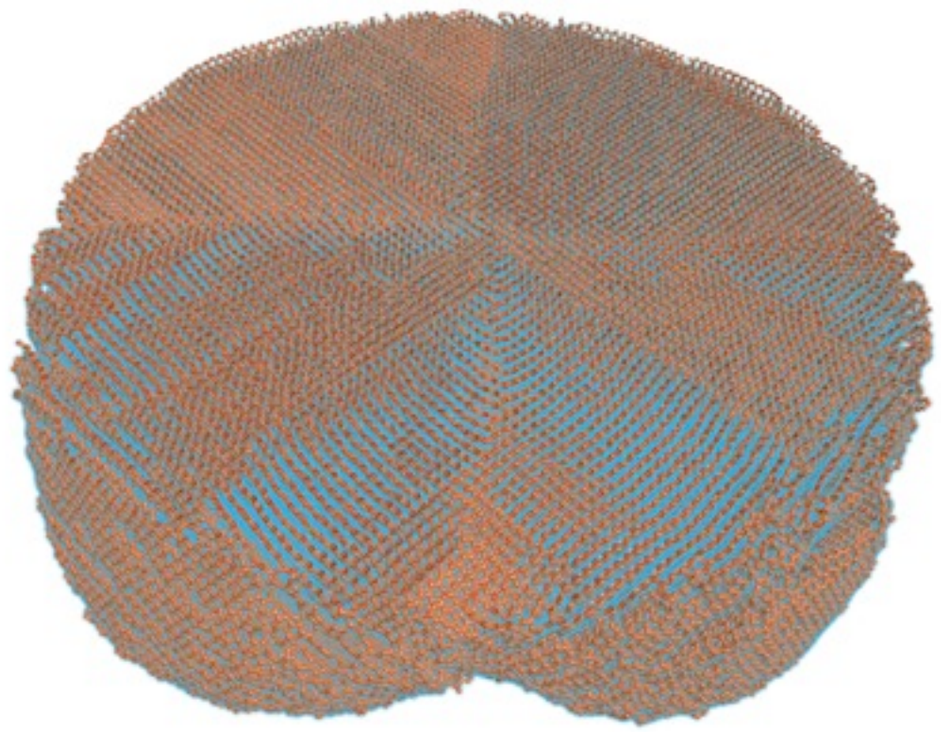


Scalable Visualization and Analysis for Computational Materials Science

Aaron Knoll, Joe Insley, Tom Uram, Venkatram Vishwanath,
Mark Hereld, Michael E Papka

Visualization Group
Argonne National Laboratory

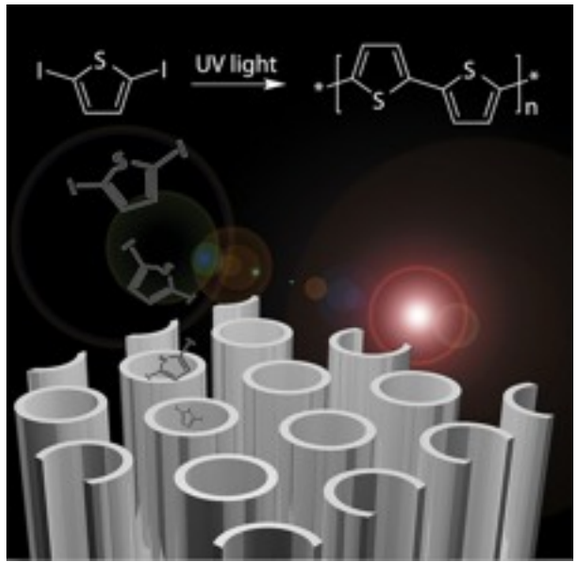
Motivation

- Computational chemistry drives **new energy technology**

- battery, photovoltaic, synthetic & biofuels, biomass conversion

- catalysis, diffusion, oxidation, heat/energy transfer, structure

- special vis/analysis needs



Petascale-exascale era

- K computer
  (3 / 5 applications in PR)

- ALCF Mira
  (7 / 16 early science projects)

- OLCF Titan
  (2 / 6 critical codes)

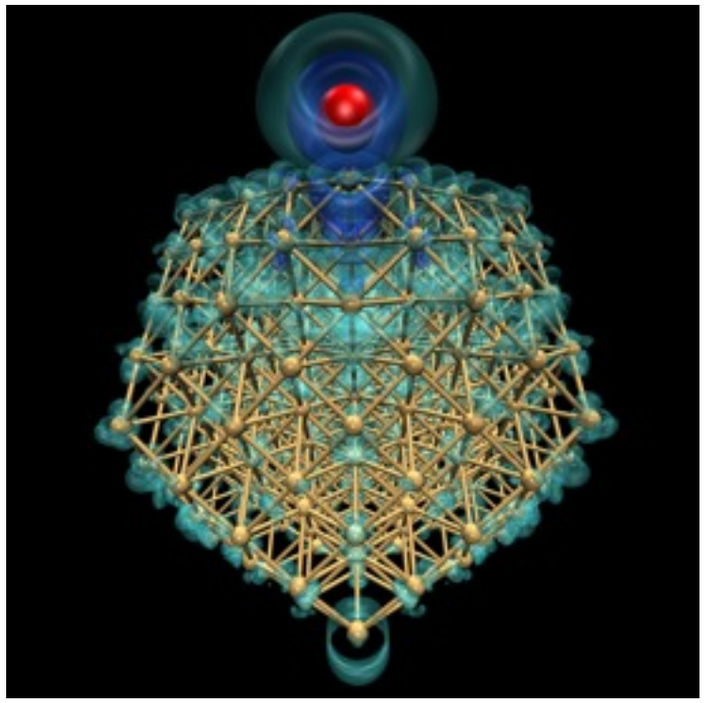
When the LINPACK benchmark program measured the K computer in its final configuration, the supercomputer system achieved a speed of 10.51 petaflops exceeding its previous speed of 8.162 petaflops, which had placed the system in first place on the TOP500 supercomputer list published June 2011, the companies said.

So what kind of applications is such a monster machine good for? Fujitsu and Riken listed the following:

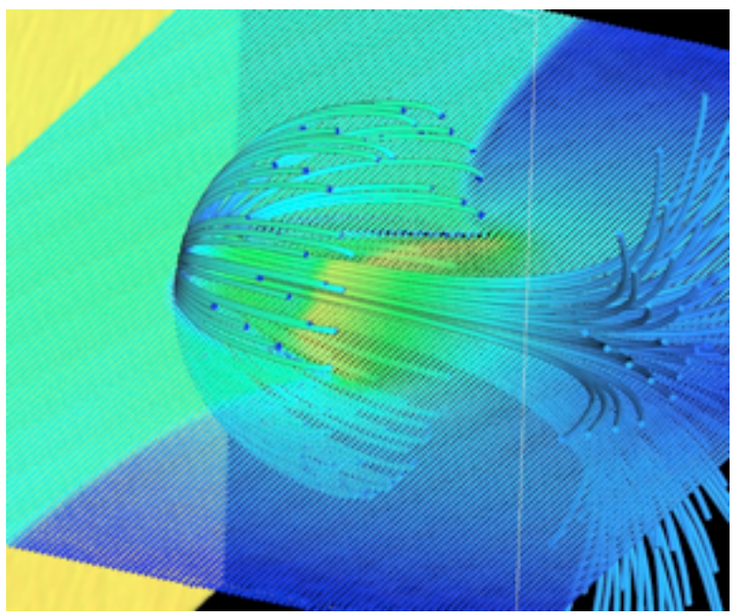
- Analyzing the behavior of nanomaterials through simulations and contributing to the early development of such next-generation semiconductor materials, particularly nanowires and carbon nanotubes, that are expected to lead to future fast-response, low-power devices.
- Predicting which compounds, from among a massive number of drug candidate molecules, will prevent illnesses by binding with active regions on the proteins that cause illnesses, as a way to reduce drug development times and costs (pharmaceutical applications).
- Simulating the actions of atoms and electrons in dye-sensitized solar cells to contribute to the development of solar cells with higher energy-conversion efficiency.
- Simulating seismic wave propagation, strong motion, and tsunamis to predict the effects they will have on human-made structures; predicting the extent of earthquake-impact zones for disaster prevention purposes; and contributing to the design of quake-resistant structures.
- Conducting high-resolution (400-m) simulations of atmospheric circulation models to provide detailed predictions of weather phenomena that elucidate localized effects, such as cloudbursts.

Computational chemistry

- **chem data is (relatively) small**
- **Density Functional Theory (DFT)** :
  simulate electrons, chemical bonds
  - catalysis, oxidation, chemical reactions
    - 100-1k electrons typical (16k electrons is big)
- **Molecular Dynamics (MD)** :
  simulate atoms, inter-molecular forces
  - diffusion, thermal annealing, structural stability
  - 10-100k atoms typical (1 million - 1 billion is big)
- *ab initio* (AIMD), QMC, others



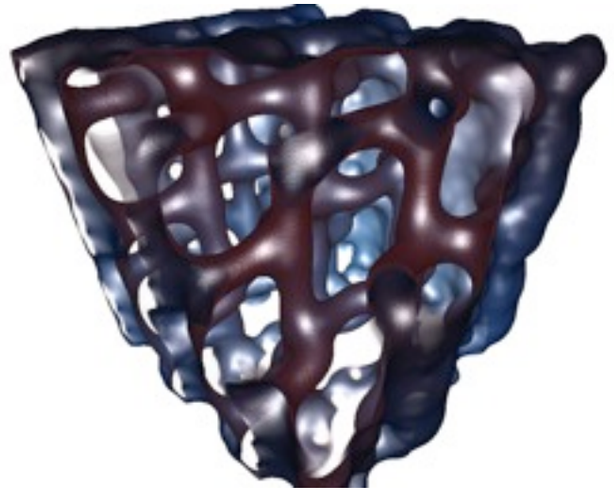
data courtesy Jeff Greeley, ANL CNM

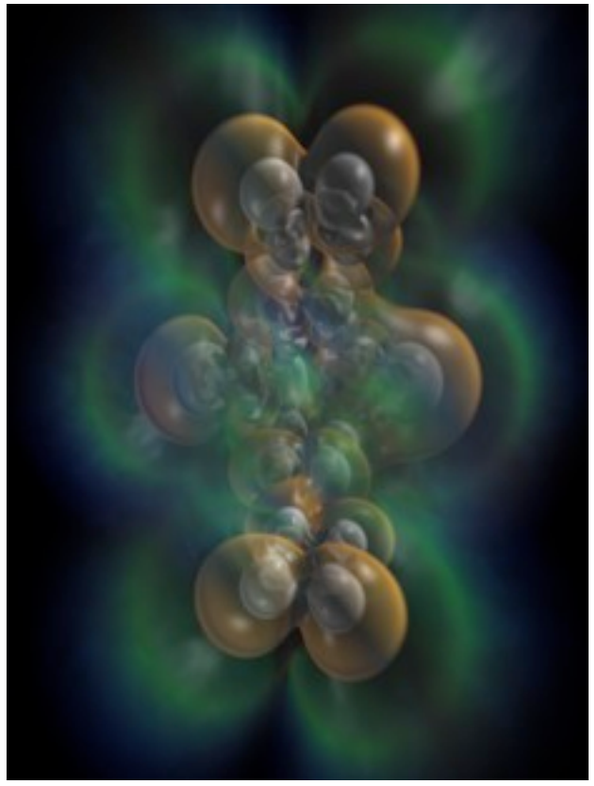


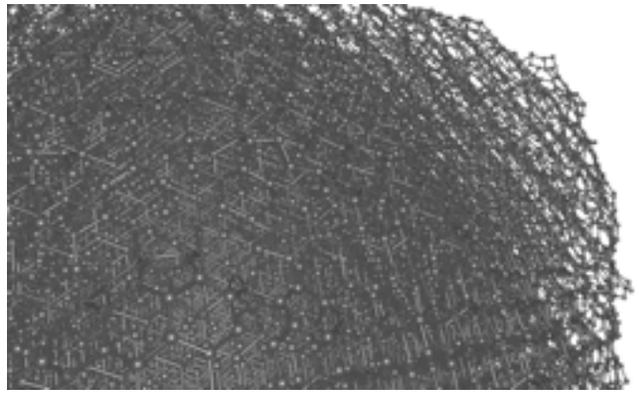
data courtesy Ken-ichi Nomura,
Priya Vashishta, USC

Chemistry vis challenges

- How do we represent molecular geometry?
- How do we interpret volume data computed from DFT?
- How do we visualize macromolecules?
- How do we compare compounds and reactions?

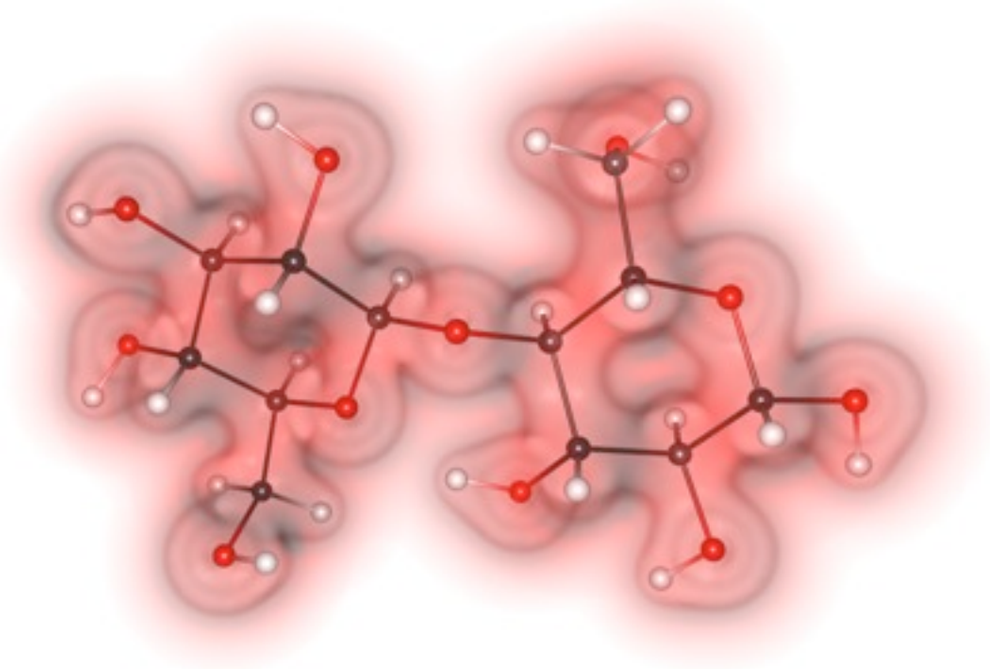






Some quantum physics

- Self-consistent field (**SCF**) theory: molecular structure is continuous



- Schroedinger Equation

$$E\psi = H\psi$$

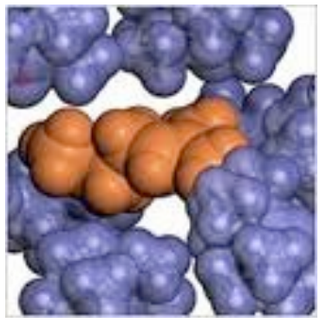
- Linear Combination of Atomic Orbitals (LCAO) DFT

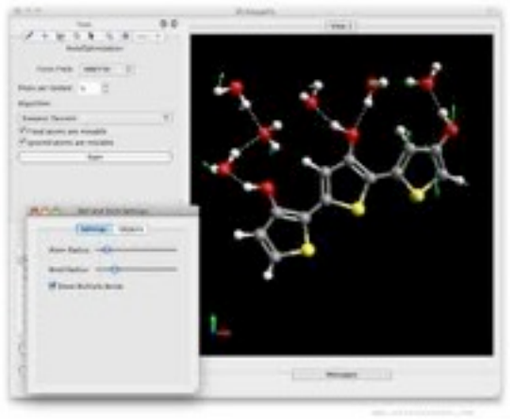
$$\theta_{nlm}(r,\theta,\phi) = R_{nl}(r)Y_{lm}(\theta,\phi)$$

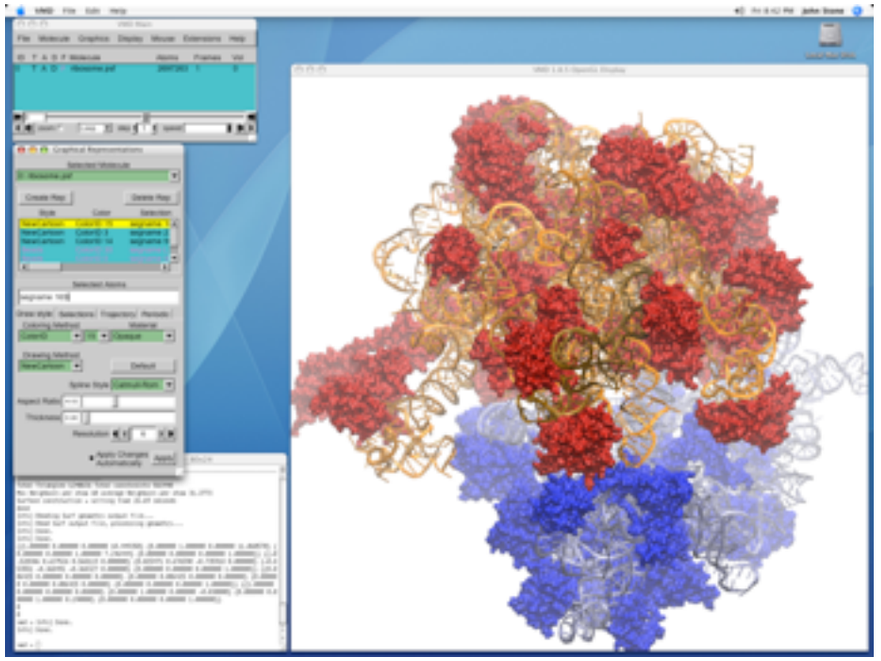
Molecular geometry is volume data.

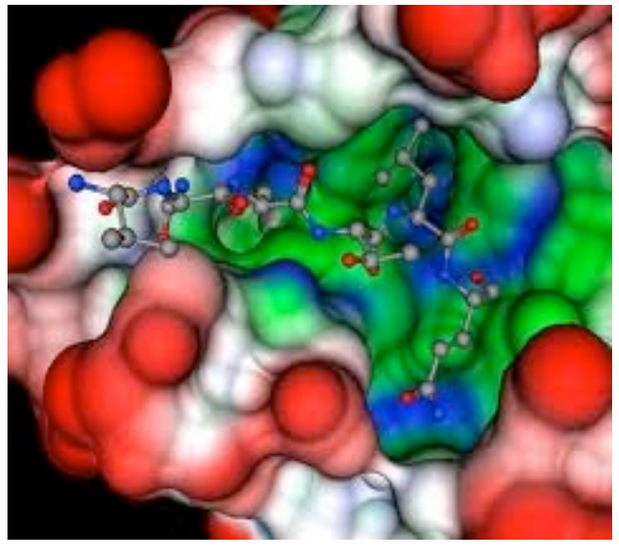
Chem vis state-of-the-art

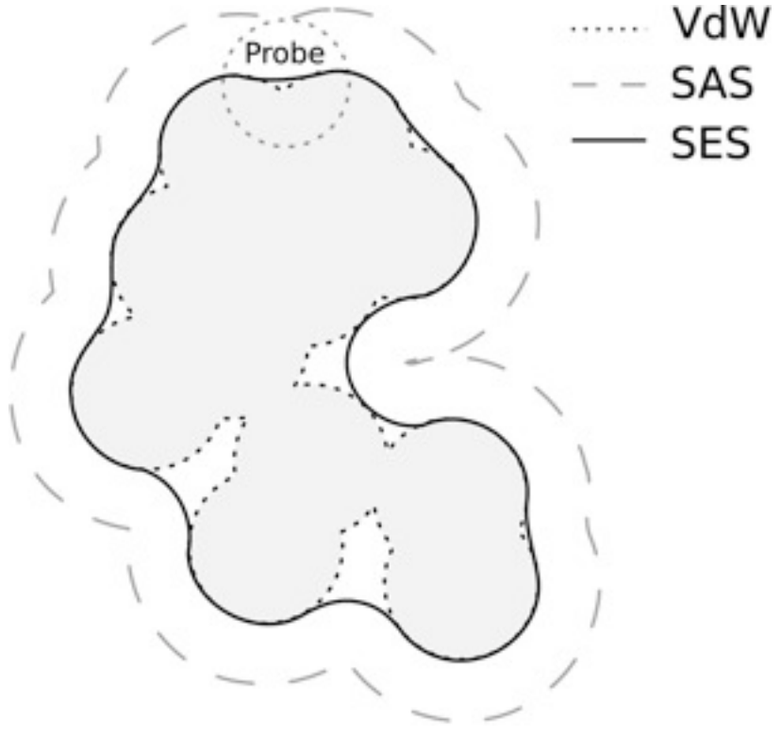
- VMD, JMol, Avogadro, MGLTools, Gaussview, Materials Studio
- Visit, Paraview
- modalities:
  - ball & stick / particles
  - molecular surfaces
  - ribbons





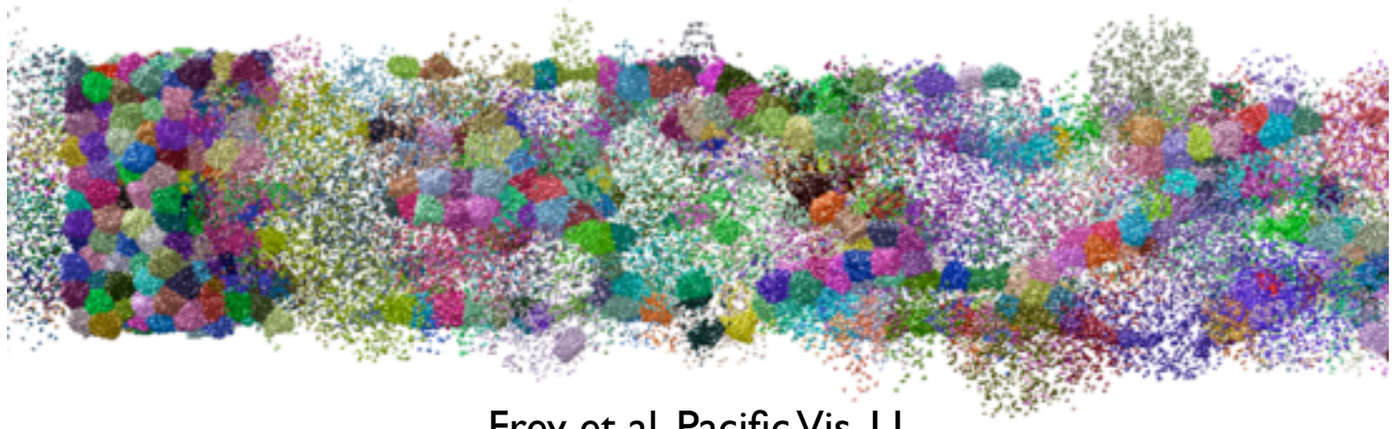




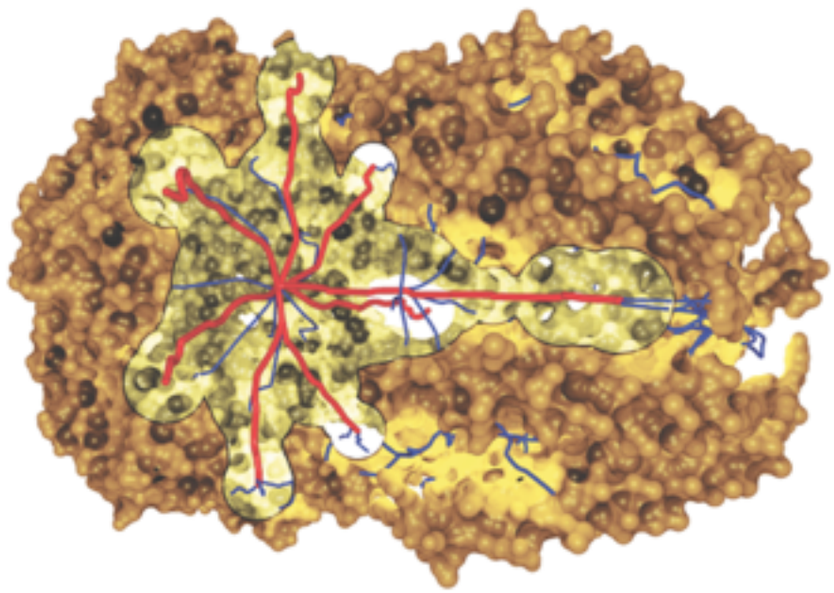


Modality matters

- desired goals:
  - scale visually
  - show bond structure
  - appropriate underlying physical model
- ball & stick, particles, molecular surfaces all have limitations
- volume representation?



Frey et al. Pacific Vis 11

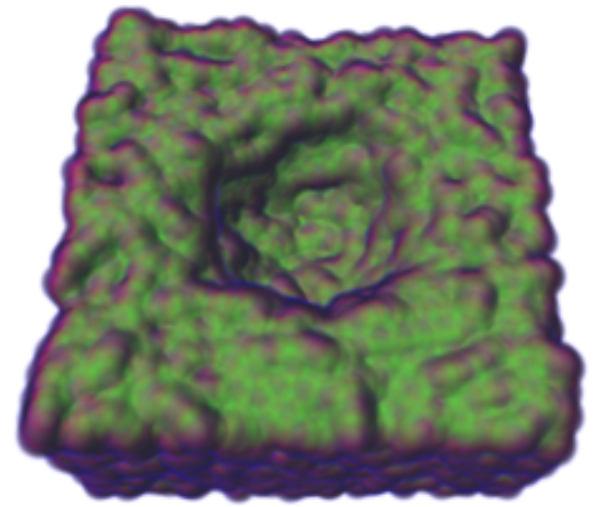


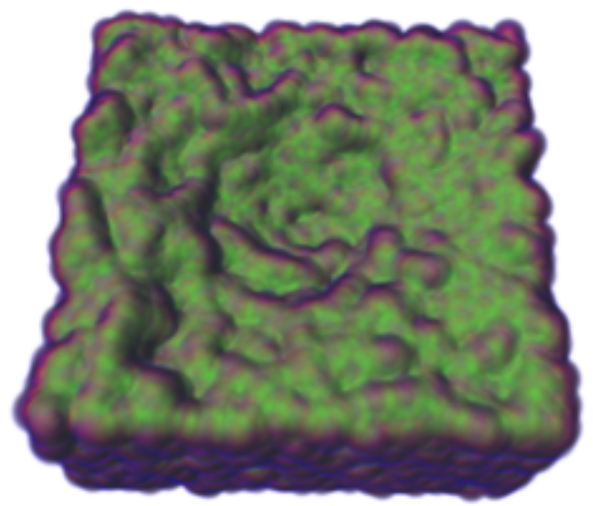
Lindow et al. IEEE Vis11

Volumetric vis/analysis for Computational Materials Science

We propose

- **use the actual SCF** from DFT computation
  - for vis, analysis of DFT data
  - model **approximate SCF's** for MD data
  - ANL booth talk, Tues at 11:30
    "Uncertainty Classification of Molecular Interfaces"



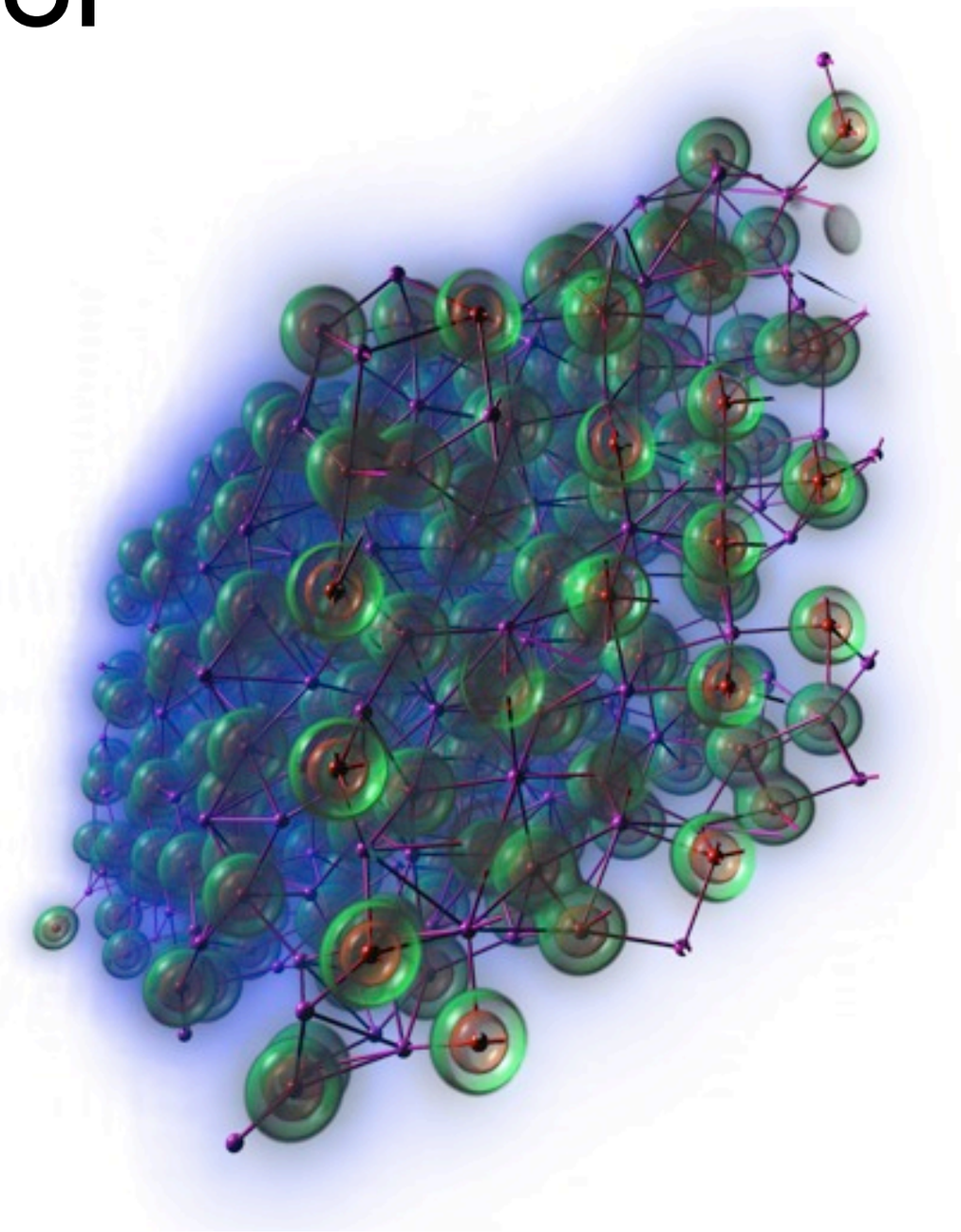


Model chemistry volumetrically.

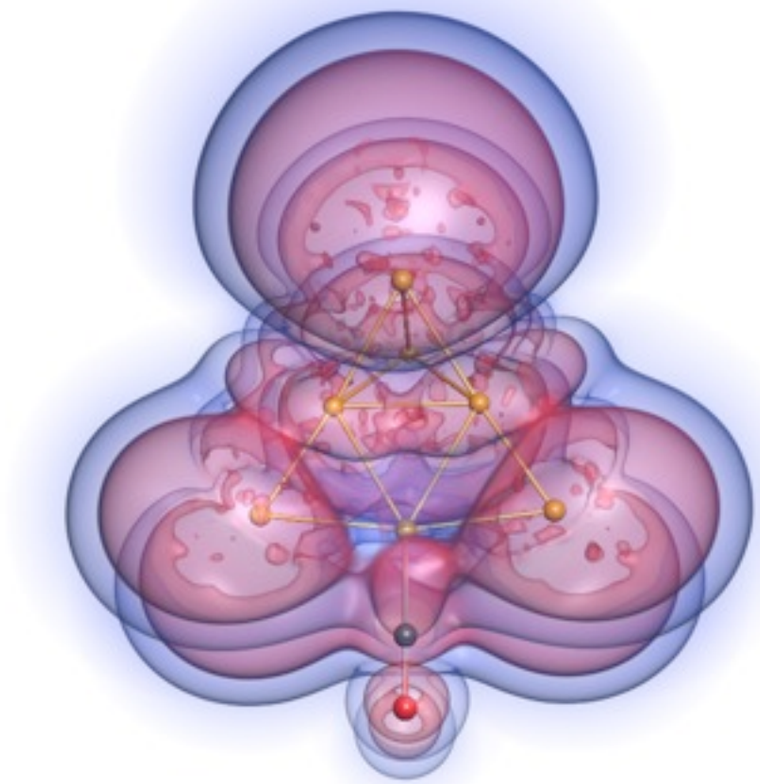
- Do vis, analysis based on first principles, not abstractions.

Nanovol

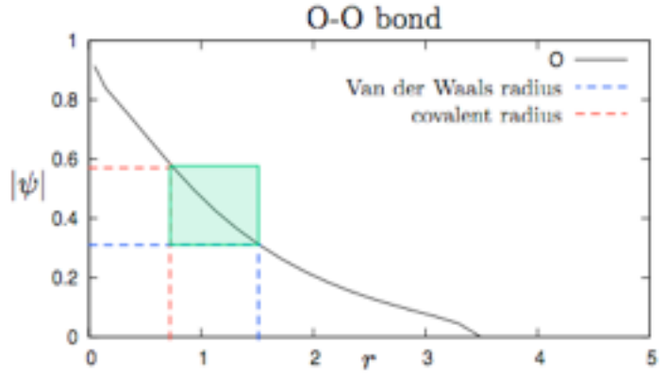
- Domain-specific vis tool
- Interactive GPU ray-casting
  - ball & stick rendering
    - scalability vs. rasterization
  - volume rendering
    (SCF, potentials, derived fields)
    - tri- cubic B-spline interpolation
- approximate SCF's for MD data
- classification / quantitative analysis

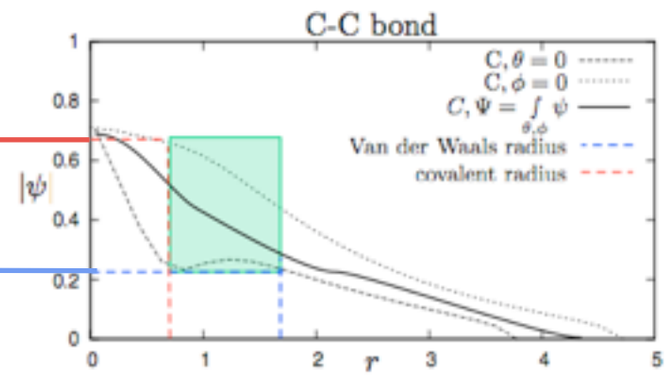


DFT classification example



electron density distribution (DFT on CO)

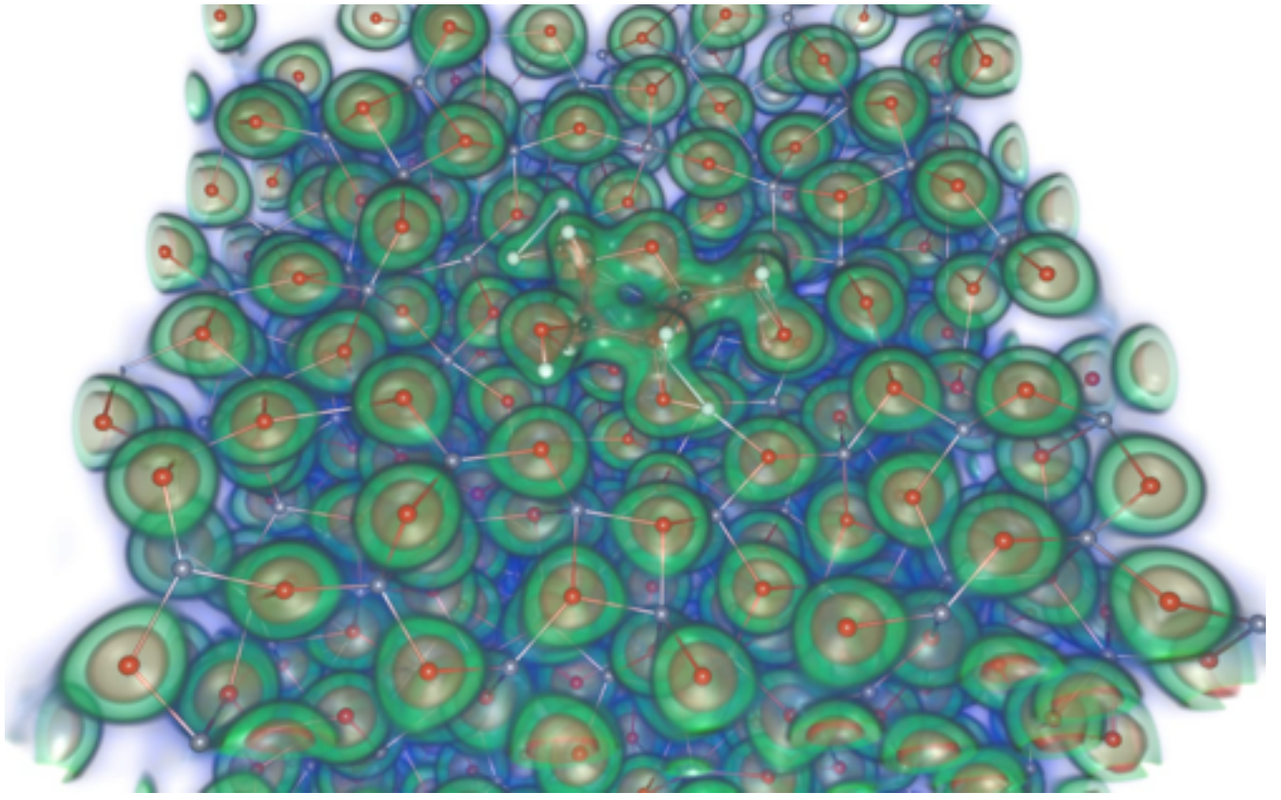




material interface (we don't know)

DFT - fructose nanobowl

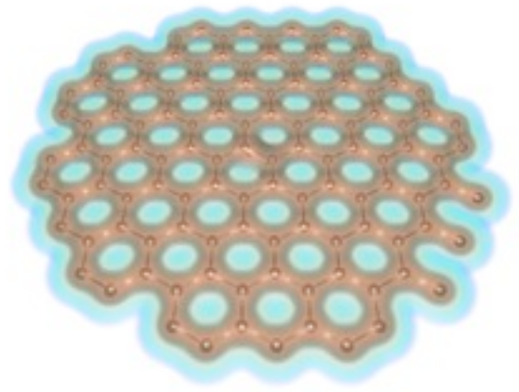
- Computed in GPAW on Intrepid (IBM BG/P), 4 million core hours
- Input: 1000 atoms (28 KB text file)
- Output: wavefunctions matrices of 9k electrons
- 55 GB per SCF, 190 SCF's, **10 TB of data**
  - (but, only wrote one equilibriated SCF to disk!)
- What we want to visualize:
  all-electron density (120^3 volume):
  2.8 MB x 190 SCF's, ~500 MB total
- What scientists want:
  activation energy of bonded compound
  (a single number!)
- Visualization is for **verification.**
  - (and PR for a big run!)



data of Lei Cheng, Larry Curtiss (MSD)
and Nick Romero (MCS) at ANL.

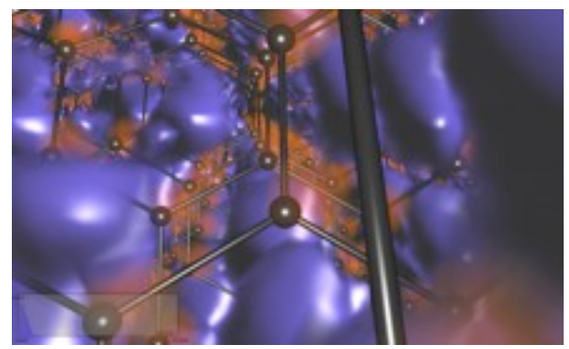
Approximate density fields (ADF)

- Use bulk DFT density distributions to approximate SCF's for MD data



  - linear combination of per-atom basis functions (kernel density estimation)

$$r_i(\vec{x}) = |\vec{x} - \vec{p}_i|$$
$$\Psi(\vec{x}) = \sum_i \Theta_i(r_i(\vec{x}))$$

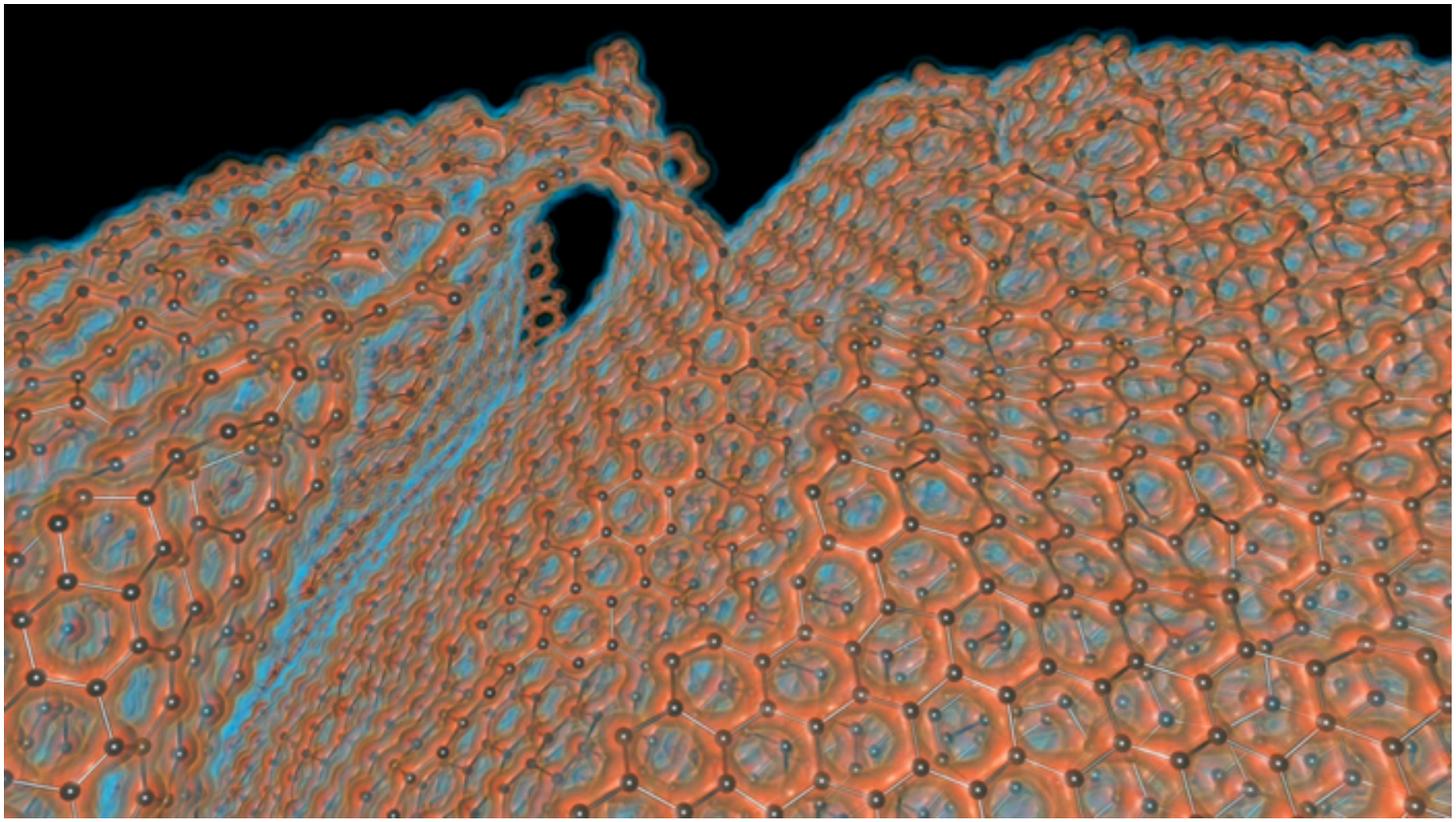


    - embarrassingly parallel

- volume rendering reduces clutter, shows structure

- image analysis of approximate SCF

MD - carbon nanospheres



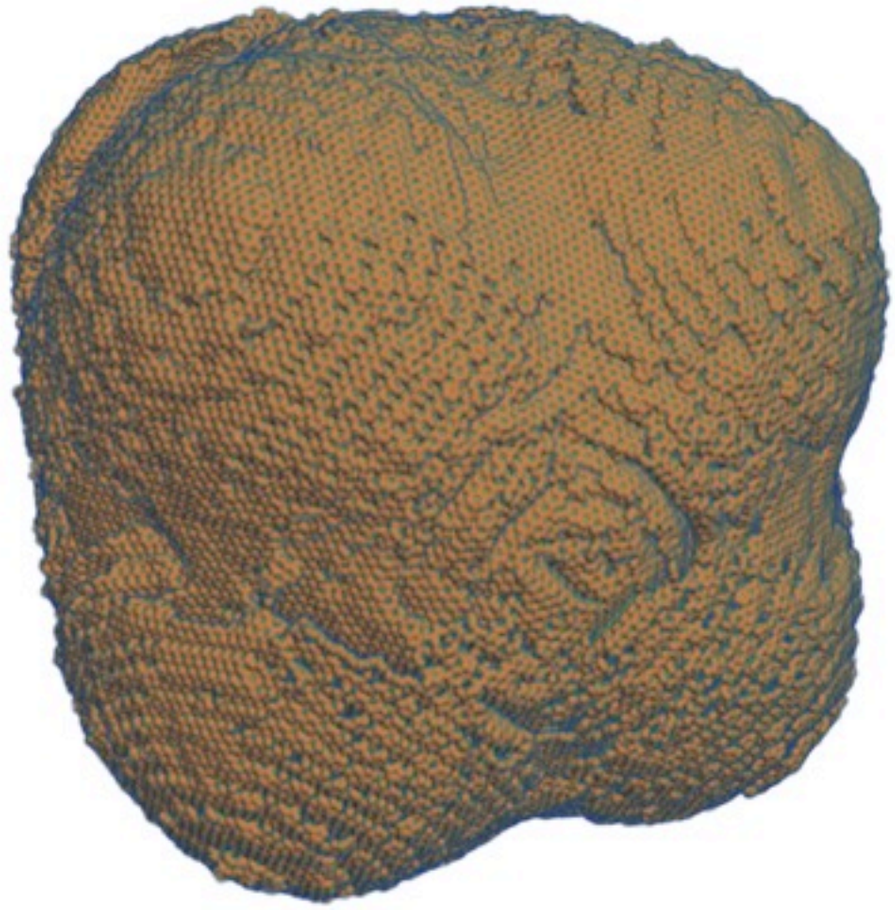
- Computed using LAMMPS, ~30,000 core hours on LCRC Fusion
- Input: amorphous carbon, 740k atoms (41 MB)
- Output: annealed geometry, 740k atoms + variables, 500k timesteps = **~20 TB** (but, only final step written to disk!)
- scientific goals: understanding / validation of structure from annealing, void space, diffusion paths

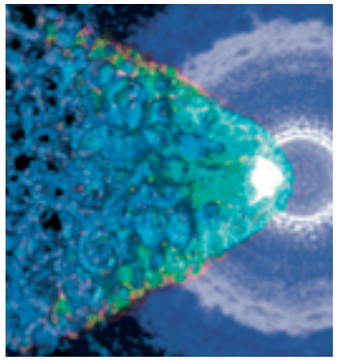
ADF scalability

- current nanosphere model:
  0.5 microns, 740k atoms, ~1k^3
  SCF, **0.5 PB**

- experimental scale (per nanosphere):
  5 microns, 10M atoms, ~2k^3 SCF,
  5 PB

- GLEAN, VL3
- Generate ADF on-the-fly
  - Vis directly from particle data
  - what do we do for analysis?
- Sacrifice temporal resolution
- Compression





Fraedrich et al. Vis 10

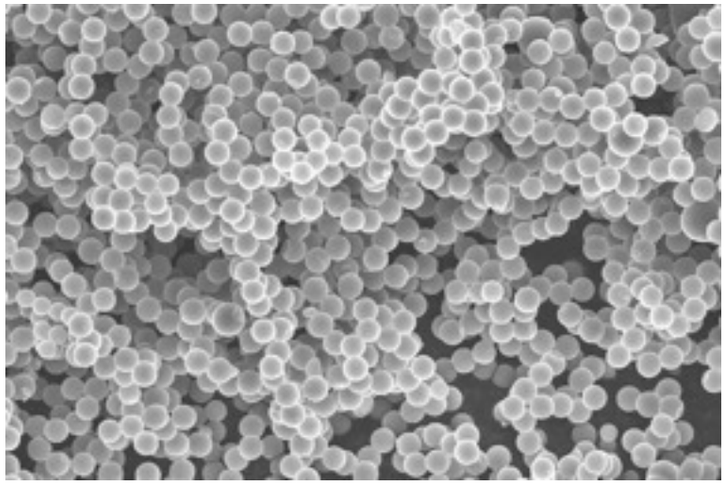
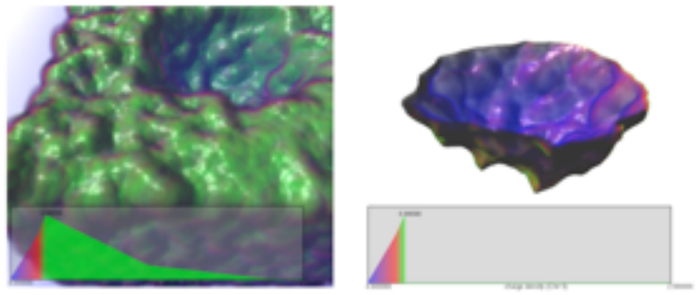
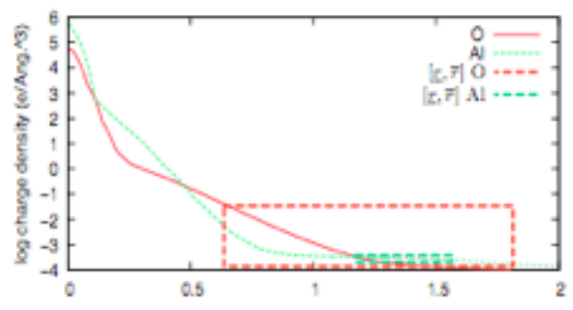


Image courtesy Vilas Pol, ANL MSD

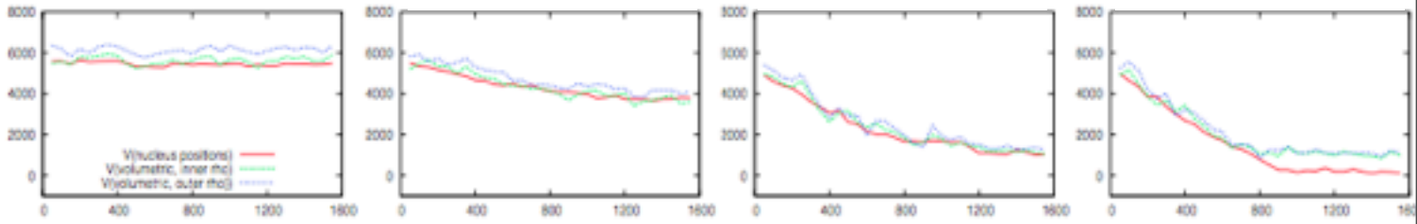
MD - nanobowls

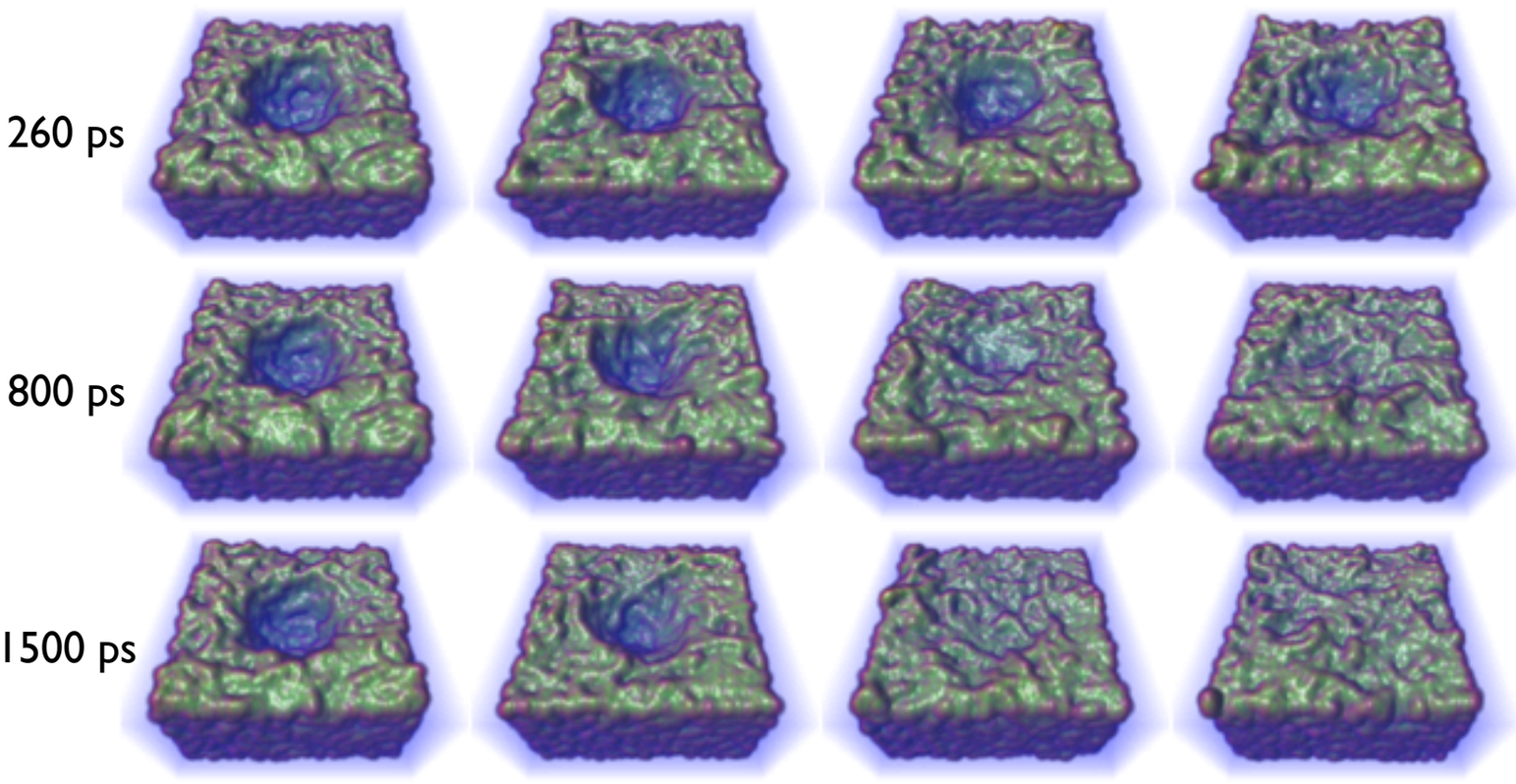
- ensembles can have many parameters.
- alumina oxide nanobowls
  20k atoms x 150,000 timesteps in DL_POLY
  - bowl radius (4-15 Angstrom)
    - temperature
      (1000K - 1500K)
    - embedded fuels, catalysts
- 50,000x temporal loss
  (400 TB *per run* with ADF's)
- **comparative vis**
  WYSIWYG analysis





1000K 1200K 1300K 1350K

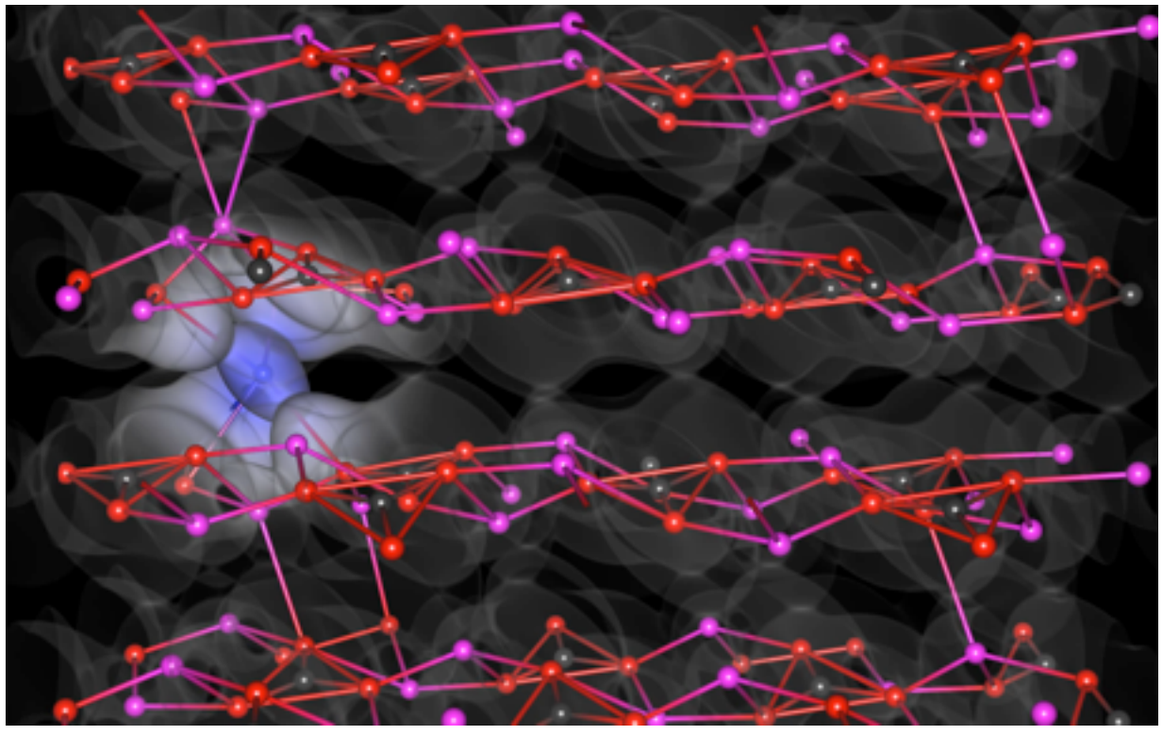


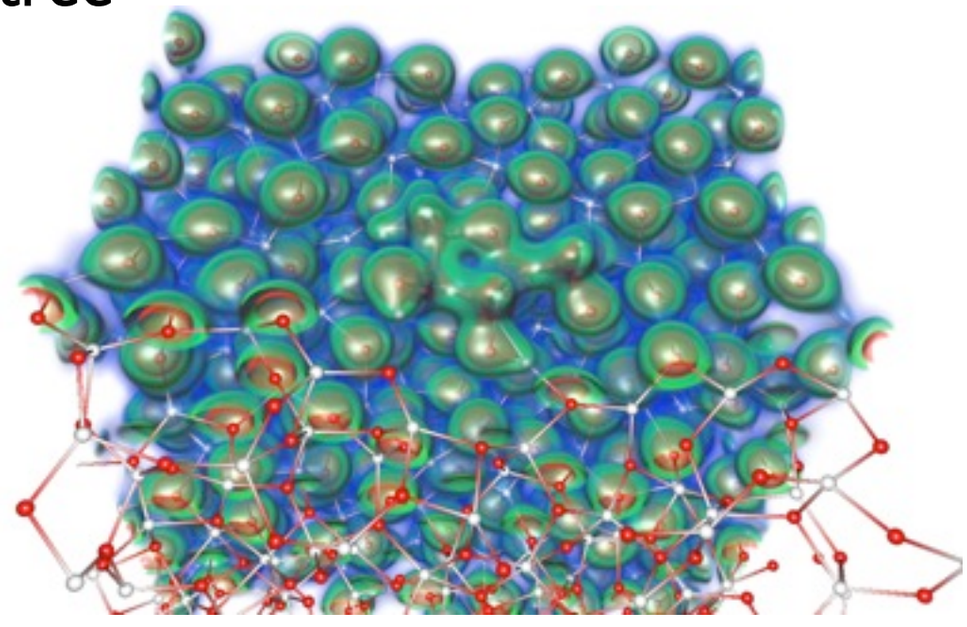


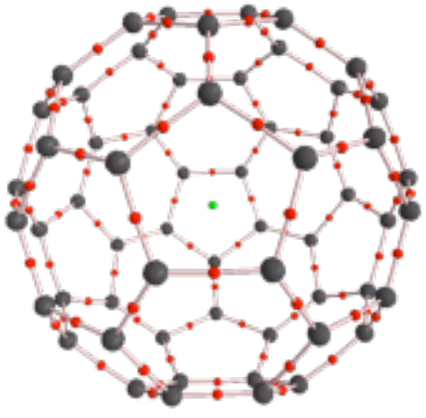
Future challenges

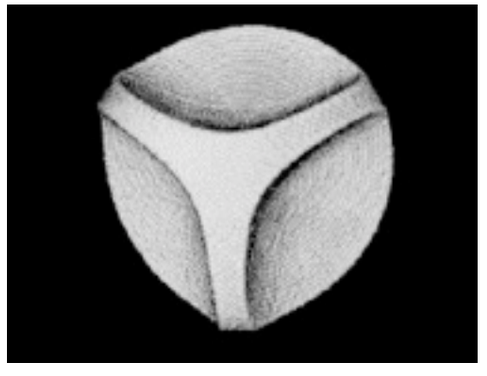
Selection / focus

- Which regions of the SCF *really* correspond to which atoms/molecules?
  - Theory of Atoms in Molecules (Bader)
  - Morse-Smale complex
- Determine chemical bonds from SCF?
  - Contour tree





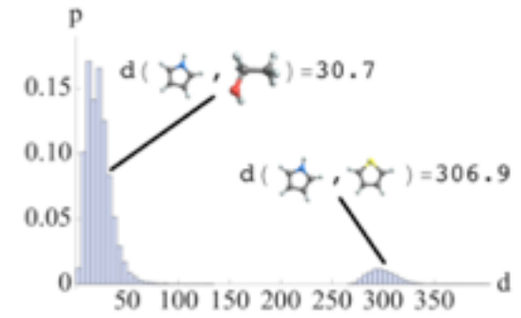


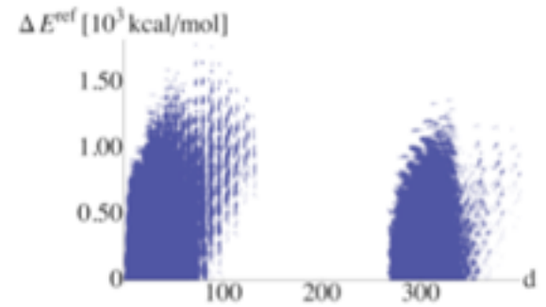


Bader, http://www.aim2000.de/

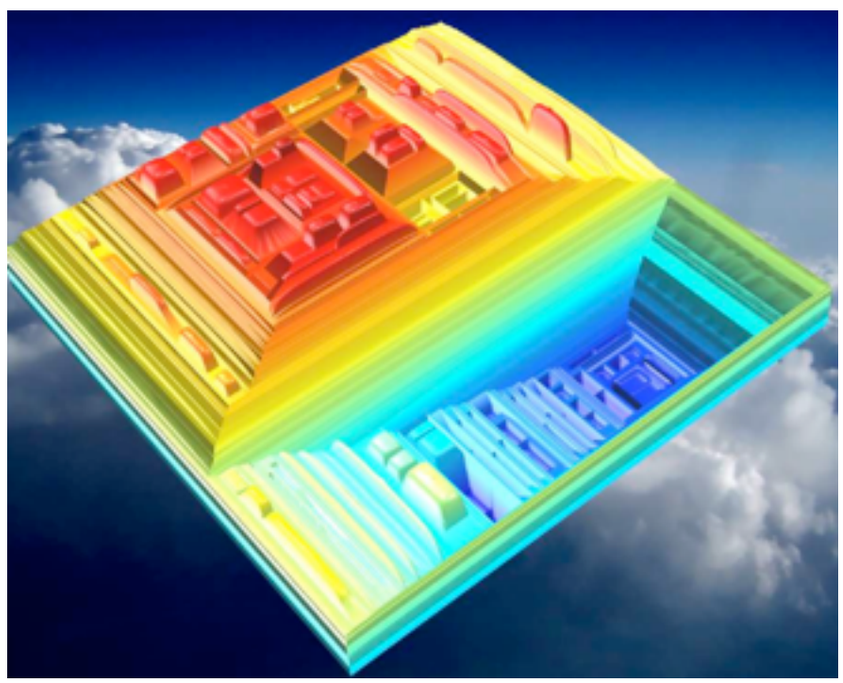
Compound spaces

- Use ML to optimize over search space
  - use DFT computations as training sequence
  - alternative to ensembles
- Visualization:
  - understanding ML metric space
  - reconstructing approximate SCF, geometry
  - vis as coanalysis alongside ML



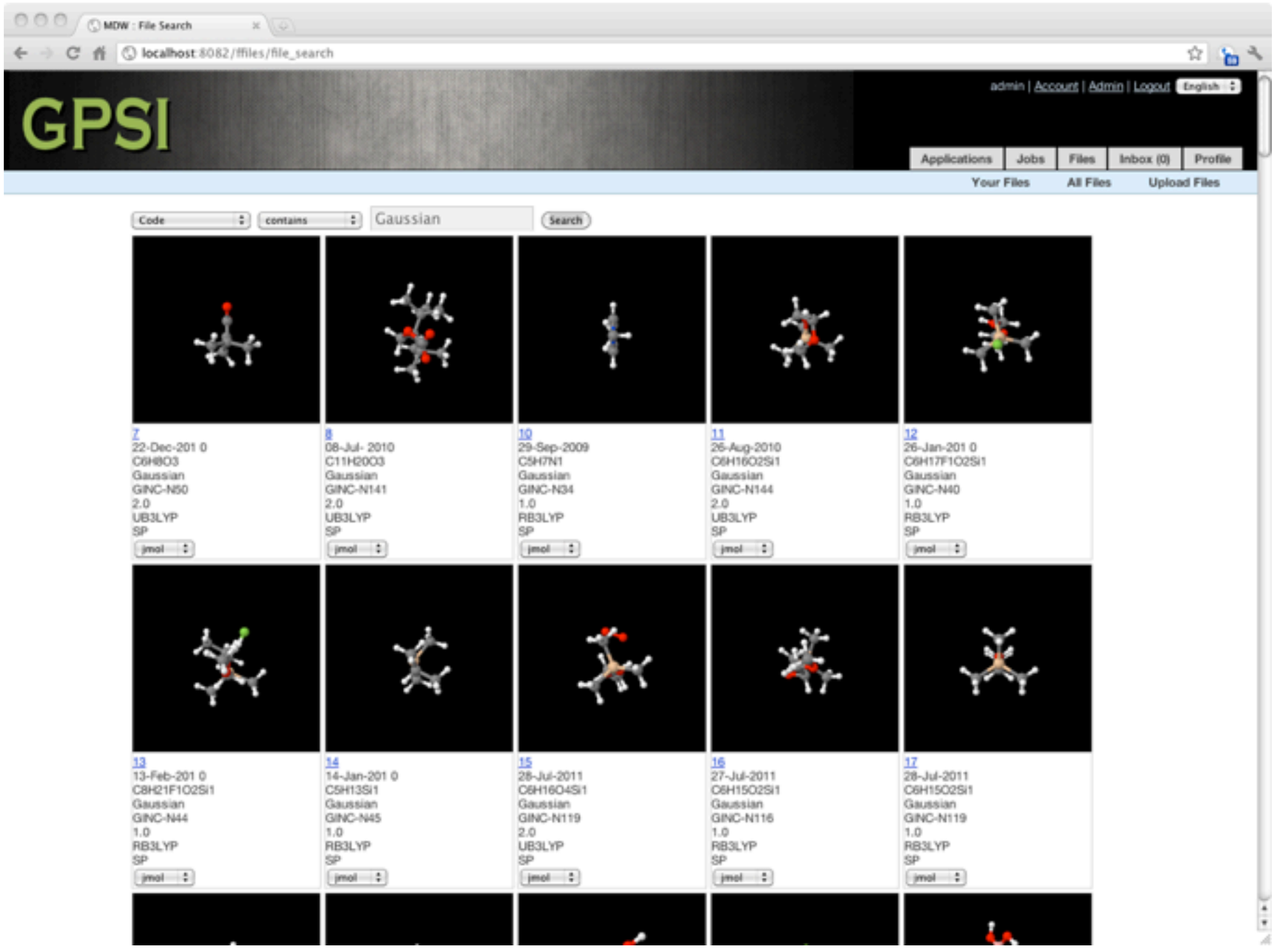


Rupp, et al. 2011 "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning"



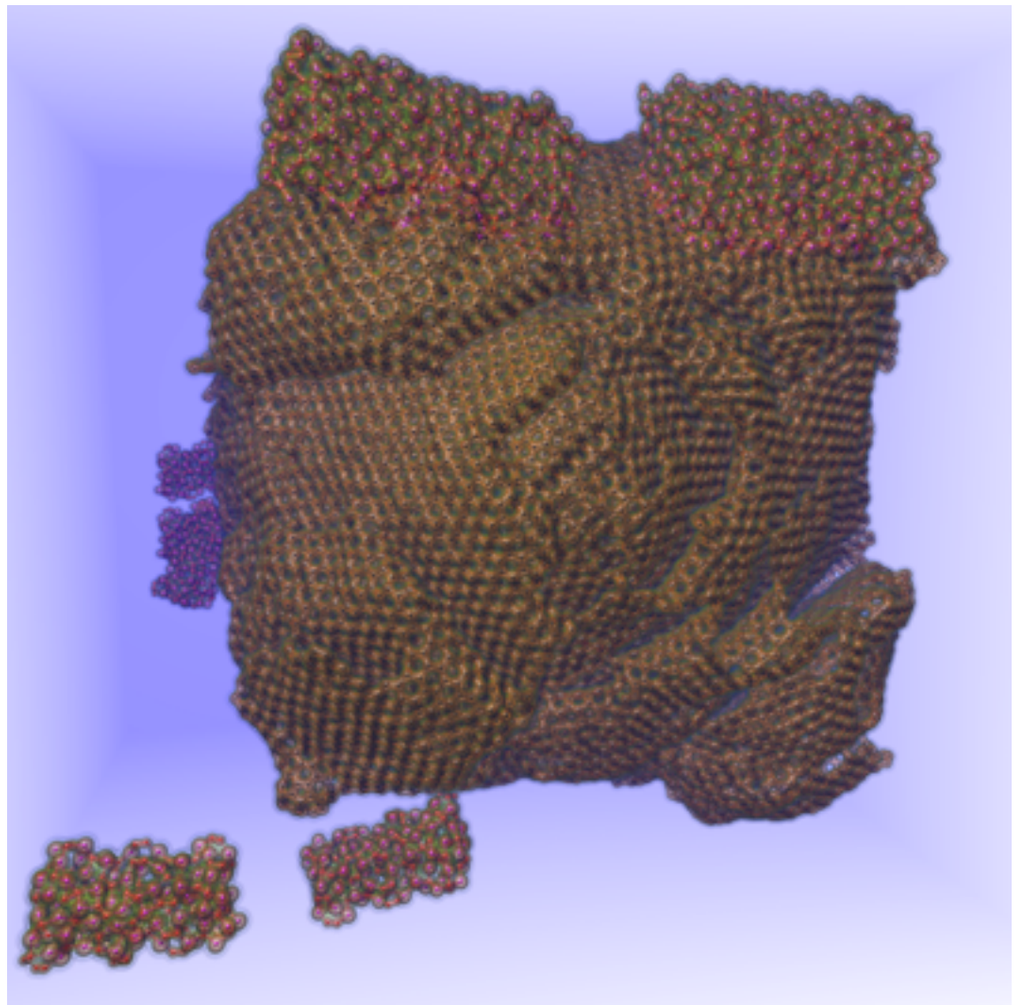
Collections of topological landscapes (Harvey and Wang, Eurovis10)

web/database vis



Conclusions

- Volumetric methods let us do molecular vis, analysis **based on first-principles**
- **High-dimension problem space, not Cartesian resolution**, is the biggest computational challenge
  - single runs, ensembles, and compound spaces
  - multi-molecule simulations
  - Postprocess / co-process is fine (currently)
  - encourage larger runs, improve IO
  - keep vis/scientists in tight loop



Thank you!

- Ultrascale vis workshop, SC 2011

- Mike Papka, Mark Hereld, Venkat Vishwanath, Joe Insley, Tom Uram, Eric Olson, Randy Hudson, Tom Peterka

- Materials/chemistry/computation collaborators at ANL:

  Bin Liu, Maria Chan, Jeff Greeley (Center for Nanoscale Materials)

  KC Lau, Lei Cheng, Hakim Iddir, Glen Ferguson, Larry Curtiss
  (Materials Science Division)

  Aslihan Sumer, Julius Jellinek (Chemical Sciences and Engineering)

  Anatole von Lilienfeld, Nick Romero, Anour Benali, Alvaro Vasquez, Jeff Hammond (ALCF)

- Funding: Office of Advanced Scientific Computing Research, Office of Science, US Department of Energy under Contract DE-AC02-06CH11357.
  Computational Postdoc Fellowship at Argonne National Laboratory supported by the American Reinvestment and Recovery Act (ARRA)