The present and future role of computers in chemical research

Markus Reiher

Laboratorium für Physikalische Chemie, ETH Zurich, Switzerland

`http://www.reiher.ethz.ch`

June 24, 2011

Overview

- Aims of modeling in chemistry
- Methods, principles, and limitations of computer simulation in chemistry
- Current hardware requirements
- Two example applications from my group
- IT in the Chemistry Curriculum
- An attempt on a future perspective

Aims of chemical modeling I

- **Ultimate goal:**

  Understanding and (true) prediction of molecular phenomena

  ⟶ on an atomistic level (reductionist view)

- **Central quantity:**

  An energy function of the 'system' under study

  ⟶ reactions on potential energy hypersurfaces

- **Target quantities:**

  Structure, energetics, properties, function

  ⟶ all derivable from the energy function

Aims of chemical modeling II

Qualitative predictions can be done on the back of the envelope,

quantitative predictions require computer simulations!

Aims of chemical modeling III

- **True predictions are possible for:**
  - molecular structures
    (intermediates of reactions, novel species)
  - reaction mechanisms and dynamical behavior
    (mechanisms, reaction networks)
  - spectroscopic data of any kind
    (underlying structure needed as input though)
- **detailed (post factum) understanding:**
  interpretation and explanation of experimental data
- **In principle, perfect time and spatial resolution:**
  e.g., mode of action of a protein

Aims of chemical modeling IV

- **Prominent applications of computational chemistry:**
  - relativistic quantum chemical optimization of processes involving nuclear waste (e.g., Hanford Site/NWChem, BNFL/U Cambridge)
  - systematic screening for new catalysts instead of time consuming synthetical searches
    (e.g., cooperation between U Heidelberg and BASF)
  - computational pre-screening for molecular drug design
    (e.g., Novartis)
- **Challenging new problems:**
  - de novo enzyme design
  - inverse methods (from function/reactivity to structure)

Foundations — Theory I

- **Modeling mostly based on physical theories:**
  - *Electrodynamics*: classical fields
    (for, e.g., single-molecule spectroscopy at surfaces)
  - *Quantum mechanics*: description of electrons and nuclei
    (for, e.g., bond breaking, spectroscopy)
  - *Classical mechanics*: description of atoms and coarser structures
    (weakly interact. large host–guest complexes, large amplitude motions)
  - *Statistical mechanics*: ensembles at fixed parameters
    (e.g., constant particle number, volume and temperature: N,V,T)
  - *Thermodynamics*: Gibbs free energies
  - occasionally other physical theories like *continuum fluid dynamics*
    (e.g., simulation of microfluidic systems, flow in confined geometries)

Foundations — Theory II

- we have seen a complete 'physicalization' of chemistry within the past 60 years employing even the most elaborate theories like quantum electrodynamics
- all with a static and a dynamic variant, depending on whether a static state-by-state picture is sufficient or whether explicit detailed dynamics are mandatory
  (e.g., few deep holes vs. flat energy surfaces that describe the system)
- Few, expert-based, 'physics-free' methods
  - chemoinformatics: quantitative-structure activity/property relationships (QSAR, QSPR)
  - data-bank based approaches (e.g., NMR chemical shift predictions)

  ⇒ usually not computationally expensive (except for combinatorial reasons: 'millions' of single evaluations)

Foundations — Algorithms I

- *Ongoing development since 1950*

  (close connection between hardware development and chemical applications since the early years)

- *Problem-driven optimization of methods*:

  reinvention of numerical methods, but also breakthroughs

  (e.g., Davidson subspace iteration for diagonally dominant huge sparse matrices from 1975; mathematically analyzed, but not improved around 1990; diagonalization for lowest eigenvectors of a $10^{10} \times 10^{10}$ matrix)

Foundations — Algorithms II

- *Current status*:

  well established solution methods for standard problems;

  new algorithms for unsatisfactorily solved or new problems;

  algorithms optimized for new hardware architectures
  (still predominantly developed in computational chemistry groups)

Principles of Molecular Modeling I

- Need to choose an **appropriate structural model** that resembles the experimental situation!
  - Gas phase (isolated molecular systems) vs. condensed phase (nonnegligible interaction with solvent/solid environment)
  - local reactions (small molecular model) vs. molecular recognition processes (huge molecular model)
- **Structural model** determines the computational resources, as the methods scale with a measure of the model size (number of atoms, number of electrons, ...)

Principles of Molecular Modeling II

- Need to **choose** the **appropriate method** (i.e., physical theory and suitable approximations) for solving the scientific problem given (determined by accuracy required)

- **Accuracy requirements have a wide range:**

  from accurate quantitative results (example: calculating the bond energy of H_2 to 9 significant figures) to possible/impossible (yes/no) questions for the feasibility of certain reaction steps

  Hence, modeling approach with highest accuracy is seldomly needed!

  All one needs to ensure is whether accuracy is sufficient
  (tough enough).

Limitations I

- **Approximations to treat structural models of increasing size**
- **Approximations introduce errors that must be well controllable**
- No approximate method that was truly useful at some point ever vanished from the tool box of computational chemistry, e.g.
  - 'traditional' molecular dynamics will not be replaced by 'first-principles' molecular dynamics
  - semiempirical quantum chemical methods (MNDO, DFTB, ...) will not be replaced by ab initio methods (CASSCF, CCSD(T), ...)

... mostly because the size of the structural model is increased in the realm of any method and offers new possibilities

Limitations II

- **'Problem of the Particular'**:

  Slightly changed structural models would, in principle, require a whole new calculation/simulation which is not desirable; **transferability** (at least regarding the insights) **would be needed**, but possible?

  Remember: Chemistry is about the tiny changes in molecules that can have a huge effect on their function/reactivity; understanding the 'essential physics' is not the primary aim in chemistry

- Every method is applicable only for a specific length and time scale, apart from its accuracy issues

Current hardware requirements I

- Good news: Research in computational chemistry can be done with a desktop computer or laptop since 2000
- Consequence: Such calculations are nowadays done routinely by experimentalists
- **However, the requirement for computational resources is endless!**
- Increase size of structural model or accuracy of method always leads to increased computational requirements (pushes any computer resource to its limits)

Hardware used in our group: highly diverse

Current hardware requirements II

- standard high thru-put calculations like DFT;

  parallelizable on slim nodes with 2-4 GByte per core and ordinary disks; fast interconnect helps but we usually parallelize on the n cores of an n-core machine; available in our group: about 600 cores; we also have trivially parallelizable problems which can use all cores)

- electron correlation calculations

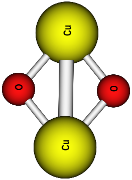
  (often not parallelized, single node with huge RAM ($>$128 GByte per job) and disk ($>$1 TByte per job); available to our group: about 100 such cores)

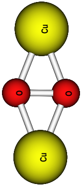
- ab initio molecular dynamics simulations:
  supercomputers of specialized centers (e.g., Manno)

Example A: Relative energy of a simple cluster

- calculation time: 2 months on 2 cores

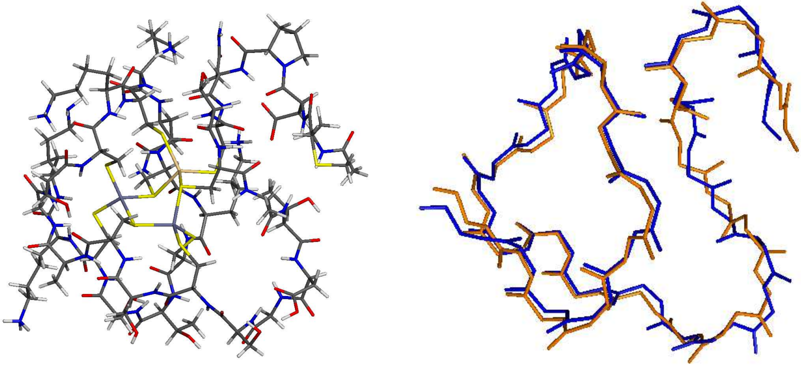
  K. Marti, M. Reiher, *Z. Phys. Chem.* *224* **2010** 583–599





m	E_{bisoxo}/E_h	E_{peroxo}/E_h	$\Delta E/\mathrm{kJmol}^{-1}$
56	−541.457 025 52	−541.487 645 94	80.39
84	−541.468 923 61	−541.504 124 70	92.42
156	−541.480 807 29	−541.518 407 11	98.72
CASSCF(16,14)	−541.467 083 19	−541.477 448 01	27.17
CASPT2(16,14)	−541.601 131 87	−541.598 756 46	−15.06
RASPT2(24,28)[a]			119.66

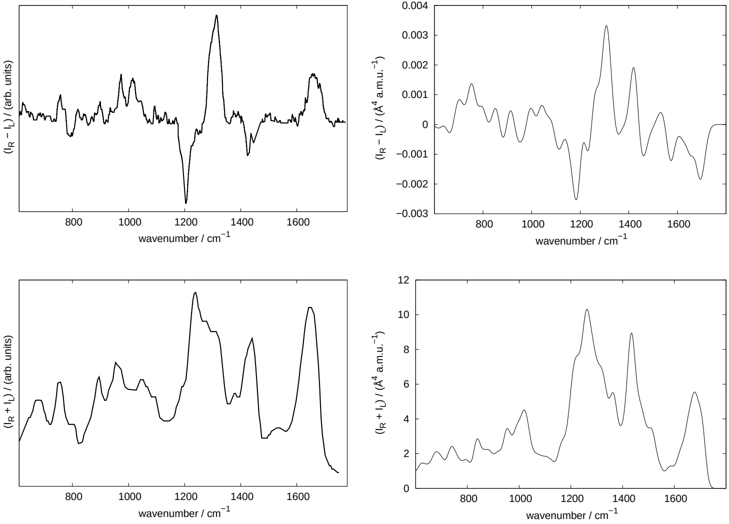
Example B: ROA spectrum of metallothionein



- >400 atoms, sulfur-rich, 2 Zn and 1 Cd atoms
- calculation time: ca. 6 months on 100 cores

S. Luber, MR, J. Phys. Chem. B **114** (2010), 1057.

ROA spectrum of β Domain of Rat Metallothionein



- Exp. (left) / Theory (right) ; ROA (top) / Raman (bottom)

Experiment: E. Smyth et al., Biopolymers **58** (2001) 138

IT in the Chemistry Curriculum at ETH

- Basic programming principles in C in 1st semester
- Algorithms as needed by Molecular Dynamics or Quantum Chemistry in advanced Bachelor or in Master courses
- This is a basis but certainly not sufficient for innovative research (no hardware-optimized programming like cache optimization, no true object-oriented programming, ...)

  ... but this can hardly be changed as the computational chemistry part of the curriculum is basically filled already with physical theories (many-particle quantum physics, statistical mechanics, Hamiltonian dynamics)
- **Truly successful PhD students in Theoretical Chemistry gained their expertise outside the chemistry curriculum**

Future perspective I

- **Predictions are difficult, of course ...**

  ⇒ This perspective is necessarily personal!
- Increasing computational power will increase our capabilities

  (more detailed and/or larger structural models, higher accuracy where needed)
- What will determine our future needs?
  - What is affordable and reasonable in terms of electricity and cooling
  - ... and what is affordable to our international competitors
- There is no unique demand for computational resources
- Massively parallel high performance computing with thin nodes will be most useful in the field of (first-principles) molecular dynamics

Future perspective II

- Moderately parallel fast fat nodes for highly accurate quantum chemical calculations
- Can we expect algorithmic and other theoretical improvements?

  Yes!

  Will that change the hardware requirements?

  No!

  No one would run a faster or otherwise optimized algorithm on a less powerful computer, e.g., to safe energy, if more powerful hardware is available

Future perspective III

- **Personal view: Because of the high maintenance cost (electricity, cooling) the community needs to agree on what type of problem is worth the tremendous effort in terms of resources and what type of simulation is just muscle play without significant scientific insight**
- We need a new 'paradigm' in computational chemistry:
  it is not about
  '*what is the largest structural model you can treat*'
  but more about
  '*what is the most intelligent modeling approach to answer a chemical question with as little resources as necessary to reach the accuracy determined by the problem*'

Future perspective IV

- We will need a balanced, reasonable, heterogeneous, state of the art equipment of computer hardware (currently to be renewed every ca. 5 years) at every level:

  clusters in research groups,

  centralized cluster at universities,

  national computing center