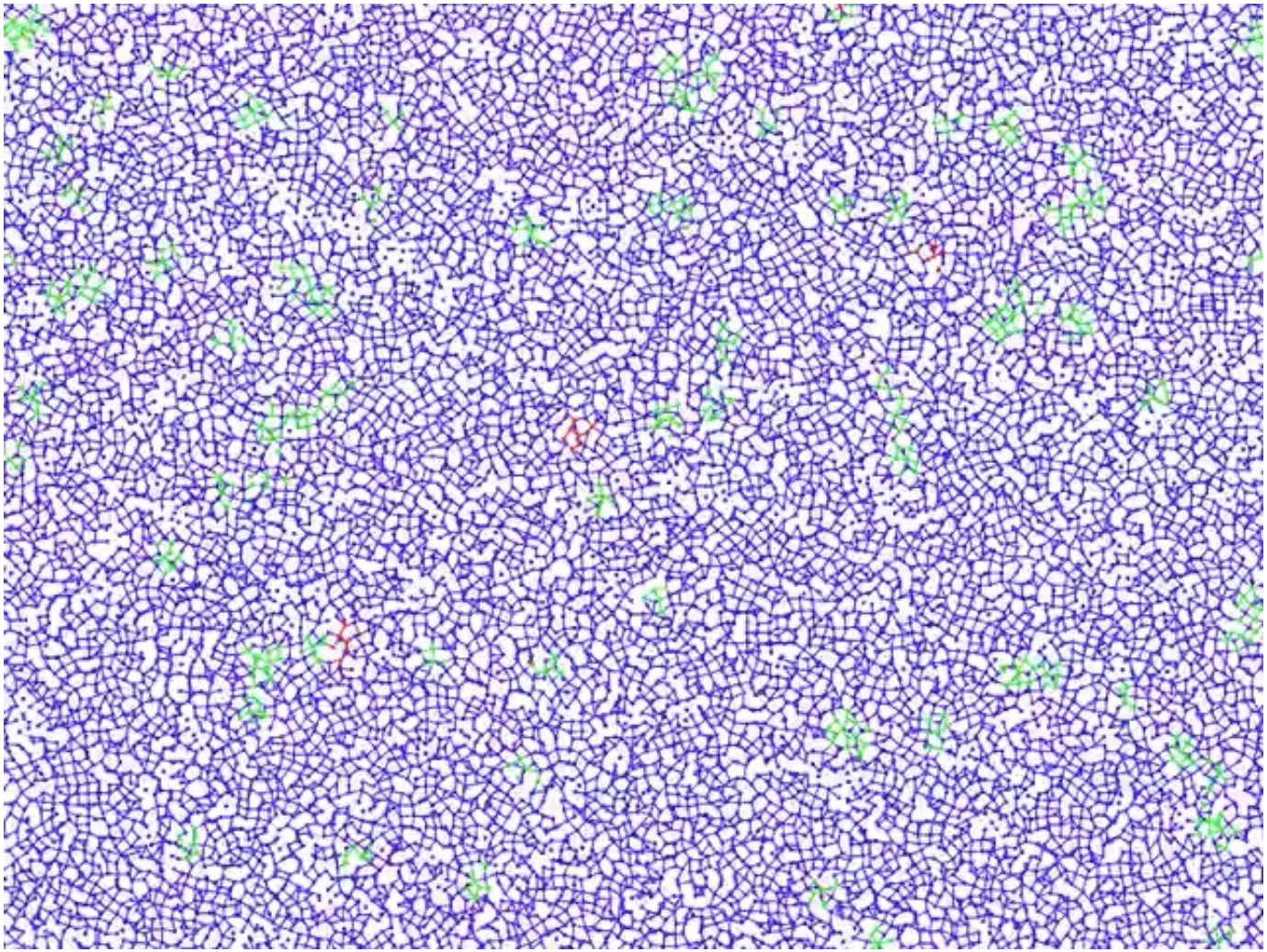
K. Liu *et al.*, *Proc. ScalA18* ('18)
S. Hong *et al.*, *JPCL* **10**, 2739 ('19)

- **Variational autoencoder to generate transformation pathways from images & latent-space algebra**

P. Rajak *et al.*, *Phys. Rev. B* **100**, 014108 ('19)



Learning Transformation Pathways



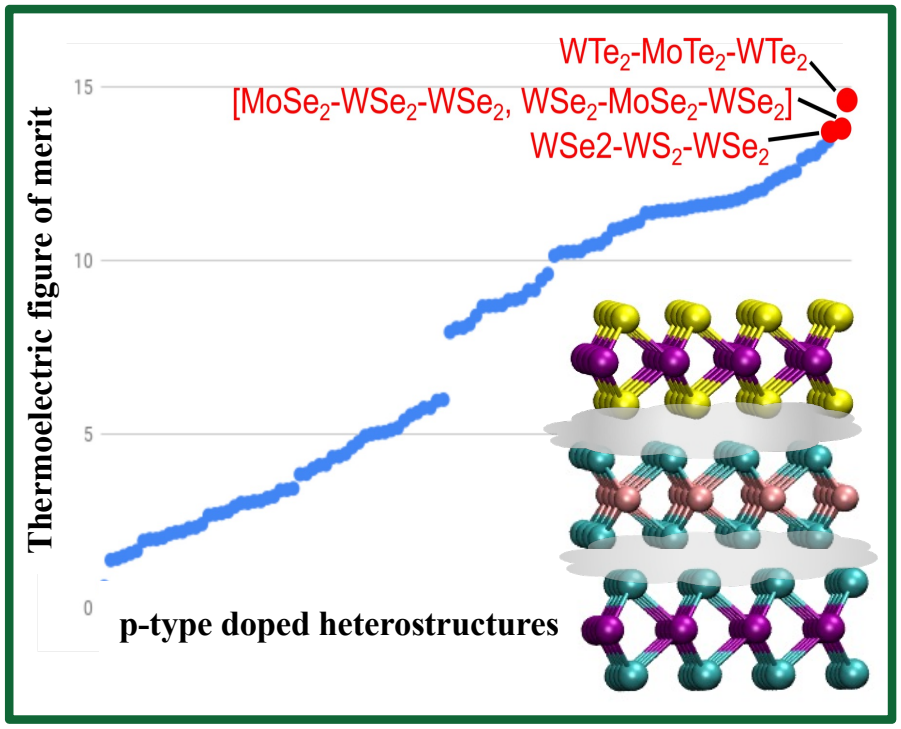
2H
1T
disorder

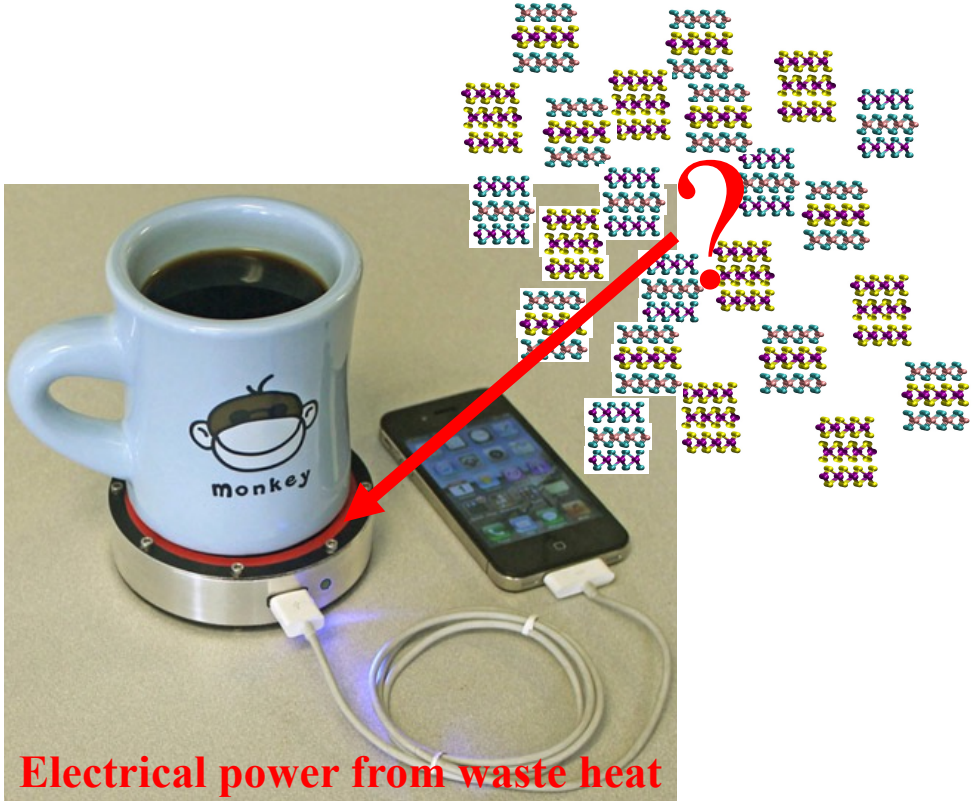
- **Found novel transformation pathways to the stable 2H phase *via* the metastable 1T phase during chemical vapor deposition (CVD) growth of MoS_2**

S. Hong *et al.*, *J. Phys. Chem. Lett.* **10**, 2739 ('19)

Active Learning of Optimal Materials

- **Bayesian optimization balances exploitation & exploration to find a structure with the desired property with a minimal number of quantum-mechanical calculations**
- **Predicted three-layered transition-metal chalcogenide (TMDC) heterostacks with the largest thermoelectric figure-of-merit**

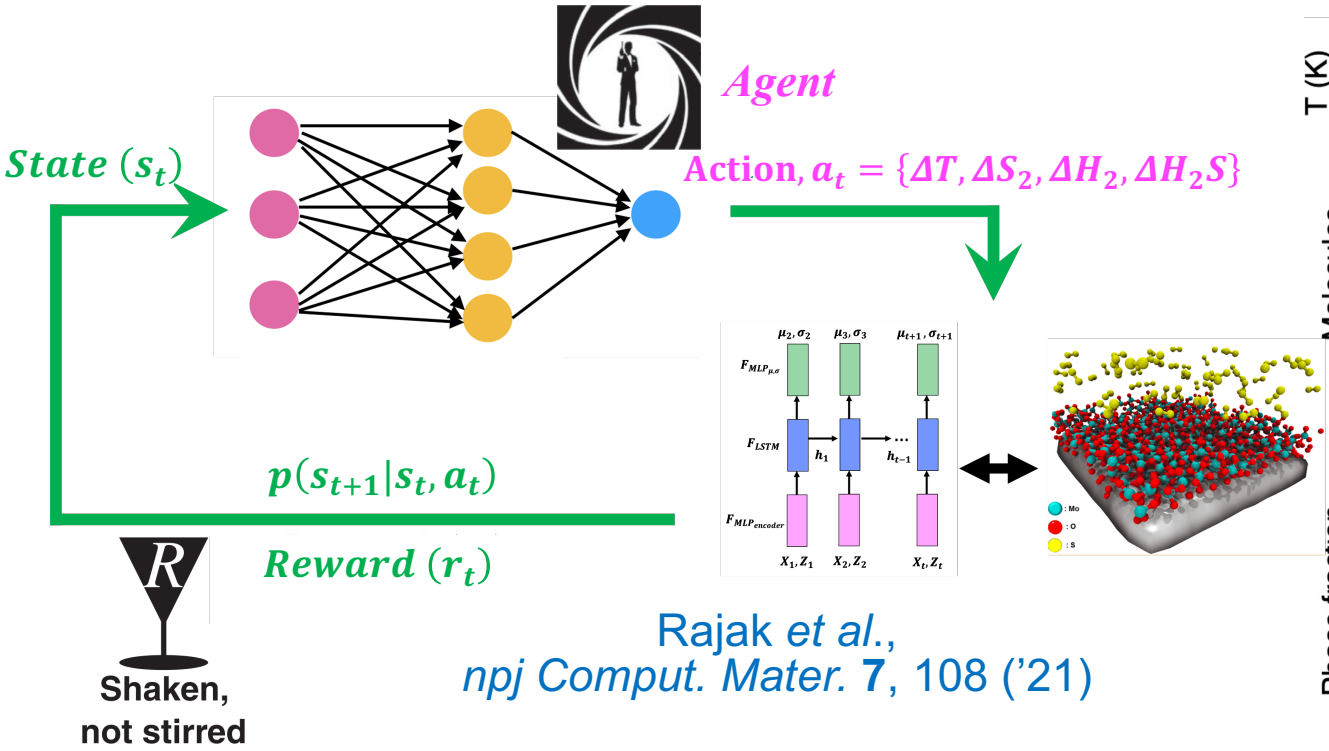




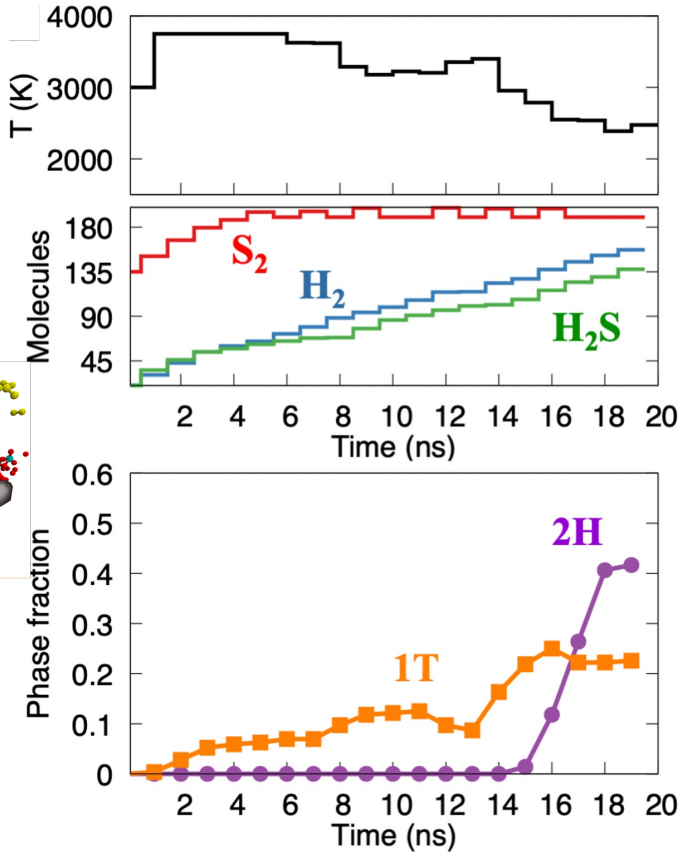
L. Bassman *et al.*, *npj Comput. Mater.* **4**, 74 ('18)

Reinforcement Learning for Growth

- In a manner AI plays a board game of Go, use reinforcement learning (RL) to design optimal growth conditions (*e.g.*, temperature & gas-pressure control) to achieve desired properties such as minimal defect density
- AI model combines:
  1. RL agent to design actions
  2. Nesural network-based dynamic model trained by molecular-dynamics (MD) simulation to predict new states



Rajak *et al.*,
npj Comput. Mater. **7**, 108 ('21)



cf. Sgroi *et al.*, *Phys. Rev. Lett.* ***126***, 020601 ('21)

AI Meets Kirigami

- **Reinforcement learning to design optimal kirigami with maximal stretchability**

Rajak *et al*., *npj Comput. Mater.* **7**, 102 ('21)

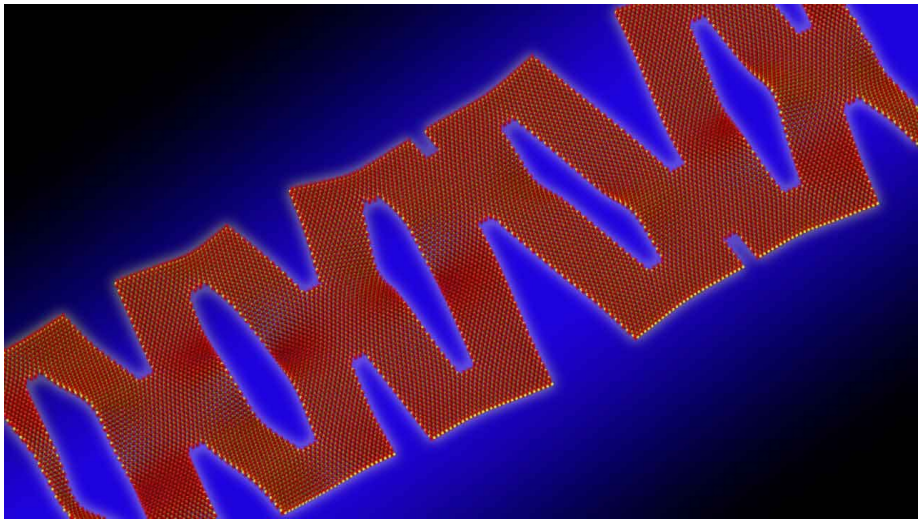
FEATURE STORY | ARGONNE NATIONAL LABORATORY

Ancient art meets AI for better materials design

BY JOHN SPIZZIRRI | APRIL 7, 2022

Ancient Japanese art of kirigami guides artificial intelligence (AI) technique for durable, wearable electronics.

Kirigami is the Japanese art of paper cutting. Likely derived from the Chinese art of jiǎnzhǐ, it emerged around the 7th century in Japan,



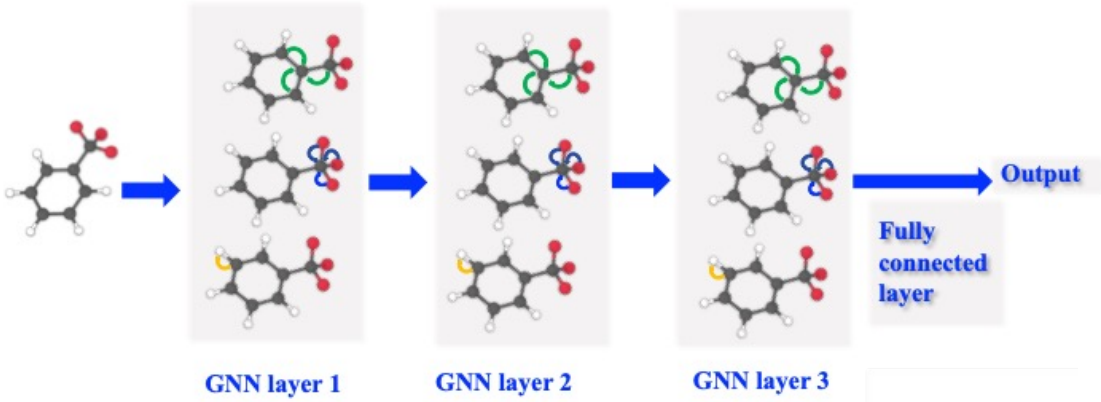
https://www.anl.gov/article/ancient-art-meets-ai-for-better-materials-design

Dielectric Polymer Genome

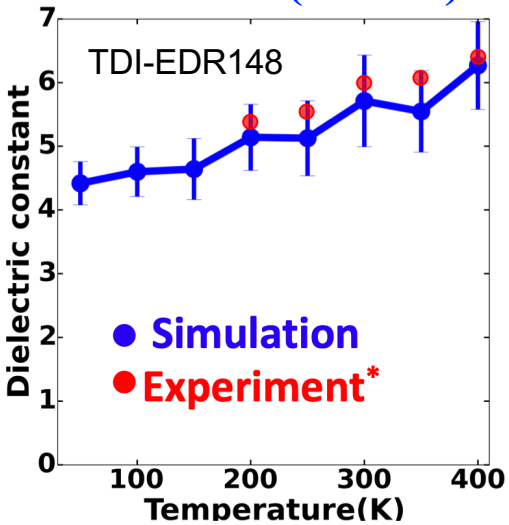
Recurrent neural network for polymer property prediction

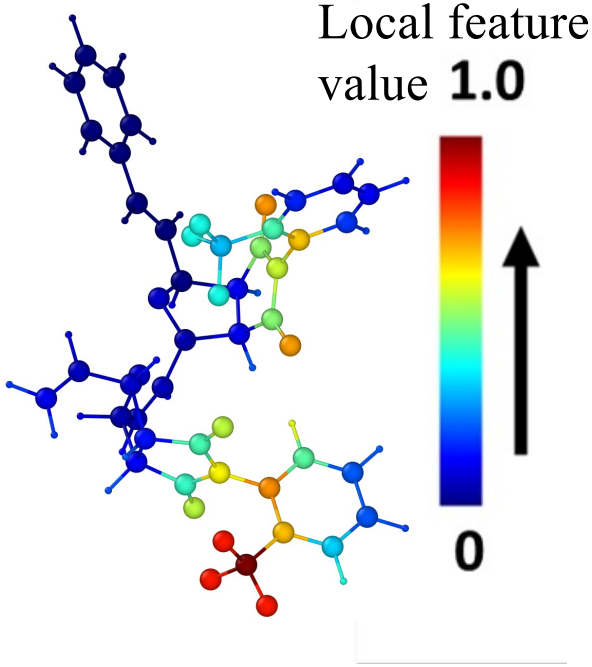
Nazarova *et al.*,
J. Chem. Info. Model. **61**, 2175 ('21)

Graph attention neural network for explainable property prediction



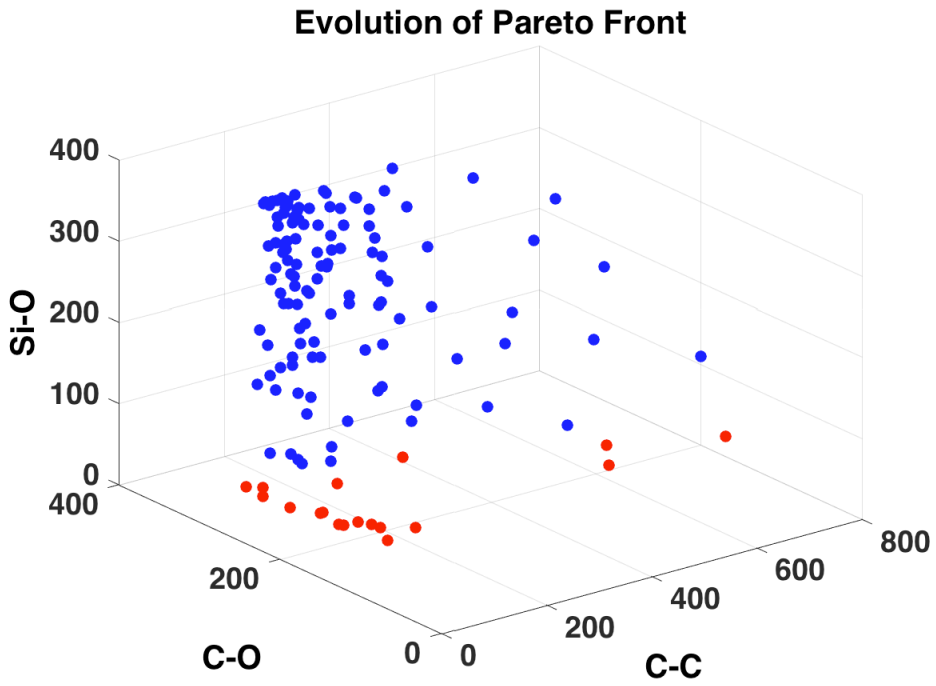
Experimental validation (Uconn)

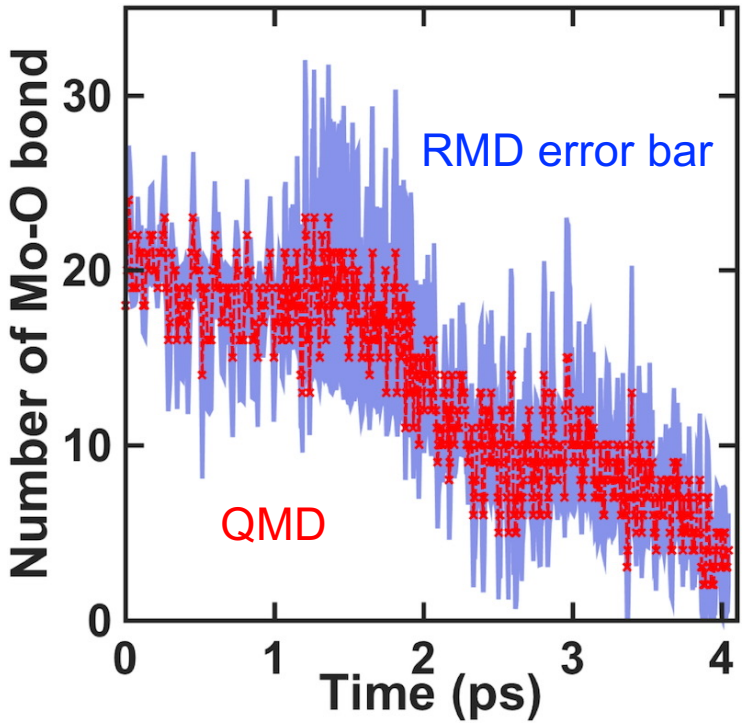




Pareto-Frontal Uncertainty Quantification

- **Train reactive force-field parameters by dynamically fitting reactive molecular dynamics (RMD) trajectories to quantum molecular dynamics (QMD) trajectories on-the-fly**
- **Pareto optimal front in multiobjective genetic algorithm (MOGA) provides an ensemble of force fields to enable uncertainty quantification (UQ)**





- Pareto-optimal solutions during genetic training (RMD errors for three quantities-of-interest)
- Converged Pareto-optimal front

A. Mishra *et al.*, *npj Comput. Mater.* **4**, 42 ('18)