The calculation of molecular properties of large systems

Trygve Helgaker

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

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

Molecular Quantum Mechanics—Analytic Gradients and Beyond

An International Conference in Honour of Professor Peter Pulay

May 29–June 3, 2007

Margaret Island, Budapest

Hungary

Overview

- This talk consists of two separate parts
- In the first part, we consider linear-scaling SCF theory
  1. the generation of a sparse, orthonormalized atomic-orbital (OAO) basis
     - Jansik *et al.*, J. Chem. Phys. **126**, 124104 (2007)
  2. energy optimization for large systems
     - Salek *et al.*, J. Chem. Phys. **126**, 114110 (2007)
  3. linear response theory for large systems
     - Coriani *et al.* J. Chem. Phys. **126**, 154108 (2007)
- In the second part, we consider Gaussian integral evaluation
  - unified scheme for undifferentiated and differentiated integrals
  - Reine *et al.*, Phys. Chem. Chem. Phys. (accepted)

The one-electron density matrix

- Traditional SCF theory rests on the concept of molecular orbitals (MOs):

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon} \leftarrow \text{Roothaan–Hall equations}$$

  - cost of diagonalization is cubic—too expensive for large systems

- Fortunately, MOs are not needed: all information is contained in the density matrix

$$\mathbf{D} = \mathbf{C}_{\mathrm{occ}}^{\mathrm{T}} \mathbf{C}_{\mathrm{occ}} \leftarrow \text{density matrix}$$

  - it is sufficient to work in terms of the one-electron density matrix
  - this is advantageous if it can be compactly (sparsely) represented

- This approach is taken in our work, but then several questions arise:
  - in what basis do we represent the density matrix?
  - how do we optimize the density matrix?
  - how do we calculate its response to perturbations?

- All tasks should be accomplished in terms of elementary matrix manipulations
  - additions, multiplications, trace operations—no diagonalizations
  - with sparsity, linear scaling is then achievable for large systems

Part I: Orthonormalization of the AO basis

- For large systems, the atomic-orbital (AO) basis is ideal in the sense that it is sparse
  - unfortunately, many computational tasks are ill-conditioned in this basis
- This may be avoided by transforming to an orthonormalized AO (OAO) basis

$$\mathbf{Z}^{\mathrm{T}}\mathbf{S}\mathbf{Z} = \mathbf{I}, \quad \mathbf{Z} = \begin{cases} \mathbf{S}^{-1/2} & \text{Löwdin} \\ \mathbf{U}^{-1} \text{ (upper triangular)} & \text{Cholesky} \end{cases}$$

  - Millam and Scuseria, JCP **106**, 5569 (1997); Challacombe, JCP **110**, 2332 (1999)
- The Löwdin basis resembles the AO basis most closely (Carlson and Keller, 1957)
  - it should therefore retain sparsity to the greatest possible extent
  - however, its construction traditionally involves diagonalization
- The Cholesky basis has therefore been preferred instead
  - many uses in quantum chemistry (rank reduction techniques)
- However, since the Löwdin basis is (presumably) the sparsest of all OAO bases, it would be surprising if it could not be generated in linear time!
- We have recently presented a robust linear-scaling algorithm for Löwdin factorization
  - Jansik *et al.*, J. Chem. Phys. **126**, 124104 (2007)

A globally convergent Newton–Schulz method for the inverse square root

- Application of Newton's elementary method for root finding gives

$$Z^{-2} - S = 0 \quad \Rightarrow \quad Z_{n+1} = \frac{3}{2} Z_n - \frac{1}{2} Z_n^3 S \;\leftarrow \text{Newton–Schulz iteration}$$

  – Z_n converges quadratically but not globally to $S^{-1/2}$; no inversion!

- In 2004, Niklasson proposed the corresponding matrix iteration

$$\mathbf{Z}_{n+1} = \frac{1}{2} \mathbf{Z}_n \left(3\mathbf{I} - \mathbf{Z}_n^{\mathrm{T}} \mathbf{S} \mathbf{Z}_n \right) \leftarrow \text{stabilized Newton–Schulz matrix iteration}$$

  – for $\mathbf{Z}_0 = \mathbf{I}$, there are severe convergence restrictions $\|\mathbf{S} - \mathbf{I}\|_2 < 1$ for such methods

  – intended for refinement of approximate inverse square roots

- However, convergence is guaranteed with the scaled $\lambda_{\min}\mathbf{S}$ that minimizes $\|\lambda\mathbf{S} - \mathbf{I}\|_2$:

$$\lambda_{\min} = \frac{2}{\epsilon_{\min} + \epsilon_{\max}} \leftarrow \text{expensive}$$

  – to avoid eigenvalues, we replace the 2-norm by a ratio of Frobenius norms

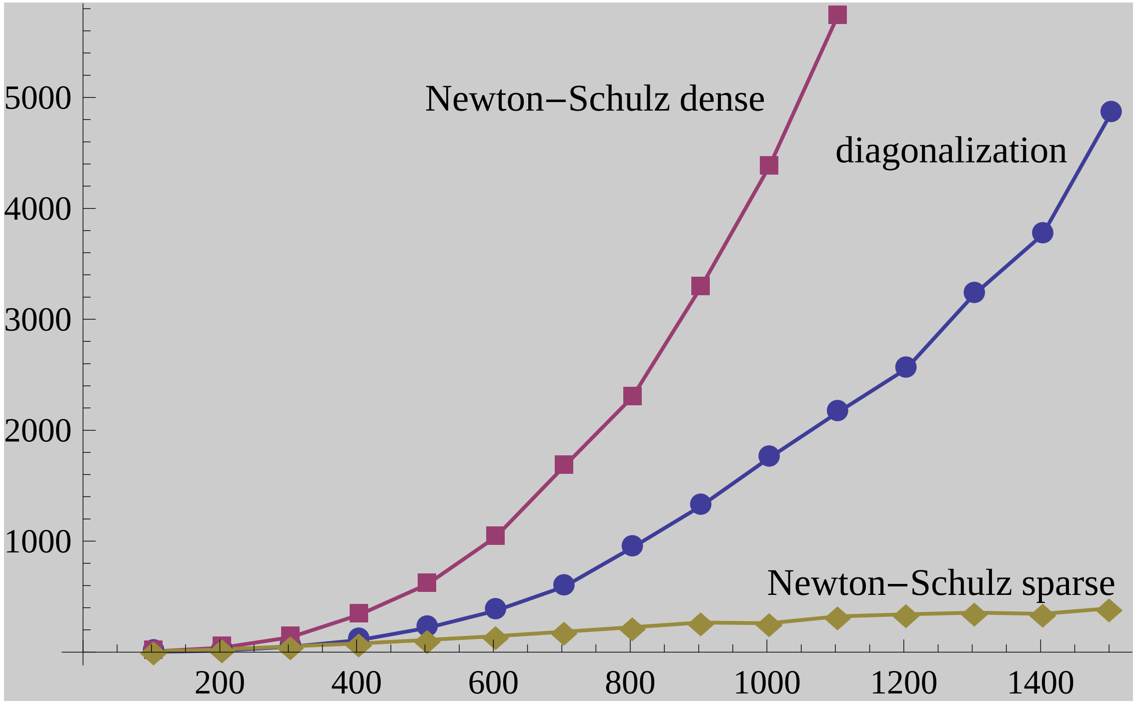
$$f(\lambda) = \sqrt{\frac{\mathrm{Tr}(\lambda\mathbf{S} - \mathbf{I})^4}{\mathrm{Tr}(\lambda\mathbf{S} - \mathbf{I})^2}} \leftarrow \text{cheap lower bound to the 2-norm}$$

  – the evaluation and minimization of $f(\lambda)$ costs only one matrix multiplication

Performance of the scaled Newton–Schulz method

- About 15 scaled Newton–Schulz iterations needed for convergence to 10^{-10}
  - less than 50 matrix multiplies
  - example: polyalanine peptide residues (6-31G) (cost vs. number of atoms)

- In dense-matrix algebra:
  - of cubic complexity
  - two–three times more expensive than diagonalization
  - easy to parallelize
  - beats diagonalization on 4–6 processors
- In sparse-matrix algebra:
  - of linear complexity
  - beats diagonalization for a few hundred atoms



Part II: Energy optimization

- In traditional SCF theory, we solve the Roothaan–Hall (RH) eigenvalue problem

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}, \quad \mathbf{D}_{\mathrm{new}} = \mathbf{C}_{\mathrm{occ}}\mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}$$

- Its solution is equivalent to minimizing the sum of the occupied orbital energies

$$\epsilon_{\mathrm{min}} = \min \textstyle\sum_i \epsilon_i = \min_{\mathbf{D}} \operatorname{Tr} \mathbf{FD} \leftarrow \text{ not all variations valid!}$$

- The valid variations may be represented by an antisymmetric matrix $\mathbf{X}$:

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX}), \quad \mathbf{X}^{\mathrm{T}} = -\mathbf{X} \leftarrow \text{ exponential parametrization}$$

  - these transformations satisfy the symmetry, trace and idempotency conditions
  - Helgaker *et al.*, CPL **327**, 397 (2000); Head-Gordon *et al.*, MP **101**, 37 (2003)

- Each RH step is therefore equivalent to the trace minimization problem

$$\epsilon_{\mathrm{min}} = \min_{\mathbf{X}} \operatorname{Tr} \mathbf{FD}(\mathbf{X}), \quad \mathbf{D}_{\mathrm{new}} = \mathbf{D}(\mathbf{X}_{\mathrm{min}})$$

  - minimization performed with Newton's method, in sparse-matrix algebra

- We are here not concerned with other aspects of the energy optimization
  - Pulay's DIIS may be applied in the same manner as for small systems

The Roothaan–Hall Newton method

- To carry out the trace minimization, we expand the energy in powers of **X**:

$$\begin{aligned}\operatorname{Tr}\mathbf{F}\mathbf{D}(\mathbf{X}) &= \operatorname{Tr}\mathbf{F}\exp(-\mathbf{X})\mathbf{D}\exp(\mathbf{X})\\ &= \operatorname{Tr}\mathbf{F}\mathbf{D} + \operatorname{Tr}\mathbf{F}\big[\mathbf{D},\mathbf{X}\big] + \tfrac{1}{2}\operatorname{Tr}\mathbf{F}\big[\big[\mathbf{D},\mathbf{X}\big],\mathbf{X}\big] + \cdots\end{aligned}$$

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

$$\left(\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}\right)\mathbf{X} + \mathbf{X}\left(\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}\right) = \mathbf{F}^{\mathrm{vo}} - \mathbf{F}^{\mathrm{ov}} \leftarrow \text{Roothaan–Hall Newton equations}$$

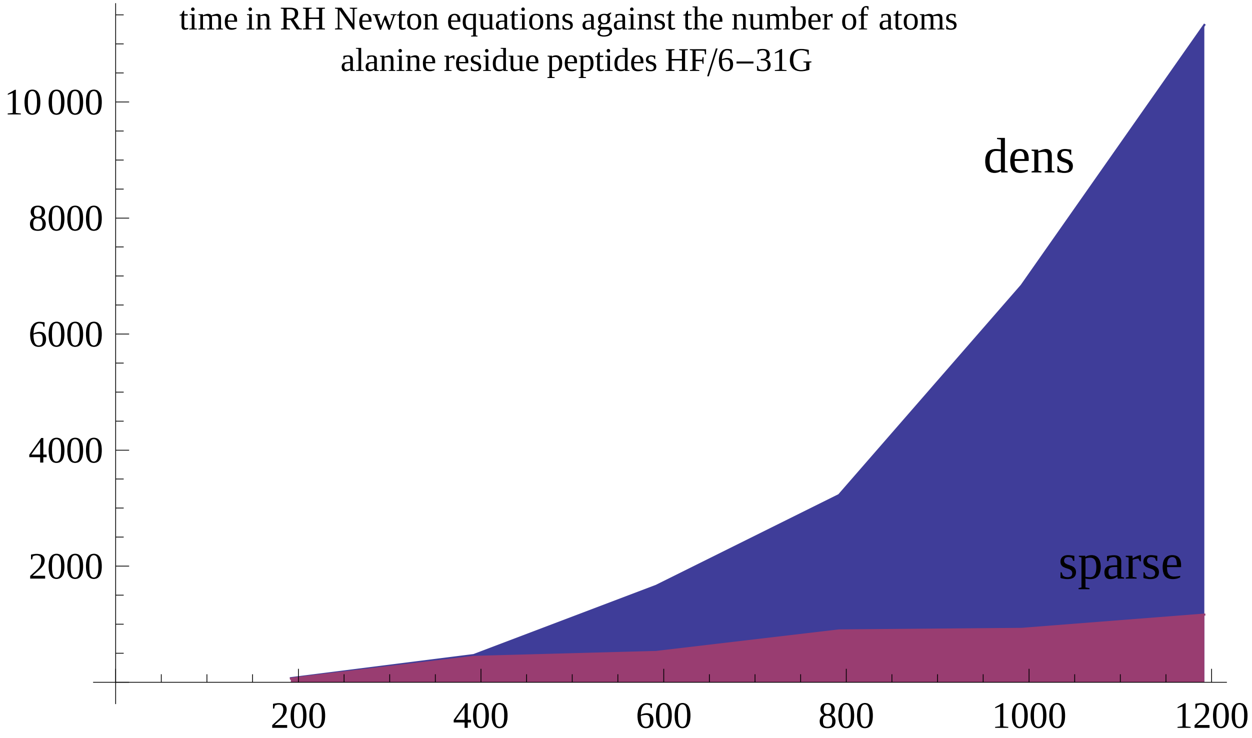
  – upper indices indicate projections onto the occupied and virtual spaces

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

- Because of their large dimensions, the Newton equations cannot be solved directly
  - solution by the preconditioned conjugate-gradient method (typically 10 iterations)
  - elementary (sparse) matrix manipulations (typically less than 100 multiplications)
- A RH diagonalization corresponds to an exact minimization (many Newton steps)
  - however, a partial minimization will do (one RH Newton step is sufficient)
- Larsen *el al.*, J. Chem. Phys. **113** 8908 (2000); Shao *et al.*, *ibid.* **118** 6144 (2003)
  Sałek *et al.*, *ibid.* **126** 114110 (2007)

Performance of the Roothaan–Hall Newton method

- We have successfully avoided Roothaan–Hall 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



Part III: 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

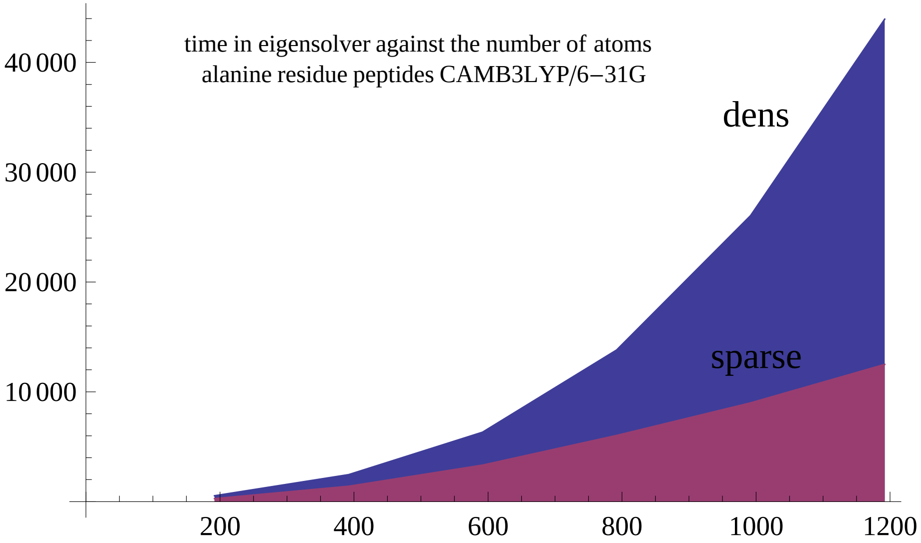
Excitation energies

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

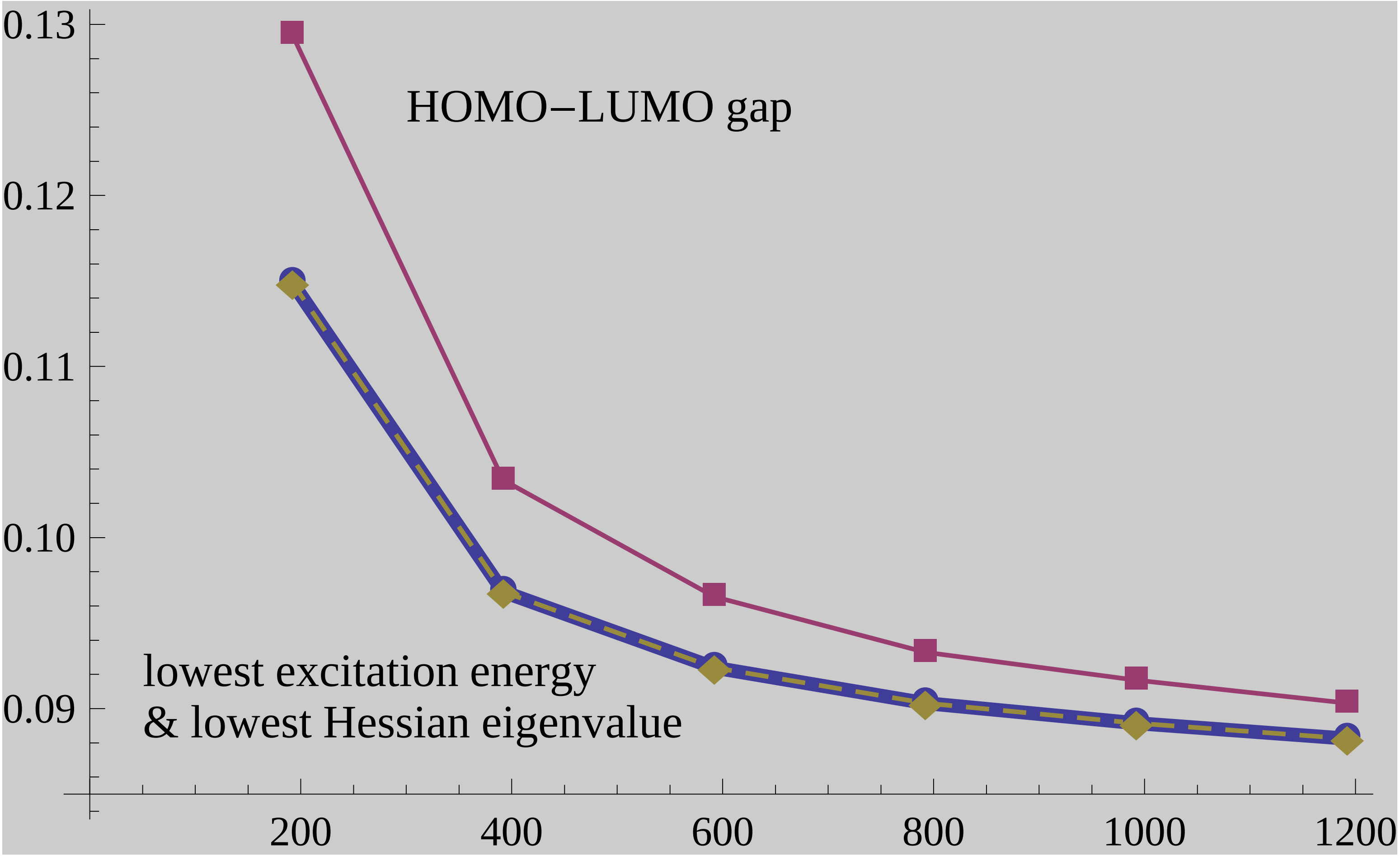
$$\mathbf{E}_{\mathrm{F}}^{(2)}(\mathbf{X}_{ia}) = (\epsilon_a - \epsilon_i)\mathbf{S}^{(2)}(\mathbf{X}_{ia}) \quad \text{where} \quad \mathbf{X}_{ia} = \mathbf{C}_i \mathbf{C}_a^{\mathrm{T}}, \quad \begin{cases} \mathbf{F}^{\mathrm{oo}}\mathbf{C}_i = \epsilon_i \mathbf{C}_i \\ \mathbf{F}^{\mathrm{vv}}\mathbf{C}_a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

  - 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)



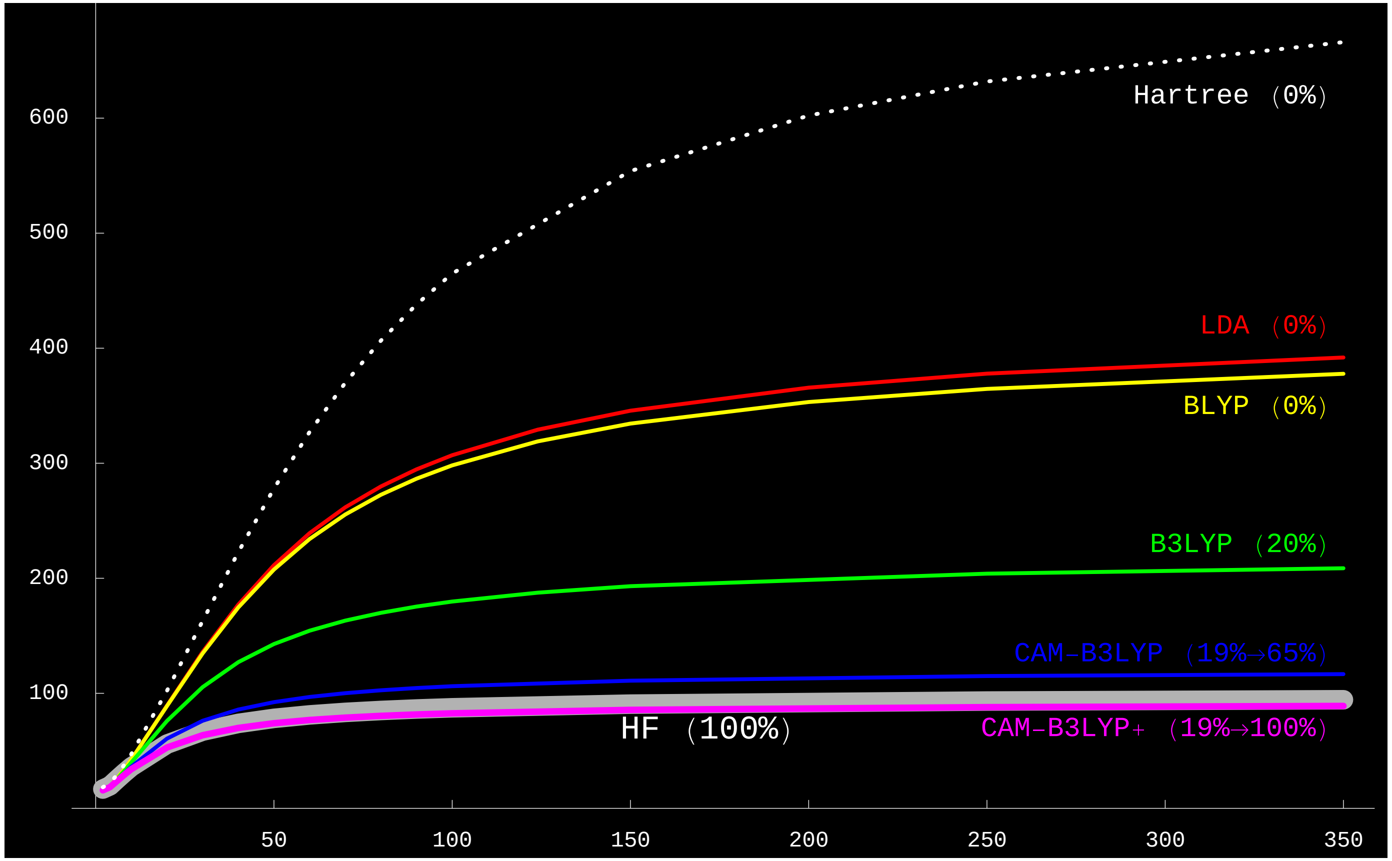
CAMB3LYP/6-31G excitation energies of alanine residue peptides



- lowest excitation energy almost identical with lowest Hessian eigenvalue

Static polarizabilities

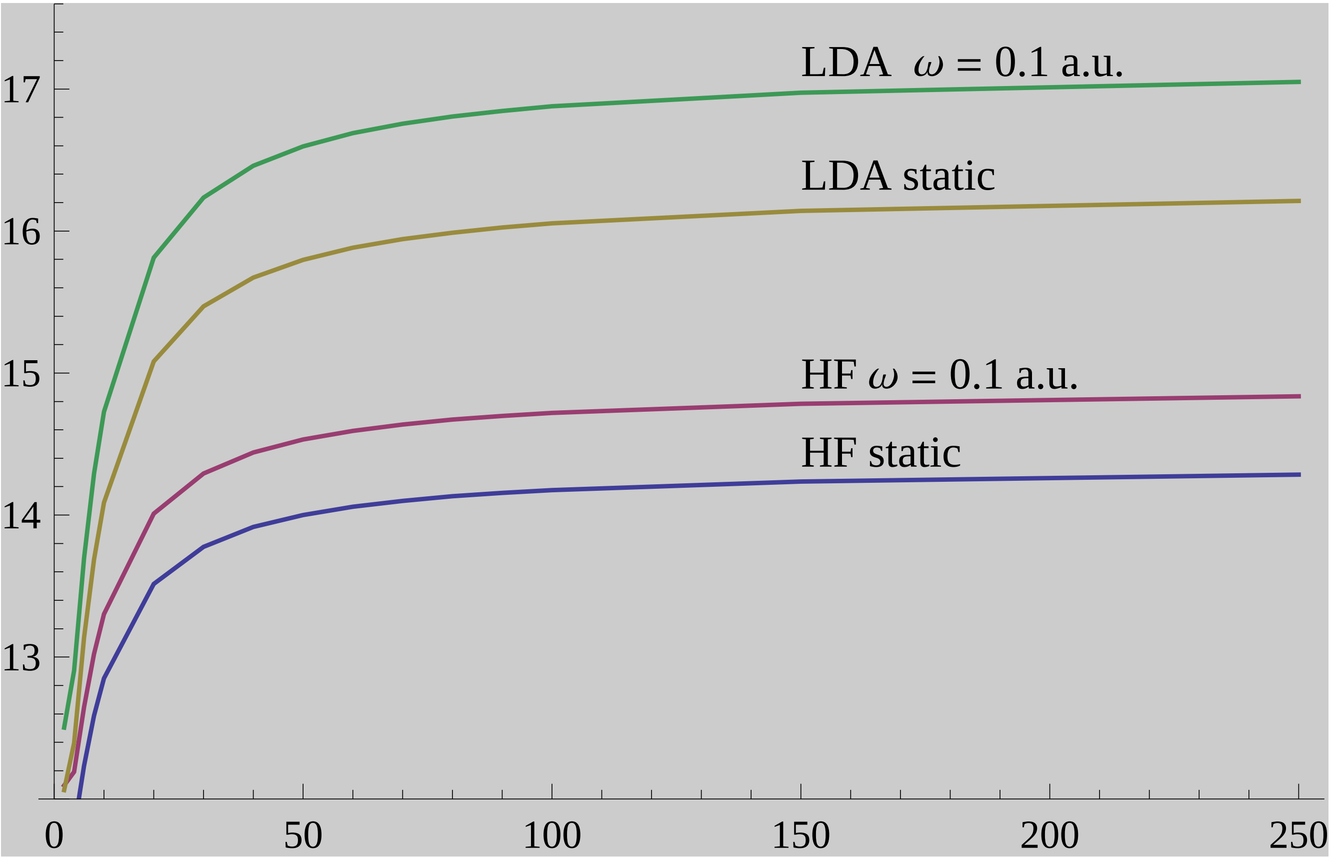
- To illustrate, we have calculated longitudinal polarizabilities of linear alkene chains
  - HF and DFT α/N in 6-31G basis, plotted against the number of carbons N



  - poor long-range exchange leads to overpolarization

Frequency-dependent polarizabilities

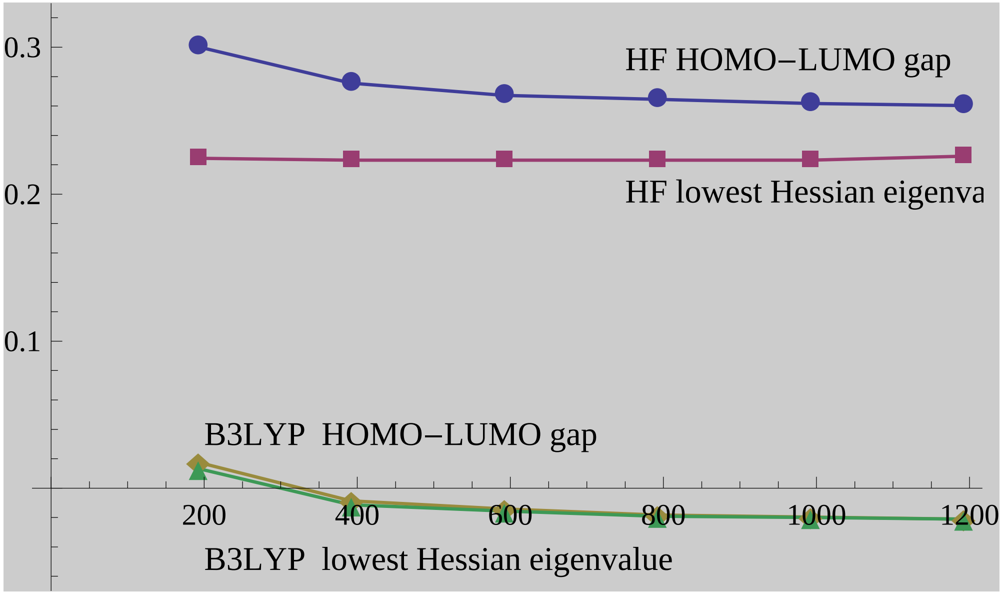
- HF and LDA longitudinal polarizabilities in linear alkane chains
  - α/N in 6-31G basis, plotted against the number of carbons N



  - long-range exchange is less important for the saturated alkanes
  - LDA overestimates dispersion as well as the static limit

Summary

- We have considered several computational tasks in self-consistent field theory
  - orthogonalization of the atomic orbitals
  - solution of the Roothaan–Hall eigenvalue problem
  - linear response theory
- For sufficiently sparse AO matrices, all these 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
- We have not considered the global convergence of the SCF method
  - standard methods may not converge to the ground state
  - better global optimization schemes must be developed
  - standard functionals often perform poorly for large systems



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

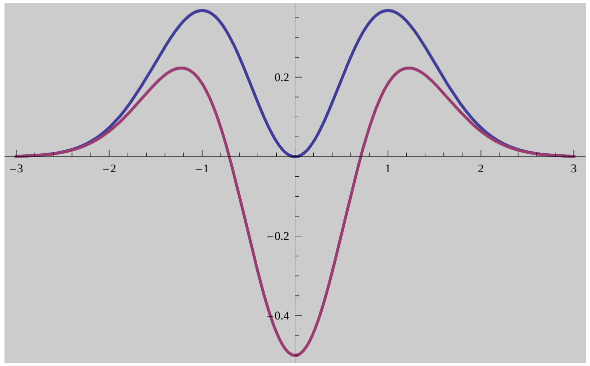
- Hermite Gaussians were introduced by Zivkovic and Maksic (1968)
  - used as intermediates by McMurchie and Davidson (1978)

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
  - all results remain the same!
  - Reine, Tellgren and Helgaker, Phys. Chem. Chem. Phys. (accepted)

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)

Integration over Hermite Gaussians by differentiation

- Let us consider the evaluation of one-electron integrals over Hermite Gaussians

$$\Omega_{ijk,lmn} = \left(H_{ijk}(\mathbf{r}_A)\middle|\hat{\Omega}\middle|H_{lmn}(\mathbf{r}_B)\right)$$

- Substituting the Hermite Gaussians

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

  and invoking Leibniz rule for differentiation under the integral sign, we obtain

$$\Omega_{ijk,lmn} = \frac{\partial^{i+\cdots+n}}{(2a\partial A_x)^i\cdots(2b\partial B_z)^n}\int\exp(-ar_A^2)\,\hat{\Omega}\,\exp(-br_B^2)\mathrm{d}\mathbf{r}$$

- Integration over Hermite Gaussians has been reduced to differentiation of s integrals
  - for example, overlap integrals become

$$S_{ijk,lmn} = \left(\frac{\pi}{a+b}\right)^{3/2}\frac{\partial^{i+\cdots+n}\exp\left(-\frac{ab}{a+b}R_{AB}^2\right)}{(2a\partial A_x)^i\cdots(2b\partial B_z)^n}$$

  - recurrence relations may be established for differentiation
  - similar to the usual Obara–Saika relations (1986) for Cartesian Gaussians

Two-electron integrals over Hermite Gaussians

- For the four-center two-electron integrals, the basic s integral was given by Boys (1950):

$$g_{\mathrm{ssss}}^{abcd} = \frac{S_{\mathrm{ss}}^{ab}(R_{AB})S_{\mathrm{ss}}^{cd}(R_{CD})}{R_{PQ}} \operatorname{erf}\left(\sqrt{\tfrac{pq}{p+q}}R_{PQ}\right)$$

  where S_{ss}^{ab} and S_{ss}^{cd} are s overlap integrals and where we have introduced

$$p = a + b, \quad p\mathbf{P} = a\mathbf{A} + b\mathbf{B}, \quad q = c + d, \quad q\mathbf{Q} = c\mathbf{C} + d\mathbf{D}$$

  - straightforward differentiation gives eight-term recurrence relations
  - similar to the standard Obara–Saika recurrences (1986), only one term differing

- When fewer centers are present, integration becomes simpler
  - two-center integrals between Gaussians $\exp(-pr_P^2)$ and $\exp(-qr_Q^2)$ become

$$g_{\mathrm{ss}}^{pq} = \left(\frac{\pi^2}{pq}\right)^{3/2} \frac{\operatorname{erf}\left(\sqrt{\tfrac{pq}{p+q}}R_{PQ}\right)}{R_{PQ}}$$

  - simpler two-term recurrence relations because only six coordinates are involved
  - same as the McMurchie–Davidson relations (1978) for intermediate Hermite integrals
  - three-center integrals give recurrences similar to those discussed by Ahlrichs (2004)

Summary: Hermite expansions of solid-harmonic Gaussians

- Solid-harmonic Gaussians may be expanded in Hermite or Cartesians Gaussians
  - all nonleading terms of the Hermite Gaussians cancel
  - the resulting functions become identical
- Some advantages of Hermite expansion
  - all integrals become derivatives of s integrals
  - recurrence relations easily established by differentiation
  - derivative codes are easy to develop, to any order
  - immediate simplifications for few-center integrals
  - translational and rotational symmetries easier to use