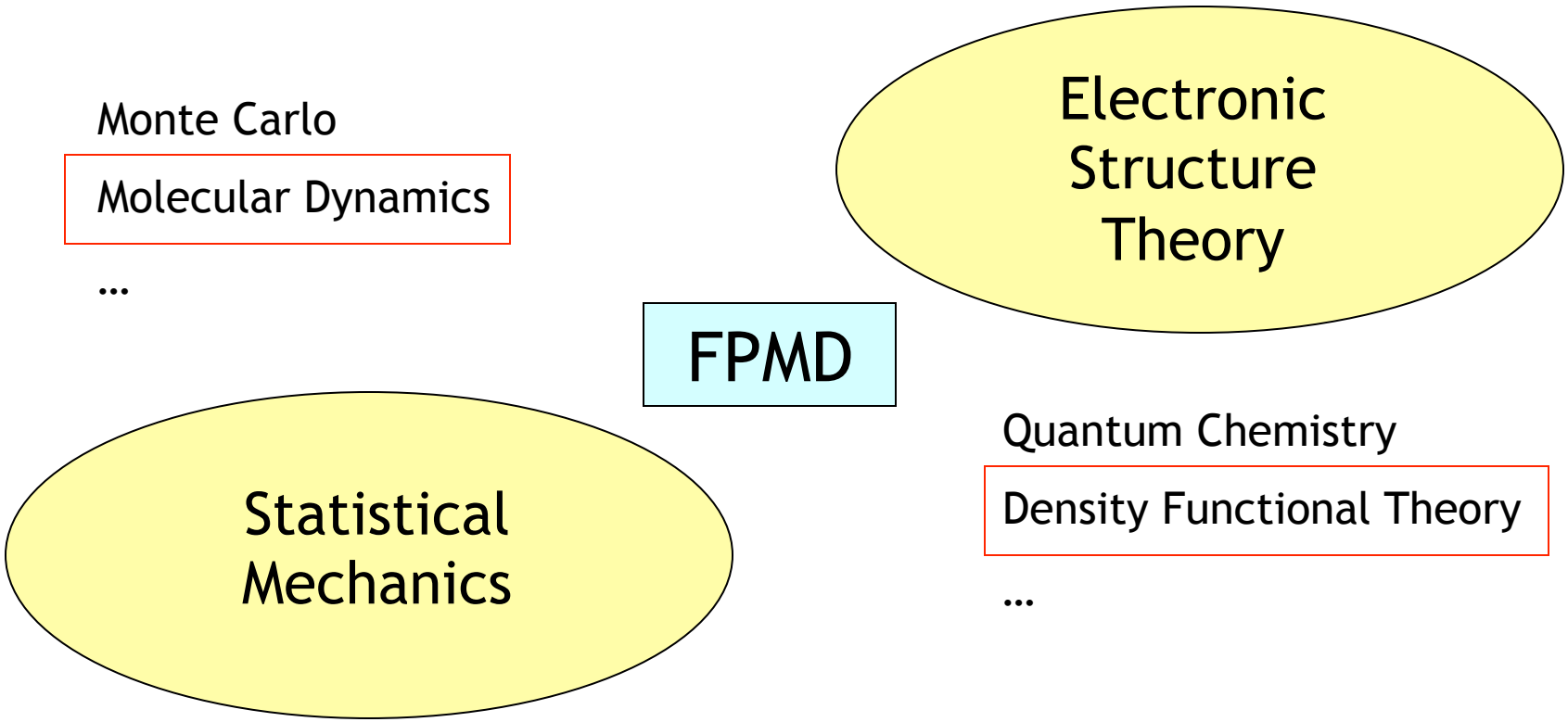
Compression Algorithms for Electronic Structure Computations

François Gygi
University of California, Davis
fgygi@ucdavis.edu
http://eslab.ucdavis.edu
http://www.quantum-simulation.org

IPAM Workshop, Computation meets Big Data, Feb 2, 2017

Supported by DOE BES DE-SC0008938 and MICCoM

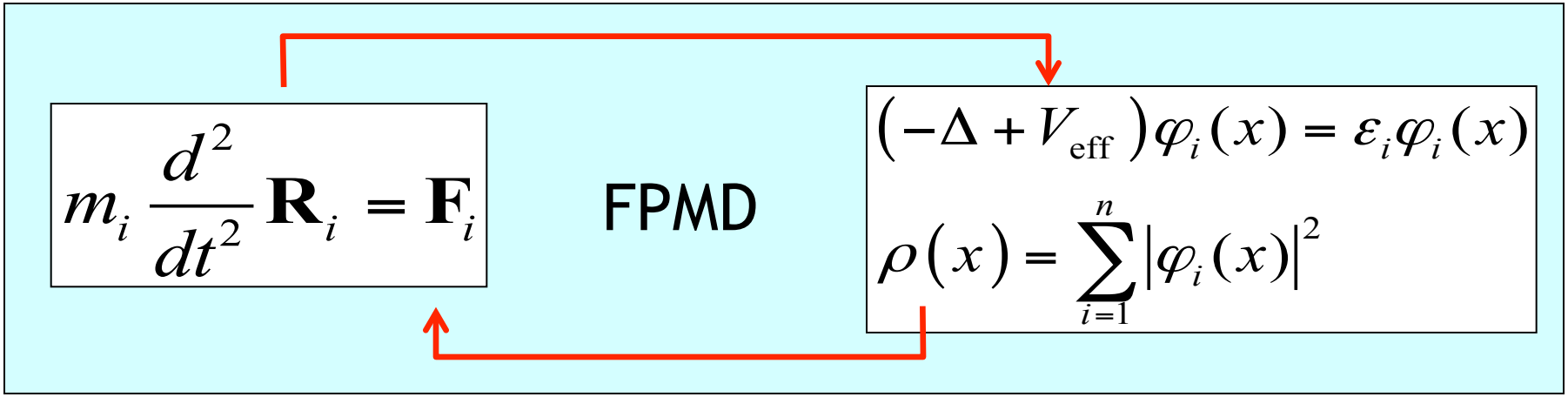
Context: First-Principles Molecular Dynamics



Context: First-Principles Molecular Dynamics

Molecular Dynamics

Density Functional Theory



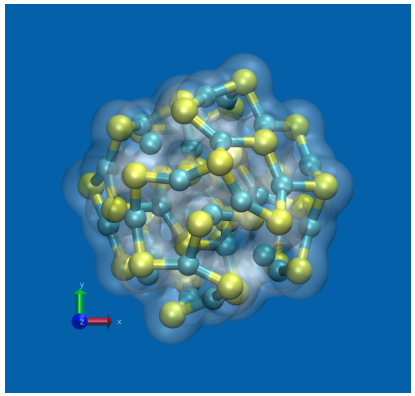
Newton equations

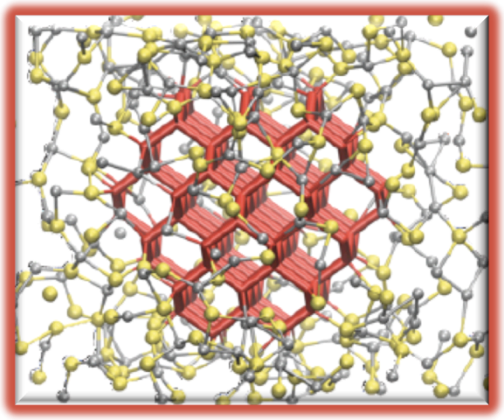
Kohn-Sham equations

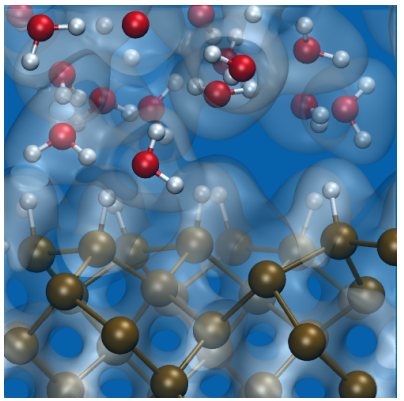
R. Car and M. Parrinello (1985)

Electronic Structure and dynamical properties of complex structures

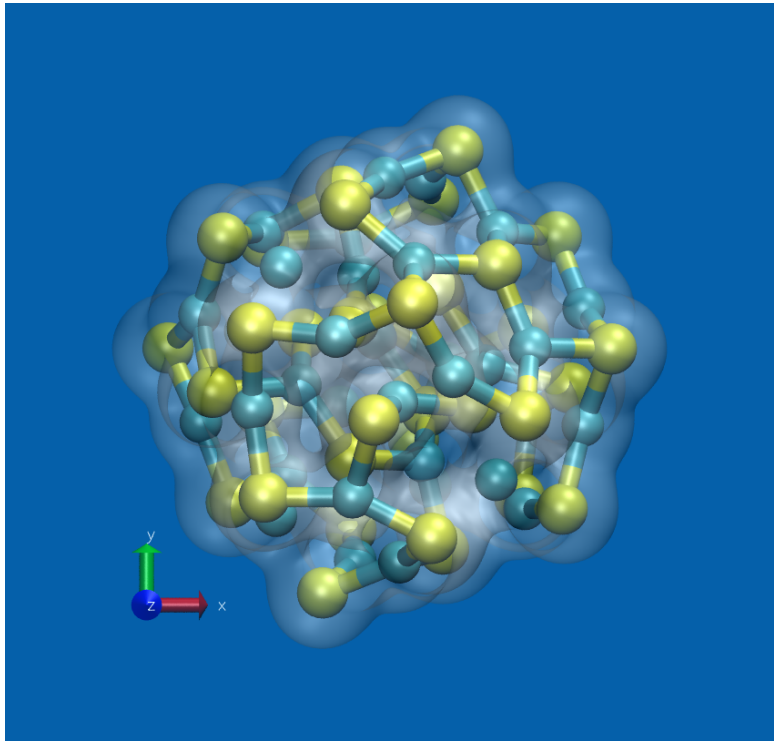
- Complex structures
  - Nanoparticles
  - Assemblies of nanoparticles
  - Embedded nanoparticles
  - Liquid/solid interfaces







Nanoparticles

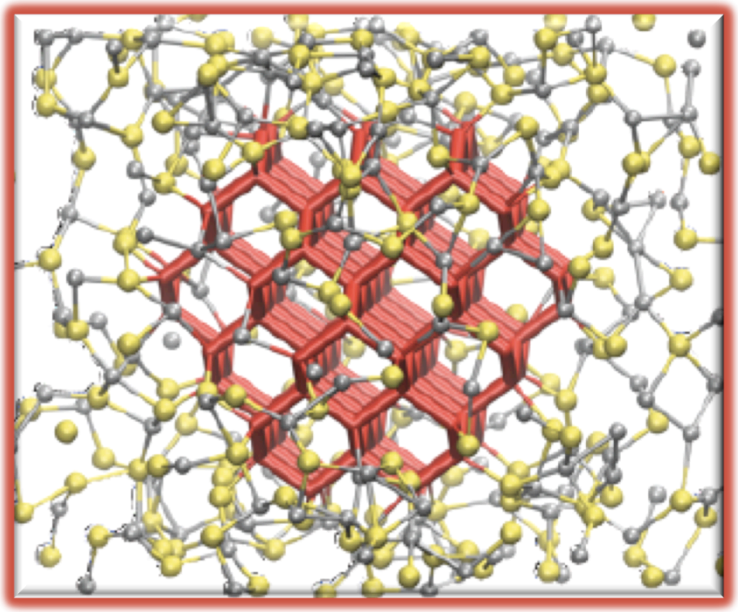


$Cd_{34}Se_{34}$

- Multiplicity of locally stable structures requires extensive sampling
- Electronic structure requires accurate methods (beyond DFT)
- Finite temperature properties require first-principles molecular dynamics

UCDAVIS
UNIVERSITY OF CALIFORNIA

Embedded nanoparticles, assemblies of nanoparticles

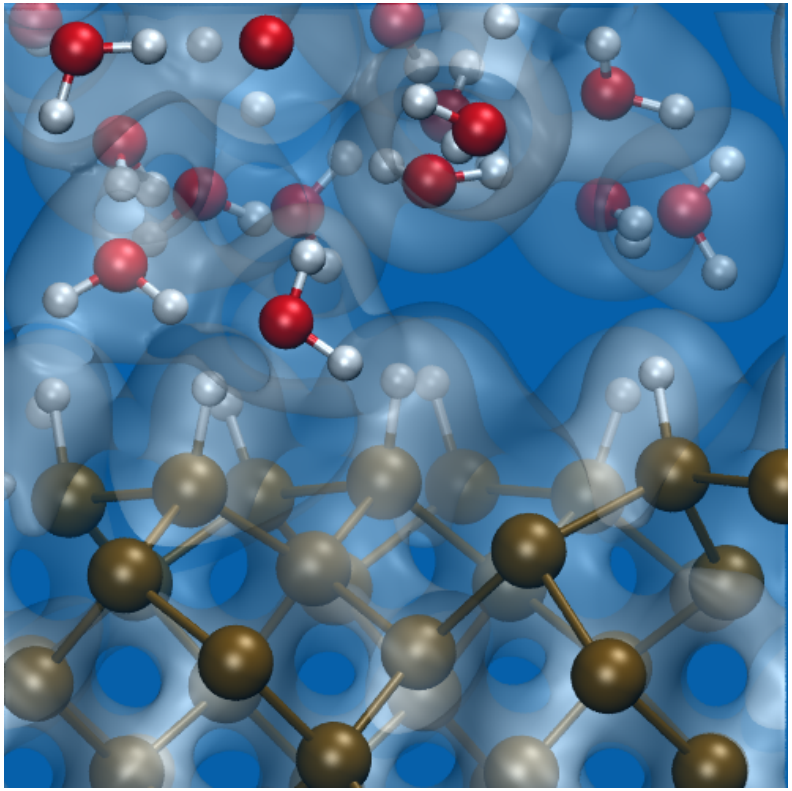


Si/ZnS

- Annealing of structures requires MD simulations
- Calculation of band gaps and band alignments requires accurate electronic structure (beyond DFT)

S. Wippermann, M. Vörös, A. Gali, F. Gygi, G. Zimanyi, G.Galli, Phys. Rev. Lett. **112**, 106801 (2014) .

Liquids and Liquid-Solid Interfaces



H_2O/Si(100)H

- Liquids require finite temperature simulations
  - ab initio MD
  - multiple samples/replica exchange simulations
- Electronic structure
  - band alignment
- Spectroscopy
  - requires calculation of IR and Raman spectra

UCDAVIS
UNIVERSITY OF CALIFORNIA

Challenges

- Multiple length and time scales
  - e.g. nanoparticle assembly process, non-equilibrium processes
- Finite temperature: MD/MC simulations
  - ab initio MD for large systems (>1000 atoms)
- Need for accurate electronic structure
  - hybrid DFT and/or GW/BSE level

Qbox code: DFT and hybrid DFT first-principles molecular dynamics

- http://qboxcode.org
- massively parallel first-principles MD
- C++/MPI/OpenMP
- DFT and hybrid DFT MD
- GPL license
- All algorithms discussed here are available in the Qbox code

Computation meets Big Data: Why data compression is necessary

- Storage of restart files in simulations
  - contain full information about wave functions
  - size is $O(N^2)$ for N atoms, reaches 100 GB-1 TB for large problems
- Acceleration of DFT and/or hybrid DFT calculations
  - Reduce cost (e.g. from $O(N^4)$ to $O(N^3)$)
  - Exploit locality of data on modern computer architectures

Desirable properties of compression schemes

- Accuracy control
  - ideally a single parameter controlling the accuracy
  - gradual reduction of the error to zero
- Efficiency
  - tradeoff between space efficiency and cost of compression algorithm

Three compression approaches

- Reduced resolution
  - Easy: in the Fourier basis: Reduce energy cutoff
  - Efficient: requires only 1 FFT per orbital
  - Problem: only moderate reduction is possible
- Compute Wannier functions
  - Efficient (if using the right algorithm)
  - Truncation procedure is ill-defined (no error control)
- Recursive Subspace Bisection (this work)
  - Efficient (as fast as Wannier function calculation)
  - Controllable error and systematic truncation scheme

Wannier functions

- Existence of exponentially localized Wannier functions
  - Kohn (1959), des Cloiseaux (1964), Nenciu (1983), Helffer *et al* (1989), Brouder *et al* (2007), Panati (2007, 2013)
- Computation of Wannier functions: find an orthogonal transformation among orbitals that minimizes the spread $\sigma_X^2 = \left\langle \left(x - \langle x \rangle \right)^2 \right\rangle$
- Optimization problem (with local minima)
  - Marzari, Vanderbilt (1997) Use conjugate gradients, etc.
- Approximate simultaneous diagonalization problem
  - F.G, Fattebert, Schwegler (2003)

F.G., J.L.Fattebert, E.Schwegler, Comput.. Phys. Comm. **155**, 1 (2003)

J.F.Cardoso and A. Souloumiac, SIAM J. Mat. Anal. Appl. **17**, 161 (1996).

Spread Functionals

- Spread of an operator $\hat{A}$ (single orbital)

$$\sigma^2_{\hat{A}}(\phi) = \left\langle \phi \left| \left(\hat{A} - \left\langle \phi \left| \hat{A} \right| \phi \right\rangle \right)^2 \right| \phi \right\rangle$$

$$= \left\langle \phi \left| \hat{A}^2 \right| \phi \right\rangle - \left\langle \phi \left| \hat{A} \right| \phi \right\rangle^2$$

- Spread of a set of orbitals

$$\sigma^2_{\hat{A}}(\{\phi_i\}) = \sum_i \sigma^2_{\hat{A}}(\phi_i)$$

Spread Functionals

- The spread is *not* invariant under orthogonal transformations among orbitals

$$\psi_i = \sum_j x_{ij}\phi_j \qquad X \in \mathbb{R}^{n\times n} \text{ orthogonal}$$

$$\sigma^2_{\hat{A}}\left(\{\psi_i\}\right) \neq \sigma^2_{\hat{A}}\left(\{\phi_i\}\right)$$

- There exists a matrix *X* that minimizes the spread

Spread Functionals

- Let

$$A, B \in \mathbb{R}^{n\times n} \qquad a_{ij} = \left\langle i \left| \hat{A} \right| j \right\rangle \qquad b_{ij} = \left\langle i \left| \hat{A}^2 \right| j \right\rangle$$

$$\sigma^2_{\hat{A}}\left(\{\psi_i\}\right) = \mathrm{tr}\left(X^T B X\right) - \sum_{i=1}^{n}\left(X^T A X\right)^2_{ii}$$

- Minimize the spread = maximize $\sum_{i=1}^{n}\left(X^T A X\right)^2_{ii}$
  = diagonalize *A*

Spread Functionals

- Case of multiple operators

$$\text{operators} \quad \hat{A}^{(k)} \; k = 1,\ldots,m$$

$$\text{matrices} \quad A^{(k)} \; k = 1,\ldots,m$$

$$\sigma^2_{\hat{A}}\left(\{\psi_i\}\right) = \sum_i \sum_k \sigma^2_{\hat{A}^{(k)}}\left(\psi_i\right)$$

- Minimize the spread = maximize $\sum_{i=1}^{n} \sum_k \left(X^T A^{(k)} X \right)^2_{ii}$
  = joint approximate diagonalization of the matrices $A^{(k)}$

Spread Functionals

- Example of multiple operators

$$\hat{A}^{(1)} = \hat{X} \qquad \left(\hat{X}\varphi\right)(x,y,z) \equiv x\,\varphi(x,y,z)$$

$$\hat{A}^{(2)} = \hat{Y} \qquad \left(\hat{Y}\varphi\right)(x,y,z) \equiv y\,\varphi(x,y,z)$$

$$\hat{A}^{(3)} = \hat{Z} \qquad \left(\hat{Z}\varphi\right)(x,y,z) \equiv z\,\varphi(x,y,z)$$

- The matrices $A^{(k)}$ do not necessarily commute, even if the operators $\hat{A}^{(k)}$ do commute

Calculation of Wannier functions

- In periodic systems

$$\hat{A}^{(1)} = \hat{C}_x \equiv \cos\frac{2\pi}{L_x}\hat{x} \qquad \hat{A}^{(2)} = \hat{S}_x \equiv \sin\frac{2\pi}{L_x}\hat{x}$$

$$\hat{A}^{(3)} = \hat{C}_y \equiv \cos\frac{2\pi}{L_y}\hat{y} \qquad \hat{A}^{(4)} = \hat{S}_y \equiv \sin\frac{2\pi}{L_y}\hat{y}$$

$$\hat{A}^{(5)} = \hat{C}_z \equiv \cos\frac{2\pi}{L_z}\hat{z} \qquad \hat{A}^{(6)} = \hat{S}_z \equiv \sin\frac{2\pi}{L_z}\hat{z}$$

Calculation of Wannier functions

- The spread is minimized by simultaneous diagonalization of the matrices $C_x, S_x, C_y, S_y, C_z, S_z$
- Positions of the center of mass of the localized solutions ("Wannier centers")

$$\tau_i = \begin{pmatrix} L_x \dfrac{\theta_i^x}{2\pi} \\ L_y \dfrac{\theta_i^y}{2\pi} \\ L_z \dfrac{\theta_i^z}{2\pi} \end{pmatrix} \qquad \theta_i^x = \arctan \frac{(S_x)_{ii}}{(C_x)_{ii}}$$

- Spreads

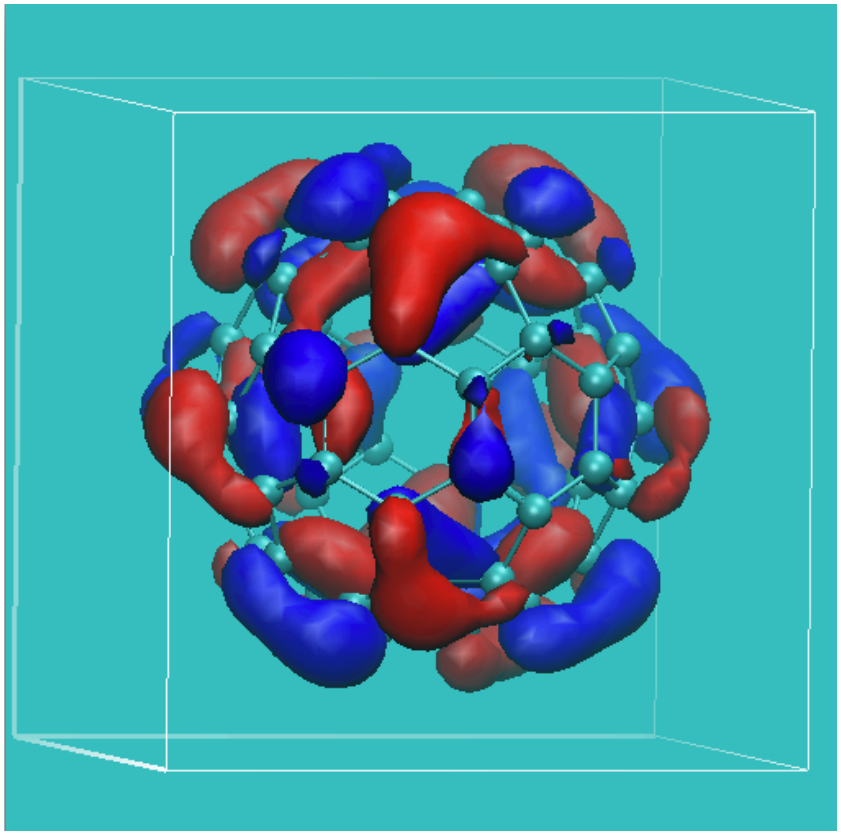
$$\left(\sigma_i^2\right)_x = L_x^2 \left(1 - (C_x)_{ii}^2 - (S_x)_{ii}^2\right)$$

F.G., J.L.Fattebert, E.Schwegler, Comput.. Phys. Comm. **155**, 1 (2003)

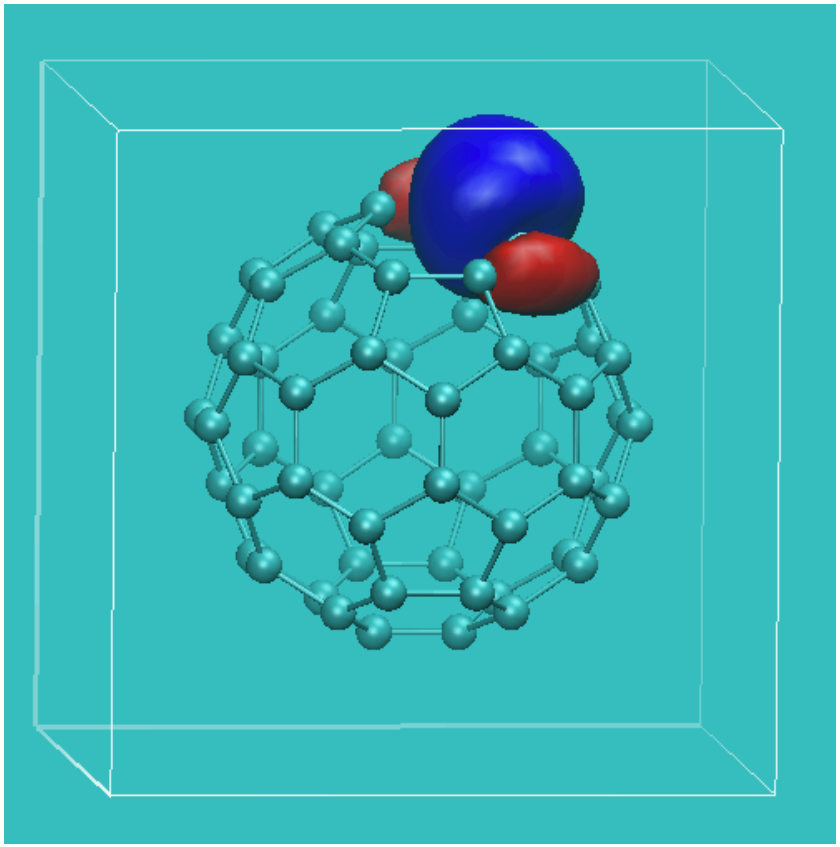
J.F.Cardoso and A. Souloumiac, SIAM J. Mat. Anal. Appl. **17**, 161 (1996).

Maximally localized Wannier functions

Extended orbital



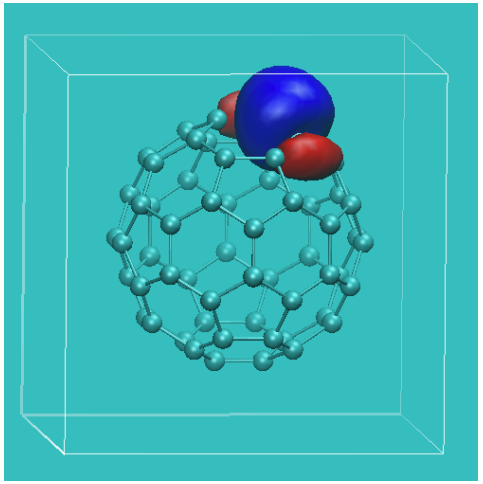
Wannier function



UCDAVIS
UNIVERSITY OF CALIFORNIA

Truncation of Wannier functions

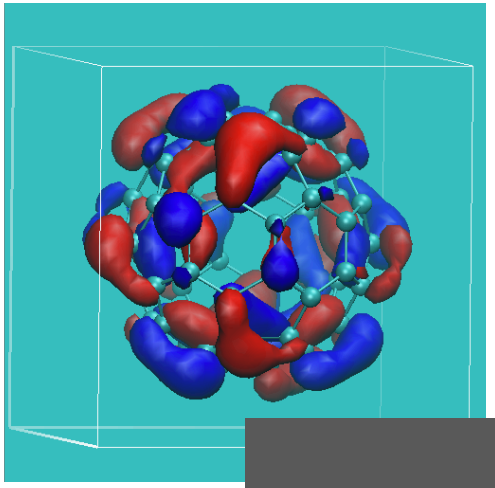
- Wannier functions (WFs) can be truncated in real space
  - truncate orbital to zero below a given threshold
  - truncate orbital to zero outside of a given radius



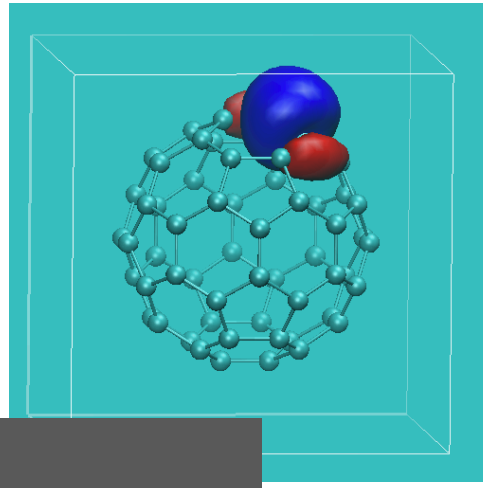
- WFs can have variable localization properties

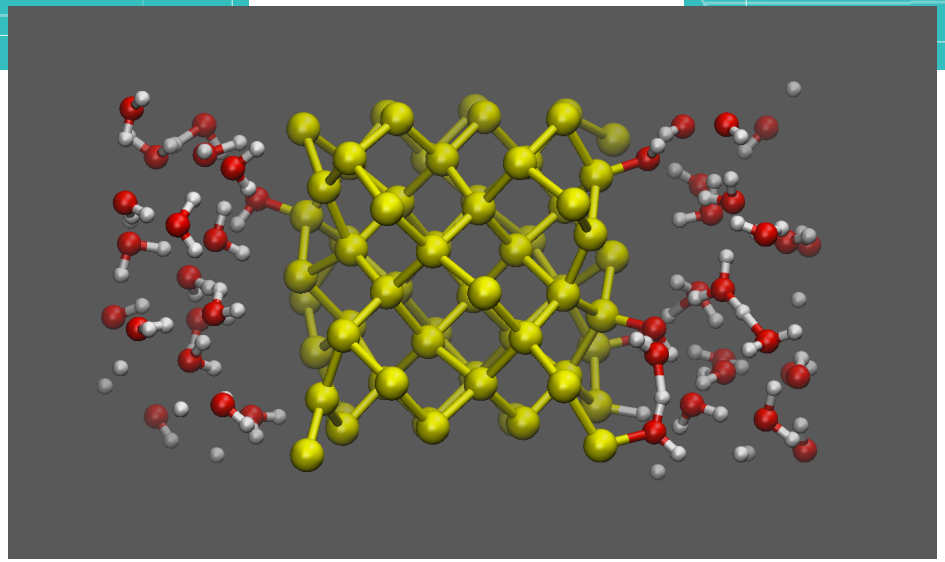
Maximally localized Wannier functions

Extended orbital



Wannier function

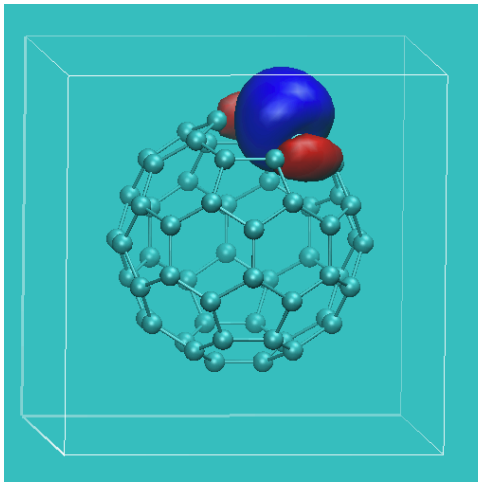




UCDAVIS
UNIVERSITY OF CALIFORNIA

Truncation of Wannier functions

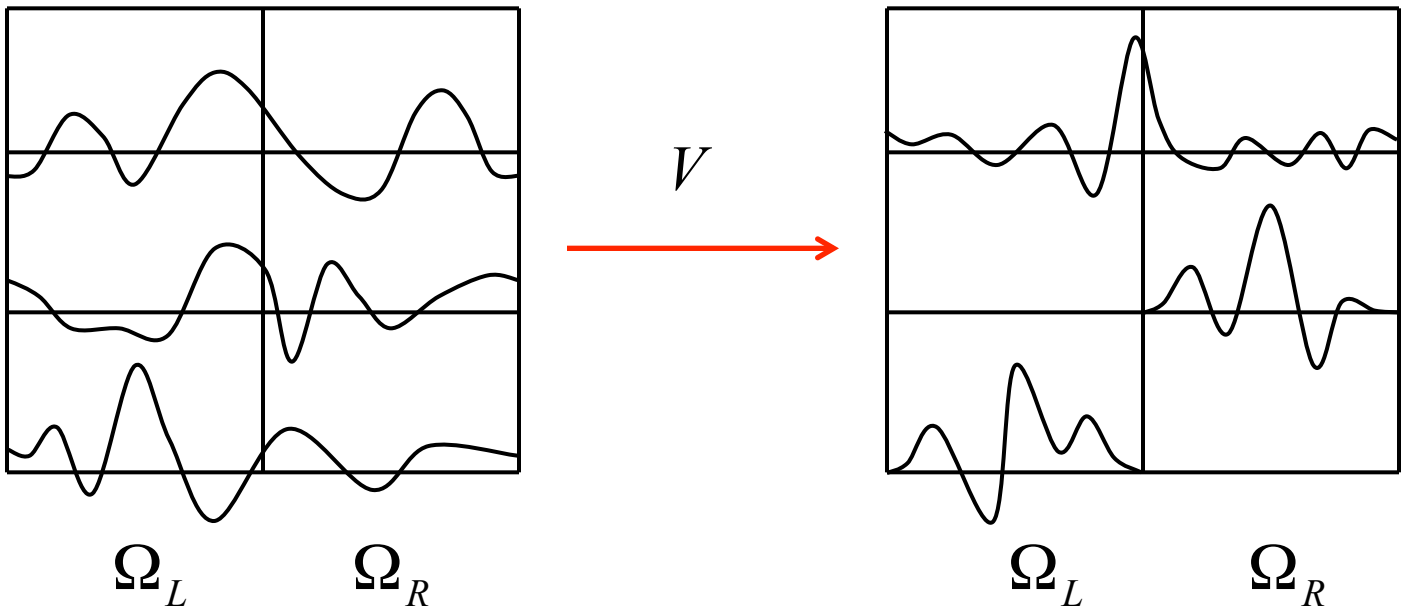
- Wannier functions (WFs) can be truncated in real space
  - truncate orbital to zero below a given threshold
  - truncate orbital to zero outside of a given radius



- WFs can have variable localization properties
- Are there other ways to localize orbitals?

Recursive subspace bisection

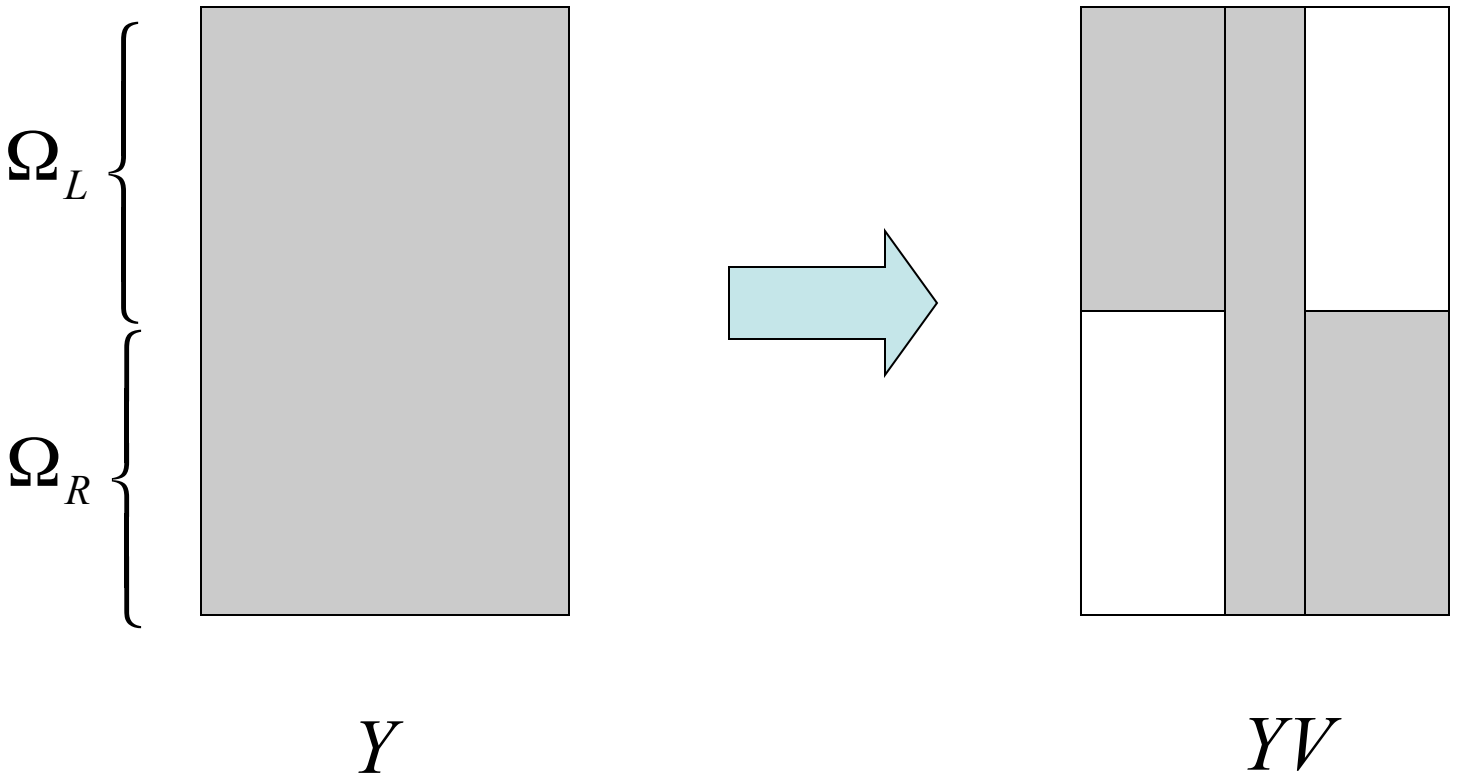
- Localize *some* orbitals on domains of decreasing size
  - divide the simulation domain into two subdomains
  - localize orbitals on 1) Ω_L 2) Ω_R or 3) keep extended on $\Omega_L \cup \Omega_R$
  - apply recursively to smaller domains



F.G. Phys. Rev. Lett. **102**, 166406 (2009)

Subspace Bisection

$$Y = [\phi_1 \ldots \phi_n]$$



The CS decomposition

- A matrix *Y* having orthogonal columns, can be decomposed as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U_1 \Sigma_1 V^T \\ U_2 \Sigma_2 V^T \end{pmatrix}$$

where U_1, U_2, V are orthogonal matrices,

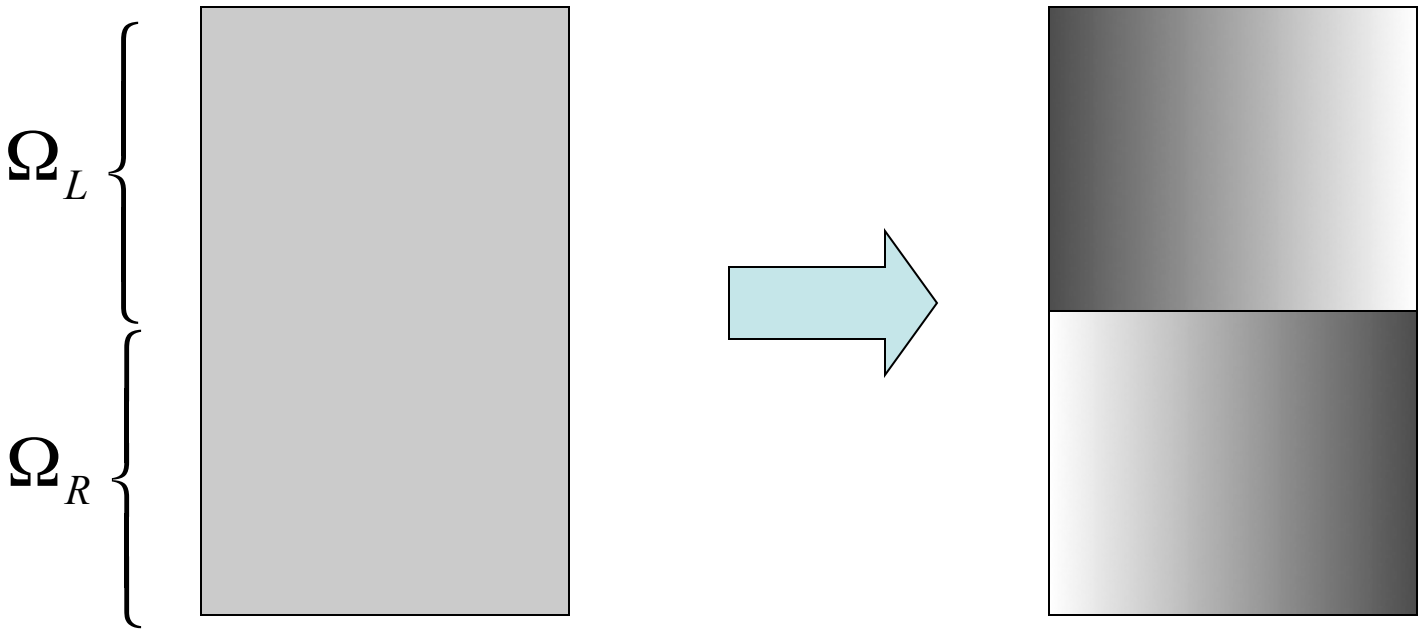
$$\Sigma_1 = \begin{pmatrix} C \\ 0 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} S \\ 0 \end{pmatrix}$$

$$C = \mathrm{diag}(c_1, \ldots, c_n) \quad S = \mathrm{diag}(s_1, \ldots, s_n)$$

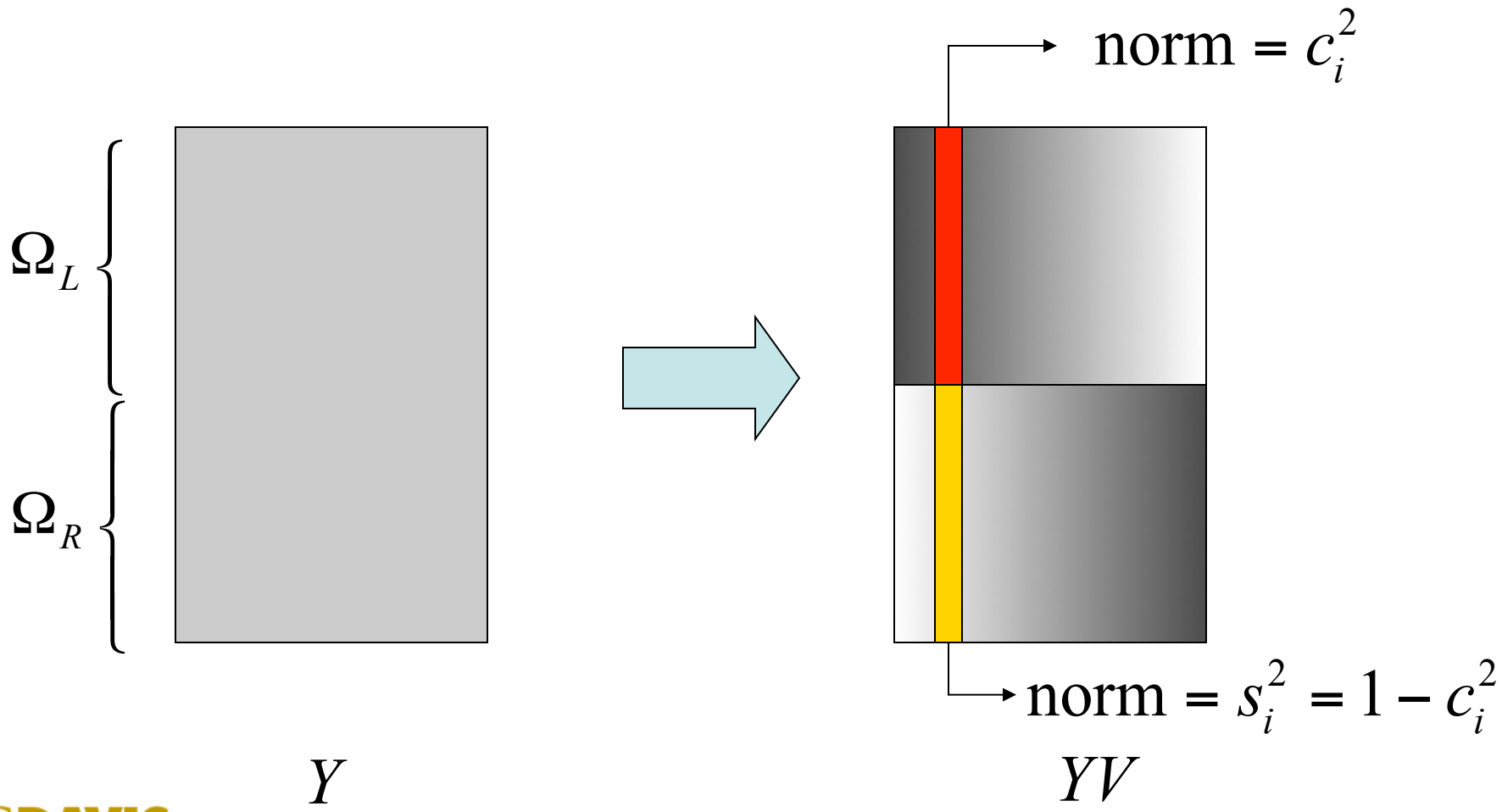
$$c_i^2 + s_i^2 = 1$$

Stewart (1982)

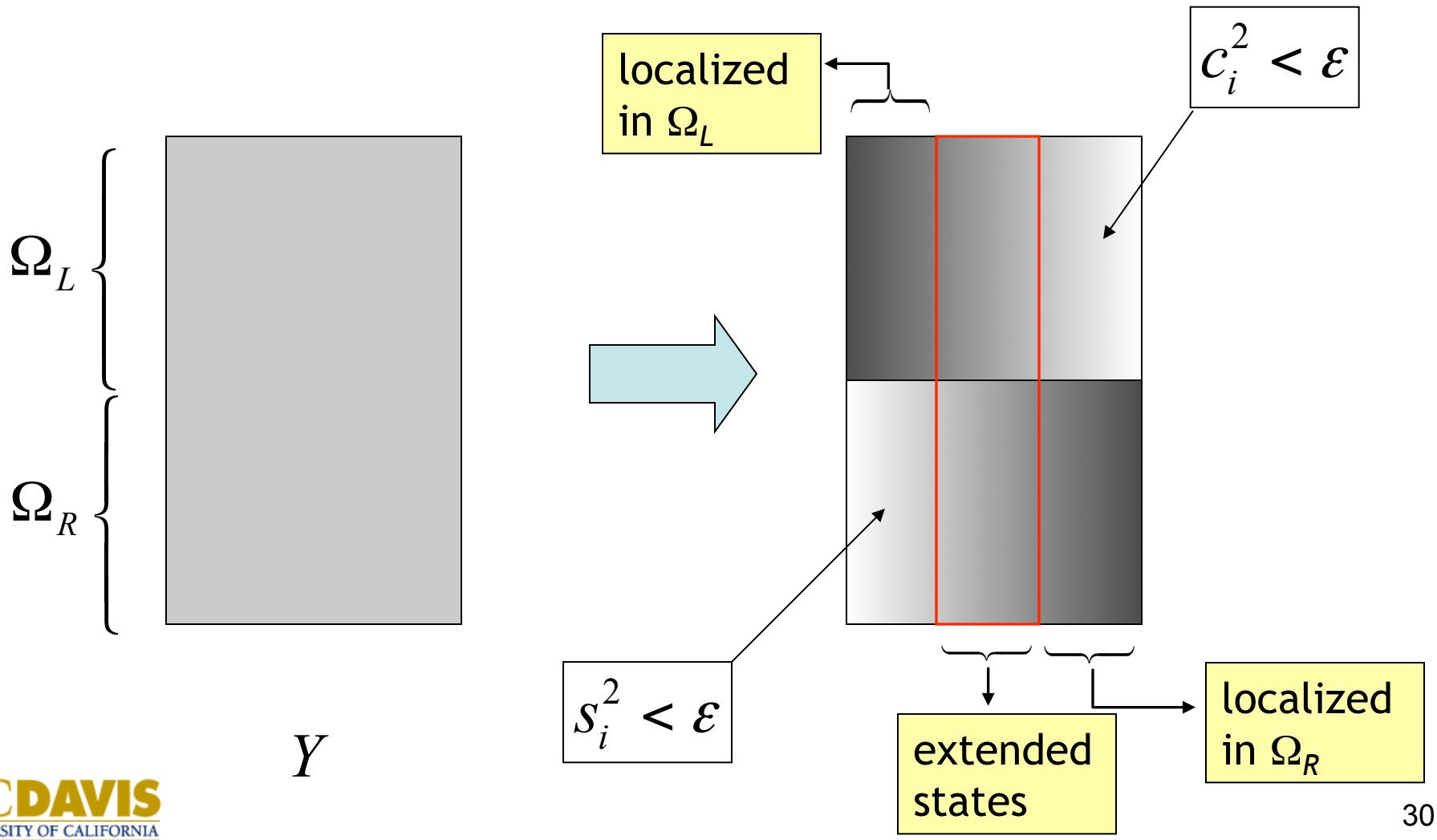
CS Decomposition



CS Decomposition



CS Decomposition



Subspace Bisection Algorithm

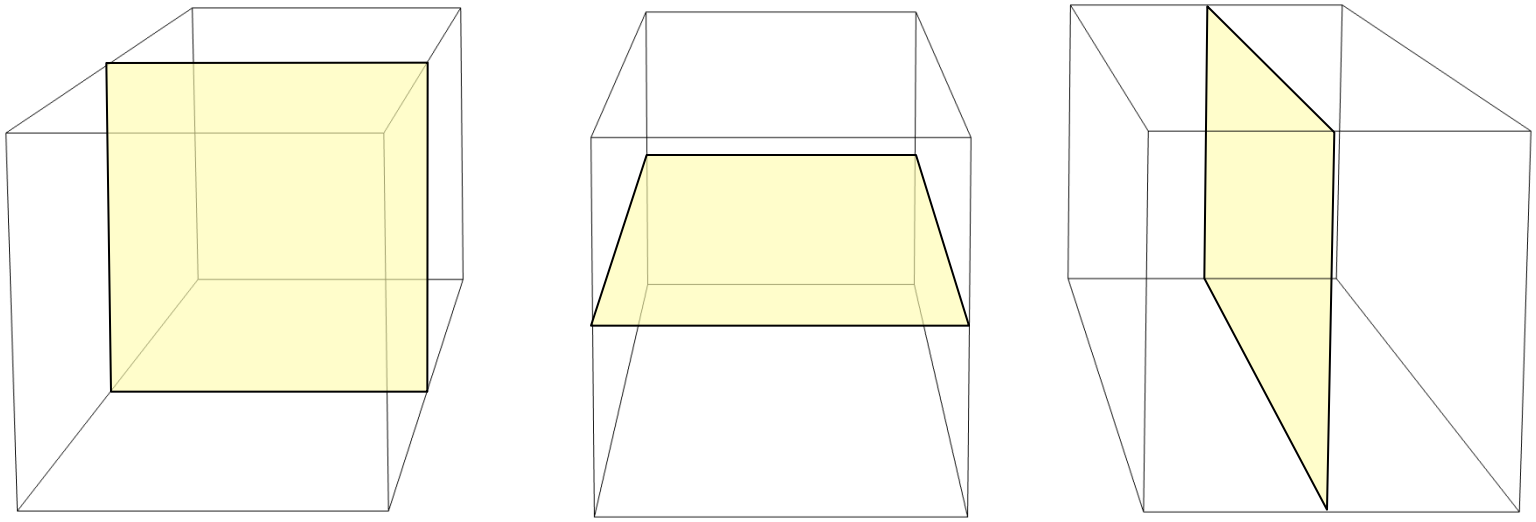
1. Choose the acceptable 2-norm error ε
2. Perform a CS decomposition of the matrix Y
3. For each vector of YV:
   - if $c_i^2 < \varepsilon$ orbital localized in Ω_R (truncate in Ω_L)
     else if $s_i^2 < \varepsilon$ orbital localized in Ω_L (truncate in Ω_R)
     else orbital is extended

Ideal limit: 2-fold data reduction

Cost: The *CS* decomposition can be achieved by diagonalization of the matrix $Y_1^T Y_1$

Recursive Subspace Bisection

- Bisection is applied simultaneously in 3 directions

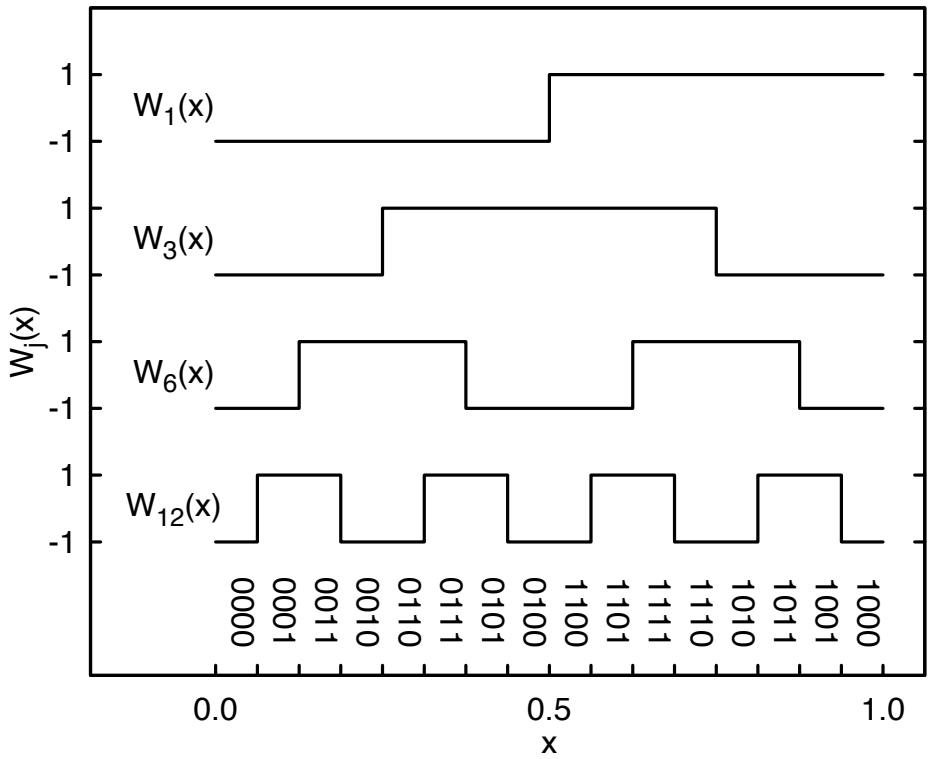


Implementation: simultaneous (approximate) diagonalization of symmetric matrices

F.G, J.-L.Fattebert, E.Schwegler, *Comp. Phys. Comm.* ***155***, *1 (2003)*.
J.F.Cardoso and A. Souloumiac, SIAM J. Mat. Anal. Appl. **17**, 161 (1996).

Recursive subspace bisection

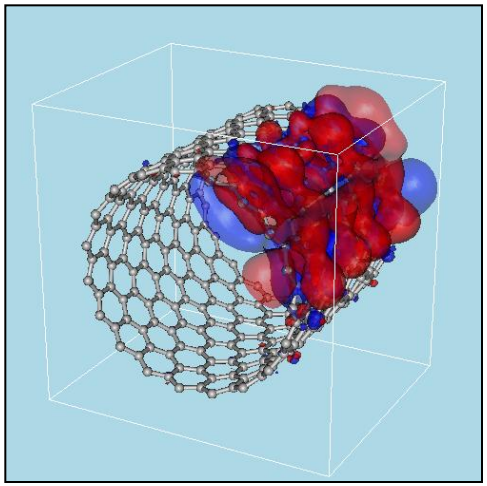
- Use multiple bisecting planes in each direction

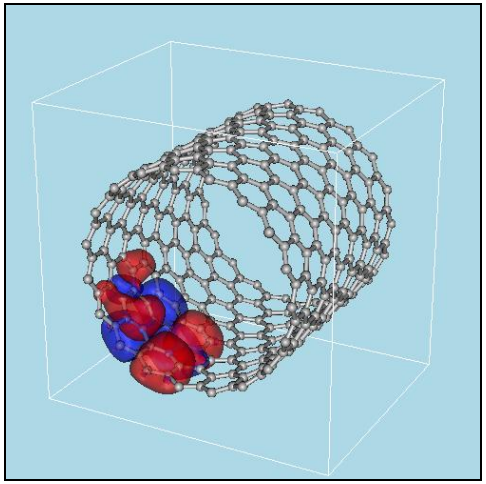


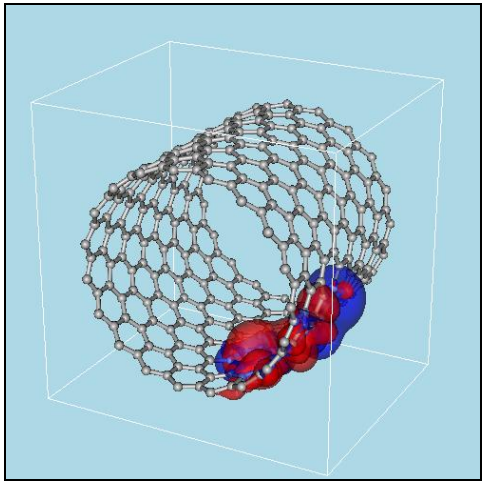
Optimal placement of bisecting planes is provided by *Walsh functions* $W_j(x)$

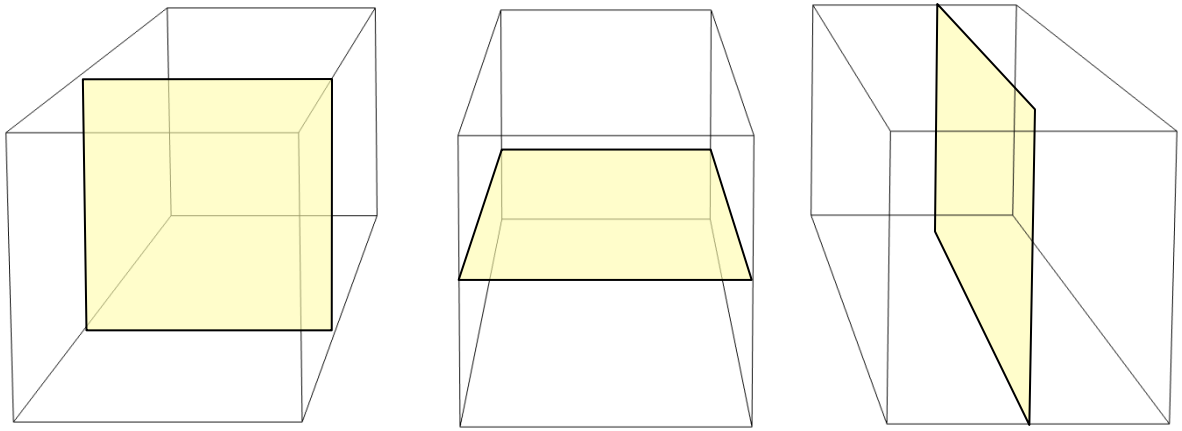
$j=1,3,6,12,\ldots$

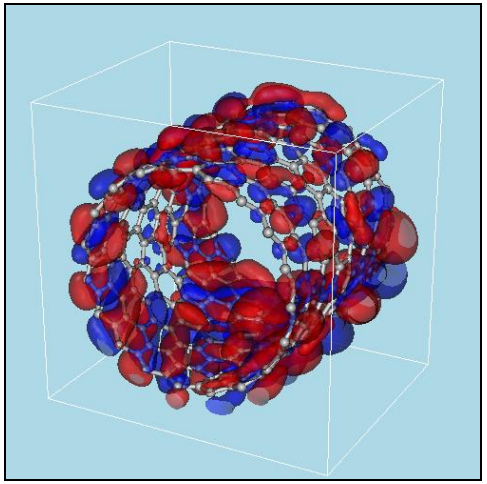
Recursive subspace bisection









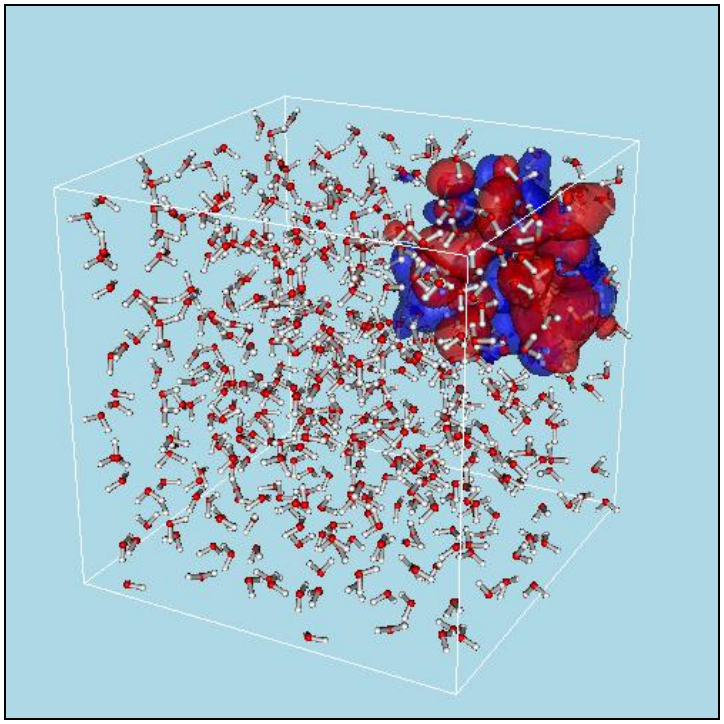


UCDAVIS
UNIVERSITY OF CALIFORNIA

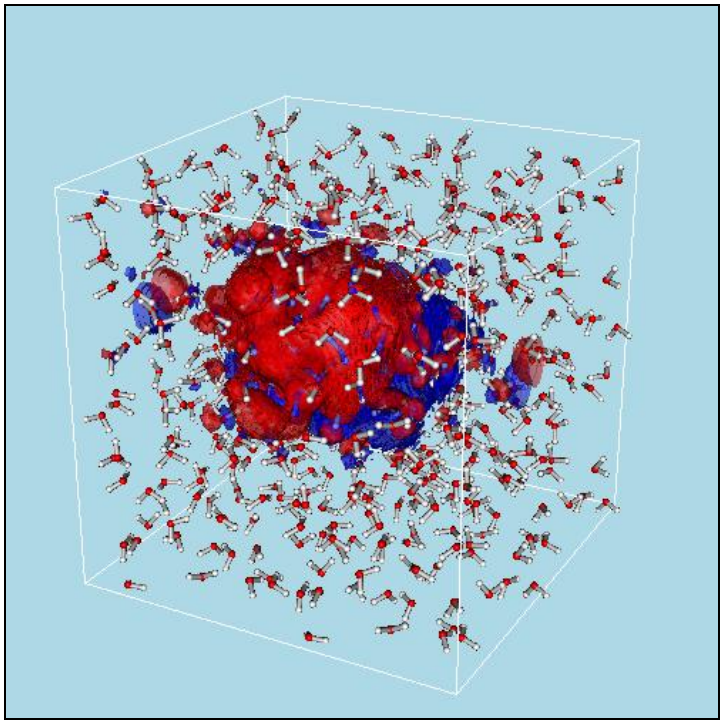
Phys. Rev. Lett. **102**, 166406 (2009).

$(H_2O)_{512}$ orbitals after bisection

localized



extended



UCDAVIS
UNIVERSITY OF CALIFORNIA

Compression ratio

- $(H_2O)_{512}$ $\varepsilon = 10^{-3}$

Data size reduction: 4.03

N_1	48	2%
$N_{1/2}$	276	14%
$N_{1/4}$	844	41%
$N_{1/8}$	880	43%

2048 orbitals

- (19x0) Carbon nanotube (304 atoms) $\varepsilon = 10^{-3}$

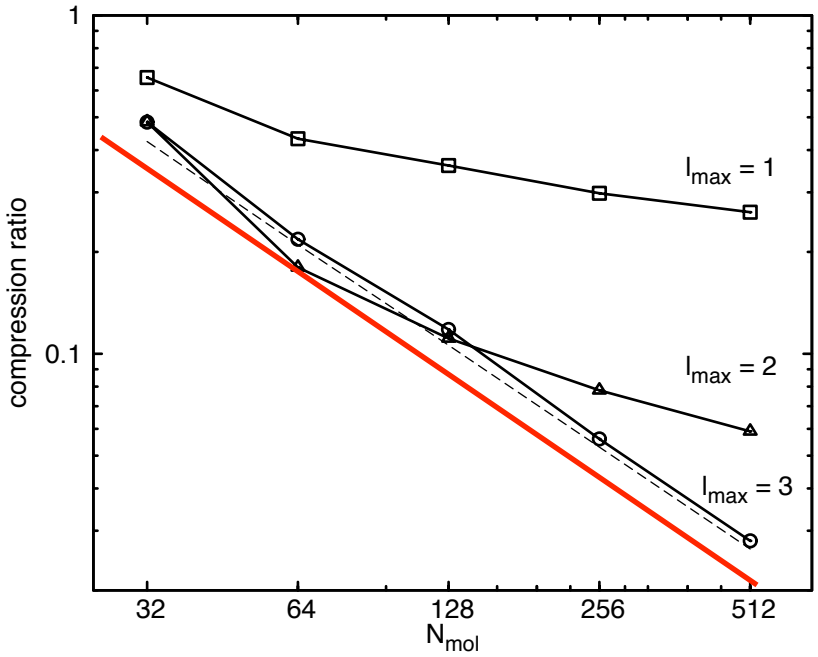
Data size reduction: 2.72

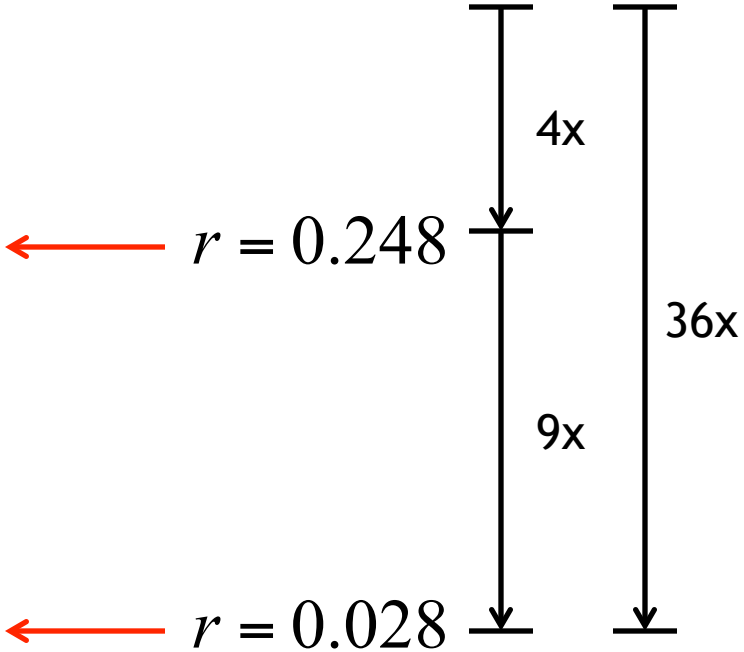
N_1	108	18%
$N_{1/2}$	80	13%
$N_{1/4}$	183	30%
$N_{1/8}$	237	39%

608 orbitals

Compression ratio using recursive bisection

- recursion on l_{max} levels, l_{max}=*1,2,3*



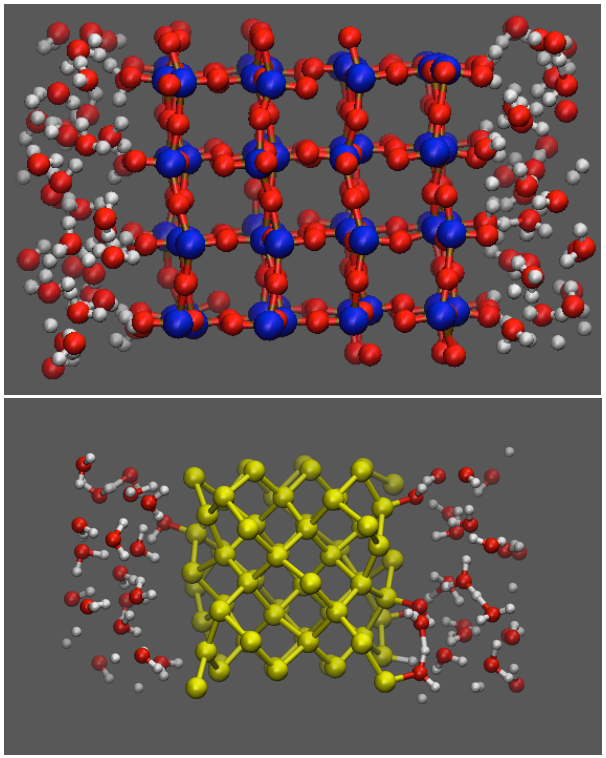


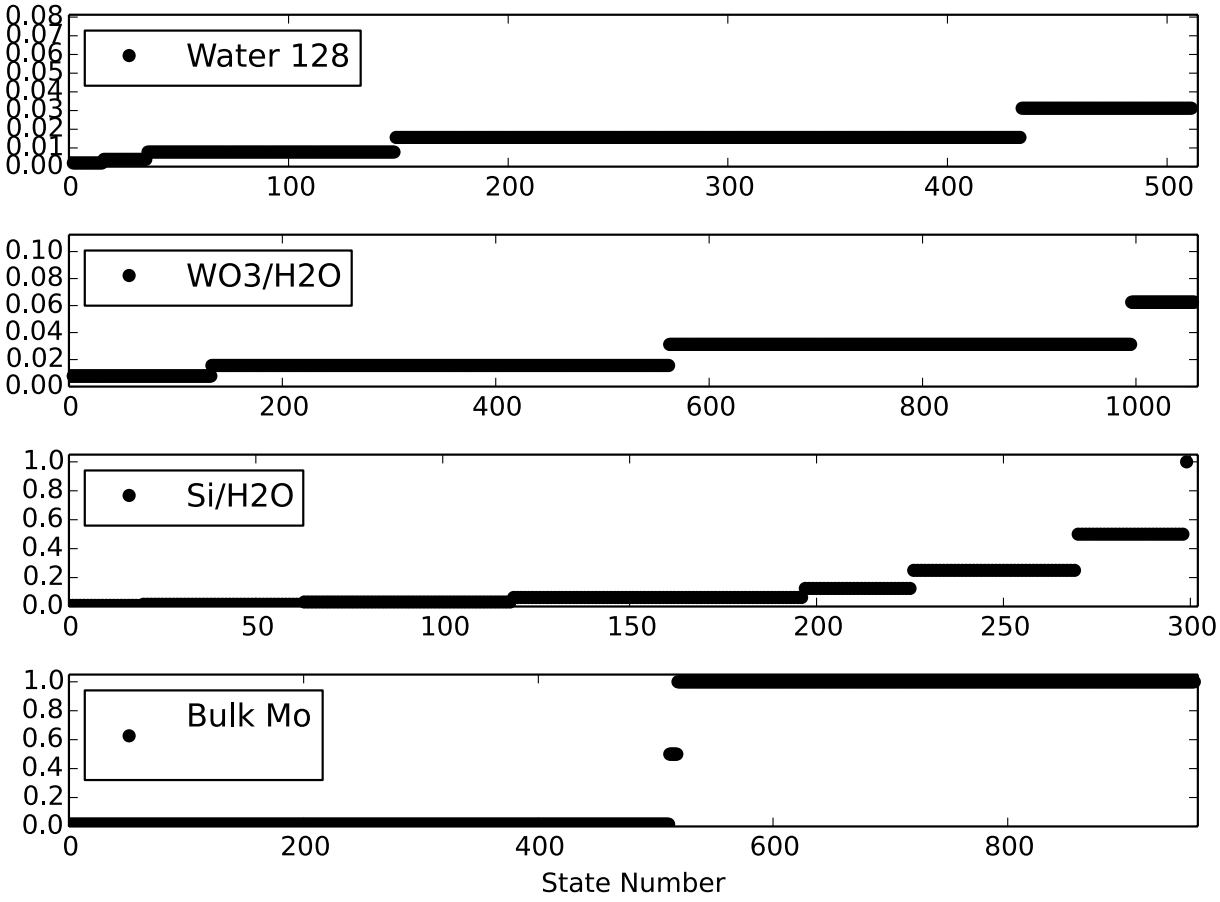
truncated MLWFs

1/(*N* log*N*) slope

F.G., Phys. Rev. Lett. **102**, 166406 (2009)

Localization of orbitals in inhomogeneous systems





W. Dawson and F.G, JCTC **11**, 4655 (2015)

Free electron gas: Distribution of CS singular values

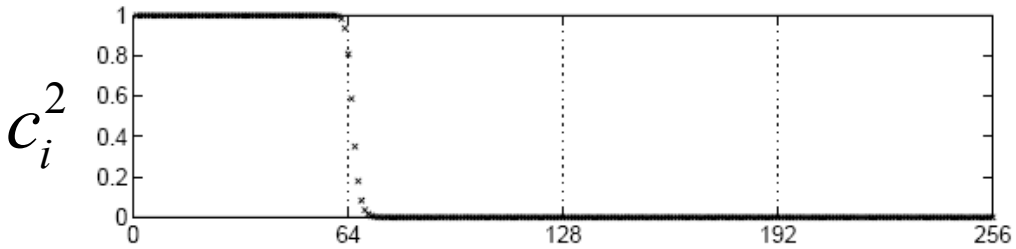


FIG. 2. *Singular values of* $F_{1024|4}$, *computed with* MATLAB.

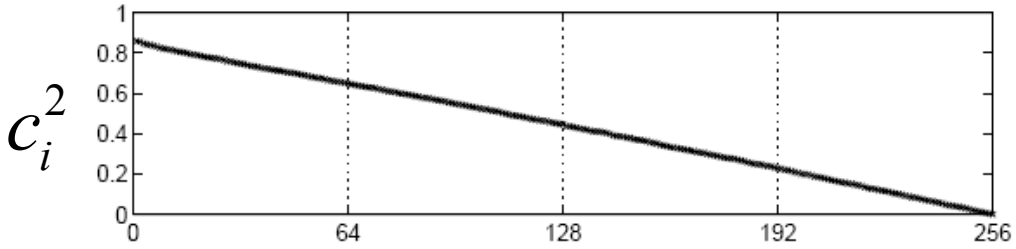


FIG. 3. *Singular values of a* 256×256 *section of a random* 1024×1024 *unitary matrix, computed with* MATLAB.

Y matrix = $F_{n|p}$ (section of the Fourier matrix, i.e. eigenvectors of the Laplacian)

Y matrix = random orthogonal matrix

A. Edelman *et al.* SIAM J. Sci. Comput. 20, 1094, (1999)

Hybrid density functionals

- Conventional density functionals are often insufficient to describe weak bonds (e.g. hydrogen bonds) or optical properties (band gap)
- *Hybrid density functionals* include a fraction of the Hartree-Fock exchange energy
- The Hartree-Fock exchange energy involves *N(N-1)/2 exchange integrals* (for all e^- pairs)

$$E_x = -\frac{1}{2}\sum_{i,j}^{N}\int \frac{\varphi_i^*(r_1)\varphi_i^*(r_2)\varphi_j(r_1)\varphi_j(r_2)}{|r_1 - r_2|} dr_1 dr_2$$

- Cost: O(N^3 log N) (with large prefactor) for plane waves
- For atom-centered basis sets: O(N) (Strain, Scuseria (1996), Burant, Scuseria, Frisch (1996), Schwegler, Challacombe, Head-Gordon (1997))

Acceleration of Hartree-Fock and hybrid DFT calculations

- *N(N-1)/2* exchange integrals (all e^- pairs)

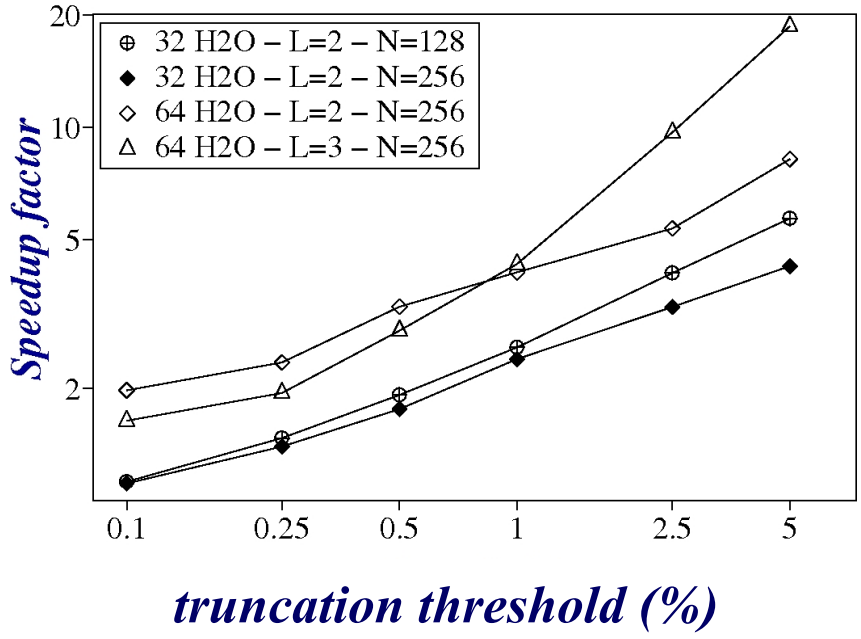
$$E_x = -\frac{1}{2}\sum_{i,j}^{N}\int \frac{\varphi_i^*(r_1)\varphi_i^*(r_2)\varphi_j(r_1)\varphi_j(r_2)}{|r_1 - r_2|} dr_1 dr_2$$

- using bisection: non-overlapping pairs can be skipped in the sum
- The error is positive

F.G., I. Duchemin, JCTC **9**, 582-587 (2013).

Speedup of hybrid-DFT calculations *vs* truncation threshold

$(H_2O)_{32}$, $(H_2O)_{64}$

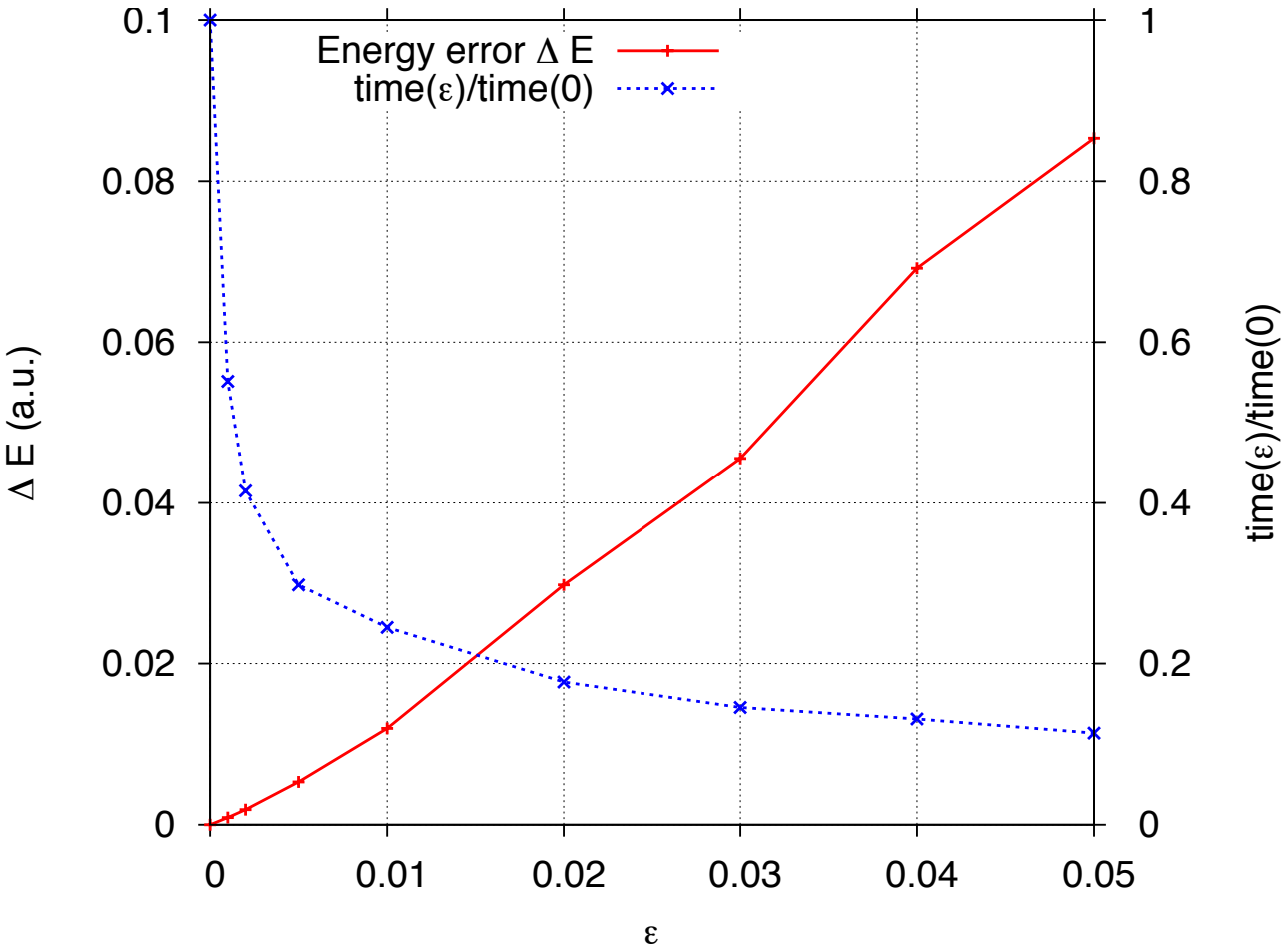


$(H_2O)_{256}$

ε	time (s)	speedup
0.0 (ref)	14000	1
0.1%	1179	11
0.25%	643	21
0.5%	425	33
1.0%	303	46
2.5%	162	86
5.0%	73	192

Truncation error due to bisection

- $(H_2O)_{63}Cl^-$



UCDAVIS
UNIVERSITY OF CALIFORNIA

Energy error per orbital vs threshold

Table 5. Energy Error (au) per Orbital

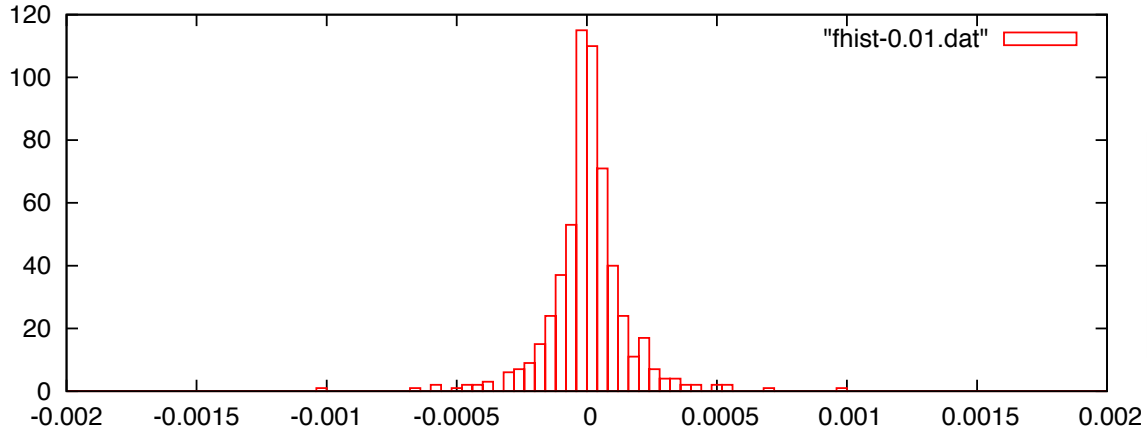
ϵ	H_2O	WO_3/H_2O	Si/H_2O	bulk Mo
0.001	2.28×10^{-6}	5.34×10^{-6}	1.17×10^{-6}	7.8×10^{-10}
0.005	2.87×10^{-5}	1.04×10^{-4}	2.35×10^{-5}	2.68×10^{-5}
0.01	7.39×10^{-5}	1.46×10^{-4}	6.08×10^{-5}	2.07×10^{-4}
0.02	1.98×10^{-4}	2.03×10^{-4}	2.95×10^{-4}	3.76×10^{-4}
0.05	4.31×10^{-4}	7.60×10^{-4}	8.28×10^{-4}	5.45×10^{-4}

Force error vs threshold

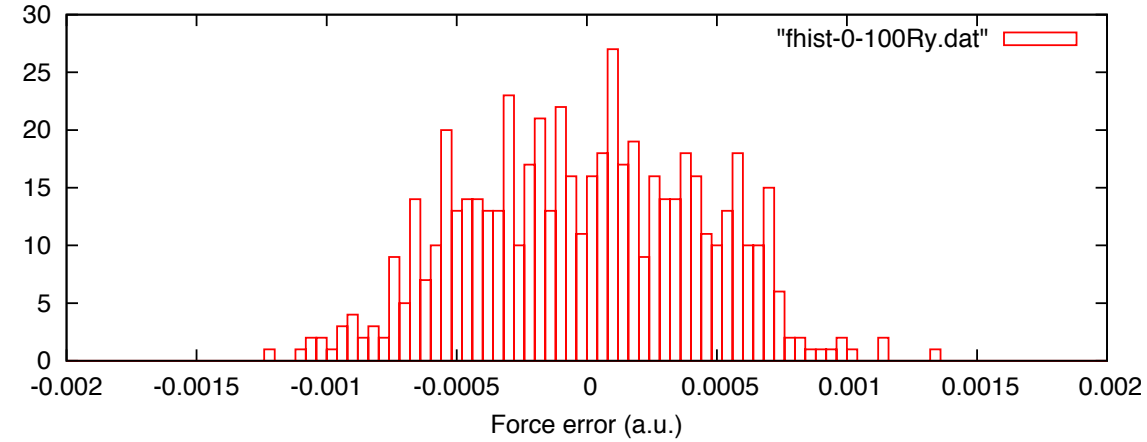
Table 7. Average Absolute Force Error (au)

ϵ	H_2O	WO_3/H_2O	Si/H_2O
0.001	9.0×10^{-6}	1.1×10^{-5}	1.0×10^{-5}
0.005	6.4×10^{-5}	2.6×10^{-4}	6.4×10^{-5}
0.01	1.5×10^{-4}	3.6×10^{-4}	1.2×10^{-4}
0.02	3.9×10^{-4}	5.0×10^{-4}	3.5×10^{-4}
0.05	8.1×10^{-4}	1.4×10^{-3}	9.5×10^{-4}

Error in ionic forces for MD applications



Error caused by bisection with 0.01 threshold

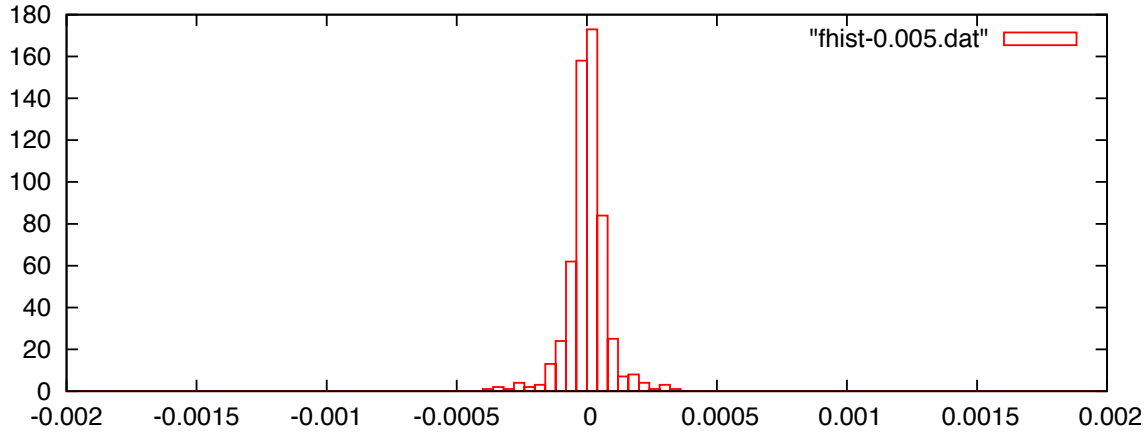


Error caused by changing Ecut from 85Ry to 100Ry

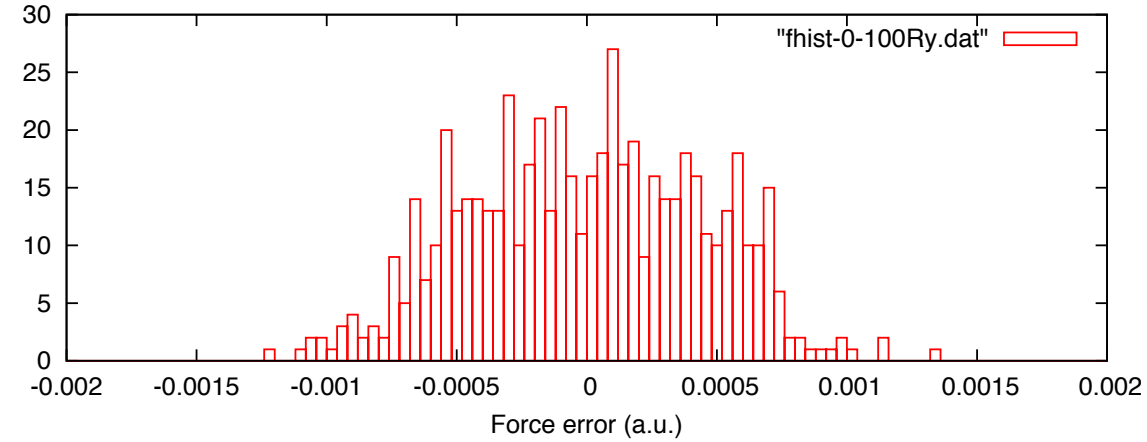
Recursive bisection affects forces in a controlled way

UCDAVIS UNIVERSITY OF CALIFORNIA

Error in ionic forces for MD applications



Error caused by bisection with 0.005 threshold



Error caused by changing Ecut from 85Ry to 100Ry

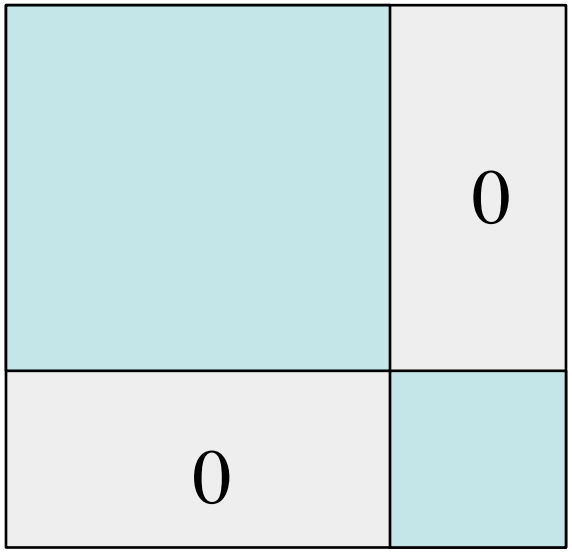
Recursive bisection affects forces in a controlled way

UCDAVIS
UNIVERSITY OF CALIFORNIA

HOMO-LUMO gap, band gaps

- Hybrid DFTs lead to large improvements in band gaps
  (Henderson, Paier, Scuseria, PhysStatSol 2011)
- Bisection algorithm: occupied and empty orbitals must not be mixed when localizing orbitals

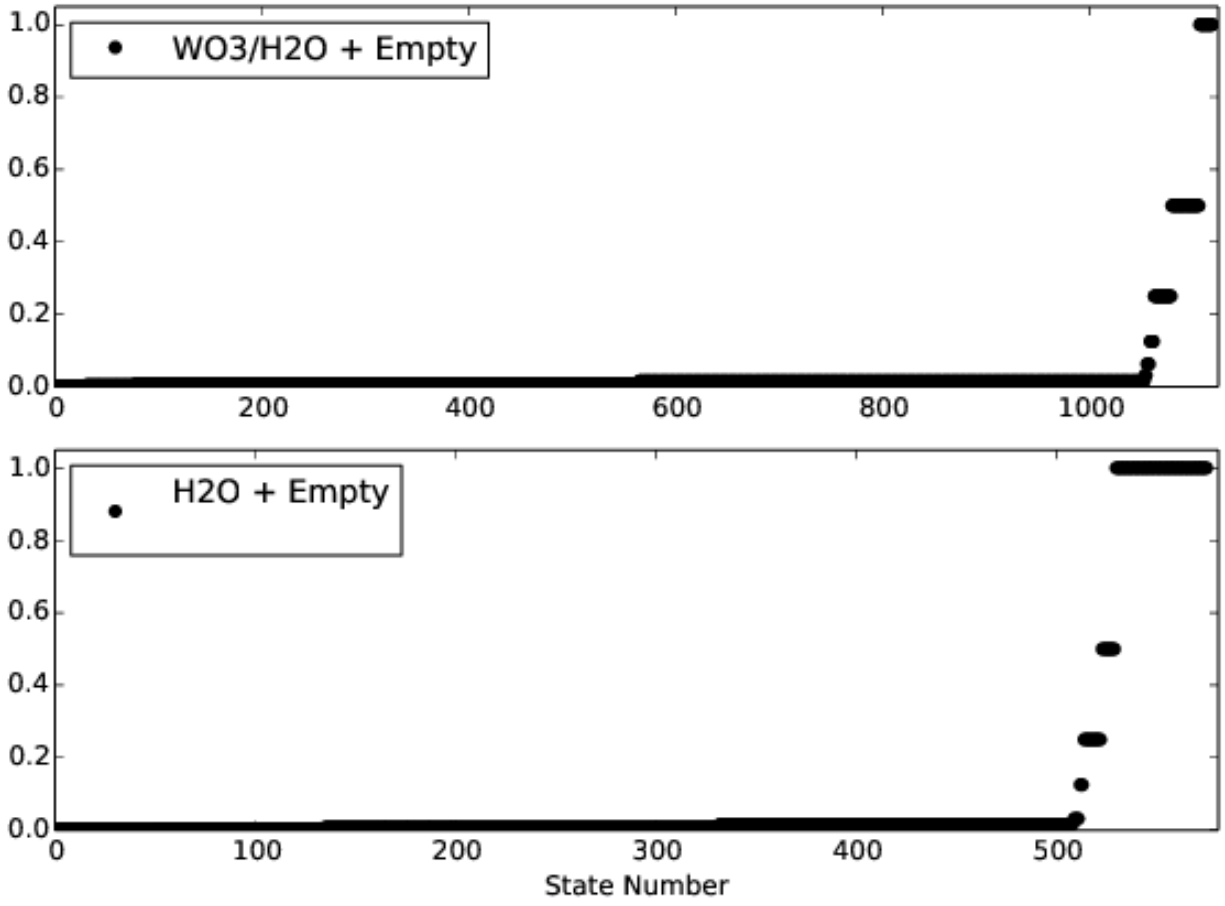
$$A^{(k)} =$$



$(H_2O)_{63}Cl^-$

threshold	Egap (eV)
0.0	7.03
0.01	7.01
0.02	6.99
0.05	6.92

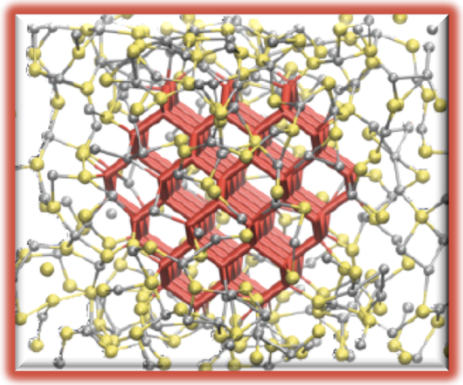
Localization of empty orbitals



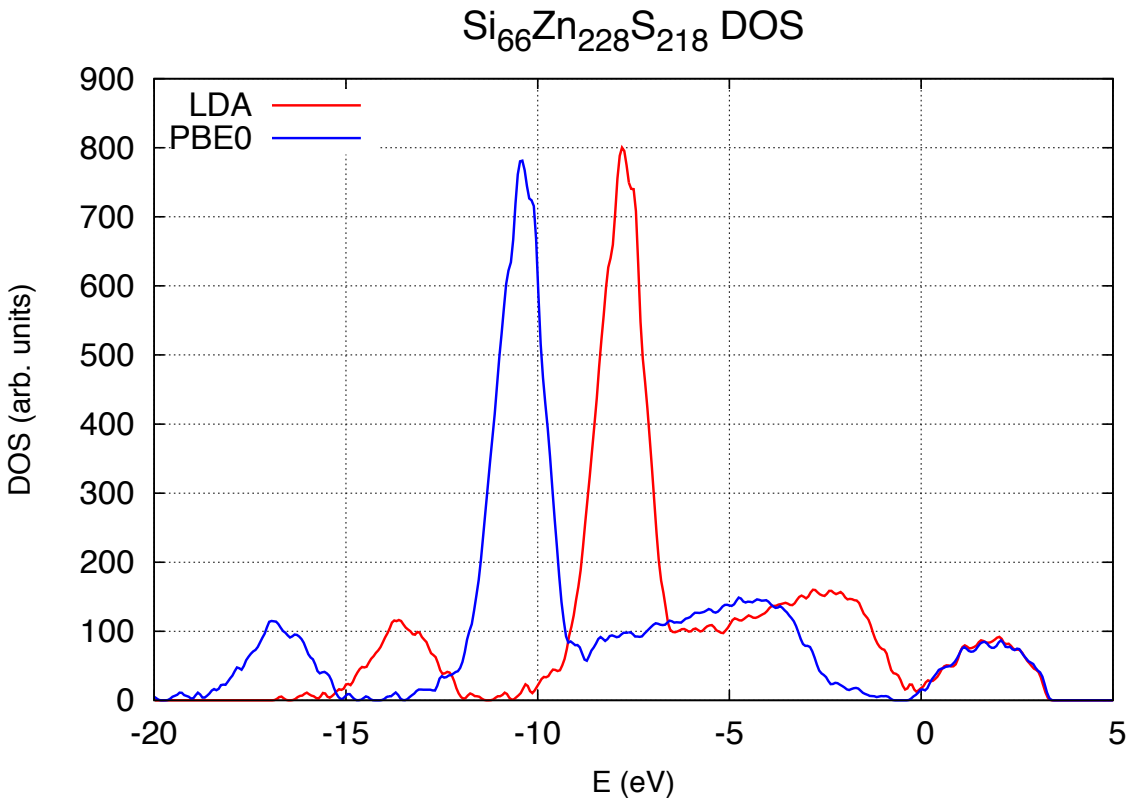
W. Dawson and F.G, JCTC **11**, 4655 (2015)

UCDAVIS
UNIVERSITY OF CALIFORNIA

Hybrid-DFT electronic structure of Si nanoparticles embedded in ZnS



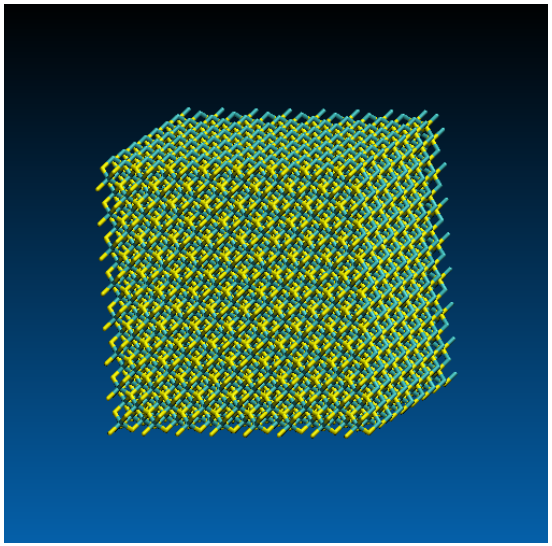
- $Si_{66}Zn_{228}S_{218}$
- 4308 electrons
- 200 empty orbitals
- PBE0
- 1 scf step, 5 iterations
- BG/Q, 16k cores, 772 s.



S. Wippermann, M. Vörös, A. Gali, F. Gygi, G. Zimanyi, G.Galli, Phys. Rev. Lett. **112**, 106801 (2014) .

UCDAVIS
UNIVERSITY OF CALIFORNIA

Hybrid-DFT electronic structure of bulk SiC (4096 atoms)



- 4096 atoms
- 16384 electrons
- PBE0 electronic structure (hybrid)
- recursive subspace bisection
- ANL Mira (BG/Q) 64k cores
- 357s/self-consistent iteration

hybrid DFT electronic structure for 4096 atoms

Where is the (Big) Data?

- Making data accessible is critical for verification and validation of simulation software
- Agreeing on data formats has proved difficult..

- http://www.quantum-simulation.org
  - XML schemas for electronic structure data
  - repository of reference simulations

Summary

- Simulation of complex materials
- Truncation of Maximally Localized Wannier functions
- Recursive subspace bisection
- Controlled error in inhomogeneous systems
- Acceleration of hybrid DFT simulations
- http://www.quantum-simulation.org
- http://qboxcode.org

Supported by DOE BES DE-SC0008938 and the MICCoM DOE center

Acknowledgements

- Giulia Galli (UChicago)
- Marco Govoni (UChicago)
- Stefan Wippermann (MPI)
- Eric Schwegler (LLNL)
- Funding
  - DOE BES
- computer time
  - DOE INCITE/ANL-ALCF
  - NSF XSEDE
  - NERSC

- William Dawson (CS, UCDavis)
- Martin Schlipf (CS, UCDavis)
- Cui Zhang (UCDavis)
- Alex Gaiduk (UChicago)
- Marton Vörös (UCDavis, ANL)
- Quan Wan (UChicago)
- T.-Anh Pham (LLNL)

Supported by DOE BES DE-SC0008938 and the DOE MICCoM center

U.S. DEPARTMENT OF ENERGY | Office of Science

UC DAVIS UNIVERSITY OF CALIFORNIA

http://qboxcode.org http://www.quantum-simulation.org

http://miccom-center.org