Linear-scaling methods for molecular energies and properties

Trygve Helgaker

Centre for Theoretical and Computational Chemistry, University of Oslo, Norway

Francesca Iozzi, Andreas Krapp, Filip Pawłowski, Simen Reine, Erik Tellgren

Stinne Høst, Branislav Jansík, Poul Jørgensen, Jeppe Olsen, University of Aarhus, Denmark

Sonia Coriani, University of Trieste, Italy

Paweł Sałek, Royal Institute of Technology, Sweden

Multiscale Modeling and Simulation in Science

AlbaNova University Campus, Stockholm, Sweden

November 24, 2009

Overview

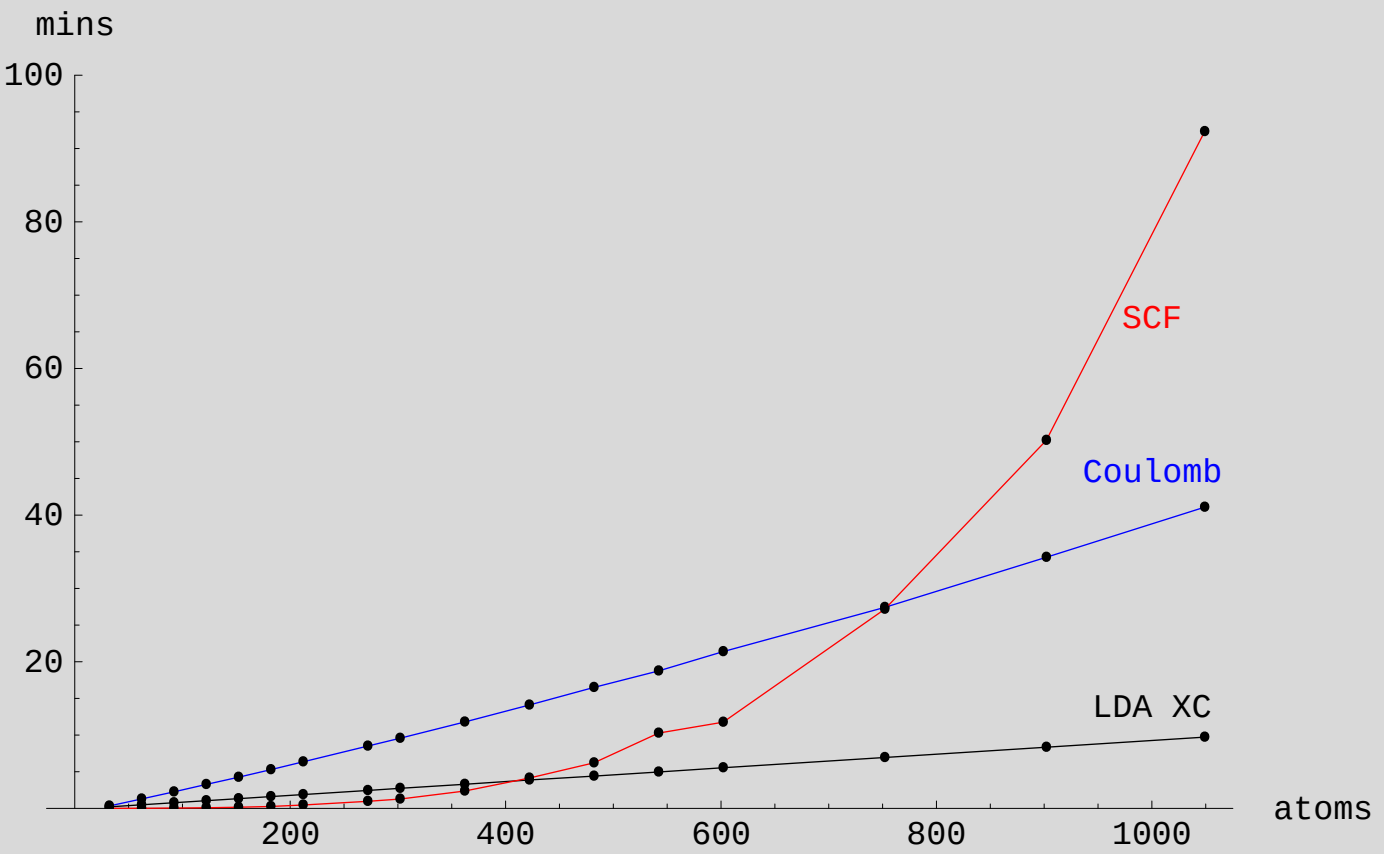
- The representation of orbitals in large systems
  - canonical molecular orbitals (MOs) should be avoided
  - iterative Löwdin orthonormalization of AOs
- The optimization of the energy
  - traditional approach: two-step Roothaan–Hall diagonalization and DIIS averaging
  - augmented Roothaan–Hall (ARH) method: one-step optimization of density matrix
- Fast integration techniques for energies and forces
  - linear-scaling density-fitting and multipole techniques
  - progress report for Dalton code
- The evaluation of response properties
  - density-matrix formulation of linear response theory
- Calculation of excitation energies in DFT
  - Kohn–Sham often fails for charge-transfer excitations
  - diagnostic for excitation energies

Calculation of energies and properties: new techniques suited large systems

- With improvements in computer technology and computational techniques, quantum chemistry is being applied to increasingly large systems
  - Hartree–Fock and Kohn–Sham theory can nowadays be applied to hundreds of atoms
- However, larger systems have always presented a problem to quantum chemistry
  - steep increase in computational cost (N^4 or worse)
    - new techniques must be introduced to curb cost (locality)
  - old models may no longer applicable
    - for example, the CI model works poorly for many electrons
    - superseded by the size-extensive coupled-cluster model
  - new physical phenomena become important
    - long-range correlation and exchange may dominate chemistry
    - traditional DFT does not account for these effects
- We shall here consider new methods, better suited for large systems
  - direct optimization of density matrix
  - integral evaluation in linear time
  - calculation of properties from density matrix

Cost of SCF optimizations in large systems

- Consider the optimization of the SCF energy (here LDA) of molecular systems:

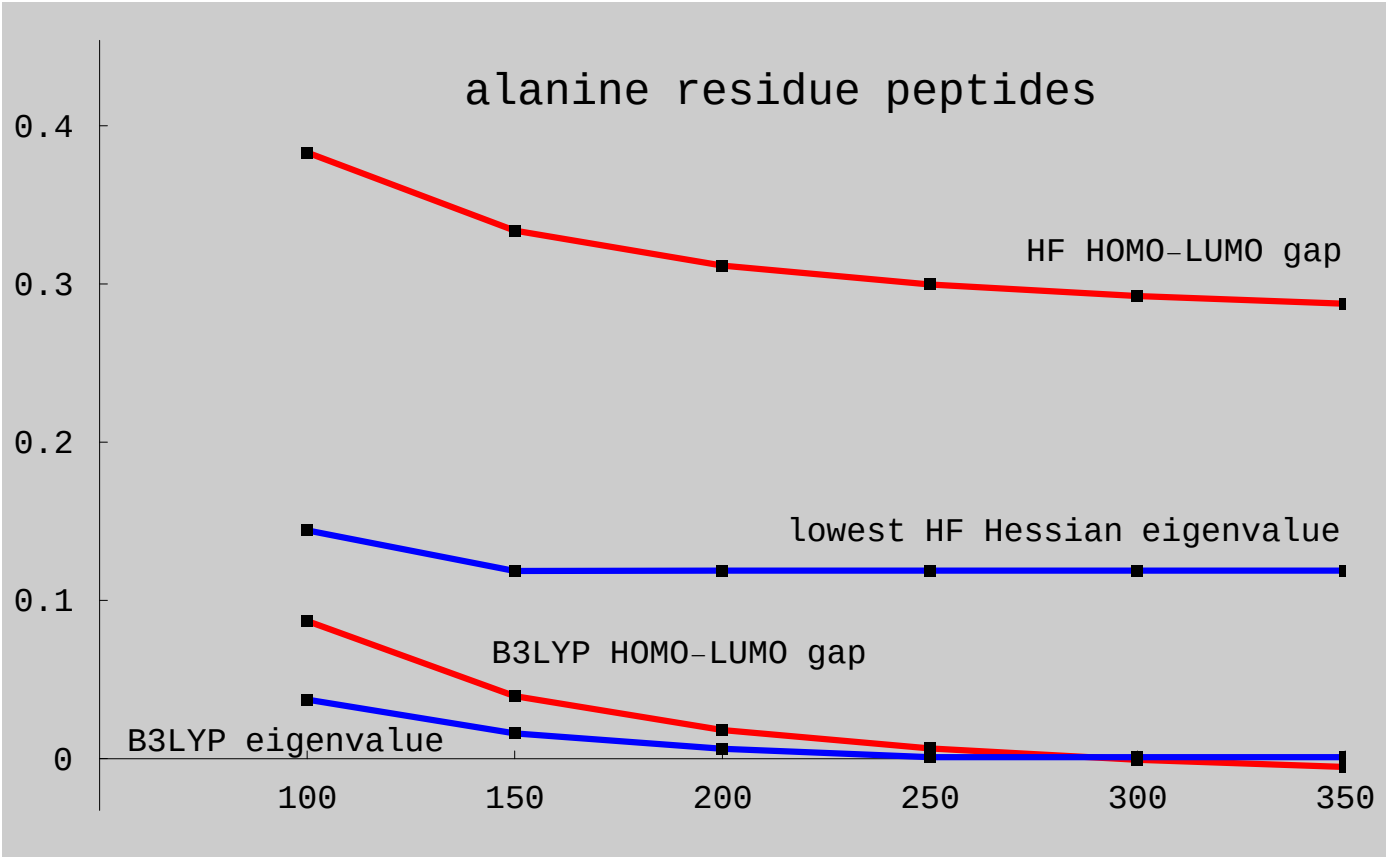


  - small systems dominated by KS-matrix evaluation, with linear scaling
  - large systems dominated by SCF diagonalization, with cubic scaling

- To achieve linear scaling, we must avoid diagonalization (and MOs)!

Difficulty of SCF optimizations in large systems

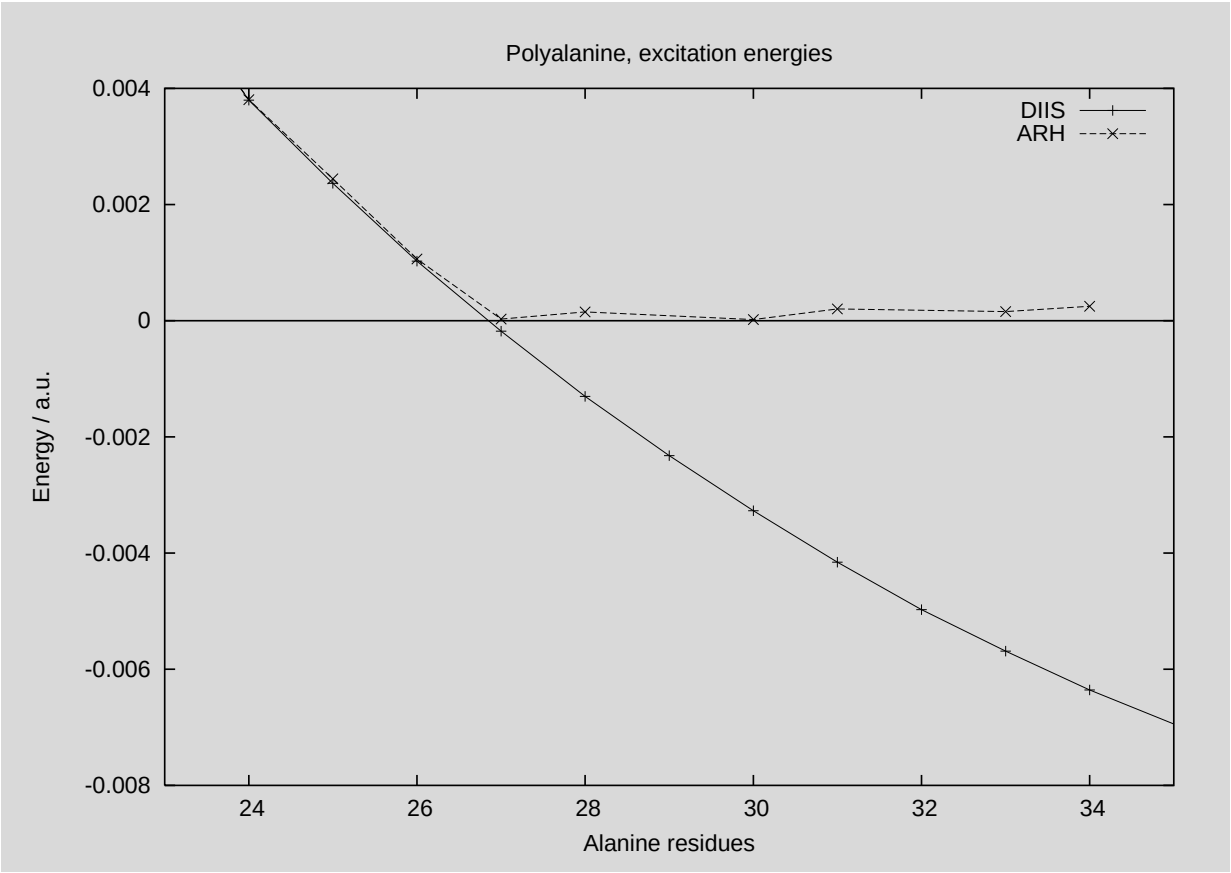
- SCF convergence is typically more difficult in larger systems
  - small (or negative) HOMO–LUMO gaps and small Hessian eigenvalues in DFT
  - 6-31G HOMO–LUMO gap and lowest Hessian eigenvalue in alanine residue peptides



- Large systems therefore force us to reconsider energy optimization

Properties for large systems

- B3LYP/6-31G excitation energies of alanine residue peptides



  - the RH/DIIS scheme converges, for sufficiently large systems, to an excited state
  - the B3LYP ground state becomes unphysical for large systems

- Why does RH/DIIS not yield a ground state? Why is the ground state so strange?

Basis for density-matrix energy minimization

- A natural basis for large molecular systems are atom-fixed atomic orbitals (AOs)
  - well suited to integration and sparse for large systems
  - unfortunately, linear equations are ill-conditioned in this nonorthogonal basis
- To obtain well-conditioned equations we orthonormalize the AO basis

$$\mathbf{Z}^{\mathrm{T}}\mathbf{S}\mathbf{Z} = \mathbf{I}$$

- The Löwdin basis retains greatest similarity with the original AOs

$$\mathbf{Z} = \mathbf{S}^{-1/2}$$

  - retains sparsity but traditionally evaluated by diagonalization
- We apply the iterative Newton–Schulz iterations (Niklasson 2004):

$$\mathbf{Z}_1 = \mathbf{I}, \quad \mathbf{Z}_{n+1} = \frac{1}{2}\mathbf{Z}_n\left(3\mathbf{I} - \mathbf{Z}_n^{\mathrm{T}}\mathbf{S}\mathbf{Z}_n\right), \quad \mathbf{S}^{-1/2} = \lim_{n\to\infty}\mathbf{Z}_n$$

  - convergence guaranteed with scaled $\lambda_{\min}\mathbf{S}$ that minimizes $\|\lambda\mathbf{S} - \mathbf{I}\|_2$
  - approximate 2-norm by Frobenius norm:

$$f(\lambda) = \sqrt{\frac{\mathrm{Tr}(\lambda\mathbf{S}-\mathbf{I})^4}{\mathrm{Tr}(\lambda\mathbf{S}-\mathbf{I})^2}} \;\leftarrow\; \text{cheap lower bound to the 2-norm}$$

  - B. Jansík *et al.*, JCP **126**, 124104 (2007)

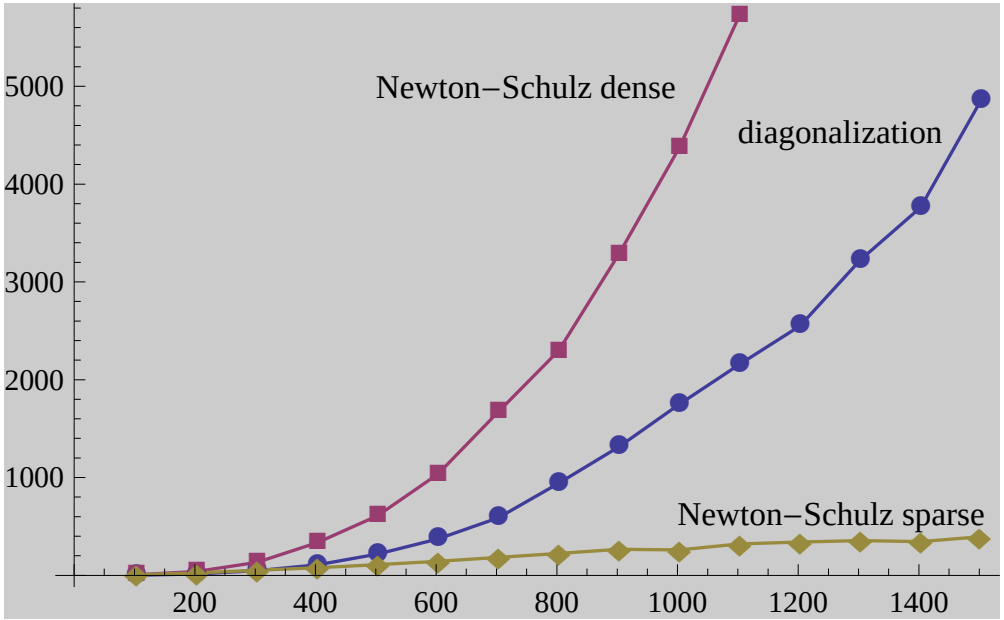
Number of iterations for convergence to 10^{-10}

- Typically 15 iterations and 50 matrix multiplications needed for convergence

	$\|\mathbf{S}-\mathbf{I}\|_2$	$N^{(2)}$	$N^{(3)}$
Fullerene C_{60}	6.0	17	11
Polysaccharide (438)	5.2	14	10
Polysaccharide (875)	5.2	14	10
DNA fragment (583)	4.9	15	11
PRC model (1082)	5.7	16	11
Red phosphorus (506)	1.9	8	6
Red phosphorus (810)	4.7	14	10
Crystobalite (984)	4.4	13	9
Tridymite (1314)	5.2	15	11
Random molecule (64)	16.0	23	27

Optimization in orthonormal Löwdin basis

- We begin all calculations by evaluating $\mathbf{S}^{\pm 1/2}$ iteratively
  - example: polyalanine peptide residues (6-31G) (cost vs. number of atoms)



- Our optimization is then carried out in the Löwdin basis

$$\mathcal{E}(\mathbf{X}) = \mathrm{Tr}\,\mathbf{D}(\mathbf{X})\widetilde{\mathbf{h}} + \tfrac{1}{2}\,\mathrm{Tr}\,\mathbf{D}(\mathbf{X})\widetilde{\mathbf{G}}(\mathbf{D}(\mathbf{X}))$$

  by transforming all integrals in the following manner

$$\widetilde{\mathbf{h}} = \mathbf{S}^{-1/2}\mathbf{h}\mathbf{S}^{-1/2}$$

- In this basis, all operations are well conditioned and no overlaps appear $\widetilde{\mathbf{S}} = \mathbf{I}$

Newton's method and self-consistent field (SCF) theory

- The calculation of electronic energies is a classical problem of optimization:

  > "Before 1940 relatively little was known about methods for numerical optimization of functions in many variables. There had been some least squares calculations carried out, and steepest descent type methods had been applied in some physics problems. The Newton method in many variables was known, and more sophisticated methods were being attempted such as the self-consistent field for variational problems in theoretical chemistry."
>
> R. Fletcher, *Practical Methods of Optimization*, Wiley, Chichester, 1980.

- We shall here explore the relationship between Newton's method and the SCF method
  - Hartree–Fock (HF) and Kohn–Sham (KS) energies
- The traditional SCF optimization method consists of two alternating steps:
  - the Roothaan–Hall (RH) diagonalization step
  - Pulay's DIIS density-averaging step
  - both steps are needed for convergence, which is nevertheless sometimes difficult
- We shall see how it is possible to merge the two steps into a single concerted step
  - the augmented Roothaan–Hall (ARH) step
  - more robust convergence

Hartree–Fock and Kohn–Sham energies

- We are interested in energy optimization in HF and KS theories

$$E(\mathbf{D}) = \mathrm{Tr}\,\mathbf{Dh} + \tfrac{1}{2}\,\mathrm{Tr}\,\mathbf{DG}(\mathbf{D}) + E_{\mathrm{XC}}(\mathbf{D})$$

  - we use standard notation, with the two-electron part given by

$$G_{pq}(\mathbf{D}) = \sum\nolimits_{rs} \langle \chi_p \chi_q | r_{12}^{-1} | \chi_r \chi_s \rangle D_{rs} + \text{exchange}$$

  - we will ignore the exchange–correlation term
  - we assume an orthonormal basis (Cholesky or Löwdin AOs) but not MOs

- For an N-electron system, there are certain constraints on the density matrix

$$\mathbf{D}^{\mathrm{T}} = \mathbf{D}, \quad \mathrm{Tr}\,\mathbf{D} = N, \quad \mathbf{D}^2 = \mathbf{D}$$

  - parameterization of the density matrix in terms of orbitals:

$$\mathbf{D} = \mathbf{C}\mathbf{C}^{\mathrm{T}}, \quad \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}$$

  - parameterization of the density matrix in terms of an antisymmetric matrix:

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

- We begin by reviewing the standard SCF optimization scheme

The Roothaan–Hall (RH) diagonalization step

- In the nth iteration, we introduce the Fock matrix constructed from $\mathbf{D}_n$:

$$\mathbf{F}_n = \mathbf{h} + \mathbf{G}(\mathbf{D}_n), \quad \delta\mathbf{D} = \mathbf{D} - \mathbf{D}_n$$

- In terms of this Fock matrix, we may rewrite the energy as:

$$E(\mathbf{D}) = \underbrace{\mathrm{Tr}\,\mathbf{D}\mathbf{h}}_{\text{1-electron}} + \tfrac{1}{2}\underbrace{\mathrm{Tr}\,\mathbf{D}\mathbf{G}(\mathbf{D})}_{\text{2-electron}} = \underbrace{\mathrm{Tr}\,\mathbf{D}\mathbf{F}_n}_{\text{large}} + \tfrac{1}{2}\underbrace{\mathrm{Tr}\,\delta\mathbf{D}\mathbf{G}(\delta\mathbf{D})}_{\text{small}} - \underbrace{\tfrac{1}{2}\mathrm{Tr}\,\mathbf{D}_n\mathbf{G}(\mathbf{D}_n)}_{\text{constant } k_n}$$

- Assuming that $\delta\mathbf{D}$ is small, we neglect the term quadratic in $\delta\mathbf{D}$:

$$E(\mathbf{D}) \approx E_{\mathrm{RH}}(\mathbf{D}) = \mathrm{Tr}\,\mathbf{D}\mathbf{F}_n + k_n \;\leftarrow\; \text{Roothaan–Hall energy}$$

- To minimize the RH energy, we use the parameterization $\mathbf{D} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$ and obtain

$$E_{\mathrm{RH}}(\mathbf{D}) = \mathrm{Tr}\,\mathbf{C}^{\mathrm{T}}\mathbf{F}_n\mathbf{C} + k_n$$

  - variation in $\mathbf{C}$ under the constraints $\mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}$ yields the RH eigenvalue problem:

$$\mathbf{F}_n\mathbf{C} = \mathbf{C}\boldsymbol{\varepsilon}$$

- In the RH step, we perform an exact minimization of an approximate HF energy
  - it may or may not converge upon iteration

Pulay's DIIS step

- To obtain the nth DIIS step, we begin with the same energy decomposition:

$$E(\mathbf{D}) = \underbrace{\mathrm{Tr}\,\mathbf{D}\mathbf{F}_n}_{\text{large}} + \tfrac{1}{2}\underbrace{\mathrm{Tr}\,\delta\mathbf{D}\mathbf{G}(\delta\mathbf{D})}_{\text{small}} + k_n$$

- However, we now retain the quadratic term but linearize the density matrix:

$$\delta\mathbf{D} = \exp(-\mathbf{X})\mathbf{D}_n\exp(\mathbf{X}) - \mathbf{D}_n \approx [\mathbf{D}_n, \mathbf{X}], \quad \mathbf{X}^{\mathrm{T}} = -\mathbf{X}$$

  – variation with respect to $\mathbf{X}$ now gives the stationary condition:

$$[\mathbf{D}_n, \mathbf{F}_n + \mathbf{G}(\delta\mathbf{D})] = \mathbf{0}$$

- To avoid the high cost of the G term, we expand $\delta\mathbf{D}$ in old density matrices:

$$\delta\mathbf{D} \approx \sum\nolimits_i c_i(\mathbf{D}_i - \mathbf{D}_n) \quad \Rightarrow \quad \mathbf{G}(\delta\mathbf{D}_n) = \sum\nolimits_i c_i(\mathbf{F}_i - \mathbf{F}_n)$$

  – the stationary condition now becomes

$$\sum\nolimits_{i=0}^{n}[\mathbf{D}_n, c_i\mathbf{F}_i] = \mathbf{0}, \quad \sum\nolimits_{i=0}^{n} c_i = 1 \;\leftarrow \text{ approx. satisfied by norm min.}$$

- In the DIIS step, we perform a quasi-Newton step with a linearized density matrix
  – it corrects the RH step for the neglect of the G term

The traditional RH–DIIS method

- The traditional SCF optimization alternates between two steps

$$E(\mathbf{D}) = \underbrace{\mathrm{Tr}\,\mathbf{D}\mathbf{F}_n}_{\text{large}} + \tfrac{1}{2}\underbrace{\mathrm{Tr}\,\delta\mathbf{D}\mathbf{G}(\delta\mathbf{D})}_{\text{small}} + k_n$$

  - RH step: minimize $E(\mathbf{D}(\mathbf{X}))$ by linearizing the energy dependence on $\delta\mathbf{D}$
  - DIIS step: minimize $E(\mathbf{D}(\mathbf{X}))$ by linearizing the density-matrix dependence on $\mathbf{X}$
  - linearization means neglect of second-order (Hessian) information
- In none of these steps do we make a full use of all available Hessian information
  - the iterations may therefore oscillate, diverge or converge to an incorrect solution
  - we shall now see how it is possible to combine these separate steps into one
- To achieve this, we return to the full energy expression as a function of $\mathbf{X}$
  - we shall straightforwardly apply Newton's method
- Learning from SCF theory, we then simplify the Newton step
  - a single step that merges the RH and DIIS steps into one
  - the augmented Roothaan–Hall (ARH) method

Newton's method applied to the SCF energy

- We expand the full SCF energy in orders of $\mathbf{X}$ abound $\mathbf{D}_n$:

$$\mathcal{E}(\mathbf{X}) = \mathrm{Tr}\,\delta\mathbf{D}\mathbf{F}_n + \tfrac{1}{2}\,\mathrm{Tr}\,\delta\mathbf{D}\mathbf{G}(\delta\mathbf{D}) + k_n$$

$$\delta\mathbf{D} = [\mathbf{D}_n, \mathbf{X}] + \tfrac{1}{2}[[\mathbf{D}_n, \mathbf{X}], \mathbf{X}] + \cdots, \quad \mathbf{X}^{\mathrm{T}} = -\mathbf{X}$$

- We truncate the expansion at second order and minimize within the trust region

$$\|\mathbf{X}\| \leq h_n \quad \leftarrow \text{trust-region radius}$$

- This constrained minimization leads to the level-shifted Newton step

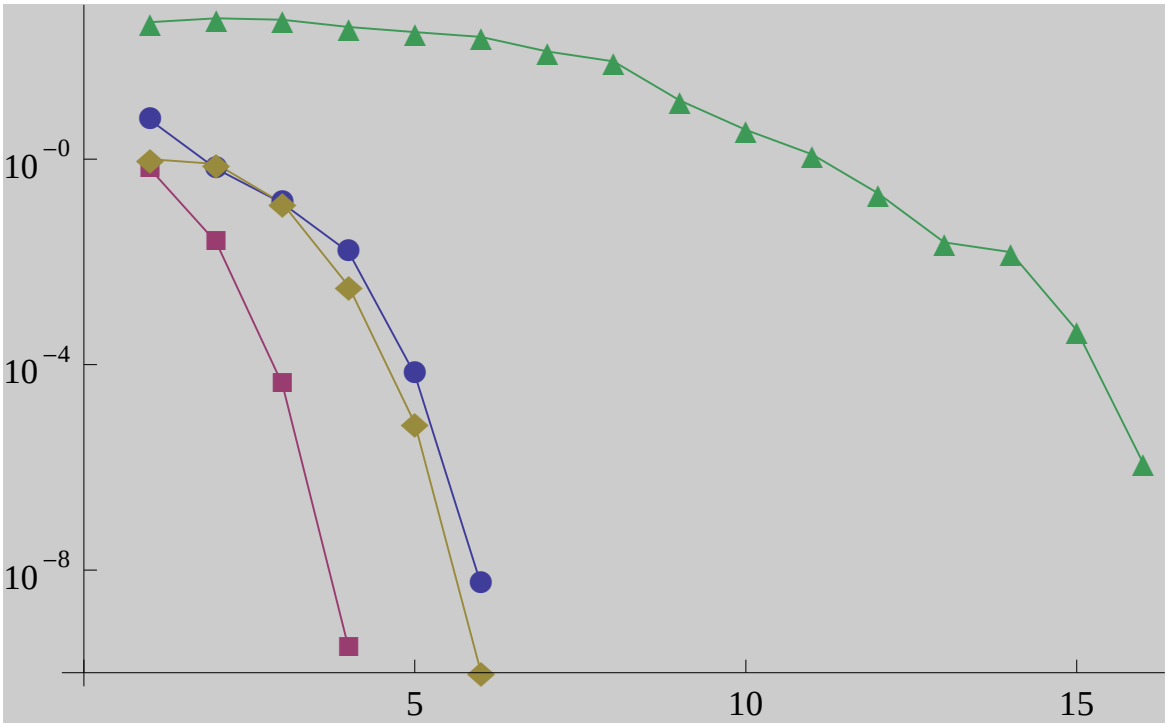
$$\sum_{\nu} \left(\frac{\partial^2 \mathcal{E}(\mathbf{X})}{\partial X_\mu \partial X_\nu} - \lambda\,\delta_{\mu\nu} \right) X_\nu = -\frac{\partial \mathcal{E}(\mathbf{X})}{\partial X_\mu}$$

  - gradient and Hessian carry information about slope and curvature, respectively
  - the level-shift parameter λ is adjusted to restrict the search inside the trust region

- Newton's method has been applied to the SCF energy by other authors
  - G. B. Bacskay (1981), H. J. Aa. Jensen and H. Ågren (1984)
  - we modify the density matrix $\mathbf{D}(\mathbf{X})$ directly
  - no MOs are introduced and diagonalization is avoided

Some examples of Newton optimizations

- Newton density-matrix optimizations of the total electronic energy
  - water HF/DZ (red), water dimer B3LYP/DZ (yellow), adenine HF/STO-3G (blue), rhodium complex HF/STO-3G+ADZ (green)



- Convergence fast in the local region, slow but smooth in global region
  - global convergence guaranteed by the trust-region method
  - a large number of Fock/KS matrix reevaluations are needed: 24, 42, 54, 211

The structure of the Newton matrix equation

- The structure of the Newton matrix equation reflects that of the energy:

$$\underbrace{(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}})\,\mathbf{X} + \mathbf{X}\,(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}})}_{\text{large 2nd-order F term}} + \underbrace{\mathbf{G}^{\mathrm{ov}}([\mathbf{D}_n, \mathbf{X}]) - \mathbf{G}^{\mathrm{vo}}([\mathbf{D}_n, \mathbf{X}])}_{\text{small 2nd-order G term}} = \underbrace{\mathbf{F}_n^{\mathrm{vo}} - \mathbf{F}_n^{\mathrm{ov}}}_{\text{1st-order F term}}$$

  - projections onto the occupied and virtual orbital spaces

$$\mathbf{F}_n = \mathbf{F}_n^{\mathrm{oo}} + \mathbf{F}_n^{\mathrm{ov}} + \mathbf{F}_n^{\mathrm{vo}} + \mathbf{F}_n^{\mathrm{vv}}, \quad (\mathbf{P}_{\mathrm{o}} = \mathbf{D}_n,\, \mathbf{P}_{\mathrm{v}} = \mathbf{I} - \mathbf{D}_n)$$

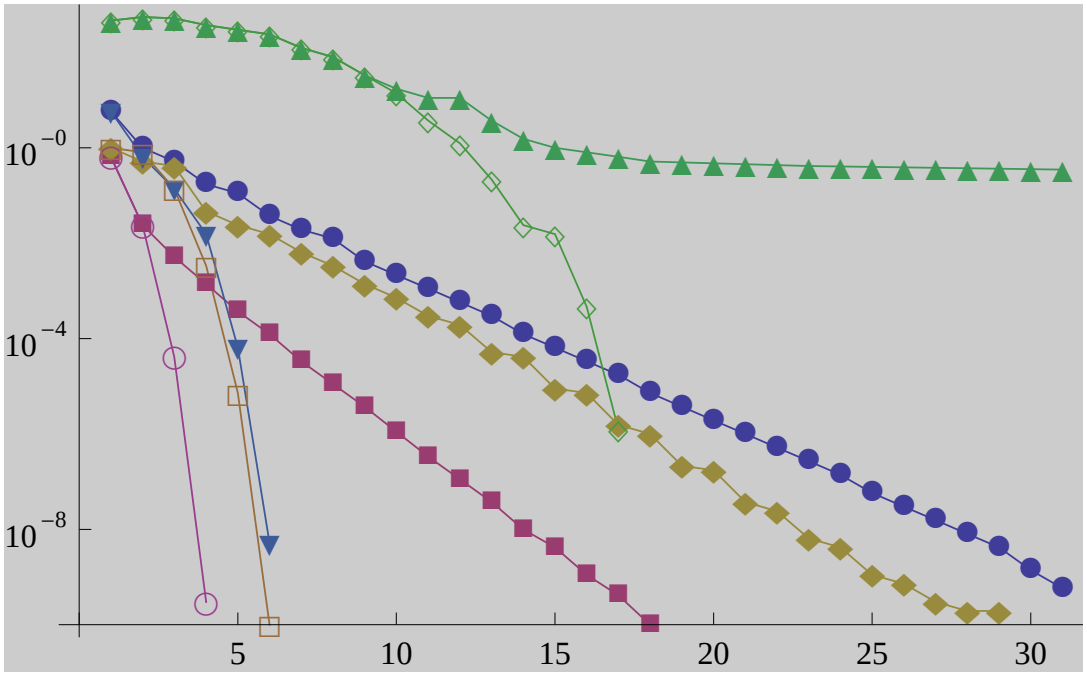
  - note: in the MO basis, the first term consists of orbital-energy differences
- The Newton matrix equation is solved iteratively, in about ten iterations
  - at each iteration, we evaluate the left-hand side with a trial $\mathbf{X}$ matrix
- The (typically large) F contribution to the Hessian transformation is inexpensive
  - requires only matrix multiplications
- The (typically small) G contribution to the Hessian transformation is expensive
  - requires a new two-electron integral evaluation
- Perhaps we can approximate the G term, in the spirit of DIIS?
  - let us first drop the G term altogether, as done in the RH step!

A (very) simplified Newton method

- Neglecting the G term, we obtain the following simple matrix equation

$$(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}})\,\mathbf{X} + \mathbf{X}\,(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}}) = \mathbf{F}_n^{\mathrm{vo}} - \mathbf{F}_n^{\mathrm{ov}}$$

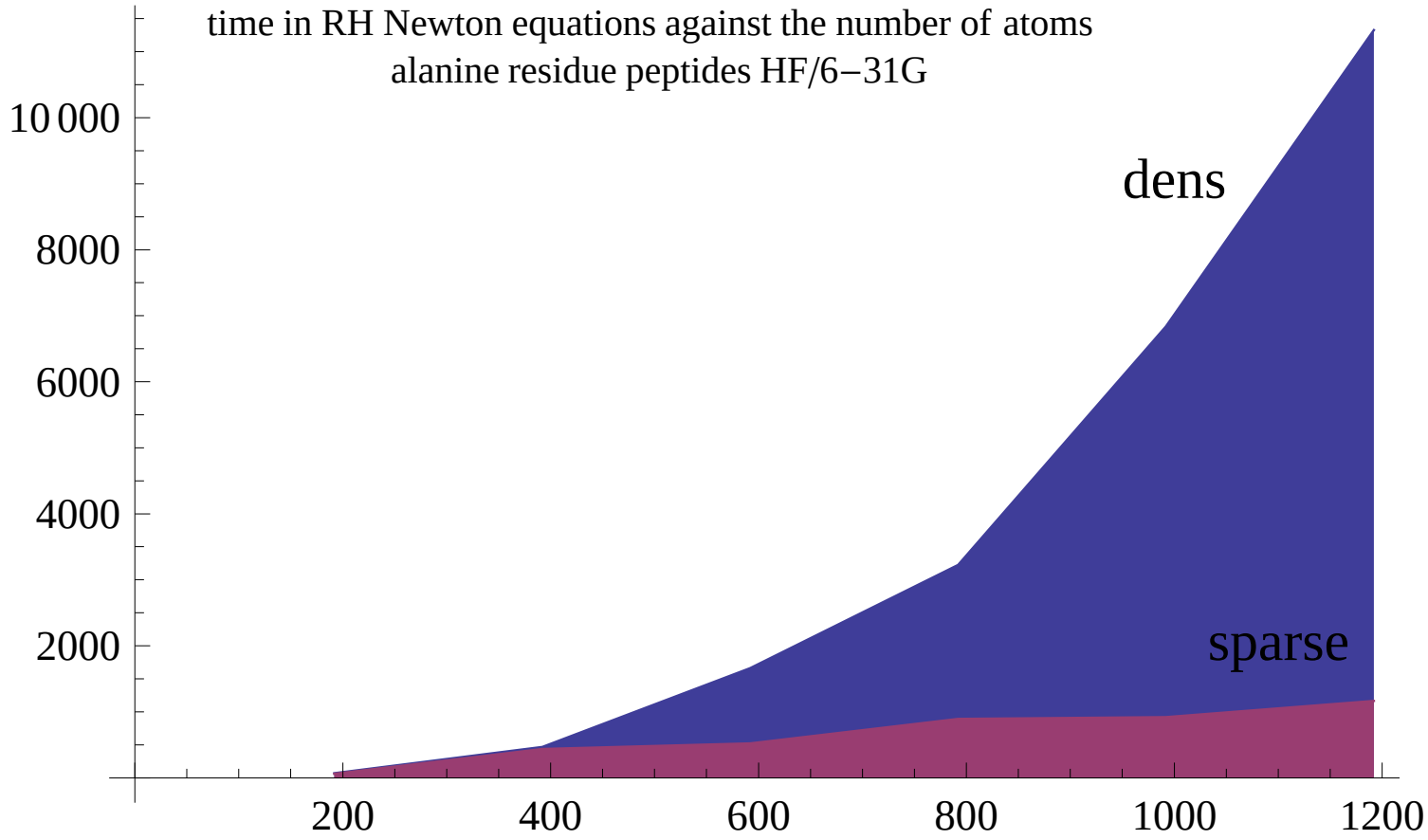
- But energy convergence suffers since the Newton step is now approximate
  - water (red), water dimer (yellow), adenine (blue), rhodium complex (green)



  - linear convergence in the local region, global convergence not guaranteed
  - the number of Fock/KS matrix reevaluations: 20 (24), 50 (42), 33 (54), ∞ (211)

The Roothaan–Hall Newton method

- A byproduct: by removing the G term, we are minimizing the RH energy
  - we are in fact carrying out a Roothaan–Hall diagonalization step!
  - an alternative to diagonalization for large systems
  - 50–100 sparse matrix multiplications



The augmented Roothaan–Hall method

- Consider the expensive part of the SCF electronic Hessian

$$\mathbf{H}_n(\mathbf{X}) = \mathbf{G}^{\mathrm{ov}}([\mathbf{D}_n, \mathbf{X}]) - \mathbf{G}^{\mathrm{vo}}([\mathbf{D}_n, \mathbf{X}])$$

- During the optimization, we have collected a set of density matrices

$$\mathbf{D}_{in} = \mathbf{D}_i - \mathbf{D}_n$$

  - in these directions, we invoke the quasi-Newton assumption (exact in HF theory)

$$\mathbf{G}(\mathbf{D}_{in}) = \mathbf{G}(\mathbf{D}_i) - \mathbf{G}(\mathbf{D}_n) \approx \mathbf{F}_i - \mathbf{F}_n = \mathbf{F}_{in} \leftarrow \text{available}$$

- To evaluate the Hessian contribution approximately, we introduce a projector onto $\mathbf{D}_{in}$:

$$\mathbf{P}_n(\mathbf{D}_{in}) = \mathbf{D}_{in}, \quad \mathbf{P}_n^2 = \mathbf{P}_n$$

  - we now obtain the following approximate Hessian contribution (JO):

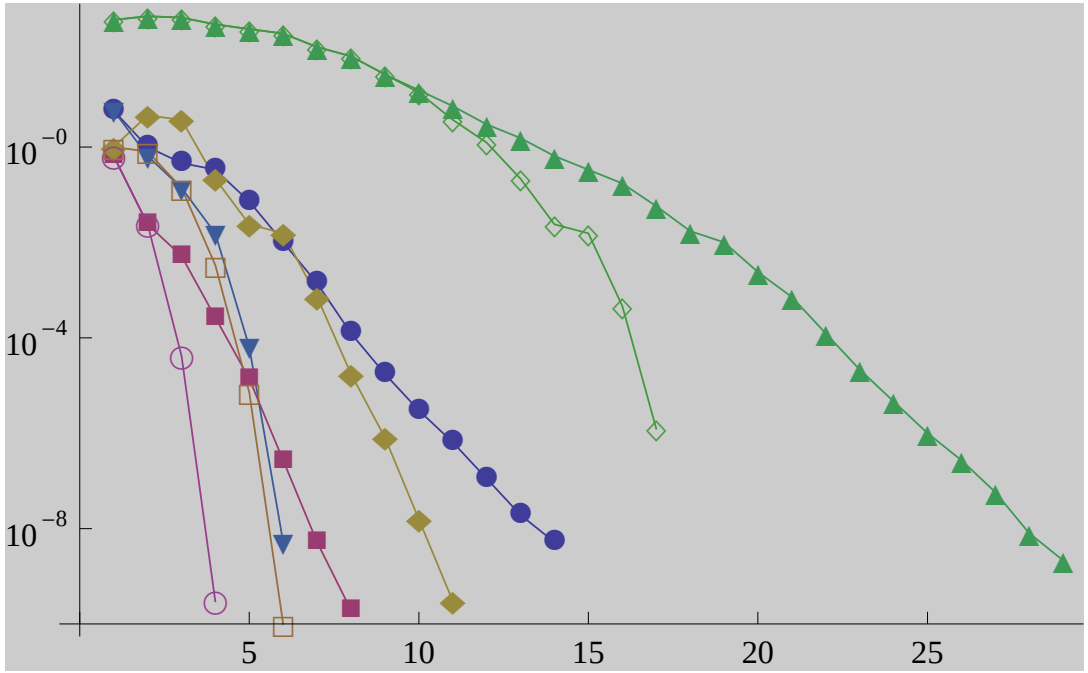
$$\overline{\mathbf{H}}_n(\mathbf{X}) = \sum\nolimits_{ij} (\mathbf{F}_{in}^{\mathrm{ov}} - \mathbf{F}_{in}^{\mathrm{vo}}) \left[\mathbf{T}^{-1}\right]_{ij} \mathrm{Tr}(\mathbf{D}_{jn}[\mathbf{D}_n, \mathbf{X}]), \quad T_{ij} = \mathrm{Tr}\, \mathbf{D}_{in}\mathbf{D}_{jn}$$

  - no new Fock/KS matrix evaluations are needed

- We term the resulting method the augmented Roothaan–Hall (ARH) method
  - a quasi-Newton method with the dominant part of the Hessian treated exactly

Examples of ARH optimizations

- The ARH method is similar to Newton globally but slower locally
  - water (red), water dimer (yellow), adenine (blue), rhodium complex (green)



- It requires the fewest number of Fock/KS evaluations (for the four systems):

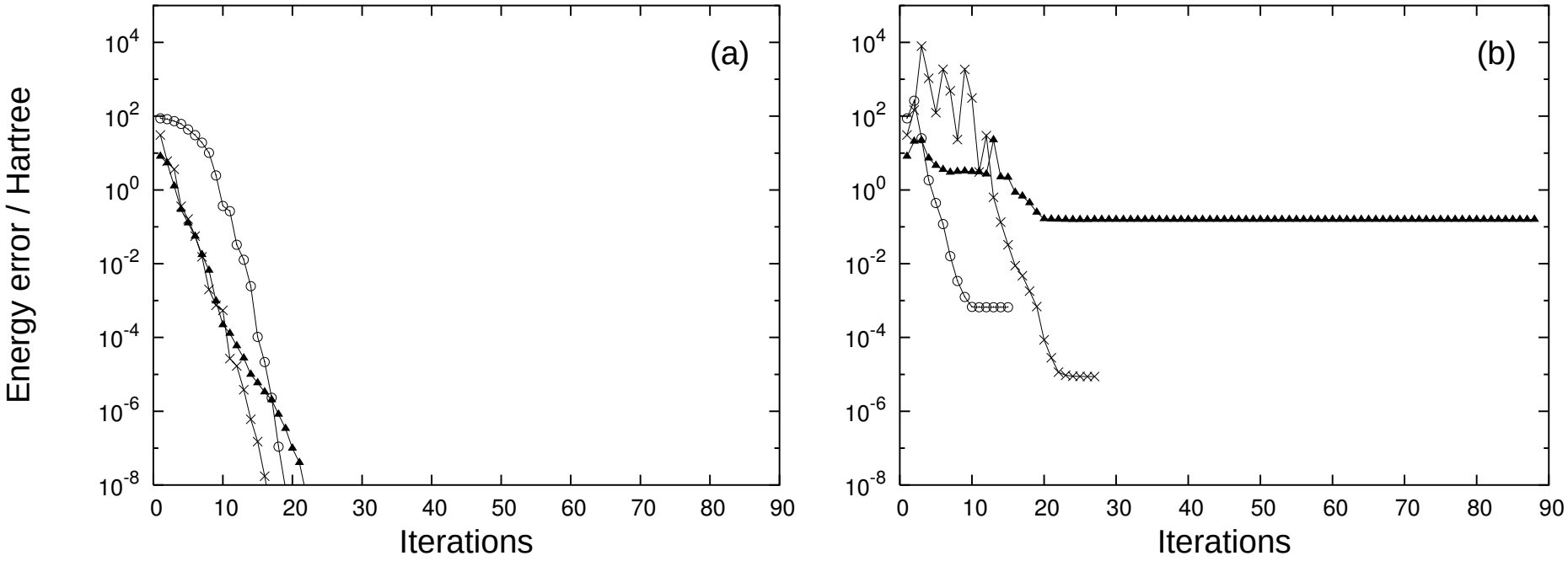
Newton	24	42	54	211
RH	20	50	33	∞
ARH	9	15	15	30

The performance of the ARH algorithm

- The ARH method appears to be robust for optimizations
  - it treats the dominant part of the Hessian exactly (like the RH step)
  - it treats the remainder in a quasi-Newton fashion (like the DIIS step)
  - it makes full use of all available Hessian information in a single, concerted step
- In the following, we shall compare it with the RH–DIIS method
  - convergence to gradient norm smaller than 10^{-4}
  - HCORE of Hückel starting guess
- Høst *et al.*, PCCP **10**, 5344–5348 (2008), JCP **129**, 124106 (2008)

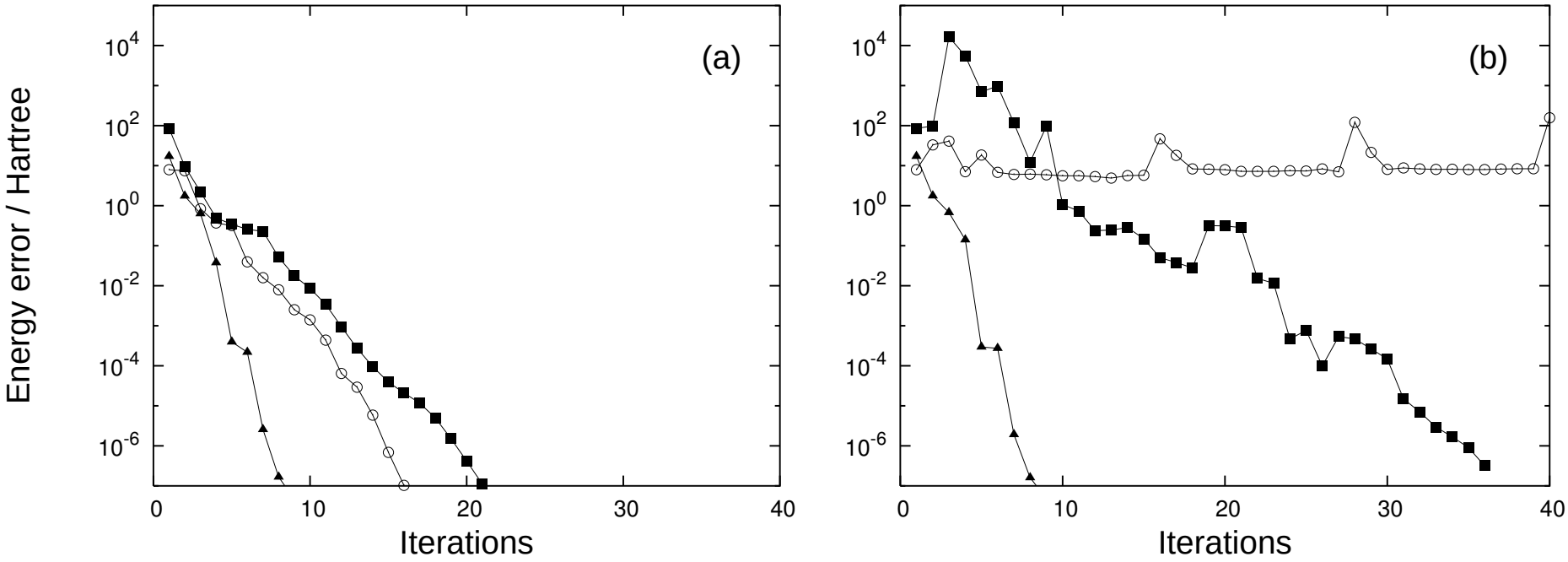
The ARH method compared with RH–DIIS (I)

- Convergence for a variety of molecules
  - cadmium–imidazole complex (circles) B3LYP/3-21G (10/89)
  - 29-residue polyalanine peptide (crosses) B3LYP/6-31G (292/1599)
  - model vitamin B12 (triangles) HF/AhlrichsVDZ (74/428)
- Energy-error on a logarithmic scale (ARH left, RH–DIIS right)
  - RH–DIIS converges to a saddle point, ARH always to a minimum
  - standard programs perform in the same manner



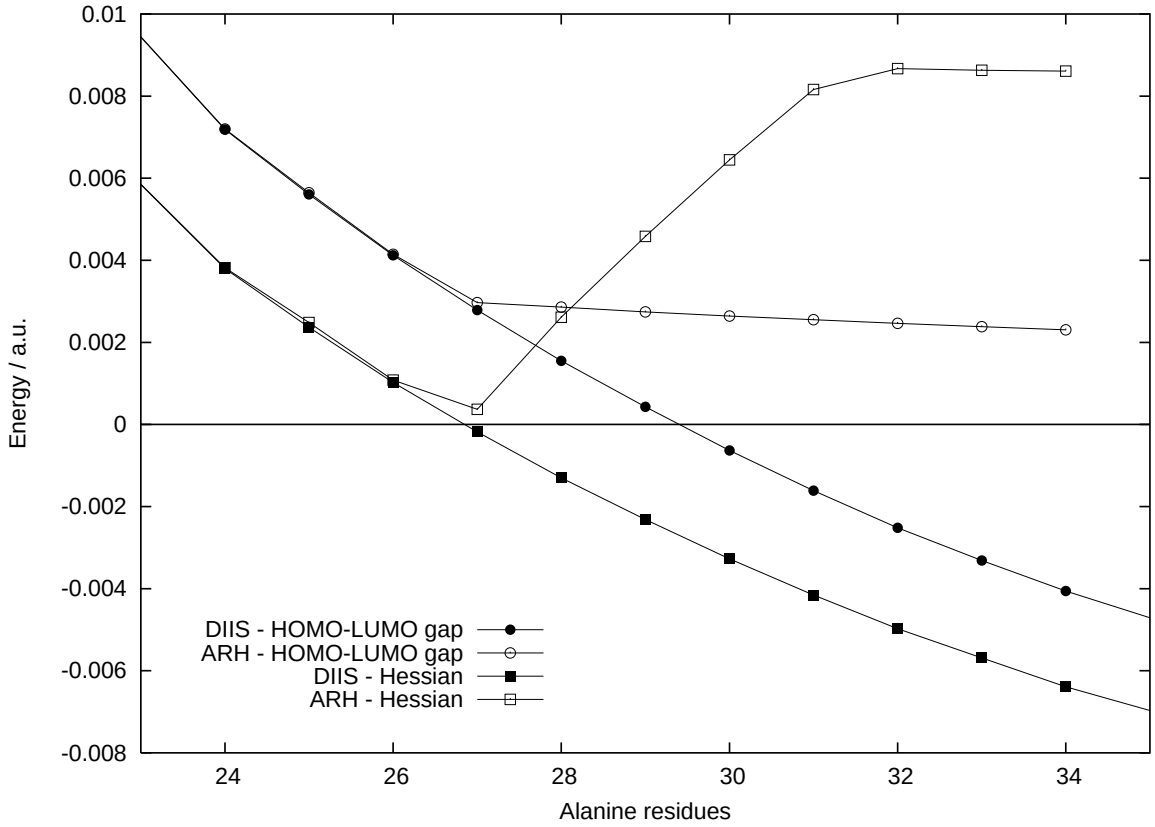
The ARH method compared with RH–DIIS (II)

- Convergence of calculations for a variety of molecules
  - cluster of 51 water molecules (full triangles) B3LYP/cc-pVTZ (153/2958)
  - insulin (full squares) B3LYP/6-21G (786/4417)
  - model vitamin B12 (empty circles) BP86/AhlrichsVDZ (74/428)
- Energy-error on a logarithmic scale (ARH left, RH–DIIS right)
  - ARH converges smoothly in all cases
  - RH–DIIS good for water cluster, oscillates for insulin, diverges for vitamin B12

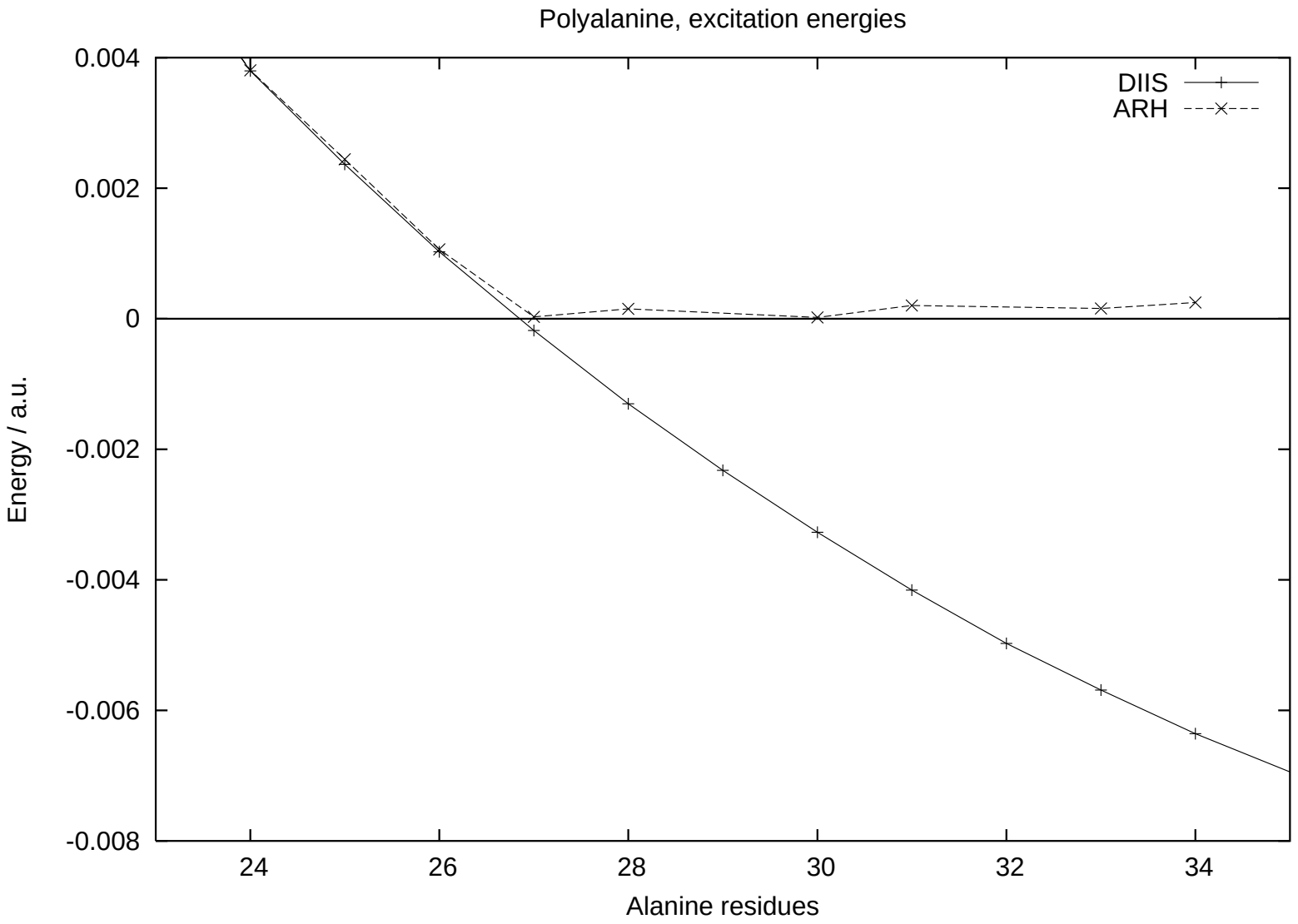


Optimization of polyalanine peptides (B3LYP)

- The RH–DIIS method converges well for less than 29 alanine residues
  - for larger peptides, a level shift of 0.01 enforces convergence
  - however, the solution then becomes a saddle point
  - the same behavior observed for standard programs
- The ARH method converges well for all systems



B3LYP/6-31G excitation energies of alanine residue peptides



Three-level optimization

- The ARH scheme has recently been implemented within a three-level (3L) scheme:
  - grand-canonical atomic optimization
  - valence-basis molecular optimization
  - full-basis molecular optimization
- Comparison with other codes for 23 transition-metal complexes
  - test set by van Lenthe *et al.* (2006), for their ATOMS scheme

	number of iterations		
	best (%)	aver. (%)	worst (%)
van Lenthe *et al.*	100	100	100
ARH–3L	64	76	100
QCP1	91	109	141
QCP2	114	157	205

- B. Jansík *et al.*, JCTC **5**, 1027 (2009)

Review and preview

- The traditional SCF scheme alternates between two steps
  - the Roothaan–Hall diagonalization step
  - the DIIS averaging step
  - neither step makes full use of the available Hessian information
- In the augmented RH (ARH) method, the two steps are merged into one concerted step
  - the dominant part of the Hessian (orbital-energy differences) is treated exactly
  - the remainder is treated accurately in the space of the previous density matrices
  - trust-region method used for step-size control
  - no diagonalization necessary—sparsity may be fully exploited
- We are now going to review integral evaluation
  - density fitting, screening, multipole expansion
  - solid harmonics from Hermite Gaussians
  - gradient evaluation

Density fitting

- Consider the DFT energy expression

$$E = T_{\rm s}[\rho] + \int \rho(\mathbf{r}) v(\mathbf{r})\, d\mathbf{r} + J[\rho] + E_{\rm xc}[\rho], \quad J[\rho] = \frac{1}{2}(\rho|\rho)$$

- Primary and secondary expansions of the density:

$$\rho(\mathbf{r}) = \sum_{ab} D_{ab}\, \chi_a(\mathbf{r})\chi_b(\mathbf{r}), \quad \widetilde{\rho}(\mathbf{r}) = \sum_{\alpha} c_\alpha \omega_\alpha(\mathbf{r}), \quad \widetilde{\rho}(\mathbf{r}) \approx \rho(\mathbf{r})$$

- We may now express the Coulomb interaction as

$$J[\rho] = \underbrace{\tfrac{1}{2}\left[(\rho\,|\widetilde{\rho}) + (\widetilde{\rho}\,|\rho) - (\widetilde{\rho}\,|\widetilde{\rho})\right]}_{\tilde{J}\ \text{(three center)}} + \underbrace{\tfrac{1}{2}\,(\rho - \widetilde{\rho}\,|\rho - \widetilde{\rho})}_{\Delta J\ \text{(four center)}} = \tilde{J} + \Delta J \geq \tilde{J}$$

- We use $\tilde{J}$ for the evaluation of the energy, avoiding four-center integrals

$$\tilde{E} = T_{\rm s}[\rho] + \int \rho(\mathbf{r}) v(\mathbf{r})\, \mathrm{d}\mathbf{r} + \tilde{J}[\rho, \tilde{\rho}] + E_{\rm xc}[\rho]$$

- We determine c_α variationally, noting that $\tilde{E}$ is concave in c_α:

$$\frac{\partial \tilde{E}}{\partial c_\alpha} = \frac{\partial \tilde{J}}{\partial c_\alpha} = 0 \;\Rightarrow\; (\alpha\,|\widetilde{\rho}) = (\alpha\,|\rho) \;\Rightarrow\; \sum_\beta (\alpha\,|\beta)\, c_\beta = (\alpha\,|\rho)$$

Density fitting: sample calculations

- benzene with no symmetry used; exact energy: $-230.208 E_{\mathrm{h}}$

	cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pV5Z
exact density late	27	374	3823	na
exact density early	10	90	687	5320
fitted density	1	7	34	166
exact energy	−230.09671	−230.17927	−230.19463	na
density-fitted energy	−230.09733	−230.17933	−230.19466	−230.20116

- Clearly, very large gains can be achieved with density fitting
- Formal scaling is cubical
  - for large systems, screening yields quadratic scaling for integral evaluation
  - the cubic cost of solving linear equations remains for large systems
- There is a long history of density fitting:
  - Whitten (1973), Baerends *et al.* (1973), Dunlap *et al.* (1979), Vahtras *et al.* (1993)

Two-electron integral scaling

- Two-electron *ssss* integrals:

$$g_{abcd} = \mathrm{erf}\left(\sqrt{\alpha} R_{PQ}\right) \frac{S_{ab} S_{cd}}{R_{PQ}}, \quad S_{ab} = \left(\frac{\pi}{a+b}\right)^{3/2} \exp\left(-\frac{ab}{a+b} R_{AB}^2\right)$$

  - the total number of integrals scales quartically with system size
  - the number of significant integrals scales quadratically

$$S_{ab}, S_{cd} \to 0 \text{ (rapidly)}, \quad R_{PQ}^{-1} \to 0 \text{ (very slowly)}$$

- Decompose integral into classical and nonclassical parts:

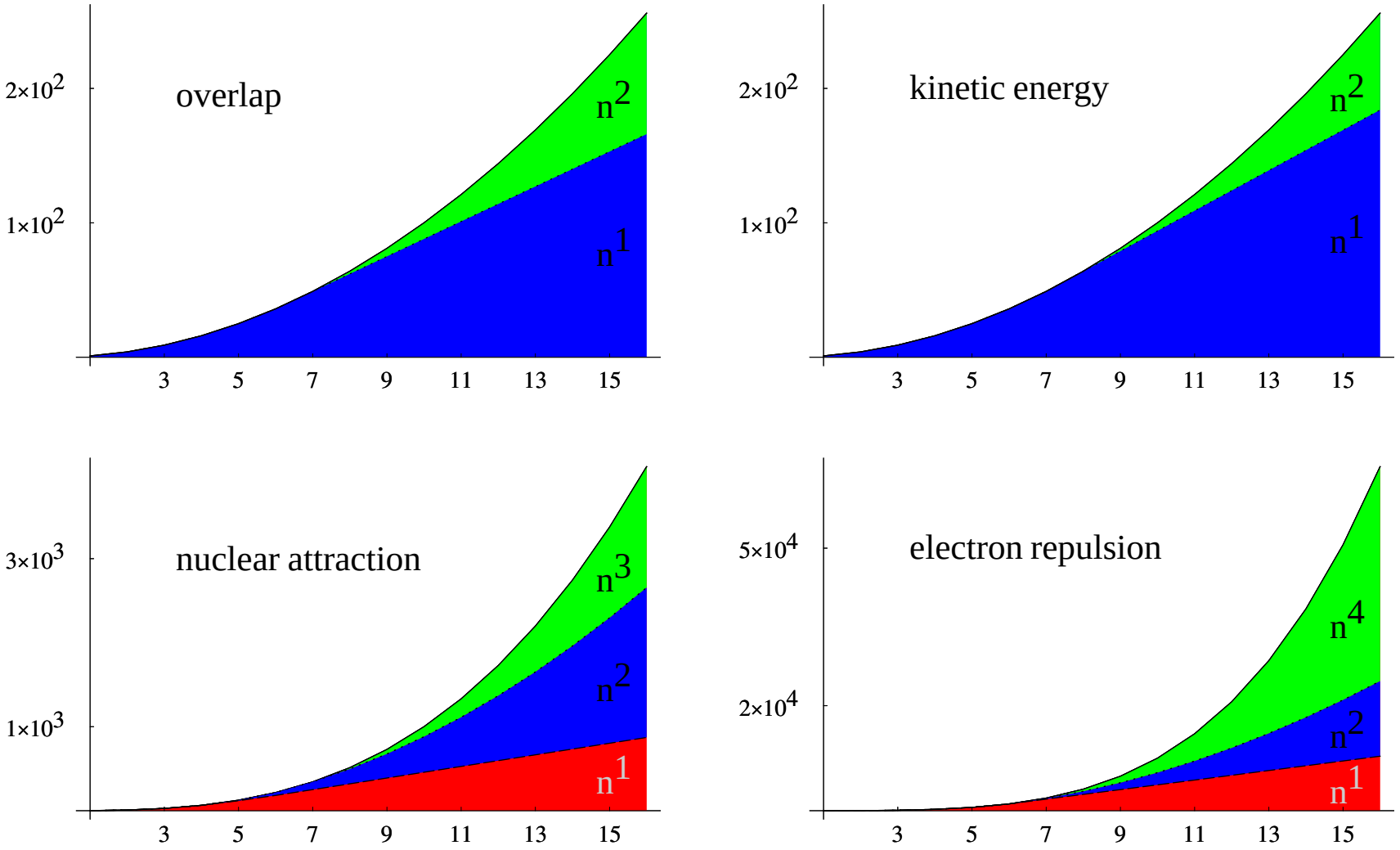
$$g_{abcd} = \underbrace{\frac{S_{ab} S_{cd}}{R_{PQ}}}_{\text{classical}} \underbrace{- \mathrm{erfc}\left(\sqrt{\alpha} R_{PQ}\right) \frac{S_{ab} S_{cd}}{R_{PQ}}}_{\text{nonclassical}}$$

  - quadratic scaling of classical part, can be treated by multipole methods
  - linear scaling of nonclassical part since $S_{ab} \to 0$, $S_{cd} \to 0$, erfc $\to 0$ rapidly

$$\mathrm{erfc}(\sqrt{\alpha} R_{PQ}) \leq \frac{\exp(-\alpha R_{PQ}^2)}{\sqrt{\pi \alpha} R_{PO}}$$

The scaling properties of molecular integrals

- linear system of up to 16 1s GTOs of unit exponent, separated by 1a_0



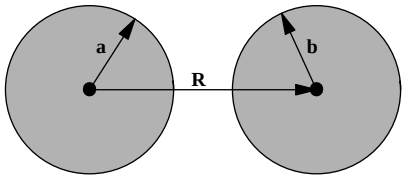
Multipole expansion

- Coulomb integral

$$g_{abcd} = \iint \frac{\Omega_{ab}(1)\Omega_{cd}(2)}{r_{12}} \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2$$

- If Ω_{ab} and Ω_{cd} do not overlap significantly, we can use multipole expansion:

$$g_{abcd} = \sum_{lmjk} q^{ab}_{lm} T_{lm,jk} q^{cd}_{jk}$$



  - multipole vector and interaction matrix

$$q^{ab}_{lm} = \int \Omega_{ab}(\mathbf{r}) R_{lm}(\mathbf{r}_P) \mathrm{d}\mathbf{r}$$

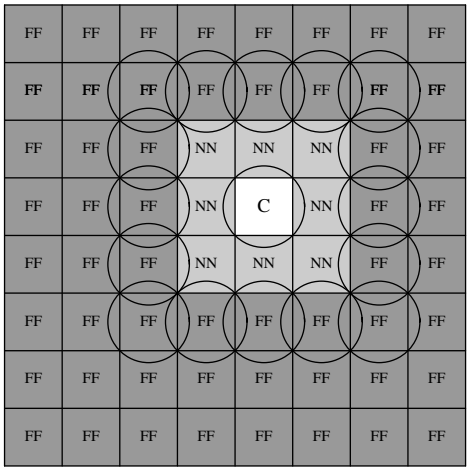
$$T_{lm,jk} = (-\mathrm{i})^j I^*_{l+j,m+k}(\mathbf{R}_{QP})$$

  - regular and irregular solid harmonics

$$R_{lm} \propto r^l Y_{lm}, \quad I_{lm} \propto r^{-l-1} Y_{lm}$$

- Can be applied to individual contracted integrals but also to more complicated electron distributions

Multipole method



- Divide molecule into boxes
- Create multipole expansions in each box
- Evaluate interactions between nonneighbouring boxes by multipole expansion

- Multipole expansions are expansions in

$$\frac{r}{R} = \frac{\text{size of boxes}}{\text{their separation}}$$

  – we can use larger boxes for more distant interactions at no loss of accuracy

- Taking advantage of this, we can calculate classical Coulomb interactions at costs

$$N \log N \quad \leftarrow \quad \text{tree code}$$
$$N \quad \leftarrow \quad \text{fast multipole method (FMM)}$$

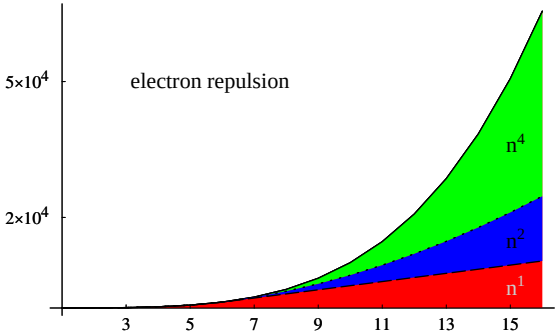
  rather than N^2, where N is the number of boxes

The fast multipole method (FMM) for Coulomb interactions

- Let us consider the evaluation of the Coulomb contribution to the Kohn–Sham matrix:

$$F_{ab} = \sum_{cd} (ab|cd)\, D_{cd} \quad \leftarrow \begin{cases} n \text{ significant orbital products } ab \text{ and } cd \\ n^2 \text{ significant integrals } (ab|cd) \end{cases}$$

- For fast evaluation, the n^2 significant integrals are traditionally divided into two classes:
  - the nonclassical integrals between overlapping charge distributions; their number scales as n
  - the classical integrals between nonoverlapping charge distributions; their number scales as n^2



- The nonclassical contribution to F_{ab} is calculated by standard techniques in linear time
- The contribution from the n^2 classical integrals is evaluated by multipole methods
  - linear scaling is achieved by the fast multipole method (FMM) (Head-Gordon, Scuseria)
- The FMM was developed for point particles, partitioned into a hierarchy of boxes
  - finite Gaussian distributions require a generalization (White & Head-Gordon, 1994)
  - in the continuous FMM (CFMM), Gaussians are divided into branches based on size
  - branches add an extra level of complexity to the FMM code

The branch-free FMM (BFMM)

- In BFMM, we avoid branches altogether by partitioning the contributions differently:
  - each integral is decomposed into a point-charge term and a size-correction term:

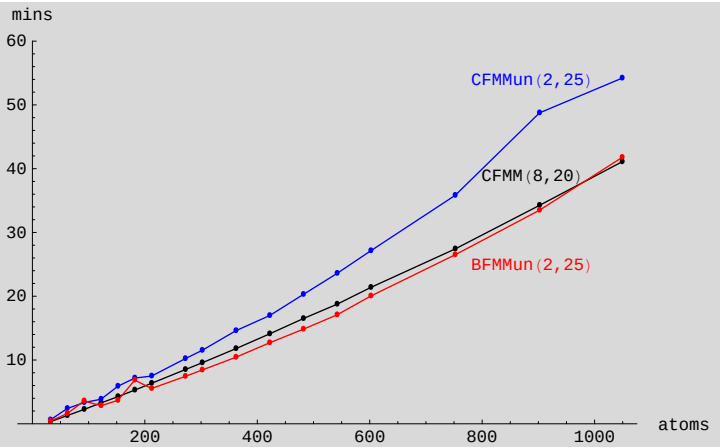
$$\left(\frac{pq}{\pi^2}\right)^{3/2} \iint \frac{\exp(-pr_{1\mathrm{P}}^2)\exp(-qr_{2\mathrm{Q}}^2)}{r_{12}}\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 = R_{\mathrm{PQ}}^{-1} + J_{pq}^{\mathrm{sc}}$$

  - the n^2 point-charge contributions R_{PQ}^{-1} are treated by FMM, without branches
  - the n local size corrections J_{pq}^{sc} are evaluated explicitly, in linear time:

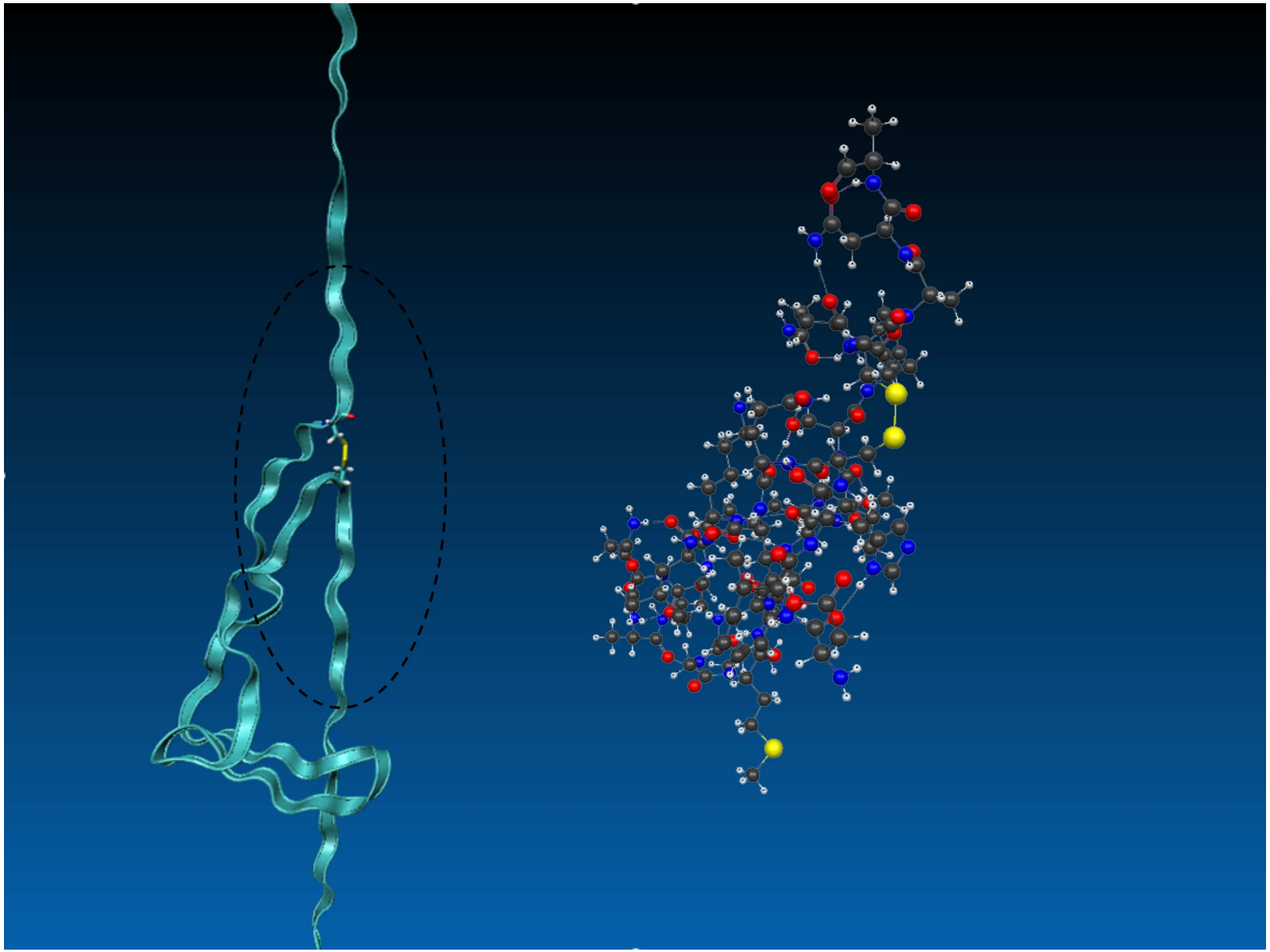
$$J_{pq}^{\mathrm{sc}} = -\frac{1}{\sqrt{\pi\alpha}}\exp\left(-\alpha R_{PQ}^2\right)\left[R_{PQ}^{-2} + \mathcal{O}\left(R_{PQ}^{-4}\right)\right], \quad \alpha = \frac{pq}{p+q}$$

- Observe: Coulomb integrals with $R_{PQ} \approx 0$ cannot be decomposed in this manner
  - such neighboor interactions are in any case not treated by multipoles in FMM

- LDA/3-21G polyethylenes
  - CFMM:
    contracted basis, $10^{-7}E_{\mathrm{h}}$
  - CFMMun and BFMMun:
    internally decontracted basis, $10^{-8}E_{\mathrm{h}}$



Titin molecule



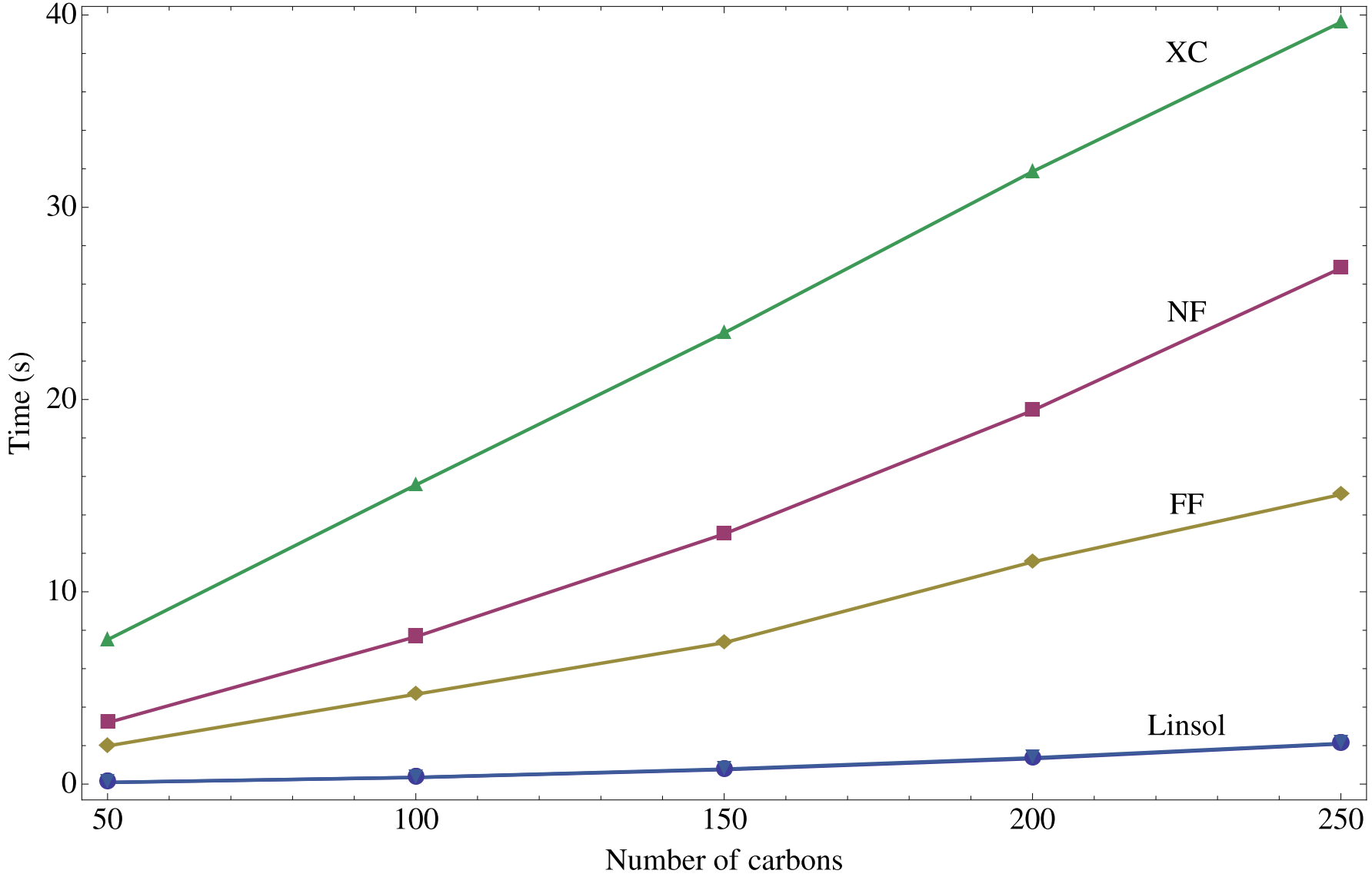
Timings (seconds) for Kohn–Sham matrix construction

- Development version of Dalton on a single Xeon 2.66 GHZ processor
- BP86 exchange–correlation functional
- First Kohn–Sham iteration, screening threshold 10^{-10}

system	# atoms	basis	# AOs	NF	FF	XC	LE	total
taxol	113	6-31G	660	19	19	27	0.1	65
		6-31G**	1123	35	31	54	0.1	120
valinomycin	169	6-31G	882	33	45	41	0.2	119
		6-31G**	1542	59	68	81	0.2	208
titin	392	6-31G	2221	99	171	109	1.3	380

- Total time for energy optimization of titin (9 iterations)
  - initialization: 812 s (XC grid 535 s)
  - NF contributions: 352 s
  - FF contributions: 1363 s (maximum order 12)
  - XC contributions: 790 s (Gauss–Chebychev with Treutler–Ahlrichs mapping, grid 4)
  - RH/DIIS optimization: 279 s

Timings for BP86/6-31G** Kohn–Sham build in linear polyene chains



Differentiated molecular integrals for gradients and beyond

- Since Boys (1950), we have used Gaussian orbitals of the general Cartesian form

$$G_{ijk}\left(\mathbf{r}, a, \mathbf{A}\right) = x_A^i y_A^j z_A^k \exp\left(-a r_A^2\right)$$

  - when we differentiate these Gaussians, we obtain linear combinations:

$$\frac{\partial G_{ijk}}{\partial A_x} = 2aG_{i+1,j,k} - iG_{i-1,j,k}$$

  - higher differentiations generate more terms, making the integration awkward

- An alternative approach would be to use Hermite Gaussians instead

$$H_{ijk}\left(\mathbf{r}, a, \mathbf{A}\right) = \frac{\partial^{i+j+k} \exp\left(-a r_A^2\right)}{(2a)^{i+j+k} \partial A_x^i \partial A_y^j \partial A_z^k}$$

  - differentiation now becomes much simpler

$$\frac{\partial H_{ijk}}{\partial A_x} = 2aH_{i+1,j,k}$$

  - only one term is generated, to any order in differentiation

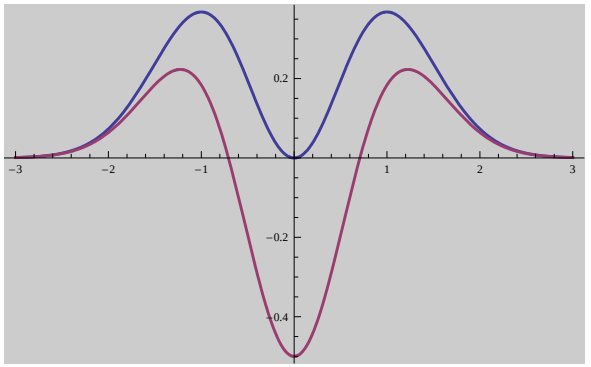
- Hermite Gaussians were introduced by Zivkovic and Maksic (1968)
  - used as intermediates by McMurchie and Davidson (1978)

Solid-harmonic Gaussians

- The Cartesian and Hermite Gaussians are different, with different radial forms
  - for example, the d_{x^2} functions are given by

$$G_{200} = x_A^2 \exp\left(-ar_A^2\right)$$

$$H_{200} = x_A^2 \exp\left(-ar_A^2\right) - \frac{1}{2a} \exp\left(-ar_A^2\right)$$



  - used individually in quantum-chemical calculations, they give different results
- However, Cartesian and Hermite Gaussians give the same solid-harmonic Gaussians

$$S_{lm}\left(\mathbf{r}, a, \mathbf{A}\right) = \sum_{i+j+k=l} S_{ijk}^{lm}\, G_{ijk}\left(\mathbf{r}, a, \mathbf{A}\right) \equiv \sum_{i+j+k=l} S_{ijk}^{lm}\, H_{ijk}\left(\mathbf{r}, a, \mathbf{A}\right)$$

  - for example, combining d_{x^2} and d_{y^2} to solid-harmonic form, we obtain

$$S_{x^2-y^2} = \frac{1}{2}\sqrt{3}\left(G_{200} - G_{020}\right) = \frac{1}{2}\sqrt{3}\left(H_{200} - H_{020}\right)$$

  - the same happens for all solid harmonics, to all orders
- In solid harmonics, we may therefore globally replace Cartesian by Hermite Gaussians
  - Reine, Tellgren and Helgaker, PCCP **9**, 4771 (2007)
  - Weniger, CCCC **70**, 1225 (2005) (spherical tensor gradient operators)

Advantages of Hermite Gaussians

- Consider the expansion of solid-harmonic Gaussians in Hermite Gaussians

$$S_{lm}\left(\mathbf{r}, a, \mathbf{A}\right) = \sum_{i+j+k=l} S_{ijk}^{lm} \, H_{ijk}\left(\mathbf{r}, a, \mathbf{A}\right)$$

- We now obtain derivatives simply by raising the Hermite quantum numbers:

$$\frac{\partial^{I+J+K} S_{lm}\left(\mathbf{r}, a, \mathbf{A}\right)}{\partial A_x^I \partial A_y^J \partial A_z^K} = (2a)^{I+J+K} \sum_{i+j+k=l} S_{ijk}^{lm} \, H_{i+I,j+J,k+K}\left(\mathbf{r}, a, \mathbf{A}\right)$$

  - the same number of terms $(i+j+k=l)$ contribute, to all orders

- A unified scheme for differentiated and undifferentiated Gaussians
  - simplifies development of derivative codes, in particular to high orders
  - useful for gradients and beyond, and for kinetically balanced basis sets

- The integration over Hermite Gaussians is no more difficult than that over Cartesians
  - all integrals may be reduced to the differentiation of s integrals
  - simplifies use of translational and rotational symmetries
  - simplifies the evaluation over two- and three-center integrals (in density fitting)

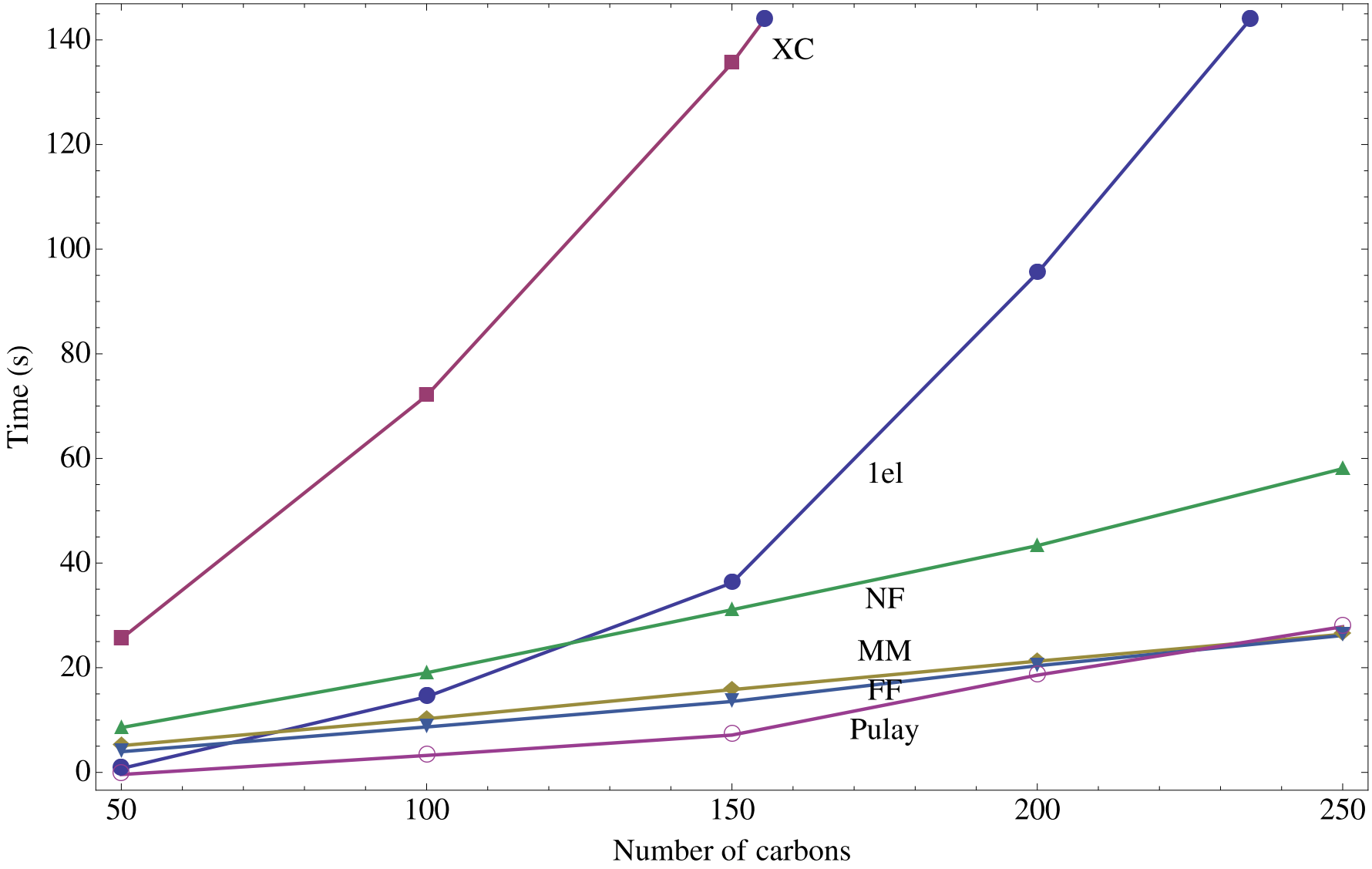
Timings (seconds) for force evaluation

- Development version of Dalton on a single Xeon 2.66 GHZ processor
- BP86 exchange–correlation functional
- Corresponding timings for energies given below

system	# atoms	basis	# AOs	1el	NF	FF	XC	PF	total
taxol	113	6-31G	660	12	47	42	89	3	193
					19	19	27		65
		6-31G**	1123	43	90	84	153	6	376
					35	31	54		120
valinomycin	169	6-31G	882	29	73	78	193	7	380
					33	45	41		119
		6-31G**	1542	111	136	144	236	12	639
					59	68	81		208
titin	392	6-31G	2221	242	206	248	420	48	1164
					99	171	109		380

- XC evaluation dominates in all cases, inefficient one-electron evaluation
- force evaluation only 2–3 times more expensive than Kohn–Sham matrix evaluation

Timings for BP86/6-31G** Kohn–Sham build in linear polyene chains



Timings (seconds) for energy and forces

- Development version of Dalton on a single Xeon 2.66 Ghz processor
- BP86 exchange–correlation functional
- Energy and force timings are averages over all geometry steps

system	# at.	basis	# AOs	# SCF	energy	force	# geo	opt.
taxol	113	6-31G	660	9.6	570	192	47	3
		6-31G**	1123	9.2	962	376	44	3
valinomycin	169	6-31G	882	7.7	879	326	38	9
		6-31G**	1542	8.6	1603	638	32	9
titin	392	6-31G	2221	9.3	3594	1164	30	106

- Force evaluation takes about one third of the time for energy optimization
- Geometry optimizer scales cubically:
  - generalized inverse of the Wilson B matrix, diagonalization of updated Hessian
- Convergence criteria with $\epsilon = 10^{-4}$ (internal coordinates):
  - rms grad (ϵ), max grad (5ϵ), rms step (3ϵ), max step (15ϵ)

Review and preview

- We have discussed energy optimization and integral evaluation
  - the augmented RH (ARH) method, merging the RH and DIIS steps
  - efficient linear-scaling evaluation of integrals and interactions
  - force evaluation from Hermite integrals
- Unlike energy optimization, response theory poses relatively few problems
  - iterative subspace algorithms used from the beginning
  - convergence of the linear equations is easy (compared with SCF iterations)
  - for large systems, most of the time is spent doing (nonsparse) linear algebra
- Large-system applications:
  - excitation energies and polarizabilities

Response theory

- We consider a system described by the one-electron density matrix

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{X})\mathbf{D}\exp(\mathbf{X}) \qquad (\mathbf{X} = \mathbf{0} \text{ for unperturbed system})$$

  and define the Hessian and metric operators in terms of their transformations

$$\mathbf{E}^{[2]}(\mathbf{X}) = (\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}})\mathbf{X} + \mathbf{X}(\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}) + \mathbf{G}^{\mathrm{vo}}([\mathbf{D}, \mathbf{X}]) - \mathbf{G}^{\mathrm{ov}}([\mathbf{D}, \mathbf{X}])$$

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

- When perturbed by $\mathbf{V}_\omega$ of frequency ω, the system responds to first order as

$$\mathbf{E}^{[2]}(\mathbf{X}_\omega) - \omega\mathbf{S}^{[2]}(\mathbf{X}_\omega) = [\mathbf{D}, \mathbf{V}_\omega] \quad \leftarrow \text{ linear response matrix equation}$$

  - perturbed density matrix: $\mathbf{D}_\omega = [\mathbf{D}, \mathbf{X}_\omega]$
  - perturbed expectation values: $\langle\langle \hat{A}; \hat{V}_\omega \rangle\rangle_\omega = \mathrm{Tr}\,\mathbf{A}\,[\mathbf{D}, \mathbf{X}_\omega]$

- In the absence of a perturbation $\mathbf{V}_\omega = \mathbf{0}$, we obtain an eigenvalue equation

$$\mathbf{E}^{[2]}(\mathbf{X}_n) = \omega_n\mathbf{S}^{[2]}(\mathbf{X}_n) \quad \leftarrow \text{RPA matrix eigenvalue equation}$$

  - transition density matrix: $\mathbf{D}_{0n} = [\mathbf{D}, \mathbf{X}_n]$
  - transition moments: $\langle 0|\hat{A}|n\rangle = \mathrm{Tr}\,\mathbf{A}\,[\mathbf{D}, \mathbf{X}_n]$

- Coriani *et al.*, J. Chem. Phys. **126**, 154108 (2007)
  - Ochsenfeld, Head-Gordon, Weber, Niklasson, and Challacombe (static properties)

Solution of the response equations

- For the solution, it is useful to consider two response equations

$$\mathbf{E}^{(2)}(\mathbf{X}) - \omega\mathbf{S}^{(2)}(\mathbf{X}) = \mathbf{B} \leftarrow \text{full response equation}$$

$$\mathbf{E}_{\mathrm{F}}^{(2)}(\mathbf{X}) - \omega\mathbf{S}^{(2)}(\mathbf{X}) = \mathbf{B} \leftarrow \text{simplified response equation}$$

  where the simplified Hessian is a good but cheap approximation to the full Hessian:

$$\mathbf{E}_{\mathrm{F}}^{(2)}(\mathbf{X}) = (\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}})\mathbf{X} + \mathbf{X}(\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}) \leftarrow \text{no two-electron part}$$

  - its eigenvalues are the orbital energy differences $\epsilon_a - \epsilon_i$

- We wish solve the full response equations iteratively

$$\mathbf{R}_i = \mathbf{E}^{(2)}(\mathbf{X}_i) - \omega\mathbf{S}^{(2)}(\mathbf{X}_i) - \mathbf{B} \quad \leftarrow \quad \text{residual}$$

  - new trial vectors are generated from the residual until it is sufficiently small

- For fast convergence, we precondition with the simplified response equation

$$\mathbf{E}_{\mathrm{F}}^{(2)}(\widetilde{\mathbf{R}}_i) - \omega\mathbf{S}^{(2)}(\widetilde{\mathbf{R}}_i) = \mathbf{R}_i$$

  - in the MO basis, $\mathbf{E}_{\mathrm{F}}^{(2)}$ is diagonal (orbital-energy differences) and solution is trivial
  - in the OAO basis, $\mathbf{E}_{\mathrm{F}}^{(2)}$ is nondiagonal and 5–20 iterations are required for solution

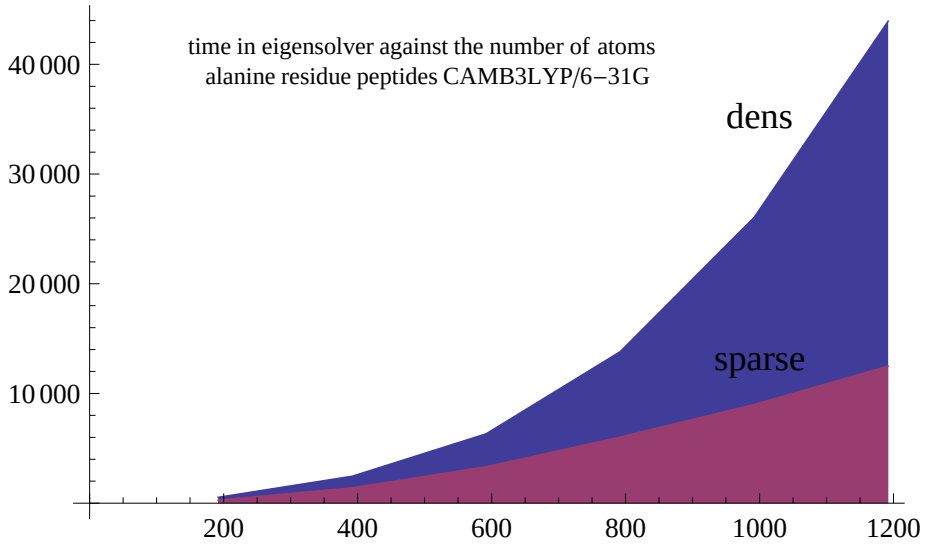
Excitation energies

- The calculation of excitation energies (RPA) more difficult than that of polarizabilities
  - the eigenmatrices of the simplified Hessian provide good starting guesses:

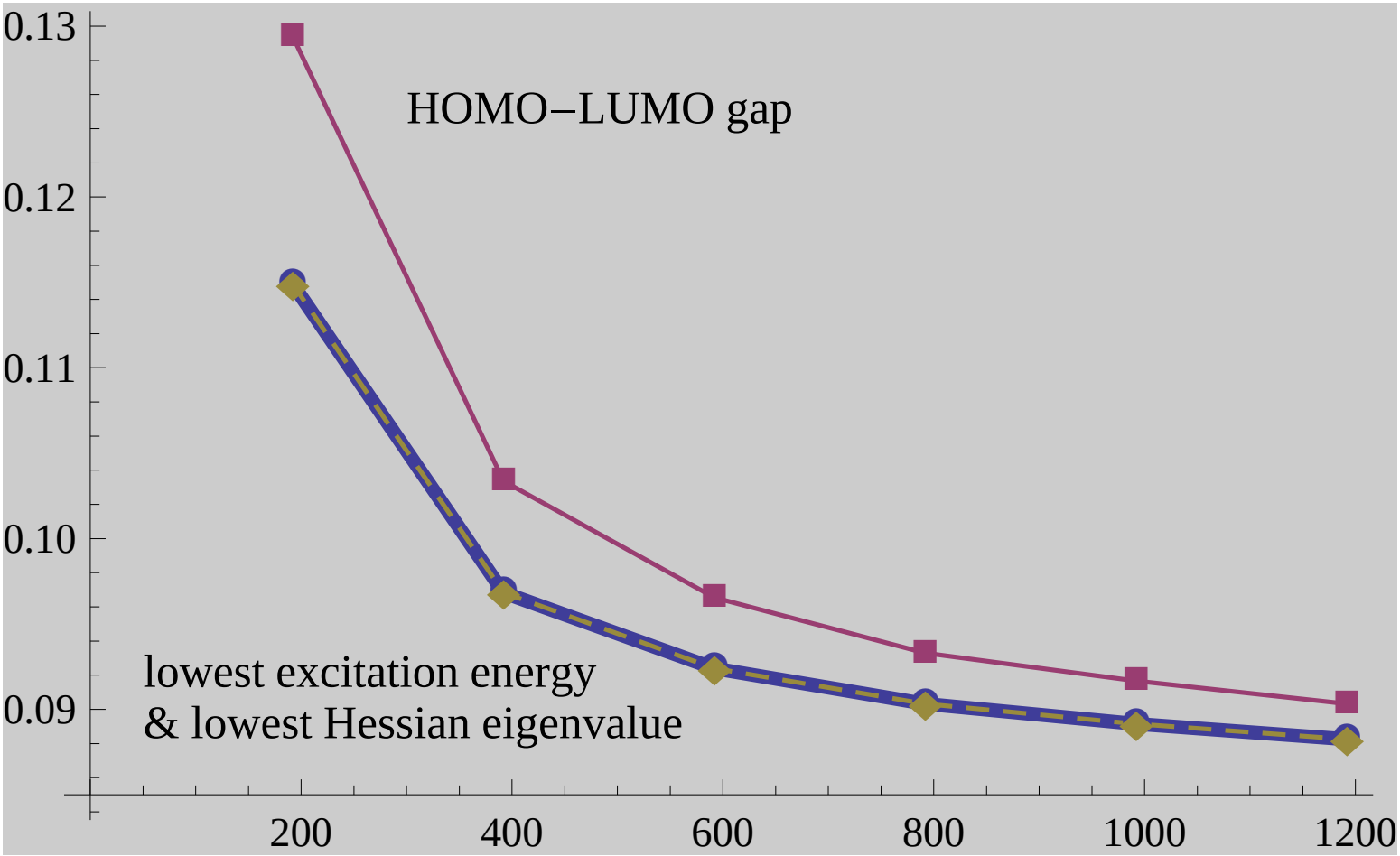
$$\mathbf{E}_{\mathrm{F}}^{(2)}(\mathbf{X}_{ia}) = (\epsilon_a - \epsilon_i)\mathbf{S}^{(2)}(\mathbf{X}_{ia}) \quad \text{where} \quad \mathbf{X}_{ia} = \mathbf{C}_i\mathbf{C}_a^{\mathrm{T}}, \quad \begin{cases} \mathbf{F}^{\mathrm{oo}}\mathbf{C}_i = \epsilon_i\mathbf{C}_i \\ \mathbf{F}^{\mathrm{vv}}\mathbf{C}_a = \epsilon_a\mathbf{C}_a \end{cases}$$

  - slow convergence of preconditioning equations (subspace problem nearly singular)

- CAMB3LYP/6-31G alanine residues
- cubic complexity with dense-matrix algebra
- linear complexity with sparse-matrix algebra
- preconditioning part dominates
- Fock/KS matrix construction dominates (not shown here)



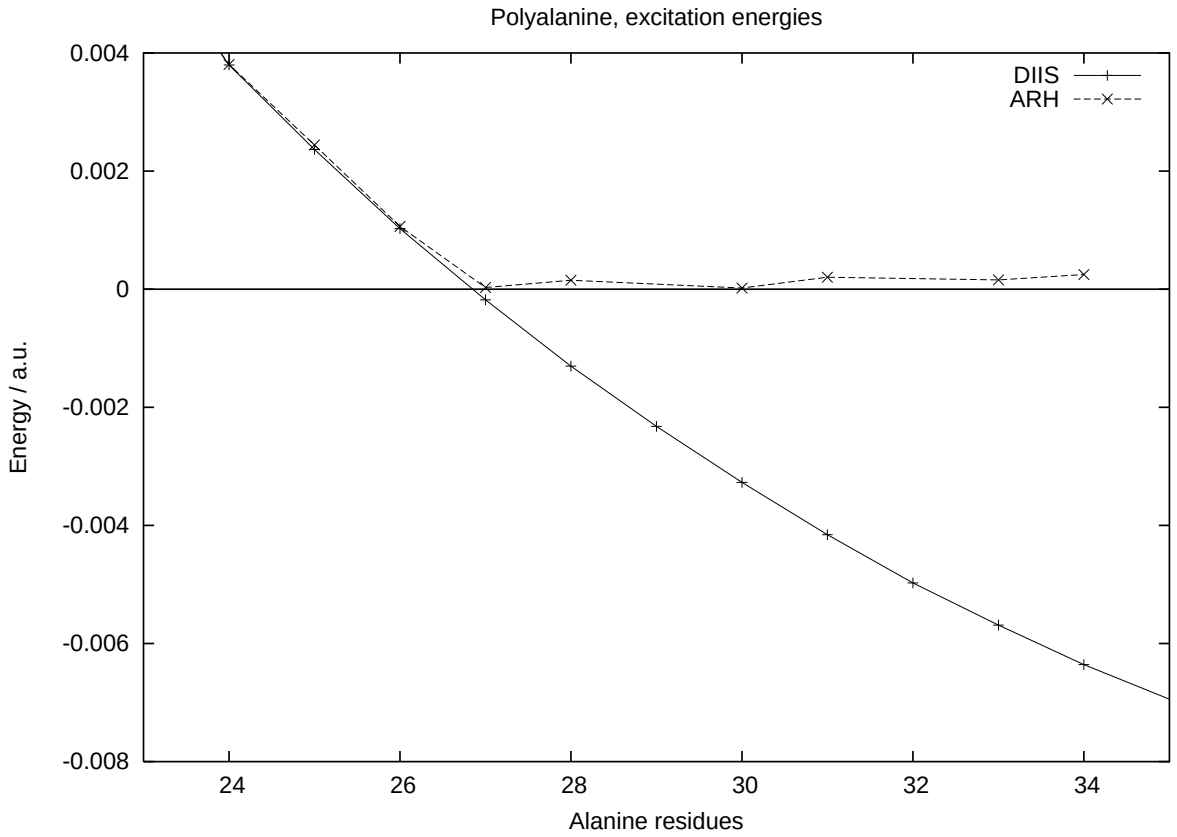
CAMB3LYP/6-31G excitation energies of alanine residue peptides



- lowest excitation energy almost identical with lowest Hessian eigenvalue
- CAMB3LYP has 19% short-range and 65% long-range exact exchange

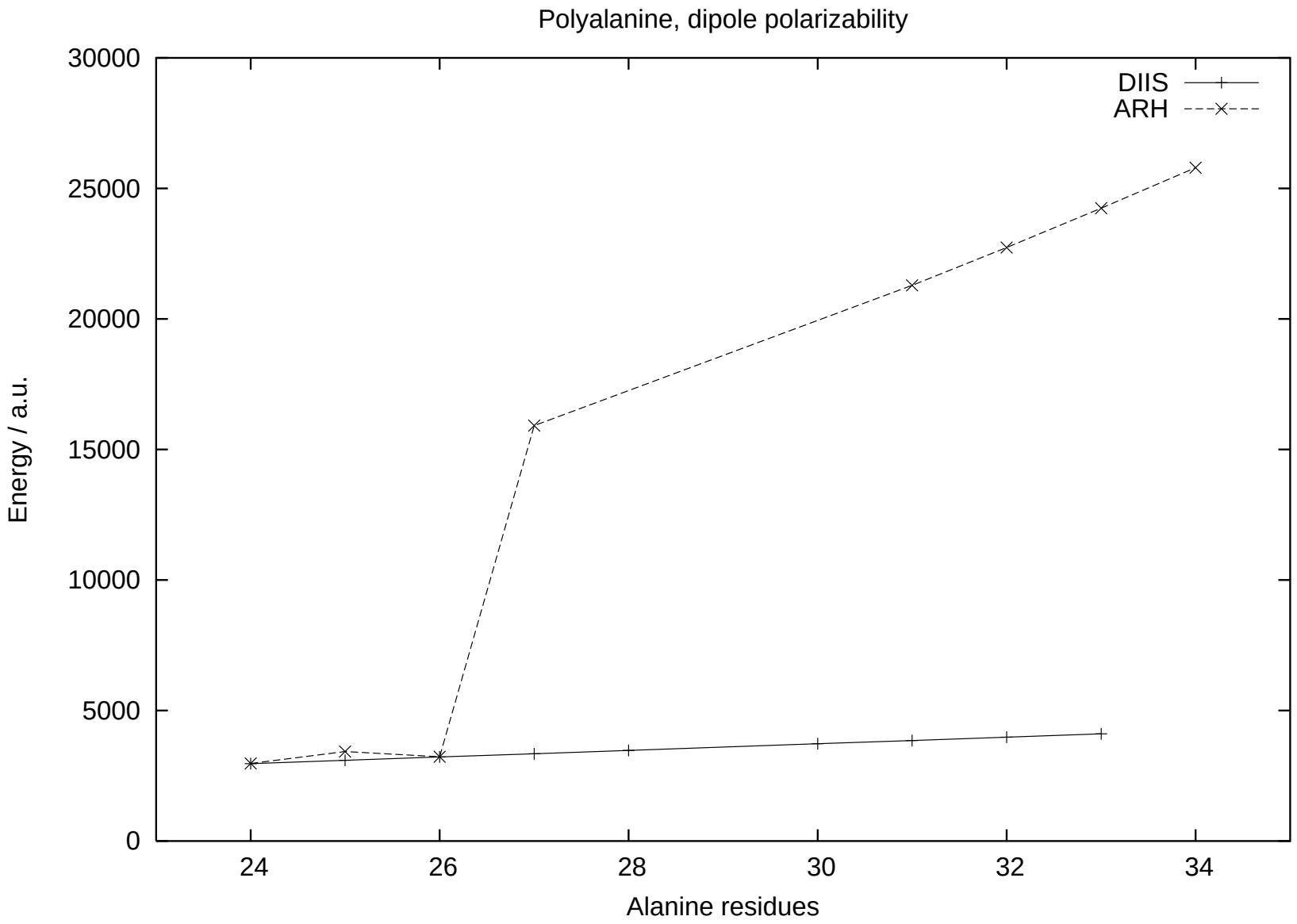
B3LYP/6-31G excitation energies of alanine residue peptides

- B3LYP (with 20% exact exchange) behaves differently, in an unphysical manner
  - beyond 27 alanine residues, the lowest excitation energy becomes very small
  - RH/DIIS converges to an excited state (first-order saddle point)



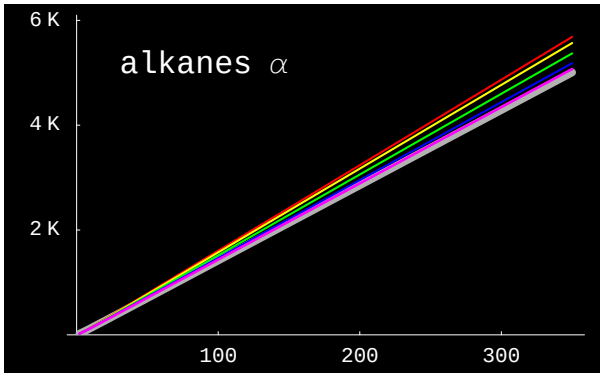
- for LDA and GGA functionals, the crossing occurs even earlier
  - local DFT functionals unable to account for long-range exchange in large molecules

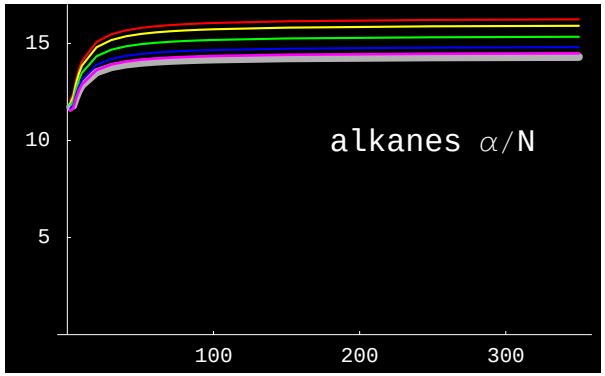
B3LYP/6-31G polyalanine polarizabilities

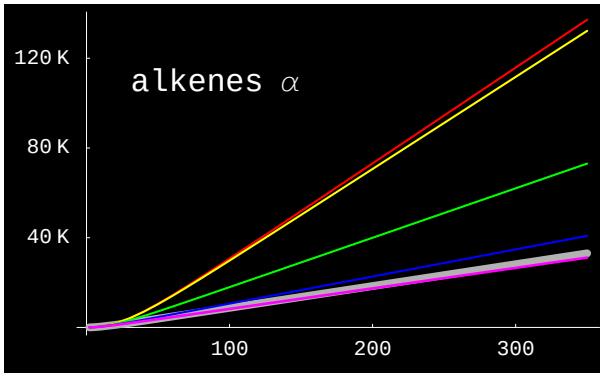


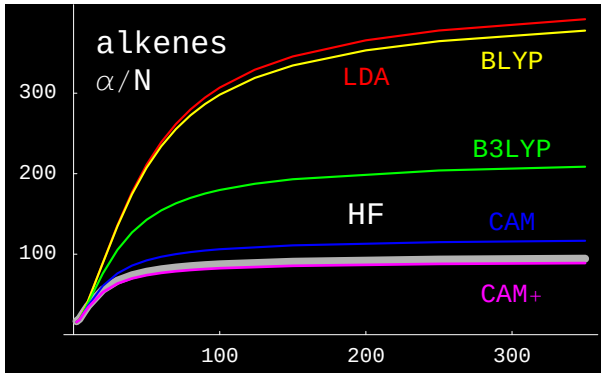
Polarizabilities of linear alkanes and alkenes

- 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





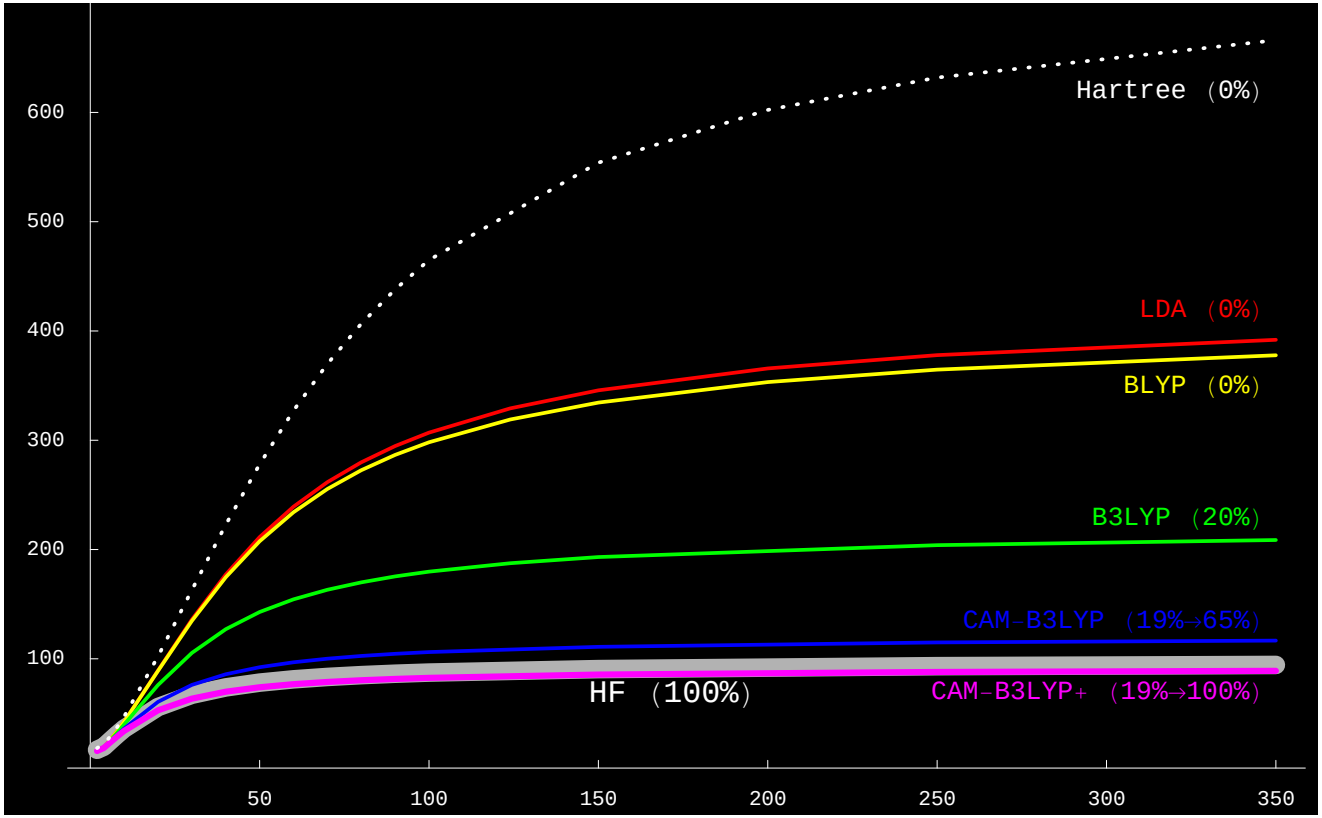




- The alkenes are about an order of magnitude more polarizable than the alkanes
  - all models agree on alkanes (α/N-limit: HF 14.4; LDA 16.3)
  - widely different results for alkenes (α/N-limit: HF 97; LDA 427)

The importance of exact exchange for longitudinal polarizabilities

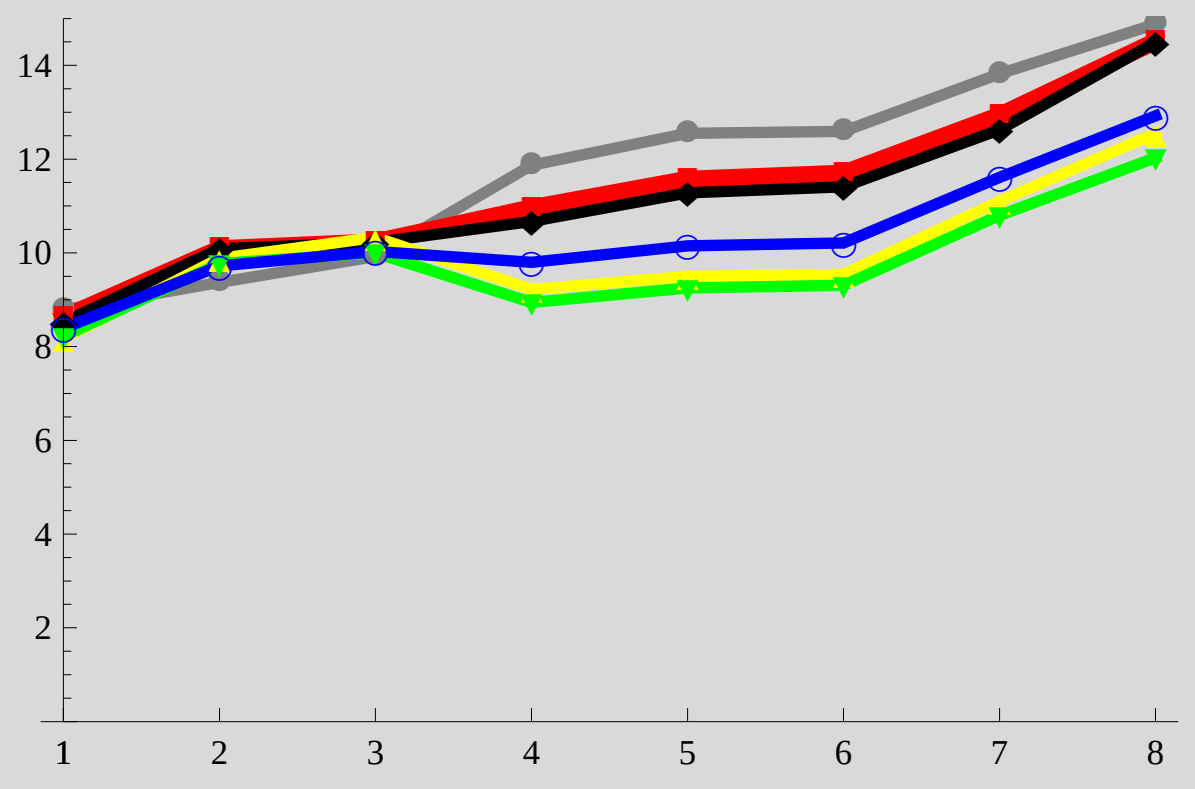
- Without a good description of long-range exchange, the systems become too polarizable



  - 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

Excitation energies and DFT

- We have observed the following typical behaviour of DFT:
  - overestimation of polarizabilities
  - underestimation of excitation energies
- Comparison of excitation energies with coupled-cluster theory for CO



- HF (grey), CCSD (red), CC3 (black) LDA (yellow), BLYP (green), B3LYP (blue)

Deficiencies of exchange–correlation functionals and potentials

- Standard pure (continuum) DFT functionals suffer from some known deficiencies
  - an incorrect asymptotic form of the exchange–correlation potential

$$\lim_{r\to\infty} v_{\rm xc}(\mathbf{r}) = \lim_{r\to\infty} \frac{\delta E_{\rm xc}[\rho]}{\delta\rho(\mathbf{r})} \neq -\frac{1}{r}$$

  - no integer discontinuity of the exchange–correlation potential

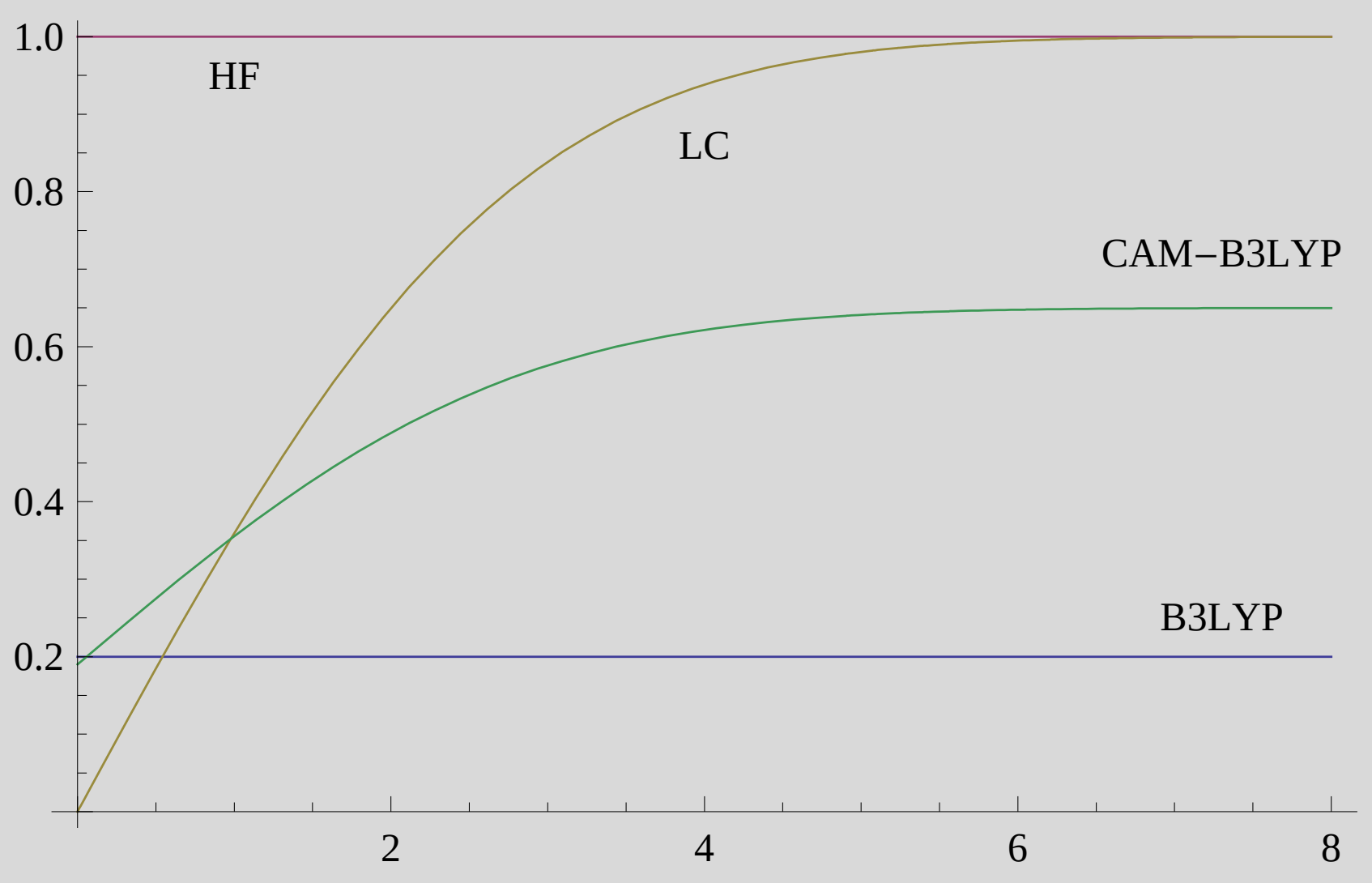
$$v^{+}_{\rm xc}(\mathbf{r}) \neq v^{-}_{\rm xc}(\mathbf{r}) + \Delta_{\rm xc}, \quad \Delta_{\rm xc} > 0$$

  - a nonvanishing self-interaction energy

$$J[\rho] + E_{\rm xc}[\rho] \neq 0 \quad \text{for } 0 \leq \int\rho(\mathbf{r})\mathrm{d}\mathbf{r} \leq 1$$

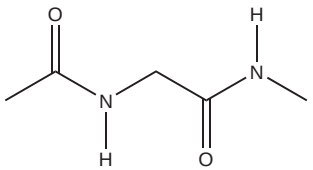
- These deficiencies give rise to problems with many properties
  - too low Rydberg excitations, charge-transfer (CT) and core excitations
  - too large polarizabilities
- These problems are alleviated by the introduction of exact exchange
  - this can be done in several different ways

Proportion of exact exchange in different functionals

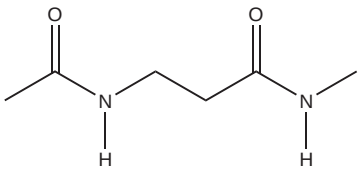


Assessment of calculated excitation energies

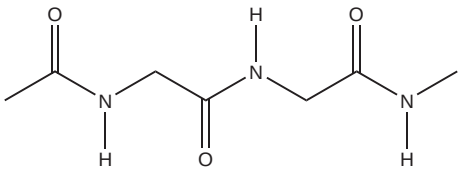
- We have made an extensive assessment of DFT excitation energies
  - Peach *et al.* JCP **128**, 044118 (2008)
  - 59 singlet excitations of 18 theoretically challenging main-group molecules



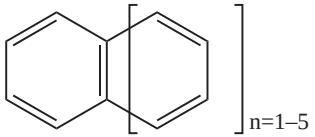
dipeptide



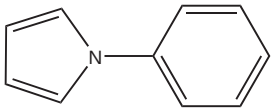
β-dipeptide



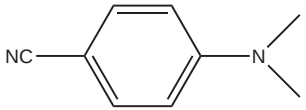
tripeptide



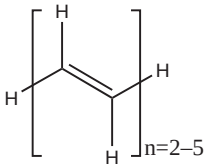
acenes (n=1–5)



N-phenylpyrrole



DMABN*



PA oligomers (n=2–5)

N_2

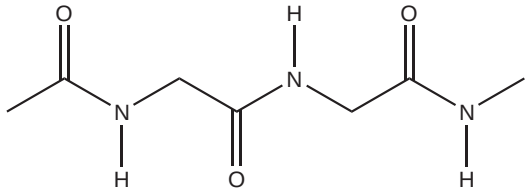
CO

H_2CO

HCl

*DMABS = 4-(N,N-dimethylamino) benzonitrile

Tripeptide excitations



excitation	type	PBE	B3LYP	CAM	exp.
$n_2 \rightarrow \pi_2^*$	local	5.58	5.74	5.92	5.61
$n_1 \rightarrow \pi_1^*$	local	5.36	5.57	5.72	5.74
$n_3 \rightarrow \pi_3^*$	local	5.74	5.88	6.00	5.91
$\pi_1 \rightarrow \pi_2^*$	CT	5.18	6.27	6.98	7.01
$\pi_2 \rightarrow \pi_3^*$	CT	5.51	6.60	7.68	7.39
$n_1 \rightarrow \pi_2^*$	CT	4.61	6.33	7.78	8.12
$n_2 \rightarrow \pi_3^*$	CT	5.16	6.83	8.25	8.33
$\pi_1 \rightarrow \pi_3^*$	CT	4.76	6.06	8.51	8.74
$n_1 \rightarrow \pi_3^*$	CT	4.26	6.12	8.67	9.30

Statistics for 59 excitation energies

		PBE	B3LYP	CAM
local	$\overline{\Delta}$	−0.31	−0.15	0.02
	$\lvert\overline{\Delta}\rvert$	0.33	0.22	0.20
	$\overline{\Delta}_{\mathrm{std}}$	0.27	0.26	0.27
	$\max\lvert\Delta\rvert$	0.83	0.61	0.71
Rydberg	$\overline{\Delta}$	−1.84	−1.11	−0.50
	$\lvert\overline{\Delta}\rvert$	1.84	1.11	0.50
	$\overline{\Delta}_{\mathrm{std}}$	0.30	0.23	0.18
	$\max\lvert\Delta\rvert$	2.24	1.43	0.80
CT	$\overline{\Delta}$	−2.60	−1.35	−0.18
	$\lvert\overline{\Delta}\rvert$	2.60	1.36	0.27
	$\overline{\Delta}_{\mathrm{std}}$	1.37	0.86	0.31
	$\max\lvert\Delta\rvert$	5.04	3.18	0.75

A diagnostic for excitation energies

- At present, DFT functionals for excitation energies are unsatisfactory
  - valence excitations are fine (off by a few tenths of an eV)
  - Rydberg, CT and core excitations are not poor
- When do we need to exercise caution? When can we trust our results?
  - problems arise when excitations occur between orbitals with no spatial overlap
- Excitations represented by rotations κ_{ai} between all occupied and virtual orbitals
  - to measure spatial overlap between s and i, we introduce

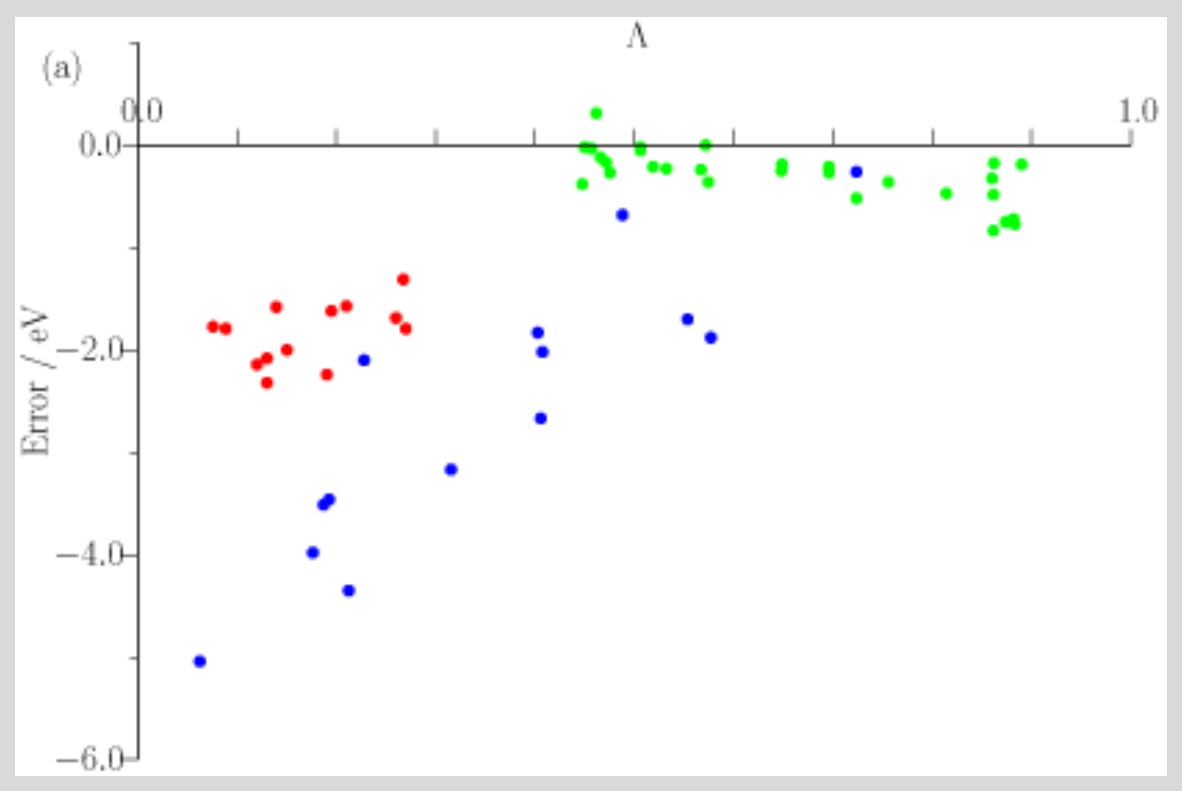
$$O_{ia} = \int |\phi_i(\mathbf{r})|\,|\phi_a(\mathbf{r})|\;\mathrm{d}\mathbf{r}$$

  - without moduli all overlaps would be trivially zero
- We now measure spatial overall using the quantity

$$\Lambda = \frac{\sum_{ai} \kappa_{ai}^2 O_{ai}}{\sum_{ai} \kappa_{ai}^2}, \quad 0 \leq \Lambda \leq 1$$

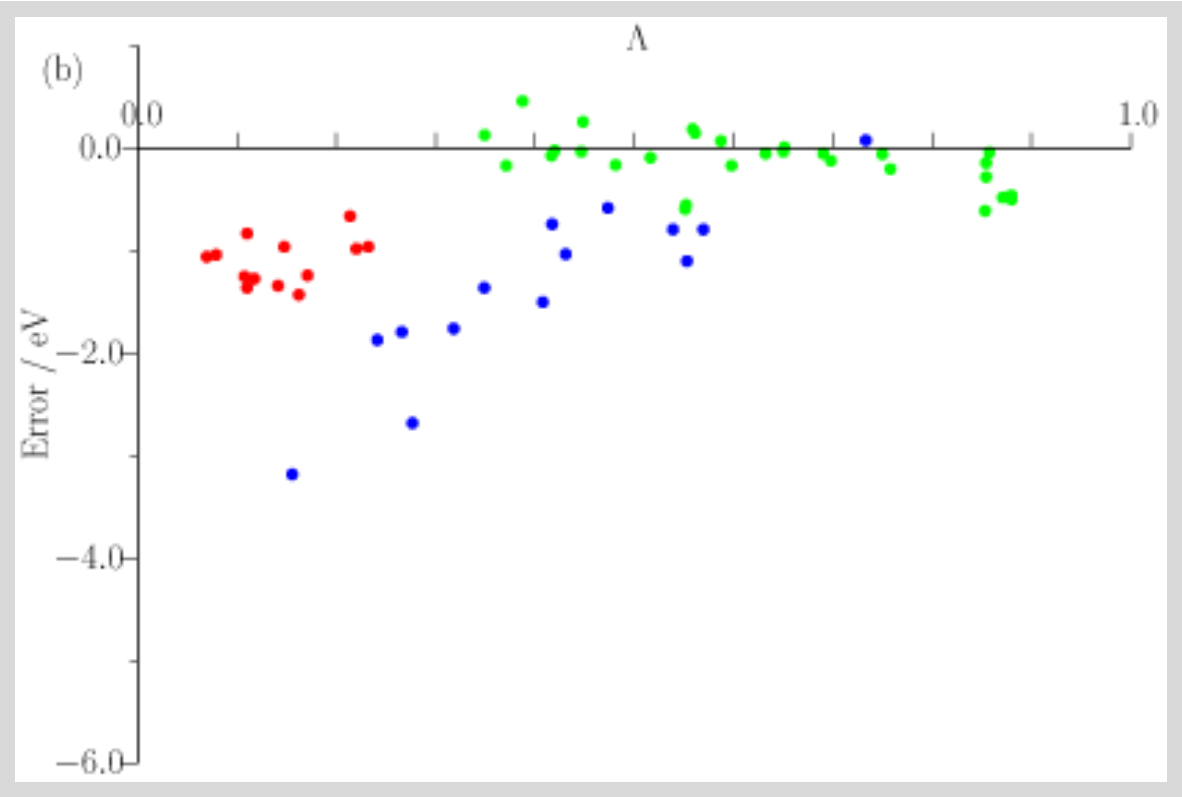
  - a small value of Λ signifies a long-range excitation
  - a large value of Λ signifies a short-range excitation
- Peach, Benfield, Helgaker, and Tozer, JCP **128**, 044118 (2008)
  - most cited among all 7656 JCP papers published since 2007

Excitation energies against Λ for PBE



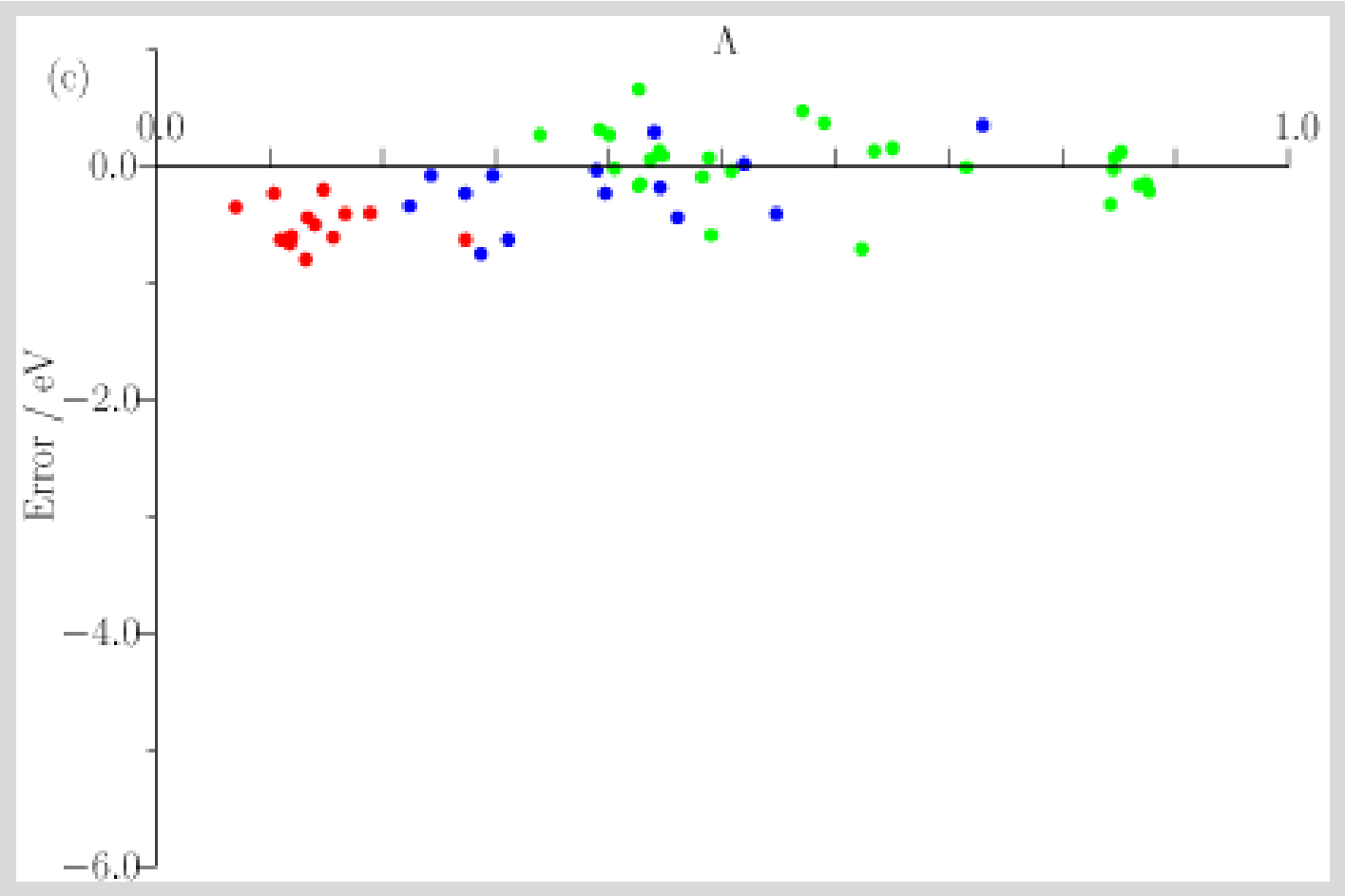
- With this GGA functional, performance for $\Lambda < 0.6$ is erratic
  - local excitations ($\Lambda > 0.4$) are well reproduced, within a few tenths of an eV
  - Rydberg excitations ($\Lambda < 0.3$) are systematically underestimated by a few eVs
  - CT excitations cover a surprisingly large Λ range (DMABN)

Excitation energies against Λ for B3LYP



- With the introduction of exact exchange, performance improves
  - local excitations are lifted and cover a slightly larger Λ range
  - errors in Rydberg excitations are reduced by a factor of two
  - likewise, an improvement in CT excitations is observed

Excitation energies against Λ for CAM-B3LYP



- Coulomb attenuation gives a uniform description of local and nonlocal excitations!
  - however, errors as large as 1 eV are still observed

Some conclusions

- Extensive assessment of PBE, B3LYP and CAM-B3LYP for excitation energies
- Errors in excitation energies correlated to spatial orbital overlap
- Best overall behaviour provided by the attenuated CAM-B3LYP functional
  - no correlation observed between errors and spatial overall
  - a good quality, balanced description
  - this functional is recommended for excitation-energy calculations
- The PBE functional exhibits a clear correlation between errors and spatial overlap
  - to a lesser extent, the B3LYP functional exhibits the same behaviour
- We propose the following diagnostic test:
  - a PBE excitation with $\Lambda < 0.4$ is likely to be in significant error
  - a B3LYP excitation with $\Lambda < 0.3$ is likely to be in significant error
- The term "charge transfer" is ambiguous
  - such excitations span a very large range of overlaps
  - small for peptide systems, large for DMABN

Summary

- We have considered some computational tasks in SCF theory
  - orthogonalization of the atomic orbitals
  - minimization of the SCF energy in an orthonormal basis
- Several approaches may be taken to the SCF energy minimization
  - full second-order Newton optimization
  - Roothaan–Hall (RH) Newton minimization (equivalent to diagonalization)
  - augmented RH (ARH) (reuses old Fock matrices)
- The ARH method is superior to RH/DIIS
  - requires one Fock/KS-matrix evaluation at each iteration
  - reuses efficiently information present in old Fock/KS matrices
  - converges by design to the ground state, smoothly and robustly
- For sufficiently sparse AO matrices, all tasks can be carried out in linear time
  - the Fock/KS matrix is reevaluated no more often than in MO theory
  - the remaining work is rich in matrix multiplication
  - well suited to parallelization