Hartree–Fock and Kohn–Sham theories for large molecular systems

Trygve Helgaker, University of Oslo, Norway

Filip Pawłowski, Simen Reine, Erik Tellgren

Stinne Høst, Branislav Jansik, Poul Jørgensen, Pekka Manninen, Jeppe Olsen,
University of Aarhus, Denmark

Sonia Coriani, University of Trieste, Italy

Paweł Sałek, Royal Institute of Technology, Sweden

The XIIth International Congress of Quantum Chemistry

May 21–26, 2006

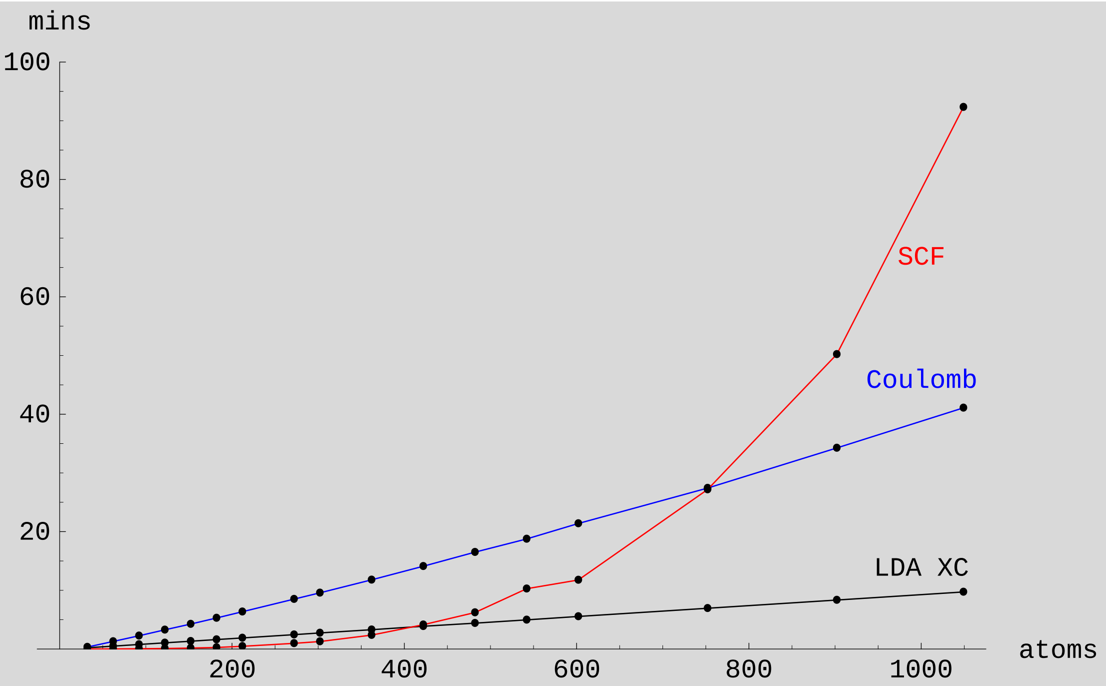
Kyoto TERRSA, Kyoto, Japan

Self-consistent field (SCF) theory

- Consider the optimization of the SCF energy (here LDA) of molecular systems:
  - small systems dominated by KS-matrix evaluation, with linear scaling
  - large systems dominated by SCF diagonalization, with cubic scaling

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}$$

$$\mathbf{D}_{\mathrm{new}} = \mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}\mathbf{C}$$



- To achieve linear scaling, we must avoid diagonalization and MOs!
- We shall here consider an alternative to diagonalization:
  - it optimizes the density matrix directly, avoiding MOs
  - involves only additions and multiplications of (sparse) one-electron matrices
  - for large (sparse systems), the calculations scale linearly with system size

Direct optimization of the density matrix

- Consider the direct optimization of the density matrix:

$$E(\mathbf{D}) = \mathrm{Tr}\,\mathbf{Dh} + \text{2-el. part}$$

  - there are constraints on the density matrix:

$$\underbrace{\mathbf{D} = \mathbf{D}^{\mathrm{T}}}_{\text{symmetry}}, \quad \underbrace{\mathrm{Tr}\,\mathbf{D} = N}_{\text{trace}}, \quad \underbrace{\mathbf{D}^2 = \mathbf{D}}_{\text{idempotency}} \quad \text{(orthonormal basis)}$$

  - any optimization must obey these constraints

- Many strategies are based on purification of the density matrix

$$\widetilde{\mathbf{D}} = 3\mathbf{D}^2 - 2\mathbf{D}^3 \quad \text{(McWeeny purification, 1960)}$$

  - Li, Nunes and Vanderbilt (1993)

$$\widetilde{E} = \mathrm{Tr}\,\widetilde{\mathbf{D}}\mathbf{h} + \mu(\mathrm{Tr}\,\mathbf{D} - N) + \text{2-el. part}$$

  - Millam and Scuseria (1997), Challacombe (1999)
  - Palser and Manolopoulos (1998), Niklasson (2002)

- We shall pursue a different approach, based on an explicitly parameterization of **D**

Exponential parameterization of the density matrix

- In a real, nonorthogonal AO basis, with $\mathbf{S} \neq \mathbf{I}$, let $\mathbf{D}$ be a valid HF/KS matrix:

$$\underbrace{\mathbf{D} = \mathbf{D}^{\mathrm{T}}}_{\text{symmetry}}, \quad \underbrace{\operatorname{Tr} \mathbf{DS} = N}_{\text{trace}}, \quad \underbrace{\mathbf{DSD} = \mathbf{D}}_{\text{idempotency}}$$

- Any other valid density matrix $\mathbf{D}(\mathbf{X})$ can then be generated from this matrix:

$$\underbrace{\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX})}_{\text{exponential parameterization}}, \quad \underbrace{\mathbf{X}^{\mathrm{T}} = -\mathbf{X}}_{\text{antisymmetric}}$$

  - Helgaker, Jørgensen and Olsen: *Molecular Electronic-Structure Theory* (Wiley, 2000)
  - Head-Gordon and coworkers, MolPhys **101**, 37 (2003), JCP **118**, 6144 (2003)

- We can obtain any valid density matrix, in the AO basis, without recourse to MOs!
  - in particular, we may optimize the energy by freely varying $X_{\mu\nu}$ with $\mu > \nu$:

$$E_{\min}(\mathbf{X}) = \min_{\mathbf{X}} \left[\operatorname{Tr} \mathbf{D}(\mathbf{X})\mathbf{h} + \text{2-el. part}\right]$$

- Is the use of $\mathbf{D}(\mathbf{X})$ a practical proposition?
  - we shall in this talk demonstrate that it is indeed so
  - we shall consider energy optimizations and property calculations

Two questions about $\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX})$

- Can it be evaluated efficiently?
  - we use a generalized Baker–Campbell–Hausdorff (BCH) expansion:

$$\mathbf{D}(\mathbf{X}) = \mathbf{D} + [\mathbf{D}, \mathbf{X}]_{\mathrm{S}} + \tfrac{1}{2}[[\mathbf{D}, \mathbf{X}]_{\mathrm{S}}, \mathbf{X}]_{\mathrm{S}} + \cdots$$

  - we have here introduced the S commutator

$$[\mathbf{D}, \mathbf{X}]_{\mathrm{S}} = \mathbf{DSX} - \mathbf{XSD}$$

  - converges rapidly (purification may be necessary), in about 10 matrix multiplications
- Are redundancies a problem?
  - the AO space consists of two parts: the occupied space and the virtual space

$$\mathbf{P} = \mathbf{DS} \text{ (onto occupied space)}, \quad \mathbf{Q} = \mathbf{I} - \mathbf{DS} \text{ (onto virtual space)}$$

  - only rotations between the occupied and virtual spaces are nonredundant;

$$\mathbf{X} = \underbrace{\mathbf{PXP}^{\mathrm{T}} + \mathbf{QXQ}^{\mathrm{T}}}_{\text{redundant}} + \underbrace{\mathbf{PXQ}^{\mathrm{T}} + \mathbf{QXP}^{\mathrm{T}}}_{\mathbf{X}_{\mathrm{ov}}}$$

  - to avoid problems with redundancies, we use the projected parameterization

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{X}_{\mathrm{ov}}\mathbf{S})\mathbf{D}\exp(\mathbf{S}\mathbf{X}_{\mathrm{ov}}), \quad \mathbf{X}^{\mathrm{T}} = -\mathbf{X}$$

Diagonalization-free Roothaan–Hall SCF optimization

- The SCF (Fock or Kohn–Sham) energy may, in principle, be optimized directly:

$$E_{\min} = \min_{\mathbf{X}} E(\mathbf{X}) \quad \Leftrightarrow \quad \underbrace{\mathbf{F}(\mathbf{D})\mathbf{D}\mathbf{S} = \mathbf{S}\mathbf{D}\mathbf{F}(\mathbf{D})}_{\text{stationary condition}}$$

  - a difficult global minimization problem!

- In MO theory, the Roothaan–Hall SCF scheme works well, especially with DIIS:

$$\mathbf{F} = \mathbf{h} + \mathbf{g}(\mathbf{D}) \quad \underset{D}{\overset{F}{\rightleftarrows}} \quad \mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}; \; \mathbf{D}_{\text{new}} = \mathbf{C}_{\text{occ}}\mathbf{C}_{\text{occ}}^{\mathrm{T}}$$

  - each diagonalization is equivalent to minimizing the sum of the (occ.) orbital energies

$$\varepsilon(\mathbf{X}) = \sum_I \epsilon_I = \operatorname{Tr} \mathbf{D}(\mathbf{X})\mathbf{F}$$

- By analogy with MO theory, we set up the following Roothaan–Hall SCF scheme:

$$\mathbf{F} = \mathbf{h} + \mathbf{g}(\mathbf{D}) \quad \underset{D}{\overset{F}{\rightleftarrows}} \quad \varepsilon_{\min} = \min_{\mathbf{X}} \operatorname{Tr} \mathbf{D}(\mathbf{X})\mathbf{F}; \; \mathbf{D}_{\text{new}} = \mathbf{D}(\mathbf{X})$$

  - at each SCF iteration, we minimize $\operatorname{Tr} \mathbf{D}(\mathbf{X})\mathbf{F}$ with respect to $\mathbf{X}$
  - the new density is then obtained by expansion of $\mathbf{D}(\mathbf{X})$ with the minimizer $\mathbf{X}_*$

- We thus avoid MOs and diagonalization but retain the SCF iterations

Newton minimization of the Roothaan–Hall energy function

- At each SCF iteration, our task is to minimize the Roothaan–Hall energy function

$$\varepsilon(\mathbf{X}) = \mathrm{Tr}\,\mathbf{D}(\mathbf{X})\mathbf{F} = \mathrm{Tr}\,\mathbf{DF} + \mathrm{Tr}\big[\mathbf{D},\mathbf{X}\big]_{\mathrm{S}}\mathbf{F} + \tfrac{1}{2}\,\mathrm{Tr}\big[\big[\mathbf{D},\mathbf{X}\big]_{\mathrm{S}},\mathbf{X}\big]_{\mathrm{S}}\mathbf{F} + \cdots$$

- Truncating to second order and setting the gradient to zero, we obtain the Newton step:

$$\mathbf{HXS} + \mathbf{SXH} = \mathbf{G} \;\leftarrow \text{the Roothaan–Hall Newton equation}$$

  - where the (negative) gradient and Hessian matrices are given by

$$\mathbf{G} = \mathbf{F}^{\mathrm{vo}} - \mathbf{F}^{\mathrm{ov}} \qquad \mathbf{F} = \mathbf{F}^{\mathrm{oo}} + \mathbf{F}^{\mathrm{ov}} + \mathbf{F}^{\mathrm{vo}} + \mathbf{F}^{\mathrm{vv}}$$

$$\mathbf{H} = \mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}$$

- A RH diagonalization corresponds to an exact minimization (many Newton steps)
  - however, a partial minimization will do
  - in fact, one RH Newton step is usually sufficient
- Because of their large dimensions, the Newton equations cannot be solved directly
  - we use an iterative scheme: the conjugate-gradient method
  - key step: repeated evaluation of the residual $\mathbf{R} = \mathbf{G} - \mathbf{HXS} - \mathbf{SXH}$
  - all operations are elementary (sparse) matrix manipulations

Solution of the Roothaan–Hall Newton equations

- At each SCF iteration, we solve the Roothaan–Hall Newton equations

$$\mathbf{HXS} + \mathbf{SXH} = \mathbf{G}$$

  - a naïve application of the CG method converges slowly
  - the equations are ill-conditioned since $\kappa(\mathbf{H})\kappa(\mathbf{S}) \gg 1$ (κ is the condition number)

- The equations may be made well-conditioned by a Löwdin orthonormalization

$$\widetilde{\mathbf{H}}\mathbf{Z} + \mathbf{Z}\widetilde{\mathbf{H}} = \widetilde{\mathbf{G}}, \quad \widetilde{\mathbf{A}} = \mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}$$

  - convergence is greatly improved since $\kappa(\widetilde{\mathbf{H}}) = \kappa(\mathbf{S}^{-1/2}\mathbf{H}\mathbf{S}^{-1/2}) \ll \kappa(\mathbf{H})\kappa(\mathbf{S})$
  - we obtain $\mathbf{S}^{-1/2}$ by Taylor expansion
  - orthogonalization is also possible by Cholesky decomposition $\mathbf{S} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$
  - Millam and Scuseria (1996), Challacombe (1998), Head-Gordon *et al.* (2003)

- Further diagonal preconditioning cuts the number of iterations by one half

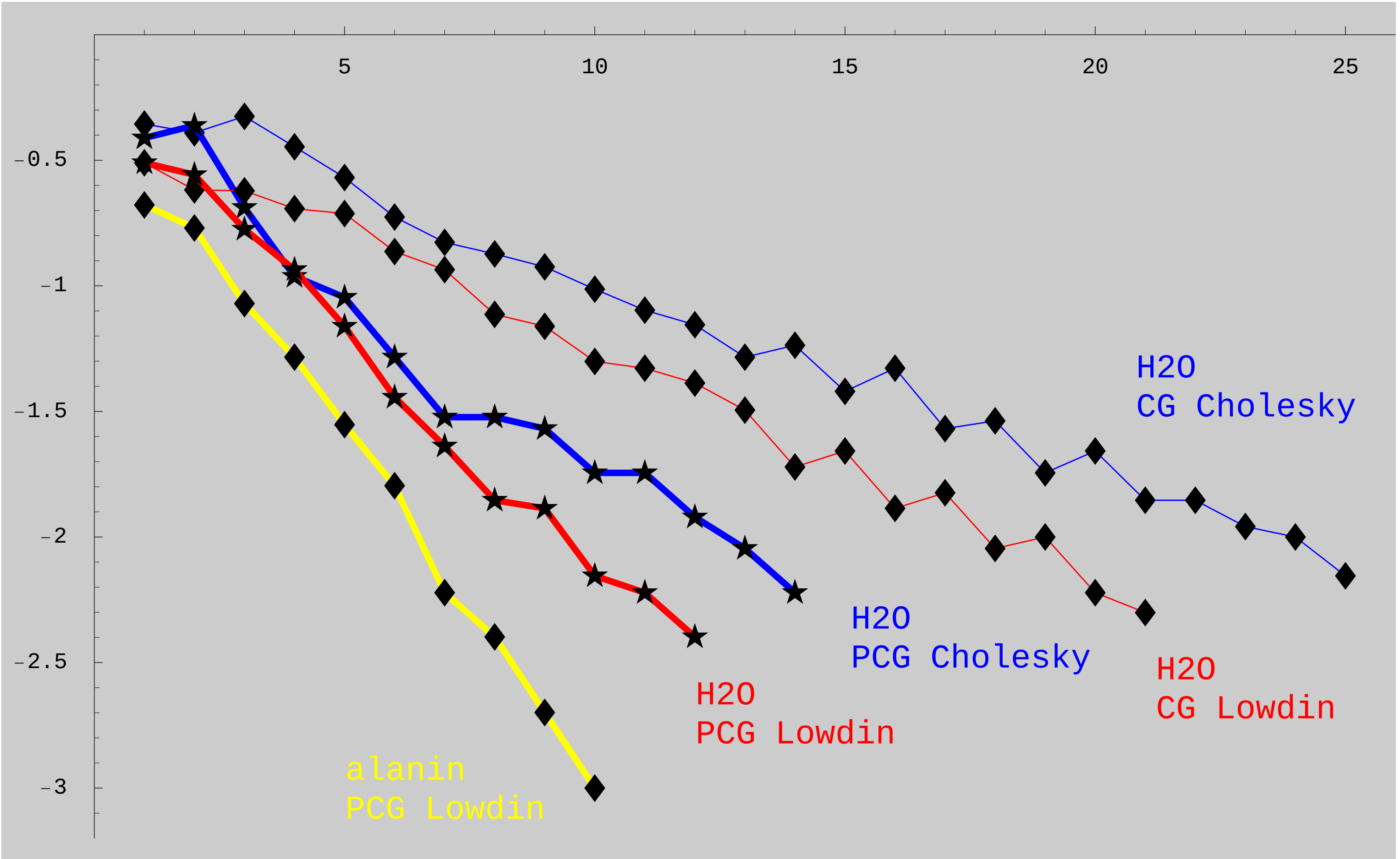
$$\widetilde{\widetilde{\mathbf{H}}} = \widetilde{\mathbf{H}}_{\mathrm{diag}}^{-1}\widetilde{\mathbf{H}}$$

  - 10 iterations typically reduce the residual by two orders of magnitude

- About 70 multiplications needed for one Newton iteration with density-matrix generation

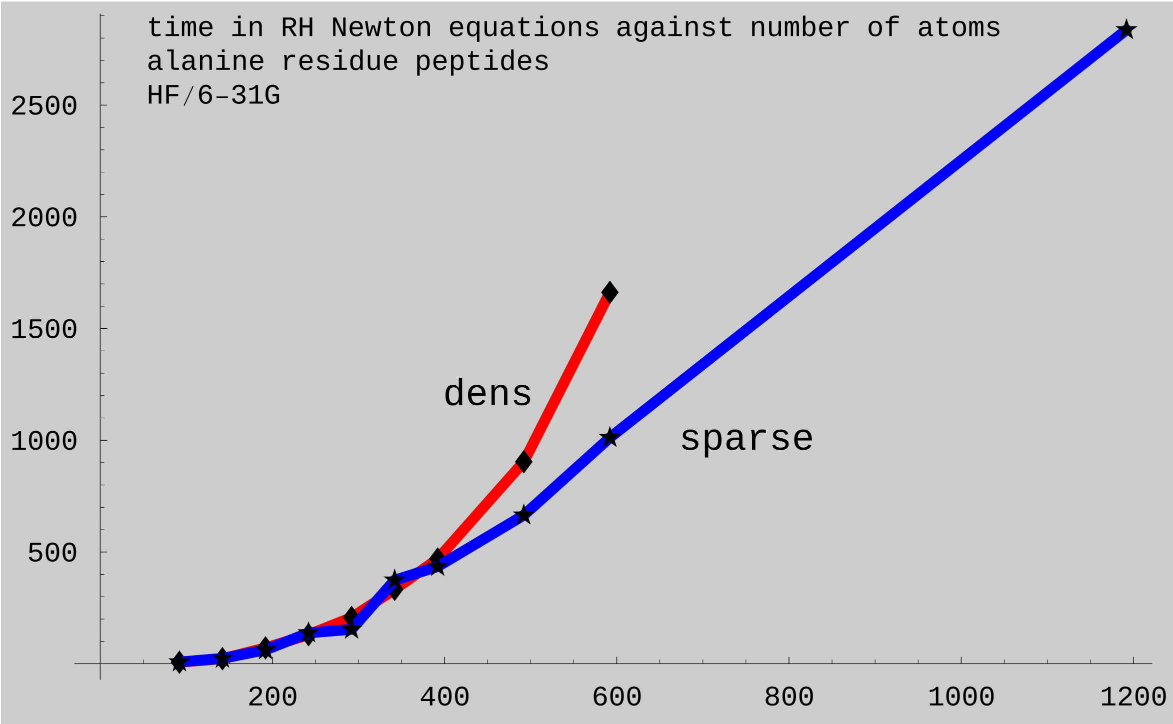
Iterative solution of Roothaan–Hall Newton equations

- Logarithmic plots of the residual against the number of iterations
  - H_2O, LDA/t-aug-cc-pVTZ
  - 99 alanine residue peptides, LDA/6-31G (5449 AOs)



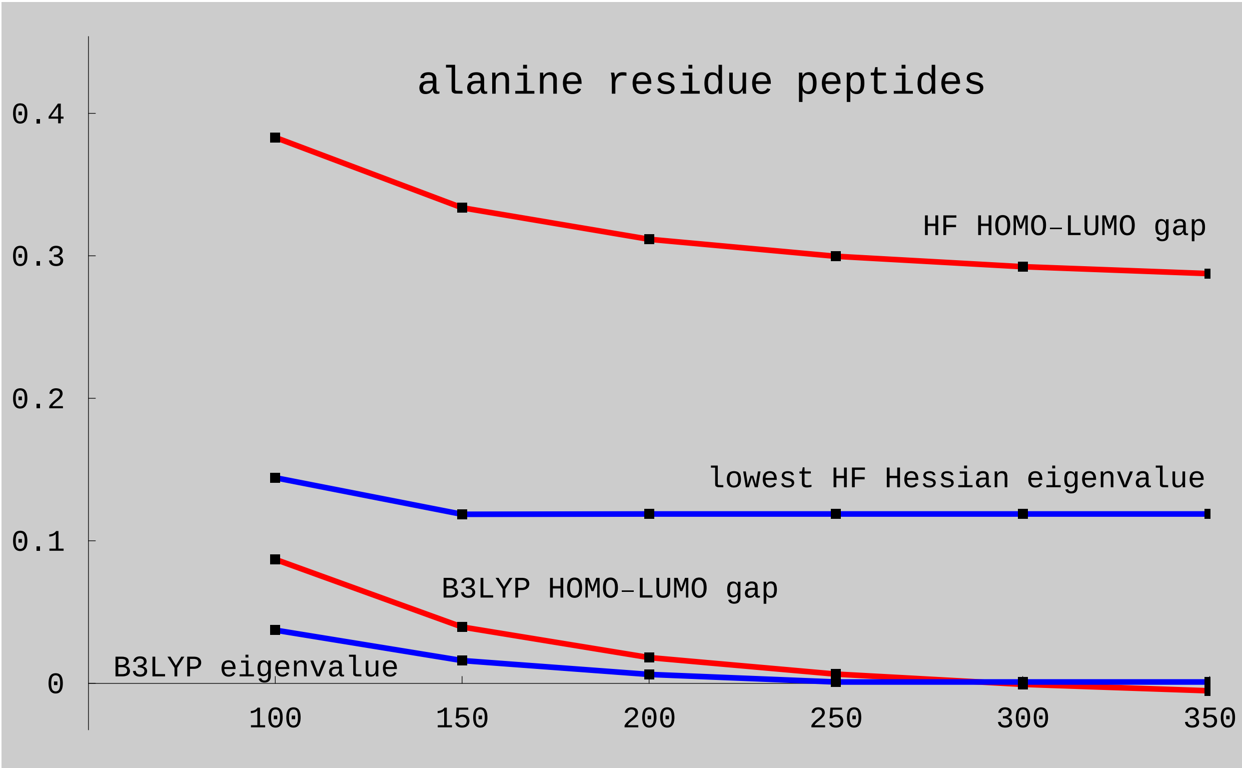
CPU time spent in Roothaan–Hall Newton equations

- We have successfully avoided Fock/Kohn–Sham diagonalization
  - minimization rather than the solution of a generalized eigenvalue problem
  - rapidly convergent: 50–100 sparse matrix multiplications needed
- Linear scaling is obtained by employing sparse-matrix algebra
  - compressed sparse-row (CSR) representation of few-atom blocks



SCF optimizations in small and large molecules

- Diagonalization can be avoided by solving Newton equations
- However, SCF convergence is typically more difficult in larger systems
  - small (or negative) HOMO-LUMO gaps and small Hessian eigenvalues in DFT
  - lowest Hessian eigenvalue and HOMO-LUMO gap in alanine residue peptides (6-31G)



- We have modified the standard SCF scheme, to make it more robust

The trust-region self-consistent field (TRSCF) method

- SCF optimizations have two ingredients
  1. Roothaan–Hall minimization (diagonalization): $\min_{\mathbf{X}} \operatorname{Tr} \mathbf{D}(\mathbf{X})\mathbf{F}$
  2. DIIS-type averaging of density matrices: $\overline{\mathbf{D}} = \sum_{i=0}^{n} c_i \mathbf{D}_i, \; \sum_i c_i = 1$
- In the Roothaan–Hall step, we minimize subject to a constraint on the step size:
  - this amounts to a simple level shifting of the Fock/Kohn–Sham matrix

$$\mathbf{F} \;\rightarrow\; \mathbf{F}(\mu) = \mathbf{F} - \mu \mathbf{S}^{\mathrm{oo}} \quad \text{(only occupied–occupied part shifted)}$$

  - for $\mu > 0$, the HOMO–LUMO gap increases, making large steps unfavourable
  - μ is adjusted until step is of desired length, during the iterative solution
- In the averaging step, we construct a second-order model of the SCF energy

$$E_{\mathrm{DSM}}(c_i) \approx E_{\mathrm{SCF}}(3\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}} - 2\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}})$$

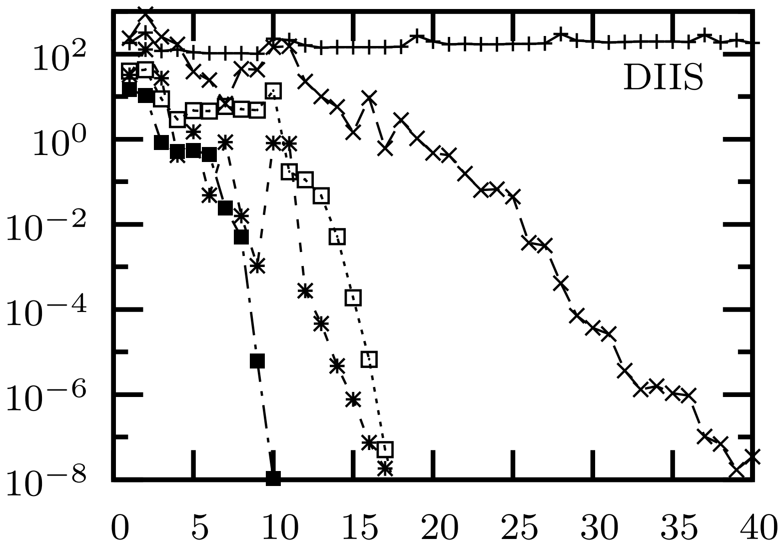
  - it has the correct gradient but an approximate Hessian
  - the best density matrix is obtained by minimization, subject to step-size constraint

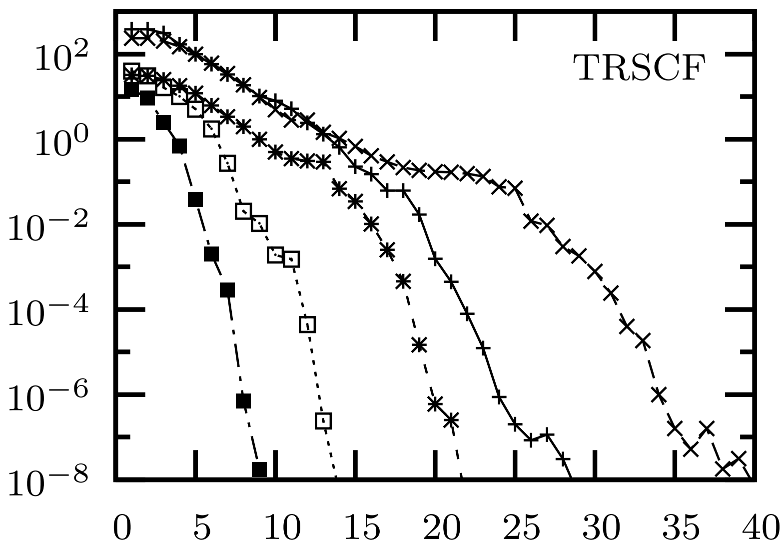
$$\min_{c_i} E_{\mathrm{DSM}}(c_i) \;\leftarrow \text{density-subspace minimization (DSM)}$$

- JCP **121**, 15 (2004); **123**, 074103 (2005)

The TRSCF method (continued)

- The TRSCF method gives a stable and uniform convergence towards the SCF minimum
- Convergence of LDA calculations for a variety of molecules
  - zinc complex(+), rhodium complex(×), cadmium complex(∗), CH_3CHO(□) and H_2O(■)



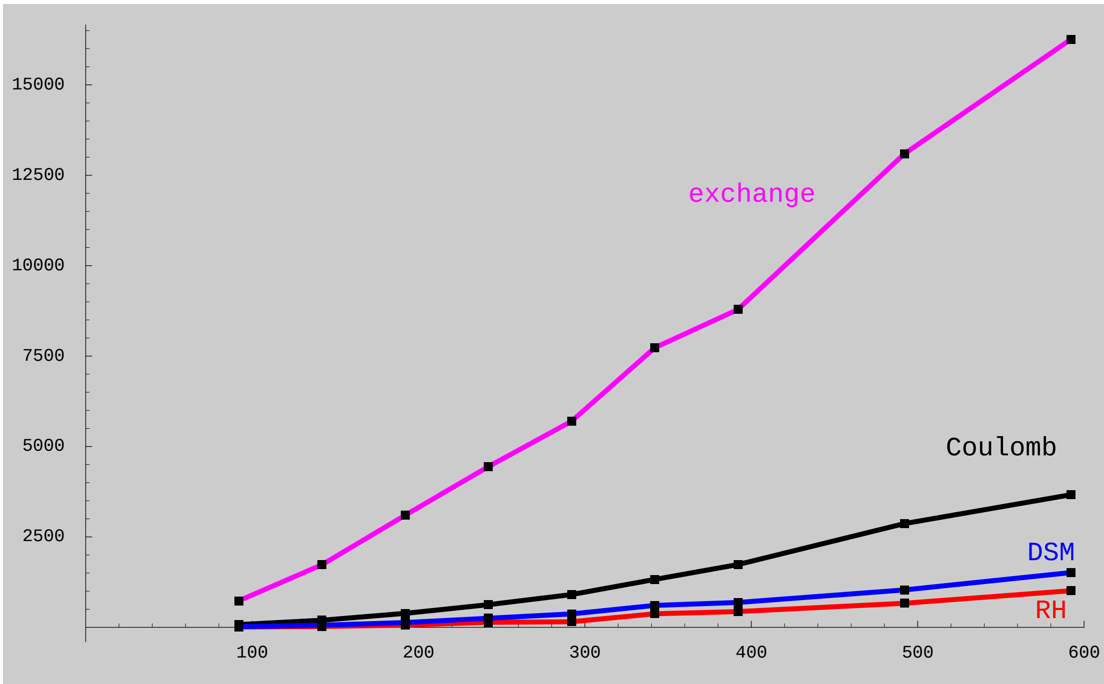


- Problems arise with small Hessian eigenvalues and small (negative) HOMO–LUMO gaps
  - Note: a Newton step of $\mathrm{Tr}\,\mathbf{FD}$ is a quasi-Newton step of $E_{\mathrm{SCF}}(\mathbf{D})$
  - we are doing a quasi-Newton optimization of the energy
  - the quasi-Newton step is poor for small or negative HOMO–LUMO gaps
  - revert to full Newton if necessary: $\min \mathrm{Tr}\,\mathbf{FD} \rightarrow \min E_{\mathrm{SCF}}(\mathbf{D})$

Illustration: alanine residue peptides

- Features of the code
  - diagonalization-free trust-region Roothaan–Hall (TRRH) energy minimization
  - trust-region density-subspace minimization (TRDSM) for density averaging
  - boxed density-fitting with FMM for Coulomb evaluation (Simen Reine)
  - LinK for exact exchange, linear-scaling exchange-correlation evaluation
  - compressed sparse-row (CSR) representation of few-atom blocks

- alanine residue peptides
  - CPU time against atoms
  - HF/6-31G
  - 5th SCF iteration
  - dominated by exchange
  - RH step least expensive
  - full lines: sparse algebra
  - dashed lines: dens algebra



Response theory

- The expectation value of $\hat{A}$ in the presence of a perturbation $\hat{V}_\omega$ of frequency ω:

$$\langle t|\hat{A}|t\rangle = \langle 0|\hat{A}|0\rangle + \int \langle\langle \hat{A}; \hat{V}_\omega \rangle\rangle_\omega \exp(-\mathrm{i}\omega t)\, \mathrm{d}\omega + \cdots$$

  – the linear-response function $\langle\langle \hat{A}; \hat{V}_\omega \rangle\rangle_\omega$ carries information about the first-order change in the expectation value

- The linear-response function may be represented compactly as:

$$\langle\langle \hat{A}; \hat{V}^\omega \rangle\rangle_\omega = -\mathbf{A}^{[1]\,\mathrm{T}} \underbrace{\left(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}\right)^{-1} \mathbf{V}_\omega^{[1]}}_{\text{linear equations}} \quad \longleftarrow \begin{cases} \mathbf{E}^{[2]} & \text{electronic Hessian} \\ \mathbf{S}^{[2]} & \text{metric matrix} \\ \mathbf{A}^{[1]} = \mathrm{vec}\left(\mathbf{ADS} - \mathbf{SDA}\right) \end{cases}$$

- In practice, the response functions are evaluated by solving linear equations

$$\left(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}\right)\mathbf{N}^{[1]} = -\mathbf{V}_\omega^{[1]}$$

$$\langle\langle \hat{A}; \hat{V}^\omega \rangle\rangle_\omega = \mathbf{A}^{[1]\,\mathrm{T}} \mathbf{N}^{[1]}$$

  – can this be accomplished efficiently in the AO basis?

Solution of the response equations

- The response equations are solved in the same manner as the RH Newton equations:

$$\left(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}\right)\mathbf{x} = \mathbf{V}^{[1]}$$

  - transformation to orthogonal basis (Cholesky or Löwdin)
  - generation of an iterative subspace until the residual is sufficiently small

$$\mathbf{R} = \left(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}\right)\mathbf{x} - \mathbf{V}^{[1]}$$

- Key step: multiplication of Hessian and metric matrices with trial vectors

$$\mathbf{E}^{[2]}(\mathbf{X}) = \mathbf{HXS} + \mathbf{SXH} + \mathbf{g}^{\mathrm{vo}}([\mathbf{D}, \mathbf{X}]_{\mathrm{S}}) - \mathbf{g}^{\mathrm{ov}}([\mathbf{D}, \mathbf{X}]_{\mathrm{S}})$$

$$\mathbf{S}^{[2]}(\mathbf{X}) = \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}} - \mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}}$$

  - requires recalculation of Fock/Kohn–Sham matrix with modified AO density matrix

- For rapid convergence, the residual vector is preconditioned

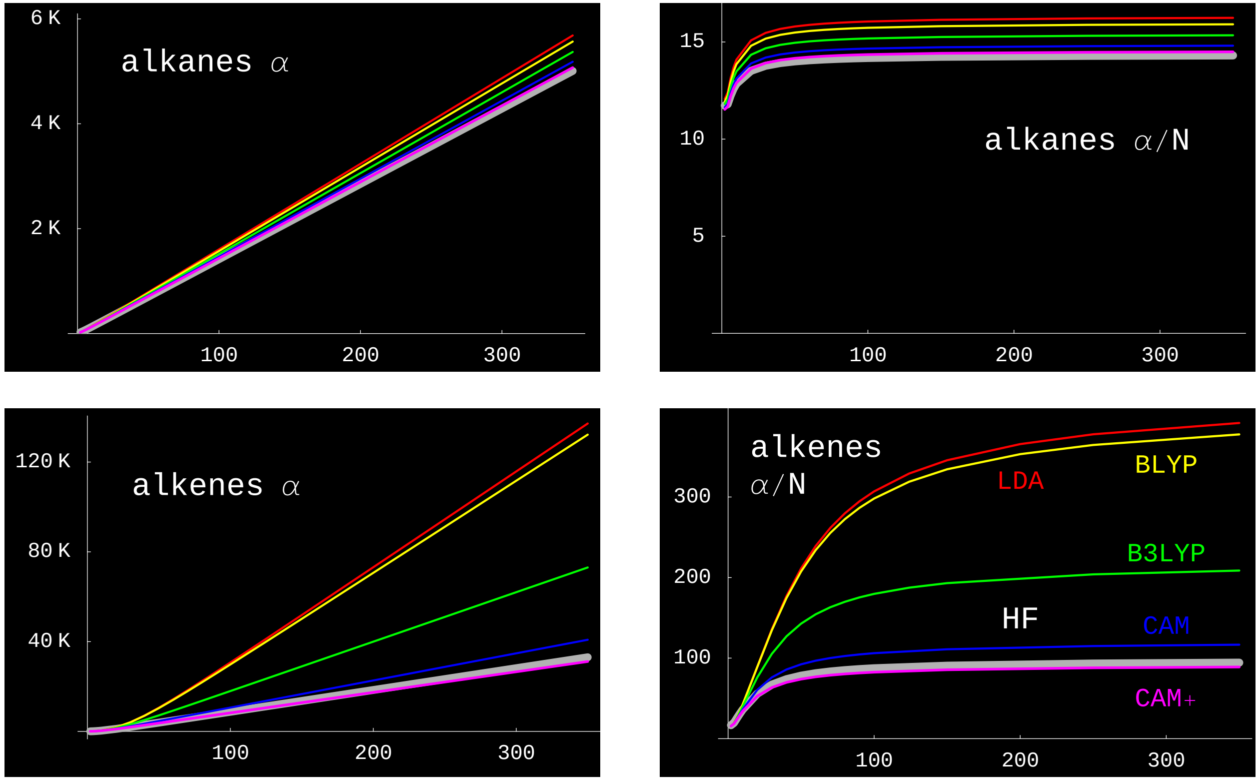
$$\widetilde{\mathbf{R}} = \mathbf{M}^{-1}\mathbf{R}, \quad \mathbf{M} = \mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]} \text{ – expensive parts}$$

  - nondiagonal preconditioning requires about 5 conjugate-gradient steps

- With this preconditioner, the response equations converge in about 4 iterations
  - indeed, this is the same convergence as in the canonical MO basis
  - total cost: 4 Fock/Kohn–Sham evaluations, 100 matrix multiplications

Polarizabilities of linear alkanes and alkenes

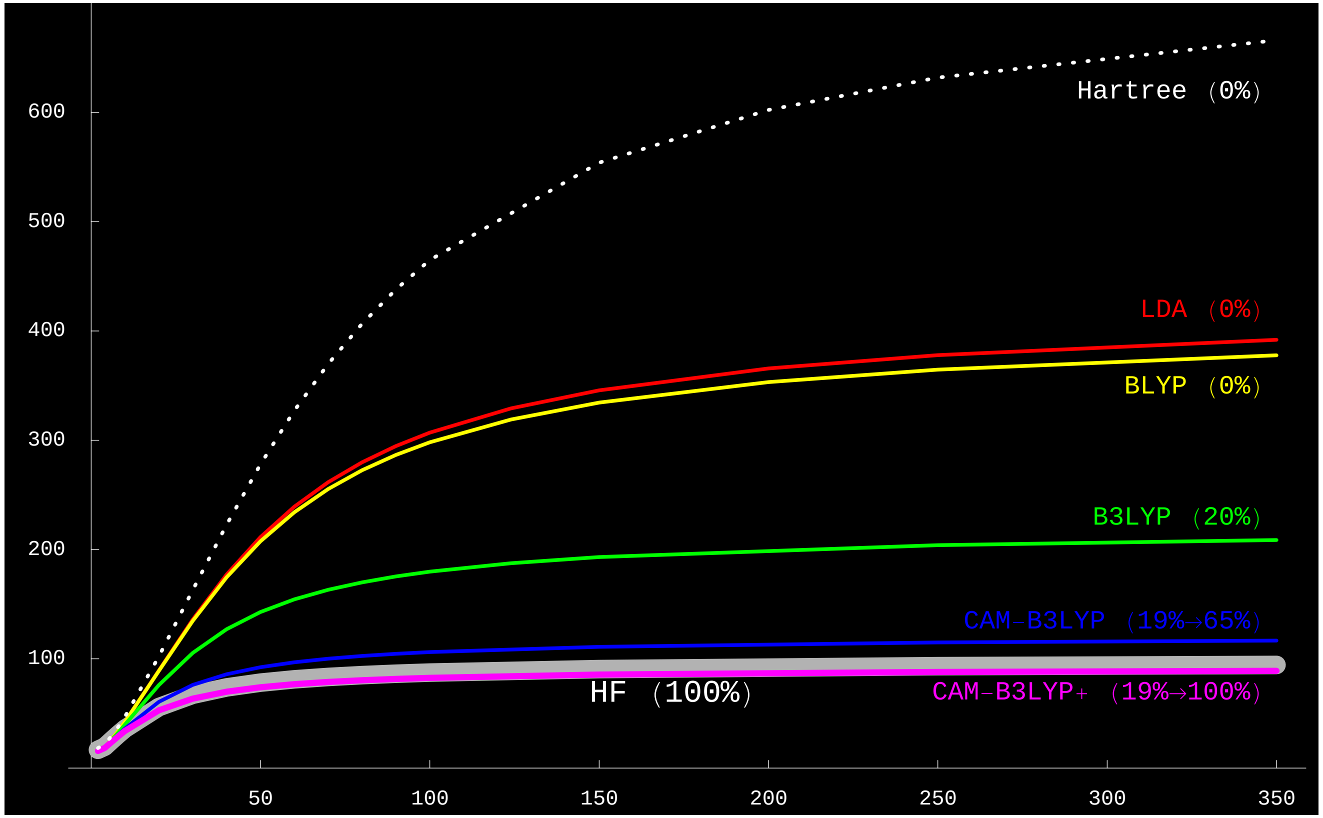
- To illustrate, we have calculated longitudinal polarizabilities in linear polymeric chains
  - HF and DFT α and α/N in 6-31G basis, plotted against the number of carbons N



- The alkenes are about an order of magnitude more polarizable than the alkanes
  - all models agree on alkanes (α/N-limit: HF 14.4; LDA 16.3)
  - widely different results for alkenes (α/N-limit: HF 97; LDA 427)

The importance of exact exchange for longitudinal polarizabilities

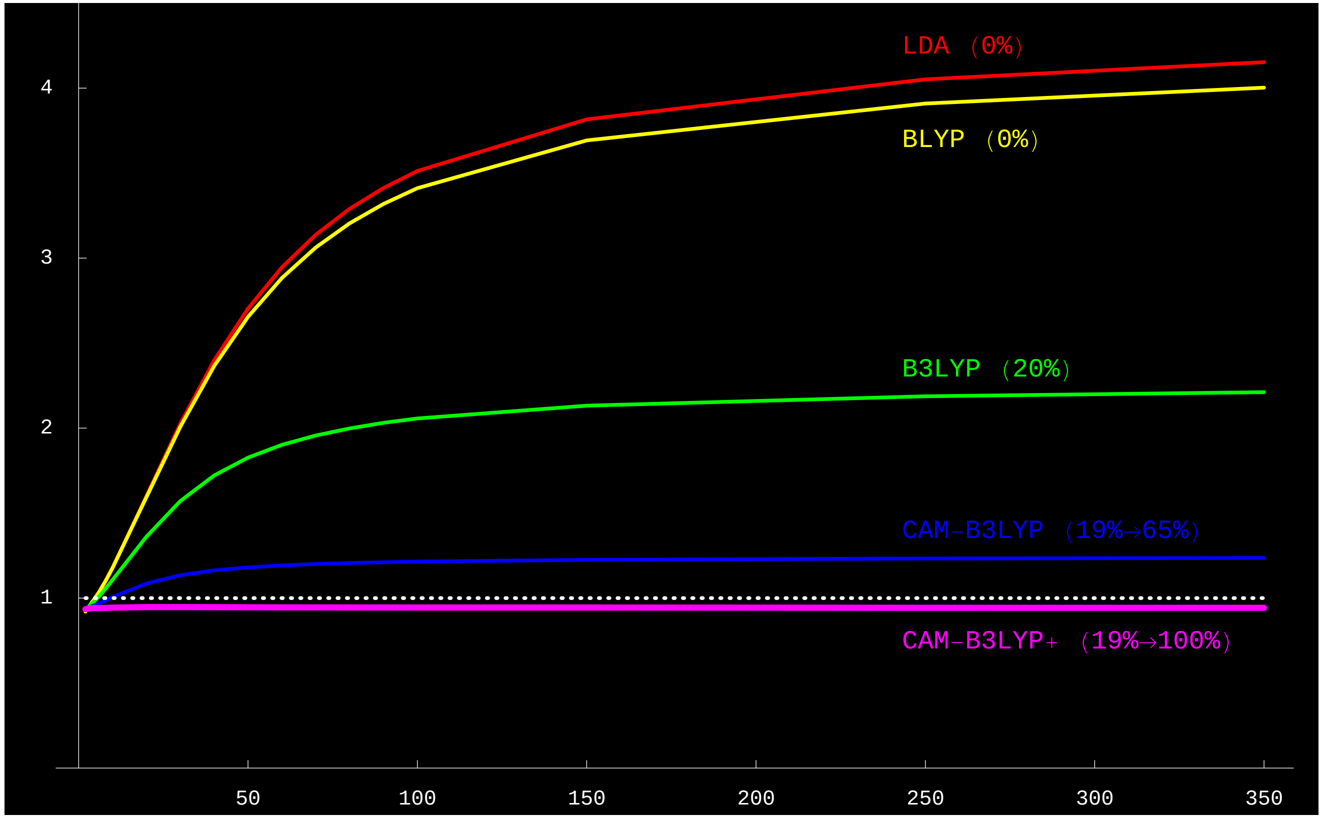
- Without a good description of long-range exchange, the systems become too polarizable



  - the Hartree model neglects all exchange and overestimates by a factor of eight
  - pure DFT has a poor long-range exchange and overestimates by a factor of four
  - hybrid functionals improve the situation, introducing some exact exchange
  - compromise solution: standard DFT at short range, full exchange at long range

The importance of exact exchange for longitudinal polarizabilities

- Without a good description of long-range exchange, the systems become too polarizable



  - the Hartree model neglects all exchange and overestimates by a factor of eight
  - pure DFT has a poor long-range exchange and overestimate by a factor of four
  - hybrid functionals improve the situation, introducing some exact exchange
  - compromise solution: standard DFT at short range, full exchange at long range

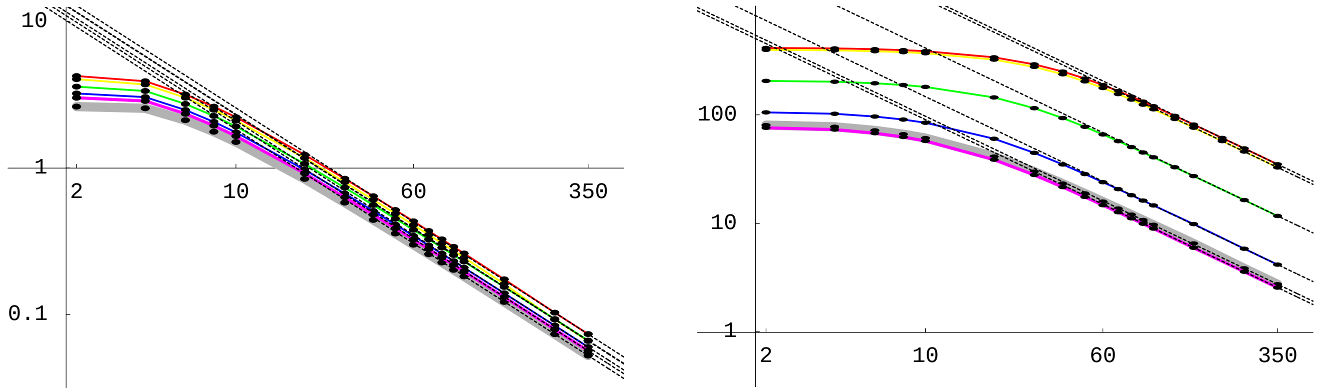
Asymptotic behaviour of group polarizabilities

- How does the group polarizability converge towards the infinite limit?

$$\overline{\alpha}_\infty - \overline{\alpha}_N = eN^{-1} + \mathcal{O}(N^{-2})$$ Kudin *et al.*, JCP **122**, 134907 (2005)

  – this behaviour is universal, holding at all levels of theory

- Log–log plots of $\overline{\alpha}_\infty - \overline{\alpha}_N$ for alkanes and alkenes:



  – limit obtained by extrapolation $\overline{\alpha}_\infty = (\alpha_N - \alpha_M)/(N - M)$

  – straight lines of slope -1 superimposed through the points at $N = 350$

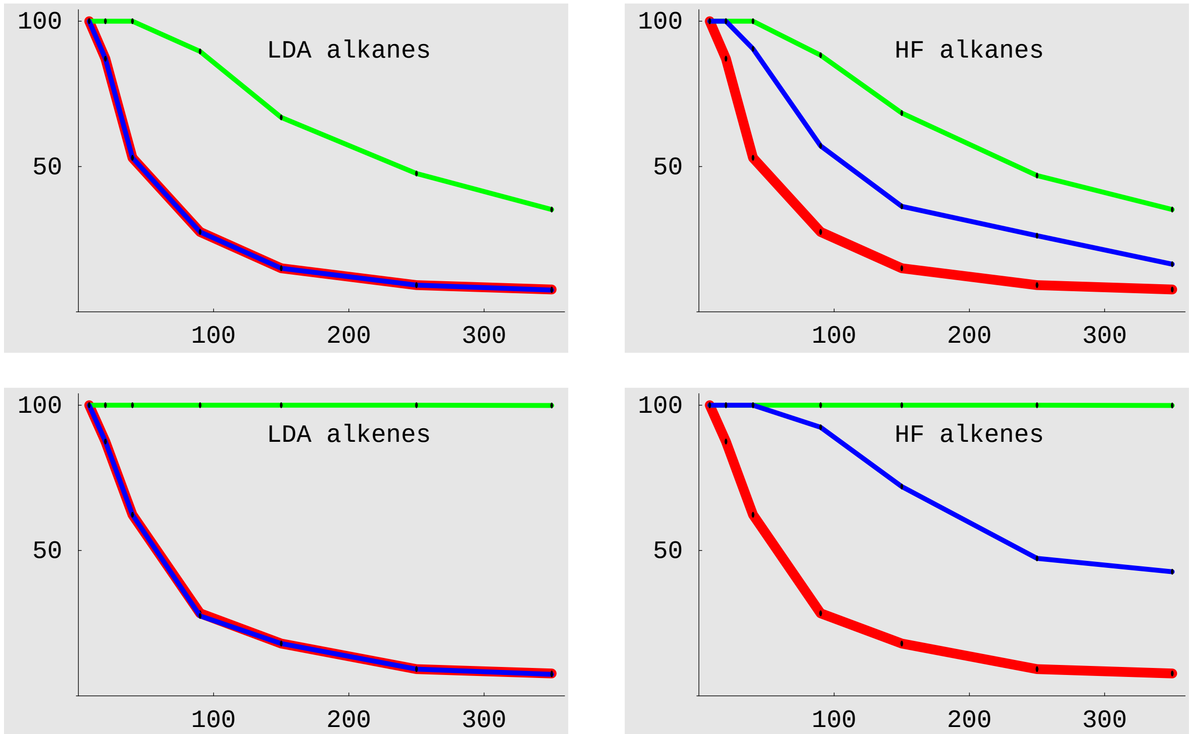
- The asymptotic region is reached with $C_{30}H_{62}$ (alkanes) and $C_{60}H_{62}$ (alkenes)

  – alkane $\overline{\alpha}_\infty$ predicted to within 1% from $C_{30}H_{62}$

  – alkene $\overline{\alpha}_\infty$ predicted to within 1% from $C_{60}H_{62}$ for HF and from $C_{150}H_{152}$ for LDA

Sparsity of linear alkanes and alkenes

- Each energy optimizations were converged in 6–14 SCF iterations
  - about 70 matrix multiplications for each TRRH step (diagonalization)
  - about 50 matrix multiplications for each TRDMS step (DIIS)
- Each polarizability component required 3–4 response iterations
  - about 20 matrix multiplications in each iteration
- Percentage of matrix elements greater than 10^{-6} in alkane and alkene chains
  - overlap matrix:
    - sparse
  - density matrix:
    - nonsparse for alkenes
    - sparse for alkanes
  - Fock/KS matrices:
    - KS matrix like overlap
    - Fock matrix intermediate between overlap and density matrices



Conclusions

- We have discussed the optimization of SCF energies without MOs
  - in each SCF iteration, we replace diagonalization by minimization
  - minimization by Newton's method, one step is usually enough
  - minimization stable and fast, highly competitive with diagonalization
  - 50–100 sparse matrix multiplications required
- Large molecules represent a more difficult minimization problem
  - small Hessian eigenvalues for pure DFT
  - trust-region SCF: careful step-size control
  - revert to second-order if necessary
- Linear-response is straightforward in the AO basis
  - one Fock/Kohn–Sham matrix build and 20 matrix multiplications pr. iteration
  - stable convergence in 3–5 iterations