TOWARDS *ACCURATE* BIOMASS CALCULATIONS

Mark S. Gordon

Department of Chemistry

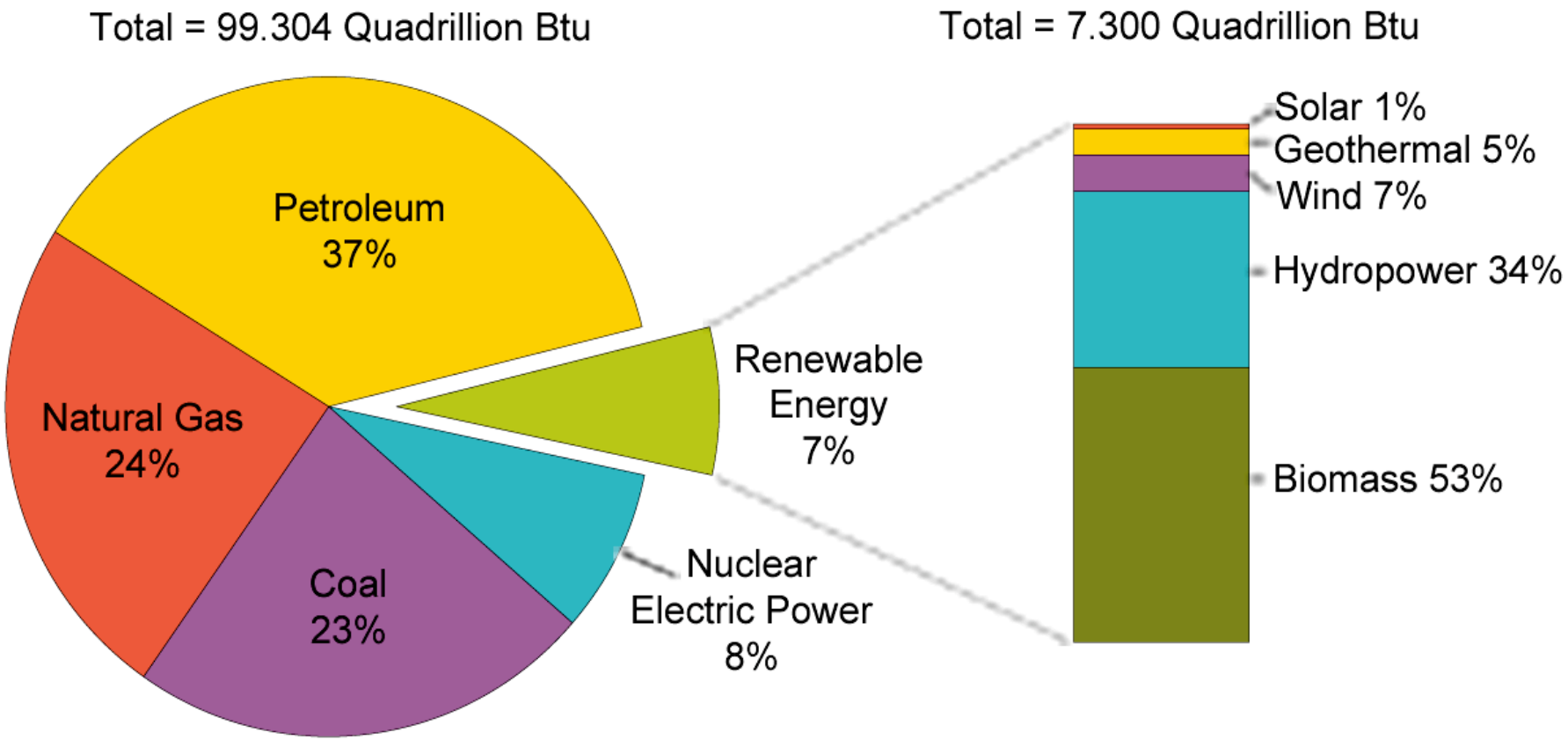
Iowa State University and Ames Laboratory

COLLABORATORS

- Theresa Windus, Yana Kholod, Ames Lab
- Monica Lamm, ISU Chemical engineering
- Jeremy Smith, ORNL

MOTIVATION

Renewable Energy in the Nation's Energy Supply



Note: Sum of components may not equal 100% due to independent rounding.
Source: U.S. Energy Information Administration, *Annual Energy Review 2008*, Table 1.3, Primary Energy Consumption by Energy Source, 1949-2008 (June 2009).

Ethanol Fuel

Nebraska,
1933

Ethanol Fuel Energy Balance*

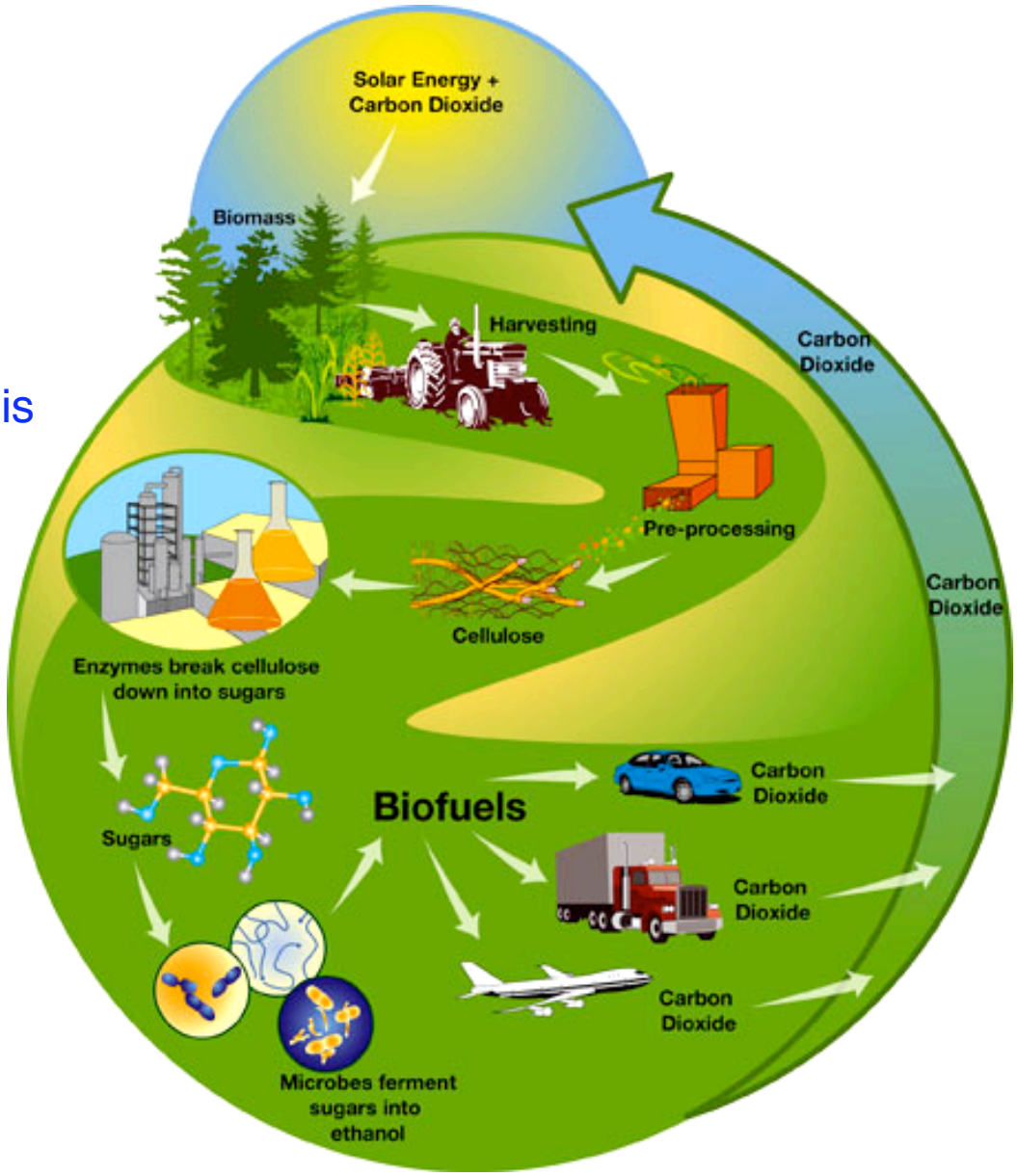
vs

Type	Energy balance
Corn ethanol	1.3
Sugarcane ethanol	8
Biodiesel	2.5
Cellulosic ethanol	2–36

***energy gain/energy drain**

Green Dreams J.K. Bourne JR, R. Clark National Geographic Magazine October 2007 p. 41

1) Hot water pretreatment
2) Enzymatic or acidic hydrolysis



Cellulosic Biofuels
via: http://www.jgi.doe.gov

QM ISSUES TO ADDRESS

- **Accurate** QM calculations typically limited to molecules of moderate size
  - Tens of atoms
- Large molecular systems usually require simpler methods
  - Modest ab initio calculations (HF, DFT)
  - Semi-empirical methods
  - Molecular mechanics (QM/MM)
- Goal: Accurate calculations on large systems

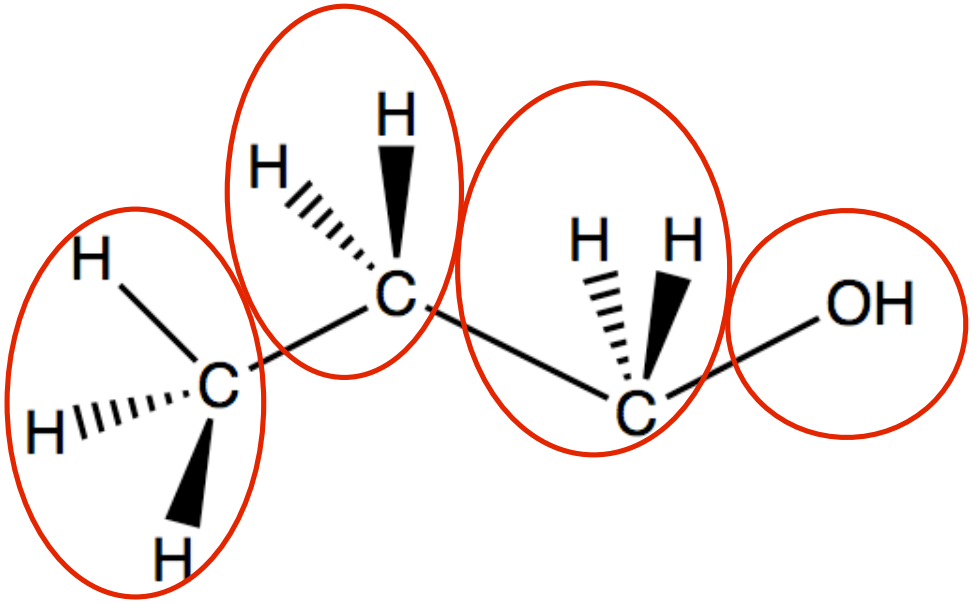
CHALLENGES

- High cost of electronic structure calculations
  - HF, hybrid DFT scale ~N^4
    - Up to 1,000 atoms
    - Analytic gradients, Hessians available
  - MP2 scales ~N^5 (N = # basis functions)
    - ~100 atoms
    - Analytic gradients, Hessians available
  - CCSD(T) & related methods scale ~N^7
    - 2^7=128
    - 10-20 atoms
    - Analytic gradients available
- What are the possible solutions?

FRAGMENTATION METHODS

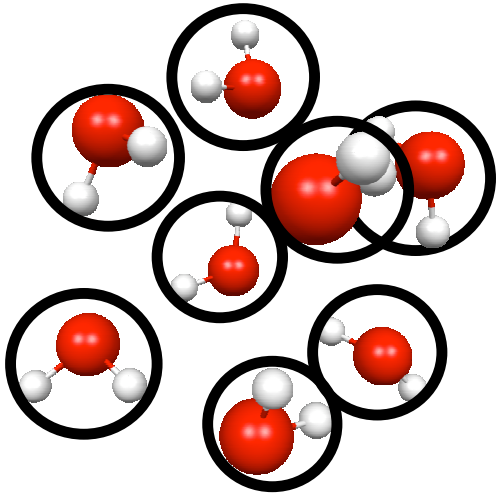
- Fragment Molecular Orbital (FMO) method
  - Fedorov & Kitaura: CPL, 313, 701 (1999)
  - Divides large molecular system into smaller fragments
    - "E Pluribus Unum" approach **["Out of Many, One"]**
    - Accounts for the influence of the entire system during each individual fragment calculation (**not** divide & conquer)
  - Inherently parallel: each fragment calculated on separate node
    - Can do tens of thousands of atoms
  - Typically kcal/mol accuracy relative to full calculation
  - Implemented for
    - HF, DFT, TDDFT, MP2, MCSCF (energies + gradients)
    - CCSD(T) energies
  - Developed for study of proteins
  - Ideal for liquids (including ionic liquids)
    - No need to cut covalent bonds
  - Interfaced with solvent methods (e.g., PCM, EFP)

FMO FRAGMENTATION



For covalently bound molecules, the fragment is divided into pieces so as not to destroy bond electron pairs.

In clusters (e.g.,water), fragmentation is easier: no covalent bond breaking.



BASIC FMO METHODOLOGY

1. Divide molecule into fragments and assign electrons to these fragments

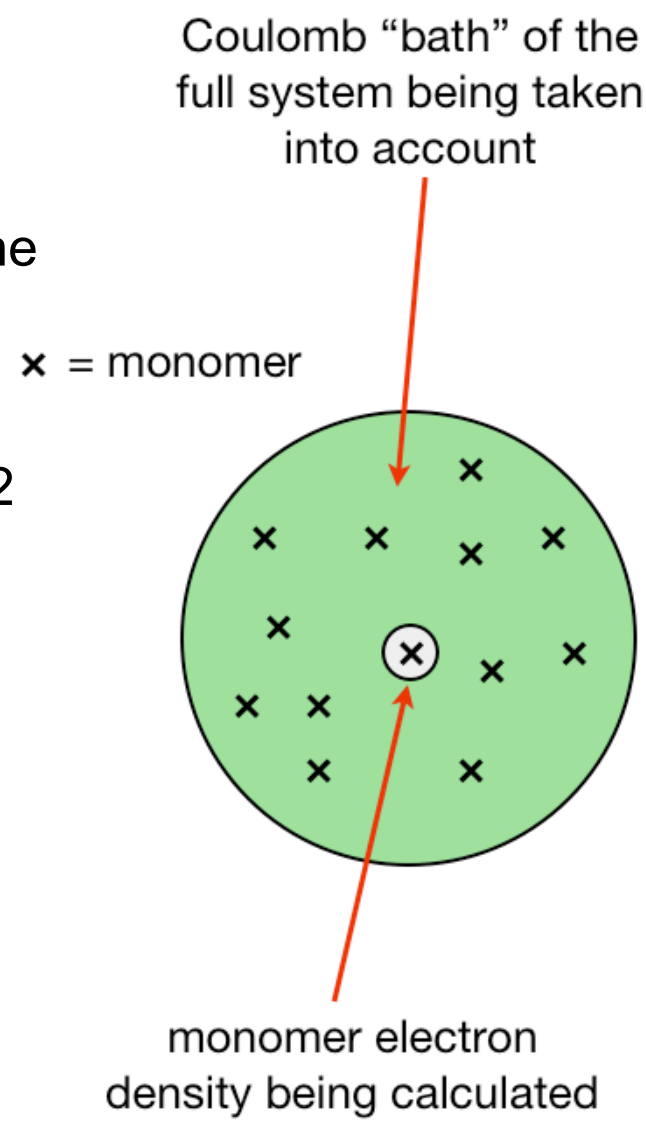
2. Calculate initial electron density distribution of the fragments in the Coulomb “bath” of the full system

3. Construct the individual fragment Fock operators using the densities calculated in 2 and solve for the fragment energies

4. Determine if the density has converged for all the fragments. If not, go back to step 3

5. Construct Hamiltonians for each dimer (trimer) calculation using the converged monomer densities from steps 3-4

6. Calculate total energy and electron density



BASIC FMO METHODOLOGY

The total energy of the system can be written as

$$E = \sum_{I}^{N} E_I + \sum_{I>J}^{N} (E_{IJ} - E_I - E_J)$$

FMO2

$$+ \sum_{I>J>K}^{N} \{(E_{IJK} - E_I - E_J - E_K) - (E_{IJ} - E_I - E_J) - (E_{JK} - E_J - E_K) - (E_{KI} - E_K - E_I)\} + \cdots,$$

FMO3

The monomer (I), dimer (IJ) and trimer (IJK) energies are obtained using the SCF method with modified Fock operators

BASIC FMO METHODOLOGY

•The Fock equation

$$\tilde{\mathbf{F}}^x \mathbf{C}^x = \mathbf{S}^x \mathbf{C}^x \tilde{\boldsymbol{\epsilon}}^x,$$

$$\tilde{\mathbf{F}}^x = \tilde{\mathbf{H}}^x + \mathbf{G}^x,$$

is modified from the standard form: The one-electron Hamiltonian $\tilde{\mathbf{H}}^x$ contains an additional term for the electrostatic potential (ESP) of the entire system and the projection operator used for fractioning of covalent bonds:

$$\tilde{H}^x_{\mu\nu} = H^x_{\mu\nu} + V^x_{\mu\nu} + B \sum_i \langle \mu | \varphi^h_i \rangle \langle \varphi^h_i | \nu \rangle,$$

$$\downarrow$$

Bath potential

BASIC FMO METHODOLOGY

$$\tilde{H}^x_{\mu\nu} = H^x_{\mu\nu} + \boxed{V^x_{\mu\nu}} + B\sum_i \langle\mu|\varphi^h_i\rangle\langle\varphi^h_i|\nu\rangle,$$

•The electrostatic potential contains nuclear attraction and 2-electron terms:

$$V^x_{\mu\nu} = \sum_{K(\neq x)} \left(u^K_{\mu\nu} + v^K_{\mu\nu}\right),$$

Nuclear attraction:

$$u^K_{\mu\nu} = \sum_{A\in K} \langle\mu|(-Z_A/|\mathbf{r}-\mathbf{r}_A|)|\nu\rangle,$$

Two-electron interaction

$$v^K_{\mu\nu} = \sum_{\lambda\sigma\in K} D^K_{\lambda\sigma}(\mu\nu|\lambda\sigma).$$

K is one of the monomers of the full system and each summation is over all monomers

BASIC FMO METHODOLOGY

$$E_x = \frac{1}{2} Tr\left\{ \mathbf{D}^x \left(\tilde{\mathbf{H}}^x + \tilde{\mathbf{F}}^x \right) \right\},$$

The energy monomer + dimer energy expression can now be used to calculate the total electronic energy of the molecule

$$\begin{aligned} E_{\text{mol}} &= \sum_{I>J} E_{IJ} - (N-2) \sum_{I} E_I \\ &= \sum_{I} E_I + \sum_{I>J} (E_{IJ} - E_I - E_J). \end{aligned}$$

FMO2

The two-electron integral evaluation for the environment is the most expensive part of the calculation.

This can be reduced by approximations. Two such approximations are implemented in FMO:

Mulliken atomic population approximation: intermediate range

Point charge approximation: long range

MULLIKEN APPROXIMATION INTERMEDIATE FIELD

For separations greater than some distance R_I > R_{ESPAP} the two-electron part of the ESP is rewritten as

$$v^K_{\mu\nu} = \sum_{\lambda\sigma\in K} D^K_{\lambda\sigma}(\mu\nu|\lambda\sigma). \quad \cong \sum_{\lambda\in K} (\mathbf{D}^K\mathbf{S}^K)_{\lambda\lambda}(\mu\nu|\lambda\lambda)$$

$\mathbf{D}^K\mathbf{S}^K$ are the orbital populations for fragment *K.* This approximation reduces the computational cost by a factor of N_{BF} (the number of basis functions).

POINT CHARGES: FAR FIELD

For far separated monomers, with separation greater than some distance R_{ESPPC}

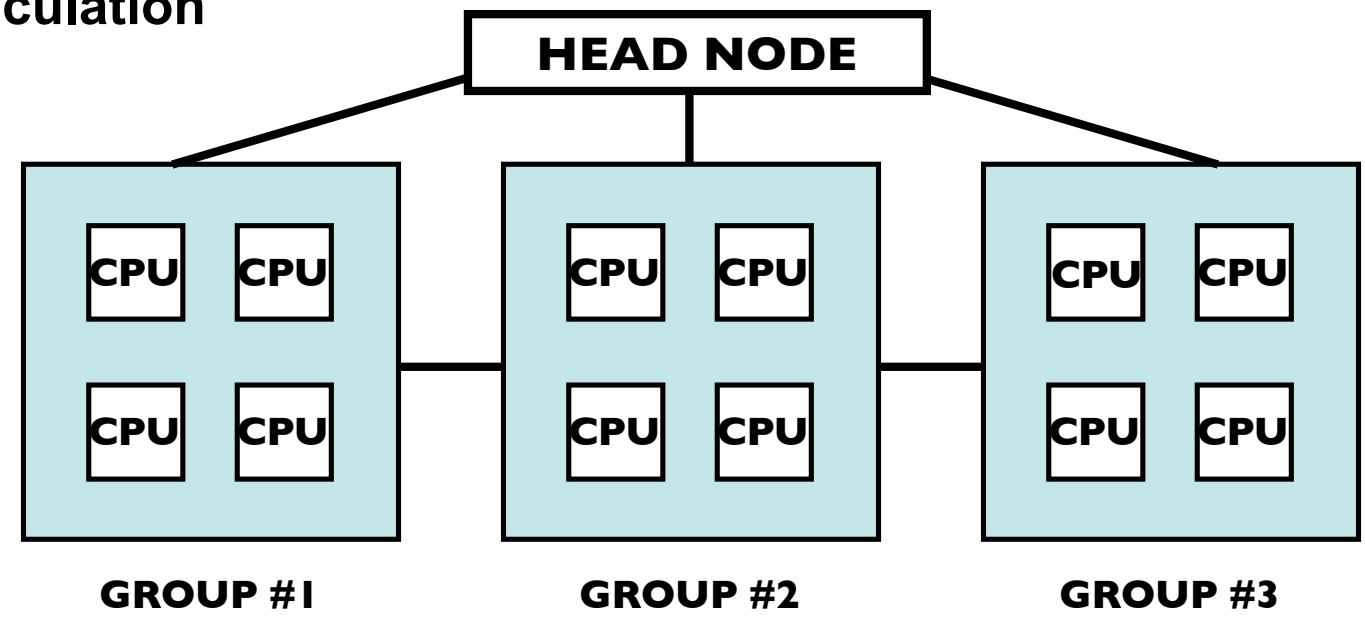
The ESP is rewritten as

$$v_{\mu\nu}^{K} = \sum_{\lambda\sigma \in K} D_{\lambda\sigma}^{K} (\mu\nu|\lambda\sigma). \quad \cong \sum_{A \in K} \langle \mu | (Q_A / |\mathbf{r} - \mathbf{r}_A|) | \nu \rangle$$

This approximation reduces the computational cost by another factor of N_{BF}

The Generalized Distributed Data Interface (GDDI)

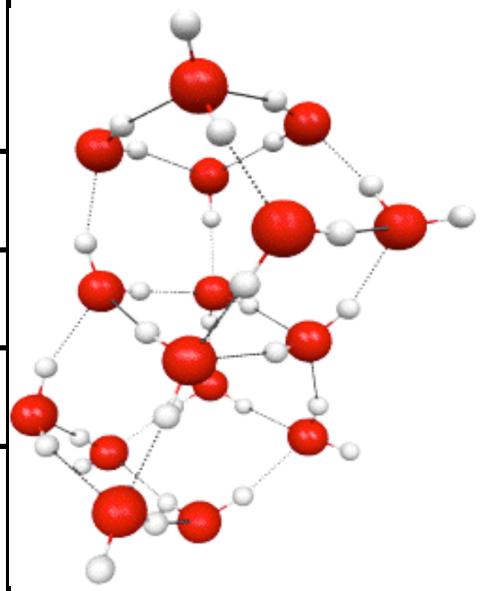
- GDDI allows for massively parallel calculations on clusters of computers or supercomputers
  - After the molecule is divided into fragments, each fragment is sent to a group which is composed of one or more processor or SMP enclosure
  - Each fragment is then run in parallel in each group
  - This provides two levels of parallelization, greatly speeding up the calculation



FMO Results

Timings using GDDI with 4 groups/2 processors per group

Method	CPU time* (min)	Error (kcal/mol)	CPU time* w/o GDDI (min)
MP2/6-31G(d)‡ *16 Waters*			
FMO2	1.3	3.3	2.6
FMO3	11.0	1.4	21.3
ab initio§	34.2	-	34.2
MP2/6-31++G(d,p)‡ *16 Waters*			
FMO2	3.3	2.9	6.3
FMO3	144.3	0.7	189.4
ab initio§	232.3	-	232.3



*CPU time error bars of ±5%

‡Calculations performed on 8 275MHz IBM Power 3 processors with 2 GB of RAM

§*ab initio* calculations performed without GDDI

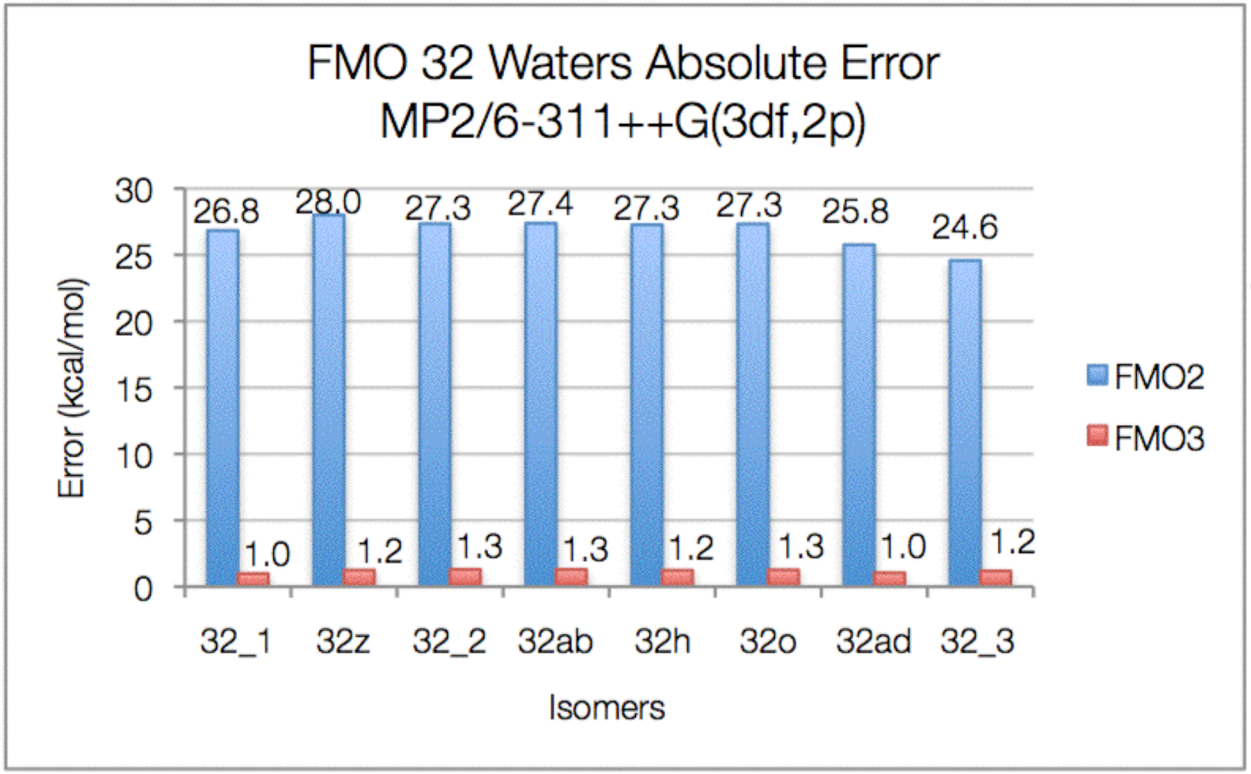
FMO Memory Requirements

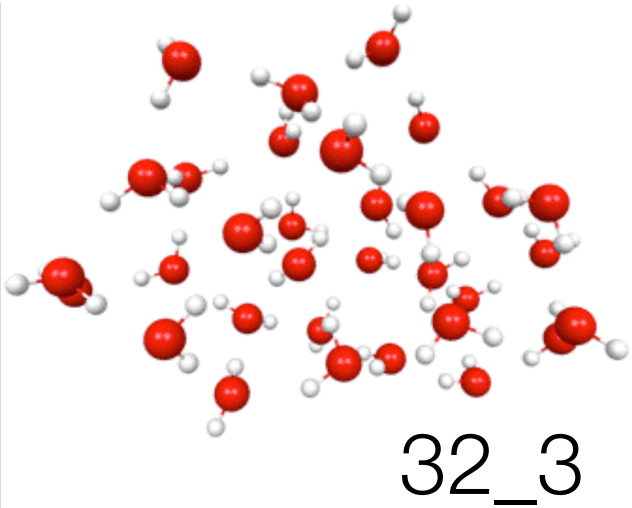
Method	Memory per Node*, MB	Memory per Node*, MB
CCSD(T) aug-cc-pVDZ§	6-Waters	32-Waters
FMO2	55	55
FMO3	166	166
ab initio	1714	1,263,762
CCSD(T) aug-cc-pVTZ§		
FMO2	547	547
FMO3	1774	1774
ab initio	17617	10,973,062

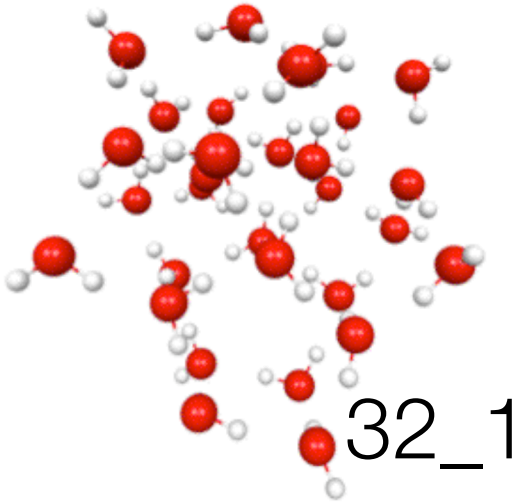
FMO memory requirement does *not* increase with system size

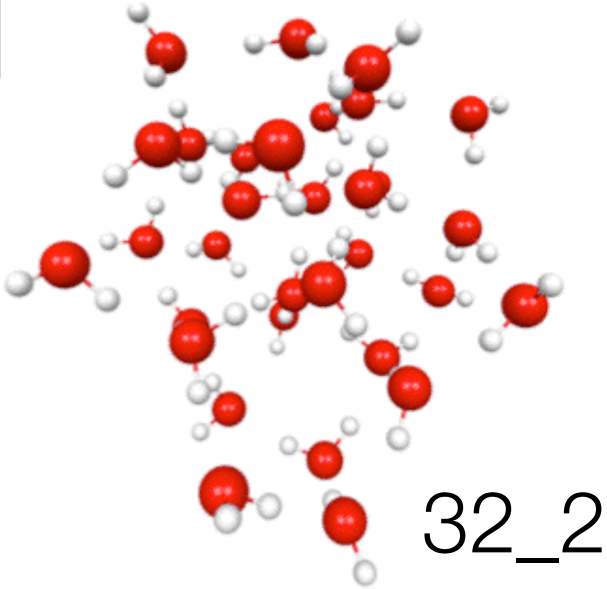
*Memory requirements for 16 processors, 4 CPUs per node

FMO For 32-Water Clusters

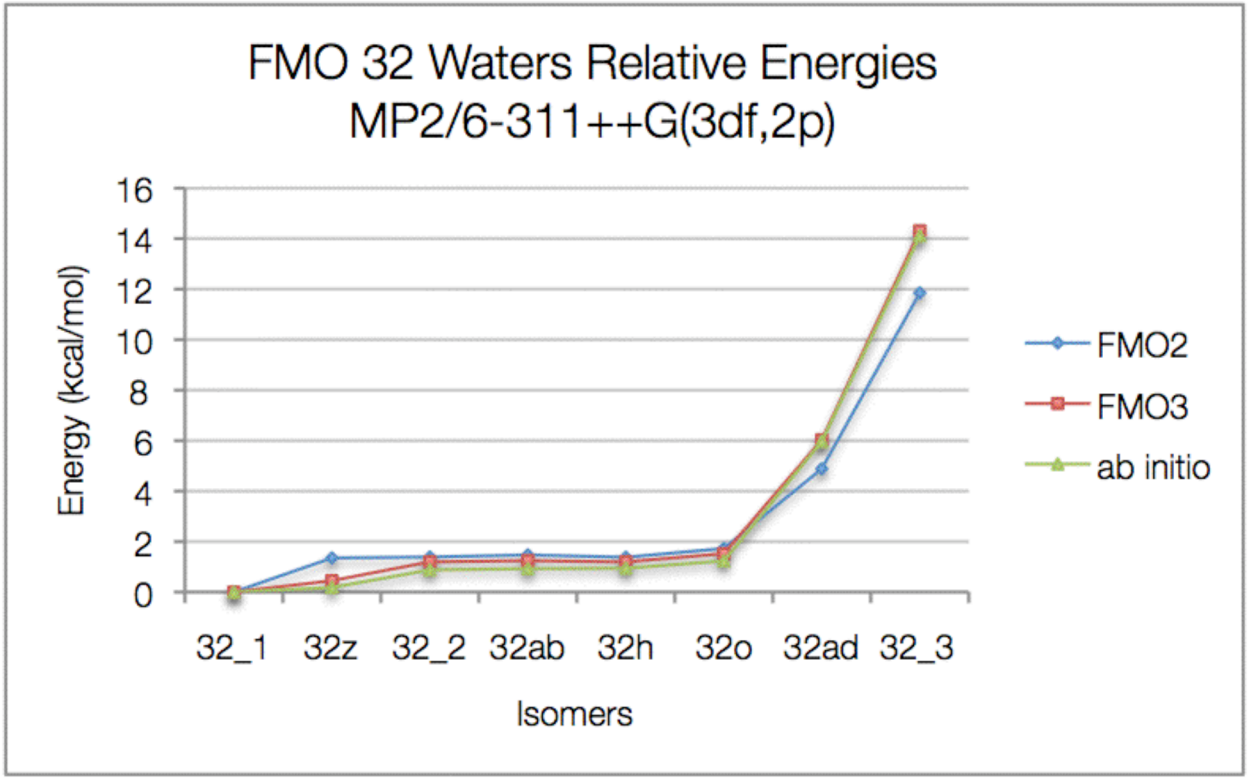


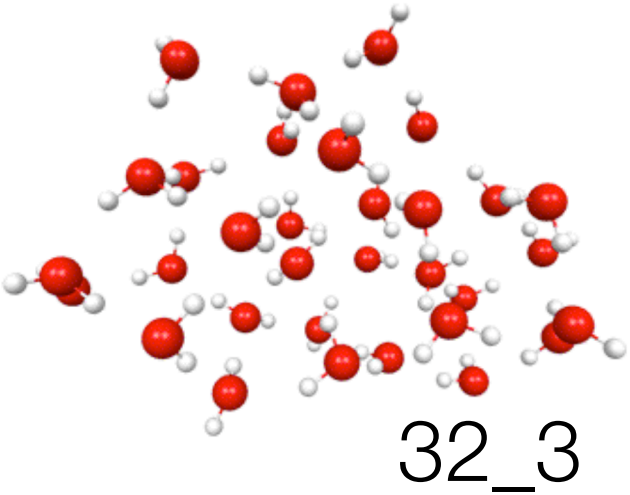


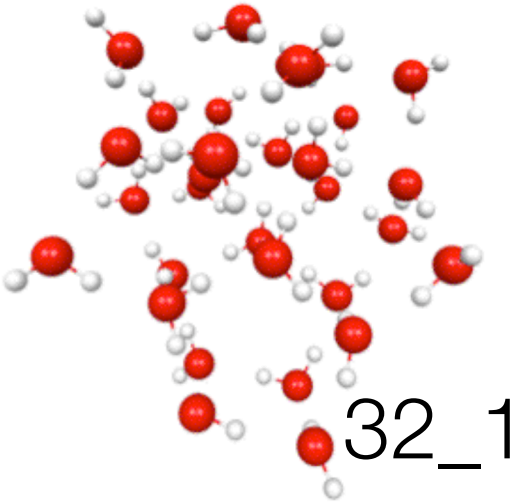


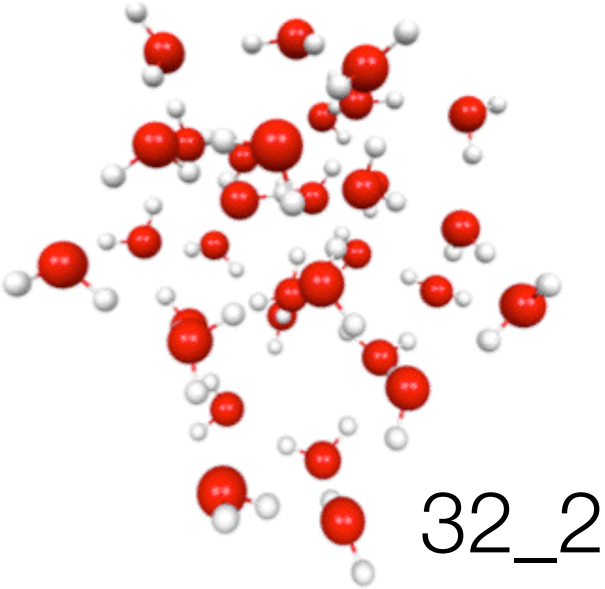


FMO For 32-Water Clusters

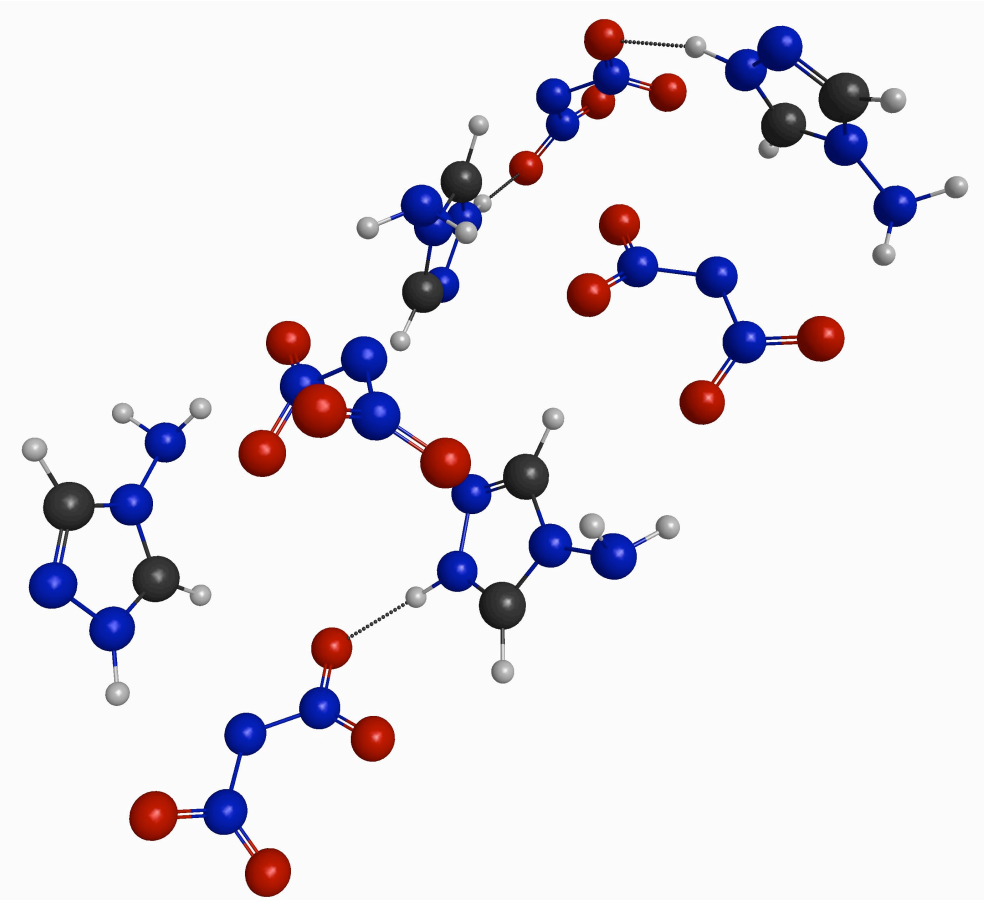




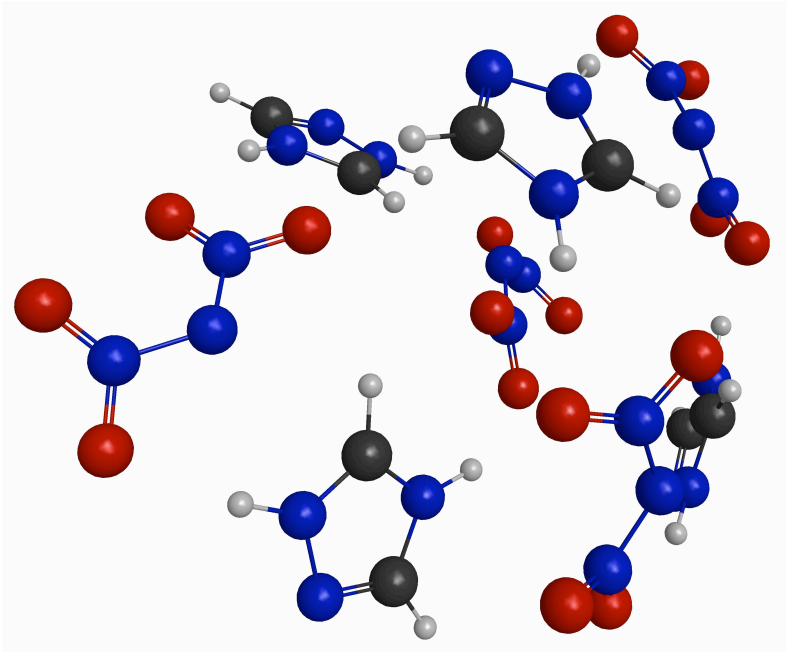




FMO for Ionic Liquid Octamers



FMO2-MP2/6-31++G(d,p)
1-amino,4-H-1,2,4-triazolium dinitramide
Error = 0.74 kcal/mol



FMO2-MP2/6-31++G(d,p)
1,4-H-1,2,4-triazolium dinitramide
Error = 0.67 kcal/mol

Water clusters/FMO1/HF/ACCD/BG/P: ANL (Graham Fletcher)

[INCITE grant, PI: Theresa Windus]

Wall time-to-solution (minutes)

		Numbers of Processors					
waters	w/f*	4096	8192	16384	32768	65536	**131072**
1024	4	14.6	7.1	4.2	3.9		
2048	8	---	81.9	41.0	22.1	17.4	11.1
4096**	16	---	---	328.0	168.0	98.8	52.1

***waters/fragment**

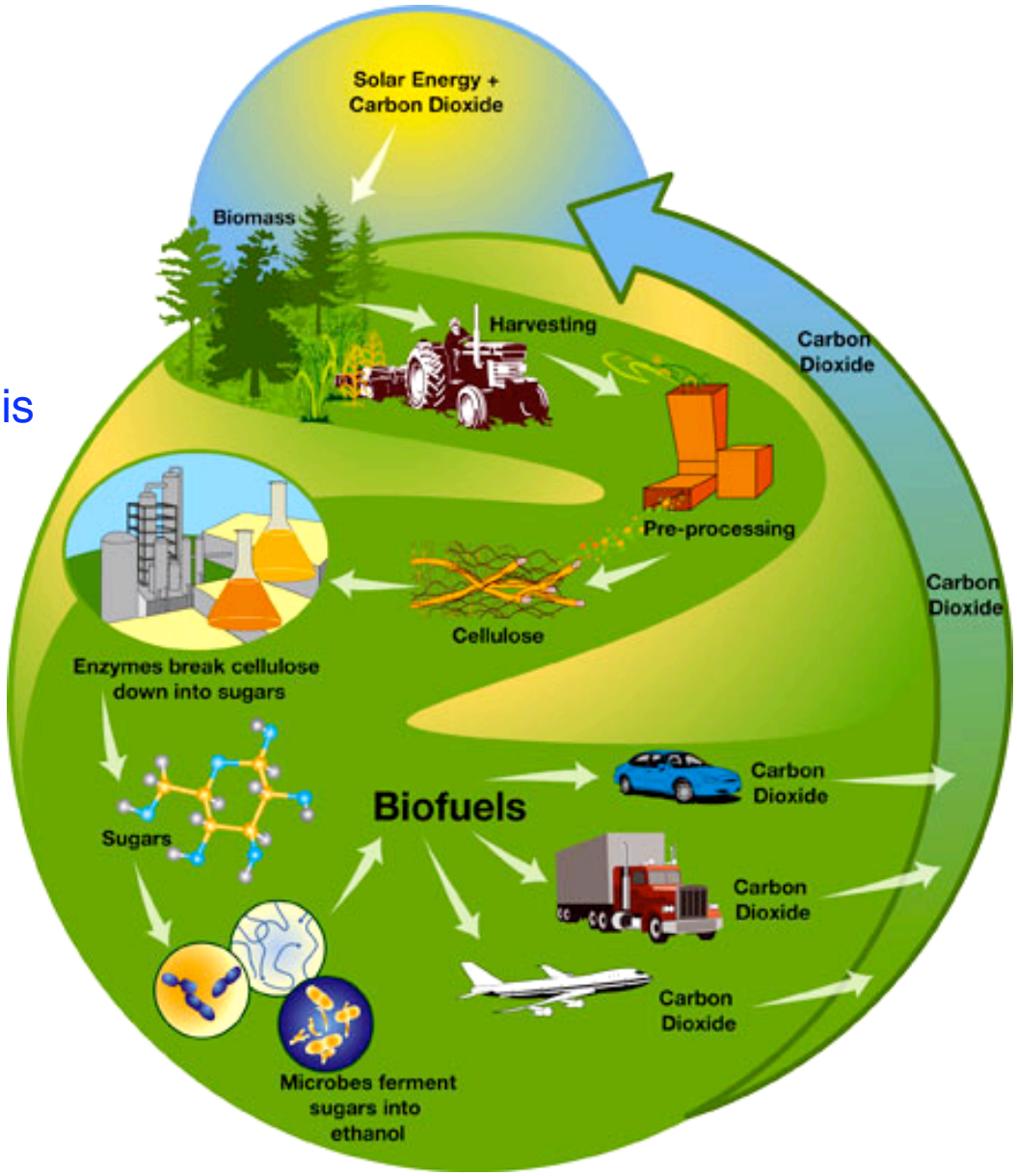
**** 170,000 basis functions!**

FMO2-MP2/aug-cc-pVDZ gradients

		Racks:	1	2	4	8	16
		Cores:	4096	8192	16,384	32,768	65,536
Waters	Atoms	Basis Functions	Wall-time-to-solution (minutes)				
128	384	5504	9.7	5.5	3.1		
256	768	11,008		11.6	6.6	4.1	
512	1536	22,016			16.8	10.1	7.4
1024	3072	44,032				24.1	16.8

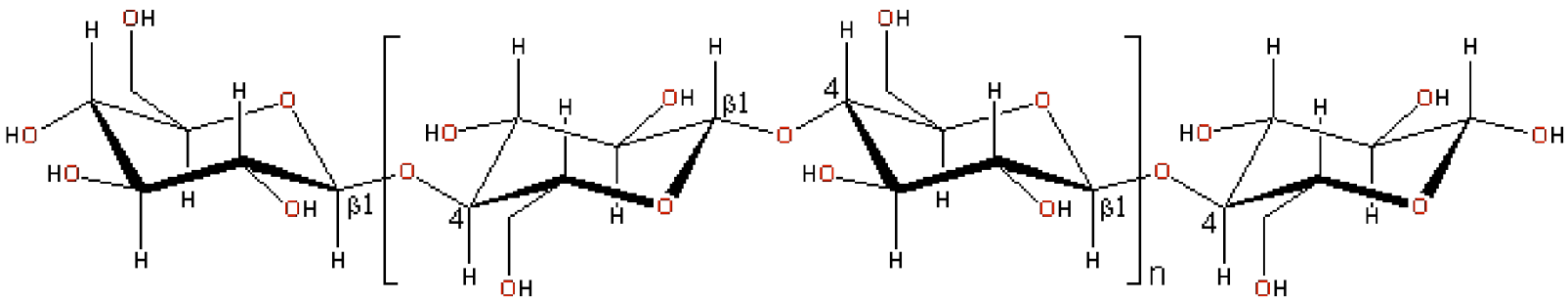
Scaling with system size =1.2

1) Hot water pretreatment
2) Enzymatic or acidic hydrolysis

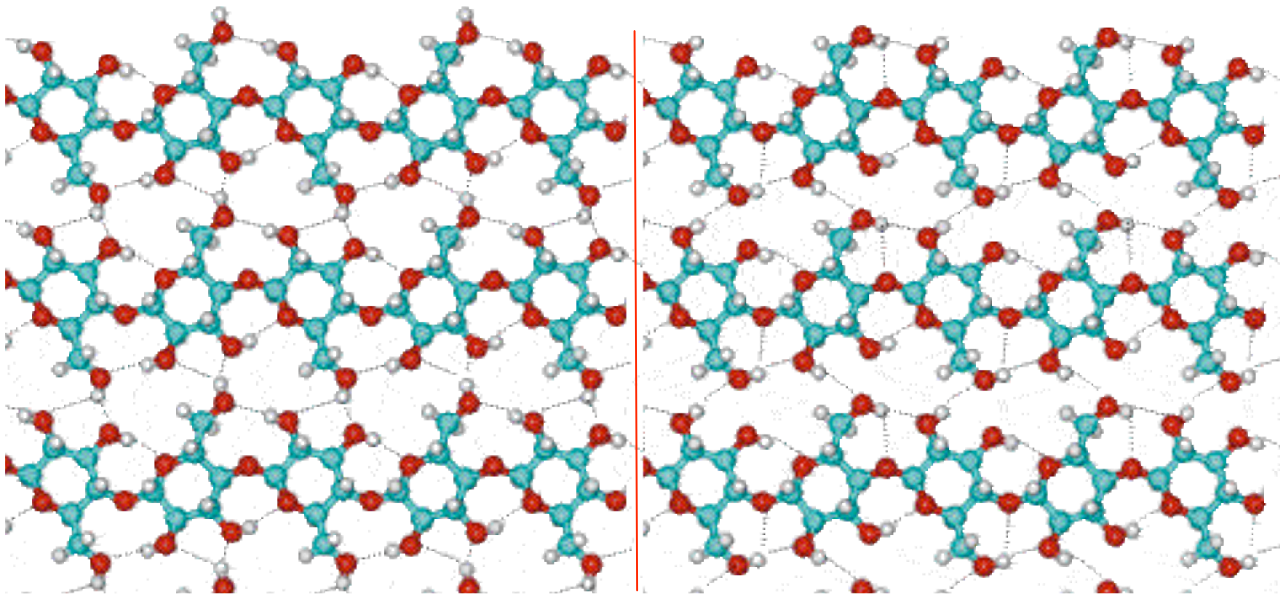


Cellulosic Biofuels
via: http://www.jgi.doe.gov

Cellulose

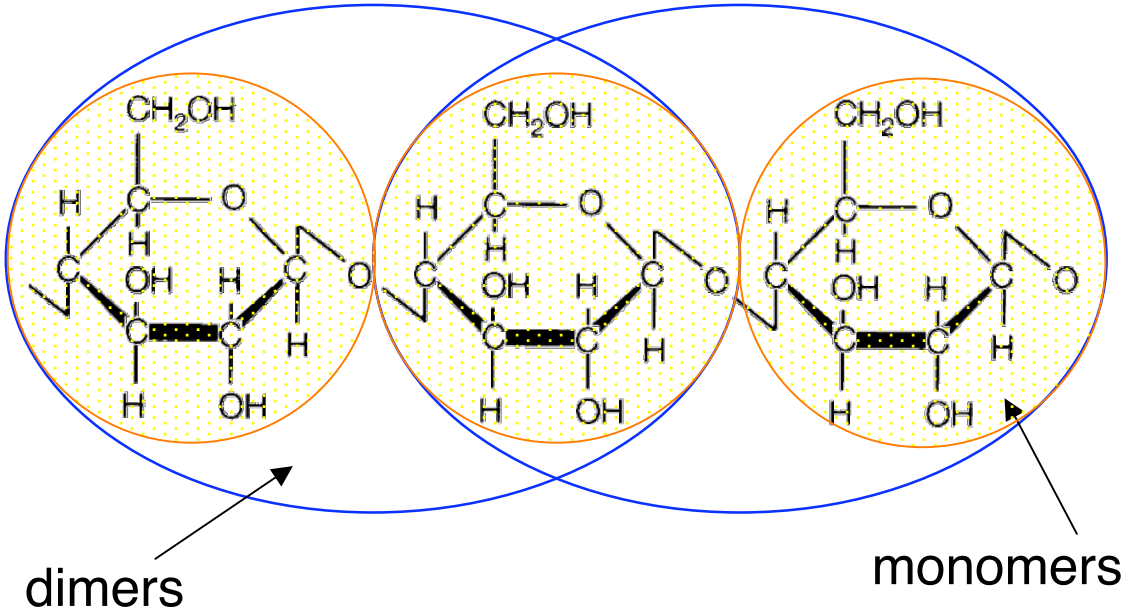


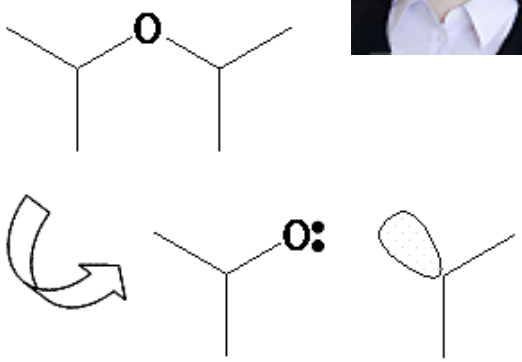
Chain: linear polymer of ß-(1 4)-D-glucopyranose units



Crystal: I-ß cellulose (in plants), monoclinic, H-bonded sheets

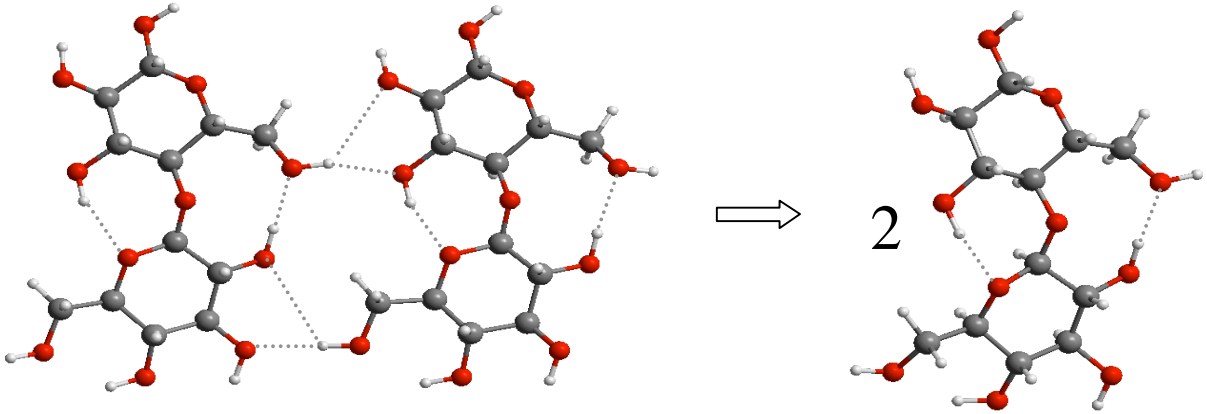
FMO2 Method for Cellulose





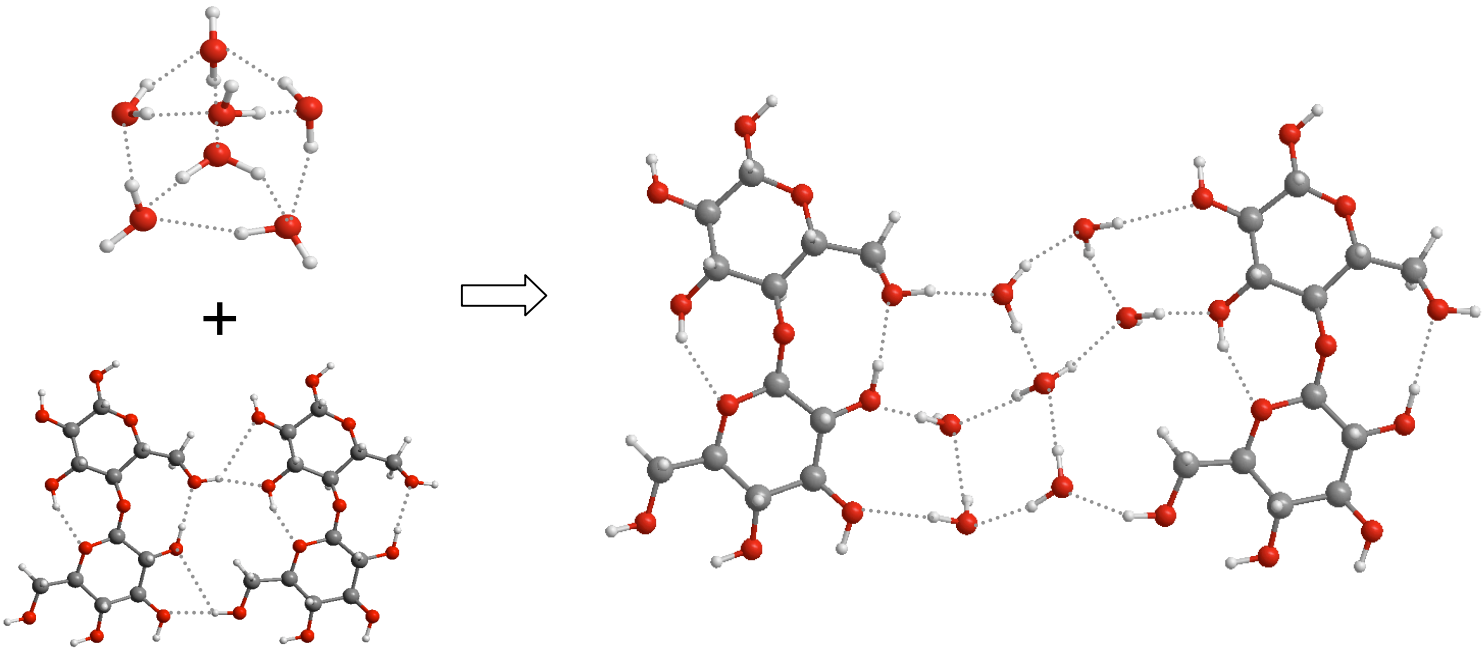
- Chains are divided into fragments by fractioning a glycosidic bond heterolytically
- Monomer energies are calculated in the electrostatic field of the whole system
- Dimer (FMO2) and trimer (FMO3) corrections are added

Intermolecular bonding in cellulose



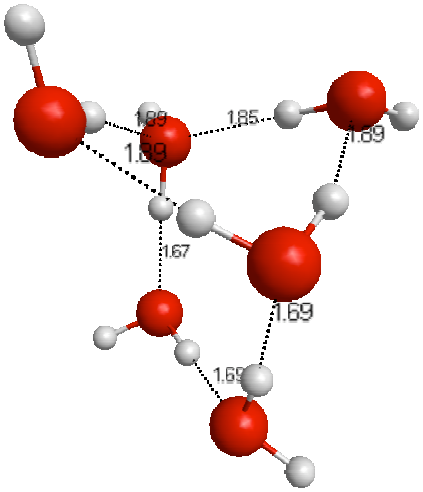
Method	Basis	Energy, kcal/mol	Energy per monomer pair, kcal/mol
HF-FMO2	6-31G(d,p)	17.9	8.9
HF-FMO3		17.1	8.5
HF		17.6	8.8
HF-FMO2	6-311G(d,p)	17.8	8.9
HF-FMO3		16.5	8.2
HF		16.8	8.4
MP2-FMO2	6-311G(d,p)	26.5	13.2
MP2-FMO3		24.6	12.3
MP2		25.3	12.6

Cellulose destruction by water

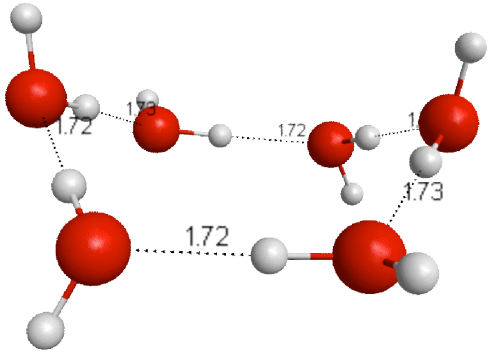


Method	Energy, kcal/mol		
	FMO2	FMO3	Full QM
HF/6-31G(d,p)	-7.6	-6.0	-7.2
HF/6-311G(d,p)	-8.8	-8.2	-
MP2/6-311G(d,p)	-11.6	-11.4	-

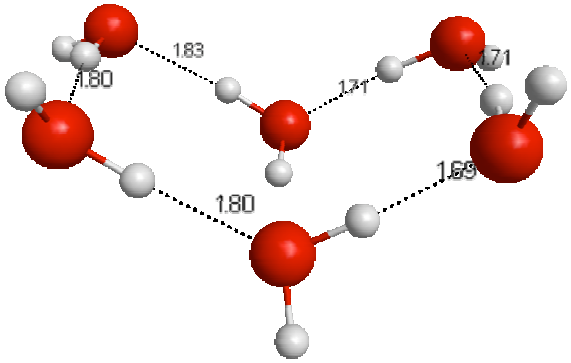
Water Hexamers



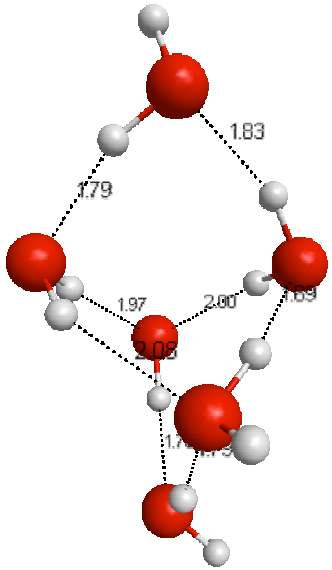
bag



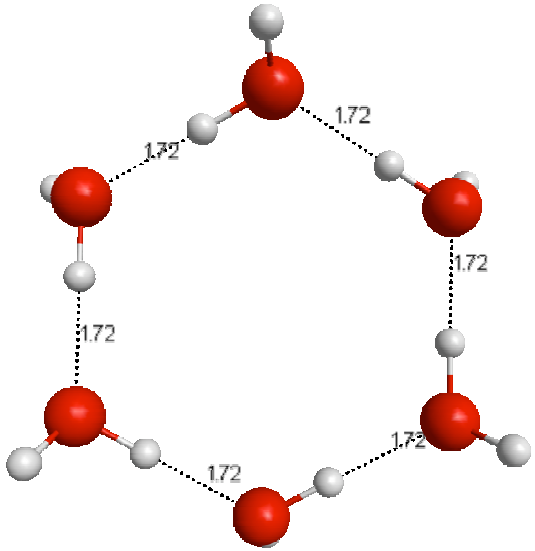
boat



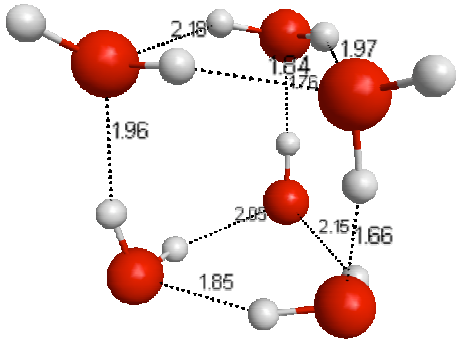
book



cage

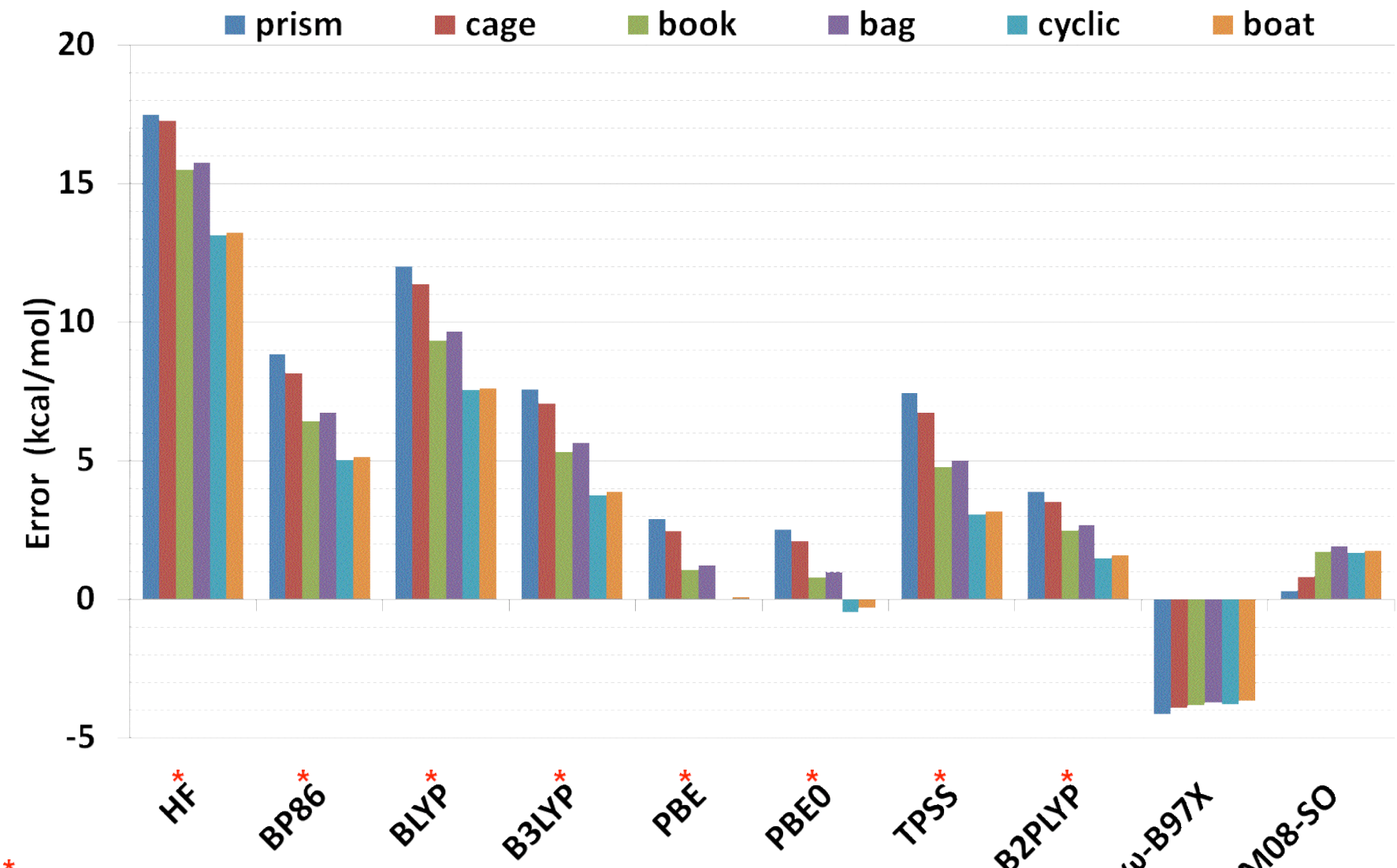


cyclic



prism

Water Hexamer
Errors in Total Interaction Energy
(relative to CCSD(T)//MP2/aug-cc-pVTZ)



* incorrect ordering of water structures

General Effective Fragment Potential (EFP)

§ **Discrete** solvent method

§ Fragment potential is one electron contribution to the QM Hamiltonian if QM part is present

§ Potentials

- are obtained by separate QM calculations
- depend on properties of **isolated** molecules
- can be systematically improved

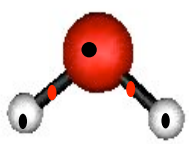
§ Internally frozen geometries

Generalized EFP Method

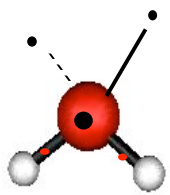
Interaction energy consists of : *electrostatic*, *polarization, exchange repulsion, dispersion, and charge transfer terms*

$$E_{interaction} = E_{Coulomb} + E_{Polarization} + E_{exrep} + E_{dispersion} + E_{Charge\ Transfer}$$

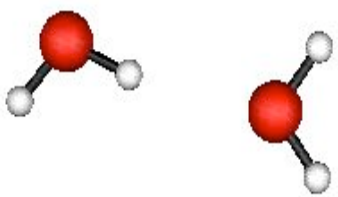
All from 1st principles: No fitted parameters



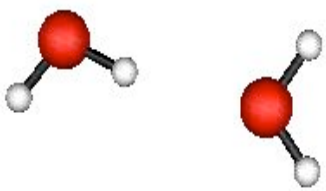
Distributed Multipolar expansion



LMO polarizability expansion



From first principles using LMO overlaps

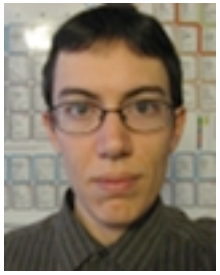


Distributed LMO dispersion from first principles

Screening: near field

Iterate to self-consistency

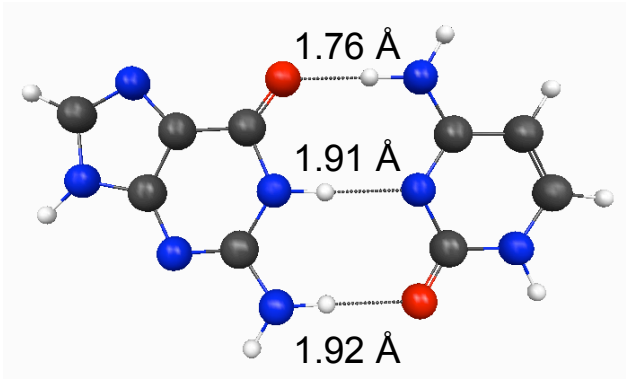
DNA BASES: TONY SMITH

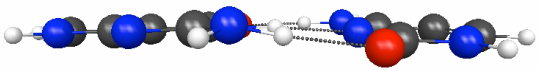


- Compare with Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608)
  - RI-MP2 / ext. cc-pVTZ or TZVPP **geometries**
  - RI-MP2/aug-cc-pVXZ (X = D, T, Q) to CBS limit **energies**
  - ΔCCSD(T) correction added
  - Counterpoise corrected for BSSE
- EFP generated with 6-311++G(3df,2p) basis set

Guanine-Cytosine H-Bonded

RI-MP2/TVZPP optimized



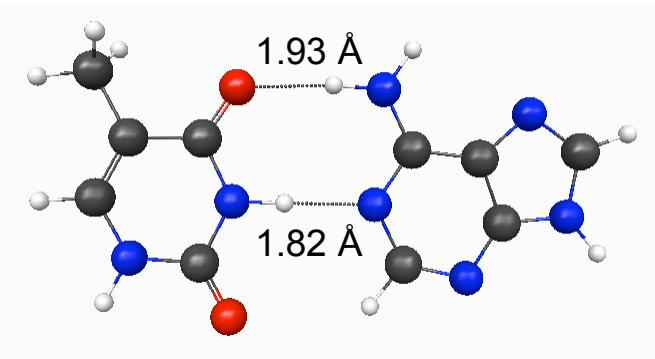


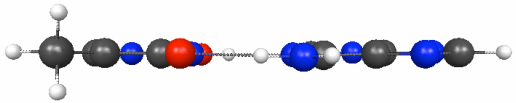
CCSD(T)/CBS energy: -32.1 kcal/mol

EFP

COUL	-43.3
REP	36.4
POL	-14.7
DISP	-10.7
TOTAL	-32.3

Adenine-Thymine H-Bonded





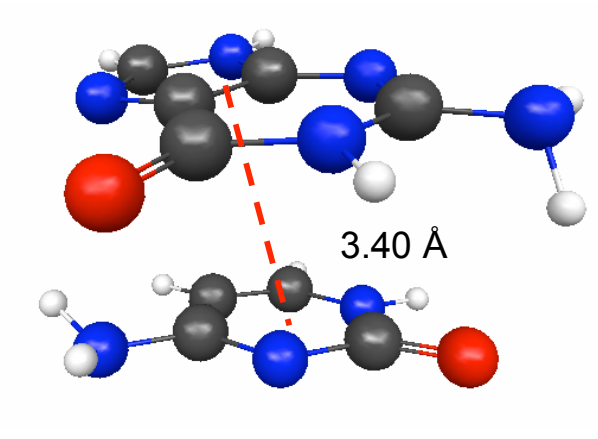
CCSD(T)/CBS energy: -16.9 kcal/mol

EFP

COUL	-27.1
REP	26.7
POL	-7.2
DISP	-8.2
TOTAL	-15.8

Guanine-Cytosine stacked

RI-MP2/TVZPP optimized



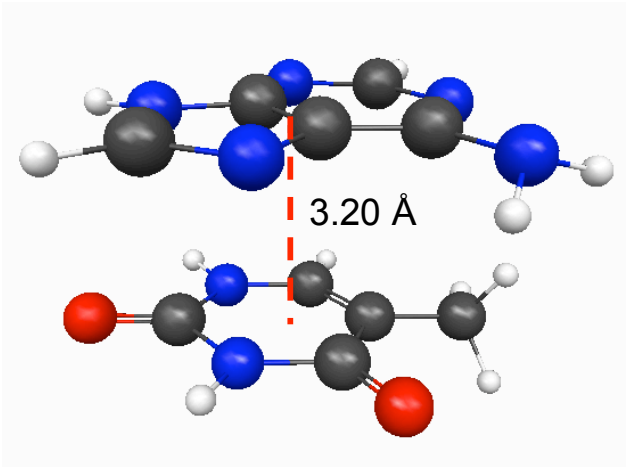
CCSD(T)/CBS energy: -19.0 kcal/mol

EFP

COUL	-19.2
REP	20.6
POL	-3.1
DISP	-17.5
TOTAL	-19.2

Adenine-Thymine Stacked

RI-MP2/TVZPP optimized



CCSD(T)/CBS energy: -12.3 kcal/mol

EFP

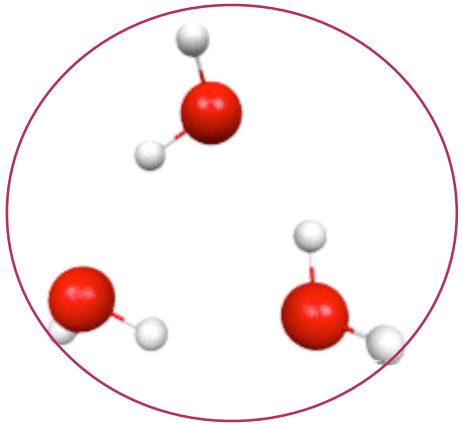
COUL	-7.8
REP	16.1
POL	-0.7
DISP	-17.6
TOTAL	-10.0

IMPROVING FMO3 EFFICIENCY

- Short-range interactions calculated fully *ab initio*
- EFP very effective for non-bonded interactions
- Replace *ab initio* non-bonded interactions with EFP
- For 3-body interactions, could save time without sacrificing accuracy
- Currently
  - Mulliken approximation used for intermediate range integrals
  - Point charges used for long-range interactions
    - EFP can account for all non-bonded interactions, not just Coulomb
- Alternative: Use EFP for all non-bonded interactions

The EFP applied to FMO

Performing EFP calculations involves the creation of the potential followed by the energy calculation



three-body interactions

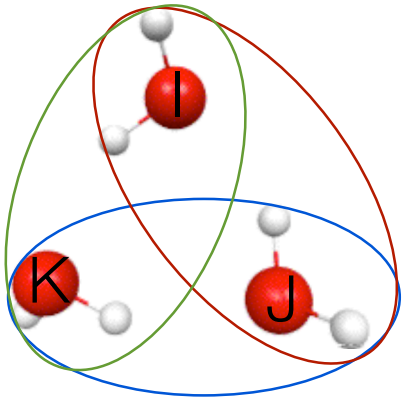
	EFP Generation	Energy Calculation	Total CPU Time (seconds)	Total Interaction Energy (kcal/mol)
MP2/6-311++G(3df,2p)	-	164.05 sec	164.05	-15.71
EFP/6-311++G(3df,2p)	13.06 sec	0.24 sec	13.30	-15.78

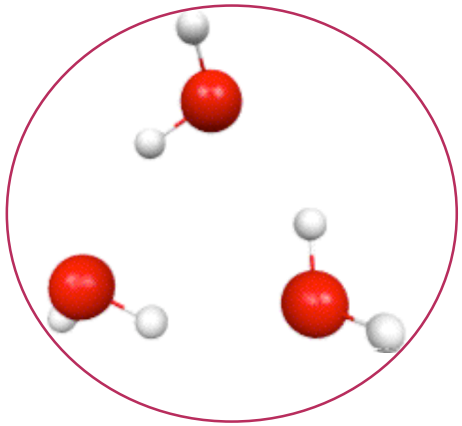
EFP can reduce 3-body computational cost by <u>at least</u> a factor of 13

*Calculations performed on a single 2.66 GHz Opteron CPU with 4GB of RAM

EFP vs. MP2 Interaction Energies

	MP2 Interaction Energy (kcal/mol)	EFP Interaction Energy (kcal/mol)
Dimers		
IJ	-4.4	-4.4
JK	-4.8	-4.7
IK	-4.7	-4.7
Trimer		
IJK	-1.7	-1.9

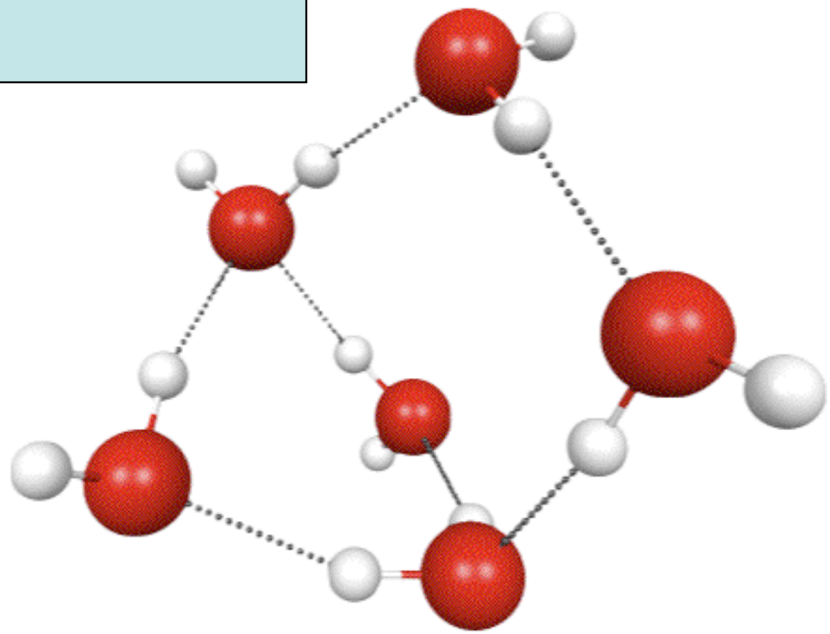




The EFP applied to FMO

- 6 monomers
- 15 dimers
- 20 trimers
  - ~ 2.7 minutes per FMO trimer*
  - ~ 1.0 second per EFP trimer*
- 54 minutes - 0.3 minutes =
- ~ 53.6 minutes saved*

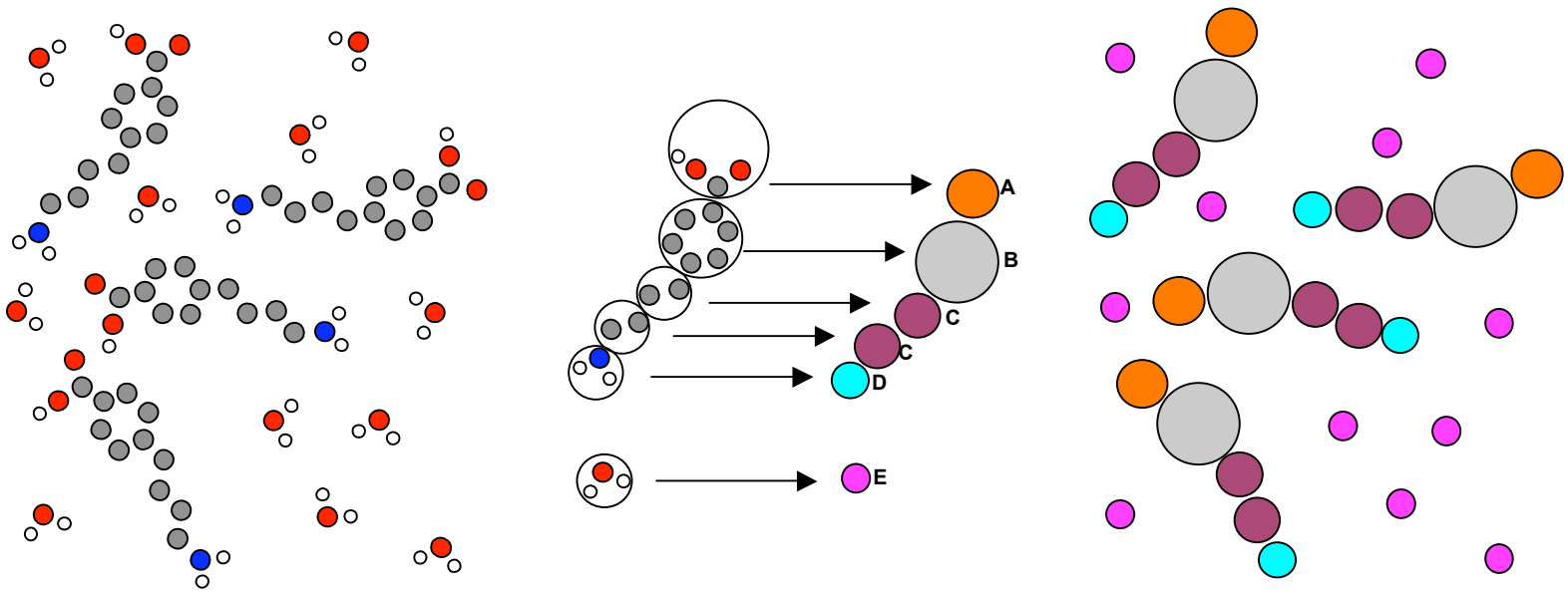
FMO2 + EFP ~ FMO3



EFP 3-body corrections can be obtained at almost no cost!

*Calculations performed on one 2.66 GHz Opteron CPU with 4GB of RAM

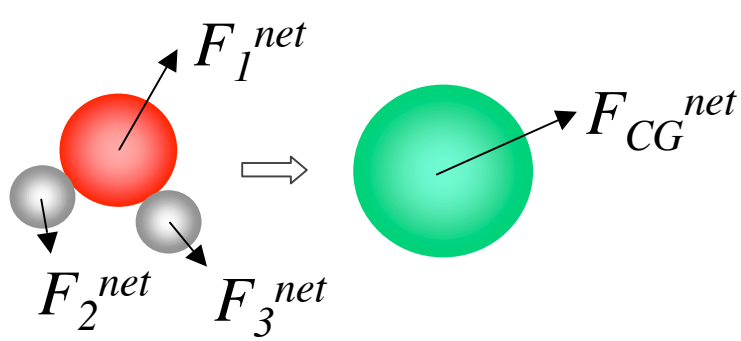
Coarse Graining Approach: Voth



(a) Atomistic system (b) Grouping scheme (c) Coarse-grained system representation

- Force matching: effective pair-forces between coarse-grained sites are derived from the net force acting on chosen CG sites along an MD trajectory obtained from a short atomistic MD

Force Matching Method



- CG pair wise effective force field built from given atomistic trajectories using force data
- A set of CG force field parameters $\{g_m\}$ is optimized by minimizing the function

$$\sum_{l,i} |\vec{F}_{i,l}^{\,\mathrm{ref}} - \vec{F}_{i,l}^{\,\mathrm{CG}}(\{g_m\})|^2$$

where $\vec{F}_{i,l}^{\,\mathrm{ref}}$ is a sum of total forces acting on atoms within CG site i in the lth frame of the reference trajectory, and $\vec{F}_{i,l}^{\,\mathrm{CG}}(\{g_m\})$ is the CG force calculated from the corresponding CG trajectory given the parameter set $\{g_m\}$.

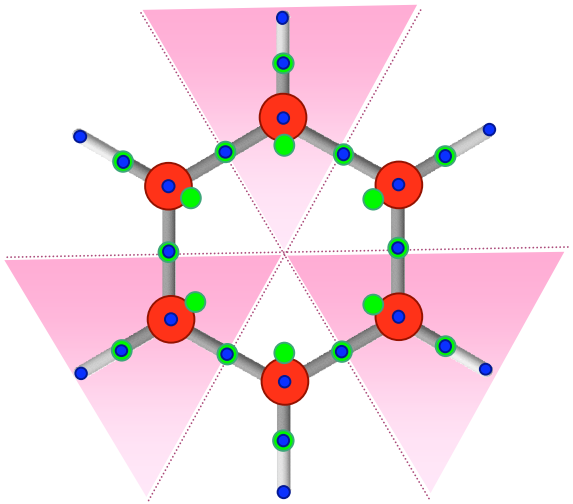
S. Izvekov and G.A. Voth, *J. Chem. Phys.* **2005** 123, 134105

EFP based coarse graining

- EFP trajectory data used to coarse-grain potentials according to the force matching scheme
- Previous study on one-site coarse graining of water, CCl_4 and benzene has shown very good match of CG RDFs with EFP MD data for H_2O and CCl_4, but slightly lower accuracy for benzene
- FMO MD in test stage
  - Long-term: FMO can replace EFP

G. Pranami, L. Slipchenko, M. H. Lamm and M S. Gordon, in Multi-scale Quantum Models for Biocatalysis: Modern Techniques and Applications", D.M. York and T.-S. Lee, Eds., Springer-Verlag, 2009

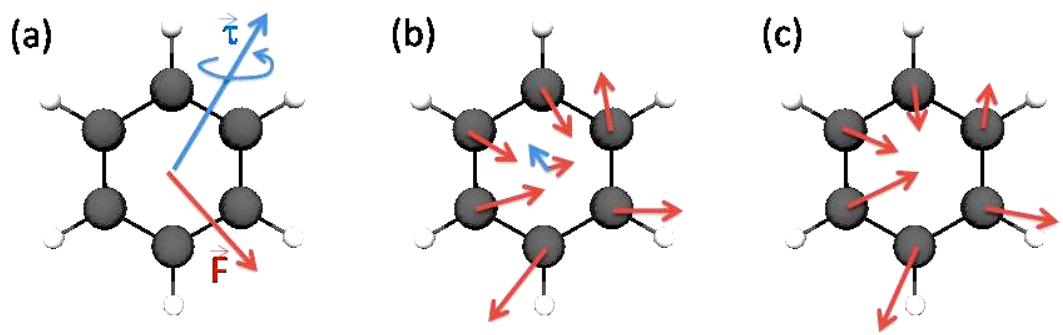
Multi-site Coarse Graining of Benzene



- electrostatic forces at multipole expansion points
- polarization dispersion and, ex-rep. forces at LMO centroids
- coarse graining sites

Force-shifting and switching functions introduced to ensure energy conservation in MD with periodic boundary conditions require an additional force and torque at fragment center of mass

$$\vec{\tau} = \Delta\vec{r} \times \vec{F}$$



EFP MD simulation

- EFP2 potential generated with HF/6-311++G(3df,2p)
- EFP molecular dynamics simulation parameters:

  64 benzene molecules

  Nose-Hoover thermostat

  NVT ensemble at T=300K

  Periodic box length 21.2 Å

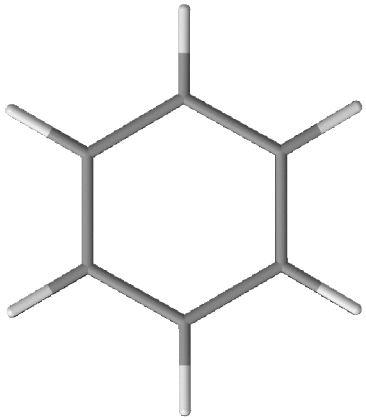
  Timestep 0.5 fs

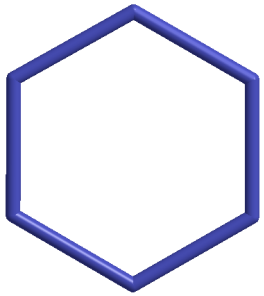
  Sampling frequency 50 fs

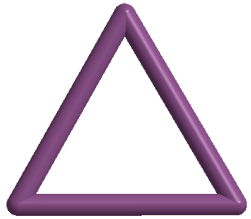
  Number of samples 2000

  Total simulation time 100 ps

Multisite Coarse Graining of Benzene

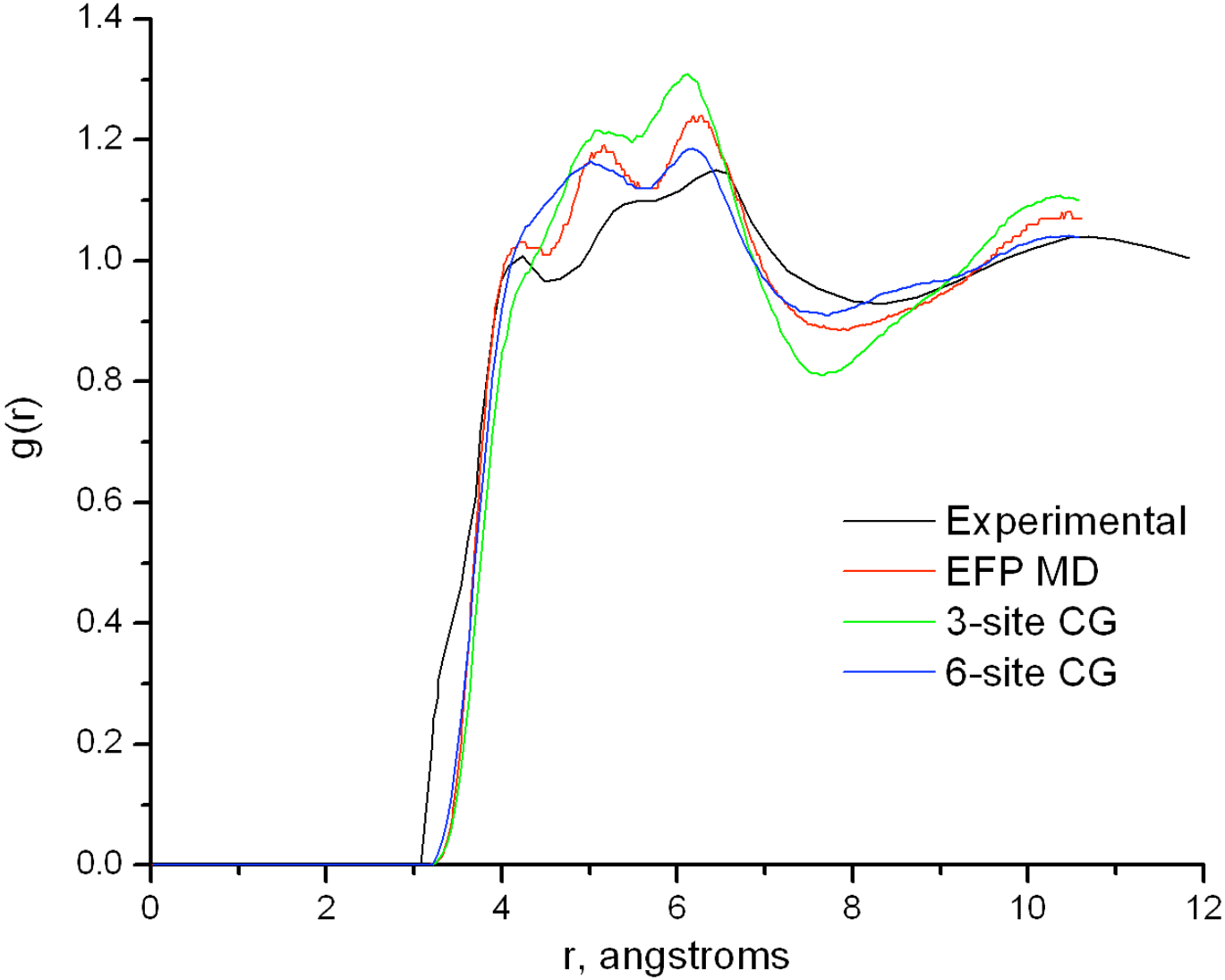




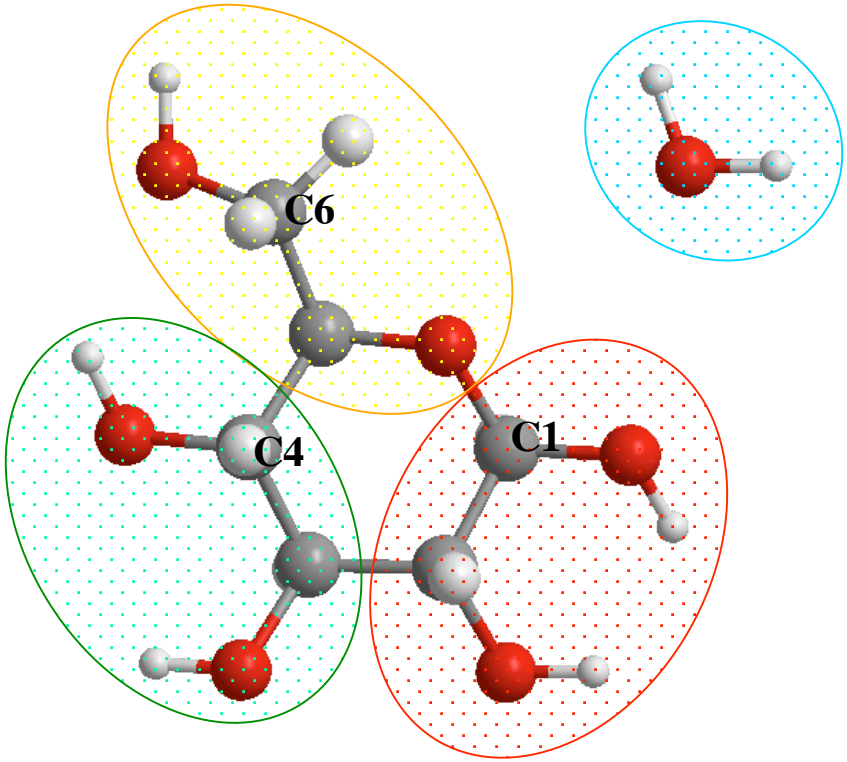


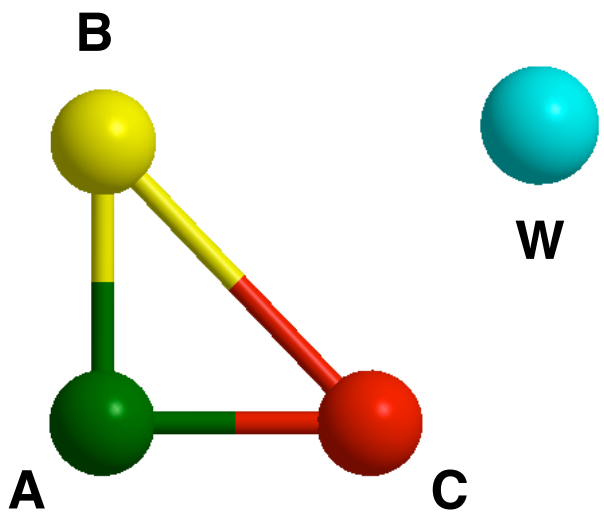
- EFP MD trajectories data used to build the coarse-grained potentials
- Multisite (3 and 6-site) CG schemes tested
- The resulting force field was employed to run coarse-grained MD
  - Simulation time: 3 ns
  - Timestep: 1 fs
- Radial distribution functions between certain atoms from CG MD were compared to atomistic EFP MD data

C-C Radial Distribution Functions



In progress: Coarse graining sugars in solution





Pair potentials: A-W
B-W
C-W

DREIDING FF, M3B CG model: V. Molinero and W.A. Goddard J. Phys. Chem. B 2004, 108, 1414
OPLS/AA FF, force matching: P. Liu, S. Izvekov, G.A. Voth *J. Phys. Chem. B* **2007,** *111,* 11566

Conclusions

- Application of high-level QM methods to large molecular systems is facilitated by the FMO approach.

- Derivation of coarse grained (CG) potentials form the first-principles based EFP may be used to study systems at large scale.

- Multiscale approach allows an efficient way to study both physical and chemical biomass processes.

"Now here comes the barbaric finale."

THANKS!

DOE SciDAC grant (ASCR/OBER)
DOE INCITE grant
US Air Force Office of Scientific Research
DoD HPC Grand Challenge Program
NSF Petascale Applications grant
ISU BG/L
ANL BG/P

WE GAMESS