Computational Chemistry
An introduction to the computational chemistry package, Gaussian 03

Matt McKenzie

LSU-HPC / LONI
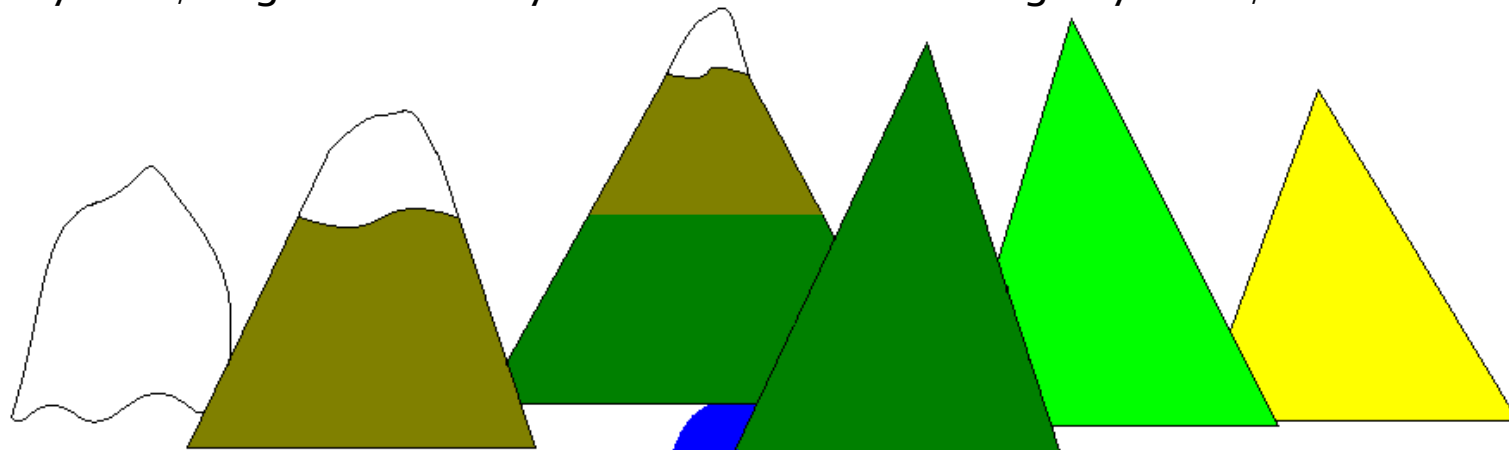
Topics

- Basic Theory
  - Gaussian specific

- Gaussian job creation and submission

- Analysis of output

Computational Chemistry Landscape

Small System/ Higher Accuracy   Large System / Lower Accuracy

Quantum Mechanics   ## Thermodynamics

MO theory, *ab initio*, DFT   Molecular Mechanics
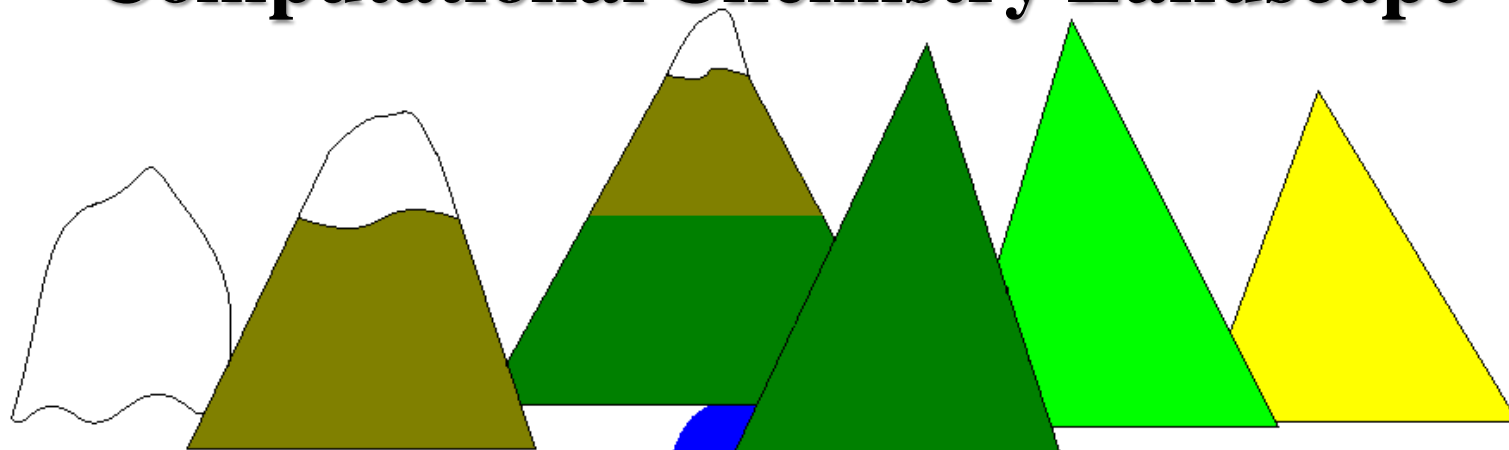
                                               MD, MC, Virtual Screening

          Basis sets   Force fields
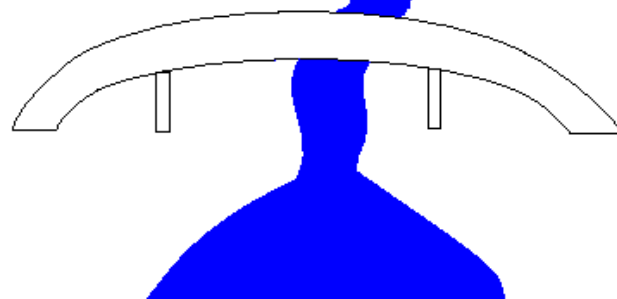
Statistical Mechanics

Computational Chemistry Landscape

Quantum Mechanics   **Thermodynamics**

Gaussian 03, Gamess
Qumax

NAMD, LAMMPS,
Gromacs*, Amber*,
AutoDock

PINY, NWChem   **Statistical Mechanics**   CPMD, QM/MM

Theory Overview

▸ Pre-req. = knowledge of quantum mechanics
▸ Goal: Obtain a <u>fundamental</u> knowledge of *ab initio* calculaitons using Gaussian
  ◦ Cannot cover the entire computational *ab intio* topic

▸ How wave functions are constructed

▸ Levels of theory used

▸ Focus on small molecules, gas phase

Theory

- Solutions to the Schrödinger equation
- $(\hat{H}-E)\Psi(r_1,r_2,\ldots,r_N)=0$

- Ways to solve:
- Numerical: Approx. partial derivatives of Ψ as finite differences of values of Ψ

- Expansion Method: model Ψ,
- $\Psi(r_1,r_2,\ldots,r_N) \approx \Sigma_i c_i \eta_i(r_1,r_2,\ldots,r_N)$
- The set of η_i forms a basis representation of Ψ

Theory – Basis Sets

- Expansion Method
- $(\hat{H}-E)\Psi(r_1,r_2,\ldots,r_N)=0$
- $\Psi(r_1,r_2,\ldots,r_N) \approx \sum_i c_i \eta_i(r_1,r_2,\ldots,r_N)$
- The *set* of η_i forms a *basis* representation of Ψ

From dictionary.com,
- Basis
- 1.   the bottom or base of anything; the part on which something *stands or rests*.
- 2.   *anything upon which something is based*; fundamental principle; groundwork.
- 3.   the principal constituent; *fundamental ingredient*.

- 4.   a *basic fact, amount*, standard, etc., *used in making computations*, reaching conclusions, or the like: The nurse is paid on an hourly basis. He was chosen on the basis of his college grades.

- 5.   Mathematics. *a set of linearly independent elements of a given vector space having the property that every element of the space can be written as a linear combination of the elements of the set.*
  ◦ $\Psi(r_1,r_2,\ldots,r_N) \approx \sum_i c_i \eta_i(r_1,r_2,\ldots,r_N)$
- Latin Origin: step, place one stands on, pedestal

Theory – Basis sets

- $\Psi(r_1, r_2, \ldots, r_N) \approx \sum_i c_i \eta_i(r_1, r_2, \ldots, r_N)$
- Instead of the whole system, let's focus on a molecular orbital

$$\varphi_i = \sum_{n=1}^{N} c_{ni} \eta_n$$

Molecular orbital coefficient
Basis function

- **Increase** η = **higher quality** of the wave function
- **Increase** η = **higher computational cost**

Basis Functions– STO's & GTO's

Slater–Type Orbitals

$$\eta^{STO}(\mathrm{r}) = x^l y^m z^n e^{-\zeta r}$$

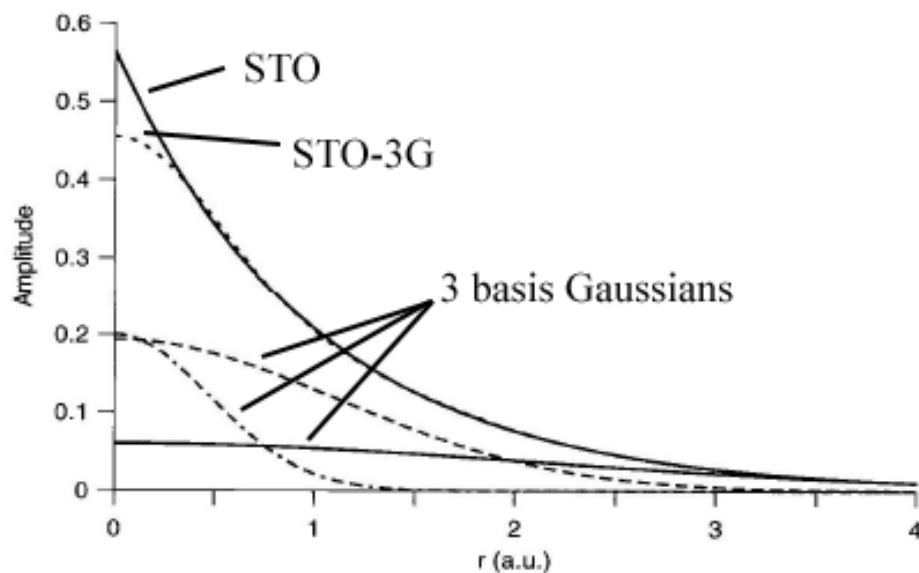Similar to angular momentum
$L = l+m+n$

Gaussian Type Orbital

$$\eta^{GTO}(\mathrm{r}) = x^l y^m z^n e^{-\alpha r^2}$$

ζ & α are the **orbital exponents.**
The radial 'size'
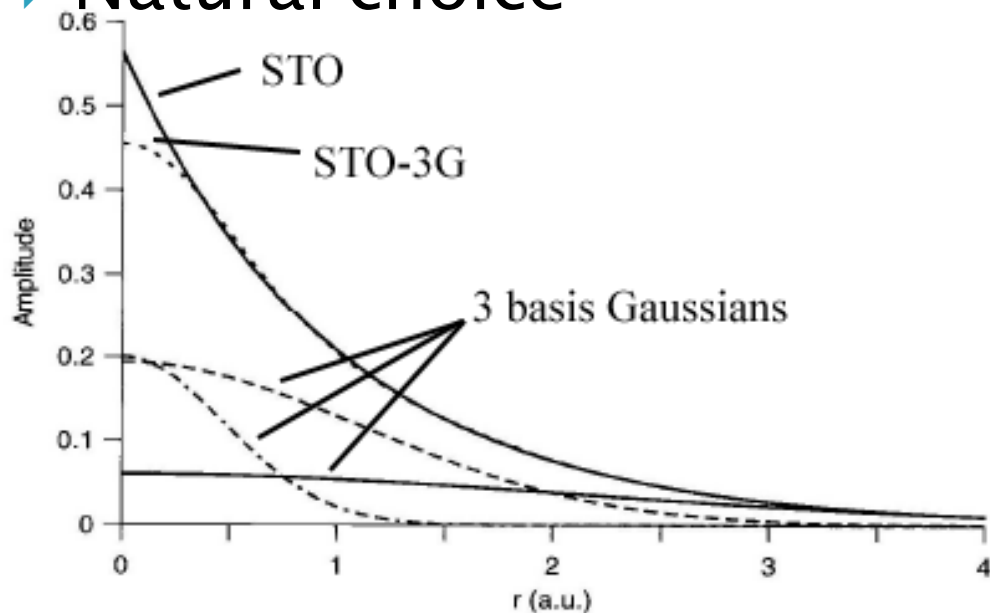 A *large* α means the electron is held *tight* to the nucleus.
  A *small* α means the electron can *roam far* from the nucleus.

STO vs. GTO

STO
- Correct cusp at r -> 0
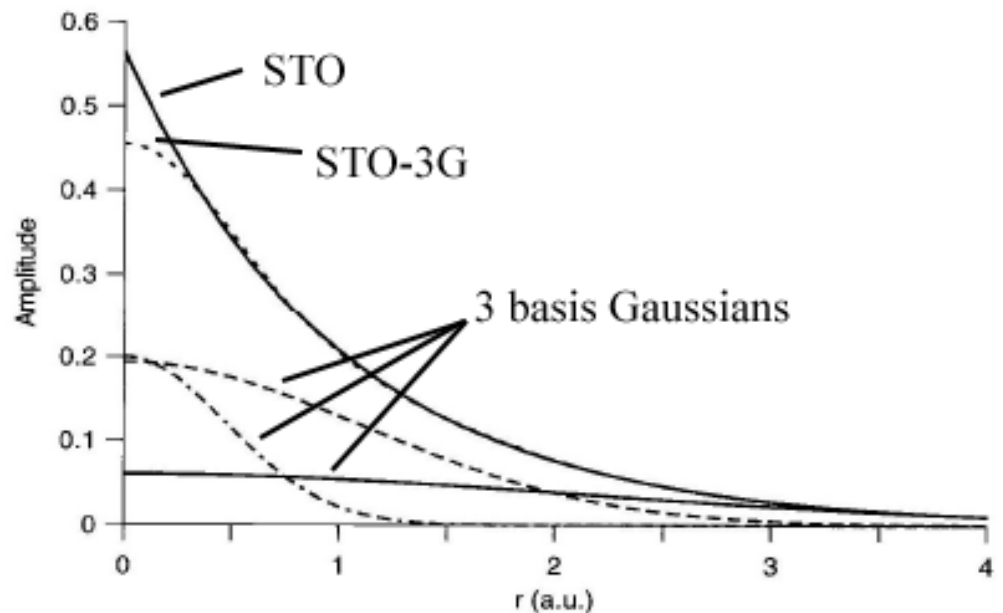- Desired decay r -> ∞

- Correctly mimics H orbitals
- Natural choice

GTO
- Wrong
- Wrong

- Gaussian x gaussian =gaussian
- Mathematically & computationally easier

Contracted Gaussian Functions (CFG)

- Product of 2 gaussians is another gaussian
- Integration is easy (versus STO's)

- Loss of accuracy BUT increase the number of gaussians used
- STO-nG, where n = # GTOs used

$$\eta_{\tau}^{CGF} = \sum_{1}^{n} d_{n\tau}\eta^{GTO}$$

```
-C
S   3 1.00
 0.1722560000D+03  0.6176690000D-01
 0.2591090000D+02  0.3587940000D+00
 0.5533350000D+01  0.7007130000D+00
SP  2 1.00
 0.3664980000D+01 -0.3958970000D+00  0.2364600000D+00
 0.7705450000D+00  0.1215840000D+01  0.8606190000D+00
SP  1 1.00
 0.1958570000D+00  0.1000000000D+01  0.1000000000D+01
        α                    d_s                  d_p
```

3-21G

$$\varphi_i = \sum_{n=1}^{N} d_{ni}\eta_n \qquad \eta^{GTO}(\mathrm{r}) = x^l y^m z^n e^{-\alpha r^2} = \varphi^{GF}$$

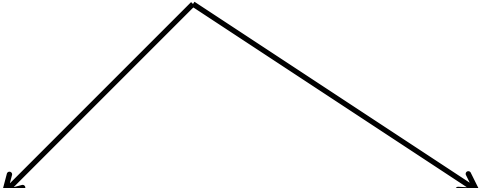$$\varphi_{2s}(r) = \sum_{i=1}^{3} d_{2s}\varphi_{2s}^{GF}(r,\alpha_{2s}) + d_{2s}'\varphi_{2s}^{GF}(r,\alpha_{2s}')$$

Sum of x G's
w/ c_i's

$$\varphi_{2s}(r) = -0.395\varphi_{2s}^{GF}(r,3.664) + 1.215\varphi_{2s}^{GF}(r,0.771)$$

$$+1.00\varphi_{2s}^{GF}(r,0.195)$$

Extended Basis Set: Split Valence

- Minimal basis set: STO-3G
  - A single CGF for each atomic orbital
- Double-Zeta (DZ) Basis Sets
  - Mulitple basis functions corresponding to each atomic orbital = greater flexibility
  - Inert core orbitals: single CGF
  - Valence orbitals: double set of CGF's

$$\varphi_{2s}(\mathrm{r}) = d\varphi_{2s}^{STO}(r, \zeta_1) + d\varphi_{2s}^{STO}(r, \zeta_2)$$

Extended Basis Sets (cont.)

3-21G

- Core = 1s orbital, a sum of 3 gaussians
- Each valence orbital is split into 2 parts
- Valence = 2s & 2p orbitals:
  - Inner part a sum of 2 gaussians
  - Outer part is 1 gaussian

Triple-Zeta (TZ)
 - Inner = 1 CGF
  - Valence = 3 CGF's
  - 6-311G

Additional Basis Functions

- Polarization Functions:
  - Distortion of molecular orbitals due to the molecular environment
  - Functions with a higher angular momentum than any occupied atomic orbital
    - p-functions for H-He **
    - d-functions for Li-Ca *
    - f-functions for transition metals
    - 6-31G(d,p) = 6-31G** = polarization functions added to elements H-Ca

Additional Basis Functions

- Diffuse Functions:
  - Electrons are bound looser in anions, neutral molecules w/ unshared pairs, zwitterions
  - Wave function tail is important

  - Remember,
  - $\phi = Constants \times exp[-\alpha r^2]$
  - For longer tails, this means
  - $0.1 > \alpha > 0.01$
  - Large radial extent

  - 6-31 ' 'G (*'s here for polar.)
  - + diffuse functions on first row
  - ++ diffuse function on H
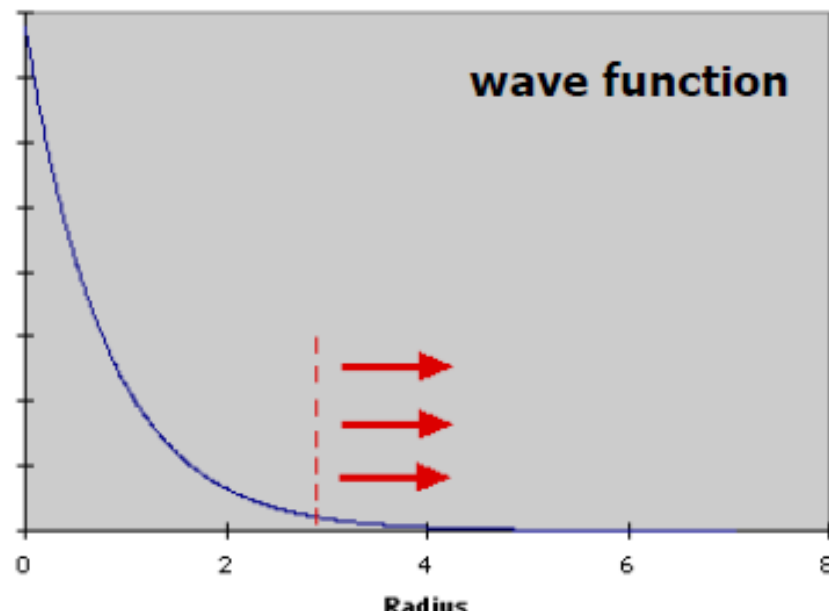
wave function

Other Basis Sets

These are "*about*" equivalent

Pople split valence	Pople valence triple zeta	Dunning correlation consistent	
6–31G	6–311G	cc–pvdz	aug–cc–pvdz
6–31+G*	6–311+G*	cc–pvtz	aug–cc–pvtz
6–31++G**	6–311++G**	cc–pvqz	aug–cc–pvqz

Diffuse functions
Long–distance interactions
anions

Polarization functions
Angular charge distribution
flexibility

Levels of theory: Hartree–Fock

- **The variational principle.**
  - If Ψ is the true wavefunction, then for any model antisymmetric wavefunction Φ
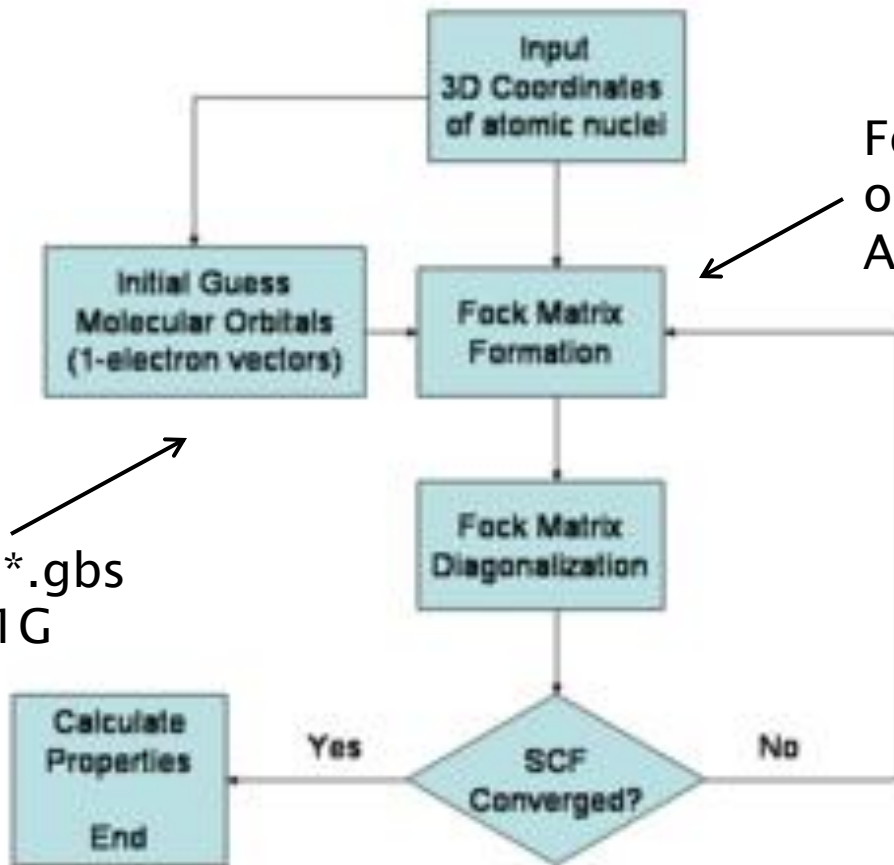 $$E(\Phi^{model}) > \text{true } E(\Psi)$$

  - To find the HF w.f.:
    - Minimize the energy w.r.t. changes in the orbitals
      - $\phi_i \rightarrow \phi_i + \delta\phi_i$

  - Adjustable parameters, the c_{ni}'s , the orbital coeff's

- E_{HF} forms an upper bound, $E_{exact} - E_{HF} = E_{correlation}$
  - HF neglects electron correlation

Simplistic & typical routine
Self-consistent field method

Self-consistent = Iterative, step by step

Next section

Input 3D Coordinates of atomic nuclei

Initial Guess Molecular Orbitals (1-electron vectors)

Fock Matrix Formation

Fock Matrix Diagonalization

Calculate Properties — Yes — **SCF Converged?** — No

End

Formulation orbital combo's Adjust -311

Field (a.k.a. mean field) where each particle is subjected to a mean field (average) created by all the other particles

From *.gbs 6-311G

Molecular Geometry Optimization

▸ Hessian matrix is the square matrix of second-order partial derivatives of a function; that is, it describes the local curvature of a function of many variables.

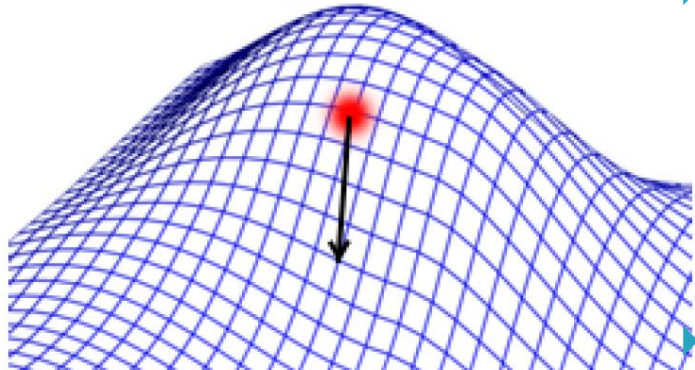$$H_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j}$$

All H eigenvalues positive = Minimum
All H eigenvalues negative = Maximum
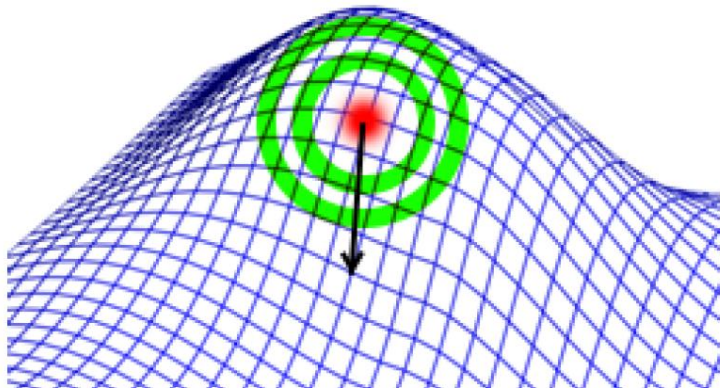Mix of pos. & peg. eigenvalues = Saddle
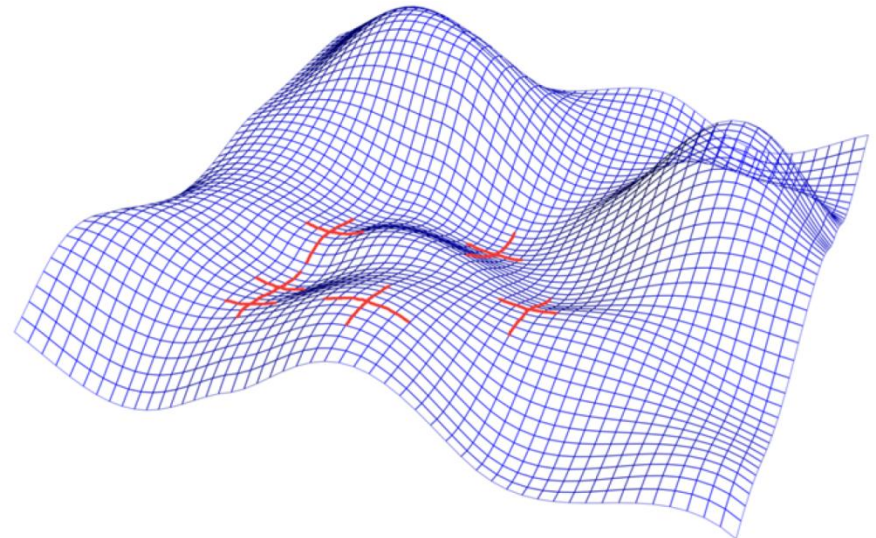
Molecular Geometry Optimization Algorithms
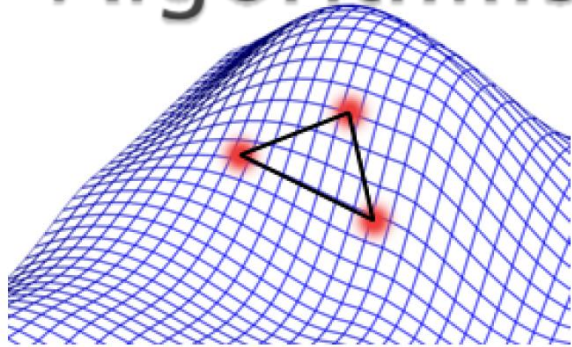
- Energy only
  - Simplest
    - Might not correspond to the true minimum

- Gradient following (1st order)
  - Line search, backtracking, gradient descent (steepest descent)
  - Less robust, gradient evaluation at every iteration step

- Higher order
  - Newton-Raphson, Berny, RFO, GDIIS, ODIIS
  - Hessians & higher order derivations at every step
  - Powerful, more time consuming, VERY dependent on initial configuration /guess

Higher orders of theory Configuration Interaction

- HF theory is limited by its neglect of electron–electron correlation.
  - Artificially, electrons interact with a SCF, not individual e's.
  - Most higher ordered levels of theory begin with HF formulation of the w.f.
- "Full CI" includes the Hartree–Fock ground–state determinant and all possible variations.

$$\Psi = a_0 \Psi_0 + \sum_{s=1}^{\infty} a_s \Psi_s$$

- The wavefunction becomes where s includes all combinations of substituting electrons into HF *virtual orbitals*.
  - CISD(T) = Config Inter. of singlets, doublets; triplets combinations
  - The a_s's are optimized; not so practical if s goes to infinity!
- Partial CI's are useful, and takes less time… ~4 months?

Levels of Theory

Semi-empirical methods
MNDO,AM1,PM3

Hartree-Fock
HF-SCF

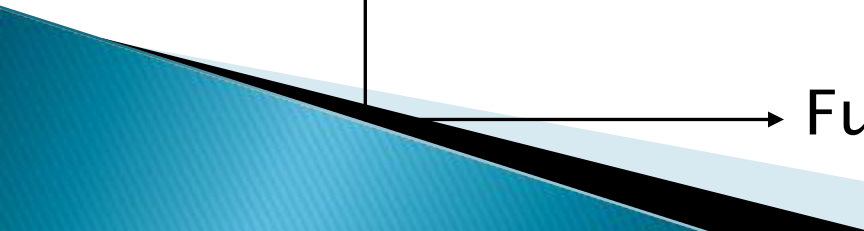Excitation Hierarchy
CIS,CISD,CISDT
CCS,CCSD,CCSDT

Perturbation Hierarchy
MP2,MP3,MP4

Multiconfigurational HF
MCSCF,CASSCF

Multireference Perturbation
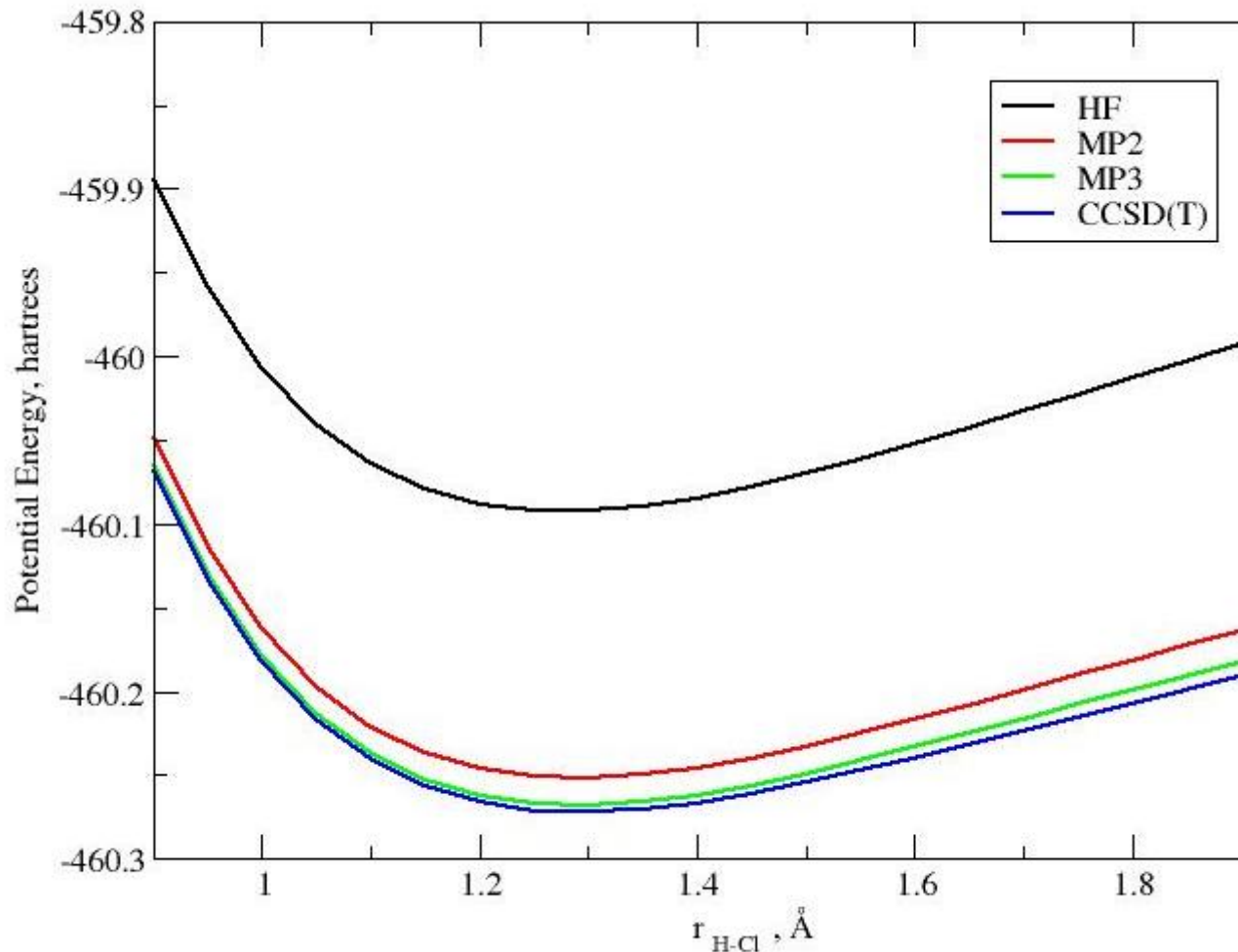CASPT2, CASPT3

Full CI

Pople's Diagram

Level of Theory
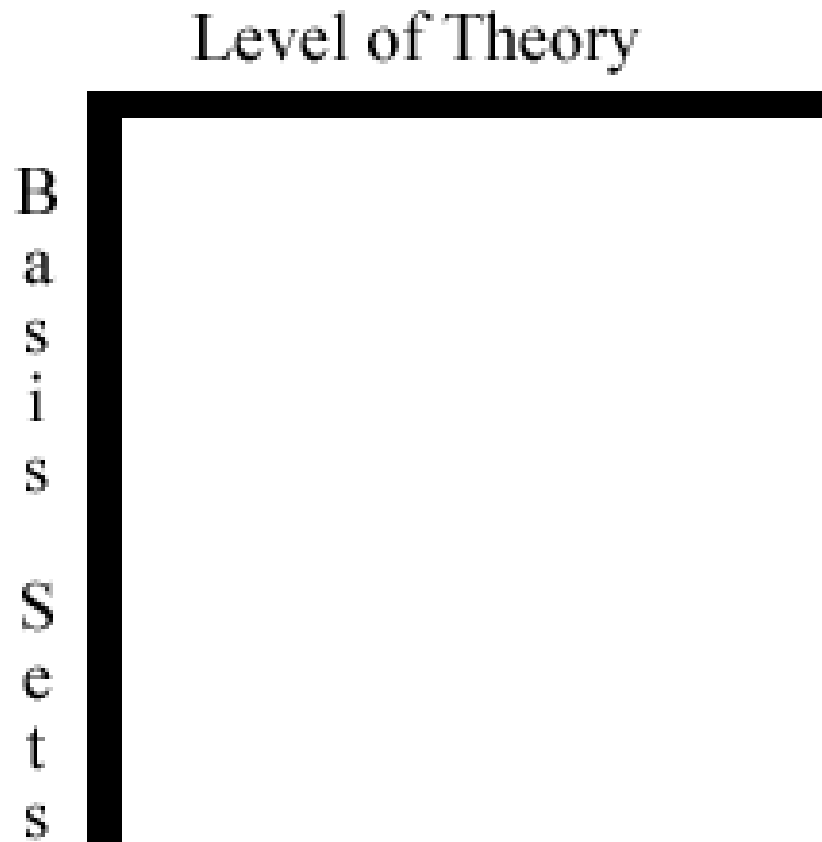
Basis Sets

HCl Potential Energy Scan, comparison of levels of theory
Basis Set: aug-cc-pvdz

$\Delta E = CCSD(T) - xx$
(Kcal /mol)

HF: 112.931
No é corr. ! !

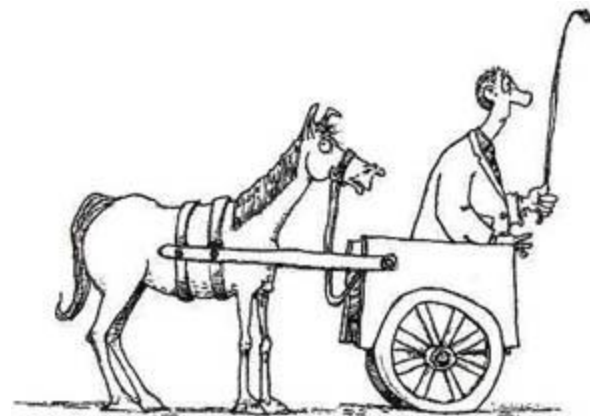MP2:   12.881

MP3:   2.672

1 a.u. =
1 hartree =
627.5095
Kcal/mol

Gaussian runtime

- Understand how the input file is constructed

- Specifying atomic coordinates

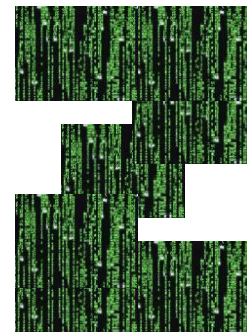- How to submit your job to a HPC

Defining Atomic Coordinates

- ▶ Cartesian Coordinates
  - ◦ René Descartes
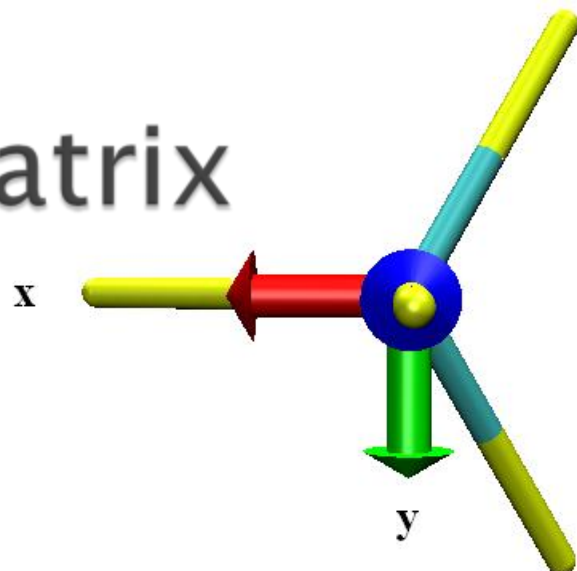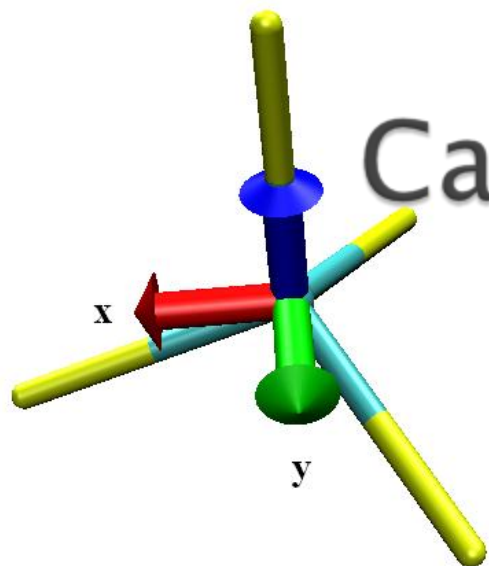 - ◦ *Atom x y z*

- ▶ Z matrix
  - ◦ Internal Coordinate System
  - ◦ *newatom*
    newatom atom1 bond-length
    newatom atom1 bond-length atom2 bondangle
    newatom atom1 bond-length atom2 bondangle atom3 dihedral-angle

Cartesian vs. Z-Matrix

```
5

C 0.000 0.000 0.000
H 0.000 0.000 1.089
H 1.027 0.000 −0.363
H −0.513 −0.889 −0.363
H −0.513 0.889 −0.363
```

```
C
H 1 1.089
H 1 1.089 2 109.47
H 1 1.089 2 109.47 3 120.000
H 1 1.089 2 109.47 3 −120.000
```

This atom is connected to atom1 having a bond-length *and* with atom2 has a bondangle *and* with atom3 has a dihedral angle

Gaussian Job Details

%chk=water.chk	Checkpt file	%chk=water.chk
%NProcLinda=2	# nodes	%NProcLinda=2
%NProcShared=4	#ppn	%NProcShared=4
# mp2/6-31g* opt=tight freq	Job description	# mp2/6-31g* opt=tight freq
<blank line>		<blank line>
Title Line	Title	Title Line
<blank line>		<blank line>
0 1	Charge Multi.	0 1
O 0.0 0.0 0.0	Coordinates	O
H 0.0 1.0 0.0	…	H 1 1.0
H 0.0 0.0 1.1	…	H 1 1.0 2 104.5
<blank line>	Blank line !	<blank line>

Gaussian Job Details

```
%chk=water.chk
%NProcLinda=2
%NProcShared=4
# mp2/6-31g* opt=tight freq
<blank line>
Title Line
<blank line>
0 1
O
H 1 1.0
H 1 1.0 2 104.5
<blank line>
```

```
%chk=water.chk
%NProcLinda=2
%NProcShared=4
# mp2/6-31g* opt=tight freq
<blank line>
Title Line
<blank line>
0 1
O
H 1 roh1
H 1 roh1 2 ahoh
<blank line>
roh1 1.0
ahoh 104.5
```

Variables section
Defining used variables

If defined by calculation type: bond stretches, isotopes, continue (serial job processing)

Gaussian on LSU HPC / LONI

- Add "+gaussian-03" to your .soft file
- Resoft

- On IBM P5 systems
- Command is
  - runG03 inputfile nprocs queue
    - runG03 input 8 MP5L
  - (LL) 'llclass' for queue types

Gaussian on Dell x86 maxchines

Add the softenv key & resoft
Sample submission script, 'qsub subscript'

```
#!/bin/tcsh
#PBS –A your_allocation      # specify the allocation. Change it to your allocation
#PBS –q checkpt              # the queue to be used.
#PBS –2 nodes=1:ppn=4       # Number of nodes and processors
#PBS –l walltime=1:00:00    # requested Wall-clock time.
#PBS –o g03_output          # name of the standard out file to be "output-file".
#PBS –j oe                  # standard error output merge to the standard output file.
#PBS –N g03test             # name of the job (that will appear on executing the qstat command).
set NPROCS= `wc –l $PBS_NODEFILE |gawk '//{print $1}'`
setenv GAUSS_SCRDIR /scratch/$USER
set NODELIST = ( –vv –nodelist "" `cat $PBS_NODEFILE` "" –mp 4)
setenv GAUSS_LFLAGS " $NODELIST "
# cd to the directory with Your input file
cd ~USERNAME/g03test
# Change this line to reflect your input file and output file
g03 < test.inp > g03job.out
```

Fighting the Queue

- **Not all simulations are equal!**
- Benchmark your simulation
  - Walltime vs. # processors
  - Walltime vs. # nodes (w/ half procs idle)
  - Find the optimum configuration
  - Decrease walltime to 1 hour
- Walltime slightly greater than time needed
- THEN: Tailor your submission script to reflect YOUR needed walltime & # nodes
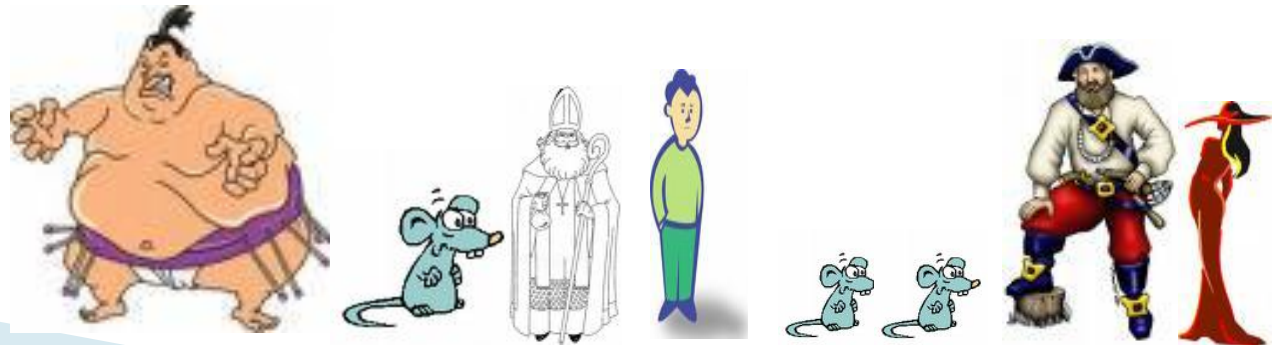
Fighting the Queue

- Check & monitor the load on *other* HPCs
- NOTE: Gaussian license is site specific
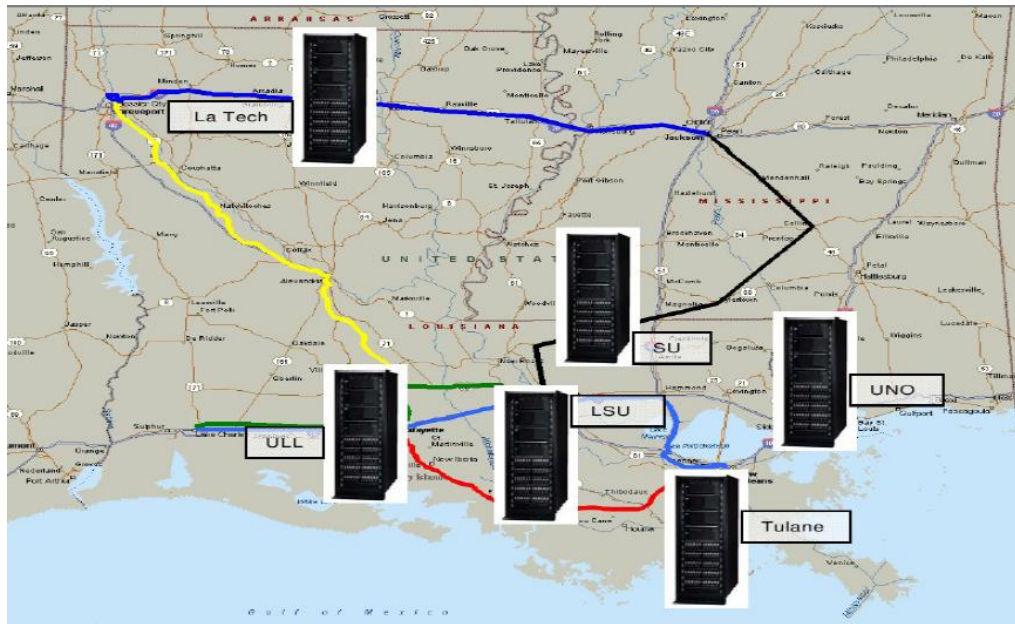
- Researchers tend to mainly work with a few clusters and submit different job types/sizes

Analysis– some pointers

Level of Theory

Theory	Basis Set	Ground State (hartrees)	Bond Length (Angstroms)	% Error
HF	3-21G*	-457.981	1.2671	-0.58%
HF	6-31G*	-460.06	1.2662	-0.65%
HF	6-311G*	-460.087	1.2713	-0.26%
HF	cc-pvDZ	-460.09	1.2776	0.24%
HF	cc-pvTZ	-460.107	1.2672	-0.58%
HF	aug-cc-pvDZ	-460.093	1.2767	0.17%
HF	aug-cc-pvTZ	-460.108	1.2675	-0.55%
B3LYP	3-21G*	-458.709	1.2888	1.12%
B3LYP	6-31G*	-460.796	1.2895	1.18%
mp2	6-311G**	-460.244	1.2734	0.09%
CCISD	3-21G*	-458.145	1.2841	0.75%
CCISD	6-31G*	-460.218	1.2854	0.85%
CCISD	cc-pVTZ	-460.371	1.2749	0.03%
Experiment			1.2746	

Geometry optimization of HCl

Is CCISD/cc–pVTZ the best?

Is bond length a good judgment of accuracy of the model?

References

- http://www.gaussian.com
- http://www.gaussian.com/g_ur/keywords.htm
  - Gaussview

- http://www.cmbi.ru.nl/molden/molden.html

- http://www.cscs.ch/molekel/