Quantum Monte Carlo Calculations of Biomacromolecule Model Systems

Martin Korth

Grimme Group, University of Münster, Germany

31/07/2006

Outline

Introduction

QMC@HOME

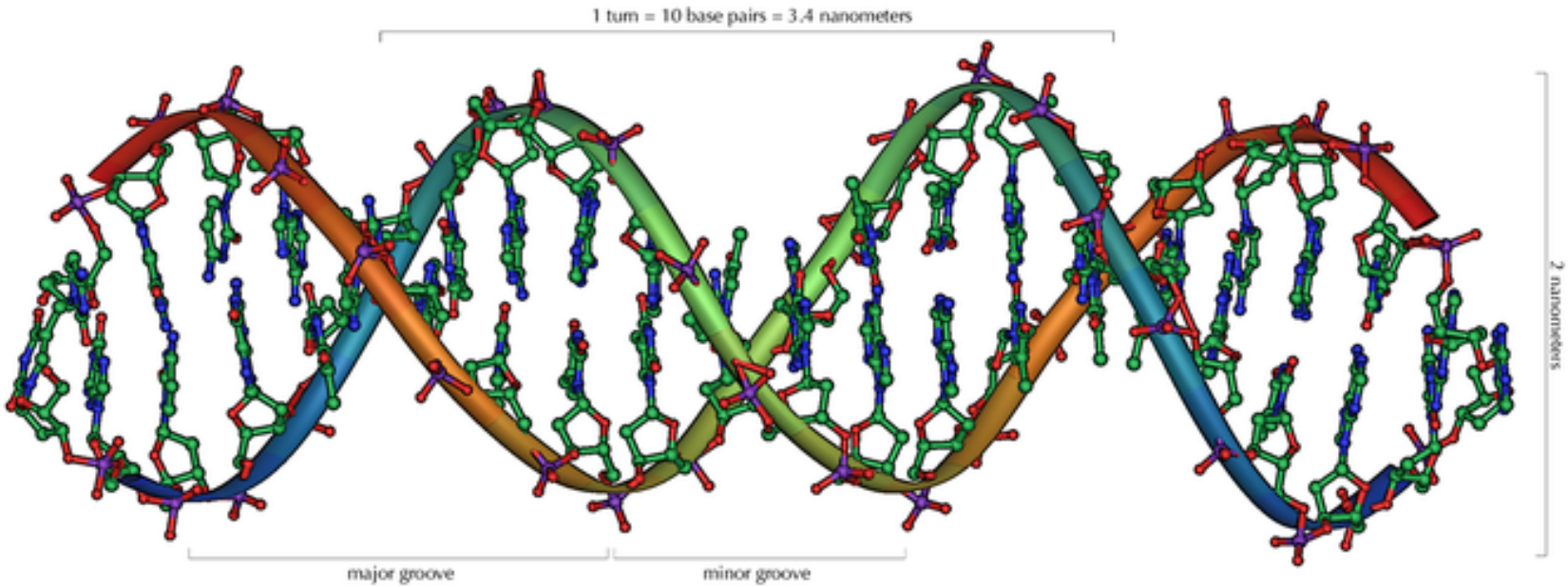
... volunteer computing for quantum chemistry!

Biomacromolecule model systems

DNA base pairs
S22 set of noncovalently bound systems
tripeptid conformer study
(alkane chain folding (IDHC7))
(non/saturated ring stacking)

Summary

Biomacromolecule interactions



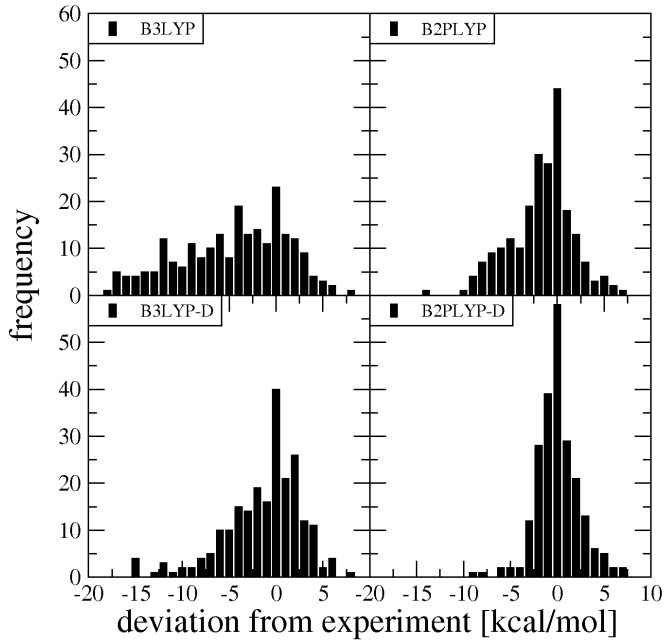
The problem

- structure of biomacromolecules (DNA, RNA, proteins) highly influenced by nonconvalent interactions between the basic building blocks (DNA/RNA bases, amino-acids)
- hydrogen bonding, electrostatic and **van der Waals interactions** important $\longrightarrow$ **very complex systems**

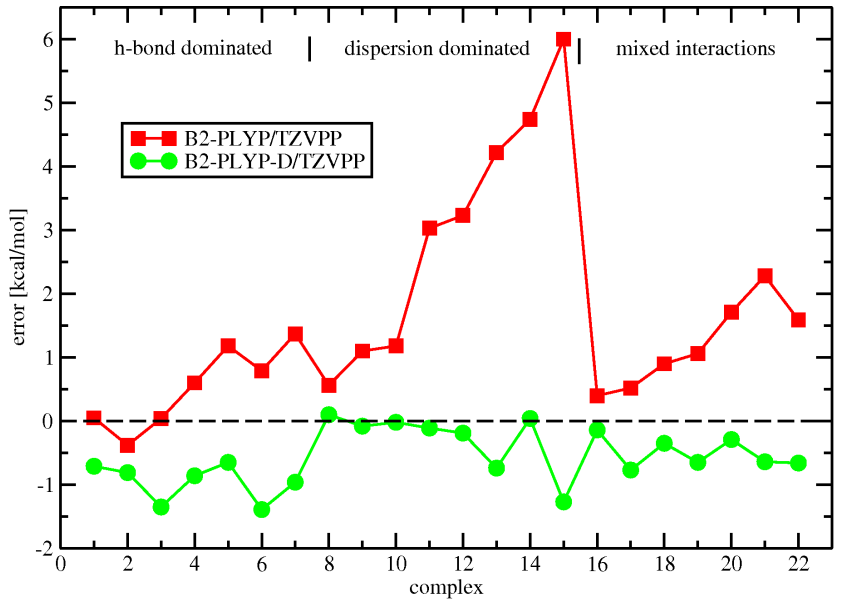
The everyday tool of quantum chemistry: DFT

Double-hybrid density functionals with long-range dispersion corrections

- outstanding accuracy, here competitive to high-level reference data
- computations for systems with 100-200 atoms are possible
- more details: Schwabe/Grimme PCCP, 2007, 9, 3397



G3/99 atomization energies



S22 noncovalent interactions

Reference methods of quantum chemistry

The 'gold standard' of quantum chemistry: CCSD(T)

- unfavorable scaling with system size and poor parallelizability → growing gap between low and high level methods
- **the future development of approximate methods needs high accuracy reference methods for large systems**
- a (temporary) solution: extrapolation of CCSD(T)/MP2 data gives high accuracy references for medium-sized systems, e.g. CCSD(T)/CBS: CC/DZ + MP2/(TQ)Z basis set extrapolation

FNDMC as reference method for noncovalently bound systems?

- advantages: scaling behavior, parallelizability
- disadvantages: large pre-factor for computation time
  → **FNDMC is promising**, but we need a lot of computing time to study biomacromolecule model systems

Part I

QMC@HOME

http://qah.uni-muenster.de

QMC@HOME

Volunteer Computing for Quantum Chemistry

- an advantage of QMC that is becoming increasingly important with high-density (multi-core, multi-socket) and distributed (cluster, grid) computing: **massively parallel calculations**
- QMC is even suited for a very special 'flavor' of distributed computing: **Volunteer Computing** (VC)

Volunteer Computing

Public resources ...

- majority of the world's computing power no longer concentrated in supercomputer centers, instead distributed in hundreds of millions of personal computers (150 millions in 2004, est. 1 billion in 2015)
- VC invites the public to donate spare computing power to science

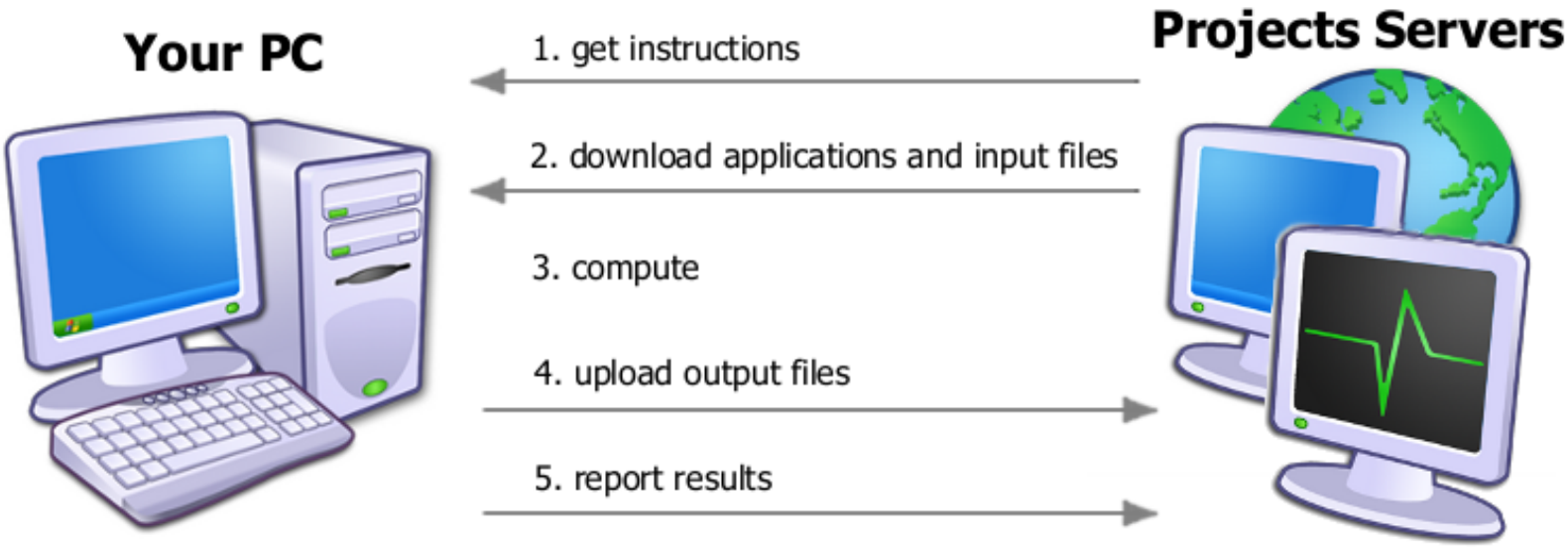
... for scientific research

- five really large (Seti, Einstein, ClimatePrediction, ...), around 10-15 mid-size (headed by QMC@home) and 30+ smaller VC projects
- together we broke the PetaFLOP barrier on Jan 31, 2008!
  - thanks to the **BOINC Volunteer Computing middle-ware**

BOINC middle-ware

Berkeley Open Infrastructure for Network Computing

- a software platform for Volunteer Computing
- developed within SETI@HOME (David P. Anderson)
- freely available via the Gnu Lesser General Public License (LGPL)



BOINC structure

Backend

- Apache web-server, MySQL databank

Middle-ware

- web-frontend (PHP), DB-Interface (Python)
- set of Daemons (C++/C)
  - scheduler daemons (getting data on it's way)
  - transitioner, validator (redundant computing)
  - work generator, assimilator (create and evaluate work)
- core-client: one client software for all projects!

Application

- communication between core client and application via MPI-like functions (boinc_init, boinc_finish, ...)

BOINC features

Project side

- project autonomy (volunteers can choose via the core client software between **independent projects**)
- failure and back-off mechanisms
- redundant computing (to avoid cheating)
- software security updates

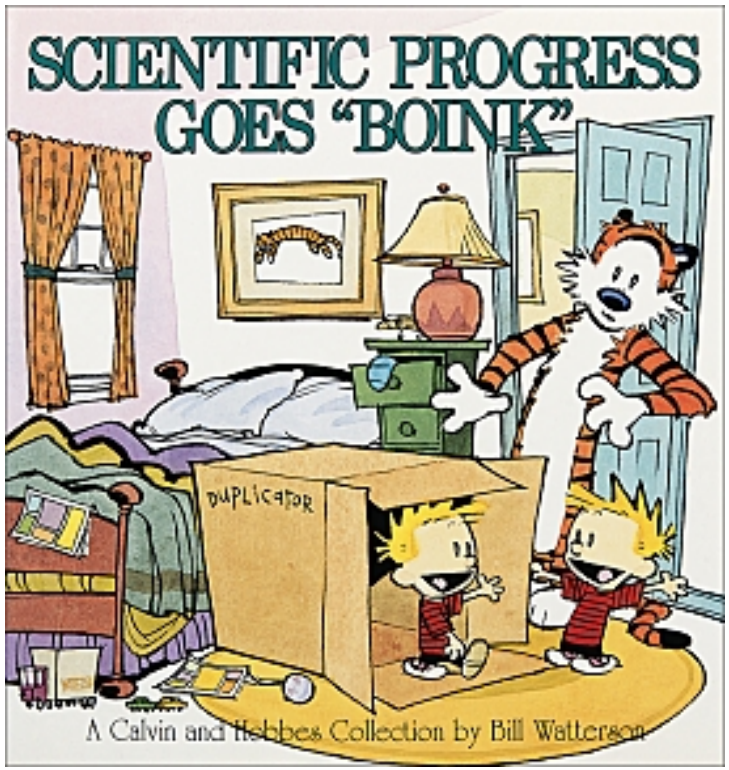
Volunteer side

- volunteer flexibility (project participation and resource allocation)
- accounting with participant preferences and **crediting**
- volunteer community features: profiles, teams, forums
- screen-saver graphics interface

AmolQC/BOINC

The QMC@HOME application

- based on the **AmolQC QMC program by Arne Lüchow et al.**
- core-client/application interface (init, file-opening, ..., finish)
- flexible checkpointing code
- screen-saver graphics code

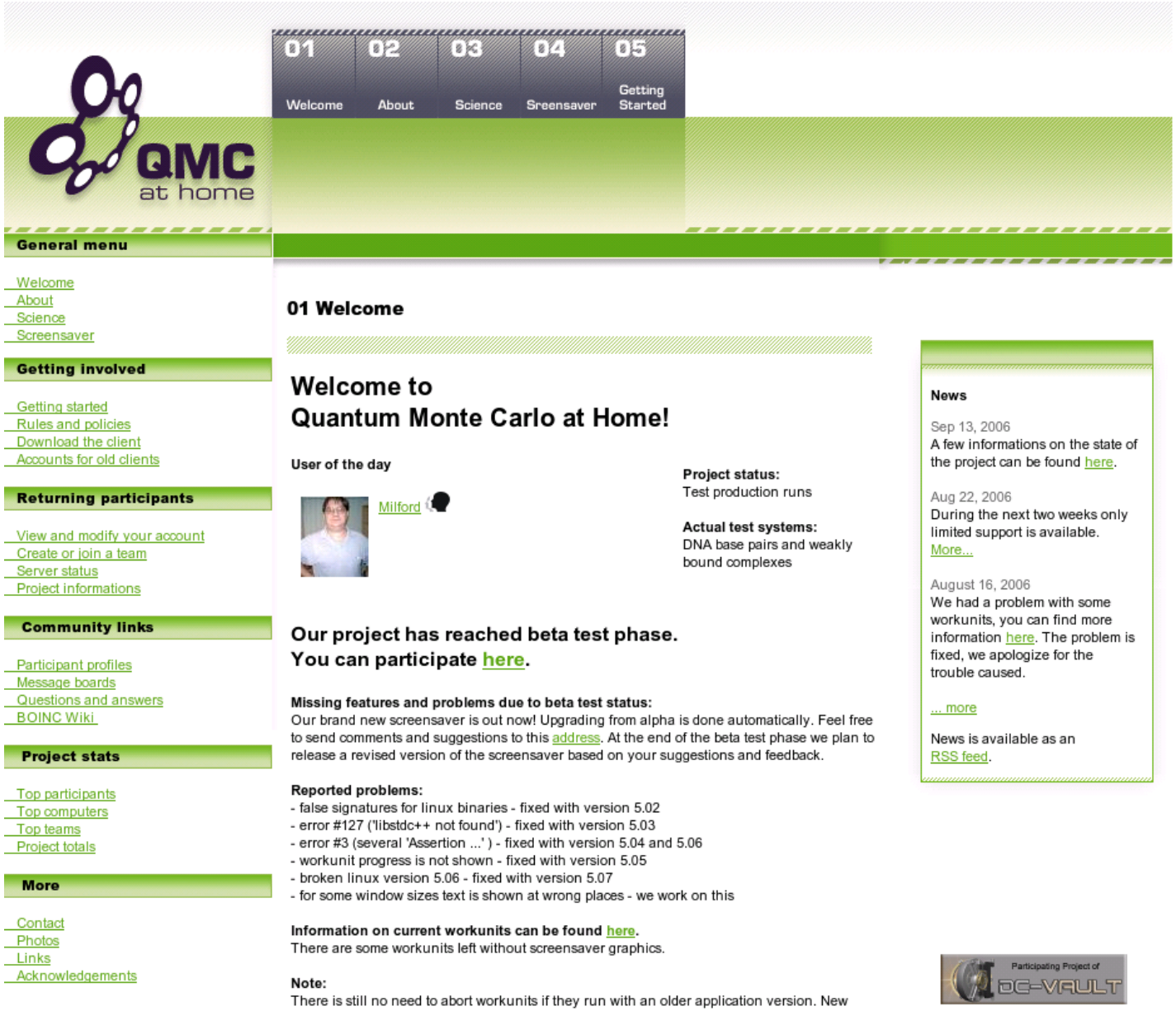


QMC@HOME web site

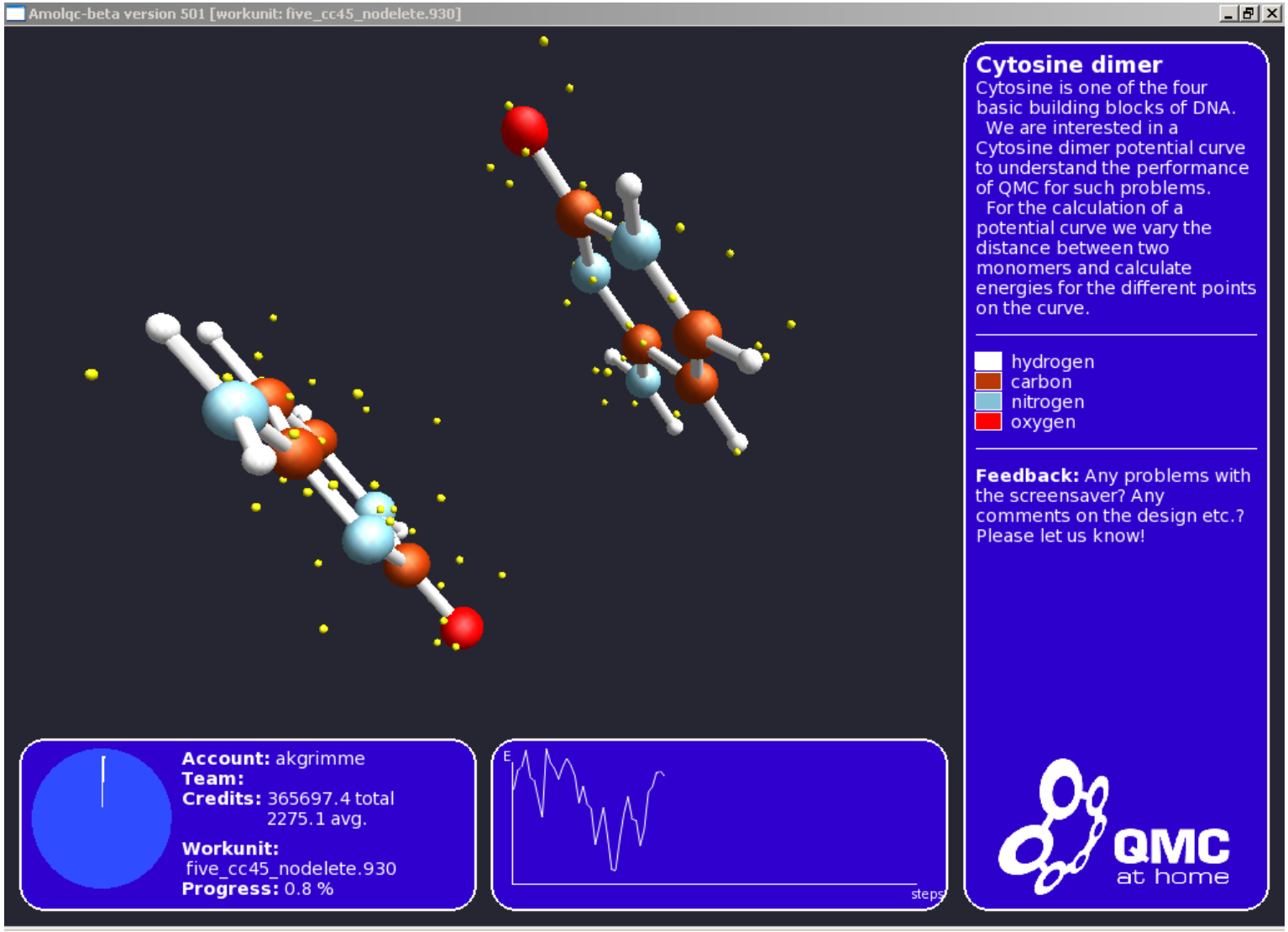
VC

BOINC

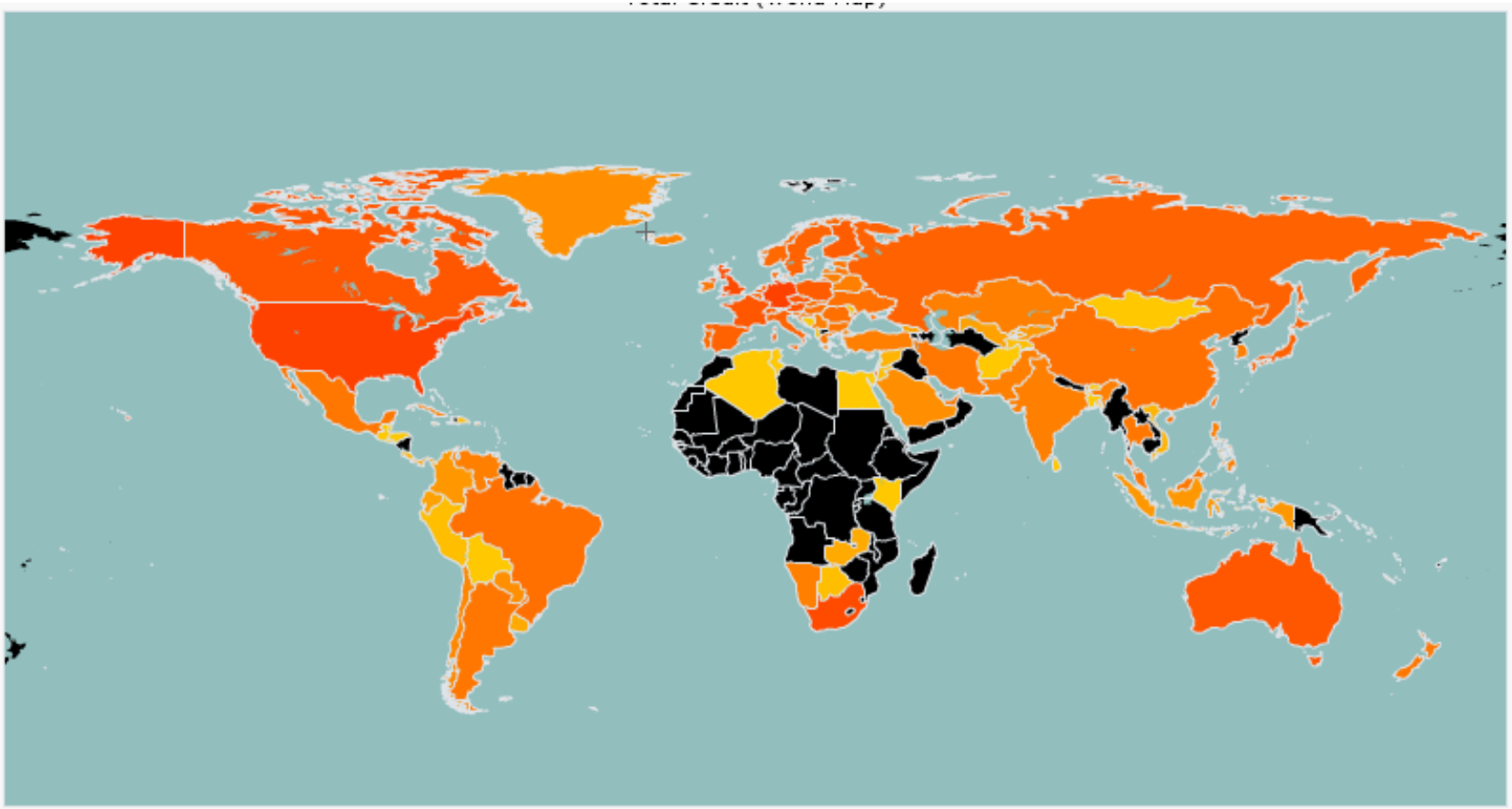
QMC
@HOME



QMC@HOME screen-saver graphics



... (nearly) all around the world ...



QMC@HOME

Statistics (08/07/2008)

- over 48000 registered users and over 107000 registered hosts
- over 11000 highly active hosts
- over **19 TeraFLOPS** average computing power
- (not really) equivalent to rank 99 (international) / 11 (Germany) on the top500.org supercomputer list - you need over 2000 brand new Xeon processors to get there!

In need for a few hundred processors?
Contact me for help setting up your BOINC project!

Part II

Biomacromolecule model systems

Technical details

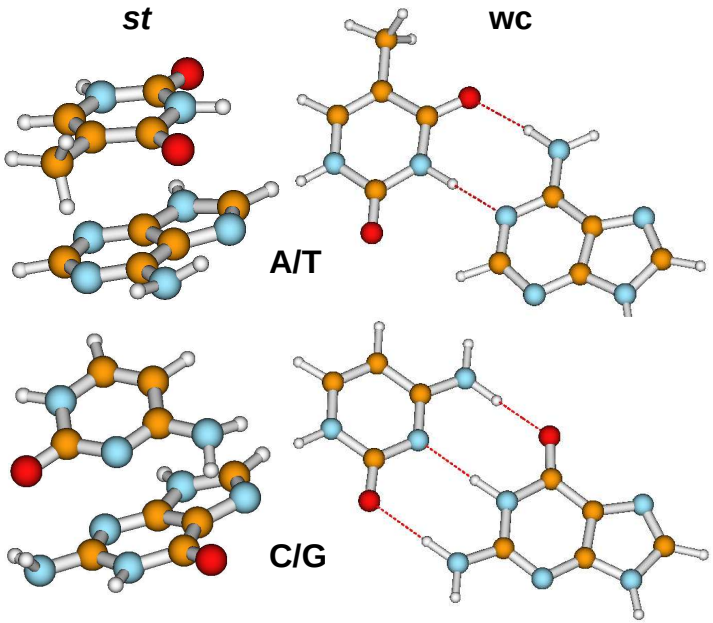
Used if not mentioned

- Slater-Jastrow type guidance functions with HF determinants and Schmidt-Moskowitz type correlation functions
- fully optimized gaussian quadruple-ζ basis sets with soft-ECPs by Ovcharenko *et al.*
- 'SM9' Jastrow type (4 ee + 3 en + 2 een), parameters optimized by variance minimization
- 250-2000 work-units, each of n*4000 steps with an ensemble of 100 walkers and a time step of 0.005

Mentioned if used

- 'EJ4' (1 ee + 3 en) and 'EJ4+L' (1 ee + 3 en + 2 long-range-en) Jastrow type, parameters optimized by variance minimization
- triple and quadruple-ζ (without g functions) basis sets and soft-ECPs by Burkatzki *et al.*, termed here 'BTZ' and 'BQZ'

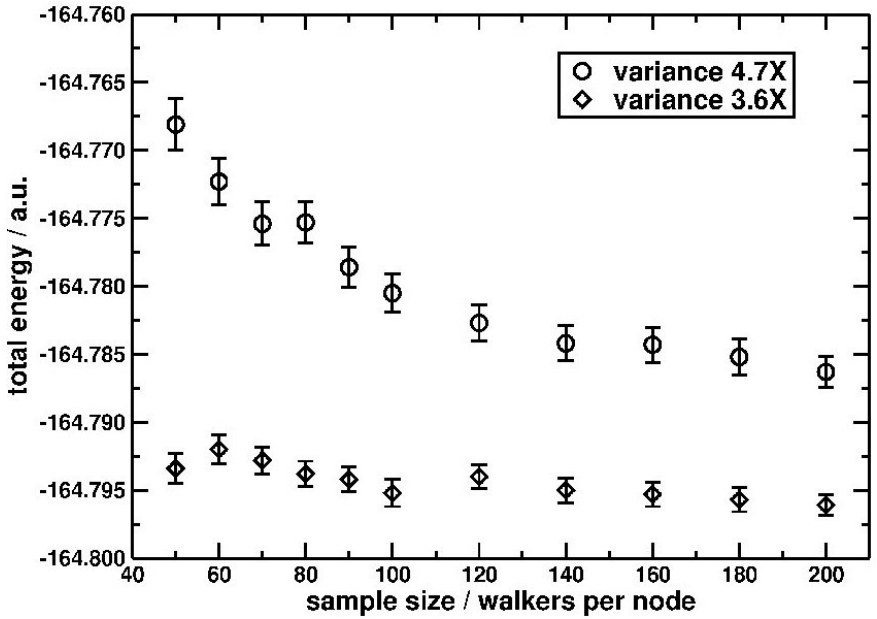
DNA base pairs



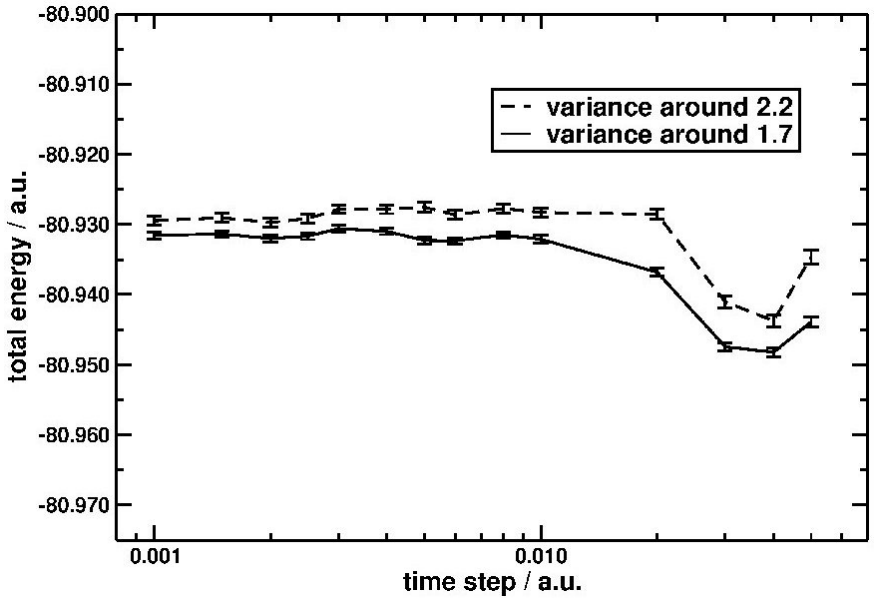
DNA model systems:

Watson-Crick bound (wc) and stacked (st) nucleic acid base pairs A/T and C/G

population control error checks for the stacked A/T complex

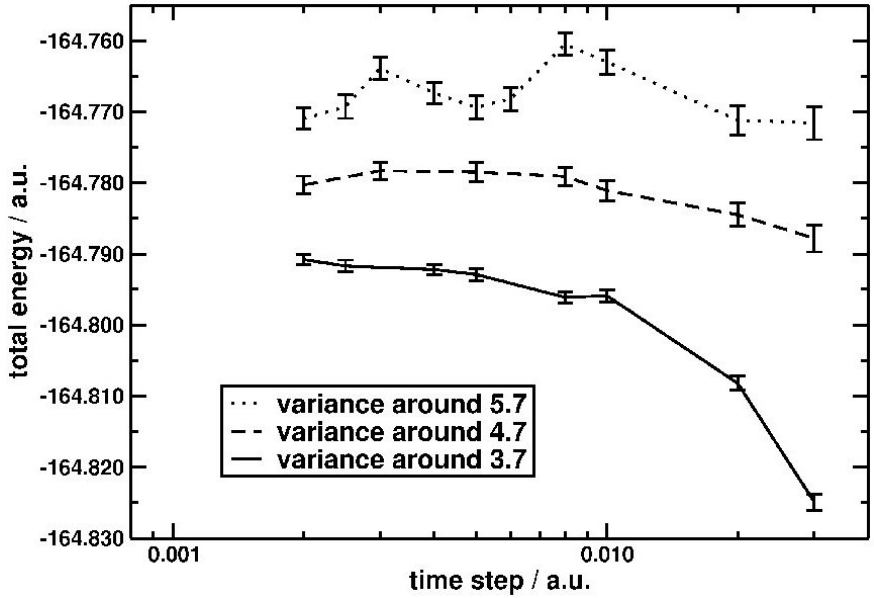


DNA base pairs - absolute energies

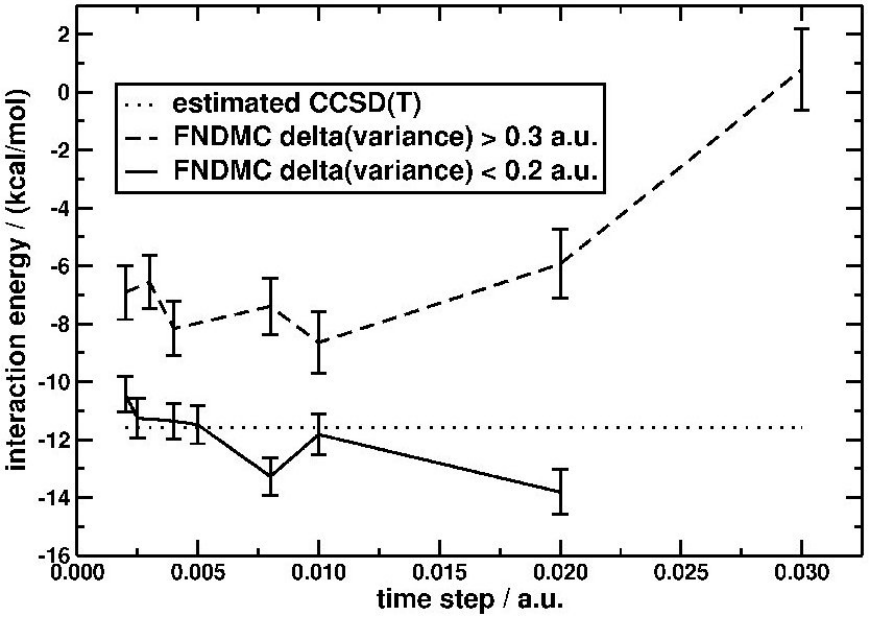


Adenine monomer

stacked A/T complex

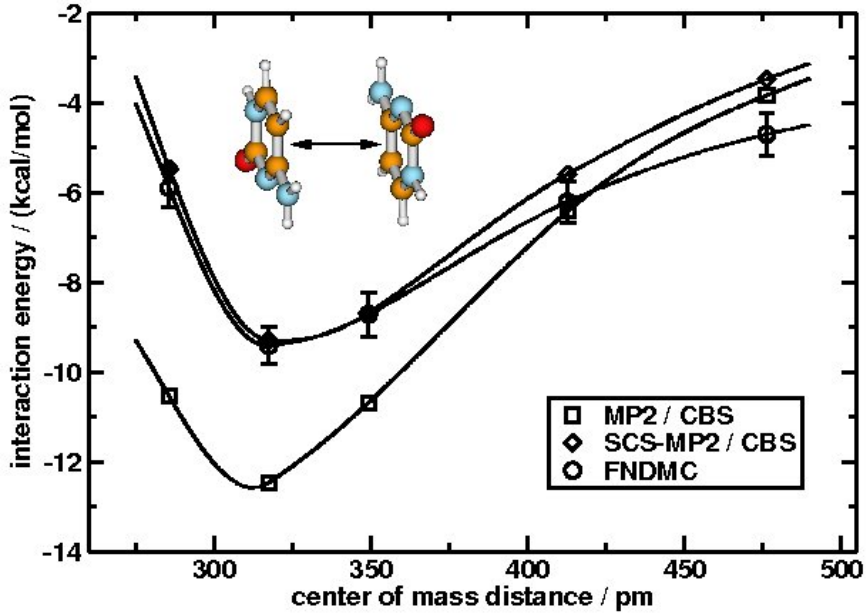


DNA base pairs - relative energies



stacked A/T complex

C/C dimer



DNA base pairs - interaction energies

QMC@HOME Results

DNA base pair interaction energies (kcal/mol):

(statistical (FNDMC) / estimated (CCSD(T)) error in parenthesis)

	FNDMC BLYP-D/TZVPP	CCSD(T)/CBS RI-MP2/TZVPP[1]
A/T, st	-13.1(8)	-11.64(50)
A/T, wc	-15.7(9)	-15.43(50)
C/G, st	-19.6(9)	-16.90(50)
C/G, wc	-30.2(9)	-28.80(50)

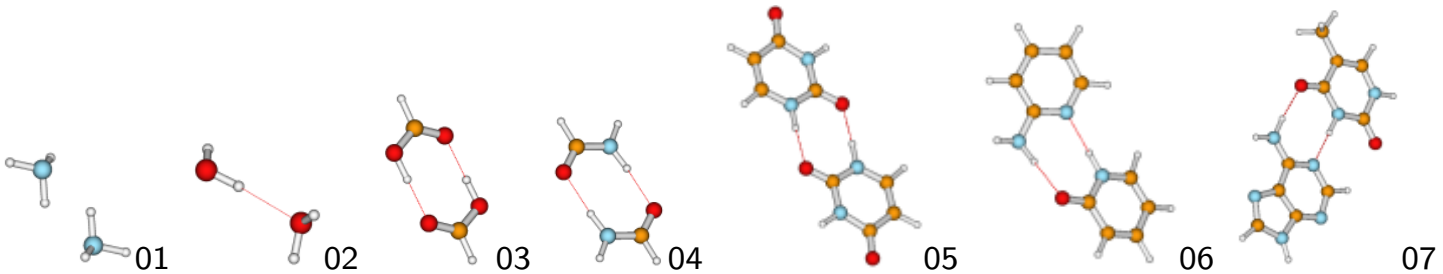
[1] Jurecka *et al.*, Phys. Chem. Chem. Phys., **2006**, 8, 1985.

The problem

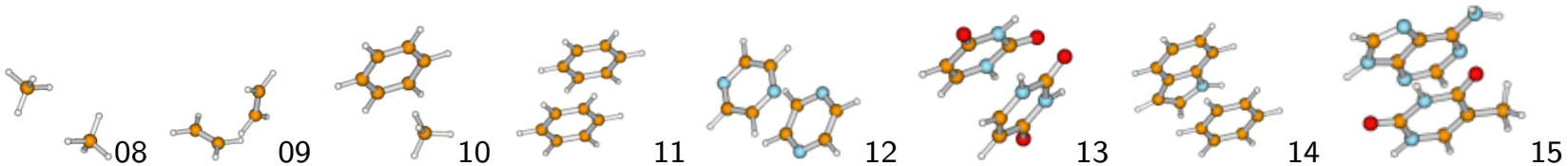
- better trial WF and smaller time steps would be nice ...
- choosing 'equally good' trial WF (measured by the variance) leads to very good results - but what general accuracy can we expect for this approach?

S22 benchmark set of noncovalently bound systems

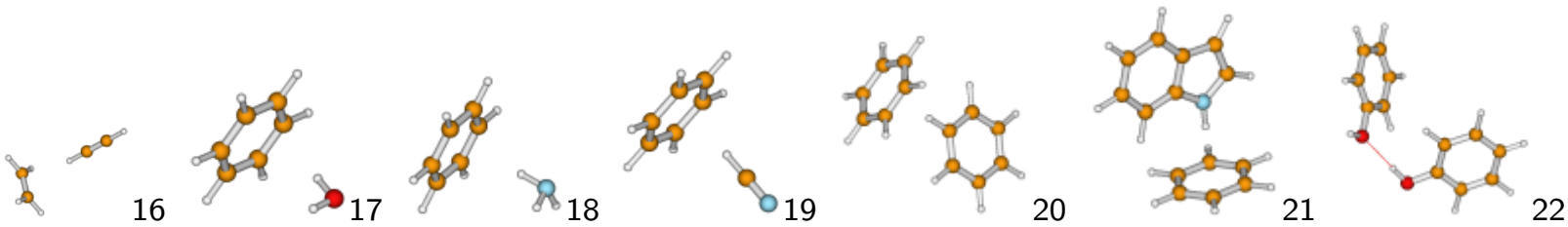
Hydrogen-bond dominated:



Dispersion dominated:

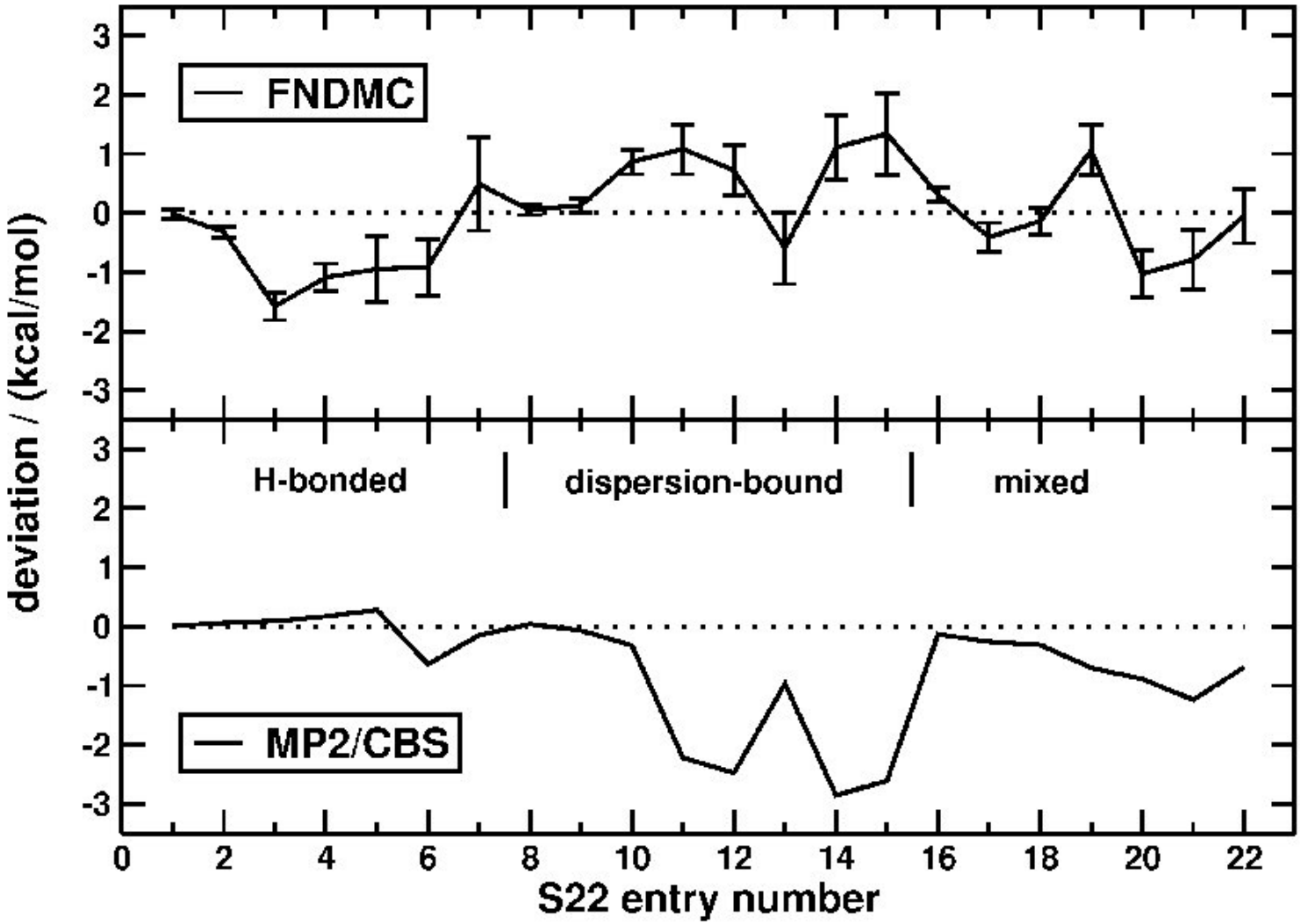


Mixed complexes:



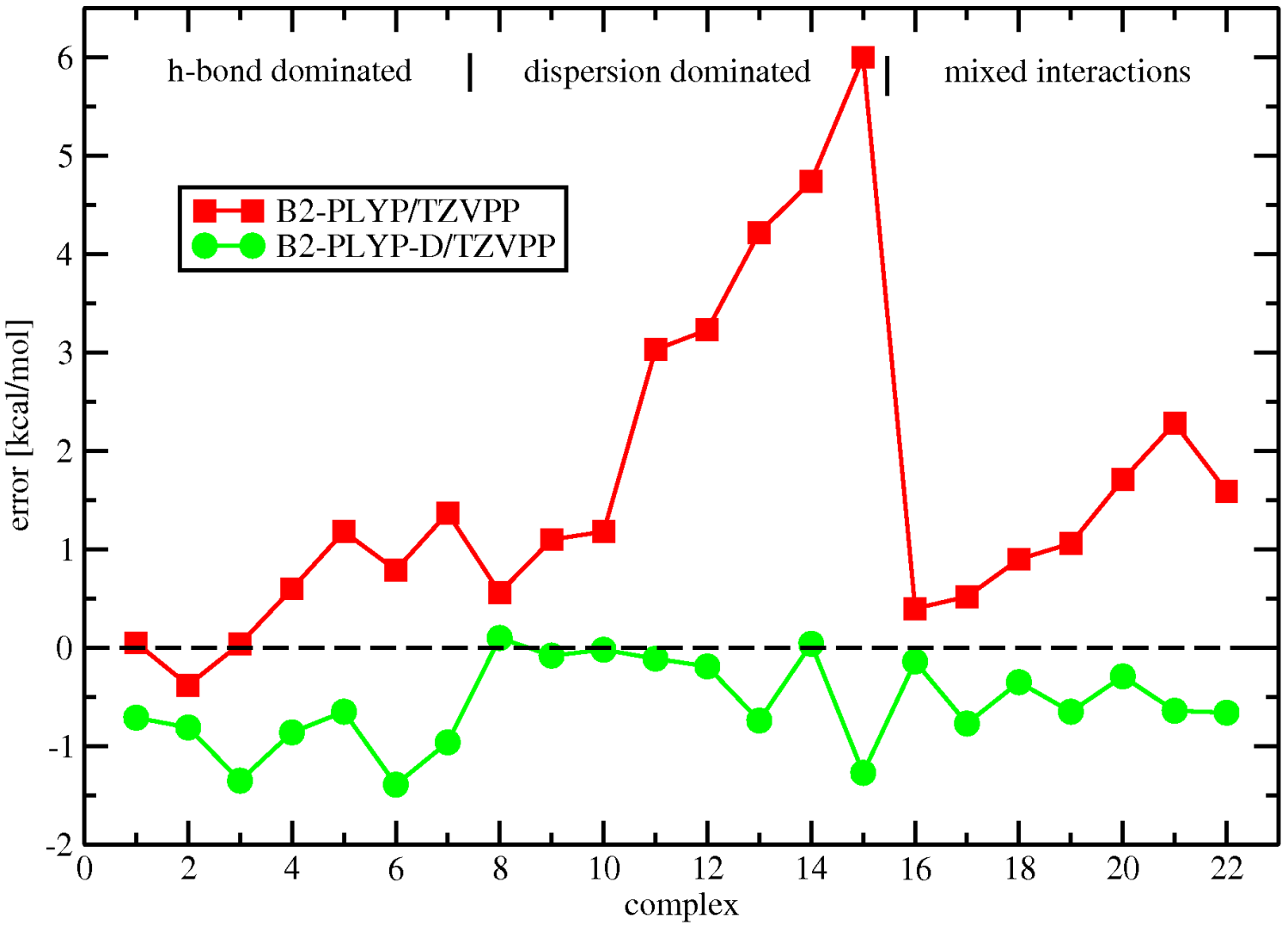
Jurecka *et al.*, Phys. Chem. Chem. Phys., **2006**, 8, 1985.

S22 benchmark set / FNDMC



MAD 0.68 kcal/mol

S22 benchmark set / DFT



MAD 0.58 kcal/mol

DNA base pairs / S22 benchmark set

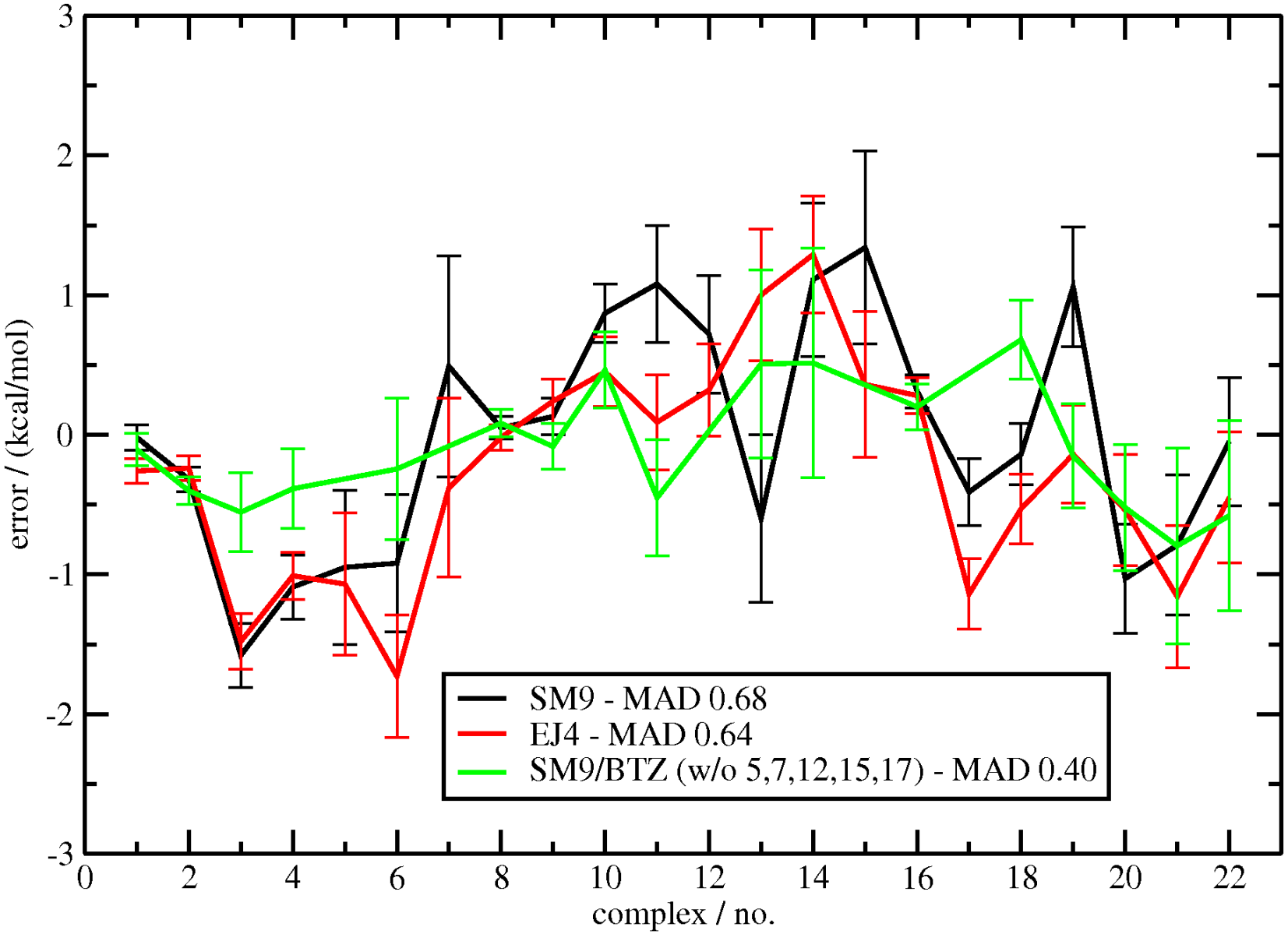
Conclusions so far

- overall good performance for the DNA base pairs and the S22 benchmark set (errors of 0.5-1.0 kcal/mol)
- more details in Korth/Lüchow/Grimme JPCA, 2008, 112, 2104
- accuracy and reliability not really sufficient for the use as reference for biomacromolecule model systems

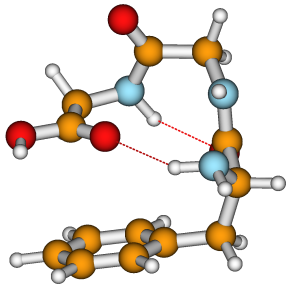
What to do next

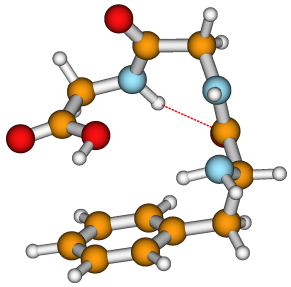
- check influence of Jastrow type, basis set and ECP on accuracy and reliability
- check performance for larger systems

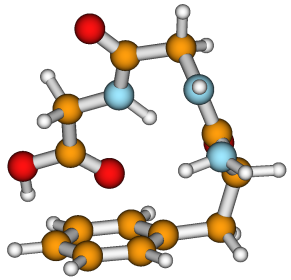
S22 benchmark set / FNDMC 2 - work in progress

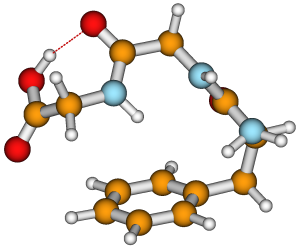


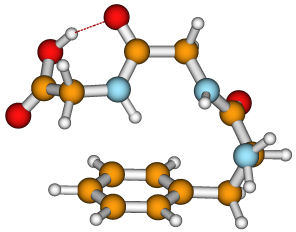
Tripeptide conformers Reha *et al.*, Chem. Eur. J., **2005**, 11, 6803.

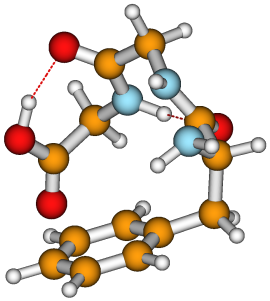

1/099

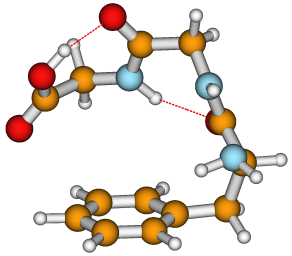

10/412

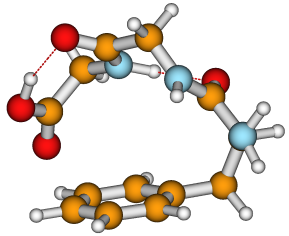

11/470

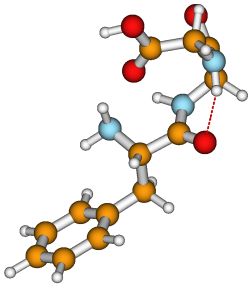

4/357

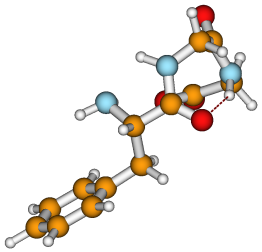

7/114

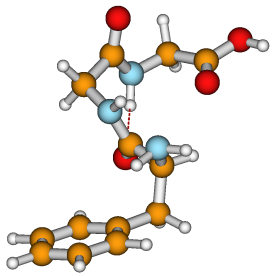

2/444


3/215

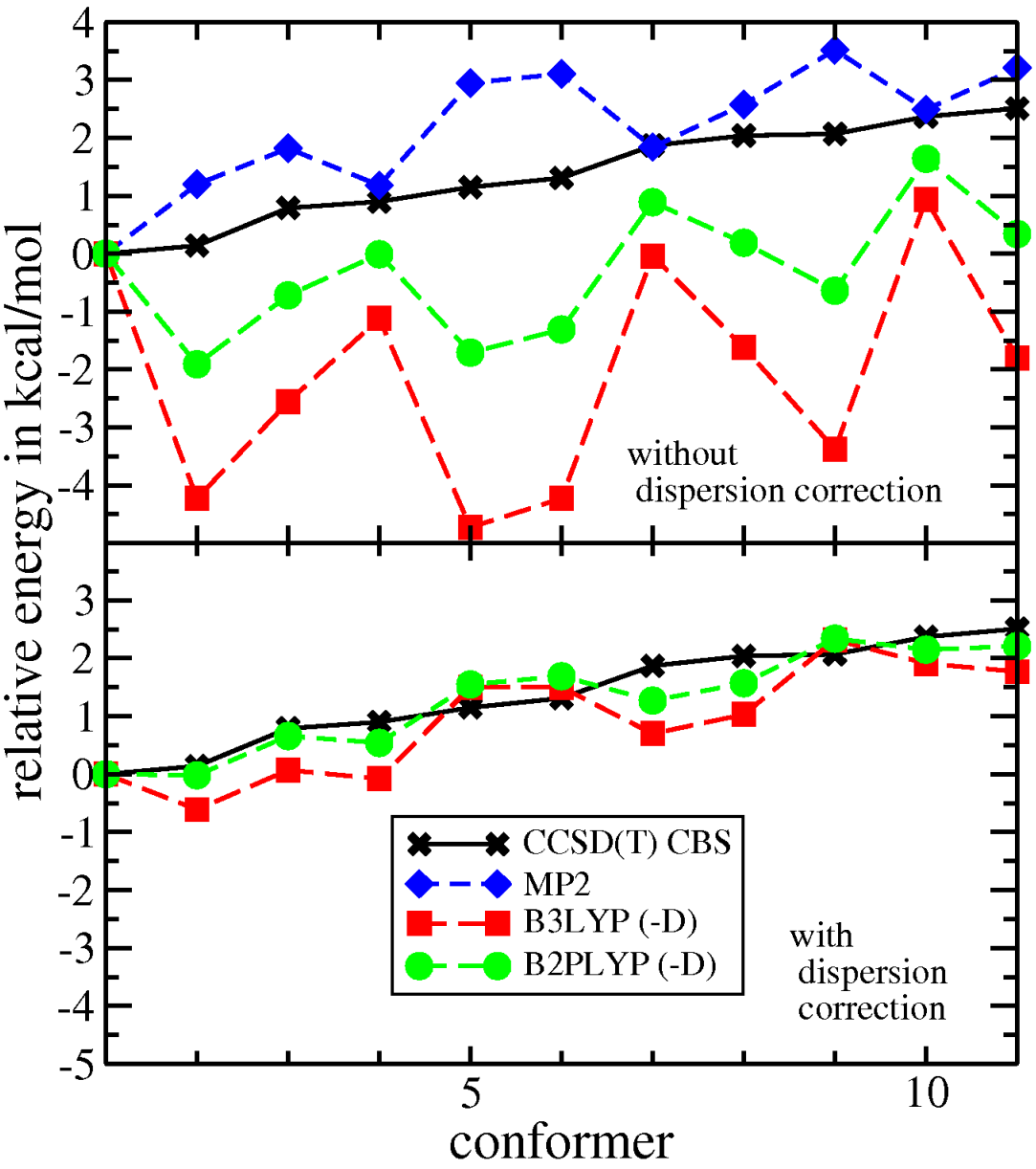

8/224


5/366

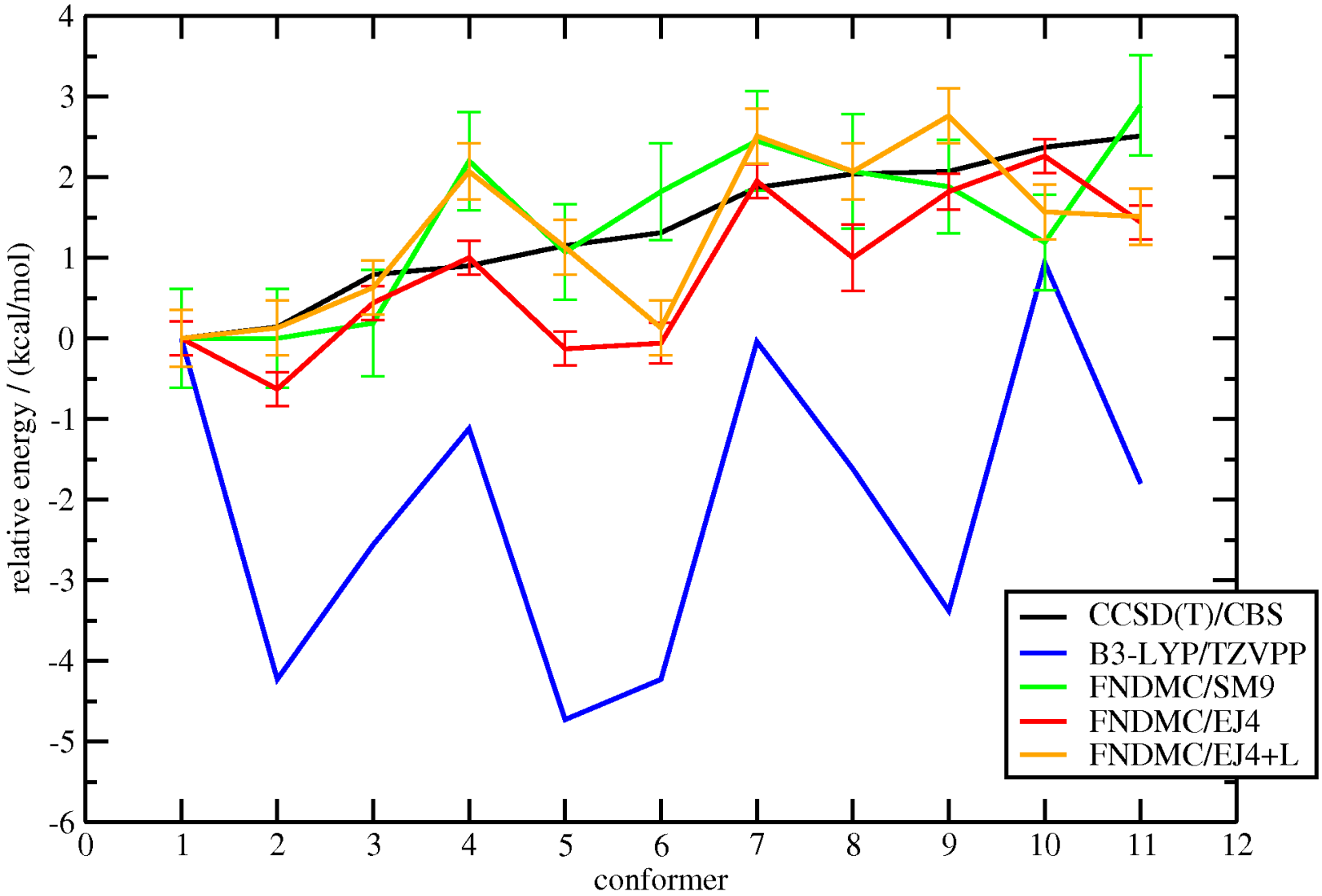

6/300


9/691

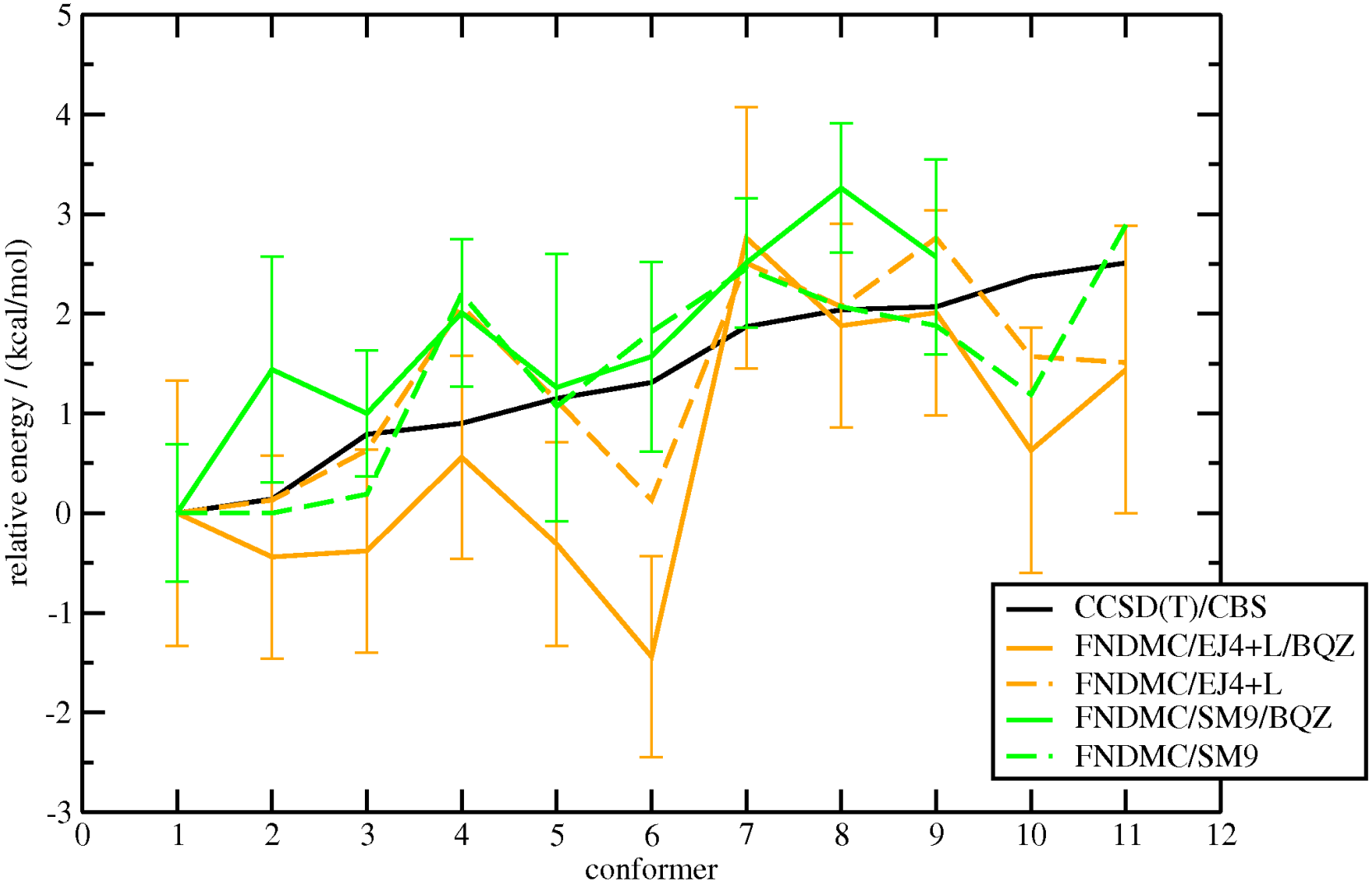
Tripeptide conformers / DFT methods



Tripeptide conformers / FNDMC



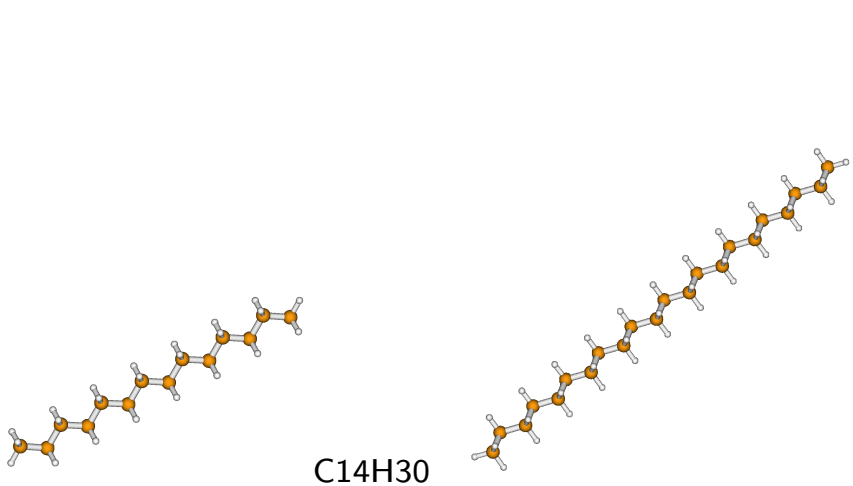
Tripeptide conformers / FNDMC 2 - work in progress



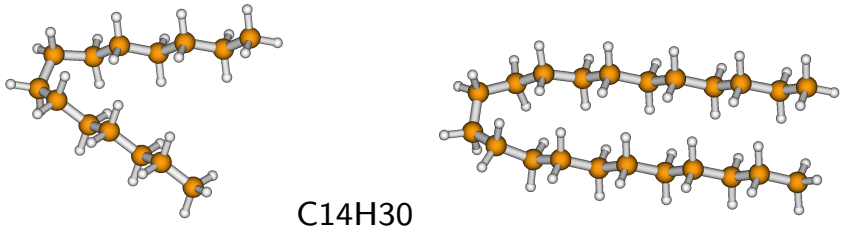
Alkane chain folding (part of the IDHC7 set)

Breaking down the interactions in the tripeptides ...

- ... into chain folding interactions and ...



C14H30 C22H46 C30H62

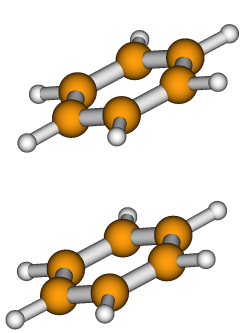


C14H30 C22H46 C30H62

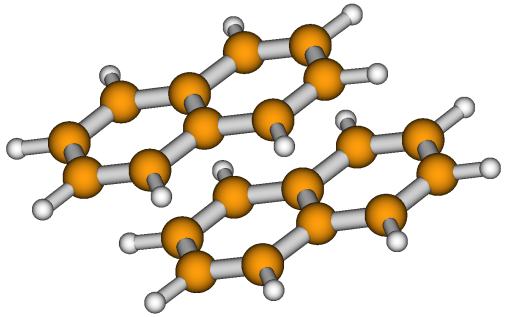
Non/saturated ring stacking

Breaking down the interactions in the tripeptides ...

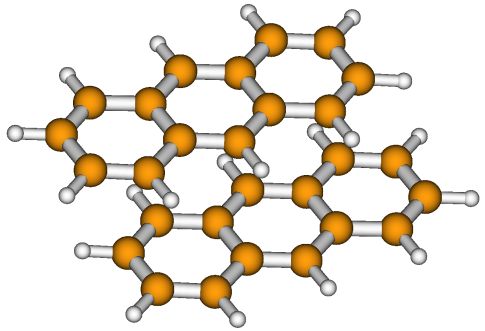
- ... into ring stacking interactions



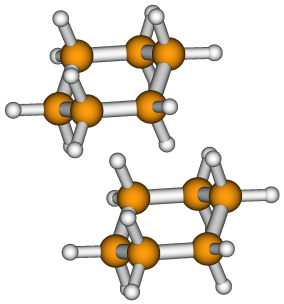
C6H6



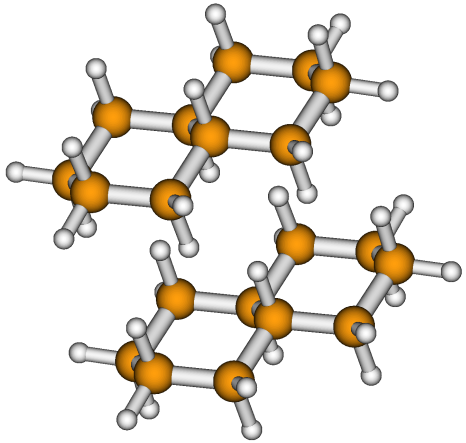
C10H10



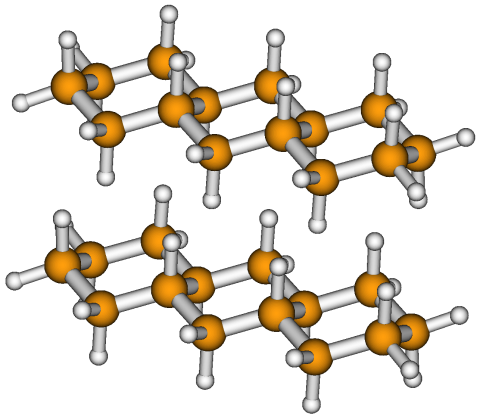
C14H14



C6H??



C22H??



C30H??

Summary

Biomacromolecule model systems

- good: comparable performance for larger and smaller systems
- interesting: long range Jastrow terms needed for dispersion (at given trial WF quality / time step level)
- great: basis sets and ECPs from Burkatzki *et al.*
  $\longrightarrow$ accuracy sufficient for the use as reference?

Acknowledgments

Cooperation partner

- Arne Lüchow

Grimme Workgroup

- especially Christian Diedrich, Christian Mück-Lichtenfeld and Stefan Grimme

Other Help

- screen-saver - Jan Budde
- web design - Martin Heinrich

Funding

- QMC@HOME is funded by the Sonderforschungsbereich 424 established by the Deutsche Forschungsgemeinschaft

Jastrow factor types

Schmidt-Moskowitz Jastrow factor

- $$U = U_{ij}(\hat{r}_{ij}) + U_{Ni}(\hat{r}_{Ni}) + U_{Nij}(\hat{r}_{ij}, \hat{r}_{Ni}, \hat{r}_{Nj}) \quad with \quad \hat{r}_{ab} = \frac{b * r_{ab}}{1 + b * r_{ab}}$$

Exponential Jastrow factor

- Short range
$$with \quad \hat{r}_{ab} = 1 - e^{-\alpha r_{ab}}$$
- Long range
$$with \quad \hat{r}_{ab} = e^{-\gamma / r_{ab}}$$

Jastrow factor types