Large scale DFT: why we need it and how to do it

David Bowler
Thomas Young Centre, UCL, Gower St, London, UK
London Centre for Nanotechnology, 17-19 Gordon St, London, UK
Department of Physics & Astronomy, UCL, Gower St, London, UK
WPI MANA, NIMS, Tsukuba, Ibaraki, Japan
david.bowler@ucl.ac.uk
http://www.atomisticsimulations.org/
http://www.order-n.org/
@MillionAtomMan

What is large scale?

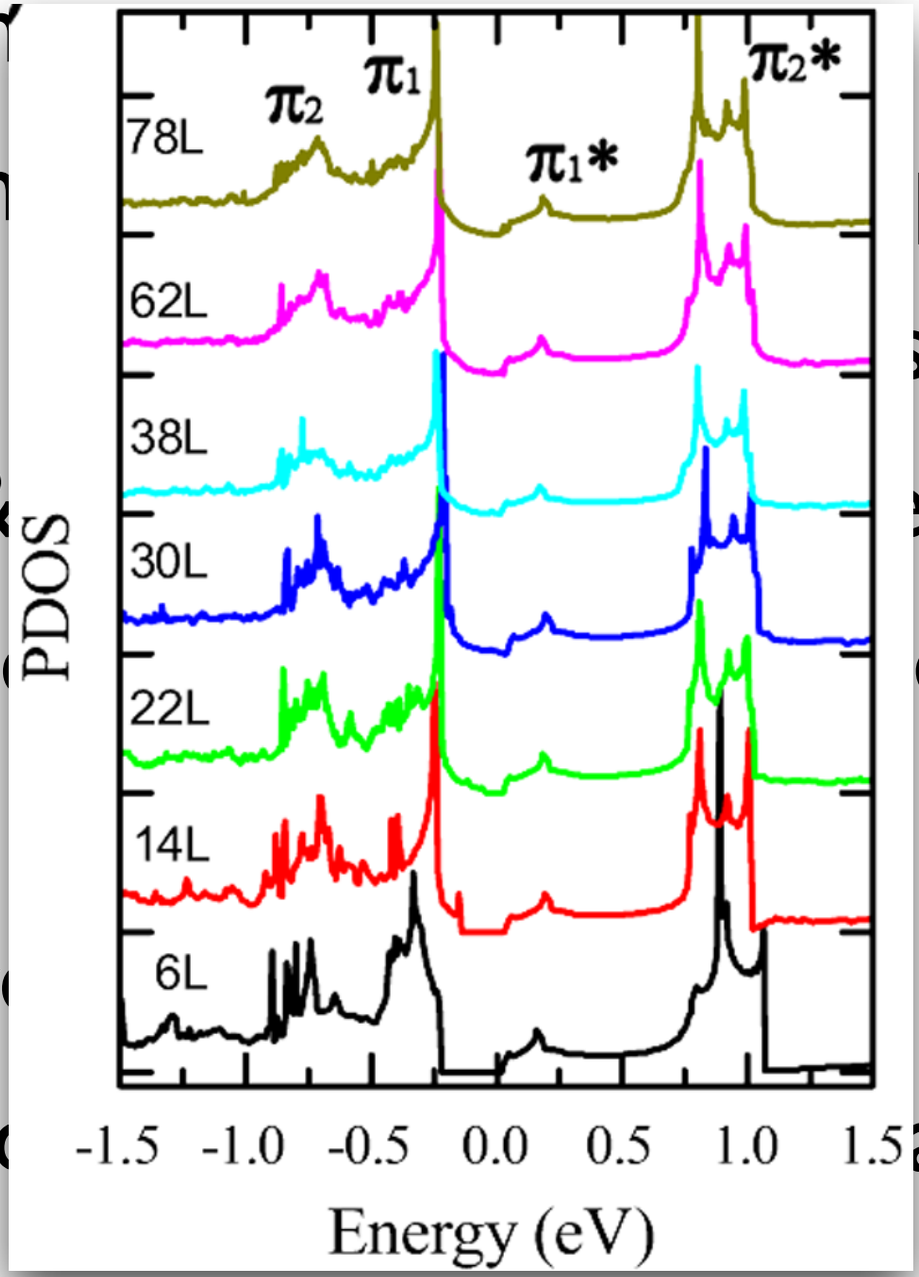
- Large scale: 500-10,000 atoms ($N_{atoms}/N_{proc} \approx 5\text{-}50$)
- 100-1,000+ processes
- Linear scaling: 1,000-1,000,000 atoms

- Many atoms: length scales
- Many processes: time scales

- Almost all DFT calculations are under 500 atoms
  - Cubic scaling
  - Habit?

Why large-scale DFT?

- We have become very good at calculations using 100-1,000 atom
- But there are in problems:
  - Large defects s)
  - Amorphous & erials
  - Realistic conc d solutions)
  - Biomolecules
  - Nanostructur
- Testing and vali als



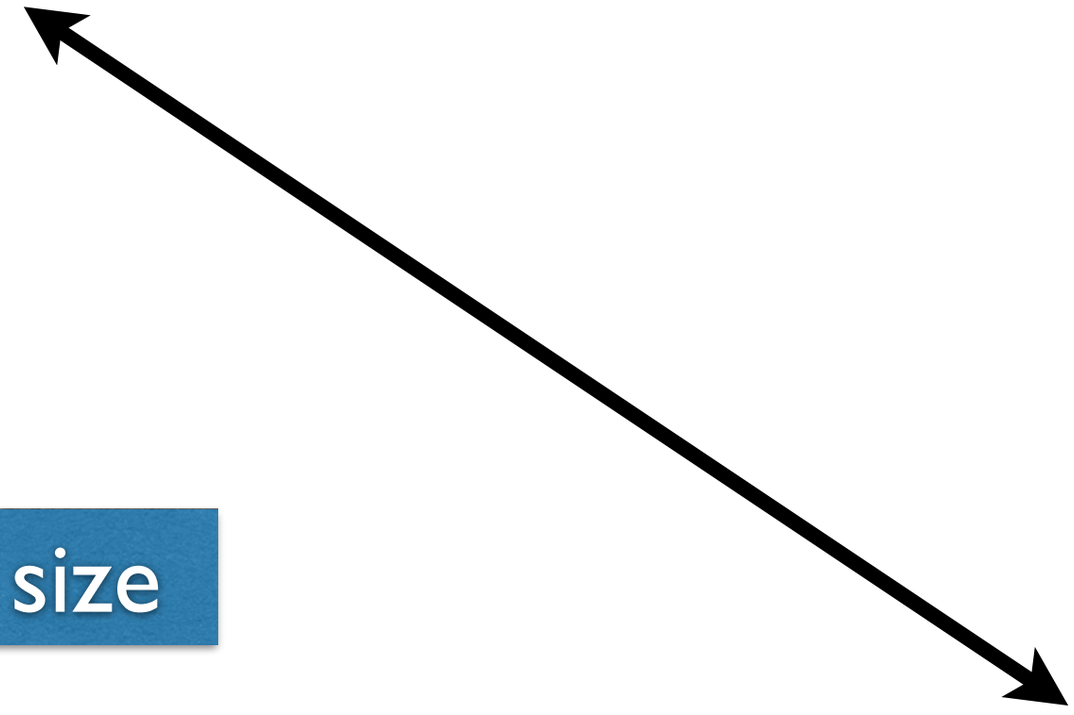
Why not large-scale DFT?

- What is stopping us?
  - Lack of a "killer application"
  - Received wisdom/habit
  - Perception that big is hard
  - Familiarity with standard codes
  - Underestimation of need

Competing requirements

Accuracy

Functional



System size

Basis size

MD time

Solver

Cost

How many atoms ?

Efficiency ⟷ Accuracy

Few — Many

Representation — Realistic

Alloys (concentration)
Doping

Perfect crystals — Defects, dislocations etc

Solving for the ground state

- N^3
  - Direct diagonalisation
  - Variational solvers (orthogonalisation)
- N^2
  - PEXSI

- N: localisation
  - Iterative
  - Variational
  - Patching/subsystem

Why large-scale DFT?

- HPC centres give many 100,000s cores (exascale)
- Effective parallel scaling is key
  - Reduce time to ground state
  - Increase system size
  - Communication (CPU-CPU, CPU-GPU)
- Use of GPUs
- Multi-threading vs message passing

Large scale calculations

- CONQUEST (1,000,000)
- BigDFT (10,000)
- CP2K (1,000,000)
- OpenMX (10,000)
- ONETEP (21,000)

- RSDFT (107,000)
- Google TPU (31,000)

- MGmol (1,000,000)
- LDC-DFT (6,000,000)
- DFT-FE (10,000)
- PARSEC (26,000)
- DGDFT (2,500,000)

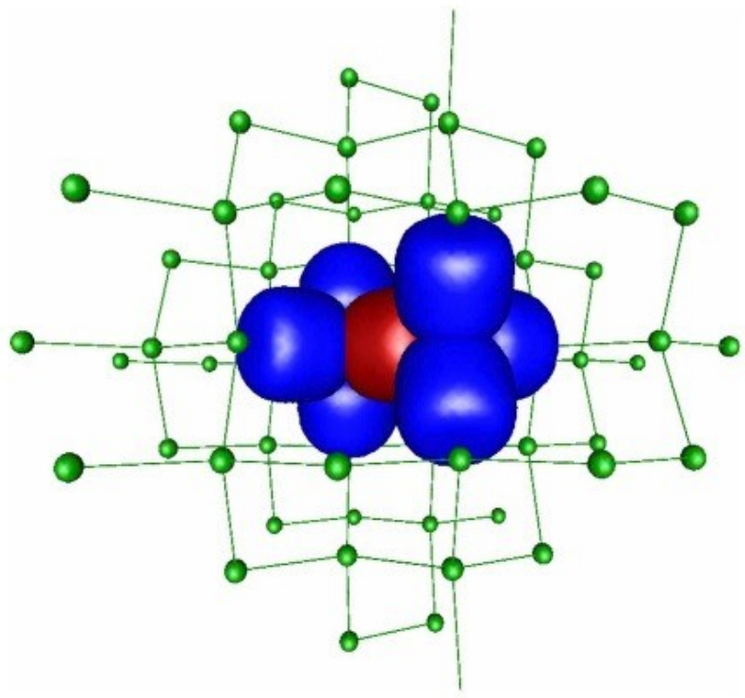
Large scale calculations

- CONQUEST (768,000)
- CP2K (47,000)
- ONETEP (100,000)
- DGDFT (35,900,000)
- RSDFT (664,000)
- MGmol (1,600,000)
- LDC-DFT (786,000)
- DFT-FE (156,000 + 22,800)
- PARSEC (16,000)

CONQUEST

Extending DFT size

- Real-space locality
- Key to parallel efficiency *and* scaling
- Standard implementations are *non-local*:
  wavefunctions span whole system
- But electronic structure is local:
  - Charge density
  - Wannier functions
  - Density matrix decay



Key concept: locality

- Local basis functions are key
- Sparse matrices; reduced scaling
- Operations local to process
- Fits with nearsightedness
- Makes efficient parallelisation easier

- Pulay forces
- Orthogonalisation?
- Systematic convergence?

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \sum_{i\alpha} c_{i\alpha,n\mathbf{k}} \phi_{i\alpha}(\mathbf{r})$$

Support Functions

- Support functions are represented in terms of a basis:

$$\phi_{i\alpha}(\mathbf{r}) = \sum_s c_{i\alpha s}\chi_s(\mathbf{r})$$

- CONQUEST can use two different basis sets:
  - Pseudo-atomic orbitals (cf OpenMX, SIESTA)
  - B-splines or blips (cf wavelets, ONETEP psincs)
- PAOs: analytic operations, small basis, intuitive
- Blips: systematic convergence (to plane-wave accuracy)

Solving for DM

- Support functions give **H** and **S** (linear scaling)
- How do we find the density matrix?
  - Exact (ScaLAPACK): 1-10,000 atoms
    - PAOs or multi-site support functions (MSSF)
  - O(N) (range): 1,000-1,000,000+ atoms
- Complementary: choose appropriate method
- Aim: efficient, accurate solution & analysis

Diagonalisation: DM

- We build **K** from wavefunction coefficients
- Efficient memory use

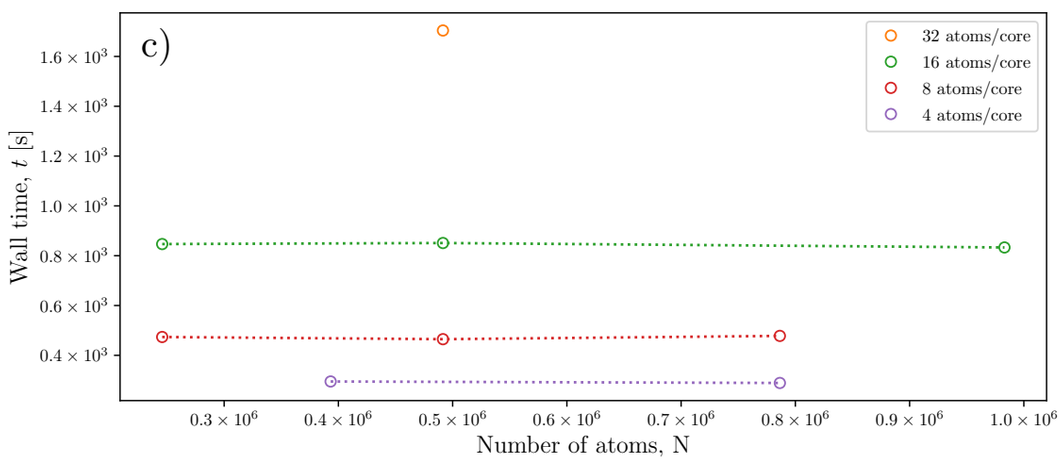
$$\psi_{n\mathbf{k}}(\mathbf{r}) = \sum_{i\alpha} c_{i\alpha,n\mathbf{k}} \phi_{i\alpha}(\mathbf{r})$$

$$K_{i\alpha,j\beta} = \sum_{n\mathbf{k}} w_{n\mathbf{k}} f_{n\mathbf{k}} c^{\star}_{i\alpha,n\mathbf{k}} c_{j\beta,n\mathbf{k}}$$

CONQUEST: Basic capabilities

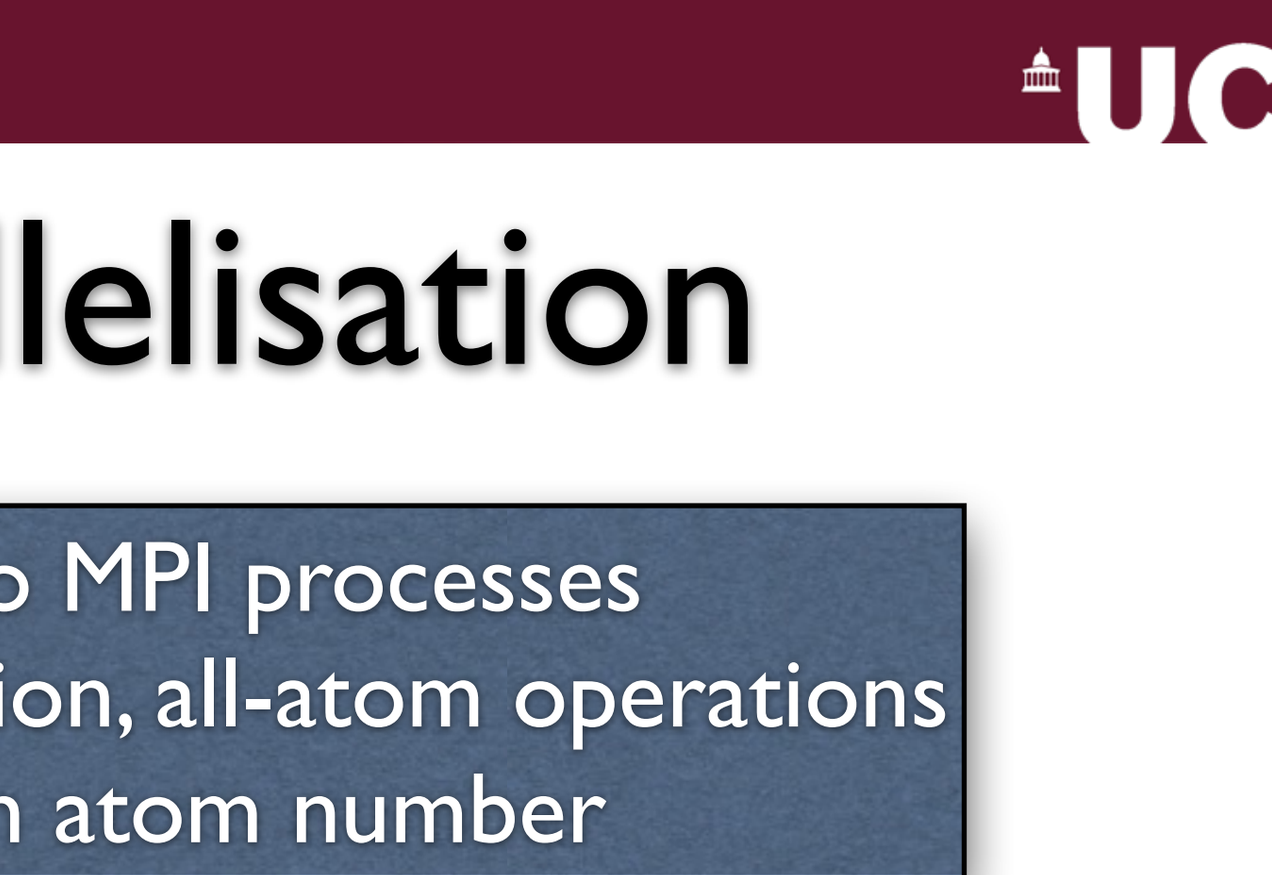
- Efficient parallelisation (from 1 to 200,000+ cores)
- Flexible (from 1 to 2,000,000+ atoms)
- Exact, analytic forces; MD on 32,000+ atoms
- ONCVPSP pseudopotentials (PseudoDojo)
- LDA, GGA, DFT-D2/3/TS, vdW-DF, EXX (partial)

Open source: https://github.com/OrderN/CONQUEST-release/



Parallelisation

- Assign areas of space to MPI processes
- Care with communication, all-atom operations
- Increase processes with atom number



Pseudo-atomic orbitals

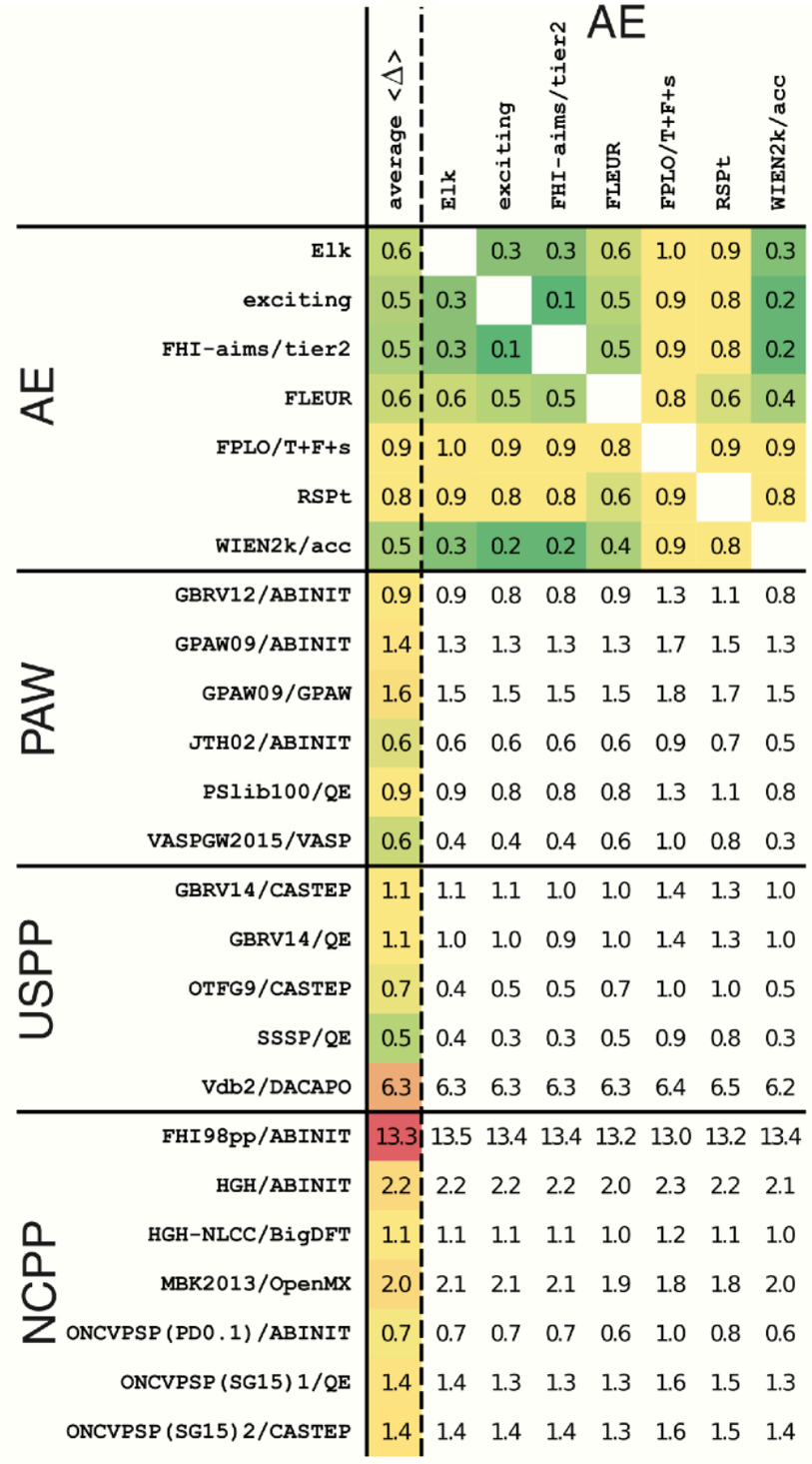
CONQUEST: PAOs

- Use PseudoDojo pseudopotentials
- Pseudo-atomic orbital basis sets
- Can we define reasonable default bases?
- How do we choose radii?
  - Siesta energy shifts
  - Adapt: large and small (2eV and 20meV)
  - Either share energies or average radii
  - Allows SZ up to TZTP automatic bases

Jpn. J. Appl. Phys. **58**, 100503 (2019)

Pseudopotentials

- Hamann NC pseudopotentials
- PseudoDojo set (0.7 meV/atom)
- As accurate as PAWs
- High quality pseudopotentials



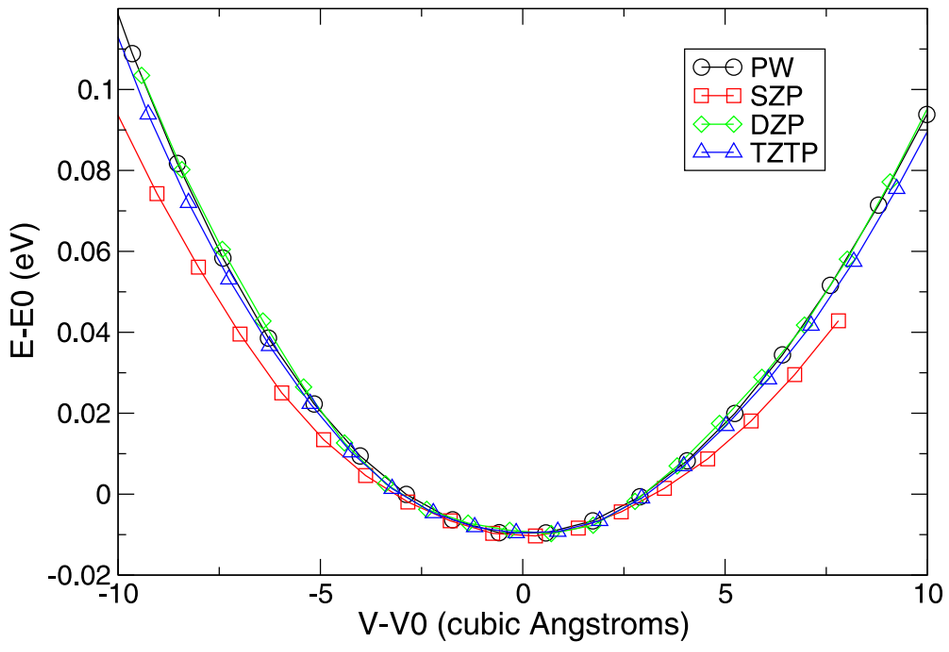
		average <Δ>	AE: Elk	AE: exciting	AE: FHI-aims/tier2	AE: FLEUR	AE: FPLO/T+F+s	AE: RSPt	AE: WIEN2k/acc
AE	Elk	0.6		0.3	0.3	0.6	1.0	0.9	0.3
	exciting	0.5	0.3		0.1	0.5	0.9	0.8	0.2
	FHI-aims/tier2	0.5	0.3	0.1		0.5	0.9	0.8	0.2
	FLEUR	0.6	0.6	0.5	0.5		0.8	0.6	0.4
	FPLO/T+F+s	0.9	1.0	0.9	0.9	0.8		0.9	0.9
	RSPt	0.8	0.9	0.8	0.8	0.6	0.9		0.8
	WIEN2k/acc	0.5	0.3	0.2	0.2	0.4	0.9	0.8	
PAW	GBRV12/ABINIT	0.9	0.9	0.8	0.8	0.9	1.3	1.1	0.8
	GPAW09/ABINIT	1.4	1.3	1.3	1.3	1.3	1.7	1.5	1.3
	GPAW09/GPAW	1.6	1.5	1.5	1.5	1.5	1.8	1.7	1.5
	JTH02/ABINIT	0.6	0.6	0.6	0.6	0.6	0.9	0.7	0.5
	PSlib100/QE	0.9	0.9	0.8	0.8	0.8	1.3	1.1	0.8
	VASPGW2015/VASP	0.6	0.4	0.4	0.4	0.6	1.0	0.8	0.3
USPP	GBRV14/CASTEP	1.1	1.1	1.1	1.0	1.0	1.4	1.3	1.0
	GBRV14/QE	1.1	1.0	1.0	0.9	1.0	1.4	1.3	1.0
	OTFG9/CASTEP	0.7	0.4	0.5	0.5	0.7	1.0	1.0	0.5
	SSSP/QE	0.5	0.4	0.3	0.3	0.5	0.9	0.8	0.3
	Vdb2/DACAPO	6.3	6.3	6.3	6.3	6.3	6.4	6.5	6.2
NCPP	FHI98pp/ABINIT	13.3	13.5	13.4	13.4	13.2	13.0	13.2	13.4
	HGH/ABINIT	2.2	2.2	2.2	2.2	2.0	2.3	2.2	2.1
	HGH-NLCC/BigDFT	1.1	1.1	1.1	1.1	1.0	1.2	1.1	1.0
	MBK2013/OpenMX	2.0	2.1	2.1	2.1	1.9	1.8	1.8	2.0
	ONCVPSP(PD0.1)/ABINIT	0.7	0.7	0.7	0.7	0.6	1.0	0.8	0.6
	ONCVPSP(SG15)1/QE	1.4	1.4	1.3	1.3	1.3	1.6	1.5	1.3
	ONCVPSP(SG15)2/CASTEP	1.4	1.4	1.4	1.4	1.3	1.6	1.5	1.4

PAO generation

- Stand-alone CONQUEST code
- Reads output from Hamann code (small patch)
- Generates pseudo-atomic orbitals and ion files
- Automated basis sets possible
- Full user control available

CONQUEST basis set tests

Bulk Ge



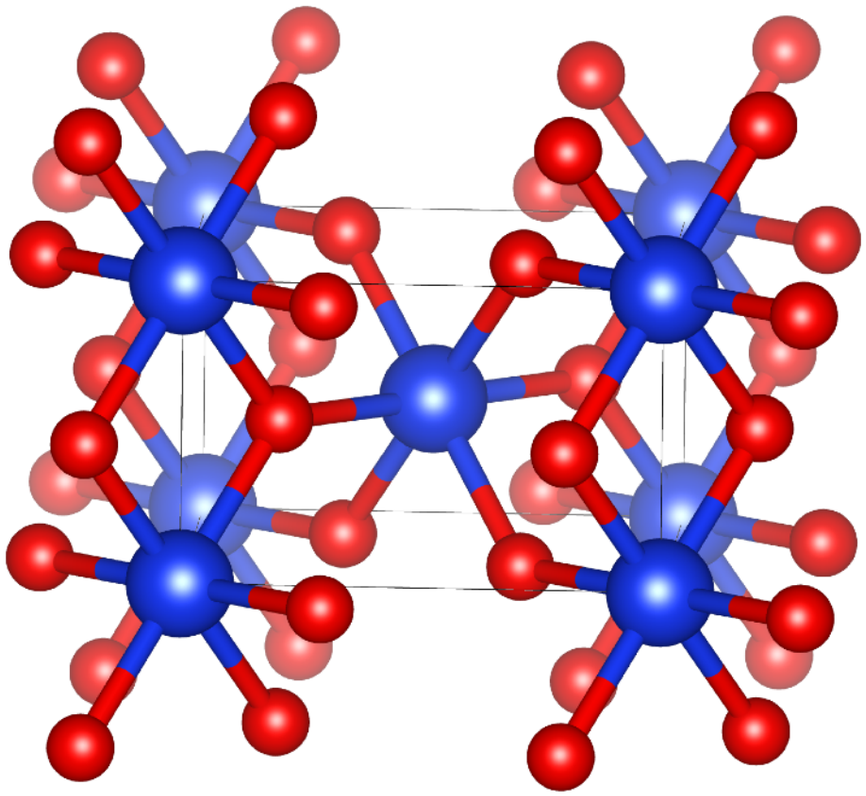
	PW		CQ	
	a_0/V	B	a_0/V	B
C	3.558	449	3.562	453
Si	5.431	93	5.44	92
Ge	5.676	67	5.682	67
STO	58.8	186	60.1	183
PTO	60.1	191	60.6	190
$MgSiO_3$	167.4	235.7	165.0	253.2

Jpn. J. Appl. Phys. **58**, 100503 (2019)

Basis set tests: SiO_2

SiO_2 (stishovite)

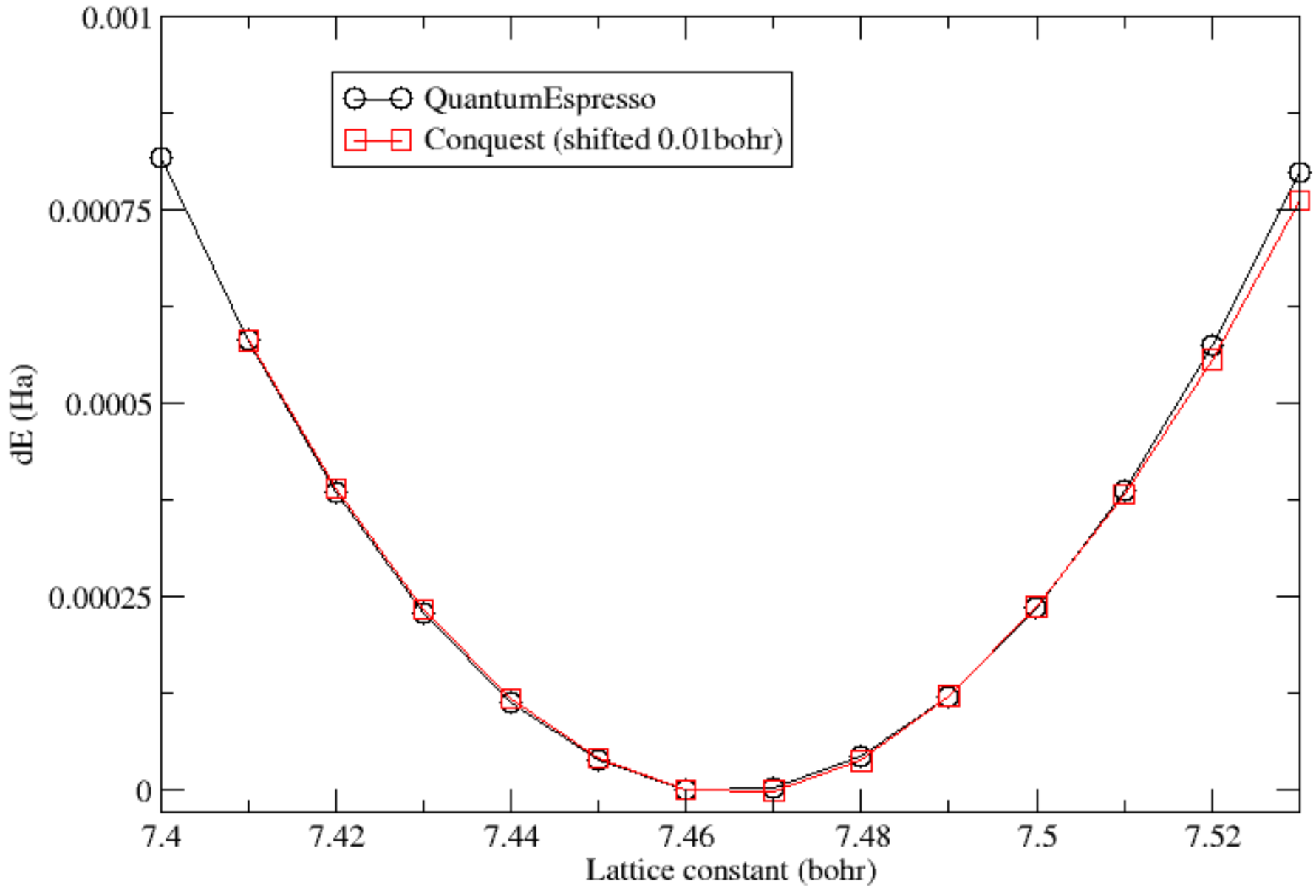
Z. Raza (NIMS)



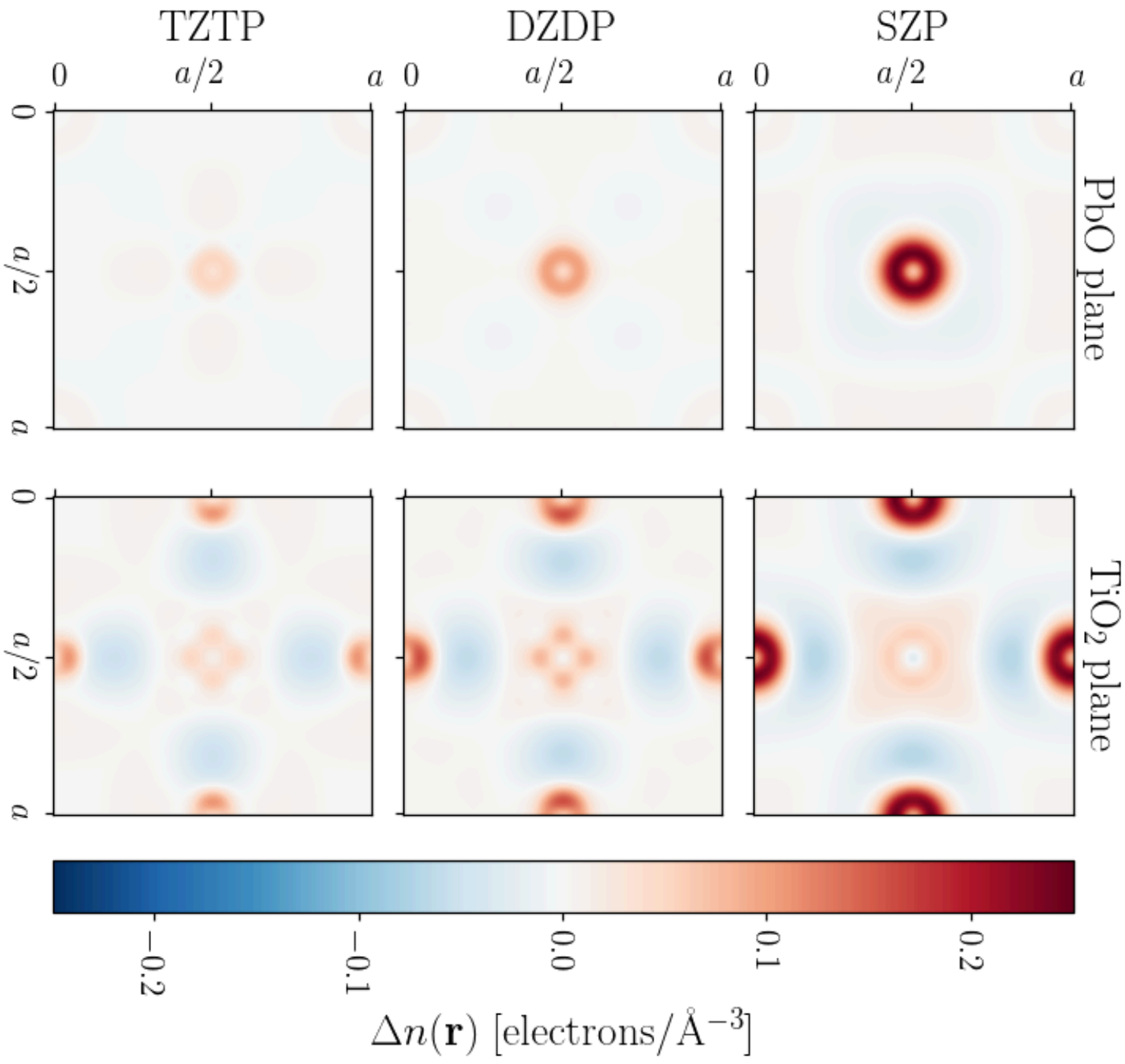
	a (Å)	c/a	% error in a
Experiment	4.177	0.638	
VASP/PBE[a]	4.225	0.637	+1.1%
SZ	4.476	0.638	+7.1%
SZP	4.278	0.638	+2.4%
DZP	4.270	0.638	+2.2%
TZTP	4.233	0.638	+1.3%

[a] *Phys. Rev. B* **88**, 184103 (2013)

Basis set tests: $SrTiO_3$



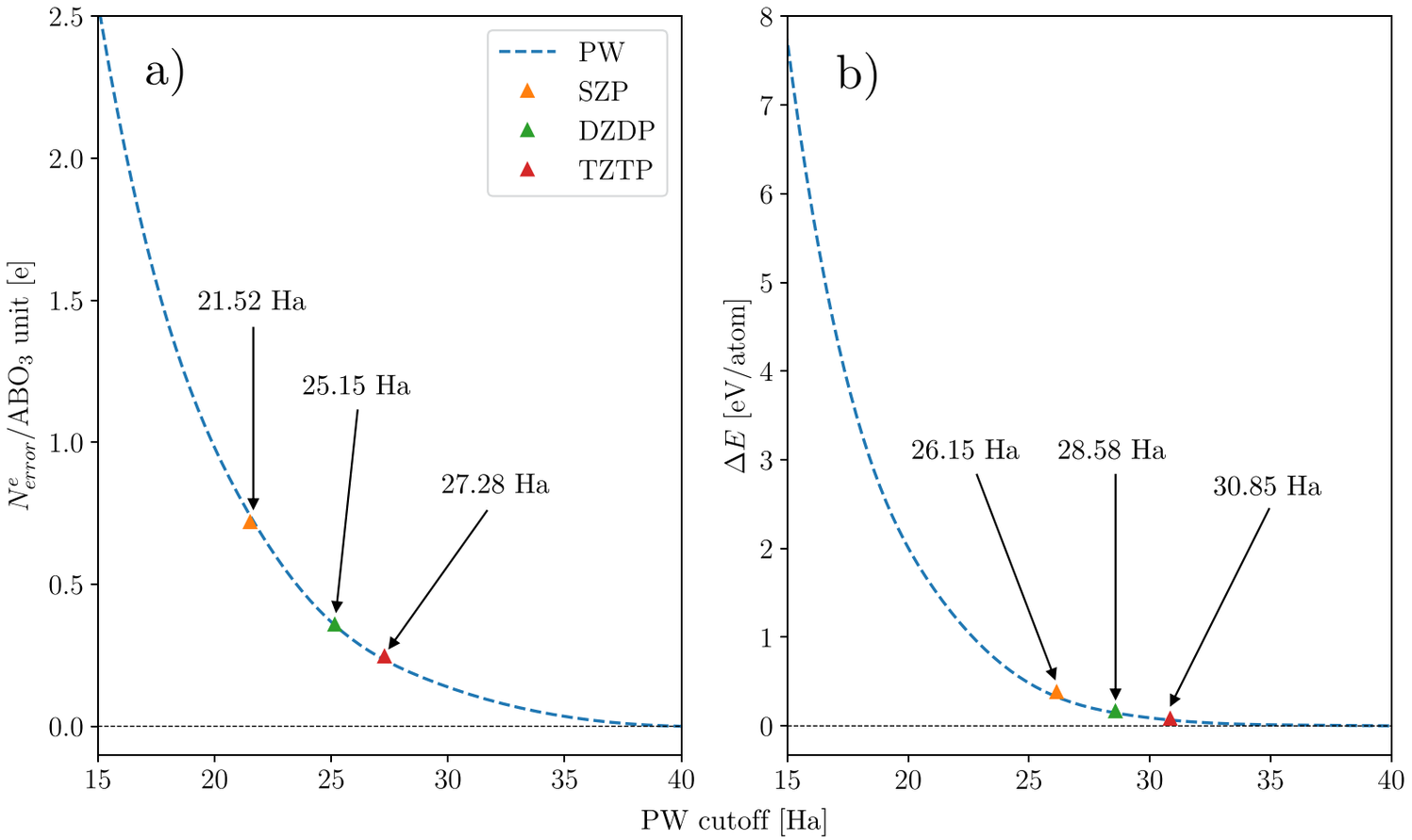
How good is the density?



- Key quantity:
  - Energy
  - Polarisation, …
- Compare to PW
- 40Ha cutoff

Electron. Struct. **2**, 025002 (2020)

Converging density & energy

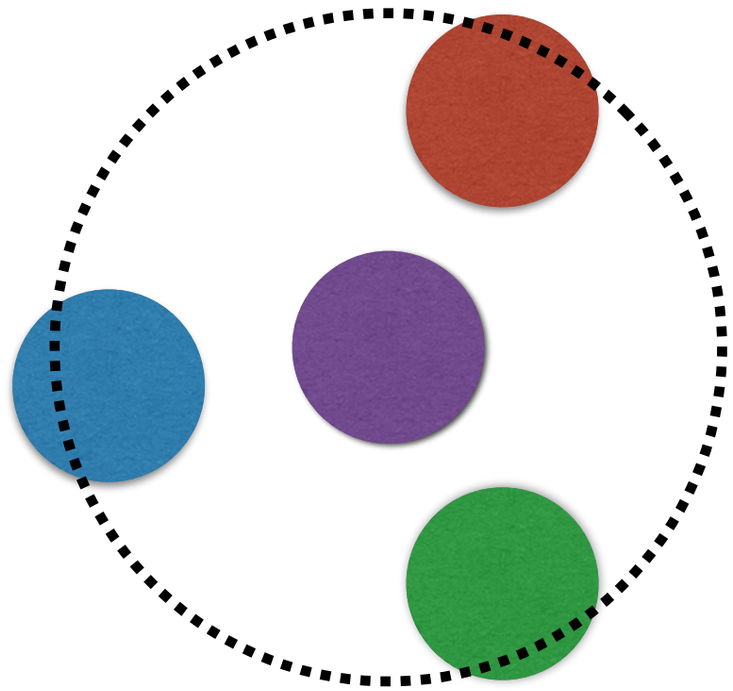


- Compare to PW
- 40Ha cutoff
- Equivalent cutoffs

Electron. Struct. **2**, 025002 (2020)

Multi-site SFs

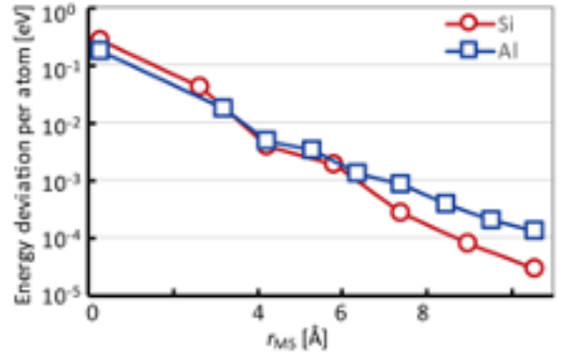
Multi-site SFs



- We want an accurate basis
- We want efficient performance
- Make a small set of SFs for each atom
- Build them from PAOs from several atoms
- Local filter diagonalisation

JCTC **10**, 4813; PCCP **17**, 31427

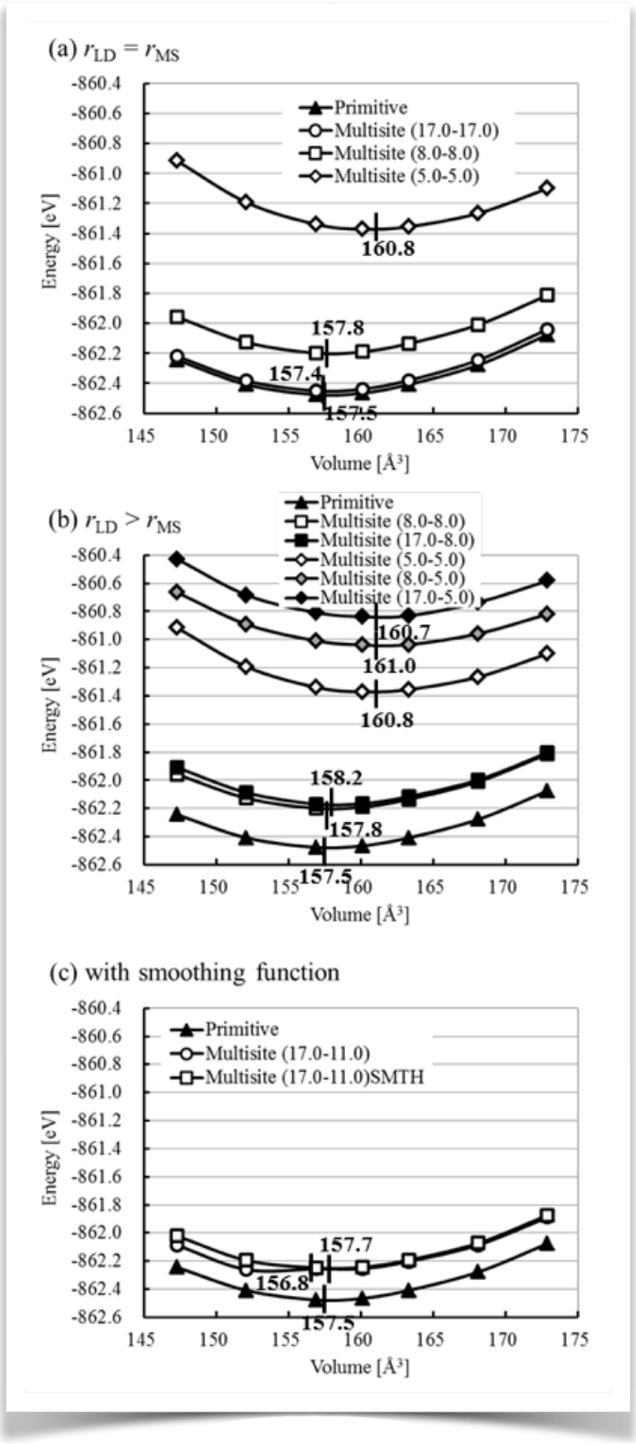
Multi-site functions

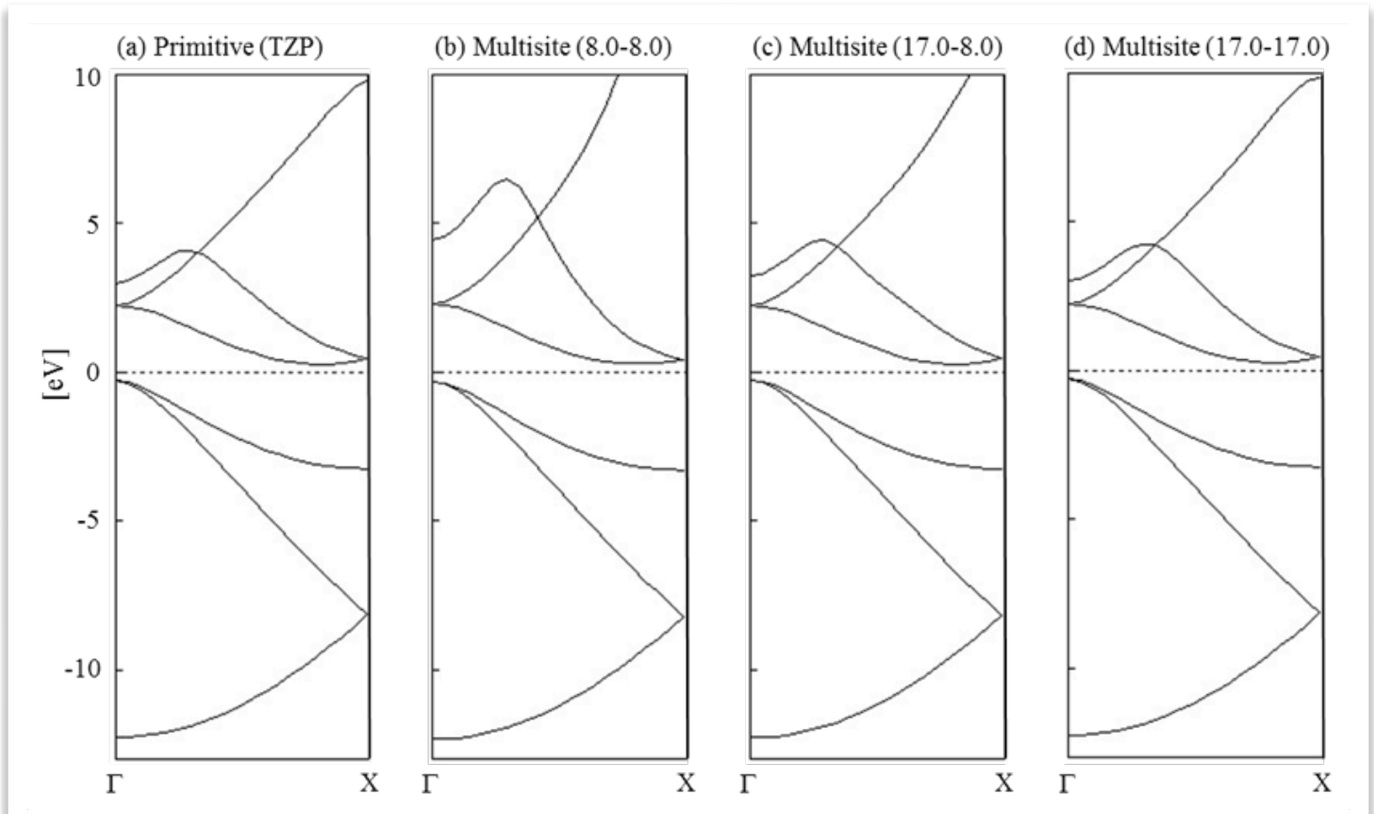


- Use PAOs from neighbouring atoms
- Effectively localised molecular orbitals
- Two ranges: MS and LFD
- This is accurate and efficient: small **H** and **S**
- Follows method of Rayson & Briddon*
- Improvements: JCTC **10**, 4813; PCCP **17**, 31427

*M. J. Rayson and P. R. Briddon, *Phys. Rev. B*, 2009, **80**, 205104

MSSF: bulk Si





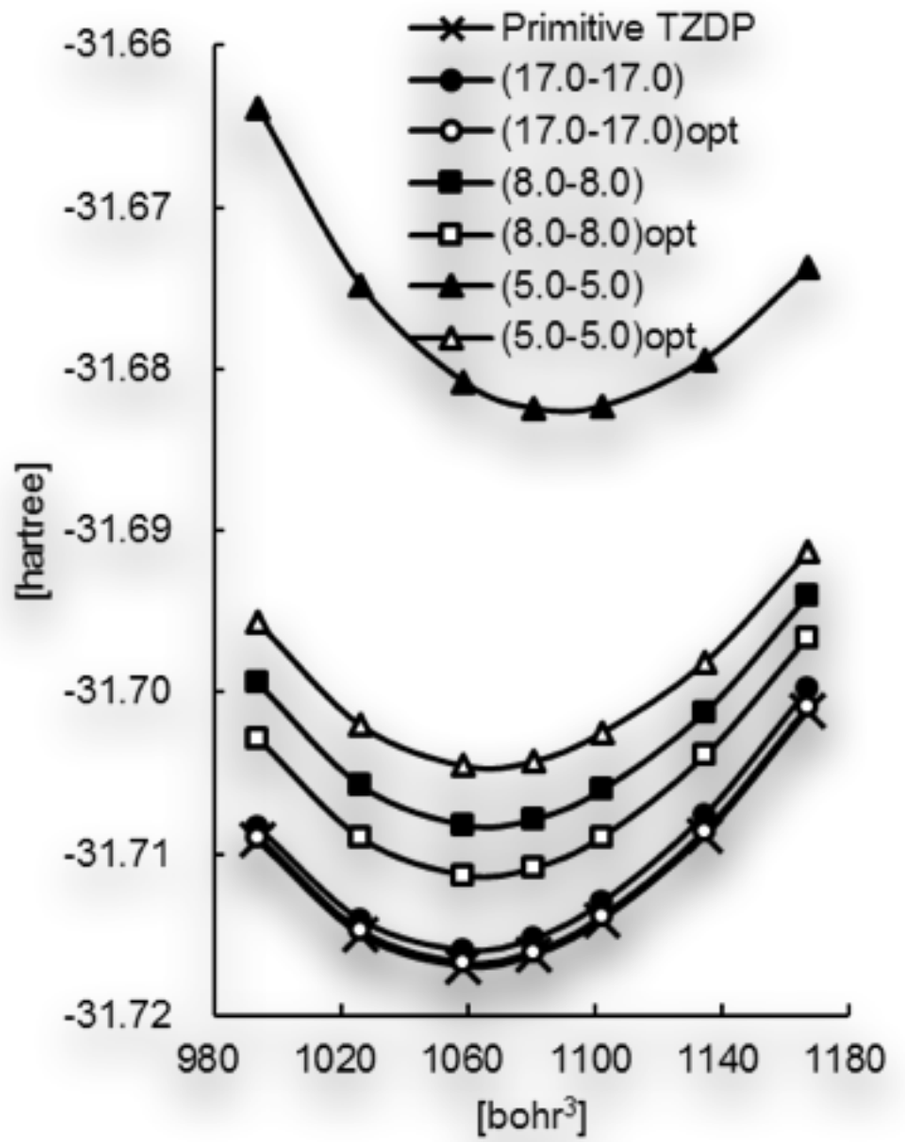
- Lattice constant and bulk modulus good
- Band structure excellent

Optimising MSSF

- Original method performs filter diagonalisation
  - Large, primitive basis; small support function basis
  - We diagonalise **H** for a small cluster
  - Filter eigenstates onto MSSF using Fermi function
- Can we improve by *optimising* energy?
- Analytic energy gradients available

Optimising MSSF: bulk Si

- Little effect for large radii
- Significant improvement for small radii
- Promising approach



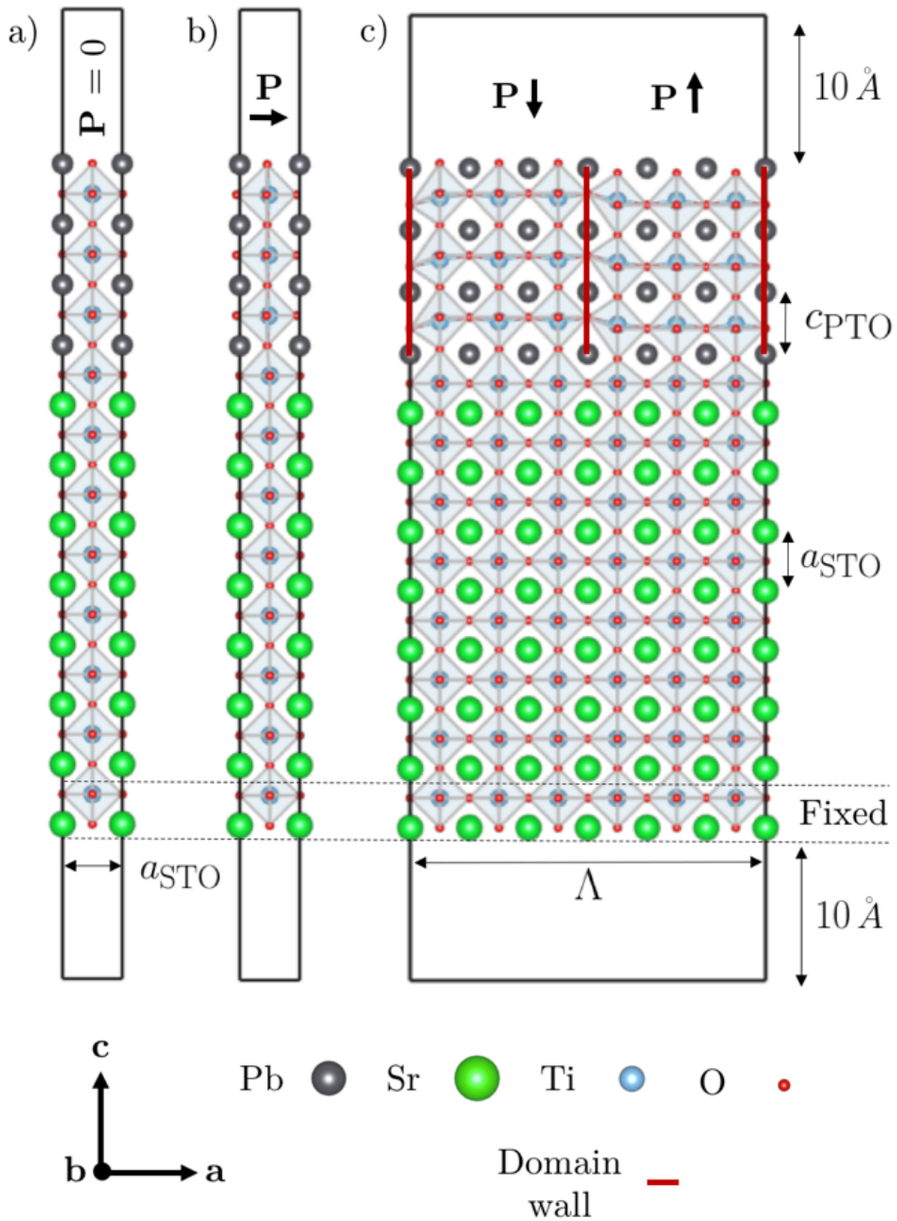
MSSF timing & scaling

	PAO	MSSF	PAO	MSSF	MSSF
# SF	27,192	8122	54,384	16,244	16,244
# MPI	432	432	108	108	864
Matrix	64	400	156	1,455	406
Diag	1193	39	37,648	701	166
Total	1257	439	37,804	2156	572

- Graphene on Rh(111); 1,544 or 3,088 atoms

MSSF: Applications

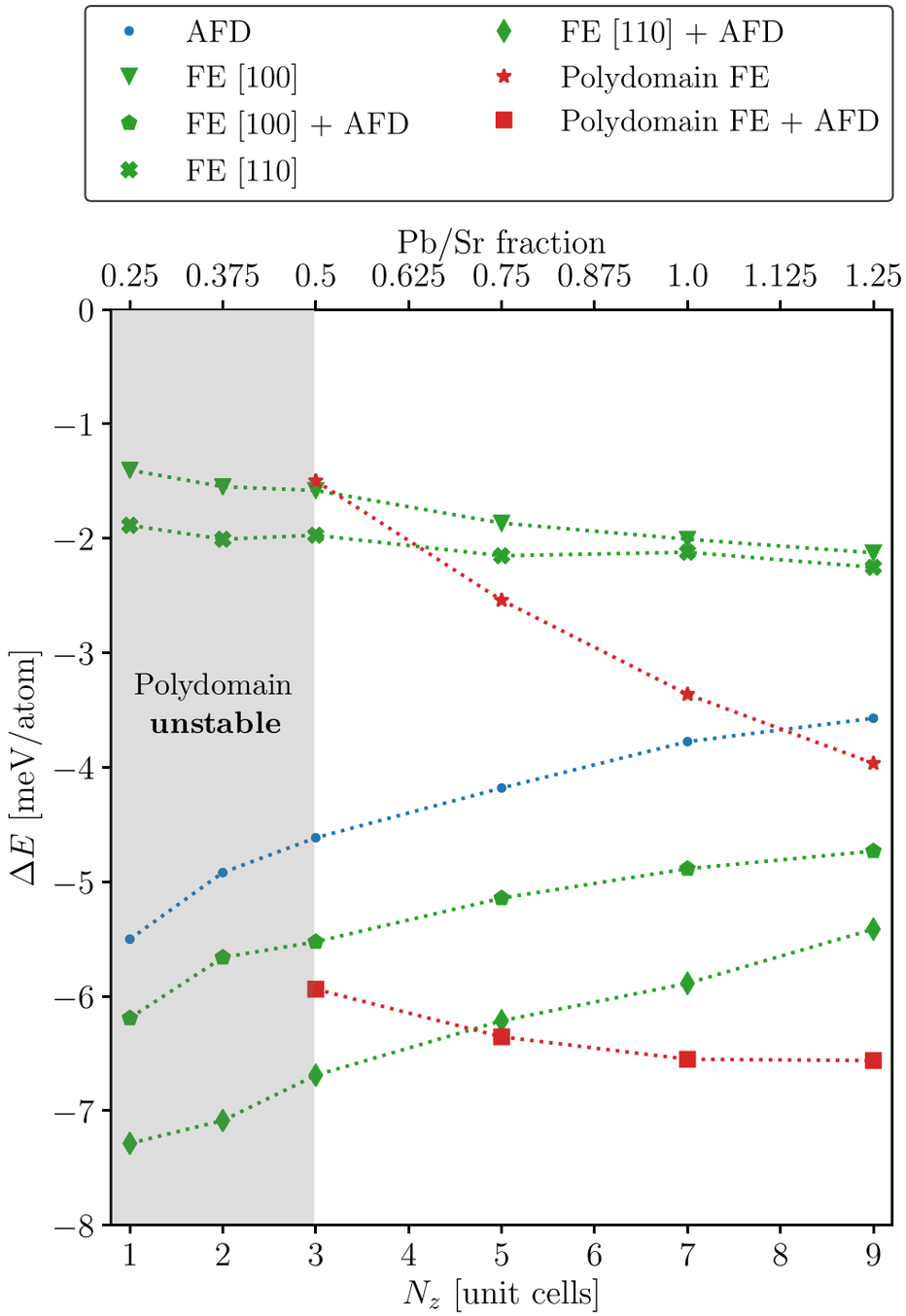
Domains in PTO/STO



- PTO on STO
- Thin film (0-9 f.u.)
- Up to ~2,000 atom cells
- 3 cell film shown
  - (a) Paraelectric
  - (b) Monodomain FE (in plane)
  - (c) Polydomain FE (out of plane)

Adv. Theory Simul. **3**, 2000154 (2020)

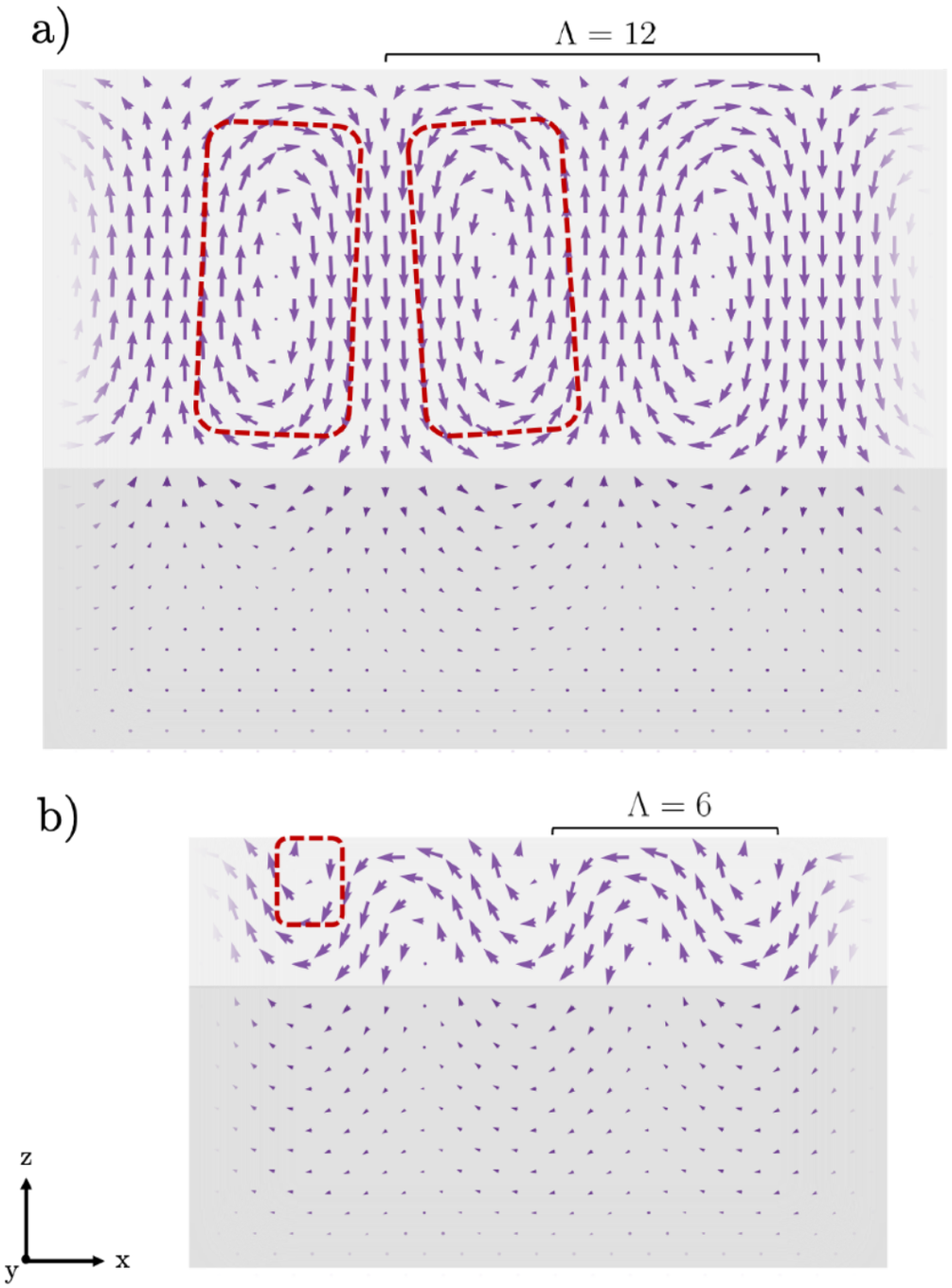
Domains in PTO/STO



- Compare FE modes to AFD
- Mono vs polydomain
- Effect of thickness
- Monodomain[110] stable (3 fu)
- Polydomain stable (>3 fu)
- Competition between:
  - Ferroelectric modes
  - Anti-ferrodistortive modes

Adv. Theory Simul. **3**, 2000154 (2020)

Domains in PTO/STO

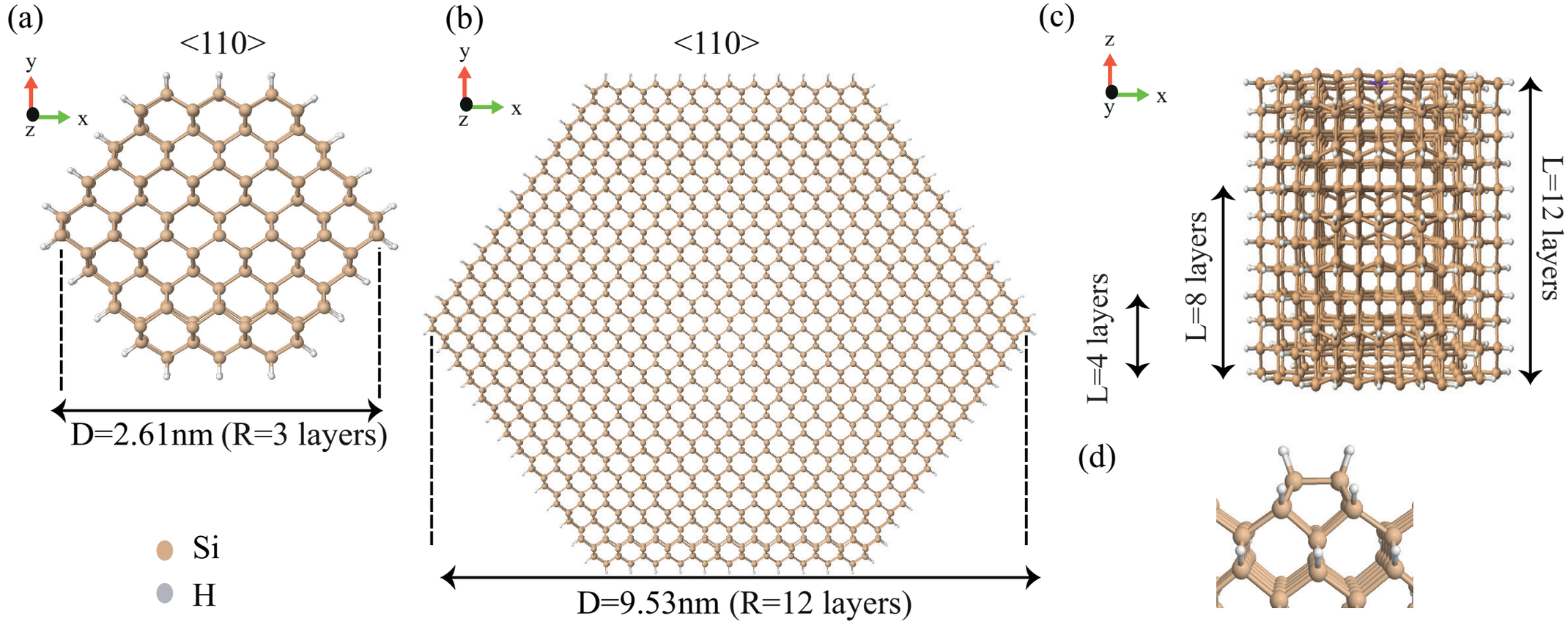


- Polarization fields
  - a) Thickness 9: flux closure
  - b) Thickness 3: polar wave
- Competition between:
  - Ferroelectric modes
  - Anti-ferrodistortive modes
- Periodicity is p(2 x Λ)

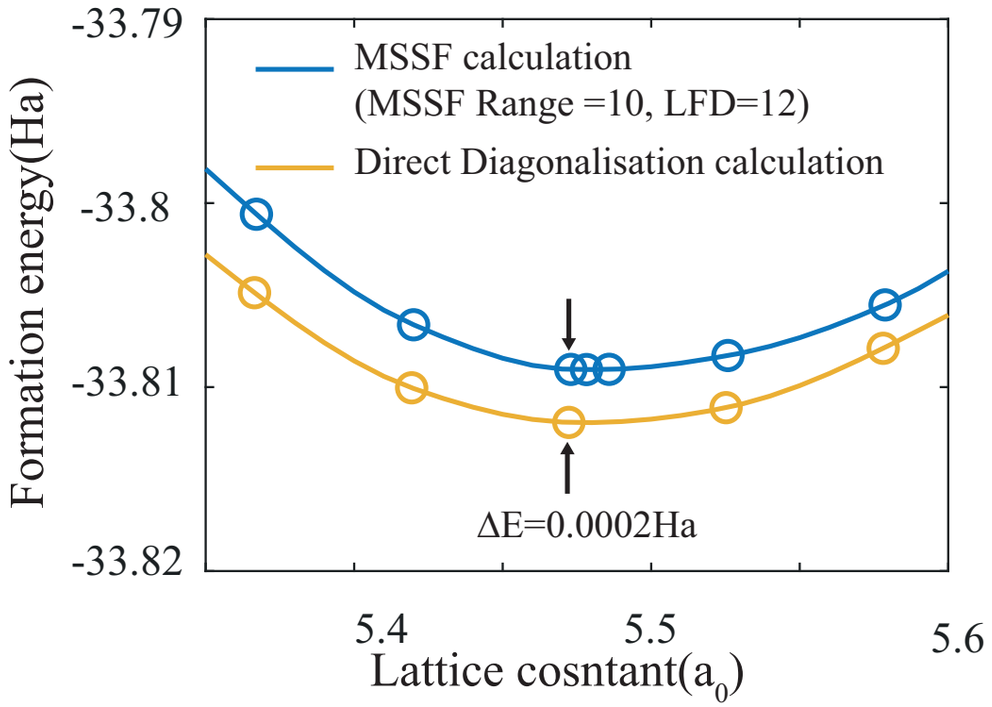
Adv. Theory Simul. **3**, 2000154 (2020)

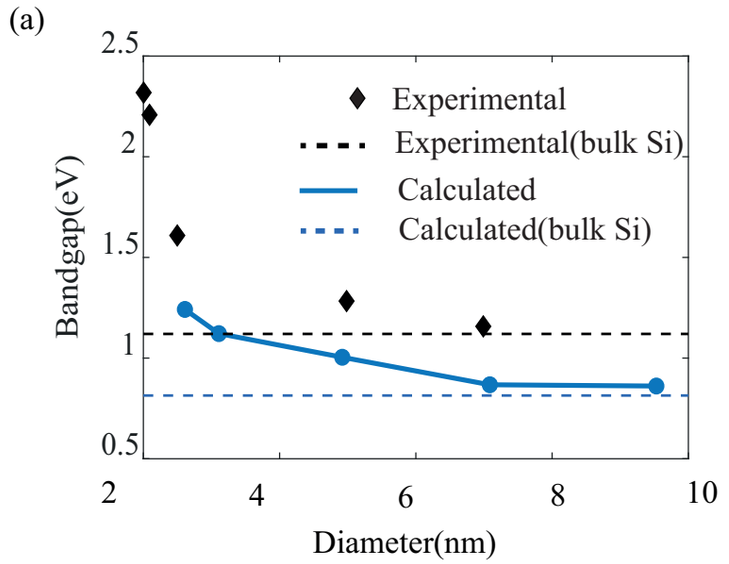
Doped Silicon Nanowires

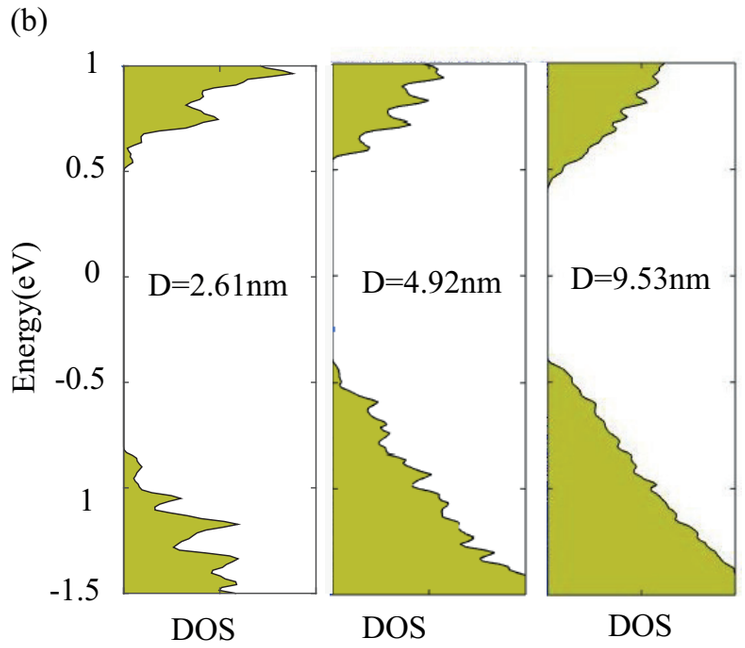
- Pure Si NWs: size (radius, repeat length), doping
- ~600-2,500 atoms (MSSF)
- Electronic structure



Pure Nanowires

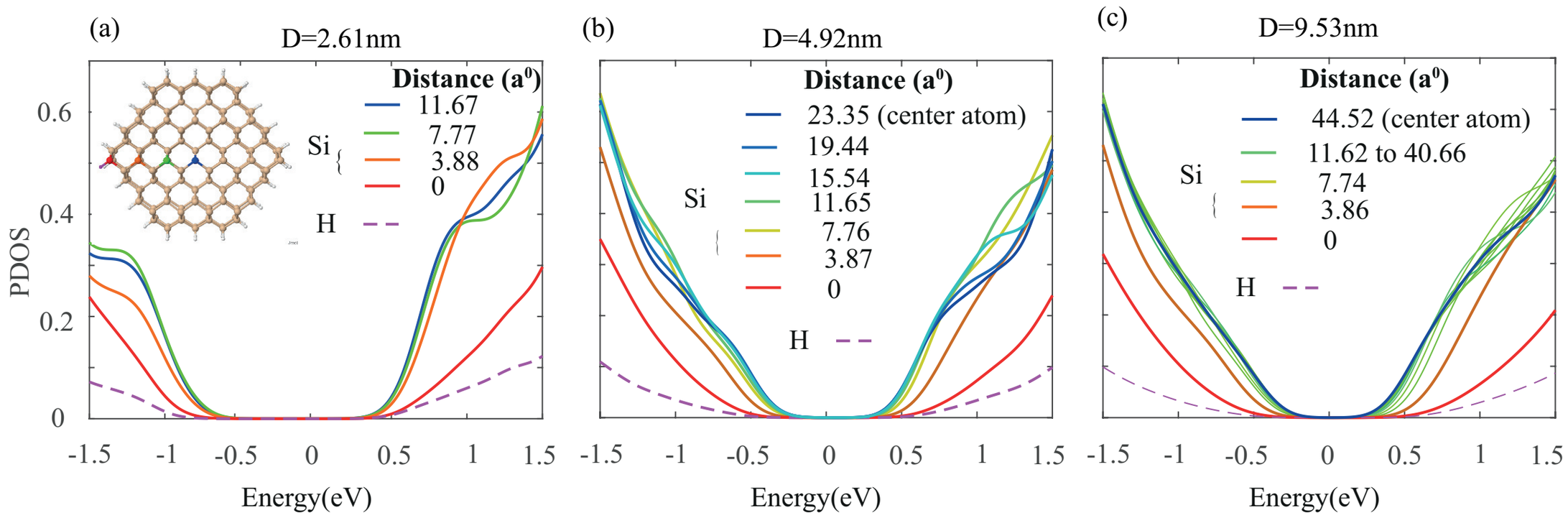






- MSSF reproduces full lattice constant
- Gap reduces with increasing radius
- Converged by ~9nm diameter

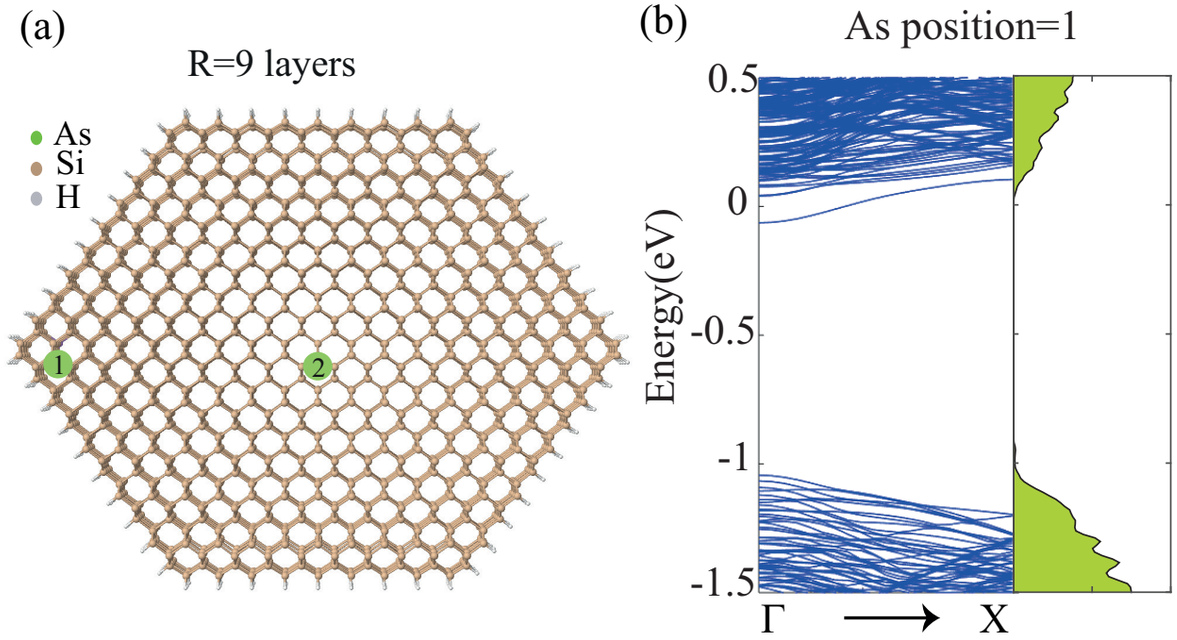
Pure Nanowires: DOS

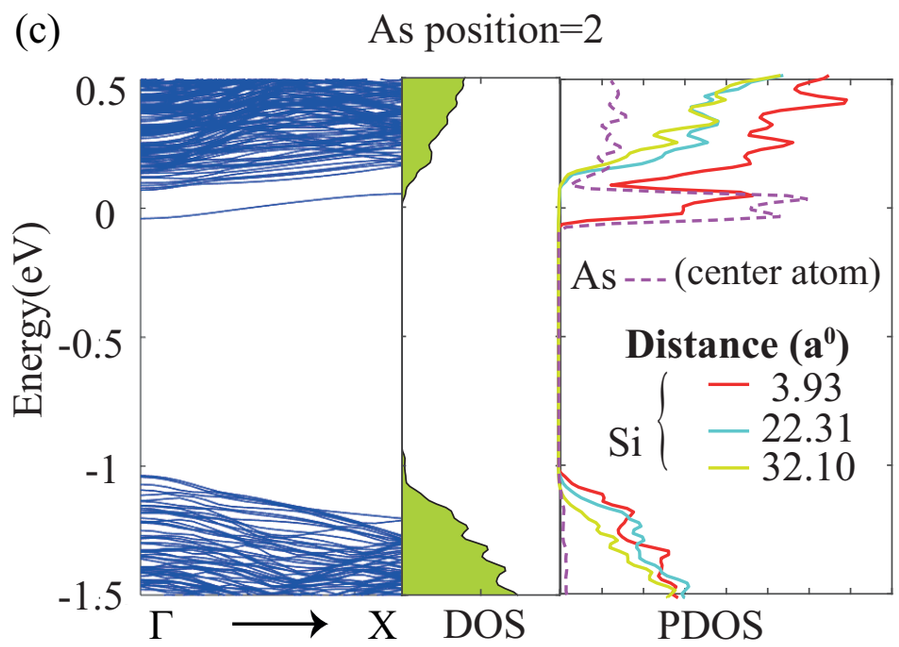


- PDOS depends on proximity to surface
- First 2-3 layers mainly

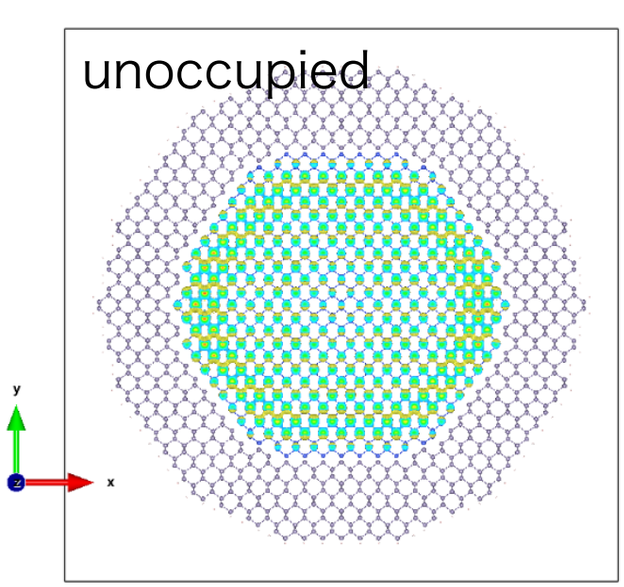
Doping: As

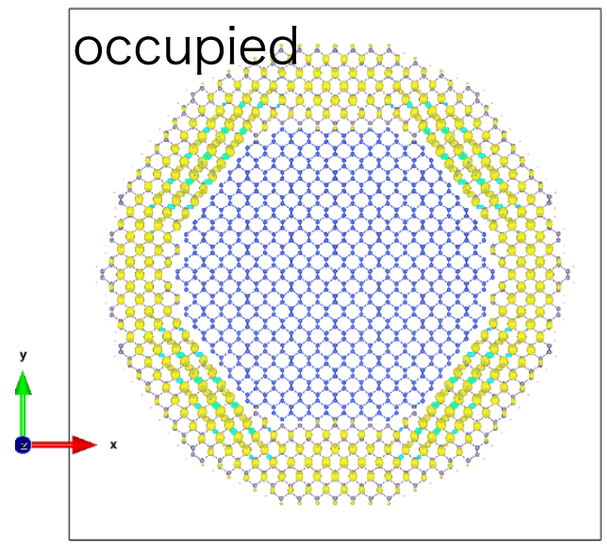
- Largest NW (7.3nm)
- Eight layers along axis
- As at edge (1) & centre (2)
- Flatter band at centre

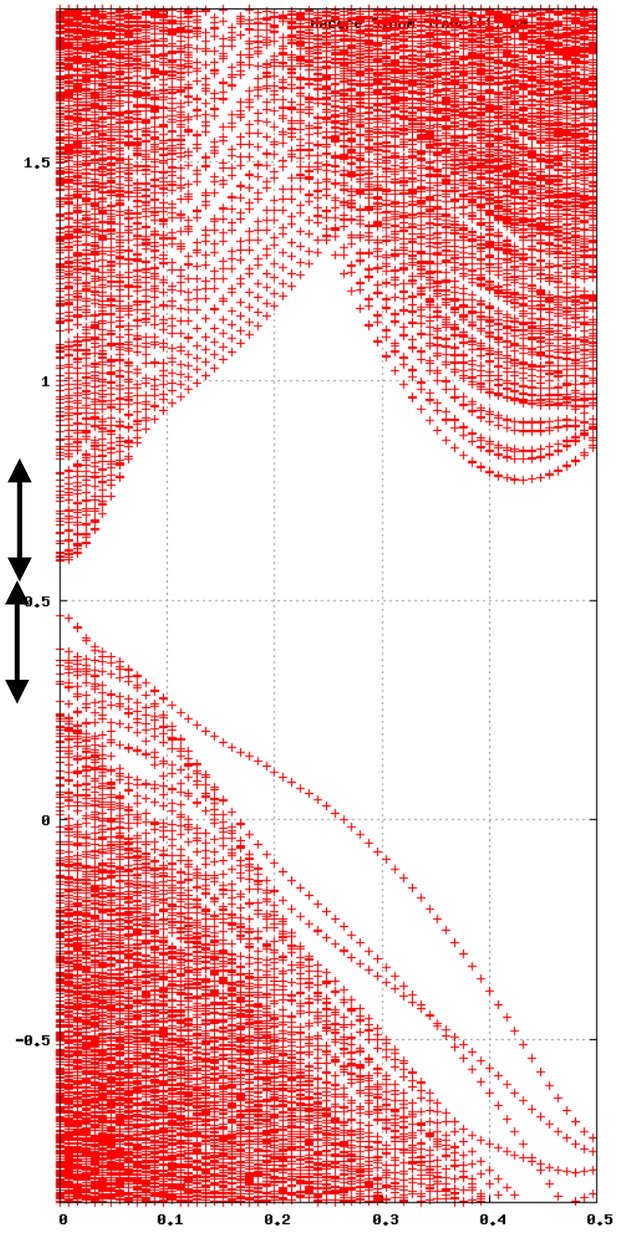




Core-shell NWs:VBM/CBM



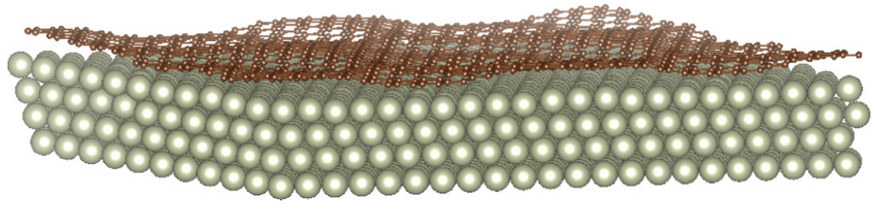


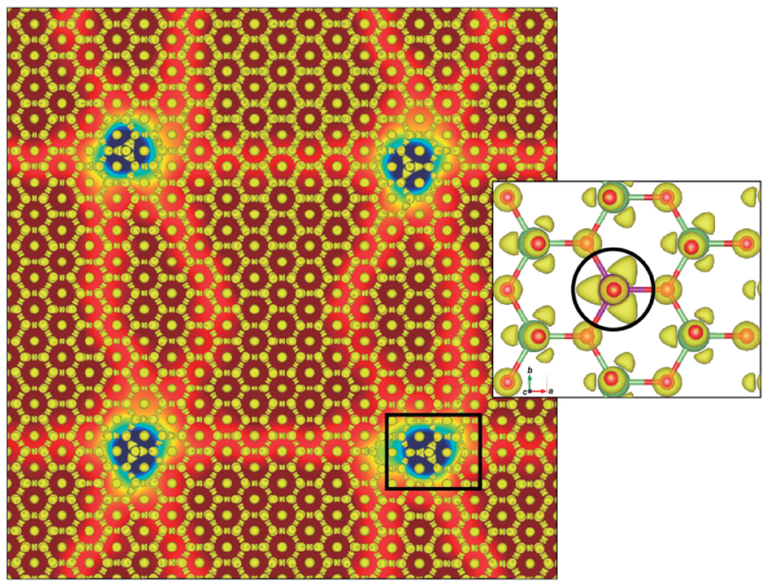


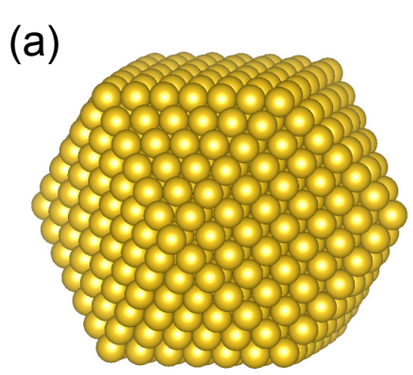
Si-Ge circular NW (large)

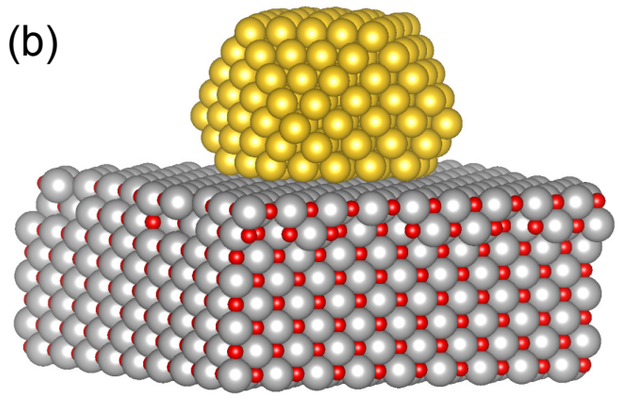
Other applications

- Graphene on Rh(111)
- Topological defects in $YGaO_3$
- Au nanoparticles









Linear scaling

How to make it local ?

Density matrix

$$\rho(\mathbf{r},\mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r}) \psi_n^\star(\mathbf{r}')$$

$$\rho(\mathbf{r},\mathbf{r}') \to 0, |\mathbf{r}-\mathbf{r}'| \to \infty$$

$$\rho(\mathbf{r},\mathbf{r}') = 0, |\mathbf{r}-\mathbf{r}'| > R_c$$

How do we represent it ?

Support functions: local in space
Ranged matrix, **K**

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i\alpha, j\beta} \phi_{i\alpha}(\mathbf{r}) K_{i\alpha, j\beta} \phi_{j\beta}(\mathbf{r}')$$

Linear scaling DFT

- Minimise total energy with respect to $\phi_{i\alpha}(\mathbf{r})$ and $K_{i\alpha,j\beta}$ subject to:

1. Correct electron number
2. Self consistency (potential, charge density)
3. Idempotency of density matrix $\rho^2 = \rho$

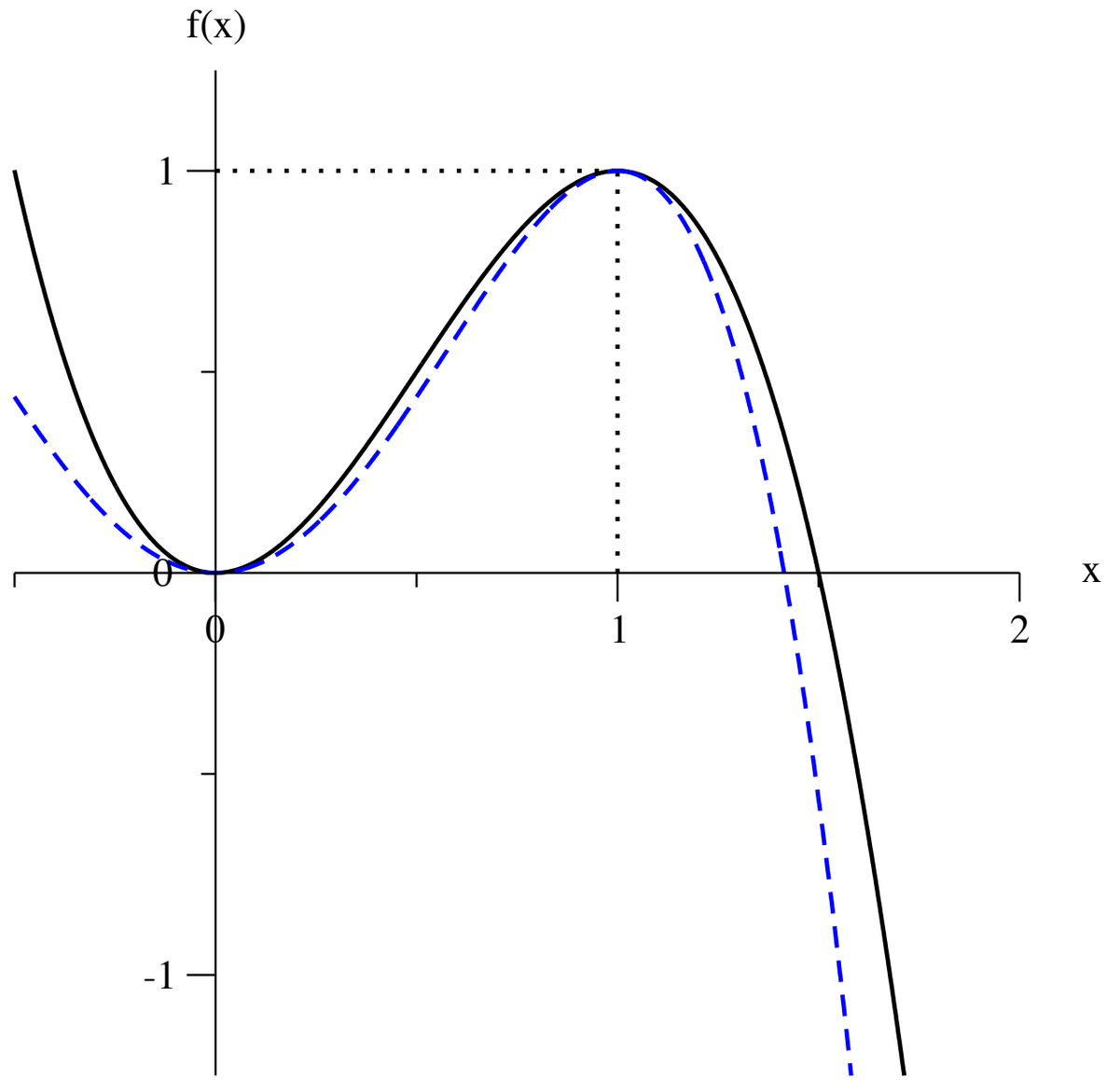
Truncating density matrix: O(N)

- The support functions are confined within a sphere, radius R_{reg}
- The **K** matrix is truncated
- All matrices are sparse, ρ is local
- By increasing radii, approach exact result

$$K_{i\alpha j\beta} = 0, |\mathbf{R}_i - \mathbf{R}_j| > R_c$$

Linear scaling DFT

- How do we find **K** ?
  - Direct methods
    - LNV
    - OMM
  - Iterative methods
    - McWeeny
    - TC etc
  - Divide-and-conquer
  - Recursion (Lanczos etc)
  - Penalty functionals
- Alternative: OFDFT



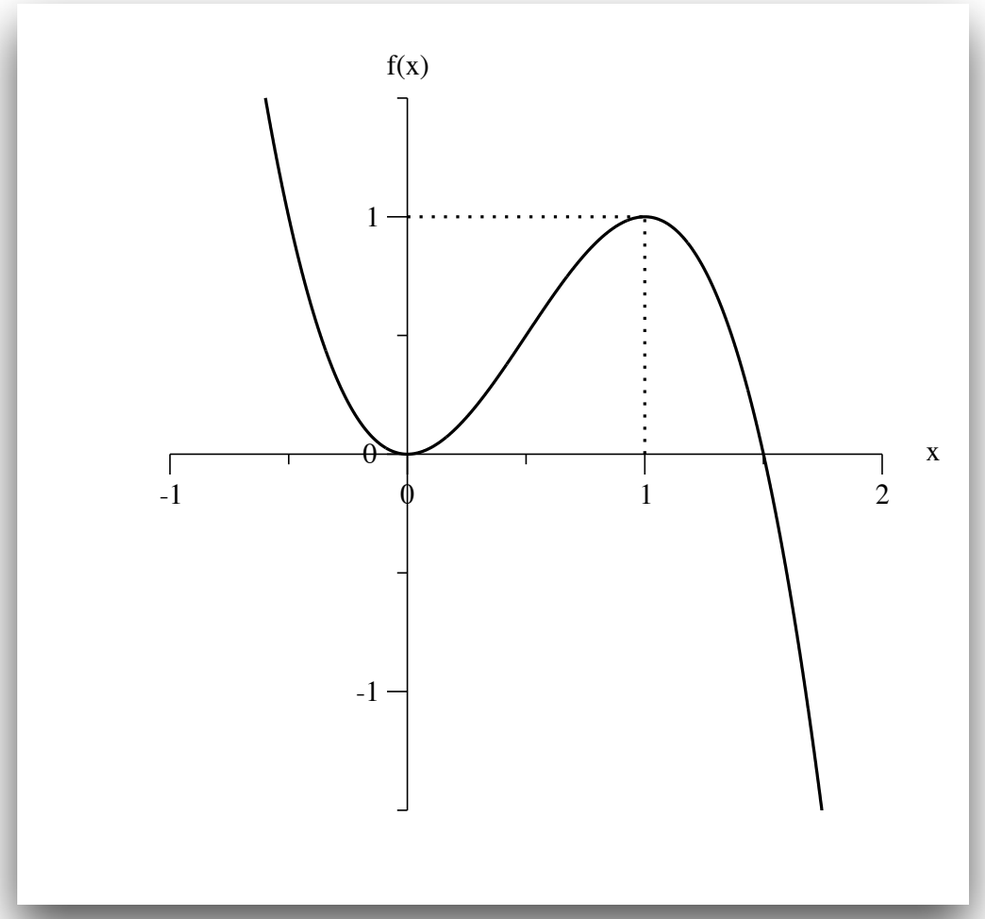
Truncation

- Spatial truncation
  - Approximate DM sparsity/range (imposed)
  - Variational
  - Consistent sparsity patterns (parallelisation)
    - LNV, OMM
- Numerical truncation
  - Approximate DM decay (impose tolerance)
  - Non-variational (?)
  - Varying sparsity patterns
  - McWeeny, TC2 etc

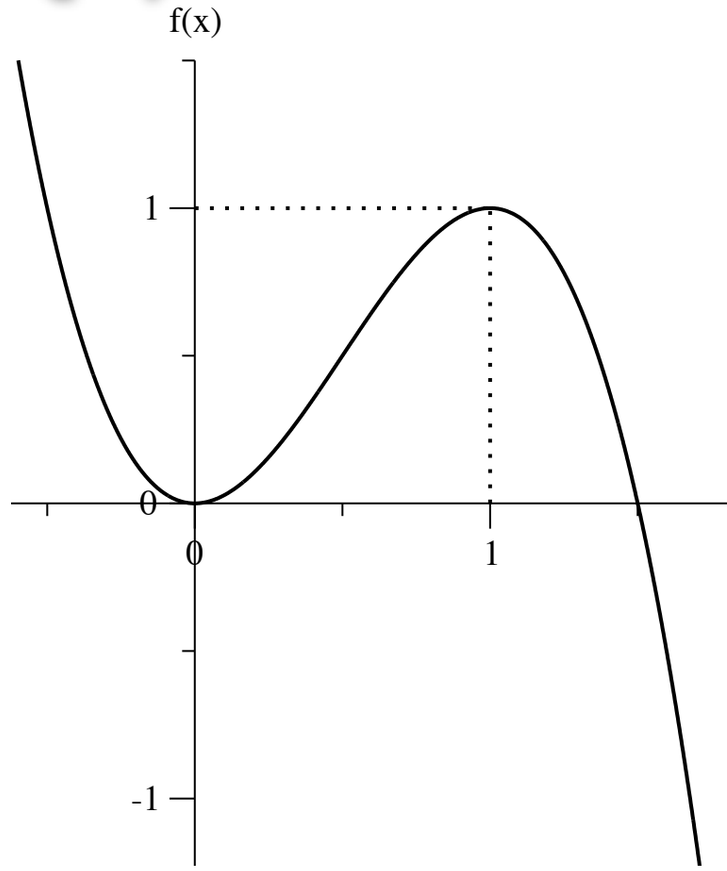
Rep. Prog. Phys. **75** 036503 (2012)

Idempotency

Hard to impose $K^2 = K$
Use approximation (McWeeny)



Minimising *K*: LNV



- Imposing idempotency is *difficult*
- We use the McWeeny transform:

$$\rho = 3\sigma^2 - 2\sigma^3$$

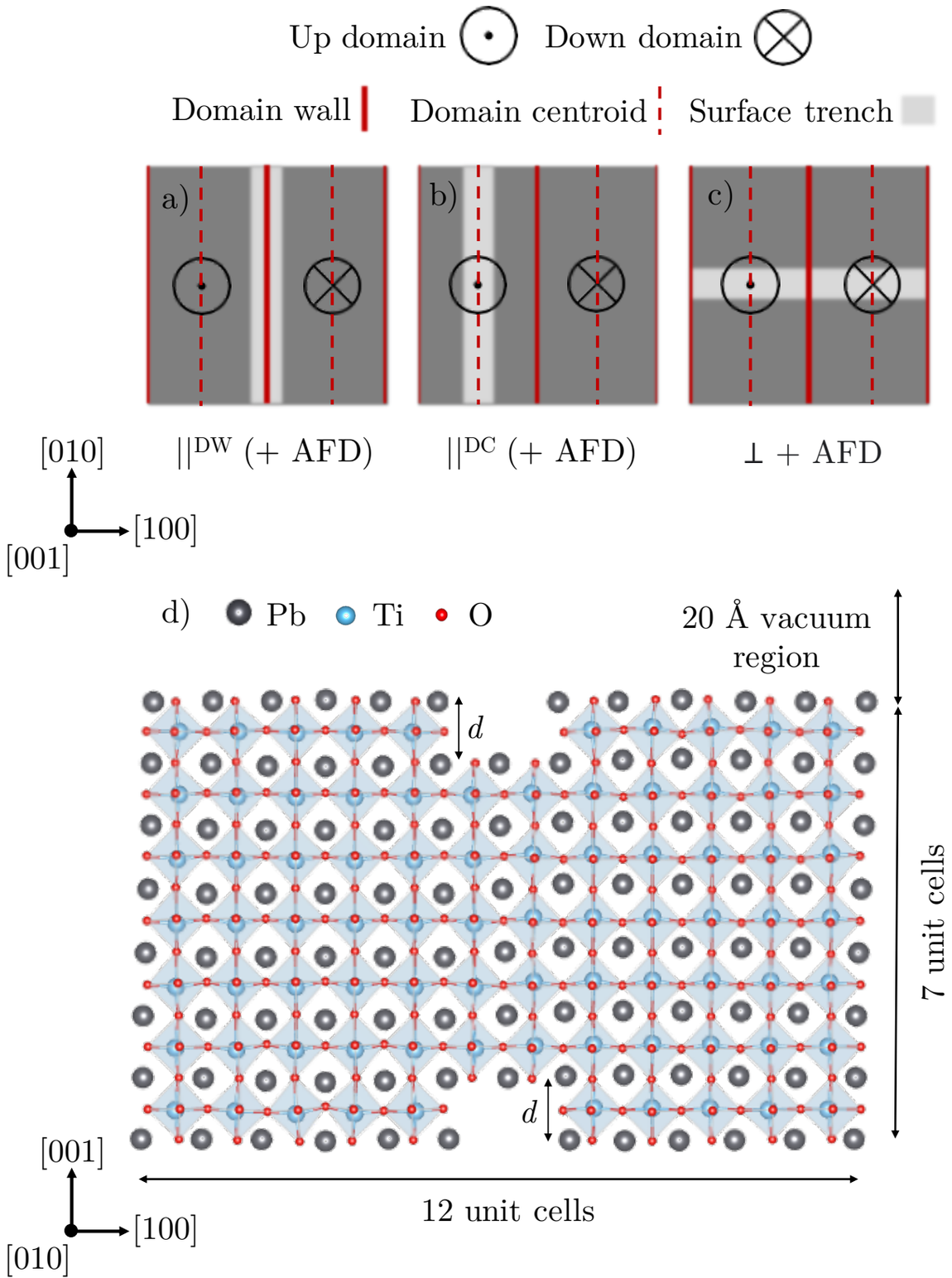
- Here σ is an *auxiliary* density matrix
- If λ_σ lie in [-0.5, 1.5] then λ_ρ will lie in [0,1]
- Vary energy with respect to elements of σ
- During minimisation, ρ tends towards idempotency

Issues to consider

- Prefactors
- Accuracy (truncation vs BZ sampling)
- Information required (eigenstates)
- Parallelisation
- Timescales (linear scaling MD and weak scaling)

Linear scaling: applications

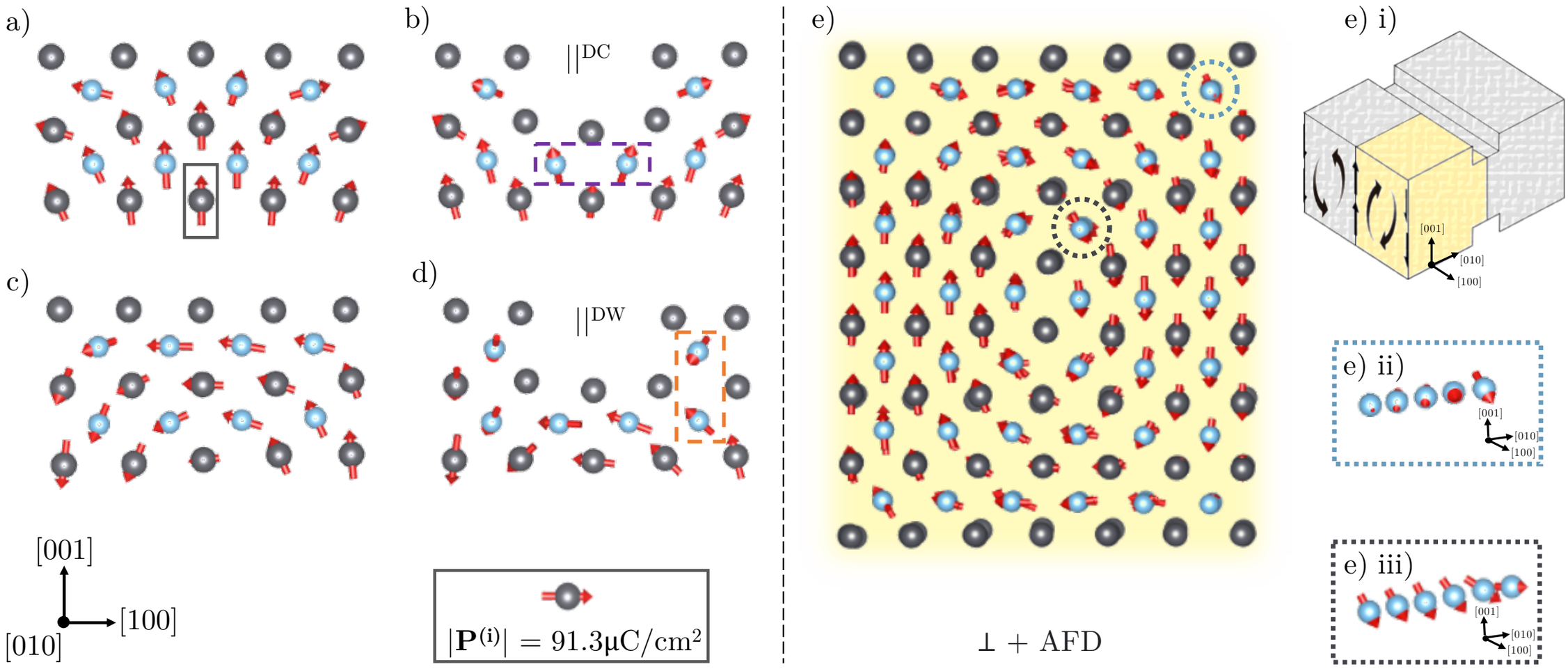
Trenches in PTO



- Domain walls align parallel to surface trenches
- PTO film strained to STO
- *O(N)* calculations
  - Up to 5,136 atoms
- Most stable parallel to DC

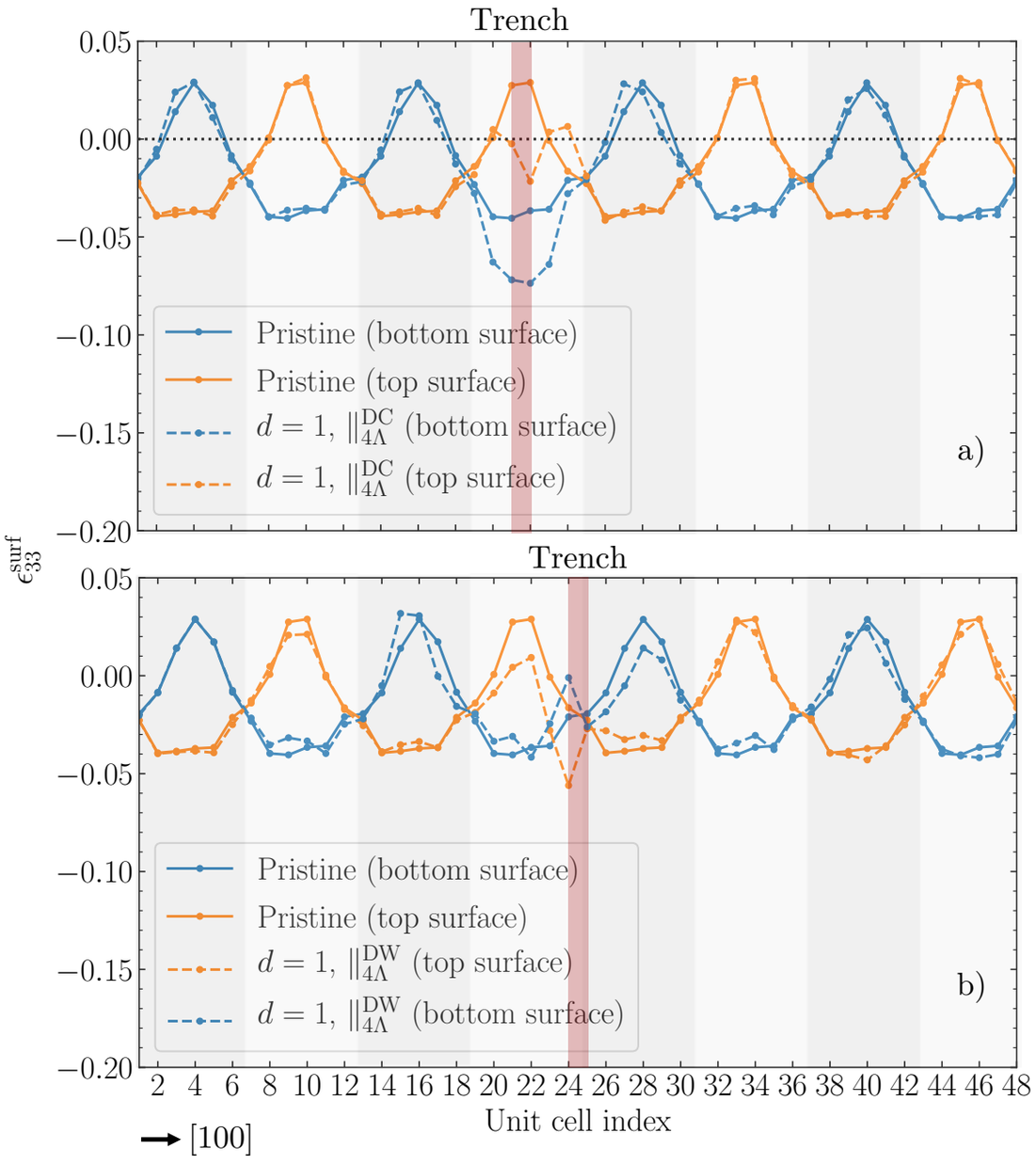
PRL **127**, 247601 (2021)

Trenches in PTO



- Local polarisation fields for the three orientations
- Key: how polar texture adapts to trench
- PRL **127**, 247601 (2021)

Trenches in PTO

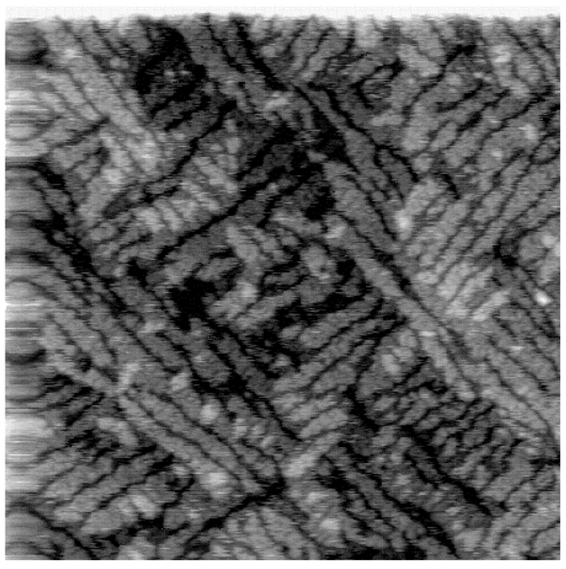


- Vertical surface strain
- DC (top) and DW (bottom)
  - Pristine (solid line)
  - Trench (dashed line)
- Cooperative strain (DC)
- More disorder (DW)

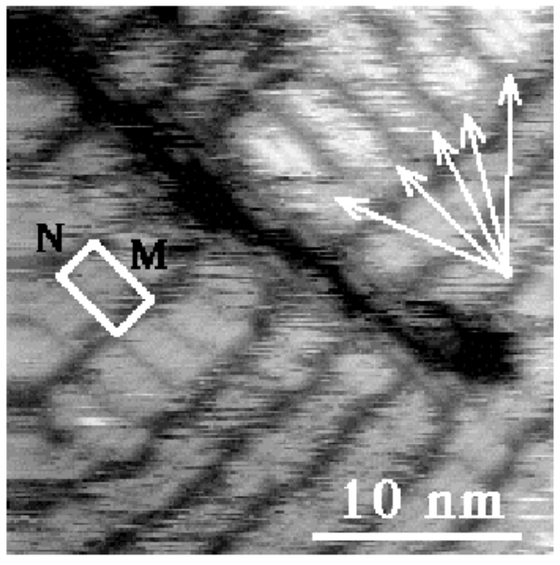
PRL **127**, 247601 (2021)

Heteroepitaxy: Ge/Si(001)

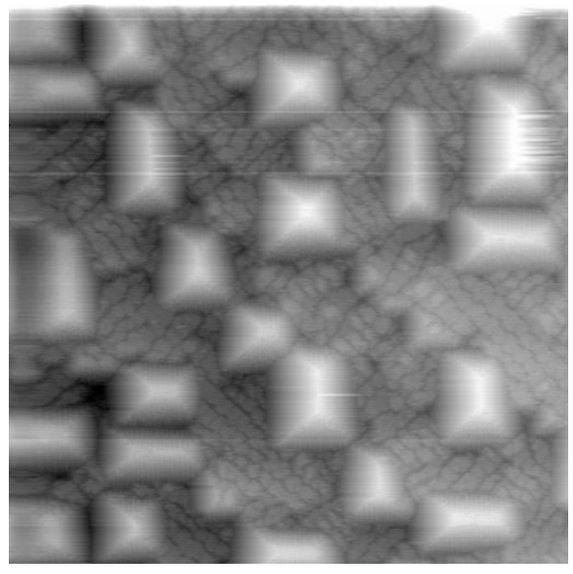
- Nanostructures & self-assembly fascinating
- Ge on Si(001) has 4.2% lattice mismatch
- After ~3ML, 3D hut clusters appear
- Is it driven by kinetics or energetics ?
- Sergiu Arapan (NIMS)



2xN



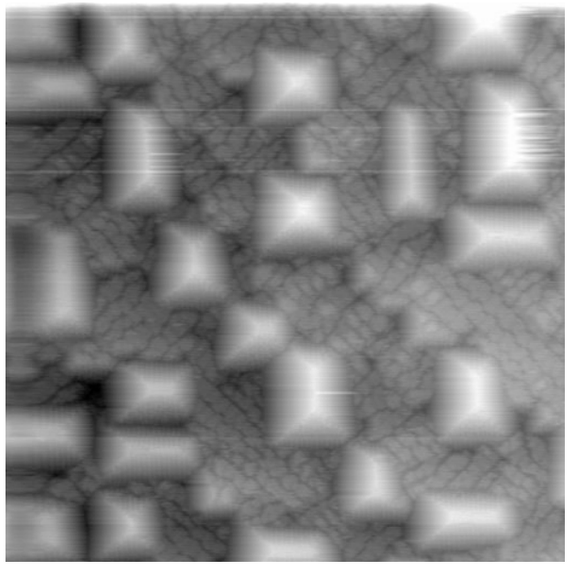
MxN



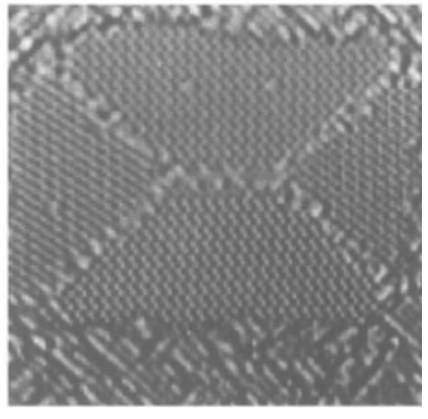
Huts

Heteroepitaxy: Ge/Si(001)

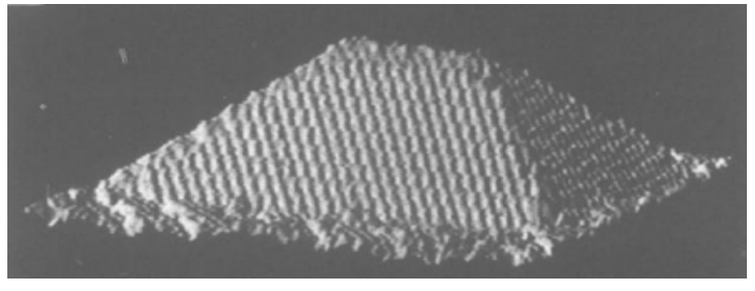
- Hut clusters are complex, 3D objects
- Experimentally facets grow top-down
- What can we learn with O(N) DFT ?



Huts



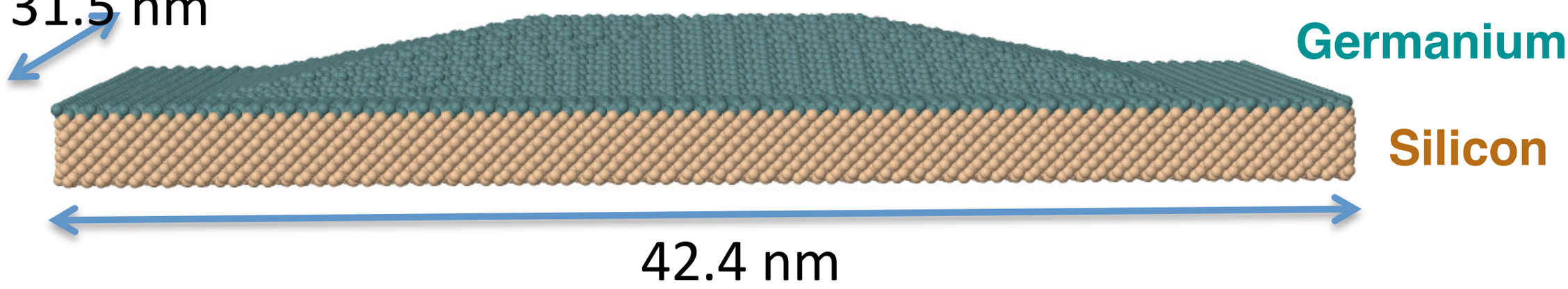
Top



Y. -W. Mo, D. E. Savage, B. S. Swartzentruber and M. G. Lagally, PRL **65**, 1020 (1990)

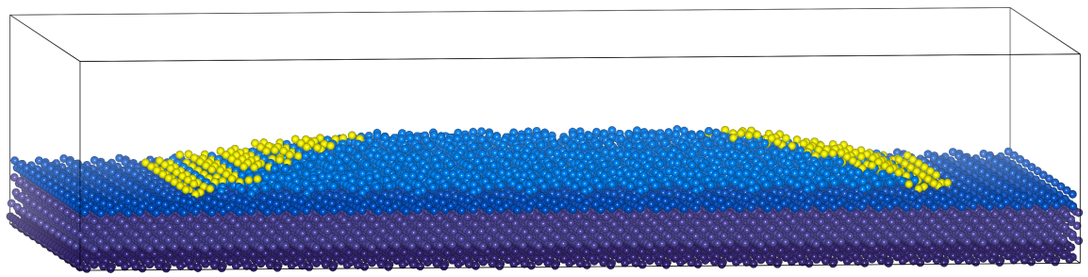
Side

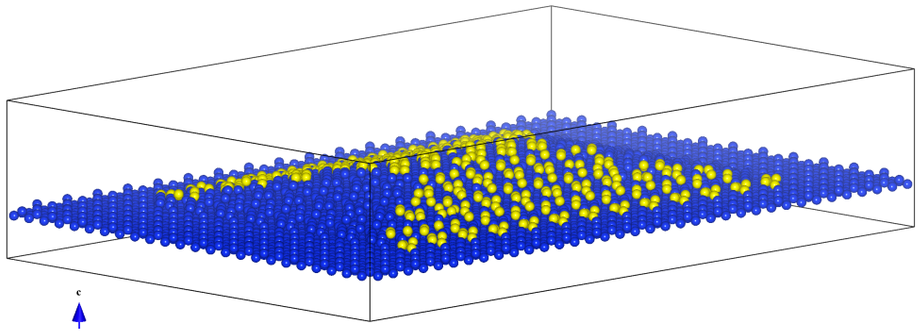
Ge/Si(001):Optimisation

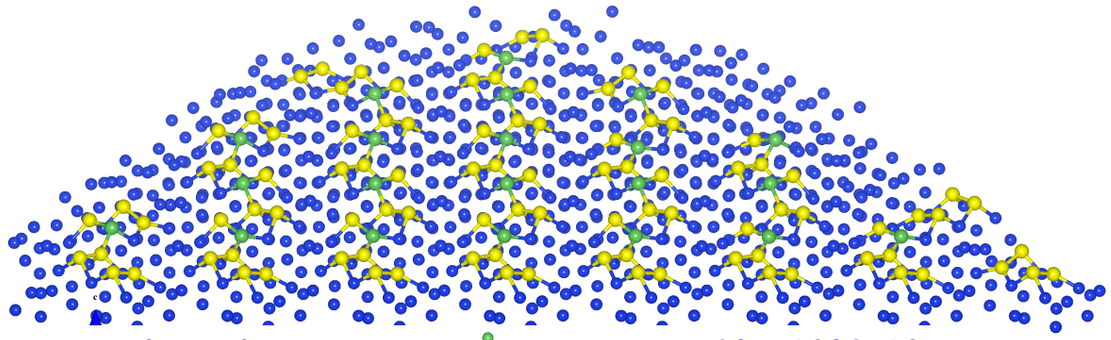


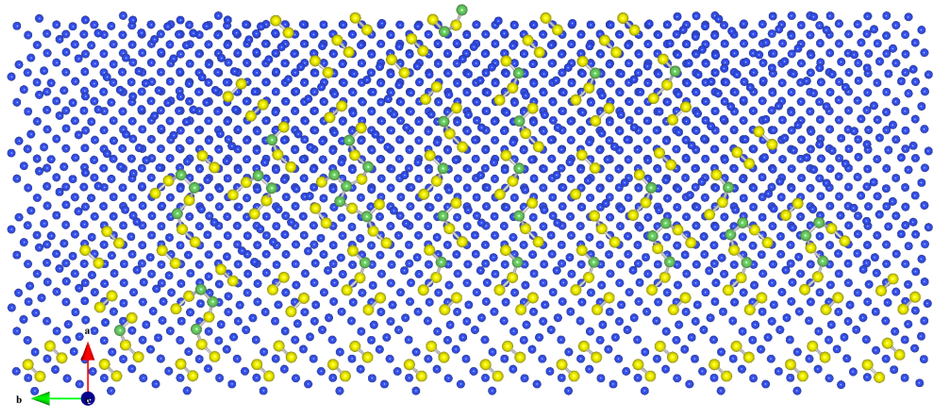
We performed structural relaxation of 200,000-atom systems with *O(N)* DFT.

Hut Cluster Growth





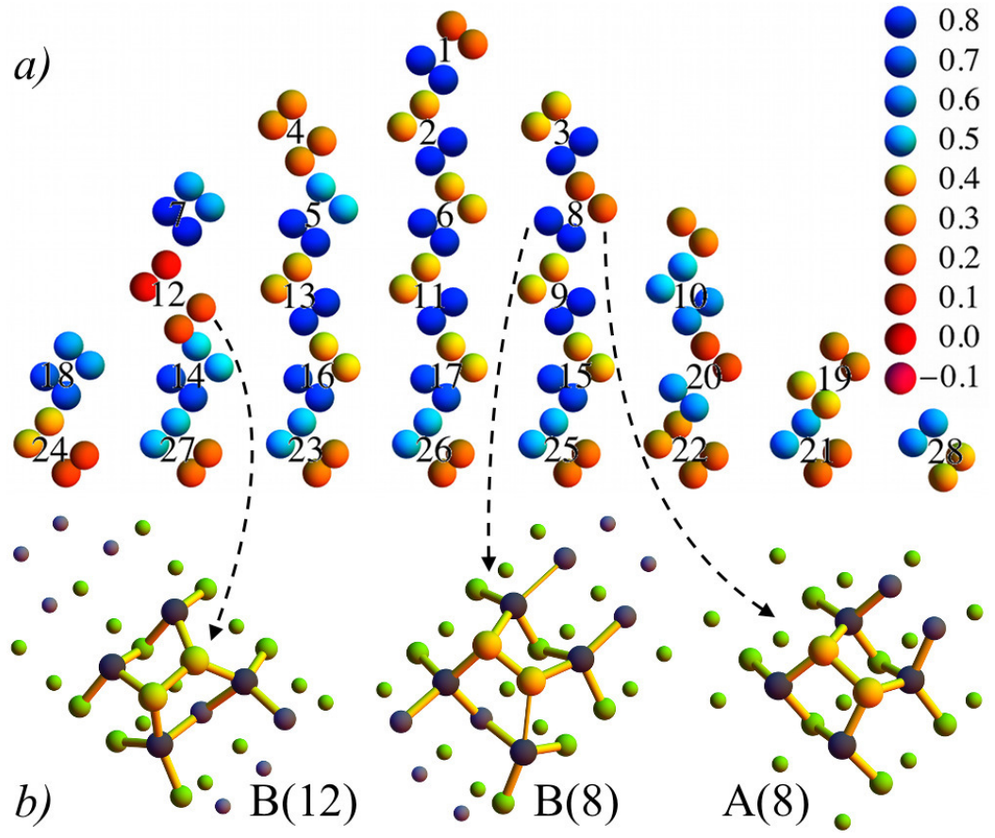




- How do huts grow ?
- Experiment shows complete new facets
- We study dimer stability: where on the face ?
- Elongated hut (blue: 8x13; yellow: 8x14)
- 6149 Ge atoms, 13824 Si atoms

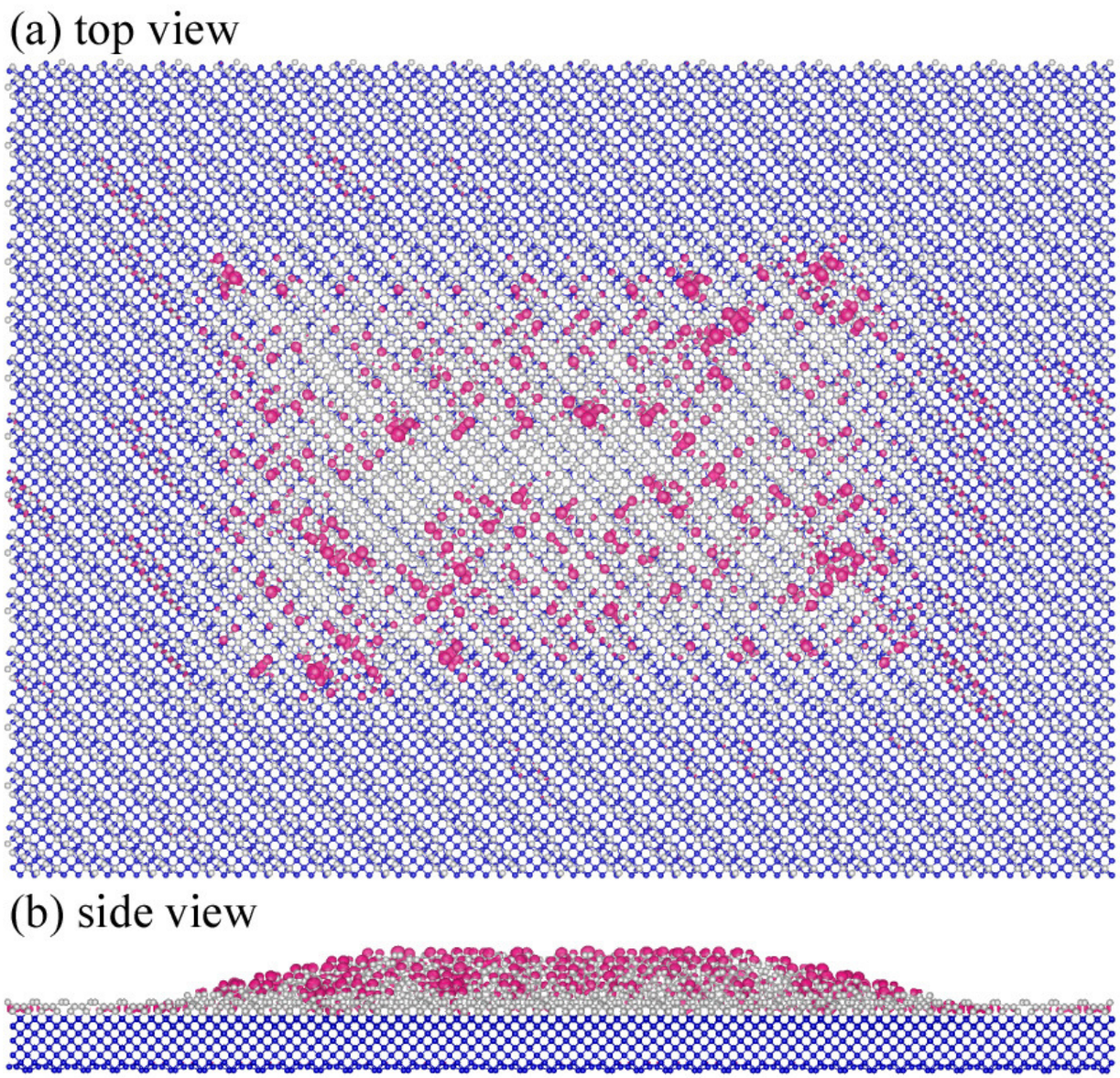
Hut Cluster Growth

- Energy map for ad-dimers
- Red is stable, blue is less stable
- Short side only
- Two types of dimer: A and B
- A (generally) more stable
- Edges, base important



Electronic structure

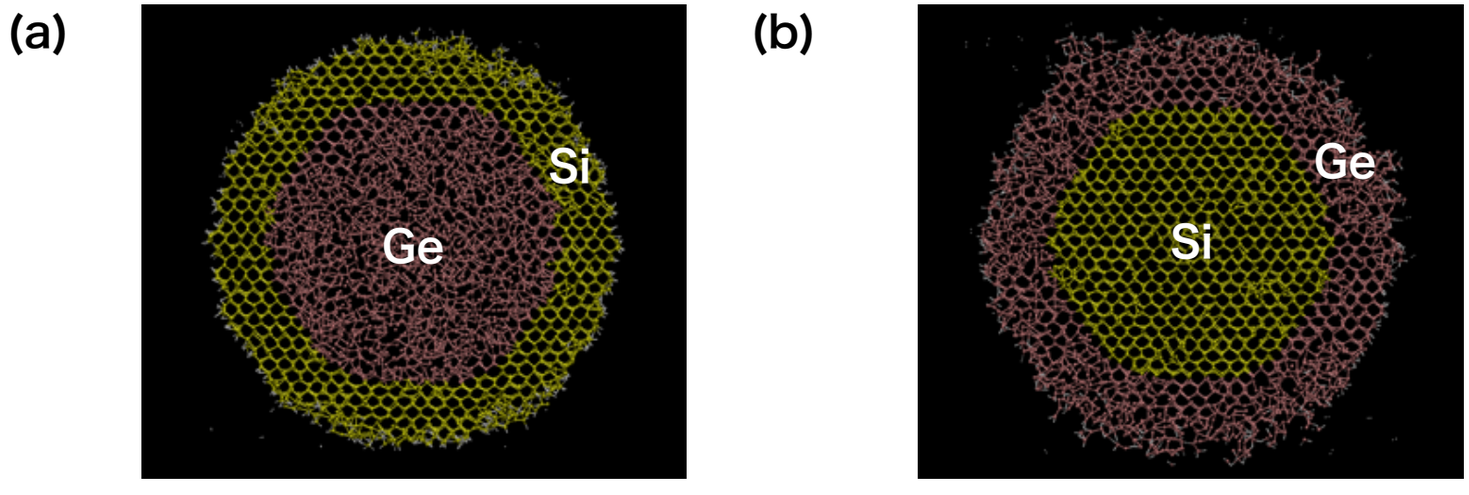
- States within 0.01eV of Fermi level
- Confined to hut
- 43 of 213,633 states



J. Chem. Theo. Comput. **13** 4146 (2018)

Linear scaling MD

- MD with *O(N)* solver
- Core-shell NWs (~5,000 atoms) at 3000K
- Ge melts first
- NPT also possible



Conclusions

- Large-scale DFT with CONQUEST
  - Up to 1,000 atoms with no restriction
  - Up to 10,000 atoms with MSSF
  - Up to 1,000,000 atoms with *O(N)* (so far)
- Application of DFT to PTO films
  - Very accurate basis sets
  - Polar domains in thin films
  - Interaction of domains with trenches

THE ROYAL SOCIETY
CELEBRATING 350 YEARS

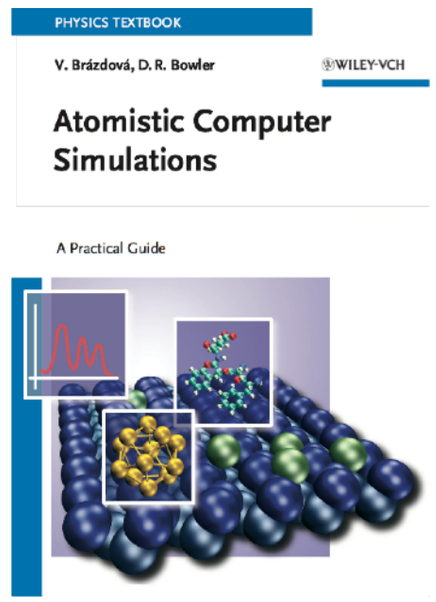
Acknowledgements

- Mike Gillan (CONQUEST originator)
- Tsuyoshi Miyazaki (CONQUEST co-leader)
- Jack Poulton, Shereif Mujahed, Zamaan Raza (PAOs)
- Ayako Nakata (MSSFs)
- Chathurangi Kumarasinghe (Si NWs)
- Jack Baker (PAOs, PTO/STO)

THE ROYAL SOCIETY
CELEBRATING 350 YEARS

References

- Linear scaling:
  - Rep. Prog. Phys. **75**, 036503 (2012)
- CONQUEST
  - J. Chem. Phys. **152** 164112 (2020)
  - PAOs: Jap. J. Appl. Phys. **58** 100503 (2019)
  - MSSF: J. Phys. Soc. Jpn. **91** 091011 (2022)
- Ferroelectrics
  - Electron. Struct. **2**, 025002 (2020)
  - Adv. Theory Simul. **3**, 2000154 (2020)
  - PRL **127**, 247601 (2021)



http://www.order-n.org/

http://www.atomisticsimulations.org/

https://github.com/OrderN/CONQUEST-release/