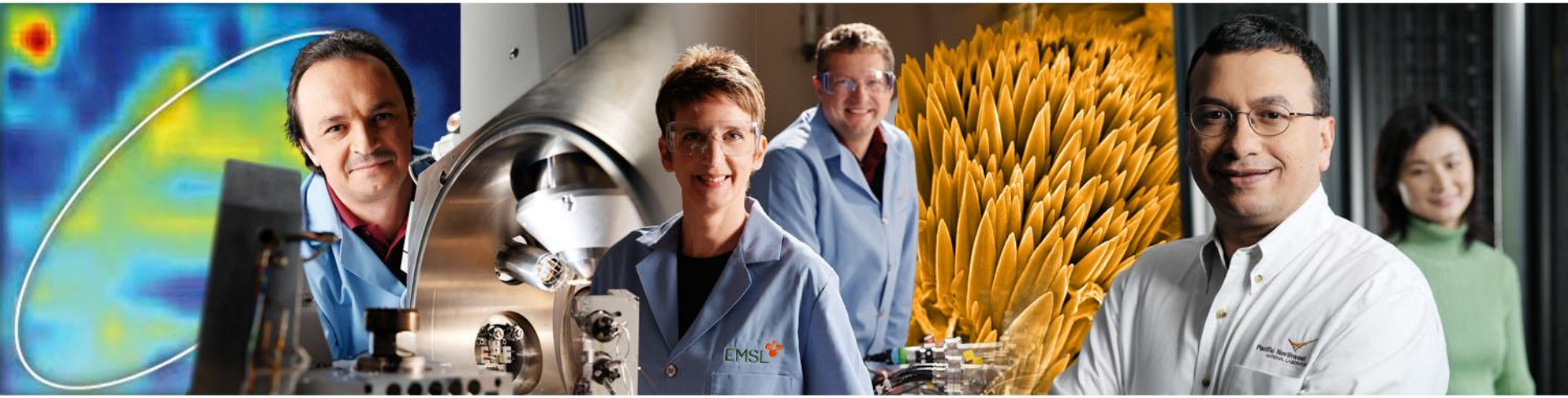
EMSL

Basic Introduction of Computational Chemistry



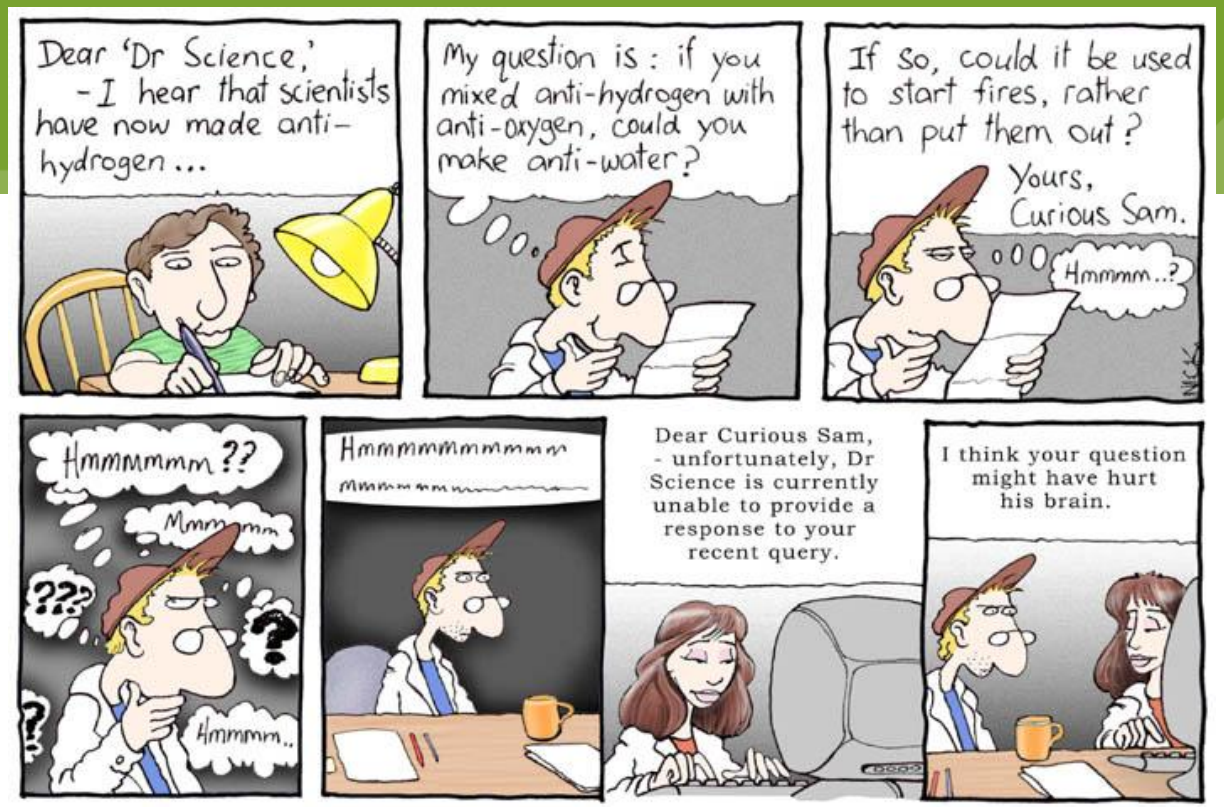
www.emsl.pnl.gov

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

U.S. DEPARTMENT OF ENERGY

What is Computational Chemistry?



- Definition
- Some examples
- The relation to the real world

Cartoons from "scientist at work; work, scientists, work!"

Computational Chemistry is…

- A branch of chemistry
- That uses equations encapsulating the behavior of matter on an atomistic scale and
- Uses computers to solve these equations
- To calculate structures and properties
- Of molecules, gases, liquids and solids
- To explain or predict chemical phenomena.

- See also:
  - Wikipedia,
  - http://www.chem.yorku.ca/profs/renef/whatiscc.html [11/10/2010]
  - http://www.ccl.net/cca/documents/dyoung/topics-orig/compchem.html [11/10/2010]

Computational Chemistry is…

EMSL

- Including:
  - Electron dynamics
  - Time independent ab initio calculations
  - Semi-empirical calculations
  - Classical molecular dynamics
  - Embedded models
  - Coarse grained models
- Not including:
  - Quantum chromo-dynamics
  - Calculations on Jellium
  - Continuum models
  - Computational fluid dynamics
  - Data mining
  - Rule based derivations

Computational Chemistry is…

EMSL

- "To explain or predict chemical phenomena":
  - Phenomenon is any observable occurrence
  - Therefore computational chemistry has to connect with practical/experimental chemistry
  - In many cases fruitful projects live at the interface between computational and experimental chemistry because:
    - Both domains criticize each other leading to improved approaches
    - Both domains are complementary as results that are inaccessible in the one domain might be easily accessible in the other
    - Agreeing on the problem helps focus the invested effort

Where do you start?

- Selection of energy expressions
- Hartree-Fock / Density Functional Theory
- Moller-Plesset Perturbation Theory
- Coupled Cluster
- Quantum Mechanics / Molecular Mechanics
- Molecular Mechanics

You start with…

- Everything starts with an energy expression
- Calculations either minimize to obtain:
  - the ground state
  - equilibrium geometries
- Or differentiate to obtain properties:
  - Infra-red spectra
  - NMR spectra
  - Polarizabilities
- Or add constraints to
  - Optimize reaction pathways (NEB, string method, ParaReal)
- The choice of the energy expression determines the achievable accuracy

Energy expressions NWChem supports

- Effective 1-Electron Models
  - Hartree-Fock and Density Functional Theory
  - Plane wave formulation
  - Local basis set formulation
- Correlated Models
  - Møller-Plesset Perturbation Theory
  - Coupled Cluster
- Combined Quantum Mechanical / Molecular Mechanics (QM/MM)
- Molecular Mechanics

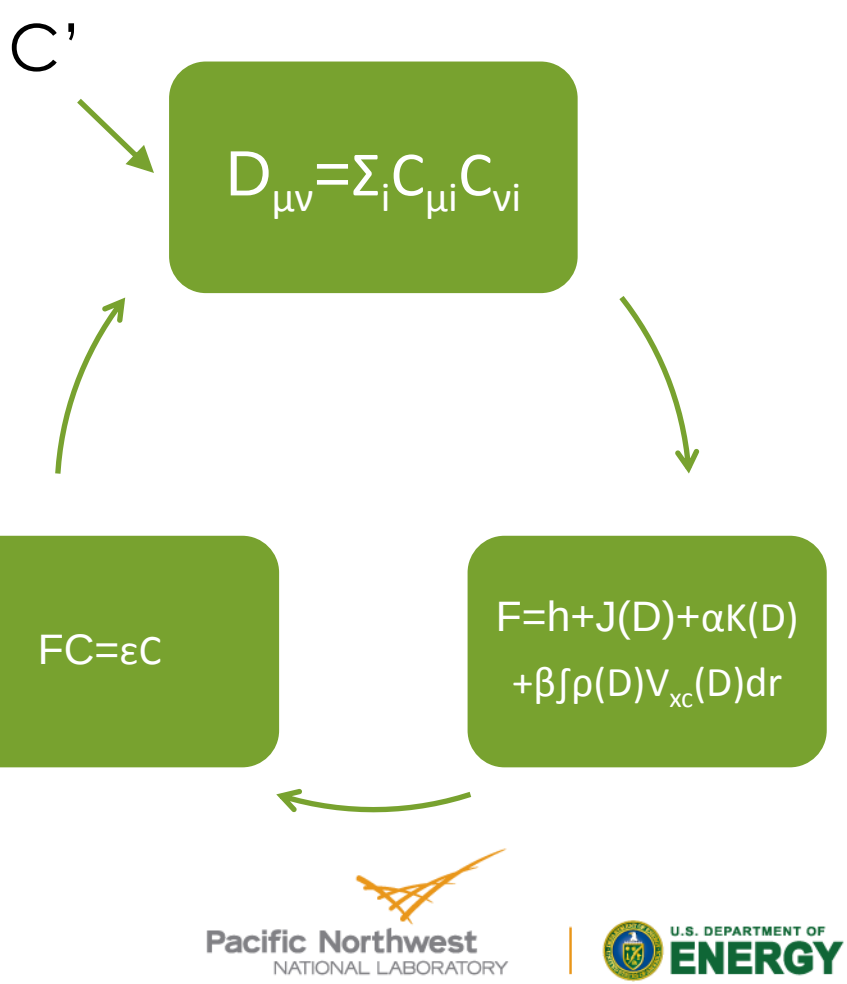
Hartree-Fock & Density Functional Theory I
Plane wave & Local basis

EMSL

- The energy expression is derived from a single determinant wave function approximation
- Replace the exchange with a functional to go from Hartree-Fock to DFT
- Use different basis sets for different problems
  - Plane waves for infinite condensed phase systems
  - Local basis sets for finite systems

Hartree-Fock & Density Functional Theory II
Plane wave & Local basis

- Minimize energy with respect to $C_{\mu i}$ and ε_i
- Iterative process cycling until Self-Consistency
- Gives
  - The total energy *E*
  - The molecular orbitals $C_{\mu i}$
  - The orbital energies ε_i



Hartree-Fock & Density Functional Theory III Local Basis Sets

EMSL

Memory requirements

- Largest quantities are the density, Fock, overlap, 1-electron matrices
- Memory needed $O(N^2)$
  - Replicated data $O(N^2)$ per node
  - Distributed data $O(N^2)$ for whole calculation

Computational Complexity

- Main cost is the evaluation of the 2-electron integrals
- Takes $O(N^2)$-$O(N^4)$ work
  - $O(N^4)$ for small systems
  - $O(N^2)$ in the large N limit
- For large N the linear algebra becomes dominant at $O(N^3)$

Hartree-Fock & Density Functional Theory IV Plane waves

EMSL

Memory requirements

- Largest quantities are the density, Fock, overlap, 1-electron matrices
- Memory needed $O(N^2)$
  - Replicated data $O(N^2)$ per node
  - Distributed data $O(N^2)$ for whole calculation

Computational Complexity

- Main cost stems from the Fourier transforms
- For small systems and large processor counts dominated by FFTs costing $O(N^2*\ln(N))$
- For large systems the non-local operator and orthogonalization are important costing $O(N^3)$

Møller-Plesset Perturbation Theory I

- Assumes that electron correlation effects are small
- The Hartree-Fock energy is the 1st order corrected energy
- The 2nd and 3rd order corrected energy can be calculated from the 1st order corrected wave function (2N+1 rule)

Møller-Plesset Perturbation Theory II

- The zeroth order energy is the sum of occupied orbital energies
- The first order energy is the Hartree-Fock energy
- The energy correction gives an estimate of the interaction of the Hartree-Fock determinant with all singly and double substituted determinants
- The total energy scales correctly with system size
- It is ill defined if the HOMO and LUMO are degenerate

$$E = E^{HF} + E^{\Delta MP2}$$

$$E^{\Delta MP2} = \sum_{\substack{i,j \in \{occ\} \\ r,s \in \{virt\}}} \frac{\left|(ij|rs)\right|^2}{\varepsilon_i + \varepsilon_j - \varepsilon_r - \varepsilon_s}$$

http://en.wikipedia.org/wiki/Moller-Plesset_perturbation_theory [11/19/2010]

Møller-Plesset Perturbation Theory III

Memory requirements

- The MO basis 2-electron integrals require the dominant amount of storage
  - This takes $O(N^4)$ storage
  - Can be reduced to $O(N^3)$ by treating the integral in batches

Computational complexity

- Transforming the integrals requires a summation over all basis functions for every integral
  - This takes $O(N^5)$ work
  - If not all transformed integrals are stored then there is an extra cost of calculating all the integrals for every batch at $O(N^4)$

Coupled Cluster I

- Sums some corrections to infinite order
- Involves singly, doubly, triply-substituted determinants
- Simplest form is CCSD
- Often necessary to include triply-substituted determinants (at least perturbatively), i.e. CCSD(T)

Coupled Cluster II

- The wavefunction is expressed in an exponential form
- The operator *T* contains single and double substitution operators with associated amplitudes
- The vector equation:
  - Solve top two lines for the amplitudes d_i^r, d_{ij}^{rs}
  - The bottom line gives the total energy

$$|\Psi\rangle = e^{\hat{T}} |\Psi^{HF}\rangle$$

$$\hat{T} \approx \hat{T}_1 + \hat{T}_2$$

$$= \sum_{\substack{i\in\{occ\} \\ r\in\{virt\}}} d_i^r a_r^+ a_i + \sum_{\substack{i,j\in\{occ\} \\ r,s\in\{virt\}}} d_{ij}^{rs} a_r^+ a_s^+ a_i a_j$$

$$\begin{pmatrix} 0 \\ 0 \\ E \end{pmatrix} = \begin{pmatrix} \langle \Psi_{ij}^{rs} | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Psi^{HF} \rangle \\ \langle \Psi_i^r | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Psi^{HF} \rangle \\ \langle \Psi^{HF} | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Psi^{HF} \rangle \end{pmatrix}$$

http://en.wikipedia.org/wiki/Coupled_cluster [11/19/2010]

Coupled Cluster III

EMSL

Memory requirements

- The main objects to store are the transformed 2-electron integrals and the amplitudes
- This costs $O(N^4)$ storage
- Local memory depends on tile sizes and level of theory
  - CCSD – $O(n_t^4)$
  - CCSD(T) – $O(n_t^6)$

Computational complexity

- The main cost are the tensor contractions
- For CCSD they can be formulated so that they take $O(N^6)$ work
- For CCSD(T) the additional perturbative step dominates at $O(N^7)$

QM/MM Models I

- Describe local chemistry under the influence of an environment
  - Quantum region treated with ab-initio method of choice
  - Surrounded by classical region treated with molecular mechanics
  - Coupled by electrostatics, constraints, link-atoms, etc.
- Crucial part is the coupling of different energy expressions

QM/MM Models II

- The scheme we are considering is an hybrid scheme (not an additive scheme like ONIOM)
- This scheme is valid for any situation that the MM and QM methods can describe, the limitation lies in the interface region

$$E = E_{QM}(r, R, \Psi) + E_{MM}(r, R)$$

$$E_{QM} = E_{QM}^{\text{internal}}(r, R, \Psi) + E_{QM}^{\text{external}}(R, \rho[\Psi])$$

$$E_{QM}^{\text{external}} = \sum_{I \in MM} \int \frac{Z_I \rho(r')}{|R_I - r'|} dr' + \sum_{I \in MM} \sum_{i \in QM} \frac{Z_I Z_i}{|R_I - R_i|} + E_{vdW}(R)$$

http://www.ccl.net/cca/documents/dyoung/topics-orig/qmmm.html [11/20/2010]
http://www.salilab.org/~ben/talk.pdf [11/20/2010]
http://www.chem.umn.edu/groups/gao/qmmm_notes/LEC_HYB.html [11/20/2010]

QM/MM Models III

EMSL

Memory requirements

- Dominated by the memory requirements of the QM method
- See chosen QM method, N is now the size of the QM region

Computational complexity

- Dominated by the complexity of the QM method
- See chosen QM method, N is now the size of the QM region

Molecular Mechanics I

- Energy of system expressed in terms of relative positions of atoms
- The parameters depend on the atom and the environment
  - A carbon atom is different than a oxygen atom
  - A carbon atom bound to 3 other atoms is different from one bound to 4 other atoms
  - A carbon atom bound to hydrogen is different from one bound to fluorine
  - Etc.

Molecular mechanics II

$$E = \tfrac{1}{2}\sum_{A,B} k_{AB}\left(r_{AB} - r^0_{AB}\right)^2$$
$$+\tfrac{1}{2}\sum_{A,B,C} k_{ABC}\left(\theta_{ABC} - \theta^0_{ABC}\right)^2$$
$$+\tfrac{1}{2}\sum_{A,B,C,D}\sum_{n} v_{ABCD;n}\left[1+\cos\left(n\phi_{ABCD} - \phi^0_{ABCD}\right)\right]$$
$$+\tfrac{1}{2}\sum_{A,B}\left[\frac{\alpha_{AB}}{r^{12}_{AB}} - \frac{\beta_{AB}}{r^{6}_{AB}}\right]$$
$$+\tfrac{1}{2}\sum_{A,B}\frac{\varepsilon_{AB} q_A q_B}{r_{AB}}$$

- Energy terms (parameters)
  - Bond distances (k_{AB}, r^0_{AB})
  - Bond angles (k_{ABC}, θ^0_{ABC})
  - Dihedral angles ($v_{ABCD;n}$,ϕ^0_{ABCD})
  - Van der Waals interactions (α_{AB},β_{AB})
  - Electrostatic interactions (ε_{AB})
- The parameters are defined in special files
- For a molecule the parameters are extracted and stored in the topology file
- This force field is only valid near equilibrium geometries

W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, Jr., D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, P. Kollman, J. Am. Chem. Soc., (1996) **118**, 2309, http://dx.doi.org/10.1021/ja955032e

Molecular Mechanics III

EMSL

Memory requirements

- Main data objects are the atomic positions
- Storage O(N)

Computational complexity

- Most terms involve local interactions between atoms connected by bonds, these cost O(N) work
  - Bond terms
  - Angle terms
  - Dihedral angle terms
- The remaining two terms involve non-local interactions, cost at worst $O(N^2)$, but implemented using the particle mesh Ewald summation it costs O(N*log(N))

Summarizing methods

Method	Memory	Complexity	Strengths
Molecular Dynamics	O(N)	O(N*ln(N))	Conformational sampling/Free energy calculations
Hartree-Fock/DFT	$O(N^2)$	$O(N^3)$	Equilibrium geometries, 1-electron properties, also excited states
Møller-Plesset	$O(N^4)$	$O(N^5)$	Medium accuracy correlation energies, dispersive interactions
Coupled-Cluster	$O(N^4)$	$O(N^6)$-$O(N^7)$	High accuracy correlation energies, reaction barriers
QM/MM	*	*	Efficient calculations on complex systems, ground state and excited state properties

* Depends on the methods combined in the QM/MM framework.

What properties might you want to calculate?

- Energies
- Equilibrium geometries
- Infrared spectra
- UV/Vis spectra
- NMR chemical shifts
- Reaction energies
- Thermodynamics
- Transition states
- Reaction pathways
- Polarizabilities

Energy evaluations

- Having chosen an energy expression we can calculate energies and their differences
  - Bonding energies
  - Isomerization energies
  - Conformational change energies
  - Identification of the spin state
  - Electron affinities and ionization potentials
- For QM methods the wavefunction and for MM methods the partial charges are also obtained. This allows the calculation of
  - Molecular potentials (including docking potentials)
  - Analysis of the charge and/or spin distribution
  - Natural bond order analysis
  - Multi-pole moments

Gradient evaluations

- Differentiating the energy with respect to the nuclear coordinates we get gradients which allows the calculation of
  - Equilibrium and transition state geometries
  - Forces to do dynamics

Hessian evaluations I

- Differentiating the energy twice with respect to the nuclear coordinates gives the Hessian which allows calculating
  - The molecular vibrational modes and frequencies (if all frequencies are positive you are at a minimum, if one is negative you are at a transition state)
  - Infra-red spectra
  - Initial search directions for transition state searches
- Hessians are implemented for the Hartree-Fock and Density Functional Theory methods

Hessian evaluation II

Memory requirements

- In the effective 1-electron models a perturbed Fock and density matrix needs to be stored for every atomic coordinate
- The memory required is therefore $O(N^3)$

Computational complexity

- To compute the perturbed density matrices a linear system of dimension $O(N^2)$ has to be solved for every atomic coordinate
- The number of operations needed is $O(N^5)$

Magnetic properties, e.g. NMR

- The chemical shift is calculated as a mixed second derivative of the energy with respect to the nuclear magnetic moment and the external magnetic field.
- Often the nuclear magnetic moment is treated as a perturbation
- Note that
  - The paramagnetic and diamagnetic tensors are not rotationally invariant
  - The total isotropic and an-isotropic shifts are rotationally invariant
- Requires the solution of the CPHF equations at $O(N^4)$

http://en.wikipedia.org/wiki/Chemical_shift
[11/23/2010]

Polarizabilities

- Adding an external electric field to the Hamiltonian and differentiating the energy with respect to the field strength gives polarizability
  - Hartree-Fock and DFT
  - CCSD, CCSDT

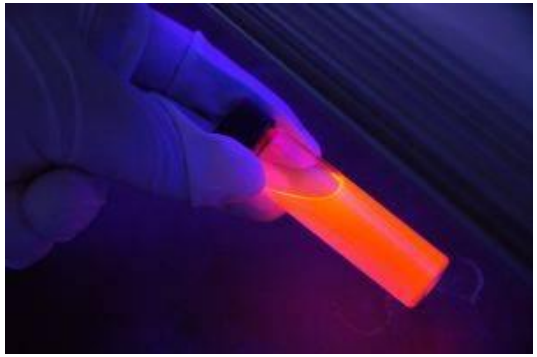
Linear response I

- Adding a time dependent electric field to the Hamiltonian, substituting it in the dependent Schrodinger equation, and expanding the time-dependent density in a series an equation for the first order correction can be obtained.
- This expression is transformed from the time domain to the frequency domain to obtain an equation for the excitation energies
- Solving this equation for every root of interest has a cost of the same order a the corresponding Hartree-Fock or DFT calculation, both in memory requirements as in the computational complexity.

http://www.physik.fu-berlin.de/~ag-gross/articles/pdf/MG03.pdf

Linear response II

- The equations have $N_{occ}*N_{virt}$ solutions
- Note that the vectors are normalized but differently so than your usual wavefunction
- The orbital energy difference is a main term in the excitation energy
- In the case of pure DFT with large molecules most of the integrals involving F_{xc} vanish as this is a local kernel



$$\begin{pmatrix} A & B \\ B^* & A^* \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \omega \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$$

$$1 = \left(X \middle| X\right) - \left(Y \middle| Y\right)$$

$$A_{ia,jb} = \delta_{ij}\delta_{ab}\left(\varepsilon_a - \varepsilon_i\right) + \left(ia\middle|F_H + F_{xc}\middle|jb\right)$$

$$B_{ia,jb} = \left(ia\middle|F_H + F_{xc}\middle|jb\right)$$

$$F_{xc}\left(r_1, r_2\right) = \frac{\partial^2 f}{\partial\rho\left(r_1\right)\partial\rho\left(r_2\right)}$$

http://www.tddft.org/TDDFT2008/lectures/IT2.pdf

EOM CCSD/CCSD(T)

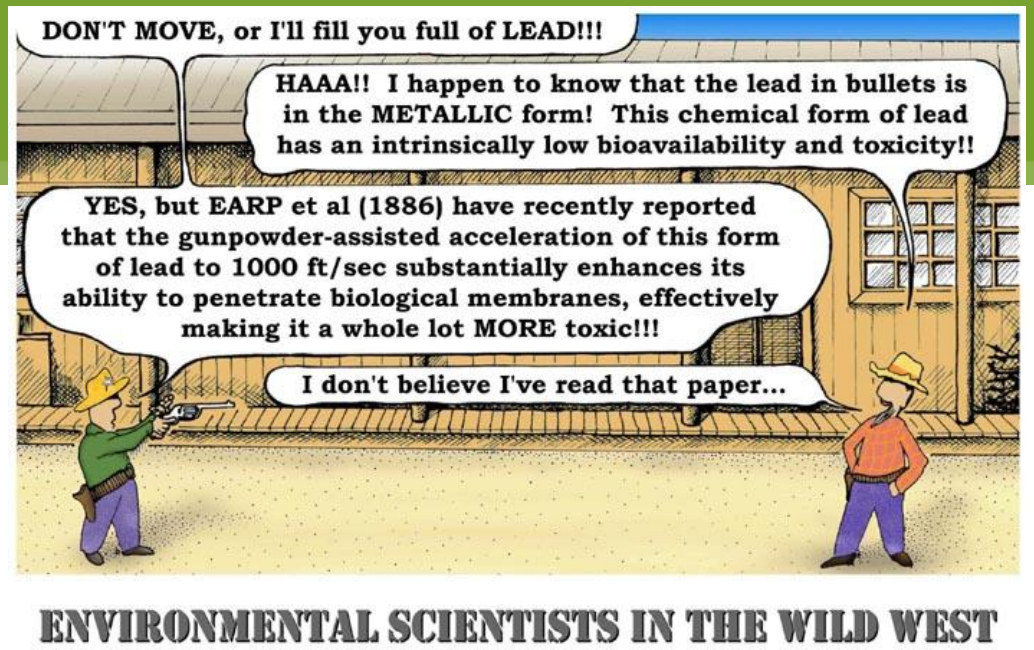
- A Coupled Cluster method for excited states
- Depends on the ground state cluster amplitudes
- Memory and computational complexity similar to corresponding Coupled Cluster method

$$\left|\Psi_k\right\rangle = R_k e^{\hat{T}} \left|\Psi^{HF}\right\rangle$$

$$R_k = r_k + \sum_{i,s} r_{k,i}^{s} a_s^{+} a_i + \sum_{i,j,s,t} r_{k,ij}^{st} a_s^{+} a_t^{+} a_i a_j$$

$$e^{-\hat{T}} \hat{H} e^{\hat{T}} R_k \left|\Psi^{HF}\right\rangle = E_k R_k \left|\Psi^{HF}\right\rangle$$

Which methods do you pick?



- Factors in decision making
- Available functionality
- Accuracy aspects

The decision making factors

- The method of choice for a particular problem depends on a number of factors:
  - The availability of the functionality
  - The accuracy or appropriateness of the method
  - The memory requirements and computational cost of the method

Available functionality

	MM	HF/DFT	MP2	CC	QM/MM
Energy	✓	✓	✓	✓	✓
Gradients	✓	✓	✓	×	✓
Hessians	×	✓	×	×	×
Polarizabilities	×	✓	×	×	✓
Excited states	×	✓	×	✓	✓
NMR	×	✓	×	×	✓

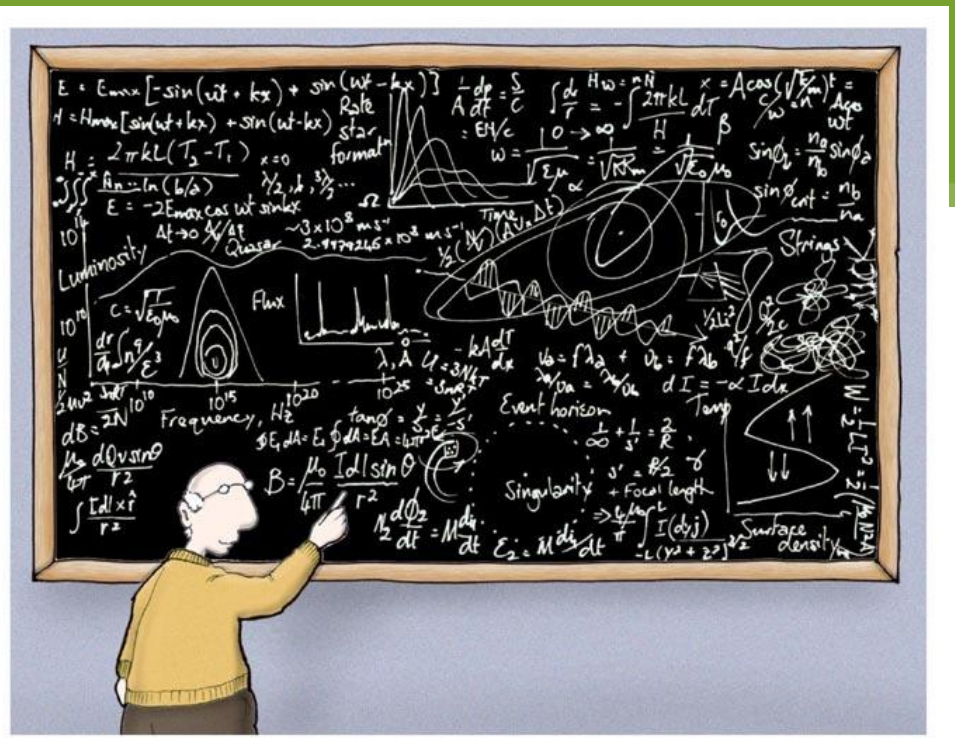
Gradients and Hessians can always be obtained by numerical differentiation but this is slow.

http://www.nwchem-sw.org/

Appropriateness of method

- Too complex a question to answer here in general
- For example consider bond breaking
  - Molecular Mechanics cannot be used as this is explicitly not part of the energy expression
  - HF/DFT can be used but accuracy limited near transition states (unrestricted formulation yields better energies, but often spin-contaminated wavefunctions)
  - Moller-Plesset cannot be used as near degeneracies cause singularities
  - CCSD or CCSD(T) can be used with good accuracy
  - QM/MM designed for these kinds of calculations of course with the right choice of QM region
- So check your methods before you decide, if necessary perform some test calculations on a small problem.
- Often methods that are not a natural fit have been extended, e.g. dispersion corrections in DFT
- Bottom line: Know your methods!

Wrapping up…



Astrophysics made simple

- Further reading
- Acknowledgements
- Questions

Further reading

- D. Young, "*Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*", Wiley-Interscience, 2001, ISBN:0471333689.
- C.J. Cramer, "*Essentials of Computational Chemistry: Theories and Models*", Wiley, 2004, ISBN:0470091827.
- R.A.L. Jones "*Soft Condensed Matter*", Oxford University Press, 2002, ISBN:0198505892.

Further reading (Books)

- **General**
  - D. Young, "*Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*"
  - C.J. Cramer, "*Essentials of Computational Chemistry: Theories and Models*"
  - F. Jensen, "*Introduction to Computational Chemistry*"
- **Molecular dynamics**
  - Frenkel & Smit, "*Understanding Molecular Simulation*"
  - Allen & Tilldesley, "*Computer Simulation of Liquids*"
  - Leach, "*Molecular Modelling: Principles & Applications*"
- **Condensed phase**
  - R.M. Martin, "*Electronic Structure: basic theory and practical methods*"
  - J. Kohanoff, "*Electronic Structure Calculations for Solids and Molecules*"
  - D. Marx, J. Hutter, "*Ab Initio Molecular Dynamics*"
- **Quantum chemistry**
  - Ostlund & Szabo, "*Modern Quantum Chemistry*"
  - Helgaker, Jorgensen, Olsen, "*Molecular Electronic Structure Theory*"
  - McWeeny, "*Methods of Molecular Quantum Chemistry*"
  - Parr & Yang, "*Density Functional Theory of Atoms & Molecules*"
  - Marques et al, "*Time-Dependent Density Functional Theory*"
- **Other**
  - Janssen & Nielsen, "*Parallel Computing in Quantum Chemistry*"
  - Shavitt & Bartlett, "*Many-Body Methods in Chemistry and Physics*"

Acknowledgement

- This work was produced using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

Questions?