

A Trust Region Direct Constrained Minimization Algorithm for the Kohn-Sham Equations

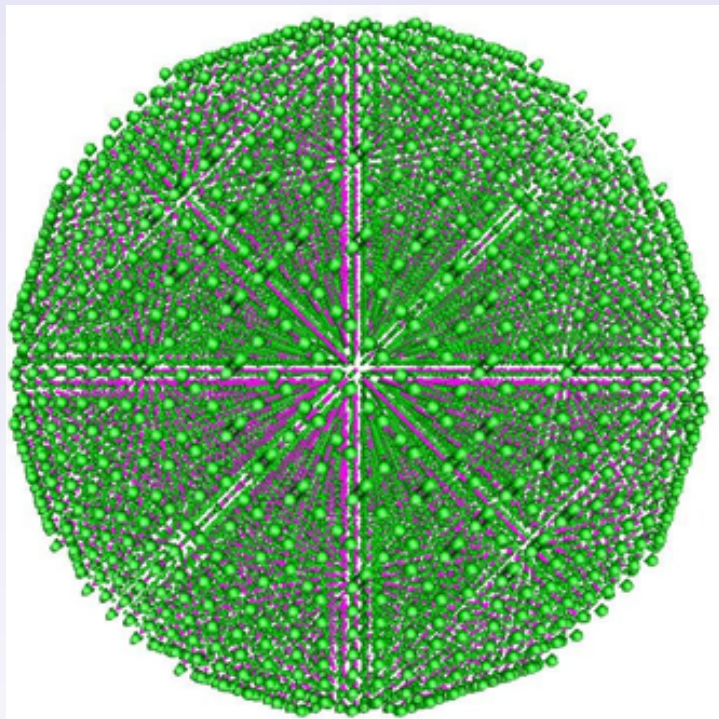
Juan Meza
School of Natural Sciences
University of California, Merced

Spring 2015

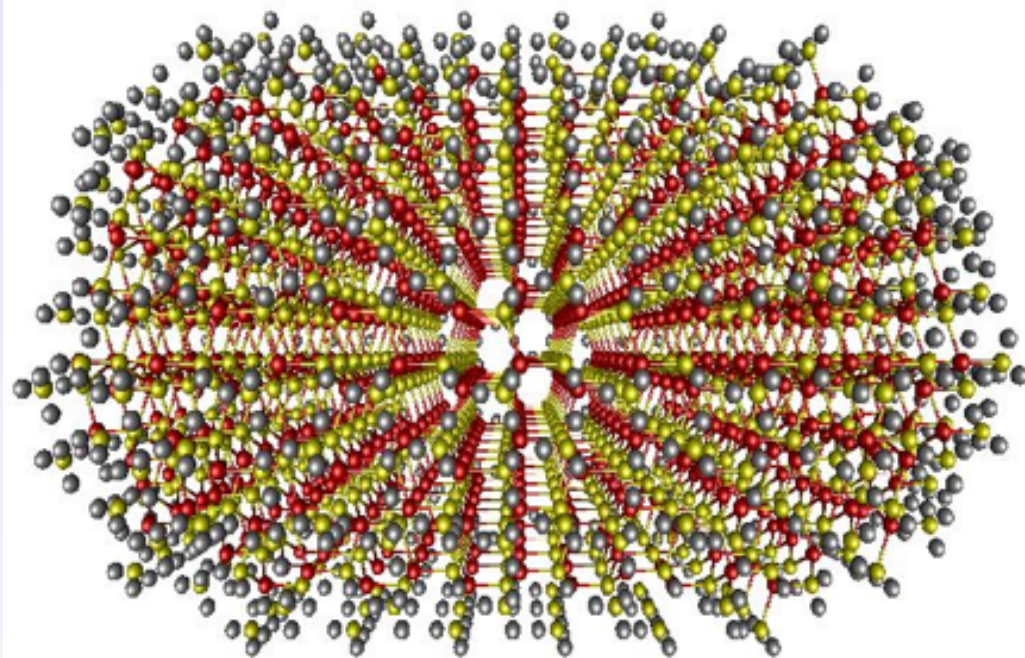
Outline

- 1 Motivation with a Short History
- 2 Density Functional Theory
 - Kohn-Sham Equations
 - Setting up the Problem
 - Solving the Kohn-Sham Equations
- 3 Self-Consistent Field (SCF) Iteration
 - Advantages/Disadvantages
 - Some Approaches for Improving Convergence
 - Trust Region SCF
- 4 Direct Constrained Minimization
 - Minimizing Energy Directly
- 5 Software
 - KSSOLV
- 6 Summary

Today we can simulate large realistic molecular structures

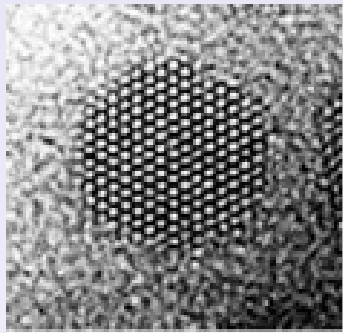


The charge density of a 15,000 atom quantum dot, $Si_{13607}H_{2236}$. Using 2048 processors at NERSC the calculation took about 5 hours.



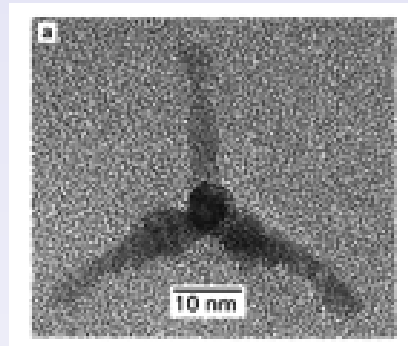
The calculated dipole moment of a 2633 atom CdSe quantum rod, $Cd_{961}Se_{724}H_{948}$. Using 2560 processors at NERSC the calculation took about 30 hours.

Materials by design



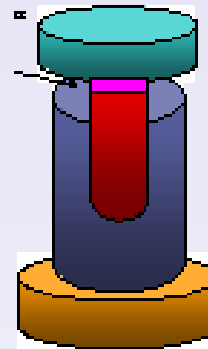
CdSe QD

1000-5000



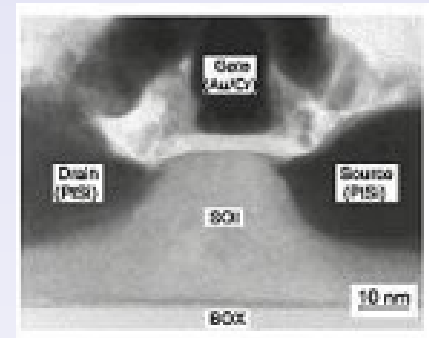
Tetrapod

10,000



Solar Cell

50,000



CMOS

1,000,000 atoms

- **Materials by design** could potentially be accelerated through the judicious use of computational and mathematical tools
- Nanosystems often involve $10^3 - 10^6$ atoms
- Atomic details are important: surface charge, impurities, dopant, symmetry, passivation, etc.

On using mathematics for chemistry

Every attempt to employ mathematical methods in the study of chemical questions must be considered **profoundly irrational** and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry — an aberration which is happily almost impossible — it would occasion a rapid and **widespread degeneration** of that science.

Auguste Comte, 1830

The problem is solved ...

...in the Schrödinger equation we very nearly have the mathematical foundation for the solution of the whole problem of atomic and molecular structure ...

G. N. Lewis, J. Chem. Phys. 1, 17 (1933).

Many-body Schrödinger equation

$$H\Psi_i(r_1, r_2, \dots, r_N) = E_i\Psi_i(r_1, r_2, \dots, r_N)$$

where the Hamiltonian is given by

$$H = -\frac{h}{2m} \sum_{i=1}^N \nabla_i^2 + \sum_{i=1}^N v(r_i) + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|r_i - r_j|}$$

- Ψ_i contains all the information needed to study a system
- $|\Psi_i|^2$ probability density of finding electrons at a certain state
- E_i quantized energy
- $3N$ unknowns, i.e. the electron positions

The problem is solved ... almost

...in the Schrödinger equation we very nearly have the mathematical foundation for the solution of the whole problem of atomic and molecular structure ...

but

the problem of the many bodies contained in the atom and the molecule cannot be completely solved without a great further development in mathematical technique.

G. N. Lewis, J. Chem. Phys. 1, 17 (1933).

Point of Reference: Computational work for solving this formulation of the Schrödinger equation is $O(10^{3N})$

Density Functional Theory

- The unknown is simple – the ground state electron density, ρ_0
- Hohenberg-Kohn Theory
 - There is a unique mapping between the ground state energy, E_0 , and the ground state density, ρ_0
 - No construction given for the functional and none known except for case of 1 electron
- DFT is an **exact theory** for many-body systems

Ideas behind Kohn-Sham Equations

- Independent particle model – electrons move independently in an average effective potential field
- Replace **many-particle** wavefunctions, Ψ_i , with **single-particle** wavefunctions, ψ_i
- We can then write Kohn-Sham total energy as:

$$E_{total}[\{\psi_i\}] = \frac{1}{2} \sum_{i=1}^{n_e} \int_{\Omega} |\nabla \psi_i|^2 + \int_{\Omega} V_{ion} \rho + \frac{1}{2} \int_{\Omega} \frac{\rho(r)\rho(r')}{|r - r'|} dr dr' + E_{xc}(\rho),$$

where $\rho(r) = \sum_{i=1}^{n_e} |\psi_i(r)|^2$, $\int_{\Omega} \psi_i \psi_j = \delta_{i,j}$, n_e is the number of electrons

Kohn-Sham Equations

- Using this formulation of the energy, it is easy to derive the **Kohn-Sham** equations as a result of minimizing the total energy with respect to the density:

$$\left[-\frac{1}{2}\nabla^2 + V_{ion}(r) + \int \frac{\rho}{|r - r'|} + V_{xc}(\rho) \right] \psi_i = \epsilon_i \psi_i, \quad i = 1, 2, \dots, n_e$$

or

$$H\psi_i = \epsilon_i \psi_i$$

with

$$H = \left[-\frac{1}{2}\nabla^2 + V_{ion}(r) + \int \frac{\rho}{|r - r'|} + V_{xc}(\rho) \right]$$

Exchange-Correlation term (short detour)

- Most of the complexity is hidden in the exchange-correlation functional, E_{xc} or equivalently V_{xc}
- Exchange arises from antisymmetry due to the Pauli exclusion principle
- Correlation accounts for other many-body effects missing from single-particle approximation, e.g. K.E. not covered by first term of Hamiltonian
- No systematic way to improve the exchange–correlation functional
- Local Density Approximation (LDA)
 - Simplest approximation to exchange–correlation term
 - Assumes energy is equal to energy from a homogeneous electron gas
 - Purely local, yet remarkably successful
 - Known limitations

Periodic Supercells (another short detour)

- **Bloch's Theorem:** In a periodic solid each electronic wave function can be expressed as the product of $\exp(ik \cdot r)$ and a periodic function, where k is a wavevector, i.e.

$$\psi_{nk}(r) = e^{ik \cdot r} \phi_{nk}(r)$$

- Can expand $\phi_{nk}(r)$ in a set of plane waves so that $\psi_{nk}(r)$ is a sum of plane waves (more in a minute)
- Bloch Theorem allows us to express the electronic wavefunctions in terms of a discrete set of plane waves

Discretization Options

- Finite difference $\psi'(r_j) \approx [\psi(r_j + h) - \psi(r_j - h)]/h$
- Finite elements

$$\psi(r) \approx \sum_j^n \alpha_j \phi_j(r), \quad \phi_j(r) \text{ nice functions with local support}$$

- Local orbital method (good for molecules)
 - Choose $\phi_j(r)$ as Gaussian or other “nice” functions
- Planewave expansion
 - Choose $\phi_j(r)$ as $e^{ig_j \cdot r}$

Leads to a plane-wave basis set for our system

- Write

$$\psi_{nk}(r) = \sum_j \alpha_j e^{i(k+g_j)\cdot r}$$

- In principle, you need an infinite plane wave basis set
- In practice, you introduce an *energy cutoff* to truncate the basis set
- As a bonus, the kinetic energy term of Hamiltonian is diagonal when using a plane-wave basis set

Finally ready to state our finite dimensional problem

Recall we want to

$$\min E[\{\psi_i\}] = \frac{1}{2} \sum_{i=1}^{n_e} \int_{\Omega} |\nabla \psi_i|^2 + \int_{\Omega} V_{ion} \rho + \frac{1}{2} \int_{\Omega} \frac{\rho(r) \rho(r')}{|r - r'|} dr dr' + E_{xc}(\rho)$$

After discretization we have

$$\min_{X^* X = I_{n_e}} E_{total}(X) \equiv E_{kinetic}(X) + E_{ion}(X) + E_{Hartree}(X) + E_{xc}(X),$$

$$E_{kinetic} = \frac{1}{2} \text{tr}(X^* L X)$$

$$E_{ionic} = \text{tr}(X^* V_{ion} X)$$

$$E_{Hartree} = \frac{1}{2} \rho(X)^T L^\dagger \rho(X)$$

$$E_{xc} = e^T (\epsilon_{xc}[\rho(X)])$$

$$\rho(X) = \text{diag}(X X^*)$$

$$L = \text{is the discretized Laplacian}$$

Pseudopotentials (final detour)

- Introduction of energy cutoff allowed us to truncate the plane-wave basis set
- Problem is still too large for systems that contain both valence and core electrons
- Pseudopotentials allow us to use a much smaller number of plane-wave basis thereby reducing the computational cost
- Based on idea that most chemistry is dependent on valence electrons rather than core electrons
- Therefore we replace the core electrons (and the ionic potential) with a weaker pseudopotential
- Rest is outside our scope

Lagrangian for the KS system

- Minimization problem can then be written as

$$\min_{X^*X=I_{n_e}} E_{total}(X)$$

- Associated Lagrangian for minimization problem can be written as

$$\mathcal{L}(X) = E_{total}(X) - \text{tr} [\Lambda^T (X^*X - I_{n_e})]$$

- Any solution must satisfy the first order KKT condition

$$\begin{aligned}\nabla_X \mathcal{L}(X, \Lambda) &= 0, \\ X^*X &= I_{n_e}.\end{aligned}$$

Which then leads to

- A little bit of algebra then leads to the following set of equations called the (discretized) **Kohn-Sham equations**:

$$\begin{aligned} H(X)X &= X\Lambda, \\ X^*X &= I_{n_e}. \end{aligned}$$

- where the Kohn-Sham Hamiltonian is given by:

$$H(X) = \frac{1}{2}L + V_{ion} + \text{Diag} (L^\dagger \rho(X)) + \text{Diag} (\mu_{xc}(\rho(X)))$$

- As before, L refers to the discretized Laplacian
- X is an $N \times n_e$ matrix
- $\text{Diag}(\cdot)$ is a diagonal matrix

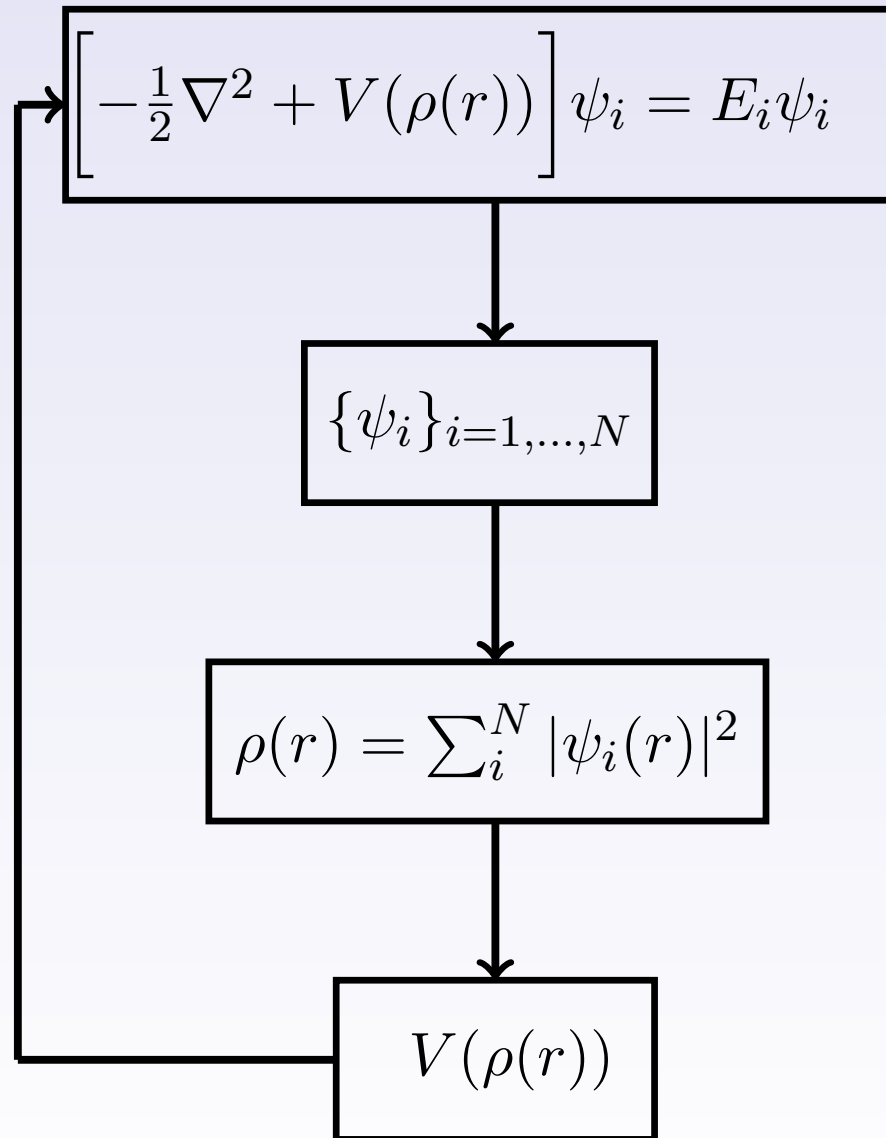
Important Observation

- If X is a solution to the **Kohn-Sham equations** then so is XQ , where $Q \in \mathbb{C}^{k \times k}$ and $Q^*Q = I_k$
- Another way of viewing this is that the solution to the constrained minimization problem is a k -dimensional **invariant subspace** in \mathbb{C}^n and not just a single matrix.
- We will come back to this property when we use trust-regions to globalize the minimization algorithms.

Two main approaches for solving the Kohn-Sham Equations

- Many methods - we don't have the time to discuss them all today
- Work with the KS equations directly
 - Self-Consistent Field (SCF) iteration
 - View as large-scale linear eigenvalue problem
 - Need to precondition
 - Need other acceleration techniques to improve convergence
- Minimize the total energy directly
 - Direct Constrained Minimization (DCM)
 - Constrained optimization problem
 - Also requires globalization techniques

The SCF Iteration in a nutshell



- Most of the work is in solving the linear eigenvalue problem
- If using reciprocal (Fourier) space, then you also have many 3D FFTs
- For large systems, the calculation of nonlocal potentials can also be expensive
- SCF does **NOT** decrease the energy monotonically

Advantages/Disadvantages of DFT

- By some counts DFT methods account for 75% of the material science simulations at supercomputer centers
- Parallel efficiencies can be quite high
 - on plane wave basis can scale to ≈ 1000 processors
 - on plane wave basis and wavefunction index can scale to $\approx 10,000$ processors
- Not systematically improvable
- Inadequate for strong and/or non-local correlations

So what's the problem?

The main workhorse in DFT codes is the SCF algorithm, which can converge slowly and sometimes doesn't converge at all!

When can we expect SCF to work?

SCF convergence properties

- Surprisingly few theoretical results
- $E(x)$ may not monotonically decrease between SCF iterations
- SCF does not always converge;

$$\lim_{i \rightarrow \infty} \|H(x^{(i+1)}) - H(x^{(i)})\| \neq 0,$$

or

$$\lim_{i \rightarrow \infty} \|\rho(x^{(i+1)}) - \rho(x^{(i)})\| \neq 0$$

- For some problems, we can show subsequence convergence;

$$\lim_{i \rightarrow \infty} \|\rho(x^{(i+1)}) - \rho(x^{(i-1)})\| = 0$$

Example

$$E(x) = \frac{1}{2}x^T Lx + \frac{\alpha}{4}\rho(x)^T L^{-1}\rho(x)$$

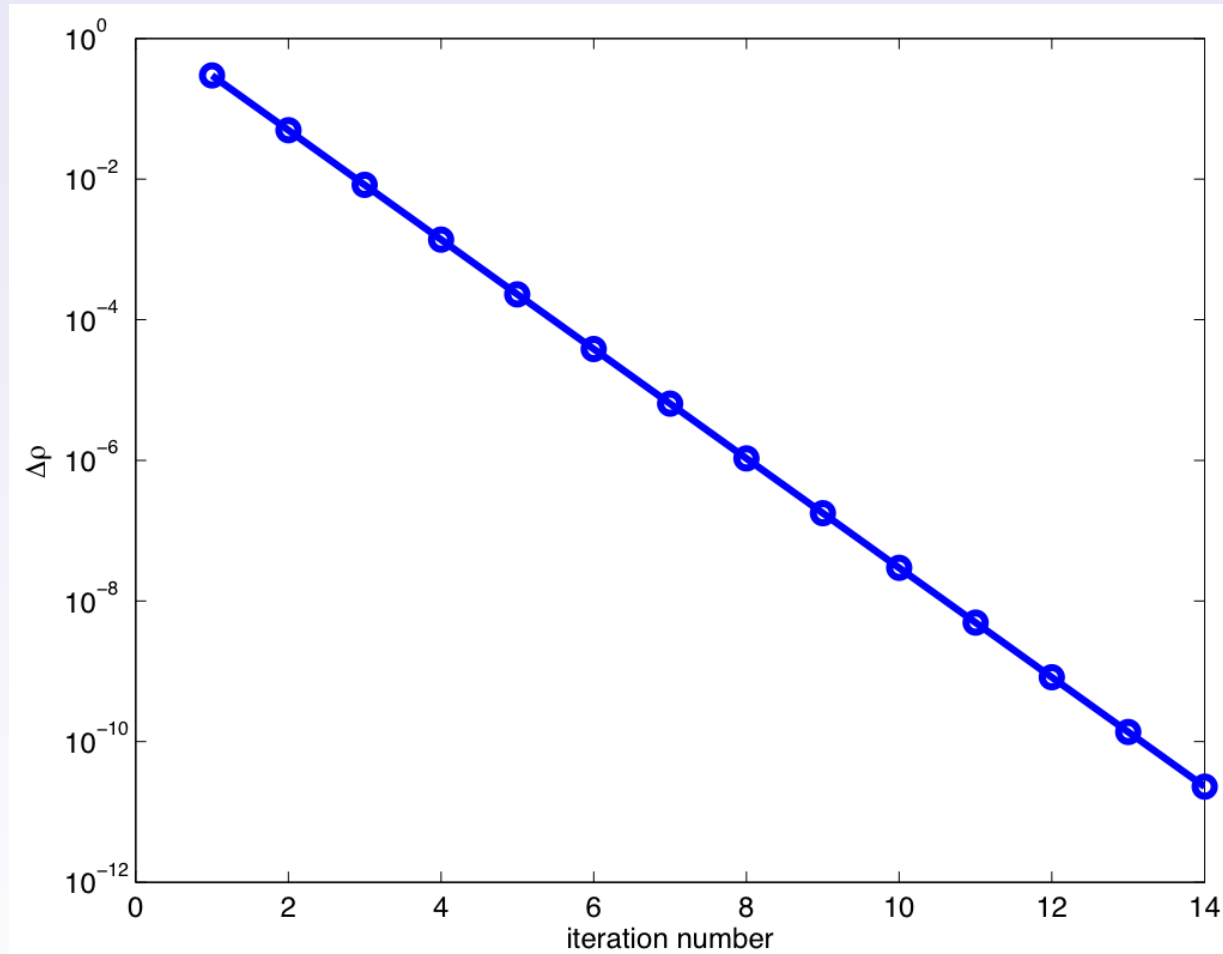
$$L = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \rho(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$$

$$\begin{array}{ll} \min E(x) \\ \text{s.t. } x_1^2 + x_2^2 = 1 \end{array}$$

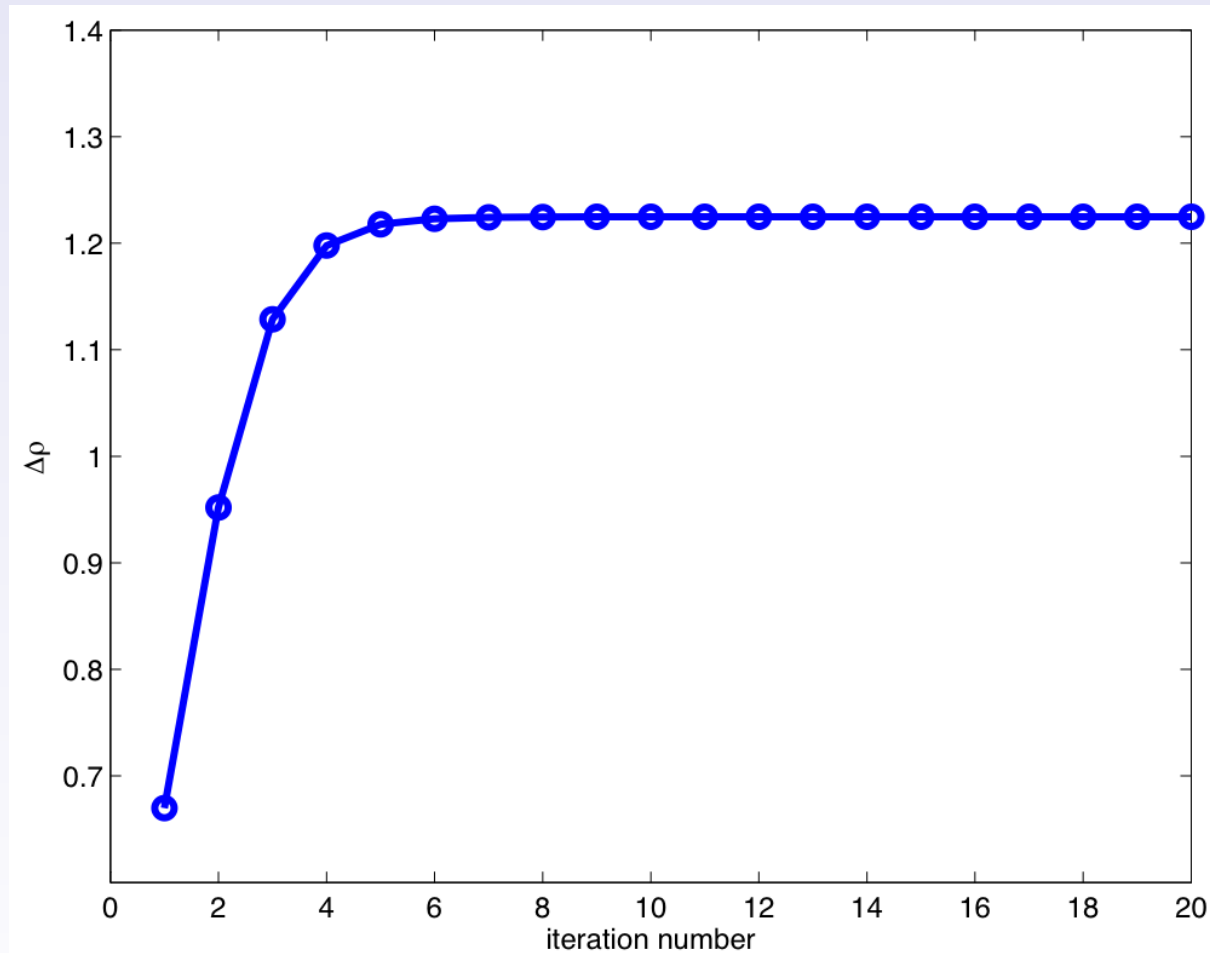
$$\left[L + \alpha \text{Diag}(L^{-1}\rho(x)) \right] x = \lambda_1 x$$

SCF Converges when $\alpha = 1.0$

$$\Delta\rho^{(i)} = \|\rho^{(i)} - \rho^{(i-1)}\|$$



SCF fails when $\alpha = 12.0$



Why does SCF fail?

- Finding k smallest eigenvalues is equivalent to solving the trace minimization problem
- SCF can be viewed as attempting to minimize a sequence of surrogate models $q(x)$:

$$\begin{aligned} \min q(x) &= \frac{1}{2} \operatorname{tr} (X^* H^{(i)} X) \\ \text{s.t. } X^* X &= I_k \end{aligned}$$

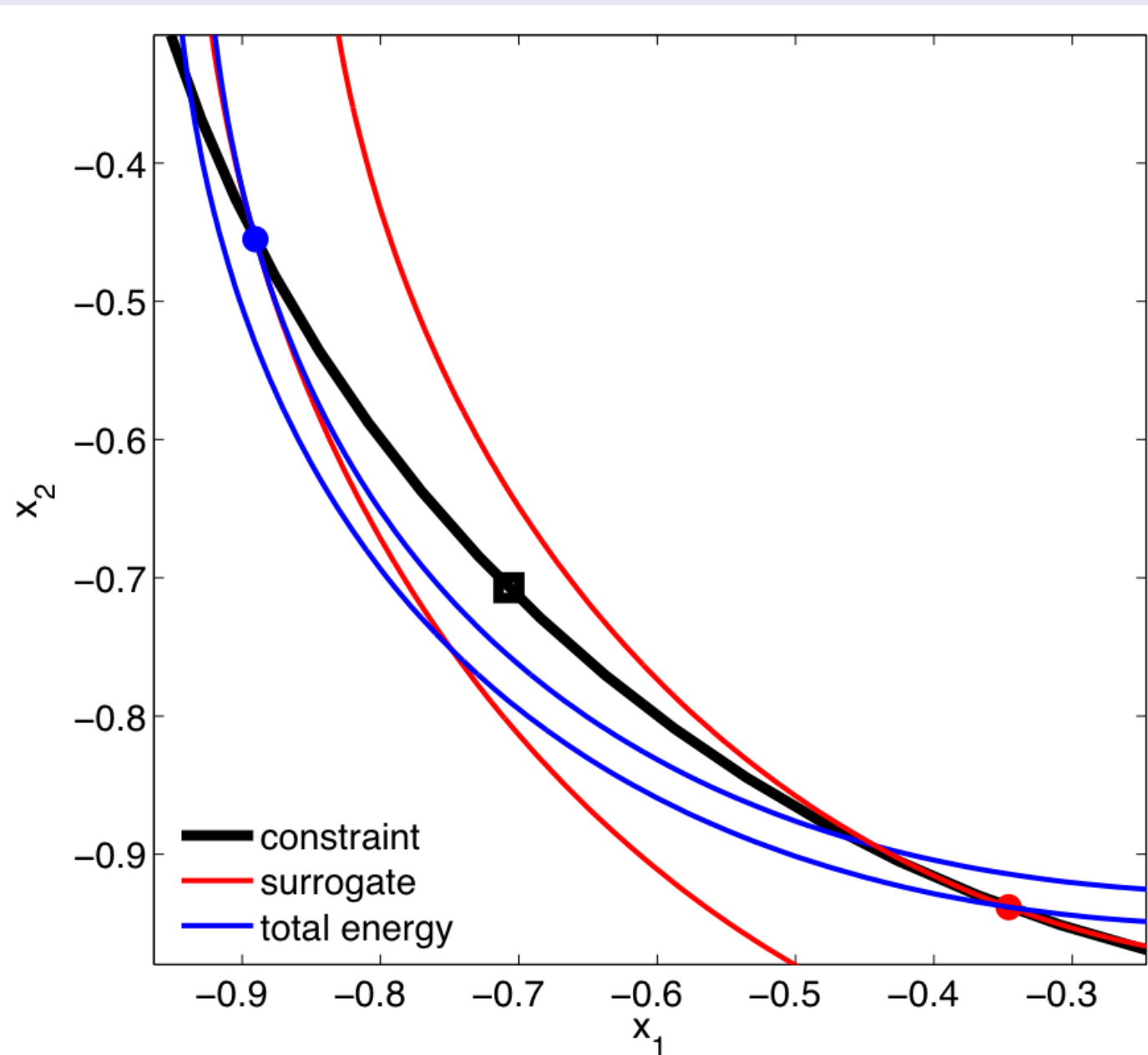
Recap: Why does SCF fail?

- True Objective and Gradient:
 - $E(x) = \frac{1}{2}x^T Lx + \frac{\alpha}{4}\rho(x)^T L^{-1}\rho(x)$
 - $\nabla E(x) = H(x)x$
- Surrogate and Gradient:
 - $E_{sur}(x) = \frac{1}{2}(x^T H(x^{(i)})x),$
 - $\nabla E_{sur}(x) = H(x^{(i)})x$

Objective and Surrogate are different BUT gradients match at $x^{(i)}$ – (by construction!)

$$\nabla E(x^{(i)}) = \nabla E_{sur}(x^{(i)})$$

Revisit $\alpha = 12$ case. Steplength is too long!



Some Approaches for Improving SCF

- Construct better surrogate – cannot afford to use local quadratic approximation (Hessian too expensive)
- Charge mixing to improve convergence (heuristic)
- Use Trust Region to restrict the update of the x in a small neighborhood of the gradient matching point. TRSCF – Thogersen, Olsen, Yeager & Jorgensen (2004)
- DCM – Yang, Meza & Wang (2007)

Using charge mixing to improve SCF

- We would call this accelerating an iterative method. Some choices include:
- Simple mixing

$$\rho^{i+1} = \tau \rho_{in}^i + (1 - \tau) \rho_{out}^i \quad 0 < \tau < 1$$

- Pulay mixing (Direct Inversion of Iterative Subspace)

$$\rho^{i+1} = \sum_{j=1}^i \alpha_j \rho^j, \quad \sum_{j=1}^i \alpha_j = 1$$

- Broyden mixing

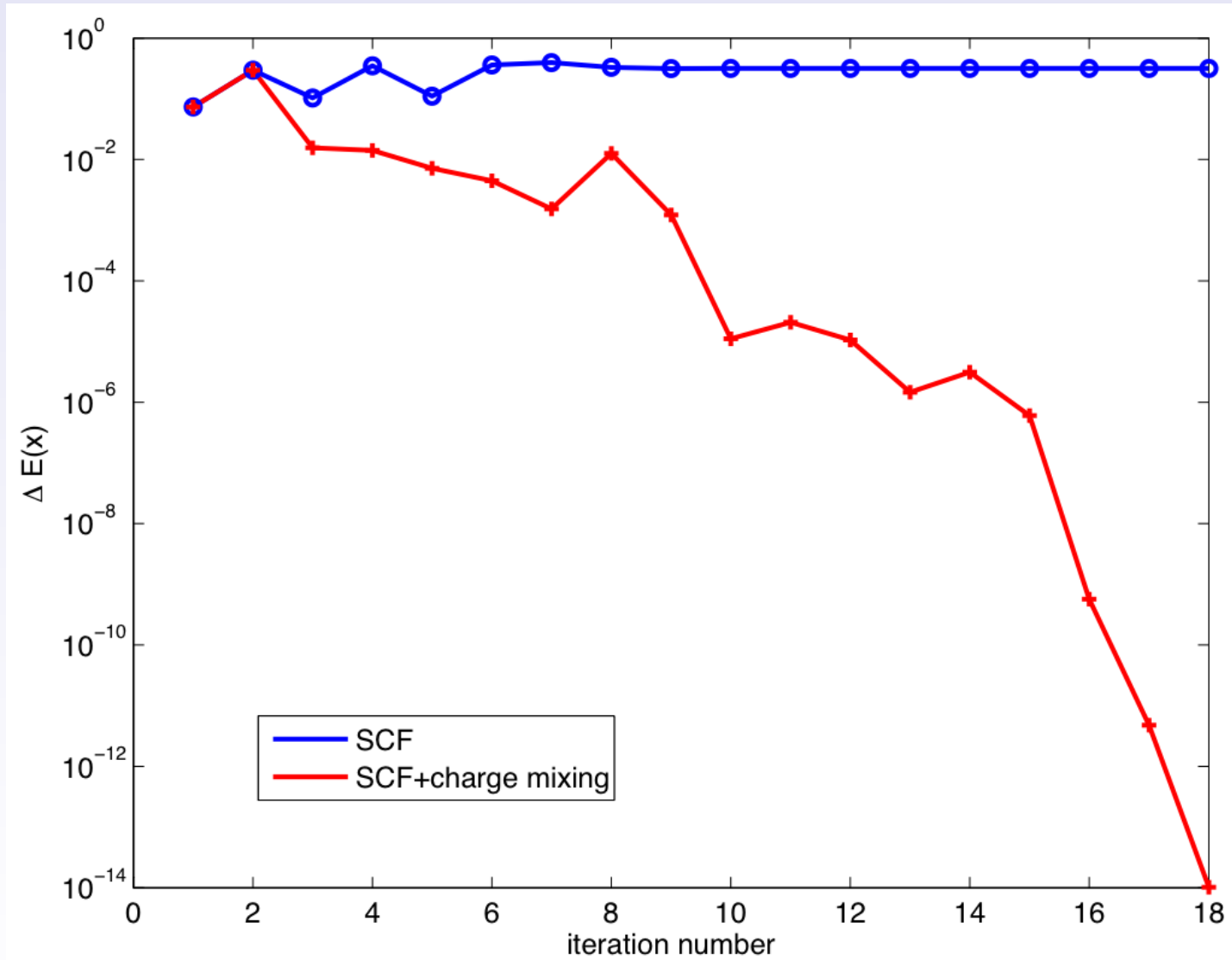
$$\rho^{i+1} = \rho^i + \tau C_{i+1} r^i$$

- Anderson mixing

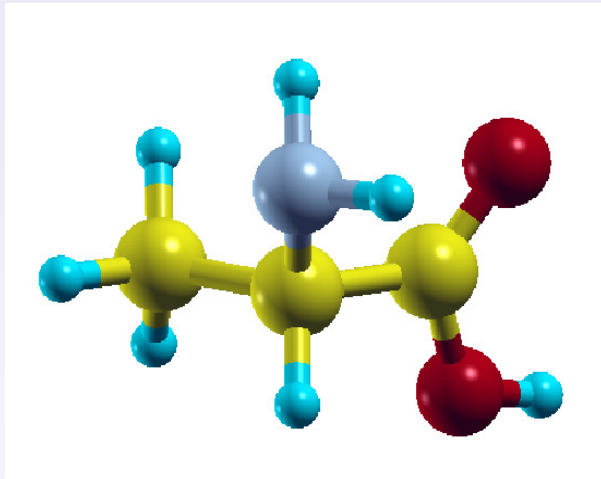
$$\rho^{i+1} = \rho^i + \tau r^i + (S_i - \tau Y_i) Y_i^\dagger r_i$$

SCF + charge mixing improves convergence

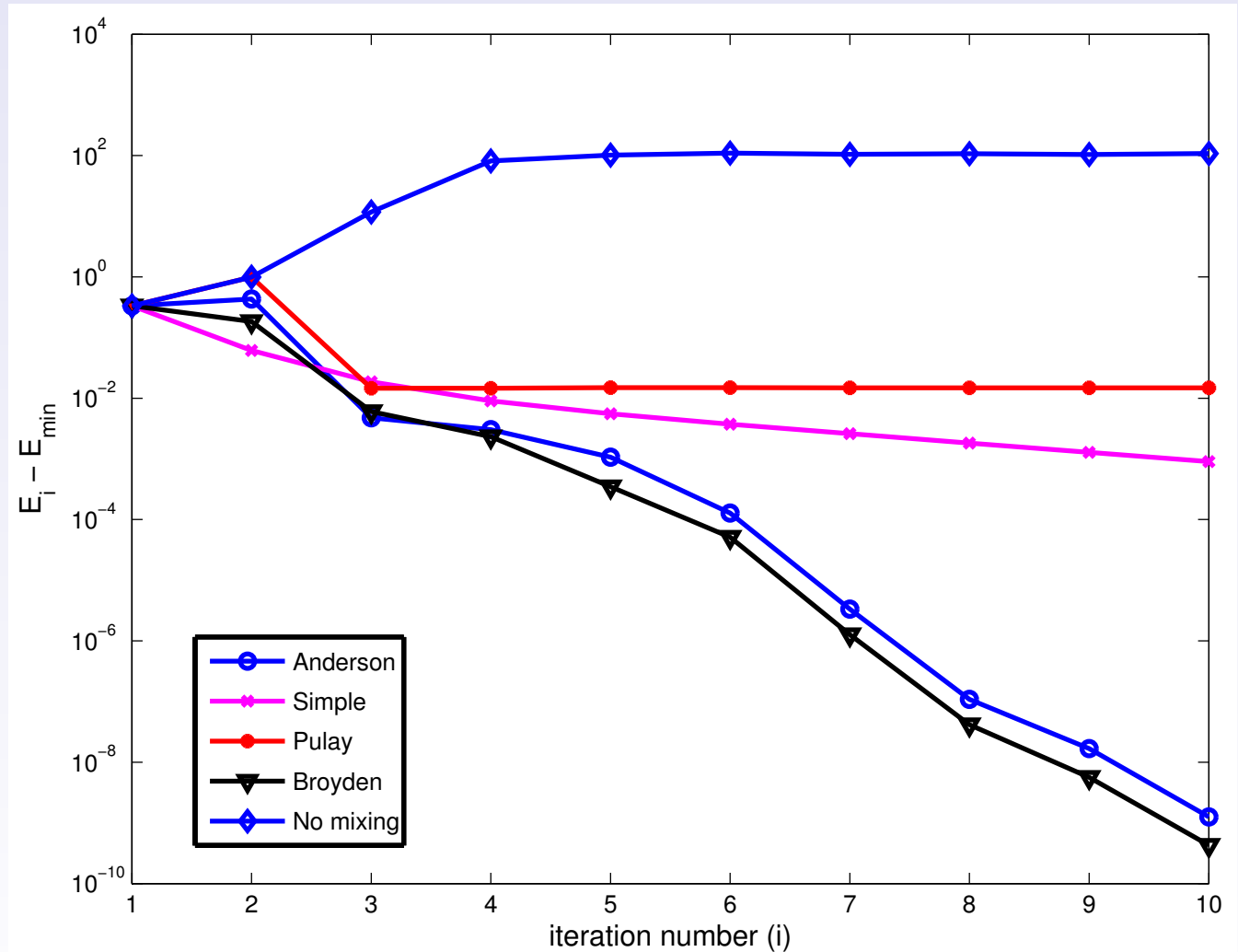
$$\Delta E(x^{(i)}) = \|E(x^{(i)}) - E_{min}\| \quad \alpha = 12, n = 10, n_e = 2$$



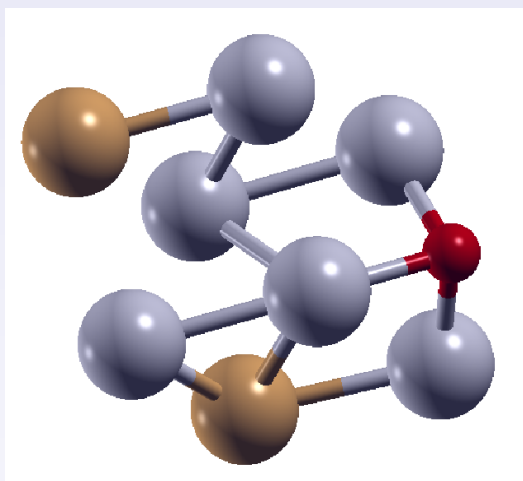
Comparison of charge mixing schemes on alanine



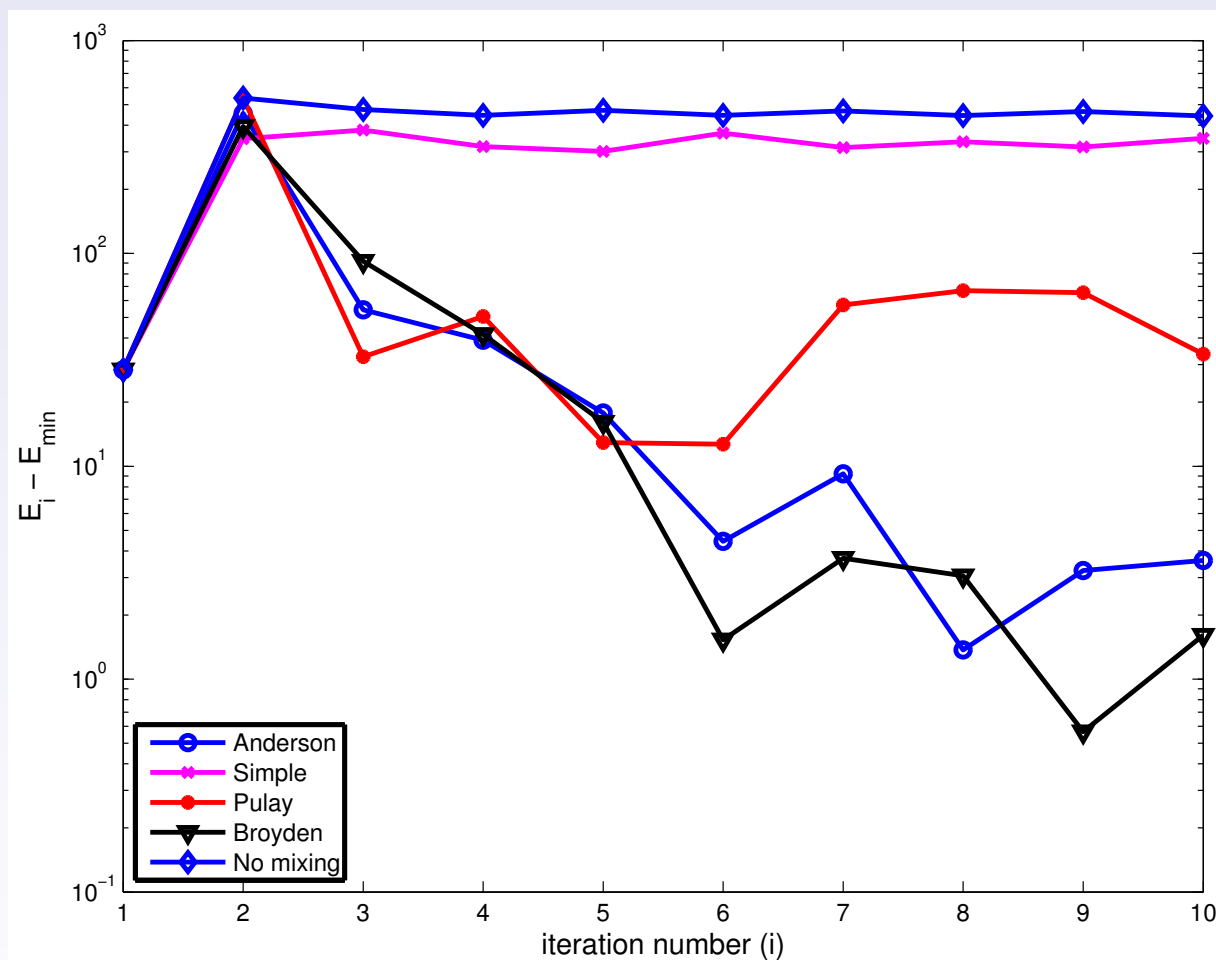
alanine



Comparison of charge mixing schemes on Pt_6Ni_2O



Pt_6Ni_2O



Trust Region Subproblem

- Simplest idea is to add a trust region constraint, e.g.

$$\min \quad E_{sur}(X)$$

s.t.

$$X^* X = I, \quad \text{orthogonality constraint}$$

$$\|X - X^{(i)}\|_F^2 \leq \Delta \quad \text{trust region constraint}$$

- Recall that if X is a solution to the [Kohn-Sham equations](#) then so is XQ , where $Q \in \mathbb{C}^{k \times k}$ and $Q^* Q = I_k$
- Unfortunately this simple constraint does not preserve the rotational invariance condition, i.e. $\|XQ - X^i\|_F^2 \leq \Delta$ for all $Q \in \mathbb{C}^{k \times k}$

Trust Region Subproblem

- Instead we will consider the following formulation

$$\min \quad E_{sur}(X)$$

s.t.

$$X^* X = I,$$

orthogonality constraint

$$\|X X^* - X^{(i)} (X^{(i)})^*\|_F^2 \leq \Delta$$

trust region constraint

- Which is equivalent to solving

$$\begin{aligned} \left[H(X^{(i)}) - \sigma X^{(i)} (X^{(i)})^* \right] X &= X \Lambda \\ X^* X &= I_k \end{aligned}$$

σ is a penalty parameter (Lagrange multiplier for the trust region constraint)

- Note: the matrix $D(X) = X X^*$ is called the **density matrix**

Detour: hints for deriving equations (4.4)–(4.5) in paper

- Let $D(X) = XX^*$ be the **density matrix**
- Can then show that

$$\begin{aligned}\|D(X) - D(X^{(i)})\|_F^2 &= \|D(X)\|_F^2 + \|D(X^{(i)})\|_F^2 - \\ &\quad 2 \operatorname{tr} \left[D(X)^* D(X^{(i)}) \right] \\ &= 2k - 2 \operatorname{tr} \left[X^* X^{(i)} X^{(i)*} X \right]\end{aligned}$$

- using the following properties of the trace operator

$$\begin{aligned}\|A\|_F^2 &= \operatorname{tr}(A^* A) \\ \operatorname{tr}(AB) &= \operatorname{tr}(BA) \\ \operatorname{tr}(A + B) &= \operatorname{tr}(A) + \operatorname{tr}(B)\end{aligned}$$

Some ideas for choosing σ

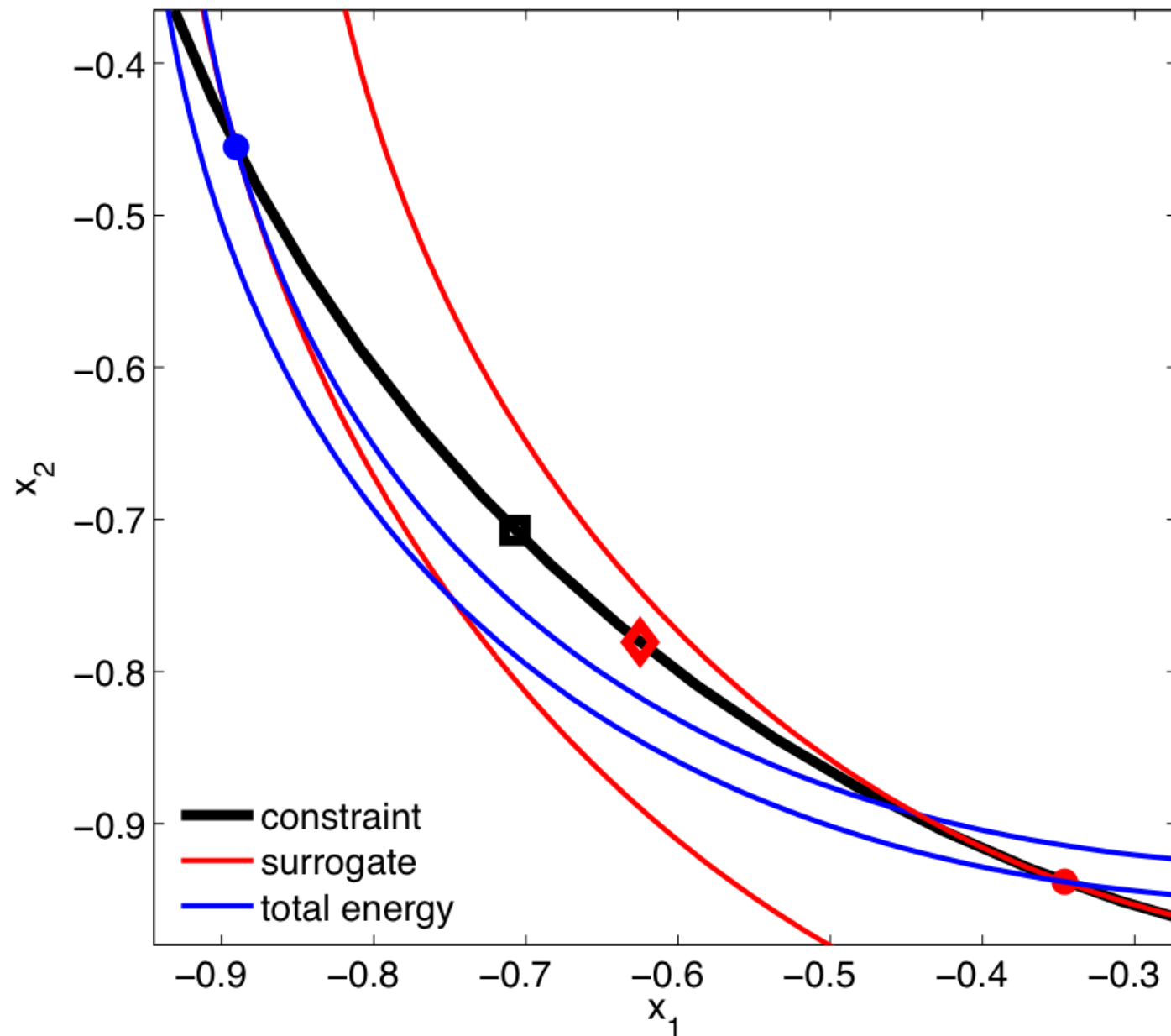
- The convergence of SCF depends on the gap between λ_k and λ_{k+1} , (Yang, Gao & Meza 2007)
- If X is the solution to $H(X)X = X\Lambda$, where $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k$ are the smallest k eigenvalues of $H(X)$ that appear in Λ , then the eigenvalues of $H(X) - \sigma XX^*$ are

$$\lambda_1 - \sigma, \lambda_2 - \sigma, \dots, \lambda_k - \sigma, \lambda_{k+1}, \dots, \lambda_n.$$

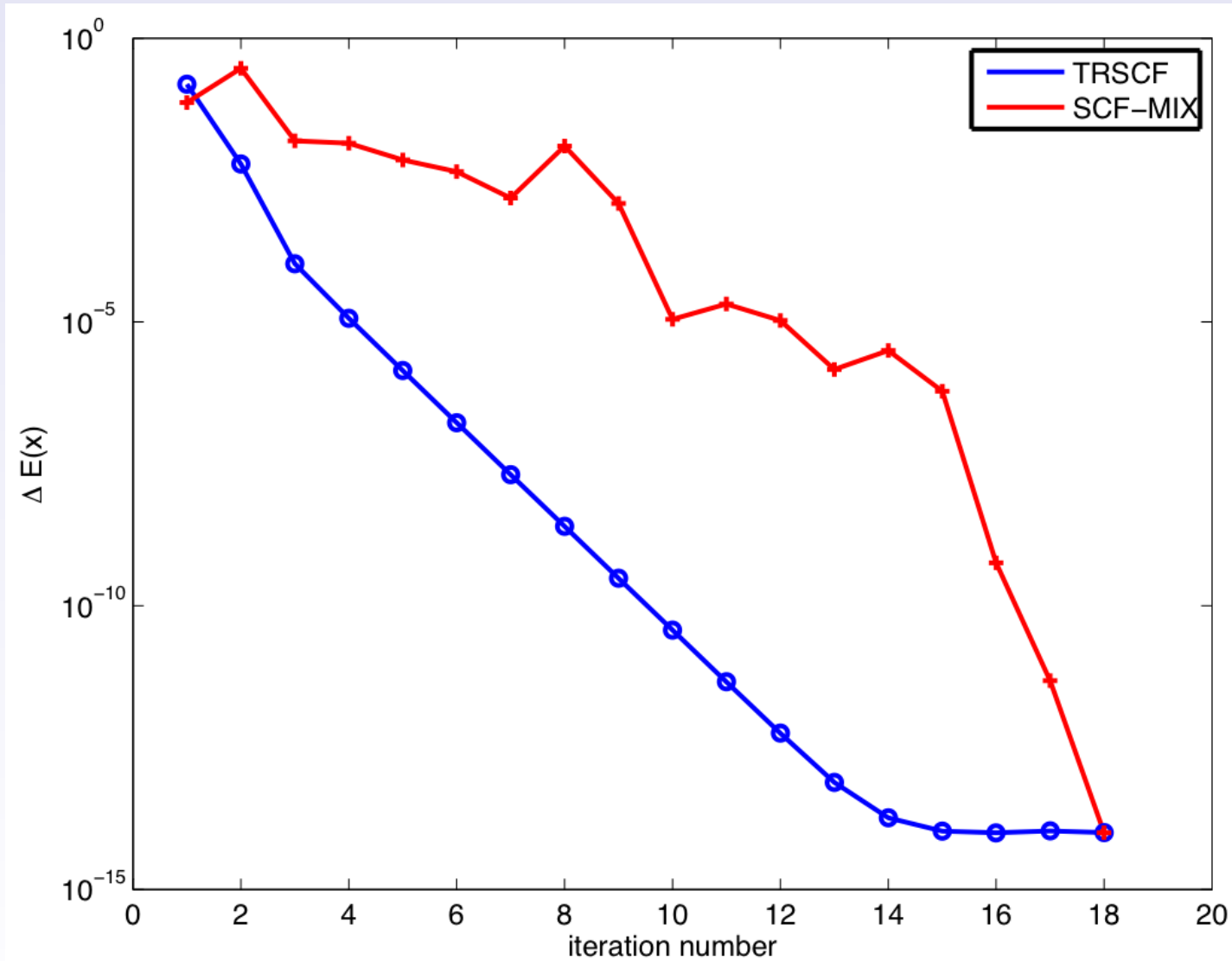
- **Strategy:** choose σ to open up the gap between $\lambda_k(H(X^{(i)}))$ and $\lambda_{k+1}(H(X^{(i)}))$, where

$$H(X^{(i)}) = H(X^{(i)}) - \sigma X^{(i)}(X^{(i)})^*.$$

Trust Region SCF (TRSCF)



TRSCF further improves convergence



Observation leading to DCM method

- Let $x^{(i)}$ is the current approximation
- Idea: minimize the energy in a certain (smaller) subspace
- Set

$$\begin{aligned}x^{(i+1)} &= \alpha x^{(i)} + \beta p^{(i-1)} + \gamma r^{(i)}, \quad \text{where} \\p^{(i-1)} &= \text{previous search direction} \\r^{(i)} &= H^{(i)} x^{(i)} - \theta^{(i)} x^{(i)} \quad (\text{gradient of Lagrangian})\end{aligned}$$

and α, β, γ are chosen so that

$$\begin{aligned}x^{(i+1)T} x^{(i+1)} &= 1 \\E(x^{(i+1)}) &< E(x^{(i)})\end{aligned}$$

N.B. Extension of LOBPCG (Knyazev, 2001) to nonlinear EV problem

Subspace Minimization

- Let $Y = (X^{(i)}, P^{(i-1)}, R^{(i)})$
- Then we can write $X^{(i+1)} = YG$, for some $G \in \mathbb{C}^{3k \times k}$
- And the minimization problem

$$\min_{X^*X=I_k} E(X^{i+1}) \equiv \min_{G^*Y^*YG=I_k} E(YG)$$

- is equivalent to solving

$$\begin{aligned} H(G)G &= BG\Omega_k \\ G^*BG &= I_k \end{aligned} \tag{*}$$

where $B = Y^*Y$ and $H(G) = Y^*H(X)Y$ is the projected Hamiltonian

- N.B. Equation (*) is much smaller than original nonlinear EV problem (k vs n)

Direct Constrained Minimization ¹

Algorithm

Input: Initial guess $X^{(0)} \in \mathbb{C}^{n \times k}$,

① $P^{(0)} = []$, $i = 0$;

② while (not converged)

① $\Theta^{(i)} = X^{(i)*} H^{(i)} X^{(i)}$;

② $R^{(i)} = H^{(i)} X^{(i)} - X^{(i)} \Theta^{(i)}$;

③ Set $Y = (X^{(i)}, P^{(i-1)}, K^{-1} R^{(i)})$;

④ Solve

$$\min_{G^* Y^* Y G = I_k} E_{tot}(Y G)$$

⑤ $X^{(i+1)} = Y G(1 : k, :)$; $P^{(i+1)} = Y G(k + 1 : 3k, :)$;

⑥ $i \leftarrow i + 1$;

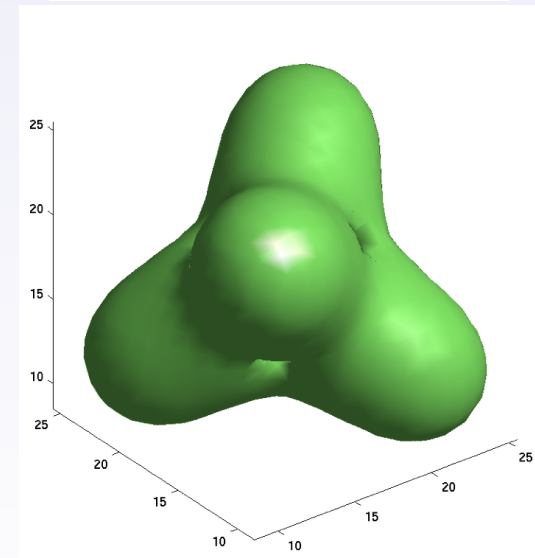
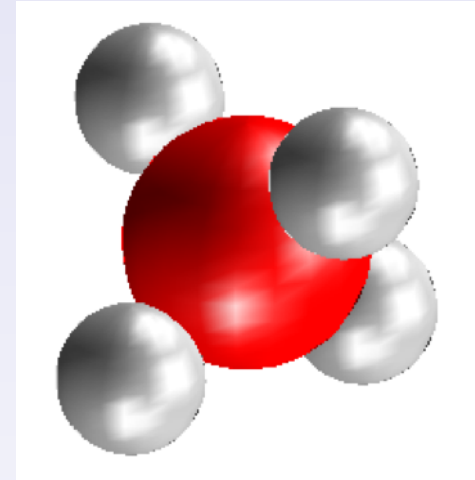
¹C. Yang, J. Meza, L. Wang, A Constrained Optimization Algorithm for Total Energy Minimization in Electronic Structure Calculation, J. Comp. Phys., 217, 709-721 (2006)

KSSOLV Matlab package

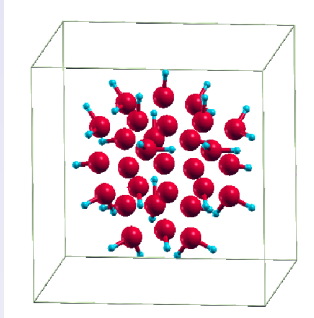
- KSSOLV Matlab code for solving the Kohn-Sham equations
- Open source package
- Handles SCF, DCM, Trust Region
- Example problems to get started with
- Object-oriented design - easy to extend
- Good starting point for students

Example: SiH_4

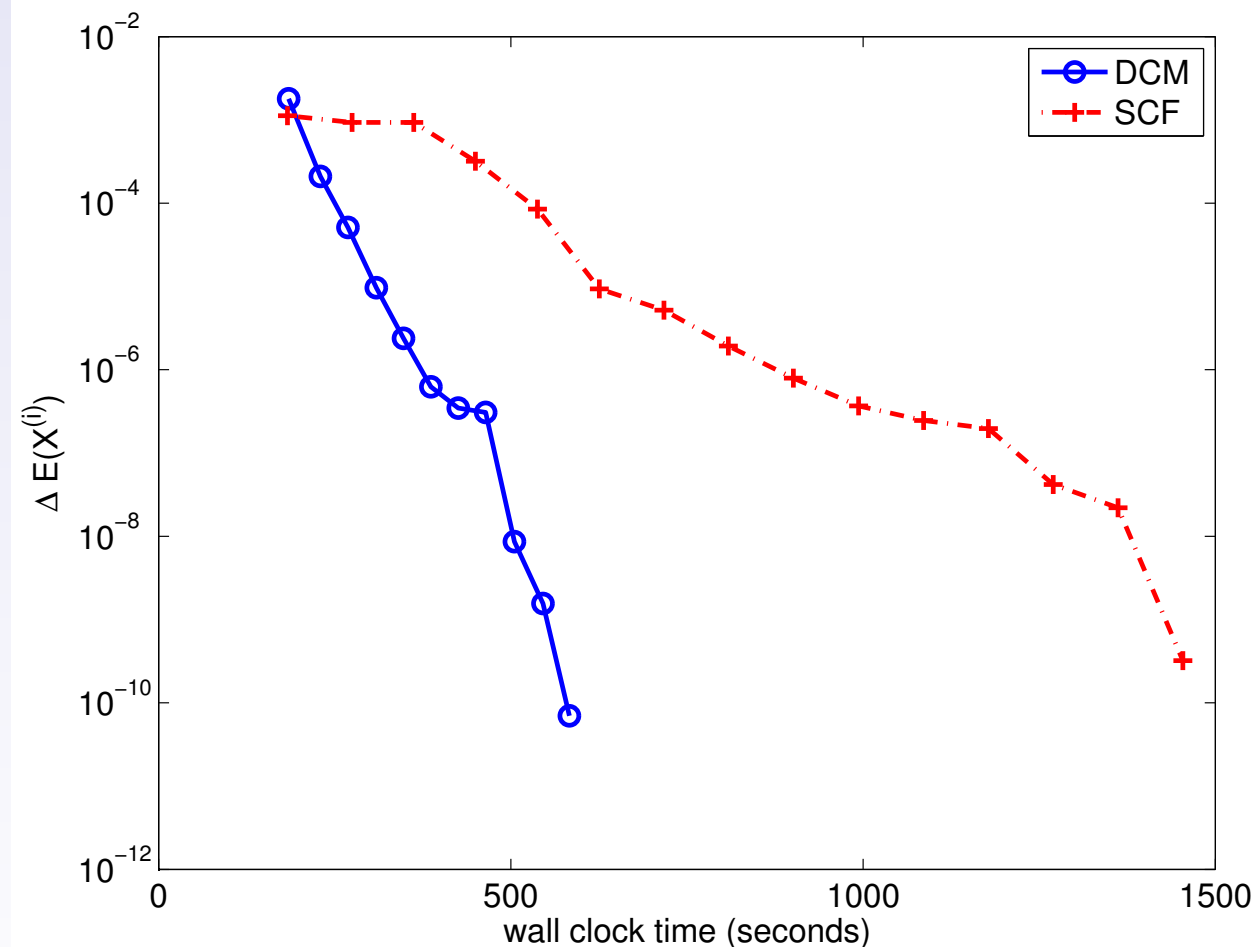
```
a1 = Atom('Si');  
a2 = Atom('H');  
alist = [a1 a2 a2 a2 a2];  
xyzlist= [  
    0.0    0.0    0.0  
    1.61   1.61   1.61  
    ...  ];  
mol = Molecule();  
mol = set(mol,'Blattice',BL);  
mol = set(mol,'atomlist',alist);  
mol = set(mol,'xyzlist' ,xyzlist);  
mol = set(mol,'ecut', 25);  
mol = set(mol,'name','SiH4');  
...  
isosurface(rho);
```



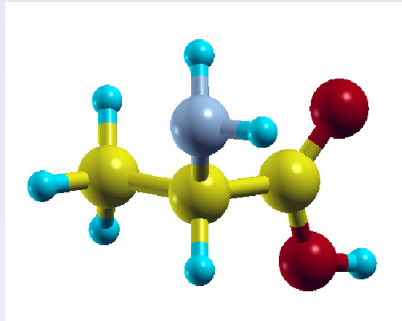
Example 1: $Si_{29}H_{36}$



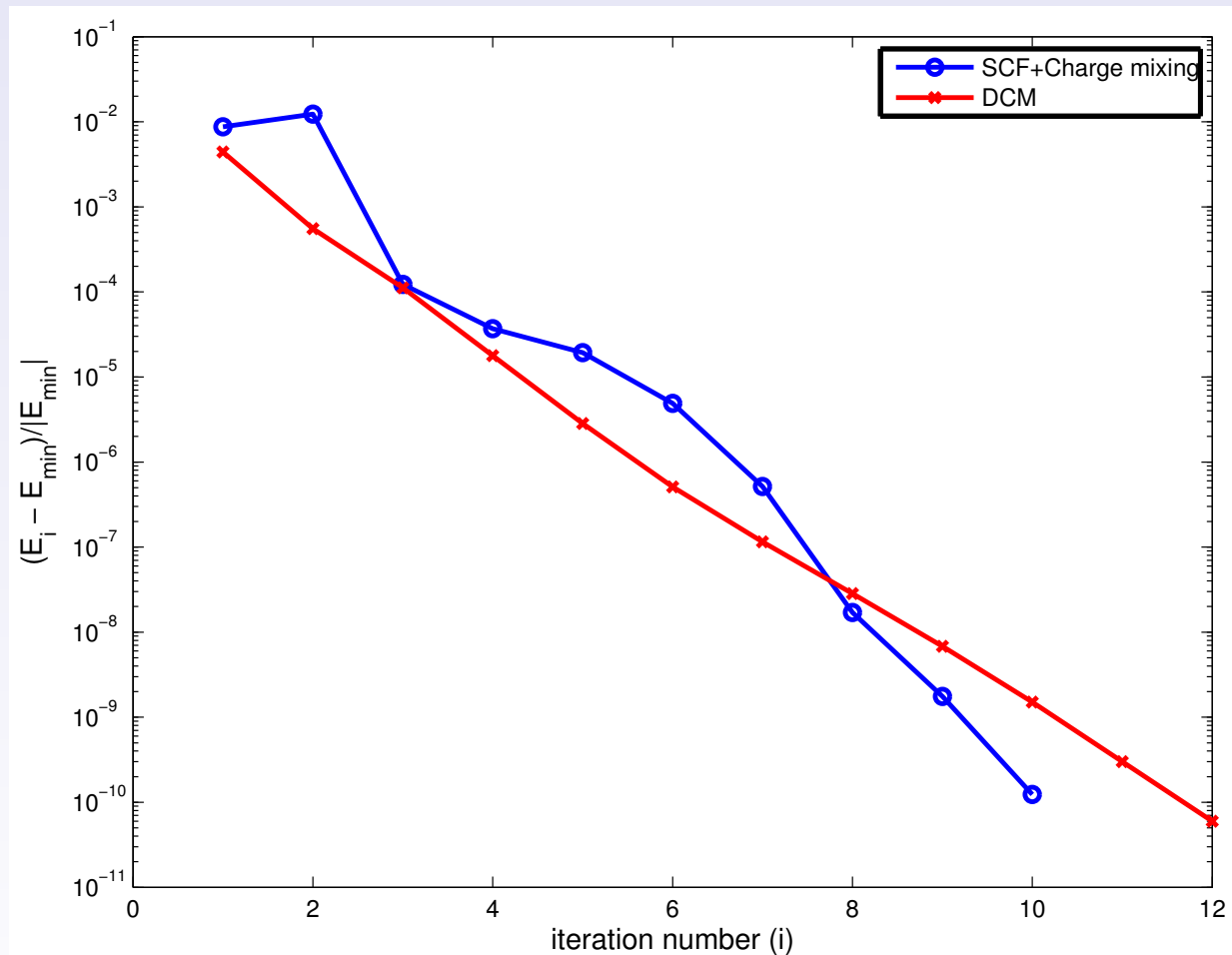
- supercell:
 $25.65 \times 25.65 \times 25.65$
- sampling grid:
 $96 \times 96 \times 96$ (ecut=25 Ryd)
- 10 PCG iterations in each SCF outer iteration.(PETOT)
- 3 inner SCF iteration in each DCM outer iteration



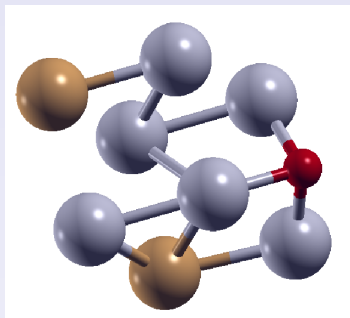
Example 2: Alanine



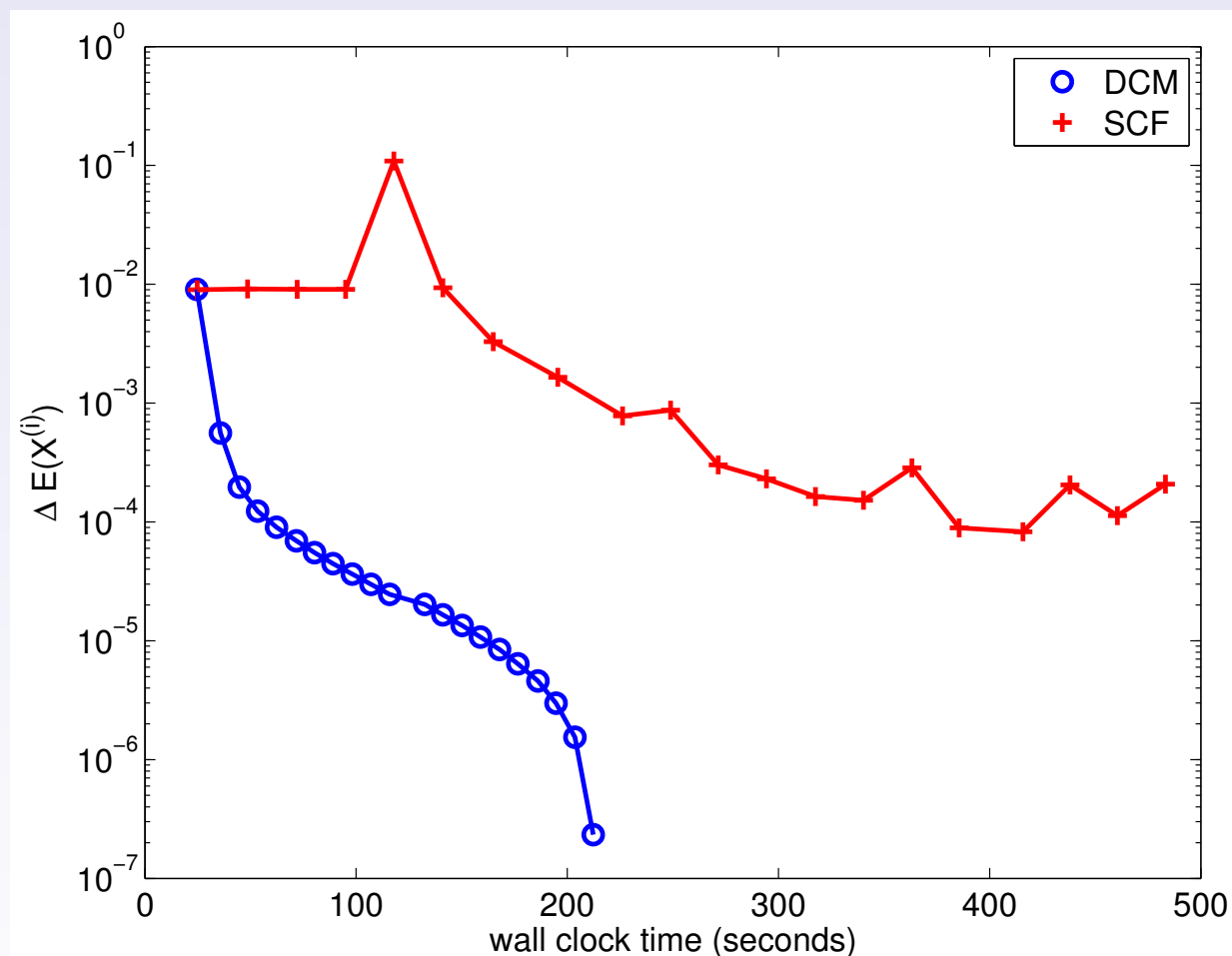
- supercell: $20 \times 15 \times 20$
- sampling grid:
 $96 \times 48 \times 96$ (ecut=25 Ryd)
- 10 PCG iterations in each SCF outer iteration.
- 3 inner SCF iteration in each DCM outer iteration



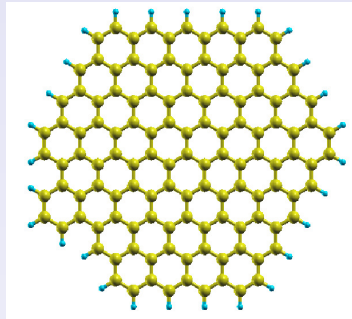
Example 3: Pt_6Ni_2O



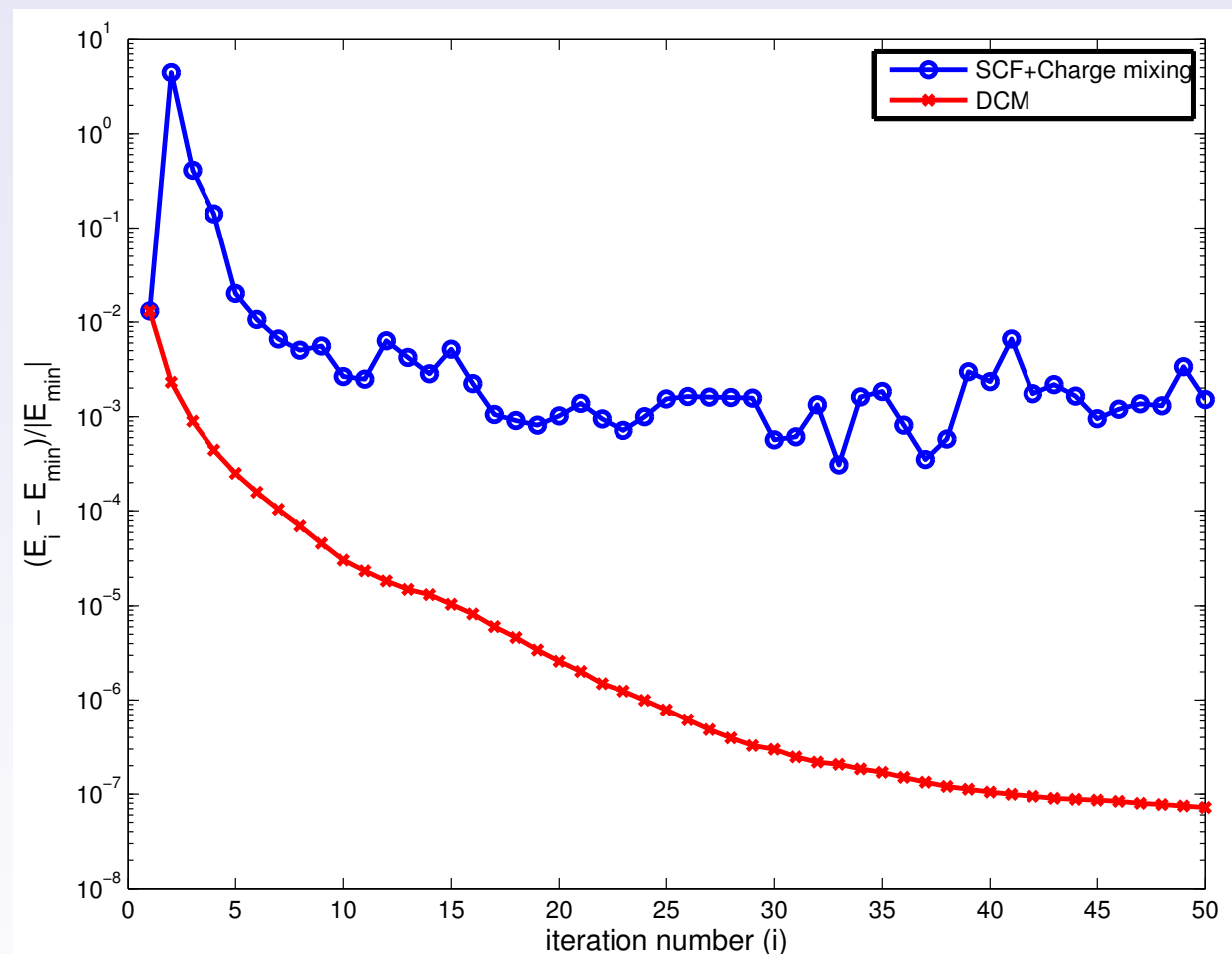
- supercell:
 $19.6 \times 10.7 \text{ \AA} \times 9.2 \text{ \AA}$
- sampling grid:
 $96 \times 48 \times 48$
- 10 PCG iterations in each SCF outer iteration.(PETOT)
- 5 inner SCF iteration in each DCM outer iteration



Example 4: Graphene




- supercell:
 $40 \times 40 \times 5$
- sampling grid:
 $114 \times 114 \times 15$
- 10 PCG iterations in each SCF outer iteration.
- 5 inner SCF iteration in each DCM outer iteration



Summary

- Despite dire warnings, mathematical techniques actually help chemistry
- New approach for solving the Kohn-Sham Equations using a direct optimization method improves convergence
- Trust region modification increases robustness of both SCF and DCM
- New computational software tools for modeling and simulation of nanosystems


References

 M. C. Payne, M. P. Teter, D. C. Allen, T. A. Arias, J. D. Joannopoulos,

Iterative minimization techniques for ab initio total energy calculation: Molecular dynamics and conjugate gradients,
Reviews of Modern Physics, Vol. 64, Number 4, pp. 1045–1097
(1992).

 Christopher J. Cramer,
Essentials of Computational Chemistry,
John Wiley and Sons (2003).

 Richard M. Martin,
Electronic Structure Basic Theory and Practical Methods,
Cambridge University Press (2005).

 J.M. Thijssen,
Computational Physics,
Cambridge University Press (2003).