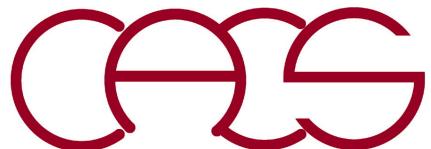


Scientific Data Mining & Machine Learning

Aiichiro Nakano

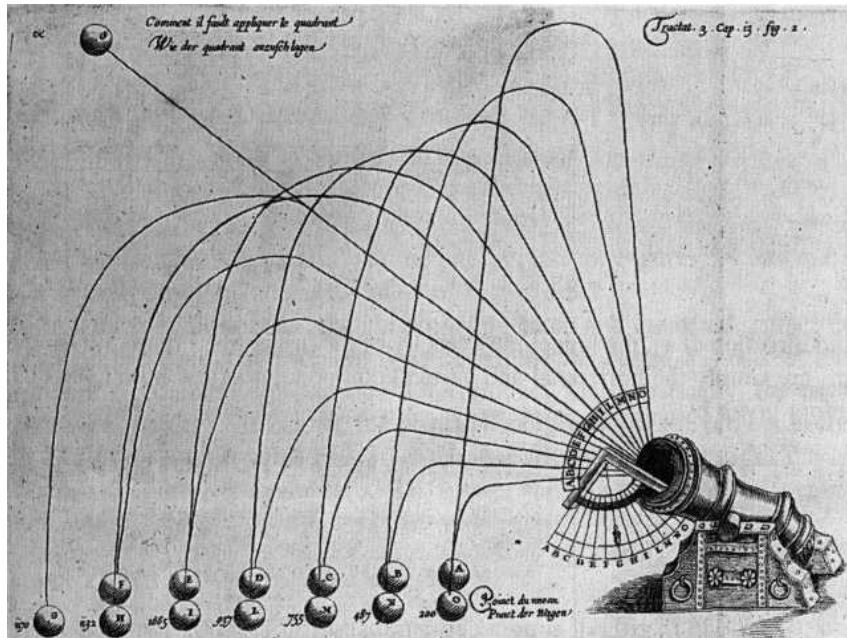
*Collaboratory for Advanced Computing & Simulations
Department of Computer Science
Department of Physics & Astronomy
Department of Quantitative & Computational Biology
University of Southern California*

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

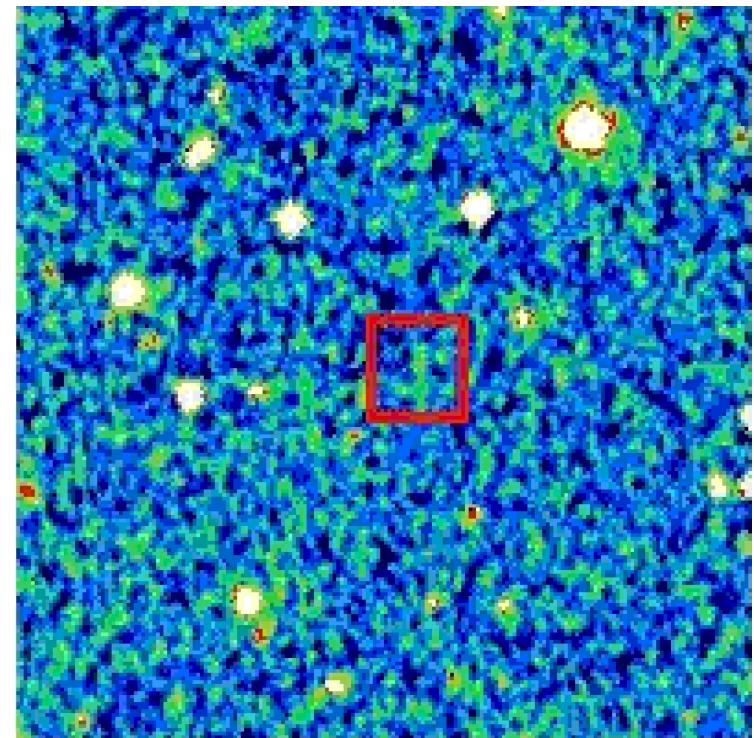
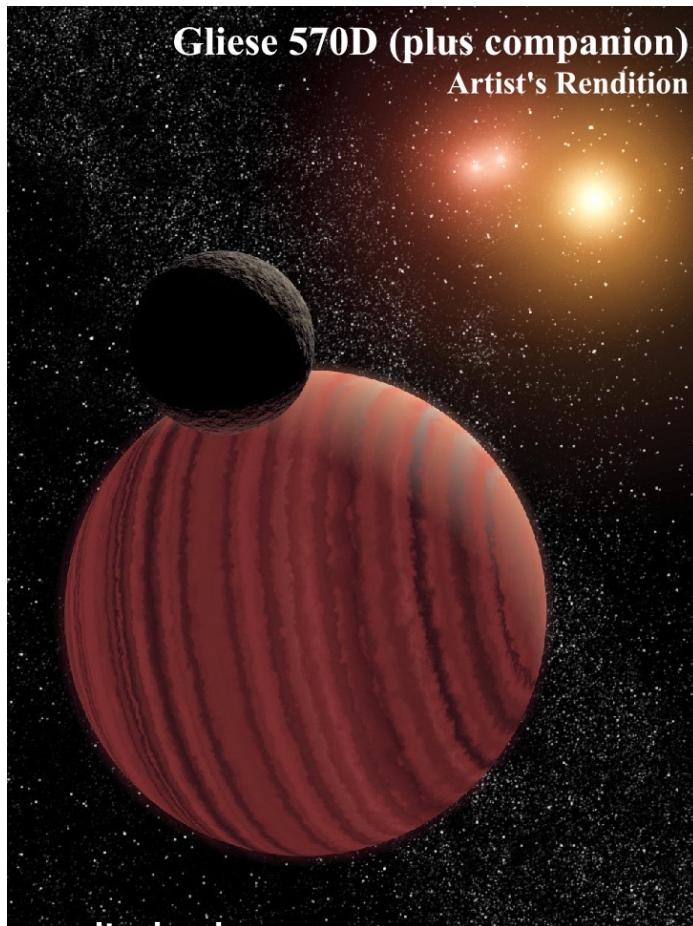


$$m \frac{d^2}{dt^2} \vec{r}(t) = \vec{F}$$

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)

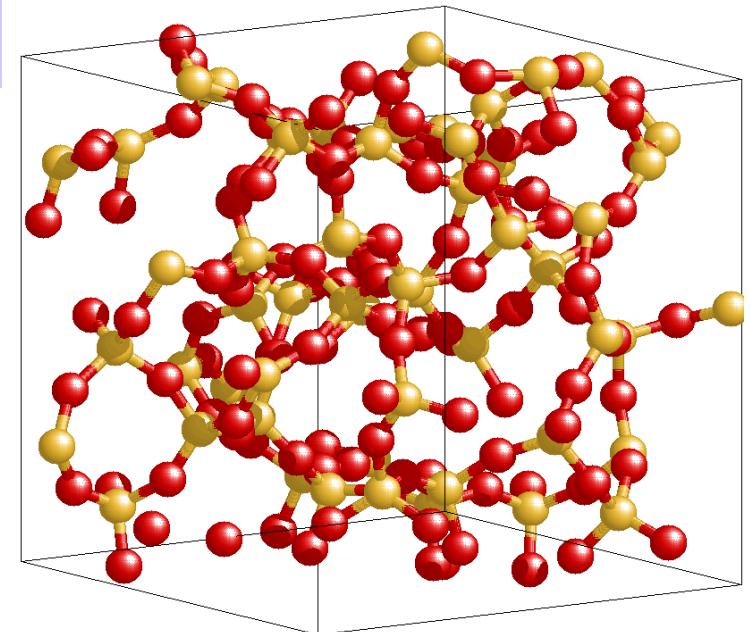


SDSS image of brown dwarf,
2MASSI J0104075-005328

Atomistic Data as a Graph

- Molecular dynamics data
 - Atomic data: species, positions, velocities, stresses, ...
$$\{\lambda_i, \vec{r}_i, \vec{v}_i, \vec{\sigma}_i, \dots | i = 1, \dots, N\}$$
 - Atomic-pair data: bond order, pair distance, ...
$$\{B_{ij}, \vec{r}_{ij}, \dots | i, j = 1, \dots, N; i \neq j\}$$
- Chemical bond network $G = (V, E)$
 - Node degrees
 - Paths
 - Rings
 - Frequently occurring subgraphs

V: Set of atoms
E: Set of bonds

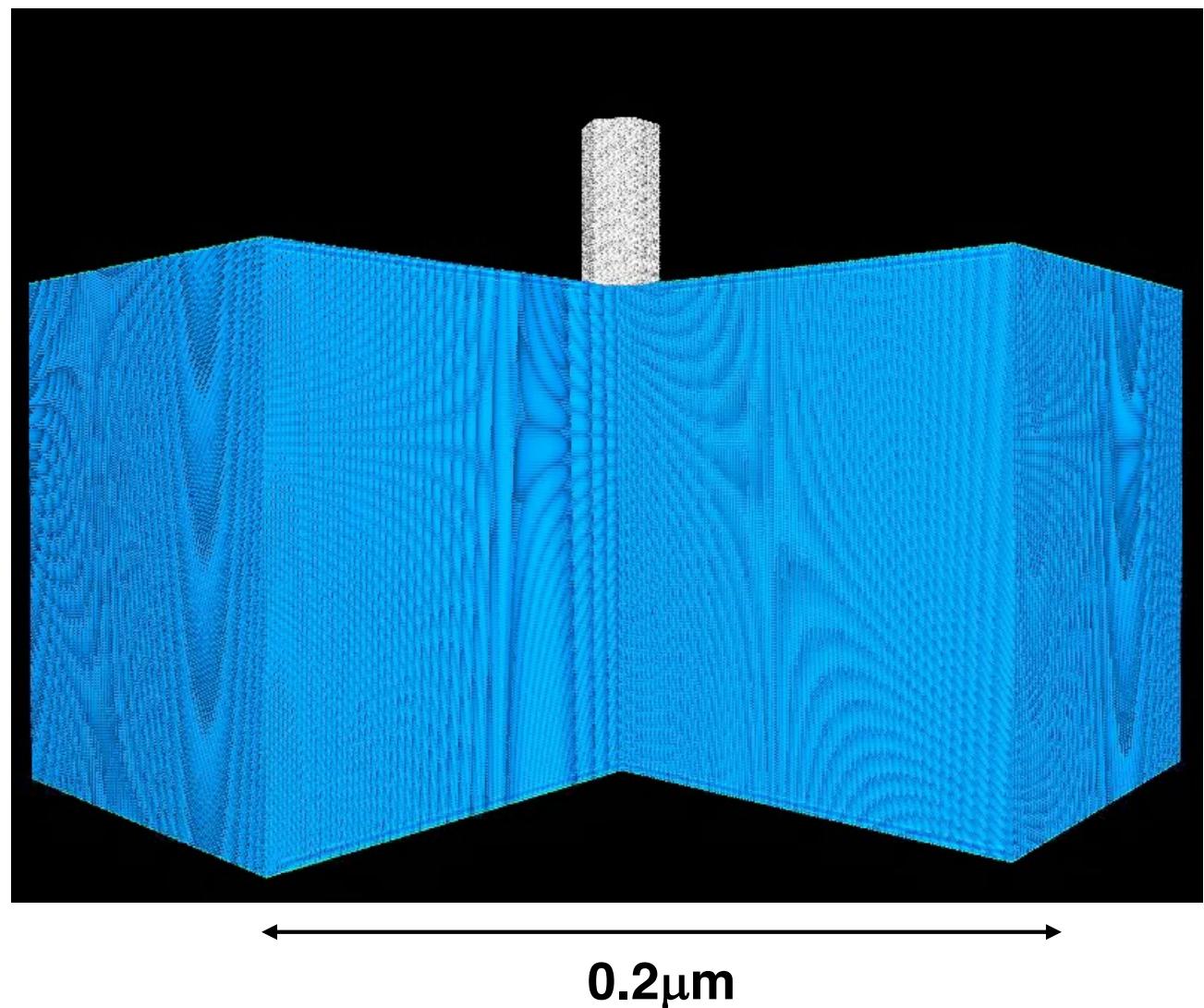


Hypervelocity Impact on Ceramics

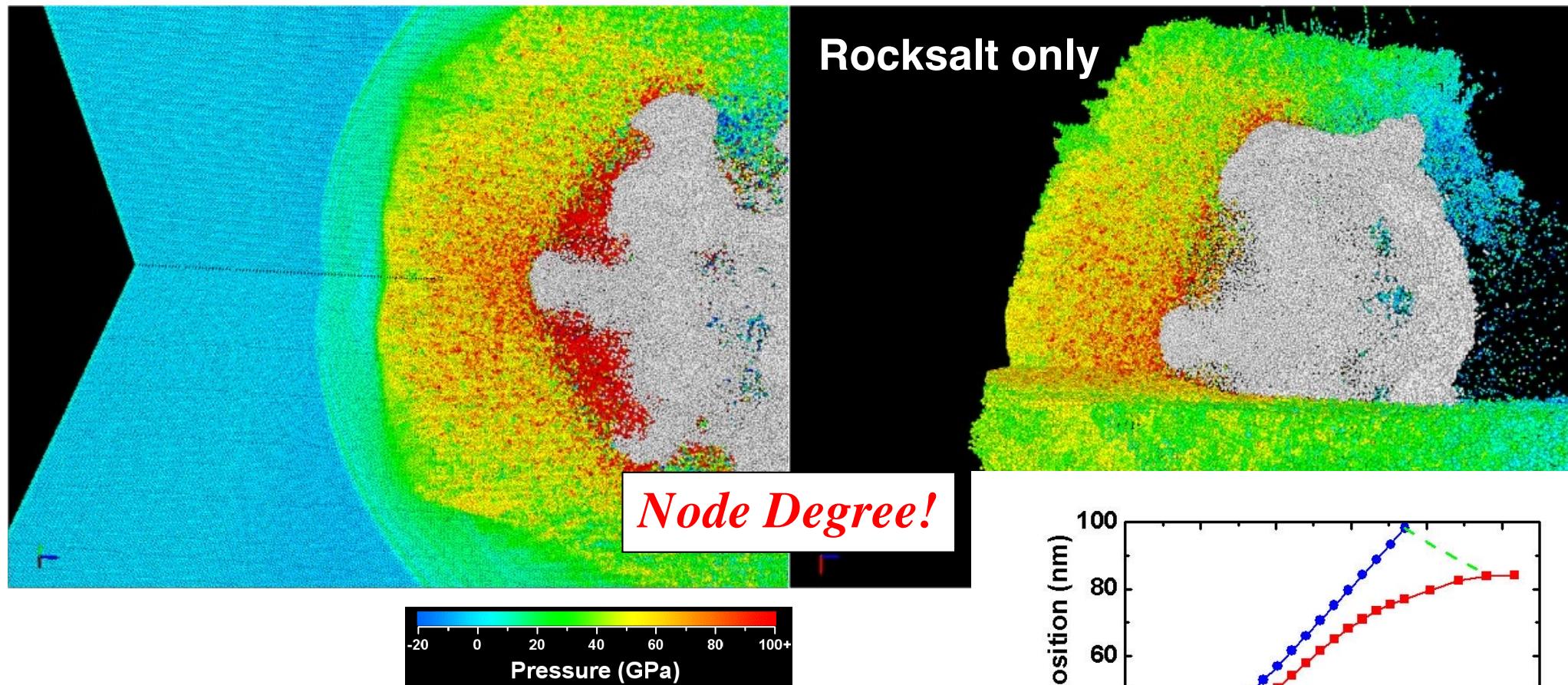
- 209M-atom MD of AlN
- 300M-atom MD of SiC
- 540M-atom MD of Al_2O_3

↑ [0001]

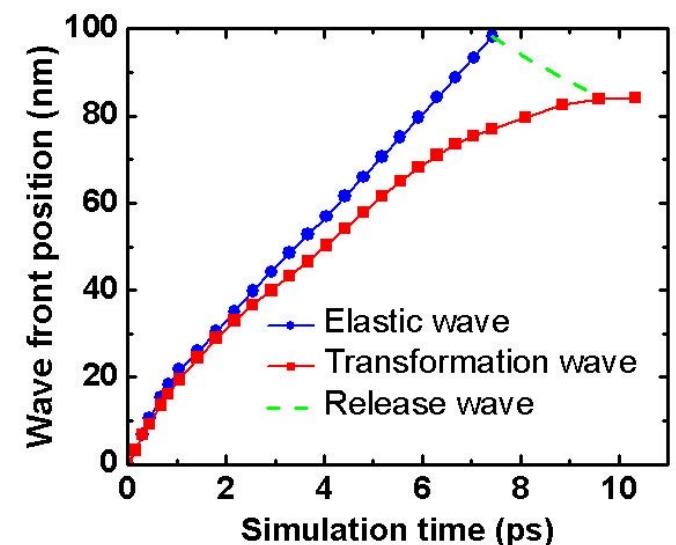
- Al_2O_3 plate
- 18 km/s impact



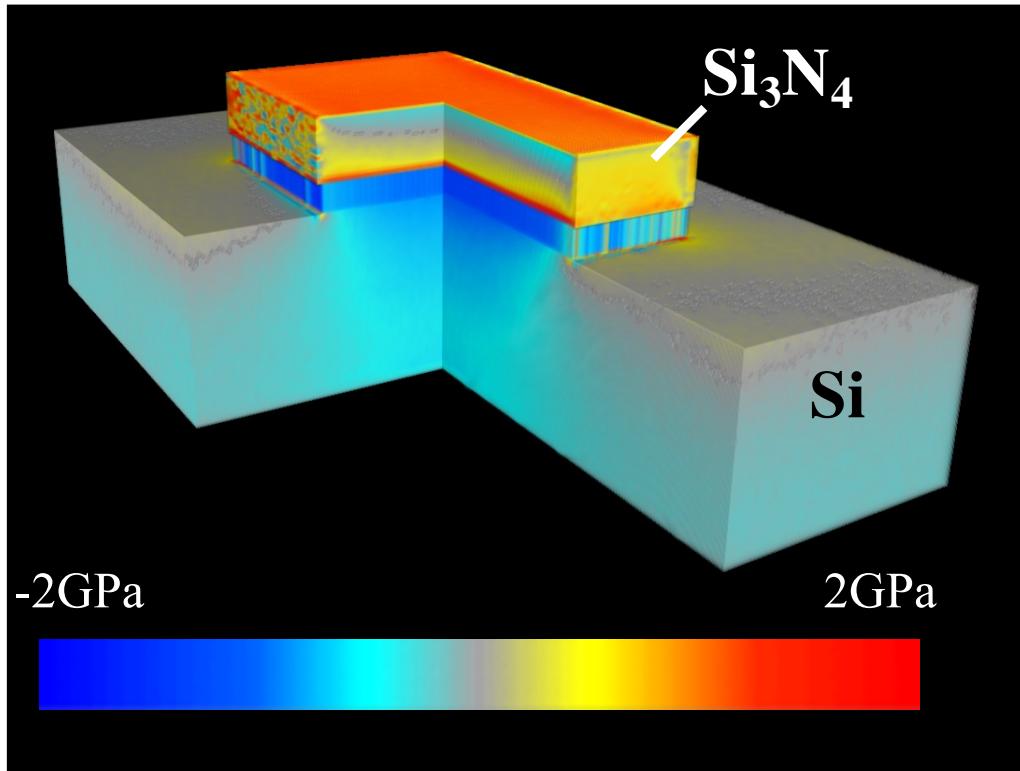
Shock-Induced Structural Phase Transformation in AlN



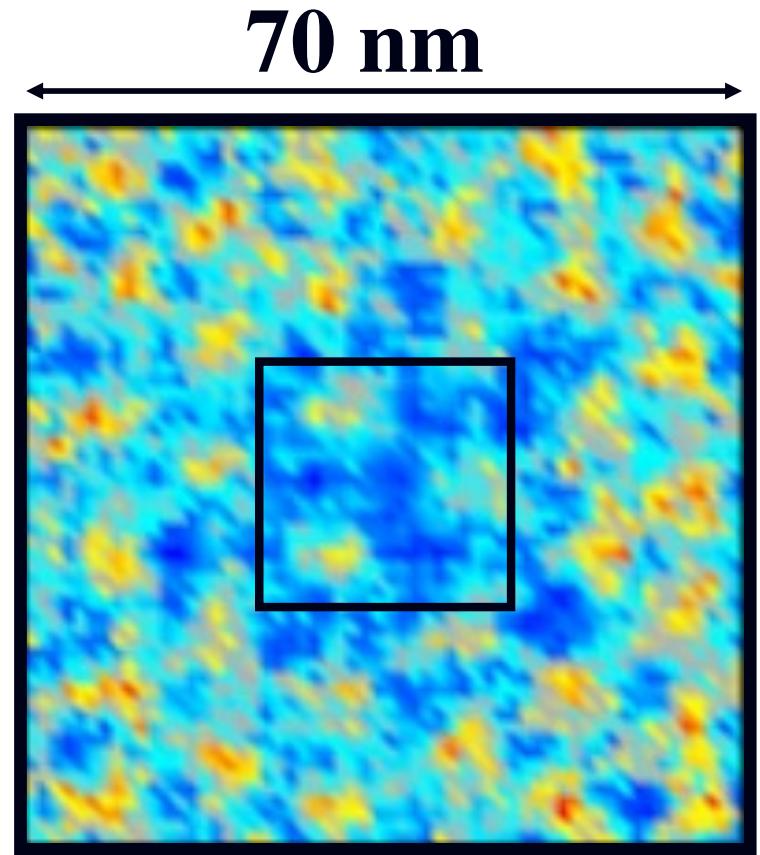
- Wurtzite (4-coordinated) to rocksalt (6-coordinated) phase transformation at 20 GPa



Stress Domains in Si_3N_4 /Si Nanopixels

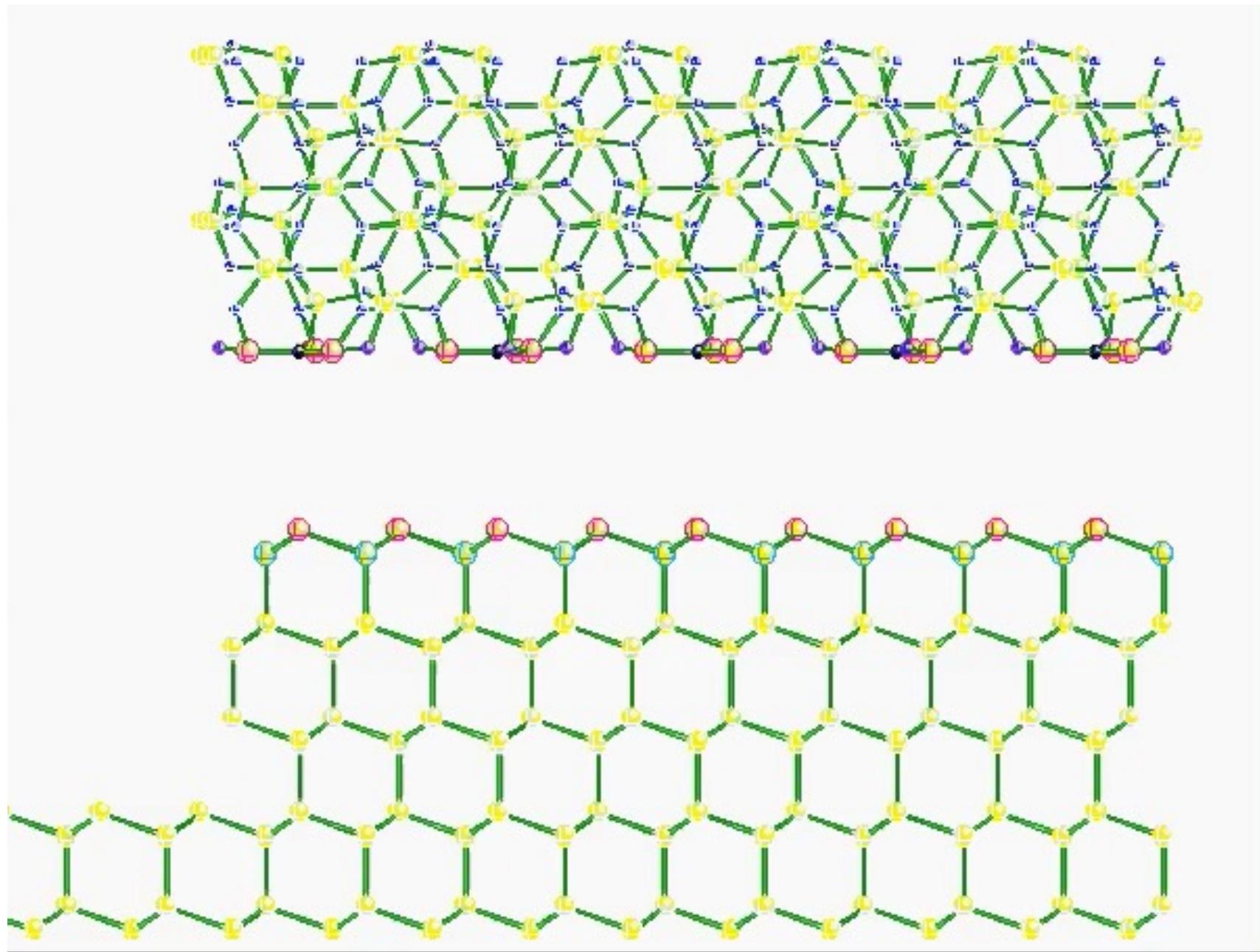


Stress well in Si with a
crystalline Si_3N_4 film
due to lattice mismatch

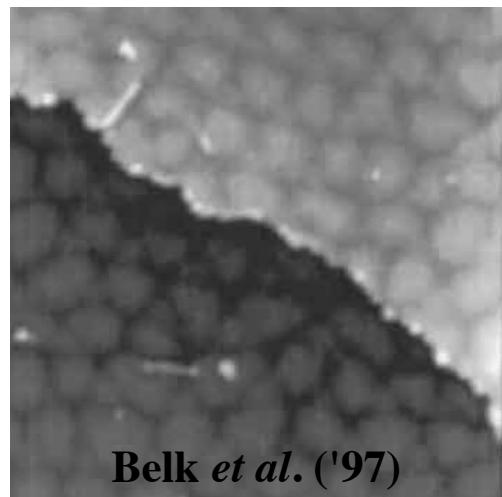
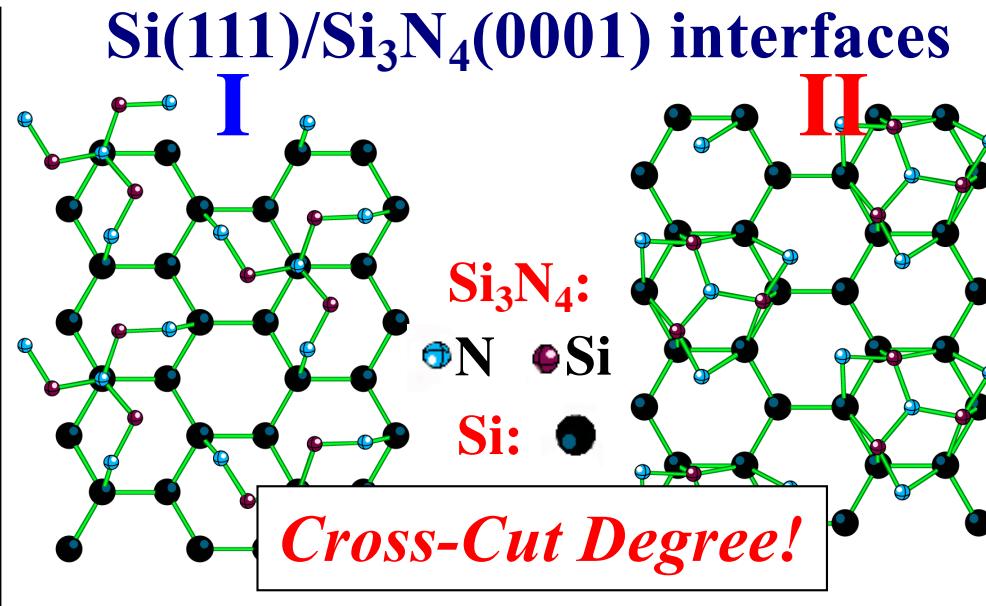
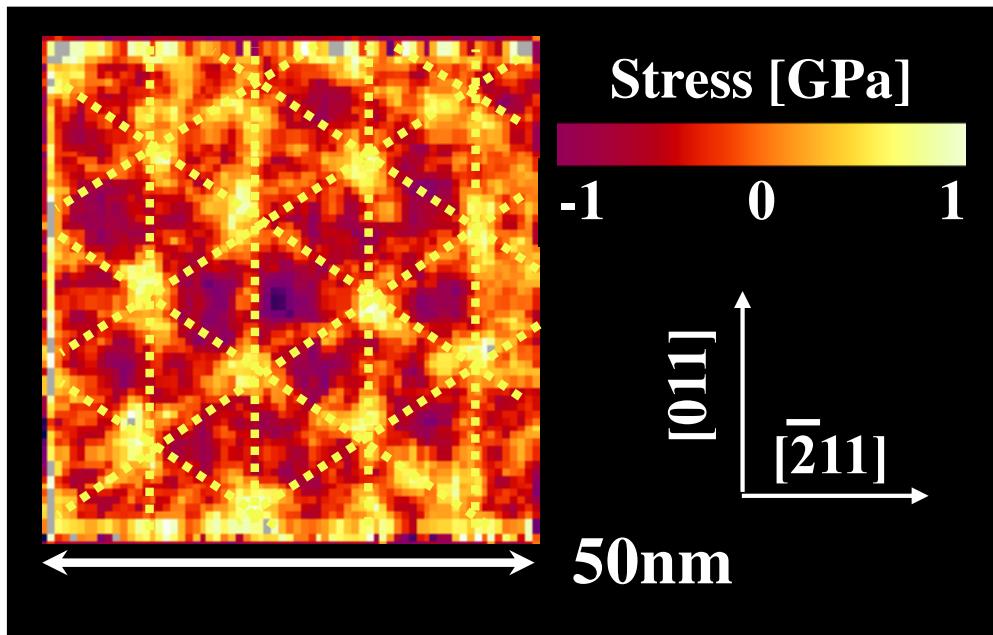


Stress domains in Si
due to an amorphous
 Si_3N_4 film

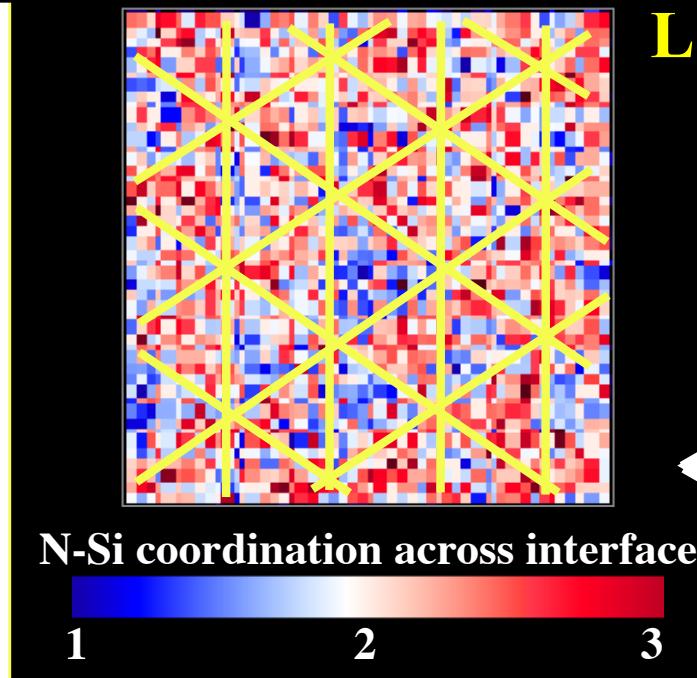
Si(111)/Si₃N₄(0001) Interface



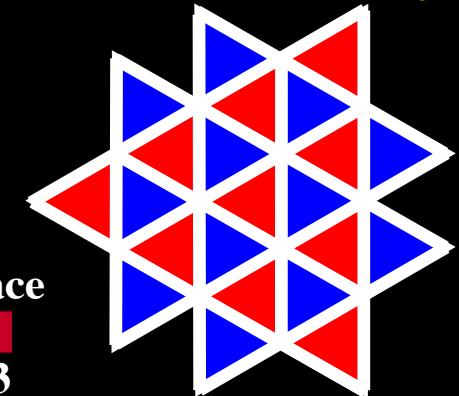
Stress Domains in Si/Si₃N₄ Nanopixel



Misfit dislocation network
in InAs/GaAs(111)

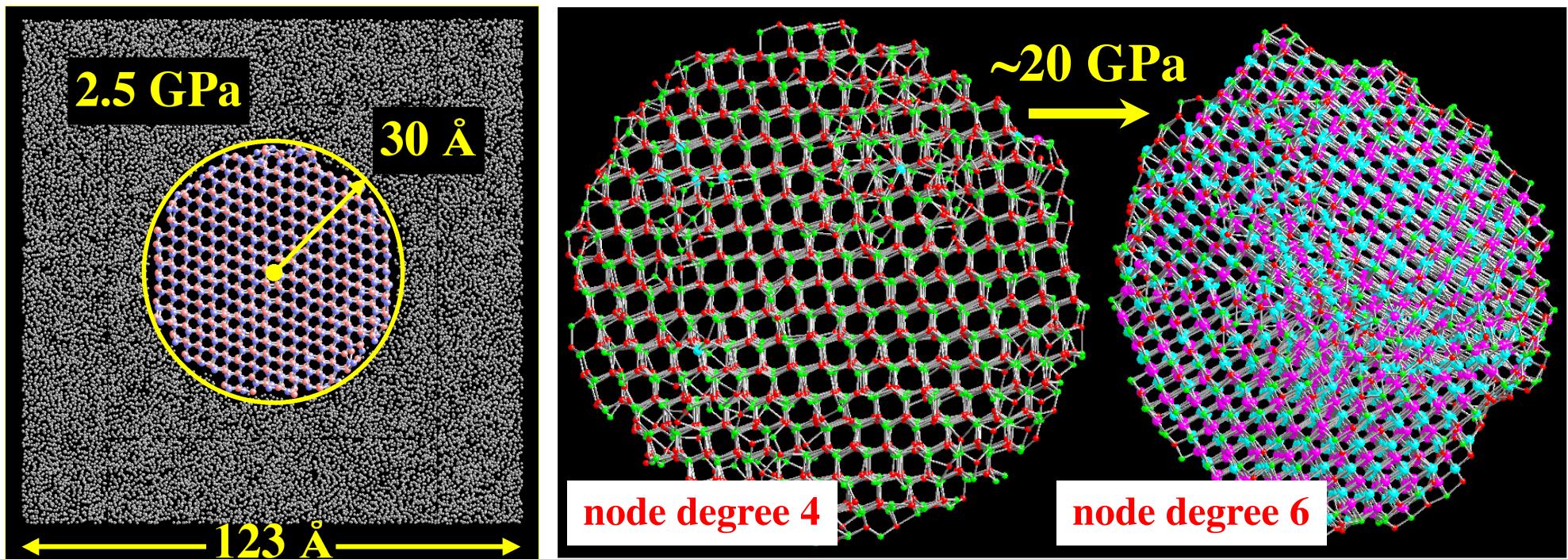


Lattice mismatch
(1%) induced
interfacial
domain array



High-Pressure Structural Transformation

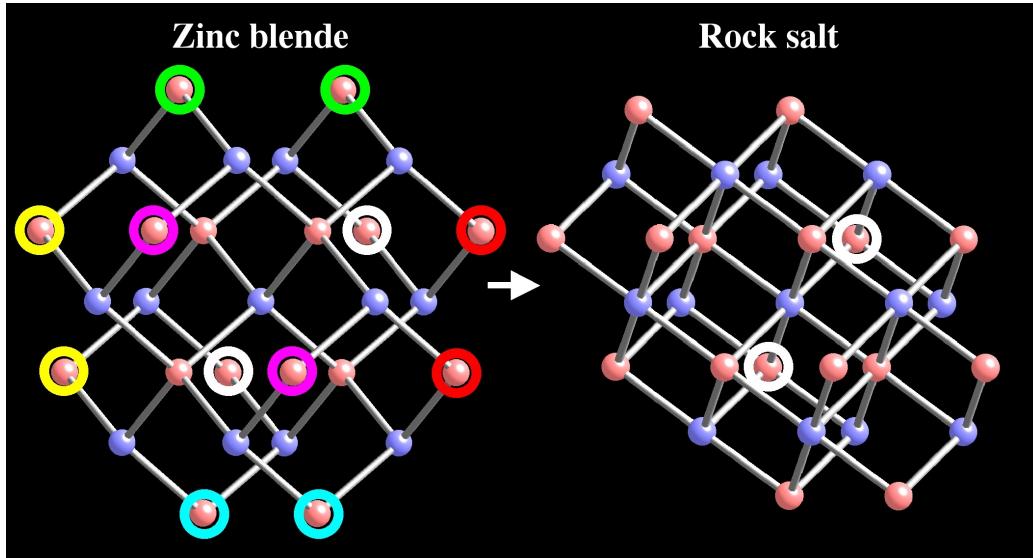
- Wurzite (node degree 4) to rocksalt (node degree 6) structural transformation of a GaAs nanoparticle under high pressure



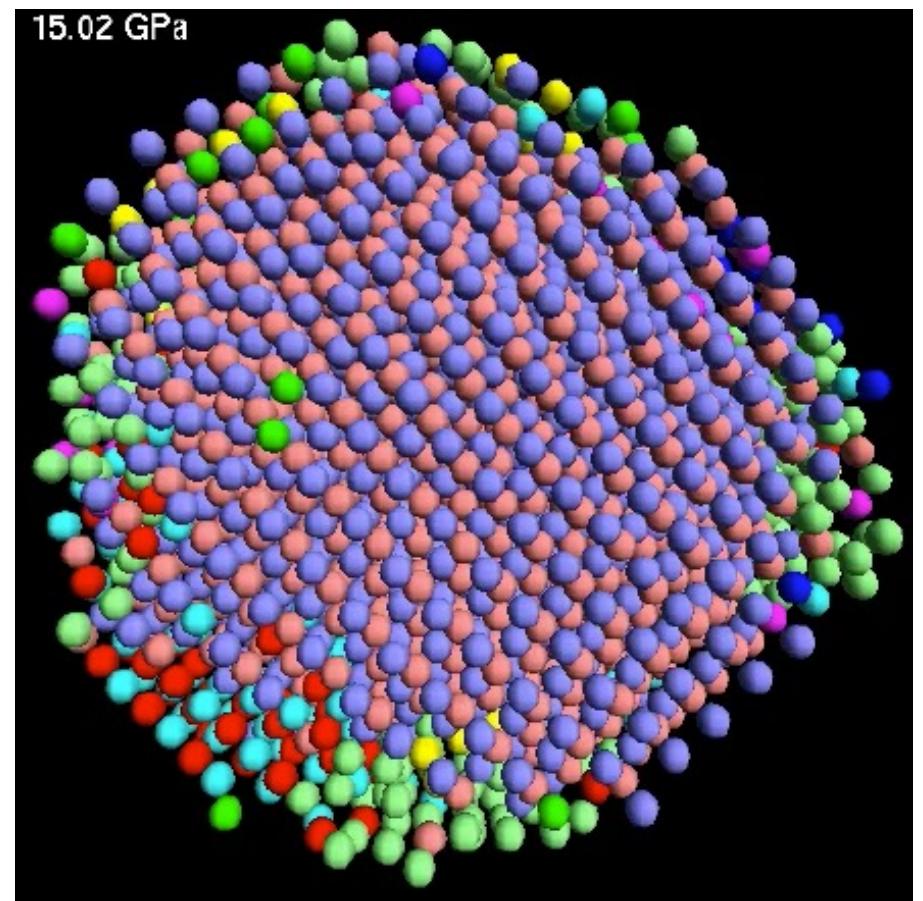
- Existence of multiple domains?

Graph-Transition Tracking

- Finite set of graph transitions as a classifier



$$\begin{array}{c} G = (V, E) \\ \downarrow \\ G' = (V, E') \\ E \subset E' \end{array}$$



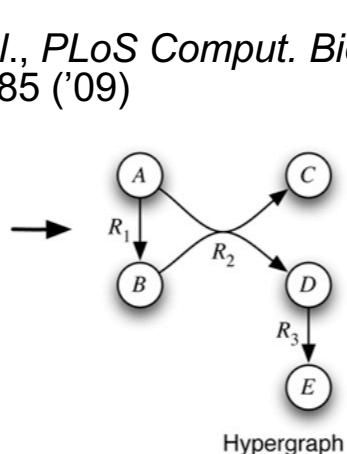
Graph Transition!

Chemical Reaction Network

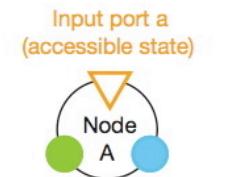
Klamt et al., PLoS Comput. Biol. 5, e1000385 ('09)

Reaction networks

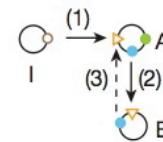
$$\begin{aligned} R_1 : A &\longrightarrow B \\ R_2 : A + B &\longrightarrow C + D \\ R_3 : D &\longrightarrow E \end{aligned}$$



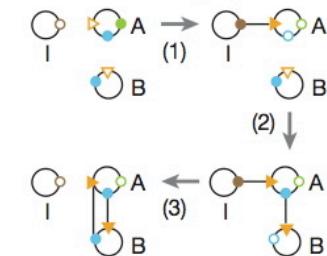
c Nodal abstraction



d Reaction graph

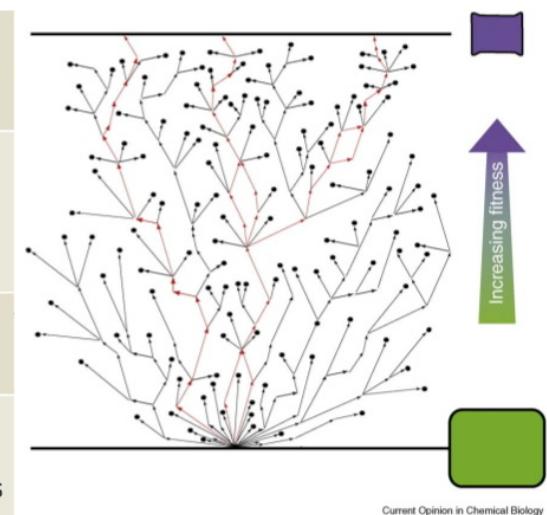
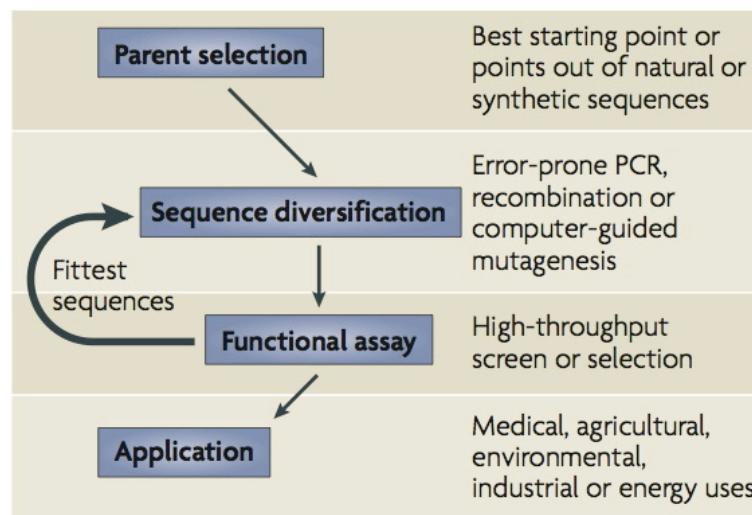
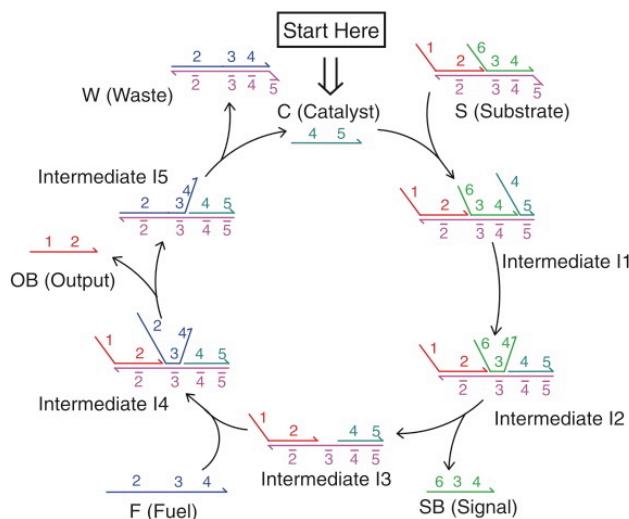


e Execution of reaction graph



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

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



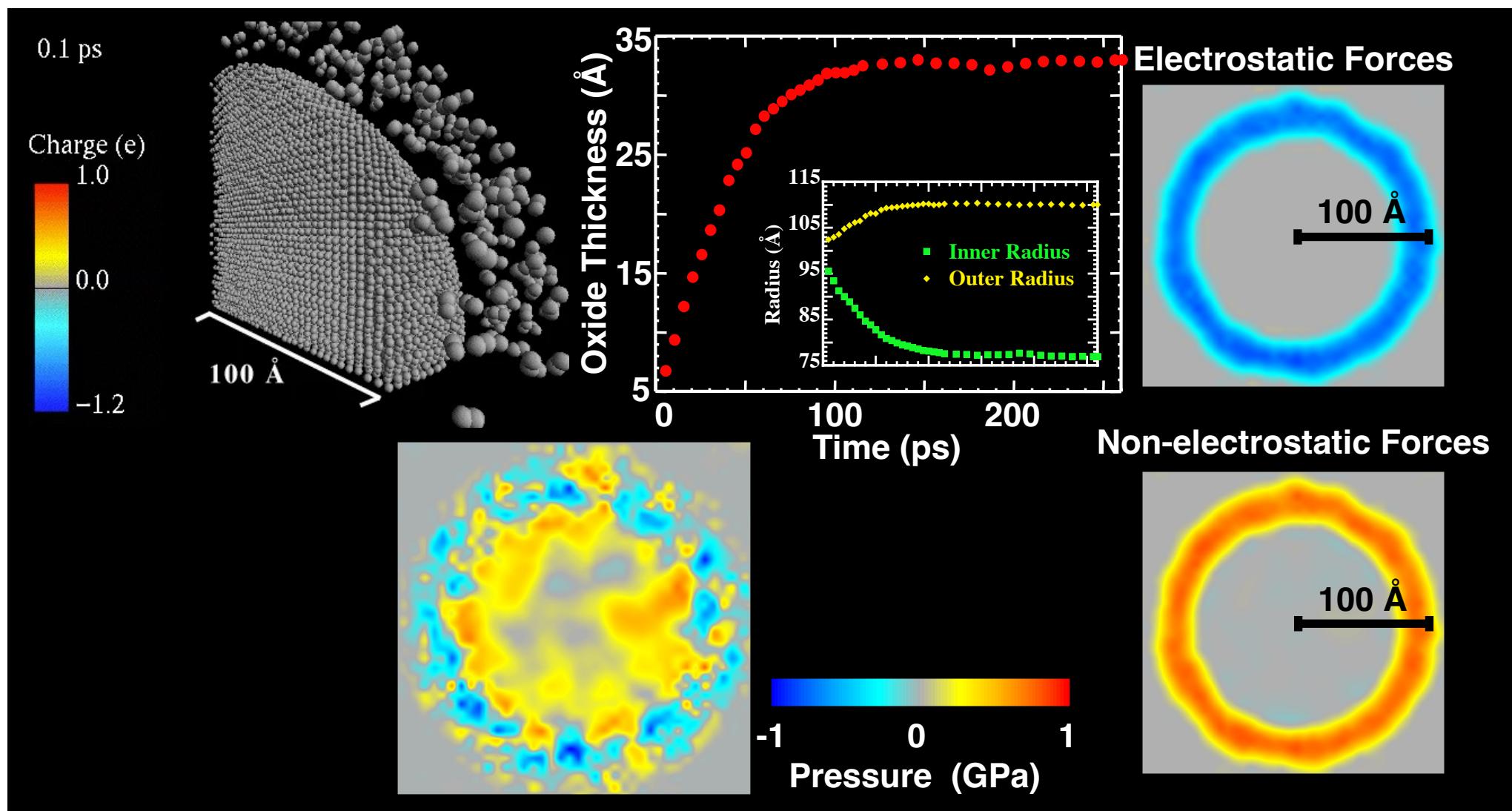
Current Opinion in Chemical Biology

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

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

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

Oxidation of an Al Nanoparticle (n-Al)

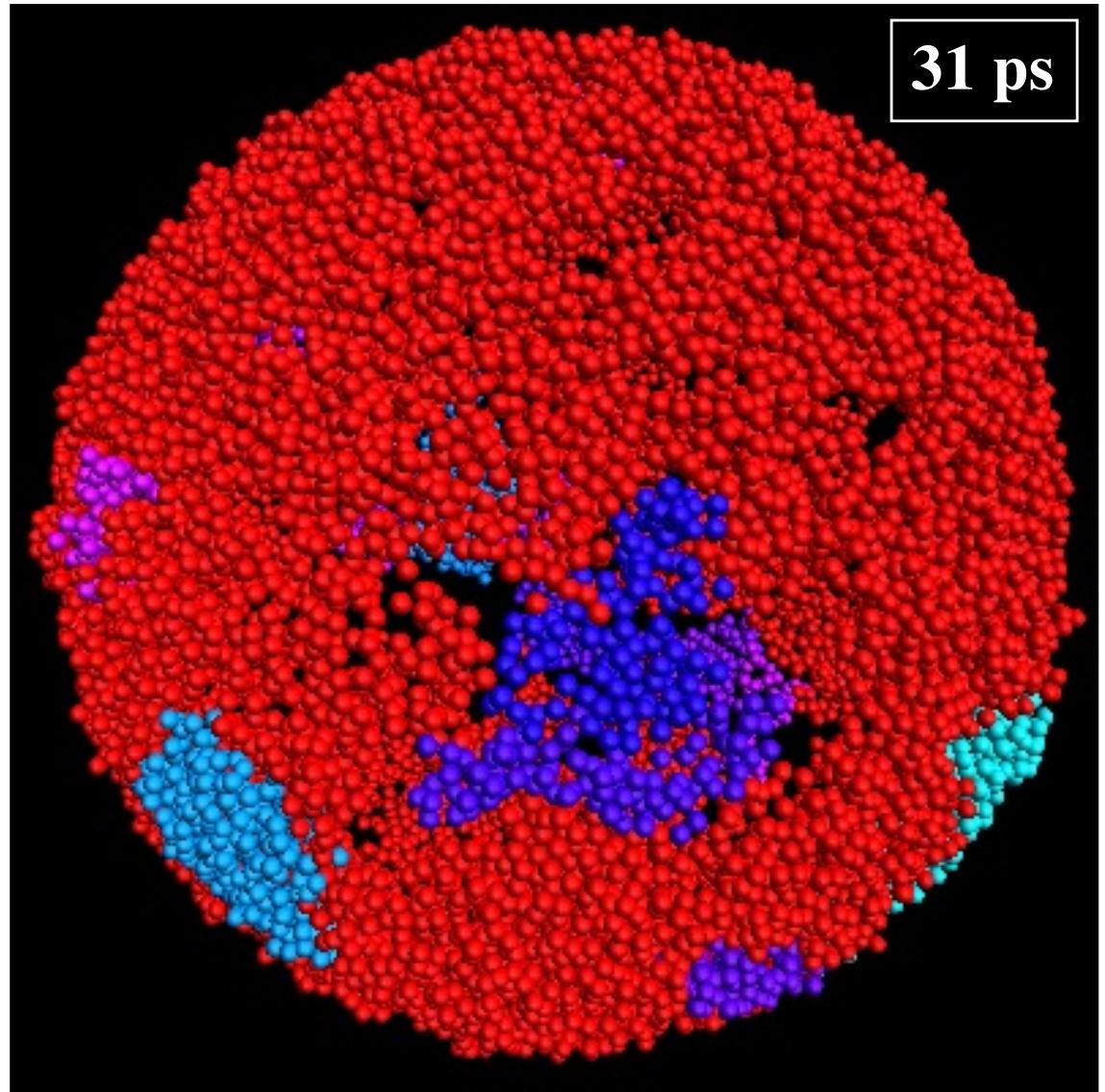


- Oxide thickness saturates at 40 Å after 0.5 ns, in agreement with experiments
- Oxide region/metal core is under negative/positive pressure
- Attractive Al-O Coulomb forces contribute large negative pressure in the oxide

Oxidative Percolation

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

*Percorative
Connected Components!*

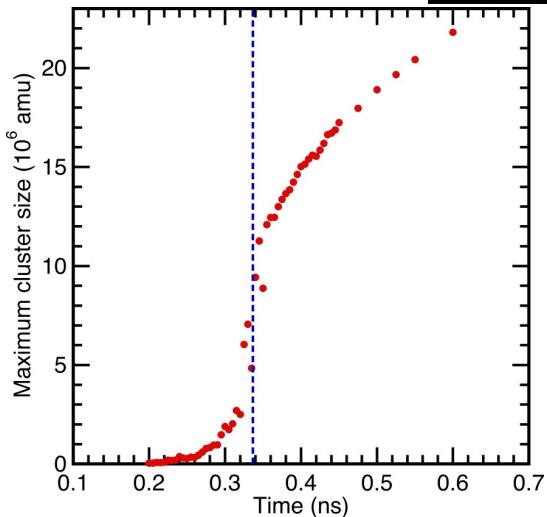
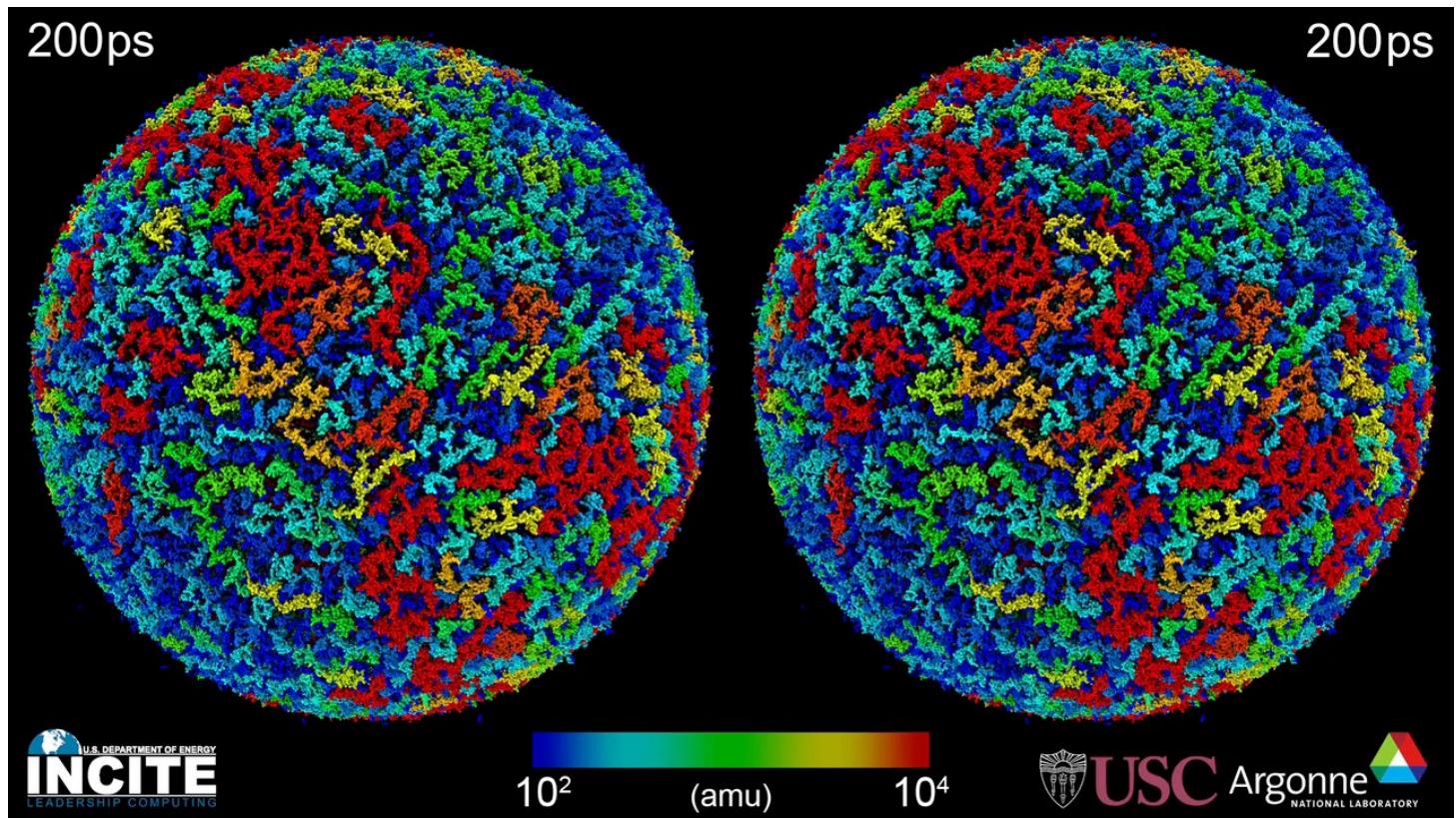


Size of Network

10^2 10^3 10^4

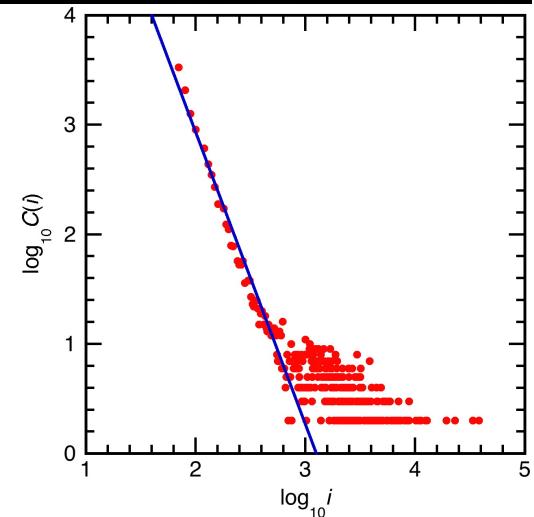
Fractal Nanocarbon Product

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



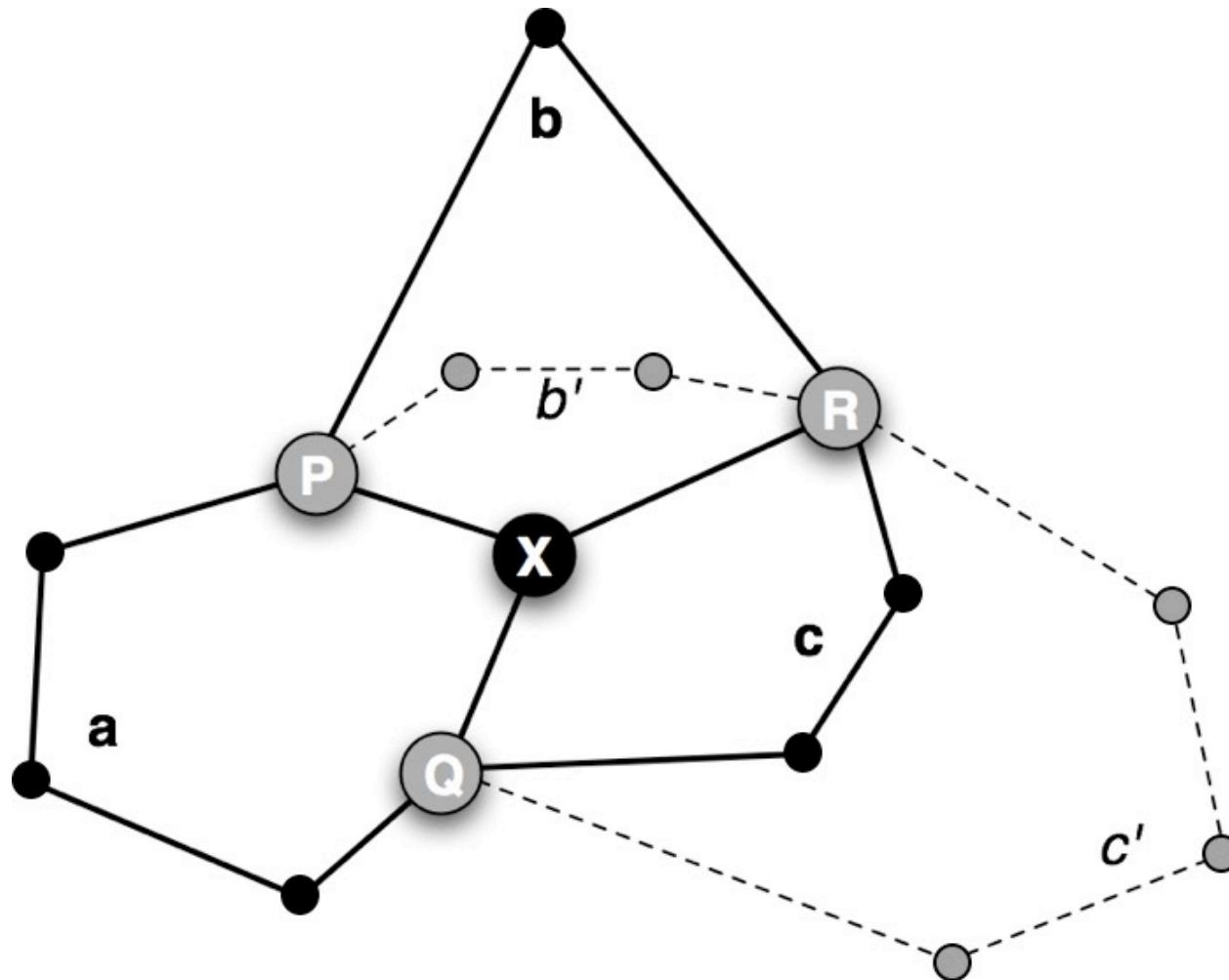
- Fractal nanocarbon product with large surface areas may find supercapacitor, battery-electrode & mechanical metamaterial applications: $d_f = d/(\tau - 1) \sim 1.85$

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



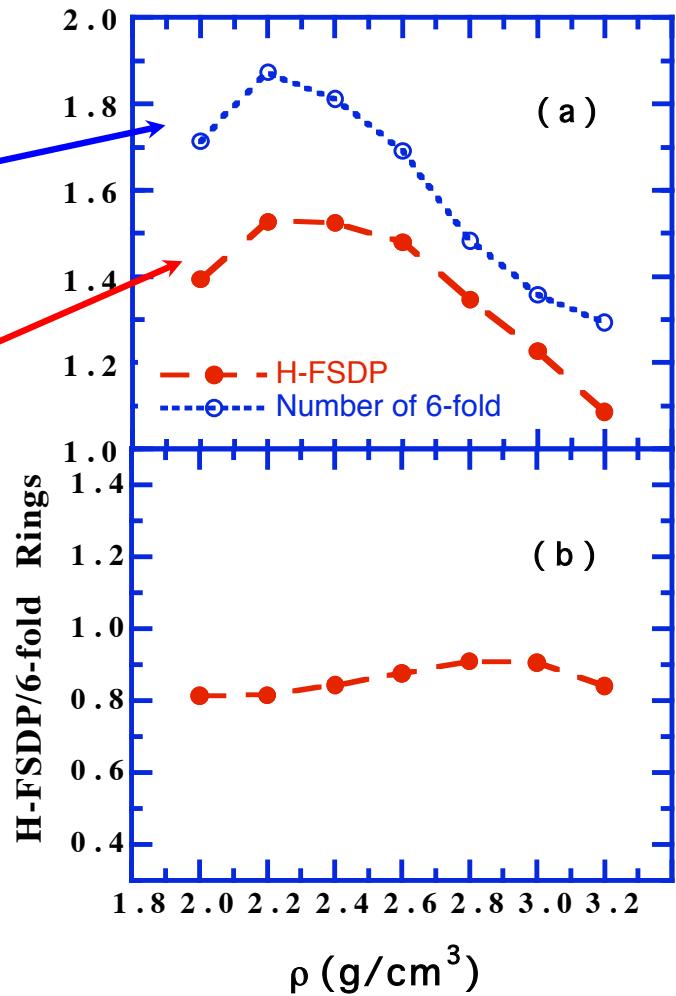
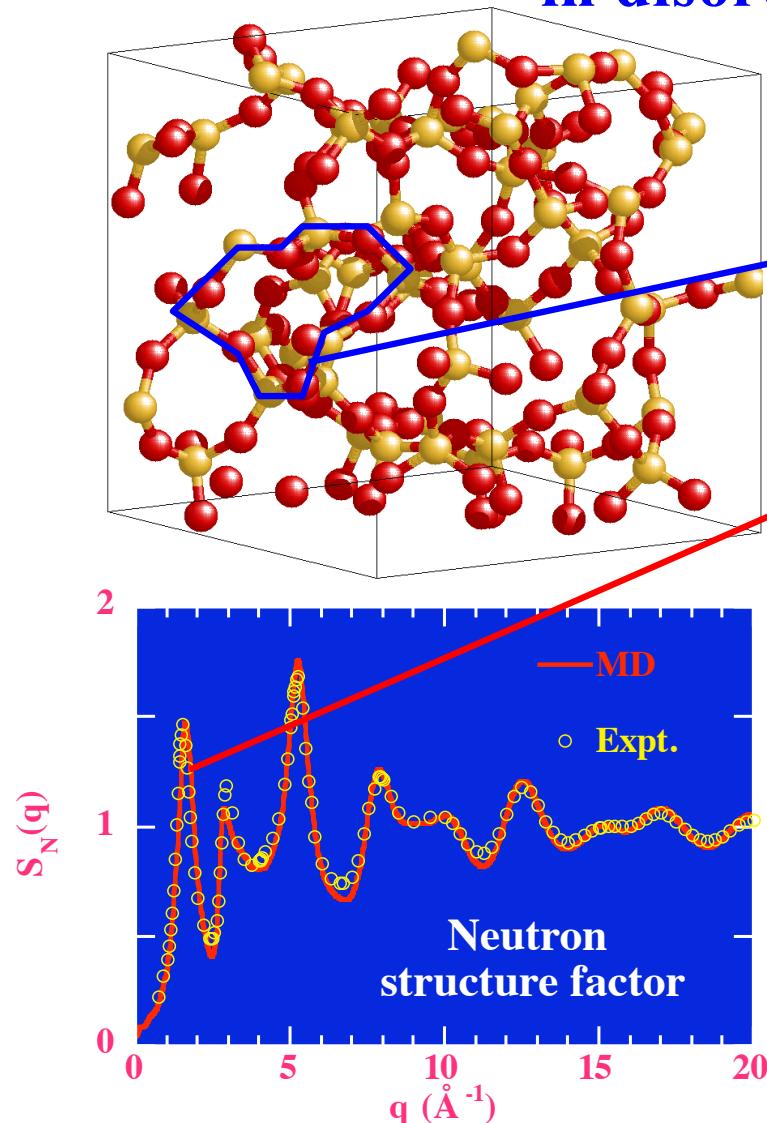
Shortest-Path Rings

- **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 $G-x$



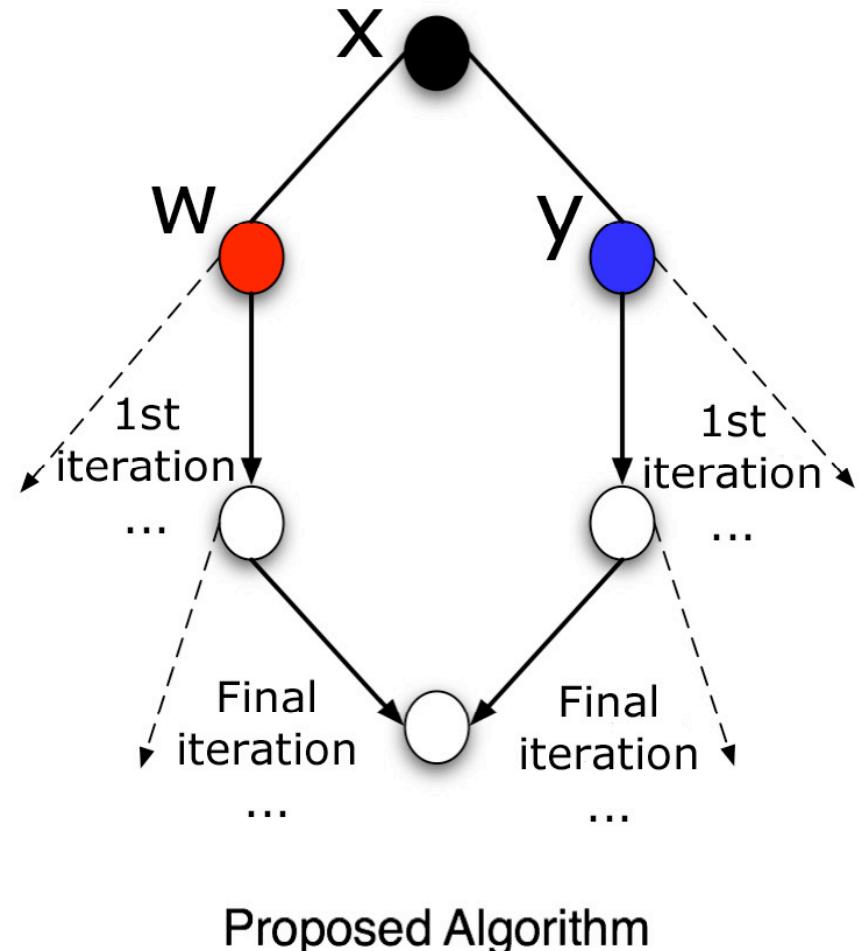
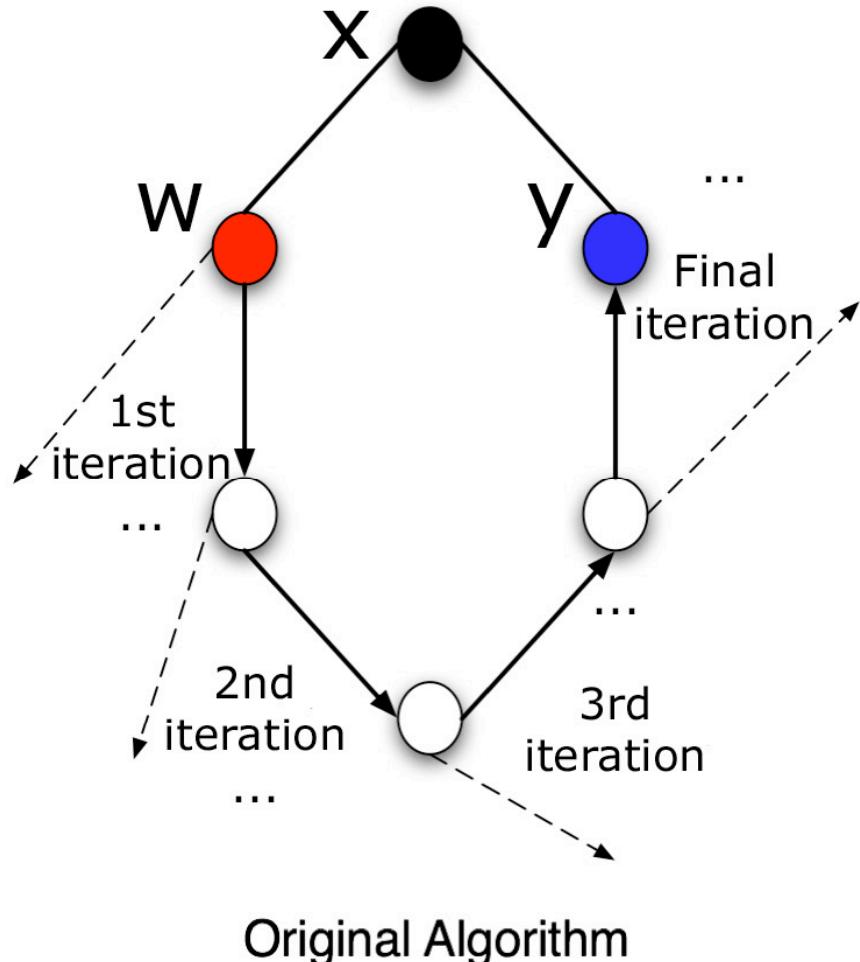
Ring-based Data Mining

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



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:

$\text{Neighbors}(V)$ = 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_o using R_{bc} as cutoff distance
- 2 for every vertex $V \in V_o$
 - for each vertex pair V_i and V_j in $\text{Neighbors}(V)$ do
 - $A_1 = \{V_i\}$
 - $A_2 = \{V_j\}$
 - $L_b = 0$
 - while $(A_1 \cap A_2 = \emptyset \text{ AND } L_b < L_{MAX})$ do
 - $L_b = L_b + 2$
 - if $(A_1 \cap \text{Neighbors}(A_2) \neq \emptyset \text{ OR } A_2 \cap \text{Neighbors}(A_1) \neq \emptyset)$
 - $L_b = L_b + 1$
 - break
 - else if $(\text{Neighbors}(A_1) \cap \text{Neighbors}(A_2) \neq \emptyset)$
 - $L_b = L_b + 2$
 - $A_1 = \text{Neighbors}(A_1)$
 - $A_2 = \text{Neighbors}(A_2)$
 - if $(L_b < L_{MAX}) \quad ++ K_v(L_b)$

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

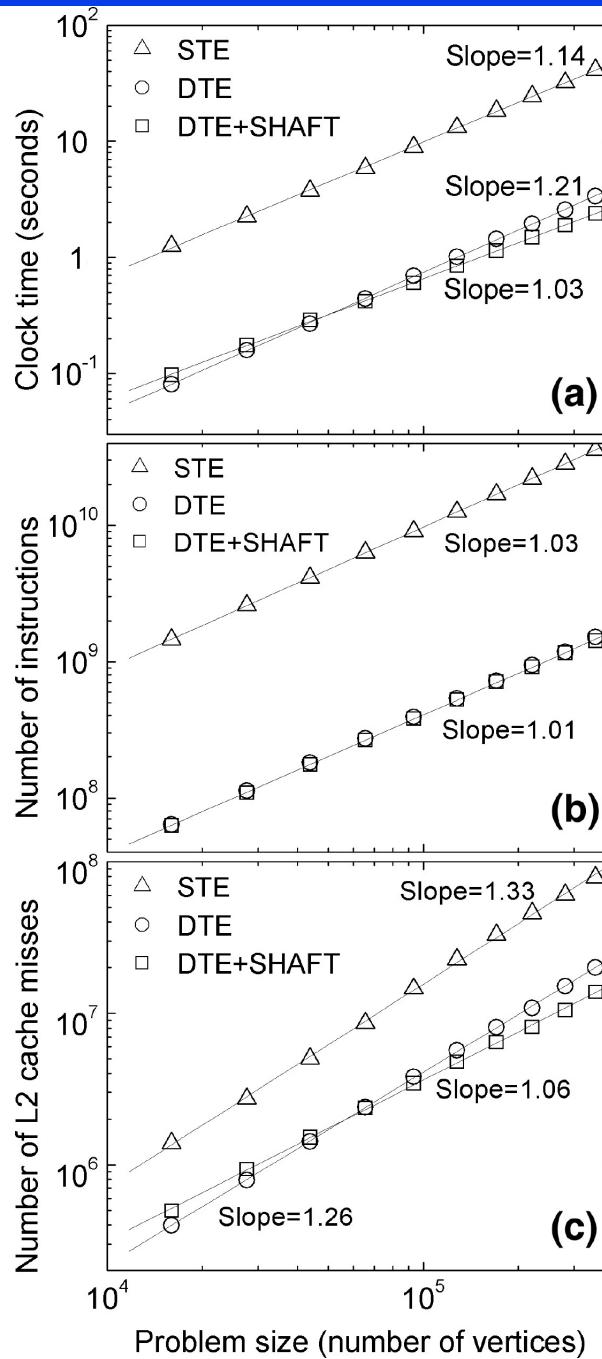
$$q_i = \lfloor C_i/b \rfloor$$

$$q_i \% = m$$

$$\text{return } q = q_3 \times m^2 + q_2 \times m + q_1$$

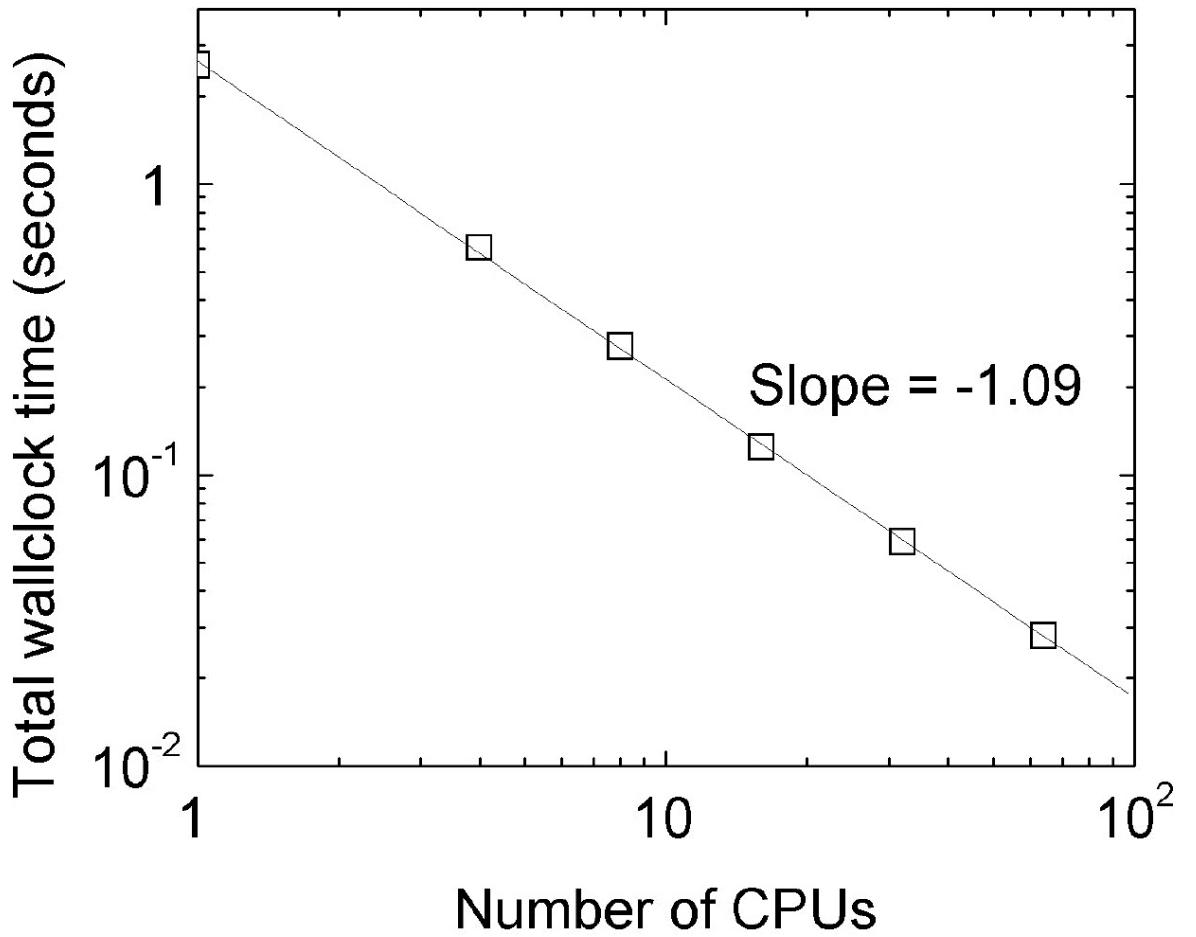
0	1	2	3	4	0	1	2	3	4
5	6	7	8	9	5	6	7	8	9
10	11	12	13	14	10	11	12	13	14
15	16	17	18	19	15	16	17	18	19
20	21	22	23	24	20	21	22	23	24
0	1	2	3	4	0	1	2	3	4
5	6	7	8	9	5	6	7	8	9
10	11	12	13	14	10	11	12	13	14
15	16	17	18	19	15	16	17	18	19
20	21	22	23	24	20	21	22	23	24

Numerical Tests



Linear scaling
on the problem size

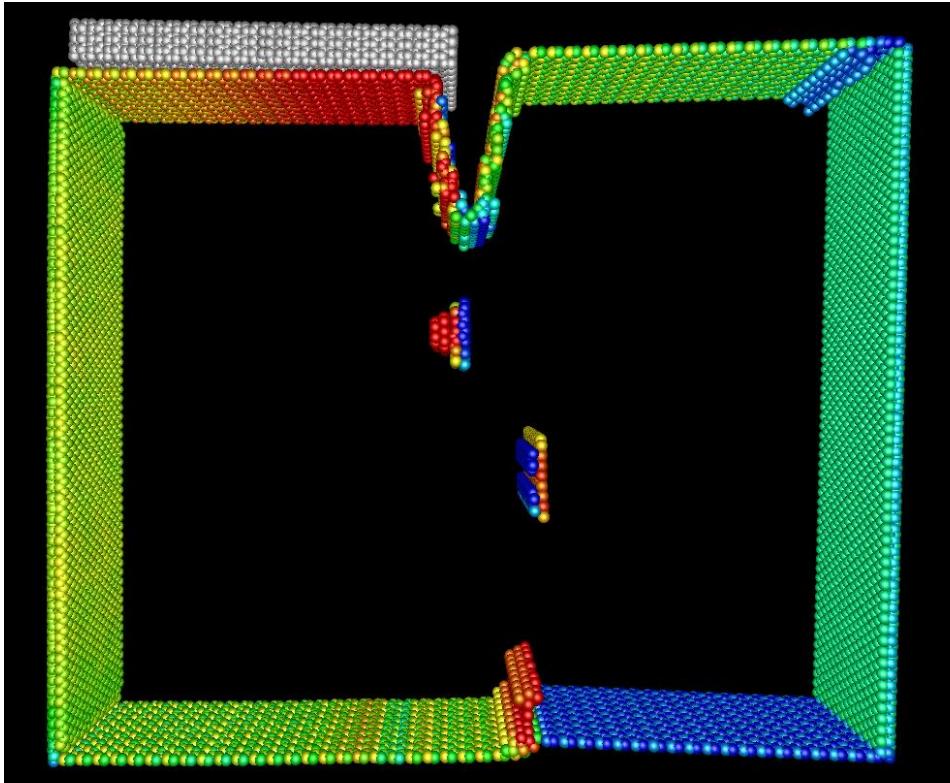
Superlinear (strong) scaling
on the number of CPUs



Dislocation Mining

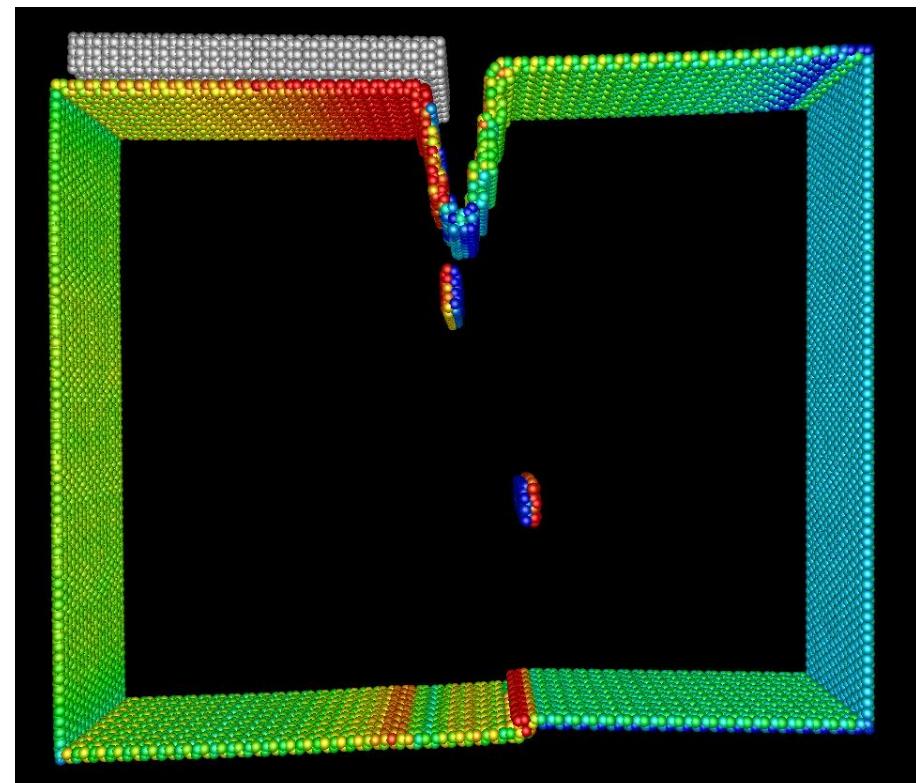
Based on
potential energy

Shown atoms with high energy compared to bulk



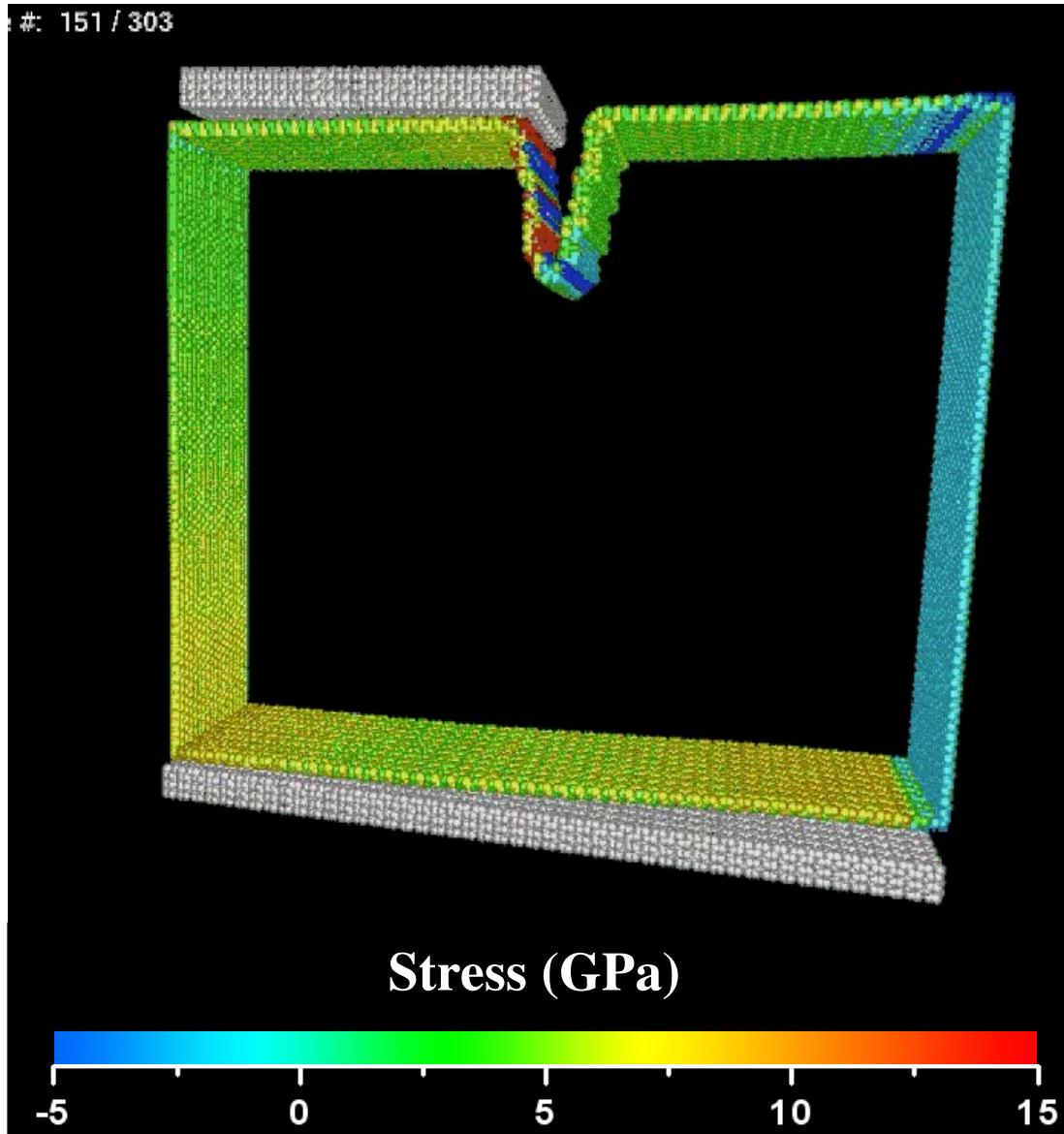
Based on
shortest-path ring statistics

Shown atoms with less than 12 6-membered rings



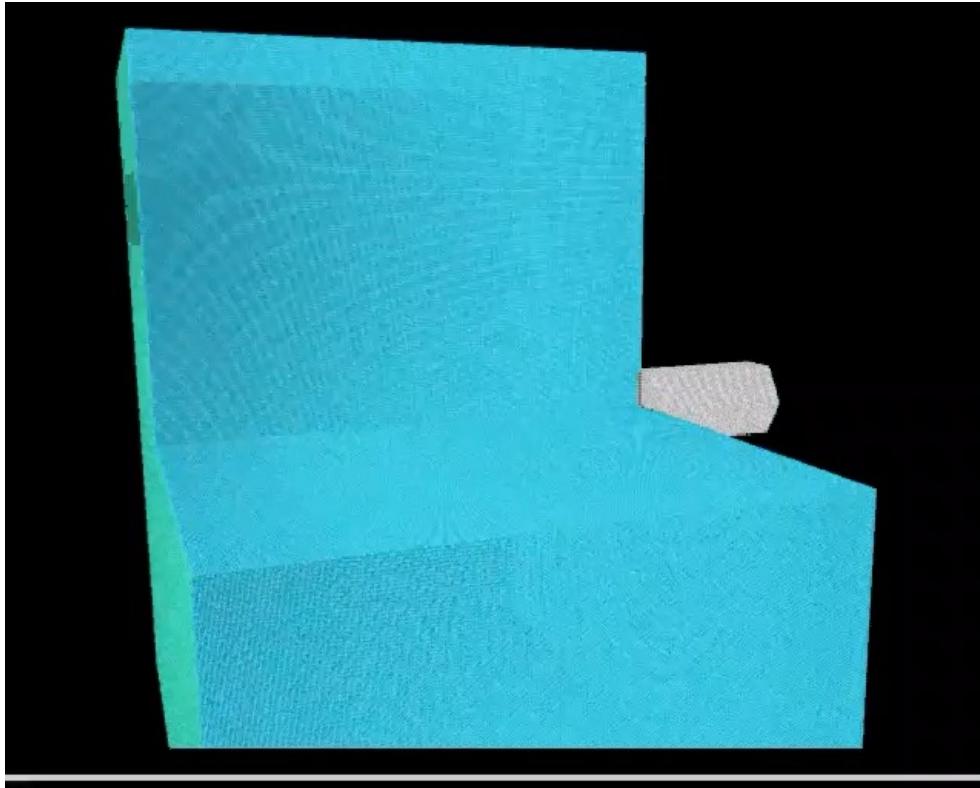
100 km/s Impact on Notched AlN

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

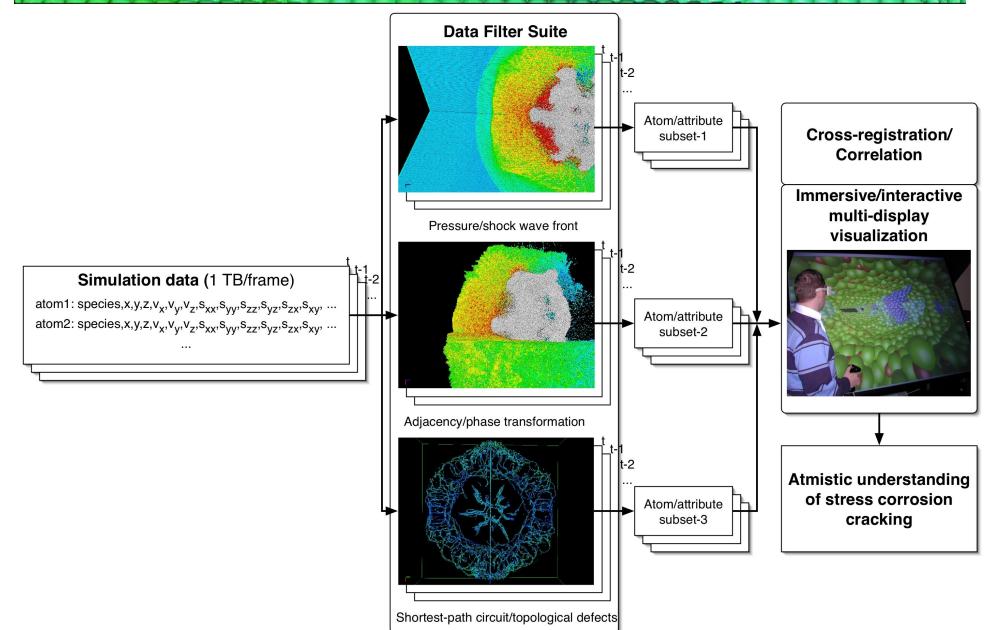
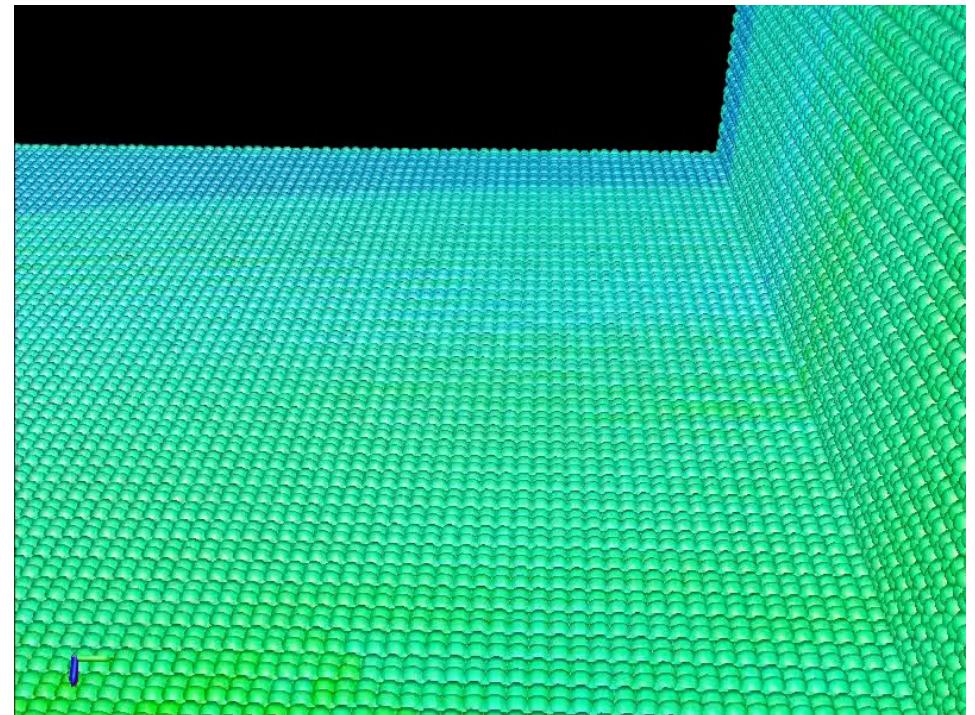


Impact-Damage Tolerant Ceramics?

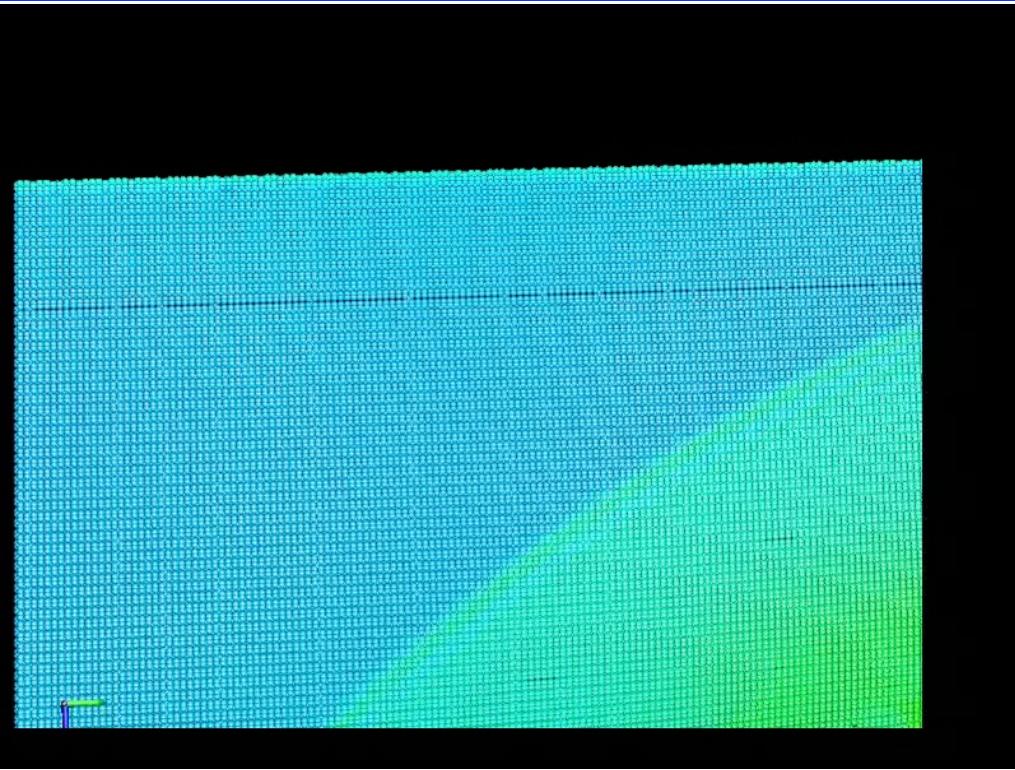
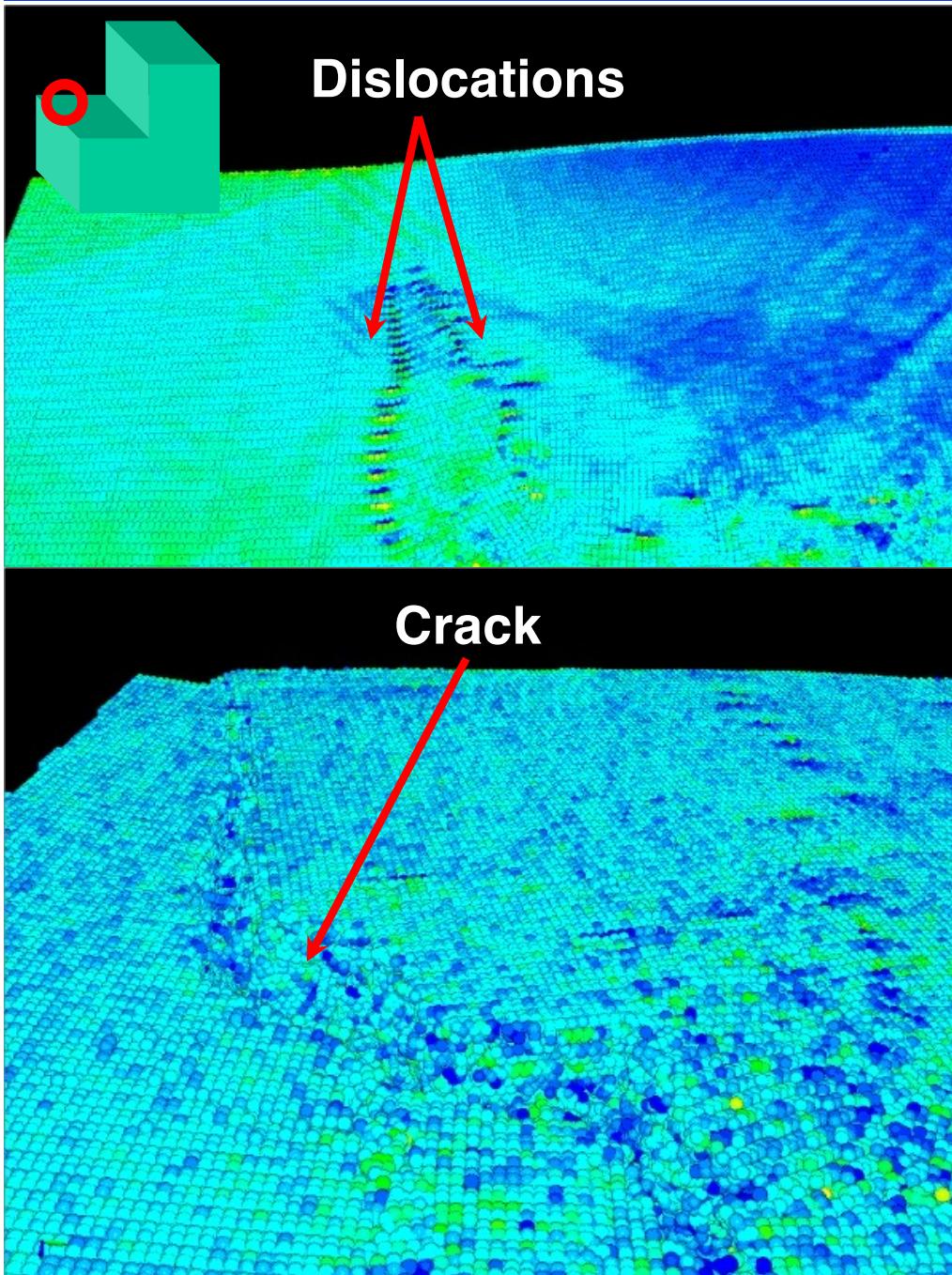
Inverse problem: design materials with desired properties



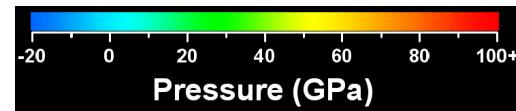
209 million atom MD of hypervelocity impact in AlN for the design of light-weight ceramic armors



Crack Nucleation at Kink Bands

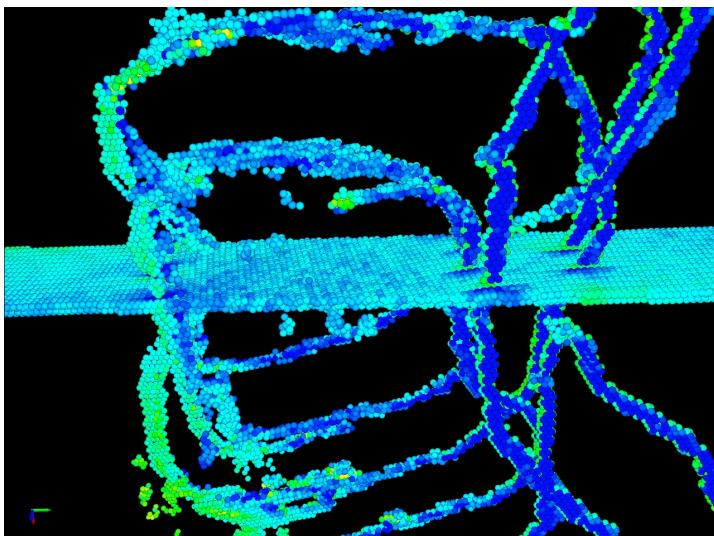
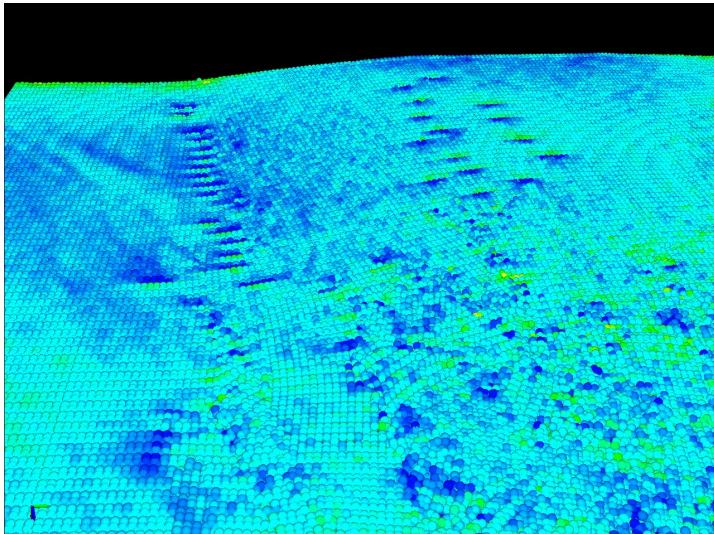


- Series of dislocation dipoles with opposite Burgers vectors form a kink band to releases stress
- Tilt grain boundaries of the kink bands act as sources of mode-II (shear) crack nucleation

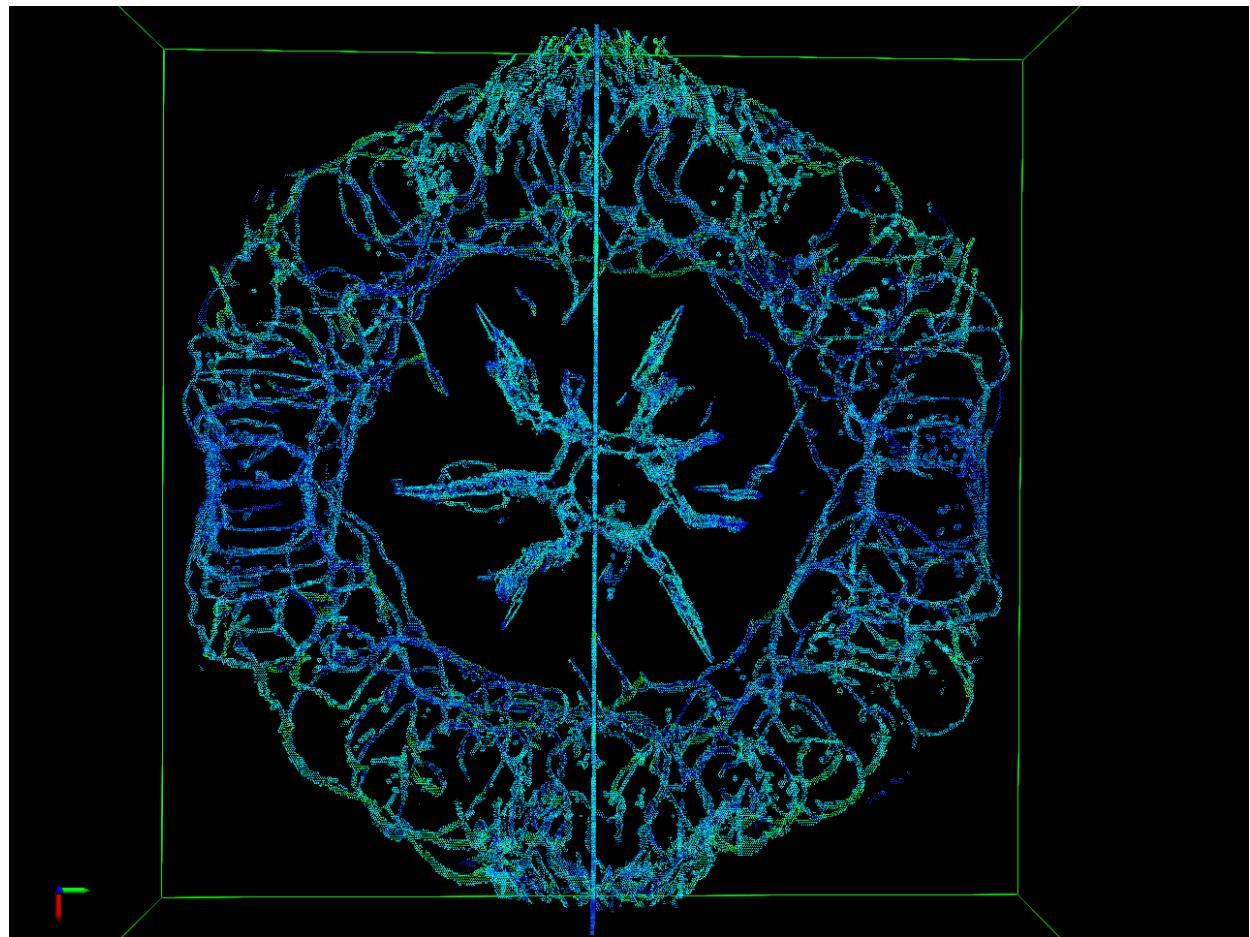


Dislocation Loops at Kink Bands

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



Atoms participating in
non-6-member circuits



Dislocation network

Nanoindentation on Nanophase SiC

Superhardness

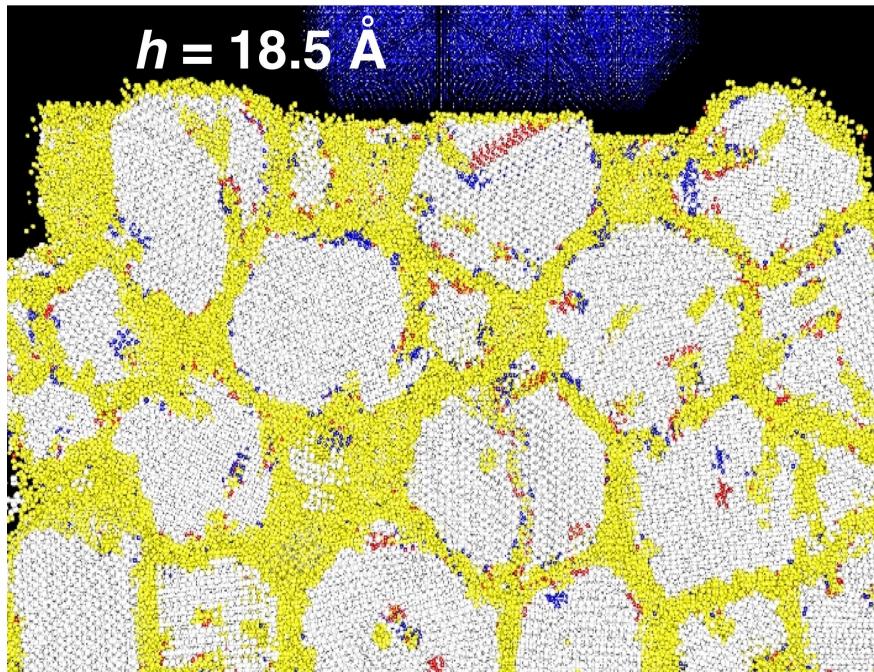
MD: 39 GPa

(grain size, $d = 8 \text{ nm}$)

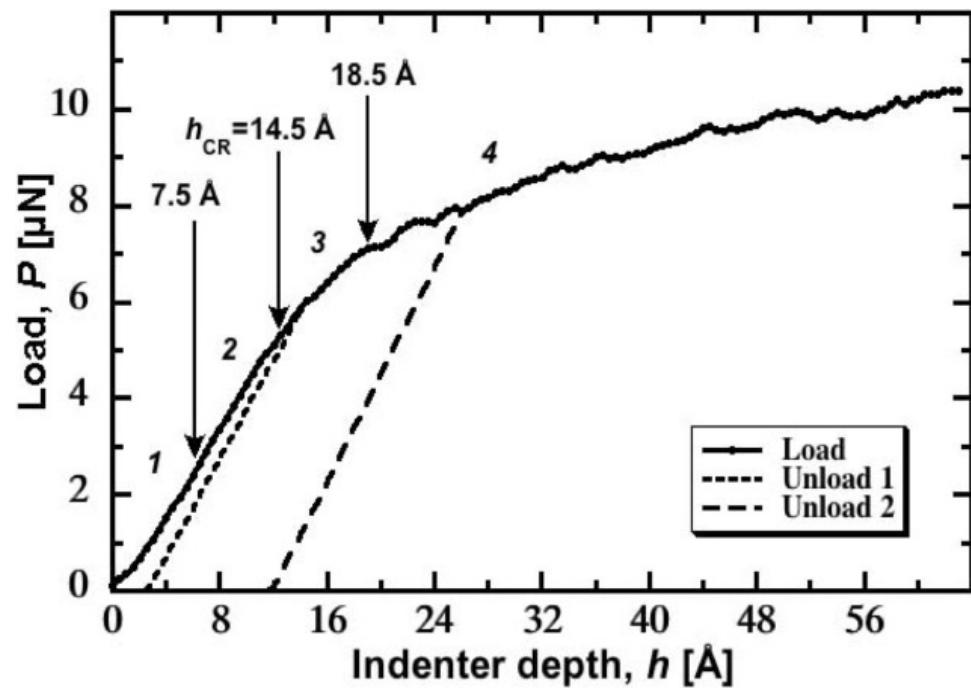
Expt.: 30-50 GPa

($d = 5\text{-}20 \text{ nm}$)

[Liao *et al.*, APL, '05]



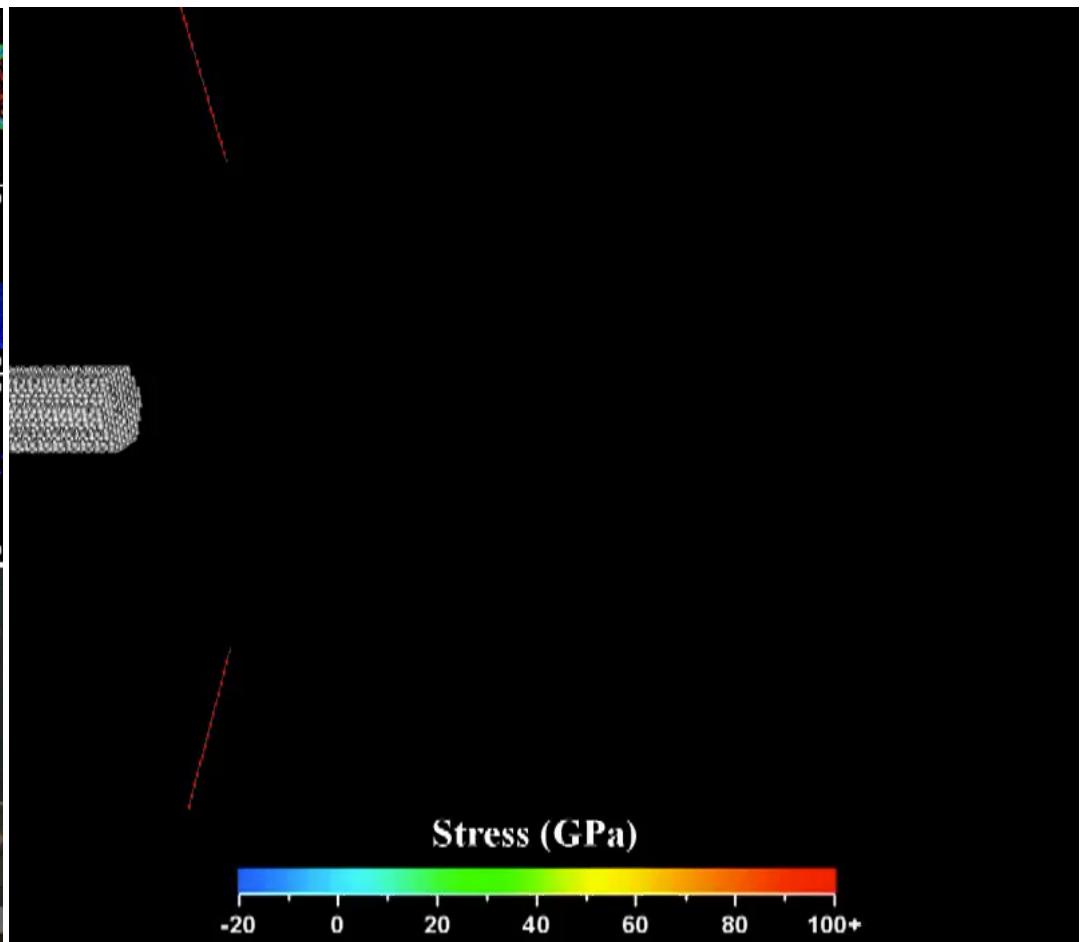
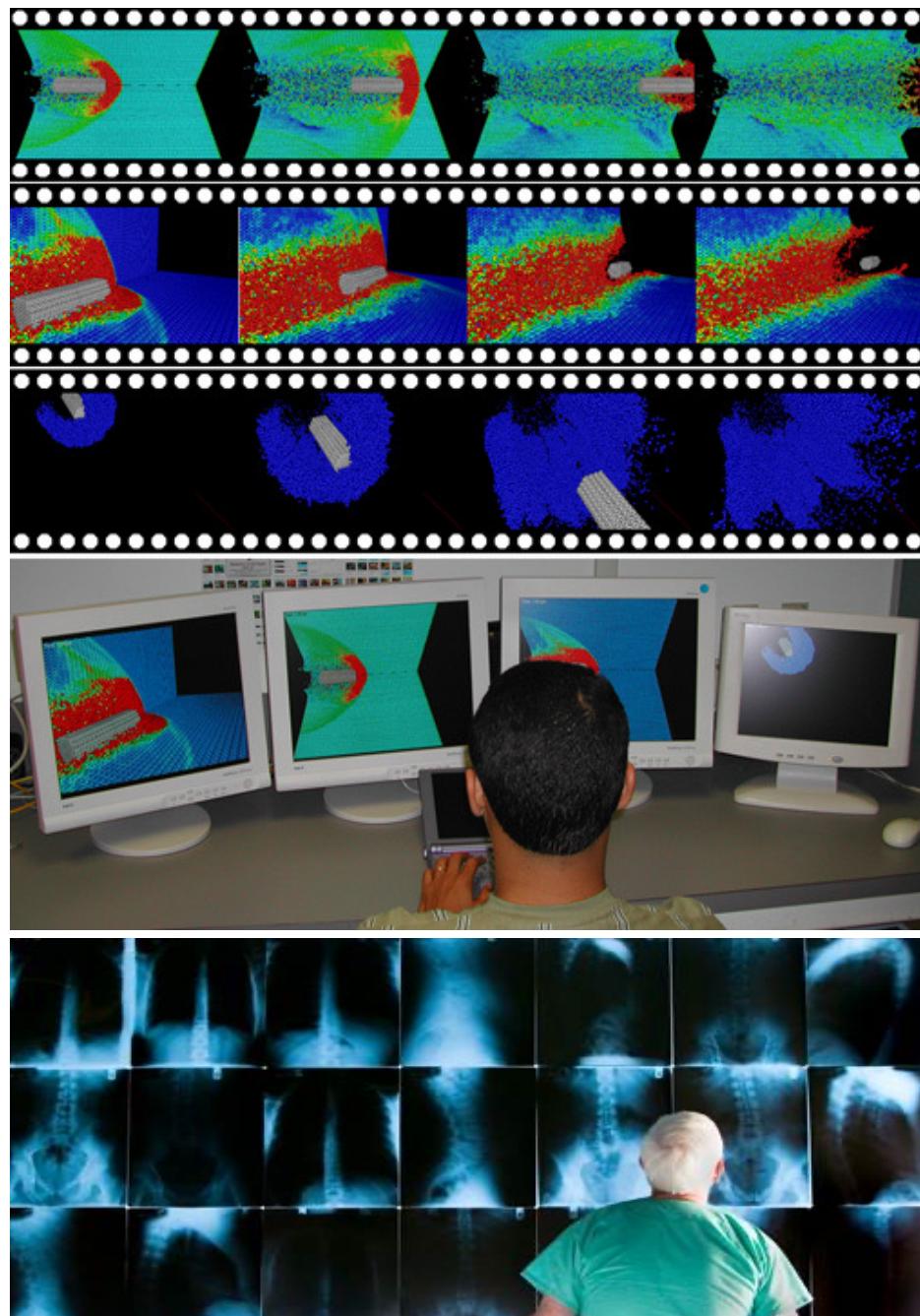
Load-displacement curve



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

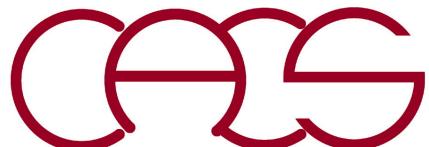
Singular Value Decomposition & Data Mining

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 \cong data compression



Rank of a Matrix

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

$$M \begin{bmatrix} 1 \\ x \end{bmatrix} \quad x \in R^M \xrightarrow{A} b = Ax \in R^N \begin{bmatrix} 1 \\ b \end{bmatrix} \quad N$$

- **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

$$b = A^\forall x = \sum_{v=1}^m c_v |v\rangle$$

Low Rank Approximations of a Matrix

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

$$N \begin{bmatrix} M \\ \psi \end{bmatrix} \cong \begin{bmatrix} u \\ v \end{bmatrix} |u\rangle\langle v| \forall x \rangle \propto |u\rangle$$

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

$$\begin{bmatrix} \psi \end{bmatrix} \cong \begin{bmatrix} u_1 \\ w_1 \end{bmatrix} \begin{bmatrix} v_1 \end{bmatrix} + \begin{bmatrix} u_2 \\ w_2 \end{bmatrix} \begin{bmatrix} v_2 \end{bmatrix}$$

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

$$\begin{bmatrix} \psi \end{bmatrix} \cong \sum_{v=1}^m \begin{bmatrix} u_v \\ w_v \end{bmatrix} \begin{bmatrix} v_v \end{bmatrix}$$

Singular Value Decomposition

- **Problem:** Optimal approximation of an $N \times M$ matrix ψ of rank- m ($m \ll N$)?
- **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 $U \in R^{N \times M}$ & $V \in R^{M \times M}$ are column orthogonal & D is diagonal

$$N \begin{bmatrix} \psi \end{bmatrix} = M \begin{bmatrix} U \end{bmatrix} \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_M \end{bmatrix} \begin{bmatrix} V^T \end{bmatrix}_{M \times M} \quad U^T U = V^T V = I_M$$

See [appendix on polar & singular decompositions](#)

- **Theorem:** Sort the SVD diagonal elements in descending order, $d_1 \geq d_2 \geq \dots \geq d_M \geq 0$, & retain the first m terms $\psi^{(m)} \equiv \sum_{\nu=1}^m u_i^{(\nu)} d_\nu v_j^{(\nu)T}$

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

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

cf. [singular.c](#) & [svdcmp.c](#)

Use the program!

SVD for Image Compression



Original Image



5 Iterations



10 Iterations

D. Richards & A. Abrahamsen



20 Iterations

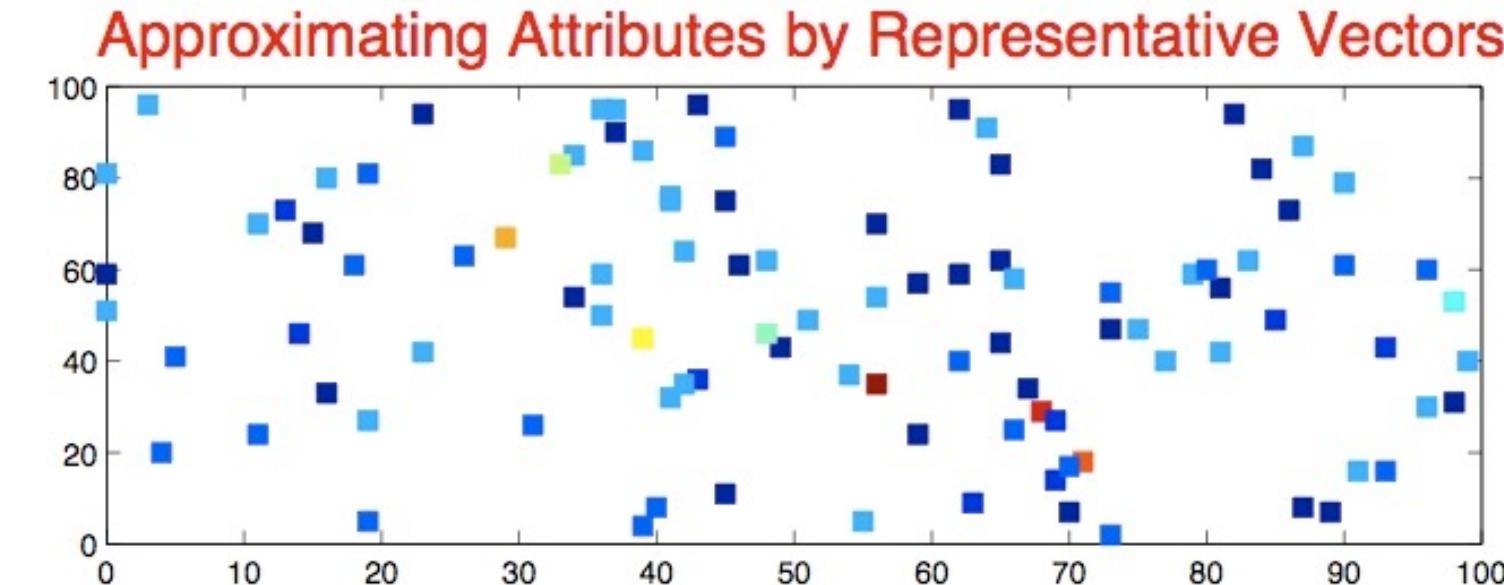
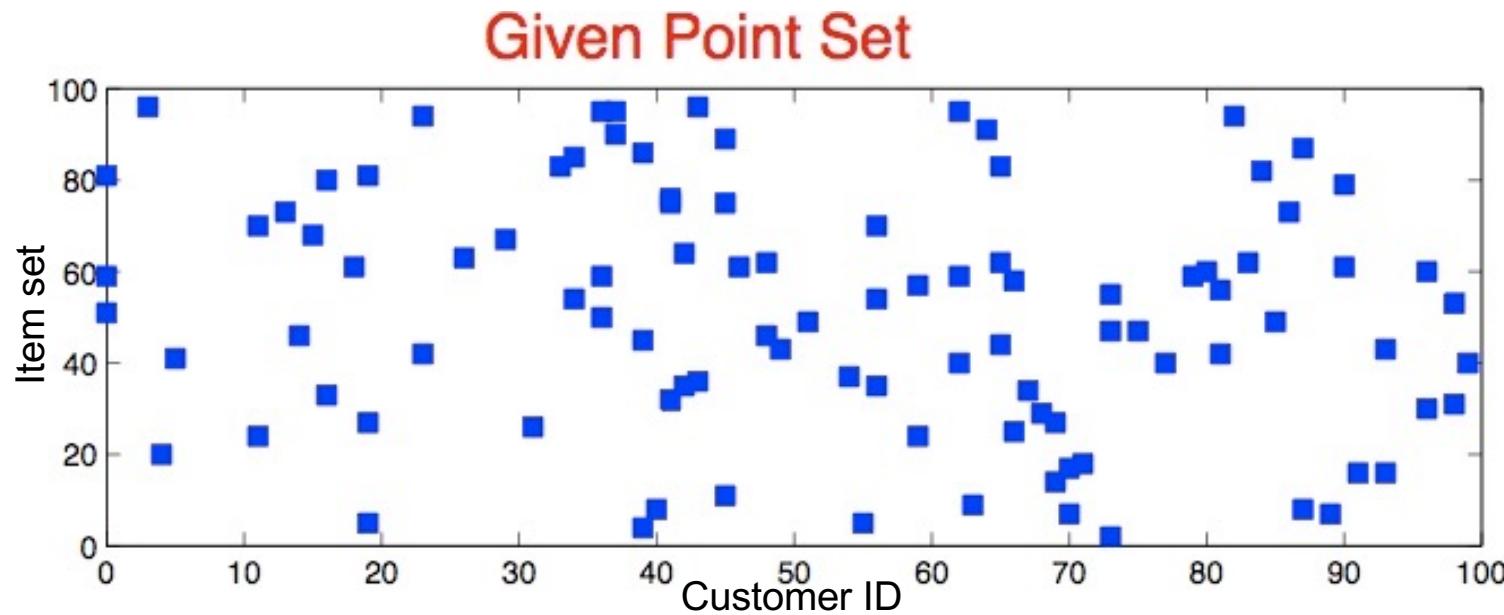


60 Iterations



100 Iterations

SVD in Data Mining

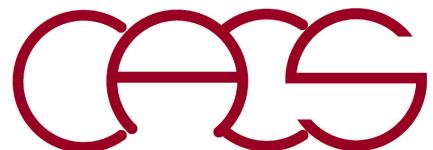


Machine Learning in Simulation

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

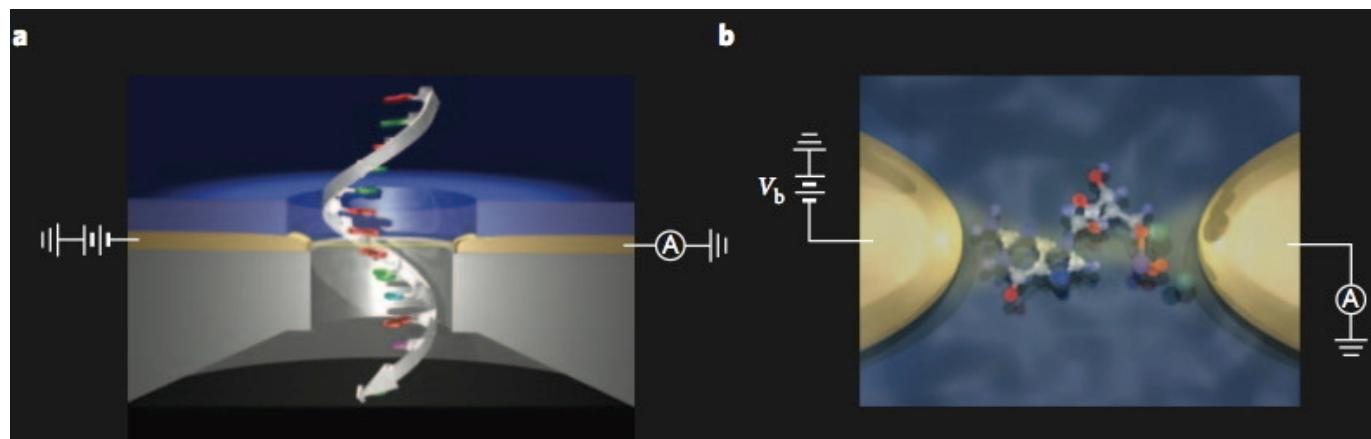


SVD for Rapid Genome Sequencing

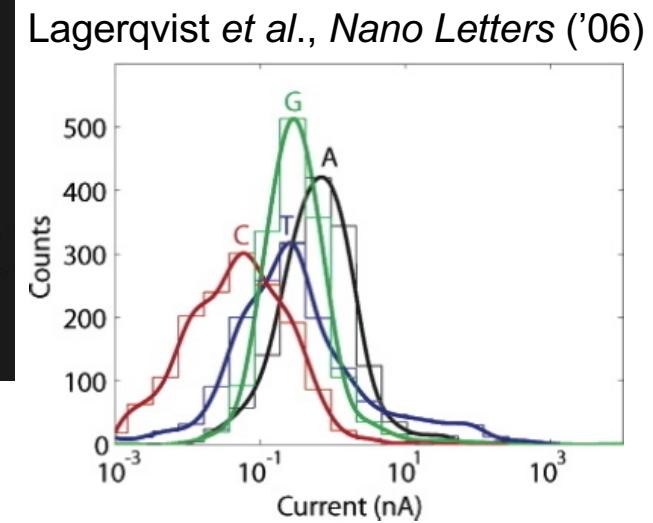
- \$10M Archon X prize for decoding 100 human genomes in 10 days & \$10K per genome (<http://genomics.xprize.org>): Preemptive attack on diseases



- Quantum tunneling current for rapid DNA sequencing?



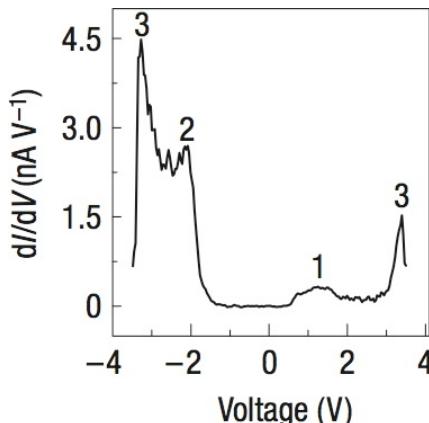
Tsutsui *et al.*, *Nature Nanotechnology* ('10)



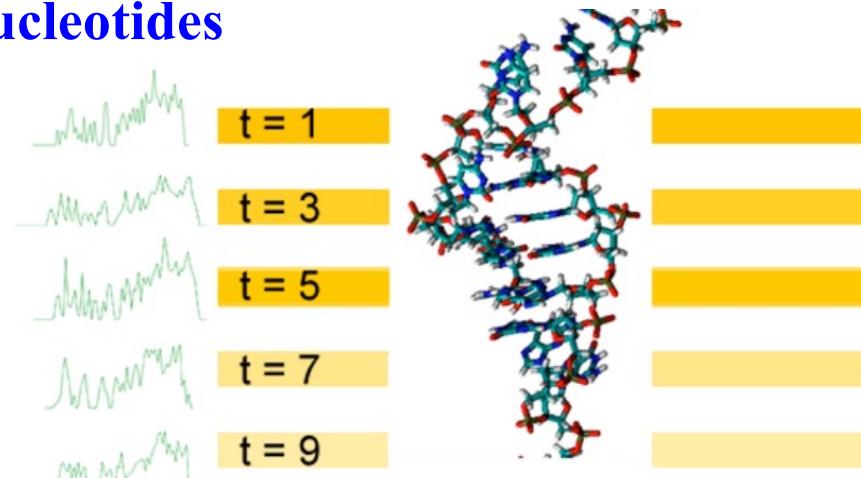
- Tunneling current alone cannot distinguish the 4 nucleotides (A, C, G, T)

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

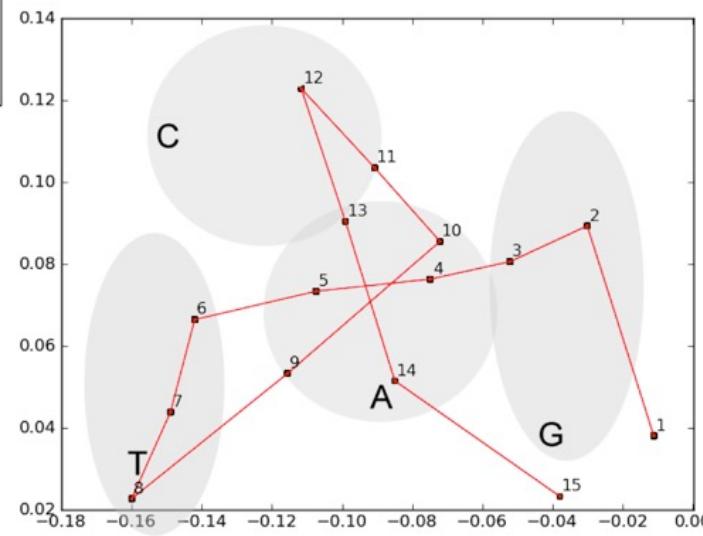
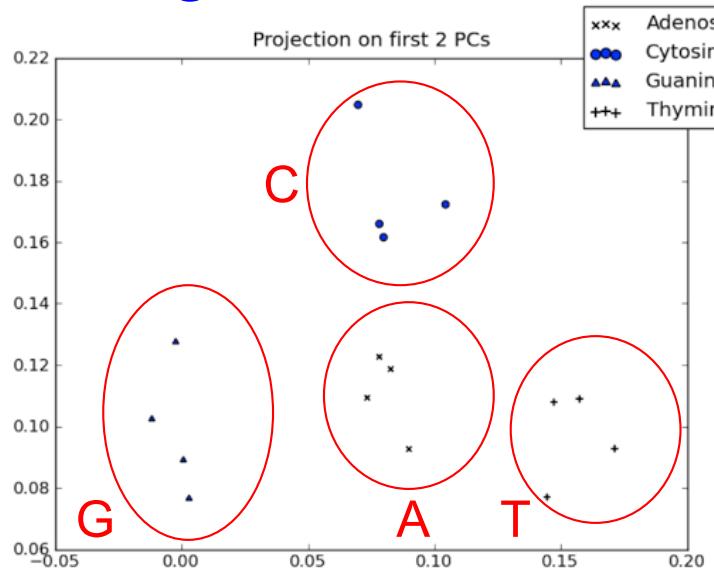


Shapir *et al.*,
Nature Materials ('08)



- 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) $\times T$ (number of time points) of stock-price time series

$$\underset{T \times N}{[\Sigma]}^T = \underset{T \times N}{[U]} \underset{N \times N}{\Sigma} \underset{N \times N}{[V]}^T$$

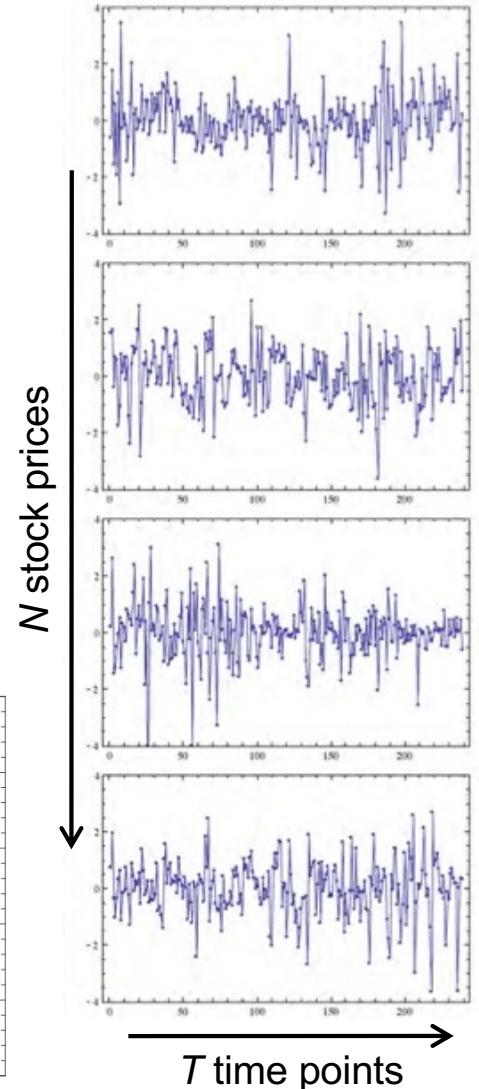
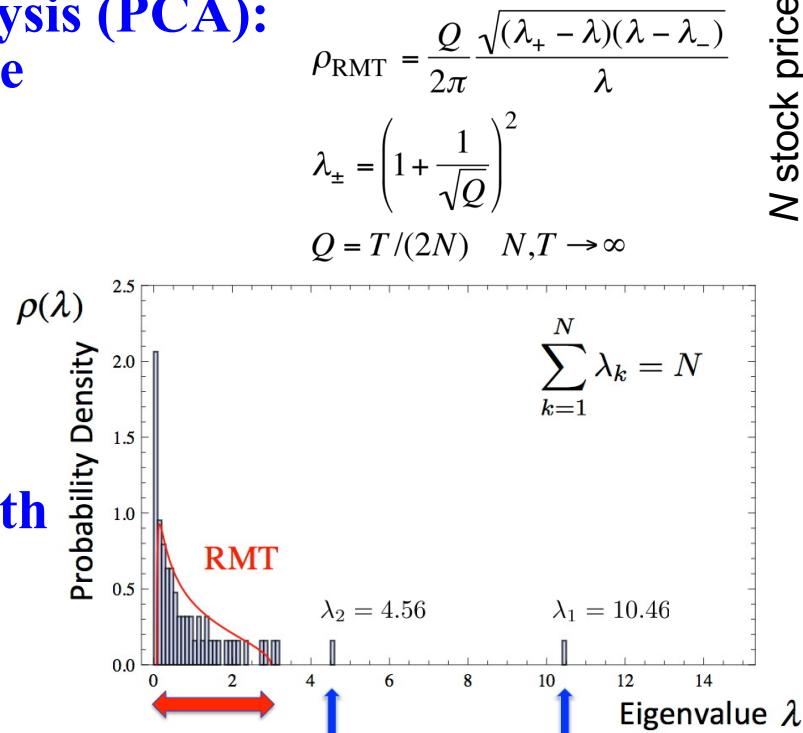
- Stock correlation matrix

$$\underset{N \times N}{[C]} = \underset{N \times T}{[\Sigma]} \underset{T \times N}{[\Sigma]}^T$$

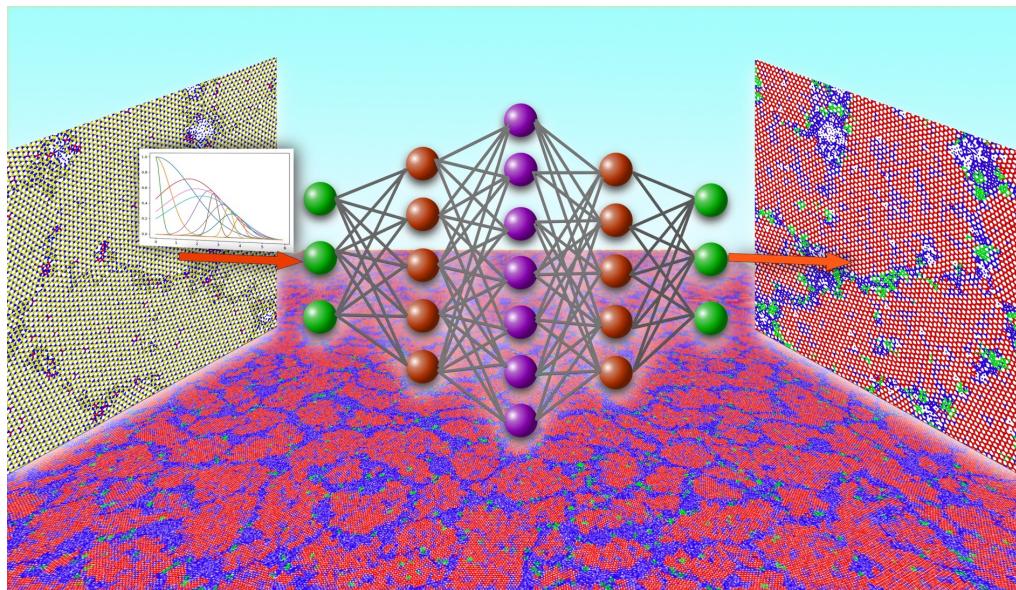
- Principal component analysis (PCA): Eigen decomposition of the correlation matrix

$$\begin{aligned} C &= \underset{I}{\Sigma} \Sigma^T \\ &= V \Sigma \widetilde{U^T U} \Sigma V^T \\ &= V \Sigma^2 V^T \end{aligned}$$

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



Learning Materials Phases & Defects

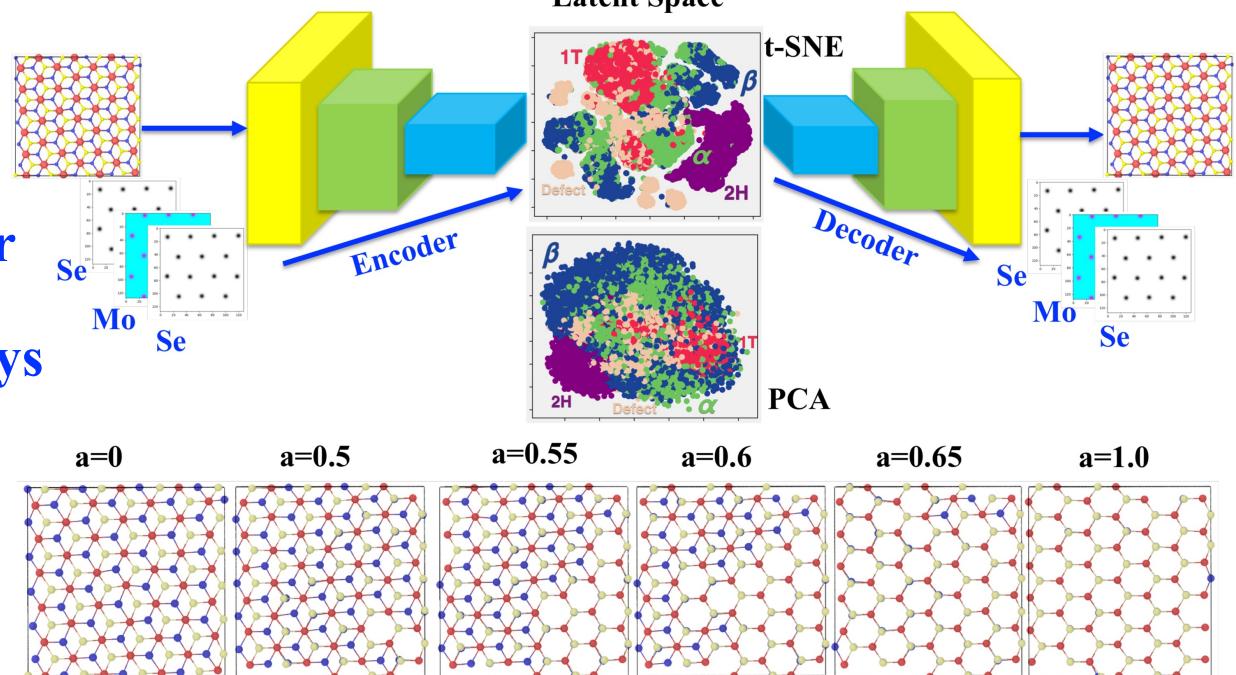


- Feedforward neural network to learn phases from local symmetry functions

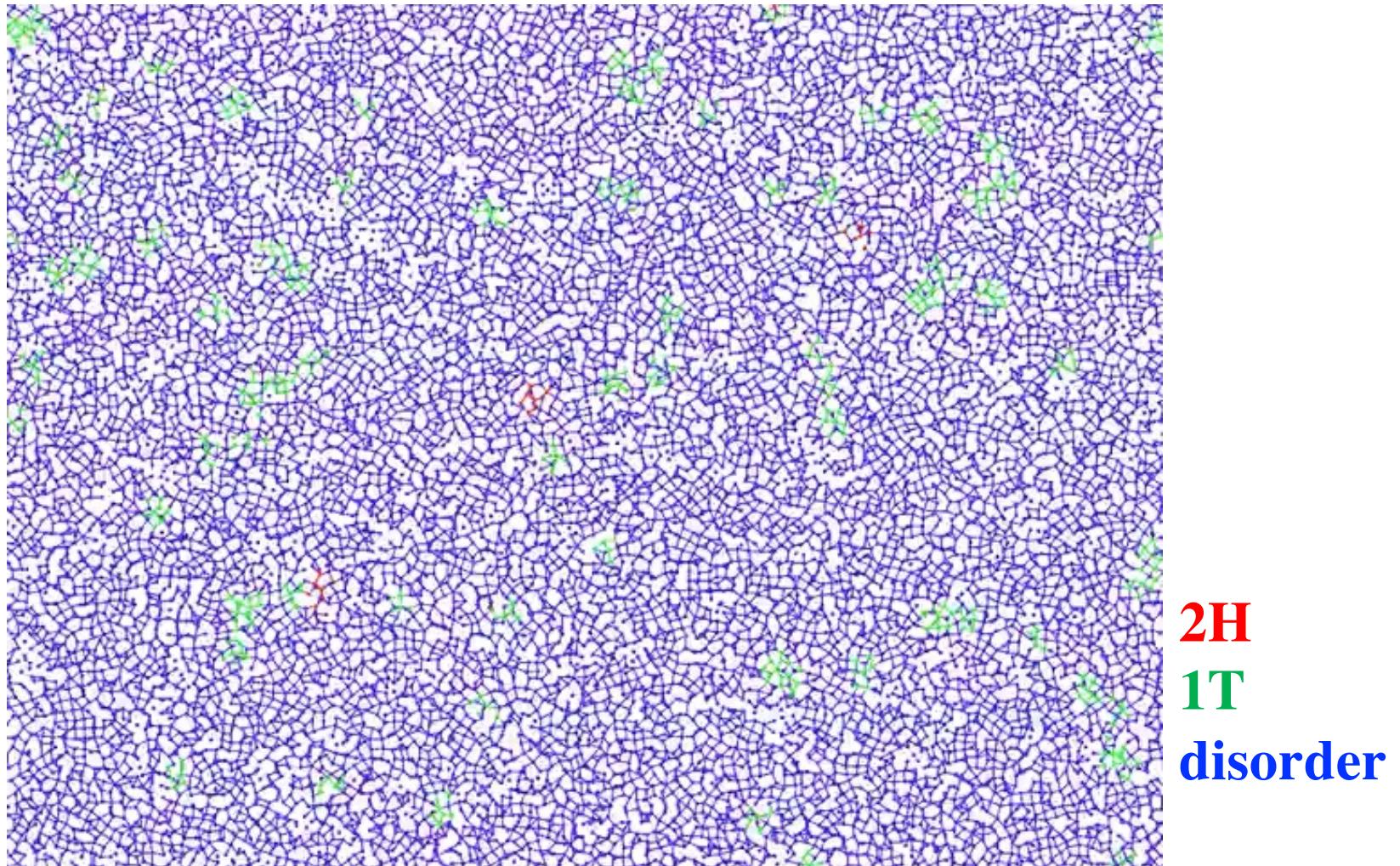
K. Liu *et al.*, Proc. ScalA18 ('18)
S. Hong *et al.*, JPCl 10, 2739 ('19)

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

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



Learning Transformation Pathways

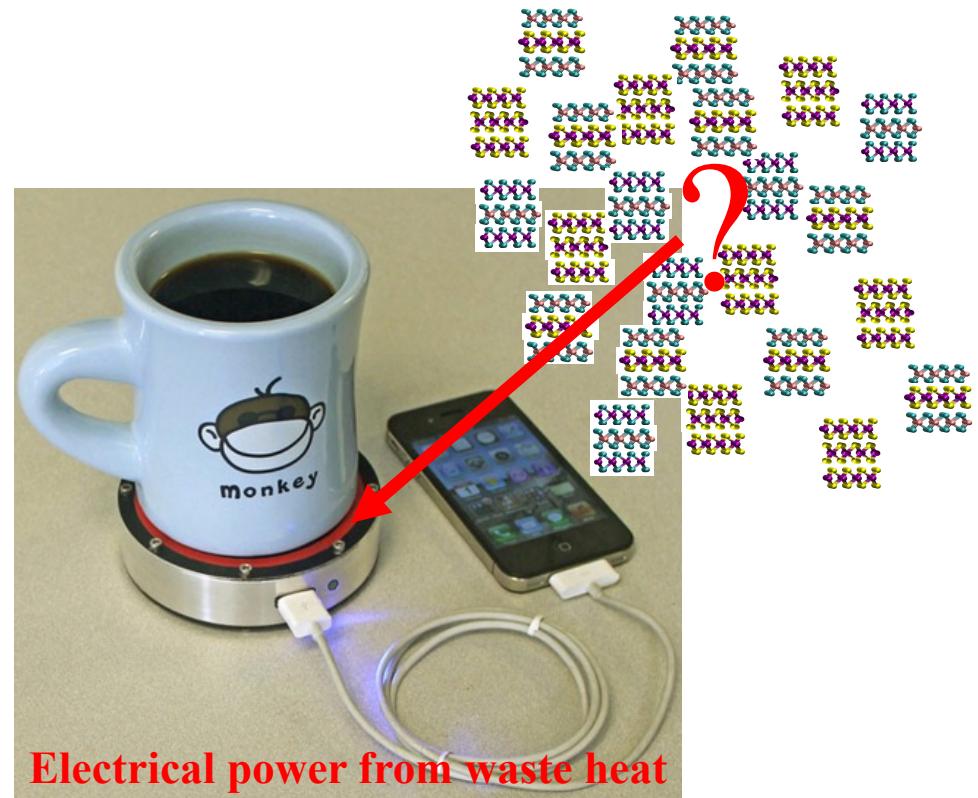
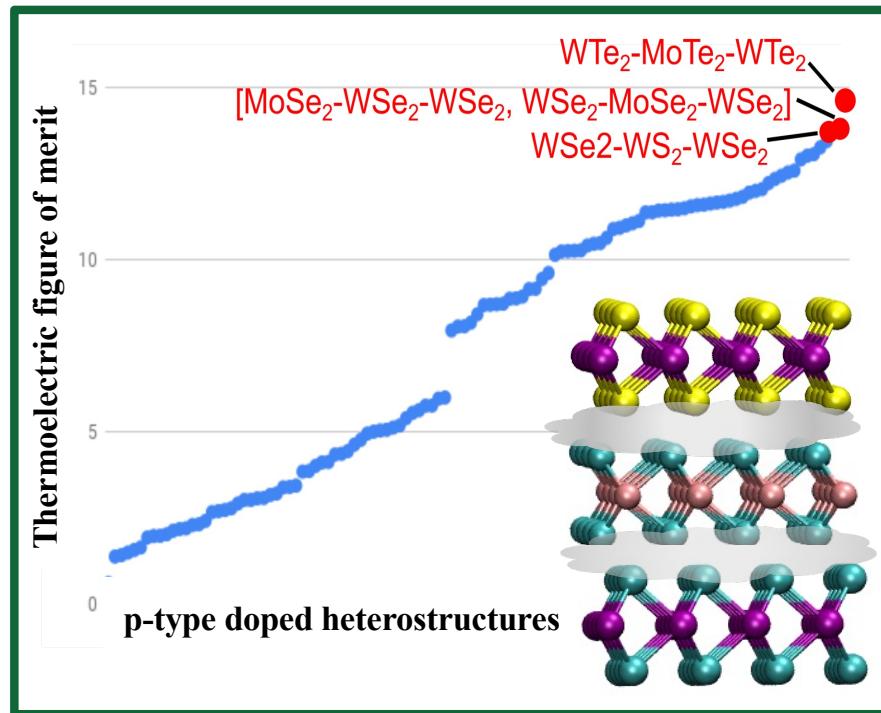


- Found novel transformation pathways to the stable 2H phase via the metastable 1T phase during chemical vapor deposition (CVD) growth of MoS_2

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

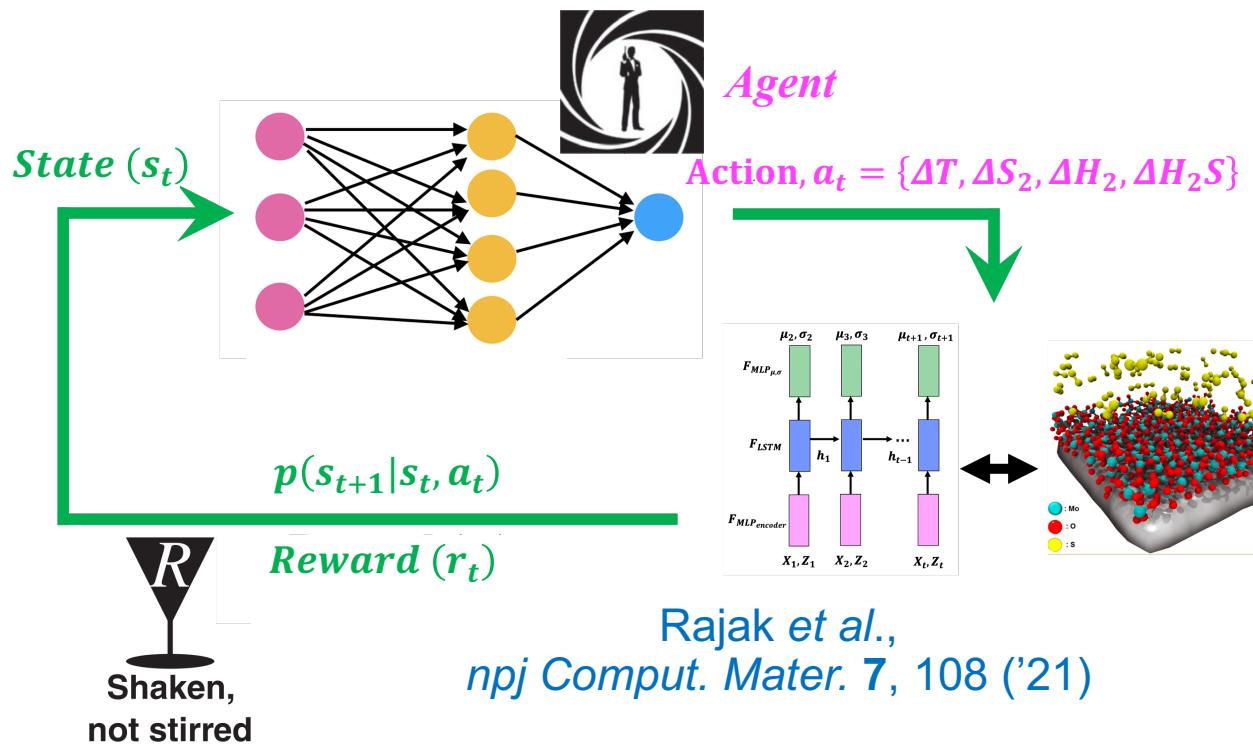
Active Learning of Optimal Materials

- 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



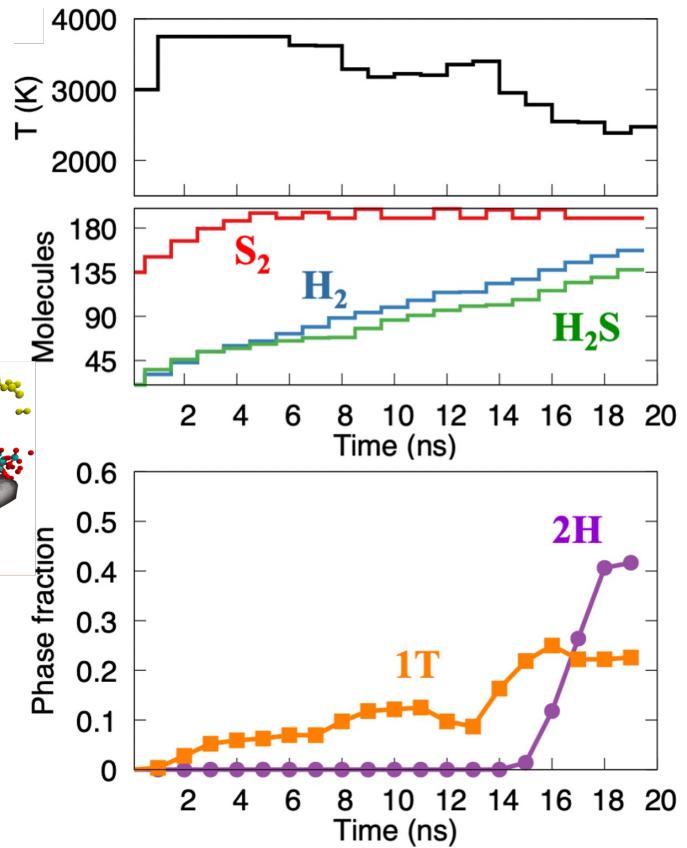
Reinforcement Learning for Growth

- 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. Neural 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)

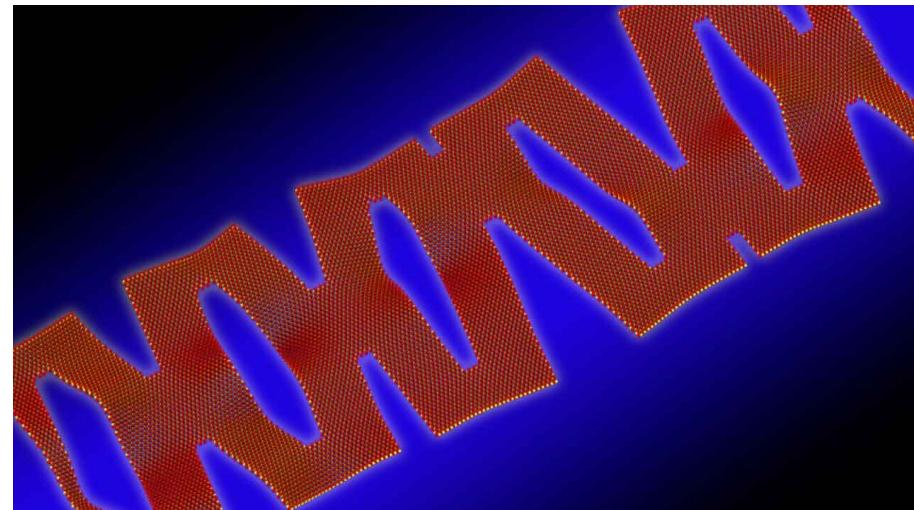
FEATURE STORY | ARGONNE NATIONAL LABORATORY

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 jiānzhǐ, it emerged around the 7th century in Japan,



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

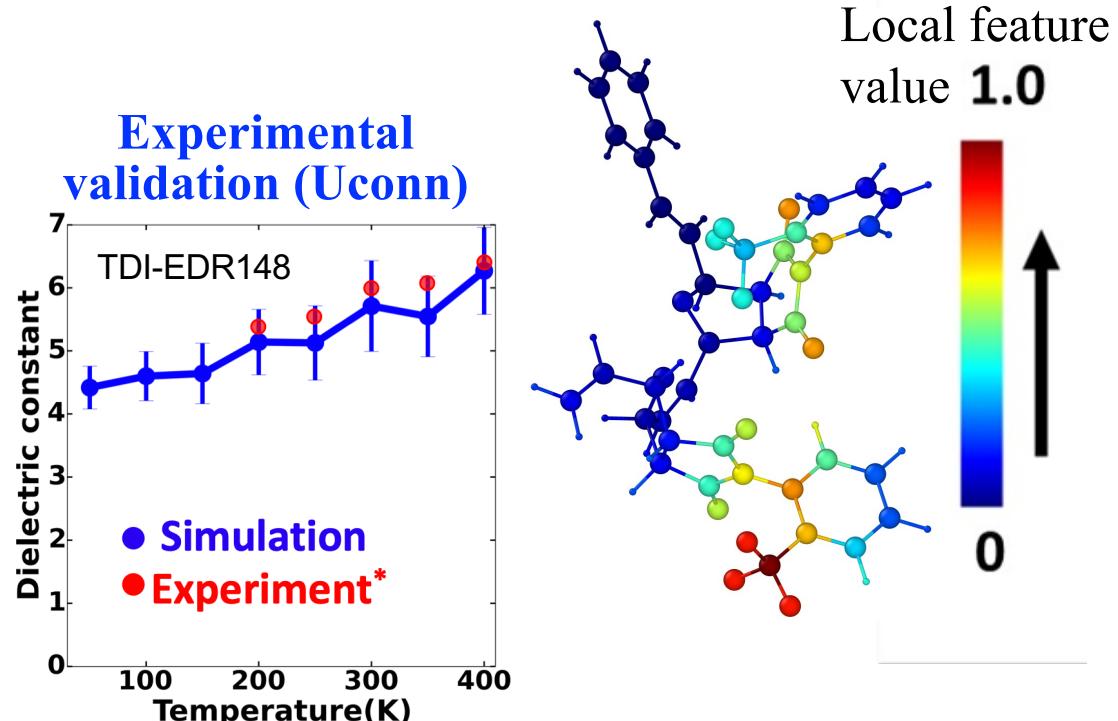
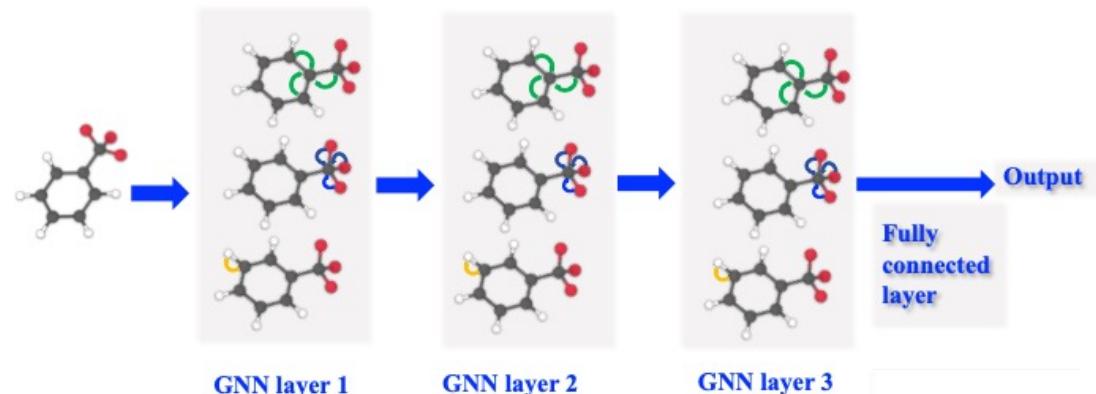
Dielectric Polymer Genome

Recurrent neural network for polymer property prediction



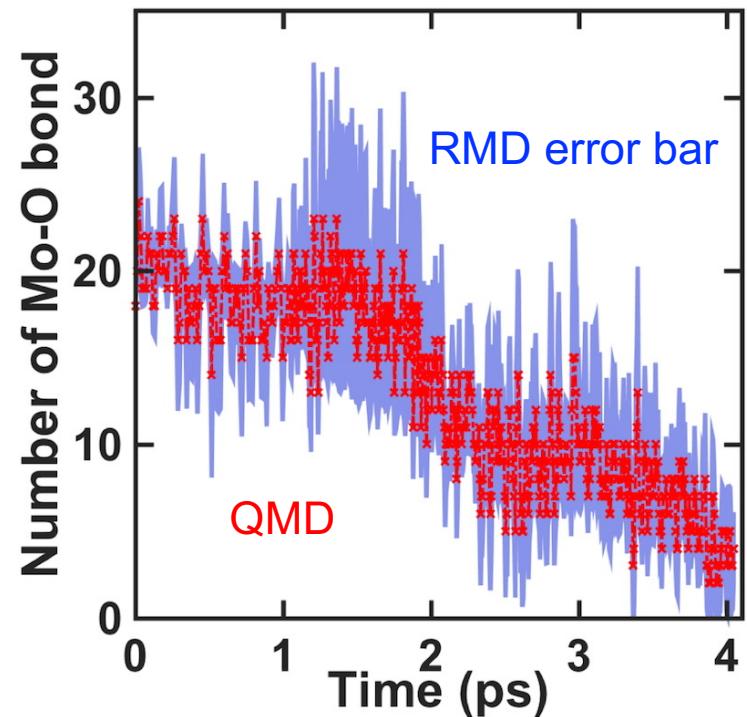
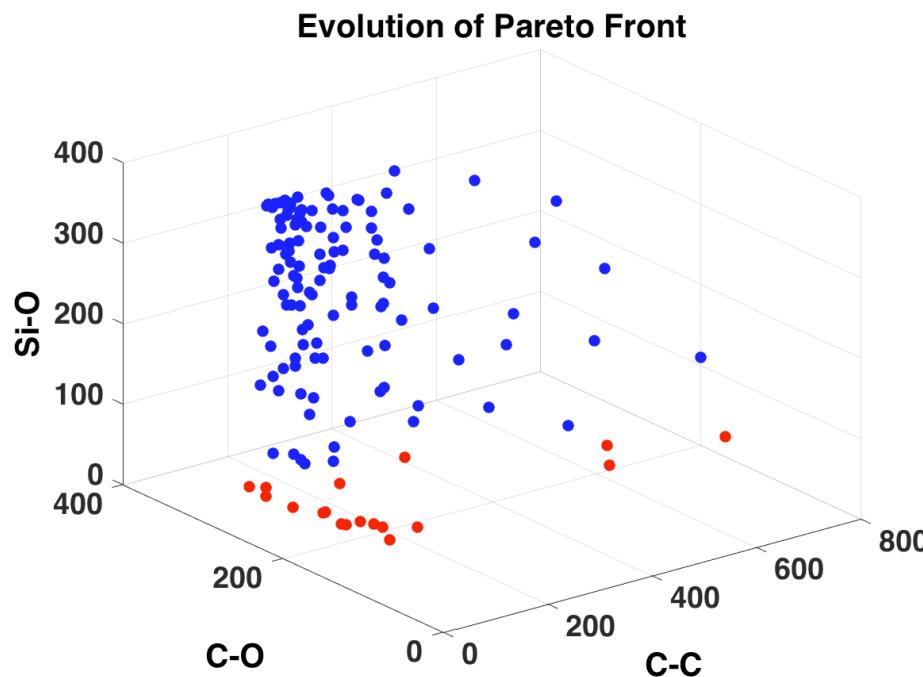
Nazarova et al.,
J. Chem. Info. Model. **61**, 2175 ('21)

Graph attention neural network for explainable property prediction



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)